LCS-EnsemNet: A Semisupervised Deep Neural Network for SAR Image Change Detection With Dual Feature Extraction and Label-Consistent Self-Ensemble

Jian Wang, Yinghua Wang, *Member, IEEE*, Bo Chen, *Senior Member, IEEE*, and Hongwei Liu, *Member, IEEE*

***Abstract*—Change detection (CD) in synthetic aperture radar (SAR) images faces two challenging problems limiting the detection performance: inherent speckle noise in SAR data causes the overlapping nature of changed and unchanged classes and, thus, affects the image understanding for inferring category of each image pixel; and adequate labeled samples are quite laborious and time-consuming to collect, which is the major limitation for supervised methods. In this article, we develop a novel deep learning-based semisupervised method to address these challenges. The method first incorporates a pixel-wise log-ratio difference image (DI) and its saliency map to produce a spatially enhanced (SE) DI using a reweighting scheme based on the fact that changed pixels exhibit higher saliency than unchanged pixels. As a result, prominent changed regions are highlighted, and the class separability is significantly increased. We construct pixel-wise and context-wise features based on the log-ratio DI and SE DI, which respectively provide image detail cue and spatial context cue, as dual input features to jointly characterize the change information at each pixel position. Second, we propose a label-consistent self-ensemble network (LCS-EnsemNet), which can take advantage of the unlabeled samples to learn discriminative high-level features for the precise identification of changed pixels. By enforcing a label consistency between dual features and a label consistency across multiple classifiers, the label-consistent self-ensemble strategy enables the proposed network to selectively transform unlabeled samples into pseudo-labeled samples in an unsupervised manner and ensures that the selected pseudo-labels are reliably and stably predicted. Finally, the cross-entropy loss is calculated with the limited labeled data and selected pseudo-labeled samples to optimize the LCS-EnsemNet in a supervised way. The proposed method is evaluated on three low/medium-resolution SAR datasets and one high-resolution SAR dataset, and experimental results have demonstrated its efficiency and effectiveness.**

***Index Terms*—Change detection (CD), deep neural network (DNN), label-consistent self-ensemble, semisupervised learning (SSL), spatially enhanced (SE) difference image (DI), synthetic aperture radar (SAR).**

Manuscript received January 27, 2021; revised June 7, 2021 and September 23, 2021; accepted October 18, 2021. Date of publication October 26, 2021; date of current version December 3, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61671354, in part by the 111 Project, and in part by the Shaanxi Innovation Team Project. *(Corresponding authors: Yinghua Wang; Hongwei Liu.)*

The authors are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: wj_xidian@163.com; yhwang@xidian.edu.cn; bchen@mail.xidian.edu.cn; hwliu@xidian.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2021.3122461

I. Introduction

CHANGE detection (CD) is one of the central problems in Earth observation as it can analyze images of the same scene acquired at different times to identify changes that may have occurred [1]–[3]. It can be utilized in numerous applications, to name a few, agricultural monitoring [4], disaster surveillance [5], and urban spatial planning [6]. In recent years, thanks to the all-weather and all-time imaging capability [7], [8], synthetic aperture radar (SAR) has played an important role in remote sensing (RS) image CD task. However, the intrinsic complexity of SAR data makes it a challenging task to identify the changed areas accurately.

CD in SAR images mainly comprises two steps: 1) the first step is the generation of a difference image (DI), which indicates the degree of changes or dissimilarities in the corresponding areas of two coregistered SAR images. This step routinely utilizes local comparative operators to quantify the dissimilarity between corresponding pixels. Commonly used operators, such as the subtraction, ratio [9], and log-ratio operators [1]–[3], [10], [11] primarily compute dissimilarity measurement pixel-by-pixel without considering nonlocal spatial information, resulting in extensive false alarms due to the interference of speckle noise [1], [12], [13]. 2) The second step is the classification of DI into changed and unchanged classes [3], [8]. In the works of literature, many approaches have emerged to perform this binary classification. Supervised classifiers receive less attention due to the difficulty of gathering ground references. By contrast, unsupervised algorithms, such as the traditional thresholding analysis [14], segmentation approaches [10], statistical modeling [3], [15], and clustering algorithms [16], as well as recently developed deep learning-based methods with the preclassification and pseudo-labeling frameworks [17]–[19], can generate a final change map without any labeled samples. However, they rely on the premise of large interclass separation or assumption of a statistical distribution of classes in discriminating the changed pixels and unchanged pixels, which cannot be held in the context of SAR data, for instance, the Gaussian distribution assumption in Gaussian mixture model based clustering [20]. In addition, the absence of labeled samples makes the unsupervised methods more intractable to achieve superior results. As a tradeoff between performance and demand of labeled training samples, semisupervised learning (SSL) may be a feasible solution. However, the existing semisupervised SAR image CD methods [21], [22] still depend on conventional shallow machine learning algorithms to a large extent, which lack hierarchical and abstract feature learning ability, thereby limiting the CD performance.

To sum up, there are two challenging problems in SAR image CD. 1) The inherent speckle noise [1], [12], [13] in SAR images causes the strong intensity variation and further leads to pixels from changed and unchanged classes mutually overlapping, which inevitably reduces intraclass compactness and interclass separation. Hence, the large similarity between some noisy unchanged pixels and real changed pixels, especially in heterogeneous areas and borders, may bring extensive false alarms in the results. The widely utilized comparative operators are almost pixel-wise descriptors, which are sensitive to speckle noise, making it hard to effectively measure the change information of bitemporal SAR images. Meanwhile, multiscale spatial information [23]–[25], especially global spatial information and context information, are seldom used to generate the DI for CD in SAR images. Therefore, spatial-contextual semantic information should be taken into account to compensate for the pixel-wise change information from conventional comparative operators to facilitate the suppression of speckle noise and the reduction of class confusion in inhomogeneous areas and borders. 2) Sufficient labeled samples are significant for excellent detection performance, which, however, are extremely expensive and time-consuming to gather by manual pixel-wise annotation for SAR data [26]. SSL can reveal underlying label information and extract discriminative features from massive unlabeled samples for the improvement of generalization ability while requiring only a few labeled training samples [27]. However, the existing semisupervised CD methods in SAR images still adopt traditional machine learning models, which lack hierarchical and abstract feature learning ability. More advanced deep SSL approaches should be introduced to achieve CD in SAR images.

To alleviate the first problem discussed above, i.e., the commonly used pixel-wise DI is not effective enough in discriminating the similar changed and unchanged pixels in the heterogeneous areas, borders, and noisy areas, we introduce the saliency map [28] to generate a spatially enhanced (SE) DI and then construct dual low-level features from both the pixel-wise DI and context-wise SE DI. Considering the defect of the traditional operators mentioned above, we try to exploit spatial context information in the saliency map to consolidate the change measure. Therefore, a reweighting scheme is especially designed to inject the spatial-contextual change information extracted from the saliency map into the log-ratio DI to construct a context-wise SE DI, achieving an improved class separability. Thereafter, by using the log-ratio DI and SE DI as well as the original SAR images, we carefully design a dual-feature extraction scheme for constructing dual input features, which respectively emphasize the pixel-wise information and spatial context information. The utilization of the dual low-level features benefits the high-level discriminative feature learning and improves the CD performance.

As for the second problem of the scarcity of labeled training data abovementioned, deep learning-based semisupervised classification provides a potential solution by learning an informative high-level feature representation from unlabeled samples with the guidance from limited labeled training samples [29]–[34]. These learned high-level representations are commonly used to assign pseudo-labels to unlabeled samples, which will be employed together with the labeled data to train the classifier further. Hence, it is important to find a suitable pseudo-labeling strategy to obtain precise pseudo-labels. Recently, a wide variety of SSL algorithms [35], [36] have been developed based on the idea of ensemble learning. These algorithms aim at capturing reliable category information (e.g., pseudo-labels and class probabilities) from unlabeled samples for the training of a deep network. The core idea behind these algorithms is to aggregate the intermediate predictions by the deep network temporally over training epochs to ensure the reliability of the captured label information. However, directly applying these algorithms may encounter a performance drop since they are designed for optical images. In addition, they utilize all the predictions of unlabeled instances for network training without any selection procedure, which may lead to incorrect predictions for hard-to-be-classified samples and then mislead the model learning, resulting in performance drop accordingly. Considering these drawbacks, we try to adopt an elaborate strategy to refine the pseudo-labels predicted by the trained network. Therefore, we develop a semisupervised label-consistent self-ensemble network (LCS-EnsemNet). During the training process, both label consistency between dual features and label consistency across multiple classifiers are imposed on the network predictions of the same input, aiming to select the trustworthy pseudo-labels for subsequent model training. The selected reliable pseudo-labels, along with their corresponding samples, help the model to learn more discriminative and robust feature representations from unlabeled samples, thereby improving the CD performance. The main contributions of this article are as follows.

1) A dual-feature representation construction that considers both pixel-wise and context-wise change feature extraction is carefully designed. Since pixel-wise feature and context-wise feature describe the change information of each pixel in terms of image details and spatial contexts, they can be regarded as information from two different modalities. Fully exploiting the dual features helps the network to better distinguish between changed and unchanged pixels, thus improving the detection performance.
2) A two-branch network, referred to as the LCS-EnsemNet, is presented to exploit unlabeled data to achieve CD in SAR data, whose two branches are trained respectively to learn image detail cue from pixel-wise features and spatial context cue from context-wise features. In this manner, high-level representations are independently learned from two different modalities.
3) A two-stage label refinement strategy termed label-consistent self-ensemble is devised by considering both label consistency between dual features and label consistency across multiple classifiers, with which only trustworthy pseudo-labels are selected for model learning. During

TABLE I
NOTATIONS AND DEFINITION

Notation	Definition
$\mathbf{I}_1$	Prechange SAR image at time1.
$\mathbf{I}_2$	Postchange SAR image at time2.
$\mathbf{D}_{\text{log}}$	Log-ratio DI.
$\mathbf{D}_{\text{se}}$	Spatially enhanced (SE) DI.
$H \times W$	Image size of the original SAR images.
$\mathbf{P}_i$	Image patch centered at pixel i .
Ω_l	Labeled training set.
Ω_u	Unlabeled training set.
$\Omega_p^{(t)}$	Pseudo labeled training set at epoch *t*.
Ω_{test}	Testing set.
$(\mathbf{x}_{i,l}^{\text{pw}}, \mathbf{x}_{i,l}^{\text{cw}})$, y_i	The *i*th labeled sample-pair and its label.
$(\mathbf{x}_{i,u}^{\text{pw}}, \mathbf{x}_{i,u}^{\text{cw}})$	The *i*th unlabeled sample-pair.
$(\hat{\mathbf{x}}_{i,p}^{\text{pw}}, \hat{\mathbf{x}}_{i,p}^{\text{cw}})$, $\hat{y}_i$	The *i*th pseudo labeled sample-pair and its label.
$(\mathbf{x}_{i,test}^{\text{pw}}, \mathbf{x}_{i,test}^{\text{cw}})$	The *i*th testing sample-pair.
Φ , Ψ	Two subnetworks of the LCS-EnsemNet.
$\mathbf{W}_n$, $\mathbf{b}_n$	Weights and bias of the *n*th layer in subnetwork.
α	Parameter of reweighting scheme.
K	Length of sliding window in self-ensemble stage.
$h \times h$	Patch size.
τ_1 , τ_2	Thresholds for selecting pseudo labels.

the training process, the label-consistent self-ensemble strategy selects credible pseudo-labels from the LCS-EnsemNet and uses them to retrain the model, therefore improving the feature learning and classification performance. In addition, the improved model further updates the pseudo-labels and this process is repeated in training, improving the CD performance progressively.

The rest of this article is organized as follows. In Section II, a few related works, including the CD methods in SAR image and the deep SSL methods, will be reviewed. Section III gives a description of the preliminary knowledge. Section IV details the dual feature extraction and the proposed semisupervised LCS-EnsemNet. In Section V, LCS-EnsemNet is tested with experiments on three low/medium-resolution SAR datasets and one high-resolution SAR dataset, and the experimental results demonstrate its effectiveness and adaptation to images with different spatial resolutions. Finally, Section VI concludes this article.

For clarity, this article uses bold letters to denote a matrix, a bold lowercase letter to denote a vector, and italic letters (both upper and lowercase) to denote scalar. We illustrate the important notations and definitions in Table I.

II. Related Work

In this section, we will briefly review the related works on CD in SAR images and deep SSL.

A. CD in SAR Image

Over recent decades, the CD has become a topic of concern to researchers in the SAR community [1]. In the literature, the CD approaches for SAR images can be summarized into three categories: unsupervised, supervised, and semisupervised approaches. Here, we quickly review them.

Since collecting labeled samples is labor-intensive and time-consuming, the unsupervised framework has drawn considerable attention. Bruzzone and Prieto [38], Bazi *et al.* [3], [39], [40], Bovolo and Bruzzone [2], [41], Inglada and Mercier [42], Moser and Serpico [14], and Celik [16] implemented pioneering works for unsupervised CD by introducing the expectation-maximization algorithm, statistical modeling, automatic thresholding, and clustering algorithms. According to the available literature, unsupervised SAR image CD has been intensively investigated from two aspects. On the one hand, some efforts have been devoted to reducing the deleterious effects of speckle and, thus, generating a high-quality DI, such as applying undecimated discrete wavelet transform to perform a multiresolution analysis of the log-ratio DI [10], [11], fusing several DIs in the wavelet domain [43], as well as exploiting a nonlocal low-rank model and the statistical characteristics of multitemporal images to reconstruct a cleaner DI [23]. Recently, Sun *et al.* [44] exploited the structure consistency of the bitemporal images for the DI generation, thus improving the robustness of the difference information to the speckle. On the other hand, improvements have been made by modifying classification algorithms, such as random field classifiers [45], kernel clustering [46], and hierarchical clustering [18], [47]. In recent years, thanks to the powerful learning abilities of deep networks, deep learning-based CD methods have attracted great attention, with which a preclassification framework is established [17]–[19], [48], [49]. Under the framework, Gong *et al.* [17], Gao *et al.* [18], [48], [49], Geng *et al.* [19], and Li *et al.* [50] utilized clustering results (i.e., preclassification results) of DI as pseudolabel information for model learning and achieved appealing performance. However, there are several problems associated with the existing conventional and deep learning-based methods. Conventional methods face the following challenges: difficulty in designing hand-crafted features, inadequate ability to model complex SAR data, and high sensitivity of the comparative operators to noise. In the deep learning-based methods, the pseudo-labels are only predicted by traditional clustering algorithms, and the label accuracy cannot be ensured. Although the deep learning-based methods have the better capability in feature learning, yet the incorrect pseudo-labels may mislead the model learning, resulting in performance degradation.

More recently, a few works have been proposed to address the problem of insufficient labeled training samples by leveraging transfer learning and self-supervised learning [24], [51]–[54]. For example, Saha *et al.* [51] proposed a novel building CD method that first trains a cycle consistent generative adversarial networks (CycleGAN) to learn transcoding between SAR and optical images in an unsupervised fashion, and subsequently, optical-like features can be extracted from SAR images. The deep change vector analysis framework [55] specialized for the optical image CD is applied to predict the CD map.

In addition, some supervised methods for CD task have been proposed. Camps-Valls *et al.* [56] employed a kernel-based support vector machine (SVM) for CD in multisource RS data. To alleviate the problem of lacking labeled training instances, Li *et al.* [57] introduced the deformable transformation to limited training samples to generate diversified patterns for dictionary learning, similar to data augmentation. Recent works [58]–[60] employed deep models as classifiers for the final classification. In [59], a supervised method is presented using a multiscale capsule network. Wang *et al.* [60] devised a lightweight network to reduce the computational complexity and achieved better results than conventional heavy networks. However, the major restriction of these supervised methods is still the scarcity of labeled training samples.

Recently, SSL [29]–[34], [61]–[66] has been introduced to solve the problem of scarcity of labeled data. Despite the great success of SSL in the field of RS imagery, few efforts have been made in the context of SAR image CD [21], [22]. In [21], Jia *et al.* proposed a kernel-based semisupervised SVM for CD. An *et al.* [22] built two discriminative models based on Markov random field (MRF), which are trained on labeled samples and unlabeled samples, respectively. Finally, these two trained MRF-based models are combined to predict the final CD map. However, these traditional shallow semisupervised models cannot capture sufficiently discriminative features from nonlinearly separable SAR data with complex scenes. To overcome this drawback, we develop a deep semisupervised network, referred to as LCS-EnsemNet, to learn informative, credible, and underlying features from both labeled and unlabeled samples for better CD performance in SAR images.

B. Deep SSL

SSL targets at leveraging a large volume of unlabeled data to mitigate the shortfall of labeled training samples, which has shown excellent capability in extracting the category information from unlabeled data and improving the generalization ability. SSL has been widely utilized in RS imagery analysis [29]–[34], [61]–[66]. Generally speaking, SSL algorithms include generative models [67], discriminative models [64], low-density separation approach [68], graph-based model [61], [69], graph convolutional network [71], and self-training algorithm [35], [36], [71], [72]. The self-training algorithm is one of the most popular methods, where the inferred pseudo-labels on unlabeled samples by self-labeling strategy are utilized as the real labels to supervise model training for better performance. Naturally, the reliability of the pseudolabel is essential for the generalization ability, whereas high reliability is usually hard to ensure.

In recent works of literature, many researchers have combined consistency regularization and deep models to extract more reliable pseudolabel information [35], [36], [72]. Specifically, Laine and Aila proposed temporal ensembling [35] to accumulate the intermediate predictions of a single network over different training epochs into more reliable predictions, which can be exploited for subsequent training. In [35], the network architecture in different training epochs differs due to dropout regularization [37]. Thus, the accumulated reliable prediction is equivalent to the ensemble prediction from many different individual classifiers. However, considering the inherent speckle noise and class confusion in SAR CD, the above one-stage self-labeling strategy may fail to ensure the high reliability of the pseudo-labels. Accordingly, the wrongly predicted labels may hinder network learning. In this article, we develop a two-stage label-consistent self-ensemble strategy to refine the captured category knowledge and generate more accurate pseudo-labels.

III. Preliminary Knowledge

This section reviews the preliminary knowledge necessary to develop the SE DI and the LCS-EnsemNet, including a brief introduction to saliency information extraction and the temporal ensembling algorithm.

A. Saliency Information Extraction

In the proposed method, we exploit the context-aware saliency detection (CASD) [28] method to get the context-wise SE DI. The saliency information extraction process is reformulated with more detail as follows and shown in Fig. 1.

First of all, the log-ratio DI is computed by $\mathbf{D}_{\log} = |\log(\mathbf{I}_1/\mathbf{I}_2)|$, where $\mathbf{I}_1 \in \mathbb{R}^{H\times W}$ and $\mathbf{I}_2 \in \mathbb{R}^{H\times W}$ represent the prechange and postchange SAR images, respectively. The image $\mathbf{D}_{\log} \in \mathbb{R}^{H\times W}$ is used as the input of the CASD method for saliency information extraction.

1) Dissimilarity Measure: In essence, the saliency value evaluates the uniqueness of pixels. For a certain pixel i, its saliency can be evaluated by comparing it with all the pixels in the image, which can be formulated as follows:

$$S_i = \sum_{j=1}^{H\times W} dis(\mathbf{P}_i, \mathbf{P}_j) \tag{1}$$

where S_i denotes the saliency value at pixel i, $H \times W$ is the number of pixels in the image, $dis(\cdot)$ represents the dissimilarity measure of paired pixels, and each pixel is represented by the 7 × 7 image patch centered at this pixel, i.e., $\mathbf{P}_i$ and $\mathbf{P}_j$ respectively denotes the 7 × 7 image patch centered at pixels i and *j*, as shown in Fig. 1(a). For pixels located around the image edges, mirror padding is performed to allow the dense patch extraction at pixels near or at the image edges. Specifically, mirror padding captures image boundaries and then supplements them around the original image through mirroring, such that the image size is increased to $(H + 6) \times (W + 6)$. To simplify the calculation, only the L_{ms} most similar pixels rather than all the pixels are used to compute the saliency value. That is to say, if the most similar pixels are highly different from the pixel i, then clearly all the pixels in the image are highly different from it. Therefore, the saliency calculation is simplified as follows:

$$S_i = \sum_{n=1}^{L_{ms}} dis(\mathbf{P}_i, \mathbf{P}_{j_n}) \tag{2}$$

where the most similar pixels $\{j_n\}_{n=1}^{L_{ms}}$ of pixel i are found in the image according to the dissimilarity measure. Through

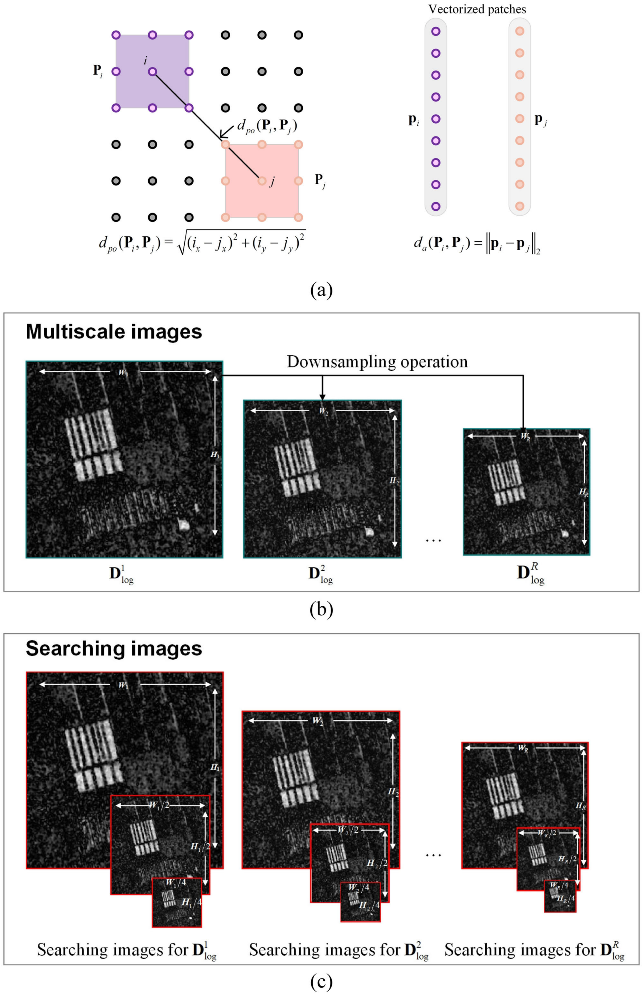


Fig. 1. Schematic illustration of the saliency detection. (a) Dissimilarity measure. (b) Multiscale images. (c) Searching images for saliency computation of each pixel.

searching for the most similar pixels throughout the entire image, global information is embedded. Then, how to define the dissimilarity measure is essential for the calculation of the saliency.

Intuitively, the dissimilarity measure evaluates how much different two pixels are and can be naturally defined as

$$d_a(\mathbf{P}_i, \mathbf{P}_j) = \|\mathbf{p}_i - \mathbf{p}_j\|_2 \tag{3}$$

where $\mathbf{p}_i$ and $\mathbf{p}_j$ denote the vectorized image patches and $d_a(\mathbf{P}_i, \mathbf{P}_j)$ refers to the Euclidean distance between $\mathbf{p}_i$ and $\mathbf{p}_j$, as shown in the right part of Fig. 1(a). It is noted that $d_a(\mathbf{P}_i, \mathbf{P}_j)$ represents the amplitude difference between image patches $\mathbf{P}_i$ and $\mathbf{P}_j$. With the dissimilarity $d_a(\mathbf{P}_i, \mathbf{P}_j)$, the L_{ms} most similar pixels $\{(j_n)\}_{n=1}^{L_{ms}}$ can be found in searching images.

In fact, there is a heuristic principle that the conspicuous pixels tend to be grouped together. Specifically, if pixel i is salient, the most similar pixels $\{(j_n)\}_{n=1}^{L_{ms}}$ will be close to it in position with a high probability; on the contrary, background nonsalient pixels tend to distribute all over the image and have similar pixels both near and far-away in entire image ("background nonsalient pixels" correspond to the unchanged pixels in the context of SAR image CD). Considering this point, positional regularization is introduced and the dissimilarity specialized for describing saliency can be rewritten as

$$dis(\mathbf{P}_i, \mathbf{P}_j) = \frac{d_a(\mathbf{P}_i, \mathbf{P}_j)}{1 + \eta \cdot d_{po}(\mathbf{P}_i, \mathbf{P}_j)} \tag{4}$$

where $dis(\mathbf{P}_i, \mathbf{P}_j)$ refers to the dissimilarity between patches $\mathbf{P}_i$ and $\mathbf{P}_j$, $d_{po}(\mathbf{P}_i, \mathbf{P}_j)$ represents the Euclidean distance between the positions of pixels i and j, and η is a balance factor controlling the influence of the positional distance. According to the above heuristic principle, in the saliency computation, the smaller positional distance between pixel i and its similar pixel j indicates a higher likelihood that pixel i is salient. Besides, as analyzed in [28], the variation of the balance factor η has little influence on the saliency detection results. Hence, η is set to 3, as suggested in [28].

2) Multiscale Saliency: As shown in Fig. 1(b), R images $\{\mathbf{D}_{\log}^r\}_{r=1}^R$ with different image sizes are obtained by applying a downsampling operation to $\mathbf{D}_{\log} \in \mathbb{R}^{H \times W}$ sequentially for multiscale saliency information extraction, where the scale 1 corresponds to the input image itself, i.e., $\mathbf{D}_{\log}^1 = \mathbf{D}_{\log}$, and $\{\mathbf{D}_{\log}^r\}_{r=2}^R$ are the $R-1$ downsampled images. Then, with the defined dissimilarity measure in (4), the saliency maps of images $\{\mathbf{D}_{\log}^r\}_{r=1}^R$ ($\mathbf{D}_{\log}^1 = \mathbf{D}_{\log}$) are computed, respectively.

Specifically, to compute the saliency map of $\mathbf{D}_{\log}^r$, for a certain pixel i in the image $\mathbf{D}_{\log}^r$, its L_{ms} most similar pixels $\{j_n\}_{n=1}^{L_{ms}}$ are found in the searching images $\{\mathbf{D}_{\log}^r, \mathbf{D}_{\log}^{r,1}, \mathbf{D}_{\log}^{r,2}\}$ according to $d_a(\mathbf{P}_i, \mathbf{P}_j)$. The image $\mathbf{D}_{\log}^{r,1}$ with a size of $H_r/2 \times W_r/2$ and the image $\mathbf{D}_{\log}^{r,2}$ with a size of $H_r/4 \times W_r/4$ are obtained by downsampling image $\mathbf{D}_{\log}^r$ with a size of $H_r \times W_r$, where $H_1 = H$ and $W_1 = W$, as shown in Fig. 1(c). The saliency value at the pixel i is computed as

$$S_i^r = 1 - \exp\left\{-\frac{1}{L_{ms}} \sum_{n=1}^{L_{ms}} dis(\mathbf{P}_i^r, \mathbf{P}_{j_n}^r)\right\} \tag{5}$$

where S_i^r is the saliency value of pixel i in the image $\mathbf{D}_{\log}^r$, $\mathbf{P}_i^r$ represents the patch centered at pixel i in the image $\mathbf{D}_{\log}^r$, and $\mathbf{P}_{j_n}^r$ represents the most similar patch centered at j_n obtained from the searching images.

To aggregate the multiscale information, the saliency maps $\{\mathbf{S}^r\}_{r=1}^R$ are upsampled to the image size of $H \times W$. Then, the across-scale saliency map aggregation can be formulated as

$$\mathbf{S} = \frac{1}{R} \sum_{r=1}^{R} \mathbf{S}^r \tag{6}$$

where $\mathbf{S}$ is the aggregated saliency map.

3) Context-Weighted Saliency: Spatial context implicitly characterizes the symbiotic relationships between the attention pixel and its surroundings by considering more positional information, which is profitable for describing the saliency information. The attention pixels can be located by searching for the pixels whose saliency value exceeds a predefined threshold τ_s, i.e., $\{j : S_j^r > \tau_s\}$. Then, for the pixel i, only its closest attention

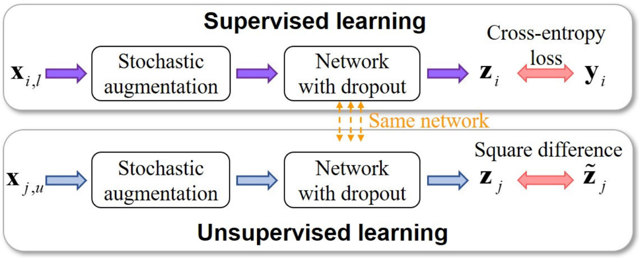


Fig. 2. Flowchart of temporal ensembling [35]. “temporal” in this article refers to the different training epochs of the neural network.

pixel $j_{closest}$ is used to define the context-related weight term, and the weighted saliency value for pixel *i* is as follows:

$$S_i = \frac{1}{R}\sum_{r=1}^{R} S_i^r \left(1 - d_{po}^r\left(i, j_{closest}\right)\right) \tag{7}$$

where $d_{po}^r(i, j_{closest})$ is the Euclidean distance between the positions of i and $j_{closest}$ in the image $\mathbf{D}_{\log}^r$. The context information can be reflected by the position-related weight term $1 - d_{po}^r(i, j_{closest})$, which means that the closer the point is to the attention pixel, the more possible the pixel is in the salient region of interest (ROI).

Thanks to its ability to capture the salient ROI and provide spatial context cues, saliency detection has been widely applied in CD, such as to segment the prominent changed regions to remove noisy background unchanged pixels and confusing pixels [19], [72], [73]. In this article, we use the publicly available code[1] and its default settings provided in [28] for salient change information extraction, which is subsequently utilized in the context-wise feature extraction.

B. Temporal Ensembling

Temporal ensembling [35] is one of the typical deep learning-based SSL algorithms. The core idea behind the algorithm is to accumulate the intermediate predictions of unlabeled samples over training epochs as category information to optimize the network. Owing to dropout regularization, the network architecture varies with the training iterations, which can be viewed as a series of individual networks. Based on these individual networks, ensemble predictions from multiple classifiers can be realized.

The network training is achieved with a labeled training set $\{\mathbf{x}_{i,l}, \mathbf{y}_i\}_{i=1}^{N_l}$ and an unlabeled training set $\{\mathbf{x}_{j,u}\}_{j=1}^{N_u}$, as illustrated in Fig. 2, where N_l and N_u denote the number of labeled and unlabeled training samples, respectively. In addition, $\mathbf{x}_{i,l}$ and $\mathbf{y}_i$ denote the *i*th labeled sample and its one-hot label vector, respectively. Stochastic augmentation [35] first imposes a random disturbance on input samples to improve robustness. Suppose that $\mathbf{z}_i$ is the predicted probability vector for the input sample $\mathbf{x}_{i,l}$. For supervised learning, the cross-entropy (CE) loss function is introduced as

$$\mathcal{J}_{\text{CE}} = -\frac{1}{N_l}\sum_{i=1}^{N_l} \mathbf{y}_i^T \log \mathbf{z}_i. \tag{8}$$

The CE loss guarantees that the network has an essential discrimination ability. It is widely used as classification loss in the deep learning community to train the network. For unsupervised learning, specifically, at the training epoch *t*, the predicted probability vector $\mathbf{z}_j^{(t)}$ for the *j*th unlabeled sample is first accumulated into an ensemble prediction $\mathbf{Z}_j^{(t)}$. The updating equation is defined as

$$\begin{aligned}\mathbf{Z}_j^{(t)} &= \lambda \mathbf{Z}_j^{(t-1)} + (1-\lambda)\mathbf{z}_j^{(t)} \\ &= (1-\lambda)\sum_{i=1}^{t} \lambda^{t-i}\mathbf{z}_j^{(i)}\end{aligned} \tag{9}$$

where $\mathbf{Z}_j^{(t)}$ is the accumulated ensemble prediction at epoch t and λ is the momentum term that controls the influence of preceding predictions on the current ensemble prediction. However, due to the initialization $\mathbf{Z}_j^{(0)} = 0$, i.e., $\mathbf{Z}_j^{(1)} = (1-\lambda)\mathbf{z}_j^{(1)}$, there is a startup bias that needs to be corrected. Thus, the corrected ensemble prediction is

$$\tilde{\mathbf{z}}_j^{(t)} = \mathbf{Z}_j^{(t)} \Big/ (1 - \lambda^t). \tag{10}$$

To enable the model to learn from more reliable samples, the mean square error between $\mathbf{z}_j^{(t)}$ and $\tilde{\mathbf{z}}_j^{(t)}$ is utilized as the unsupervised loss

$$\mathcal{J}_{\text{MSE}} = \omega(t)\frac{1}{N_u}\sum_{j=1}^{N_u}\left\|\mathbf{z}_j^{(t)} - \tilde{\mathbf{z}}_j^{(t)}\right\|^2 \tag{11}$$

where $\omega(t)$ is a Gaussian ramp-up function to weigh the loss $\mathcal{J}_{\text{MSE}}$. For a certain unlabeled input $\mathbf{x}_j$, the predictions at different training epochs are equivalent to classification results from different base classifiers since the dropout regularization randomly varies the network architecture during training. Significantly, through the loss function in (11), accumulated predictions $\tilde{\mathbf{z}}_j^{(t)}$, in which category information is more stable and reliable, are utilized as a target or teacher prediction of $\mathbf{z}_j^{(t)}$ to minimize the difference between them, assisting the network to learn category information from the reliable accumulated predictions and iteratively improving the generalization ability.

IV. Proposed Methodology

In this section, we propose a novel SAR image CD method, as depicted in Fig. 3. The proposed method contains two main modules: 1) the construction of the dual feature representation; and 2) an SSL module. The first module, illustrated in Fig. 3 (left), constructs pixel-wise and context-wise features to jointly describe the change information at each pixel. The dual features emphasize the image details and the spatial contexts, respectively. The second module, illustrated in Fig. 3 (middle), is the presented semisupervised label-consistent self-ensemble

[1] [Online]. Available: https://cgm.technion.ac.il/Computer-Graphics-Multimedia/Software/Saliency/Saliency.html

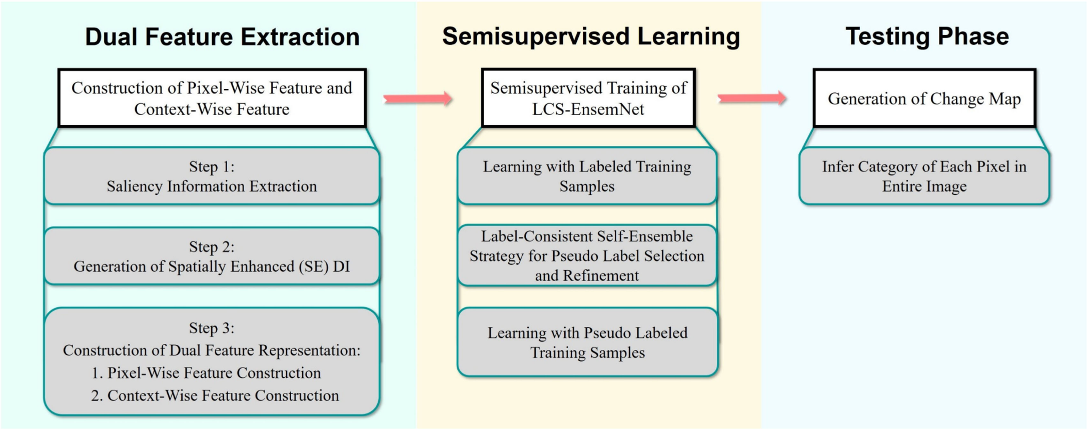

Dual Feature Extraction
Construction of Pixel-Wise Feature and Context-Wise Feature
Step 1: Saliency Information Extraction
Step 2: Generation of Spatially Enhanced (SE) DI
Step 3: Construction of Dual Feature Representation:
1. Pixel-Wise Feature Construction
2. Context-Wise Feature Construction
Semisupervised Learning
Semisupervised Training of LCS-EnsemNet
Learning with Labeled Training Samples
Label-Consistent Self-Ensemble Strategy for Pseudo Label Selection and Refinement
Learning with Pseudo Labeled Training Samples
Testing Phase
Generation of Change Map
Infer Category of Each Pixel in Entire Image

Fig. 3. Schematic overview of the proposed semisupervised SAR image CD method.

network, LCS-EnsemNet for short. This network can independently learn knowledge from the pixel-wise and context-wise features and, concurrently, refine the pseudo-labels for model learning, thus boosting the CD performance. These modules will be elaborated in the following subsections.

A. Dual Feature Representation Construction

The widely utilized comparative operators for the SAR image CD task, represented by the log-ratio operator, routinely measure dissimilarity pixel-by-pixel in the image domain for change information extraction. Such methods have high sensitivity to speckle noise, slight variation in viewing angles, and geometrical deformation. These problems could be mitigated by taking into account the spatial contextual information. Several previous works [19], [73], [74] make use of the saliency map for CD in SAR images by applying hard segmentation or thresholding to it to locate the conspicuous changed areas. However, directly segmenting the saliency map into a binary image would discard the useful spatial context information at each pixel and overlook many image details as well as tiny yet important changed areas. In the hyperspectral image CD works of literature [25], [51], [55], the spatial context of pixels has been exploited by using the multiscale nature of hierarchical feature representations [25] and multilayer convolutional neural network (CNN) architectures [51], [55], enabling effective detection and precise localization. Bovolo [25] presented a method to model the spatial context of pixels through multilevel hierarchical segmentation of multitemporal images. Recently, Saha *et al.* [51], [55] attempt to introduce deep learning techniques into DI generation by concatenating multiple deep feature maps from CNN to encode spatial context and then computing DIs in the deep feature domain. Such a deep-feature-based DI generation scheme can overcome the defects of the operators that compute DI in the image domain and achieve superior results when there are sufficient labeled training samples. However, sufficient labeled training samples are hard to acquire in practice. In the case of few labeled samples, training CNN for discriminative features is difficult. Considering the necessity of effective change features in SAR CD, we propose to exploit the saliency map to capture the spatial context cue and to consolidate the difference measure in the image domain with high efficiency, which is independent of network training.

Inspired by the works [25], [73], [74], a reweighting scheme is devised to effectively and quicky incorporate the pixel-wise log-ratio DI and its saliency map to produce an SE DI such that the spatial context in the saliency map can be encoded into the generated SE DI. Then, pixel-wise and context-wise features are extracted on the basis of the log-ratio DI and the SE DI, respectively, for a joint description of pixels in the bitemporal SAR images. The dual feature extraction is made up of three steps (see Fig. 3).

1) Saliency Information Extraction: To capture the spatial-contextual information, the CASD method is first applied to the log-ratio DI $\mathbf{D}_{\text{log}}$, as described in detail in Section III-A. By combining the multiscale spatial and context information, the saliency map $\mathbf{S}$ is derived, which models the spatial context information of pixels.

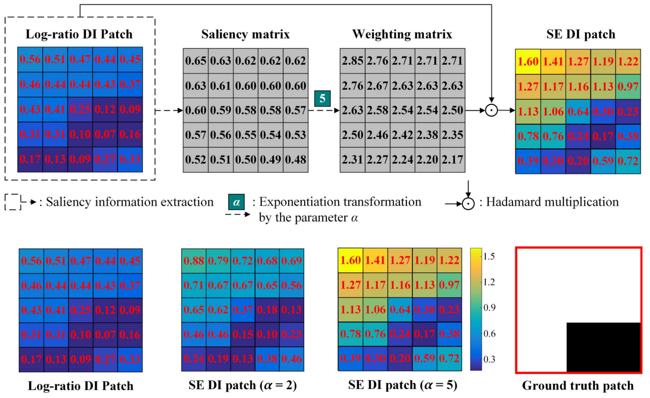
2) Generation of SE DI: Considering that the prominent changed regions have already been located in $\mathbf{S}$ with a high saliency value, we propose the DI reweighting scheme that is able to further highlight spatially salient changed pixels while suppressing background unchanged pixels as well as the speckle effects in the image. Instead of using the simple hard segmentation in [19], [73], and [74], the proposed reweighting scheme assigns varying weights to pixels in the DI $\mathbf{D}_{\text{log}}$. Specifically, the reweighting scheme produces the SE DI $\mathbf{D}_{\text{se}}$ by multiplying the input DI $\mathbf{D}_{\text{log}}$ with a weighting map $\mathbf{W}$ in an element-wise manner

$$\mathbf{D}_{\text{se}} = \mathbf{W} \odot \mathbf{D}_{\text{log}} \tag{12}$$

where $\mathbf{D}_{\text{se}}$ refers to the SE DI and symbol $\odot$ denotes the Hadamard product. The weighting map $\mathbf{W}$ is calculated by applying a nonlinear and monotonically increasing exponentiation transformation function to the saliency map $\mathbf{S}$

$$\mathbf{W} = \alpha^{\mathbf{S}} \, (\alpha > 1) \, . \tag{13}$$

Here, α is a parameter in the exponentiation transformation, which adjusts the intensity of the weighting map. We instantiate



Log-ratio DI Patch

0.56	0.51	0.47	0.44	0.45
0.46	0.44	0.44	0.43	0.37
0.43	0.41	0.25	0.12	0.09
0.31	0.31	0.10	0.07	0.16
0.17	0.13	0.09	0.27	0.33

Saliency matrix

0.65	0.63	0.62	0.62	0.62
0.63	0.61	0.60	0.60	0.60
0.60	0.59	0.58	0.58	0.57
0.57	0.56	0.55	0.54	0.53
0.52	0.51	0.50	0.49	0.48

Weighting matrix

2.85	2.76	2.71	2.71	2.71
2.76	2.67	2.63	2.63	2.63
2.63	2.58	2.54	2.54	2.50
2.50	2.46	2.42	2.38	2.35
2.31	2.27	2.24	2.20	2.17

SE DI patch

1.60	1.41	1.27	1.19	1.22
1.27	1.17	1.16	1.13	0.97
1.13	1.06	0.64	0.30	0.23
0.78	0.76	0.24	0.17	0.38
0.39	0.30	0.20	0.59	0.72

Log-ratio DI Patch

0.56	0.51	0.47	0.44	0.45
0.46	0.44	0.44	0.43	0.37
0.43	0.41	0.25	0.12	0.09
0.31	0.31	0.10	0.07	0.16
0.17	0.13	0.09	0.27	0.33

SE DI patch (α = 2)

0.88	0.79	0.72	0.68	0.69
0.71	0.67	0.67	0.65	0.56
0.65	0.62	0.37	0.18	0.13
0.46	0.46	0.15	0.10	0.23
0.24	0.19	0.13	0.38	0.46

SE DI patch (α = 5)

1.60	1.41	1.27	1.19	1.22
1.27	1.17	1.16	1.13	0.97
1.13	1.06	0.64	0.30	0.23
0.78	0.76	0.24	0.17	0.38
0.39	0.30	0.20	0.59	0.72

Fig. 4. Examples illustrating the reweighting scheme. First row: flowchart for reweighting on a localized region in Farmland A dataset, with an exponentiation parameter α equal to 5. The figure includes the log-ratio DI patch, the saliency matrix, the weighting matrix, and the SE DI patch. Second row: the log-ratio DI patch, the corresponding SE DI patches with parameter α equal to 2 and 5, and the ground truth patch, orderly. From the second row, it is clear that the reweighting scheme can effectively highlight the real changes, implying the increased class separability compared to the log-ratio DI. Moreover, under a larger value of parameter α, the highlighting effect is more significant. Refer Section V-A for more detailed descriptions of the Farmland A dataset.

the reweighting scheme in Fig. 4 by depicting the reweighting process and the corresponding results on a localized region in Farmland A dataset.

By leveraging this nonlinear exponentiation transformation, the reweighted difference values will be influenced by the saliency of the pixels such that the varying change information in different pixels can be modeled adaptively according to the corresponding saliency. The reweighting scheme aims to assign higher weights to those pixels that are more salient. In other words, if the saliency of a pixel is high, the reweighting scheme will assign a higher weight to the pixel and vice versa. Consequently, since the weights are exponentially proportional to the saliency value, the contrast between changed pixels (with high saliency values) and unchanged pixels (with low saliency values) dramatically increases. With respect to α, (12), (13) and Fig. 4 show that if the value of α is larger, then the intensity of the entire weighting map will be higher, and the contrast between the changed and unchanged regions can be improved further. Finally, the reweighted SE DI is obtained by choosing the suitable value of the parameter α. To further demonstrate the effectiveness of the proposed reweighting scheme, more results are shown in Fig. 5, in which an ROI is enlarged and shown with a red box for better visualization of the difference.

The reweighting scheme is task-oriented, which manages to effectively highlight the prominent changed pixels and increase the interclass separability. As a result, pixel-wise information in $\mathbf{D}_{\text{log}}$ and spatial-contextual information in $\mathbf{S}$ are effectively incorporated, making the pixels in the SE DI more spatially distinguishable. Compared to the existing pixel-wise comparative operators in SAR CD, spatial-contextual information is explored in an effective manner to quickly suppress the effects of speckle noise and enhance the contrast between changed and unchanged regions, thereby highlighting the changed regions and improving interclass separability.

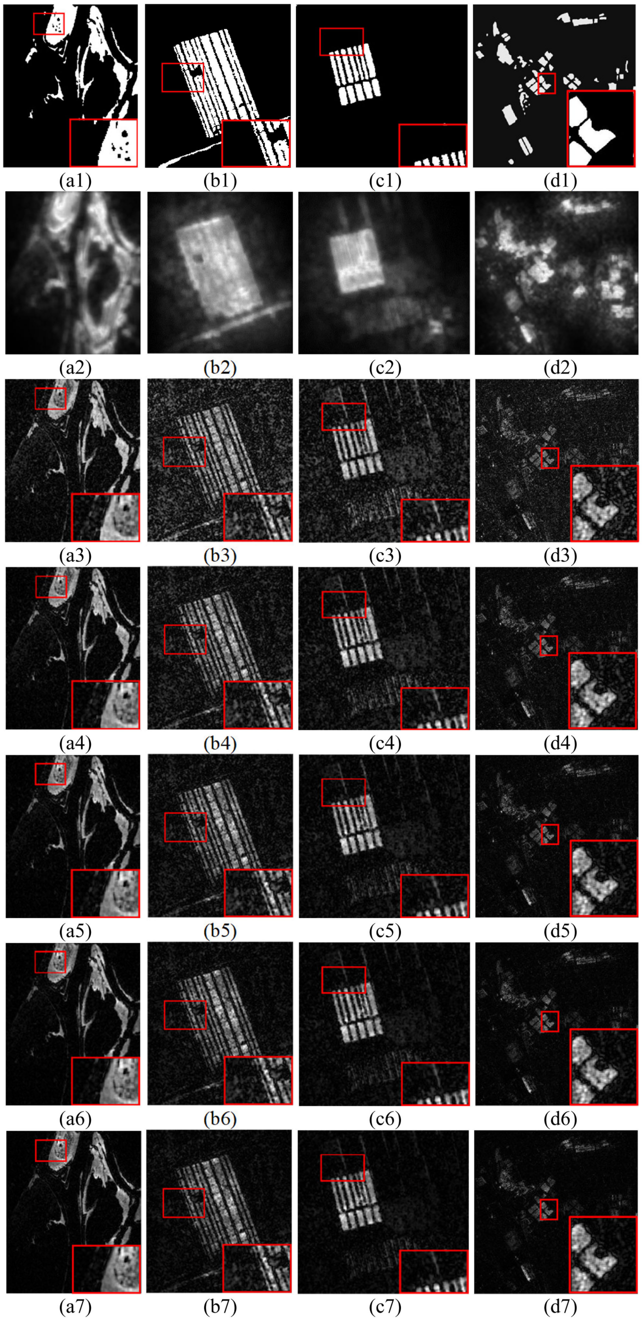


Fig. 5. SE DI on four real SAR image pairs. An ROI is enlarged and shown with a red box for better visualization of the difference. (a1)–(a7) Ottawa images. (b1)–(b7) Farmland A images. (c1)–(c7) Farmland B images. (d1)–(d7) Foshan city images. (a1)–(d1) Ground truth maps. (a2)–(d2) Saliency maps produced by the CASD method. (a3)–(d3) Log-ratio DI. (a4)–(d4) SE DI when $\alpha = 2$. (a5)–(d5) SE DI when $\alpha = 3$. (a6)–(d6) SE DI when $\alpha = 4$. (a7)–(d7) SE DI when $\alpha = 5$. Refer to Section V-A for more detailed descriptions of the four SAR datasets.

3) Construction of Dual Feature Representation: Thanks to the introduction of the saliency map, the reweighting scheme can well preserve the prominent changed regions, whereas the image details are sacrificed to a certain extent, resulting in ambiguity on borders. By contrast, the log-ratio operator calculates the difference values pixel-by-pixel such that the delicate image details can be well preserved. Accordingly, the information in the log-ratio DI and SE DI can be regarded as coming from different modalities, both of which are beneficial to better discriminate the changed pixels from unchanged backgrounds. As mentioned in [75], directly concatenating the features from different modalities into a single feature would discard the particularity of each modality.

In this work, we construct a dual feature representation for each sample. The dual feature representation is, in fact, a couple of feature vectors, including the pixel-wise feature vector and the context-wise feature vector. The pixel-wise feature vector is formed by the vectorized patches of the original SAR images and the log-ratio DI $\mathbf{D}_{\log}$, while the context-wise feature vector is formed by the vectorized patches of the original SAR images and the SE DI $\mathbf{D}_{\text{se}}$. Such cascade design makes it easier for discriminative feature extraction by providing indispensable change information such as log-ratio DI and SE DI and prevents information loss by integrating original SAR data. Moreover, the constructed dual feature vectors will be independently explored to extract their respective high-level semantic features. Complementary information in these two high-level features will be fused in decision to enforce the label consistency, refining pseudo-labels.

B. Proposed SSL Model: LCS-EnsemNet

Since usually only a limited number of labeled samples can be gathered manually through expert knowledge in real-world applications, SSL, which manages to make full use of both few labeled samples and abundant unlabeled samples, is a proper solution. Inspired by recent developments in deep learning-based semisupervised approaches [35], [36], [72], a novel network, namely, LCS-EnsemNet, is proposed to reliably predict pseudo-labels for unlabeled samples and efficiently learn more discriminative and generalized features from both the labeled and pseudo-labeled training data. In this manner, network optimization and pseudolabel refinement are coupled together to benefit each other in an iterative way.

In this section, we denote the labeled training set by $\Omega_l = \{(\mathbf{x}_{i,l}^{\text{pw}}, \mathbf{x}_{i,l}^{\text{cw}}), y_i\}_{i=1}^{N_l}$, where the subscript *i* represents the sample index, the subscript *l* refers to a labeled sample, and the superscript pw and cw are short for pixel-wise feature and context-wise feature, respectively. Similarly, we denote the unlabeled training set by $\Omega_u = \{(\mathbf{x}_{i,u}^{\text{pw}}, \mathbf{x}_{i,u}^{\text{cw}})\}_{i=1}^{N_u}$, where the subscript *u* refers to an unlabeled sample.

*1) Network Architecture:*The proposed LCS-EnsemNet, as depicted in Fig. 6, is composed of two multilayer perceptron subnets Φ and Ψ with identical structures. The first subnet Φ aims to learn the image detail knowledge, using the pixel-wise feature $\mathbf{x}_{i,l}^{\text{pw}}$ and $\mathbf{x}_{i,u}^{\text{pw}}$. The second subnet Ψconcentrates on learning spatial context knowledge, using the context-wise feature $\mathbf{x}_{i,l}^{\text{cw}}$ and $\mathbf{x}_{i,u}^{\text{cw}}$. To be specific, for the subnet in each branch, batch normalization (BN) layers first standardize the distributions of network inputs to smooth the optimization landscape of the loss function and then promote network training [77]. After the BN

Algorithm 1: Training Strategy.

```
Input:  labeled training set $\Omega_l = \{(\mathbf{x}_{i,l}^{\text{pw}}, \mathbf{x}_{i,l}^{\text{cw}}), y_i\}_{i=1}^{N_l}$;
        unlabeled training set $\Omega_u = \{(\mathbf{x}_{i,u}^{\text{pw}}, \mathbf{x}_{i,u}^{\text{cw}})\}_{i=1}^{N_u}$;
        moving window length $K$; threshold parameter $\tau_1$
        and $\tau_2$, number of training epoch $N_{epoch}$.
Initialization:  subnet weights $\theta_\Phi^0$ and $\theta_\Psi^0$, memory block.
  Supervised learning (only using $\Omega_l$):
  for $t = 1, 2, \ldots, K$ do
      use data $\Omega_l$ to train the subnet $\Phi$ and $\Psi$
      compute loss via (16)
      update $\theta_\Phi^{t-1}$ and $\theta_\Psi^{t-1}$ to $\theta_\Phi^t$ and $\theta_\Psi^t$
      compute predictions $\{\tilde{y}_i^{(t)}\}_{i=1}^{N_u}$ for $\Omega_u$ via (18),
      (19)
      store $\{\tilde{y}_i^{(t)}\}_{i=1}^{N_u}$ into the memory block
  end for
  Semisupervised learning:
  for $t = K+1, \ldots, N_{epoch}$ do
      initialize the pseudoset $\Omega_p^{(t)}$.
    for $i = 1, 2, \ldots, N_u$ do
        extract predictions $\{\tilde{y}_i^{(k)}\}_{k=t-K}^{t-1}$ from memory
        block.
        compute pseudolabel $\hat{y}_i^{(t)}$ via (20)–(22)
        add $\{(\mathbf{x}_{i,u}^{\text{pw}}, \mathbf{x}_{i,u}^{\text{cw}}), \hat{y}_i^{(t)}\}$ into $\Omega_p^{(t)}$
    end for
    use data $\Omega_l$ and $\Omega_p^{(t)}$ to train the subnet $\Phi$ and $\Psi$
    compute loss via (16), (23), (24)
    update $\theta_\Phi^{t-1}$ and $\theta_\Psi^{t-1}$ to $\theta_\Phi^t$ and $\theta_\Psi^t$
    compute prediction set $\{\tilde{y}_i^{(t)}\}_{i=1}^{N_u}$ via (18), (19)
    store $\{\tilde{y}_i^{(t)}\}_{i=1}^{N_u}$ into the memory block
  end for
Output:  trained subnet $\Phi$ and $\Psi$.
```

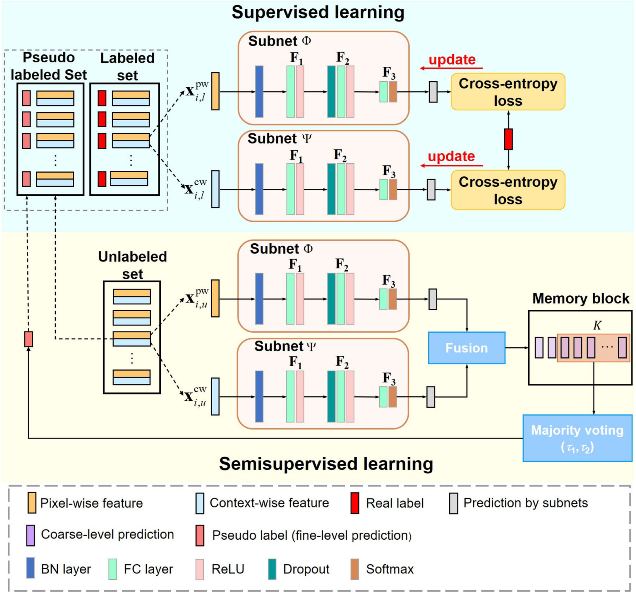


Fig. 6. Framework of the LCS-EnsemNet for SAR image CD.

layer, three fully connected (FC) blocks are cascaded to extract hierarchical features from inputs. For the blocks F_1 and F_3, the hidden features are extracted as follows:

$$\mathbf{h}_n = \sigma\left(\mathbf{W}_n \cdot \mathbf{h}_{n-1} + \mathbf{b}_n\right) \tag{14}$$

where $\mathbf{W}_n$ and $\mathbf{b}_n$ denote the weight matrix and bias vector of the *n*th FC block, respectively, $\mathbf{h}_n$ represents the output features of the *n*th FC block, and σ is a nonlinear activation function. The rectified linear unit (ReLU) [77] nonlinear function is used. Particularly, dropout regularization [37] is embedded in the F_2 block to change subnetwork architectures randomly. The hidden feature extracted by the block F_2 is as follows:

$$\mathbf{h}_2 = \sigma\left(\mathbf{W}_2 \cdot \left(\mathbf{d} \otimes \mathbf{h}_1\right) + \mathbf{b}_2\right) \tag{15}$$

where $\mathbf{d}$ is a random binary vector with which hidden units are dropped stochastically at probability β, and symbol $\otimes$ refers to the dropout operation. Thanks to dropout regularization, network architectures vary with the training epochs, which is equivalent to using multiple individual feature extraction and classification networks during the training. Similar to the bagging algorithm [77], predictions of multiple individual networks in training can be accumulated to realize a more reliable prediction for the pseudolabel.

2) Supervised Learning With Limited Labeled Data: Limited labeled data are used to supervise network learning for initial and essential discrimination ability. Given the labeled training set $\Omega_l = \{(\mathbf{x}^{\mathrm{pw}}_{i,l}, \mathbf{x}^{\mathrm{cw}}_{i,l}), y_i\}_{i=1}^{N_l}$, where $y_i = 0$ if the current sample is supposed to be unchanged and $y_i = 1$ if it is supposed to be changed, the subnetworks Φ and Ψ are trained on $\{\mathbf{x}^{\mathrm{pw}}_{i,l}, y_i\}_{i=1}^{N_l}$ and $\{\mathbf{x}^{\mathrm{cw}}_{i,l}, y_i\}_{i=1}^{N_l}$, respectively. The objective function of supervised training can be expressed as

$$\begin{aligned} J_{\mathrm{sup}} = \frac{1}{N_l}\sum\nolimits_{i=1}^{N_l} L_{bce}\left(\mathbf{x}^{\mathrm{pw}}_{i,l}, y_i; \Phi\right) \\ + \frac{1}{N_l}\sum\nolimits_{i=1}^{N_l} L_{bce}\left(\mathbf{x}^{\mathrm{cw}}_{i,l}, y_i; \Psi\right) \end{aligned} \tag{16}$$

where N_l is the number of labeled samples and L_{bce} denotes the binary cross-entropy loss, which is defined as

$$\begin{aligned} L_{bce}\left(\mathbf{x}^{\mathrm{pw}}_{i,l}, y_i\right) = & -y_i \log p\left({y'}_i = 1|\mathbf{x}^{\mathrm{pw}}_{i,l}\right) \\ & -(1-y_i)\log p\left({y'}_i = 0|\mathbf{x}^{\mathrm{pw}}_{i,l}\right). \end{aligned} \tag{17}$$

Here y'_i represents the predicted label, $p(y'_i = 1|\mathbf{x}^{\mathrm{pw}}_i)$ is the probability that the sample $\mathbf{x}^{\mathrm{pw}}_{i,l}$ is assigned to the changed class, and $p(y'_i = 0|\mathbf{x}^{\mathrm{pw}}_i)$ is the probability that the sample $\mathbf{x}^{\mathrm{pw}}_i$ is assigned to the unchanged class.

Through updating network parameters using the loss J_{sup}, the network gets the initial discrimination capability. However, training with a limited number of labeled samples has the risk of overfitting. In the next subsection, we will present how to use the label-consistent self-ensemble strategy to realize the SSL, to mitigate overfitting, and to improve generalization ability.

3) SSL With Label-Consistent Self-Ensemble Strategy: Reliable pseudo-labels are vital for SSL and the improvement of performance. Existing representative methods, such as the temporal ensembling [35], the mean teacher [36], and the MixMatch [72], usually utilize the predictions of all the unlabeled training samples as targets or teacher predictions to guide network training without any selection or refinement stage. In this way, incorrectly inferred predictions of hard-to-be-classified samples will hinder model learning and result in performance degradation. For this reason, it is necessary to improve the quality and reliability of pseudo-labels.

Toward this, we design a two-stage self-labeling strategy called label-consistent self-ensemble. The strategy aims to discover reliable category information from unlabeled data and refine the pseudo-labels by imposing both the label consistency between dual features and the label consistency across multiple classifiers. With this proposed strategy, LCS-EnsemNet can dynamically select a subset of reliable predictions along with their corresponding samples as pseudo-labeled samples, which are employed together with labeled training samples to further train the network in the next training epoch, improving the performance. The improved network can be used to update the pseudo-labels. In this way, the network can effectively learn useful discriminative knowledge and generalize to more diversified changed and unchanged patterns. Meanwhile, the network optimization and pseudolabel refinement are integrated into one framework and iteratively facilitate each other in a positive way. Next, we will introduce the details of the ensemble strategy.

a) Label Consistency Between Dual Features: Due to insufficient training with limited labeled samples, the network output may contain trustless predictions on unlabeled samples, especially for the hard-to-be-classified samples. To solve this problem, label consistency is imposed on the dual features. The subnets Φ and Ψ, respectively, take the pixel-wise and context-wise features as input to get their respective predictions. To make full use of the complementary information from the dual features, these two predictions are fused to get a single consistent prediction, which explicitly enforces the label consistency between the pixel-wise and context-wise features. Intuitively, combining the two predicted class probabilities at the decision level can improve the reliability of predicted labels to some extent.

Given the unlabeled set $\Omega_u = \{(\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u})\}_{i=1}^{N_u}$, the subnets Φ and Ψ output predicted probabilities $p_{\Phi_t}(y|\mathbf{x}^{\mathrm{pw}}_{i,u})$ and $p_{\Psi_t}(y|\mathbf{x}^{\mathrm{cw}}_{i,u})$, respectively, for the unlabeled sample pair $(\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u})$, where Φ_t and Ψ_t indicate the architectures of subnets Φ and Ψ at the epoch *t*. Using the soft majority voting, a coarse-level prediction is generated by aggregating the probability predictions at the class level as follows:

$$p_{\Phi_t,\Psi_t}\left(y'|\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u}\right) = \frac{p_{\Phi_t}({y'}_i|\mathbf{x}^{\mathrm{pw}}_{i,u}) + p_{\Psi_t}\left({y'}_i|\mathbf{x}^{\mathrm{cw}}_{i,u}\right)}{2} \tag{18}$$

$$\tilde{y}_i^{(t)} = \arg\max_{j\in\{0,1\}} p_{\Phi_t,\Psi_t}\left(y' = j|\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u}\right) \tag{19}$$

where $p_{\Phi_t}(y'_i|\mathbf{x}^{\mathrm{pw}}_{i,u})$ and $p_{\Psi_t}(y'_i|\mathbf{x}^{\mathrm{cw}}_{i,u})$ denote the predicted class probabilities by subnets Φ and Ψ, $p_{\Phi_t,\Psi_t}(y'|\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u})$ indicates the jointly predicted probability of the two subnets, and $\tilde{y}_i^{(t)}$ is the finally predicted label for $(\mathbf{x}^{\mathrm{pw}}_{i,u}, \mathbf{x}^{\mathrm{cw}}_{i,u})$ at the training epoch *t*, termed coarse-level prediction in this article.

With the decision fusion, the label consistency between the dual features is well preserved, ensuring the reliability of the predictions for unlabeled samples. More importantly, the pixel-wise and context-wise information is integrated at the class level, which is beneficial to suppress the effect of speckle while maintaining the image detail. Note that the coarse-level predictions of the same input during training are recorded in a memory block for subsequent refinement, as shown in Fig. 6.

b) Label Consistency Across Multiple Classifiers: Despite the improved reliability in the first stage, incorrect predictions may still exist and further refinement is required. Motivated by the consistency regularization in SSL that encourages consistent predictions when inputs or models are perturbed [78], we introduce the label consistency across multiple classifiers to further refine the predictions. Multiple classifiers are obtained by using dropout regularization in the two subnets Φ and Ψ. The coarse-level predictions of the two subnets under different dropout units are ensembled into a consistent prediction, i.e., the pseudolabel. In other words, the label consistency across multiple classifiers is intrinsically similar to the ensemble of multiple classifiers.

To achieve this, inspired by the temporal ensemble, dropout regularization is embedded to randomly change the network architectures, as shown in Fig. 6, which means that the network structures at consecutive training epochs are equivalent to multiple different classifiers. As mentioned earlier, a memory block is constructed to record a series of coarse-level predictions of each input that are generated by the networks at each training epoch. The recorded predictions are used to achieve the ensemble of multiple classifiers. To be specific, a moving window is built to progressively leverage *K* coarse-level predictions from the recent *K* training epochs to generate the pseudolabel, which shares the same spirit as the moving average in the temporal ensemble method. That is to say, at the epoch *t*, only the coarse-level predictions from recent *K* training epochs (i.e., training epochs $[t-K, t-1]$) are used since the predictions in early training epochs may be incorrect due to the insufficient training of the network.

In the *t*th epoch, for an unlabeled samples-pair $(\mathbf{x}_{i,u}^{\mathrm{pw}}, \mathbf{x}_{i,u}^{\mathrm{cw}})$, we will have $t-1$ coarse-level labels, i.e., $\{\tilde{y}_i^{(1)}, \tilde{y}_i^{(2)}, \ldots, \tilde{y}_i^{(t-1)}\}$. Like the moving average of accumulated predictions in the temporal ensemble [35], in our method, a pseudolabel is estimated using the prediction series $\{\tilde{y}_i^{(t-K)}, \tilde{y}_i^{(t-K+1)}, \ldots, \tilde{y}_i^{(t-1)}\}$ only from the recent *K* epochs. In order to get reliable pseudo-labels, we devise a majority voting rule to find the samples with consistent labels in the prediction series and reject those with inconsistent labels

$$V_c = \sum\nolimits_{k=t-K}^{t-1} \mathbb{I}\left(\tilde{y}_i^{(k)} = 1\right) \tag{20}$$

$$V_u = \sum\nolimits_{k=t-K}^{t-1} \mathbb{I}\left(\tilde{y}_i^{(k)} = 0\right) \tag{21}$$

$$\hat{y}_i^{(t)} = \begin{cases} 1, & V_c \geq \tau_1 \\ 0, & V_u \geq \tau_2 \\ reject, & otherwise \end{cases} \tag{22}$$

where $\hat{y}_i^{(t)}$ is the refined pseudolabel for the *i*th unlabeled sample pair at epoch *t*. $\mathbb{I}(\cdot)$ is an indicator function. Particularly, it equals 1 if the input is true; otherwise, it equals 0. Besides, V_c and V_u represent the number of votes for changed and unchanged classes, respectively. τ_1 and τ_2 refer to the thresholds for V_c and V_u, respectively. It is noticed that V_c and V_u satisfy $V_c + V_u = K$.

Naturally, the thresholds τ_1 and τ_2 need to be set to a value larger than $K/2$. Thus, $V_c \geq \tau_1$ or $V_u \geq \tau_2$ indicates that the recorded coarse-level predictions within the prediction series $\{\tilde{y}_i^{(t-K)}, \tilde{y}_i^{(t-K+1)}, \ldots, \tilde{y}_i^{(t-1)}\}$ are remarkably consistent. In other words, the *K* classifiers agree on their predictions of the current unlabeled sample pair. Thus, the corresponding label is given to the sample pair. In another case, when $V_c < \tau_1$ and $V_u < \tau_2$, the coarse-level predictions within the prediction series are inconsistent, indicating that the predictions are highly uncertain and should be discarded to prevent incorrect prediction. Finally, assisted with the two-stage refinement strategy, the pseudo-labeled training set $\Omega_p^{(t)} = \{(\hat{\mathbf{x}}_{i,p}^{\mathrm{pw}}, \hat{\mathbf{x}}_{i,p}^{\mathrm{cw}}), \hat{y}_i\}_{i=1}^{N_p^{(t)}}$ at epoch *t* is built, where $(\hat{\mathbf{x}}_{i,p}^{\mathrm{pw}}, \hat{\mathbf{x}}_{i,p}^{\mathrm{cw}})$ is the *i*th selected pseudo-labeled sample pair, $\hat{y}_i$ is the corresponding pseudolabel, and $N_p^{(t)}$ is the number of pseudo-labeled sample pairs.

The reliable pseudo-labeled set $\Omega_p^{(t)} = \{(\hat{\mathbf{x}}_{i,p}^{\mathrm{pw}}, \hat{\mathbf{x}}_{i,p}^{\mathrm{cw}}), \hat{y}_i\}_{i=1}^{N_p^{(t)}}$ can be used for model learning, likewise labeled training set. The objective function for the pseudo-labeled sample set can be formulated as

$$J_{\mathrm{semi}}^{(t)} = \frac{1}{N_p^{(t)}} \sum\nolimits_{i=1}^{N_p^{(t)}} L_{bce}\left(\hat{\mathbf{x}}_{i,p}^{\mathrm{pw}}, \hat{y}_i; \Phi_t\right) + \frac{1}{N_p^{(t)}} \sum\nolimits_{i=1}^{N_p^{(t)}} L_{bce}\left(\hat{\mathbf{x}}_{i,p}^{\mathrm{cw}}, \hat{y}_i.; \Psi_t\right) \tag{23}$$

Then the LCS-EnsemNet is optimized using the following total loss:

$$J^{(t)} = J_{\mathrm{sup}} + J_{\mathrm{semi}}^{(t)}. \tag{24}$$

It should be noted that $J^{(t)}$ varies with the training epochs because the pseudotraining set is changeable during the training procedure. The training strategy is described in detail in Algorithm 1, where θ_Φ^t and θ_Ψ^t represent all the weights and bias of the subnets Φ and Ψ at the epoch *t*.

By taking advantage of label consistency in estimating and selecting the pseudo-labels, more reliable category information in unlabeled data can be captured, and concurrently, the reliable pseudosamples are utilized to learn more generalized features and more accurate decision boundary. Compared to the existing semisupervised algorithms, the two-stage strategy imposes both the label consistency between dual features and the label consistency across multiple classifiers on the network predictions such that pseudo-labels can be reliably refined and selected and then provided to the next training epoch of the model to improve the detection performance. On the other hand, the sample pairs with inconsistent recorded predictions are regarded as uncertain ones and discarded to prevent the network from learning with wrongly predicted labels and avoid performance drop.

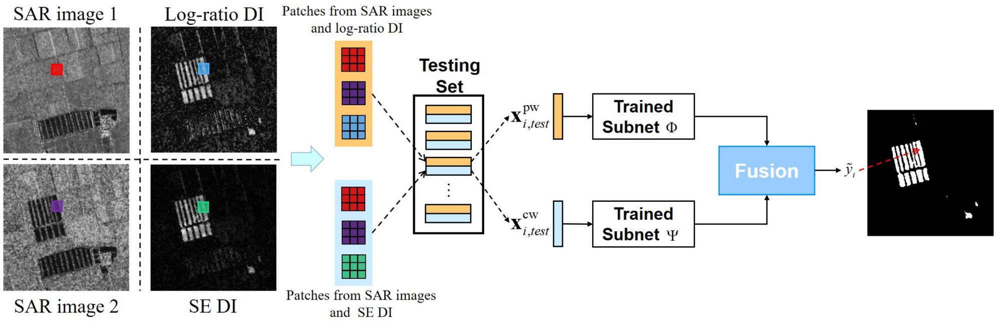

SAR image 1
Log-ratio DI
Patches from SAR images and log-ratio DI
Testing Set
$\mathbf{x}^{\text{pw}}_{i,test}$
Trained Subnet Φ
Fusion
$\tilde{y}_i$
$\mathbf{x}^{\text{cw}}_{i,test}$
Trained Subnet Ψ
SAR image 2
SE DI
Patches from SAR images and SE DI

Fig. 7. Inference of the pixel category in the testing phase. The cross-entropy loss and the memory block are removed. The category of each testing sample pair is inferred by the trained network with decision fusion.

C. Generation of the Final Change Map

As shown in Fig. 7, to generate the final change map, the testing set $\Omega_{test} = \{(\mathbf{x}^{\text{pw}}_{i,test}, \mathbf{x}^{\text{cw}}_{i,test})\}_{i=1}^{H\times W}$ is constructed in the same way as the training set construction, where the sample pair of all the pixels in the image is included. With the trained LCS-EnsemNet, the inferred label $\tilde{y}_i$ for each sample pair is obtained by fusing the predicted probabilities from the two subnets, i.e.,

$$p_{\Phi,\Psi}(y|\mathbf{x}^{\text{pw}}_{i,test}, \mathbf{x}^{\text{cw}}_{i,test}) = \frac{p_\Phi(y|\mathbf{x}^{\text{pw}}_{i,test}) + p_\Psi(y|\mathbf{x}^{\text{cw}}_{i,test})}{2} \tag{25}$$

$$\tilde{y}_i = \arg\max_{j\in\{0,1\}} p_{\Phi,\Psi}(y=j|\mathbf{x}^{\text{pw}}_{i,test}, \mathbf{x}^{\text{cw}}_{i,test}). \tag{26}$$

The inferred label $\tilde{y}_i$ at each pixel position forms the final change map.

V. Experimental Results

In this section, we evaluate the performance of the proposed semisupervised method through extensive experiments on four real SAR datasets. The datasets, evaluation criteria, parameter analysis, and experimental results are described in detail. All the experiments are conducted on a workstation with an Intel(R) Core(TM) i7-8750 H CPU (6 cores, 2.2 GHz, 32 GB RAM) and an Nvidia Quadro P2000 graphical processing unit (GPU) (4 GB RAM). The proposed model is performed using the Chainer-GPU (ver. 7.2.0) deep learning platform [79] and MATLAB 2016a in Windows 10 environment. The corresponding code of the proposed method will be made available at https://github.com/CATJianWang/LCS-EnsemNet.

A. Dataset Description

The performance of the proposed method is evaluated on four real SAR image datasets, which were acquired by Radarsat-1, Radarsat-2, and TerraSAR-X sensors. They are described in the following.

1) *Ottawa* Dataset: The dataset contains a pair of real SAR images of spatial size 290 × 350. They have 10m spatial resolution. The image pairs were acquired in July and August 1997, respectively, by the Radarsat-1 sensor. Fig. 8(a) and (b) shows the bitemporal SAR images and Fig. 8(c) shows the ground truth map. The dataset reflects the flooded areas over Ottawa, Canada.

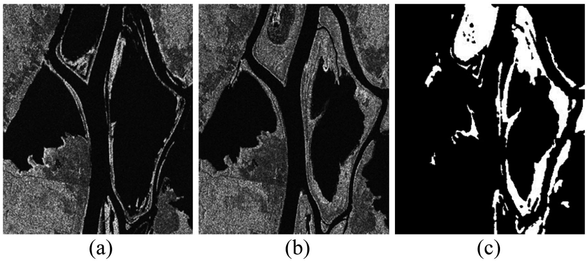


Fig. 8. Ottawa dataset. (a) Image acquired in July 1997. (b) Image acquired in August 1997. (c) Ground truth.

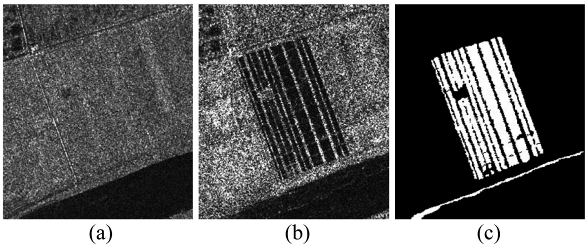


Fig. 9. Farmland A dataset. (a) Image acquired in June 2008. (b) Image acquired in June 2009. (c) Ground truth.

2) *Farmland A* dataset: This dataset was acquired in June 2008 and June 2009 by the Radarsat-2 sensor over a farmland area at the Yellow River Estuary in China. The bitemporal SAR images have 8m spatial resolution and the size of the bitemporal SAR images is 287 × 259 pixels. The SAR images and the ground truth map are shown in Fig. 9. This dataset mainly contains farmland changes near the Yellow River Estuary.

3) *Farmland B* dataset: The dataset contains a pair of SAR images acquired in June 2008 and June 2009. They have 8m resolution and a size of 291 × 306 pixels. SAR images and the ground truth map are shown in Fig. 10. Similar to the Farmland A dataset, the dataset reflects land cover changes near the Yellow River Estuary in China.

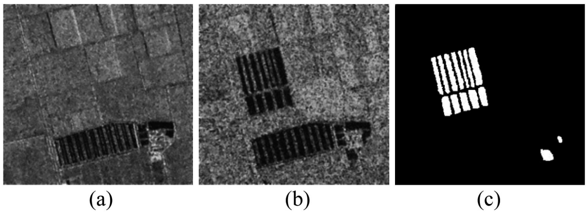


Fig. 10. Farmland B dataset. (a) Image acquired in June 2008. (b) Image acquired in June 2009. (c) Ground truth.

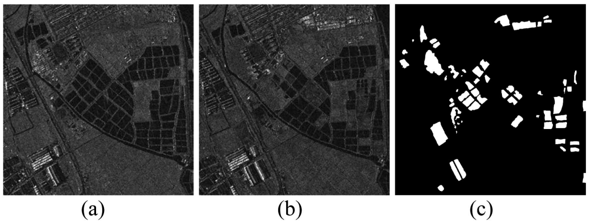


Fig. 11. Foshan City dataset. (a) Image acquired in May 2008. (b) Image acquired in December 2008. (c) Ground truth.

TABLE II
DATASETS DESCRIPTION

Data Set	Satellite	Spatial resolution	Change Type
Ottawa	Radarsat-1	10m	Flooding
Farmland A	Radarsat-2	8m	Cultivation
Farmland B	Radarsat-2	8m	Cultivation
Foshan City	TerraSAR-X	3m	Landfills of farmland; Farmland changed to buildings or reversed

4) *Foshan City* dataset: This dataset consists of two SAR images. They were acquired over Foshan City (Guangdong Province, China) in May 2008 and December 2008 by the TerraSAR-X sensor. The size of the SAR images is 1536 × 1536 pixels and the spatial resolution is 3m. As shown in Fig. 11, they present the landfills of farmland and farmland changes.

Note that in the Farmland A and Farmland B datasets, the images acquired in 2008 are four-look, but the ones obtained in 2009 are single-look, indicating the discrepancy of the impact of speckle noise on the bitemporal images. Naturally, the discrepancy increases the difficulty in the CD task. The descriptions for all the datasets are given in Table II.

In our experiments, sample pairs are constructed in the way described in Fig. 7 to represent each pixel position in SAR datasets and then utilized for training and testing. For each dataset, 10 000 sample pairs (5000 sample pairs for each class) are selected randomly for training. The proportion of labeled sample pairs is set to 0.3% to verify the performance of the proposed method in the case of a few labeled data (performance variation with different proportions of labeled and unlabeled sample pairs is presented in Sections V-D and V-E). For testing, the final change map is formed by the inferred category of each pixel in the image.

TABLE III
SUBNETWORK ARCHITECTURE

Layer	Description
Input layer	Feature vector with the size of $3h^2$
Normalization layer	BN layer
FC 1	1500-dim FC layer
Activation layer 1	ReLU
Dropout layer	Dropout with $\beta = 0.2$
FC 2	1500-dim FC layer
Activation layer 2	ReLU
FC 3	2-dim FC layer
Output layer	Softmax layer

B. Experimental Setup

1) Evaluation Criteria: To quantitatively evaluate the proposed method and compare the performance of different approaches, false positives (FP), false negatives (FN), percentage of correct classification (PCC), overall error (OE), and Kappa coefficient (κ) [80] are employed as the evaluation criteria. The PCC, OE, and κ are calculated as follows:

$$\mathrm{PCC} = (\mathrm{TP} + \mathrm{TN})/(\mathrm{TP} + \mathrm{FP} + \mathrm{TN} + \mathrm{FN}) \tag{27}$$

$$\mathrm{OE} = 1 - \mathrm{PCC} \tag{28}$$

$$\kappa = (\mathrm{PCC} - \mathrm{PRE})/(1 - \mathrm{PRE}) \tag{29}$$

$$\mathrm{PRE} = \frac{(\mathrm{TP}+\mathrm{FP}) \times (\mathrm{TP}+\mathrm{FN}) + (\mathrm{TN}+\mathrm{FN}) \times (\mathrm{TN}+\mathrm{FP})}{(\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN})^2}. \tag{30}$$

Here, TP is the number of pixels that belong to the changed class in the ground truth and are also correctly detected as the changed class, while TN is the number of pixels that belong to the unchanged class in the ground truth and are also correctly detected as the unchanged class.

2) Network Architecture: The architectures of the two subnetworks in the LCS-EnsemNet are made to be identical, as displayed in Table III. Due to the small size of the input samples, the network is composed of only three FC blocks, each comprising an FC layer as well as a rectified linear unit (ReLU) activation function [77], [81] (for the first two blocks) or a softmax function [77] (for the last block). Especially, the dropout regularization is embedded in the second FC block, enabling the network to have different architectures at each training epoch.

C. Parameter Analysis

In the proposed method, there are five essential parameters that need to be fixed, including the image patch size h, the weighting parameter α, the length of sliding window K, and the refinement thresholds τ_1 and τ_2. Extensive experiments are conducted to determine the value of these parameters.

Patch Size: As described in Section IV, each pixel-wise or context-wise feature vector is formed by three vectorized image patches of the same size. Thus, the determination of the feature

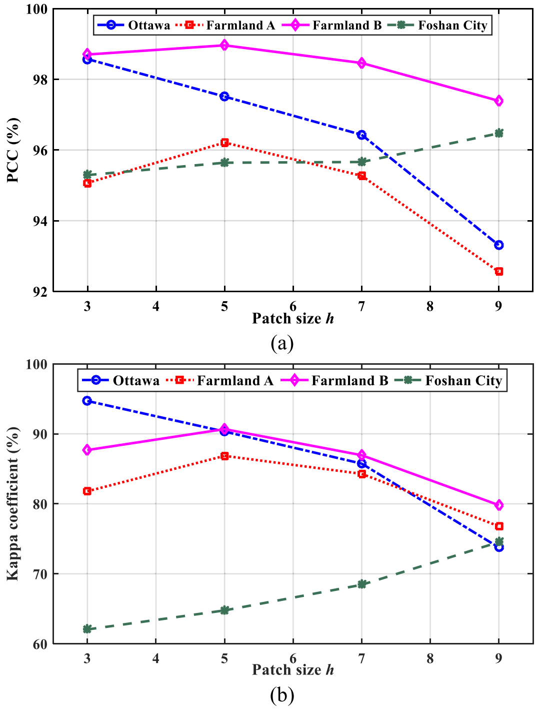


Fig. 12. PCC and Kappa coefficient (κ) vary with different patch size h.

TABLE IV
ENL ON DIFFERENT DATASETS

Data Set	Ottawa	Farmland A	Farmland B	Foshan City
ENL 1	17.89	14.18	20.77	4.61
ENL 2	14.09	5.76	9.28	4.29

TABLE V
PARAMETER SETTING IN EXPERIMENTS

Data Set	Ottawa	Farmland A	Farmland B	Foshan City
$h \times h$	3×3	5×5	5×5	9×9
α	3	5	5	5
K	10	10	10	10
τ_1	10	10	10	10
τ_2	7	9	7	9

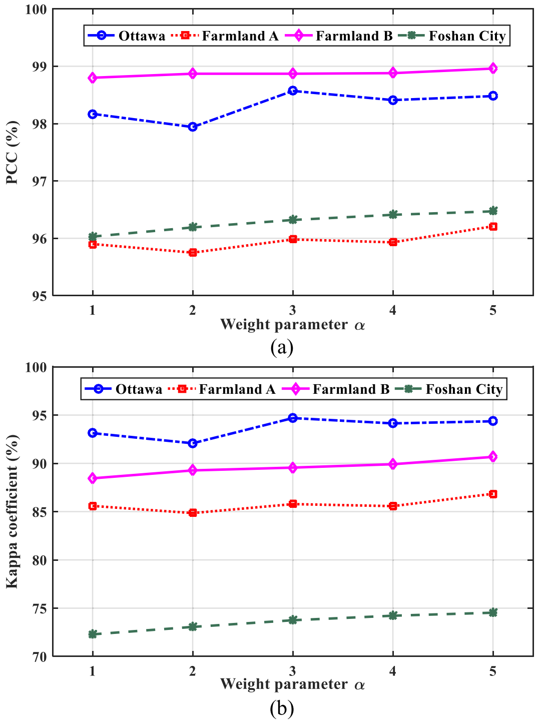


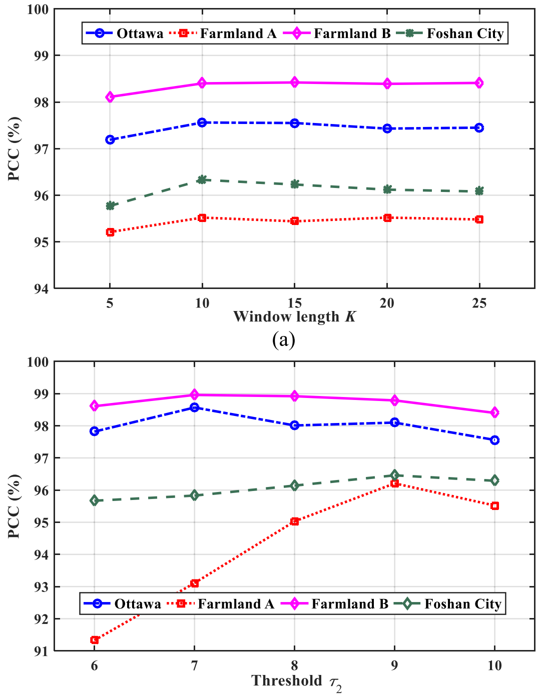
Fig. 13. Relation of the (a) PCC and (b) Kappa coefficient (κ) with the parameter α.

vector size can be implemented by investigating the effect of patch size h on detection results. Here, we set the patch size h as $3, 5, 7, 9$ in the feature construction and analyze the performance on the four SAR datasets. We evaluate the performance of the proposed method under different patch sizes in terms of the Kappa coefficient κ and PCC. The corresponding results are shown in Fig. 12. The results are compliant with the intuitive understanding that detection performance is sensitive to the variation of patch size. When the level of speckle is strong, a large patch size would weaken the speckle but remove image details. By contrast, a small patch size will well preserve image details but cause false alarms in results. To analyze the effect of speckle noise level on the patch size, the level of speckle noise in SAR images is measured by the equivalent number of looks (ENL) [13]. ENL value is given by

$$\mathrm{ENL} = \frac{\mu^2}{\mathrm{var}} \tag{31}$$

where μ denotes the mean value and var denotes the variance in the homogeneous region in the image. The ENL values of paired images in the four SAR datasets are listed in Table IV. According to the ENL values in Table IV, it can be seen that the images in Farmland A, Farmland B, and Foshan City dataset have stronger speckle noise. And it is also noticed in Fig. 12 that a larger patch size yields better performance for these three datasets. This implies that a comparatively larger patch size should be used for the dataset with stronger speckle noise. Furthermore, according to the results shown in Fig. 12, the patch size is also associated with spatial resolution. In high-resolution SAR images, there are many inhomogeneous regions, and the inhomogeneity may cause false detections and should be suppressed. In this case, a larger patch size should be selected. To sum up, the results in Fig. 12 illustrate that small patch size is appropriate for the low/medium-resolution images with a lower level of speckle noise, whereas a large patch size is necessary for large-scale high-resolution SAR images with stronger speckle noise. Accordingly, the suitable values of patch size for different datasets are listed in Table V.

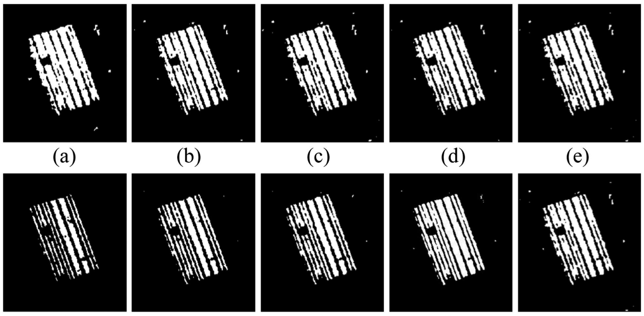
Parameter α: The parameter α in (13) is important for the quality of SE DI. To select the value of parameter α, the results of the proposed method on four real datasets are analyzed. Fig. 13 shows the impact of α on detection performance. By combining with the ENL values given in Table IV, it is noticed that the selection of parameter α has a close relationship with



(a)

(b)

Fig. 14. Performance by varying (a) window length K and (b) threshold parameter τ_2.



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

(f) (g) (h) (i) (j)

Fig. 15. Examples of CD results for Farmland A dataset with varied K and τ_2. (Top row) Results by varying K. (a) $K = 5$. (b) $K = 10$. (c) $K = 15$. (d) $K = 20$. (e) $K = 25$. (Bottom row) Results by varying τ_2. (f) $\tau_2 = 6$. (g) $\tau_2 = 7$. (h) $\tau_2 = 8$. (i) $\tau_2 = 9$. (j) $\tau_2 = 10$.

the level of speckle noise in bitemporal images. According to the results in Figs. 5 and 13, for the Farmland A, Farmland B, and Foshan City datasets corrupted by strong speckle noise, a large value of α can lead to the strong contrast between changed and unchanged classes, increase interclass separability, and weaken the speckle noise effect significantly; for the Ottawa dataset with weak speckle noise, a small value of α can mitigate the information loss induced by the saliency-based reweighting scheme. Consequently, the value of α should be determined according to the speckle noise level in images. For data suffering from strong noise, such as Farmland A, Farmland B, and Foshan City, α is set to 5 to enforce the effect of noise suppression; otherwise, an intermediate value of 3 is preferred. The suitable value of α is listed in Table V.

Parameters in the Label-Consistent Self-Ensemble Strategy: The selection of the window length K and thresholds τ_1 and τ_2 in the label-consistent self-ensemble strategy is important in SSL, directly impacting the reliability of pseudo-labels. For the changed class, the corresponding threshold τ_1 should be set to a large value to ensure the reliability and accuracy of the pseudo-labels of the changed class, preventing the false alarms caused by wrong predictions. Accordingly, the threshold τ_1 is set to be the same value as the window length K in our experiments. Extensive experiments are conducted to investigate the effect of the sliding window length K and the threshold τ_2 on the proposed network.

We varied the sliding window length K by fixing $\tau_2 = K$. The results in Fig. 14(a) indicate that the length of the sliding window affects the network learning and detection performance. A window length K of 10 provides the best performance for most datasets. A long window would contain more predictions by the early trained network, which may be incorrect and result in performance degradation. A short window would contain less recorded label information for pseudolabel refinement, which also has negative effect on network training mainly due to the reduced number of ensemble classifiers [see Figs. 14(a) and 15(a)–(e)]. Therefore, an intermediate value of 10 is selected for the window length K.

We varied the threshold τ_2 by fixing $K = 10$. According to the results in Fig. 14(b), the selection of τ_2 is associated with the scene complexity and speckle noise of the original SAR data. For the Farmland A and Foshan City datasets with a complicated scene and a strong level of speckle, the threshold τ_2 should have a large value of 9 to prevent the incorrect predictions from hard-to-be-classified background unchanged sample pairs that are easily classified into changed class. Due to the weak speckle noise in the Ottawa dataset and the simple scene in the Farmland B dataset, the hard-to-be-classified sample pairs appear relatively less. Therefore, a slack value of 7 is more appropriate for τ_2 in the Ottawa and Farmland B datasets to add more unchanged sample pairs into the pseudo-labeled set [see Figs. 14(b) and 15(f)–(j)].

D. Experiments With Different Proportions of Labeled Samples

To quantify the performance gain of the LCS-EnsemNet compared to the related models, including the subnetwork Φ (trained with labeled pixel-wise features solely), subnetwork Ψ (trained with labeled context-wise features solely), and network $\Phi + \Psi$ (only with fusion on the outputs of supervised models Φ and Ψ), the CD results with different proportions of labeled training samples are obtained. In experiments, we vary the proportion of labeled training samples to investigate the effect of labeled and unlabeled samples on the detection performance. The related models Φ, Ψ and $\Phi + \Psi$ are trained on labeled features in a supervised way.

As given in Tables VI–IX, the proposed LCS-EnsemNet yields better detection results than supervised models, which

TABLE VI
CHANGE DETECTION ACCURACY (PCC) (%) COMPARISON ON OTTAWA DATASET UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES

Proportion (10000 samples)	LCS-EnsemNet	Φ	Ψ	$\Phi + \Psi$
0.3%	98.57	97.76	97.68	97.78
1.0%	98.66	98.11	98.07	98.13
2.0%	98.82	98.23	98.14	98.24
3.0%	98.91	98.35	98.31	98.39
4.0%	98.92	98.34	98.32	98.38
5.0%	98.81	98.49	98.45	98.54

TABLE VII
CHANGE DETECTION ACCURACY (PCC) (%) COMPARISON ON FARMLAND A DATASET UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES

Proportion (10000 samples)	LCS-EnsemNet	Φ	Ψ	$\Phi + \Psi$
0.3%	96.21	95.36	95.41	95.44
1.0%	96.20	95.47	95.55	95.57
2.0%	96.31	95.53	95.57	95.61
3.0%	96.36	95.64	95.67	95.69
4.0%	96.43	95.65	95.69	95.72
5.0%	96.43	95.76	95.81	95.84

TABLE VIII
CHANGE DETECTION ACCURACY (PCC) (%) COMPARISON ON FARMLAND B DATASET UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES

Proportion (10000 samples)	LCS-EnsemNet	Φ	Ψ	$\Phi + \Psi$
0.3%	99.02	96.41	96.47	96.55
1.0%	99.02	97.51	97.49	97.53
2.0%	99.04	98.09	98.16	98.18
3.0%	99.01	98.22	98.28	98.33
4.0%	99.02	98.33	98.39	98.43
5.0%	99.06	98.52	98.56	98.57

TABLE IX
CHANGE DETECTION ACCURACY (PCC) (%) COMPARISON ON FOSHAN CITY DATASET UNDER DIFFERENT PROPORTIONS OF LABELED SAMPLES

Proportion (10000 samples)	LCS-EnsemNet	Φ	Ψ	$\Phi + \Psi$
0.3%	96.47	95.26	95.41	95.52
1.0%	96.55	95.35	95.56	95.63
2.0%	96.63	95.41	95.59	95.66
3.0%	96.77	95.52	95.67	95.72
4.0%	96.79	95.63	95.66	95.71
5.0%	96.82	95.69	95.72	95.75

validate the improvement of the label-consistent self-ensemble strategy and the capability of the proposed network in discovering category information from unlabeled data. According to the results, the LCS-EnsemNet can achieve high accuracy at the labeled data proportion of 0.3% and outperform the supervised models. The improvement of the proposed network is more significant when the proportion or number of labeled samples is smaller. The results are better with the increase of labeled samples, which also indicate that labeled data have a significant impact on the performance in the semisupervised model.

E. Sensitivity Analysis on Unlabeled Training Set Size

SSL is known for its capability of exploring discriminative information from unlabeled samples. In this section, to analyze the influence of unlabeled samples on the proposed method, 15 labeled samples per class and different numbers of unlabeled samples are randomly selected as the training set. The CD results are presented in the form of PCCs and Kappa statistics κ in Table X.

From Table X, we observe that the PCC and κ increase rapidly along with the number increase of unlabeled samples on the considered datasets, which indicates that the proposed method can effectively discover meaningful information from unlabeled samples for the network training. Furthermore, the PCC and κ become stable when the unlabeled training set size reaches a certain extent, which implies that the discriminative information and the diversity of underlying sample patterns in unlabeled samples are close to saturation.

F. Quality Assessment of the SE DI

The quality of DI is crucial for detection performance in CD task. In this section, we assess the quality of the SE DI and compare its performance with other comparison operators in the forms of the receiver operating characteristic (ROC) curves and the area under the curve (AUC). Moreover, the ROC and AUC can also be utilized as indicators in analyzing the parameter value selection of α.

The SE DI obtained using the proposed reweighting scheme is compared to the log-ratio operator (LR), mean-ratio operator (MR), neighborhood-ratio (NR) method [82], and INLPG method [44], respectively. For the NR method, the neighborhood size is set to 3×3. For the INLPG, parameters follow the default values in the original paper [44]. The DIs produced by different methods are shown in Fig. 16 using the "jet" color map in MATLAB. The corresponding ROC curves of the SE DI and other DIs produced by different methods are shown in Fig. 17 and the corresponding AUC values are listed in Table XI.

The ROC curves show that the SE DIs are more stable and almost on the top of the other curves on the Ottawa, Farmland B, and Foshan City datasets, indicating that the SE DIs have better quality than other DIs. In Table XI, it is obvious that the SE DIs provide larger AUC values than other DIs. Particularly, the AUC values of SE DIs with different values of parameter α are entirely consistent with the results in Fig. 13. That is to say, the SE DI with the selected values of parameter α has a better quality than other SE DIs. Therefore, the AUC value can be utilized as an indicator in the analysis of the parameter value selection for parameter α.

G. Comparison With Counterpart Methods

To verify the effectiveness of the proposed method, we compare it to other counterparts on four real SAR datasets, which are typical SSL models and widely utilized SAR image CD methods. In this group of experiments, the proportion of labeled samples is set to 0.3%, as described earlier.

TABLE X
PCCs (%) AND KAPPA STATISTICS κ(%) ACHIEVED AFTER APPLYING THE PROPOSED METHOD TO THE OTTAWA, FARMLAND A, FARMLAND B, AND FOSHAN CITY DATASETS USING 30 LABELED SAMPLES (15 SAMPLES PER CLASS) AND VARIED NUMBERS OF UNLABELED SAMPLES

Data Set	Number of Unlabeled Samples Per Class											
	500		1000		2000		3000		4000		4985	
	PCC	κ	PCC	κ	PCC	κ	PCC	κ	PCC	κ	PCC	κ
Ottawa	97.92	92.33	97.94	92.51	98.22	93.54	98.31	93.60	98.45	94.24	98.57	94.69
Farmland A	95.61	85.02	95.56	84.66	95.81	85.52	96.03	86.11	96.17	86.24	96.21	86.85
Farmland B	97.72	82.23	98.64	87.09	98.77	88.51	98.69	87.32	98.91	88.94	99.02	90.90
Foshan City	95.65	71.53	95.86	72.73	95.99	73.21	96.19	74.18	96.24	74.32	96.47	74.53

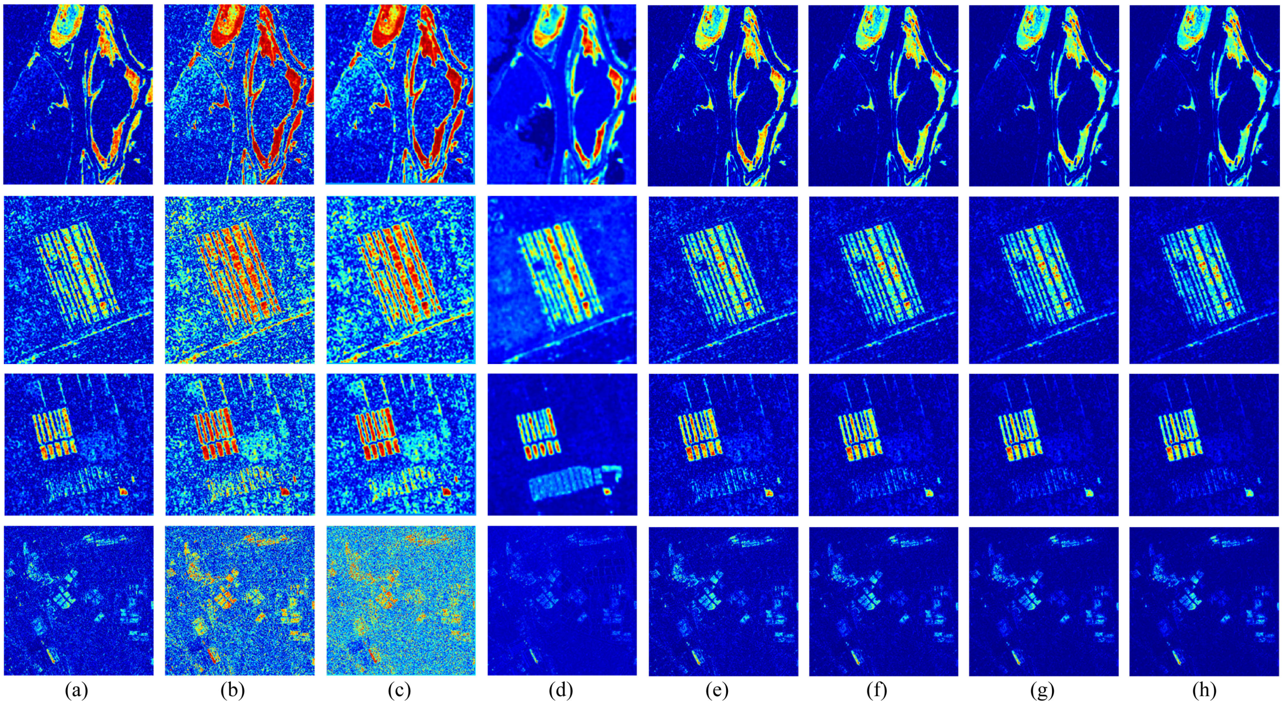


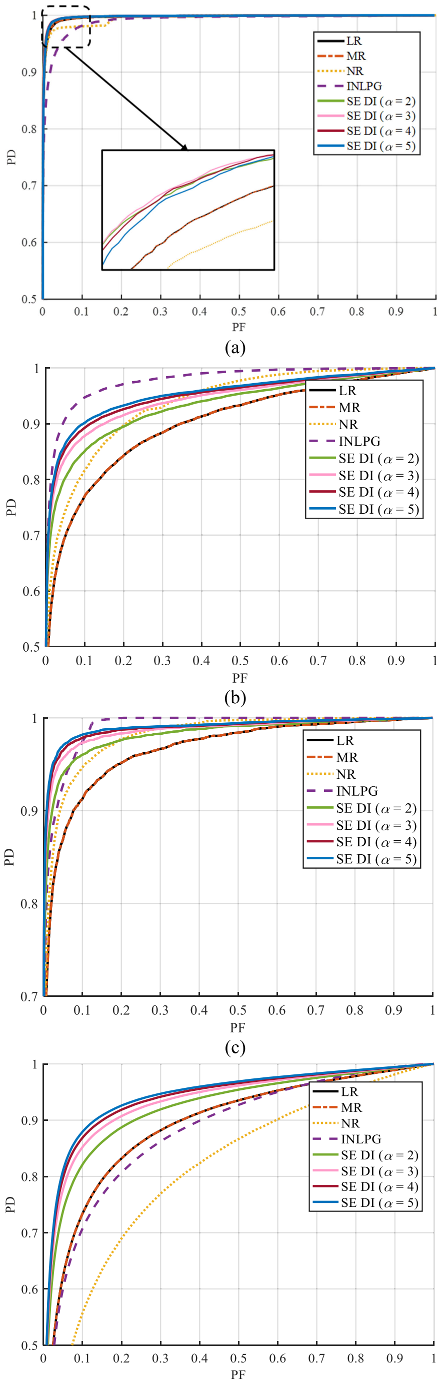
Fig. 16. DIs produced by different comparison methods. DIs on Ottawa (the first row), Farmland A (the second row), Farmland B (the third row), and Foshan City (the fourth row). (a) DIs by LR. (b) DIs by MR. (c) DIs by NR. (d) DIs by INLPG. (e) SE DIs ($\alpha = 2$). (f) SE DIs ($\alpha = 3$). (g) SE DIs ($\alpha = 4$). (h) SE DIs ($\alpha = 5$).

1) *Pseudo-Label* [83], which is a classical deep learning-based semisupervised model. In experiments, Pseudo-Label + PF (PLPF) represents the model trained with the pixel-wise features solely, and Pseudo-Label + CF (PLCF) represents the model trained with the context-wise features solely.
2) *Temporal Ensembling* [35], which is the baseline of the developed LCS-EnsemNet. Likewise, Temporal Ensembling + PF (TEPF) and Temporal Ensembling + CF (TECF) represent the model trained with pixel-wise features and context-wise features, respectively.
3) *PCA-kmeans* [16], which is a simple but effective SAR image CD method.
4) *PCANet* [18], a deep-learning method, has achieved stable and excellent results in the CD of SAR images.
5) *CWNN* [48], a preclassification-based unsupervised CD method in SAR images using the convolutional wavelet neural network (CWNN).

TABLE XI
AUC COMPARISON OF DIFFERENT DI GENERATION METHODS

Methods	Ottawa	Farmland A	Farmland B	Foshan City
LR	0.9970	0.9019	0.9661	0.8897
MR	0.9970	0.9019	0.9661	0.8897
NR	0.9952	0.9379	0.9811	0.8110
INLPG	0.9909	**0.9783**	0.9901	0.8804
$\alpha = 2$	**0.9979**	0.9334	0.9923	0.9218
$\alpha = 3$	**0.9979**	0.9446	0.9869	0.9335
$\alpha = 4$	**0.9979**	0.9505	0.9890	0.9399
$\alpha = 5$	0.9978	**0.9541**	**0.9902**	**0.9439**

6) *DDNet* [49], which uses features from spatial domain and frequency domain as the input of a two-branch CNN network. The method also uses the preclassification step for the pseudo-labeled samples selection.
7) *SGDNNs* [19], which segments the saliency map of the log-ratio DI as a label prior to guide the pseudo-labeled



(a)

(b)

(c)

(d)

Fig. 17. ROC curves on (a) Ottawa, (b) Farmland A, (c) Farmland B, and (d) Foshan City datasets. PF represents the probability of false alarm and PD represents the probability of detection.

samples acquisition. A DNN is pretrained in an unsupervised way and fine-tuned on the pseudo-labeled samples.

8) *INLPG-CWNN* [44], which replaces the DIs in the CWNN method with the DIs generated by the INLPG method.

The visual and quantitative results of the reference methods on four real SAR datasets are shown in Figs. 18 –21 and Tables XII–XV.

TABLE XII
EXPERIMENTAL RESULTS ON OTTAWA DATASET

Method	FP	FN	OE	PCC (%)	κ (%)
PCA-kmeans	589	1898	2487	97.55	90.49
PCANet	995	853	1848	98.18	93.22
PLPF	3378	224	3602	96.45	87.67
PLCF	3199	224	3423	96.62	88.19
TEPF	598	452	1050	98.83	95.64
TECF	720	598	1318	98.70	95.14
CWNN	1291	434	1725	98.30	93.75
DDNet	622	1186	1808	98.22	93.21
SGDNNs	0	1067	1067	**98.95**	**95.94**
INLPG-CWNN	0	7039	7039	93.07	68.31
Proposed method	919	530	1449	**98.57**	**94.69**

TABLE XIII
EXPERIMENTAL RESULTS ON FARMLAND A DATASET

Method	FP	FN	OE	PCC (%)	κ (%)
PCA-kmeans	4211	3365	7576	89.80	66.40
PCANet	1716	1686	3402	95.42	84.55
PLPF	2982	1187	4169	94.39	81.99
PLCF	2874	1157	4031	94.57	82.55
TEPF	2065	1917	3982	94.64	81.97
TECF	1940	1778	3718	94.99	83.18
CWNN	837	1690	2527	**96.60**	**88.23**
DDNet	952	1846	2798	96.23	86.95
SGDNNs	862	2894	3756	94.94	81.86
INLPG-CWNN	510	2409	2919	96.07	85.96
Proposed method	953	1864	2817	**96.21**	**86.85**

TABLE XIV
EXPERIMENTAL RESULTS ON FARMLAND B DATASET

Method	FP	FN	OE	PCC (%)	κ (%)
PCA-kmeans	2235	623	2858	96.79	74.78
PCANet	356	1265	1300	98.54	85.40
PLPF	1733	85	1820	97.95	84.00
PLCF	1755	85	1840	97.93	83.84
TEPF	3507	193	3700	95.84	71.63
TECF	2682	167	2849	96.80	76.85
CWNN	225	863	1088	98.78	88.37
DDNet	155	1211	1366	98.47	84.80
SGDNNs	18	1848	1866	97.90	77.53
INLPG-CWNN	7007	22	7029	91.88	54.53
Proposed method	267	606	873	**99.02**	**90.90**

TABLE XV
EXPERIMENTAL RESULTS ON FOSHAN CITY DATASET

Method	FP	FN	OE	PCC (%)	κ (%)
PCA-kmeans	417002	79598	496600	78.95	20.93
PCANet	139325	85198	110184	90.48	42.37
PLPF	92976	28642	121618	94.85	69.46
PLCF	80235	23469	103704	95.60	73.53
TEPF	48006	61686	109692	95.35	67.02
TECF	39544	62862	102406	95.66	68.43
CWNN	529397	41854	571251	75.79	24.28
DDNet	335119	62040	397159	83.17	30.80
SGDNNs	1235	143907	145142	93.35	32.92
INLPG-CWNN	335055	38110	373165	84.18	37.18
Proposed method	31366	51875	83241	**96.47**	**74.53**

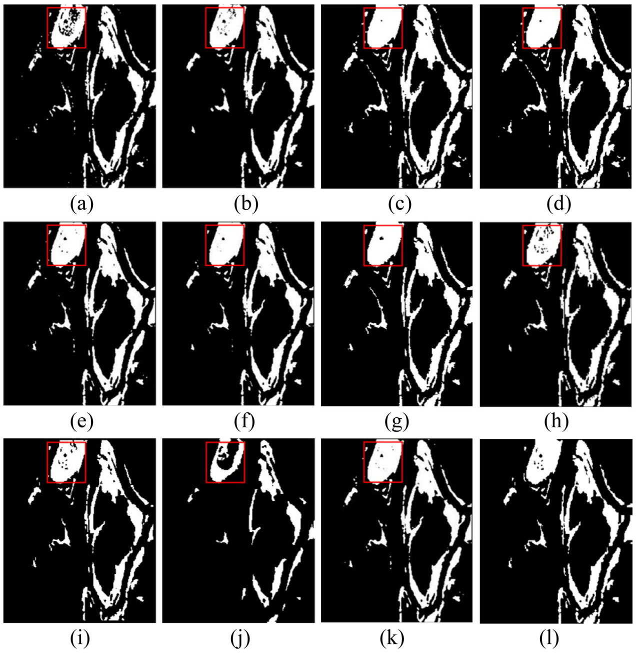


Fig. 18. Ground truth and change map by different methods for the Ottawa dataset. (a) PCA-kmeans. (b) PCANet. (c) PLPF. (d) PLCF. (e) TEPF. (f) TECF. (g) CWNN. (h) DDNet. (i) SGDNNs. (j) INLPG-CWNN. (k) Proposed method. (l) Ground truth.

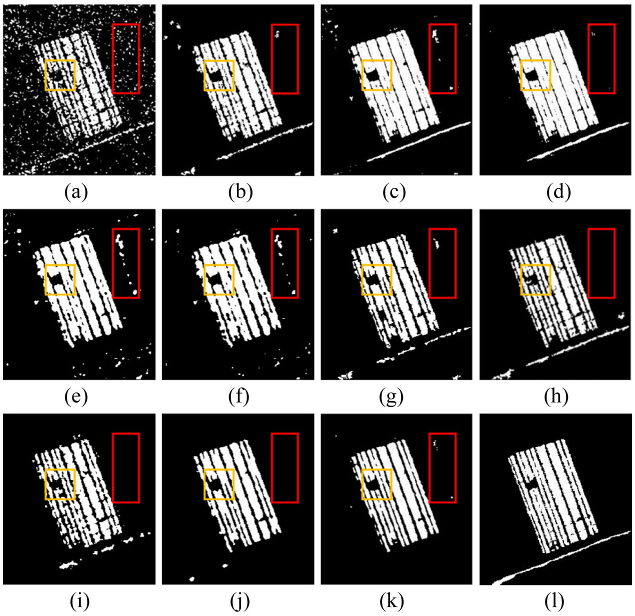


Fig. 19. Ground truth and change map by different methods for the Farmland A dataset. (a) PCA-kmeans. (b) PCANet. (c) PLPF. (d) PLCF. (e) TEPF. (f) TECF. (g) CWNN. (h) DDNet. (i) SGDNNs. (j) INLPG-CWNN. (k) Proposed method. (l) Ground truth.

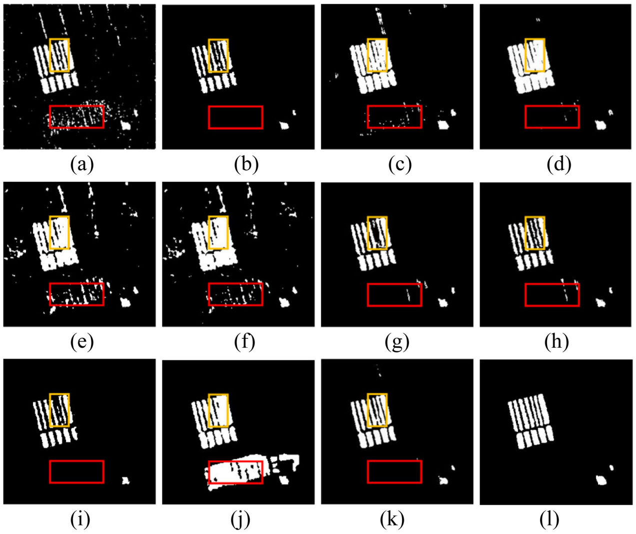


Fig. 20. Ground truth and change map by different methods for the Farmland B dataset. (a) PCA-kmeans. (b) PCANet. (c) PLPF. (d) PLCF. (e) TEPF. (f) TECF. (g) CWNN. (h) DDNet. (i) SGDNNs. (j) INLPG-CWNN. (k) Proposed method. (l) Ground truth.

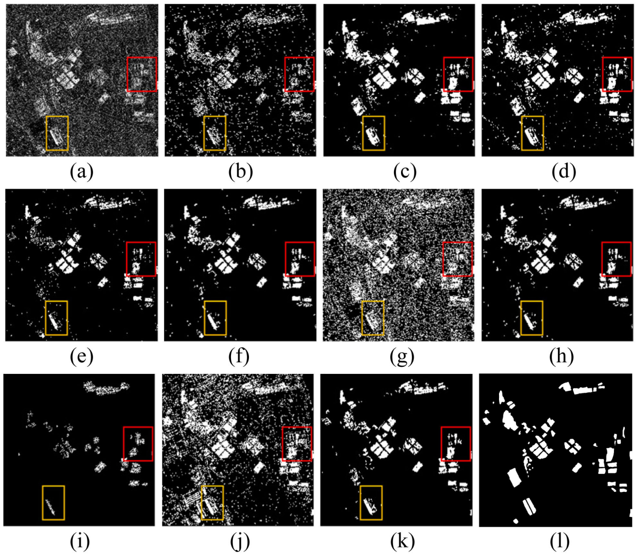


Fig. 21. Ground truth and change map by different methods for the Foshan City dataset. (a) PCA-kmeans. (b) PCANet. (c) PLPF. (d) PLCF. (e) TEPF. (f) TECF. (g) CWNN. (h) DDNet. (i) SGDNNs. (j) INLPG-CWNN. (k) Proposed method. (l) Ground truth.

The results of the Ottawa dataset are shown in Fig. 18 and Table XII. For the Ottawa dataset, there are many changed areas, notably minor isolated changed areas. Thus, the primary challenge is to detect these minor regions with high accuracy. It can be seen from the results that the proposed method acquires the PCC of 98.57% and the Kappa value of 94.69%, which is quite competitive to the best results by SGDNNs on this dataset. The ROI within a red box in Fig. 18 shows that the result by the proposed method is more compact with less missed detections, which achieves a relative balance between detail preservation and speckle suppression. PCA-kmeans obtains inferior visual and quantitative results on account of the inconsistency of the hand-crafted features with SAR images and the poor performance of traditional *k*-means clustering. Although PCANet, PLPF, and PLCF use the pseudo-labels for model learning, PLPF and PLCF acquire the worst results, whereas PCANet yields better results due to the refinement process of the pseudo-labels.

Because of the weak level of speckle noise in this dataset, by incorporating all the accumulated predictions of unlabeled samples for network training, TEPF and TECF achieve an appealing result. The recently proposed SGDNNs achieve the best results, including the smallest FP and highest Kappa value. From the results in Table XII and Fig. 18, although the FP and OE of the results by the proposed method are a little higher than SGDNNs, it yields a similar visual effect.

The results on the Farmland A dataset are shown in Fig. 19 and Table XIII. This dataset is suffered from a strong level of speckle noise, and the interclass separability of the changed and unchanged classes is small. As shown in Fig. 19 and Table XIII, the results of the counterpart methods are affected severely by the speckle noise. Particularly, changed ROI within the yellow box in Fig. 19 validates that the result by the proposed method is complete and more accurate. In Fig. 19, the result in the red box predicted by our method has few false detections. Due to the interference of strong noise in data and weak representation ability of traditional clustering algorithm, FP and FN of the results by PCA-kmeans are the highest. Since the number of hard-to-be-classified samples is increased in the heterogeneous and noisy regions, both the FN and FP results by the PLPF, PLCF, TEPF, and TECF are large, such as the false alarms within the red box in Fig. 19(e) and (f). The deep learning-based methods, such as the PCANet, CWNN, DDNet, INLPG-CWNN, and the proposed method, achieve better results than other counterpart methods, mainly due to the powerful feature abstraction ability of deep networks. The CWNN provides the best PCC (96.60%) and Kappa value (88.23%). For our method, the accumulated label information selected by the proposed two-stage label-consistent self-ensemble strategy improves the reliability and stability, reducing the negative effect of speckle noise and improving the generalization ability. In addition, the introduced spatial context information also helps to remove the confusing pixels in heterogeneous and noisy areas. Therefore, the proposed method achieves comparable results with the CWNN and DDNet, including the PCC (96.21%) and Kappa value (86.85%), demonstrating its robustness to speckle noise and verifying its effectiveness under complex scenes and a strong level of speckle noise. Even though the result of the proposed method missed the linear changed regions, the OE is low and the prominent changed regions are visually complete with less false alarms.

The results of Farmland B are shown in Fig. 20 and Table XIV. The proposed method obtains better results than other counterpart methods. As shown in Fig. 20 and Table XIV, the best results are achieved by the proposed method at the largest PCC (99.02%) and Kappa value (90.90%). Although PCA-kmeans obtains a similar FN result with the proposed method, its FP is much higher, such as the false alarms within the red box in Fig. 20(a). PCANet obtains better results than other conventional reference methods due to its robustness and learning ability of the deep network. Due to the heterogeneous regions and noisy regions in this data, more hard-to-be-classified samples exist, resulting in the performance degradation of the PLPF, PLCF, TEPF, and TECF. For the SSL methods relying on the pseudolabel information without any selection stage, incorrect predictions from hard-to-be-classified samples lead to high FP values, such as the regions within the red box in Fig. 20(c), (e), and (f). The recently proposed CWNN method provides better results than other compared methods. Nevertheless, the deep learning-based methods, such as the DDNet, SGDNNs, and the INLPG-CWNN methods, achieve worse results than our method because the wrongly predicted pseudo-labeled samples by conventional clustering algorithms mislead the model learning, despite the excellent learning ability of the deep networks. In our method, using the two-stage selection strategy, the reliability of the selected accumulated pseudo-labels is improved and the hard-to-be-classified sample pairs are removed. The result within the yellow box in Fig. 20(k) is more accurate than other methods. Through the experiments, we can see that the proposed method achieves the leading performance since the designed two-stage selection strategy prevents the hard-to-be-classified sample pair from harming the model learning and improves the reliability of pseudo-labels.

TABLE XVI
COMPUTATIONAL TIME (SECONDS) OF DIFFERENT COMPARISON METHODS

Methods	Ottawa	Farmland A	Farmland B	Foshan City
LR	0.87	0.61	0.84	0.72
MR	1.07	0.53	1.09	0.69
NR	2.05	1.44	1.70	50.63
INLPG	33.25	10.56	14.16	1726
SE DI	11.05	14.43	17.47	202.55

The results of Foshan City are shown in Fig. 21 and Table XV. The data consist of two high-resolution SAR images, which have extensive inhomogeneous regions and also contain a strong level of speckle noise. In Fig. 21 (two changed ROIs are marked by red and yellow boxes, respectively), we can see that changed regions are detected by the proposed method with more image detail and less false alarms. In Table XV, the proposed method is shown to provide higher PCC and Kappa values than other methods when applied to the high-resolution SAR images. Due to the inhomogeneous signature in the high-resolution SAR images, a large number of pixels are wrongly classified into the changed category, which causes extensive false alarms in the CD maps. Especially for the preclassification-based methods, the extensive hard samples cause the performance degradation because the incorrectly pseudo-labeled samples would have an adverse impact on the model learning. However, thanks to the two-stage refinement strategy in the proposed method, a large number of hard samples are prevented from network training, avoiding significant performance degradation. The experimental results demonstrate the effectiveness of the proposed network in the task of processing large-scale high-resolution SAR images.

H. Running Time Analysis

In this section, we compare the running time of the SE DI generation with other comparison methods, as given in Table XVI. From Table XVI, we can see that the running time of the conventional DI generation methods, such as the LR, MR, and NR, is less than the methods considering the global information, such as INLPG and the proposed method. Due to the global

TABLE XVII
INFERENCE TIME (SECONDS) ON DIFFERENT DATASETS

Method	Ottawa	Farmland A	Farmland B	Foshan City
LCS-EnsemNet	9.13	8.37	7.83	211.32

information exploitation in saliency map calculation, the SE DI generation has slightly more computational time than other methods. Nonetheless, the performance of the SE DI is generally better than the DIs of other methods. Thus, the proposed SE DI generation can achieve a relatively preferable balance between the performance and the time complexity. Additionally, we also list the inference time of the well-trained LCS-EnsemNet on the constructed features of each pixel in the image for the four real datasets in Table XVII. From Table XVII, it is noticed that the proposed network can infer the category of all the image pixels in a short time for the small-scale SAR images. However, for the large-scale high-resolution SAR images such as Foshan City, the inference time is a little higher due to the large number of pixels in the entire image.

VI. CONCLUSION

In this article, a novel semisupervised SAR image CD method is proposed to reveal and refine the underlying category information from abundant unlabeled data to enhance the detection performance and generalization ability.

The proposed method includes the construction of feature representation and the semisupervised LCS-EnsemNet with the label-consistent self-ensemble strategy. First, an SE DI is generated by incorporating the spatial-contextual information in the CASD map and the pixel-wise log-ratio DI, resulting in improved separability of changed and unchanged regions. Based on the log-ratio DI and SE DI, a couple of feature vectors are constructed to represent the same pixel position in SAR images, expressing pixel-wise feature and context-wise feature, respectively. They are utilized as the input of the customized two-branch LCS-EnsemNet. Second, the LCS-EnsemNet with the label-consistent self-ensemble strategy is devised especially for the SAR image CD to alleviate the problems arising from the lack of sufficient labeled samples. The devised label-consistent self-ensemble strategy is the core of the network training, improving the reliability and accuracy of pseudo-labels through the two-stage ensemble. The selected pseudo-labeled samples are helpful to improve the generalization ability. The experiments on real SAR datasets demonstrate that the proposed method outperforms the reference methods, which validate the effectiveness of the proposed semisupervised SAR image CD method. As future work, an SE DI selection scheme will be explored by evaluating the quality of SE DI under different parameter values of α and using the image texture information to automatically select the best SE DI for subsequent CD task.

ACKNOWLEDGMENT

The authors would like to thank the Editor, the Associate Editor, and the anonymous reviewers for their valuable comments and suggestions to improve this article. The authors would also like to give special thanks to Prof. W. Yang of Wuhan University for providing the Foshan City dataset [23] and Dr. J. Geng of Northwestern Polytechnical University for providing the CD code [19].

REFERENCES

[1] F. Bovolo and L. Bruzzone, “The time variable in data fusion: A change detection perspective,” *IEEE Geosci. Remote Sens. Mag.*, vol. 3, no. 3, pp. 8–26, Sep. 2015.

[2] F. Bovolo and L. Bruzzone, “A detail-preserving scale-driven approach to change detection in multitemporal SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 12, pp. 2963–2972, Dec. 2005.

[3] Y. Bazi, L. Bruzzone, and F. Melgani, “An unsupervised approach based on the generalized Gaussian model to automatic change detection in multitemporal SAR images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 43, no. 4, pp. 874–887, Apr. 2005.

[4] L. Bruzzone and S. B. Serpico, “An iterative technique for the detection of land-cover transitions in multispectral remote-sensing images,” *IEEE. Trans. Geosci. Remote Sens.*, vol. 35, no. 4, pp. 858–867, Jul. 1997.

[5] F. Bovolo, C. Marin, and L. Bruzzone, “A hierarchical approach to change detection in very high resolution SAR images for surveillance applications,” *IEEE Trans. Geosci. Remote Sens.*, vol. 51, no. 4, pp. 2042–2054, Apr. 2013.

[6] Y. Ban and O. A. Yousif, “Multitemporal spaceborne SAR data for urban change detection in China,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1087–1094, Aug. 2002.

[7] G. Moser and S. B. Serpico, “Unsupervised change detection from multichannel SAR data by Markovian data fusion,” *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 7, pp. 2114–2128, Jul. 2009.

[8] M. G. Gong, P. Z. Zhang, L. Su, and J. Liu, “Coupled dictionary learning for change detection from multisource data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 12, pp. 7077–7091, Dec. 2016.

[9] E. J. M. Rignot and J. J. Van Zyl, “Change detection techniques for ERS-1 SAR data,” *IEEE Trans. Geosci. Remote Sens.*, vol. 31, no. 4, pp. 896–906, Jul. 1993.

[10] T. Celik and K. Ma, “Multitemporal image change detection using undecimated discrete wavelet transform and active contours,” *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 2, pp. 706–716, Feb. 2011.

[11] T. Celik, “Multiscale change detection in multitemporal satellite images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 820–824, Oct. 2009.

[12] R. J. Dekker, “Speckle filtering in satellite SAR change detection imagery,” *Int. J. Remote Sens.*, vol. 19, no. 6, pp. 1133–1146, Apr. 1998.

[13] C. Oliver and S. Quegan, *Understanding Synthetic Aperture Radar Images*. Rijeka, Croatia: SciTech, 2004.

[14] G. Moser and S. B. Serpico, “Generalized minimum-error thresholding for unsupervised change detection from SAR amplitude imagery,” *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 10, pp. 2972–2982, Oct. 2006.

[15] F. Chatelain, J.-Y. Tourneret, and J. Inglada, “Change detection in multisensor SAR images using bivariate gamma distributions,” *IEEE Trans. Image Process.*, vol. 17, no. 3, pp. 249–258, Mar. 2008.

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

[17] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, “Change detection in synthetic aperture radar images based on deep neural networks,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[18] F. Gao, J. Dong, B. Li, and Q. Xu, “Automatic change detection in synthetic aperture radar images based on PCANet,” *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[19] J. Geng, X. Ma, X. Zhou, and H. Wang, “Saliency-guided deep neural networks for SAR image change detection,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 10, pp. 7365–7377, Oct. 2019.

[20] M. H. C. Law, M. A. T. Figueiredo, and A. K. Jain, “Simultaneous feature selection and clustering using mixture models,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[21] L. Jia, M. Li, Y. Wu, P. Zhang, H. Chen, and L. An, “Semisupervised SAR image change detection using a cluster-neighborhood kernel,” *IEEE Geosci. Remote Sens. Lett.*, vol. 11, no. 8, pp. 1443–1447, Aug. 2014.

[22] L. An, M. Li, Y. Wu, L. Jia, and W. Song, “Discriminative random fields based on maximum entropy principle for semisupervised SAR image change detection,” *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3395–3404, Aug. 2016.

[23] W. Yang, H. Song, X. Huang, X. Xu, and M. Liao, "Change detection in high-resolution SAR images based on Jensen–Shannon divergence and hierarchical Markov model," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 8, pp. 3318–3327, Aug. 2014.

[24] L. Bergamasco, S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised change-detection based on convolutional-autoencoder feature extraction," *SPIE Remote Sens.*, vol. 11155, 2019, Art. no. 1115510, doi: 10.1117/12.2533812.

[25] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[26] Y. Sun, L. Lei, D. Guan, X. Li, and G. Kuang, "SAR image change detection based on nonlocal low-rank model and two-level clustering," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 293–306, Jan. 2020.

[27] U. Maulik and D. Chakraborty, "Remote sensing image classification: A survey of support-vector-machine-based advanced techniques," *IEEE Geosci. Remote Sens. Mag.*, vol. 5, no. 1, pp. 33–52, Mar. 2017.

[28] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[29] H. Bi, J. Sun, and Z. Xu, "A graph-based semisupervised deep learning model for PolSAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 4, pp. 2116–2132, Apr. 2019.

[30] C. Liu, J. Li, and L. He, "Superpixel-based semisupervised active learning for hyperspectral image classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 357–370, Jan. 2019.

[31] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 241–245, Feb. 2019.

[32] S. Zhou, Z. Xue, and P. Du, "Semisupervised stacked autoencoder with cotraining for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3813–3826, Jun. 2019.

[33] 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.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.

[34] J. Liu *et al.*, "Semi-supervised change detection based on graphs with generative adversarial networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Yokohama, Japan, 2019, pp. 74–77.

[35] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. 5th Int. Conf. Learn. Represent*, 2017, pp. 1–6.

[36] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *Proc. Adv. Neural Inf. Process. Syst*, 2017, pp. 1195–1204.

[37] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, Jun. 2014.

[38] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000.

[39] Y. Bazi, L. Bruzzone, and F. Melgani, "Automatic identification of the number and values of decision thresholds in the log-ratio image for change detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 3, no. 3, pp. 349–353, Jul. 2006.

[40] Y. Bazi, L. Bruzzone, and F. Melgani, "Image thresholding based on the EM algorithm and the generalized Gaussian distribution," *Pattern Recognit.*, vol. 40, no. 2, pp. 619–634, Feb. 2007.

[41] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size multitemporal images: Application to Tsunami-damage assessment," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 6, pp. 1658–1670, Jun. 2007.

[42] J. Inglada and G. Mercier, "A new statistical similarity measure for change detection in multitemporal SAR images and its extension to multiscale change analysis," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 5, pp. 1432–1445, May 2007.

[43] M. Gong, Z. Zhou, L. X, and J. Ma, "Change detection in synthetic aperture radar images based on image fusion and fuzzy clustering," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2141–2151, Apr. 2012.

[44] 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.*, to be published, doi: 10.1109/TGRS.2021.3053571.

[45] O. Yousif and Y. Ban, "Improving SAR-based urban change detection by combining MAP-MRF classifier and nonlocal means similarity weights," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 10, pp. 4288–4300, Oct. 2014.

[46] L. Jia, M. Li, P. Zhang, Y. Wu, and H. Zhu, "SAR image change detection based on multiple kernel K-means clustering with local-neighborhood information," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 6, pp. 856–860, Jun. 2016.

[47] H.-C. Li, T. Celik, N. Longbotham, and W. J. Emery, "Gabor feature based unsupervised change detection of multitemporal SAR images based on two-level clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 12, pp. 2458–2462, Dec. 2015.

[48] Y. Gao, F. Gao, and J. Dong, "Sea ice change detection in SAR images based on convolutional-wavelet neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1240–1244, Oct. 2019.

[49] X. Qu, F. Gao, J. Dong, Q. Du, and H.-C. Li, "Change detection in synthetic aperture radar images using a dual-domain network," *IEEE Geosci. Remote Sens. Lett.*, to be published, doi: 10.1109/LGRS.2021.3073900.

[50] Y. Li, C. Peng, Y. Chen, L. Jiao, L. Zhou, and R. Shang, "A deep learning method for change detection in synthetic aperture radar images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 5751–5763, Aug. 2019.

[51] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021.

[52] S. Saha, F. Bovolo, and L. Bruzzone, "Change detection in image time-series using unsupervised LSTM," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1240–1244, Oct. 2019.

[53] Y. Gao, F. Gao, J. Dong, and S. Wang, "Transferred deep learning for sea ice change detection from synthetic-aperture radar images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 10, pp. 1655–1659, Oct. 2019.

[54] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, "Selective adversarial adaptation-based cross-scene change detection framework in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 8, pp. 2188–2203, Mar. 2021.

[55] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[56] G. Camps-Valls and L. Gomez-Chova, "Kernel-based framework for multitemporal and multisource remote sensing data classification and change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 6, pp. 1822–1835, Jun. 2008.

[57] L. Li *et al.*, "Deformable dictionary learning for SAR image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4605–4617, Aug. 2018.

[58] R. Wang, J. Zhang, J. Chen, L. Jiao, and M. Wang, "Imbalanced learning-based automatic SAR images change detection by morphologically dupervised PCA-Net," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 4, pp. 554–558, Apr. 2019.

[59] Y. Gao, L. Lin, F. Gao, J. Dong, and H.-C. Li, "SAR image change detection based on multiscale capsule network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 3, pp. 484–488, Mar. 2021.

[60] R. Wang, F. Ding, J.-W. Chen, L. Jiao, and L. Wang, "A lightweight convolutional neural network for bitemporal image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Waikoloa, HI, USA, 2020, pp. 2551–2554.

[61] G. Camps-Valls, T. Bandos, and D. Zhou, "Semi-supervised graph-based hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 45, no. 10, pp. 3044–3054, Oct. 2007.

[62] F. Ratle, G. Camps-Valls, and J. Weston, "Semisupervised neural networks for efficient hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 5, pp. 2271–2282, May 2010.

[63] J. Li, J. M. Bioucas-Dias, and A. Plaza, "Semisupervised hyperspectral image segmentation using multinomial logistic regression with active learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 48, no. 11, pp. 4085–4098, Nov. 2010.

[64] L. Bruzzone, M. Chi, and M. Marconcini, "A novel transductive SVM for semisupervised classification of remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 11, pp. 3363–3373, Nov. 2006.

[65] Z. Wang, N. M. Nasrabadi, and T. S. Huang, "Semisupervised hyperspectral classification using task-driven dictionary learning with laplacian regularization," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 3, pp. 1161–1173, Mar. 2015.

[66] L. Wan, K. Tang, M. Li, Y. Zhong, and A. K. Qin, "Collaborative active and semisupervised learning for hyperspectral remote sensing image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 5, pp. 2384–2396, May 2015.

[67] A. Blum and T. Mitchell, *Combining Labeled and Unlabeled Data With Co-Training*. San Mateo, CA, USA: Morgan Kaufmann, pp. 92–100, 1998.

[68] V. Vapnik, *Statistical Learning Theory*. Hoboken, NJ, USA: Wiley, 1998.

[69] Y. Gao, R. Ji, P. Cui, Q. Dai, and G. Hua, "Hyperspectral image classification through bilayer graph-based learning," *IEEE Trans. Image Process.*, vol. 23, no. 7, pp. 2769–2778, Jul. 2014.

[70] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 607–611, Apr. 2021.

[71] U. Maulik and D. Chakraborty, "A self-trained ensemble with semisupervised SVM: An application to pixel classification of remote sensing imagery," *Pattern Recognit.*, vol. 44, no. 3, pp. 615–623, Feb. 2011.

[72] D. Berthelot, N. Carlini, I. Goodfellow, N. Papernot, A. Oliver, and C. A. Raffel, "Mixmatch: A holistic approach to semi-supervised learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 5050–5060.

[73] Y. Zhang, S. Wang, C. Wang, J. Li, and H. Zhang, "SAR image change detection using saliency extraction and shearlet transform," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 12, pp. 4701–4710, Dec. 2018.

[74] M. Li, M. Li, P. Zhang, Y. Wu, W. Song, and L. An, "SAR image change detection using PCANet guided by saliency detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 3, pp. 402–406, Mar. 2019.

[75] B. Wang, L. Yang, and Y. Zhao, "POLO: Learning explicit cross-modality fusion for temporal action localization," *IEEE Signal Process. Lett,* vol. 28, pp. 503–507, Feb. 2021.

[76] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization? (no, it is not about internal covariate shift)," in *Proc. Conf. Neural Inf. Process. Syst.*, 2018, pp. 2483–2493.

[77] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016. [Online]. Available: http://www.deeplearningbook.org

[78] X. Li, L. Yu, H. Chen, C. -W. Fu, L. Xing, and P. -A. Heng, "Transformation-consistent self-ensembling model for semisupervised medical image segmentation," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 2, pp. 523–534, Feb. 2021.

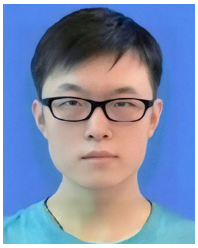
[79] S. Tokui *et al.*, "Chainer: A deep learning framework for accelerating the research cycle," 2019, *arXiv:1908.00213*. [Online]. Available: http://arxiv.org/abs/1908.00213

[80] G. H. Rosenfield and A. Fitzpatrick-Lins, "A coefficient of agreement as a measure of thematic classification accuracy," *Photogramm. Eng. Remote Sens.*, vol. 52, no. 2, pp. 223–227, Feb. 1986.

[81] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

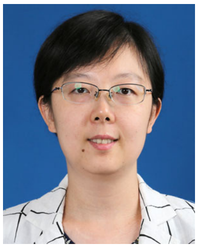
[82] M. Gong, Y. Cao, and Q. Wu, "A neighborhood-based ratio approach for change detection in SAR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 9, no. 2, pp. 307–311, Mar. 2012.

[83] H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. ICML Workshop*, 2013, pp. 1–6.



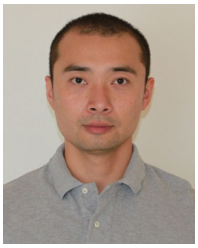
Jian Wang received the B.S. degree in electronic information science and technology from the North University of China, Taiyuan, China, in 2015. He is currently working toward the Ph.D. degree with Xidian University, Xi'an, China.

His research interests include synthetic aperture radar (SAR) image change detection and SAR image processing.



Yinghua Wang (Member, IEEE) received the B.S. degree in information engineering and the Ph.D. degree in control science and engineering from Xi'an Jiaotong University, Xi'an, China, in 2004 and 2010, respectively.

In 2007, she joined the Image and Signal Processing Department, Telecom Paris, Paris, France, as a Visiting Student. She is an Associate Professor with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China. Her research interests include synthetic aperture radar (SAR) automatic target recognition, polarimetric SAR data analysis and interpretation, and SAR image processing.



Bo Chen (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2003, 2006, and 2008, respectively.

From 2008 to 2012, he was a Postdoctoral Fellow, a Research Scientist, and a Senior Research Scientist with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. Since 2013, he has been a Professor with the National Laboratory for Radar Signal Processing, Xidian University. His current research interests include statistical machine learning, statistical signal processing, and radar automatic target detection and recognition.

Dr. Chen was the recipient of the Honorable Mention for the 2010 National Excellent Doctoral Dissertation Award and is selected into Overseas Talent by the Chinese Central Government in 2014.

Hongwei Liu (Member, IEEE) received the M.S. and Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 1995 and 1999, respectively.

From 2001 to 2002, he was a Visiting Scholar with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. He is a Professor with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an. His research interests include radar automatic target recognition, radar signal processing, and adaptive signal processing.