Target Discrimination Based on Weakly Supervised Learning for High-Resolution SAR Images in Complex Scenes

Lan Du, *Senior Member, IEEE*, Hui Dai, Yan Wang, Weitong Xie, and Zhaocheng Wang

***Abstract*— To design a highly automatic and practical discrimination method for high-resolution synthetic aperture radar (SAR) images in complex scenes, a novel target discrimination framework based on weakly supervised learning (WSL) of the mid-level features is proposed in this article. First, we extract the dense SAR scale-invariant feature transform (SAR-SIFT) features of the candidate regions obtained from the detected SAR images. Then, the dense SAR-SIFT descriptors are transformed into richer mid-level features by coding and pooling. Finally, the mid-level features are input into a WSL-based target discrimination method, where the training set is initially selected by the unsupervised latent Dirichlet allocation (LDA) and iteratively updated by the linear support vector machine (SVM) discriminator. In the proposed method, only the image-level annotations (weak labels), which indicate whether the images containing the targets of interest or not, are required. By introducing WSL, the manual annotations of target regions from SAR images can be avoided, which is generally expensive in complex scenes and may tend to be less accurate and unreliable for the occluded or camouflaged targets. The comprehensive and specific experiments on the measured SAR data have demonstrated the effectiveness of the proposed method in benchmarking with the supervised learning-based linear SVM and linear support vector data description (SVDD) discriminators.**

***Index Terms*— Latent Dirichlet allocation (LDA), mid-level features, synthetic aperture radar (SAR), target discrimination, weakly supervised learning (WSL).**

I. Introduction

TARGET discrimination for synthetic aperture radar (SAR) images is a subtask in SAR automatic target recognition (ATR). A typical SAR ATR system often consists of three stages: detection, discrimination, and classification or recognition [1]–[3]. The detection stage searches the entire SAR image to find out the candidate regions, which consists of the real target regions as well as numerous clutter false alarms (CFAs), and provides them to the discrimination stage. The discrimination stage removes both natural and artificial clutter, and maintains the real target regions. Finally, the classification stage classifies the remaining targets of discrimination stage into different types. Target discrimination, essentially as a binary classification problem, is a postprocessing for the candidate regions extracted in the detection stage, with the primary purpose of reducing classification or recognition cost by removing the CFAs [3], [4]. Moreover, as the stage linking detection and classification or recognition, it directly influences the performance of the whole ATR system [3], [5]. As a result, designing an effective and practical target discrimination method for SAR images is of great importance. In this article, we focus our attention on SAR target discrimination.

Since the past two decades, SAR target discrimination has gained increasing attention, and many target discrimination methods have been developed [3], [4], [6]–[9]. The common target discrimination method consists of two steps: 1) features used for discrimination are extracted and selected to describe the candidate regions and 2) discriminator is designed to make the decision.

The classical target discrimination features, e.g., the Lincoln features [3], [10], [11], mainly consider the difference between target and clutter in texture, size, contrast, and shape. Some of these classical discrimination features have been utilized into the real SAR ATR system [5], [11], and the performance is impressive in some simple scenes, such as the moving and stationary target acquisition and recognition (MSTAR) public release data set [4], [6]. Nevertheless, they have several disadvantages in practice. First, most classical discrimination features are only effective for the candidate regions containing only one integrated target and CFAs being highly different from the target. However, in complex scenes, the candidate regions obtained in the detection stage may contain multiple targets or partial targets with complex clutter. Fig. 1 shows some candidate region examples in the complex scene. Second, some classical discrimination features, e.g., the features based on target's shape and size, need segmentation processing to obtain the target-shaped blob. However, the segmented target-shaped blob usually varies with segmentation methods, and thereby, the corresponding features are not robust especially for some complex scenes. Third, there exists redundancy among these traditional discrimination features, which would influence the performance of discrimination [12]. Moreover, the optimal feature sequences can vary with different imaging conditions [13]. As a result, feature selection is necessary.

Manuscript received April 24, 2019; accepted August 13, 2019. Date of publication September 16, 2019; date of current version December 27, 2019. This work was supported in part by the National Science Foundation of China under Grant 61771362, Grant U1833203, and Grant 61671354 and in part by 111 Project under Grant B18039. *(Corresponding author: Lan Du.)*

The authors are with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an 710071, China (e-mail: dulan@mail.xidian.edu.cn).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TGRS.2019.2937175

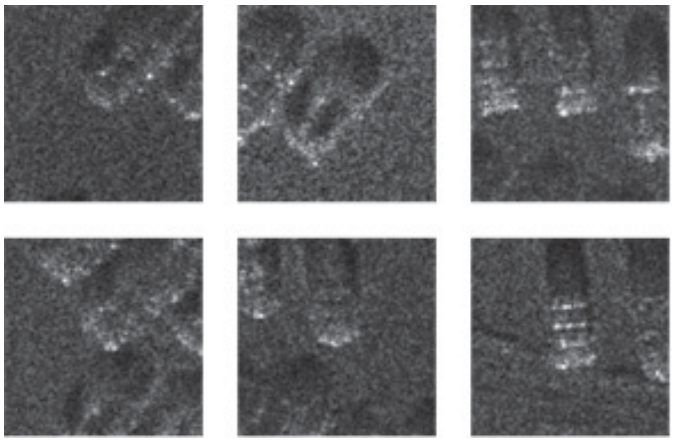


Fig. 1. Some candidate region examples.

However, developing an optimal feature selection method is quite difficult and complex.

After obtaining the optimal feature subset, the discriminator is designed to discriminate the targets from CFAs. The usually used discriminators fall into two categories: two-class and one-class discriminators. The two-class discriminator can achieve great discrimination performance when it is trained on a large number of training data including targets and CFAs with manual annotations, i.e., region-level annotations. However, manual annotation of candidate regions in complex scenes is generally expensive. Moreover, the manual annotation may tend to be less accurate and unreliable when the targets are occluded or camouflaged. The one-class discriminator can be trained only on one class training set (targets or CFAs). Compared with the tediousness, laboriousness, and ambiguousness of annotation on target training set, the annotations of the CFAs are simper and easier. Thus, the one-class discriminator trained on CFAs can be obtained more easily. However, it is a difficult task to train a one class discriminator to match the CFAs in a feature space, since the CFAs for SAR images in complex scenes are of wide variety, which would degrade the discrimination performance.

To summarize what has been mentioned above, the common target discrimination methods for SAR images in the complex scenes have two disadvantages. First, the traditional discrimination features [6], [10], [14] are insufficiently powerful to describe the spatial and structural information of the candidate regions for SAR images in complex scenes, and selecting the optimal feature set from these traditional discrimination features is of great difficulty and perplexity. Second, the region-level annotations of target training data are of tediousness, laboriousness, and toughness.

To tackle the first disadvantage of the common target discrimination methods for SAR images, this article proposed to use mid-level features. Some papers have studied different mid-level features for the target detection and classification or recognition in optical images [15], [16], which has demonstrated that mid-level features can well catch the spatial and structural information of the targets. Lazebnik *et al.* [15] extracted dense scale-invariant feature transform (SIFT) descriptors as low-level features and then generate mid-level features via vector quantization (VQ) of SIFT features for natural scene classification. Wang *et al.* [16] extracted the histogram of oriented gradient (HOG) features as low-level features and then generated mid-level features via locality-constrained linear coding (LLC) of HOG features for optical image classification. It can be seen that the mid-level features are usually obtained by two steps: low-level feature extraction and low-level features coding. How to choose the appropriate low-level features and coding manner for special application is the key for mid-level features extraction. In our SAR target discrimination task, we first extract the dense SAR-SIFT descriptors as the low-level features for its outstanding ability in describing the local property of SAR candidate regions and its robustness to the speckle noise. Our earlier work [17] also demonstrates the effectiveness of dense SAR-SIFT descriptors in describing the SAR candidate regions. Since the LLC has better reconstruction and is of high computational efficiency compared with the conventional coding manners, we use the LLC to encode the dense SAR-SIFT descriptor to obtain the mid-level features, which can describe the global structural information of targets in the candidate regions. Nowadays, some high-level features have been proposed to mine the semantic information of an image. For example, a number of papers have used full-blown objectors as features to describe and reason about images [18]–[20]. Others have employed discriminative part detectors, such as poselets [21], attribute detectors [22], or "visual phrases" [23], as features. Recently, inspired by the human visual system, deep learning builds hierarchical layers of visual representation to extract the high-level features of an image. Lv *et al.* [24] utilized the deep belief networks (DBNs) model to extract the high-level feature for urban land use and land cover classification. The work in [25] investigates the use of the CNN model for classifying remote sensing scenes. Yao *et al.* [26] proposed to learn the high-level feature with a stacked discriminative sparse autoencoder (SDSAE) for the semantic annotation of high-resolution satellite images. Although these high-level features have been proved to be effective, the majority of them require amounts of labeled training data. Compared to these high-level features, the mid-level feature is extracted in a fully unsupervised manner, which is of great significance in the weakly supervised SAR target discrimination task where no explicit region-level annotations are available. There are also some unsupervised deep learning models, including deep Boltzmann machine (DBM) and autoencoder (AE). However, the high-level features extracted with these unsupervised deep models are not discriminative enough because these models aim to retain all the information that is needed to perfectly reconstruct the input samples and pay little attention to the important information that is useful to discriminate targets and clutter.

To tackle the second disadvantage aforementioned, we introduce weakly supervised learning (WSL) [27] for our discrimination. WSL means a machine learning framework where the model is trained using examples that are only partially annotated or labeled. Unlike conventional supervised learning approaches, WSL only requires a weak label for the training images to specify whether the image contains the targets of interest or not, which is referred to as image-level annotation,

rather than relies on manually labeled target regions, which is called region-level annotation. The image with image-level annotation can be regarded as a "bag" [28]. Each bag is considered as a collection of many examples: the image with targets can be considered as a positive bag which is assumed to contain at least one positive example; the image without target can be considered as a negative bag which only consists of negative examples. As one of the most cost-effective learning approaches, WSL is widely used in computer vision and have achieved promising results, especially for object detection in optical images [29]–[39]. Rochan and Wang [29] proposed a learning framework, which took advantage of selective search, multiple instance learning, and bag splitting to learn an object localization model from large-scale data with image-level annotations. Wang *et al.* [30] proposed the latent category learning (LCL), which combined convolutional neural network (CNN), probabilistic latent semantic analysis (PLSA), and bag-of-words (BoW) in WSL framework, to achieve object localization in large-scale cluttered optical images. Han *et al.* [31] proposed an improved WSL method, which integrated saliency, intra-class compactness, and inter-class separability in a Bayesian framework, to realize object detection in optical remote sensing images. Although these WSL approaches have achieved satisfactory results, they cannot be directly used to the SAR target discrimination for two reasons. On the one hand, most of the WSL frameworks designed for the optical images are always based on the fact that there are large amount of positive bags in optical images analysis [35]–[39]. For example, Nguyen *et al.* [37] extracted a new feature via linearly combining the features of all the examples in a positive bag, and then adopts the new feature as a positive example to train the classifier associated with the negative examples from the negative bags. Cinbis *et al.* [38] utilized the previously trained classifier to select a target example with high confidence from each positive bag, and then uses these selected positive examples and all negative examples from the negative bags to update the classifier until reaching the stop condition. However, for the SAR target discrimination task, there are only a small number of positive bags can be exploited. On the other hand, the existing WSL approaches usually deal with detection and classification simultaneously, making the framework considerably complex. In SAR target discrimination, we only need to discriminate the targets from clutter. Therefore, in this article, we propose a novel WSL framework for target discrimination in SAR images with complex scenes. Based on the CFAs from the negative images, the unsupervised latent Dirichlet allocation (LDA) [42] is first used to initially select the training examples based on the learned likelihood value of each candidate region, and then a linear support vector machine (SVM) discriminator is trained with the iteratively updated training sets.

In summary, the main contributions of this article are twofold.

1) We propose the mid-level features appropriate for the SAR target discrimination. The learned mid-level features can capture the spatial and structural information of the candidate regions, which leads to the improvement of target discrimination performance.

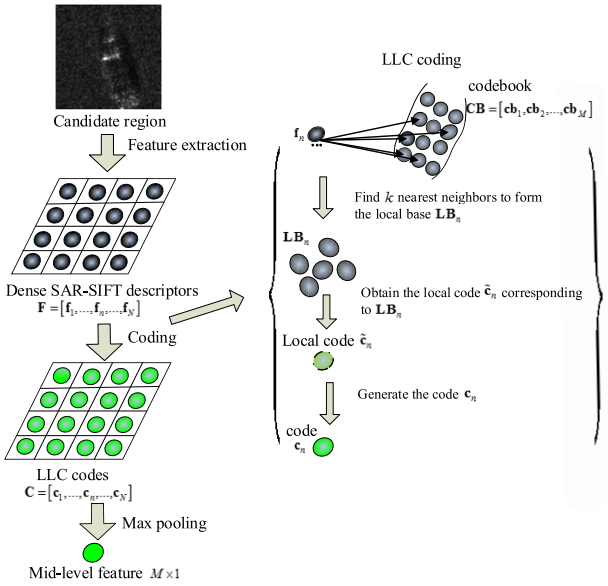


Fig. 2. Procedure of mid-level feature generation.

2) We propose a novel WSL framework based on LDA and iteratively updated SVM for SAR target discrimination. By exploiting the LDA model's likelihood differences between negative examples and positive examples to initialize training set and then taking advantage of the information of both positive and negative examples to gradually refine the target discriminator, our proposed WSL method can obtain satisfactory discrimination results.

The rest of this article is organized as follows. Section II describes the mid-level feature generation procedure. Section III introduces the WSL approach for SAR target discrimination. Experimental results are shown in Section IV and the conclusions are drawn in Section V.

II. Mid-Level Feature Generation

The performance of the traditional target discrimination features for SAR images in complex scenes is still far from satisfactory. The main issue lies in the insufficiently powerfulness to characterize the structural information of the target regions. With the advancement of the SAR imagery technology, it is possible for SAR image with high spatial resolution to capture spatial and structure information. At present, the accurate interpretation of SAR images relies on effective spatial feature representation to capture the structural and informative property of the regions. In our paper, the mid-level features are used to represent the candidate regions. The procedure of mid-level feature generation is shown in Fig. 2.

A. Low-Level Feature Extraction

Due to its ability to handle variations in terms of intensity, rotation, scale, and affine projection, the SIFT descriptor has been widely used and many SIFT variations have been proposed since it was proposed by Lazebnik *et al.* [15] and Lowe [41]. The dense SIFT descriptors that use the dense points instead of keypoints could more rapidly produce

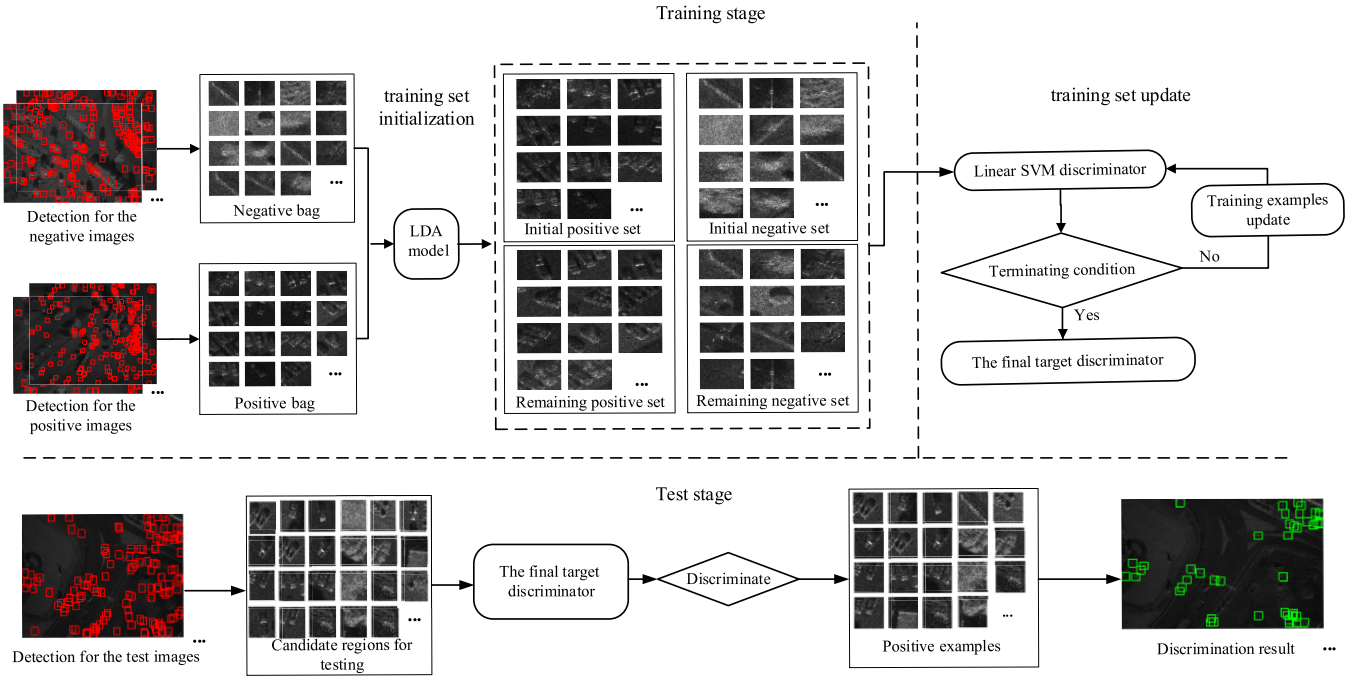


Fig. 3. Framework of WSL-based target discrimination.

descriptors and Li *et.al.* [42] have demonstrated that the dense SIFT performs better than the original SIFT in classification. Therefore, the dense SIFT descriptors are used in many applications and obtain promising results. However, the original SIFT descriptor cannot be directly used in the SAR image because of the speckle noise. To deal with the multiplicative speckle noise in SAR images, Dellinger *et.al.* [43] proposed the SAR-SIFT descriptor by defining the gradient of a pixel to be the log-ratio of its neighbor regions. SAR-SIFT has been used in SAR target discrimination [17], [44]. In our paper, the dense SAR-SIFT descriptors [17] are adopted in the proposed algorithm as the low-level features to characterize the candidate regions obtained in the detection stage.

The dense SAR-SIFT descriptors are extracted as the following steps. We divide a candidate region into many small patches using an overlapping, fixed size sliding window over the region. The center pixel of each patch is defined as the key point, and the SAR-SIFT descriptor is computed for each patch. The patch is divided into 16 components that contain 4×4 grids. We calculate an orientation histogram of the gradient with eight bins, which cover the 360 range of rotations for each component. When distributing each gradient value into neighboring bins, we use a trilinear interpolation to avoid the boundary effect of histogram binning. The SAR-SIFT descriptor of the patch is obtained by combining the $4 \times 4 \times 8$ bins. The low-level features of the region are obtained by combining the SAR-SIFT descriptors of all the patches.

B. Coding and Pooling

To alleviate the unrecoverable loss of discriminative information, we apply the LLC model [16] to encode the SAR-SIFT descriptors into region representation. Specifically, all the extracted low-level features for the training set are clustered to generate a codebook $\mathbf{CB} = [\mathbf{cb}_1, \mathbf{cb}_2, \ldots, \mathbf{cb}_M]$ with M entries by using the K-means method. Let $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_n, \ldots, \mathbf{f}_N]$ denotes a set of N extracted SAR-SIFT descriptors in one region.

LLC convents each SAR-SIFT descriptor into a M-dimensional code shown in Fig. 2, which is implemented by the following three steps.

1) For the input SAR-SIFT descriptor $\mathbf{f}_n$, $n \in [1, N]$, its five nearest neighbors in $\mathbf{CB}$ are found to form the local base $\mathbf{LB}_n$.
2) Its local code $\tilde{\mathbf{c}}_n$ is obtained using the $\mathbf{LB}_n$ by solving the objective function

$$\min \|\mathbf{f}_n - \mathbf{LB}_n \tilde{\mathbf{c}}_n\|^2 \text{s.t.} \|\tilde{\mathbf{c}}_n\|_1 = 1. \tag{1}$$

3) The code $\mathbf{c}_n$, which is an M-dimensional vector with five non-zero elements whose values are the corresponding local code $\tilde{\mathbf{c}}_n$, is acquired. Finally, the final mid-level features of the region are generated by max pooling all the obtained codes within the region.

III. WSL-Based Target Discrimination

Fig. 3 illustrates the framework of the WSL-based target discrimination. Given a training set with weak labels only indicating whether an image contains the targets of interest or not, the candidate regions can be obtained via the log-normal-based constant false alarm rate (CFAR) and target clustering method [2], which are denoted by red rectangles in the detected SAR images shown in Fig. 3. The candidate regions extracted from the positive training images are referred to as positive bag, which contains not only target regions but also CFAs. The candidate regions extracted from the negative

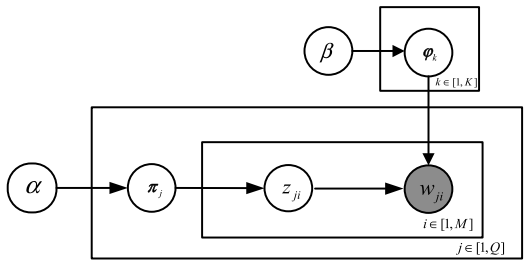


Fig. 4. Generated model of LDA.

training images are called negative bag, where all regions are CFAs. After generating the candidate regions, the next step is to construct feature representation for them. In the later discrimination task, targets and CFAs are regarded as positive examples and negative examples respectively. We use the mid-level features described in Section II to represent the candidate regions.

Based on the mid-level features, the WSL-based target discrimination contains two stages: training and test. The goal of the training stage is to learn a target discriminator. In the test stage, the learned target discriminator is applied to discriminate the targets and CFAs. The training stage includes two steps: training set initialization and training set update.

A. Training Set Initialization

By applying log-normal-based CFAR and the clustering method as preprocessing, we utilize the images with targets to construct a positive bag $S^+ = \{\mathbf{s}_p^+ | p \in [1, P]\}$ and the images without target to form a negative bag $S^- = \{\mathbf{s}_q^- | q \in [1, Q]\}$. As shown in Fig. 3, the positive bag contains not only a few positive examples but also numerous negative examples. Therefore, we first need to select potential positive examples from S^+ to generate the initial positive set S_0^+. The mid-level feature we design for the SAR target discrimination is based on the statistical histogram of local SAR-SIFT descriptors and has the meaning of word occurrence. It can turn the candidate region into a vector of word occurrence, thus obtaining the statistical information of the candidate region which is beneficial for the statistical modeling. Therefore, we utilize the statistical property of mid-level feature to select the initial positive samples. Specifically, the LDA [42] is used to initialize the target training set. We use the negative bag S^- from the negative training images for LDA learning. We denote each candidate region in the negative bag as a document in a corpus, and the M entries in the codebook as the visual words, thus the corresponding mid-level features can be denoted as the occurrence frequency of the M words.

LDA is a generative model shown in Fig. 4. The generative process for j-th candidate region in the S^- is as follows, where $j \in \{1, 2, ..., Q\}$:

1) Sample $\boldsymbol{\pi}_j \sim \text{Dir}(\alpha)$;
2) For $i \in \{1, 2, ..., M\}$:
   a) Sample a topic $z_{ji} = k \sim \text{Mult}(\boldsymbol{\pi}_j)$, $z_{ji} \in T = \{1, 2, ..., K\}$, where T is the set of topics;
   b) Sample $\boldsymbol{\varphi}_{z_{ji}} \sim \text{Dir}(\beta)$;
   c) Sample a word w_{ji} from $w_{ji} \sim Mult(\boldsymbol{\varphi}_{z_{ji}})$.

where $Dir(\cdot)$ and $Mult(\cdot)$ indicate the Dirichlet distribution and multinomial distribution, respectively. The joint probability distribution of LDA is given by:

$$p(\mathbf{s}_j^-, \boldsymbol{\pi}_j, \mathbf{z}_j, \boldsymbol{\varphi} | \alpha, \beta) = \prod_{i=1}^{M} p(\boldsymbol{\pi}_j | \alpha) p(z_{ji} = k | \boldsymbol{\pi}_j) p(\boldsymbol{\varphi} | \beta) p(w_{ji} | \varphi_{z_{ji}=k}) \quad (2)$$

where $\mathbf{z}_j = \{z_{ji}\}_{i=1}^M$, $\boldsymbol{\varphi} = \{\boldsymbol{\varphi}_{z_{ji}}\}_{i=1}^M$. Marginalizing over the latent variables $\boldsymbol{\pi}_j \mathbf{z}_j$ and $\boldsymbol{\varphi}$ determines the likelihood function $p(\mathbf{s}_j^- | \alpha, \beta)$ by

$$p(\mathbf{s}_j^- | \alpha, \beta) = \int_{\boldsymbol{\pi}_j} \int_{\boldsymbol{\varphi}} \sum_{\mathbf{z}_i} p(\mathbf{s}_j^-, \boldsymbol{\pi}_j, \mathbf{z}_i, \boldsymbol{\varphi} | \alpha, \beta) \quad (3)$$

Learning a LDA model from a corpus of negative bag $S^- = \{\mathbf{s}_q^- | q \in [1, Q]\}$ involves finding α and β that maximize the log likelihood of the data

$$l(\alpha, \beta) = \sum_{j=1}^{Q} \log p\left(\mathbf{s}_j^- | \alpha, \beta\right) \quad (4)$$

This parameter estimation problem can be solved by the variational Bayesian EM algorithm developed in [42].

Having learned the LDA using the negative bag, we can calculate the likelihood of each candidate region in the positive bag via (3), which measures how likely it is a negative example. It is a fact that the latent topic distributions of different classes are different. Therefore, a threshold Th, which is set as the minimal log likelihood value of the example in the negative bag multiplied by a proportion coefficient ε, can be used to select initial positive training examples set $\mathbf{s}_p^+$ as

$$S_0^+ = \left\{\mathbf{s}_p^+ | \log p(\mathbf{s}_p^+ | \alpha, \beta) < Th, p \in [1, P]\right\} \quad (5)$$

where S_0^+ denotes the initial positive training set. We hope the true positive examples in the selected examples are as many as possible to confirm the purity of initial positive training set, which is beneficial for the classifier learning. As for negative training set S_0^-, we select examples with some minimal log likelihood values of S^-, which has the same as S_0^+; In Section IV-C, the experimental analysis will show that how the proportion coefficient ε affects the discrimination performance.

B. Training Set Update

In order to avoid bringing negative examples into the initial positive set, we only choose a few training examples by LDA. However, a small number of examples cannot express the diversity of examples, and may result in overfitting. To tackle this problem, we train the SVM with iteratively updated training set until obtaining the optimal classifier. The initial training set is first used to learn a SVM. Then, the learned SVM is used to select a fixed number of new training examples, in which the examples classified to be positive are added to the training set and the remained examples will be put back. The updated training set is, in turn, used to retrain the SVM. This iteration process goes until the number of remained training examples is

less than the fixed number of new training examples selected at each iteration or the training set remains unchanged, i.e., there are no positive examples in the selected training examples. Due to the low training cost and high efficiency, the linear SVM is used as the discriminator in our method [45]. Although SVM is a strong learner which is opposite to the weak learner, weak learners and WSL are different concepts. WSL is opposite to the fully supervised learning. WSL means a machine learning framework where the model is trained using examples that are only partially annotated or labeled [28]. Weak learners are opposite to the strong learners. The weak learner, such as cart, is referred to as a classifier whose accuracy is a little better than random guess [46].

The proposed training algorithm is shown in Algorithm 1. As for the input parameters, S_0^+, S_U^+, and S^- correspond to the initial positive training set, the remaining examples from the positive bag and the negative bag, respectively. P_0^+, P_U^+, and P^- are the log likelihood values corresponding to S_0^+, S_U^+, and S^-, respectively. As for the output parameters, $\mathbf{w}^*$ and b^* are the model parameters and bias of the final target discriminator.

From the above description, we can see that, different from the WSL methods for optical images which need a large number of positive bags [35]–[39], our WSL method only requires a single positive bag and a single negative bag both in the training set initialization and the training set update steps. It is consistent with the fact that a large number of positive bags cannot be obtained for the SAR target discrimination application.

Algorithm 1 Procedure of Iterative Discriminator Training

Input:

S_0^+ : the initial positive training set
S_U^+ : the remaining examples from positive bag
S_0^- : the initial negative training set
S^- : the negative bag
P_0^+ : the log likelihood values corresponding to S_0^+
P_U^+ : the log likelihood values corresponding to S_U^+
P^- : the log likelihood values corresponding to S^-

Output:

$\mathbf{w}^*$: the mode parameters of linear SVM
b^* : the bias of linear SVM

1.Initialize:

Set the number of selected examples u

2. for the number of $S_U^+ > u$ **do**

2.1 Update the negative training set S_0^- by selecting examples with some minimal log likelihood values of S^-, which has the same size as S_0^+;

2.2 Learn a linear SVM based on the training set S_0^+ and S_0^-, and obtain the parameter $\mathbf{w}$ and b of the discriminator;

2.3 Create a pool $S_{U'}^+$ of examples by choosing u examples with minimal log likelihood values from the S_U^+;

2.4 Using the learned SVM to label the examples of $S_{U'}^+$, and count the number of positive examples n_p;

2.5 if $n_p > 0$ then

2.6 Update the parameters $\mathbf{w}^* = \mathbf{w}$ and $b^* = b$;

2.7 Add the $n'\left(n' \le n_p\right)$ positive examples with $\left(w^* * s_p + b^*\right) > \tau$ to the S_0^+, then remove them and their log likelihood values from S_U^+ and P_U^+;

2.8 else then

break;

2.9 end if

2.10. **end for**

IV. Experiments

A. A. Data Set Description

The miniSAR[1] real data set is acquired by Sandia National Laboratories of America. The SAR images in the miniSAR real data set are with complex scenes, in which the vehicles are the positive examples to be detected and the others, such as buildings, trees, grasslands, and roads, are treated as the negative examples, and so on. We select nine typical SAR images with complex scenes from the miniSAR real data set for our experiments. Their resolutions are 0.1 m × 0.1 m and their sizes are 1638 × 2510 pixels. Five randomly selected images with weak labels are used as the training set (three images containing vehicles as positive training images and two images not containing any vehicles as negative training images), and the remaining four images are used as the test images. Fig. 5 shows two training images, including a positive image and a negative image, and two test images used in our experiments with manually marked annotation diagrams showing vehicle targets and different kinds of clutter.

In the experiments, we use the log-normal-based CFAR and the target clustering method to obtain the candidate regions from the training and test images. For the convenience of the following evaluations, we also manually label all candidate regions in both training data and test data. After the detection stage, there are large amount of CFAs in the positive bag. We randomly select some examples from positive and negative bags and show them in Fig. 6. In Fig. 6(a), the examples in the third and fourth rows are CFAs. Table I gives the detection results of the nine SAR images. The number of dictionary $M = 1024$ empirically when generating the mid-level features for each candidate region, and the number of topics $K = 9$ in LDA.

TABLE I
Detection Results of Nine SAR Images

Detection results	Number of candidate regions	Number of positive examples	Number of negative examples
Positive training images	469	156	313
Negative training images	333	0	333
Test images	426	153	273

B. Evaluation of the Mid-Level Features

The mid-level features are used to describe the candidate regions obtained from the detection stage in this article.

[1] http://www.sandia.gov/radar/imagery

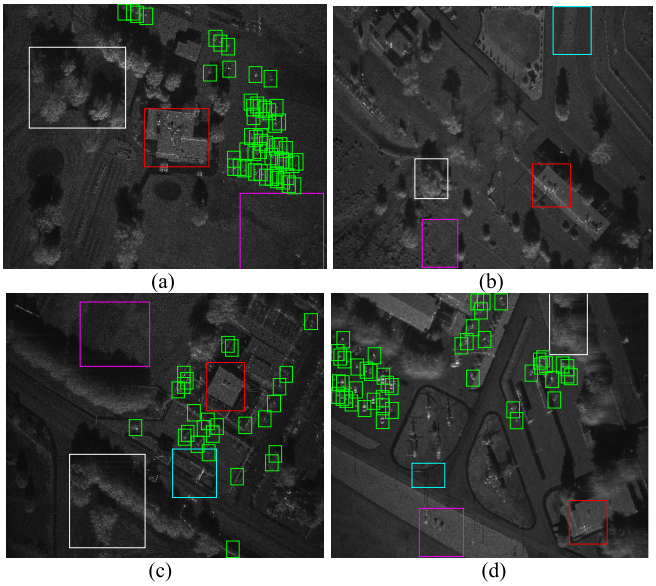


Fig. 5. Some images used in our experiments. (a) and (b) Training images, in which (a) is a positive image and (b) is a negative image. (c) and (d) Test images, where the green rectangles represent vehicle targets, the red rectangles represent buildings, the white rectangles represent trees, the cyan rectangles represent roads and the magenta rectangles represent grasslands.

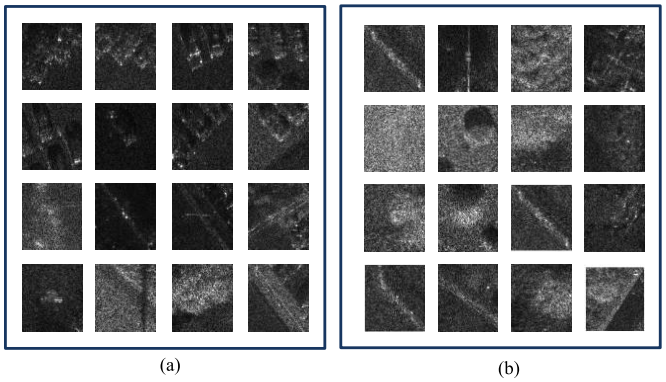


Fig. 6. Some examples of (a) positive and (b) negative examples.

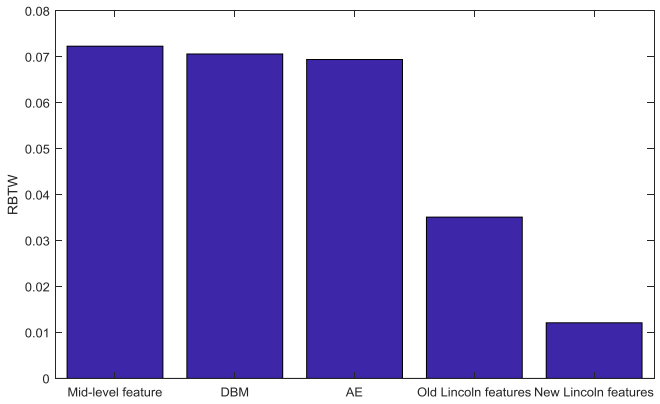


Fig. 7. Average RBTW values of the mid-level features, the two classical discrimination feature sets and high-level features extracted with DBM and AE under 50 runs with randomly selecting the samples.

TABLE II

COMPARISON OF THE DISCRIMINATION PERFORMANCE BETWEEN THE CLASSICAL DISCRIMINATION FEATURES, THE MID-LEVEL FEATURE, AND HIGH-LEVEL FEATURES EXTRACTED WITH DBM AND AE UNDER FULLY SUPERVISED LEARNING WITH THE LINEAR SVM CLASSIFIERS

Discrimination features	Probability of detection *pd*	Probability of false alarm *pfa*	F1-score
Mid-level features	0.8077	**0.1022**	**0.8115**
Old Lincoln features	0.8390	0.2880	0.7133
New Lincoln features	0.4640	0.2007	0.5089
High-level feature extracted with DBM	0.8889	0.2930	0.7371
High-level feature extracted with AE	**0.9420**	0.3630	0.7275

In order to demonstrate the effectiveness of the mid-level features, we compare them with two classical discrimination feature sets, i.e., the old Lincoln features [3] and the new Lincoln features [10], and two high-level features extracted with unsupervised deep models, including DBM and AE.

1) Analysis of Feature Separability: To measure the linear separability of the discrimination features, the ratio of between-class distance to within-class distance (RBTW) [46] is adopted as the criterion in this article. We randomly select 150 positive examples and 150 negative examples from the detection results of positive images to compare the linear separability of the mid-level features and the classical discrimination feature set. Fig. 7 shows the average RBTW values of the mid-level features, the two classical discrimination feature sets and the high-level features extracted with DBM and AE from 50 runs with randomly selecting the examples. From Fig. 7, we can see that the RBTW value of the mid-level features is much larger than those of the two classical discrimination feature sets and a little larger than those of high-level features extracted with DBM and AE. That is to say, the mid-level features can linearly separate the positive examples from the negative examples more easily than the classical discrimination feature sets and high-level features extracted with DBM and AE.

2) Analysis of Discrimination Results: To validate the discrimination performance of the mid-level features, 150 positive examples and 150 negative examples from the detection results of positive images are randomly selected as the training set, and the candidate regions from the detection results of test images are used for test. A linear SVM is adopted to classify positive examples and negative examples. The discrimination experiment is executed 100 runs. Table II gives the comparison of average discrimination results of the mid-level features, two classical Lincoln feature sets and high-level features extracted with DBM and AE to demonstrate the advantage of our proposed mid-level feature. For the sake of fairness, these three features are compared under fully supervised learning with the linear SVM classifiers. In Table II, the probability of detection pd, the probability of false alarm pfa, and F1-score

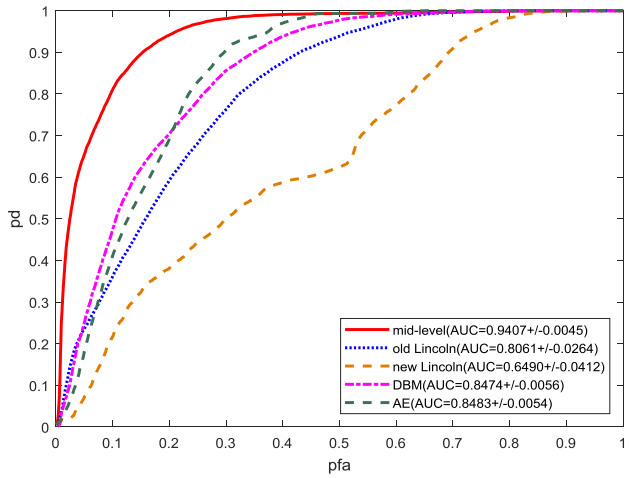


Fig. 8. Average ROC curves and the corresponding AUCs of different discrimination features for 100 runs using the linear SVM.

are calculated as follows:

$$pd = \frac{\text{TP}}{\text{NP}} \tag{6}$$

$$pfa = \frac{\text{FP}}{\text{NN}} \tag{7}$$

$$\text{F1-score} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{recall} = \text{pd} \tag{9}$$

where TP is the number of positive examples detected as positive examples, NP is the number of positive examples, FP is the number of negative examples detected as positive examples, and NN is the number of the negative examples. Since pd and pfa cannot fully describe the performance of the discriminator, we also use the F1-score to evaluate the discrimination performance. F1-score can evaluate the discrimination performance more comprehensively for taking into account both pd and precision. The higher the F1-score is, the more robust the discrimination is. From Table II, we can see that the mid-level features get lower pd than the old Lincoln features and high-level features extracted with DBM and AE, but get much lower pfa than the two classical Lincoln feature sets and high-level features. In addition, the mid-level can get the highest F1-score. To compare the performance of different discrimination features more comprehensively, we give the average receiver operating characteristic (ROC) curves and the corresponding area under the ROC curves (AUCs) of different discrimination features for 100 runs using the linear SVM in Fig. 8. As shown in Fig. 8, the average ROC curve of the mid-level feature is the best among all discrimination features, and the corresponding AUC of the mid-level feature is also much larger than the other discrimination features. Based on these results, we can draw the conclusion that the mid-level features are more competitive than the two classical feature sets, and high-level features extracted with DBM and AE to discriminate the positive examples and negative examples.

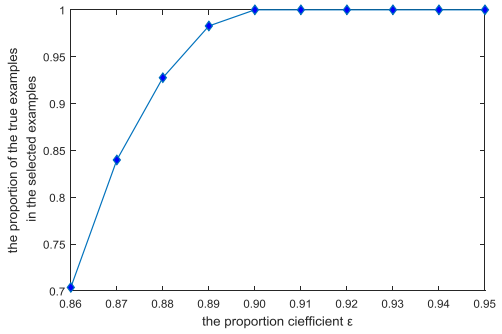
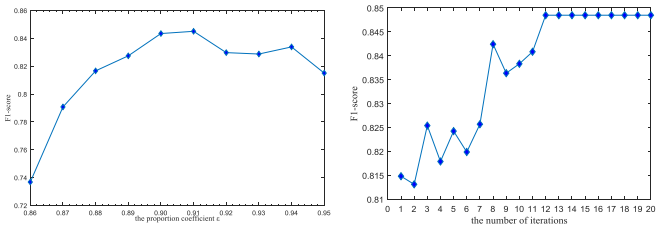


Fig. 9. Influence of the coefficient ε to the ratio of the number of selected examples to the number of true positive examples in the selected examples.



(a) (b)

Fig. 10. Influence of the (a) coefficient ε on discrimination performance and (b) Number of iterations on discrimination performance when $\epsilon = 0.91$.

C. Analysis of the Parameter

As described in Section III, the proposed WSL method consists of two steps, training set initialization and training set update. In the implementation of training set initialization, the selected threshold Th in (5) may affect the initialization result and then affect the following discriminator training. Since the threshold Th is determined by the proportion coefficient ε, we perform experiments with different values of ε to show how the value of ε affects the discrimination result. Fig. 9 shows the ratio of the number of selected examples to the number of true positive examples in the selected examples versus the proportion coefficient ε. As mentioned in Section III, it is expected that the true positive examples in the selected examples are as many as possible, which is beneficial for the classifier learning. From Fig. 9, we can see that the ratio is close to 1 when $\varepsilon \in (0.86, 0.95]$ and especially equal to 1 when $\varepsilon \geq 0.9$. This indicates that the influence of proportion coefficient ε on the ratio of the number of selected examples to the number of true positive examples in the selected examples is weak. Fig. 10(a) shows the variations of the discrimination performance with different proportion coefficients. We randomly select some examples from initial positive and negative training sets when $\varepsilon = 0.91$ and show them in Fig. 11. From Fig. 11(a), we can see that there are no CFAs in the initial positive training set, which demonstrates the effectiveness of using the LDA model to select positive examples. From Fig. 10(a), we can see that the discrimination performance change is small as ε increase. When $\varepsilon = 0.91$, the F1-score is the highest and the discrimination performance is the best. Therefore, we set the coefficient $\varepsilon = 0.91$ in our final experiment. Fig. 10(b) shows the variations of the

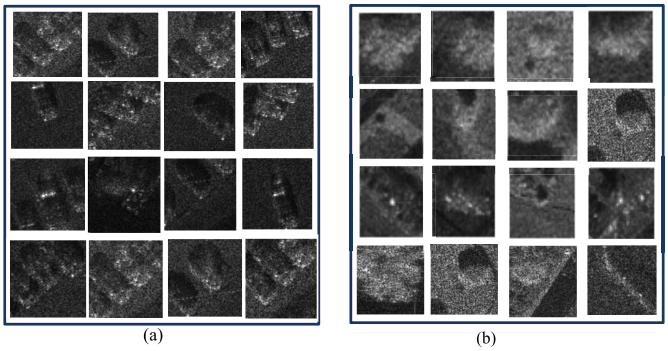


Fig. 11. Some examples in (a) Initial positive and (b) negative training sets.

discrimination performance with iteration when $\varepsilon = 0.91$. In each iteration, new selected samples are added into the training set. From Fig. 11(b), we can see that the F1-score is almost on the rise as the iterations increases, and then stops growing at the 12th iteration for the training examples have stopped updating at this point, which indicates the update of training examples is beneficial for training the discrimination. In our method, the SVM stops updating when the number of remained training examples is less than the fixed number of new training examples selected at each iteration or the training set remains unchanged, i.e., there are no positive examples in the selected training examples. In this experiment, the training set stops updating at the 12th iteration when the second termination condition is reached. In order to analyze the effect of updating the training set, here we give the results of 20 iterations in Fig.11(b).

D. Evaluation of the Target Discriminator

We evaluate the performance of the WSL-based target discriminator by comparing it with eight supervised discrimination methods and a WSL method [32]. The WSL method in [32] also consists of two stages: the training set initialization and the iterative update of target detector. The main difference of our method and WSL method in [32] is that we use the LDA model learned with the negative examples to select the initial positive examples, while the latter selects the initial positive examples based on the Manhattan distance. The manner of selecting positive samples in [32] does not take the distribution of samples into consideration and is easily disturbed by data sparsity and noise. While we employ the LDA model to mine the underlying information of samples and thus can reduce the influence of data sparsity and noise. For the fair comparison, in these experiments, we use the same the same test images. The setup of the reference WSL method is the same as that of our proposed WSL method. The setups of reference supervised methods are as follows, in which linear SVM, the quadratic polynomial discriminator (QPD) [48] and the quadratic distance discriminator (QDD) [1] are supervised binary classifiers, and linear support vector data description (SVDD) is a supervised one-class classifier. Moreover, in our proposed method, the negative examples are accurate. Therefore, linear SVDD is trained on the negative examples to ensure the fair comparison.

TABLE III

DISCRIMINATION PERFORMANCE COMPARISON BETWEEN OUR PROPOSED METHOD AND THE REFERENCE METHODS FOR THE ALL TEST IMAGES

Methods		Probability of detection *pd*	Probability of false alarm *pfa*	F1-score
Our proposed weakly supervised method		**0.9150**	0.1355	**0.8485**
WSL method in [26]		0.8301	0.1868	0.7674
The supervised methods	Linear SVM 1	0.7843	0.0956	0.8023
	Linear SVM 2	0.8150	0.1044	0.8142
	QDD1	0.6997	0.1881	0.6869
	QDD2	0.5670	0.1295	0.6287
	QPD1	0.7448	0.3078	0.6491
	QPD2	0.7367	0.3228	0.6370
	Linear SVDD 1	0.3922	0.0916	0.3625
	Linear SVDD 2	0.1373	**0.0147**	0.1355

Linear SVM 1: Linear SVM trained on the 156 negative examples randomly selected from the negative bag and 156 positive examples from the positive bag.

Linear SVM 2: Linear SVM trained on the 156 negative examples and 156 positive examples both from the positive bag.

QPD 1: QPD trained on the 156 negative examples randomly

selected from the negative bag and 156 positive examples from the positive bag.

QPD 2: QPD trained on the 156 negative examples and 156 positive examples both from the positive bag.

QDD 1: QDD trained on the 156 negative examples randomly selected from the negative bag and 156 positive examples from the positive bag.

QDD 2: QDD trained on the 156 negative examples and 156 positive examples both from the positive bag.

Linear SVDD 1: Linear SVDD trained on the 333 negative examples from the negative bag.

Linear SVDD 2: Linear SVDD trained on the 313 negative examples from the positive bag.

It should be noted that SVM and SVDD are based on the mid-level features, the QPD and QDD are based on the traditional Lincoln features rather than the mid-level features. This is because the QDD maps the n-dimension feature space into the 1-D feature space and its performance greatly decreases when the feature dimension is high. The mid-level feature is with high dimension and the discrimination performance is considerably bad when applying the mid-level features into QDD. For QPD, the parameters needed to be learned are proportional to the square of feature dimension and therefore, it is always applied to features with low dimensions.

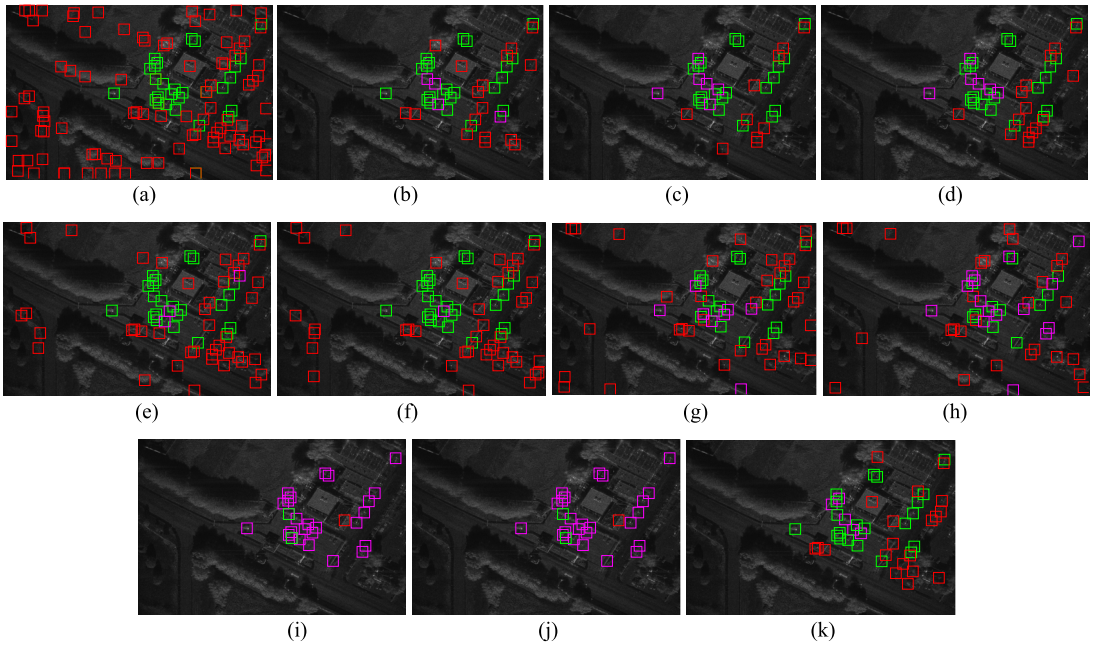


Fig. 12. Annotation diagram of discrimination results on the images shown in Fig. 5(c) via different discrimination methods, where the green rectangles represent correctly discriminated vehicle targets, the red rectangles represent false alarms and the magenta rectangles represent missed targets. (a) Reference image. (b) The proposed method. (c) Linear SVM 1. (d) Linear SVM 2. (e) QDD1. (f) QDD2. (g) QPD1. (h) QPD2. (i) Linear SVDD 1. (j) Linear SVDD 2. (k) WSL in [32].

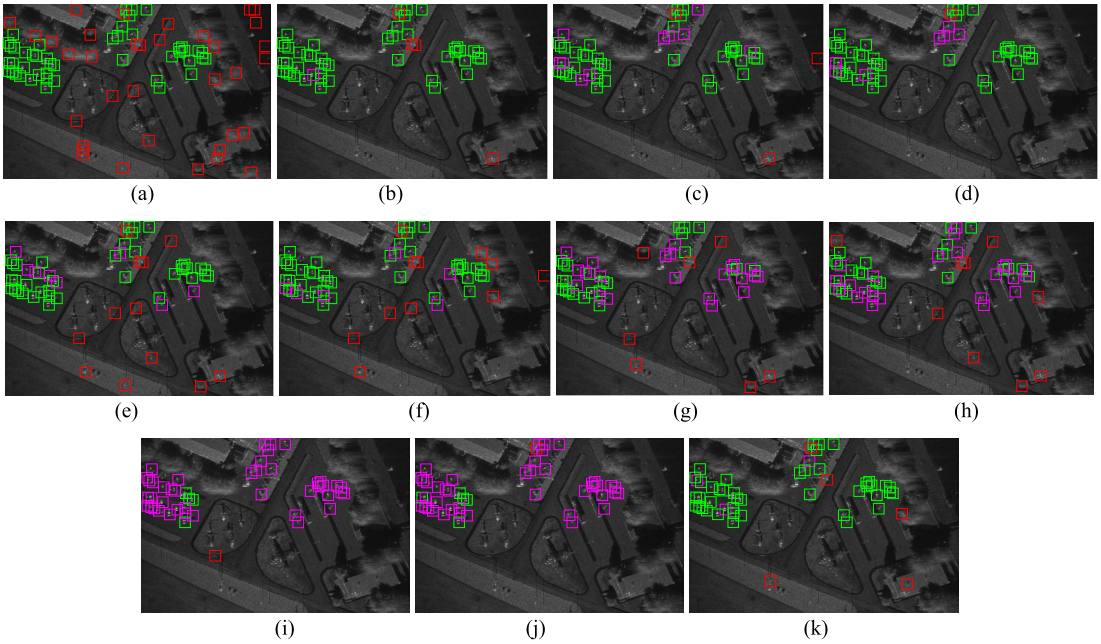


Fig. 13. Annotation diagram of discrimination results on the images shown in Fig. 5(d) via different discrimination methods, where the green rectangles represent correctly discriminated vehicle targets, the red rectangles represent false alarms, and the magenta rectangles represent missed targets. (a) Reference image. (b) The proposed method. (c) Linear SVM 1. (d) Linear SVM 2. (e) QDD1. (f) QDD2. (g) QPD1. (h) QPD2. (i) Linear SVDD 1. (j) Linear SVDD 2. (k) WSL in [32].

In addition, the old Lincoln features, new Lincoln features, and their combination are used as the inputs of QDD and QPD. In the following experiments, we only give the best results.

Table III gives the comparison of our proposed WSL framework, WSL framework in [32], the supervised SVM, QDD, QPD, and one-class classifier SVDD to illustrate the effectiveness of our proposed WSL framework. From Table III, we can see that pfa of our proposed method is less than those of QPDs 1 and 2, and a little more than those of other methods. However, our proposed method can obtain much higher F1-score and pd than other methods, which demonstrates the superiority of our method compared with the reference methods. In addition, the supervised linear SVMs 1 and 2 could achieve the equivalent results when different negative examples are used to train the linear SVM; while for SVDD, the results are very different when using different

negative examples. That is to say the linear SVM is insensitive to the source of negative examples, but the linear SVDD is sensitive to the source of negative examples. This is because the linear SVM, as a two-class classifier, is trained with both positive examples and negative examples; while the linear SVDD is a one-class classifier and trained only with negative examples. In this experiment, the similarity between positive examples and negative examples from positive images is higher than that between positive examples and negative examples from negative images, thus the linear SVDD 2 performs worse than the linear SVDD 1. Although our proposed discrimination method uses the same supervised information as the linear SVDD 1, it also exploits the information of negatives examples to initialize positive examples by LDA, and then takes advantage of the information for both of them to gradually refine the target discriminator. Therefore, our proposed method can achieve better performance than the linear SVDD. In addition, from Table III, we can see that the discrimination performance of our proposed method is better than that of WSL method in [32], which demonstrates that the manner of selecting initial positive samples in our proposed method is superior to that in [32].

To better illustrate the effectiveness of our approach, in Figs. 12 and 13, we give the annotation diagrams of discrimination results of the test images showed in Fig. 5(c) and (d). Figs. 12 and 13(a) show the reference images with the candidate regions manually annotated, where the green rectangles represent vehicle targets and the red rectangles represent clutters. When we label the candidate regions by hand, the candidate regions containing vehicle targets are regarded as the targets and those containing no vehicle targets are regarded as the false alarms. Figs. 12 and 13(b)–(k) show the annotation diagrams of discrimination results via our method, linear SVMs 1 and 2, QDDs 1 and 2, QPDs 1 and 2, Linear SVDDs 1 and 2, and WSL method in [32], respectively, where the green rectangles represent correctly discriminated vehicle targets, the red rectangles represent false alarms and the magenta rectangles represent missed targets. From Figs 12 and 13, we can see that the missed targets of our proposed method are much less than those of other methods£" but the false alarms of our proposed method are a little more than those of other method. As discussed in Section IV-B2, compared with separately using pd and pfa to evaluate a discrimination method, F1-score can evaluate the discrimination performance more comprehensively for taking into account both pd and precision. As given in Table III, based on the F1-scores, we can obtain the conclusion that the discrimination ability of our proposed method is better than those of the supervised linear SVMs, QDD, QPD, SVDDs, and WSL method in [32].

V. Conclusion

In this article, we have proposed a novel framework to tackle the problem of target discrimination for high-resolution SAR images in complex scenes. There are two main contributions compared with the previous target discrimination methods for SAR images. First, the mid-level features are used for SAR target discrimination, which offer sufficiently powerfulness to characterize the structural information of the image regions. Second, the proposed WSL discrimination method can not only avoid the manual annotation of target regions from SAR images but also achieve the promising discrimination performance.

References

[1] L. M. Novak, G. J. Owirka, and C. M. Netishen, "Performance of a high-resolution polarimetric SAR automatic target recognition system," *Lincoln Lab. J.*, vol. 6, no. 1, pp. 11–24, 1993.

[2] G. Gao, L. Liu, L. Zhao, G. Shi, and G. Kuang, "An adaptive and fast CFAR algorithm based on automatic censoring for target detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 6, pp. 1685–1697, Jun. 2009.

[3] D. E. Kreithen, S. D. Halversen, and G. J. Owirka, "Discriminating targets from clutter," *Lincoln Lab. J.*, vol. 6, no. 1, pp. 25–51, 1993.

[4] G. Gao, "An improved scheme for target discrimination in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 49, no. 1, pp. 277–294, Jan. 2011.

[5] L. M. Novak, G. J. Owirka, W. S. Brower, and A. L. Weaver, "The automatic target-recognition system in SAIP," *Lincoln Lab. J.*, vol. 10, no. 2, pp. 187–202, 1997.

[6] J.-I. Park, S.-H. Park, and K.-T. Kim, "New discrimination features for SAR automatic target recognition," *IEEE Geosci. Remote Sens. Lett.*, vol. 10, no. 3, pp. 476–480, May 2013.

[7] T. Li and L. Du, "Target discrimination for SAR ATR based on scattering center feature and K-center one-class classification," *IEEE Sensors J.*, vol. 18, no. 6, pp. 2453–2461, Mar. 2018.

[8] N. Wang, Y. Wang, H. Liu, Q. Zuo, and J. He, "Feature-fused SAR target discrimination using multiple convolutional neural networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1695–1699, Oct. 2017.

[9] Z. Wang *et al.*, "Visual attention-based target detection and discrimination for high-resolution SAR images in complex scenes," *IEEE Trans. Geosci. Remote Sens.*, vol. 14, no. 10, pp. 1855–1872, Oct. 2017.

[10] S. M. Verbout, A. L. Weaver, and L. M. Novak, "New image features for discriminating targets from clutter," *Proc. SPIE*, vol. 3395, pp. 120–137, Aug. 1998.

[11] M. Greenspan, L. Pham, and N. Tardella, "Development and evaluation of a real time SAR ATR system," in *Proc. IEEE RADAR*, May 1998, pp. 38–43.

[12] G. Gao, G. Kuang, Q. Zhang, and D. Li, "Fast detecting and locating groups of targets in high-resolution SAR images," *Pattern Recognit.*, vol. 40, no. 4, pp. 1378–1384, Apr. 2007.

[13] Y. Lin, "Feature synthesis and analysis by evolutionary computation for object detection and recognition," Ph.D. dissertation, Univ. California, Riverside, CA, USA, 2003.

[14] B. Bhanu and Y. Lin, "Genetic algorithm based feature selection for target detection in SAR images," *Image Vis. Comput.*, vol. 21, pp. 591–608, Jul. 2003.

[15] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. CVPR*, Jun. 2006, pp. 2169–2178.

[16] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3360–3367.

[17] Z. Wang, L. Du, and H. Su, "Superpixel-level target discrimination for high-resolution SAR images in complex scenes," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 11, no. 9, pp. 3127–3143, Sep. 2018.

[18] L. J. Li, H. Su, L. Fei-Fei, and E. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *Proc. NIPS*, 2010, pp. 1378–1386.

[19] L. Torresani, M. Szummer, and A. Fitzgibbon, "Efficient object category recognition using classemes," in *Proc. ECCV*, 2010, pp. 776–789.

[20] N. Payet and S. Todorovic, "Scene shape from texture of objects," in *Proc. CVPR*, Jun. 2011, pp. 2017–2024.

[21] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. ICCV*, Sep. 2009, pp. 1365–1372.

[22] A. Farhadi, I. Endres, and D. Hoiem, "Attribute-centric recognition for cross-category generalization," in *Proc. CVPR*, Jun. 2010, pp. 2352–2359.

[23] M. A. Sadeghi and A. Farhadi, "Recognition using visual phrases" in *Proc. CVPR*, Jun. 2011, pp. 1745–1752.

[24] Q. Lv, Y. Dou, X. Niu, J. Xu, J. Xu, and F. Xia, "Urban land use and land cover classification using remotely sensed SAR data through deep belief networks," *J. Sens.*, vol. 2015, Jan. 2015.

[25] O. A. Penatti, K. Nogueira, and J. A. dos Santos, "Do deep features generalize from everyday objects to remote sensing and aerial scenes domains?" in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2015, pp. 44–51.
[26] X. Yao, J. Han, G. Cheng, X. Qian, and L. Guo, "Semantic annotation of high-resolution Satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3671, Jun. 2016.
[27] Z.-H. Zhou, "A brief introduction to weakly supervised learning," *Nat. Sci. Rev.*, vol. 5, no. 1, pp. 44–53, 2017.
[28] L. Torresani, *Weakly Supervised Learning*. New York, NY, USA: Springer, 2014.
[29] M. Rochan and Y. Wang, "Weakly supervised localization of novel objects using appearance transfer," in *Proc. CVPR*, Jun. 2015, pp. 4315–4324.
[30] C. Wang, K. Huang, W. Ren, J. Zhang, and S. Maybank, "Large-scale weakly supervised object localization via latent category learning," *IEEE Trans. Image Process.*, vol. 24, no. 4, pp. 1371–1385, Apr. 2015.
[31] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 6, pp. 3325–3337, Jun. 2015.
[32] D. Zhang, J. Han, G. Cheng, Z. Liu, S. Bu, and L. Guo, "Weakly supervised learning for target detection in remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 4, pp. 701–705, Apr. 2015.
[33] X. Yao, J. Han, G. Cheng, X. Qian, and L. Gao, "Semantic annotation of high-resolution satellite images via weakly supervised learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 6, pp. 3660–3670, Jun. 2016.
[34] L. Lin, K. Wang, D. Meng, W. Zuo, and L. Zhang, "Active self-paced learning for cost-effective and progressive face identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 1, pp. 7–19, Jan. 2018.
[35] S. Vijayanarasimhan and K. Grauman, "Keywords to visual categories: Multiple-instance learning forweakly supervised object categorization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[36] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr. 2015.
[37] M. H. Nguyen, L. Torresani, F. De La Torre, and C. Rother, "Weakly supervised discriminative localization and classification: A joint learning process," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep. 2010, pp. 1925–1932.
[38] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly supervised object localization with multi-fold multiple instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 1, pp. 189–203, Jan. 2017.
[39] R. Cinbis, J. Verbeek, and C. Schmid, "Multi-fold MIL training for weakly supervised object localization," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2409–2416.
[40] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.
[41] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
[42] F.-F. Li and P. Perona, "A Bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2005, pp. 524–531.
[43] F. Dellinger, J. Delon, Y. Gousseau, J. Michel, and F. Tupin, "SAR-SIFT: A SIFT-like algorithm for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 53, no. 1, pp. 453–466, Jan. 2015.
[44] Y. Wang and H. Liu, "SAR target discrimination based on BOW model with sample-reweighted category-specific and shared dictionary learning," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 11, pp. 2097–2101, Nov. 2017.
[45] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learn. Res.*, vol. 9, pp. 1871–1874, Aug. 2008.
[46] C.-X. Zhang, J.-S. Zhang, and G.-Y. Zhang, "An efficient modified boosting method for solving classification problems," *J. Comput. Appl. Math.*, vol. 214, no. 2, pp. 381–392, May 2008.
[47] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York, NY, USA: Springer, 2006.
[48] J. C. Principe, A. Radisavljevic, J. Fisher, M. Hiett, and L. M. Novak, "Target prescreening based on a quadratic gamma discriminator," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 34, no. 3, pp. 706–715, Jul. 1998.

Lan Du (M'11–SM'17) received the B.S., M.S., and Ph. D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2001, 2004, and 2007, respectively. Her doctoral dissertation was granted by the National Excellent Doctoral Dissertation of China in 2009.

From September 2007 to September 2009, she was a Post-Doctoral Research Associate with the Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA. She is currently a Professor with the National Laboratory of Radar Signal Processing, Xidian University. Her research work is supported by NSFC for the Excellent Young Scholars and Chang Jiang Scholars Program for Young Scholars. Her main research interests include statistical signal processing and machine learning with application to radar target recognition.

Hui Dai received the B.S. and M.S. degrees in electronic engineering from Xidian University, Xi'an, China, in 2014 and 2017, respectively.

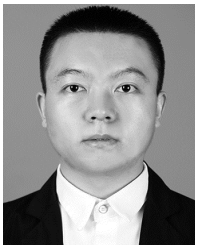
Her main research interests include synthetic aperture radar (SAR) target detection, image processing, and pattern recognition.

Yan Wang received the B.S. degree in information and communication from the Guilin University of Electronic Technology, Guilin, China, in 2013. She is currently pursuing the Ph.D. degree in signal processing with the National Laboratory of Radar Signal Processing, Xidian University, Xi'an, China.

Her research interests include synthetic aperture radar (SAR) image change detection, SAR target discrimination, image processing, and pattern recognition.

Weitong Xie received the B.S. and M.S. degrees in electronic engineering from Xidian University, Xi'an, China, in 2016 and 2019, respectively.

Her main research interests include synthetic aperture radar (SAR) automatic target recognition, image processing, and pattern recognition.



Zhaocheng Wang received the Ph.D. degree in electronic engineering from Xidian University, Xi'an, China, in 2018.

He is currently with the School of Electronic and Information Engineering, Hebei University of Technology, Tianjin, China. His research interests include synthetic aperture radar (SAR) automatic target recognition, remote sensing image processing, and pattern recognition.