Physics-oriented adversarial attacks on SAR image target recognition

Jiahao Cui [1] **Wang Guo** [1] **Run Shao** [1] **Tiandong Shi** [1] **Haifeng Li** [1]

Abstract

SAR target recognition algorithms based on deep neural networks are widely used in key tasks such as wartime reconnaissance, environmental monitoring, but the security of SAR systems is also vulnerable to adversarial examples. The imaging process for SAR images in the physical world is dissimilar to that of optical images because SAR imaging is solely regulated by imaging equations rather than the what-you-see-is-what-you-get principle. As a result, generating SAR adversarial examples in the physical world requires considering the changes in SAR imaging equations that happen after deploying physical devices. Thus, this study proposes a Physics-oriented adversarial attacks on SAR image target recognition. The proposed algorithm distinguishes itself through two key features: (1) SAR-BagNet is utilized to identify the salient regions of SAR targets recognized by classifiers, allowing for the exact position and size determination of the adversarial scatterers and enhancing interpretability; (2) Dynamic step size optimization, which is based on the difference equation, continuously refines the electromagnetic parameters, structural parameters, and texture parameters of the adversarial scatterers, leading to a higher search efficiency. In the simulation experiment, the generated adversarial examples can reduce the accuracy of the classifier to recognize the simulated image from 100 % to 14.4 %, thus verifying the method proposed in this paper.

1. Introduction

Synthetic Aperture Radar (SAR) target recognition using deep learning is a prominent research area in radar image interpretation. Leveraging the benefits of end-to-end feature learning, this approach significantly enhances the target recognition rate. Consequently, it finds extensive application in various domains, including military reconnaissance, marine monitoring, and geological exploration. Nevertheless, previous research has demonstrated that deep neural network models are susceptible to adversarial examples. The existence of adversarial examples raises a significant threat to SAR target recognition tasks. Consequently, studying adversarial examples of SAR remote sensing images is vital in enhancing the security and the robustness of SAR systems and mitigating potential risks originating from adversarial attacks.

SAR remote sensing images can be modified using two categories of adversarial attacks: digital domain attacks and physical domain attacks. Digital domain attacks include gradient optimization(Li et al., 2020; Huang et al., 2020; Du et al., 2023), constrained optimization(Chen et al., 2018; Du & Zhang, 2021), decision boundary estimation(Peng et al., 2022a; Qin & Wang, 2022), and Generative Adversarial Networks (GAN)(Wang et al., 2021; Du & Zhang, 2021). These methods add imperceptible perturbations to SAR remote sensing images in the time or frequency domain to mislead target recognition models(Zhang et al., 2022). Physical domain attacks use an electromagnetic scattering parameterization model based on physical optics and multiple reflection processes to perturb imaging parameter model structural parameters(Dang et al., 2021; Peng et al., 2022b). In addition to disturbing the structural parameters, some researchers borrowed from the idea of target scattered wave modulation and applied scattered wave modulation interference and periodic two-phase phase modulation to SAR remote sensing image adversarial attacks. The phase parameters of SAR echo signal are perturbed to generate SAR image adversarial examples.(Liu et al., 2021; Xia et al., 2022).

Optical images and SAR images differ fundamentally in imaging principles and characteristics. Optical images use the energy superposition principle to store grayscale information of multiple gray segments. The pixel values of optical images represent their corresponding physical meanings. In contrast, SAR images use microwave scattering characteristics to create images. The pixel values of SAR images convey amplitude phase and other backscatter infor-

[1]School of Geosciences and Info-Physics, Central South University, Changsha, China. Correspondence to: Haifeng Li <lihaifeng@csu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

mation. Therefore, borrowing attack methods from optical images limits the interpretability of adversarial examples. In the physical domain, SAR remote sensing image adversarial attack methods can obtain electromagnetic scattering characteristics and phase information with clear physical implication. However, the electromagnetic scattering characteristics of targets may experience considerable variations due to changes in target internal configuration, environmental factors, and observation conditions. Therefore, electromagnetic scattering models created through practical measurements and simulation calculations may fail to accurately depict the multidimensional space of electromagnetic scattering variation. Additionally, the phase modulation method may not precisely control the size of the disturbance value due to constraints caused by factors such as time delay and amplitude mismatch. These challenges make it difficult for existing attack methods to generate adversarial examples that are applicable in real-world scenarios.

SAR image is a microwave reflection of ground objects captured by radar, with the echo signal dependent on radar system parameters, ground object surface characteristics, terrain conditions, and atmospheric meteorological conditions. These factors affect the image quality, and it is possible to interfere with the imaging through physical attacks involving changes in the ground object surface characteristics. Therefore, it is intuitively possible to interfere with imaging by changing the surface characteristics of ground objects to generate SAR adversarial examples in the real world. Based on this, we propose a Physics-oriented adversarial attacks on SAR image target recognition. In this method, firstly, we construct a parameter model of the target and its scatterer from a physical model (e.g., CAD model + material), after which a ray tracer is utilized to simulate the electromagnetic wave scattering process involving the target, predicting its features in varying configuration and orientation. Following this, the SAR-BagNet network identifies the salient regions of the target, allowing us to determine the position and size of adversarial scatterers that will be added(Li et al., 2022). Finally, we utilize a dynamic step size optimization strategy based on difference equations to optimize the parameter model by adjusting its electromagnetic, structural, and texture parameters to generate SAR image adversarial examples.

2. Related Work

2.1. Universal Adversarial Perturbations

Universal Adversarial Perturbations (UAP) was originally proposed by Moosavi-Dezfooli with the aim of causing misclassification of deep neural network models through small modifications to images that are not perceptible to the human eye(Moosavi-Dezfooli et al., 2017). UAP is a general method that produces perturbations, which are valid for multiple image samples, as opposed to a single one. The generation of UAP is rooted in the concept of constraint optimization. Assuming that the universal adversarial perturbation is u, most of the samples in the target data set follow a distribution μ, for sample x, machine learning model $f(x)$, find a perturbation u such that:

$$f(x+u) \neq f(x), \text{for “most” } x \sim \mu \tag{1}$$

The perturbations satisfying (1) are called universal adversarial perturbations, as they represent a consistent, image-agnostic perturbation that induces a change in the label for most images sampled x from the data distribution μ. For the universal adversarial perturbation u, the following two constraints exist:

1. $\|u\|_p \leq \varepsilon$, that is, limited to space ε by perturbing the p norm of u.

2. $\mathbb{P}_{x\sim\mu}(f(x+u) \neq f(x)) \geq 1-\delta, \delta$ measures the success rate of an expected universal adversarial perturbation attack on a dataset sampled from distribution μ. When the generated universal adversarial perturbation is added to the data set, the attack on the SAR remote sensing image target recognition model can be completed with a high probability.

2.2. Adversarial Scatterers

Adversarial scatterers are materials or structures designed to interfere with or conceal targets. By adjusting their scattering properties, the effective signal return of synthetic aperture radar (SAR) detection can be reduced. These scatterers can be created using various methods, such as surface texture, complex geometric shapes, or special materials, to absorb, scatter, or deflect radar signals, thereby achieving the goal of hiding targets or blurring echo signals.

In practical applications, corner reflectors have the potential to act as adversarial scatterers, which can be used to conceal or confuse targets. The design of corner reflectors is intended to focus and direct the radar beam back to enhance the detection and imaging effect of the target. However,the scattering properties of corner reflectors can be changed by properly designing their geometric shape, surface texture, or material properties to make them adversarial. For example, corner reflector structures using special materials or coatings can be used to absorb or scatter radar waves, thereby reducing the probability of target detection. Therefore, by adjusting the design parameters of corner reflectors, they can not only enhance the signal return but also counteract radar detection to some extent, achieving the effect of adversarial scatterers.

2.3. Salient Regions

In the field of computer vision, salient regions refer to local regions in an image that are closely related to the target and exert significant influence on the classification decisions made by deep neural networks. These regions contain crucial semantic information, playing a vital role in understanding and explaining how deep learning models perceive and make decisions about objects. In recent years, various methods have been proposed for detecting and analyzing salient regions. Among them, the Class Activation Map (CAM) method has emerged as a widely adopted approach(Zhou et al., 2016). CAM generates salient regions associated with specific target categories by combining feature maps and weights derived from deep neural networks. Specifically, CAM utilizes the feature maps extracted from the last convolutional layer of a CNN, which capture abstract image representations across different spatial locations. By linearly combining these feature maps with the weights of the classification layer, a visualized saliency map S_{ij}^c is generated, facilitating the localization of salient regions corresponding to the target class within the image. The generation method of class activation mapping of class c is expressed as follows:

$$S_{ij}^c = \sum_k \omega_k^c A_{ij}^k \tag{2}$$

Where A_{ij}^k represents the value of the k-th feature map of the last convolutional layer in coordinates (i, j); ω_k^c is the weight that corresponds to class c for the unit pooled from the feature map in the k-th channel. The saliency map generated by the CAM method can enable us to understand the object attention and decision-making process of the deep convolutional neural network model in the object recognition task, thereby providing interpretable and intuitive explanations of the model predictions.

3. Method

The overall design process of the network model that meets the requirements of the target task is illustrated in Figure 1. This paper proposes a comprehensive network model consisting of three modules: SAR image simulation, salient region extraction, and parameter model optimization. The SAR image simulation module adopts Ray-SAR (Ray-Tracing Synthetic Aperture Radar) technology, a method that uses the principle of ray tracing to simulate the imaging process of the SAR system and generates high-quality synthetic aperture radar images(Auer et al., 2016; Niu et al., 2020; 2021). We apply the SAR-BagNet interpretable recognition framework to identify and extract the salient regions of SAR targets in the salient region extraction stage. The framework provides accurate saliency maps for each part of the SAR image, thereby determining the location and size of the added adversarial scatterer. The parameter model optimization process proposes a pseudo-gradient dynamic step size optimization strategy based on difference equations to optimize the electromagnetic parameters, structural parameters, and texture parameters of the parameter model. The optimized model generates optimal adversarial examples that carry attack effects on SAR images.

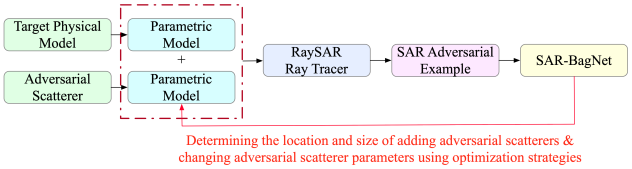

Figure 1. Overall flow chart of the network

3.1. SAR Image Simulation Module

This section mainly introduces the SAR simulation image generation part in the proposed network framework (Figure 2). In the simulation process, we used the RaySAR simulation software developed and open-sourced by Dr. Stefan Auer. This software is an advanced ray-tracing-based SAR image simulator that can effectively simulate the multiple reflections in SAR images, thus generating more realistic SAR images.

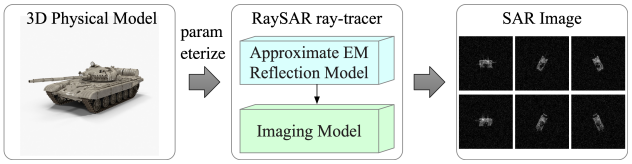

Figure 2. SAR simulation image generation module

The simulation process of RaySAR software can be divided into two main parts. The first part calculates the echo signal strength received by the radar receiver after multiple reflections of the transmitted electromagnetic waves. The second part determines the position where the signal echo is focused on the distance-azimuth plane. These two parts correspond to the electromagnetic scattering model and imaging model in the simulation method, respectively. The electromagnetic scattering model uses specular reflection and diffuse reflection models in the optical field to approximate the echo strength of the radar signal when reflected between or from objects onto the radar antenna. The formulas for the two reflection models are described below:

$$I_s = F_s \cdot (\overrightarrow{N} \cdot \overrightarrow{H})^{\frac{1}{F_r}} \tag{3}$$

$$I_d = F_d \cdot I_{\text{sig}} \cdot (\vec{N} \cdot \vec{L})^{F_b} \tag{4}$$

For the specular reflection model I_s, where F_s is the specular reflection coefficient; F_r is the surface roughness; $\vec{N}$ is the surface normal vector; and $\vec{H}$ is the bisection vector. For the diffuse reflection model I_d, where F_d is the diffuse reflection coefficient; I_{sig} represents the intensity of the incident signal; $\vec{L}$ is the normalized signal vector from the surface point to SAR; and F_b is the surface brightness factor. In the imaging model, the location of the signal echo is calculated by projecting the starting point of the focused ray onto the azimuth and elevation directions. As shown in Figure 3, we use the example of second reflection to illustrate how to calculate the position of the signal echo, and the formula can be expressed as follows:

$$a = \frac{a_o + a_p}{2} \tag{5}$$

$$r = \frac{1}{2}(r_1 + r_2 + r_3) \tag{6}$$

Among them, a and r respectively represent the coordinates of the azimuth direction and the distance direction; a_o and a_p respectively are the coordinates of the main light ray incident point projected on the azimuth direction and the focal light ray exit point; r represents the sum of the length of the light ray emanating from the sensor plane and undergoing secondary reflection and returning to the sensor plane, that is, $r_1 + r_2 + r_3$.

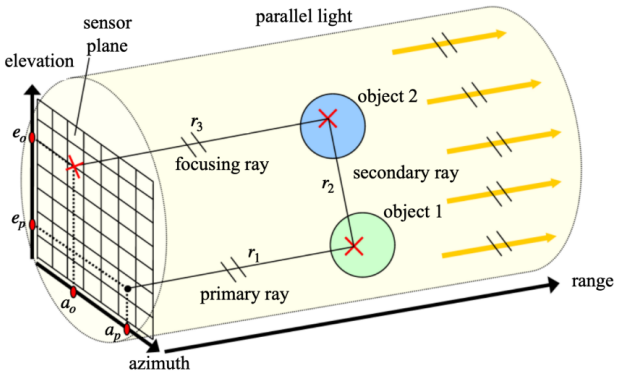

Figure 3. Schematic diagram of imaging model

3.2. Salient Region Extraction Module

After obtaining the SAR image via simulation experiments, the image must undergo training to obtain a classifier model that allows us to precisely grasp the correlations between the image components and the classification decision. In turn, this will enable us to accurately identify the location and magnitude of the adversarial scattering entities that should be added. In previous related work, we introduced how the CAM method works to acquire salient regions. However, their latent representations are extracted from the whole image, and it is unclear how the heatmaps in the latent space are related to the pixel space. In order to accurately reflect the impact of various parts of the SAR image on the final network decision, we choose to apply the SAR-BagNet network to obtain clear saliency maps. Aside from its outstanding interpretability, the SAR-BagNet network also boasts precision high enough to merit consideration.

Figure 4 depicts the SAR-BagNet network, which employs a ResNet-18 comprising global average pooling and a linear classifier to retrieve category-discriminating saliency maps. Unlike CAM, the receptive field of CNN in SAR-BagNet network is confined to small image patches, thereby assuring precise tracking of the contribution of each image patch to the ultimate decision. Subsequently, these small image patches are fed into the SAR-BagNet network, where they undergo feature extraction and generate activations in the corresponding saliency map regions. In this case, a specific class c activations L_c of an image patch can be expressed by Equation (7).

$$L_c = \frac{1}{n}\sum_k \omega_k^c \sum_i \sum_j A_{ij}^k \tag{7}$$

Among them, the activation L_c represents the classification score for class c; n represents the number of pixel units in the feature map, the appearance of 1/n in the equation is due to the global average pooling layer following the last convolution layer. After obtaining the class-specific activation scores for each image patch, we concatenate all the image patches together to obtain the saliency map of the entire image. The SAR-BagNet network generates a saliency map for each class. These saliency maps are spatially averaged and the final class probabilities are obtained by a softmax layer.

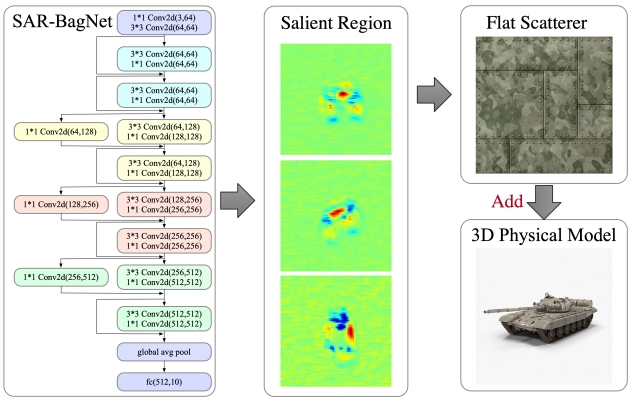

Figure 4. Salient Region Extraction Module

3.3. Parameter Model Optimization Module

After extracting modules from salient regions and obtaining the locations and sizes of required adversarial scatterers, we continuously optimize the physical parameters, including the reflection coefficient, scattering coefficient, and surface roughness, in the adversarial scatterer parameter model. The objective is to identify a set of parameters resulting in the highest classification error rate for a particular category in the classifier. To achieve this, we suggest a pseudogradient dynamic step size algorithm based on differential equations. Our algorithm updates only one parameter component at a time, calculates the differential, and estimates the gradient at that point, grounded on the current parameter value of the image's classification score and cross-entropy loss of the ground truth. Accordingly, the parameter is updated according to the estimated gradient. We use the scattering coefficient F_d as an example to illustrate our optimization process. According to the definition of universal adversarial perturbation, the SAR image adversarial example is represented by the following equation:

$$\begin{cases} p = f\left(g\left(F_d\right)\right) \\ p' = f\left(g\left(F_d + \xi\right)\right) \\ p \neq p' \quad \text{s.t.} \quad 0 \leq F_d + \xi \leq 1 \end{cases} \tag{8}$$

Where p is the predicted output of the original image, and p' is the predicted output of the perturbed image; $g(\cdot)$ represents the SAR image generated by the ray tracer simulation; $f(\cdot)$ represents the classifier model; ξ represents the added perturbation. When the scattering coefficient changes, it will lead to a corresponding change in the classification results. Since $f\left(g\left(F_d\right)\right)$ is continuous in the parameter search space, it is feasible to use the gradient method for optimization. For optimization problems using the cross-entropy loss function, the loss can be expressed as:

$$\begin{cases} loss = -\sum_{i=1}^{N} ylog(p) + (1-y)log(1-p) \\ loss' = -\sum_{i=1}^{N} ylog(p') + (1-y)log(1-p') \end{cases} \tag{9}$$

Where y is the target label (desired output). According to the definition of derivative, when approximating the gradient at each point F_d, using difference instead of derivative is utilized:

$$grad = \frac{\partial Loss}{\partial F_d} \approx \frac{loss' - loss}{\xi} \tag{10}$$

After obtaining the approximate gradient, the perturbation ξ can be iteratively updated:

$$F_d = F_d + \alpha \cdot grad \tag{11}$$

The variable α represents the step size. In order to ensure the efficiency of the search algorithm and reduce ineffective calculation, a larger initial step size is used, which decreases as the loss function increases. This approach balances the algorithm's ability to fully search the parameter space and its ability to quickly converge to the optimal solution. It is theoretically feasible to use a smaller ξ to approximate partial derivatives. We utilize GPUs to accelerate the training process, and if the computed gradient is valid, we can quickly obtain the optimal solution.

4. Experiment

The primary objective of this study is to simulate the T-72 main battle tank model from the MSTAR dataset. The T-72 target, with its complex geometrical structures, provides an excellent testing ground to evaluate the effectiveness of the RaySAR model in simulating complex targets. Initially, a parameter model of the T-72 target is created by parameterizing its physical model, which includes its CAD model and material properties. Then, a ray tracer is utilized to simulate the scattering process of electromagnetic waves on the target. The original dataset utilized SAR with a phased array operation, an X-band frequency, and HH polarization, while employing a 0.3m × 0.3m resolution. We implemented a point light source in the simulation experiment and adjust the imaging parameters, including azimuth angle, elevation angle, and resolution, to ensure they are consistent with the real image. We used the simulation conditions of an elevation angle of 15° and generated a training set with an azimuth angle interval of 1°. In addition, we created a test set under the simulation conditions of an elevation angle of 17° with an azimuth angle interval of 1°. The simulation results have been presented in Figure 5.

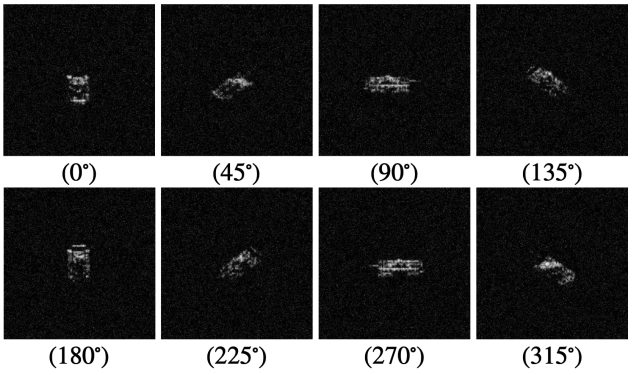

Figure 5. Training Set Simulation Results

After obtaining simulated images, we replaced the corresponding categories in the MSTAR dataset and trained the SAR-BagNet network with this modified dataset. The SAR-BagNet uses the ResNet-18 backbone network with global average pooling and a linear classifier to extract class-

Table 1. Adversarial Attack Results

	AVERAGE CONFIDENCE (T72)	AVERAGE CONFIDENC (BMP2)	RECOGNITION ACCURACY	ATTACK SUCCESS RATE
ORIGIN_OBJECT	0.975	0.024	100%	——
ADD SCATTER STRUCTURE	0.852	0.148	87.8%	12.2%
CHANGE REFLECTIVITY	0.727	0.273	84.4%	15.6%
CHANGE SCATTER RATE	0.263	0.736	27.8%	72.2%
CHANGE SURFACE ROUGHNESS	0.165	0.833	14.4%	85.6%

specific saliency maps and train a well-performing ten-class classifier. Salient regions of SAR-BagNet's recognition of simulated images are shown in Figure 6. The figure shows that the importance weights for recognizing the simulated images are concentrated around the edge of the target and vary depending on the angle of observation. To alter the salient features of the generated simulated images, it may be considered to add an inflatable planar corner reflector to cause a change in the echo signal of the T-72 main battle tank model. By masking the edge salient regions continuously changing with the angle of observation, the added corner reflector structure could deceive the classifier into producing incorrect classification results. Generated adversarial examples with the added inflatable planar corner reflector are demonstrated in Figure 6. The figure indicates that the added corner reflector is capable of effectively masking the salient regions, leading the classifier to produce incorrect classification decisions.

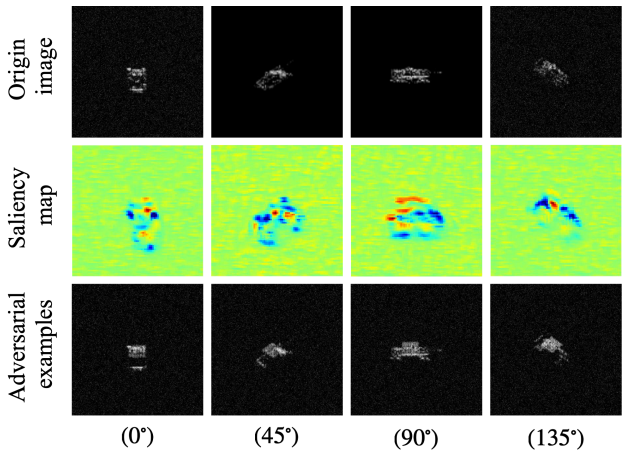

Figure 6. Show the salient regions of the original image and the corresponding generated adversarial examples

In the parameter model optimization module, we adjust the design parameters of the adversarial scatterer by implementing a pseudo-gradient dynamic step size optimization strategy based on differential equations. Table 1 presents the experimental outcomes of adding an inflatable planar corner reflector as an adversarial scatterer and modifying its reflection coefficient, scattering coefficient, and surface roughness through optimization strategies. The second column of the table reflects the average classification score of the simulated image, while the third column represents the same metric attributed to the BMP2 category in the MSTAR dataset. The fourth column depicts the classification accuracy of the ResNet-18 classification network in recognizing the simulated image. Finally, the fifth column indicates the success rate of the attack after adjusting the adversarial scatterer and its design parameters using optimization strategy. The results indicate that by optimizing the scattering coefficient of the adversarial scatterer, we can efficiently transform the classifier's classification outputs. More precisely, adding the adversarial scatterer and modifying its design parameters in the original physical model generates adversarial examples that can reduce the classifier's accuracy from 100% to 14.4% in recognizing simulated images. Hence, our experimental outcomes suggest that our strategy offers high feasibility for examining adversarial attacks in the SAR physical domain.

5. Discussion and Conclusions

This paper proposes a Physics-oriented adversarial attacks on SAR image target recognition. Firstly, the method obtains the simulated image through the SAR image simulator based on ray tracing, and constructs a hybrid dataset of simulated images and MSTAR measured data. Then, utilizing the SAR-BagNet network to identify salient regions of SAR targets to determine the location and size of added adversarial scatterers. Finally, the parameters in the adversarial scatterer are optimized by the pseudo-gradient dynamic step size optimization strategy based on the difference equation to generate SAR image adversarial examples. Experimental results demonstrate that our model has a strong comprehensive performance in terms of attack success rate and computational efficiency. Future research will investigate different types of corner reflectors, their placement methods, and our algorithm's attack performance under black-box conditions with the introduction of random latent variables.

References

Auer, S., Bamler, R., and Reinartz, P. Raysar-3d sar simulator: Now open source. In *2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 6730–6733. IEEE, 2016.

Chen, P.-Y., Sharma, Y., Zhang, H., Yi, J., and Hsieh, C.-J. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

Dang, X., Yan, H., Hu, L., Feng, X., Huo, C., and Yin, H. Sar image adversarial samples generation based on parametric model. In *2021 International Conference on Microwave and Millimeter Wave Technology (ICMMT)*, pp. 1–3. IEEE, 2021.

Du, C. and Zhang, L. Adversarial attack for sar target recognition based on unet-generative adversarial network. *Remote Sensing*, 13(21):4358, 2021.

Du, C., Huo, C., Zhang, L., Chen, B., and Yuan, Y. Fast c&w: A fast adversarial attack algorithm to fool sar target recognition with deep convolutional neural networks. *IEEE Geoscience and Remote Sensing Letters*, 19:1–5, 2021.

Du, M., Sun, Y., Sun, B., Wu, Z., Luo, L., Bi, D., and Du, M. Tan: A transferable adversarial network for dnn-based uav sar automatic target recognition models. *Drones*, 7 (3):205, 2023.

Huang, T., Chen, Y., Yao, B., Yang, B., Wang, X., and Li, Y. Adversarial attacks on deep-learning-based radar range profile target recognition. *Information Sciences*, 531:159–176, 2020.

Li, H., Huang, H., Chen, L., Peng, J., Huang, H., Cui, Z., Mei, X., and Wu, G. Adversarial examples for cnn-based sar image classification: An experience study. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:1333–1347, 2020.

Li, P., Feng, C., Hu, X., and Tang, Z. Sar-bagnet: An ante-hoc interpretable recognition model based on deep network for sar image. *Remote Sensing*, 14(9):2150, 2022.

Liu, Z., Xia, W., and Lei, Y. Sar-gpa: Sar generation perturbation algorithm. In *2021 3rd International Conference on Advanced Information Science and System (AISS 2021)*, pp. 1–6, 2021.

Moosavi-Dezfooli, S.-M., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1765–1773, 2017.

Niu, S., Qiu, X., Lei, B., Ding, C., and Fu, K. Parameter extraction based on deep neural network for sar target simulation. *IEEE Transactions on Geoscience and Remote Sensing*, 58(7):4901–4914, 2020.

Niu, S., Qiu, X., Lei, B., and Fu, K. A sar target image simulation method with dnn embedded to calculate electromagnetic reflection. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14: 2593–2610, 2021.

Peng, B., Peng, B., Yong, S., and Liu, L. An empirical study of fully black-box and universal adversarial attack for sar target recognition. *Remote Sensing*, 14(16):4017, 2022a.

Peng, B., Peng, B., Zhou, J., Xie, J., and Liu, L. Scattering model guided adversarial examples for sar target recognition: Attack and defense. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022b.

Qin, W.-B. and Wang, F. A universal adversarial attack on cnn-sar image classification by feature dictionary modeling. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 1027–1030. IEEE, 2022.

Wang, M., Wang, H., and Wang, L. Adversarial examples generation and attack on sar image classification. In *2021 the 5th International Conference on Innovation in Artificial Intelligence*, pp. 87–91, 2021.

Xia, W., Liu, Z., and Li, Y. Sar-pega: A generation method of adversarial examples for sar image target recognition network. *IEEE Transactions on Aerospace and Electronic Systems*, 2022.

Zhang, L., Jiang, T., Gao, S., Zhang, Y., Xu, M., and Liu, L. Generating adversarial examples on sar images by optimizing flow field directly in frequency domain. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pp. 2979–2982. IEEE, 2022.

Zhou, B., Aditya, K., Agata, L., Aude, O., and Antonio, T. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2921–2929. IEEE, 2016.