Adversarial Deception Against SAR Target Recognition Network

Fan Zhang, *Senior Member, IEEE*, Tianying Meng, Deliang Xiang, *Member, IEEE*, Fei Ma, *Member, IEEE*, Xiaokun Sun, and Yongsheng Zhou, *Member, IEEE*

***Abstract*—Synthetic aperture radar (SAR) automatic target recognition (ATR) technology is one of the key technologies to achieve intelligent interpretation for SAR images. With the rapid development of deep learning, deep neural networks have been successively used in SAR ATR and show priority in comparison with the conventional methods. Recently, more and more attention is paid to the robustness of deep learning-based SAR ATR methods. The reason is that maliciously modified and imperceptible adversarial images can deceive the SAR ATR methods, which are based on the deep neural networks. In this article, we propose a novel SAR ATR adversarial deception algorithm, which fully considers the characteristics of SAR data. Our method can obtain the satisfactory perturbations with a higher deception success rate, higher recognition confidence, and smaller perturbation coverage than other state-of-the-art methods for the SAR images. Experimental results using the MSTAR dataset and OpenSARShip dataset demonstrate the effectiveness of our method. The proposed adversarial deception method can be used in the applications, such as SAR dataset protection, SAR sensor design, and SAR image quality evaluation.**

***Index Terms*—Adversarial attack, automatic target recognition (ATR), deep learning, synthetic aperture radar (SAR).**

I. Introduction

DUE to the imaging ability of day-and-night and weather independence, synthetic aperture radar (SAR) has been widely used for remote sensing for more than 30 years. It plays a significant role in the geographical survey, climate change research, environment monitoring, military information processing, and other applications [1]. With the wide applications of SAR in the remote sensing field, target information extraction from SAR data has become a hot research topic, especially the automatic target recognition (ATR).

Traditional SAR ATR can be composed by three steps, which are preprocessing of SAR images, feature extraction, and classification. The preprocessing step mainly includes filtering [2]–[5] and target detection [6]–[8], which are employed to provide the region of interests (ROIs) in SAR images. The feature extraction aims to reduce redundant information of target regions while providing discriminative representation information [9], [10]. Based on the artificially designed features, the classifiers can be used to determine the exact category information of the ROIs, such as template matching [11], model-based method [12], [13], neural networks [14], machine learning [15]–[17], and deep learning [18]–[21].

In recent years, with the rapid development of deep learning technology, its conspicuous learning ability and classification ability have attracted increased attention in the field of optical remote sensing image processing and SAR target recognition. Compared with the traditional SAR ATR methods, deep learning based methods have achieved satisfactory performance, e.g., deep convolutional autoencoder [22], deep belief network [23], restricted Boltzmann machine [24], convolutional neural networks (CNNs) [18], recurrent neural networks [19], [25], and their derived methods, e.g., spatial-temporal ensemble convolution [26]. Up to now, the state-of-the-art methods can reach a 100% recognition rate on the 10-class dataset of the Moving and Stationary Target Acquisition and Recognition (MSTAR) Program [27].

Deep learning has achieved remarkable results and become the dominant approach for the SAR ATR. However, the interpretability and robustness of deep learning-based SAR ATR are worthy of further discussion. The internal working mechanism and decision-making of the deep learning are relatively complicated, making it difficult to determine the decision-making boundaries. Therefore, the deep learning is fragile. Szegedy *et al.* [28] discovered the existence of this specific disturbance against the neural networks for the first time. Brown *et al.* [29] designed the specific perturbations according to different scenarios of the physical world. Ekholt *et al.* [30] added specific interference stickers on the road signs to deceive the autonomous vehicles in different scenarios, which can make the automatic driving recognition model generate incorrect instructions. Hence, the definition of adversarial sample is given, that is, a modified version of the original image, which can deceive the machine learning classification technique [31].

Adversarial sample can also weaken the performance of various deep neural networks (DNNs), such as the detection

Manuscript received February 20, 2022; revised April 23, 2022; accepted May 20, 2022. Date of publication May 30, 2022; date of current version June 14, 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871413, Grant 62171015, and in part by the Fundamental Research Funds for the Central Universities under Grant XK2020-03. *(Corresponding author: Deliang Xiang.)*

Fan Zhang, Tianying Meng, Fei Ma, Xiaokun Sun, and Yongsheng Zhou are with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: zhangf@mail.buct.edu.cn; mty752559908@vip.qq.com; mafei@mail.buct.edu.cn; sunxk@mail.buct.edu.cn; zhyosh@mail.buct.edu.cn).

Deliang Xiang is with the Beijing Advanced Innovation Center for Soft Matter Science and Engineering, Beijing 100191, China, and also with the Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing 100029, China (e-mail: xiangdeliang@gmail.com).

Digital Object Identifier 10.1109/JSTARS.2022.3179171

networks, segmentation networks, and so on [32]–[34]. For the deep learning-based SAR ATR models, the maliciously modified and imperceptible perturbations can also cause its misjudgments [35]. According to different attack scenarios and objectives, there are several adversarial attack routes. As far as the prior knowledge of deep model is concerned, it can be classified into white box and black box, namely, whether the model parameters are known. In terms of the attack objective, it can be divided into targeted attack and untargeted attack, that is, the target's misclassification category is determined or uncertain. In the early stages of theoretical research, these methods are always combined to implement an adversarial attack. Huang *et al.* [36] generated SAR adversarial samples using the iterative fast gradient sign method (I-FGSM), iterative least likely class method (ILCM), and the decision-based attack (DBA) [37] against different SAR target recognition networks, respectively. Among them, the I-FGSM and ILCM attack the SAR ATR deep models in white-box mode, whereas the I-FGSM achieves targeted attack and ILCM achieves untargeted attack. As for the black-box mode, the DBA [37] is employed to attack the SAR ATR models under the targeted attack. Meanwhile, it can be seen from the experimental results that the SAR adversarial samples can alleviate the recognition accuracy of SAR ATR model by more than 90% [36]. In addition, this work also demonstrates the vulnerability of the DNN-based SAR ATR methods. Li *et al.* [38] used the FGSM [39] and basic iteration method (BIM) methods to generate SAR adversarial samples against the SAR ATR white-box models. Similarly, the experiments in this article prove that the SAR ATR deep model will make misjudgments when recognizing the adversarial image samples.

Although the previous studies verify that the deep learning-based SAR ATR is susceptible to malicious perturbations, there are still several unresolved issues. First, the fooling rate is not promising and can be further improved. Second, as for a successful adversarial sample, the confidence probability is relatively low, especially in black-box mode. Third, the coverage of adversarial perturbation (AP) is relatively wide and can be further decreased for better imperceptibility. Therefore, the following two criteria should be considered to improve the deceptiveness of the adversarial samples in SAR ATR.

1) *Keep strong deceptive ability:* Maximize the fooling rate and the recognition confidence of SAR adversarial samples.
2) *Reduce the visual perception:* Minimize the deceptive perturbation coverage in the SAR image without dramatically changing the backscattering.

Generally, existing SAR adversarial attack methods cannot completely meet the above criteria. In order to further improve the deceptive performance, we propose a new SAR ATR adversarial deception method in this article, which introduces three specific constraints as the improved optimization strategy to generate SAR adversarial samples. Compared with other methods, the SAR adversarial samples generated by our method are expected to be more deceptive and robust. In this article, the proposed method is based on three optimization strategies, utilizing an iterative solution to yield malicious perturbations and further generate SAR adversarial images. The main contributions are given as follows.

1) A higher fooling rate of the adversarial deception method is achieved by minimizing the image differences and maximizing the feature differences between the original and the adversarial images.
2) A higher recognition confidence of the adversarial sample is reached through the recognition probability constraints among classes.
3) A smaller perturbation coverage of the adversarial sample is meet by introducing the nonzero element constraint of the perturbation image.

The rest of this article is organized as follows. In Section II, we introduce the recent studies on adversarial attacks on SAR target recognition network. In Section III, the proposed deception method for SAR ATR is introduced, including the basic generation, the high misjudgment probability generation, and the small perturbation coverage generation of SAR perturbation images. Section IV presents the experimental results and analysis based on the MSTAR and OpenSARShip dataset, as well as the comparisons with other state-of-the-art methods. Finally, Section V concludes this article.

II. Related Works

In this section, we briefly introduce the newly published adversarial attack methods against the SAR target recognition networks. Currently, the I-FGSM (BIM), ILCM, and DBA methods have been utilized for the attack of SAR target recognition network. The basic ideas of these methods are similar, i.e., make the network model produce wrong labels by iteratively adjust the perturbation.

Given an image $X \in \mathbb{R}^n$, there will find a perturbation δ, which will produce a roughly similar image $X' = X + \delta$ (adversarial example). For the images X and X', different labels are marked by the well-trained target recognition network. According to the definition of adversarial example, $C\&W$ model the process of generating δ into as a following constrained minimization problem [40]:

$$\begin{aligned}&\text{minimize} \quad \|\delta\|_2^2 + c \cdot f(X', l)\\ &\text{such that} \quad X' \in [0,1]^n\end{aligned} \tag{1}$$

where c presents a hyperparameter, $f(\cdot)$ is a loss function that reflects the level of adversarial attacks, and l indicates the class label of X'.

In detail, $X + \delta \in [0,1]^n$ is rewritten as $\frac{1}{2}\tanh(W) + 1$. Since $-1 \leq \tanh(W) \leq 1$, so that $0 \leq X + \delta \leq 1$. The perturbation δ can be expressed by W. In this way, we can formulate the problem as follows [40]:

$$\begin{aligned}\text{minimize } \mathcal{L} &= \|\delta\|_2^2 + c \cdot f(X', l))\\ \delta &= \frac{1}{2}(\tanh(W) + 1) - X\end{aligned} \tag{2}$$

where $\mathcal{L}$ presents the loss function of the specific constraint. By minimizing the loss function, we can find the optimal perturbation image δ.

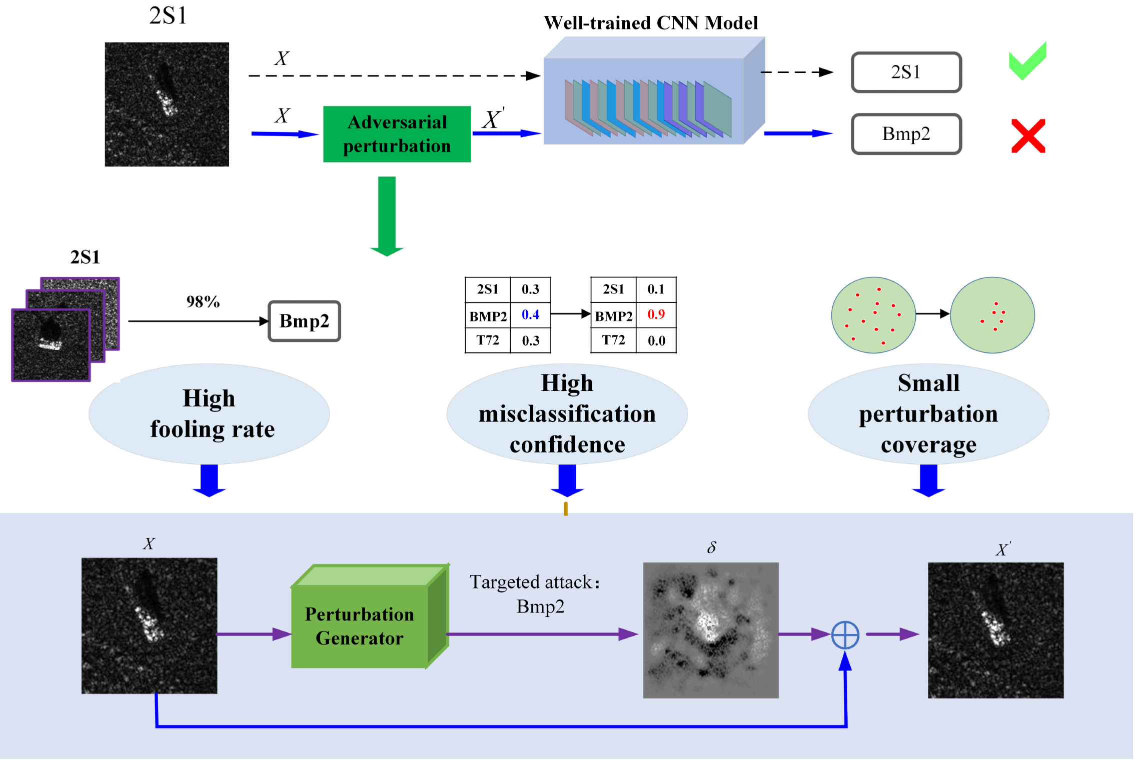


Fig. 1. Illustration of adversarial deception against SAR ATR networks.

So far, the loss function is still an indeterminate expression and is highly related to the deep recognition network's output. In order to realize a high-quality deception in two different ways, namely, targeted attack mode and untargeted attack mode, two objective functions are used for the optimal perturbation generation [40].

1) Objective Function of Targeted Attack Mode: In the targeted attack model, the adversarial sample X' will be determined to a specific target class label t. Accordingly, the objective function in targeted attack mode are as follows:

$$f_1\left(X', t\right) = \max_{i \neq t}(Z(X')_i - Z(X')_t) \tag{3}$$

where $Z(X') \in \mathbb{R}^K$ is the logit layer representation of the recognition network, and $Z(X')_i$ indicates the predicted probability that X' belongs to class i, $i \in (1, K)$.

2) Objective Function of UnTargeted Attack Mode: In the untargeted attacks mode, the adversarial sample X' can be determined to any label, as long as the condition $C(X') \neq C^{(}X)$ can be met. The objective function is set as follows:

$$f_2\left(X', t_0\right) = Z(X')_{t_0} - \max_{i \neq t_0}(Z(X')_i) \tag{4}$$

where t_0 is the class label of the original image X. The objective function $f_2(\cdot)$ intends to increase the probability difference between the non-t_0 class and the t_0 class to realize a high confidence DNN deception.

In general, the perturbation image δ will be calculated by minimizing the loss function $\mathcal{L}$. By constructing different objective functions, adversarial image samples for different application scenarios will be generated to fool the specific DNNs.

III. Proposed Method

In this section, we will introduce the framework of SAR target recognition network deception method, as shown in Fig. 1. Its basic idea is to superimpose an imperceptible perturbation image δ to the original SAR image X to generate an adversarial sample image X', which can trick the SAR ATR model into producing wrong labels. According to the original SAR image X and specific constraints, the perturbation δ will be yielded after iterative optimization. Under the premise of fully considering the characteristics of SAR images (speckles, backscattering, and geometric distortions), three attack objectives are considered for the perturbation generation, which are high fooling rate, high misclassification confidence (MC), and small perturbation coverage. Correspondingly, the generation process is divided into two parts, i.e., perturbation generation and adversarial image generation.

Compared with the low confidence probability in [36], the high confidence constraint can make the classification of the generated SAR adversarial samples with dominant confidence. Compared with the noticeable perturbation in [38], the low perturbation range constraint only allows a few pixels to be disturbed, thus preserving the texture information of the SAR image. In this article, we set two attack modes (targeted and untargeted attacks) for the proposed method under different application scenarios.

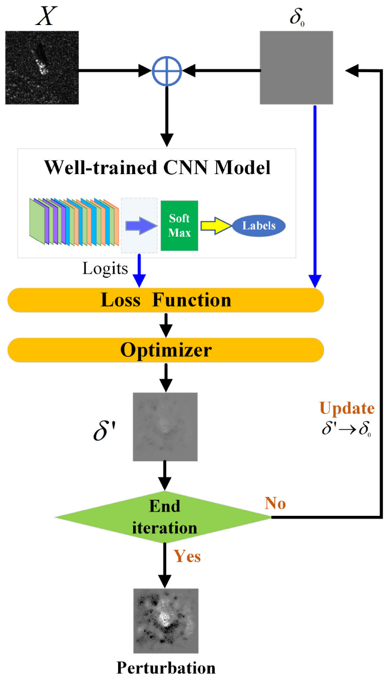


Fig. 2. Generation of SAR adversarial image.

Algorithm 1: Method of Perturbation Image δ Generation.

```
Input: SAR image $X$ and its label $t_0$, initial perturbation
  $\delta_0$ and its target label $t$ (for targeted attack model),
  number of iterations $N$.
Output: Perturbation image $\delta$.
while $(\nabla\mathcal{L}(\delta_0) > 0)||n < N)$ do
    calculate the objective function $f_1(X,'t)$ or
    $f_2(X,'t_0)$;
    calculate the loss function $\mathcal{L}(\delta_0)$ according to (2);
    minimize $\mathcal{L}(\delta_0)$ to obtain $\delta'$;
    Update $\delta_0 \leftarrow \delta'$;
    $n$++;
end while
$\delta = \delta_0$;
```

A. Framework of SAR Adversarial Image Generation

1) Perturbation Generation: As shown in Fig. 2, the generation of SAR perturbation image is an iterative optimization process, through which an AP that meets the objectives can be generated. In each iteration, the SAR image X and the iteration variable δ_0 are superimposed to yield the initial adversarial image X', which will be input into a well-trained CNN model to obtain the inference label. According to the logits of X' and the δ_0, the loss value of $\mathcal{L}(\delta_0)$ is calculated by the loss function (2). Then, a perturbation δ' will be calculated by the optimizer (Adam Optimizer is employed to minimize $\mathcal{L}(\delta_0)$). If the iteration termination condition is not met, the variable δ' will be updated to the iteration variable δ_0, and the next round of iterative optimization will start. When the loss value no longer drops, or the number of iterations is reached, the final perturbation image δ is obtained. The detailed process of perturbation generation is shown in Algorithm 1.

2) Adversarial Image Generation: By superimposing δ on the SAR image X, a SAR adversarial image sample X' is produced. Besides, the best perturbation may not necessarily correspond to the best adversarial image. So, the matrix δ' of each iteration could be saved to generate more adversarial image samples, which can be verified by the well-trained network to find the optimal SAR adversarial image with high fooling rate, high MC and low visual difference.

B. Loss Function of High Fooling Rate

As the first step of the proposed SAR ATR adversarial deception method, the goal is to achieve a high fooling rate, which means the high ratio of the misclassified adversarial samples to the total adversarial examples. According to the basic optimization strategy and two attack modes, we set the fundamental constraint of SAR adversarial image as the following:

$$\begin{cases} \text{Targeted} \ : \mathcal{L}_{B_t} = \|\delta\|_2^2 + c \cdot f_1(X,'t) \\ \text{Untargeted} \ : \mathcal{L}_{B_u} = \|\delta\|_2^2 + c \cdot f_2(X,'t_0) \end{cases} \tag{5}$$

where $\mathcal{L}_{B_t}$ is the loss function under targeted attack mode and $\mathcal{L}_{B_u}$ is the loss function under untargeted attack mode. We can generate SAR adversarial samples by minimizing the loss function under different attack modes.

C. Loss Function of High MC

Technically speaking, a high fooling rate does not mean a successful deception attack. If the MC level is too low, it may not pass the manual or algorithmic rechecks. Therefore, based on the basic loss function, a new constraint that can maximize the misjudgment probability is introduced to improve the adversarial performance. The main idea of the constraint is to increase the logit difference between the misclassified category and the others, and the constraint strategy is depicted as follows:

$$\mathcal{L}_H = \min\{\underset{i \neq l}{\text{sum}}(Z(X')_i) - Z(X')_l\} \tag{6}$$

where l represents the label of misclassified category. It can be seen that the smaller the value of the high constraint function, the higher the confidence of the misclassified perturbation sample. Since the misjudgment category label of the adversarial sample image should be given in advance, only the target attack mode is considered in this constraint optimization.

D. Loss Function of Small Perturbation Coverage

The previous two constraints are proposed from the view of data results and intend to achieve a good and highly credible recognition result. However, the adversarial image sample may be easily checked if such a deceptive image is more differentiated and violates the physics of electromagnetic scattering, e.g., there are some significant perturbations on airport runways where should exhibit low backscattering. Compared with the targets, the scattering distribution of the background clutter is generally more homogeneous. Once this homogeneous texture changes, it will easily attract visual attention. Therefore, a better way to attack SAR recognition networks with low influence on the image texture is to narrow the perturbation coverage.

$C\&W\ l_0$ algorithm is employed to generate the SAR adversarial images with a small perturbation coverage. Its basic idea is to use l_0-norm to constraint the coverage of perturbations. The definition of l_0-norm is depicted as follows:

$$\|\delta\|_0 = \#(\mathrm{i}|\delta_\mathrm{i} \neq 0) \tag{7}$$

which represents the total number of nonzero elements in the image matrix. By minimizing the l_0-norm of perturbation (L2P) image, the number of nonzero perturbation pixels can be significantly decreased. Thus, the definition of the constraint with small perturbation coverage is described as follow:

$$\mathcal{L}_S = \|\delta\|_0. \tag{8}$$

For the l_0-norm, $C\&W\ l_0$ is used to achieve the minimization. $C\&W\ l_0$ is an iterative algorithm. The process of $C\&W\ l_0$ algorithm is explained as follows. First, in each iteration, the basic optimal strategy is used to generate the SAR adversarial image without considering a small range coverage. Second, these pixels, which have less impact on the range coverage, can be found through the gradient of the loss function to the input δ. Then, we set the perturbation of these pixels as zero for the further iteration. In this procedure, the number of zero-pixel in the SAR adversarial image will increase in each iteration. Eventually, the SAR adversarial image with a small coverage can be obtained.

By minimizing the constraint, the SAR perturbation image with a small perturbation coverage is generated. Eventually, $\mathcal{L}_{B_{t,u}}$, $\mathcal{L}_S$, and $\mathcal{L}_H$ are integrated to optimize the generation of SAR adversarial image, which can meet the goals of high fooling rate, high misclassified confidence, and small perturbation coverage. Therefore, the final loss function can be depicted as follows:

$$\begin{cases} \text{Targeted } : \mathcal{L}_T = \|\delta\|_2^2 + f_t(X+\delta) + \mathcal{L}_S + \mathcal{L}_H(X+\delta) \\ \text{Untargeted } : \mathcal{L}_U = \|\delta\|_2^2 + f_u(X+\delta) + \mathcal{L}_S \end{cases}. \tag{9}$$

E. Minimization of Loss Function

As for minimizing the loss function, Adam optimizer [41] is employed to find the optimal δ' through the gradient of the loss functions $\mathcal{L}$. The process of minimizing loss function to get δ is shown in Algorithm 2. The gradient of the SAR target recognition network is hard to solve. In the white box and the black box, the network information we obtain is different, so the way to solve the gradient is also different.

For the SAR target recognition network, its expression can be depicted as follows:

$$\begin{cases} z^i = \varphi_i\left(W^i \cdot a^{i-1} + b^i\right) \\ Z = z^l = \phi(X; W, b) \end{cases} \tag{10}$$

where z^i is the expression of ith layer, W^i and b^i represent the weights and biases of ith layer, respectively. a^{i-1} is the input of ith layer. At the same time, it is also the output of (i-1)th layer. z^l is the logit layer Then, the expression of the logit layer Z can be expressed as a composite function between the input layer and logit layer.

Algorithm 2: Adam Algorithm: Minimizing $\mathcal{L}$ to Get δ.

```
Input: $\mathcal{L}$: Loss function with parameters $\delta \in \mathbb{R}^n$ (n pixels)
$\alpha = 0.01$: StepSize, $\beta_1 = 0.9$, $\beta_2 = 0.999$: Exponential
  decay rates for the moment estimates in Adam algorithm.
$\delta_0$: Initial parameter
$m_0 \longleftarrow 0$, $v_0 \longleftarrow 0$, $t \longleftarrow 0$, Same as $\delta's$ dimension
$m_0$, $v_0$ are the first and second moments, respectively
Output Perturbation image $\delta$.
while $\delta_t$ is not converged do
    $t$++;
    $g_t \leftarrow \nabla_\delta \mathcal{L}(\delta_{t-1})$;
    White box: Estimate it by backpropagation of (10);
    Black box: Estimate it by calculating (11).
    $m_t \leftarrow \beta_1 \cdot m_{t-1} + (1-\beta_1) \cdot g_t$, Update $m_t$;
    $v_t \leftarrow \beta_2 \cdot v_{t-1} + (1-\beta_2) \cdot {g_t}^2$, Update $v_t$;
    $\widehat{m}_t \leftarrow m_t/(1-\beta_1^t)$, Bias-corrected $m_t$;
    $\widehat{v}_t \leftarrow v_t/(1-\beta_2^t)$, Bias-corrected $v_t$;
    Update $\delta_t \leftarrow \delta_{t-1} - \alpha \cdot \widehat{m}_t/(\sqrt{\widehat{v}_t} + \epsilon)$;
end while
$\delta$=$\delta_t$;
```

In the white box, weights and biases of each layer of the SAR target recognition network are known. So, we can obtain the gradient by the composite function of the white-box network.

However, in the black box, we can only get the classification score $F(.)$ of the network. The symmetrical difference is employed to get each pixel's gradient of the black-box network. The definition of symmetrical difference is

$$\nabla_{\delta_\mathrm{i}} F(\delta) \approx \frac{F\left(\delta + he_i\right) - F\left(\delta - he_i\right)}{2h} \tag{11}$$

where h is a small constant, which is set to 0.0001 in the experiments, e_i is a standard basis vector that has the same dimension with δ, and the ith component of e_i is 1, and the rests are 0.

Compared with the black-box network, the white-box network has a faster gradient solution speed and more accurate gradient information. However, it requires more information about the SAR target recognition network. In this way, we can get the optimal δ by the Adam optimizer.

IV. Experimental Results and Analysis

In this section, four different scenarios are designed to discuss the effectiveness of the proposed method for the experiments, namely, targeted attack on white-box CNN, targeted attack on black-box CNN, untargeted attack on white-box CNN, and untargeted attack on black-box CNN. In the experiments, the deception success rate is selected as an indicator and three state-of-the-art methods are chosen for the comparison. The employed CNN networks are implemented under the Keras architecture and TensorFlow backend. Table I gives the basic configuration of the platform.

TABLE I
CONFIGURATION OF EXPERIMENT PLATFORM

Programming language	Python 3.5
Deep learning API	tensorflow 1.4 & Keras 2.1.6
CPU	Intel core i5-8500
Memory	16GB
GPU	GeForce GTX 2070

TABLE II
NUMBER OF SAMPLES IN THE MSTAR DATASET

Class	Name	Train	Test	Adversarial	
				Untargeted	Targeted
0	2S1	299	275	275	275*9
1	BRDM2	298	274	274	274*9
2	BTR60	256	195	195	195*9
3	D7	299	274	274	274*9
4	T62	299	273	273	273*9
5	ZIL131	299	274	274	274*9
6	ZSU234	299	274	274	274*9
7	BMP2	233	195	195	195*9
8	BTR70	233	296	296	296*9
9	T72	232	196	196	196*9
Total	-	2747	2526	2526	2526*9

TABLE III
NUMBER OF SAMPLES IN THE OPENSARSHIP DATASET

Class	Name	Train	Test	Adversarial	
				Untargeted	Targeted
1	BulkCarrier	261	261	261	261*2
2	CargoShip	525	525	525	525*2
3	Fishing	372	372	372	372*2
Total	-	1158	1158	1158	1158*2

TABLE IV
MODEL ARCHITECTURES FOR THE CNN MODEL

Layer type	Layer structure
Convolution + ReLU	3×3×32
Max Pooling	2×2
Convolution + ReLU	3×3×32
Max Pooling	2×2
Convolution + ReLU	3×3×64
Max Pooling	2×2
Convolution + ReLU	3×3×64
Max Pooling	2×2
Fully Connected + ReLU	200
Fully Connected + ReLU	200
Fully Connected + ReLU + Softmax	10(MSTAR), 3(OpenSARship)

TABLE V
TRAINING AND TESTING ACCURACY OF THE CNN MODEL

Dataset	Model	Training	Testing
MSTAR	CNN Model	100%	97.32%
OpenSARship	CNN Model	99.90%	98.45%

A. Experimental Data and Model

1) MSTAR Dataset: In this article, we adopt the widely used MSTAR dataset, which includes ten types of vehicle targets and has different image sizes. We use the dataset with 17 depression angles as the training set and the dataset with 15 depression angles as the test set. The training set and validation set include 2547 and 200 images, respectively. The details of the dataset are given in Table II. In addition, in order to make the image size uniform for the CNN input, we center-crop each image to 128×128.

2) OpenSARship Dataset: In order to verify the robustness of the method, the OpenSARShip dataset is also employed in the experiments. The dataset includes three types of ship targets, which have the same image size (512 × 512). The training set and test set include 1050 and 108 images, respectively. We crop each image to 128 × 128 for the network input. The details of the dataset are given in Table III.

3) Adversarial Sample Set: For these two datasets, their test samples are selected as the original images of the adversarial samples. In untargeted attack mode, the label will be confused as any category other than its own category. For each original image, a random label from other categories is distributed, and accordingly, the number of adversarial samples is the same as the number of the test set. In targeted attack mode, the label will be confused as a specific class other than its own category. For each original image, adversarial samples are generated by assigning the corresponding labels form other classes. Therefore, the number of adversarial samples in the MSTAR dataset is nine times the number of the test set. Similarly, the number of adversarial samples in the OpenSARShip dataset is two times the number of the test set.

4) Well-Trained CNN Model: The CNN model architecture is given in Table IV. The CNN model is composed of four convolution layers and three fully connected layers. In addition, the ReLU activation function and the maximum pooling method are adopted in the CNN model. The ReLU activation function can suppress the gradient diffusion phenomenon of the network model to a certain extent, and it has high computational efficiency.

In the experiments, the input image size is 128×128. We use the SGD optimizer to train it. The hyperparameters are set as follows: learningrate $= 0.01$, decay $= 1e$-6, momentum $= 0.9$, epoch $= 30$ and batchsize $= 32$. The final training results are given in Table V.

5) Evaluation Metrics for the SAR Adversarial Images: To better evaluate the performance of the proposed method, four evaluation metrics are used for the SAR adversarial images. First, the fooling rate is introduced to verify the effectiveness of the proposed method. It represents the ratio of the number of samples that can successfully fool the SAR ATR model (N_{success}) to the number of all adversarial samples (N_{all}), and can be expressed as the following:

$$\text{fooling rate } = \frac{N_{\text{success}}}{N_{\text{all}}}. \tag{12}$$

Second, the MC is employed to verify the deception performance of SAR adversarial samples against SAR ATR models. It indicates the confidence probability that a SAR image of a particular class is misclassified into other class, and can be acquired from softmax unit of ATR model.

Third, the structural similarity (SSIM) and L2P are introduced to evaluate the imperceptibility of the SAR adversarial sample.

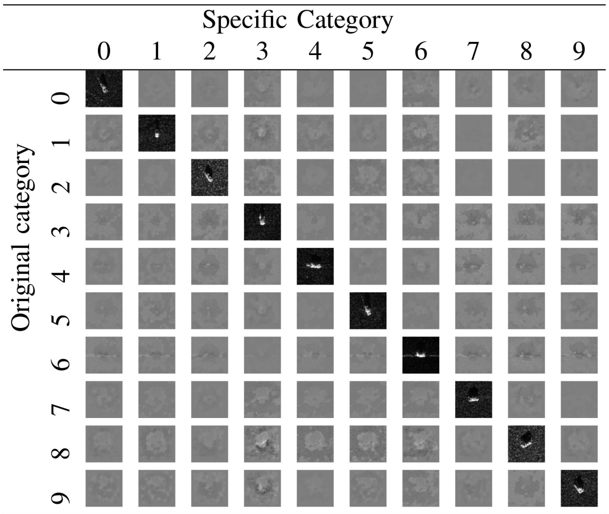


Fig. 3. SAR adversarial samples generated in the white-box state under targeted attack mode.

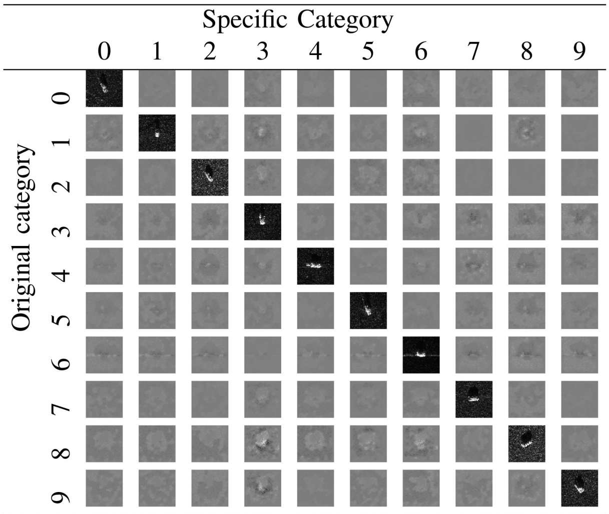


Fig. 4. SAR adversarial samples generated in the black-box state under targeted attack mode.

The L2P is expressed as the following:

$$\mathrm{L2P} = \|\delta\|_2 = \sqrt{\delta^2_{(1,1)} + \delta^2_{(1,2)} + \cdots + \delta^2_{(n,n)}}, n = 128. \tag{13}$$

At last, the number of disturbing pixels (NDP) is used to verify the controlling effect of the perturbation coverage. For each perturbation image, pixels whose pixel gray value is greater than 0 are regarded as the disturbing pixels.

B. SAR Adversarial Deception With MSTAR Dataset

In this section, attack experiments are conducted under four different scenarios. First, the construction of adversarial deception set will be introduced. Meanwhile, the visual effect of these perturbations will be presented. Second, the detailed evaluation on adversarial deception will be discussed among these four scenarios.

In targeted attack mode, for each SAR image, we set the other nine categories as their specific target categories to generate nine adversarial samples separately, as shown in Figs. 3 and 4. The nine perturbation images of each SAR image are quite different regardless of whether it is in the white-box or black-box CNN models. Take the image in the left corner as an example, it will be misclassified to the fifth class if the fifth perturbation in the first row is superimposed on itself.

In the untargeted attack mode, for each SAR image, the class label of adversarial image is an arbitrary label generated at random except for its own label, as shown in Figs. 5 and 6. In each three columns, the left column lists the original images, the middle column lists the corresponding perturbations, and the right column lists the adversarial images. It can be seen that the perturbations of white-box CNN is weaker than that of black-box CNN. The reason may be that the white-box way acquires more information than the black-box way.

Furthermore, to verify the imperceptibility of the SAR adversarial image, the SSIM is used to evaluate the similarity between the original SAR image and the SAR adversarial image, as shown in Tables VI and VII. As for the experimental data, one SAR image is selected for each class in the targeted and untargeted experiments, ten SAR images are selected for each class in the average SSIM experiment, including the targeted and untargeted attack. In the experiment of targeted attack, a SAR image is selected from each are employed to generate 90 adversarial images The SSIMs of the white-box attack are slightly bigger than that of the black-box attack, and the SSIMs of untargeted attack are bigger than that of targeted attack. Compared with the untargeted mode, the extra constraint condition in targeted attack mode will increase the noisy level of AP. In general, the adversarial image generated by the proposed method is highly similar to the original SAR image, thus achieving high concealment.

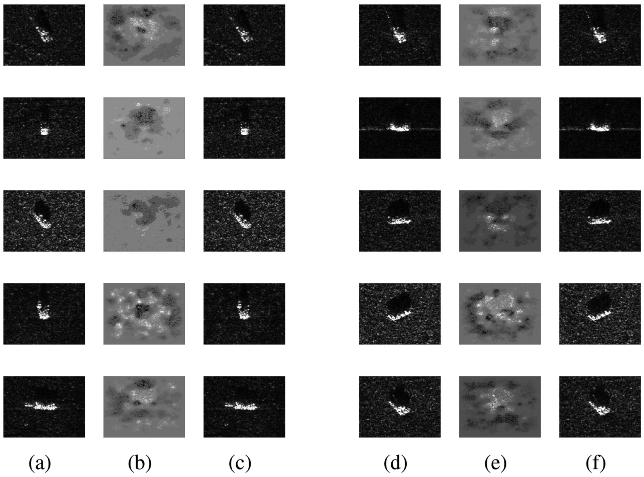


Fig. 5. SAR adversarial samples generated in the white-box state under untargeted attack mode. (a) Original image of class 0–4. (b) Perturbation image. (c) Adversarial image of class 0–4. (d) Original image of class 5–9. (e) Perturbation image. (f) Adversarial image of class 5–9.

In order to evaluate the adversarial performance of the proposed method, nine adversarial image samples of 2S1 category

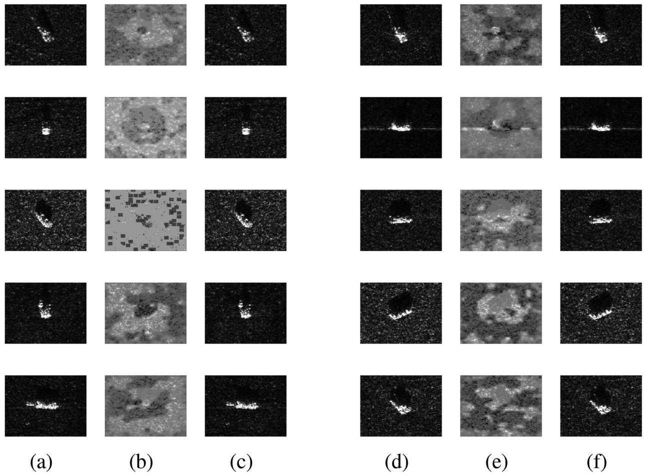


Fig. 6. SAR adversarial samples generated in the black-box state under untargeted attack mode. (a) Original image of class 0–4. (b) Perturbation image. (c) Adversarial image of class 0–4. (d) Original image of class 5–9. (e) Perturbation image. (f) Adversarial image of class 5-9.

TABLE VI
SSIM OF THE ORIGINAL SAR IMAGE AND SAR ADVERSARIAL IMAGE IN THE WHITE-BOX STATE

SSIM under Targeted Attack										
ori-adv	0	1	2	3	4	5	6	7	8	9
0	-	0.993	0.989	0.972	0.979	0.994	0.975	0.983	9.976	0.968
1	0.987	-	0.982	0.962	0.986	0.986	0.969	0.996	0.961	0.987
2	0.991	0.989	-	0.966	0.987	0.977	0.973	0.998	0.999	0.994
3	0.968	0.979	0.953	-	0.979	0.987	0.973	0.919	0.926	0.870
4	0.969	0.981	0.917	0.967	-	0.986	0.969	0.916	0.894	0.908
5	0.969	0.969	0.974	0.980	0.987	-	0.984	0.963	0.945	0.959
6	0.917	0.976	0.919	0.983	0.978	0.969	-	0.827	0.848	0.808
7	0.961	0.978	0.992	0.974	0.974	0.981	0.977	-	0.981	0.997
8	0.954	0.952	0.955	0.891	0.944	0.941	0.940	0.960	-	0.973
9	0.982	0.979	0.974	0.950	0.985	0.981	0.971	0.968	0.962	-
SSIM under Untargeted Attack										
ori	0	1	2	3	4	5	6	7	8	9
	0.994	0.996	0.999	0.977	0.985	0.988	0.979	0.992	0.990	0.991
Average SSIM										
ori	0	1	2	3	4	5	6	7	8	9
	0.986	0.975	0.983	0.981	0.993	0.987	0.982	0.984	0.978	0.980

TABLE VII
SSIM OF THE ORIGINAL SAR IMAGE AND SAR ADVERSARIAL IMAGE IN THE BLACK-BOX STATE

SSIM under Targeted Attack										
ori-adv	0	1	2	3	4	5	6	7	8	9
0	-	0.996	0.996	0.977	0.991	0.998	0.974	0.989	0.988	0.985
1	0.978	-	0.972	0.951	0.979	0.987	0.956	0.999	0.961	0.998
2	0.997	0.997	-	0.983	0.997	0.985	0.985	0.999	0.999	0.998
3	0.972	0.977	0.937	-	0.981	0.980	0.958	0.949	0.924	0.918
4	0.980	0.983	0.951	0.968	-	0.990	0.978	0.936	0.934	0.966
5	0.986	0.972	0.979	0.984	0.994	-	0.991	0.976	0.958	0.983
6	0.926	0.945	0.939	0.993	0.968	0.952	-	0.920	0.852	0.890
7	0.984	0.987	0.995	0.974	0.982	0.985	0.973	-	0.982	0.997
8	0.989	0.981	0.986	0.913	0.963	0.966	0.945	0.994	-	0.992
9	0.985	0.985	0.981	0.937	0.995	0.984	0.969	0.983	0.975	-
SSIM under Untargeted Attack										
ori	0	1	2	3	4	5	6	7	8	9
	0.990	0.992	1.000	0.947	0.976	9.981	0.971	0.988	0.986	1.000
Average SSIM										
ori	0	1	2	3	4	5	6	7	8	9
	0.989	0.981	0.983	0.979	0.988	0.985	0.984	0.979	0.978	0.988

are tested by the CNN model, and their confidence probabilities are shown in Fig. 7. Basically, the confidence of white-box attack is much higher than that of the black-box attack, and can be stabilized above 70%. As for the black-box attacks, although they are not dominant, the confidences still exceed 50%.

C. Performance of Improved Loss Function

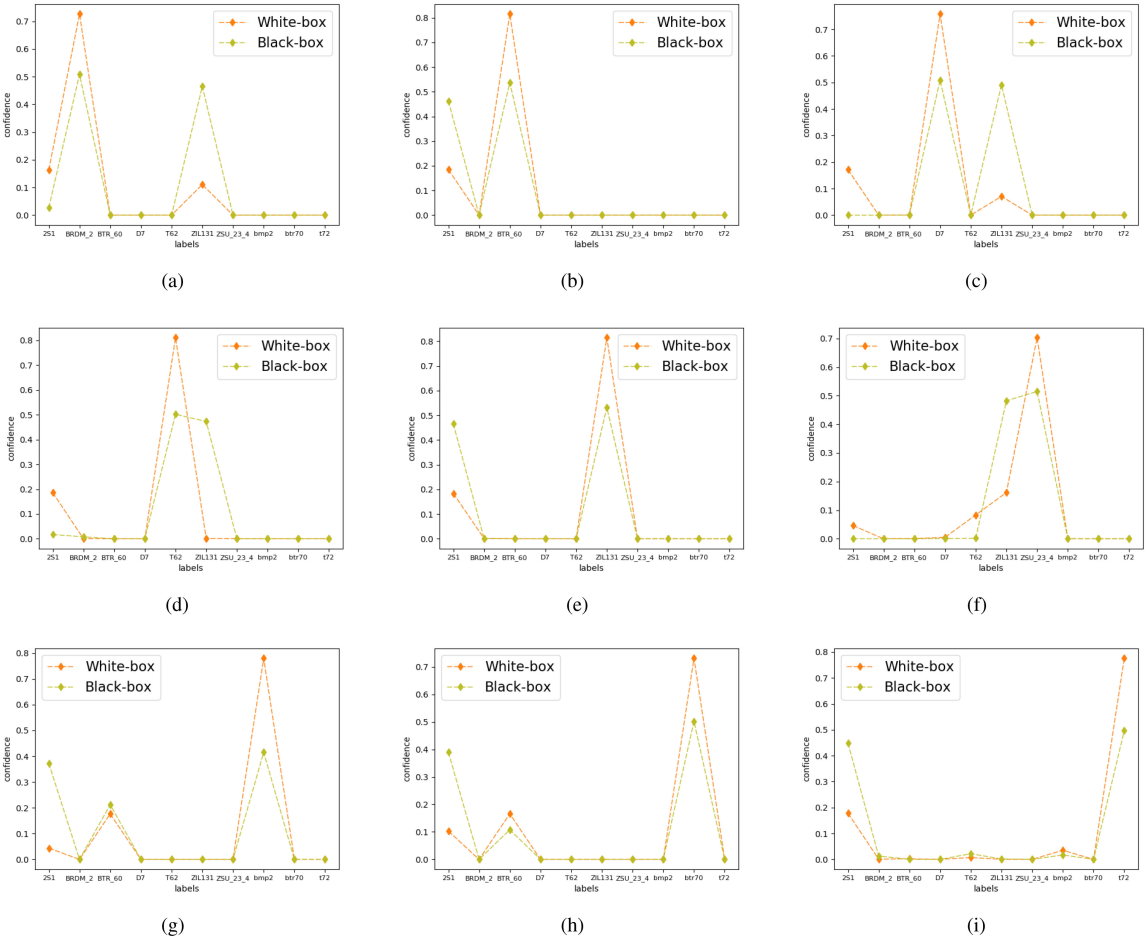
In this article, we propose the constraints of high fooling rate, high MC, and small perturbation coverage for the deception algorithm to reach the criteria (reduce the change of texture and keep strong deceptive ability of SAR adversarial image). This section will verify the positive performance of the constraints on SAR adversarial deception algorithm.

First, an original SAR image of 2S1 category is employed to perform the ablation experiments, which are the attack with high fooling rate constraint, attack with high fooling rate and confidence probability constraints, and attack with all three constraints. Second, in targeted attack mode, nine adversarial images are generated under the high fooling rate constraint, nine adversarial images are generated under the first two constraints, and nine adversarial images are generated under the whole constraints, respectively. Finally, in the case of a successful adversarial attack, we compare the effects of SAR adversarial images under the three constraints from different perspectives, as shown in Table VIII.

Each row of Table VIII lists the perturbation images of different constraint combinations and their evaluation indicators, which include MC, L2P, NDP, and SSIM. From the results, it can be seen that every additional constraint will bring performance improvement. As far as the first constraint is concerned, L2P indicates that the perturbation intensity level is satisfactory, but the confidence is rather low, and the number of contaminated pixels is relatively high. In this situation, the average MC is less than 40%. After adding the confidence constraint, it can be seen that the average MC is improved to around 60%. However, the average NDP is still greater than 12 000 pixels. For L2P, it and MC are proportional. In this sense, a higher MC will be achieved as the perturbation intensity increases. But too high perturbation intensity will cause apparent visual changes, which may not pass the human visual inspection. As for the coverage constraint, its advantage is obvious that the average NDP is reduced to a level of less than 100 pixels, and the high confidence is still maintained. Visually, the perturbation is concentrated on a small part of the pixels (less than 1% of the whole SAR image). Compared with the previous two cases, the proposed method with three constraints is easier to achieve SAR adversarial deception.

D. SAR Adversarial Image Discussion With the OpenSARship

To further validate the robustness of the proposed method, the OpenSARShip dataset is used for evaluation. For verifying the deceptive ability of the proposed method, the adversarial samples are generated under four different scenarios, which



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Fig. 7. Confidence of SAR adversarial samples generated in the white-box state and the black-box state under targeted attack mode. The category of the original sample is 2S1. The category of the SAR adversarial example is: (a) BRDM2, (b) BTR60, (c) D7, (d) T62, (e) ZIL131, (f) ZSU234, (g) BMP2, (h) BTR70, and (i) T72.

TABLE VIII
PERFORMANCE OF SAR ADVERSARIAL IMAGE UNDER DIFFERENT CRITERIA

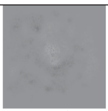
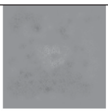
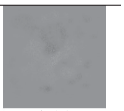
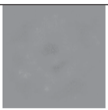
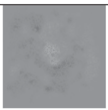
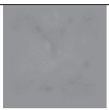
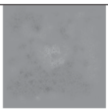
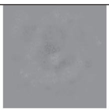
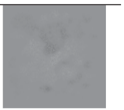
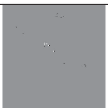
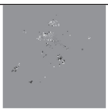
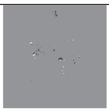
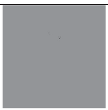
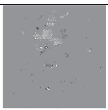
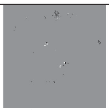
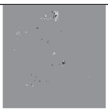
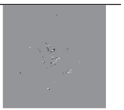
	Specific category								
	BRDM2	BTR60	D7	T62	ZIL131	ZSU234	bmp2	btr70	t72
✖									
MC	33.64%	50.24%	32.25%	25.45%	50.23%	30.89%	33.75%	35.49%	25.66%
L2P	0.65776	0.65915	1.79322	1.44070	0.37793	2.03865	1.07145	1.17561	1.30916
NDP	12249	11533	13501	13130	10442	13958	12225	12395	12436
SSIM	0.9983	0.9980	0.9870	0.9947	0.9992	0.9838	0.9955	0.9937	0.9921
✓									
MC	58.48%	73.19%	54.72%	47.73%	74.01%	57.65%	57.83%	56.89%	54.52%
L2P	1.01452	0.95452	2.25545	1.50455	0.60340	2.56139	1.50534	1.44186	1.57297
NDP	12963	12362	13858	13548	11212	14300	13511	13057	13122
SSIM	0.9962	0.9965	0.9783	0.9921	0.9983	0.9745	0.9929	0.9909	0.9891
✓									
MC	57.72%	73.11%	56.06%	50.04%	73.06%	57.59%	55.83%	57.13%	44.43%
L2P	4.30094	3.43184	8.29439	5.82525	2.16334	7.65638	6.17669	4.76509	5.08823
NDP	34	25	312	85	6	482	107	142	94
SSIM	0.9973	0.9980	0.9731	0.9892	0.9993	0.9644	0.9953	0.9961	0.9906

TABLE IX
Fooling Rates of SAR Adversarial Images Under Four Scenarios With OPENSARSHIP Dataset

	Untargeted	Targeted
White box	99.01%	98.01%
Black box	80.80%	67.00%

are white-box CNN with targeted attack, white-box CNN with untargeted attack, black-box CNN with targeted attack, and black-box CNN with untargeted attack. As given in Table IX, the fooling rate of white-box attack is bigger than that of black-box attack, and untargeted attack is better than the targeted attack. It can be concluded that a successful adversarial deception requires more CNN model information and less dependence on target information.

In order to better analyze the performance of the proposed method, an original OpenSARship image is randomly selected from the bulk carrier category, and is employed to generate SAR adversarial samples under four different scenarios. As given in Table X, the generated SAR adversarial samples are highly similar to the original images from the visual check and the SSIM indicator. From its perturbation images, we can analyze the differences among the SAR adversarial samples of the four different scenarios.

Under the untargeted attack mode, the L2P and NDP are smaller than that of the targeted attack mode. From the aspect of visual effect, the SSIM of untargeted attack is better than the targeted attack. In terms of the perturbation image, the white-box attack can realize a smaller coverage of perturbation while the balck-box attack requires a larger coverage of disturbance.

For the white-box SAR ATR model, based on the model prior, the impact of different SAR adversarial samples on the ATR model can be obtained for the iterative optimization. Therefore, the white-box attack method can quickly generate the SAR adversarial samples that concentrate the perturbation in the target region. However, for the black-box SAR ATR network, the impact of SAR perturbation is hard to acquire. The limited information lies on whether the perturbation can achieve a successful attack, namely, the classification probabilities and the determined labels. The experimental results show that the NDP in the white-box attack is much lower than that of the black-box attack, and the former attack also can reach adversarial deception with higher confidence.

E. Analysis of the SAR Adversarial Image

In this section, in order to better represent the effectiveness of the proposed method, we analyze the SAR adversarial samples from the following three aspects, namely, the feature map analysis, transferability analysis, and physical realizability analysis. Among them, the feature map part analyzes the key information of the SAR adversarial sample that leads to its misjudgment in the CNN-based SAR ATR model. The transferability part analyzes the effect of adversarial samples on different networks and initially discusses the applicability of SAR adversarial samples. The physical realizability section analyzes the perturbation distribution of the SAR adversarial samples and discusses the mechanism of physical camouflage realization of the SAR target in detail.

1) Feature Map Analysis: Generally, one important advantage of CNN is to automatically extract or learn the features from the training samples. Thus, the key point of its decision-making is the feature maps of SAR images from the convolutional layers. Therefore, we conduct an experiment to demonstrate the recognition state inside the so-called black-box model of CNN for disclosing the recognition mechanism. In addition to the original and adversarial samples, the experiment also adds the Rayleigh noise sample, which is generated by multiplying Rayleigh noise with the original image, to simulate the severe speckle effect.

As shown in Fig. 8, the first row lists the feature maps of the original sample, the second row lists the feature maps of the original sample multiplied by Rayleigh noise, and the third row lists the feature maps of the original sample with AP added. It can be seen that the feature maps of the original and adversarial samples are significantly different from the third layer to the last layer. Accordingly, the recognition results are confused from Bmp2 to ZIL131 for the CNN model, and misclassified from T72 to BTR60 for the AconvNet. On the other hand, for the original and Rayleigh noise samples, their feature maps are basically similar in visual representation. Thus, the simulated severe speckle effect on the original sample does not affect the recognition results.

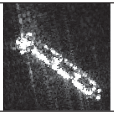
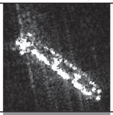
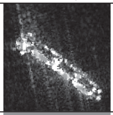
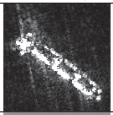
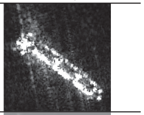
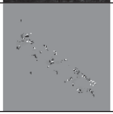
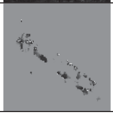
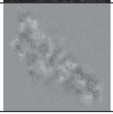
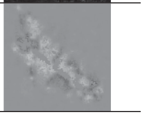
Although the adversarial sample maintains the similarity with the original SAR image, its perturbation changes the critical information of the SAR image representation, and brings the obvious changes in feature maps, resulting in misclassification. For SAR adversarial sample, its perturbation is controlled in a small range, so the key regions of SAR images for the CNN-based model can be reflected in the perturbation images.

2) Transferability Analysis: In order to disclose the transferability of the adversarial samples, two experiments are designed, respectively, white-targeted attack and white-untargeted attack. The adversarial samples are generated from MSTAR datasets with the proposed method and the employed CNN model (see Table XI). The well-trained classical shallow and moderate DNN, namely AConvNet and ResNet18, are employed to evaluate the transferability of the adversarial samples.

In the experiment on white-targeted attack, the baseline fooling rate for the CNN model is 98.11%, and the fooling rates decrease by about 27% for AConvNet and 15% for ResNet18. In another experiment, the baseline fooling rate is 98.95%, and the fooling rates decrease by about 58% for AConvNet and 20% for ResNet18. For different adversarial attack modes, the fooling rate of targeted attack mode is better than that of untargeted attack mode. For the adversarial sample set generated under targeted attack mode, the generation process is more targeted, thus having better transferability in other CNN-based SAR ATR models. From the results, it can be seen that the adversarial samples can achieve successful adversarial attacks against other neural networks to a certain extent, and are more likely to fool the DNN, such as ResNet18.

3) Physical Realizability Analysis: In this section, we analyze the SAR adversarial image at the perspective of physical

TABLE X
PERFORMANCE OF SAR ADVERSARIAL IMAGE UNDER FOUR SCENARIOS WITH OPENSARSHIP DATASET

	Original image	White untargeted	White targeted	Black untargeted	Black targeted
SAR image					
Perturbation image					
MC	Bulk Carrier: 99.99%	Cargo Ship: 70.58%	Fishing: 69.70%	Cargo Ship: 66.96%	Fishing: 68.88%
L2P		3.75992	4.22674	4.77711	5.20934
NDP		873	953	16384	16384
SSIM		0.9988	0.9952	0.9971	0.9909

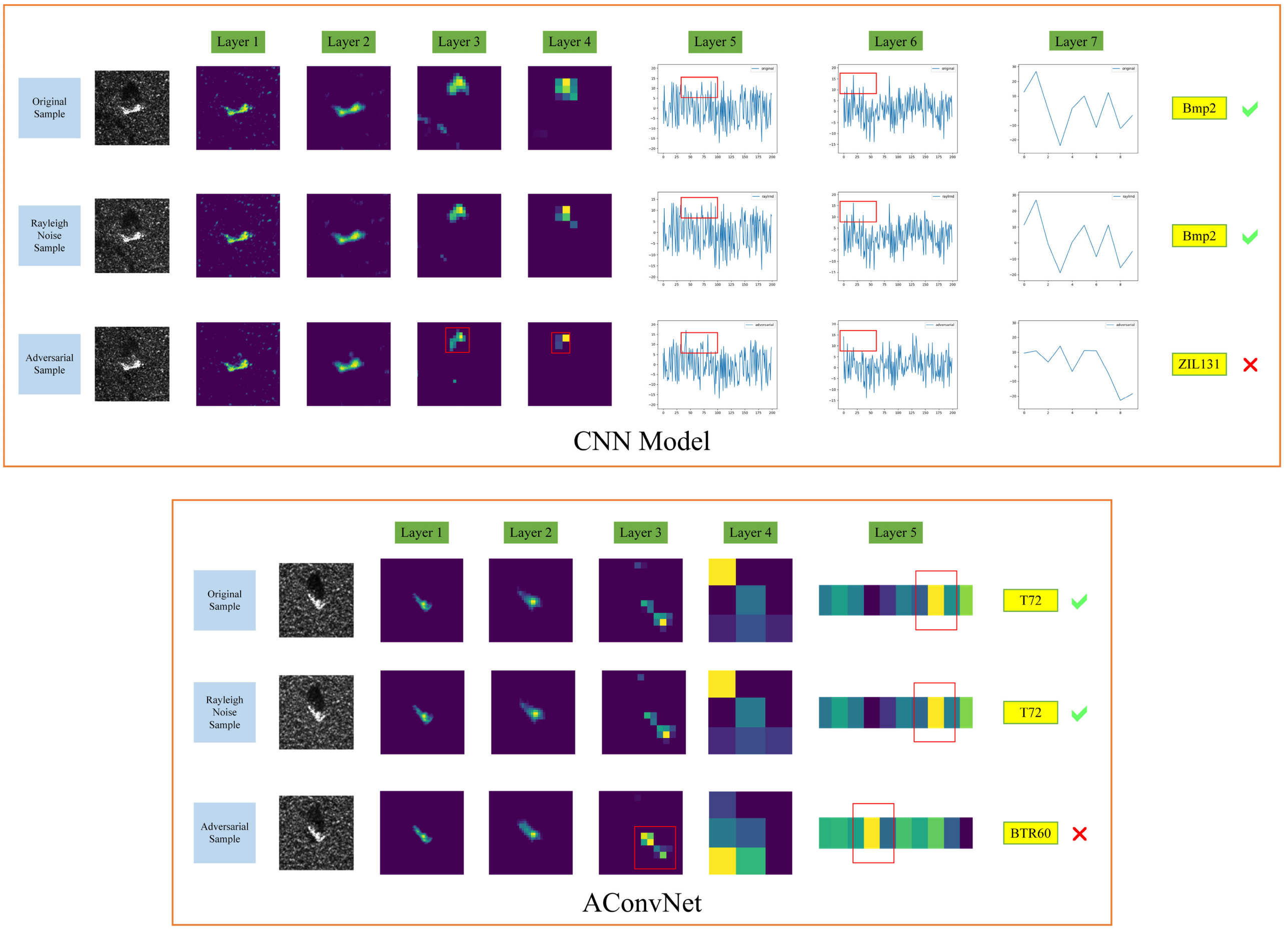


Fig. 8. Feature maps of each layer in CNN Model and AConvNet.

realization. As shown in Fig. 9, each image in the MSTAR dataset consists of three parts: the vehicle area, shadow area, and background area.

Due to the limitation of electromagnetic waves, the electromagnetic waves around the vehicle will be blocked by the vehicle, so the corresponding location in the SAR image will be imaged as a shadow area. Therefore, it is difficult to add perturbation in the shadow area. And due to the irradiation angle of electromagnetic waves and the characteristics of the target vehicle, some bright spots (strong scattering points) will be formed in the vehicle area, as shown in the SAR image. Therefore, for the vehicle area, a corresponding perturbation can be added to change the distribution of bright spots (scattering information) in the vehicle area. As for the background area of the vehicle, the background is usually complex and changeable, such as an ocean, urban area, farmland, and desert. Compared with the background area, the types of vehicles of interest are limited. So it is relatively easy to add perturbations to the target.

TABLE XI
FOOLING RATES OF ADVERSARIAL SAMPLES ON DIFFERENT NETWORK MODELS

	CNNModel	AConvNet	ResNet18
white-targeted attack	98.11%	71.70%	83.86%
white-untargeted attack	98.95%	41.50%	78.95%

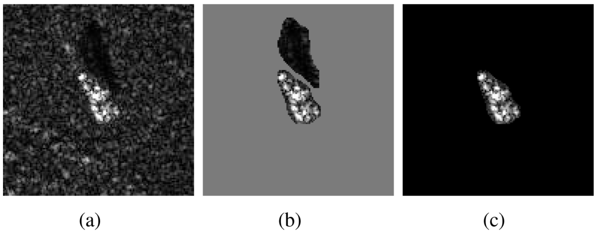

(a) (b) (c)

Fig. 9. Examples of extracting vehicle areas from SAR image. (a) Original SAR image. (b) Three areas of image (vehicle, shadow, and background). (c) Vehicle area.

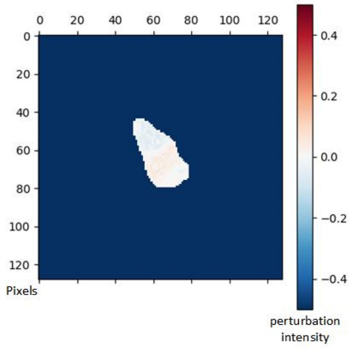


Fig. 10. Heatmap of perturbation image on vehicle area. Perturbation intensity: the gray level difference between the SAR adversarial sample and the original sample.

Therefore, considering the subsequent physical realization of the SAR adversarial image, we only care about the perturbation of the vehicle area.

We select an image in the 2S1 category and use it to generate SAR adversarial image. In addition, we extract the perturbation in the vehicle area and display it in the heat map.

As shown in Fig. 10, for SAR adversarial image, the red pixel indicates that its gray level has increased relative to the original SAR image, and the blue area indicates that its gray level has decreased relative to the original SAR sample. It can be seen from the experimental results that the perturbation of the SAR adversarial image is concentrated on the boundary and the inner side of the vehicle. In particular, at the boundary of the vehicle, the brightness of the SAR disturbance becomes brighter, whereas at the inner side of the vehicle, the brightness of the SAR disturbance becomes darker. Therefore, according to the physical characteristics of the SAR image and the law of SAR disturbance, some strong scattering objects, which can change the SAR backscattering can be added to the vehicle to achieve the effect of physically deceiving the SAR target recognition network.

TABLE XII
FOOLING RATES OF SAR ADVERSARIAL IMAGES GENERATED BY DIFFERENT METHODS WITH MSTAR DATASET

Generating SAR adversarial image				
Method	FGSM [38]	BIM [36]	DBA [36]	Ours
AConvNet	83.66%	94.44%	88.91%	98.87%
ResNet18	79.35%	91.32%	83.66%	98.31%

F. Performance Comparison

A comparative experiment with state-of-the-art methods are designed to evaluate the performance of the proposed method. Huang *et al.* [36] and Li *et al.* [38], respectively, used the mainstream methods to attack SAR ATR networks, including FGSM [38], BIM [36], [38], and DBA [36]. In the experiments, the shallow SAR ATR model (AConvNet) and deep SAR ATR model (ResNet) are attacked and compared based on the MSTAR dataset. After training, the classification accuracy of the AConvNet model reached 97.61% and the classification accuracy of the ResNet18 model reached 98.39%.

The experimental results are given in Table XII. For shallow SAR ATR model, the fooling rate of FGSM, BIM, and DBA are 83.66%, 94.44%, and 88.91%, respectively. The proposed method achieves a fooling rate of 98.87%, outperforming other methods by an average of 9.87%. For deep SAR ATR model, the fooling rate of FGSM, BIM, and DBA are 79.35%, 91.32%, and 83.66%, respectively. The proposed method achieves a fooling rate of 98.31%, outperforming other methods by an average of 13.53%. The experimental results show that the proposed method can implement a better adversarial deception for both shallow and deep SAR ATR models. In addition, as the depth of the SAR ATR model increases, the proposed method still maintains its superior fooling rate.

V. CONCLUSION

Aiming at the shortcomings of deep learning networks that are susceptible to small perturbation, this article employs MSTAR and OpenSARship dataset in the SAR ATR network model and uses a regularization constraint method to generate SAR adversarial images in different modes. Specifically, considering the speckle and backscattering characteristics of SAR images, three different constraints, namely, high fooling rate, high confidence, and small perturbation coverage, are introduced to generate the SAR adversarial images that can deceive the well-trained SAR target recognition network. These SAR adversarial images are deceptive and are more conducive to subsequent physical disturbance analysis. Not only that, the previous method applied to the SAR ATR network is compared with the proposed method. The experimental results show that the proposed method has a higher fooling rate. In the follow-up work, we will further study the generation rules of SAR adversarial images and control the perturbation coverage on the vehicle area and the area around the vehicle. At last, combining the imaging characteristics of the SAR, the perturbation can physically realize, thereby forming the camouflage of the target object.

References

[1] A. Moreira, P. Prats-iraola, M. Younis, G. Krieger, I. Hajnsek, and K. Papathanassiou, "A tutorial on synthetic aperture radar," *IEEE Geosci. Remote Sens. Mag.*, vol. 1, no. 1, pp. 6–43, Mar. 2013.

[2] A. Alonso-González, C. López-Martínez, and P. Salembier, "Filtering and segmentation of polarimetric SAR data based on binary partition trees," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 2, pp. 593–605, Feb. 2012.

[3] F. Gao, X. Xue, J. Sun, J. Wang, and Y. Zhang, "A SAR image despeckling method based on two-dimensional S transform shrinkage," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 5, pp. 3025–3034, May 2016.

[4] F. Ma, F. Zhang, Q. Yin, D. Xiang, and Y. Zhou, "Fast SAR image segmentation with deep task-specific superpixel sampling and soft graph convolution," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5214116.

[5] F. Ma, F. Zhang, D. Xiang, Q. Yin, and Y. Zhou, "Fast task-specific region merging for SAR image segmentation," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5222316.

[6] P. Wang, X. Sun, W. Diao, and K. Fu, "FMSSD: Feature-merged single-shot detection for multiscale objects in large-scale remote sensing imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3377–3390, May 2020.

[7] X. Sun, P. Wang, C. Wang, Y. Liu, and K. Fu, "PBNet: Part-based convolutional neural network for complex composite object detection in remote sensing imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 173, pp. 50–65, 2021.

[8] Y. Zhou, F. Zhang, F. Ma, D. Xiang, and F. Zhang, "Small vessel detection based on adaptive dual-polarimetric feature fusion and sea–land segmentation in SAR images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 2519–2534, Mar. 2022.

[9] H. Li, V. A. Krylov, P. Fan, J. Zerubia, and W. J. Emery, "Unsupervised learning of generalized gamma mixture model with application in statistical modeling of high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 4, pp. 2153–2170, Apr. 2016.

[10] L. Wang, F. Zhang, W. Li, X. Xie, and W. Hu, "A method of SAR target recognition based on gabor filter and local texture feature extraction," *J. Radars*, vol. 4, no. 6, pp. 658–665, 2015.

[11] A. O. Knapskog, "Classification of ships in TerraSAR-X images based on 3D models and silhouette matching," in *Proc. Eur. Conf. Synthetic Aperture Radar*, 2010, pp. 1–4.

[12] R. Hummel, "Model-based ATR using synthetic aperture radar," in *Proc. Eur. Conf. Synthetic Aperture Radar*, 2000, pp. 856–861.

[13] J. A. O'Sullivan, M. D. DeVore, V. Kedia, and M. I. Miller, "SAR ATR performance using a conditionally gaussian model," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 1, pp. 91–108, Jan. 2001.

[14] S. K. Rogers, J. M. Colombi, C. E. Martin, and J. C. Gainey, "Neural networks for automatic target recognition," *Neural Netw.*, vol. 8, no. 7, pp. 1153–1184, 1995.

[15] Q. Zhao and J. C. Principe, "Support vector machines for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 37, no. 2, pp. 643–654, Apr. 2001.

[16] Y. Sun, Z. Liu, S. Todorovic, and J. Li, "Adaptive boosting for SAR automatic target recognition," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 43, no. 1, pp. 112–125, Jan. 2007.

[17] G. Dong and G. Kuang, "Classification on the monogenic scale space: Application to target recognition in SAR image," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2527–2539, Aug. 2015.

[18] S. Chen, H. Wang, F. Xu, and Y. Q. Jin, "Target classification using the deep convolutional networks for SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 54, no. 8, pp. 4806–4817, Aug. 2016.

[19] F. Zhang, C. Hu, Q. Yin, W. Li, H. Li, and W. Hong, "Multi-aspect-aware bidirectional LSTM networks for synthetic aperture radar target recognition," *IEEE Access*, vol. 5, pp. 26880–26891, 2017.

[20] F. Zhang, Y. Wang, J. Ni, Y. Zhou, and W. Hu, "SAR target small sample recognition based on CNN cascaded features and AdaBoost rotation forest," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 6, pp. 1008–1012, Jun. 2020.

[21] J. Ni, F. Zhang, Q. Yin, Y. Zhou, H.-C. Li, and W. Hong, "Random neighbor pixel-block-based deep recurrent learning for polarimetric SAR image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7557–7569, Sep. 2021.

[22] J. Geng, J. Fan, H. Wang, X. Ma, B. Li, and F. Chen, "High-resolution SAR image classification via deep convolutional autoencoders," *IEEE Geosci. Remote Sens. Lett.*, vol. 12, no. 11, pp. 2351–2355, Nov. 2015.

[23] Q. Lv, Y. Dou, X. Niu, J. Xu, and B. Li, "Classification of land cover based on deep belief networks using polarimetric RADARSAT-2 data," in *Proc. IEEE Geosci. Remote Sens. Symp.*, 2014, pp. 4679–4682.

[24] C. Liu, J. Yin, and J. Yang, "Application of deep learning to polarimetric SAR classification," in *Proc. IET Int. Radar Conf.*, 2015, pp. 1–4.

[25] L. Wang, X. Bai, R. Xue, and F. Zhou, "Few-shot SAR automatic target recognition based on Conv-BiLSTM prototypical network," *Neurocomputing*, vol. 443, no. 12, pp. 235–246, 2021.

[26] W. Fan, F. Zhou, Z. Zhang, X. Bai, and T. Tian, "Deceptive jamming template synthesis for SAR based on generative adversarial nets," *Signal Process.*, vol. 172, 2020, Art. no. 107528.

[27] T. Burns, "Moving and stationary target acquisition and recognition," in *Proc. DARPA Image Understanding Technol. Prog. Rev.*, 1996, pp. 265–281.

[28] C. Szegedy *et al.*, "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[29] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," 2017, *arXiv:1712.09665*.

[30] K. Eykholt *et al.*, "Robust physical-world attacks on deep learning visual classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 1625–1634.

[31] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14410–14430, 2018.

[32] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1378–1387.

[33] A. Arnab, O. Miksik, and P. H. Torr, "On the robustness of semantic segmentation models to adversarial attacks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 888–897.

[34] Y. Lin, Z. Hong, Y. Liao, M. Shi, M. Liu, and M. Sun, "Tactics of adversarial attack on deep reinforcement learning agents," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, *arXiv:1703.06748*.

[35] S. Hao, C. Jin, and L. Lin, "Adversarial robustness of deep convolutional neural network-based image recognition models: A review," *J. Radars*, vol. 10, no. 7, pp. 571–594, 2021.

[36] T. Huang, Q. Zhang, J. Liu, R. Hou, X. Wang, and Y. Li, "Adversarial attacks on deep-learning-based SAR image target recognition," *J. Netw. Comput. Appl.*, vol. 162, 2020, Art. no. 102632.

[37] W. Brendel, J. Rauber, and M. Bethge, "Decision-Based adversarial attacks: Reliable attacks against Black-Box machine learning models," 2017, *arXiv:1712.04248*.

[38] H. Li *et al.*, "Adversarial examples for CNN-based SAR image classification: An experience study," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1333–1347, Nov. 2020.

[39] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[40] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy*, 2017, pp. 39–57.

[41] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

Fan Zhang (Senior Member, IEEE) received the B.E. degree in communication engineering from the Civil Aviation University of China, Tianjin, China, in 2002, the M.S. degree in signal and information processing from Beihang University, Beijing, China, in 2005, and the Ph.D. degree in signal and information processing from Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2008.

He is currently a Full Professor of electronic and information engineering with the Beijing University of Chemical Technology, Beijing. His research interests include remote sensing image processing, high-performance computing, and artificial intelligence.

Dr. Zhang is an Associate Editor for IEEE Access and a Reviewer of the IEEE Transactions on Geoscience and Remote Sensing, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Geoscience and Remote Sensing Letters, and *Journal of Radars*.

Tianying Meng is currently working toward the master's degree, majoring in control engineering, with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China.

Her research interests include control engineering, image processing, and deep learning-based adversarial attack research.

Deliang Xiang (Member, IEEE) received the B.S. degree in remote sensing science and technology from Wuhan University, Wuhan, China, in 2010, the M.S. degree in photogrammetry and remote sensing from the National University of Defense Technology, Changsha, China, in 2012, and the Ph.D. degree in geoinformatics from the KTH Royal Institute of Technology, Stockholm, Sweden, in 2016.

Since 2020, he has been a Full Professor with Interdisciplinary Research Center for Artificial Intelligence, Beijing University of Chemical Technology, Beijing, China. His research interests include urban remote sensing, synthetic aperture radar (SAR)/polarimetric SAR image processing, artificial intelligence, and pattern recognition.

Dr. Xiang is a Reviewer for the *Remote Sensing of Environment*, *ISPRS Journal of Photogrammetry and Remote Sensing*, IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, and several other international journals in the remote sensing field. In 2019, he was the recipient of the Humboldt Research Fellowship.

Fei Ma (Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electronic and information engineering from the Beijing University of Aeronautics and Astronautics, Beijing, China, in 2013, 2016, and 2020, respectively.

He is currently with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, as an Associate Professor. His research interests include radar signal processing, image processing, machine learning, and target detection.

Xiaokun Sun received the B.S. degree in communication engineering from the Xidian University, Xi'an, China, in 2001, and the Ph.D. degree in electronic science and technology from the University of National Defense Technology, Changsha, China, in 2008.

She is currently with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing, China, as an Associate Research Fellow. Her main research interests include synthetic aperture radar (SAR) satellite calibration and quality assessment, and SAR satellite applications.

Yongsheng Zhou (Member, IEEE) received the B.E. degree in communication engineering from Beijing Information Science and Technology University, Beijing, China, in 2005, and the Ph.D. degree in signal and information processing from the Institute of Electronics, Chinese Academy of Sciences, Beijing, in 2010.

He was with Academy of Opto-Electronics, Chinese Academy of Sciences, in 2010 and 2019, and currently a Professor of electronic and information engineering with the College of Information Science and Technology, Beijing University of Chemical Technology, Beijing. His research interests include target detection and recognition from microwave remotely sensed image, digital signal, and image processing.