Improving Reliability of Heterogeneous Change Detection by Sample Synthesis and Knowledge Transfer

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

***Abstract*— Detecting changes in heterogeneous images without the supervision of changed label is a challenging yet critical task for quick responding natural disaster relief. Nevertheless, most of available unsupervised heterogeneous change detection (CD) methods strong rely on the quality of pseudo-labels, and they suffer from performance degradation, even irreversible model collapse, when encounter the low-quality pseudo-labels, leading to unreliable detection results. In order to improve the reliability of unsupervised heterogeneous CD, in this article, we propose a novel CD paradigm based on sample synthesis and knowledge transfer. We address the issue of label reliability by artificially creating a changed region and assigning labels rather than constructing pseudo-labels. These constructed labels guide the network in automatically learning the correspondence between heterogeneous images, confirming the reliability of changed regions. Moreover, an augmentation with synthetic samples on real samples makes it possible to generate more transferable samples while reducing the domain gap coarsely. A dual-branch joint training with feature contrastive learning is further developed to transfer the knowledge of changes from the synthetic sample domain to real sample domain. Experimental results on five public datasets demonstrate that our proposed method has superior performance when compared with available state-of-the-art (SOTA) methods. Our code is available at https://github.com/zhangqiiii/SS-KT.**

***Index Terms*— Change detection (CD), heterogeneous, knowledge transfer, sample synthesis.**

Manuscript received 16 November 2023; revised 31 January 2024 and 18 March 2024; accepted 25 March 2024. Date of publication 27 March 2024; date of current version 8 April 2024. This work was supported in part by the National Natural Science Foundation of China (NFSC) under Grant 62201467 and Grant 62101453, in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2021A1515110544, in part by the Project funded by China Postdoctoral Science Foundation under Grant 2022TQ0260, in part by the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313088, in part by the Natural Science Basic Research Program of Shaanxi under Grant 2024JC-YBQN-0719 and Grant 2022JC-DW-08, and in part by the Natural Science Foundation of NingBo under Grant 2023J262. *(Corresponding author: Lingyan Ran.)*

Yinghui Xing is 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, and also with the Research and Development Institute, Northwestern Polytechnical University in Shenzhen, Shenzhen 518057, China.

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-Ground-Ocean Big Data Application Technology, School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China (e-mail: xyh_7491@nwpu.edu.cn).

Digital Object Identifier 10.1109/TGRS.2024.3382474

I. Introduction

REMOTE sensing image change detection (CD) is of significant importance in monitoring human activities and natural resources [1], such as ecological environment monitoring [2], [3], urban development planning [4], and disaster rescue and assessment [5], [6]. It involves comparing remote sensing images of the same location captured at different times in order to identify changes.

With the continuous progress of remote sensing technology, an increasing number of satellites have been launched, which provide diverse remote sensing images, thus facilitating the remote sensing interpretations via multisource images [7], [8], [9]. In the field of CD, heterogeneous CD [10], [11], [12] allows the utilization of images acquired from any sensor to detect changes. Due to the limited spatial resolution, accurately labeling remote sensing data is essentially labor-intensive [13], and the heterogeneity also raises the difficulties of labeling samples, especially for SAR images that require extensive extra expert knowledge. Therefore, most of available heterogeneous CD methods are based on unsupervised learning [14]. Unsupervised heterogeneous CD technology does not require a large number of labeled images, demonstrating great potential in practice.

The challenges of unsupervised heterogeneous CD are two folds. The first one is unquestionable heterogeneity. Due to the imaging difference, heterogeneous images are visually unique for the same geographical area [15], bringing about interference for models to clearly discriminate the truly changed regions. To improve the discrimination ability, Touati et al. [16] proposed to estimate the pixel-pairwise distributions in a Bayesian framework. Liu et al. [17] transformed the heterogeneous images into a common space where images share the same statistical properties. The second challenge for unsupervised heterogeneous CD lies in ambiguous prior. It is difficult for models to accurately learn target changes without label supervision. Therefore, most of researchers first train the models several iterations to obtain an initial change map, also known as pseudo-labels, and then use these pseudo-labels to guide the model training [18], [19]. Though effective, the reliability of them is skeptical. Since the detected difference between bitemporal images may come from different imaging mechanism, dissimilar resolutions, distinct appearance, as well as target changes, and the models

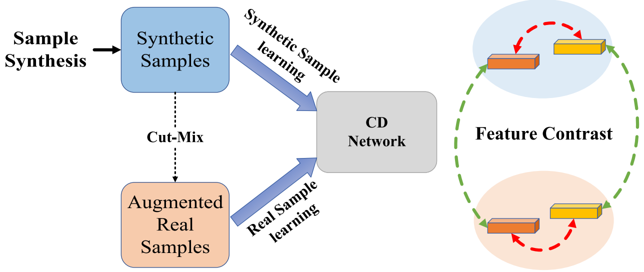


Fig. 1. Overview of the proposed approach, which mainly contains sample synthesis process, a dual-branch joint learning process, and a CD network with feature contrast constraint.

cannot differentiate them, which possibly leads to irreversible false predictions. Some advanced CD methods focus on edge integrity and internal holes phenomenon [20], and improve detection accuracy through sample balance strategies [21]. However, these designs depend on accurate label information. To cope with above problems, there are some CD methods based on sample generation. Zheng et al. [22] utilized an annotated building extraction dataset to synthesize bitemporal images for training. It bypassed the manual annotation process of pixel-wise change masks from bitemporal images, but required a large number of building extraction labels, which is also difficult to obtain [23]. Aiming at the task of building CD, Chen et al. [24] proposed an instance-level change augmentation strategy to generate bitemporal images that contain changes of diverse buildings, and leveraged the generative adversarial training to detect changes. Similarly, Sun et al. [23] designed a pseudo-bitemporal data generator to generate a large number of pseudo-bitemporal images with CD labels. All of them concentrate on the homogeneous CD task, and they generate data for the changes of building objects. Owing to the heterogeneity and the diverse changing objects, these methods cannot be directly taken into the heterogeneous CD task.

To make the model discriminate between real changes and heterogeneous differences. In this article, we design a simple changed sample synthesis strategy to simulate the changes. Specifically, the cut–paste operation [25] is utilized to synthesize changed samples. We first construct a bank of pasting pieces, and then select an appropriate region to paste a piece from the bank. The changed labels of synthetic samples are obtained during this process. The introduction of synthetic samples compels the CD model to be aware of the real changes rather than appearance differences produced by different imaging mechanism. Since the process of sample synthesis only considers regular shapes with distinct boundaries, which is inconsistent with realistic change scenes, we further propose a dual-branch joint training to facilitate knowledge transfer from synthesis samples to real samples. As illustrated in Fig. 1, the scheme contains a synthetic-sample branch and a real-sample branch, where the synthetic samples are first used to augment real samples to coarsely reduce domain gap. And the dual-branch samples pass through a CD network simultaneously to calculate the feature contrastive loss. The contrastive loss is used to constrain the feature similarity between two branches, aiming to improve the model's domain generalization ability and feature discrimination ability.

Through the cut–paste of simple shapes like circle and rectangle, the model can obtain a direct prompt to recognize real changes. And the dual-branch joint training reduces the domain gap between synthetic domain and real domain, which further improves the reliability of detection results.

The main contributions of this article are as follows.

1) We develop a novel unsupervised CD approach based on sample synthesis. The proposed method constrains the model to learn changes and suppresses the interference of other factors by a naive and direct sample synthesis strategy.
2) A dual-branch joint training strategy is designed, which enables the synthetic samples to guide the learning process of real samples through knowledge transfer.
3) Through the designed architecture, a feature contrastive mechanism is introduced, greatly improving the detection performance. The distinction between classes enhances the discrimination of features, while the similarity between two branches facilitates the extraction of change-aware features, i.e., domain-agnostic information.

II. Related Works

A. Heterogeneous Change Detection

Heterogeneous CD is a hot topic in the field of remote sensing. Earlier methods, like statistical models [26], [27], energy models [28], [29], and evidence theory [30], are on the basis of the assumption that heterogeneous images share similar structure features for the same ground targets [10]. In recent years, three groups of methods have gradually emerged: feature alignment-based, image translation-based, and image structural similarity-based. Feature alignment-based methods project heterogeneous images to higher-dimensional space, and then try to find some invariant relationships between them. Liu et al. [14] proposed an asymmetrically coupled network to project two images into a comparable feature space. Li et al. [31] applied the self-paced learning theory to systematically improve the projection distance between heterogeneous images. Xing et al. [18] achieved progressive alignment of bitemporal image features by refining the network's projection results. Similarly, PRBCD [32] and SGAE [33] involved a coarse prediction module and an iterative refining module to extract discriminative features and then generate a refined change map by change map optimizers. Wei et al. [34] utilized the Transformer model to develop an interactive mapping encoder, enhancing performance by leveraging the global modeling capability of Transformer. However, excavating the correspondence between heterogeneous images is difficult and unstable, which limits the performance of these methods. Image translation-based methods try to obtain homogeneous images through style transformation or regression estimation. HPT [17] achieved mutual image translation through the pixel-level regression algorithm. Nonetheless, the detection accuracy deeply relies on the quality of translated images. cGAN [35] utilized a conditional GAN to translate images and subsequently improved the results using an approximation network. Gong et al. [36] proposed a coupling translation network [36] based on CycleGAN with

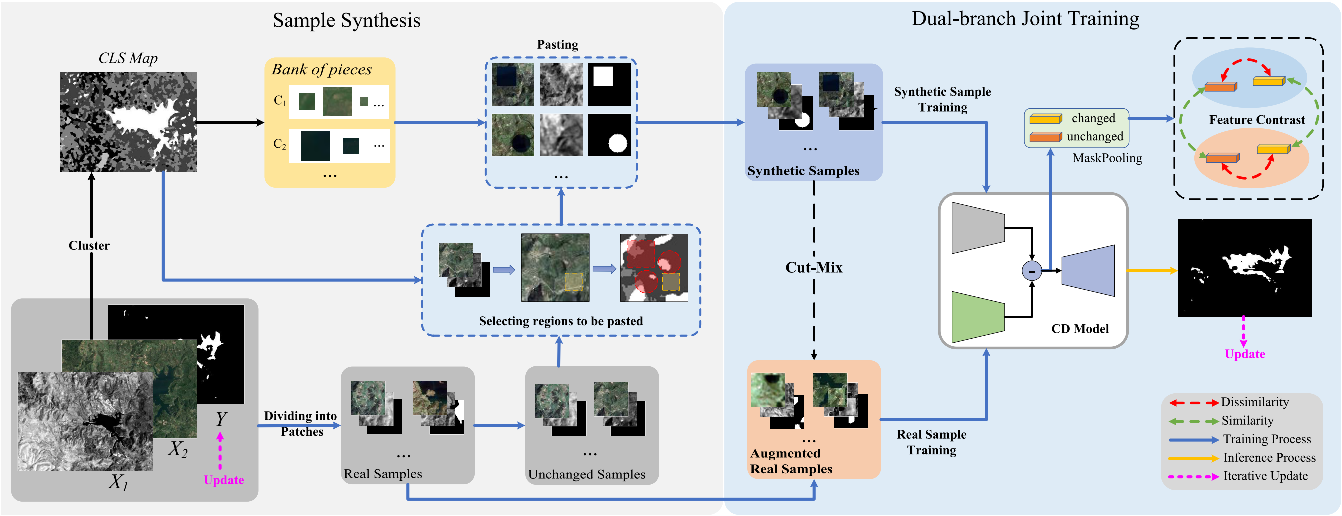


Fig. 2. Detailed architecture of the proposed method. In the sample synthesis process, the pasting region is selected from image pairs of bitemporal images (X_1 and X_2) and updated change map Y, while the pasted pieces are constructed and selected from single-temporal image. In the dual-branch joint training, cut-mix is first applied to coarsely reduce the domain gap between synthetic and real samples, and a feature contrastive learning is also used to further facilitate the knowledge transfer.

a partially shared generator. Though some of them [37], [38] have successfully achieved high-quality image translation through the utilization of a well-designed loss function, the direct comparison in the image domain is easily disturbed by noises. Methods based on image structural similarity require to manually extract features to describe the structure of the image. The image structural similarity is usually measured by graph structure. NLPG [39] and INLPG [40] established the similarity between each target patch and other patches. They evaluate the similarity of graph structure by aligning graph relationships to the same phase in order to assess the changes in the target. Similarly, Sun et al. [41] designed an iterative graph structure model, and the change graph is obtained through conditional random field segmentation. However, in general, manually extracted features are inflexible and not robust enough to apply to different datasets.

B. Data Synthesis Through Cut–Paste

In many real-world scenarios, obtaining real task-related data can be expensive or even impossible [42], [43]. Synthesizing data is a frequently utilized way in model training [44], [45], [46]. There are various data synthesis methods [47], [48], [49], [50], [51]. Among them, cut–paste [25], a data augmentation method, is widely used in deep learning to improve the generalization ability of models by increasing the amount of data samples. At the same time, it is also utilized to synthesize data in various computer vision tasks. Wang et al. [52] proposed an unsupervised pre-training method for semantic segmentation, which pasted the same image onto different backgrounds, making the model learn semantics of different regions by modeling the similarity of foreground features. Dwibedi et al. [53] synthesized training data by automatically cutting object instances and then pasting them onto a random background. In [54], different anomaly regions were generated on perfect images by cut–paste, which enabled the model to detect unknown anomaly patterns without anomaly data. Cut–paste has a natural correlation with CD task, because the cut patches is essential a changed region. Therefore, it is utilized to directly synthesize training data in some semi-supervised and unsupervised CD methods. Chen et al. [24] synthesized new data by cutting and pasting instance-level labels of a single phase to train the detector. Based on cut–paste, Seo et al. [55] made it possible to train a CD model on two spatially unrelated images. However, both of them aim to generate realistic synthetic datasets, which is laborious and also requires external datasets or labels. In this article, we simulate the changed regions by randomly generating some simple shapes to cut and paste.

III. Methodology

A. Overview

In this article, we improve the reliability of heterogeneous CD by first synthesizing changed samples and then reducing domain gap between synthetic and real samples. The detailed architecture is illustrated in Fig. 2, where we have bitemporal images, including the pre-event image X_1 and the post-event image X_2, as well as the prior change map Y that is zero-initialized and can be updated during the iteration. Our sample synthesis depends only on single-temporal images, i.e., cutting and pasting are conducted on the same temporal phase of images. Specifically, we first use the clustering method to obtain a classification (CLS) map from a certain of temporal phase image, and then construct a bank of pieces with multiple scales and classes according to the CLS map. At the same time, image pairs together with updated change map are divided into patches to form a real sample set, from which we further pick out the unchanged samples to select regions to be pasted. Note that the regions can be squares, rectangles, or even cycles, and CLS map also guides the process of selection. Then, we randomly choose some pieces to be pasted on the selected regions from bank of pieces. Since the cut and paste are executed on the same temporal phase, the synthetic changed regions are located in the position where we select the regular shapes. Therefore, we can obtain the paired synthetic samples with corresponding changed labels.

Unlike many homogeneous CD methods based on image generation, we synthesize samples in a naive way, where the changed regions are constructed with regular shapes, which is inconsistent with practical scenes. In other words, there is a large domain gap between synthetic sample set and real sample set. To reduce the domain gap and realize reasonable knowledge transfer, we adopt two strategies. The first one is cut-mix with synthetic samples, and the second one is dual-branch joint training. We first use cut-mix to augment real samples. Then, both the synthetic sample set and the augmented real sample set pass through a CD model to obtain the change map. This change map is on one hand combined with pre-event and post-event images to update real sample set, and on the other hand, it acts as a mask to distinguish changed and unchanged prototype vectors. These prototype vectors are considered in feature contrastive learning to further reduce the domain gap.

B. Sample Synthesis

We use synthetic samples to simulate changed region. To maintain consistency within the dataset, we select the pasting pieces and the regions to be pasted on the same temporal phase image.

1) Construction of Pasting Pieces: It is essential to ensure that the pasting piece and the pasted region belong to distinct semantic classes. Therefore, we extract uniform regions as far as possible to form a bank of pieces. Specifically, the whole post-event image is first classified into C categories by the K-means clustering algorithm, and the classification map is then filtered by a median filter, which confirms the uniformity within class regions, and also reduces the speckle noises of SAR images. This process can be formulated as

$$\mathrm{cls} = \mathrm{Median}(\mathrm{Cluster}(X_t)), \quad X_t \in \mathbb{R}^{H\times W} \tag{1}$$

where H and W denote the height and width of the image. X_t can either be pre-event image X_1 or post-event image X_2. Since the changed objects always appear in the post-event phase, we choose pasting pieces within post-event phase image, i.e., $X_t = X_2$. Then, we search for pieces by sliding three sizes of windows on classification map cls, i.e., small (S), medium (M), and large (L)

$$\mathrm{cls}_w = \left\{\mathrm{cls}_{i,j}^{\mathrm{size}} \middle| i, j \in (0, s, 2s, \ldots, ns), \mathrm{size} \in (S, M, L)\right\} \tag{2}$$

where s is search stride. After that, we count the number of pixels belonging to each class, and find the most dominant class c^*. When the number of pixels belongs to c^*th class in the window is greater than a predefined threshold, we crop the window in original post-event image and put it into the c^*th class of the bank

$$c^* = \arg\max_c\left(\mathrm{count}_c\left(\mathrm{cls}_{i,j}^{\mathrm{size}}\right)\right), \quad c \in (1, 2, \ldots, C)$$
$$p^{c^*} = \begin{cases} X_{t,(i,j)}^{\mathrm{size}}, & \dfrac{\mathrm{count}_{c^*}}{\sum \mathrm{count}_c} > 0.95 \\ \mathrm{Null}, & \text{otherwise} \end{cases} \tag{3}$$

where $\mathrm{count}_c(\cdot)$ is the function to count the number of pixels belonging to cth class, and p^{c^*} is a pasting piece with class c^*. $X_{t,(i,j)}^{\mathrm{size}}$ indicates the selected piece on image X_t centered in the position (i, j). Note that we choose a relatively higher threshold 0.95 to confirm the uniformity of pieces, and the model is not sensitive to the threshold. The whole bank of pieces are expressed as

$$\mathrm{Bank} = \left\{p_k^c \middle| c \in (1, 2, \ldots, C), k \in (1, 2, \ldots, N_c)\right\} \tag{4}$$

where N_c denotes the number of pasting pieces for class c.

2) Selection of Pasted Regions: Ideally, the selected pasting piece should be placed in the unchanged regions to keep the distinctiveness of boundaries between changed and unchanged regions, or the boundaries between pasting and pasted regions. Therefore, we divide the pre-event, post-event images, and the prior change map into patches, and select the unchanged samples based on a prior change map Y, which indicates regions likely to be changed or unchanged, and can be updated by the CD model. Specifically, we consider the patch samples where the proportion of changed pixels is below 0.01 within the patch as unchanged samples.

To ensure diversity of synthetic samples, we randomly select the pasted region's location, size, and shape in each unchanged sample. Meanwhile, to mitigate the potential impact of class inconsistency caused by pasted regions, we only select regions, whose class consistency measurement is higher than T_{pasted} as pasted regions.

3) Construction of Synthetic Samples: After obtaining the bank of pasting pieces and the pasted regions, we can construct synthetic samples through cut–paste. To further confirm the distinctiveness of boundaries between pasting and pasted regions, we first use the mean value $\bar{x} = (1/hw)\sum x_{i,j}$ of pasted region to represent the whole region. And then calculate the distances between $\bar{x}$ with C center of clustering to choose the Top-n farthest categories. Finally, we randomly select one pasting piece from n classes to paste to the selected region. After construction of the synthetic samples, the corresponding change map y^s can also be obtained.

C. Dual-Branch Joint Training

Since the synthetic samples have regular changed shapes, which is inconsistent with practical scenes, we use dual-branch joint training to reduce the domain gap between synthetic and real samples. First, real samples are augmented by cut-mix with synthetic samples to reduce the gap between synthetic sample set and real sample set. Then the synthetic samples and augmented real samples pass through the CD network in parallel to obtain detection results.

As Fig. 2 shows, the CD model contains two encoders and a decoder, where two encoders extract features of bitemporal images. Then the difference between extracted bitemporal features is obtained. These difference features are used on one hand to pass through the decoder to obtain results, on the other hand, they are processed by the mask-pooling to acquire changed and unchanged class prototypes, which are used in the dual-branch contrastive learning.

1) Synthetic Sample Branch: The synthetic sample branch utilizes synthetic samples to make the model learn real changes and at the same time surpasses the interference of other factors. In Section III-B3), we merge the prior change information with that of the pasted region to obtain the labels of the synthetic samples. During the selection of pasted regions, we select the unchanged regions as far as possible; however, inevitably, there are fewer changed pixels. To avoid their negative impact, we assign a small weight to these changed pixels in the loss function. Because the initial prior change map is not accurate and it is updated during training, the value of this weight α can be gradually increased. The loss function $\mathcal{L}_s$ is calculated by

$$\mathcal{L}_s = \sum_{k\in\text{prior}} \alpha\text{CE}\left(\hat{y}_k, y_k^s\right) + \sum_{k\notin\text{prior}} \text{CE}\left(\hat{y}_k, y_k^s\right) \tag{5}$$

where prior represents the pixels who are not in the pasted region and labeled as changed in prior change map. $\text{CE}(\cdot)$ indicates the cross-entropy loss. $\hat{y}_k$ and y_k^s denote the prediction and the synthetic change label of kth pixel.

2) Real Sample Branch: This branch is trained in parallel with the synthetic sample branch to reinforce the knowledge learned from synthetic sample branch. And it shares the same network with synthetic sample branch. We first augment the real samples by utilizing synthetic samples through cut-mix augmentation to reduce domain gap and at the same time obtain more generalized knowledge. Then the sample prior change map y^r is used as pseudo-labels to calculate the cross-entropy loss $\mathcal{L}_r$

$$\mathcal{L}_r = \sum_k \text{CE}\left(\hat{y}_k, y_k^r\right) \tag{6}$$

where y_k^r is pseudo-label of the kth pixel in y^r.

3) Dual-Branch Contrastive Learning: In order to enhance the feature discrimination between changed and unchanged classes, and further reduce the domain gap between synthetic and real samples, we design the dual-branch contrastive learning to the output of encoders.

Specifically, we apply mask-pooling to the encoded features based on the synthetic or pseudo-labels of synthetic or real sample branches, respectively. This process generates two prototype vectors for the changed class: v_c^s (synthetic) and v_c^r (real), as well as two prototype vectors for the unchanged class: v_{uc}^s (synthetic) and v_{uc}^r (real). Then, we utilize cosine similarity $\mathcal{D}(\mathbf{a}, \mathbf{b}) = -|\mathbf{a}\cdot\mathbf{b}/\|\mathbf{a}\|\|\mathbf{b}\||$ as the distance measurement to establish the following four distance relationships:

$$d_{s_{\text{dissim}}} = \mathcal{D}\left(v_{uc}^s, v_c^s\right), \quad d_{r_{\text{dissim}}} = \mathcal{D}\left(v_{uc}^r, v_c^r\right) \tag{7}$$

$$d_{uc_{\text{sim}}} = -\mathcal{D}\left(v_{uc}^s, v_{uc}^r\right), \quad d_{c_{\text{sim}}} = -\mathcal{D}\left(v_c^s, v_c^r\right) \tag{8}$$

where $d_{s_{\text{dissim}}}$ and $d_{r_{\text{dissim}}}$ denote the dissimilarity of changed and unchanged features in synthetic sample learning and real sample learning. $d_{uc_{\text{sim}}}$ represents the similarity of unchanged features, while $d_{c_{\text{sim}}}$ denotes the similarity of changed features. By minimizing (7), the model can have high discrimination between changed and unchanged regions. Equations in (8) help to reduce the domain gap between synthetic and real samples, constraining the network to extract domain-invariant features.

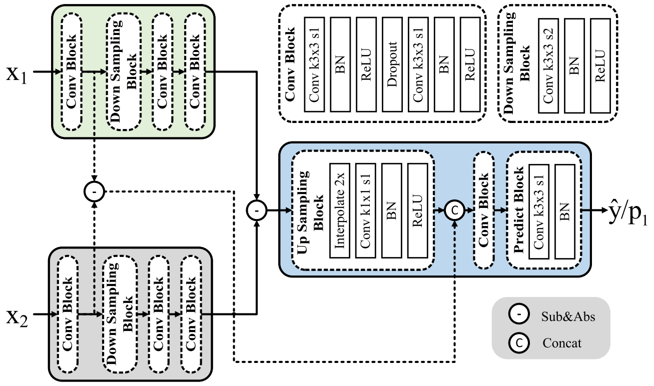


Fig. 3. Details of the CD network.

We keep these prototype vectors for the changed and unchanged classes, and update them by taking the mean values of the new feature vector acquired from each iteration.

4) Total Training Process: The synthetic and real sample branches are trained simultaneously. At the beginning, the model trained on the synthetic samples may be unreliable. In order to prevent the misleading of training on real samples, we set a weight β for the real sample branch training, and the weight is increased during training. The total training loss is

$$\begin{aligned}\mathcal{L} &= \beta\mathcal{L}_r' + \mathcal{L}_s + d_{s_{\text{dissim}}}\\ \mathcal{L}_r' &= \mathcal{L}_r + d_{r_{\text{dissim}}} + d_{uc_{\text{sim}}} + d_{c_{\text{sim}}}.\end{aligned} \tag{9}$$

After several training epochs, the bitemporal images are input into CD model to output the change probability map p_1. This change probability map is binarized to obtain a new change map, which is then used to update the prior change map.

D. Model Architecture and Inference Process

As shown in Fig. 3, our CD network is on the basis of a dual-branch UNet, which contains two encoders and a decoder. The encoders extract discriminative features of bitemporal images and the decoder predicts change map from difference features.

During inference, the CD network directly takes bitemporal images as inputs to obtain change probability map p_1. At the same time, the changed and unchanged prototype vector are maintained during training. We preserve these prototypes, and further calculate the distance between them and the encoder's output, to obtain changed map p_2 and unchanged map p_3. Finally, p_1, p_2, and p_3 are fused to obtain the final probability map p

$$p = \lambda_1 p_1 + \lambda_2 p_2 + \left(1 - \lambda_1 - \lambda_2\right)(1 - p_3). \tag{10}$$

p is then binarized with threshold 0.5 to obtain the change map.

IV. Experiments

In this section, we first clarify the experimental settings, then test the proposed method on five public datasets, and compare it with ten state-of-the-art (SOTA) methods, i.e., FPMS [56], NACCL [57], INLPG [57], SCASC [58], IRG-Mcs [41], SCCN [14], cGAN [35], CAAE [38], PMA [18],

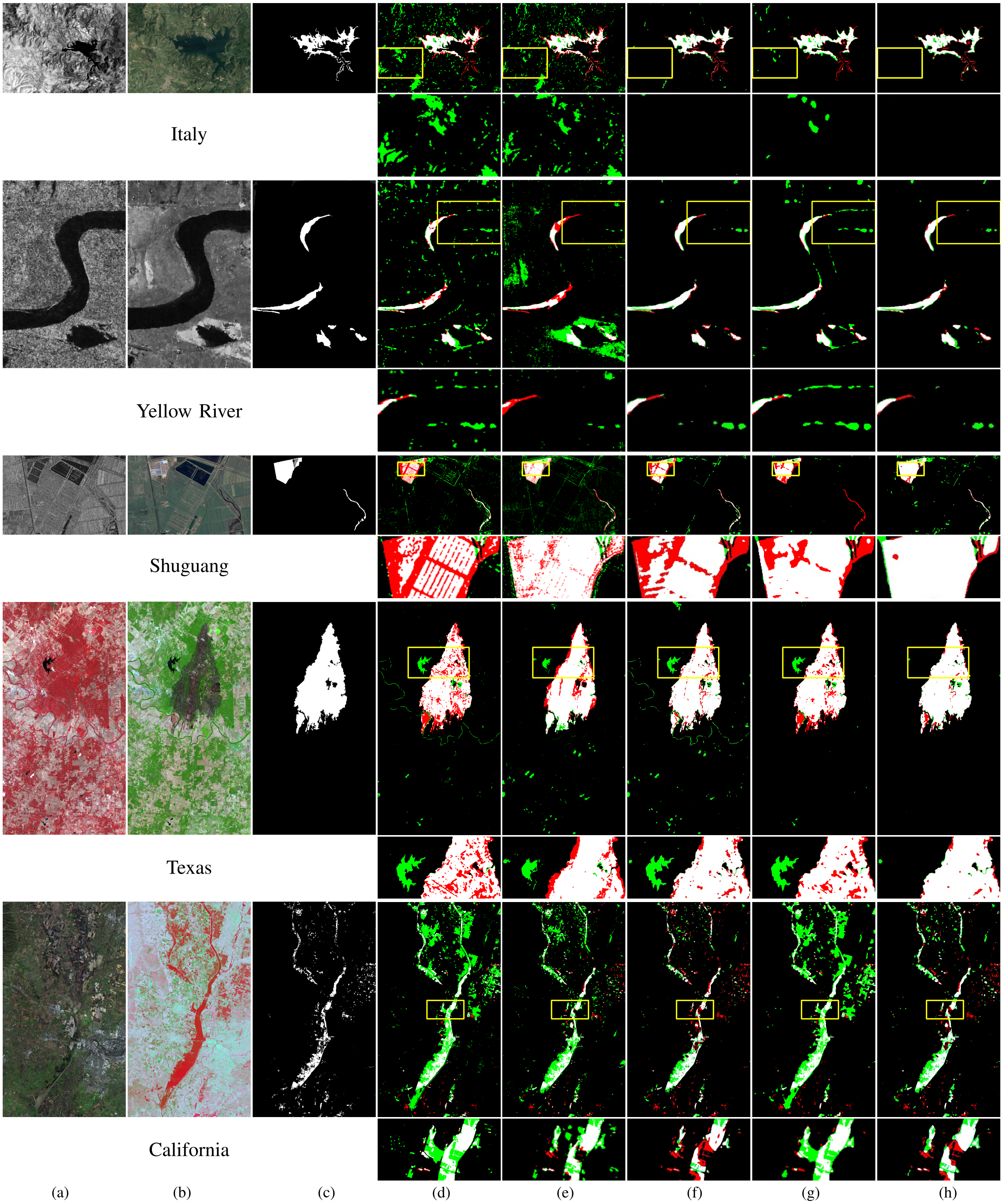


Fig. 4. CD results of different methods on five datasets. In the confusion map, TP, TN, FP, and FN are represented in white, black, green, and red colors, respectively. (a) Pre-event. (b) Post-event. (c) GT. (d) SCCN. (e) CAAE. (f) PMA. (g) PRBCD. (h) Ours.

and PRBCD [32]. In addition, optimal hyperparameters are selected by further experiments.

A. Datasets and Evaluation Metrics

In the experiment, we use five heterogeneous datasets to verify the effectiveness of our method, including optical RGB, near-infrared, multispectral, and SAR images. The pre-event, post-event, and corresponding ground truth of each dataset are shown in Fig. 4(a)–(c). The details of datasets are described as follows.

1) Italy dataset consists of a near-infrared image and an RGB image, which are taken in Sardinia, Italy, where a lake flooding event occurred. The images have the size of 412 × 300.

Algorithm 1 Model Training Process

```
Input: Bi-temporal images $X_1$, $X_2$, total number of training epochs $N$
Output: Change map $CM$
Generate the classification map $cls$ from $X_2$.
Construct $BankofPieces$ based on $cls$.
while $current\_epoch \leq N$ do
    Divide $(X_1, X_2, Y)$ into patches $(x_1, x_2, y)_i$.
    Construct synthetic sample set $(x_1^s, x_2^s, y^s)_i$.
    Construct real sample set $(x_1^r, x_2^r, y^r)_i$.
    Training with Eq. (9).
    Deduce $p_1$ and update CD map.
end while
Calculate distance map $p_2$ and $p_3$.
Fuse $p_1$, $p_2$, $p_3$ and get final change map $CM$.
return Change map $CM$
```

2) Yellow River dataset consists of a SAR image and an optical image. They are captured in Yellow River, China, and the image size is 291 × 444.
3) Shuguang dataset consists of a SAR image and an RGB image, captured in Shuguang village, DongYing city, China. It has the farmland changes to buildings. Images have the size of 921 × 594.
4) Texas dataset consists of two multispectral images from different sensors. They are captured in Texas, America, where a forest fire occurred. The size of images is 808 × 1534.
5) California dataset consists of a multispectral image and a SAR image. A river flooding occurred during the time intervals. The size of images is 2000 × 3500, which is resized to 1000 × 1750 at training.

We use area under the ROC curve (AUC), $F1$-score, overall accuracy (OA), and kappa coefficient (κ) to evaluate the performance of different methods. In the confusion matrix, true positive (TP) denotes positive samples that are also detected as positive, and true negative (TN) denotes those negative samples detected as negative. False positives (FPs) are negative samples that are detected as positive, and false negatives (FNs) are those positive samples detected as negative. In the confusion map of Fig. 4, TP, TN, FP, and FN are represented in white, black, green, and red colors, respectively.

B. Experimental Settings

In the process of generating classification map, the number of clustered class C is set to 5. n is used to select the number of classes farthest from the pasted area, which is set by experiment to 3. When constructing a synthetic sample, the threshold T_{pasted} for the pasted area is set to increase linearly from 0.80 to 0.85. The sizes of three search windows for constructing pasting pieces are $S = \text{patch\_size}/8$, $M = \text{patch\_size}/4$, and $L = \text{patch\_size}/2$, where patch_size is the patch size of datasets. Since different pairs of images have different sizes, we use different patch size in our experiment. Specifically, for Italy, Yellow River, Shuguang, Texas, and California datasets, the patch size is set to 64, 48, 64, 128, and 128, respectively. The balanced weights in loss function are $\alpha = 0.2 * e/e_{\max}$ and $\beta = e/e_{\max}$, where e indicates current epoch and $e_{\max}$ is the total number of epochs.

During training, the batch size is set to 16 and the learning rate is 2×10^{-3}. We use SGD optimizer to update the parameters. All experiments are conducted on the NVIDIA GeForce GTX 1080Ti GPU, and the PyTorch framework is used to construct our model. The prior change map is zero-initialized, and is updated every five epochs. Our model is trained totally 100 epochs.

C. Experimental Results

From Table I, we can observe that our proposed method obtains superior results on most datasets, except for California dataset. Since the ground objects in California dataset are very complex, there are many small and scattered change regions, which leads to a lot of missing detection and our method only achieves the second-best accuracy. It should be noted that our proposed method shows impressive performance on Texas dataset, which has been universally acknowledged to be a "difficult" dataset [18]. Due to the guidance of synthetic samples and their "labels," our method outperforms the comparison methods by a large margin in the AUC, $F1$-score, OA, and kappa coefficient indices. The standard deviations (s.t.d) of our method are listed in the last row of Table I.

We also provide detection results of several deep learning methods in Fig. 4. It can be clearly observed that the results of our method are significantly superior than those of other methods. Some of them are prone to small false detections, presented as green spots in the confusion map. The typical methods like SCCN and PMA, whose feature alignment is achieved in pixel-level, cannot pay much attention to region-level information extraction due to the absence of reliable changed label-supervision. In addition, those image translation-based methods, such as CAAE, apply change comparisons also in the pixel level without considerations of the regional correlation. In contrast, our method performs much better in the nonuniform regions, which we believe is due to the fact that we consider the consistency within the region when pasting the changed regions, allowing for the tolerance of local disunity. The model trained on such synthetic data can focus on a larger range of features to avoid the small false detections. In particular, on the Shuguang dataset, other methods show discontinuities in changed regions, while ours can recognize the whole region as changed, achieving 10% better than other methods in the kappa coefficient index.

D. Hyperparameter Analysis

1) Threshold T_{pasted} in Selection of Pasted Regions: T_{pasted} is a very important hyperparameter that largely determines the rationality and accuracy of synthetic samples. In the experiment, we set it within an interval, and it is increased dynamically along with training process. In order to select an appropriate value, we fix the interval range as 0.05, and make the interval start from 0.7.

The results are shown in Fig. 5. As can be seen, the accuracy of all datasets begins to decline after [0.85, 0.90],

TABLE I

METRICS OF DIFFERENT METHODS

Methods	Italy				Yellow River				Shuguang				Texas				California			
	AUC	F1	OA	κ	AUC	F1	OA	κ	AUC	F1	OA	κ	AUC	F1	OA	κ	AUC	F1	OA	κ
FPMS	0.9138	0.6109	0.9398	0.5798	0.9221	0.5000	0.9763	0.4897	0.9938	0.5711	0.9317	0.5412	0.2517	0.1017	0.8960	0.0850	0.9132	0.2917	0.8540	0.2464
NACCL	-	0.1218	0.8230	0.0396	-	0.5778	0.9685	0.5616	-	0.4984	0.9444	0.4700	-	0.2490	0.9004	0.2154	-	0.3575	0.8879	0.3182
INLPG	0.9238	0.4081	0.9136	0.3471	0.9795	0.5563	0.9579	0.5363	0.9827	0.7099	0.9706	0.6945	0.9569	0.0941	0.8969	0.0813	0.9227	0.3886	0.8971	0.3477
SCASC	0.6876	0.2406	0.8953	0.1850	0.9272	0.4879	0.9435	0.4631	0.9629	0.9733	0.6587	0.6543	0.9594	0.9289	0.5917	0.5548	0.9136	0.4671	0.9466	0.4408
IRG-Mcs	0.8927	0.5109	0.9158	0.4688	0.9030	0.3500	0.9011	0.3147	0.9739	0.6755	0.9654	0.6576	0.9462	0.1767	0.8986	0.1521	0.9232	0.4333	0.9266	0.4010
SCCN	0.9186	0.5385	0.9231	0.5034	0.9186	0.5873	0.9231	0.5691	0.9163	0.4838	0.9445	0.4551	0.9604	0.8127	0.9621	0.7927	0.9532	0.4304	0.9110	0.3956
cGAN	0.9050	0.4830	0.8966	0.4299	0.9100	0.3849	0.9231	0.3541	0.8323	0.3007	0.8920	0.2560	0.9107	0.5716	0.9094	0.5194	0.8305	0.1485	0.8359	0.2749
CAAE	0.9119	0.5537	0.8359	0.5188	0.9210	0.3861	0.9278	0.3530	0.9655	0.5686	0.9425	0.5425	0.9903	0.8777	0.9748	0.8641	**0.9471**	0.5254	0.9360	0.5585
PMA	0.9711	0.8136	0.9777	0.8016	0.9920	0.8156	0.9877	0.8110	0.9720	0.6736	0.9708	0.6552	0.9923	0.8984	0.9777	0.8838	0.9443	0.5964	**0.9703**	**0.6023**
PRBCD	0.9444	0.7786	0.9716	0.7634	0.9903	0.7319	0.9785	0.7211	0.8978	0.7460	0.9801	0.7360	0.9856	0.8817	0.9764	0.8687	0.9420	0.4238	0.9123	0.3888
Proposed	**0.9771**	**0.8262**	**0.9799**	**0.8193**	**0.9921**	**0.8184**	**0.9878**	**0.8120**	**0.9837**	**0.7718**	**0.9768**	**0.7597**	**0.9977**	**0.9567**	**0.9908**	**0.9519**	0.9382	**0.5987**	0.9633	0.5815
s.t.d	0.0107	0.0079	0.0007	0.0082	0.0083	0.0041	0.0041	0.0042	0.0118	0.0721	0.0139	0.0779	0.0113	0.0079	0.0021	0.0088	0.0042	0.0030	0.0006	0.0033

TABLE II

DETECTION ACCURACY (KAPPA COEFFICIENT) UNDER DIFFERENT n

n	Yellow River	Shuguang	Texas
1	0.4795	0.7115	0.2085
2	0.8117	0.5205	0.2117
3	0.8120	0.7597	0.9519
4	0.7950	0.6778	0.9385

TABLE III

TIME COST (SECONDS) OF DIFFERENT DEEP LEARNING-BASED METHODS

Methods	Italy	Yellow River	Shuguang	Texas	California	Average
SCCN	27.13	24.19	72.06	176.27	243.22	108.57
cGAN	133.30	114.66	283.5	546.33	700.63	355.70
CAAE	126.21	135.45	165.63	283.70	354.59	213.12
PMA	45.46	46.29	82.36	293.11	100.60	113.56
PRBCD	460.67	548.69	2024.09	4477.19	1518.02	1805.73
Proposed	113.05	31.52	120.58	245.70	153.38	132.85

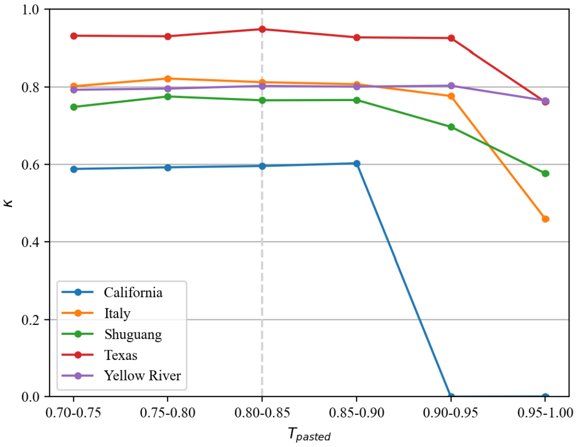


Fig. 5. Investigation of hyperparameter T_{pasted}.

and relatively better performances are obtained in [0.80, 0.85]. Therefore, we finally select the range of T_{pasted} as [0.80, 0.85]. The nearly stable curves also illustrate that the model is insensitive to this parameter.

2) Number of Selected Categories in Construction of Synthetic Samples (n): The hyperparameter n controls the number of classes that can be selected in the construction of synthetic samples. Since the number of classes in the K-means clustering is $C = 5$, here we investigate optimal value of n by setting $n = \{1, 2, 3, 4\}$ respectively. The results are shown in Table II. It can be seen that when $n = 1$, the detection accuracy is generally very low. We think this is due to the limited combination types of the changed regions. With the increase of n, the combination types are more diverse and the detection accuracy is improved and reaches a peak until $n = 3$. Therefore, we set $n = 3$ in our experiment.

3) Fusion Weights in Final Probability Map: The fusion weights λ_1 and λ_2 in (10) are also important hyperparameters. We explore the influence of them through grid search, where λ_1 and λ_2 change with a step size of 0.05. The experimental results are shown in Fig. 6. It can be observed that the detection accuracy tends to stable when λ_1 is above 0.6 with a relatively small λ_2. According to Fig. 6, we finally set λ_1 to 0.7 and λ_2 to 0.2 for all datasets. In fact, if we carefully adjust them on each of the dataset, a slightly higher precision can be obtained. However, in order to confirm the universality of our model, we use the same hyperparameter setting for all datasets.

E. Time Cost Comparison

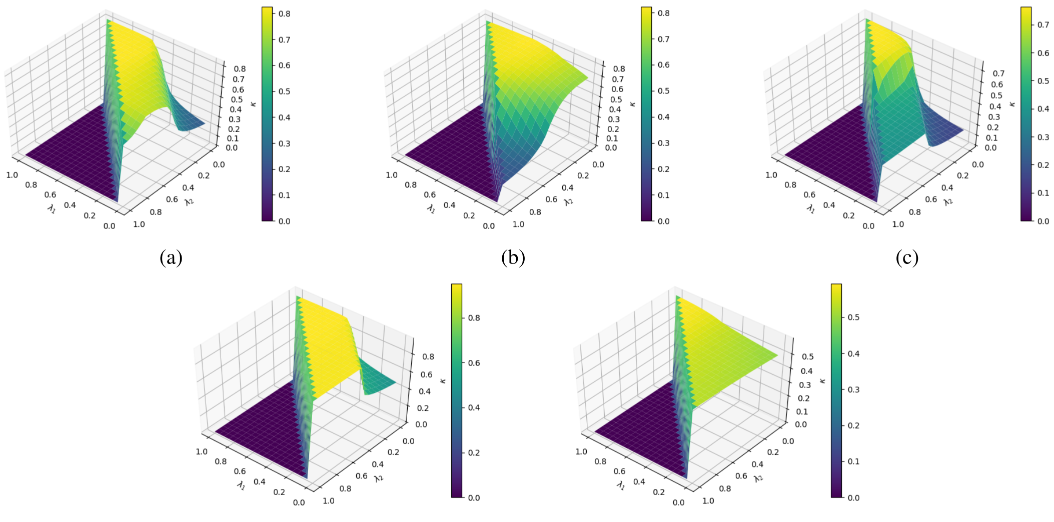
We count the time costs of different deep learning-based methods. Since the processing time is related to the size of images of each dataset, we also compute the average value of them. The results are provided in Table III. In order to make the comparison as fair as possible, we unify the total number of training epochs to 50, which ensures all methods to be fully trained. Our method only obtains the third-best in the time cost comparison, since the construction of synthetic sample requires a lot of serial computations on the CPU. Nonetheless, our proposed method also has a relative acceptable time consumption.

F. Ablation Study

In this section, we conduct experiments to verify the effectiveness of different components of our method. The first

TABLE IV
ABLATION STUDY

	Modules				Datasets				
	Real Branch	Cut-Mix	Feature Contrast	Fusion Prototype	Italy	Yellow River	Shuguang	Texas	California
1					0.6604	0.7976	0.6011	0.9321	0.4466
2	✓				0.6546	0.7973	0.5714	0.7407	0.4246
3	✓	✓			0.6628	0.7940	0.6720	0.9400	0.4651
4	✓		✓		0.7970	0.7815	0.7649	0.9278	0.5827
5	✓	✓	✓		0.8147	0.8029	**0.7657**	0.9485	**0.5957**
6	✓	✓	✓	✓	**0.8193**	**0.8120**	0.7597	**0.9519**	0.5815



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

Fig. 6. Investigation of fusion weights λ_1 and λ_2. (a) Italy. (b) Yellow River. (c) Shuguang. (d) Texas. (e) California.

row in Table IV shows the *baseline*, where the model is trained with only synthetic samples. Compared with previous comparison methods in Table I, *baseline* already has satisfactory results in almost all the datasets, which demonstrates the superiority of our proposed synthetic sample learning paradigm. By comparing the first and the second rows of Table IV, we can find that the performance decreases after adding the real branch, which we think is disturbed by inaccurate knowledge learned in the early training. Therefore, some auxiliary measurements are needed to help the model learn more generalized knowledge. We introduce cut-mix and find an overall performance improvement especially on the Texas dataset. We further add the feature contrastive constraint, and an increase of accuracy appears, demonstrating the effectiveness of feature contrastive learning for all the datasets. The fusion of prototype can boost the results of our method on Itay, Yellow River, and Texas datasets, but seems to be useless on Shuguang and California datasets, since we think this phenomenon appears due to that single changed or unchanged prototype cannot well model the complex mapping relation in some cases, leading to inaccurate CDs based on prototype distance map.

V. CONCLUSION

This article presents a novel unsupervised heterogeneous CD paradigm based on sample synthesis and knowledge transfer. To improve the reliability of CD models, we propose to simulate the real changes by artificially constructing a synthetic sample set. These synthetic samples are then used to guide model training, making the model to learn changes and suppress interference of other factors. In parallel to the synthetic sample branch, a real sample branch is preserved to improve the model discrimination capability. Through the dual-branch joint training with feature contrastive learning, our model obtains superior performance compared with other SOTA methods. In this article, we attempt to provide a possible solution to unsupervised heterogeneous CD on sample synthesis. The performance gains are obtained even through a naive generation manner. In the future, we will explore more advanced image generation methods based on single-temporal images to release the label-dependency and paired-images-dependency dilemmas in heterogeneous CD.

REFERENCES

[1] J. Wang, W. Li, M. Zhang, R. Tao, and J. Chanussot, “Remote-sensing scene classification via multistage self-guided separation network,” *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5615312.

[2] B. Yuan et al., “Spatiotemporal change detection of ecological quality and the associated affecting factors in Dongting Lake basin, based on RSEI,” *J. Cleaner Prod.*, vol. 302, Jun. 2021, Art. no. 126995.

[3] Z. Lv, H. Huang, W. Sun, M. Jia, J. A. Benediktsson, and F. Chen, “Iterative training sample augmentation for enhancing land cover change detection performance with deep learning neural network,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, pp. 1–14, 2023, doi: 10.1109/TNNLS.2023.3282935.

[4] J. Bolorinos, N. K. Ajami, and R. Rajagopal, "Consumption change detection for urban planning: Monitoring and segmenting water customers during drought," *Water Resour. Res.*, vol. 56, no. 3, Mar. 2020, Art. no. e2019WR025812.

[5] D. Brunner, G. Lemoine, and L. Bruzzone, "Earthquake damage assessment of buildings using VHR optical and SAR imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2403–2420, May 2010.

[6] 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 Trans. Geosci. Remote Sens.*, vol. 48, no. 2, pp. 882–894, Feb. 2010.

[7] J. Zhang, "Multi-source remote sensing data fusion: Status and trends," *Int. J. Image Data Fusion*, vol. 1, no. 1, pp. 5–24, Mar. 2010.

[8] B. Adriano et al., "Learning from multimodal and multitemporal Earth observation data for building damage mapping," *ISPRS J. Photogramm. Remote Sens.*, vol. 175, pp. 132–143, May 2021.

[9] J. Wang, W. Li, Y. Gao, M. Zhang, R. Tao, and Q. Du, "Hyperspectral and SAR image classification via multiscale interactive fusion network," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 10823–10837, Jun. 2023.

[10] Z. Lv et al., "Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective," *Proc. IEEE*, vol. 110, no. 12, pp. 1976–1991, Dec. 2022.

[11] R. Shao, C. Du, H. Chen, and J. Li, "SUNet: Change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network," *Remote Sens.*, vol. 13, no. 18, p. 3750, Sep. 2021.

[12] Z. Lv, H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi, "Simple multiscale UNet for change detection with heterogeneous remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[13] F.-Q. Liu and Z.-Y. Wang, "Automatic 'ground truth' annotation and industrial workpiece dataset generation for deep learning," *Int. J. Autom. Comput.*, vol. 17, pp. 539–550, Jan. 2020.

[14] 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 Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[15] 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 Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6402–6416, Nov. 2022.

[16] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based Markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[17] Z. Liu, G. Li, G. Mercier, Y. He, and Q. Pan, "Change detection in heterogenous remote sensing images via homogeneous pixel transformation," *IEEE Trans. Image Process.*, vol. 27, no. 4, pp. 1822–1834, Apr. 2018.

[18] Y. Xing, Q. Zhang, L. Ran, X. Zhang, H. Yin, and Y. Zhang, "Progressive modality-alignment for unsupervised heterogeneous change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5614612.

[19] T. Zhan, M. Gong, J. Liu, and P. Zhang, "Iterative feature mapping network for detecting multiple changes in multi-source remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 146, pp. 38–51, Dec. 2018.

[20] C. Han, C. Wu, H. Guo, M. Hu, J. Li, and H. Chen, "Change guiding network: Incorporating change prior to guide change detection in remote sensing imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 8395–8407, 2023.

[21] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, 2023.

[22] Z. Zheng, A. Ma, L. Zhang, and Y. Zhong, "Change is everywhere: Single-temporal supervised object change detection in remote sensing imagery," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15193–15202.

[23] C. Sun, H. Chen, C. Du, and N. Jing, "SemiBuildingChange: A semi-supervised high-resolution remote sensing image building change detection method with a pseudo bitemporal data generator," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5622319.

[24] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5603216.

[25] M. Xu, S. Yoon, A. Fuentes, and D. S. Park, "A comprehensive survey of image augmentation techniques for deep learning," *Pattern Recognit.*, vol. 137, May 2023, Art. no. 109347.

[26] G. Mercier, G. Moser, and S. B. Serpico, "Conditional copulas for change detection in heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 5, pp. 1428–1441, May 2008.

[27] J. Prendes, M. Chabert, F. Pascal, A. Giros, and J.-Y. Tourneret, "A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 799–812, Mar. 2015.

[28] Y. Bazi, F. Melgani, and H. D. Al-Sharari, "Unsupervised change detection in multispectral remotely sensed imagery with level set methods," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 8, pp. 3178–3187, Aug. 2010.

[29] R. Touati and M. Mignotte, "An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1046–1058, Feb. 2018.

[30] Z.-G. Liu, G. Mercier, J. Dezert, and Q. Pan, "Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning," *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 1, pp. 168–172, Jan. 2014.

[31] H. Li, M. Gong, M. Zhang, and Y. Wu, "Spatially self-paced convolutional networks for change detection in heterogeneous images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4966–4979, 2021.

[32] 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. 5620717.

[33] J. Shi, T. Wu, A. K. Qin, Y. Lei, and G. Jeon, "Self-guided autoencoders for unsupervised change detection in heterogeneous remote sensing images," *IEEE Trans. Artif. Intell.*, early access, pp. 1–13, 2024, doi: 10.1109/TAI.2024.3357667.

[34] L. Wei, G. Chen, Q. Zhou, C. Liu, and C. Cai, "Cross-mapping net: Unsupervised change detection from heterogeneous remote sensing images using a transformer network," in *Proc. 8th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Apr. 2023, pp. 1021–1026.

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

[36] M. Gong, X. Niu, T. Zhan, and M. Zhang, "A coupling translation network for change detection in heterogeneous images," *Int. J. Remote Sens.*, vol. 40, no. 9, pp. 3647–3672, 2019.

[37] M. Jia, C. Zhang, Z. Lv, Z. Zhao, and L. Wang, "Bipartite adversarial autoencoders with structural self-similarity for unsupervised heterogeneous remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[38] L. T. Luppino et al., "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 60–72, Jan. 2024.

[39] Y. Sun, L. Lei, X. Li, H. Sun, and G. Kuang, "Nonlocal patch similarity based heterogeneous remote sensing change detection," *Pattern Recognit.*, vol. 109, Jan. 2021, Art. no. 107598.

[40] 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 Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4700221.

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

[42] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," 2012, *arXiv:1206.5538*.

[43] S. A. Assefa, D. Dervovic, M. Mahfouz, R. E. Tillman, P. Reddy, and M. Veloso, "Generating synthetic data in finance: Opportunities, challenges and pitfalls," in *Proc. 1st ACM Int. Conf. AI Finance*, New York, NY, USA, Oct. 2020, pp. 1–8.

[44] G. Ros, L. Sellart, J. Materzynska, D. Vazquez, and A. M. Lopez, "The SYNTHIA dataset: A large collection of synthetic images for semantic segmentation of urban scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3234–3243.

[45] G. Varol et al., "Learning from synthetic humans," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4627–4635.

[46] A. Sobral and A. Vacavant, "A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos," *Comput. Vis. Image Understand.*, vol. 122, pp. 4–21, May 2014.

[47] B. Yu, Y. Wang, L. Wang, D. Shen, and L. Zhou, "Medical image synthesis via deep learning," in *Deep Learning in Medical Image Analysis: Challenges and Applications*. Cham, Switzerland: Springer, 2020, pp. 23–44.

[48] D. A. B. Oliveira, "Controllable skin lesion synthesis using texture patches, Bézier curves and conditional GANs," in *Proc. IEEE 17th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2020, pp. 1798–1802.

[49] N. Dvornik, J. Mairal, and C. Schmid, "Modeling visual context is key to augmenting object detection datasets," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*. Berlin, Germany: Springer, 2018, pp. 375–391.

[50] J. Rao and J. Zhang, "Cut and paste: Generate artificial labels for object detection," in *Proc. Int. Conf. Video Image Process.* New York, NY, USA: Association for Computing Machinery, Dec. 2017, pp. 29–33.

[51] C. Kwan, B. Ayhan, J. Larkin, L. Kwan, S. Bernabé, and A. Plaza, "Performance of change detection algorithms using heterogeneous images and extended multi-attribute profiles (EMAPs)," *Remote Sens.*, vol. 11, no. 20, p. 2377, Oct. 2019.

[52] F. Wang, H. Wang, C. Wei, A. Yuille, and W. Shen, "CP^2: Copy-paste contrastive pretraining for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Berlin, Germany: Springer-Verlag, 2022, pp. 499–515.

[53] D. Dwibedi, I. Misra, and M. Hebert, "Cut, paste and learn: Surprisingly easy synthesis for instance detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1310–1319.

[54] C.-L. Li, K. Sohn, J. Yoon, and T. Pfister, "CutPaste: Self-supervised learning for anomaly detection and localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 9664–9674.

[55] M. Seo, H. Lee, Y. Jeon, and J. Seo, "Self-pair: Synthesizing changes from single source for object change detection in remote sensing imagery," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 6363–6372.

[56] M. Mignotte, "A fractal projection and Markovian segmentation-based approach for multimodal change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 11, pp. 8046–8058, Nov. 2020.

[57] M. Mignotte, "MRF models based on a neighborhood adaptive class conditional likelihood for multimodal change detection," *AI, Comput. Sci. Robot. Technol.*, vol. 2022, pp. 1–20, Mar. 2022.

[58] Y. Sun, L. Lei, D. Guan, M. Li, and G. Kuang, "Sparse-constrained adaptive structure consistency-based unsupervised image regression for heterogeneous remote-sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4405814.

Yinghui Xing (Member, IEEE) received the B.S. and Ph.D. degrees from the School of Artificial Intelligence, Xidian University, Xi'an, China, in 2014 and 2020, respectively.

She is an Associate Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an. Her research interests include remote sensing image processing, image fusion, and image super-resolution.

Qi Zhang received the B.S. degree from the Northwestern Polytechnical University, Xi'an, China, in 2021, where he is currently pursuing the M.S. degree with the School of Computer Science.

His research interests include remote sensing image object detection and change detection.

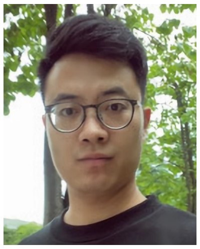
Lingyan Ran received the B.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 2011 and 2018, respectively.

From 2013 to 2015, he was a Visiting Scholar with the Stevens Institute of Technology, Hoboken, NJ, USA. His research interests include image classification and semantic segmentation.

Dr. Ran is currently a member of CSIG.

Xiuwei Zhang received the B.S., M.S., and Ph.D. degrees from the School of Computer Science, Northwestern Polytechnical University, Xi'an, China, in 2004, 2007, and 2011, respectively.

She is now an Associate Professor with the School of Computer Science, Northwestern Polytechnical University. Her research interests include remote sensing image processing, multimodel image fusion, image registration, and intelligent forecasting.



Hanlin Yin received the B.S. degree in control science and engineering from Harbin Institute of Technology, Harbin, China, in 2009, and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2016.

He is currently an Assistant Professor with the School of Computer Science, Northwestern Polytechnical University, Xi'an. His research interests include rainfall-runoff modeling, time series prediction, information fusion, and performance evaluation.

Yanning Zhang (Senior Member, IEEE) received the B.S. degree from Dalian University of Science and Engineering, Dalian, China, in 1988, and the M.S. and Ph.D. degrees from Northwestern Polytechnical University, Xi'an, China, in 1993 and 1996, respectively.

She is a Professor with the School of Computer Science, Northwestern Polytechnical University. She has published more than 200 articles in international journals, conferences, and Chinese key journals. Her research interests include signal and image processing, computer vision, and pattern recognition.

Dr. Zhang is also the Organization Chair of the Ninth Asian Conference on Computer Vision (ACCV2009).