CutMix-CD: Advancing Semi-Supervised Change Detection via Mixed Sample Consistency

Qidi Shu, Xiaolin Zhu, *Senior Member, IEEE*, Luoma Wan, Shuheng Zhao, Denghong Liu, Longkang Peng, Xiaobei Chen

***Abstract*—Change detection is an important task in earth observation. In the past few years, significant progress has been made in supervised change detection research. However, change labels are extremely expensive. Semi-supervised change detection has attracted increasing attention. In semi-supervised change detection, the problem of scarcity of positive samples is magnified. Moreover, the imbalance of change types (e.g., disappearance and appearance) exacerbates the missing detection phenomenon. To address the above problems, we propose a semi-supervised change detection method: CutMix-CD, which incorporates the change-aware CutMix augmentation into the consistency framework of change detection. The semi-supervised learning framework enriches change contexts and places special emphasis on the comparative process, facilitating more robust representations of changes with improved generalization capabilities. Firstly, mixed samples are synthesized using the change-aware CutMix operation. Then, we developed a student path and a teacher path to predict the changes of original samples and mixed samples respectively. Finally, the consistency loss is conducted between the two predictions to help the model learn change information of unlabeled samples. In addition, an unsupervised feature constraint loss is proposed to further optimize the change features. Experiments on four datasets validate the effectiveness of CutMix-CD. It can effectively alleviate the overfitting problem for unbalanced types of changes, and even outperforms the fully-supervised methods for some challenging samples. The code will be released in https://github.com/SQD1/CutMixCD.**

***Index Terms*—Change detection, semi-supervised learning, consistency learning, deep learning.**

I. Introduction

WITH the rapid development of remote sensing technology, increasing satellite images are available for earth dynamic observations. Change detection (CD) is one of challenging tasks in remote sensing image understanding, which aims to identify change areas by comparing co-registered images. CD has broad applications in urban planning [1, 2], environment monitoring [3] and damage assessment [4].

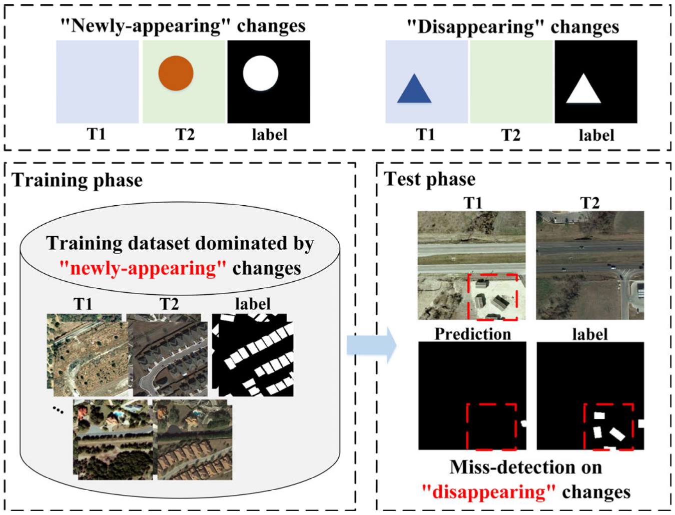

"Newly-appearing" changes
T1
T2
label
"Disappearing" changes
T1
T2
label
Training phase
Training dataset dominated by "newly-appearing" changes
T1
T2
label
Test phase
T1
T2
Prediction
label
Miss-detection on "disappearing" changes

Fig. 1. The overfitting problem in change detection. There are two basic types of changes: newly-appearing changes and disappearing changes (upper). If training samples are dominated by newly-appearing changes (lower left), there will be severe miss-detection for disappearing changes in test phase (lower right).

In recent years, deep learning-based CD methods have made great progress. Convolutional neural network [5, 6] and transformer [7-9] structures are widely used to extract deep features. However, most deep learning-based methods rely heavily on labeled samples. Labeling changes is time-consuming which requires frequent comparisons and exclusion of pseudo-changes. To reduce training costs, increasing research focus on semi-supervised change detection (Semi-supervised CD). Semi-supervised CD utilizes limited change labels and a large number of unlabeled samples to train a CD model, which significantly expands the scope of applications for CD in real-world scenarios.

Currently, semi-supervised CD is still in the developing stage. The key to semi-supervised CD lies in learning the latent change information from unlabeled image pairs. In general, semi-supervised methods can be summarized into three categories [10]: self-training methods [3, 11], adversarial learning-based methods [12-14] and consistency learning-based methods [15-18].

This work was supported by Guangdong Basic and Applied Basic Research Foundation (No. 2022B1515130001) and the Department of Land Surveying and Geo-Informatics at The Hong Kong Polytechnic University (WZ83). *(Corresponding author: Xiaolin Zhu)*

Qidi Shu, Xiaolin Zhu, Luoma Wan, Shuheng Zhao, Denghong Liu, Longkang Peng, Xiaobei Chen are with the Department of Land Surveying and Geo-Informatics, Hong Kong Polytechnic University, Hong Kong (e-mail: qidi17.shu@connect.polyu.hk; xiaolin.zhu@polyu.edu.hk).

Self-training methods transform the unsupervised phase to a supervised mode by assigning pseudo labels to unlabeled samples. In this way, supervised loss can be calculated on unlabeled samples. Normally, a model is first trained with labeled samples. Then, pseudo change labels can be acquired by predicting on unlabeled samples. However, pseudo labels often contain errors. The trained models often have difficulty correcting its own mistakes [19]. Adversarial learning-based methods use a discriminator to distinguish predictions from the ground truth [20]. Adversarial loss allows learning the distribution of unlabeled data, but the training process is usually unstable. Consistency learning requires a model to output consistent results for perturbated versions of inputs [21]. It is based on the idea that similar input data should have similar output. The supervision signals are provided by minimizing the distance between corresponding predictions of unlabeled data. Consistency learning is stable in training and has solid theoretical basis [22], which makes it a popular method in semi-supervised learning.

In semi-supervised CD, improving the generalization capability is a paramount focus of current research. There exist two challenges in this regard. Firstly, the scarcity of positive samples is exacerbated in semi-supervised CD. Limited in quantity and monotonic in type, the positive samples fail to represent the overall distribution of changes, resulting in a lack of representativeness in the extracted change features. Secondly, the imbalance of change types (newly-appearing and disappearing changes) leads to the overfitting to a specific type of change, further aggravating the omission of detections. As shown in Fig. 1, based on the distribution of change objects in the bi-temporal images, change can be classified into two fundamental categories: "newly-appearing" and "disappearing" changes. When the majority of changes in the training samples belong to the newly-appearing category, overfitting tends to occur, resulting in severe miss-detection for disappearing changes. Unfortunately, the existing research on semi-supervised CD exhibits limited focus on the problem caused by imbalanced change types.

To address the above problems, a novel semi-supervised CD framework named CutMix-CD is proposed. The core idea is to learn generalized change features under limited-label condition by embedding the synthesized samples in a consistency learning framework. Specifically, for unlabeled samples, we propose a change-aware CutMix operation to synthesize novel samples with changes. Then, we design two paths (i.e., teacher path and student path) in consistency learning framework to predict the changes of the mixed samples respectively. Finally, consistency loss is conducted between the two predictions to learn the change information in the unlabeled samples.

The proposed CutMix-CD framework alleviates the overfitting problems from two perspectives. Firstly, we introduce synthetic change sample pairs at the data level. Data augmentation is a direct and efficient method in addressing overfitting issues. Synthesized samples have been proved effective in unsupervised CD [23]. Our proposed change-aware CutMix operation pastes the regions of change from one set of sample pairs onto another, thereby increasing the number of positive samples and expanding the context of change objects. This is crucial because the backgrounds of the two sample groups differ. The new context of change objects contributes to the change learning process. Secondly, the training focuses on the comparison process, mitigating the impact of imbalanced change types on the detection of certain types of changes. Although the mixed samples exhibit discontinuity between the masked region and the surrounding objects within a single image, the positions of the mask in bi-temporal images still correspond to each other. Therefore, during the training process, the novel samples enable the model to not be concerned with the continuity of the context within a single image, but rather to prioritize whether the corresponding regions have undergone changes. This allows the model to learn more generalized change features and alleviate overfitting to specific types of changes. In addition, combined with consistency regularization between the teacher path and the student path, change information of unlabeled data are well explored thus further improving the CD performance.

The proposed CutMix-CD framework enriches change contexts and places special emphasis on the comparative process to alleviate the overfitting problem, thus improving the generalization capability of semi-supervised CD. The main contributions are as follows:

(1) A semi-supervised CD framework CutMix-CD is proposed. Consistency learning with synthetic samples largely benefit the learning of change information from unlabeled data.

(2) A change-aware CutMix operation is proposed to synthesize mixed samples with changes, which effectively alleviates the severe overfitting issue in semi-supervised CD.

(3) Experiments on four datasets validate the effectiveness of our method. CutMix-CD shows significant improvement under the unbalanced distribution conditions.

The rest of this paper is organized as follows. Section II presents the related work. In section III, the proposed method is presented in detail. Section IV shows the experimental results. The conclusions are drawn in Section V.

II. Related work

A. Deep learning based supervised CD

Supervised CD based on deep learning has achieved state-of-the-art performance after undergoing a series of developments. Initially, the introduction of Fully Convolutional Networks (FCN) [5] greatly facilitated the progress of pixel-level prediction tasks, resulting in a significant leap in accuracy. Subsequently, FCN-based CD methods rapidly evolved, categorized into early fusion and late fusion approaches based on the position and manner of fusion between two temporal features [24]. Early fusion methods stack the two temporal images during the image input stage, while late fusion methods extract features separately from the two temporal images and fuse them at the feature level. The research focus lies in how to extract discriminative features and efficiently fuse the two temporal features. Regarding feature extraction, approaches such as UNet++ [25], transformer [26] and Mamba [27] have

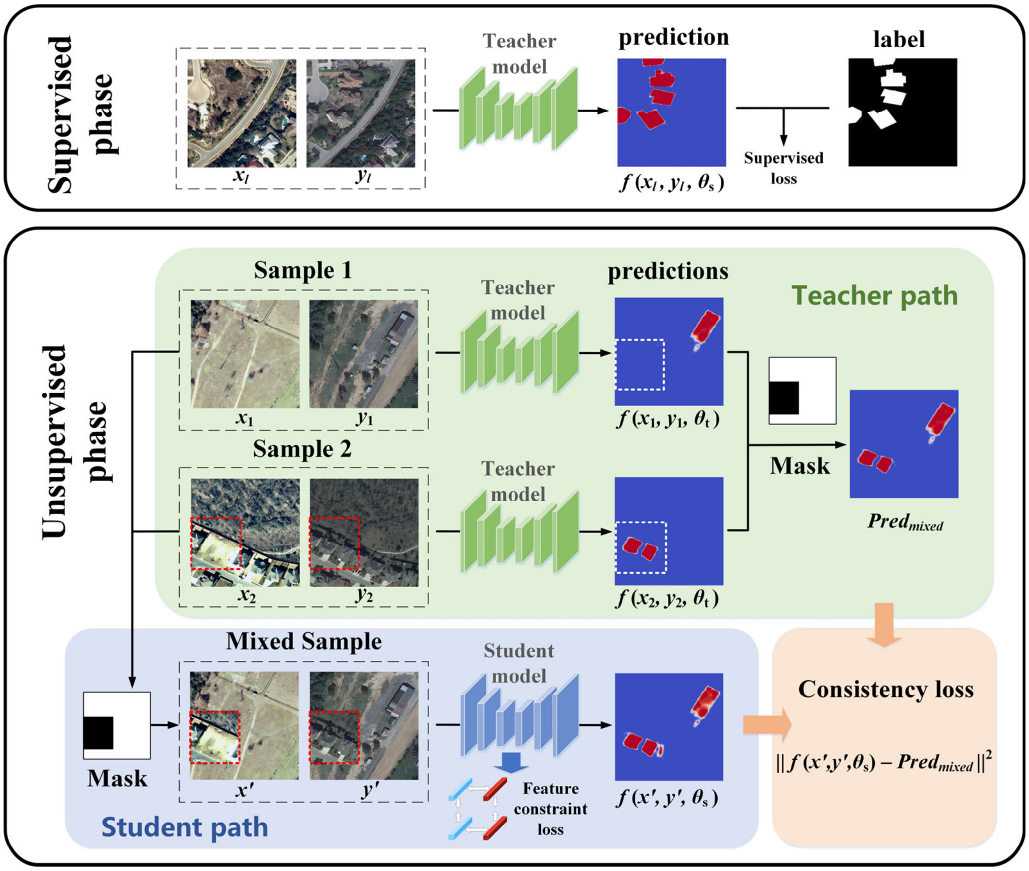


Fig. 2. The framework of CutMix-CD.

been introduced to CD. In terms of temporal feature fusion, methods based on attention mechanisms have achieved impressive results [28]. For example, SAAN [29] proposes a similarity-guided attention flow mechanism to enhance spatial-channel feature representation. DPCCNet [30] proposes a dual-perspective fusion module to extract change features from each temporal phase. However, supervised CD methods heavily rely on annotated change labels [31], which are labor-intensive, thus limiting the practical application in real-world scenarios.

B. Semi-supervised learning

Semi-supervised learning leverages limited labeled data and large proportion of unlabeled data for training, thus reducing the training cost. In semi-supervised learning, the most crucial aspect is how to provide reliable supervision signals for the unlabeled data. Three main branches of methods are proposed, namely self-training methods, adversarial learning-based methods and consistency learning-based methods. Self-training based methods generate pseudo-labels for unlabeled samples [32]. The core is to improve the quality of the pseudo-label. For example, ST++ [33] selects reliable unlabeled data based on holistic prediction-level stability. Adversarial learning-based methods learn the distribution of unlabeled data through adversarial loss. S4GAN [13] uses an extra discriminator to match the distribution of the prediction and the ground truth. Consistency learning-based methods introduce perturbations to the original data and enforce consistency between the outputs of both the perturbed and unperturbed versions. The success of consistency learning framework heavily relies on the design of strong data augmentations. Strong perturbation such as CutMix and CutOut can effectively enhance the consistency learning and alleviate the overfitting in semi-supervised semantic segmentation [21]. Cross-consistency training (CCT) [15] first applies different perturbations on the feature level, which enforces a consistency between the outputs of the main decoder and auxiliary decoders. FixMatch [34] uses the weakly perturbed branch to supervise the strongly perturbed branch for semi-supervised classification. Based on FixMatch, UniMatch [35] further explores a broader perturbation space by an auxiliary stream and achieves state-of-the-art performance on semi-supervised semantic segmentation. Recently, visual language models (VLMs) are introduced to provide additional supervision signals for unlabeled data and achieve promising performance [36, 37].

However, the above perturbations used in single image do not consider the characteristic of change detection task. Different from traditional perturbations (color jitter, flip, random crop) used in consistency learning, we incorporate a specially designed perturbation for unlabeled image pairs into a consistency framework which largely boosts the model learning in semi-supervised CD.

C. Semi-supervised CD

Due to the huge costs of pixel-level annotation of CD task,

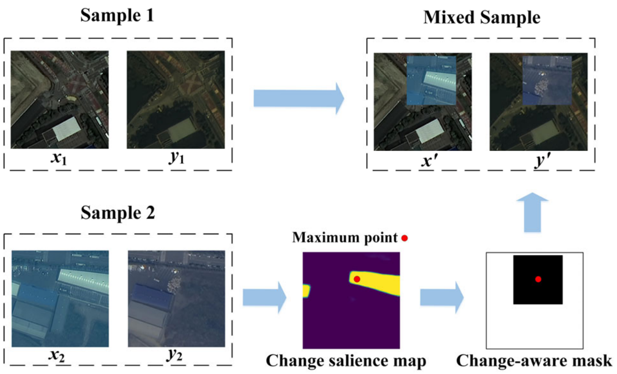


Fig. 3. Change-aware CutMix operation.

efforts have been made in recent years to explore semi-supervised CD. Adversarial learning-based SemiCDNet [12] adopts entropy and segmentation discriminators to distinguish features from labeled and unlabeled samples, using two adversarial losses to learn the distribution of unlabeled data. Reliable contrastive learning (RCL) acquires reliable pseudo labels considering the uncertainty [38]. Consistency-based method FPA [39] applies different data augmentation techniques as perturbations at the input level, imposing consistency constraints on both the feature and pixel-level prediction results. RCR [40] introduces data perturbations at the feature level. TCNet [41] considers the consistency of different input orders. C2F-SemiCD [42] proposes a novel coarse-to-fine network in a Mean teacher framework. Previous work MTCNet [43] proposes task-level consistency learning, utilizing the existing semantic information of the single-temporal phase to derive a new Diff-T1 branch. This approach significantly improves the CD results by promoting consistency between the predictions of the Diff-T1 branch and the original T1 branch.

The aforementioned semi-supervised CD methods have made improvements in the generalization ability with limited labeled data. However, existing methods do not pay enough attention on the overfitting problem caused by imbalanced change types (newly-appearing and disappearing changes), as shown in Fig. 1. To address this issue, this paper proposes a novel semi-supervised CD framework called CutMix-CD, which effectively detects various types of changes under imbalanced conditions while efficiently making advantage of unlabeled data.

III. Methodology

A. Overview

CutMix-CD is a consistency learning framework which incorporates synthesized samples to learn the change information of unlabeled data, thus enhancing the generalization ability of CD. The overall structure of CutMix-CD is shown in Fig. 2.

CutMix-CD consists of two phases: supervised phase and unsupervised phase. In the supervised phase, change labels are available for supervised cross-entropy loss. This process endows the model with a preliminary change detection capability. However, the model is prone to overfitting due to limited labels and poorer sample diversity, resulting in poor generalization performance. In the unsupervised phase, the proposed change-aware CutMix operation is first utilized to synthesize new sample pairs. Then, a student path and a teacher path are designed to detect the changes in the mixed samples respectively. The change information in the unlabeled samples is learned through consistency regularization between the results of two paths. In this way, the model can learn more generalized change features.

B. Change network

In the CutMix-CD framework, the student and teacher model are two change networks with the same structure used for the change prediction. In this paper, we adopt the change detection network ResNet-CD [40]. ResNet-CD adopts the structure of late fusion and consists of three main parts: A weight-shared encoder to extract bi-temporal features, a feature difference module to extract the change features, and a decoder to output the change map. The ImageNet pre-trained ResNet50 [44] is used as the encoder. The feature difference module contains a pyramid pooling module (PPM) [45] to extract multi-scale change features. The decoder consists of convolutional upsampling modules to restore the spatial resolution to the input images.

It should be noted that the focus of this study is on semi-supervised learning for CD, rather than the CD network structure. ResNet-CD is uniformly used as the underlying change network in all experiments to fairly compare with other the semi-supervised CD methods (see in Section IV. D). Other CD networks can also be easily incorporated into the proposed semi-supervised learning framework and have the potential to improve overall CD performance.

C. Change-aware CutMix operation

CutMix [46] is a data augmentation method that is widely applied in various tasks such as object detection, scene classification, and semantic segmentation. As for CD task, the inputs are more complicated which consists of bi-temporal images. Different from normal CutMix operation in classification or semantic segmentation tasks, we adapt CutMix for CD on pair-level. As shown in Fig. 3, Pair-level CutMix synthesizes a new pair of temporal images (Mixed sample) by combining two sets of bi-temporal images (Sample 1 and sample 2). This combination introduces more positive samples and broadens the change contextual information. The synthetic sample pairs are more conducive to the model's understanding of changes.

The augmented images may not necessarily contain target objects due to the random generation of masks. This can adversely affect the accuracy and efficiency of model learning. Therefore, we further propose a change-aware CutMix operation with the aim of maximizing the inclusion of changed objects in mixed samples. Specifically, in contrast to random generation in standard CutMix, the mask in our method is determined based on the change features. This largely increases the possibility that the mask region contains change objects, and subsequently, new samples are synthesized using this mask region.

As shown in Fig 3, The change-aware CutMix operation consists of the following steps. First, the change salience map is obtained by extracting the output of the teacher network. The pixel values in this map reflect the likelihood of changes, with higher values indicating a higher probability of change. As the training iterations progress, the precision of the change salience map also improves. Next, gaussian noise is then added to enhance the robustness of the change salience map. Afterward, the position of the pixel with the maximum value in the change salience map is identified as the center point of the mask.

The aforementioned operation fulfils two requirements. Firstly, it determine the position of the mask, ensuring that it contains the changed objects without the need for precise identification of their locations. Secondly, the operation maintains a low computational load, thus ensuring efficient training.

The generated mask is then utilized to synthesize new pairs of samples. P1 $[x_1, y_1]$ and P2 $[x_2, y_2]$ represent two image pairs, where x and y represent the pre- and post-temporal images respectively. M denotes the change-aware mask. A mixed image pair Pm $[x', y']$ can be obtained using the following formula:

$$x' = x_1 \odot M + x_2 \odot (1 - M) \tag{1}$$

$$y' = y_1 \odot M + y_2 \odot (1 - M) \tag{2}$$

$\odot$ represents element-wise multiplication.

D. CutMix-CD framework

CutMix-CD involves transferring the identical mask regions from one set of image pairs to their corresponding regions in another set of image pairs. This process results in the synthesis of a new set of image pairs.

For individual image (x' or y') within the mixed sample, the masked region is discontinuous with the surrounding image, promoting diversity in the contextual information of the image. Meanwhile, the mixed pair Pm $[x', y']$ still maintains strict geospatial registration. The mixed samples promote the model to focus more on the comparative process between the two temporal phases instead of focusing on objects in either temporal phase, thereby mitigating the problem of overfitting in one temporal phase.

For P1, P2 and Pm, we incorporate them into a consistency learning framework [47] to learn the change information in unlabeled samples. The process is illustrated in Fig. 2.

The teacher model and student model are structurally identical change detection networks. Initially, P1 $[x_1, y_1]$ and P2 $[x_2, y_2]$ are separately fed into the teacher model to get the change probability predictions for these two sets of samples denoted as $f(x_1, y_1, \theta_t)$ and $f(x_2, y_2, \theta_t)$ respectively. Here, θ_t represents the parameters of the teacher model.

Subsequently, the mixed probability predictions, denoted as $Pred_{mixed}$, can be obtained by employing the mask M based on the following equation.

$$Pred_{mixed} = f(x_1, y_1, \theta_t) \odot M + f(x_2, y_2, \theta_t) \odot (1 - M) \tag{3}$$

Next, the mixed image pair Pm $[x', y']$ is directly fed into the student model, resulting in the predicted change probability for the mixed sample, denoted as $f(x', y', \theta_s)$, where θ_s represents the parameters of the student model.

Finally, based on consistency regularization, the prediction $f(x', y', \theta_s)$ of the mixed image pair $[x', y']$ should converge towards a combination of the predictions $Pred_{mixed}$ from two original samples. Therefore, the consistency loss L_c is calculated using $f(x', y', \theta_s)$ and $Pred_{mixed}$ as follows:

$$L_c = D(f(x', y', \theta_s), Pred_{mixed}) \tag{4}$$

$D(.)$ Represents a distance function. We use mean squared error (MSE) in consistency loss.

Additionally, to make full use of the diverse samples generated by change-aware CutMix operation, we further propose a feature constraint loss for unlabeled samples. Inspired by class-aware feature alignment in FPA [39], the idea lies in that the change (or unchanged) features should be aligned in different samples. Conversely, the difference between changed and unchanged features should be enlarged. This loss can take advantage of diverse synthetic samples to better optimize change features.

Firstly, we need to obtain the change/no-change feature vector. Following FPA, we use the output of the feature difference module from the change network as the initial change features $F \in \mathbb{R}^{(H/8)\times(W/8)\times C}$. C represents the channel number. Using mixed probability predictions from the teacher model, we derive the change/no-change map and resize it to match the spatial dimensions of F, resulting in $m_c \in \mathbb{R}^{(H/8)\times(W/8)\times 2}$. By applying the change map m_c as weights, we perform a weighted sum over the spatial dimension of F to obtain the change/no-change feature vector $v \in \mathbb{R}^{2\times C}$. $v(0)$ and $v(1)$ deonte the no-change and change vectors, respectively.

Based on the change/no-change feature vector $v \in \mathbb{R}^{2\times C}$, the feature constraint loss consists of two components (Eq. 5). The first component aligns the change feature vectors between different sample pairs within a batch, while the second component pushes apart the change and no-change feature vectors within a sample pair.

$$L_f = \frac{1}{BB} \sum_{i=1}^{B} \sum_{j=1}^{B} D(v_i, v_j) - \frac{1}{B} \sum_{k=1}^{B} D(v_k(0), v_k(1)) \tag{5}$$

Similarly, $D(.)$ Represents a distance function. Cosine similarity function is used in Eq. (5) to measure the distance between two vectors.

In the training process of CutMix-CD, different gradient updating strategies are employed for the teacher model and the student model. First, in the supervised phase, the teacher model is trained using change labels. The backward propagation is performed by the supervised cross-entropy loss. Then, both student and teacher models are loaded with the best model after supervised training. Next, in the unsupervised phase, the consistency loss L_c (Eq. 4) and the feature constraint loss L_f (Eq. 5) are utilized to update the parameters of the student model. At the same time, labeled samples are also used to conduct the supervised cross-entropy loss L_s for student model. In this way, labeled samples can be optimally leveraged in both the supervised and unsupervised phases. The overall loss $L_{overall}$ for student model is the sum of the above three elements:

$$L_{overall} = L_c + L_f + L_s \tag{6}$$

It is crucial to note that the gradients for the teacher model are blocked in unsupervised phase. Therefore, the consistency loss L_c is not directly used to update its parameters. Instead, the exponential moving average (EMA) technique is employed, allowing the parameters of the teacher model to be updated using the parameters of the student model (Eq. 7). α is a smoothing coefficient hyperparameter.

$$\theta_t = \alpha\,\theta_t + (1-\alpha)\,\theta_s \tag{7}$$

Through the aforementioned training strategy, the teacher model and the student model play distinct roles in learning unlabeled change features. The student model is exposed to synthesized samples, which are more challenging because the mixed samples are not present in the real world. But it enables the student model to capture more generalizable change features. The teacher model, on the other hand, is consistently provided with original samples to ensure that the inputs during prediction align closely with the real-world distribution. Simultaneously, the EMA technique is employed to merge the general features learned by the student model, thereby further enhancing the generalization capability of teacher model. The outputs of the teacher model serve as the final CD results.

IV. Experiments

A. Datasets

LEVIR-CD dataset [48] and S2looking dataset [49] are used to evaluate the performance of CutMix-CD. Moreover, we construct an unbalanced version of S2looking dataset named Unbalanced-S2looking, where the distribution of change types (appearance and disappearance) in training and test phase are largely unbalanced. In addition, we perform the cross-dataset experiments on WHU-CD dataset [50] to further assess the generalization capability. Details of the four datasets are as follows.

LEVIR-CD: LEVIR-CD contains 637 very high-resolution Google Earth image pairs, mainly focusing on building changes, each with a resolution of 0.5 meters and a size of 1024 × 1024. The original image is cropped into 256 × 256. Follow the original split scheme, 7120, 1024, 2048 image pairs are obtained for train, validation and test respectively. Noted that the dominate change type in this dataset is newly-appearing changes. Only a few samples contain disappearing changes.

S2looking: S2looking subdivides change labels into two types: appearance and disappearance. The two types of changes are relatively more evenly distributed. The original dataset contains 5000 pairs of 1024 by 1024 samples with a resolution of 0.5-0.8m, and the training, validation, and test sets are divided according to the ratio of 7:1:2. In this experiment, the images are cropped to 256 by 256 size without overlapping and the unchanged samples are eliminated, resulting in 4328 samples for the training set, 607 samples for the validation set and 1159 samples for the test set.

Unbalanced-S2looking: An Unbalanced-S2looking dataset is constructed based on S2looking dataset to test the generalization ability of different methods. It is challenging because there is a huge difference in the distribution of change types in train and test sets. The training set contains 2000 samples with only the disappearing changes. The test set contains 800 samples: 300 samples with only disappearing changes, 300 samples with only newly-appearing changes, and 200 samples with both disappearing and newly-appearing changes. The aim of this dataset is to test whether a model trained on the full disappearing change data can extend its ability to detect changes in the case of newly-appearing changes. The experiment results on this dataset are shown in Section IV. E.

WHU-CD: WHU-CD comprises bi-temporal high-resolution images with the size of 32507 × 15345. It is captured in the Christchurch region of New Zealand with a resolution of 0.2 meters. It mainly focuses on building changes. The original images are cropped into non-overlapping 256 × 256 patches. We use WHU-CD to perform the cross-dataset experiments with LEVIR-CD and assess the generalization capability of different methods.

B. Evaluation Metrics

The performance of CD is evaluated using five metrics: F1-score (F1), Intersection over Union (IoU), overall accuracy (OA), Precision and Recall. IoU specifically refers to the IoU of the change category. OA takes into account the classification accuracy of both change and no-change categories. The Recall metric can assess the missing detections of various methods. F1 is a comprehensive metric that weighs Precision and Recall. The five metrics are calculated by:

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

$$IoU = \frac{TP}{TP + FP + FN} \tag{11}$$

$$OA = \frac{TP + TN}{TP + FP + FN + TN} \tag{12}$$

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

TABLE I
QUANTITATIVE PERFORMANCE OF DIFFERENT METHODS ON THE LEVIR-CD DATA SET
(THE BEST VALUES ARE BOLDED AND THE SECOND BEST VALUES ARE UNDERLINED.)

Method	5%					10%					20%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Sup-only	79.09	65.41	98.02	82.14	76.25	81.85	69.27	98.08	79.10	84.79	85.31	74.38	98.59	91.27	80.08
SemiCDNet	80.69	67.62	98.19	88.15	74.38	83.97	72.38	98.41	86.57	81.53	86.11	75.61	98.66	91.22	81.54
RCR	84.03	72.46	98.39	84.68	83.40	85.51	74.69	98.55	87.23	83.85	87.35	77.55	98.73	88.99	85.78
RCL	84.74	73.52	98.47	86.00	83.51	85.29	74.35	98.53	87.24	83.42	86.86	76.77	98.68	88.08	85.66
FPA	85.07	74.01	98.52	88.09	82.24	86.17	75.70	98.64	89.53	83.05	87.40	77.62	98.76	90.85	84.21
UniMatch	85.21	74.23	98.54	88.03	82.57	87.57	77.90	98.77	89.87	85.40	88.61	79.55	98.84	88.94	**88.28**
CutMix-CD	**87.77**	**78.21**	**98.78**	**89.70**	**85.92**	**88.76**	**79.79**	**98.87**	**90.00**	**87.56**	**89.44**	**80.90**	**98.95**	**91.56**	87.41
Fully-sup	F1=90.37, IoU=82.44, OA=99.04, Pre=91.90, Rec=88.90														

C. Experimental Settings

In the experiments, the student model and the teacher model utilized the same CD network as described in [40]. The student network employed the Adam optimizer with a learning rate set to 10-4. For the teacher network, the smoothing coefficient hyperparameter α was set to 0.99. Batch size is set to 8. The mask window was set to 1/4 of the original image size. The sensitivity analysis of mask sizes is presented in Section IV.G. The training process was divided into two phases. For supervised phase, the teacher model was trained using only the labeled samples for 30 epochs. For unsupervised phase, the student and teacher model were first initialized with the best model from supervised phase and then trained on the whole dataset for 100 epochs. The ablation studies on the initialization strategy are presented in Section IV.G. All the experiments were conducted on a single NVIDIA GeForce RTX 3090 GPU.

D. Comparison Experiments

The comparison experiments are conducted on LEVIR-CD and S2looking datasets. Five semi-supervised methods, including SemiCDNet [12], RCR [40], FPA [39] , RCL [38] and UniMatch [35] are employed for comparison. SemiCDNet is an adversarial learning-based method. RCR, FPA and UniMatch are three advanced consistency learning-based methods. RCR fits the cluster assumption for CD task and utilizes feature-level perturbations in consistency learning. FPA proposes feature alignment and prediction alignment strategies in consistency learning. UniMatch utilizes an auxiliary feature perturbation stream for an expanded perturbation space, which demonstrates its superiority in natural, remote sensing and medical image segmentation. RCL is a self-training method, which selects reliable pseudo labels by calculating the uncertainty. Additionally, two groups of experiments, namely Sup-only and Fully-sup are conducted for comparison. Sup-only indicates only training using different proportions of labeled data while unlabeled data is not used. Fully-sup refers to using all the labels in the training dataset.

To ensure fairness in the comparison, the basic CD network used in the aforementioned semi-supervised CD methods is the same as the change network employed in this approach (as described in Section III.B). Based on the same setting, the ability to learn from unlabeled data can be fairly compared.

On LEVIR-CD dataset:

Table I shows the quantitative performance on the LEVIR-CD dataset with the labeled ratio of 5%, 10% and 20%. CutMix-CD achieves the best performance across different labeled ratios in terms of F1, IoU and OA, which demonstrates the effective learning of change information from unlabeled samples by CutMix-CD. We notice that consistency learning-based methods (RCR, FPA, UniMatch and CutMix-CD) outperform adversarial learning-based method (SemiCDNet), indicating higher uncertainty in the learning process of GAN models. Comparing to the Sup-only approach, CutMix-CD exhibits larger improvements as the labeled ratio decreases. Under the labeled ratios of 5%, 10%, and 20%, CutMix-CD achieves F1 improvements of 8.68%, 6.91%, and 4.13%, respectively. Furthermore, with only 20% of the available labels, CutMix-CD achieves an F1 that is only 0.93% lower than that obtained by training with all the labels, which shows the efficient utilization of unlabeled data by CutMix-CD.

Fig. 4 shows the visual results on the LEVIR-CD dataset at a labeled ratio of 10%. In the LEVIR-CD dataset, the majority of changes are "newly-appearing", with only a small number of disappearing changes. Consequently, CD models trained on this dataset often suffer from severe overfitting issues, accurately predicting newly-appearing changes but exhibiting significant detection failure for disappearing changes.

To test the generalization ability in term of this issue, we specifically select two groups of samples containing disappearing changes within the test dataset (as shown in the first two rows of Fig. 4). The first row involves both disappearing and added objects. CutMix-CD yields the best visual outcomes. At the top of the images, other methods exhibit higher degrees of missing detection for buildings disappearing. Meanwhile, CutMix-CD has superior ability to

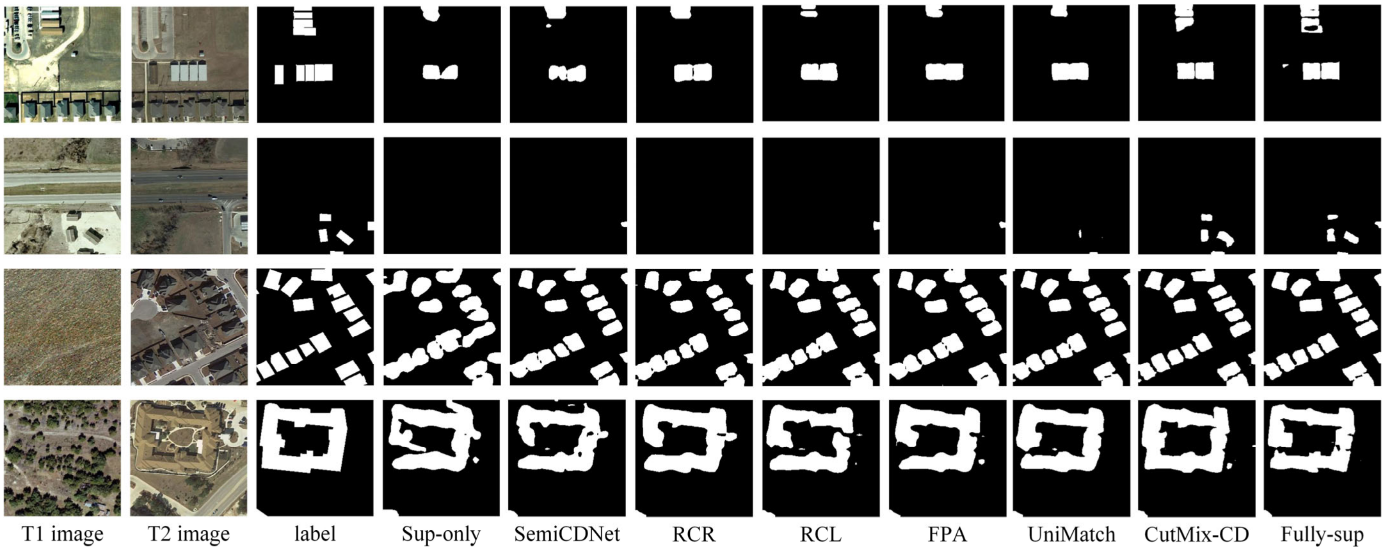


Fig. 4. Examples of CD results on the LEVIR-CD dataset at the labeled ratio of 10%.

handle added objects as well, which produces sharper outlines of the newly-built buildings in the middle of the images.

The samples in the second row features disappearing small buildings. Apart from CutMix-CD and Fully-sup, other methods all fail to detect the disappearing buildings. While SemiCDNet and FPA can detect the small newly-built building on the far right, their detection ability for disappearing changes is insufficient. This indicates that their limitation does not lie in detecting small objects, but in understanding the concept of "change" itself. Even in scenarios where the vast majority of training samples are newly-appearing changes, CutMix-CD still manages to identify disappearing changes, demonstrating its capability to enhance the change network's understanding of the abstract notion of "change" and alleviate the overfitting issue. Consequently, CutMix-CD exhibits stronger generalization abilities for various types of changes.

The third row presents samples depicting changes in dense building areas. CutMix-CD extracts clearer boundaries of buildings and effectively avoids sticking between buildings, which closely resemble the ground truth of the changes. The fourth row shows a sample with large-scale changed objects. CutMix-CD achieves the best completeness in detecting these large-scale changed objects even compared with Fully-sup method. None of the other methods are able to detect the buildings in their entirety.

On S2looking dataset:

Table II shows the quantitative performance on the S2looking dataset with the labeled ratio of 10%, 20% and 40%. Compared to the LEVIR-CD dataset, S2looking dataset exhibits more complex and diverse changed objects. The Fully-sup approach achieves a F1 of 67.64% and an IoU of 51.10% on this dataset. CutMix-CD achieves the best performance in terms of F1 and IoU across all three labeled ratio conditions. We observe that some semi-supervised methods exhibit lower accuracy than the Sup-only approach under certain labeled ratio conditions. For instance, at a labeled ratio of 40%, SemiCDNet and RCR achieve an IoU that is 1.98% and 0.67% lower than that of the Sup-only approach respectively. However, their OA still surpasses the Sup-only approach by 0.35% and 0.26% respectively. Although the overall classification accuracy of changed and unchanged pixels improves, the IoU for changed objects decreases. This suggests that in these methods, the improvement provided by unlabeled data is not sufficiently stable when the labeled ratio changes. In contrast, CutMix-CD exhibits significant improvements in both metrics compared to the Sup-only approach, highlighting the stability of this method.

Fig. 5 shows visual results on S2looking dataset at a labeled ratio of 20%. The first and second rows display samples of newly-appearing changes. In the first row, the Sup-only method, RCL and FPA exhibit missing detection, while SemiCDNet and RCR demonstrate false detections. In the second row, the new buildings in T2 image are similar to bare soil in T1 image. None of the compared methods except UniMatch (Sup-only, SemiCDNet, RCR, RCL, and FPA) can completely detect the changes. CutMix-CD, on the other hand, is capable of detecting changes more comprehensively while minimizing false positives. The third and fourth rows showcase samples with disappearing changes. In the third row, CutMix-CD detects each individual small building at the top of the image. In the fourth row, SemiCDNet, FPA, UniMatch and even Fully-sup fail to identify the disappearing changes. CutMix-CD achieves change contours that closely resemble the change labels. The proposed method consistently achieves the best visual outcomes across various complex change scenarios.

E. Experiments on extreme change condition

When the types of changes in a dataset are imbalanced, models tend to favor detecting the dominant type of change, resulting in significant omissions for the other types of changes (as shown in Fig. 1). To verify the generalization ability of the proposed method, we conduct experiments on the Unbalanced-S2looking dataset under extreme change condition. Training samples are all disappearing changes, while the test set contains various types of changes. Table III shows the quantitative evaluation on Unbalanced-S2looking dataset. Similarly, we

TABLE II
QUANTITATIVE PERFORMANCE OF DIFFERENT METHODS ON THE S2LOOKING DATA SET
(THE BEST VALUES ARE BOLDED AND THE SECOND BEST VALUES ARE UNDERLINED.)

Method	10%					20%					40%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Sup-only	47.79	31.40	96.42	74.61	35.15	57.93	40.78	96.91	68.16	50.37	61.76	44.67	96.76	68.21	56.42
SemiCDNet	46.79	30.54	96.48	61.23	37.85	55.28	38.19	96.66	61.06	50.49	59.84	42.69	97.11	69.09	52.77
RCR	52.53	35.93	96.82	57.46	48.02	59.26	41.87	96.93	58.20	**60.36**	61.01	44.00	97.02	65.57	57.05
RCL	57.12	39.98	96.83	63.86	51.66	61.59	44.50	97.11	67.38	56.72	64.38	47.47	97.19	66.69	62.23
FPA	56.80	39.66	96.96	67.83	48.85	60.02	42.88	97.11	69.18	53.00	62.32	45.27	97.25	70.81	55.65
UniMatch	50.52	33.80	**96.98**	**76.19**	37.79	60.69	43.57	**97.18**	**70.66**	53.19	63.20	46.2	**97.47**	**78.07**	53.09
CutMix-CD	**60.57**	**43.44**	96.79	60.74	**60.40**	**63.44**	**46.45**	97.16	66.89	60.32	**65.92**	**49.17**	97.23	66.27	**65.58**
Fully-sup	F1=67.64, IoU=51.10, OA=97.59, Pre=74.85, Rec=61.69														

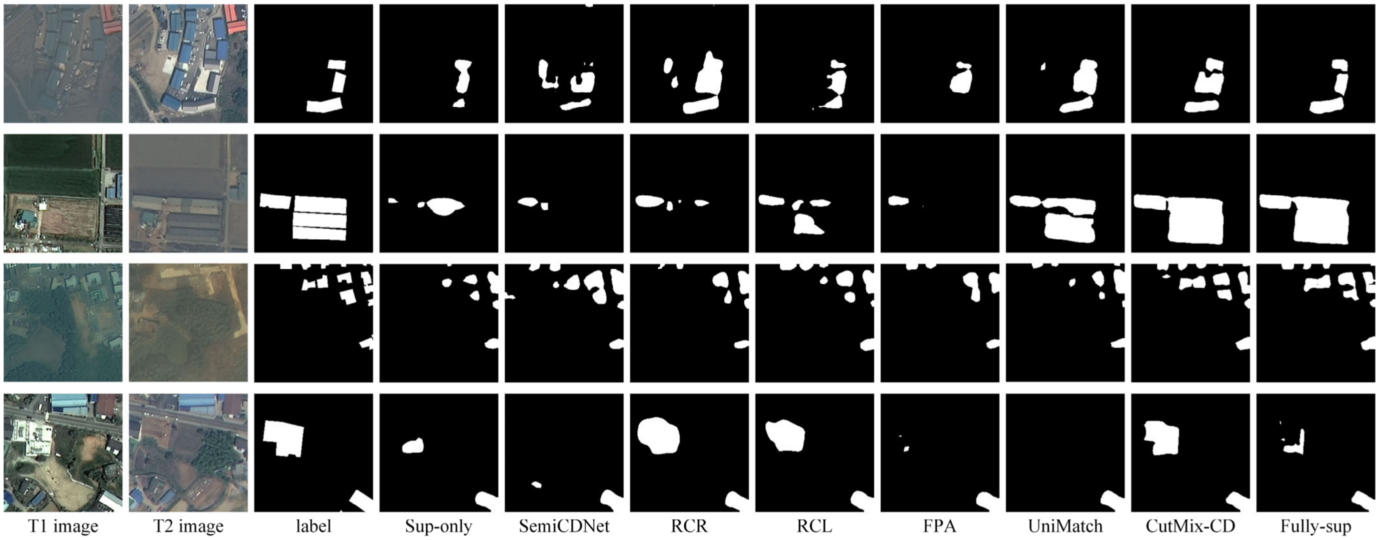


Fig. 5. Examples of CD results on the S2Looking dataset at the labeled ratio of 20%.

compare the following methods: Sup-only, SemiCDNet, RCR, RCL, FPA, UniMatch, CutMix-CD and Fully-sup. As shown in Table III, the overall accuracy on the Unbalanced-S2looking dataset is relatively low. The F1 and IoU in Fully-sup setting is 61.80% and 44.72% respectively, which indicates that unbalanced changes seriously affect the CD performance. Under such extreme training samples conditions, the proposed CutMix-CD achieves the best results on three comprehensive metrics of F1, IoU and OA. With 20% and 50% of training samples, F1 is improved by 7.3% and 6.71%, and Recall is improved by 10.68% and 10.12% compared to Sup-only, respectively. The omission situation significantly improves. RCL achieves the second-best performance in F1 and IoU. Other semi-supervised CD methods show a limited improvement. UniMatch even performs worse than Sup-only. Additionally, it is found that the performance of CutMix-CD with the label ratio of 50% is close to Fully-sup. This indicates that the CutMix-CD framework is able to learn more generalized change features and effectively mitigate the overfitting issue.

Fig. 6 shows visual comparisons on Unbanlanced-S2looking dataset at a labeled ratio of 50%. It includes representative samples from the test set with different types of changes. The first row displays the change of disappearance. Since the training set exclusively consists of disappearing changes, all methods successfully detect the disappearing changes. The second and third rows showcase newly-appearing changes. In the second row, two dense columns of new buildings appear in T2 image. Sup-only, SemiCDNet, RCR, FPA and UniMatch methods miss more than half of these new buildings, indicating that training on a dataset with only disappearing changes significantly reduces the effectiveness of recognizing newly-appearing changes. CutMix-CD demonstrates a remarkable improvement in detecting newly-appearing changes. The third row shows newly-appearing buildings under poor image quality of T2. Only CutMix-CD, UniMatch and Fully-sup can identify the change in the upper left. CutMix-CD extracts building boundaries more accurately. This indicates that the proposed method not only generalizes better across different types of changes but also performs more stable under challenging image condition. The fourth row presents samples containing both disappearing and newly-appearing changes. Nearly all the methods can identify the disappearing building on the right but with different degrees of missing detection for newly-built buildings. CutMix-CD is capable of detecting both types of changes simultaneously, which proves its generalization ability under extreme change conditions.

TABLE III
QUANTITATIVE PERFORMANCE OF DIFFERENT METHODS ON THE UNBALANCED-S2LOOKING DATA SET
(THE BEST VALUES ARE BOLDED AND THE SECOND BEST VALUES ARE UNDERLINED.)

Method	20%					50%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Sup-only	50.22	33.53	96.87	72.59	38.39	54.54	37.49	97.04	**73.95**	43.20
SemiCDNet	48.18	31.73	96.74	69.54	36.85	53.02	36.07	96.94	72.04	41.94
RCR	55.03	37.96	96.68	61.99	49.47	55.55	38.45	97.03	72.24	45.12
RCL	55.44	38.35	96.71	62.50	**49.81**	59.08	41.92	96.91	64.75	**54.32**
FPA	52.68	35.76	96.83	68.23	42.90	53.22	36.25	96.96	72.61	42.00
UniMatch	42.67	27.12	96.89	**87.86**	28.18	47.36	31.03	96.32	57.41	40.31
CutMix-CD	**57.52**	**40.38**	**97.02**	69.51	49.07	**61.25**	**44.14**	**97.23**	71.95	53.32
Fully-sup	F1=61.80, IoU=44.72, OA=97.35, Pre=76.02, Rec=52.07									

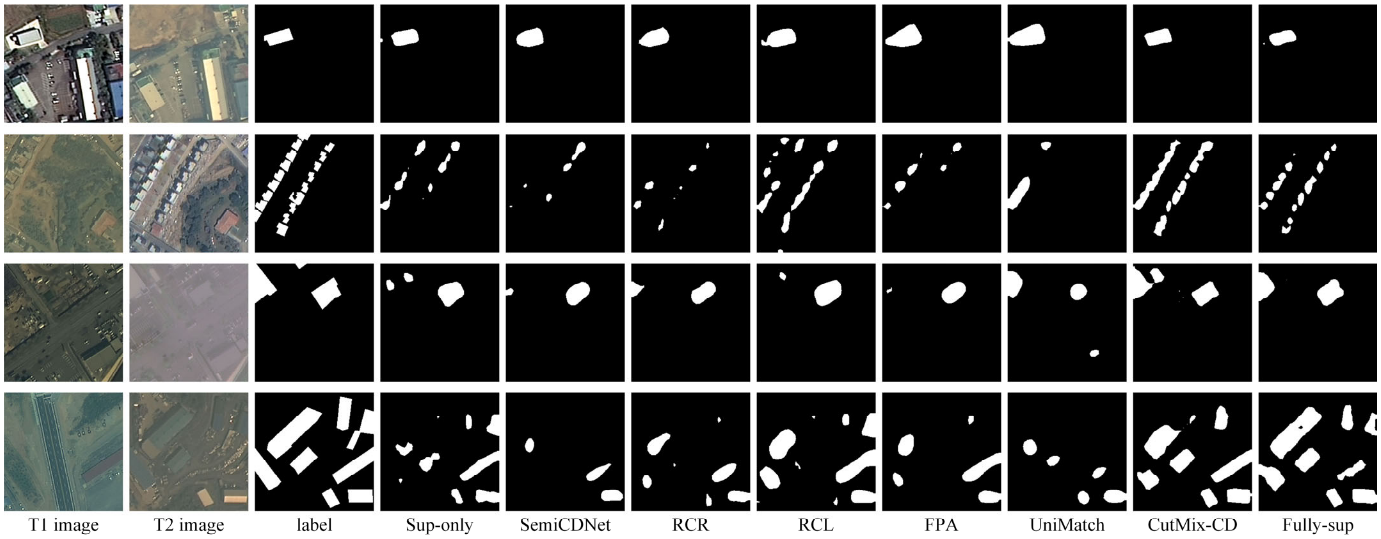


Fig. 6. Examples of CD results on the Unbalanced-S2Looking dataset at the labeled ratio of 50%.

F. Generalizability experiments

In this section, we perform cross-dataset experiments to test generalizability and transferability. In particular, we perform two groups of experiments. For the first experiment, a model is trained on the LEVIR-CD dataset with different labeled ratios and tested on the WHU-CD dataset. The experiment is denoted as LEVIR→WHU, which assesses the applicability of different semi-supervised CD methods across varied regions and architectural styles. The second experiment uses the labeled data from LEVIR-CD, combined with WHU-CD as unlabeled data, to train the model. This experiment is denoted as: [LEVIR (sup. %), WHU (unsup)]→LEVIR. It assesses whether the unlabeled images from WHU-CD can enhance performance on the LEVIR-CD dataset.

Table IV presents the quantitative performance of the LEVIR→WHU experiments. For F1 and IoU metrics, CutMix-CD achieves the best performance across all labeled ratios. CutMix-CD outperforms other methods by more than 1.92%, 0.71%, and 0.42% in F1 for 5%, 10%, and 20% labeled ratios, respectively. It indicates that CutMix-CD provides more robust supervisory signals from unlabeled data. The lower the labeled ratio, the more pronounced the improvement of CutMix-CD. Additionally, CutMix-CD consistently attains the best performance on Recall metric, showing the ability to detect a wider variety of changes. Interestingly, when the LEVIR-CD labeled ratio increases from 10% to 20%, most methods show a decline in performance. This phenomenon confirms the differences in distribution between the LEVIR-CD and WHU-CD datasets, indicating that more labeled data from LEVIR-CD does not necessarily improve performance on the WHU dataset. Despite significant architectural style differences, the change features learned by CutMix-CD from LEVIR-CD transfer more effectively to the WHU-CD dataset.

Table V shows the quantitative performance of the [LEVIR (sup. %), WHU (unsup)]→LEVIR experiments. For the comprehensive metrics F1, IoU, and OA, CutMix-CD is the only method to consistently achieve top two performance across all labeled ratios. The self-training method RCL and the consistency learning-based methods UniMatch and RCR also demonstrate stable improvements by utilizing unlabeled WHU data. However, this improvement diminishes as the LEVIR-CD labeled ratio increases. Additionally, SemiCDNet and FPA do not show improvement when trained with unlabeled WHU data compared to Sup-only, indicating that different distributions of changed building styles may negatively impact model training

TABLE IV
QUANTITATIVE PERFORMANCE FOR LEVIR→WHU EXPERIMENTS WITH PERCENTAGE OF LABELED DATA
(THE BEST VALUES ARE BOLDED AND THE SECOND BEST VALUES ARE UNDERLINED.)

Method	5%					10%					20%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Sup-only	44.26	28.42	96.70	67.12	33.02	52.93	35.99	95.87	48.27	58.59	54.96	37.89	97.30	**81.26**	41.52
SemiCDNet	47.81	31.41	96.93	73.35	35.46	55.27	38.19	97.13	72.36	44.71	56.87	39.74	97.34	<u>79.51</u>	44.27
RCR	49.31	32.72	96.94	<u>71.98</u>	37.50	65.98	49.23	97.57	74.16	<u>59.43</u>	61.36	44.25	97.33	72.00	53.46
RCL	<u>52.15</u>	<u>35.27</u>	<u>97.01</u>	71.47	41.05	58.03	40.87	97.38	79.37	45.73	59.24	42.08	97.03	65.06	<u>54.38</u>
FPA	50.73	33.99	96.90	68.60	40.25	<u>66.07</u>	<u>49.34</u>	**97.74**	**81.87**	55.39	<u>63.36</u>	<u>46.37</u>	**97.57**	78.76	52.99
UniMatch	52.14	35.26	96.88	66.73	<u>42.78</u>	63.91	46.96	<u>97.63</u>	<u>80.81</u>	52.85	60.77	43.65	97.37	74.55	51.29
CutMix-CD	**54.07**	**37.06**	**97.09**	**72.14**	**43.24**	**66.78**	**50.12**	97.14	61.97	**72.40**	**63.78**	**46.82**	<u>97.46</u>	73.50	**56.33**

TABLE V
QUANTITATIVE PERFORMANCE FOR [LEVIR (SUP. %), WHU(UNSUP)] → LEVIR EXPERIMENTS WITH PERCENTAGE OF LABELED DATA
(THE BEST VALUES ARE BOLDED AND THE SECOND BEST VALUES ARE UNDERLINED.)

Method	5%					10%					20%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Sup-only	79.09	65.41	98.02	82.14	76.25	81.85	69.27	98.08	79.10	**84.79**	85.31	74.38	98.59	<u>91.27</u>	80.08
SemiCDNet	78.51	64.62	97.94	83.85	73.81	81.62	68.95	98.21	85.74	77.88	83.41	71.54	98.43	90.24	77.54
RCR	81.22	68.37	98.05	79.58	**82.89**	83.04	71.00	98.33	86.21	80.10	85.64	74.89	98.58	88.19	<u>83.24</u>
RCL	**83.65**	**71.89**	**98.42**	<u>88.24</u>	<u>79.51</u>	84.03	72.46	98.39	85.04	<u>83.05</u>	85.57	74.78	98.53	85.32	**85.82**
FPA	73.65	58.29	97.57	82.10	66.77	79.95	66.60	98.07	85.01	75.46	83.32	71.41	98.38	87.53	79.50
UniMatch	82.05	69.57	98.29	88.23	76.68	<u>85.03</u>	<u>73.96</u>	<u>98.52</u>	<u>87.87</u>	82.37	**86.31**	**75.92**	<u>98.66</u>	90.24	82.72
CutMix-CD	<u>82.92</u>	<u>70.82</u>	<u>98.39</u>	**90.14**	76.76	**85.19**	**74.20**	**98.83**	**91.05**	80.03	<u>86.04</u>	<u>75.50</u>	**98.67**	**92.17**	80.67

Fig. 7 presents the visualization results of the LEVIR (sup 10%)→WHU experiments (first two rows) and the [LEVIR (sup 10%), WHU (unsup)]→LEVIR experiments (third and fourth rows). Compared to other semi-supervised CD methods, CutMix-CD consistently achieves the best visual results. In the first two rows, it even results in fewer false detections than Fully-sup. In the third row, CutMix-CD successfully detects all changed buildings. In the fourth row, it most comprehensively identifies the entire large new building. This demonstrates that the proposed method possesses strong generalization capabilities, effectively applying learned change information from one dataset to another.

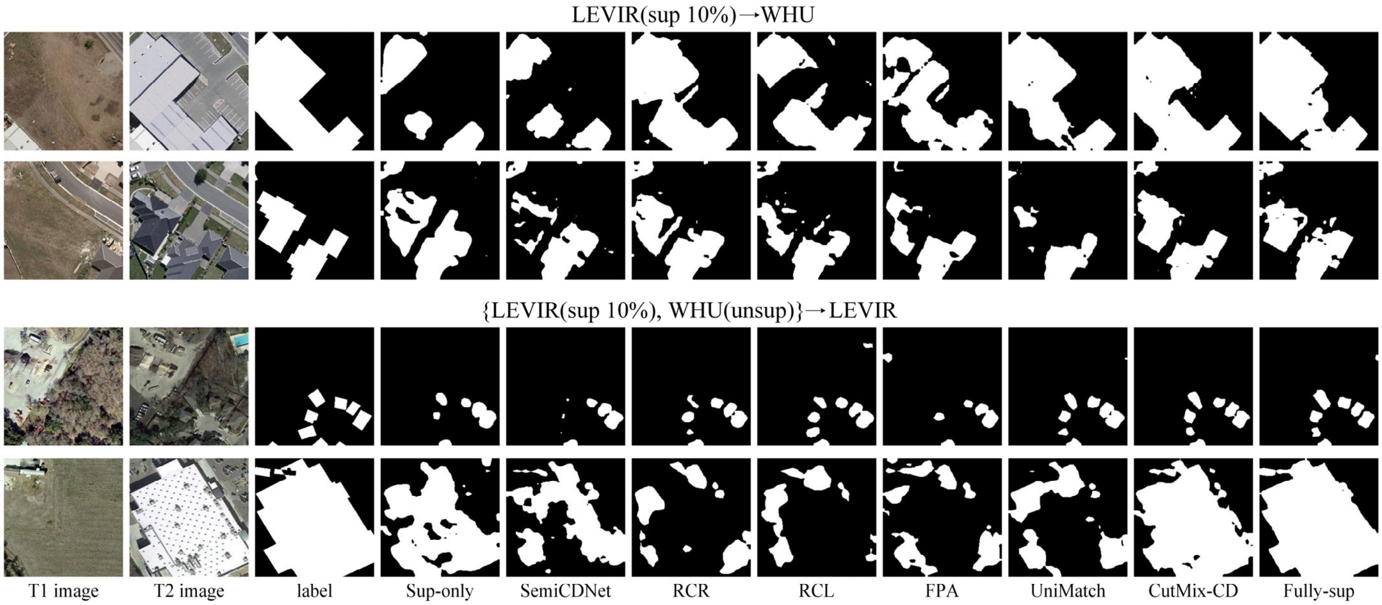


Fig. 7. Visualizations of generalizability experiments on the WHU and LEVIR dataset at the labeled ratio of 10%.

TABLE VI
Ablation Study of Change-Aware Mask and Feature Constraint Loss on S2looking Data Sets

Change-aware mask	Feature constraint loss	S2looking 20%				
		F1	IoU	OA	Pre	Rec
×	×	61.01	43.89	97.01	**77.39**	50.35
×	√	63.01	45.99	97.12	67.71	58.91
√	×	62.63	45.59	97.01	63.95	**61.36**
√	√	**63.44**	**46.45**	**97.16**	66.89	60.32

TABLE VII
Sensitivity of Mask Size on Two Data Sets

Mask size	LEVIR (10%) F1 (%)	S2looking (20%) F1 (%)	Unbalanced S2looking (50%) F1(%)
32×32	88.34	63.13	60.02
64×64	**88.76**	**63.44**	**61.25**
128×128	88.32	63.09	61.15
192×192	88.12	62.25	60.24

G. Ablation experiments

In this section, we firstly validate the effectiveness of change-aware mask and feature constraint loss. Then, we investigate the sensitivity of mask size and the initialization strategies for unsupervised phase.

Effectiveness of change-aware mask and feature constraint loss. Table VI presents the ablation study of change-aware mask and feature constraint loss on S2looking dataset. Compared with the base method, change-aware mask and feature constraint loss improve the F1 with 1.62% and 2.00%, respectively. Moreover, change-aware mask and feature constraint loss together improve the performance of 2.43% in F1. The new samples synthesized using the change-aware mask are more likely to include change regions. The increase in positive samples aids the consistency regularization process between the teacher path and the student path. The results demonstrate that such samples can more effectively promote the model's ability to complex changes. In addition, the feature constraint loss can further take advantages of the diverse samples. The feature-level alignment and distancing enhance the boundaries between changes and no-change categories.

Sensitivity of mask size. Table VII presents the performance of the CutMix-CD with different mask sizes on three datasets. For the LEVIR-CD, S2Looking and Unbalanced S2looking datasets, all the input image size is 256. The F1 shows little difference on LEVIR-CD datasets when the mask sizes are 32, 64, 128 and 192, where the performance fluctuation is within 0.64%, indicating that the performance is relatively stable across different mask sizes. However, the performance drops 1.19% with mask size of 192 on S2looking dataset. On the Unbalanced S2looking dataset, the F1 declines by more than 1% when the mask sizes are 32 and 192. This decline indicates that an excessively large or small size may not be suitable for model learning. Consequently, setting the mask size to one-quarter of the original image size, i.e., 64, maintains stable results across three datasets.

Initialization strategies. In unsupervised phase, the model initialization is important for learning the distribution of unlabeled samples. In our method, both the student and teacher models are initialized with the best model from supervised training (refer as supervised initialization strategy). We further investigate two other initialization strategies in unsupervised phase: 1) The student model is initialized with pretrained model on ImageNet [51]. 2) The student and teacher share parameters throughout the training process, which means a single model is used to extract changes from both original image pairs and mixed image pairs.

Table VIII shows the performance of three initialization strategies on three datasets. Our supervised initialization strategy achieves the best performance on the comprehensive metrics F1 and IoU, followed by ImageNet initialization and shared parameter strategy. On the LEVIR-CD and S2looking datasets, the performance gap between ImageNet initialization and supervised initialization is within 0.5% for F1, and widens to 1.28% on the more challenging Unbalanced S2looking dataset. This suggests that for more difficult tasks, there is a higher requirement for initialization. Our initialization strategy enables more stable supervisory signals for subsequent consistency learning. Additionally, we observe that sharing parameters between student and teacher models degrades the performance, which indicates that using the same model to learn original and mixed image pairs may cause confusion during consistency training. It further validates that using teacher and student models to learn original image pairs and mixed image pairs separately can fully leverage the advantages of CutMix data perturbation.

TABLE VIII
Ablation Study of Initialization Strategies on Three Data Sets

Method	LEVIR 10%					S2looking 20%					Unbalanced S2looking 50%				
	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec	F1	IoU	OA	Pre	Rec
Shared parameter	86.82	76.71	98.71	**90.68**	83.27	59.22	42.06	97.08	**69.02**	51.86	58.87	41.71	**97.31**	**79.20**	46.84
ImageNet-initial	88.26	78.99	98.83	90.54	86.10	62.99	45.98	97.00	63.44	**62.55**	59.97	42.82	97.27	75.43	49.77
Supervised-initial	**88.76**	**79.79**	**98.87**	90.00	**87.56**	**63.44**	**46.45**	**97.16**	66.89	60.32	**61.25**	**44.14**	97.23	71.95	**53.32**

V. Conclusion

In this paper, a novel semi-supervised CD framework, CutMix-CD, is proposed to learn the distribution of unlabeled data and enhance the generalization ability of change detection. The proposed method integrates the change-aware CutMix augmentation into a consistency learning framework, enriching the change samples and placing special emphasis on the comparative process. It effectively mitigates the issues of overfitting caused by the scarcity of positive samples and the imbalance of change types in semi-supervised CD. Competitive results were achieved on three datasets. Notably, our method showed significant improvements for the minority type of changes in the training set. This study can inspire researchers to consider the impact of the distribution of different change types in CD and promote the development of more generalizable methods under limited label condition. The rapid advancement of generative AI will further promote the development of generalized CD. This study primarily explores the effectiveness of synthesizing new change samples through augmentation methods. Future work will focus on generating AI-synthesized change data to create more diverse and realistic samples that align with real-world scenarios, thereby reducing model training costs and expanding the applicability in practical settings.

References

[1] H. Guo, Q. Shi, A. Marinoni, B. Du, and L. Zhang, “Deep building footprint update network: A semi-supervised method for updating existing building footprint from bi-temporal remote sensing images*,” Remote Sensing Environ.*, 264: p. 112589, 2021.

[2] S. Shi, Y. Zhong, Y. Liu, L. Zhang, and D. Li, “Cross-temporal high spatial resolution urban scene classification and change detection based on a class-weighted deep adaptation network,” Urban Inf., 3(1): p. 3, 2024.

[3] J. Chen, B. Sun, L. Wang, B. Fang, Y. Chang, Y. Li, J. Zhang, X. Lyu, and G. Chen, “Semi-supervised semantic segmentation framework with pseudo supervisions for land-use/land-cover mapping in coastal areas,” Int. J. Appl. Earth Observ. Geoinf., 112: p. 102881, 2022.

[4] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, “Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters,” Remote Sensing Environ., 265: p. 112636, 2021.

[5] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2015.

[6] J. Li, W. He, W. Cao, L. Zhang, and H. Zhang, “UANet: An Uncertainty-Aware Network for Building Extraction From Remote Sensing Images,” IEEE Trans. Geosci. Remote Sens., 62, 2024.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Proc. Adv. Neural Inf. Process. Syst., 30, 2017.

[8] H. Chen, Z. Qi, and Z. Shi, “Remote sensing image change detection with transformers,” IEEE Trans. Geosci. Remote Sens., 60: p. 1-14, 2021.

[9] M. Hao, S. Chen, H. Lin, H. Zhang, and N. Zheng, “A prior knowledge guided deep learning method for building extraction from high-resolution remote sensing images,” Urban Inf., 3(1): p. 6, 2024.

[10] Z. Ke, D. Qiu, K. Li, Q. Yan, and R.W. Lau, “Guided collaborative training for pixel-wise semi-supervised learning,” in Proc. Eur. Conf. Comput. Vis. 2020. Springer.

[11] J. Li, B. Sun, S. Li, and X. Kang, “Semisupervised semantic segmentation of remote sensing images with consistency self-training,” IEEE Trans. Geosci. Remote Sens., 60: p. 1-11, 2021.

[12] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, “SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images,” IEEE Trans. Geosci. Remote Sens., 59(7): p. 5891-5906, 2020.

[13] S. Mittal, M. Tatarchenko, and T. Brox, “Semi-supervised semantic segmentation with high-and low-level consistency,” IEEE Trans. Pattern Anal. Mach. Intell., 43(4): p. 1369-1379, 2019.

[14] W.-C. Hung, Y.-H. Tsai, Y.-T. Liou, Y.-Y. Lin, and M.-H. Yang, “Adversarial learning for semi-supervised semantic segmentation,” arXiv preprint arXiv:1802.07934, 2018.

[15] Y. Ouali, C. Hudelot, and M. Tami, “Semi-supervised semantic segmentation with cross-consistency training,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2020.

[16] J.-X. Wang, S.-B. Chen, C.H. Ding, J. Tang, and B. Luo, “RanPaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation,” IEEE Trans. Geosci. Remote Sens., 60: p. 1-16, 2021.

[17] Q. Xie, Z. Dai, E. Hovy, T. Luong, and Q. Le, “Unsupervised data augmentation for consistency training,” Adv. Neural Inf. Process. Syst., 33: p. 6256-6268, 2020.

[18] S. Hafner, Y. Ban, and A. Nascetti, “Semi-Supervised Urban Change Detection Using Multi-Modal Sentinel-1 SAR and Sentinel-2 MSI Data,” Remote Sens., 15(21): p. 5135, 2023.

[19] Y. Yang, X. Tang, J. Ma, X. Zhang, S. Pei, and L. Jiao, “ECPS: Cross Pseudo Supervision Based on Ensemble Learning for Semi-Supervised Remote Sensing Change Detection,” IEEE Trans. Geosci. Remote Sens., 2024.

[20] C. Wu, B. Du, and L. Zhang, “Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection,” IEEE Trans. Pattern Anal. Mach. Intell., 45(8): p. 9774-9788, 2023.

[21] G. French, S. Laine, T. Aila, M. Mackiewicz, and G. Finlayson, Semi-supervised semantic segmentation needs strong, varied perturbations, in Br. Mach. Vis. Conf. 2019.

[22] S. Laine and T. Aila, “Temporal ensembling for semi-supervised learning,” arXiv preprint arXiv:1610.02242, 2016.

[23] H. Chen, J. Song, C. Wu, B. Du, and N. Yokoya, “Exchange means change: An unsupervised single-temporal change detection framework based on intra-and inter-image patch exchange,” ISPRS J. Photogramm. Remote Sens., 206: p. 87-105, 2023.

[24] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, “A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images,” ISPRS J. Photogramm. Remote Sens., 166: p. 183-200, 2020.

[25] D. Peng, Y. Zhang, and H. Guan, “End-to-end change detection for high resolution satellite images using improved UNet++,” Remote Sens., 11(11): p. 1382, 2019.

[26] J. Pan, Y. Bai, Q. Shu, Z. Zhang, J. Hu, and M. Wang, “M-Swin: Transformer-based Multi-scale Feature Fusion Change Detection Network Within Cropland for Remote Sensing Images,” IEEE Trans. Geosci. Remote Sens., 2024.

[27] H. Chen, J. Song, C. Han, J. Xia, and N. Yokoya, "Changemamba: Remote sensing change detection with spatio-temporal state space model," arXiv preprint arXiv:2404.03425, 2024.

[28] 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., 2023.

[29] H. Guo, X. Su, C. Wu, B. Du, and L. Zhang, "SAAN: Similarity-aware attention flow network for change detection with VHR remote sensing images," IEEE Trans. Image Process., 2024.

[30] Q. Shu, J. Pan, Z. Zhang, and M. Wang, "DPCC-Net: Dual-perspective change contextual network for change detection in high-resolution remote sensing images," Int. J. Appl. Earth Observ. Geoinf., 112: p. 102940, 2022.

[31] H. Chen, W. Li, and Z. Shi, "Adversarial instance augmentation for building change detection in remote sensing images," IEEE Trans. Geosci. Remote Sens., 60: p. 1-16, 2021.

[32] X. Zhang, X. Huang, and J. Li, "Joint self-training and rebalanced consistency learning for semi-supervised change detection," IEEE Trans. Geosci. Remote Sens., 2023.

[33] L. Yang, W. Zhuo, L. Qi, Y. Shi, and Y. Gao, "St++: Make self-training work better for semi-supervised semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2022.

[34] K. Sohn, D. Berthelot, N. Carlini, Z. Zhang, H. Zhang, C.A. Raffel, E.D. Cubuk, A. Kurakin, and C.-L. Li, "Fixmatch: Simplifying semi-supervised learning with consistency and confidence," Adv. Neural Inf. Process. Syst., 33: p. 596-608, 2020.

[35] L. Yang, L. Qi, L. Feng, W. Zhang, and Y. Shi, "Revisiting weak-to-strong consistency in semi-supervised semantic segmentation," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2023.

[36] L. Hoyer, D.J. Tan, M.F. Naeem, L. Van Gool, and F. Tombari, "Semivl: Semi-supervised semantic segmentation with vision-language guidance," in Proc. Eur. Conf. Comput. Vis. 2025. Springer.

[37] K. Li, X. Cao, Y. Deng, J. Song, J. Liu, D. Meng, and Z. Wang, "SemiCD-VL: Visual-Language Model Guidance Makes Better Semi-supervised Change Detector," arXiv preprint arXiv:2405.04788, 2024.

[38] J.-X. Wang, T. Li, S.-B. Chen, J. Tang, B. Luo, and R.C. Wilson, "Reliable contrastive learning for semi-supervised change detection in remote sensing images," IEEE Trans. Geosci. Remote Sens., 60: p. 1-13, 2022.

[39] X. Zhang, X. Huang, and J. Li, "Semisupervised change detection with feature-prediction alignment," IEEE Trans. Geosci. Remote Sens., 61: p. 1-16, 2023.

[40] W.G.C. Bandara and V.M. Patel, "Revisiting consistency regularization for semi-supervised change detection in remote sensing images," arXiv preprint arXiv:2204.08454, 2022.

[41] Q. Shu, J. Hu, J. Pan, Y. Bai, Z. Zhang, and Z. Li, "TCNet: Temporal consistency network for semisupervised change detection," in Proc. Int. Conf. Artif. Intell. Comput. Inf. Technol. (AICIT). 2022. IEEE.

[42] C. Han, C. Wu, M. Hu, J. Li, and H. Chen, "C2F-SemiCD: A coarse-to-fine semi-supervised change detection method based on consistency regularization in high-resolution remote-sensing images," IEEE Trans. Geosci. Remote Sens., 2024.

[43] Q. Shu, J. Pan, Z. Zhang, and M. Wang, "MTCNet: Multitask consistency network with single temporal supervision for semi-supervised building change detection," Int. J. Appl. Earth Observ. Geoinf., 115: p. 103110, 2022.

[44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2016.

[45] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2017.

[46] S. Yun, D. Han, S.J. Oh, S. Chun, J. Choe, and Y. Yoo, "Cutmix: Regularization strategy to train strong classifiers with localizable features," in Proc. IEEE/CVF Int. Conf. Comput. Vis. 2019.

[47] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," Adv. Neural Inf. Process. Syst., 30, 2017.

[48] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," Remote Sens., 12(10): p. 1662, 2020.

[49] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, and B. Jiang, "S2Looking: A satellite side-looking dataset for building change detection," Remote Sens., 13(24): p. 5094, 2021.

[50] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," IEEE Trans. Geosci. Remote Sens., 57(1): p. 574-586, 2018.

[51] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. 2009. Ieee.

Qidi Shu received the B.E. degree in spatial informatics and digitalized technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2020, and the M.E degree in photogrammetry and remote sensing from State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing, Wuhan University, Wuhan, China, in 2023. He is currently pursuing the Ph.D. degree with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

His research interests include change detection and remote sensing data fusion.

Xiaolin Zhu (Senior member, IEEE) received the B.Sc. degree in resource science and engineering and M.Sc. degree in civil engineering from Beijing Normal University, Beijing, China, in 2007 and 2010, respectively, and the Ph.D. degree in geography from Ohio State University, Columbus, OH, USA, in 2014. He was a postdoctoral researcher with Colorado State University and University of California, Davis in 2015 and 2016 respectively.

He is currently an Associate Professor at Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. His research interests include remote sensing image processing, data fusion, vegetation remote sensing, nighttime light remote sensing, and urban remote sensing.

Luoma Wan received the B.Sc. degree in computer science and technology from Wuhan Institute of Technology, Wuhan, China, in 2012, the M.Sc. degree in software engineering from South China Normal University, Guangzhou, China, in 2015, and the Ph.D. degree in earth system and geoinformation science from the Chinese University of Hong Kong, Hong Kong, in 2021.

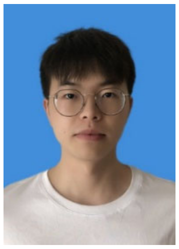
From 2015 to 2018, he worked as a research assistant in the Institute of Space and Earth Information Science, Chinese University of Hong Kong. He is currently a Postdoctoral Researcher with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University, Hong Kong. His research interests include mangrove forests monitoring with remote sensing, carbon cycle, and deep learning.

Shuheng Zhao received the B.S. degree in geodesy and geomatics engineering and M.S. degree in photogrammetry and remote sensing from the Wuhan University, Wuhan, China, in 2018 and 2021, respectively. She is currently pursuing the Ph.D. degree in The Hong Kong Polytechnic University, Hong Kong, China.

Her research interest is hyperspectral image processing.

Denghong Liu received the B.S. degree in geodesy and geomatics engineering and M.S. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2019 and 2022, respectively. He is currently pursuing the Ph.D. degree with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

His research interests include hyperspectral image processing and land surface mapping.



Longkang Peng received the B.E. degree in spatial-informatics & digitalized technology from University of Electronic science and Technology of China, Chengdu, China, in 2019 and the M.S. degree with Beijing Normal University, Beijing, China, in 2022.

He is pursuing the Ph.D. degree with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University. His research interests include remote sensing classification and segmentation.

Xiaobei Chen received her B.A. degree in journalism and psychology from Shenzhen University, China, in 2022. She is currently pursuing the Ph.D. degree with the Department of Land Surveying and Geo-Informatics, The Hong Kong Polytechnic University.

Her research interest is mainly focused on the impact of urban geographic space on human spatial cognition.