IET Radar, Sonar & Navigation

Research Article

IET Journals
The Institution of Engineering and Technology

SAR object classification using the DAE with a modified triplet restriction

ISSN 1751-8784
Received on 8th September 2018
Revised 9th February 2019
Accepted on 26th February 2019
E-First on 28th March 2019
doi: 10.1049/iet-rsn.2018.5413
www.ietdl.org

Sirui Tian[1] ✉, Chao Wang[2,3], Hong Zhang[2], Bir Bhanu[4]

[1]*Department of Electronic Engineering, School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China*

[2]*Key Laboratory of Digital Earth Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100094, People's Republic of China*

[3]*College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China*

[4]*Department of Electrical and Computer Engineering, University of California, Riverside, Riverside, CA 92521, USA*

✉ *E-mail: tiansirui@njust.edu.cn*

Abstract: Although deep learning methods have made great progress in synthetic aperture radar (SAR)-based remote sensing, lack of training data has often been the major obstacle while they are adopted for SAR automatic target recognition. In this study, a new deep network in the form of a restricted three-branch denoising auto-encoder (DAE) is proposed to take the full advantage of limited training samples. In this model, a modified triplet restriction that combines the semi-hard triplet loss with the intra-class distance penalty is devised to learn discriminative features with a small intra-class divergence and a large inter-class divergence. Besides, the reconstruction distortion is measured between the model outputs and the images filtered by the improved Lee Sigma filter rather than the original inputs to suppress clutter in the background. Furthermore, a batch-based triplet loss, which calculates the modified triplet loss in a batch-based manner, is proposed to tackle the difficulties in implementation and reduce its computation complexity. The simplified version of the three-branch Triplet-DAE is subsequently devised as a one-branch DAE restricted by the batch-based triplet loss. Experimental results with the MSTAR data demonstrate the effectiveness of the proposed method on real SAR images.

1 Introduction

As the crucial task of automatic target recognition with synthetic aperture radar images (SAR ATR), feature engineering is aimed at finding discriminative features to distinguish objects in high-resolution SAR images [1, 2]. Although numerous papers have been published in the past decades, it is still a challenging task because of the complicated imaging mechanism, the increasing resolution of images, the complicated structures of targets and the complex land form of observing areas.

In general, the conventional hand-designed feature extraction methods include two categories: the generalised methods and SAR-specialised methods. The first type employs methods that have been successfully applied in other domains, to generate features for ATR considering few characteristics of SAR images [3–6]. The other one utilises the estimated parameters of scattering models as features for ATR [7–9]. Although these methods have achieved good accuracies with the benchmark dataset, they suffer from significant performance degradations when complicated situations occur including complex scattering structures in higher-resolution images, target distortion and speckle variation in complicated observing environments and missing pose in the training data. Accordingly, if these conditions are considered, it is necessary to find new approaches that can adaptively learn features from varying raw data [10].

With the recent theoretical progress, deep networks, which are proving adept at mining intrinsic information from raw data, have been introduced to SAR-based remote sensing tasks. Various deep learning (DL) models such as the auto-encoder (AEs) [10], the restricted Boltzmann machine (RBM) [11, 12] and the convolutional neural network (CNN) [13–21], were introduced to tackle with SAR image classification tasks and obtained comparable or even better results in comparison with the state-of-the-art results obtained by the hand-designed features. Motivated by their success, new DL models were developed for better performance and low computation complexity in SAR ATR. Malmgren-Hansen *et al.* [22] utilised the simulated images with various translations to train the translation invariant CNN. Andrew *et al.* [23] developed a CNN initialised by the pre-trained LeNet and trained it with augmented data created by resampling. Kechagias-Stamatis *et al.* [24] divided the AlexNet into eight clusters of layers, of which four trained clusters were employed to learn the features. Wagner [25] combined the CNN with support vector machine (SVM), providing a remarkable improvement for the classification result. Despite these successes, insufficient data and overfitting arising therefrom should be the main constraint on their feature learning capability, bringing in performance degradation when data with complex conditions are used.

On this account, various schemes are designed to deal with problems caused by limited labelled training samples. Geng *et al.* [26] devised an improved stacked AE (SAE) with reduced training parameters where the first two hidden layers were fixed layers successively computing the grey-level co-occurrence matrix and the Gabor transformation. Li *et al.* [27] decomposed the CNN into a series of convolutional AE and shallow neural networks, which were trained separately so that the requirement for labelled samples is decreased and the training speed is increased. Although it was reported that the model worked well for MSTAR dataset, it is obvious that using unsupervised AEs without restrictions as a substitution for supervised CNN would lead to model deterioration. Chen *et al.* [28] proposed an all-convolutional network (A-Convnet), where the fully connected layers were replaced by the convolutional layers to reduce the number of trainable parameters and alleviate overfitting caused by insufficient training samples. However, lack of sufficient experiments made it difficult to evaluate whether the substitution of the fully connected layers would affect the model's generalisation performance on other datasets.

Data augmentation (DA) strategies were also devised to create simulated training samples. Ding *et al.* [29] generated new training samples (by translating, adding speckle and pose synthesis) to train the CNN. Wagner *et al.* [30] and Wagner [31] employed similar

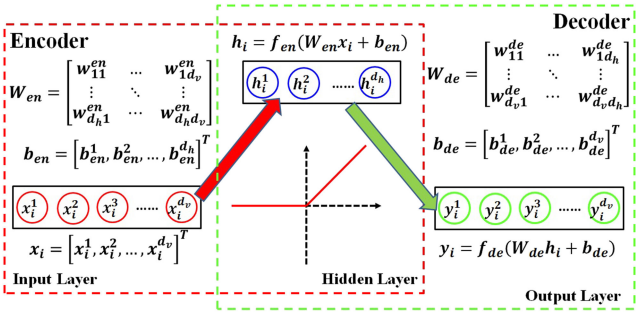

Fig. 1 *Structures of the simplest AE*

schemes, where affine transformation and elastic distortion were applied to produce additional samples. Even in the A-Convnet [28], DA process that randomly generated the training patches from the MSTAR dataset played a crucial role in alleviating overfitting caused by insufficient samples. Despite its advantages of creating sufficient and diversiform samples, the DA processes bring about certain problems including the increased cost of computation and the difficulties associated with determining the augmentation configurations (i.e. the size of the augmented dataset and the optimal manners of augmentation). Besides, the features learnt from the augmented data can amplify the inconsistency between the distribution of the measured dataset and that of the actual problem. Accordingly, it is expected that schemes without much man-made uncertainty can be devised. Deng *et al.* [32] proposed a new AE model restricted by the Euclidean distance (ED-AE). The restriction encouraged the intra-class distance of features to be a small value near zero and the inter-class distance to be close to a constant. Although the ED-AE is reported to have a significant improvement over the classical AE, the inexplicable constant for the inter-class distance made it difficult to be optimised.

In this paper, a new three-branch denoising AE (DAE) restricted by a modified triplet loss [triplet-DAE (TDAE)] is proposed for SAR ATR to take the full advantage of limited training samples. The modified triplet loss that combines the semi-hard triplet loss with the intra-class distance penalty encourages the model to learn discriminative features with small intra-class divergence and large inter-class divergence. Besides, the reconstruction distortion of the proposed model is measured by comparing the reconstructed data with the improved Lee Sigma (ILS) filtered [33] inputs rather than the original inputs to suppress the clutter and speckle in the background during training. A simplified version of the proposed TDAE is also devised to reduce the computation complexity and tackle with the difficulties in implementation, which is in the form of a one-branch DAE restricted by a batch-based triplet loss and the modified reconstruction loss.

The rest of this paper is organised in four sections. Section 2 discusses the related work. In Section 3, the three-branch TDAE and its one-branch simplified version will be discussed in detail. Experimental results and comparison with other ATR methods are shown and discussed in Section 4. Section 5 concludes this paper.

2 Related work

In this paper, the three-layer AE is utilised as the prototype of the proposed model. The major reason for this is that it is one of the simplest DL models, which is easy to be implemented and has great potential for improvement. It has limited trainable parameters in comparison with other complicated DL models that require less training samples, especially when tied weights and dropout scheme are used. Moreover, the multilayer networks can be conveniently constructed by stacking the pre-trained three-layer AEs and fine-tuning the parameters with limited samples. Furthermore, the classical AE has much lower computation complexity than the convolution-based architectures. In addition, the AE has a dimensionality reduction capability that produces features in a lower-dimensional space than the original data. In this section, a brief introduction of related work on AE will be presented including the principle of the AE, the SAE and the dropout scheme.

2.1 Principle of the three-layer AE model

As shown in Fig. 1, the three-layer AE is a symmetrical network consisting of an encoder and a decoder. The encoder maps the inputs into a representation space to generate the latent features while the decoder approximately reconstructs the inputs from the learnt features. The objective of the AE is minimising the distortion between the inputs and the reconstruction to guarantee the mapping process preserves the information of the inputs.

Considering a dataset $\boldsymbol{X} = \{x_i\}_{i=1}^{N}$ with N samples, let the d_{v}-dimensional vector $\boldsymbol{x}_i = \left[x_i^1, x_i^2, \ldots, x_i^{d_{\mathrm{v}}}\right]^{\mathrm{T}}$ be the ith sample. The latent feature $\boldsymbol{h}_i$ is generated by the encoder in the red-dashed line box of Fig. 1, i.e.

$$\boldsymbol{h}_i = f_{\mathrm{en}}(\boldsymbol{W}_{\mathrm{en}} x_i + \boldsymbol{b}_{\mathrm{en}}) \tag{1}$$

where $\boldsymbol{W}_{\mathrm{en}}$ and $\boldsymbol{b}_{\mathrm{en}}$ are the $d_{\mathrm{h}} \times d_{\mathrm{v}}$ weight matrixes and the d_{h}-dimensional bias of the encoder, respectively; f_{en} is the activation function that is usually the sigmoid function or the rectified linear unit (ReLU) function.

The decoder represented in the green-dashed line box of Fig. 1 subsequently maps the learnt representation back to the reconstruction $\boldsymbol{y}_i$ such that the reconstruction approximates the input, i.e. $\boldsymbol{y}_i \simeq \boldsymbol{x}_i$. The decoder is formulated as

$$\boldsymbol{y}_i = f_{\mathrm{de}}(\boldsymbol{W}_{\mathrm{de}} \boldsymbol{h}_i + \boldsymbol{b}_{\mathrm{de}}) \tag{2}$$

where f_{de} is the activation function of the decoder that can be the sigmoid function, the linear function or the ReLU function; $\boldsymbol{W}_{\mathrm{de}}$ and $\boldsymbol{b}_{\mathrm{de}}$ are the $d_{\mathrm{v}} \times d_{\mathrm{h}}$ weight and the d_{v}-dimensional bias of the decoder, respectively. Specifically, if $\boldsymbol{W}_{\mathrm{de}} = \boldsymbol{W}_{\mathrm{en}}^{\mathrm{T}}$ with $(\cdot)^{\mathrm{T}}$ being the transpose operation, the weight is called tied. In this paper, the tied weight is utilised. The AE is trained by finding the optimal parameters $\theta_{\mathrm{AE}} = \{\boldsymbol{W}_{\mathrm{en}}, \boldsymbol{b}_{\mathrm{en}}, \boldsymbol{b}_{\mathrm{de}}\}$ that minimise the reconstruction distortion $J_{\mathrm{AE}}(\theta_{\mathrm{AE}})$ on the training dataset $\boldsymbol{X}$, which is defined as

$$J_{\mathrm{AE}}(\theta_{\mathrm{AE}}) = \frac{1}{N} \sum_{i=1}^{N} L_{\mathrm{R}}(\boldsymbol{x}_i, \boldsymbol{y}_i) \tag{3}$$

where $L_{\mathrm{R}}(\cdot)$ is the reconstruction error of a given sample that is usually the root-mean-square error (MSE) or the MSE. The gradient descent algorithm is used to optimise the parameters by minimising $J_{\mathrm{AE}}(\theta_{\mathrm{AE}})$.

2.2 Stacking strategy and SAE

The hierarchical architecture of the SAE can be generated by removing the decoders in the trained AEs and stacking the encoders layer by layer. The structure of the SAE constructed by M trained encoders is depicted in Fig. 2. The input data $\boldsymbol{x}_i$ in Layer 0 is fed to train the first encoder, i.e. the first hidden layer (Layer 1) in the three-layer AE. Subsequently, the output $\boldsymbol{h}_i^{(1)}$ of Layer 1 is utilised as the input to train the second hidden layer. In such a manner, each hidden layer in the SAE is separately trained with the output of the previous layer. Finally, we obtain the high-level representation of the raw input as the output of the model, which is reported to be more abstract and robust than those learnt by the shallow models.

2.3 Denoising scheme with dropout

Without any restrictions, the AE can possibly learn an identity mapping and overfitting will frequently occur. It can be alleviated by using the denoising scheme, which forces the AE to learn a deterministic encoder–decoder pair by minimising the distortion between the reconstructions and the interrupted inputs. There are numerous schemes to generate contaminated inputs, among which 'dropout' is the most popular one. This scheme randomly removes units along with all its incoming and outgoing connections in each training epoch. Neurones that are dropped out do not contribute to

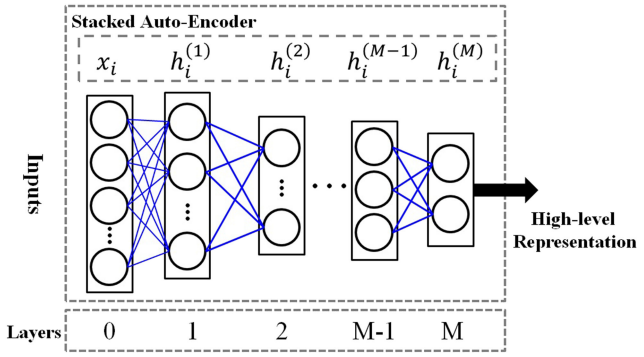

Fig. 2 *Typical M-layer SAE constructed by stacking M encoders together*

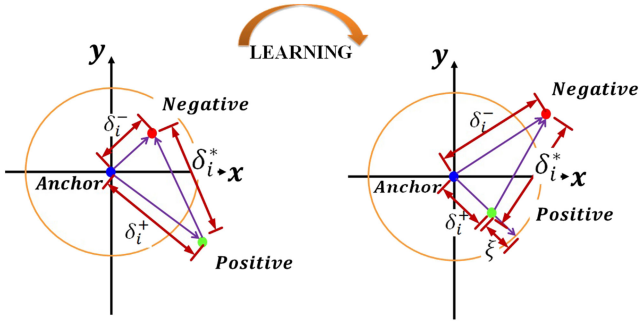

Fig. 3 *Principle of how the triplet loss helps to learn the discriminative features*

the forward and backward procedures. This technique reduces the complex co-adaptation of neurones and provides a manner of approximately combining various neural network architectures efficiently, thus overfitting is prevented. While utilising this technique, the input $\boldsymbol{x}_i = \left[x_i^1, x_i^2, \ldots, x_i^{d_v}\right]^{\mathrm{T}}$ is randomly interrupted by multiplying the masking vector $\boldsymbol{r}$ as follows:

$$\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i \boldsymbol{r} = \left[x_i^k r^k\right]_{k=1}^{d_v} \tag{4}$$

where $\tilde{\boldsymbol{x}}_i$ is the interrupted version of $\boldsymbol{x}_i$; $\boldsymbol{r} = \left[r^k\right]_{k=1}^{d_v}$ is the masking vector with $r^k \sim \text{Bernoulli}(p)$ being a Bernoulli random variable which has a probability p of being 1. The objective function of the AE with dropout (DAE) is

$$J_{\text{DAE}}(\theta_{\text{DAE}}) = \frac{1}{N}\sum_{i=1}^{N} L_{\text{R}}(\boldsymbol{x}_i, \tilde{\boldsymbol{y}}_i) \tag{5}$$

where $\tilde{\boldsymbol{y}}_i$ is the data reconstructed from $\tilde{\boldsymbol{x}}_i$.

3 TDAE model

Although deep networks have superior performance over the conventional feature extraction methods, lacking training data is a major obstacle in SAR ATR tasks. DA can partly solve this problem but it leads to additional computation complexity, difficulties in scheme selection and the possibility of amplifying the sampling bias in the training dataset. To overcome these problems, the TDAE was proposed to take the full advantage of limited training samples by introducing the supervised information. Details on the proposed model are illustrated and discussed in this section.

3.1 Modified triplet loss

Given a labelled dataset $D_{\text{train}}^{\text{triplet}} = \{\boldsymbol{x}_i, l_i\}_{i=1}^{N}$ with $\boldsymbol{x}_i$ and l_i being the ith sample and its label, respectively, the triplet loss is measured based on the triplet $T_i = (\boldsymbol{x}_i, {}^{p}\boldsymbol{x}_i, {}^{n}\boldsymbol{x}_i)$ of $\boldsymbol{x}_i$. In the triplet, $\boldsymbol{x}_i$ is the anchor; ${}^{p}x_i$ is the positive sample that has the same label as $\boldsymbol{x}_i$ and ${}^{n}x_i$ is the negative sample which belongs to a different class of $\boldsymbol{x}_i$. Feeding elements of T_i into the DAE separately, we can get the triplet of the encoded features $\tilde{h}_i^{\mathrm{T}} = \left(\tilde{\boldsymbol{h}}_i, {}^{p}\tilde{\boldsymbol{h}}_i, {}^{n}\tilde{\boldsymbol{h}}_i\right)$ and the reconstructed triplet $\tilde{y}_i^{\mathrm{T}} = (\tilde{\boldsymbol{y}}_i, {}^{p}\tilde{\boldsymbol{y}}_i, {}^{n}\tilde{\boldsymbol{y}}_i)$. For classification tasks, it is expected that $\tilde{h}_i$ is more similar to the positive encoded feature ${}^{p}\tilde{h}_i$ than the negative encoded feature ${}^{n}\tilde{h}_i$. On this account, the similarity measurement of the triplet in the margin ranking form is

$$L_{\text{Triplet}}(T_i) = \max(0,\ \delta_i^+ + \xi - \delta_i^-) \tag{6}$$

where $\|\cdot\|_2$ is the $\ell-2$ norm; $\delta_i^+ = \|\tilde{\boldsymbol{h}}_i - {}^{p}\tilde{\boldsymbol{h}}_i\|_2$ is the distance between the anchor and the positive sample in the feature space; $\delta_i^- = \|\tilde{\boldsymbol{h}}_i - {}^{n}\tilde{\boldsymbol{h}}_i\|_2$ is the distance between the anchor and the negative sample in the feature space; and ξ is the margin. As illustrated in Fig. 3, for each triplet the restriction in (6) pushes the feature learnt from the negative sample out of the circular region (i.e. the golden circle) of the anchor determined by δ_i^+ and ξ, while simultaneously pulls that learnt from the positive sample inside the golden circle, thereby providing representations with larger inter-class diversity than the intra-class diversity.

A problem of the triplet loss in (6) is that if $\delta_i^+ + \xi - \delta_i^- \le 0$ for a specific anchor, the triplet loss would have little contribution to the learning process leading to performance degradation and slower convergence. Hence, the anchor swap scheme is introduced to solve this problem and a harder restriction called the semi-hard triplet loss based on the scheme is

$$L_{\text{Triplet}}^{\text{semi-hard}}(T_i) = \max(0,\ \delta_i^+ + \xi - \min(\delta_i^-, \delta_i^*)) \tag{7}$$

where $\delta_i^* = \|{}^{p}\tilde{\boldsymbol{h}}_i - {}^{n}\tilde{\boldsymbol{h}}_i\|_2$ is the distance between the positive sample and the negative sample in the feature space as illustrated in Fig. 3. Accordingly, if $\delta_i^- > \delta_i^*$, the anchor and the positive sample are swapped, i.e. ${}^{p}\boldsymbol{x}_i$ is the anchor and $\boldsymbol{x}_i$ is the positive sample. This ensures that the harder inter-class distance is utilised in the optimisation, and it improves the performance and the convergence speed without computational overhead.

Another problem is that it does not restrict the samples belonging to the same class close to each other, which is possible to result in clusters with large intra-class divergence in the feature space. In other words, the learnt features are susceptible to the condition variations caused by object pose, heavy speckle and occlusion. To solve this problem, the intra-class distance penalty is employed to diminish largely δ_i^+, and the modified triplet loss L_{Triplet}^{m} is

$$L_{\text{Triplet}}^{m}(T_i) = \alpha L_{\text{Triplet}}^{\text{semi-hard}}(T_i) + \beta\delta_i^+ \tag{8}$$

where α and β are the weights of the semi-hard triplet loss and the intra-class distance penalty, respectively. The modified triplet restriction in (8) guarantees that the learnt representation will be discriminative features with large inter-class diversity and small intra-class diversity.

3.2 Three-branch TDAE

The proposed TDAE model restricted by the modified triplet loss in (8) is presented in Fig. 4. The TDAE consists of three branches with shared parameters, i.e. the anchor branch, the positive branch and the negative branch processing the corresponding element in the input triplet $T_i = (\boldsymbol{x}_i, {}^{p}\boldsymbol{x}_i, {}^{n}\boldsymbol{x}_i)$, respectively. Elements in T_i are processed simultaneously in similar encoding–decoding procedures to generate the reconstructed triplet $\tilde{y}_i^{\mathrm{T}} = (\tilde{\boldsymbol{y}}_i, {}^{p}\tilde{\boldsymbol{y}}_i, {}^{n}\tilde{\boldsymbol{y}}_i)$. Subsequently, the reconstruction error is computed by averaging the reconstruction errors of all the three branches using the output triplet $\tilde{y}_i^{\mathrm{T}}$. However, in most conditions, there are clutter and speckle in the target patches extracted from large scenes which not only have little information about the target but affect the performance of the learnt features. Consequently, in order to

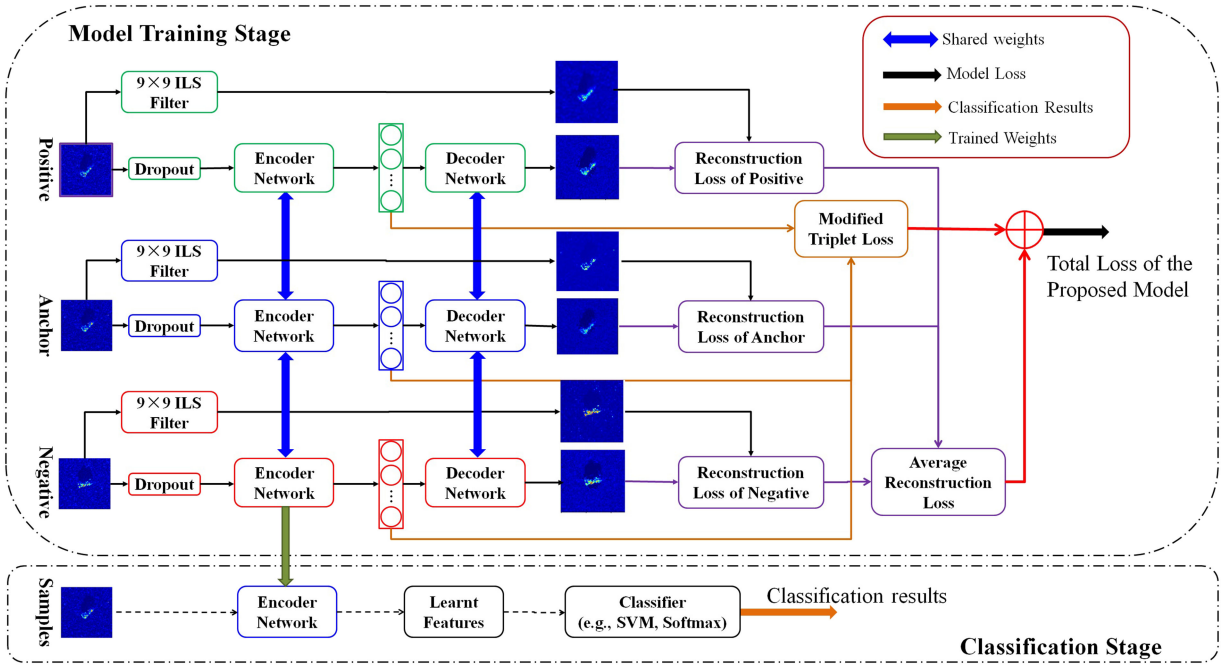

Fig. 4 *Principle of the proposed three-branch TDAE*

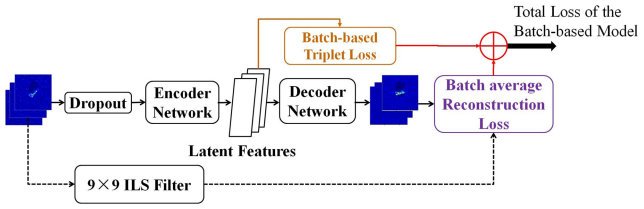

Fig. 5 *Structure of the one-branch DAE restricted by the batch-based triplet loss*

alleviate their influence, the reconstruction distortion of each branch is measured by the ED between the output and the input filtered by a 9×9 ILS filter ϕ_{ILS}. The ILS filter is selected due to its effectiveness in suppressing speckle and maintaining target details [33]. Accordingly, the modified reconstruction loss is

$$L^{\text{R}}_{\text{Triplet}}(T_i) = \frac{1}{3} \sum_{z \in Ti} \sum_{\hat{z} \in \tilde{y}_i^{\text{T}}} \| \hat{z} - \phi_{\text{ILS}}(z) \|_2 \tag{9}$$

Binding the modified reconstruction error in (9) and the modified triplet restriction in (8), the total loss of the proposed TDAE with the input triplet T_i is

$$L_{\text{Triplet-DAE}}(T_i) = L^{m}_{\text{Triplet}}(T_i) + L^{R}_{\text{Triplet}}(T_i) \tag{10}$$

When using the mini-batch approach to optimise the model, the objective function of a batch with M triplets is

$$J_{\text{Triplet-DAE}}(\theta_{\text{Triplet}}) = \frac{1}{M} \sum_{i=1}^{M} L_{\text{Triplet-DAE}}(T_i) \tag{11}$$

where θ_{Triplet} is the set of the trainable parameters in the proposed model. The proposed model can then be trained by minimising (11) utilising the gradient descent method. Once the model training is finished, the encoder network of the anchor branch with trained parameters can be utilised to extract features from the training dataset and the test dataset. The learnt features are subsequently fed to a classifier such as the SVM or the Softmax for object classification as the classification stage of Fig. 4.

A vital task of the proposed model is to construct the training set of triplets. Although traversing all the possible combinations of samples will take the full advantage of the limited training samples, it is infeasible due to a large amount of computation during the model training stage. Accordingly, to strike a balance between the computation complexity and the required triplets, a new strategy is developed. At each iteration, every training sample is used as the anchor for more than one time, which depends on the size of the training dataset. Subsequently, for each anchor, positive images and negative images are randomly selected from the rest of the samples. To ensure that we have a good representation of every class, an equal number of negative samples will be selected from every other class with an anchor from one class. By traversing all the training samples, the training set of triplets is generated. Comparisons are carried out to remove triplets with similar elements and new ones are added. Finally, shuffling the triplets in the set and splitting the dataset into batches, the model is optimised by the mini-batch gradient descent method (MBGD).

3.3 Batch-based simplified implementation of the TDAE

Although the TDAE guarantees that learnt features have a large inter-class diversity and a small intra-class diversity, it is difficult for it to be implemented utilising popular DL libraries due to the triplets constructing procedure discussed above. Besides, it also brings about high computation complexity since every sample will be processed by the encoding–decoding procedure for several times as different roles (i.e. the anchor, the positive sample and the negative sample) at each iteration. To solve the problem, the batch-based triplet loss is proposed while the MBGD is utilised to optimise the model. By applying the batch-based restriction, the encoding–decoding process is executed only once for each sample in the training dataset at every iteration. In addition, a harder triplet restriction combines the semi-hard positive batch-triplet loss and the semi-hard negative batch-triplet loss is devised in a batch-based manner. The simplified one-branch TDAE is illustrated in Fig. 5.

While using the MBGD method, the training dataset is divided into K batches and the trainable parameters are updated after processing all samples in a batch at every iteration. Considering a batch with M samples $\{\boldsymbol{x}_i\}_{i=1}^{M}$, the corresponding encoded features are $\{\tilde{\boldsymbol{h}}_i\}_{i=1}^{M}$ and the outputs of the decoder network are $\{\tilde{\boldsymbol{y}}_i\}_{i=1}^{M}$. The reconstruction loss of the batch-triplet restricted DAE is

$$L^{\text{R}}_{\text{restricted}}(\boldsymbol{x}_i) = \frac{1}{M} \sum_{i=1}^{M} \| \tilde{y}_i - \phi_{\text{ILS}}(\boldsymbol{x}_i) \|_2 \tag{12}$$

Table 1 Details of the targets used in the experiment

Type	Sl. no.	Number of samples	
		17° Depr.	15° Depr.
BMP-2	9563	233	195
	9566	232	196
	c21	233	196
BTR-70	c71	233	196
T-72	132	232	196
	812	231	195
	s7	233	191
2S1	b01	299	274
T-62	A51	299	273
BRDM-2	E-71	298	274
BTR-60	K10yt7532	256	195
D-7	92v13015	299	274
ZIL-131	E12	299	274
ZSU-234	d08	299	274

Depr. denotes the depression angle.

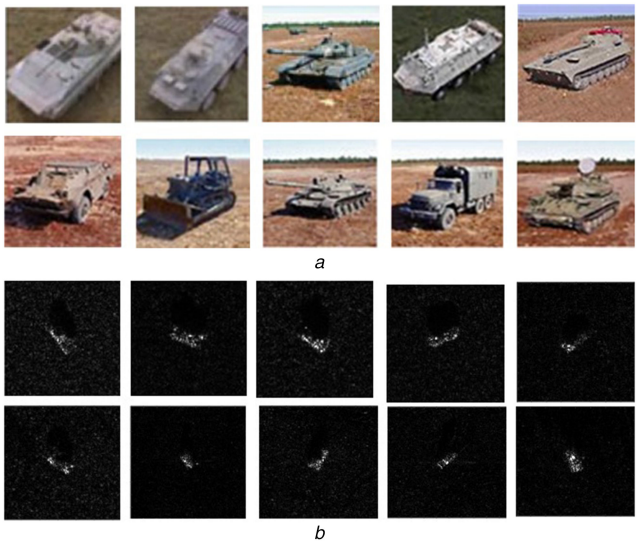

Fig. 6 *Photographs and SAR imagery examples of the MSTAR dataset for model evaluation*

(a) photographs of the ten targets in the MSTAR dataset which are BMP-2, BTR-70, T-72, BTR-60, 2S1, BRDM2, D7, T62, ZIL131 and ZSU-234 from the top left to the right bottom, ***(b)*** SAR images of the ten vehicles with the similar layout as (a)

The batch-based triplet restriction can be computed by using the encoded features $\{\tilde{\boldsymbol{h}}_i\}_{i=1}^{M}$ and their labels $l = \{l_i\}_{i=1}^{M}$. Instead of traversing all the triplet combinations in the batch, the triplet loss can be calculated by applying a harder criterion that utilises both the maximum intra-class distance and the minimum inter-class distance. The pairwise distance matrix $\boldsymbol{A} = [\delta_{ij}]_{i,j=1}^{M}$ is calculated at first, with δ_{ij} being the ED between the encoded feature $\tilde{\boldsymbol{h}}_i$ and $\tilde{\boldsymbol{h}}_j$. Subsequently, the mask matrixes of similar class pairs $\boldsymbol{\Lambda}_{\text{sim}} = \{\sigma_{ij}\}_{i,j=1}^{M}$ and the mask matrixes of different-class pairs $\boldsymbol{\Lambda}_{\text{diff}} = \{\rho_{ij}\}_{i,j=1}^{M}$ are generated according to the labels, respectively. In $\boldsymbol{\Lambda}_{\text{sim}}$, $\sigma_{ij} = 1$ if and only if $\tilde{\boldsymbol{h}}_i$ and $\tilde{\boldsymbol{h}}_j$ belong to the same class with $i \neq j$. In $\boldsymbol{\Lambda}_{\text{diff}}$, the different-class pair $\rho_{ij} = 1$ if $\tilde{\boldsymbol{h}}_i$ and $\tilde{\boldsymbol{h}}_j$ belong to different classes. Thereby the batch-based triplet loss is

$$L_{\text{batch-triplet}}(\boldsymbol{x}_i) = \alpha \max(L_{\text{sim}},\ L_{\text{diff}}) + \beta E(\sigma_{ij}\delta_{ij}) \tag{13}$$

where $E(\cdot)$ is the average operation; L_{sim} is the semi-hard positive batch-triplet loss based on the minimum inter-class distance; and L_{diff} is the semi-hard negative batch-triplet loss based on the maximum intra-class distance

$$L_{\text{sim}}(\boldsymbol{x}_i) = E\left[\max\left(\sigma_{ij}\delta_{ij} + \xi - \min\left(\rho_{ij}\delta_{ij}\right),\ 0.0\right)\right] \tag{14}$$

$$L_{\text{diff}}(\boldsymbol{x}_i) = E\left[\max\left(\max\left(\sigma_{ij}\delta_{ij}\right) + \xi - \rho_{ij}\delta_{ij},\ 0.0\right)\right] \tag{15}$$

Combining (12) and (13), the objective function of the proposed one-branch TDAE is

$$L_{\text{restricted}} = L_{\text{restricted}}^{\text{R}}(\boldsymbol{x}_i) + L_{\text{batch-triplet}}(\boldsymbol{x}_i) \tag{16}$$

Finally, the model can be optimised by minimising (16) with the MBGD method.

4 Experimental results

4.1 Dataset description and experiment setup

In this paper, the MSTAR dataset [3] is utilised to evaluate the performance of the proposed model. There are ten distinct types of ground vehicles in the dataset including the armoured personnel carrier BMP-2, BRDM-2, BTR-60 and BTR-70; the tanks T-62 and T-72; the rocket launcher 2S1; the air defence unit ZSU-234; the truck ZIL-131; and the bulldozer D7. In this dataset, patches centred on the target and surrounded by background clutter provide full aspect coverage from 0° to 360° and different views at various depression angles. The details of the ten targets used in the experiments are listed in Table 1 including their type, serial number, number of samples and their photographs and SAR image examples are depicted in Fig. 6.

A two-layer TDAE implemented as the one-branch structure is utilised in the evaluation experiments. There are 1000 units in the first hidden layer and 400 units in the second hidden layer which are similar to the experiment of the ED-AE [32]. The trainable parameters of the model are initialised by the Xavier scheme [34] while the hyper-parameters are set as follows: the triplet margin $\xi = 0.02$, the triplet loss weight $\alpha = 0.9$, the intra-class penalty weight $\beta = 0.2$ and the dropout fraction $\rho = 0.25$. In this paper, both the linear SVM (LSVM) and non-linear SVM (NSVM) are adopted as the classifier to evaluate the performance of the proposed model. The first one is utilised to demonstrate the linear separability of the learnt features while the NSVM is employed to evaluate the best classification performance. To avoid the fluctuations in the results caused by the random steps in model initialisation and optimisation, each experiment is repeated ten times and the average results are utilised for performance evaluation. Before evaluating the proposed model with the dataset, the normalisation is adopted to alleviate the amplitude variation in target patches, which possibly conceals the differences between targets and thus affect the performance of the learnt features. Except for the normalisation, no other pre-processes such as augmentation or target segmentation are applied.

To comprehensively assess the performance, the proposed TDAE is tested under standard operating conditions (SOCs) and extended operation conditions (EOCs). The SOC refers to that the target configurations in the test set are the same as those in the training set but with different aspects and depression angles. In the EOC scenario, there are large differences between the training and test sets including substantial variations in the signal-to-noise ratio (SNR), resolution and version variants. In our experiments, the proposed method is first tested on three similar targets, namely BMP-2, BTR-70 and T-72, to validate its performance under SOC and version variants. Subsequently, sensitivity analysis of the hyper-parameters is discussed based on the three-target dataset. The robustness of the proposed model under various conditions including noise corruption and resolution variance is also evaluated with the three-target dataset. Finally, experiments are conducted on ten-class MSTAR data to evaluate the performance under the extension of the target type.

4.2 Evaluation on three-target classification

In this experiment, similar to the experimental setting in Deng *et al.*'s paper [32], only the patches of BMP2-9563, BTR70-c71 and T72-132 at depression angle 17° are used to train the TDAE. All images of BMP2-c21, BMP2-9566, BMP2-9563, BTR70-c71, T72-132, T72-s7 and T72-812 at depression angle 15° are used as the test data to evaluate its performance with and without version variants.

4.2.1 Discrimination capability analysis: The t-Distributed Stochastic Neighbour Embedding (t-SNE) [35] method, which maps the high-dimensional data to a low-dimensional space according to their structure, is employed to illustrate the discrimination capability of the features. As shown in Fig. 7, the original data, the features learnt by the DAE and the TDAE are projected into a two-dimensional (2D) space. The proposed method significantly improves the classification capability of the learnt features in comparing with the original DAE due to the modified triplet restriction which pushes the samples belonging to different classes away from each other and pulls the samples of the same type together. Accordingly, the batch-based modified triplet restriction guarantees large diversity in different classes and a small diversity in the same class as demonstrated in Fig. 7.

To quantitatively analyse its performance, the ratio of inter-class distance to within-class distance (BWR) [32] is adopted. The BWR is defined as

$$\mathrm{BWR} = \frac{D_{\text{inter-class}}}{D_{\text{intra-class}}} = \frac{\sum_{i=1}^{c} \frac{1}{N} \sum_{j=1}^{N_i} \left(x^{ij} - \mu^i\right)^{\mathrm{T}} \left(x^{ij} - \mu^i\right)}{\sum_{i=1}^{c} \left(\mu^i - \mu\right)^{\mathrm{T}} \left(\mu^i - \mu\right)} \quad (17)$$

where $D_{\text{intra-class}}$ and $D_{\text{inter-class}}$ are the average intra-class distance and the average inter-class distance in the feature space, respectively; c is the number of classes; N_i is the number of samples in the ith class; x^{ij} is the jth sample of the ith class; μ^i is the average feature vector of the ith class; and μ is the average feature vector of all classes.

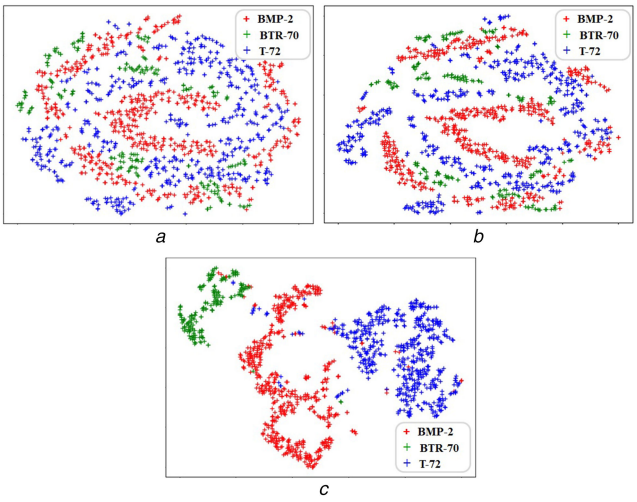

Fig. 7 *Visualisation of features learnt from the three-target dataset*
(a) projection of the original data, ***(b), (c)*** projections of the features learnt by the DAE and the proposed method, respectively

Table 2 BWR of the original images and the learnt features

Models	Original image	DAE	ED-AE	Proposed
BWR	1.03	1.10	2.50	3.02

The BWR indicates the linear classification capability of the learnt features which is proportional to the inter-class distance and inversely proportional to the intra-class distance. The BWR values of the original images and the features extracted from the DAE, the ED-AE [32] and our proposed method are listed in Table 2. According to Table 2, the BWR values of the original images and the features extracted by the DAE have little difference indicating that the DAE has limited contributions to the discrimination capability of learnt features. Although the ED-AE gains a significant improvement on the BWR value, the proposed model has the highest BWR value. It means that the TDAE learns features with a smaller intra-class distance and a larger inter-class distance than the ED-AE due to the modified triplet restriction.

4.2.2 Classification results: The average classification results of the ten executions are presented in Table 3. In this paper, we measure the performance through the probability of correct classification (P_{CC}), which is calculated through the number of targets recognised correctly divided by the number of all the targets. The results only with and without version variants are listed in the fifth and sixth columns of the table. In addition, results of the reference methods used for comparison in [32] are also listed including directly using the original images, the classic AE and the ED-AE with the LSVM classifier.

As shown in Table 3, for the experiment without version variants (i.e. under SOC), the proposed model has the highest accuracy in comparison with the reference methods. Specifically, the result obtained by the TDAE with NSVM is quite close to the state-of-the-art results which are reported in the open-published literature. Meanwhile, in the case with variants only, the proposed method is considerably improved in comparison with the classic AE and the ED-AE algorithms (11.72 and 1.88%, respectively), indicating a better generalisation performance. The major reason is that the proposed triplet restriction guarantees that the proposed model can learn features with large inter-class diversity and small intra-class distances. Besides, the dropout scheme, which prevents overfitting caused by limited samples and the hierarchical structure of the DL model, can also help the model learn robust information of the targets in the version variants condition. The average accuracies of these methods with all the test data are listed in the seventh column of Table 3. It is found that even with LSVM, the proposed model outperforms those reference methods. Especially, when the NSVM classifier is applied to mapping the extracted features into a high-dimension classification space, the average P_{cc} of 98.67% is comparable with the state-of-the-art results obtained by the complicated CNNs, with less computation complexity and no DA.

Other features extraction methods are also compared with the proposed method for further evaluation including the conventional ones and the deep networks. The conventional methods for comparison include the principal component analysis-kernel SVM (PCA-KSVM) [4], the shadow-contour (SC)-based method [5], the joint sparse representation-based method (JSRC) [7], the particle swarm optimisation with Hausdorff distance (PSO-HD) [6], the non-negative matrix factorisation (NMF) method [3] and the decision fusion-based multi-scale scattering centre matching (SCM) [9]. The deep networks utilised for comparison comprises of the CNN with DA (DA-CNN) [29], the CNN with SVM (CNN

Table 3 Classification results on the three-target dataset

Method	BMP2, %	BTR70, %	T72, %	Without variants, %	Variants only, %	Average P_{cc}, %
original image + LSVM	80.75	98.47	93.30	94.68	84.14	88.64
AE + LSVM	87.56	94.39	83.51	93.14	82.10	86.81
ED-AE + LSVM	94.21	93.88	94.16	97.08	91.94	94.14
TDAE + LSVM	92.37	98.92	96.90	98.64	93.82	95.32
TDAE + NSVM	98.30	99.54	98.74	99.66	97.93	98.67

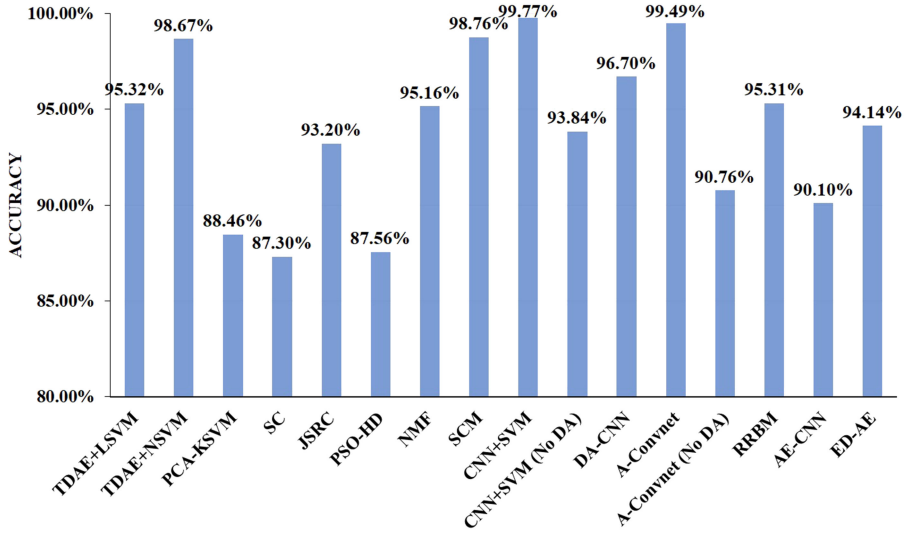

Fig. 8 *Performance comparison with different methods*

Table 4 Evaluation experiments of processing time

Method	Average processing time, s	
	Pre-process	Model training
TDAE	0.21	274.17
A-Convnet	0.30	635.63
CNN + SVM	6.21	391.79

+ SVM) [30, 31], the A-Convnet [28], the restricted RBM (RRBM) [12], the sparse AE pre-trained CNN (AE-CNN) [13] and the ED-AE [32]. Among these methods, the A-Convnet and the CNN + SVM are implemented by our own code because they are not evaluated with the three-target dataset in the open-published literature. Besides, since the DA schemes are applied to provide sufficient training samples in both the CNN + SVM and the A-Convnet, their results with and without DA are all evaluated. Moreover, in the CNN + SVM method, the input data should be rotated according to the estimated azimuth angles. However, in the evaluation experiment, the actual values of the aspect angle are directly used ignoring the errors caused by angle estimation. Consequently, the actual P_{cc} of the CNN + SVM model would be lower than that presented in the experiment as reported by Wagner [31].

The accuracies of all the methods are shown in Fig. 8. With the linear classifier, the features learnt by the proposed method have a better classification capability than most conventional methods (e.g. the JSRC, the SC-based method, the PSO-HD method, the NMF method and the PCA-KSVM) because of the modified triplet restriction and the dropout scheme. Comparison to the deep networks including the DA-CNN, the RRBM, the AE-CNN and the ED-AE also indicates that the proposed model outperforms most of the DL models which have specialised restrictions for finding discriminative features. Even compared with the SCM, the CNN + SVM and the A-Convnet that have achieved the state-of-the-art results, the proposed method with NSVM obtains a comparable result – a bit lower average accuracy, no requirement of DA or model assumption and much less computation complexity.

Among these methods, the SCM confronts the difficulties of establishing highly vivid 3D computer-aided design models of targets and devising electromagnetic (EM) code for simulating accurate backscattering data. Even now they are still challenging tasks and sometimes infeasible in some remote sensing applications. Besides, the SCM method has the highest computation complexity due to the EM code-based SCM construction step and the complicated parameter estimation algorithm which iteratively estimates the parameters of scattering centres of a target. All these prevent it from being applied in many practical earth observation tasks. For the A-Convnet and the CNN + SVM, their outstanding performance mainly relies on the DA operations as demonstrated by the accuracies achieved without DA step in Fig. 8. Although their DA processes do improve the performance, they will induce certain problems including amplifying the sampling biases and high computation complexity.

Further evaluation on the processing time of the proposed method, the CNN + SVM and the A-Convnet are conducted. The experiment is conducted on the PC platform with an Intel i7-7200QM, 16 GB random access memory and an NVIDIA GTX960M (fourth-generation memory) graphics processing unit (GPU) supported by the 64 bit Windows 10. All the models are developed in Python v3.5 supported by the Google TensorFlow v1.4.0 and CUDA v8.0. The MBGD optimisation algorithm is adopted for model training and the batch size is 64 for both the A-Convnet and CNN + SVM. The maximum training epoch is 500 and the early-stopping scheme is enabled to terminate the training if the improvement of the training loss is less than the threshold. The average times of pre-processing and model training are presented in Table 4. Obviously, the proposed method consumes much less time than the rest two DL networks in both the pre-processing and the training steps. The major reason is that the proposed method only requires limited pre-processes and training samples. Besides, the fully connected layers of the proposed model also have less computation complexity than the convolutional layers of the two methods.

4.2.3 Analysis of training set size: Furthermore, the experiment, which evaluates the model performance with limited training samples, is conducted by randomly removing a part of samples in the training set. In this experiment, only $1/n$ images are randomly selected from the dataset as training samples with n varying from one to ten. The average accuracies and their standard deviations with both the LSVM and the NSVM are presented in Fig. 9. Besides, the results of the CNN + SVM and the A-Convnet are also depicted in Fig. 9 for comparison. As shown in this figure, the accuracies of all the models reduce when the number of training samples decreases. However, the accuracies of the CNN + SVM and the A-Convnet decrease more rapidly than the proposed model. The P_{cc} of the CNN + SVM falls below 90% when only 1/4 samples in the dataset are used as training samples, while for the A-Convnet critical value of n is 1/5. For the proposed model, even when only 1/6 samples (i.e. about 116 samples) are utilised to train the network which is considered as an extreme case for most machine learning tasks, the two variants of the proposed model still provides accuracy higher than 90%. It demonstrates that the proposed model has a robust representation learning capability for limited training data to some extent.

4.3 Sensitivity analysis of hyper-parameters

In the proposed model, there are four additional hyper-parameters in comparison with the classic AE: the triplet margin ξ, the weight of triplet loss α, the weight of the intra-class distance penalty β and

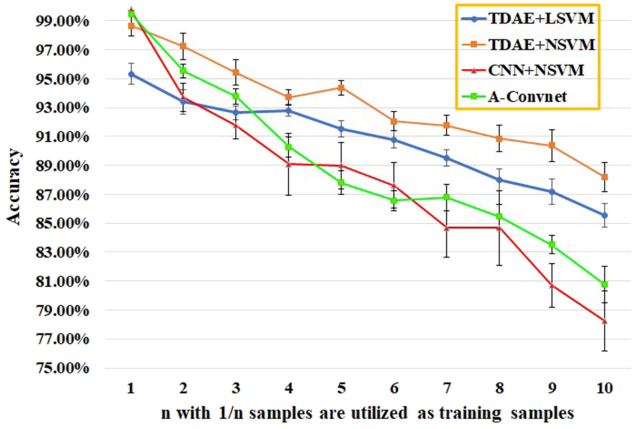

Fig. 9 *Comparison of classification results when 1/n samples are randomly selected to train the model*

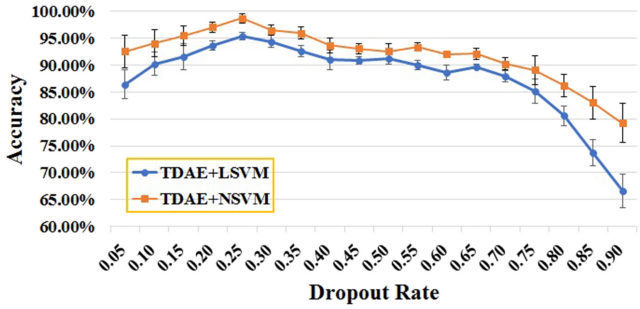

Fig. 10 *Classification accuracy affected by ρ*

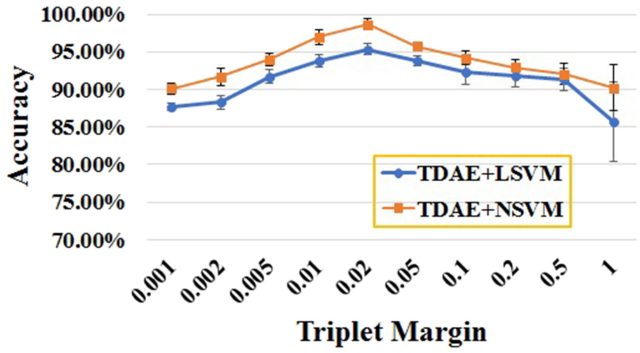

Fig. 11 *Classification accuracy due to different ξ*

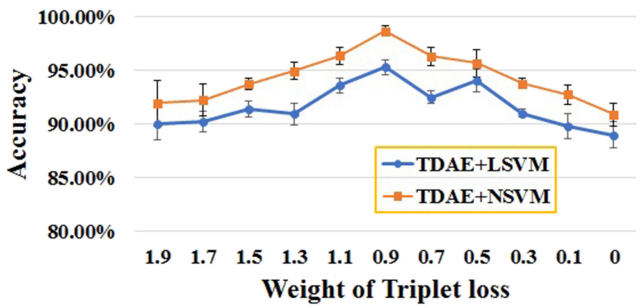

Fig. 12 *Classification results with varying α*

the dropout rate ρ. Experiments are conducted to evaluate the manner, in which they affect accuracy. Both the average accuracies and the corresponding standard deviations are used for evaluation.

The dropout rate ρ represents the probability of each unit being dropped in the network, which varies from 0 to 1. Improper dropout rate, neither too small nor too large, will lead to performance degradation because the former one cannot prevent overfitting while the latter one will remove too many units resulting in loss of useful information. The manner, in which the dropout rate affects the classification is tested and presented in Fig. 10 by varying the dropout rate from 0.05 to 0.9 while keeping the rest of the parameters unchanged. According to the influence curves of the two variants, the optimal dropout rate is 0.25. According to Fig. 10, the variation of the dropout rate in a small neighbourhood of the optimal value does not affect the accuracy much. However, for the LSVM, setting the dropout rate neither smaller than 0.1 nor larger than 0.5 will result in an accuracy below 90% with fierce fluctuations due to overfitting and loss of information. For the NSVM, the dropout rate will fall below 90% with fierce fluctuation when the dropout rate is larger than 0.7.

The triplet margin ξ is a slack parameter determining the minimum value that the inter-class distance is larger than the intra-class distance. The maximum value of the triplet margin is usually smaller than 1.0 for normalised inputs. Accordingly, the binary search method is employed to analyse the variation of the classification accuracy with different triplet margins $\xi \in [0.001, 1]$, which is shown in Fig. 11. As presented in this figure, the optimal value of the triplet margin is 0.02. Neither too large margin nor too small margin will result in a much lower classification accuracy with either LSVM or NSVM. That is because a large margin will result in fierce fluctuations in triplet loss which affect the convergence of the model while too small margin provides a weak restriction to the learnt features and thus deteriorates the performance.

In the objective function of the TDAE, there are three parts of losses, i.e. the modified reconstruction loss, the batch-triplet loss and the intra-class distance penalty. The reconstruction loss guarantees that the important information on the targets will be preserved in the learnt features. The triplet loss guarantees that the learnt features will have low inter-class coupling while the intra-class distance compels the model to learn features with higher intra-class cohesion, meaning a small divergence of samples from the same target in the feature space. The weight of the triplet loss α and the weight of the intra-class distance penalty β are parameters which balance the learning object among the three parts. Both of them have specific influences on the classification results. Herein, we will give a detailed discussion on the value of the two parameters.

Fig. 12 shows the varying classification results with different weights α. As shown in Fig. 12, the optimal value of α is 0.9. When the weight increases or decreases, the classification rate reduces. If the weight is larger than 1.3, the classification rates of both TDAE + LSVM and TDAE + SVM reduce slowly, along with slight fluctuation, indicating that the triplet loss becomes the major restriction of the model which is optimised by the backpropagation process. When the weight is smaller than 0.3, the accuracy of the TDAE + LSVM quickly reduces below 90% because the triplet loss only has a weak restriction on the objective function. However, the results of the TDAE + NSVM are still larger than 90% due to its non-linear projection capability. Specifically, when $\alpha = 0$ which means that the triplet loss will no longer restrict the objective function, the classification result of the TDAE + LSVM is 89.00%, which is quite close to the 87.56% obtained by the AE [32]. This phenomenon demonstrates that the batch-triplet loss is the major restriction that guarantees the model's capability of learning discriminative features. The 1.44% improvement in the average classification rate indicates that both the modified reconstruction loss and the intra-class distance penalty also provide some restrictions to the model.

Fig. 13 presents the relationship between the varied weight β and the accuracies obtained by LSVM and NSVM. According to this figure, the optimal value of β is 0.2. When β is larger than the optimal value, the classification rate reduces indicating that the balance between the reconstruction loss, the triplet loss and the intra-class distance penalty is broken. However, since in the proposed model the major restriction term is the batch-triplet loss, the reduction is quite slow as illustrated by the curves. When β is smaller than the optimal value, it will have fewer contributions to the objective function and the classification performance has a slight degradation. Specifically, when $\beta = 0$ which means that the intra-class distance penalty will be removed from the objective function, the classification rates obtained by LSVM and NSVM reduce to 92 and 94.9%, respectively, which are more than 3%

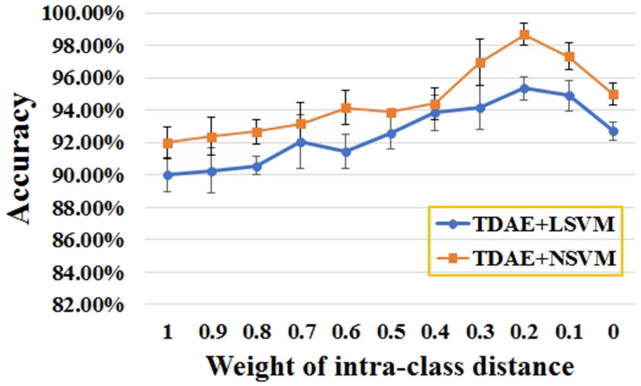

Fig. 13 *Classification results with varying β*

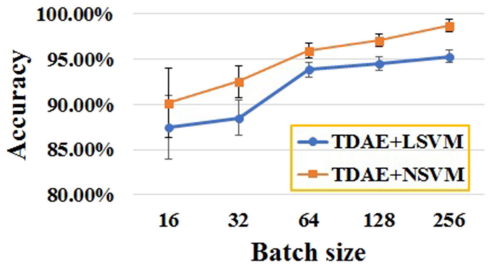

Fig. 14 *Classification results with different batch sizes*

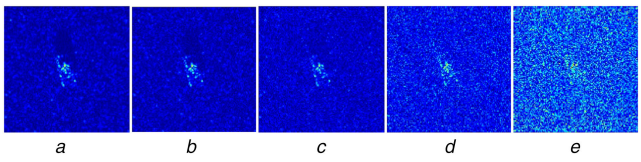

Fig. 15 *MSTAR data corrupted by different levels of noises*
(a) 10 dB, ***(b)*** 5 dB, ***(c)*** 0 dB, ***(d)*** −5 dB, ***(e)*** −10 dB

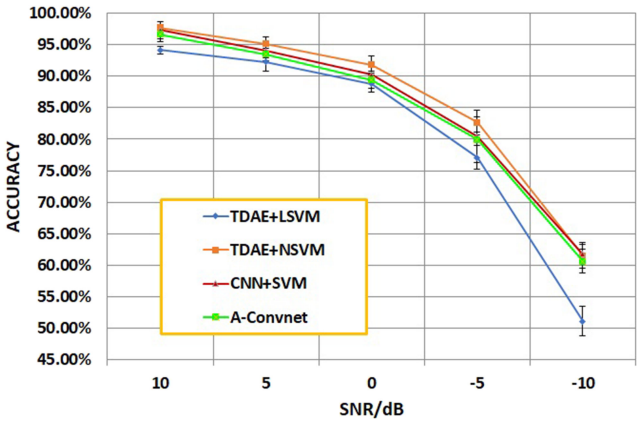

Fig. 16 *Classification results at different SNR levels*

lower than the best result. It demonstrated the effectiveness of the intra-class distance penalty for learning discriminative features.

Moreover, in the proposed batch-based model, the batch size is also an important parameter that can significantly affect the classification results since the batch-based triplet loss is computed based on all features in a batch. Insufficient samples in the batch can result in overfitting while a large batch will result in a slow convergence speed, a huge amount of computation and the requirement of large memory resources. Typically, the batch size can usually be 32, 64, 128 or 256. The classification results obtained by the two variants with different batch sizes are presented in Fig. 14. When the batch size is smaller than 64, the average accuracy obtained by the LSVM is smaller than 90% with a large standard deviation caused by randomly shuffling during the training process. Similar fierce fluctuations caused by randomly shuffling can also be observed from the results obtained by the NSVM though the average result is higher than that obtained by the LSVM. When the batch size is larger than 64, the average accuracy obtained by the LSVM increases to 94% with a small fluctuation, which is close to the best result of 95.32%. The optimal batch size in this experiment is 256 and no larger batch is evaluated because the larger batch will lead to the 'out of memory' error due to the limited memory of the GPU used in the experiment.

4.4 Classification under noise corruption

The MSTAR images have an SNR over 30 dB which is too ideal for classification. However, in SAR ATR tasks, serious noise is a major factor causing performance deterioration. Therefore, SAR images corrupted by different levels of SNR are simulated to evaluate the model's robustness to noise. The noise-corrupted images are generated by adding Gaussian noise to the frequency domain of the MSTAR images [8]. The original MSTAR images are considered noise free and different levels of noises are added according to the SNR defined as

$$\mathrm{SNR(dB)} = 10\log_{10}\frac{\sum_{h=1}^{H}\sum_{w=1}^{W}|f(h,\,w)|^2}{\mathrm{HW}\sigma^2} \tag{18}$$

where $f(h, w)$ is the complex frequency at (h, w); W and H are the numbers of bins in the range and azimuth frequency, respectively; and σ^2 is the variance of the noise.

The noise is added to the frequency data and the noise-corrupted images are obtained by transforming the frequency data back to the spatial domain. Fig. 15 shows some images with different SNR levels. The results of the two variants as well as the CNN + SVM and the A-Convnet at SNR varying from 10 to −10 dB are evaluated and plotted in Fig. 16. In the experiment, the CNN + SVM directly utilises the actual azimuth angle for aspect angle rotating ignoring the estimation error caused by the noised images. Accordingly, the actual accuracy of the CNN + SVM should be lower than that presented in Fig. 16.

With the deterioration of the SNR, the P_{cc} obtained by each model experiences a decrease with increasing standard deviations and the TDAE + NSVM achieves the highest accuracy at every SNR level. When the SNR is higher than 0 dB, in which condition the shapes and characteristics of the targets are not seriously affected by the noise, the classification rates of all the models are close to 90% and the TDAE + NSVM achieves the highest accuracy of 91.8%. Even when the SNR is −5 dB that the targets can only be partly observed in the images as shown in Fig. 15*d*, the classification rate of the TDAE + NSVM is 82.6% that is higher than other reference models, which demonstrated that the proposed method is robust to noise interruption in comparison with the reference models.

4.5 Classification under resolution variance

Ideally, the resolution of SAR imagery is only determined by the bandwidth of transmitting wave and the synthetic aperture angle. However, due to the instability of the sensors, the resolution of the measured SAR images is possible to be at variance with the theoretical values. Moreover, it is infeasible to train models corresponding to every possible resolution. Consequently, the robustness of resolution variation is also an important factor for model evaluation.

In this section, the performance of the two variants of the proposed model, the CNN + SVM and the A-Convnet are evaluated with SAR images whose resolution deteriorated from 0.3m × 0.3 m to 0.7m × 0.7 m. The SAR images with varied resolutions are simulated by extracting the low-frequency sub-band of the original images. To generate images with the same size as the original data, the sub-band data are resampled by zero padding in the frequency domain and transformed back to the spatial domain. Fig. 17 presents some images at a different resolution. The experimental results are plotted in Fig. 18. As shown in this figure, all the DL models are not seriously affected by the resolution deterioration. Even when the resolution is 0.6 × 0.6 m^2 (i.e. twice lower than the original images), their accuracies are still higher than 90%.

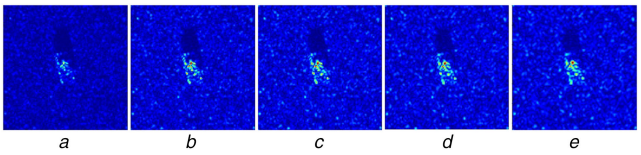

Fig. 17 *MSTAR data at different resolutions*
(a) 0.3 m × 0.3 m, ***(b)*** 0.4 m × 0.4 m, ***(c)*** 0.5 m × 0.5 m, ***(d)*** 0.6 m × 0.6 m, ***(e)*** 0.7 m × 0.7 m

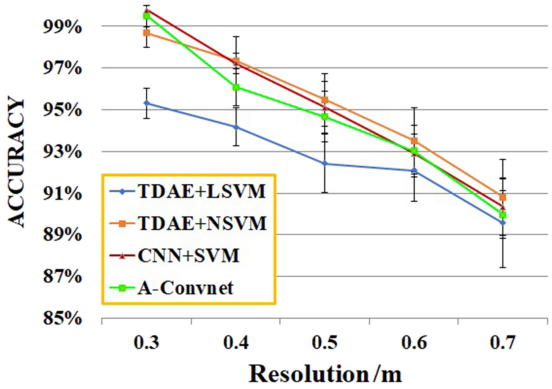

Fig. 18 *Classification results at different resolutions*

Table 5 Classification results on the ten-target dataset

Methods	2S1, %	BMP-2, %	BRDM-2, %	BTR-60, %	BTR-70, %	D7, %	T-62, %	T-72, %	ZIL-131, %	ZSU-234, %	Average, %
TDAE + LSVM	91.97	94.62	92.38	94.36	95.80	96.27	93.16	99.22	95.50	97.81	95.46
TDAE + NSVM	96.80	98.41	97.53	98.29	98.64	98.78	97.52	99.77	98.34	99.23	98.47
AE + LSVM	91.61	82.45	95.62	87.18	90.82	97.08	94.51	67.35	92.34	98.91	87.04
NMF	100	91.01	97.91	94.87	97.87	98.32	96.38	94.70	97.27	95.54	94.20
ED-AE	93.80	87.90	96.72	91.79	92.86	98.91	94.14	79.55	94.53	99.64	91.29
A-Convnet	98.54	98.81	98.91	96.92	99.49	98.91	100.00	99.66	99.27	98.91	99.03
CNN + SVM	97.81	99.83	100.00	98.97	99.49	99.64	99.27	99.83	98.91	99.27	99.41

However, among these DL networks, the proposed model with the NSVM gains the highest accuracy in comparison with other reference models when the resolution is worst than 0.4×0.4 m^2 demonstrating its excellent robustness in the case of resolution variance.

4.6 Evaluation on ten-target classification

Further evaluation of the proposed method is conducted on the ten-target dataset which consists of all the ten types of targets listed in Table 1. Similar to the experiment with the three-target dataset, images acquired at 17° depression angle are utilised as training samples while all the samples obtained at 15° depression angle construct the test set. Besides, only the data of BMP2-9563, BTR70-c71 and T72-132 are used as the samples of the BMP-2, BTR-70 and T-72 to construct the training dataset. However, in the test dataset, images of all serial numbers (i.e. version variants) are used to test the performance of the proposed method. In addition, to avoid the fluctuations caused by the random steps in model initialisation and optimisation, the experiment is repeated for ten times and the average accuracy is utilised for performance evaluation.

Besides, the proposed method is also compared with several reference methods including the original AE with LSVM (AE + LSVM), the NMF method [3], the ED-AE [32], the A-Convnet [28] and the CNN + SVM [30, 31]. Among these methods, the A-Convnet and CNN + SVM methods are implemented by our own code. The average accuracy of the proposed methods and the results of the reference methods are listed in Table 5. According to the results, the classification accuracy of the proposed method is higher than most of the reference methods including the AE, the NMF and the ED-AE. When the NSVM is utilised as the classifier, the P_{cc} of the proposed method is even comparable with the state-of-the-art results. Although both the A-Convnet and the CNN + SVM have higher accuracy than the proposed method, they all require complex DA process and have a higher computation complexity.

5 Conclusions

In this paper, a new DAE restricted by the modified triplet loss is proposed for SAR ATR to take the full advantage of limited training samples. The major contributions of this paper include:

(i) a three-branch DAE restricted by the modified triplet loss which combined the semi-hard triplet loss with the intra-class distance penalty;
(ii) a simplified version of the proposed three-branch model by devising a batch-based triplet restriction which combines the semi-hard positive triplet loss, the semi-hard negative triplet loss and the intra-class distance penalty; and
(iii) a modified reconstruction loss, in which the original inputs are replaced by the ILS filtered data to suppress clutter and speckle in the background.

The MSTAR dataset is utilised to evaluate the performance of the proposed model. The proposed method is evaluated under both SOC and several EOCs with the three-target data including noise corruption, resolution variants and version variants. Further evaluation experiment is also conducted with the ten-target data (with version variants) in the MSTAR dataset. Besides, the significant parameters of the proposed model are also discussed. Feature visualisation and evaluation experiments both demonstrated that the proposed method outperforms most conventional and DL algorithms and achieved comparable accuracy with the state-of-the-art results without DA and much additional computation.

6 Acknowledgments

This work was supported in part by the National Natural Science Foundations of China under Grants nos. 41501356 and 41331176 and the Natural Science Foundation of Jiangsu Province under Grants no. BK20150774. The authors thank the US Air Force Research Lab for providing the public MSTAR data at https://www.sdms.afrl.af.mil/index.php?collection=mstar. In addition, we are grateful to the anonymous referees for their instructive comments.

7 References

[1] Li, Y., Zhou, C., Wang, N.: 'A survey on feature extraction of SAR images'. Proc. Int. Conf. Computer Application and System Modelling, Taiyuan, China, October 2010, pp. V1-312–V1-317

[2] Zhai, Y., Li, J., Guo, C., *et al.*: 'SAR automatic target recognition based on local phase quantization plus biomimetic pattern recognition'. Proc. Int. Conf. Signal Processing, Beijing, China, October 2012, pp. 1885–1889

[3] Cui, Z., Cao, Z., Yang, J., *et al.*: 'Target recognition in synthetic aperture radar images via non-negative matrix factorisation', *IET Radar Sonar Navig.*, 2015, **9**, (9), pp. 1376–1385

[4] Mishra, A.K., Motaung, T.: 'Application of linear and nonlinear PCA to SAR ATR'. Proc. Int. Conf. Radioelektronika, Pardubice, Czech Republic, April 2015, pp. 349–354

[5] Yin, K., Lin, J., Zhang, C., *et al.*: 'A method for automatic target recognition using shadow contour of SAR image', *IETE Technical Review*, 2013, **30**, (4), pp. 313–323

[6] Dungan, K.E.: 'Feature-based vehicle classification in wide-angle synthetic aperture radar'. PhD thesis, The Ohio State University, 2010

[7] Zhang, H., Nasrabadi, N., Zhang, Y., *et al.*: 'Multi-view automatic target recognition using joint sparse representation', *IEEE Trans. Aerosp. Electron. Syst.*, 2012, **48**, (3), pp. 2481–2497

[8] Ding, B., Wen, G., Huang, X., *et al.*: 'Target recognition in synthetic aperture radar images via matching of attributed scattering centres', *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 2017, **10**, (7), pp. 3334–3347

[9] Ding, B., Wen, G.: 'Target reconstruction based on 3-d scattering centre model for robust SAR ATR', *IEEE Trans. Geosci. Remote Sens.*, 2018, **56**, (7), pp. 3772–3785

[10] Xie, H., Wang, S., Liu, K., *et al.*: 'Multilayer feature learning for polarimetric synthetic radar data classification'. Proc. Int. Geoscience Remote Sensing Symp., Quebec City, QC, Canada, July 2014, pp. 2818–2821

[11] Ni, J., Xu, Y.: 'SAR automatic target recognition based on a visual cortical system'. Proc. Int. Congress on Image and Signal Processing, Hangzhou, China, December 2013, pp. 778–782

[12] Cui, Z., Cao, Z., Yang, J., *et al.*: 'Hierarchical recognition system for target recognition from sparse representations', *Math. Probl. Eng.*, 2015, **2015**, (527095), pp. 1–6

[13] Chen, S., Wang, H.: 'SAR target recognition based on deep learning'. Proc. Int. Conf. Data Science and Advanced Analytics, Shanghai, China, 30 October–1 November 2015, pp. 541–547

[14] Morgan, D.: 'Deep convolutional neural networks for ATR from SAR imagery'. Proc. SPIE Volume 9475, Algorithms for Synthetic Aperture Radar Imagery XXII, Baltimore, MD, USA, May 2015, pp. 94750F-1–94750F-13

[15] Kreucher, C.: 'Modern approaches in deep learning for SAR ATR'. Proc. SPIE. 9843, Algorithms for Synthetic Aperture Radar Imagery XXIII, Baltimore, MD, USA, May 2016, pp. 98430N-98431–98430N-98410

[16] He, H., Wang, S., Yang, D., *et al.*: 'SAR target recognition and unsupervised detection based on convolutional neural network'. Proc. Chinese Automation Congress (CAC), Jinan, China, October 2017, pp. 435–438

[17] Li, Y., Wang, J., Xu, Y., *et al.*: 'DeepSAR-Net: deep convolutional neural networks for SAR target recognition'. Proc. Int. Conf. Big Data Analysis, Beijing, China, March 2017, pp. 740–743

[18] Shao, J., Qu, C., Li, J.: 'A performance analysis of convolutional neural network models in SAR target recognition'. Proc. SAR in Big Data Era: Models, Methods and Applications, Beijing, China, November 2017, pp. 1–6

[19] Zhao, J., Guo, W., Cui, S., *et al.*: 'Convolutional neural network for SAR image classification at patch level'. Proc. Int. Geoscience Remote Sensing Symp., Beijing, China, July 2016, pp. 945–948

[20] Zhou, Y., Wang, H., Xu, F., *et al.*: 'Polarimetric SAR image classification using deep convolutional neural networks', *IEEE Geosci. Remote Sens. Lett.*, 2016, **13**, (12), pp. 1935–1939

[21] Li, J., Wang, C., Wang, S., *et al.*: 'Classification of very high-resolution SAR image based on convolutional neural network'. Proc. Int. Workshop on Remote Sensing with Intelligent Processing, Shanghai, China, May 2017, pp. 1–4

[22] Malmgren-Hansen, D., Engholm, R., Pedersen, M.O.: 'Training convolutional neural networks for translational invariance on SAR ATR'. Proc. European Conf. Synthetic Aperture Radar, Hamburg, Germany, June 2016, pp. 459–462

[23] Andrew, P., Andres, R., Clouse, H.S.: 'Convolutional neural networks for synthetic aperture radar classification'. Proc. SPIE 9843, Algorithms for Synthetic Aperture Radar Imagery XXIII, Baltimore, MD, USA, May 2016, pp. 98430M-1–98430M-10

[24] Kechagias-Stamatis, O., Aouf, N., Belloni, C.: 'SAR automatic target recognition based on convolutional neural networks'. Proc. Int. Conf. Radar Systems, Belfast, UK, October 2017, pp. 1–4

[25] Wagner, S.: 'Combination of convolutional feature extraction and support vector machines for radar ATR'. Proc. Int. Conf. Information Fusion, Salamanca, Spain, July 2014, pp. 1–6

[26] Geng, J., Fan, J., Wang, H., *et al.*: 'High-resolution SAR image classification via deep convolutional auto-encoders', *IEEE Geosci. Remote Sens. Lett.*, 2015, **12**, (11), pp. 2351–2355

[27] Li, X., Li, C., Wang, P., *et al.*: 'SAR ATR based on dividing CNN into CAE and SNN'. Proc. Asia-Pacific Conf. Synthetic Aperture Radar, Singapore, Singapore, September 2015, pp. 676–679

[28] Chen, S., Wang, H., Xu, F., *et al.*: 'Target classification using the deep convolutional networks for SAR images', *IEEE Trans. Geosci. Remote Sens.*, 2016, **54**, (8), pp. 4806–4817

[29] Ding, J., Chen, B., Liu, H., *et al.*: 'Convolutional neural network with data augmentation for SAR target recognition', *IEEE Geosci. Remote Sens. Lett.*, 2016, **13**, (3), pp. 364–368

[30] Wagner, S., Barth, K., Brüggenwirth, S.: 'A deep learning SAR ATR system using regularization and prioritized classes'. Proc. Radar Conf., Seattle, WA, USA, May 2017, pp. 772–777

[31] Wagner, S.A.: 'SAR ATR by a combination of convolutional neural network and support vector machines', *IEEE Trans. Aerosp. Electron. Syst.*, 2017, **52**, (6), pp. 2861–2872

[32] Deng, S., Du, L., Li, C., *et al.*: 'SAR automatic target recognition based on Euclidean distance restricted autoencoder', *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, 2017, **10**, (7), pp. 3323–3333

[33] Lee, J., Wen, J., Ainsworth, T.L., *et al.*: 'Improved Sigma filter for speckle filtering of SAR imagery', *IEEE Trans. Geosci. Remote Sens.*, 2009, **47**, (1), pp. 202–213

[34] Glorot, X., Bengio, Y.: 'Understanding the difficulty of training deep feedforward neural networks'. Int. Conf. Artificial Intelligence and Statistics, Chia Laguna Resort, Sardinia, Italy, May 2010, pp. 249–256

[35] Maaten, L.V., Hinton, G.: 'Visualizing data using t-SNE', *J. Mach. Learn. Res.*, 2008, **9**, (2605), pp. 2579–2605