Progressive Modality-Alignment for Unsupervised Heterogeneous Change Detection

Yinghui Xing, *Member, IEEE*, Qi Zhang, Lingyan Ran, Xiuwei Zhang, Hanlin Yin, Yanning Zhang, *Senior Member, IEEE*

***Abstract*—Change detection based on heterogeneous images is of great importance in some applications, such as disaster monitoring and damage assessment. However, due to the huge modality discrepancy in heterogeneous images, it is difficult to accurately detect the changed regions. In this paper, we analyze the interference of modality-alignment and changed areas to each other, and propose a progressive modality-alignment based unsupervised change detection model for heterogeneous images. Specifically, the modality alignment is achieved in an iterative manner, which can improve the detection accuracy progressively. To reduce the influence of modality discrepancy and the changed regions to each other, a pseudo-label self-learning strategy is designed, where the pseudo-labels learned by the model itself are used to act as a guidance of change detection, and they are in turn refined by the proposed progressive model. Experimental results on different real heterogeneous images verify the effectiveness and robustness of proposed method.**

***Index Terms*—Heterogeneous images, change detection, auto-encoder, pseudo-label, progressive.**

I. Introduction

CHANGE detection refers to identifying changes by analyzing images acquired over the same geographical location but at different times [1]. It has been widely used in damage assessment [2], [3], land-cover monitoring [4] and environmental investigation [5].

With the development of remote sensing technology, plenty of remote sensing data have been available for earth observation, which boosts the progress of change detection. These remote sensing data can be obtained from various types of sensors [6], such as optical, Synthetic Aperture Radar (SAR), and Light Detection and Ranging (LiDAR) *etc*. In general, detecting changes between images from different types of sensors (heterogeneous) is more difficult than that from the same sensor (homogeneous), because heterogeneous change detection should consider not only the spatial interference factors, like noise, weather and illumination, but also the modality discrepancy. In some time-bounded scenarios, such as fast response for disaster management, one can not obtain images from the same sensor [7], thus heterogeneous change detection attracts more attention in recent years. However, in heterogeneous case, the bi-temporal images to be compared exhibit large intensity and geometric differences, making the traditional direct comparison infeasible [8]. Therefore, some specific designs should be taken into consideration to realize the change detection of heterogeneous images.

Due to the huge modality discrepancy, some methods firstly classify multi-temporal images and then detect changes based on the classification results [9]–[11]. They circumvent the modality alignment operation and are easy to be implemented, but the detection results heavily depend on the performance of classifiers.

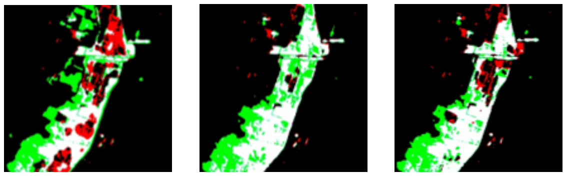
Another straightforward manner is to distinguish the changed and unchanged regions [12] by extracting modality-invariant features [13], [14]. Nevertheless, it is nontrivial to extract modality-invariant features since the image contents are strongly affected by the imaging conditions, which brings difficulties in decoupling of modality-dependent and modality-independent features, especially for SAR images with speckle noise.

Considering the semantic similarity between bi-temporal images, some other heterogeneous CD methods eliminate the modality differences first and then calculate the change map. They either transform the image from one domain to the other [13], [15]–[17], or take bi-temporal images into a common latent feature space to align them [18]–[20]. The former tries to align heterogeneous images in the image domain, and then utilizes homogeneous change detection models to obtain change maps [13], [15]–[17]. But the imaging mechanism of sensors are quite different, transforming one image to the other domain inevitably introduce some noises, leading to suboptimal results. The latter assumes that bi-temporal images can be transformed into a common latent space learned from unchanged regions. In order to obtain such a feature space, a plenty of training samples are required to guide the training process. However, the data annotation is inherently cumbersome, especially for heterogeneous images. Hence, unsupervised models have been developed and become popular in heterogeneous image change detection [13], [16], [21], [22].

In general, unsupervised heterogeneous images change detection faces two main challenges: heterogeneity and label-scarcity. Since the differences between the pre-event and post-event images come from not only the changed regions, but

This work was supported in part by the National Natural Science Foundation of China (NFSC) under Grant U19B2037 and 62201467; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110544; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2022JQ-686; in part by the Project funded by China Postdoctoral Science Foundation under Grant 2022TQ0260, and in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313088. *(Corresponding author: Lingyan Ran.)*

Yinghui Xing, Qi Zhang, Lingyan Ran, Xiuwei Zhang, Hanlin Yin and Yanning Zhang are with the Shaanxi Provincial Key Laboratory of Speech and Image Information Processing, and the National Engineering Laboratory for Integrated Aerospace-GroundOcean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China. Yinghui Xing is also with the Research Development Institute of Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China. (e-mail: xyh_7491@nwpu.edu.cn).



(a) (b) (c)

Fig. 1. Confusion map of (a) initialization, (b) first iteration and (c) last iteration of our method.

also modality discrepancy, the models should distinguish real changed regions from the spatial interferences on the condition of label-scarity and huge modality difference.

For the unavailability of labels, most of current unsupervised methods iteratively update the pseudo-labels to detect changed regions, therefore, it is crucial to obtain high-quality initial change maps to facilitate the selection of pseudo-labels. Some of them generate the initial change map through random initialization [18], [23], [24], failing to provide a proper guidance for feature alignment. Zhang *et al*. [25] used OBCD [26], a method for homogeneous CD task, to obtain the initial change map. Li *et al*. [27] utilized an extra three-layer network to obtain the initial change information. The selection of pseudo labels strongly depends on the quality of initial change map, while their initialization disregarded the modality discrepancy, thus limiting their performance.

In this paper, we propose a progressive modality-alignment method, which aligns heterogeneous image features and refines pseudo-labels alternatively. Our model reduces the interference of changed regions to modality alignment and also diminishes the impact of modality alignment to change detection result. Specifically, we use the model itself to generate pseudo-labels without introducing additional parameters. Then, a pseudo-label self-learning strategy is designed to guide the training process and at the same time further refine the pseudo-labels themselves. Finally, the pseudo-label generation and pseudo-label self-learning are achieved in a progressive manner. Since the model has the ability of producing pseudo-labels and then updating them, we can take the results of the last iteration as the final change map without any post-processing. As can be seen in Fig. 1, the change maps are effectively refined through our model.

The main contributions of this paper are summarized as follows:

1) We propose a progressive refinement framework for unsupervised heterogeneous change detection, where the subtle changed details can be detected in an alternatively progressive manner.
2) A pseudo-label self-learning strategy is designed. The pseudo-labels generated during the modality-alignment process are used to guide the detection of changed regions, and the detection results also refine the pseudo-labels in turn.
3) We comprehensively compared our model with several state-of-the-art unsupervised heterogeneous change detection methods, including FPMS [28], NACCL [29], INLPG [30], IRG-Mcs [12], SCCN [18], cGAN [31] and CAAE [16]. Our proposed one was verified to be effective and advanced in experiments by achieving a higher overall accuracy (OA), Area Under the ROC Curve (AUC), and Kappa coefficient than these counterparts.

The rest of the paper is organized as follows. Section II provides related works. Section III introduces our proposed method in detail. Experimental results and corresponding analysis are demonstrated in Section IV. Section V concludes this paper.

II. Related Works

Over the last two decades, many change detection (CD) methods have been proposed. In this section, we first briefly review the unsupervised change detection methods, and then introduce the related heterogeneous change detection models.

A. Unsupervised Change Detection

There are a number of supervised CD methods [32]–[34], which rely on a large amount of labeled data. However, due to the fact that the annotation of labels is cumbersome, unsupervised change detection methods are promising.

Bruzzone *et al*. [35] has summarized that unsupervised change detection mainly comprises difference image (DI) generation and change analysis. Following this paradigm, Celik *et al*. [36] used PCA to map local neighborhoods in differential images to high-dimensional space defined using non-overlapping image blocks. Then the K-means algorithm is used to automatically separate the changed regions from the whole regions. Hao *et al*. [37] proposed an unsupervised CD method based on the level set method of expectation maximization. Li *et al*. [38] combined the fuzzy C-means algorithm with the nearest neighbor rule to classify Gabor features extracted from SAR images. Above methods all use traditional feature extraction manners and the detection accuracy is not always satisfying. Saha *et al*. [39] extracted bitemporal features using an untrained model and further compared the features using deep change vector analysis to distinguish changed pixels. Bergamasco *et al*. [40] proposed a convolutional autoencoder (CAE) [41] to detect changes in SAR images. By using a feature selection method based on variance, this strategy selects the features with the most change information to generate the difference map.

B. Heterogeneous Change Detection

Owing to the large modality discrepancy, the annotation of changed regions in heterogeneous images is difficult. Therefore, most of heterogeneous CD methods are unsupervised.

Some heterogeneous CD methods are based on image statistical characteristics, and utilize the image regression model to detect changes [13], [15], [42]. Mercier *et al*. [42] modeled the dependence of data statistics of bi-temporal images using quantile regression. It yielded an estimation of the local statistics, which were compared with that of the groundtruth by KL divergence. HPT [15] is a kernel regression method. It selected the K-nearest neighbors of target pixels from a small

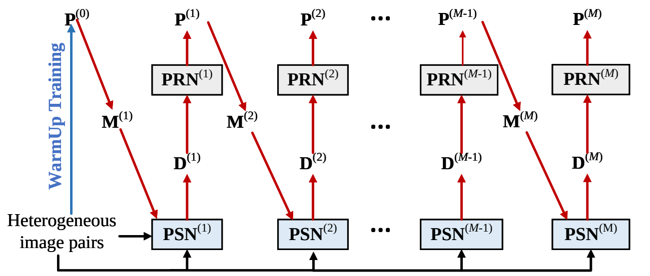


Fig. 2. Overall framework, where $\mathbf{P}^{(\mathbf{i})}$, $\mathbf{M}^{(\mathbf{i})}$, and $\mathbf{D}^{(\mathbf{i})}$ are the pseudo-labels, mask, and the difference map in i-th iteration, respectively. PSN and PRN are the abbreviations of Pseudo Siamese Network and the Pseudo-label Refinement Network.

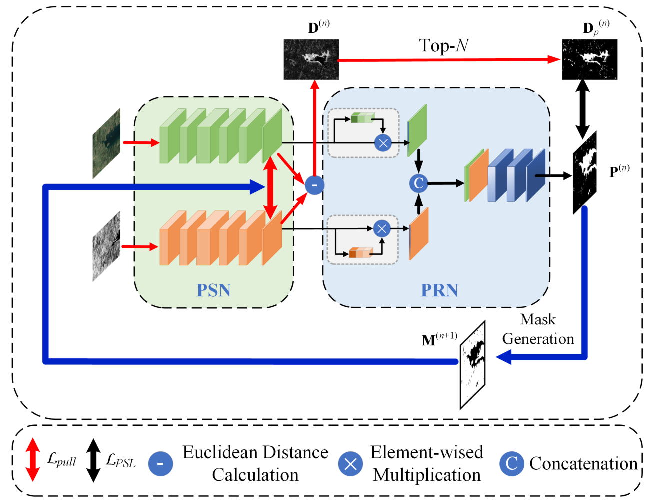


Fig. 3. Details of the n-th iteration, where the red and black arrows indicate the pseudo-label generation process and the pseudo-label self-learning process, respectively, and the blue arrows denote the mask generation and transfer to the $(n+1)$-th iteration.

number of unchanged samples to obtain the predicted value of target pixels through weighted addition. Luppino *et al.* [13] proposed an unsupervised method based on affinity matrix and image regression. The method identified the pixels that may be invariant by the similarity of affinity matrix, then used them to learn the mapping relationship between bi-temporal images.

There are some methods that based on the similarity metric [12], [18], [24], [30], [43], [44]. SCCN [18] is a coupled symmetric network, who used two stack DAEs [43] to extract features and constrained them to the same space. Yang *et al.* [24] utilized a method similar to SCCN [18], but they integrated multi-scale strategy into feature extraction model to obtain more accurate detection results. NPSG [44], and its modified versions [12], [30] constructed a graph based on the nonlocal patch similarity, and then defined the degree of change by measuring the consistency of the graph structure.

Benefiting from the development of generative adversarial networks (GANs), some researchers detect changes on the basis of image translation or style transfer methods. To the best of our knowledge, cGAN [31] is the first attempt. The authors used conditional GAN [45] model to transform an image to another style, and then utilized an approximation network to narrow the gap of image domains. Liu *et al.* [46] realized the image translation through cycleGAN [47], and selected the partially changed and unchanged pixels to form the training set to train a random forest classifier. While Jiang *et al.* [48] directly used VGGNet to extract image features, and computed Gram matrix to represent image style. Achieving change detection by retaining content information and transforming style information in the hidden space.

Similar to our proposed one, Zhan *et al.* [49] designed an iterative method to map heterogeneous images to the same space, where the direct comparison can be conducted. However, the same structure is used to generate and reuse the change map, and such a framework depends heavily on the results of the first mapping. Our method adds a pseudo-label self-learning strategy to alternately detect changes and refine pseudo-labels, which makes up for the defects of Zhan's method [49].

III. Methodology

In this section, we first formulate the heterogeneous change detection task, and then provide an overview of our proposed approach, followed by a detailed introduction of pseudo-label generation, pseudo-label self-learning and progressive refinement. The framework is shown in Fig. 2.

A. Problem Formulation

Change detection task takes two coregistered images $\mathbf{X} \in \mathbb{R}^{m\times n\times b}$ and $\mathbf{Y} \in \mathbb{R}^{m\times n\times b}$ as inputs to generate a change map $\mathbf{C} \in \mathbb{R}^{m\times n}$. $\mathbf{C}$ is a binary matrix, whose elements indicate the changed areas. In heterogeneous change detection task, the Pre-event image $\mathbf{X}$ and the Post-event image $\mathbf{Y}$ come from different type of sensors. Owing to the modality discrepancy, some recent models first map two images into the same hidden space $\mathcal{Z}$, and then obtain the change map by some change analysis methods $A(\cdot,\cdot)$.

$$\mathbf{C} = A(\mathbf{Z}_X, \mathbf{Z}_Y), \tag{1}$$

where $\mathbf{Z}_X$ and $\mathbf{Z}_Y$ are the mapped features of $\mathbf{X}$ and $\mathbf{Y}$ in the same hidden space $\mathcal{Z}$.

B. Overview

On the whole, our model is built upon an alternative optimization framework. It contains a warm-up training stage and a progressive training stage. The aim of warm-up training is to obtain a prior change map $\mathbf{P}^{(0)}$ to guide the progressive training. In the warm-up training, the Pseudo Siamese Network (PSN) takes bi-temporal images as inputs to extract features in a common space, which is trained a few epochs with a warm-up loss $\mathcal{L}_{warm}$. The progressive training stage consists of a pseudo-label generation process and a pseudo-label self-learning process, which are learned alternatively. Taking the n-th iteration as an example, as illustrated in Fig. 3, in the pseudo-label generation process, we first generate the mask $\mathbf{M}^{(n)}$ from $\mathbf{P}^{(n-1)}$, and then calculate the masked loss term $\mathcal{L}_{pull}$ to align the features extracted by PSN. Combined with the reconstruction loss term, PSN is trained to extract aligned and informative features. To refine the pseudo-labels, we obtain the pseudo-label map $\mathbf{D}_p^{(n)}$ from the difference map $\mathbf{D}^{(n)}$ produced by PSN. Then the predictions of current iteration are obtained through the Pseudo-label Refinement Network (PRN). The pseudo-label map and the predictions are used to train PRN. Because the calculation of both pseudo-labels and the predictions are achieved by the model itself, we call this process as pseudo-label self-learning process. Finally, the predictions of n-th iteration are used to generate the mask $\mathbf{M}^{(n+1)}$ for the next iteration. After several alternative iterations, we take the predictions of the last iteration as our final detection results.

Note that although the training of our model is composed of two stages, the inference is achieved directly through the cascade of PSN and PRN.

C. Warm-Up Training Stage

Since our model detect changed regions in a progressive manner, an initial change map should be taken as starting point. The simplest method is to utilize the random initialization or the identity initialization. However, we have conducted thorough experiments and found that such initializations would result in unstable performance on different datasets. Actually, a rough change map produced by available change detection methods can be treated as initial change map to guide the progressive training. Therefore, in our paper, we use the Pseudo Siamese Network (PSN) itself, whose branches have the same structure but different parameters, to extract image features. The extracted features are represented by

$$\begin{aligned} \mathbf{Z}_X &= F_e(\mathbf{X}), \\ \mathbf{Z}_Y &= G_e(\mathbf{Y}), \end{aligned} \tag{2}$$

where $F_e(\cdot)$ and $G_e(\cdot)$ denote the encoders of the Pre-event and Post-event images, and they can transform the heterogenous images to the same latent space $\mathcal{Z}$.

The warm-up loss $\mathcal{L}_{warm}$ is then defined as:

$$\mathcal{L}_{warm} = \frac{1}{mn}\sum_{i,j} \|\mathbf{Z}_X(i,j) - \mathbf{Z}_Y(i,j)\|_1, \tag{3}$$

where m and n denote the height and width of image feature. After warm-up training, the initial convergence of the model produces relatively reliable predictions, which can be used as priors to further improve change detection. We then use the classical threshold method OTSU [50] to obtain an initial change map $\mathbf{P}^{(0)}$ by $\mathbf{P}^{(0)} = OTSU(\mathbf{D}^{(0)})$, where $\mathbf{D}^{(0)}$ denotes the difference map in warm up training. Experimental results show that such a warm-up training not only helps to obtain an initial change map, but also enables the stable training of our model.

D. Progressive Training Stage

The progressive training stage contains a pseudo-label generation process and a pseudo-label self-learning process.

1) Pseudo-Label Generation: Because heterogeneous images have huge distribution difference *e.g.* optical and SAR images, we should align heterogeneous image features to well locate the changed regions. In general, we can minimize their difference in feature space. However, the differences come from not only the modality discrepancy, but also the changed regions. Therefore, we use a masked loss term defined as follows to reduce the interference of changed pixels to the alignment:

$$\mathcal{L}_{pull} = \frac{1}{mn}\sum_{i,j} \mathbf{M}^{(n)}(i,j)\|\mathbf{Z}_X(i,j) - \mathbf{Z}_Y(i,j)\|_1, \tag{4}$$

where $\mathbf{M}^{(n)} = 1 - \mathbf{P}^{(n-1)}$ is the calculated mask for the n-th iteration, which is also detailed later. Ideally, $\mathbf{M}^{(n)}(i,j) = 1$ if the position (i,j) belongs to unchanged regions, and vice verse. Such an mask enables the encoders to learn mapping function without the interference of changed regions.

At the same time, in order to make the encoders learn the most informative features, we also use a reconstruction loss term $\mathcal{L}_{rec}$ for two encoders with the assistance of a pair of decoders,

$$\mathcal{L}_{rec} = \|\mathbf{X} - F_d(\mathbf{Z}_X)\|_1 + \|\mathbf{Y} - G_d(\mathbf{Z}_Y)\|_1, \tag{5}$$

where $F_d(\cdot)$ and $G_d(\cdot)$ are the decoders corresponding to encoders $F_e(\cdot)$ and $G_e(\cdot)$.

The loss function for PSN is

$$\mathcal{L}_{PSN} = \mathcal{L}_{pull} + \lambda\mathcal{L}_{rec}, \tag{6}$$

where λ denotes the balance weight of two loss terms.

After several training epochs, the heterogeneous images are mapped to the same feature space, therefore, we can obtain the difference images by calculating the Euclidean distance between these features

$$\mathbf{D}(i,j) = \|\mathbf{Z}_X(i,j) - \mathbf{Z}_Y(i,j)\|_2, \quad D(i,j) \in [0,1]. \tag{7}$$

Theoretically, changed regions have larger distance than unchanged regions if the features are strictly aligned. However, due to the heterogeneity of images, learning such a well aligned latent feature space is not an easy task. Though we can use classical clustering or threshold methods to obtain pseudo-labels, it strongly depends on the alignment ability of encoders. As a result, we adopt an intuitive manner that only select Top-N percent of pixels as changed pixels and others as unchanged to obtain a reliable pseudo-label map. According to the setting of N, we can calculate a threshold α by

$$\alpha = percentile(\mathbf{D}, N), \tag{8}$$

where $percentile(\mathbf{D}, N)$ denotes the calculation of N-th percentile in $\mathbf{D}$. Then the pseudo-label map $\mathbf{D}_p$ is obtained through

$$\mathbf{D}_p(i,j) = \begin{cases} 1, & \mathbf{D}(i,j) > \alpha \\ 0, & otherwise \end{cases}. \tag{9}$$

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3294300

2) Pseudo-label Self-Learning: Starting from the difference image, we have obtained the pseudo-label map. Then, the pseudo-labels generated by the model itself are used to guide the learning of change maps, and inversely refined themselves. Since the features are obtained by different encoders, they still contain modality-specific characteristics. To reduce the heterogeneity, channel attention module is applied before the feature fusion to adaptively emphasize important features while suppressing irrelevant features across the channel. As shown in Fig. 3, We first utilize channel attention [51] to two feature representations $\mathbf{Z}_X$ and $\mathbf{Z}_Y$. After that, the attentive features are concatenated and then fused by several convolutional layers to obtain a prediction map $\hat{\mathbf{p}}$, this process can be expressed as

$$\hat{\mathbf{p}} = f\big(concat(attn(\mathbf{Z}_X), attn(\mathbf{Z}_Y))\big), \tag{10}$$

where $attn(\cdot)$ denotes channel attention operation [51], and $concat(\cdot,\cdot)$ represents the concatenation of features along channel dimension. $f(\cdot)$ is the Pseudo-label Refinement Network (PRN). We use pseudo-labels generated by the network itself to guide the learning process. Here the cross-entropy loss together with a weight decay term is used to achieve the pseudo-label self-learning,

$$\mathcal{L}_{PRN} = CE(\mathbf{D}_p, \hat{\mathbf{p}}) + \frac{1}{2}\beta\|\mathbf{W}_f\|_2^2, \tag{11}$$

where $\mathbf{W}_f$ denotes the weights of f, and β is the balance parameter. Since the change detection can be treated as binary classification task, the predictions $\hat{\mathbf{p}}$ has two channels, representing the probabilities of change and unchange. We only take the channel representing changed probability to obtain change map. In order to make the threshold p_t general and applicable to different datasets, we first use min-max normalization on changed probability map to obtain $\hat{\mathbf{p}}^*$, and the change map is obtained through

$$\mathbf{P}^{(n)}(i,j) = \begin{cases} 1, & \hat{\mathbf{p}}^*(i,j) > p_t \\ 0, & otherwise \end{cases}. \tag{12}$$

where p_t is a manually assigned hyper-parameter.

3) Progressive Refinement: Although above operations can produce a change detection result, it is trained with the assistance of mask, which strongly relies on the quality of pseudo-labels, and the model can be failed when the noise ratio of pseudo-labels is high. Therefore, we propose a progressive refinement strategy to improve the quality of pseudo-labels and also refine the change map in an iterative manner.

We can first obtain the prior change map $\mathbf{P}^{(0)}$ after the warm-up training. The initial mask is computed by $\mathbf{M}^{(1)} = 1 - \mathbf{P}^{(0)}$, then the mask is taken into PSN to obtain difference map $\mathbf{D}^{(1)}$, from which, we select the reliable pseudo-labels to guide the learning of PRN. The output of PRN is a change map $\mathbf{P}^{(1)}$ of the first iteration, and the mask is then updated through $\mathbf{M}^{(2)} = 1 - \mathbf{P}^{(1)}$ for the second iteration. The change map is refined progressively with the update of pseudo-labels, guiding to a more accurate change map.

The whole algorithm is shown in Algorithm 1. The predictions of the last iteration are taken as final detection results.

Algorithm 1 Procedure of proposed model.

Input: Heterogeneous image pairs $\{\mathbf{X}_j, \mathbf{Y}_j\}_{j=1}^N$.
Output: Binary change map $\mathbf{P}$.

Warm-up training:
Train Pseudo Siamese Network (PSN) a few epochs with $\mathcal{L}_{warm}$ to obtain the prior change map $\mathbf{P}^{(0)}$;
Progressive training:
for $i = 1$ to M **do**
  Pseudo-label Generation:
  Obtain the mask by $\mathbf{M}^{(i)} = 1 - \mathbf{P}^{(i-1)}$;
  Train the PSN with $\mathcal{L}_{PSN}$ to obtain difference map $\mathbf{D}^{(i)}$;
  Obtain reliable pseudo-label map $\mathbf{D}_p^{(i)}$ from difference map $\mathbf{D}^{(i)}$ with Top-N method for the training of PRN.
  Pseudo-label self-learning:
  Train Pseudo-label Refinement Network (PRN) by $\mathcal{L}_{PRN}$.
  Obtain the change map $\mathbf{P}^{(i)}$ by Eq. (12).
end for
return $\mathbf{P} = \mathbf{P}^{(M)}$.

IV. Experiments

A. Datasets

We evaluate the performance of proposed method on five datasets, including Italy, Yellow River, Shuguang, Texas, and California datasets. The bi-temporal images and the corresponding groundtruth are shown in Figs. 4-7.

Italy Dataset. Italy dataset consists of a near-infrared image and an RGB optical image, both of which was taken in Sardinia, Italy, where a lake flooding event occurred. Fig. 4(a) is Pre-event image, near-infrared band of the Landsat-5 TM acquired in September 1995, and Fig. 4(b) is the Post-event optical image acquired in July 1996 from Google Earth. The image size is 412×300, and the spatial resolution is 30m.

Yellow River Dataset. As Fig. 5 shows, Yellow River dataset consists of a SAR image acquired by Radarsat-2 in June, 2008, and an optical image acquired from Google Earth in September, 2010. It is captured in Yellow River, China. The spatial resolution of them is 8m, and their size is 291×444.

Texas Dataset. Fig. 6 shows the false color image of Texas dataset. It is captured in Texas, America, and a forest fire occurred in Fig. 6(b). Both of Pre-event and Post-event images are multispectral images caputured through differet sensors, where the Pre-event image is acquired by Landsat-5 TM in August 2011, and the Post-event image is acquired by EO-1 ALI in September 2011, with 7 and 10 channels respectively. Their spatial resolution is 30m, and the spatial size is 808×1534.

California Dataset. California dataset is a multispectral/SAR image pair. As Fig. 7 shows, a blood happened within the time period. The Pre-event multispectral image with 11 bands is acquired by Landsat-8 in January 2017, and the Post-event SAR image is acquired by Sentinel-1A in February 2017. The size of them is 2000×3500.

Shuguang Dataset. It is also a SAR/optical image pair, captured in Shuguang village, DongYing city of China. As Fig. 8 shows, parts of farmland in Fig. 8(a) was changed to buildings in Fig. 8(b). The Pre-event image is the SAR image acquired in June, 2008, and the Post-event image is optical acquired in September 2012. The spatial size of them is 921×594, and the spatial resolution is 8m.

B. Experimental Settings and Evaluation Metrics

1) Details of Hyperparameter settings: In PSN, the two auto-encoders have exactly the same structure and each layer is a cascade of "*Conv-BN-LeakyReLU*" except for the last layer, which uses the "*Conv-BN-Sigmoid*" structure. The kernel size of all convolutional layers is 3x3. We set the number of channels of the last layer as 5. In PRN, each layer, except for the last one, is a cascade of "*Conv-BN-ReLU*" structure, and the activation of the last layer is "*Sigmoid*".

Through experiments, we found that the parameter N used to generate pseudo-labels in Eq. (9) is not sensitive to model performance, and then it is set to 0.08. The threshold p_t in Eq. (12) is set to 0.95, and the balance weights λ and β in the loss function are set to 2 and 1, respectively. It should be noted that the model architecture and the hyperparameters of all datasets is consistent, and the sensitivity of hyperparameters will be discussed in Section IV-E.

2) Training Details: Images are sliced to overlapped patches to form a training set, and the batch size is set to 32. Data augmentations, such as flipping, scaling, rotation, adding noise, etc. is used in training except for the warm-up training. We use SGD as our optimizer, and the learning rate is initialized to 1e-4. Our model is implemented by Pytorch with an Intel Core i7-7700 CPU at 3.6 GHz and NVIDIA GTX1080Ti GPU with 11GB memory and 32GB RAM.

3) Evaluation Criteria: In the experiments, we use *OA* (Overall Accuracy), κ (kappa coefficient) and *AUC* (Area Under the ROC Curve) curve to evaluate the performance of different methods. The first two metrics are used to comprehensively evaluate the final binary change map, and the last one is to evaluate the difference map.

In confusion matrix, *TP* (True Positive) denotes the number of samples that are positive and are also detected as positive, and *TN* (True Negative) denotes those negative samples that are detected as negative. Both *TP* and *TN* are correct detections. On the contrary, *FP* (False Positive) are negative samples that are detected as positive, and *FN* (False Negative) are those positive samples detected as negative. *FP* and *FN* represent wrong detections. Then the *OA* and kappa coefficient κ are defined as:

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

$$PRE = \frac{(TP+FP)(TP+FN)+(FN+TN)(FP+TN)}{(TP+TN+FP+FN)^2}$$

$$\kappa = \frac{OA-PRE}{1-PRE}.$$

Besides, false alarm (FA) and miss alarms (MA) are used to assist in quantifying detection results:

$$FA = \frac{FP}{FP+TN},$$

$$MA = \frac{FN}{TP+FN}.$$

C. Comparisons with State-of-the-Art Methods

Our proposed method is compared with seven unsupervised methods, *i.e.* INLPG [30], IRG-Mcs [12], FPMS [28], NACCL [29], SCCN [18], cGAN [31], and CAAE [16]. We reproduce them by the source code provided from the authors. In the following, we first introduce the comparison methods, and then provide the results on above five datasets. Note that in Figs. 4-7, the pixels with white color stand for the *TP*, black, green and red are for *TN*, *FP*, and *FN*, respectively.

INLPG [30] is a structure consistency based method, which detects changes by comparing the structure features of two images, rather than simply comparing the pixel values. IRG-Mcs [12] constructs a robust K-nearest neighbor graph to represent the structure of each image, and detects the changes through a Markovian co-segmentation model by comparing the constructed graphs within the same image domain. FPMS [28] utilizes a new parametric mapping strategy based on the modified geometric fractal decomposition and a contractive mapping approach to project two images to the same modality, then binarized the difference map under the unsupervised Bayesian framework. NACCL [29] is a Bayesian statistical approach, which relies on spatially adaptive class conditional likelihoods to be adaptive to the considered heterogeneous image pairs. The change map is then obtained based on the model for each pixel and each image modality. SCCN [18] constructs a deep convolutional coupling network to project heterogeneous images to the same feature space, and then uses threshold method to get change map. cGAN [31] make use of a conditional GAN to translate Pre-event image to Post-event one, and then utilizes an approximate network to further narrow the gap between feature representations. CAAE [16] takes a variety of constraints into consideration to realize image translation under the auto-encoders framework. Compared with GAN-based methods, CAAE [16] is more portable and easy to train.

1) Results on Italy Dataset: As Figs. 4(a)-(b) show, there are many mountain textures in this dataset, and the mountain shadows of the bi-temporal images are not consistent, which greatly hampers the detection of regions with real changes. The final confusion maps of different change detection methods are shown in Fig. 4, from which we can observe that there are a large number of false alarms (FAs) of speckles in NACCL [29], INLPG [30], cGAN [31] methods. CAAE [16] exhibits better performance than others on this dataset, due to the fact that it filters the difference map. Our proposed method effectively alleviate this phenomenon thanks to the pseudo-label self-learning and the progressive refinement strategy. The quantitative evaluations are shown in Table I, which is in consistent with the visual inspections.

This article has been accepted for publication in IEEE Transactions on Geoscience and Remote Sensing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TGRS.2023.3294300

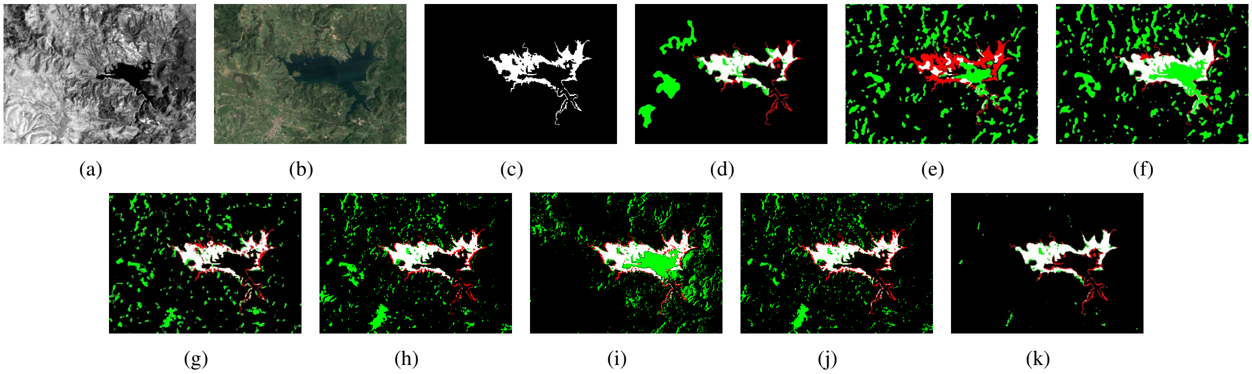


Fig. 4. Results of different methods on Italy dataset. (a) Pre-event image. (b) Post-event image. (c) GroundTruth. (d) FPMS. (e) NACCL. (f) INLPG. (g) IRG-Mcs. (h) SCCN. (i) cGAN. (j) CAAE. (k) Proposed. The TP, TN, FP and FN are represented in white, black, green and red colors.

TABLE I
QUANTITIVE EVALUATIONS OF DIFFERENT METHOD ON ITALY DATASET

Method	*FA*	*MA*	*OA*	*AUC*	κ
FPMS [28]	0.0487	0.2345	0.9398	0.9138	0.5798
NACCL [29]	0.1360	0.8011	0.8230	-	0.0396
INLPG [30]	0.1504	**0.1572**	0.9136	0.9238	0.3471
IRG-Mcs [12]	0.0708	0.2876	0.9158	0.8927	0.4688
SCCN [18]	0.0660	0.2616	0.9231	0.9186	0.5034
cGAN [31]	0.0949	0.2225	0.8966	0.9050	0.4299
CAAE [16]	0.0591	0.2733	0.9278	0.9119	0.5188
Ours	**0.0097**	0.2127	**0.9777**	**0.9711**	**0.8016**

2) Results on Yellow River Dataset: We visualize the confusion maps of different methods on the Yellow River dataset in Fig. 5. In contrast to the Italy dataset, there are no complex effects of light and shadow on this dataset. The main detection difficulty comes from the land around the river in the lower right corner, which is clearly visible in optical modality while is faint in the SAR modality. It can be seen that INLPG [30], IRG-Mcs [12], cGAN [31] and CAAE [16] are caught in such interferences, leading to a huge number of FAs. Although SCCN [18] can successfully learn the feature mapping in bi-temporal images, it still has many FAs. Compared with these methods, our proposed one is able to detect the changes more accurately. The quantitative evaluations are shown in Table II, where our method shows superiority than other compared methods.

TABLE II
QUANTITIVE EVALUATIONS OF DIFFERENT METHOD ON YELLOW RIVER DATASET

Method	*FA*	*MA*	*OA*	*AUC*	κ
FPMS [28]	**0.0027**	0.6402	0.9763	0.9221	0.4897
NACCL [29]	0.0209	0.3450	0.9685	-	0.5616
INLPG [30]	0.0367	0.1991	0.9579	0.9795	0.5363
IRG-Mcs [12]	0.0958	0.1911	0.9011	0.9030	0.3147
SCCN [18]	0.0332	0.1778	0.9231	0.9186	0.5691
cGAN [31]	0.0712	0.2504	0.9231	0.91	0.3541
CAAE [16]	0.0633	0.3144	0.9278	0.9210	0.3530
Ours	0.0068	**0.1666**	**0.9877**	**0.9940**	**0.8110**

3) Results on Texas Dataset: On Texas dataset, both the changed and the unchanged regions are differ in color, which brings difficulty in detecting the changes. If we have accurate labels, it will be easy to detect the changes. However, the acquisition of accurate labels is essentially cumbersome. As Figs. 6(d)-(g) show, many unsupervised methods can only detect a fraction of the changes, which due to that the more discriminative changed regions, like the lower left side of changed region, suppress other regions with less discrimination in implicit feature learning. Since the features in CAAE [16] method is mapped in an explicit manner, the features of the whole changed region are explicitly retained, reflecting the better detection results in Fig. 6(k). Our method utilizes a couple of auto-encoders to encode the input images, and the information of the source images is well preserved. Therefore, it even obtains higher accuracy than CAAE [16]. The same conclusion can be obtained through the quantitative evaluations shown in Table III.

TABLE III
QUANTITIVE EVALUATIONS OF DIFFERENT METHOD ON TEXAS DATASET

Method	*FA*	*MA*	*OA*	*AUC*	κ
FPMS [28]	0.0045	0.9444	0.8960	0.2517	0.0850
NACCL [29]	0.0115	0.8440	0.9004	-	0.2154
INLPG [30]	**0.0023**	0.9497	0.8969	0.9569	0.0813
IRG-Mcs [12]	0.0066	0.8977	0.8986	0.9461	0.1521
SCCN [18]	0.0102	0.2568	0.9621	0.9604	0.7927
cGAN [31]	0.0546	0.4162	0.9094	0.9107	0.5194
CAAE [16]	0.0106	0.1483	0.9748	0.9903	0.8641
Ours	0.0141	**0.0882**	**0.9777**	**0.9923**	**0.8838**

4) Results on California Dataset: The California dataset, as Figs. 7(a)-(b) show, has complex ground objects, such as farmland, mountains, rivers and towns. Due to the lower spatial resolution of this dataset, the artificial buildings in the towns have little influence on the detection results. However, various grid-shaped farmlands have totally different mappings in the bi-temporal images, which brings difficulties for change detection, and many methods thus generate too many false alarms (FAs), for example, in Figs. 7(d), (f) and (h). The complex image content also brings difficulties to the image translation of cGAN [31], resulting in unsatisfactory detection results, as Fig. 7(i) shows. In Fig. 7(g), there are some rounded speckles, which may be derived from the superpixel segmentation of IRG-Mcs [12]. Our method achieves

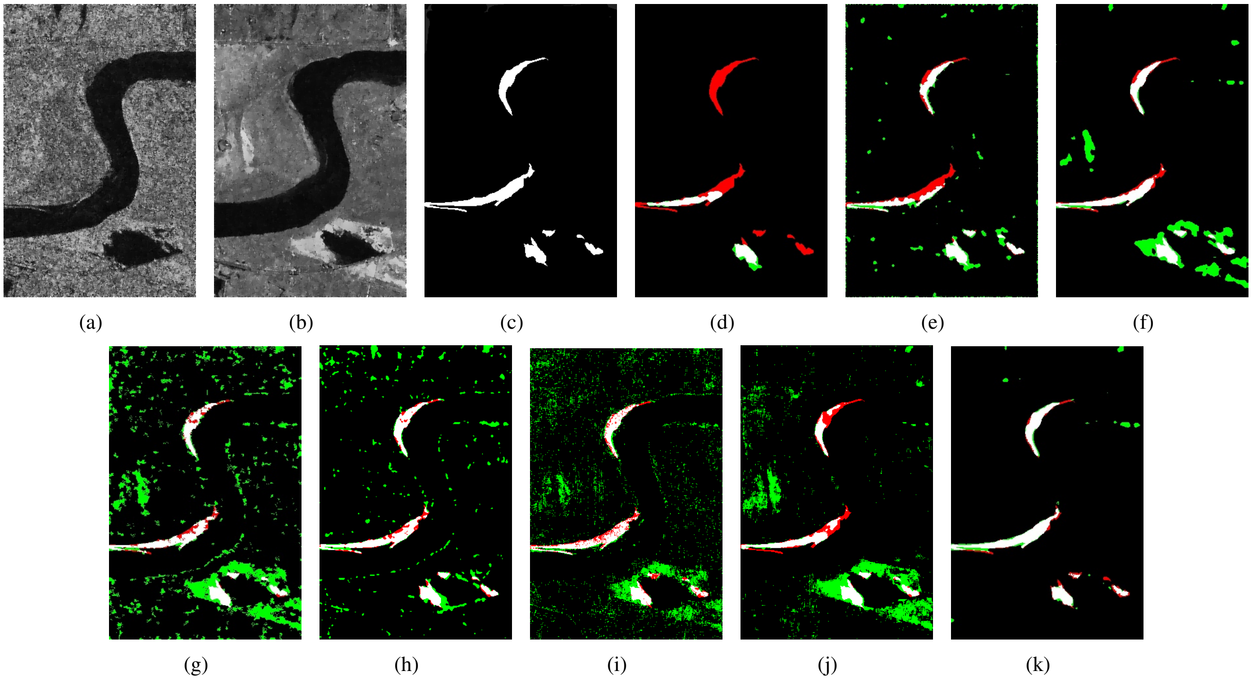


Fig. 5. Results of different methods on Yellow River dataset. (a) Pre-event image. (b) Post-event image. (c) GroundTruth. (d) FPMS. (e) NACCL. (f) INLPG. (g) IRG-Mcs. (h) SCCN. (i) cGAN. (j) CAAE. (k) Proposed. The TP, TN, FP and FN are represented in white, black, green and red colors.

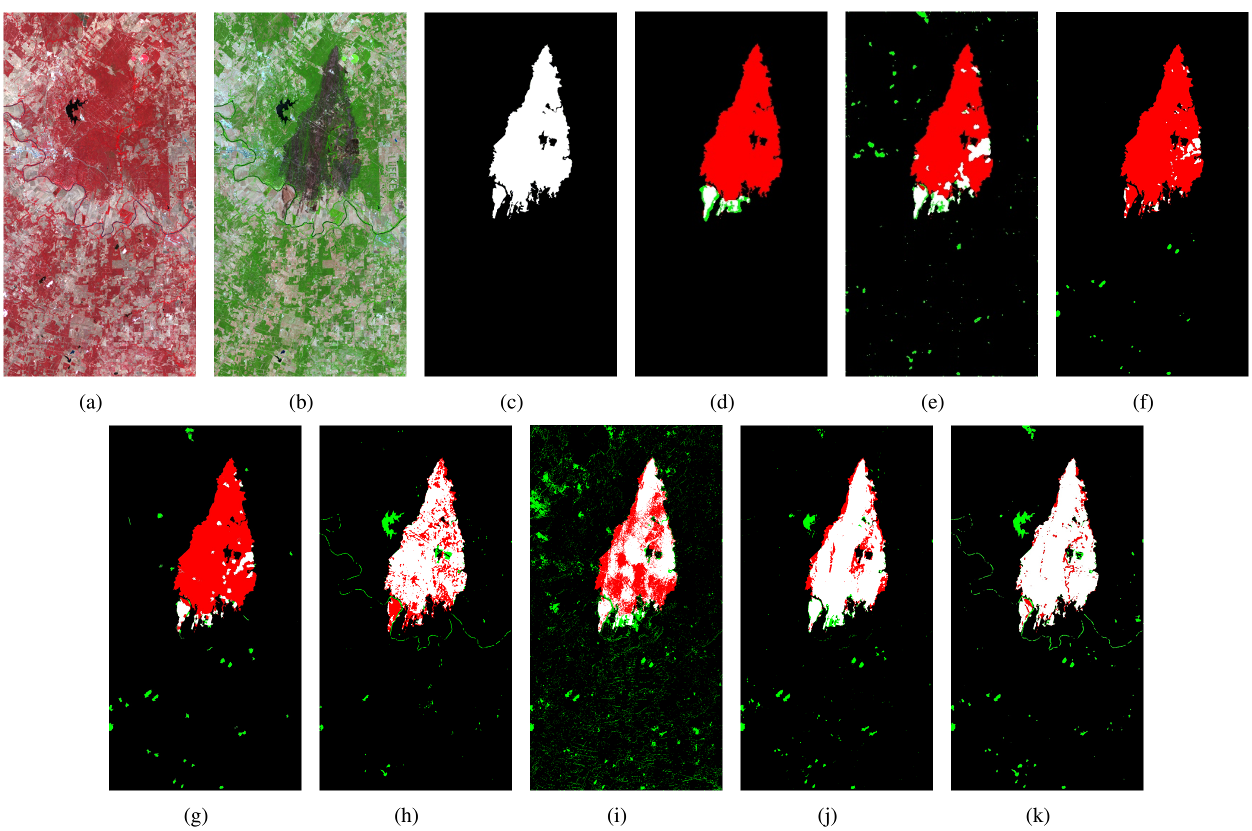


Fig. 6. Results of different methods on Texas dataset. (a) Pre-event image. (b) Post-event image. (c) GroundTruth. (d) FPMS. (e) NACCL. (f) INLPG. (g) IRG-Mcs. (h) SCCN. (i) cGAN. (j) CAAE. (k) Proposed. The TP, TN, FP and FN are represented in white, black, green and red colors.

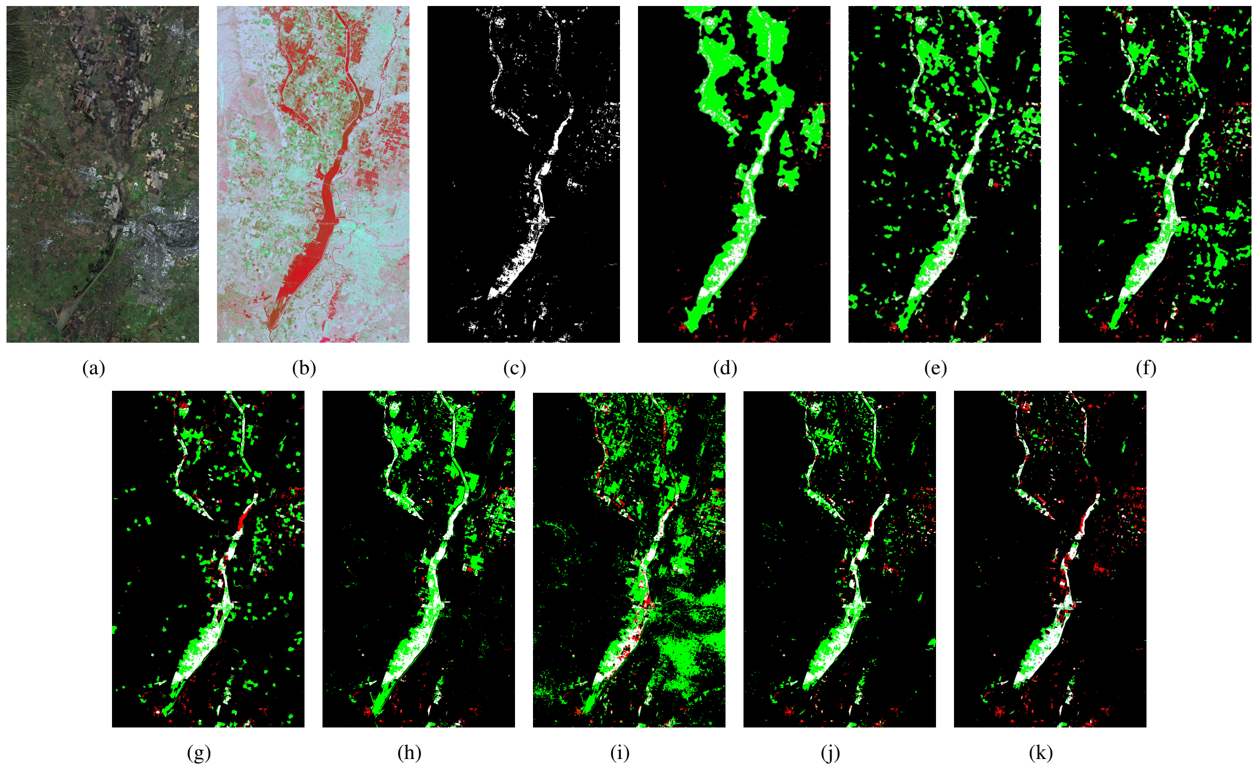


Fig. 7. Results of different methods on California dataset. (a) Pre-event image. (b) Post-event image. (c) GroundTruth. (d) FPMS. (e) NACCL. (f) INLPG. (g) IRG-Mcs. (h) SCCN. (i) cGAN. (j) CAAE. (k) Proposed. The TP, TN, FP and FN are represented in white, black, green and red colors.

superior result compared with other unsupervised methods. The quantitive evaluations are shown in Table IV.

TABLE IV
QUANTITIVE EVALUATIONS OF DIFFERENT METHOD ON CALIFORNIA DATASET

Method	*FA*	*MA*	*OA*	*AUC*	κ
FPMS [28]	0.1448	0.1755	0.8540	0.9132	0.2464
NACCL [29]	0.1109	0.1440	0.8879	-	0.3182
INLPG [30]	0.0918	0.2139	0.8971	0.9227	0.3477
IRG-Mcs [12]	0.0659	0.2635	0.9266	0.9232	0.4010
SCCN [18]	0.0879	**0.1174**	0.9110	**0.9532**	0.3956
cGAN [31]	0.3295	0.2527	0.8359	0.8305	0.2749
CAAE [16]	0.0502	0.1923	0.9360	0.9471	0.5585
Ours	**0.0201**	0.3596	**0.9703**	0.9443	**0.6023**

5) Results on Shuguang Dataset: This dataset exhibits more details of ground, due to its higher spatial resolution. The changed areas range from farmland to the artificial buildings. Artificial architecture is difficult to detect because it is more complex than natural landscape. Almost all methods are inaccurate in edge detection, especially the black farmlands in the original image. Probably because their mapping relationship in the bi-temporal image is obviously different from those around them. There are a lot of noise spots in the results of cGAN [31] and CAAE [16], which mainly attribute to the inaccurate image translation. The binarized segmentation algorithm, like in FPMS [28], can suppress these noises. Our method also performs well in noise suppression as Fig. 8(k) shows. Nevertheless, our method can not achieve the best quantitive evaluations in Table V in terms of AUC and kappa coefficient κ, due to the lack of prior knowledge.

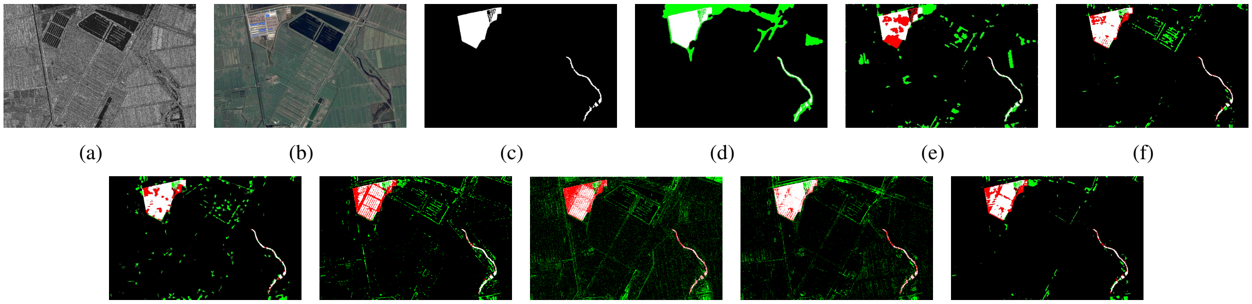
TABLE V
QUANTITIVE EVALUATIONS OF DIFFERENT METHOD ON SHUGUANG DATASET

Method	*FA*	*MA*	*OA*	*AUC*	κ
FPMS [28]	0.0715	**0.0027**	0.9317	**0.9938**	0.5412
NACCL [29]	0.0394	0.3946	0.9444	-	0.4700
INLPG [30]	0.0203	0.2176	0.9706	0.9827	**0.6945**
IRG-Mcs [12]	0.0258	0.2163	0.9654	0.9739	0.6576
SCCN [18]	0.0373	0.4333	0.9445	0.9163	0.4551
cGAN [31]	0.0890	0.4933	0.8920	0.8323	0.2560
CAAE [16]	0.0518	0.1750	0.9425	0.9655	0.5425
Ours	**0.0142**	0.4324	**0.9708**	0.9720	0.6552

D. Ablation Studies

In this section, we conduct experiments incrementally to demonstrate the effectiveness of different components or strategies. We take the model that only has the PSN as our baseline, and the experimental results are shown in Table VI.

1) Necessity of Warm-up Training: In our progressive refinement strategy, warm-up training is very important since it provides adequate initialization for the iteration structure. From the first and the second rows of Table VI, we can observe that warm-up training dramatically improves the detection accuracy except for the Texas dataset. The prior map generated by the warm-up training on Texas dataset is inaccurate, leading to poorer initialization. Due to the absence of PRN and



(a) (b) (c) (d) (e) (f)

(g) (h) (i) (j) (k)

Fig. 8. Results of different methods on Shuguang dataset. (a) Pre-event image. (b) Post-event image. (c) GroundTruth. (d) FPMS. (e) NACCL. (f) INLPG. (g) IRG-Mcs. (h) SCCN. (i) cGAN. (j) CAAE. (k) Proposed. The TP, TN, FP and FN are represented in white, black, green and red colors.

progressive refinement, the model training in this case may be misled to a false optimization direction. From the last two rows in Table VI, we find that such a misleading phenomenon can be eliminated by PRN with attention and the progressive refinement.

2) Necessity of Pseudo-label Self-Learning: After warm-up training, we can obtain several pseudo-labels through PSN. These generated pseudo-labels are utilized to guide the model learning. To verify the necessity of pseudo-label self-learning, we conduct ablative experiments and the results are shown in Table VI. It should be noted that there is a channel attention structure in our PRN, we list the results with/without channel attention to further demonstrate the effectiveness of attention mechanism in our model. It can be seen from the table that the introduction of pseudo-label self-learning strategy brings great improvements on the detection accuracy, and our model also benefit from the channel attention.

3) Effectiveness of Progressive Refinement Strategy: We know that the pseudo-labels are not accurate even though they are generated by PSN after the warm-up training. Therefore, training the model only by generated coarse pseudo-labels is infeasible, and we should refine them by a progressive refinement strategy. The last row in Table VI shows the results on five datasets. We can observe from the table that the progressive refinement strategy brings the increase of kappa coefficient between 1.5% and 8.3%, except for Texas dataset that decreases about 1.3%.

E. Hyper-parameter Analysis

There are two critical hyperparameters in our model, *i.e.* the proportion of changed pixels in pseudo-labels (N) and the segmentation threshold (p_t). In this section, we provide the selection principle of them.

Based on the assumption that changed regions are always small proportion of the whole image, we choose a small value of N. Keeping other configurations unchanged, we investigate the influence of N by experiments. Fig. 9 draws the variation of kappa coefficient with the increase of N. The kappa coefficient dramatically increases up to $N = 5\%$. But when $N > 5\%$, kappa coefficient varies faintly. For Texas and Italy datasets, the best results are obtained under $N = 8\%$, while for other three datasets, the variations are small, which proves that the proposed method is insensitive to pseudo-labels. After comprehensive consideration, we select $N = 8\%$ in our model.

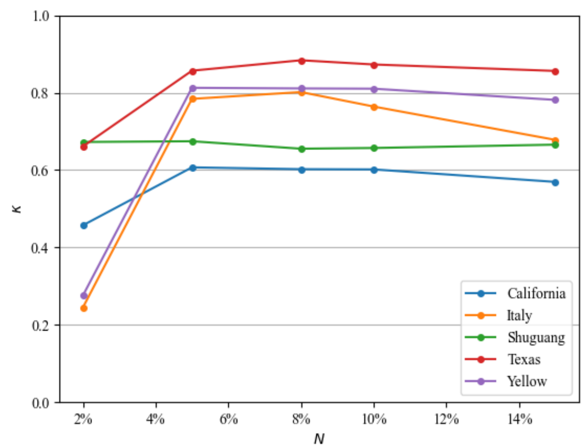


Fig. 9. Kappa coefficient κ of proposed method on five datasets with different value of N.

p_t is a threshold used to obtain the binary change map. If we choose a higher value, the MAs rate will be high. On the contrary, a lower p_t will increase the FAs rate. The detection performance of our model varies on five different datasets, as depicted in Fig. 10. We find that when $p_t = 0.95$, our model can achieve the highest kappa coefficient on almost all datasets. Therefore, p_t is set to 0.95.

V. Conclusion

In this paper, we propose an unsupervised progressive modality-alignment heterogeneous change detection method. The proposed model achieves modality-alignment and pseudo-label refinement alternately, improving the accuracy of change detection. A pseudo Siamese network is firstly used to map and align the features of bi-temporal images, and then the pseudo-labels are generated and refined by the model itself. After that, these pseudo-labels are taken as guidance to learn the change map. Such a pseudo-label self-learning strategy can effectively suppress false alarms, thus further improving the detection accuracy. The whole model is under an iterative framework, enabling it to well locate some subtle details.

TABLE VI
ABLATION STUDIES

Modules				Kappa Coefficient				
Warm Up Training	PRN w/o Attention	PRN w/ Attention	Progressive Refinement	Italy	Yellow River	Shuguang	Texas	California
				0.2146	0.4224	0.2205	0.6642	0.1223
✓				0.7492	0.7586	0.4777	0.1703	0.3946
✓	✓			0.7969	0.7943	0.5553	0.6143	0.4450
✓		✓		0.7850	0.7960	0.5722	**0.8965**	0.5314
✓		✓	✓	**0.8016**	**0.8110**	**0.6552**	0.8838	**0.6023**

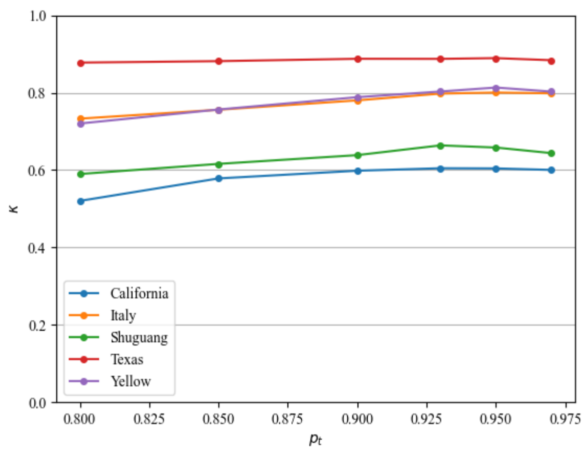


Fig. 10. Kappa coefficient κ of proposed method on five datasets with different value of p_t.

Experimental results also validate the effectiveness of the proposed model. However, our proposed method contains a two-stage training, which may be inefficient in some situations. Moreover, it has many hyper-parameters to be manually set, such as the proportion in producing pseudo-labels and the threshold in generating the binary map, which is inflexible and requires some priors. In the future, we will explore a simpler and general unsupervised framework for heterogeneous image change detection.

REFERENCES

[1] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *International journal of remote sensing*, vol. 10, no. 6, pp. 989–1003, 1989.

[2] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using vhr optical and sar imagery," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 5, pp. 2403–2420, 2010.

[3] D. C. Mason, R. Speck, B. Devereux, G. J.-P. Schumann, J. C. Neal, and P. D. Bates, "Flood detection in urban areas using terrasar-x," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 48, no. 2, pp. 882–894, 2009.

[4] J. L. Gil-Yepes, L. A. Ruiz, J. A. Recio, Á. Balaguer-Beser, and T. Hermosilla, "Description and validation of a new set of object-based temporal geostatistical features for land-use/land-cover change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 121, pp. 77–91, 2016.

[5] Z. Zhu, "Change detection using landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 130, pp. 370–384, 2017.

[6] C. Toth and G. Jóźków, "Remote sensing platforms and sensors: A survey," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 115, pp. 22–36, 2016, theme issue 'State-of-the-art in photogrammetry, remote sensing and spatial information science'.

[7] S. Saha, P. Ebel, and X. X. Zhu, "Self-supervised multisensor change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–10, 2021.

[8] G. Camps-Valls, L. Gómez-Chova, J. Muñoz-Marí, J. L. Rojo-Álvarez, and M. Martínez-Ramón, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 6, pp. 1822–1835, 2008.

[9] L. Wan, Y. Xiang, and H. You, "A post-classification comparison method for sar and optical images change detection," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 7, pp. 1026–1030, 2019.

[10] ——, "An object-based hierarchical compound classification method for change detection in heterogeneous optical and sar images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9941–9959, 2019.

[11] Y. Wu, Z. Bai, Q. Miao, W. Ma, Y. Yang, and M. Gong, "A classified adversarial network for multi-spectral remote sensing image change detection," *Remote. Sens.*, vol. 12, p. 2098, 2020.

[12] Y. Sun, L. Lei, D. Guan, and G. Kuang, "Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images," *IEEE Transactions on Image Processing*, vol. 30, pp. 6277–6291, 2021.

[13] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 12, pp. 9960–9975, 2019.

[14] L. Wan, T. Zhang, and H. J. You, "Multi-sensor remote sensing image change detection based on sorted histograms," *International Journal of Remote Sensing*, vol. 39, no. 11, pp. 3753–3775, 2018.

[15] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogenous remote sensing images via homogeneous pixel transformation," *IEEE Transactions on Image Processing*, vol. 27, no. 4, pp. 1822–1834, 2017.

[16] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, F. M. Bianchi, G. Moser, R. Jenssen, and S. N. Anfinsen, "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2022.

[17] S. Jia, S. Jiang, Z. Lin, M. Xu, W. Sun, Q. Huang, J. Zhu, and X. Jia, "A semisupervised siamese network for hyperspectral image classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–17, 2022.

[18] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 3, pp. 545–559, 2018.

[19] W. Zhao, Z. Wang, M. Gong, and J. Liu, "Discriminative feature learning for unsupervised change detection in heterogeneous images based on a coupled neural network," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 12, pp. 7066–7080, 2017.

[20] T. Zhan, M. Gong, X. Jiang, and S. Li, "Log-based transformation feature learning for change detection in heterogeneous images," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 9, pp. 1352–1356, 2018.

[21] S. N. Anfinsen, L. T. Luppino, M. A. Hansen, G. Moser, and S. B. Serpico, "Unsupervised heterogeneous change detection in radar images

by cross-domain affinity matching," in *2020 IEEE Radar Conference (RadarConf20)*. IEEE, 2020, pp. 1–6.

[22] D. Wang, F. Zhao, H. Yi, Y. Li, and X. Chen, "An unsupervised heterogeneous change detection method based on image translation network and post-processing algorithm," *International Journal of Digital Earth*, vol. 15, no. 1, pp. 1056–1080, 2022.

[23] T. Zhan, M. Gong, J. Liu, and P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 38–51, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618302417

[24] M. Yang, L. Jiao, F. Liu, B. Hou, S. Yang, and M. Jian, "Dpfl-nets: Deep pyramid feature learning networks for multiscale change detection," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2021.

[25] P. Zhang, M. Gong, L. Su, J. Liu, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 116, pp. 24–41, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271616000563

[26] F. Bovolo, S. Marchesi, and L. Bruzzone, "A framework for automatic and unsupervised detection of multiple changes in multitemporal images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 50, no. 6, pp. 2196–2212, 2012.

[27] H. Li, M. Gong, M. Zhang, and Y. Wu, "Spatially self-paced convolutional networks for change detection in heterogeneous images," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 4966–4979, 2021.

[28] M. Mignotte, "A fractal projection and markovian segmentation-based approach for multimodal change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 58, no. 11, pp. 8046–8058, 2020.

[29] ——, "Mrf models based on a neighborhood adaptive class conditional likelihood for multimodal change detection," *AI, Computer Science and Robotics Technology*, vol. 2022, pp. 1–20, 03 2022.

[30] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–21, 2022.

[31] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 1, pp. 45–49, 2019.

[32] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *2018 25th IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 4063–4067.

[33] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved unet++," *Remote Sensing*, vol. 11, no. 11, 2019. [Online]. Available: https://www.mdpi.com/2072-4292/11/11/1382

[34] 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 Transactions on Geoscience and Remote Sensing*, vol. 58, no. 4, pp. 2848–2864, 2020.

[35] L. Bruzzone and D. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 38, no. 3, pp. 1171–1182, 2000.

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

[37] M. Hao, W. Shi, H. Zhang, and C. Li, "Unsupervised change detection with expectation-maximization-based level set," *IEEE Geoscience and Remote Sensing Letters*, vol. 11, no. 1, pp. 210–214, 2014.

[38] H.-C. Li, T. Celik, N. Longbotham, and W. J. Emery, "Gabor feature based unsupervised change detection of multitemporal sar images based on two-level clustering," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 12, pp. 2458–2462, 2015.

[39] S. Saha, L. Kondmann, Q. Song, and X. X. Zhu, "Change detection in hyperdimensional images using untrained models," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 11 029–11 041, 2021.

[40] L. Bergamasco, S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised change-detection based on convolutional-autoencoder feature extraction," 10 2019, p. 34.

[41] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *Artificial Neural Networks and Machine Learning – ICANN 2011*, T. Honkela, W. Duch, M. Girolami, and S. Kaski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 52–59.

[42] G. Mercier, G. Moser, and S. B. Serpico, "Conditional copulas for change detection in heterogeneous remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 46, no. 5, pp. 1428–1441, 2008.

[43] V. Pascal, L. Hugo, L. Isabelle, B. Yoshua, and M. Pierre-Antoine, "Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion," *Journal of Machine Learning Research*, 2010.

[44] Y. Sun, L. Lei, X. Li, H. Sun, and G. Kuang, "Nonlocal patch similarity based heterogeneous remote sensing change detection," *Pattern Recognition*, vol. 109, p. 107598, 2021.

[45] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[46] Z.-G. Liu, Z.-W. Zhang, Q. Pan, and L.-B. Ning, "Unsupervised change detection from heterogeneous data based on image translation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–13, 2022.

[47] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2242–2251.

[48] X. Jiang, G. Li, Y. Liu, X.-P. Zhang, and Y. He, "Change detection in heterogeneous optical and sar remote sensing images via deep homogeneous feature fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 1551–1566, 2020.

[49] T. Zhan, M. Gong, J. Liu, and P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 146, pp. 38–51, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924271618302417

[50] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 9, no. 1, pp. 62–66, 1979.

[51] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018.