Scattering Model Guided Adversarial Examples for SAR Target Recognition: Attack and Defense

Bowen Peng®, Bo Peng®, Jie Zhou, Jianyue Xie, and Li Liu®, Senior Member, IEEE

Abstract—Deep neural network (DNN)-based synthetic aperture radar (SAR) automatic target recognition (ATR) systems have been shown to be highly vulnerable to adversarial perturbations that are deliberately designed yet almost imperceptible but can bias DNN inference when added to targeted objects. This leads to serious safety concerns when applying DNNs to high-stakes SAR ATR applications. Therefore, enhancing the adversarial robustness of DNNs is essential for applying DNNs to modern real-world SAR ATR systems. Toward building more robust DNN-based SAR ATR models, this article explores the domain knowledge of the SAR imaging process and proposes a novel scattering model guided adversarial attack (SMGAA) algorithm, which can generate adversarial perturbations in the form of electromagnetic scattering response (called adversarial scatterers). The proposed SMGAA consists of two parts: 1) a parametric scattering model and corresponding imaging method and 2) a customized gradient-based optimization algorithm. First, we introduce the effective attributed scattering center model (ASCM) and a general imaging method to describe the scattering behavior of typical geometric structures in the SAR imaging process. By further devising several strategies to take the domain knowledge of SAR target images into account and relax the greedy search procedure, the proposed method does not need to be prudentially fine-tuned and can efficiently find the effective ASCM parameters to fool the SAR classifiers and facilitate the robust model training. Comprehensive evaluations on the moving and stationary target acquisition and recognition (MSTAR) dataset show that the adversarial scatterers generated by SMGAA are more robust to perturbations and transformations in the SAR processing chain than the currently studied attacks and are effective to construct a defensive model against the malicious scatterers.

Index Terms—Adversarial attack, automatic target recognition (ATR), deep neural networks (DNNs), physical attack, robustness, synthetic aperture radar (SAR).

I. INTRODUCTION

SYNTHETIC aperture radar automatic target recognition (SAR ATR) is a fundamental and challenging problem in SAR image interpretation and has various applications in both civil and military fields [1], [2], [3]. Therefore, for the past several decades, SAR ATR has received significant attention. Recently, deep neural networks (DNNs) have brought considerable progress in remote sensing image analysis [4], [5], [6], [7], [8], [9] but are largely limited to the analysis of optical imagery. Certainly, DNNs have been introduced in SAR image interpretation tasks, including SAR ATR [10], [11], [12], and have also achieved promising progress. It is reasonable to expect that the huge potential of DNNs for SAR image interpretation will be unlocked in the future [4].

Due to the high-stakes nature of military and homeland security applications, SAR ATR techniques need to be highly reliable and safe, in addition to being accurate. However, there are many challenges leading to safety concerns of DNNs when used in many security-critical applications, such as SAR ATR. A critical problem is that DNNs are shown to be highly vulnerable to adversarial examples, i.e., images added with intentionally designed yet imperceptible perturbations that are usually unnoticed by the human visual system but can fool DNNs [13]. For instance, aircraft could be maliciously fooled as birds via adversarial attack. There has been a great amount of work finding [14], [15], understanding [16], [17], and defending digital and physical adversarial examples1 [18], [19]. There exists an attack–defense race in the deep learning community, and it has been demonstrated that guaranteeing adversarial robustness is a big challenge [20].

Research on adversarial attacking and defending DNN-based systems has also drawn significant attention in the remote sensing community. However, most of existing studies along this thread are also limited to optical images [20], [21], [22], [23] with only a few initial attempts on SAR images [24], [25], [26]. For example, Chen et al. [24] and Li et al. [25] carried out a comprehensive evaluation of the adversarial vulnerability in the SAR scene classification and target recognition tasks utilizing the existing digital attack algorithms, such as the fast gradient sign method (FGSM) [27], DeepFool [28], and Carlini and Wagner (C&W) [29]. More recently, Du et al. [26] designed an accelerated C&W algorithm for SAR ATR to pursue a balance between time consumption and attack ability. In these studies, DNN-based SAR ATR has been reported as extremely vulnerable in the face of extremely small perturbations.

However, to the best of our knowledge, the aforementioned works just directly transfer adversarial attacks against optical imagery to SAR imagery and have serious limitations. On the practical side of the research, digital attacks need to be realized physically. Then, the question of the applicability of the

1With the online API, digital adversarial examples can be directly fed into the victim DNN. In contrast, physical attacks design the adversarial examples that remain highly aggressive after being actively imaged by the victim’s own sensing system.
adversarial perturbations against physical SAR ATR systems should be given special attention as it is significantly different from generating adversarial perturbations for natural images. Physical attacks against optical imagery have mostly been studied in the problems of facial recognition, natural object recognition, and so on, where the natural images are usually captured at small distances, such as within the sensing range of a camera in an autonomous car or an intelligent video surveillance system [14], [30]. By contrast, the applicability of physical attacks against SAR ATR systems is challenged by many factors, such as prior information on the targeted SAR ATR systems (such as platform, resolution, or imaging mode), the domain knowledge of SAR image acquisition processes, such as the geometric scattering mechanism, where perturbations can be realized in the signal processing chain, and how to make the adversarial perturbations robust to real-world perturbations inherent in the sensing chain, such as imaging conditions, imaging viewpoint, and postprocessing. For instance, as contrasted in Fig. 1, the scattering responses of the SAR target are intrinsically different from the currently studied perturbations and exhibit sidelobe diffusion due to the limitations of the SAR imaging mechanism [31]. For another instance, perturbations with very small bound are sensitive to the attributed scattering center model (ASCM) [37], a well-defined scattering model that provides concise and physically relevant attributes for both localized and distributed scattering mechanisms. Thus, it is difficult to study physical adversarial threats. Moreover, it can directly serve the defense against malicious scatterers with limited additional computational cost.

In this article, a novel electromagnetic-relevant attack framework, namely, scattering model-guided adversarial attack (SMGAA), is proposed to better highlight and further prevent the adversarial risks in real-world circumstances. The SMGAA aims at extending the optimization chain between the perturbation image and the model’s prediction of the electromagnetic objects and eventually generating more feasible and robust perturbations. Fig. 1(b) demonstrates an example of SMGAA-generated perturbations, which makes a well-trained AConvNet [36] misclassify the rocket launcher as a tank with confidence score up to 97.43%. These perturbations are strictly generated by the parametric scattering model at the signal level, carry clear electromagnetic attributes, and, thus, are termed adversarial scatterers in this article. The illustration of the proposed framework can be found in Fig. 2. With the attributed scattering center model (ASCM) [37], a well-defined scattering model that provides concise and physically relevant attributes for both localized and distributed scattering mechanisms, we devise a customized optimization process to search for the most effective adversarial scatterers in a softened greedy manner. First, the candidate scatterers will be initialized at the target and shadow region since it contains most of the structural information of the SAR target. Second, with the randomness and progressive adaptation provided by a Gaussian stepsize generator, the optimization can escape the local optima and utilize the memorization of previous steps to be more efficient and effective. The failed updates will also be adopted with a certain probability during the process to alleviate the overfitting of greedy search. Furthermore, inspired by the random restart strategy and population-based algorithms, the SMGAA is shown to be robust to random noise, as well as Gaussian and median filters.

The remainder of this article is organized as follows.

1) We propose a novel SMGAA framework for DNN-based SAR ATR, which leverages the geometric scattering mechanism to generate parametric adversarial scatterers with well-defined physical attributes. It provides high feasibility for studying physical adversarial threats. Moreover, it can directly serve the defense against malicious scatterers with limited additional computational cost.

2) We devise a customized optimization process to exploit the most effective ASCM parameters to fool the ATR models and facilitate robust model training. The optimization takes the distribution of SAR target images into account and softens the greedy search process, which is highly prone to stick to local optima.

3) We demonstrate that the SMGAA is generalizable across a variety of DNNs by comprehensive evaluations on eight cutting-edge DNN structures both in computer vision and SAR ATR applications, for example, 53.7%, 77.1%, and 87.2% average fooling rates can be, respectively, achieved by just one, two, and three adversarial scatterers. Compared with the currently studied attack algorithms, the SMGAA is shown to be robust to random noise, as well as Gaussian and median filters.

4) We demonstrate the necessity and efficacy of the SMGAA in defending malicious scatterers. The limitedly exploited SMGAA can bring robust accuracy improvements above 40% on average, distrust to the induced category, and consistent discriminative evidence to perturbed images, by evaluations on AConvNet. By contrast, current digital attack-based defense is shown to be almost ineffective against malicious scatterers.
Adversarial Example
Target Model
Prediction
Optimize with Customized Strategies
\( L(x + I(\zeta, \Theta), y^p) \)

\( \Delta \)

Electromagnetic Scattering Model

SAR para. \( \zeta \)
Scattering para. \( \Theta_N \)

Adversarial Scatterer

(a)

(b)

Fig. 2. Framework of the SMGAA and an example correlates the digital adversarial scatter with the physical scatterer deployment. (a) Pipeline of the proposed SMGAA. In this instance, a trihedral adversarial scatterer was generated for attacking the T62 tank. (b) Example of a physical attack implementation on the T62 tank guided by the SMGAA.

where function \( D(\cdot) \) and value \( \epsilon \) serve as a distance constraint to guarantee that the perturbation \( \delta \) is imperceptible to the human visual system. There are several metrics have been studied, such as \( \ell_p \) norm, geodesics in data manifold [40], the Wasserstein distance [41], or perceptibility metrics [42]. In the remainder of this article, we focus on the most commonly studied \( \ell_p \) attacks of which the perturbations are restricted in the \( \ell_p \) norm ball with radius \( \epsilon \), i.e., \( \|\delta\|_p \leq \epsilon \), and the \( \ell_p \) norm \( \|\cdot\|_p \) is defined as

\[
\|\delta\|_p = \left( \sum_{0 \leq i < h} \sum_{0 \leq j < w} |\delta_{ij}|^p \right)^{\frac{1}{p}} \quad \text{for } p = \infty \\
\|\delta\|_p = \max_{0 \leq i < h} \sum_{0 \leq j < w} (\delta_{ij} \neq 0) \quad \text{for } p = 0 \\
\|\delta\|_p = \sum_{0 \leq i < h} \sum_{0 \leq j < w} \|\delta_{ij}\| \quad \text{for } 0 < p < \infty
\]

where \( \|\cdot\| \) represents the indicator function. Generally, \( p \) is selected as 0 for sparsity (number of pixels to perturb), 2 for stealthiness (Euclidean distance), and \( \infty \) for efficiency (maximum absolute value for every pixel). When \( p \neq 0 \), the classification loss can be efficiently magnified via a normal gradient ascent process

\[
x_{i+1}^\text{adv} = x_i^\text{adv} + \epsilon_i \frac{\nabla_i L(x_i^\text{adv}, y^{gt})}{\|\nabla_i L(x_i^\text{adv}, y^{gt})\|_p}
\]

where \( L(\cdot) \) calculates the loss of classifier \( F \) with respect to the input and its ground-truth label. In the case of \( \ell_\infty \) attacks, such as FGSM, the basic iterative method (BIM) [43], and the projected gradient descent (PGD) [18], the normalization can be simply replaced by the sign function. In addition to finding perturbations with specific magnitude, some attacks are designed more delicately and pursue minimal perturbations. For instance, C&W [29] proposed several customized loss functions to jointly optimize the \( D(x, x + \delta) \) and the classification loss. DeepFool [28] analytically derives the minimal perturbation that pulls the sample across the decision hyperplane with the linear assumption of DNNs. The aforementioned attacks generate full-scale perturbations for the DNN inputs. \( \ell_0 \) attacks, by contrast, carefully perturb the input with limited pixels. There are currently sparse attacks based on variants of gradient-based methods integrating the \( \ell_0 \) constraint, such as the PGD (\( \ell_0 \)) [44] and Jacobian-based saliency map attack (JSMA) [45], or search-based methods, such as the Sparse-RS [35]. Fig. 1(a) exhibits the adversarial examples generated by different values of \( p \) for a SAR target image.

Up until now, the state-of-the-art (SOTA) DNNs are still extremely vulnerable to adversarial examples [13], [46], highlighting the importance of adversarial robustness on the security-critical tasks, such as autonomous driving and homeland security applications. There are many hypotheses tried to explain this nonrobust behavior. Szegedy et al. [39] proposed that the adversarial examples represent the low-probability pockets in the high-dimensional manifold. It was also proposed that the existence of such examples is due to the strong linear nature of DNNs [27]. Tanay and Griffin [47] suggested that this odd behavior may be caused by the tilting decision boundary in high-dimensional space. Although being a security threat, the existence of adversarial examples exhibits

works of this study are introduced. In Section III, the ASCM and the correlated image formation process are elaborated. Our proposed adversarial scatterer generation algorithm is described in Section IV. Then, the experimental results and analysis are reported in Section V, including the attack performance and application of defense, comparison to the \( \ell_p \) attacks, and ablation study. In the end, Section VI summarizes this article and designs future work.

II. BACKGROUND AND RELATED WORKS

In order to better comprehend the article and the related works in SAR ATR, we introduce the background linked to the adversarial attack

A. Adversarial Attack in General Object Recognition

1) Adversarial Attack: For the image classification task, adversarial attack aims at finding a visually imperceptible perturbation \( \delta \) with respect to an input image \( x \in \mathbb{R}^{h \times w} \) and its ground-truth label \( y^{gt} \in \mathbb{R}^I \) to mislead the classifier \( F: \mathbb{R}^{h \times w} \rightarrow \mathbb{R}^I \) into outputting a wrong prediction \( F(x + \delta) \neq y^{gt} \) [27], [38], [39]. This motivation can be described as the following problem:

\[
F(x + \delta) \neq y^{gt} \quad \text{s.t.} \quad D(x, x + \delta) \leq \epsilon
\]
important value for enhancing the understanding [48], robustness, and trustworthiness of DNN-based high-dimensional classification [13].

2) Adversarial Attack in Physical Setting: Many works in the field of optical imaging are devoted to revealing and describing real-world threats. The access to physically realizable perturbation mainly lies in three paths: 1) generating the localized perceivable patch, e.g., pasting an adversarial sticker on the traffic sign [49], [50] or wearing a customized glass coated with adversarial pattern [51]; 2) using infrared [30] and laser beam [14] or exploiting the radiometric effect, such as rolling shutter [52], to deploy perturbations that are invisible to the human visual system; and 3) constructing differentiable renderer [53] to guide physical attack, which modifies the physical properties, such as the material, light, and geometry. Physical attacks in the optical imaging setting provide many insightful ideas; however, they cannot be directly utilized in SAR applications due to the different imaging mechanisms.

B. Adversarial Attack in SAR ATR

In contrast to the considerable studies in computer vision, the research on the adversarial robustness of DNN-based SAR ATR is quite restricted. Typically, the study on the adversarial attack in SAR ATR has gone through two steps: 1) exploration and 2) optimization. Primeval research has focused on verifying the vulnerability of SAR models [24], [25], [54] by using attack algorithms borrowed from the computer vision community. Works in this stage constructed comprehensive evaluations to reveal the adversarial vulnerability and also concluded some fundamental characteristics. For example, Chen et al. [24] concluded that the adversarial examples transfer well across disparate SAR ATR models, and Li et al. [25] suggested that the vulnerability of models is proportional to their complexity. In addition, it has been demonstrated that the moving and stationary target acquisition and recognition (MSTAR) targets have the intriguing trait of having misclassified classes of various models and attacks that are highly clustered. Afterward, specific requirements for fooling SAR ATR continuously motivate scholars to optimize the existing attack algorithms. For real-time perturbation generation, Du et al. [26] accelerated the C&W attack by introducing the generative models to learn the mapping from the original images to the adversarial examples. Peng et al. [55] proposed to enhance the attacks by manipulating the speckle-noise pattern for a more noncooperative condition. However, current research concentrated on the \( \ell_p \) attack framework, which generates strictly pixelwise perturbations and is less relevant to SAR imaging. As most related, the initial application of the parametric model in designing perturbation was of Dang et al. [56] published in 2021, where the authors reported a 42% fooling rate against a support vector machine (SVM)-based binary classification task.

III. IMAGE FORMATION OF THE ASCM

Based on high-frequency approximations [57], the total backscatter can be decomposed into responses of individual scatterers. Developed from the geometric theory of diffraction (GTD) [57], many parametric models were proposed for modeling the individual scatterers, including the point scattering model (PSM) [58], the damped exponential model (DEM) [59], ASCM [37], and so on. Among them, the ASCM characterizes both localized and distributed scattering mechanisms on frequency and aspect dependence, as well as the physical attributes (such as the structure and size), provides a concise and physically relevant description for scatterers, is compatible with most scattering models, and has shown great effectiveness in the SOTA SAR ATR techniques [11], [60], [61].

In this section, the ASCM and its image formation process are introduced.

A. ASCM

According to GTD [57], [62], the summation of the responses from individual scatterers can well approximate the total backscattered field at the high-frequency region. Then, the ASCM describes the total scattered field as a function of frequency \( f \) and aspect angle \( \phi \) as [37]

\[
E(f, \phi; \Theta_N) = \sum_{i=1}^{N} E_i(f, \phi; \theta_i) \tag{4}
\]

where \( \Theta_N = \{ \theta_i \mid \theta_i = [A_i, x_i, y_i, \alpha_i, \gamma_i, L_i, \phi_i], 1 \leq i \leq N \} \) is the parameter set of \( N \) individual scatterers and

\[
E_i(f, \phi; \theta_i) = A_i \cdot \left( \frac{f}{f_c} \right)^{\alpha_i} \cdot \exp\left( -\frac{4\pi f}{c} (x_i \cos \phi + y_i \sin \phi) \right)
\cdot \sin(\frac{2\pi f}{c} L_i \sin(\phi - \phi_i)) \cdot \exp(-2\pi f \gamma_i \sin \phi). \tag{5}
\]

Herein, \( f_c \) is the center frequency of the radar wave and \( c \) is the velocity of light. For the \( i \)-th scatterer, \( A_i \) is amplitude, and \( x_i \) and \( y_i \) are range and cross-range locations. \( \alpha \in [-1, -0.5, 0, 0.5, 1] \) models the frequency dependence. For a localized scattering mechanism that has localized returns in the SAR image, \( L = \phi = 0 \), and \( \gamma \) describes the aspect dependence. For distributed scattering mechanism, whose responses span several image pixels along cross-range, \( \gamma = 0 \) and \( L, \phi \) respectively, model the length and orientation angle. Table I lists the most common scattering structures in the real world, demonstrating the generalization ability of the ASCM. It can be seen that different types of geometric scattering structures can be distinguished by frequency dependence and length. For example, the dihedral degrades to trihedral when \( L = 0 \). The parameter setting is important to ensure the physical feasibility, e.g., too large frequency response may not be achievable with existing materials or normal structures. Therefore, the parameter sensitivity will be further analyzed in the experiments.

B. Image Formation

The response data calculated by (5) need to be further coped to form the SAR image. Fig. 3 provides a general SAR image formation process of ASCM that was given in [63]. In the frequency domain, the polar format data are uniformly sampled in frequency \( f \in [f_c - B/2, f_c + B/2] \) and aspect angle \( \phi \in [-\phi_m/2, \phi_m/2] \), where \( B \) is the bandwidth of radar wave and \( \phi_m \) is the aperture accumulation angle. Using \( E_{m \times n}(f, \phi) \) to represent the total response in polar plane,
TABLE I
GEOMETRIC SCATTERING TYPES DIFFERENTIATED
BY FREQUENCY AND ASPECT DEPENDENCE

| Geometric scattering type | $\alpha$ | $\gamma$ | $L$ | $\bar{\phi}$ |
|---------------------------|---------|---------|-----|-------------|
| Trihedral                 | 1       |         |     |             |
| Localized                 | 0.5     | $>0$    | 0   | 0           |
| Top Hat (TH)              | 0       | $>0$    | 0   | $\bar{\phi}$ |
| Sphere                    | 0       |         |     |             |
| Corner Diffraction (CD)   | 1       |         |     |             |
| Distributed               | 0.5     | $>0$    | $\neq$ | 0           |
| Dihedral                  |         |         |     |             |
| Edge Broadside (EB)       | 0       |         |     |             |
| Edge Diffraction (ED)     | -0.5    |         |     |             |

which is formed by sampling $m$ and $n$ points in frequency and aspect angle sections, respectively, first, $E_{m\times n}(f_x, f_y)$ is obtained by uniformly resampling $E_{m\times n}(f, \phi)$ to the Cartesian plane

$$f_c - B/2 \leq f_x = f \cos(\phi) \leq f_c + B/2$$

$$-f_c \cdot \sin(\phi_m/2) \leq f_y = f \sin(\phi) \leq f_c \cdot \sin(\phi_m/2).$$

To improve the quality of the resultant image, $E_{m\times n}(f_x, f_y)$ will be multiplied by a window function $W(f_x, f_y)$ and then be zero-padded to $m^* \times n^*$ samples. Finally, the SAR image $I_{m^*\times n^*}(x, y)$ is obtained by 2-D inverse discrete Fourier transform (2-D-IDFT). Through this process, the pixel spacing of range and cross-range is given by

$$p_x = \frac{c}{2B} \cdot \eta_x, \quad p_y = \frac{c}{f_c \cdot 4 \sin(\phi_m/2)} \cdot \eta_y$$

where $\eta_x = (m - 1/m^* - 1)$ and $\eta_y = (n - 1/n^* - 1)$ are the zero-padding effects.

C. Parameter Normalization

In (5), $x$, $y$, and $L$ model the range and cross-range locations, and length of the scatterer in meters, and $\bar{\phi}$ measures the orientation angle in degrees. In addition, the aspect dependence factor $\gamma$ is smaller than other parameters by almost ten orders of magnitude, which is not conducive to being updated. For better numerical properties and to provide a pixel-level description of the scatterers’ location and length, the model parameters are normalized using the following transformations [63, 64]:

$$(x_p, y_p, L_p) = \left(\frac{x}{p_x}, \frac{y}{p_y}, \frac{L}{p_y}\right)$$

$$\gamma_p = \gamma \cdot 2\pi f_c, \quad \bar{\phi}_p = \frac{\bar{\phi}}{\phi_m/2}.$$ (8)

In this way, $x_p$ and $y_p$ are the pixel location and $L_p$ models the length in pixel level, which is consistent with the normal pixel index approach. The orientation index $\bar{\phi}_p$, then lies into $[-1, 1]$, and the normalized ASCM can be tidied as follows:

$$E(f_x, f_y; \theta) = A \cdot \left(\frac{j \sqrt{(f_x^2 + f_y^2)}}{f_c} \cdot \exp\left(-\frac{f_x}{f_c} \gamma_p\right)
\cdot \text{sinc}\left(\frac{\pi \sqrt{(f_x^2 + f_y^2)}}{f_c \cdot 2 \sin(\phi_m/2)}\right)\right.
\cdot \exp\left(-j \frac{4\pi}{c} \left(p_x f_x + p_y f_y\right)\right).$$ \hspace{1cm} (9)

To summarize at the end of this section, for a given parameter set $\Theta_N = \{\theta_i\} \setminus \theta = \{A_i, x_{pi}, y_{pi}, \bar{\phi}_i, L_{pi}, \phi_{pii}\}, 1 \leq i \leq N$ of the normalized ASCM, and the SAR imaging parameters $\zeta = \{f_c, B, \phi_m, m, n, n^*, W(\cdot)\}$, the corresponding SAR image $I_{m^*\times n^*}(x, y)$ can be obtained using the imaging process illustrated in Fig. 3.

IV. CRAFTING ADVERSARIAL SCATTERERS

In this section, we detail the second part of the proposed SMGAA, including the optimization objective of the normalized ASCM parameters and the customized strategies.

A. Objective

Recall the motivation of adversarial attack described in (1) and our motivation to generate adversarial scatterers. The perturbation $I(\zeta, \Theta_N)$ is controlled by the parameter set of normalized ASCM. Therefore, the goal is to optimize the parameter set $\Theta_N$ that makes the adversarial example $x^{\text{adv}} = x + I(\zeta, \Theta_N)$ classified by the SAR target classifier at a low cost, i.e., by adding very few adversarial scatterers. In addition to limiting the number of scatterers, a pair of vectors $[\theta_{\text{max}}, \theta_{\text{min}}]$ is selected to restrict the parameter set in a proper range. Then, adversarial scatterers could be calculated by maximizing the classification loss

$$\arg \max_{\Theta_N} L(x + I(\zeta, \Theta_N), y^{\text{gt}})$$

s.t. $\theta_{\text{min}} \leq \Theta_N \leq \theta_{\text{max}}$

$$x + I(\zeta, \Theta_N) \in \mathbb{R}^{h \times w}$$

$$N \leq \epsilon.$$ (10)

Here, the term $x + I(\zeta, \Theta_N) \in \mathbb{R}^{h \times w}$ guarantees that the adversarial example falls within the reasonable input interval for the model.

$^{2}$This image-level fusion is equivalent to signal-level superposition due to the linear nature of the discrete Fourier transform.
B. Method

Generally, we can obtain approximate solutions of (10) simply using gradient-based methods, such as BIM [43] and PGD [18], since the imaging function $I(\cdot)$ can be formulated differentiable. For a given scattering type, ASCM allows us to manipulate the scatterer parameters (attributes), such as amplitude, location, size, and aspect angle. However, adjusting the images from these many modalities requires taking into account the impact of each parameter on the loss surface, which is a highly nonlinear and nonconvex task. Meanwhile, the optimization is prone to be stuck into local optima. To efficiently find adversarial scatterers, three simple yet effective strategies are proposed to exploit the gradient. Algorithm 1 elaborates on the suggested procedure, which will be expanded upon in the following.

1) Parameter Initialization: To assure the feasibility of the generated scatterers, we choose the scattering types in Table I to initialize the ASCM parameters, and the corresponding vectors are defined to guarantee the proper update. Specifically, when initializing the candidate scatterers, the scattering types are determined with random types from Table I. Meanwhile, a corresponding vector $\mathbf{r}$ is generated to freeze the type-related parameters during the optimization, as such the frequency dependence and length (for localized scatterers). The influence parameters during the optimization, such as the frequency $v$, are determined with random types from Table I. Meanwhile, when initializing the candidate scatterers, the scattering types are assigned to the initial scatterers. The other parameters are simply using gradient-based methods, such as BIM [43] and PGD [18], since the imaging function $I(\cdot)$ can be formulated differentiable. For a given scattering type, ASCM allows us to manipulate the scatterer parameters (attributes), such as amplitude, location, size, and aspect angle. However, adjusting the images from these many modalities requires taking into account the impact of each parameter on the loss surface, which is a highly nonlinear and nonconvex task. Meanwhile, the optimization is prone to be stuck into local optima. To efficiently find adversarial scatterers, three simple yet effective strategies are proposed to exploit the gradient. Algorithm 1 elaborates on the suggested procedure, which will be expanded upon in the following.

2) Stepsize Adaption: The iterative gradient ascent marches by softened greedy search steps in which the fail update would also be randomly adopted to escape from the local optima

$$\Delta \mathbf{\Theta} = s \cdot \text{sign}(\nabla \mathbf{\Theta} \mathcal{L})$$

$$\mathbf{\Theta}^*_{N} = \text{Clip}(\mathbf{\Theta}_{N} + \mathbf{r} \cdot \Delta \mathbf{\Theta}, \mathbf{\Theta}_{\min}, \mathbf{\Theta}_{\max})$$

$$\mathbf{\Theta}_{N} = \begin{cases} 
\mathbf{\Theta}_{N}^* , & \text{if loss increases or randu}(1) > 0.5 \\
\mathbf{\Theta}_{N}, & \text{otherwise} 
\end{cases}$$

where $\mathcal{L} = \mathcal{L}(x + I(\mathbf{\zeta}, \mathbf{\Theta}_{N}), y^{\text{gt}})$ denotes the cross-entropy loss, sign(·) determines in which direction the parameter set should be updated, and $s$ is the stepsize. Function Clip(·, a, b) is applied to clip the input to range $[a, b]$. A Gaussian stepsize generator $\mathcal{N}(S, \sigma)$ is designed to introduce randomness and capability of adaption during the optimization process. In each iteration, the stepsize $s$ is generated by

$$s = \text{randn}(S, \sigma)$$

The initial mean value $S_0$ and standard deviation $\sigma$ are predefined, and the mean value adapts across the iterations

$$S_{i+1} = \begin{cases} 
\lambda \cdot S_i + (1 - \lambda) \cdot |\Delta \mathbf{\Theta}| , & \text{if loss increases} \\
S_i, & \text{otherwise} 
\end{cases}$$

where factor $\lambda$ is the balance factor and learning rate to relax potential overfitting of the adaption and is set to 0.5 in the experiments. By utilizing this adaption strategy, a roughly defined stepsize can adapt diverse initialization and optimization processes, and the randomness brought by the generator can help escape from the local optimum.

3) Population-Style Generation Process: A population-style generation process is applied to accelerate the attack and fit the batch training process. Specifically, a batch of $B_N$ parameter sets, i.e., a $B_N \times N \times 7$ tensor, would be initialized and optimized at once. Such a process reduces time consumption by increasing memory footprint. Notice that there has no information interaction between each individual since the generation process is considered mainly related to the initialization. During the updates, we control the amplitude of each adversarial scatterer lower than 1 (for normalized images), , i.e., the response of the adversarial scatterers is not allowed stronger than the target. Practically, when the amplitude of each scatterer exceeds 1, the parameter $A$ would

---

Algorithm 1 SMGAA

**Input:** Classifier $F$; Image $x$, corresponding category $y^{\text{gt}}$, and the target&shadow mask $m$; Imaging parameters $\mathbf{\zeta}$; Available range for ASC parameters $[\mathbf{\Theta}_{\min}, \mathbf{\Theta}_{\max}]$; Maximum iteration $n_{\text{max}}$; Batch capacity $B_N$; Confidence threshold $v_{\text{th}}$; Normal distribution $\mathcal{N}(S, \sigma)$; Learning rate $\lambda$; Number of scatterers $N$;

**Output:** A parameter set $\mathbf{\Theta}_{N}$;

1: for $j \leftarrow 1$ to $B_N$ and $i \leftarrow 1$ to $N$ do
2: Initialize $a_i^j$ and corresponding adjustable vector $\mathbf{r}_i^j$;
3: Initialize $x_i^j, y_i^j$ with random coordinate pairs in $m$;
4: Initialize other parameters with randu($\mathbf{\Theta}_{\min}, \mathbf{\Theta}_{\max}$);
5: end for
6: $n \leftarrow 1$
7: $v \leftarrow C_F(x + I(\mathbf{\zeta}, \mathbf{\Theta}_{N}), y^{\text{gt}})$
8: $\mathcal{L} \leftarrow \mathcal{L}(x + I(\mathbf{\zeta}, \mathbf{\Theta}_{N}), y^{\text{gt}})$
9: while $n < n_{\text{max}}$ or min($v$) $< v_{\text{th}}$ do
10: $n \leftarrow n + 1$
11: $\Delta \mathbf{\Theta} \leftarrow \text{randn}(S, \sigma) \cdot \text{sign}(\nabla \mathbf{\Theta} \mathcal{L})$
12: $\mathbf{\Theta}_i = \text{Clip}(\mathbf{\Theta}_{N} + \mathbf{r} \cdot \Delta \mathbf{\Theta}, \mathbf{\Theta}_{\min}, \mathbf{\Theta}_{\max})$
13: for $i, j \in \{\text{max}(I(\mathbf{\zeta}, \mathbf{\Theta}_i^j)) \geq 1\}$ do
14: $A_{i}^{j} \leftarrow A_{i}^{j} \cdot (\lambda \cdot S_i + (1 - \lambda) \cdot |\Delta \mathbf{\Theta}|)$
15: end for
16: $x^{\text{adv}} \leftarrow \text{Clip}(x + I(\mathbf{\zeta}, \mathbf{\Theta}_i), 0, 1)$
17: $L_a \leftarrow \mathcal{L}(x^{\text{adv}}, y^{\text{gt}})$
18: for $k \leftarrow 1$ to $B_N$ do
19: if $L_a > L_k$ or randu$(1) > 0.5$ then
20: $\mathbf{\Theta}_k^i \leftarrow \mathbf{\Theta}_k^i$
21: if $L_k > L^k$ then
22: $S_k \leftarrow \lambda \cdot S_k + (1 - \lambda) \cdot |\Delta \mathbf{\Theta}|$
23: end if
24: end if
25: end for
26: $\mathcal{L} \leftarrow L_a$
27: $v \leftarrow C_F(x + I(\mathbf{\zeta}, \mathbf{\Theta}_{N}), y^{\text{gt}})$
28: idx $\leftarrow \text{argmin}(v)$
29: end while
30: Return $\mathbf{\Theta}_{N}^{\text{adv}}$
be divided by $A + \text{randu}(1)$ to project the response back into a lower level. Meanwhile, the adversarial examples will be clipped to $[0, 1]$ before being fed into the model to satisfy the input range of DNNs. The clip operation mainly works when the adversarial scatterer is located at the strong response of the target, which is visualized in Fig. 4. It can be observed that the scatterer response losses after clipping, leading to very limited additive perturbation to the original image. Nonetheless, this mild perturbation can be implemented by modifying the target, such as by appending coatings that strengthen the scattering. Otherwise, the whole target will be suppressed by the local strong response across the maximum normalization just as shown in Fig. 4(d), which is investigated in Section V-D4. Attack will stop when reaching the max iteration number $n_{\text{max}}$, and early stopping will be triggered when the confidence score of ground-truth label $C_F(x + I(\xi, \Theta_N), y^{\text{gt}})$ is lower than a threshold $\theta_{\text{th}}$.

V. EXPERIMENTS

Experiments on the publicly accessible dataset were conducted to evaluate the proposed method. Section V-A summarizes the dataset and experimental setup. Section V-B demonstrates the performance, analysis, and defense application of the SMGAA. In Section V-C, the SMGAA is compared with the currently studied digital attacks in the SAR community and the typical physical attack in the optical object recognition tasks. In Section V-D, we investigate the efficacy of the proposed strategies and the parameter sensitivity of the generation algorithm. The discussion on the previously reported attack selectivity of the MSTAR dataset is given in Section V-E.

A. Dataset and Experimental Setup

1) Dataset: The MSTAR Program, with funding from the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL), released a dataset [65] for the public study of SAR ATR. The measured SAR data were collected using the Sandia National Laboratory SAR sensor platform with the X-band imaging ability in 1-ft resolution. The resulted dataset comprises SAR imagery with 360° articulation with a 1° spacing, an image size of 128² pixels, and several depression angles. The MSTAR dataset contains ten types of ground vehicle targets (rocket launcher: 2S1; armored personnel carrier: BMP2, BRDM2, BTR70, and BTR60; bulldozer: D7; tank: T62 and T72; truck: ZIL131; and air defense unit: ZSU234), and the optical and SAR images of these targets are listed in Fig. 5.

![Fig. 4. Influence of the clip operation to the maximum-normalization process: (a) target image, (b) 20° central patch of (a) (normalized magnitude is 1), (c) point scatterer response with a normalized magnitude of 0.18, (d) fused image without clipping (the whole image is suppressed by the local strong superposition), and (e) fused image with clipping.](image)

![Fig. 5. Examples of ten types of ground vehicle targets in the MSTAR dataset: (Top) optical images and (Bottom) corresponding SAR images.](image)

There are four commonly studied partitions, including one standard operating condition (SOC) and three extended operating conditions (EOCs), respectively, for the classification tasks of typical condition (ten targets), large depression variation (four targets), configuration variants (five variants of one target), and version variants (seven variants of two targets). The SOC subset was selected to carry out evaluations since this article mainly concentrates on category confusion. In the SOC subset, the images that had been collected at 17° depression angle were used for training and 15° were used for testing. More specifics about the SOC subset are concluded in Table II.

2) DNN Structures: In the experiments, eight typical DNN structures were involved to access comprehensive results. Among them, AlexNet³ [66], VGG11³ [67], ResNet50³ [68], and DenseNet121³ [69] are generally utilized as feature extraction backbones and have achieved eye-catching performance in various domains. In addition, MobileNetV2⁴ [70], AConvNet⁴ [36], ShuffleNetV2⁵ [71], and SqueezeNet⁶ [72] are lightweight designs of DNNs that reduce the number of

³https://github.com/pytorch/vision/tree/master/torchvision/models
⁴We build the AConvNet according to the original paper with Pytorch framework, and the authors’ released Caffe code can be found at https://github.com/fudanxu/MSTAR-AConvNet

| Class  | Serial Number | Training Set (17°) | Test Set (15°) |
|--------|---------------|--------------------|----------------|
| 2S1    | b01           | 299                | 274            |
| BMP2   | 9563          | 233                | 195            |
| BRDM2  | B71           | 298                | 274            |
| BTR70  | c71           | 233                | 196            |
| BTR60  | k10y7532      | 256                | 195            |
| D7     | 92v13015      | 299                | 274            |
| T72    | 132           | 232                | 196            |
| T62    | A51           | 299                | 273            |
| ZIL131 | B12           | 299                | 274            |
| ZSU234 | d08           | 299                | 274            |

Total   | 2747          | 2425               |
parameters, model size, and the requirement for computing resources. The lightweight models are more appropriate for circumstances with limited computing resources, such as the edge device and onboard SAR image processing task.

3) Data Processing and Implementation Details: The proposed method and the eight involved DNN models were implemented using the Python (v3.6) and Pytorch [73] deep learning framework (v1.10.1). All the experiments were supported by an NVIDIA DGX-1 server, which is powered by a dual 20-core Intel Xeon E5-2698 v4 CPU and equipped with eight Tesla-V100 GPUs. In addition, all the experiments were accelerated by CUDA Toolkit (v10.2).

The DNN models were trained following the well-known preprocessing of MSTAR data in [36]. Specifically, all the target slices were normalized to $[0, 1]$ to accelerate the convergence of the loss function. The single-channel data were fed into the models as gray-scale images. The random 882 patches of training images were used for training as data augmentation, and the central patches of test images were for testing the models’ accuracy. The cross-entropy loss function and the stochastic gradient descent (SGD) optimizer were employed to train all the studied models, with more details reported in Table III. The publicly accessible MSTAR segmentation annotation SARBake5 [74] was utilized as coordinates pool $m$ to initialize the scatterers’ location.

The parameter setting of the generation algorithm is listed in Table IV. Among them, $n_{\text{max}}$ was set to 90 according to a simple search that is described in Fig. 6. Generally, the more likely the entire optimization process will succeed when batch capacity $B_N$ is larger. It becomes an exhaustive search when $B_N \to \infty$. To seek a balance between efficacy and computational burden, $B_N$ was set to 100 to obtain candidate scatterers of good quality. Notice that $S_0$ and $\sigma$ were just roughly set as $(\theta_{\text{max}} - \theta_{\text{min}})/200$. The analyses on the effect of ASCM parameters on the adversarial scatterers and the generation algorithm are provided, respectively, in Sections V-B3 and V-D. The SAR parameter set $\zeta$ was calculated according to the file of the MSTAR dataset.

4) Evaluation Measurements: The attack performance was measured using the fooling rate, which is defined as

$$\text{Fooling rate} = \frac{\sum_{i=1}^{N_{\text{total}}} \mathbb{1}(F(x_{\text{adv}}) \neq y_i^{\text{gt}})}{N_{\text{total}}}$$

where $N_{\text{total}}$ is the capacity of the evaluated images in the experiments. The fooling rate denotes the ratio of misclassification after being attacked, and the value is between 0 and 1. Fundamentally, with fair constraints and interference, the larger fooling rate indicates a stronger attack algorithm.

B. Results

A 1000-image subset was equally sampled from the test set to ensure that the misclassification is, indeed, caused by attacks and eliminate the effect of category imbalance, which contains a hundred images for each category and is correctly classified by all the DNNs. All the following results were obtained on this subset. For abbreviation, the SMGAA with $N_{\text{adversarial}}$ scatterers is denoted by SMGAA-$N$ in the rest of this article.

1) Quantitative Results: We report the fooling rates achieved by SMGAA-$N$ attacks against the eight evaluated models in Table V, from which the following summaries can be obtained. First, the SMGAA is capable of achieving significant fooling rates against both the studied general and lightweight DNN structures that even a single scatterer can achieve an average fooling rate of 53.7%. Second, it shows that the more available scatterers to deploy the higher fooling
Table V

|                  | AlexNet | VGG11 | ResNet50 | DenseNet121 | MobileNetV2 | AConvNet | ShuffleNetV2 | SqueezeNet | Average |
|------------------|---------|-------|----------|-------------|-------------|----------|--------------|------------|---------|
| SMGAA-1          | 67.3    | 49.5  | 50.2     | 43.8        | 39.5        | 60.4     | 59.2         | 59.6       | 53.7    |
| SMGAA-2          | 87.4    | 78.8  | 73.2     | 69.0        | 62.8        | 82.4     | 79.4         | 83.9       | 77.1    |
| SMGAA-3          | 93.0    | 89.8  | 85.3     | 81.7        | 74.8        | 91.9     | 88.8         | 92.3       | 87.2    |

Fig. 7. Transfer attack fooling rates of SMGAA-3. The surrogate model refers to the one used to craft adversarial examples, and the target model denotes the unknown target victim classifier.

rates can be achieved, illustrating the efficacy of objective (10) in fooling the DNN-based SAR ATR models.

The transfer attack was performed to verify the cross-model transferability of the SMGAA. Fig. 7 depicts the transfer fooling rates achieved by SMGAA-3 attack, which were obtained by attacking the black-box target model using the adversarial examples generated by the surrogate model. What stands out in the matrix is that the SqueezeNet is fooled by black-box SMGAA-3 with a 35.7% fooling rate on average, and the adversaries crafted on VGG11 achieve an average fooling rate of 33.0% against the other seven black-box models. It indicates that there exists an adversarial threat present even in a highly noncooperative setting and even with the simple addition of a few extra scatterers.

2) Visualizations: Some results of the SMGAA are visualized in Fig. 8. In each subgroup, the original image, adversarial scatterer, and perturbed image are arranged from left to right, and the corresponding prediction and confidence score given by AConvNet, as well as the geometric scattering types and normalized ASCM parameters of the scatterers, are listed below the images. It can be observed that our attack framework is capable of achieving satisfactory visual stealthiness and powerful deception, that is, can fool the DNN classifier with a high confidence score (over 90%, depending on the value of \( v_0 \)). Moreover, the adversarial scatterers carry explicit physical attributes compared with the digital perturbations shown in Fig. 1.

3) Parameter Sensitivity: To verify the parameter sensitivity of the scatterers, we modified their parameters and then retested the fooling rate. Fig. 9 depicts the sensitivities to \( A, \gamma_p, L_p, \) and \( \phi_p \), of which results were obtained on the resulted scatterers of SMGAA-1 for AConvNet. The effect of the location variation on both range and azimuth directions is shown in Fig. 10. Several observations can be summarized. First, the attack performance is not affected by the frequency and aspect dependence. The reason is that \( \alpha \) and \( \gamma_p \) have too little effect on the resulted image. Second, the fooling rate decreases when the amplitude turns too larger or smaller, indicating that a certain intensity of scatterer is required to perform a successful attack. Third, as the length increases, the fooling rate remains level and eventually decreases since increasing length will disperse the total energy and lead to a lower intensity. Finally, the adversarial scatterers are more sensitive to excursion along the range direction than the azimuth direction. The reason is that the distributed scatterer has an extension in the azimuth direction and is more robust to small displacements.

Overall, the results show that the adversarial scatterers are mainly affected by their intensity, location, length, and aspect angle. The attack performance is stable within a certain range of parameter fluctuation. It demonstrates the high practical relevance of SMGAA in designing physical adversarial scatterers, that is, attackers do not need to carefully engineer material properties to meet frequency and angle dependence requirements while also having certain tolerance for other attributes.

4) Application of Defense: It is our intention and the most direct application of the proposed method to defend against randomly destructive or intentionally malicious scatterers. With the adversarial training framework [18], [75], our proposal can be utilized as the worst case scatterer generator to facilitate the min–max optimization of robust model training. Particularly, every training SAR target image will be fed into the training model with the adversarial scatterers. The general adversarial training procedure is expensive; thus, we set an efficient SMGAA-3 attack as the surrogate generator whose batch capacity and iteration were set to 1 and 20. The comparisons between the normally and adversarially trained AConvNets are presented in Table VI and Fig. 11.

In Table VI, we compare the SMGAA with the most famous adversarial training procedure with the PGD attack. It reports the accuracy against SMGAA-N and PGD attacks, and the confidence score on the successful adversarial examples. We can observe from the table that the adversarially trained model by SMGAA exhibits significant adversarial examples.

6When \( \alpha \) fluctuates to each of the five available values, the fooling rate remains constant.
improvements and distrust against SMGAA-N attacks even just trained with limitedly exploited SMGAA-3 attack. We can also observe that the PGD-trained model is not capable of defending the adversarial scatterers, highlighting the necessity of the proposed method for defense purposes. Fig. 11 further visualizes the attention maps (extracted by guided-backpropagation [76]), contrasting the discriminative evidence of the normally and adversarially trained models. As can be seen, the robust model gives consistent attention to the adversarial examples, while the nonrobust one is significantly confused.

To summarize, our proposal is shown to be promising and effective for defense purposes. In addition to a compatible defense efficacy against more expensive attacks than the surrogate attack, the robust model also shows well generalizability to defend against attacks with different numbers of scatterers.
Table VI

| Attacks  | Normal | Defense (SMGAA) | Defense (PGD) |
|----------|--------|----------------|---------------|
| Clean    | 98.1   | 97.4\*         | 96.5\*        |
| SMGAA-1  | 39.6   | 80.8\*         | 39.3\*        |
| SMGAA-2  | 17.6   | 62.5\*         | 19.3\*        |
| SMGAA-3  | 8.1    | 46.3\*         | 10.9\*        |
| PGD      | 41.9   | 13.4\*         | 86.5\*        |

\(\epsilon = 4/255\) is set for PGD, and other parameter setting can be found in Table IX.

Confidence on the adversarial category

| Attacks  | Normal | Defense (SMGAA) | Defense (PGD) |
|----------|--------|----------------|---------------|
| SMGAA-1  | 86.5   | 71.6\*         | 77.9\*        |
| SMGAA-2  | 90.4   | 77.2\*         | 82.9\*        |
| SMGAA-3  | 91.0   | 80.7\*         | 85.3\*        |
| PGD      | 89.6   | 87.2\*         | 66.6\*        |

Note: The highest accuracies against the adversarial examples and the lowest confidences on the adversarial category are highlighted in bold.

C. Comparative Experiments

Although our proposal already takes advantage of the physical relevance, in this section, we focus on comparing the proposed method to the currently studied \(\ell_p\) attacks in SAR ATR, as well as the typical physical attack framework in the optical object recognition tasks. It should be as fair as possible since there are different distance measurements. Therefore, equivalent constraints were sought for the \(\ell_p\) attacks. Specifically, the \(\ell_\infty\) norm constraint was set as the average \(\ell_1\) norm perturbation caused by SMGAA-N attack, that is, \(\mathbb{E}_{x_N \sim X_N} [\|\Delta_{x_N}\|/\|x_N\|^2]\), where \(X_N\) represents the adversarial set generated by SMGAA-N for all the models. The \(\ell_0\) norm constraint was set to the average number of pixels that contain 99% energy of \(N\) adversarial scatterers (calculated by \(\ell_2\) norm).

Table VII summarizes the corresponding constraints for different input sizes \(s\), as well as values of \(p\) and \(N\). For a fair comparison, we set no constraint for the physical attack Localized and Visible Adversarial Noise (LaVAN) [50]. Detailed parameter settings of all the studied attacks can be found in Table IX.

The fooling rates and average transfer rates (for instance, the mean value of the matrix in Fig. 7) of all the studied attack algorithms are reported in Table VIII, from which we can conclude the following observations. First, the SMGAA can achieve competitive attack performance although constrained by the extremely low-dimensional ASCM and outperform the physical attack. Second, our proposal can achieve relatively universal attack performance against all the models than the \(\ell_\infty\) and \(\ell_0\) attacks. The performances of the later attacks are shown highly dependent on the target models. For example, they can achieve high fooling rates for some of the studied models (such as DenseNet121 and ShuffleNetV2) but extremely low fooling rates for others (such as VGG11, AConvNet, and SqueezeNet). Third, while \(\ell_0\) attacks are capable of fooling the models through modifying limited pixels, the resulting sparse perturbations are visually abrupt and have no physical relevance to the SAR target images (as shown in Fig. 12). Finally, the \(\ell_2\) attacks can effectively fool the models with finite perturbations and have a universal attack performance to different DNN structures. It should be noted that the \(\ell_2\) variant of the SVA [55] attack accumulates perturbations to maximize utilization of the given constraints and transfers well across different models, while DeepFool and C&W pursue the smallest perturbation and perform poorly in transfer attack conditions.

We then investigate the robustness of these attacks. Various interference was further applied to test the perturbations’ robustness. Table X elaborates on the details and results that show that the minimal perturbations generated by DeepFool...
TABLE VIII
FOOLING RATES AND AVERAGE TRANSFER RATES (%) AGAINST EIGHT DNN MODELS ACHIEVED BY VARIOUS ATTACK ALGORITHMS, INCLUDING:
1) $\ell$\textsubscript{$\infty$}: FGSM [27], BIM [43], and PGD [18]; 2) $\ell$\textsubscript{0}: JSMA [45], SPARSE-RS [35], and PGD ($\ell$\textsubscript{0}) [44]; 3) $\ell$\textsubscript{2}: SVA ($\ell$\textsubscript{2}) [55], DEEPFOOL [28], and C&W [29]; and 4) PHYSICAL ATTACK: LAVAN [50]. THE $\ell$\textsubscript{$\infty$} AND $\ell$\textsubscript{2} CONSTRAINTS FOR THREE CONDITIONS WERE SET TO THE MEAN PERTURBATION CAUSED BY SMGAA-N. FOR $\ell$\textsubscript{0} ATTACKS, THE MAXIMUM NUMBER OF PIXELS PERTURBED WAS SET TO THE NUMBER OF PIXELS THAT CONTAIN MORE THAN 99% OF THE ADVERSARIAL SCATTERERS’ ENERGY.

| Propose | Attacks | AlexNet | VGG11 | ResNet50 | DenseNet121 | MobileNetV2 | AConvNet | ShuffleNetV2 | SqueezeNet | Average |
|---------|---------|---------|-------|---------|-------------|-------------|----------|-------------|------------|---------|
| **Proposed** | | | | | | | | | | | |
| SMGAA-1 | 67.3 | 49.5 | 50.2 | 50.1 | 43.8 | 39.5 | 60.4 | 59.2 | 59.6 | 53.7 |
| FGSM [27] | 24.1 | 5.2 | 10.1 | 9.9 | 79.8 | 7.2 | 73.4 | 4.8 | 32.2 | |
| BIM [43] | 26.9 | 6.0 | 10.2 | 9.9 | 99.6 | 2.3 | 96.5 | 5.1 | 40.8 | |
| PGD [18] | 28.0 | 6.1 | 10.6 | 87.8 | 100.0 | 2.3 | 99.2 | 5.2 | 42.4 | |
| JSMA [45] | - | - | - | - | 53.3 | - | - | 53.3 | |
| SPARSE-RS [35] | 74.9 | 20.1 | 24.9 | 97.7 | 79.0 | 17.6 | 94.0 | 39.0 | 55.9 | |
| PGD ($\ell$\textsubscript{0}) [44] | 80.1 | 21.0 | 30.4 | 92.7 | 76.8 | 40.0 | 96.6 | 77.1 | 65.6 | |
| SVA ($\ell$\textsubscript{2}) [55] | 85.0 | 89.3 | 78.2 | 92.9 | 78.6 | 80.2 | 76.5 | 79.6 | 80.0 | |
| **Proposed** | | | | | | | | | | | |
| SMGAA-2 | 87.4 | 28.8 | 73.2 | 69.0 | 62.8 | 82.4 | 79.4 | 83.9 | 77.1 | |
| FGSM [27] | 40.6 | 10.0 | 17.9 | 77.3 | 85.3 | 3.6 | 83.0 | 7.9 | 40.7 | |
| BIM [43] | 47.7 | 11.3 | 19.2 | 95.0 | 100.0 | 3.7 | 99.8 | 8.4 | 48.1 | |
| PGD [18] | 51.0 | 11.6 | 19.5 | 97.9 | 100.0 | 3.7 | 99.9 | 8.9 | 49.1 | |
| JSMA [45] | - | - | - | - | 66.9 | - | - | 66.9 | |
| SPARSE-RS [35] | 80.3 | 28.4 | 33.6 | 95.5 | 85.8 | 25.5 | 97.9 | 45.1 | 61.5 | |
| PGD ($\ell$\textsubscript{0}) [44] | 93.5 | 36.4 | 44.8 | 92.0 | 88.5 | 67.3 | 99.2 | 88.2 | 76.2 | |
| SVA ($\ell$\textsubscript{2}) [55] | 87.7 | 75.0 | 81.3 | 93.9 | 83.2 | 84.9 | 79.5 | 82.8 | 83.5 | |
| **Proposed** | | | | | | | | | | | |
| SMGAA-3 | 93.0 | 89.8 | 85.3 | 81.7 | 74.8 | 91.9 | 88.8 | 92.3 | 87.2 | |
| FGSM [27] | 54.3 | 13.6 | 25.4 | 85.7 | 89.1 | 4.9 | 86.4 | 11.0 | 46.3 | |
| BIM [43] | 62.3 | 16.3 | 27.7 | 98.3 | 100.0 | 5.1 | 99.9 | 12.4 | 52.8 | |
| PGD [18] | 66.1 | 17.2 | 29.0 | 99.6 | 100.0 | 5.3 | 100.0 | 12.9 | 53.8 | |
| JSMA [45] | - | - | - | - | 71.7 | - | - | 71.7 | |
| SPARSE-RS [35] | 81.8 | 32.6 | 35.9 | 94.6 | 87.8 | 29.7 | 97.7 | 47.2 | 63.4 | |
| PGD ($\ell$\textsubscript{0}) [44] | 92.2 | 44.7 | 53.0 | 91.8 | 90.7 | 77.6 | 100.0 | 91.2 | 80.4 | |
| SVA ($\ell$\textsubscript{2}) [55] | 88.7 | 78.1 | 82.9 | 95.2 | 85.2 | 87.3 | 82.2 | 85.0 | 85.6 | |
| **Minimal $\ell$\textsubscript{2}** | | | | | | | | | | | |
| DeepFOOL [28] | 99.7 | 98.3 | 90.5 | 99.6 | 99.5 | 92.4 | 99.4 | 89.4 | 96.1 | |
| C&W [29] | 100.0 | 99.1 | 99.6 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 99.8 | |
| **Proposed** | | | | | | | | | | | |
| SMGAA-3 | 25.7 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | 1.3 | |
| FGSM | | | | | | | | | | | |
| BIM | | | | | | | | | | | |
| PGD | | | | | | | | | | | |
| JSMA | 34.5 | 27.8 | 44.3 | | | | | | | |
| SPARSE-RS | | | | | | | | | | | |
| PGD ($\ell$\textsubscript{0}) | 5.1 | 3.1 | 38.3 | | | | | | | |
| SVA ($\ell$\textsubscript{2}) | | | | | | | | | | | |
| LAVAN | | | | | | | | | | | |

**D. Ablation Study**

In this section, we investigate how effective is the proposed generation algorithm and its sensitivity to the main ASCM attributes, such as the scattering type, amplitude, and location. For efficiency, we selected a 100-image subset to perform the SMGAA-1 attack with $n^\text{max}$ and $B_Y$ both set to 50. Other parameters were consistent with Table IV unless otherwise stated.

1) Efficacy of the Proposed Optimization Strategies: Fig. 13 depicts the attack performance (average fooling rate on eight models) of the different combinations of the proposed initialization and adaption strategies. When $n^\text{max}$ varies from 10 to 130, all three conditions exhibit stable improvements, and the initialization strategy-based attack performs significantly better than the vanilla attack. The reason is that the optimization is extremely prone to be stuck in local optima subject to the location modality; thus, a good initial location can significantly promote the whole process. The adaption is also helpful to find an available solution in limited steps.

and C&W attacks are the most nonrobust. The results align with the widely held belief that extremely small perturbations would easily be disrupted by noise or transformations. We can also observe that the SMGAA and the local accumulation-based attack [SVA ($\ell$\textsubscript{2})] exhibit robustness in the face of every interference, with distinctly higher remaining average fooling rates than other algorithms. It should be pointed out that the resize operation was considered assuming that the scaled adversarial examples will be implemented to their original size. In this context, our algorithm just generates 882 adversarial images and is free of resizing.
Moreover, the proposed two strategies are compatible to bring more advances to the optimization.

2) Influence of Location: We further investigate the influence of the initial location, with results summarized in Table XI. Our strategy that initializes the candidate scatterers in the target and shadow region succeeds in achieving the highest fooling rates on each of the models. It can be observed that the scatterers initialized in background clutter perform poorly than the uniformly random initial location on the whole image. The results support the intuition that the target and shadow parts give rise to most of the informative features as they carry the structural and scattering information.

In Fig. 14, we show the sensitive region of the SAR target images by performing a hundred attacks (SMGAA-1 with \( \nu_{th} = 0.2 \)) for each image and superposing the scatterers’ location. In the figure, the brighter regions are more vulnerable to perturbations in the form of scatterers. The results show that the vulnerable regions are mainly the electromagnetic shadowing parts both on the target and background, as well as the sides of the target. The result leads to the assumption that the DNN may implicitly learn the structural information from the training data and is highly sensitive to structural deformations.

3) Effect of Scattering Type: Fig. 15 summarizes the fooling rates achieved by different scattering types. In the experiment, each batch of candidate scatterers was initialized by a single scattering type. There is an overview that the distributed scattering mechanism is more aggressive due to it alters more pixels of the resulted image. The average fooling rate of distributed scatterers is 57.5%, higher than the one of located scatterers by 35.1%, and the fooling rates within the two groups are not significantly different. It suggests that no careful configuration of frequency dependence is required when initializing the perturbation.

4) Effect of Amplitude: The SAR images are usually maximum-normalized at the end of the imaging process, which allows a locally strong scatterer to suppress the whole target response. We take this suppression as a weakness the SMGAA can leverage. Table XII exhibits the fooling rates when the amplitude of the adversarial examples (relative to the largest

![Fig. 14. Heatmaps of the vulnerable region of SAR target images, which are the superposition of a hundred attacks to AConvNet (SMGAA-1 with \( \nu_{th} = 0.2 \)).](image)

![Fig. 15. Fooling rates (%) against eight DNN models achieved by SMGAA-1 with different scattering types.](image)
TABLE X

AVERAGE FOOLING RATES (%) AGAINST EIGHT DNN MODELS WITH AND WITHOUT VARIOUS EXTRA PERTURBATIONS. THE EXTRA INTERFERENCE INCLUDES ADDITIVE WHITE NOISE ($\sigma = 10^{-2}$), GAUSSIAN FILTERING ($\sigma = 1$), MEDIAN FILTERING, AND RESIZING (BILINEAR INTERPOLATION, $s^2 \rightarrow 32^2$, $s^2 \rightarrow 88^2$, $s^2 \rightarrow 224^2$). THE KERNEL SIZE OF FILTERS WAS, RESPECTIVELY, SET TO $3^2$, $7^2$, AND $11^2$ FOR $s = 88$ AND 224.

| Attack | Proposed $\ell_\infty$ | Proposed $\ell_0$ | Proposed $\ell_2$ & Physical |
|--------|------------------------|------------------|-------------------------|
| No Interference | SMGAA-1 | SMGAA-2 | SMGAA-3 |
| Add. Noise | 53.7 | 55.7 | 67.2 |
| Gau. Filtering | 43.0 | 44.0 | 57.2 |
| Med. Filtering | 40.1 | 42.1 | 51.2 |
| Resizing | 53.7 | 55.7 | 67.2 |
| Add. Noise | 77.1 | 79.1 | 83.1 |
| Gau. Filtering | 62.6 | 65.6 | 72.1 |
| Med. Filtering | 56.5 | 59.5 | 66.5 |
| Resizing | 77.1 | 79.1 | 83.1 |
| Add. Noise | 87.2 | 89.2 | 93.2 |
| Gau. Filtering | 70.7 | 73.7 | 79.7 |
| Med. Filtering | 62.7 | 65.7 | 71.7 |
| Resizing | 87.2 | 89.2 | 93.2 |

Note: The most and second robust results are respectively bolded and underlined.

TABLE XI

FOOLING RATES (%) AGAINST EIGHT DNN MODELS WITH ACHIEVED BY SMGAA-1 WITH DIFFERENT INITIALIZATION LOCATIONS.

| Initialization of location | AlexNet | VGG11 | ResNet50 | DenseNet121 | MobileNetV2 | AConvNet | ShuffleNetV2 | SqueezeNet | Average |
|---------------------------|---------|-------|----------|-------------|-------------|----------|--------------|------------|---------|
| Whole Image               | 48      | 33    | 37       | 23          | 21          | 32       | 33           | 45         | 34.0    |
| Background                | 28      | 21    | 30       | 8           | 14          | 19       | 18           | 29         | 20.9    |
| Target & Shadow           | 57      | 45    | 48       | 33          | 34          | 48       | 45           | 55         | 45.6    |

Note: Best result in each column is highlighted in bold.

TABLE XII

FOOLING RATES (%) AGAINST EIGHT DNN MODELS WITH ACHIEVED BY SMGAA-1 WITH DIFFERENT AMPLITUDE CONSTRAINTS.

| Maximum amplitudes (dB) | AlexNet | VGG11 | ResNet50 | DenseNet121 | MobileNetV2 | AConvNet | ShuffleNetV2 | SqueezeNet | Average |
|-------------------------|---------|-------|----------|-------------|-------------|----------|--------------|------------|---------|
| 0                       | 57      | 45    | 48       | 33          | 34          | 48       | 45           | 55         | 45.6    |
| 0.5                     | 75      | 57    | 66       | 66          | 60          | 75       | 76           | 82         | 69.0    |
| 1.0                     | 81      | 65    | 78       | 77          | 75          | 76       | 82           | 81         | 76.9    |
| 1.5                     | 81      | 68    | 81       | 80          | 77          | 78       | 86           | 83         | 79.3    |

amplitude of original targets) ranges from 0 to 1.5 (dB). It can be seen that the average fooling rate against eight models increases by 23.4%, 31.3%, and 33.7% at 0.5, 1, and 1.5 dB, respectively. Note that this suppression differs from conventional suppression jamming. It mainly works by strengthening the adversarial features and destructing original features in high-dimensional space. It also may take effect by causing the statistical excursion. Practically, all the studied models are sensitive to this mild suppression, suggesting carefully distinguishing the strong scattering centers in the circumstance that exists potential adversarial risk.

E. Discussion on the Attack Selectivity of MSTAR Data

It has shown that the MSTAR data have attack selectivity [24], [25], that is, the misclassified categories of the adversarial examples are concentrated and exhibit high consistency across different attack algorithms and victim models. For instance, most of the adversarial examples of D7 would be recognized as ZIL131, ZSU234, and T62 for various attack algorithms and victim models. The distribution of adversarial categories caused by the proposed method is reported in Fig. 16. Each element in the subfigure denotes the original adversarial (row to column) category pairs under the
SMGAA-3 attack. It illustrates that each original category is mainly misclassified into two to five adversarial categories, which means that different data points of the same category share similar vulnerable categories. Moreover, there is a certain symmetry across all the matrices, and all the matrices exhibit a certain similarity, suggesting that the attack selectivity is not only related to the model structure and attack method but also to the distribution of the MSTAR data and, more fundamentally, the targets themselves. Since it is difficult for a nonexpert to distinguish the details of the SAR targets, the proposed method may be utilized as an analytical tool to learn the intercategories’ similarity and its imaging representation.

VI. Conclusion

In this article, we have reported the framework, algorithm, results, and analysis of the SMGAA. To generate more feasible and physically informative adversarial perturbations for DNN-based SAR ATR, the ASCM was introduced as an electromagnetic constraint for the attack process, ensuring the physical feasibility of the resulted adversarial perturbations. It is hard to efficiently find solutions on the extremely intricate loss surface of DNNs. To achieve that, several customized, simple, yet effective strategies were proposed to facilitate the general gradient-based optimization, showing significant improvements. Compared with the currently studied $\ell_p$ attacks in this field, our proposal exhibited universal robustness in attacking various DNN structures and against diverse perturbations. More importantly, the proposed method is promising to serve as a surrogate to construct robust ATR models that are defensive against malicious scatterers. In addition, the intercategory similarity and the sensitive region of SAR target images were reported, which are also the potential to design defense methods.

For future work, we plan to validate the proposed method in the real world and exploit the proposed method for more powerful robust model training. We would also like to improve the transferability of our method to better describe the noncooperative attack threats.

ACKNOWLEDGMENT

The authors thank the anonymous referees for their valuable comments and Dr. Shuaifeng Zhi for his suggestions to improve the presentation.

REFERENCES

[1] X. Zhu et al., “Deep learning meets SAR: Concepts, models, pitfalls, and perspectives,” IEEE Geosci. Remote Sens. Mag., vol. 9, no. 4, pp. 143–172, Feb. 2021.
[2] D. Yue, F. Xu, A. C. Frery, and Y. Jin, “Synthetic aperture radar image statistical modeling: Part I—Single-pixel statistical models,” IEEE Geosci. Remote Sens. Mag., vol. 9, no. 1, pp. 82–114, Mar. 2021.
[3] O. Kechagias-Stamatis and N. Aouf, “Automatic target recognition on synthetic aperture radar imagery: A survey,” IEEE Aerosp. Electron. Syst. Mag., vol. 36, no. 3, pp. 56–81, Mar. 2021.
[4] X. Zhu et al., “Deep learning in remote sensing: A comprehensive review and list of resources,” IEEE Geosci. Remote Sens. Mag., vol. 5, no. 4, pp. 8–36, Dec. 2017.
[5] L. Liu et al., “Deep learning for generic object detection: A survey,” Int. J. Comput. Vis., vol. 128, no. 2, pp. 261–318, 2020.
[6] L. Liu, J. Chen, G. Zhao, P. Fieguth, X. Chen, and M. Pietikainen, “Texture classification in extreme scale variations using GalNet,” IEEE Trans. Image Process., vol. 28, no. 8, pp. 3910–3922, Aug. 2019.
A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

D. Malmgren-Hansen and M. Nobel-Jorgensen, “Convolutional neural networks for SAR image segmentation,” in *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

X. Dang, H. Yan, L. Hu, X. Feng, C. Huo, and H. Yin, “SAR image adversarial samples generation based on parametric model,” in *Proc. Int. Conf. Microw. Millim. Wave Technol. (ICMMT)*, May 2021, pp. 1–3.

J. B. Keller, “Geometrical theory of diffraction,” *J. Opt. Soc. Amer.*, vol. 52, no. 2, pp. 116–130, 1962.

D. R. Wehner, *High Resolution Radar*, 2nd ed. Norwood, MA, USA: Artech House, 1995.

J. Sacchini, W. Steedly, and R. Moses, “Two-dimensional Prony modeling and parameter estimation,” in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, vol. 3, Nov. 1992, pp. 333–336.

T. Li and L. Du, “SAR automatic target recognition based on attribute scattering center model and discriminative dictionary learning,” *IEEE Sensors J.*, vol. 19, no. 12, pp. 4598–4611, Jun. 2019.

J. Zhang, M. Xing, and Y. Xie, “FEC: A feature fusion framework for SAR target recognition based on electromagnetic scattering features and deep CNN features,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2174–2187, Mar. 2020.

R. G. Kouyoumjian and P. H. Pathak, “A uniform geometrical theory of diffraction for an edge in a perfectly conducting surface,” *Proc. IEEE*, vol. 62, no. 11, pp. 1448–1461, Nov. 1974.

Y. Akyildiz and R. L. Moses, “Scattering center model for SAR imagery,” *Proc. SPIE*, vol. 3869, Dec. 1999, pp. 76–85.

A. Zhang, “Attributed scattering center feature extraction of complex target from high resolution SAR imagery,” M.S. thesis, Graduate School, Nat. Univ. Defense Technol., Changsha, China, 2009.

The Air Force Moving and Stationary Target Recognition (MSTAR) Database. Accessed: Oct. 29, 2020. [Online]. Available: https://www.sdms.afrl.af.mil/datasets/mstar/

A. Krizhevsky, I. Sutskever, and G. Hinton, “ImageNet classification with deep convolutional neural networks,” in *Proc. NIPS*, 2012, pp. 1097–1105.

K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in *Proc. ICLR*, 2015, pp. 1–14.

K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. CVPR*, Jun. 2016, pp. 770–778.

G. Huang, Z. Liu, L. Van Der Maaten, and K. Weinberger, “Densely connected convolutional networks,” in *Proc. CVPR*, Jul. 2017, pp. 4700–4708.

M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, “ShuffleNet V2: Practical guidelines for efficient CNN architecture design,” in *Proc. ECCV*, Sep. 2018, pp. 116–131.

F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, “SqueezeNet: AlexNet-level accuracy with 50X fewer parameters and < 0.5 MB model size,” 2016, arXiv:1602.07360.

A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d’Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019.

D. Malmgren-Hansen and M. Nobel-Jorgensen, “Convolutional neural networks for SAR image segmentation,” in *Proc. IEEE Int. Symp. Signal Process. Inf. Technol. (ISSPIT)*, Dec. 2015, pp. 231–236.

N. Inkawhich, E. Davis, U. Majumder, C. Capraro, and Y. Chen, “Advanced techniques for robust SAR ATR: Mitigating noise and phase errors,” in *Proc. IEEE Int. Radar Conf. (RADAR)*, Apr. 2020, pp. 844–849.

J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, “Striving for simplicity: The all convolutional net,” in *Proc. ICLR*, 2015, pp. 1–14.

Bo Peng received the B.S., M.S.E.E., and Ph.D. degrees from the National University of Defense Technology (NUDT), Changsha, China, in 2008, 2010, and 2014, respectively. He is currently an Associate Researcher with the College of Electronic Science, NUDT. His research interests include nonstationary signal processing, micro-Doppler analysis, detection, estimation and imaging of nonstatic targets for radar systems, and radar target recognition.

Jie Zhou received the B.S. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2020, where she is currently pursuing the Ph.D. degree with the College of Electronic Science. Her research interests include radar target recognition and adversarial attacks.

Jianyue Xie received the B.S. degree from the Hangzhou University of Electronic Science and Technology, Hangzhou, China, in 2021. He is currently pursuing the M.S. degree with the College of Electronic Science, National University of Defense Technology, Changsha, China. His research interests include synthetic aperture radar (SAR) deceptive jamming and radar signal processing.

Li Liu (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from the National University of Defense Technology (NUDT), Changsha, China, in 2012. During her Ph.D. study, she spent more than two years as a Visiting Student at the University of Waterloo, Waterloo, ON, Canada, from 2008 to 2010. From 2015 to 2016, she spent ten months visiting the Multimedia Laboratory, The Chinese University of Hong Kong, Hong Kong. From December 2016 to November 2018, she was a Senior Researcher with the Machine Vision Group, University of Oulu, Oulu, Finland. She is currently a Full Professor with the College of System Engineering, NUDT. Her articles have currently over 7600 citations in Google Scholar. Her research interests include computer vision, machine learning, trustworthy artificial intelligence, and synthetic aperture radar.

Dr. Liu has served as a Guest Editor for special issues in *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* and *International Journal of Computer Vision (IJCV)*. She serves as an Associate Editor for *IEEE Transactions on Circuits and Systems for Video Technology, Pattern Recognition*, and *Pattern Recognition Letters*. 

Bowen Peng received the B.S. degree from the National University of Defense Technology (NUDT), Changsha, China, in 2020, where he is currently pursuing the M.S. degree with the Comprehensive Situation Awareness Group, College of Electronic Science. His research interests include machine learning and trustworthy radar signal processing.