

雷达像智能识别对抗研究进展

高勋章* 张志伟* 刘梅 龚政辉 黎湘

(国防科技大学电子科学学院 长沙 410073)

摘要: 基于深度神经网络的雷达像智能识别技术已经成为雷达信息处理领域的前沿和热点。然而, 神经网络模型易受到对抗攻击的威胁。攻击者可以在隐蔽的条件下误导智能目标识别模型做出错误预测, 严重影响其识别精度和鲁棒性。该文梳理了近年来雷达像智能识别对抗技术发展现状, 总结分析了现有雷达一维/二维像识别对抗攻击方法和防御方法的特点, 最后讨论了当前雷达像智能识别对抗研究领域值得关注的5个开放问题。

关键词: 雷达像识别; 神经网络; 对抗攻击; 后门攻击; 对抗防御

中图分类号: TP753

文献标识码: A

文章编号: 2095-283X(2023)04-0696-17

DOI: 10.12000/JR23098

引用格式: 高勋章, 张志伟, 刘梅, 等. 雷达像智能识别对抗研究进展[J]. 雷达学报, 2023, 12(4): 696-712. doi: 10.12000/JR23098.

Reference format: GAO Xunzhang, ZHANG Zhiwei, LIU Mei, *et al.* Intelligent radar image recognition countermeasures: A review[J]. *Journal of Radars*, 2023, 12(4): 696-712. doi: 10.12000/JR23098.

Intelligent Radar Image Recognition Countermeasures: A Review

GAO Xunzhang* ZHANG Zhiwei* LIU Mei GONG Zhenghui LI Xiang

(College of Electronic Science and Technology, National University of Defense Technology, Changsha 410073, China)

Abstract: Intelligent radar image recognition based on Deep Neural Networks (DNN) has become an important topic in radar information processing. However, DNN models are susceptible to adversarial attacks. Malicious attackers can cause intelligent image recognition models to make incorrect predictions, considerably reducing their recognition accuracy and robustness. This article reviews recent research progress on intelligent radar image recognition countermeasures. Then it summarizes the adversarial attack methods on one/two-dimensional radar image recognition models and adversarial defense methods. Finally, it discusses five open questions worthy of in-depth research in intelligent radar image recognition countermeasures.

Key words: Radar image recognition; Deep Neural Networks (DNN); Adversarial attack; Backdoor attack; Adversarial defense

1 引言

随着人工智能技术的发展, 基于深度神经网络的雷达像智能识别算法取得了优异的性能^[1]。然而现有研究表明, 神经网络模型通常存在鲁棒性缺陷, 易受到对抗攻击^[2]的威胁。利用这种技术,

攻击者可以在隐蔽的条件下诱导雷达智能目标识别模型做出错误的预测, 比如, 在一张干净的自行榴弹炮样本(图1(a))上添加微小的对抗扰动(图1(b))后, 神经网络以高置信度将其误判为推土机(图1(c))。如何设计和防御这一类对抗样本, 将成为部署雷达目标智能识别系统时需要考虑的重要问题。

近年来, 针对神经网络分类模型的对抗攻击与对抗防御研究呈现出快速发展的趋势, 雷达像智能识别对抗技术也引起了研究者的关注。文献[3]系统分析了雷达目标识别模型的对抗鲁棒性, 总结了常用的对抗攻击与防御方法。在文献[4]中, 作者综述了遥感图像智能识别面临的安全性问题。最近, 一些结合了雷达目标特性的对抗攻击与对抗防御新方法相继被提出, 这些方法考虑了雷达目标散射中心

收稿日期: 2023-05-29; 改回日期: 2023-07-13; 网络出版: 2023-07-26

*通信作者: 高勋章 gaoxunzhang@nudt.edu.cn;

张志伟 514131141@qq.com

*Corresponding Authors: GAO Xunzhang, gaoxunzhang@nudt.

edu.cn; ZHANG Zhiwei, 514131141@qq.com

基金项目: 国家自然科学基金(61921001)

Foundation Item: The National Natural Science Foundation of China (61921001)

责任编辑: 徐丰 Corresponding Editor: XU Feng

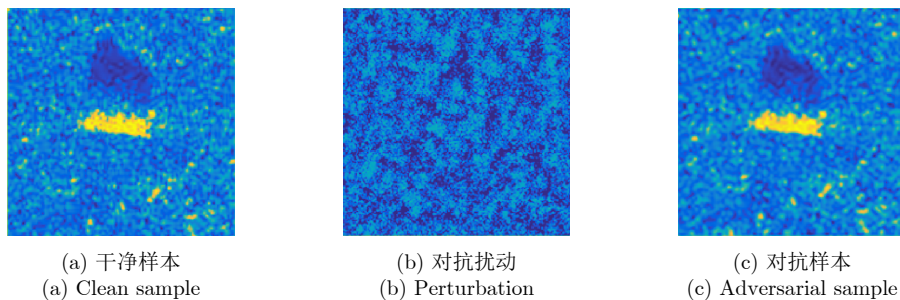


图1 雷达目标对抗样本示例

Fig. 1 Radar target adversarial sample

和距离单元的特点，具有一定程度的语义可解释性，推动了雷达目标智能识别对抗技术的快速发展。

本文系统梳理了雷达像识别对抗领域的最新研究成果，总结了雷达一维像和二维像对抗攻击方法以及对抗防御方法，并讨论了目前雷达像智能识别对抗领域中的开放问题。

2 雷达像智能识别

传统的雷达像识别方法通常利用特征工程构建模板库，并采用合适的分类器^[5-7]进行识别，其识别效果依赖人工设计特征的质量。基于深度神经网络的雷达像智能识别算法通过卷积、池化等操作自动获取雷达像特征，其识别性能优于传统人工方法。

雷达像包括雷达一维像和雷达二维像。雷达一维像通常指高分辨距离像(High-Resolution Range Profile, HRRP)，反映目标散射中心在雷达视线方向上的投影，具有姿态敏感性、幅度敏感性和平移敏感性等特点。基于HRRP的雷达目标智能识别模型通常采用一维卷积网络^[8]和循环神经网络(Recurrent Neural Network, RNN)^[9]。雷达二维像反映目标散射中心在二维成像平面中的投影，可分为合成孔径雷达(Synthetic Aperture Radar, SAR)图像和逆合成孔径雷达(Inverse Synthetic Aperture Radar, ISAR)图像。目前针对SAR图像的智能识别研究较多，所用模型主要采用二维深度卷积网络及各种改进模型。相比于光学图像，雷达像的获取成本较高，在实际应用中难以获取充足的训练样本，导致深度网络模型出现过拟合问题。从简化模型结构的角度出发，复旦大学徐丰团队^[10]提出了仅有5个全卷积层的轻量化神经网络A-ConvNets，在MSTAR^[11]数据集上的识别率达到99%以上的同时大幅减少了网络参数。还有一部分学者从数据增强的角度出发，对有限训练样本进行精细化处理^[12]或者利用生成式模型扩充训练集^[13]，实现训练数据受限条件下的目标识别。在非合作场景下，识别方可获取的训练样本更加有限，通常只能获取少量甚至

若干个训练样本。这种少样本条件下的雷达像识别通常采用迁移学习^[14]和元学习^[15]等方法。

基于深度神经网络的雷达像识别方法一般采用梯度下降的方式更新网络参数，使模型对训练数据的预测分布与其真实分布的交叉熵最小。这种基于数据驱动的智能识别方法存在潜在的鲁棒性缺陷。比如，只需对图像中若干个特定位置的像素施加扰动便可显著增大样本在模型上的交叉熵损失，从而导致深度识别模型误判样本的类别。这种缺陷为深度学习系统在雷达像识别中的应用带来了极大的安全隐患。

3 雷达像智能识别对抗攻击

2014年，Szegedy等人^[16]发现深度神经网络易遭受对抗攻击的威胁。通过在一张干净图片上加入一些精心设计的微小扰动，攻击者可以生成对抗样本，并在人眼难以察觉的条件下误导神经网络分类模型做出错误的预测。设计并实现误导深度神经网络模型的对抗性扰动的过程，称为对抗攻击。

3.1 对抗攻击原理

深度识别模型的训练过程如图2所示。

给定一个尺寸为 $h \times w$ 的样本 x 及其真实标签 y ，一个网络参数为 θ 的 k 分类网络 f_θ 的训练阶段可视作在已知标签 y 的前提下，寻找最小化交叉熵损失 l 的模型参数 θ 的过程：

$$\operatorname{argmin}_{\theta} l(f_{\theta}(x), y) \quad (1)$$

模型 f_θ 的测试阶段则是在固定模型参数 θ 的前提下，为待测样本寻找损失最小的类别标签 i 的过程，表示为

$$\operatorname{argmin}_i l(f_{\theta}(x), i), i \in (1, 2, \dots, k) \quad (2)$$

对抗攻击将样本 x 视作待优化量，沿着梯度上升的方向对 x 添加扰动来增大其与真实标签 y 之间的交叉熵，进而改变目标式(2)的预测结果，即

$$\max_x l(f_\theta(x), y) \tag{3}$$

令 η 和 x_{adv} 分别表示样本在扰动前后的差异和生成的对抗样本，并用 L 范数来对 η 的大小进行约束，则对抗攻击的目标函数可转化为

$$\min_{\eta} \|\eta\|_L \text{ s.t. } \operatorname{argmax}_{x_{adv}} (f(x + \eta)) \neq y \tag{4}$$

扰动范数 L_p 定义了目标函数(1)中扰动的约束形式，其表达式为

$$\|\eta\|_p = \left(\sum_{i=1}^n |\eta_i|^p \right)^{\frac{1}{p}}, \eta = x - x' \tag{5}$$

常用的约束范数有 L_0 范数、 L_2 范数和 L_∞ 范数。 L_0 范数衡量了 $x \neq x_{adv}$ 的像素总和，此约束下的扰动具有稀疏性； L_2 范数衡量了 x 和 x_{adv} 之间的欧氏距离，此约束下的扰动在视觉上难以察觉； L_∞ 衡量了 x 和 x_{adv} 之间的最大像素深度差异， $\|x - x'\|_\infty = \max(|x_1 - x_1'|, |x_2 - x_2'|, \dots, |x_n - x_n'|)$ ， L_∞ 范数下的扰动方向与梯度方向一致，运算更加便捷。

3.2 对抗攻击分类

图3给出了对抗攻击从扰动机理、攻击者先验、攻击特异性和攻击阶段4个方面的分类。

从产生机理来看，对抗攻击可分为基于梯度的攻击、基于优化的攻击和生成式攻击。几乎所有的神经网络都通过梯度下降的方式来优化损失函数，损失函数值越小，代表分类误差越小，模型识别效果越好。从这一机制出发，基于梯度的攻击沿着梯度上升的方向对干净样本施加对抗扰动，使得新样本在模型上的损失函数值增大，以实现诱导模型误判的目的，典型的方法有快速梯度符号法^[17](Fast Gradient Sign Method, FGSM)、基础迭代法^[18](Basic Iteration Method, BIM)、DeepFool法^[19]等。基于优化的攻击将对抗扰动的生成转化为约束条件下的寻优问题，通过在限定条件下寻找最能影响分类结果的像素点进行扰动实现对抗样本的生成，典型的方法有CW法^[20]、雅可比显著图法^[21]、单像素法^[22]等。不同于在干净样本上添加对抗扰动的方法，生成式攻击^[23]利用生成对抗网络直接生成对抗样本，具有生成速度快、无需获取真实目标样本的优势。

依据攻击者对目标模型的先验信息获取程度，对抗攻击可分为白盒攻击、黑盒攻击和灰盒攻击。白盒攻击是指攻击者对深度模型的网络种类、节点权重、训练集等参数完全已知，且可以与模型进行

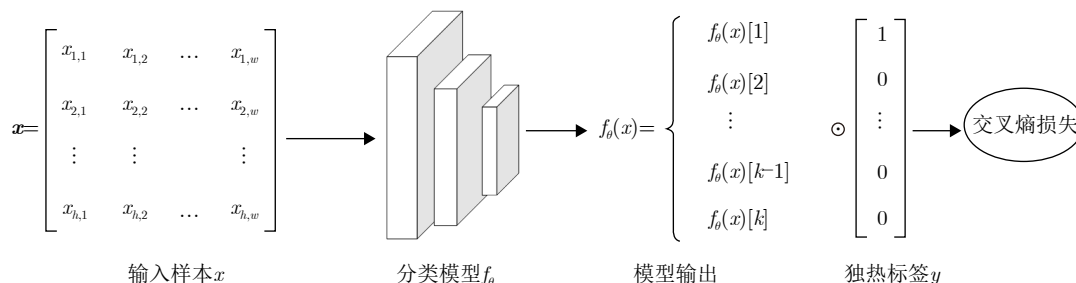


图 2 深度模型识别原理图

Fig. 2 Recognition process of deep neural network

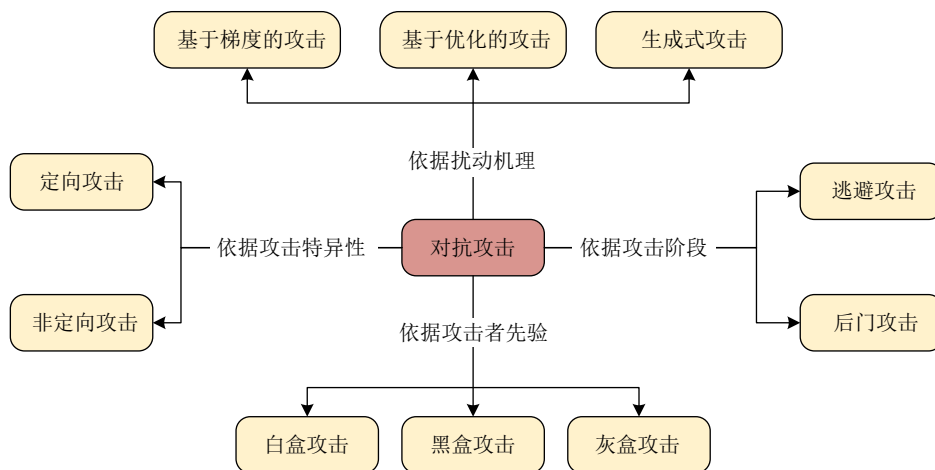


图 3 对抗攻击方法分类

Fig. 3 Categories of adversarial attack

输入与输出的交互。黑盒攻击是指攻击者除了可以与模型交互外，无法获知模型的其他任何信息。灰盒攻击是指攻击者知道模型的种类和结构，可以与之交互，但节点权重未知。通过白盒攻击^[17-25]生成的对抗样本通常在目标模型上具备较高的欺骗率，但在不同模型之间的迁移性较差。黑盒对抗样本通常在某个代理白盒模型上生成，再通过特定的寻优方式^[26-28]增强自身跨模型的攻击性能，在牺牲了白盒欺骗率的情况下提高了迁移性。灰盒对抗样本的性能介于白盒对抗样本和黑盒对抗样本之间。当前，雷达像智能识别对抗研究大都基于白盒假设，即攻击者完全获取了目标模型的所有信息。然而在非合作场景下的雷达目标识别中，识别方与被识别方均缺乏彼此的先验信息，且在识别过程中双方无法交互信息，因而攻击方难以获取目标模型的反馈来指导对抗扰动的生成。在黑盒条件下，对抗样本的扰动生成需要耗费较长的运算时间，难以实时地应用于雷达目标识别系统，且攻击的成功率通常也较低。

依据攻击特异性，对抗攻击可分为定向攻击和非定向攻击。定向攻击要求模型将对抗样本误判为攻击者指定的类别。非定向攻击生成的对抗样本只要求模型的预测类别与真实类别不同，无需指定具体的错误类别。定向攻击通常需要减小对抗样本在指定类别上的损失，直到样本在指定类别上的预测概率超过真实类别的预测概率。非定向攻击只需增大对抗样本与其真实类别的损失，直到真实类别的预测概率低于某一阈值或者被其他类别的预测概率超过。

依据攻击发生的阶段，对抗攻击可分为逃避攻击和后门攻击。逃避攻击在模型的推理阶段生成对抗样本进行攻击，后门攻击发生在模型的训练阶段，攻击者可篡改一部分训练数据或者对训练过程进行恶意操纵，使模型对含有特定图案(称为触发器)的图像样本预测为攻击者指定的类别，而对干净样本正常预测。从原理上看，后门攻击利用的漏洞来源于异常数据，这种漏洞是人为构造的而非网络自然形成的，逃避攻击所用漏洞来源于神经网络与人类认知的差异性。从攻击者的权限上看，逃避攻击通常需要在推理阶段结合待测样本与目标模型，经过一定的优化过程在线产生，而后门攻击需要干预模型的训练阶段，具有投毒和木马两种形式，投毒攻击通常只篡改训练集数据，木马攻击者权限则可扩展至模型参数、模型结构、训练方法等。从攻击目的来看，后门攻击关注触发器对模型行为的干扰，而逃避攻击更加关注样本对模型预测的影响。

目前后门攻击已在人脸识别^[29]、交通路牌识别^[30]领域造成安全威胁，在卫星遥感^[31]和无人机侦察^[32]领域，后门攻击也引起了研究者的关注。鉴于目前雷达像识别领域尚未有后门攻击研究的公开文献，下文所述雷达二维/一维像对抗攻击方法均属于逃避攻击。

3.3 雷达二维像智能识别对抗攻击

早期的雷达像智能识别对抗攻击方法将雷达像视作单通道的灰度图像，借鉴光学图像中的对抗攻击方法逐像素地生成对抗扰动，这类方法在雷达像识别对抗攻击的起步阶段研究较多，通常为经验性探索和实证性研究。文献^[33]首次对雷达数据集上的对抗样本进行了经验性分析，指出蕴含特征信息越丰富的雷达像越容易受到对抗噪声的扰动，并且对抗样本通常分布在几种特定的类别上。文献^[34]认为雷达像与光学图像的主要区别在于雷达像具有稀疏性，并在MSTAR数据集上复现了多种基于 L_0 范数的稀疏攻击方法。文献^[35]将光学图像识别中经典的通用对抗扰动方法迁移到雷达像识别中，并在MSTAR数据集上评价了主流分类识别网络的对抗鲁棒性，指出结构越复杂的网络越容易受到对抗样本的攻击。

在迁移、复现光学对抗攻击方法的基础上，一些研究进一步考虑了雷达像在数字域的特点，比如特征的稀疏性^[36]，以及适用于雷达目标识别的攻击场景，比如黑盒场景^[37]和融合识别场景^[38]。文献^[36]指出深度模型在SAR图像中提取到的高维特征存在大量冗余信息，并提出使用U-net^[39]码网络进行前向映射来代替传统CW方法的大范围寻优，在攻击效果基本不变的情况下大幅度提高了运算速度。文献^[37]针对雷达二维像提出了一种扰动幅度更小的攻击方法，且该方法在黑盒条件下仍保持了在错误类别上的较高置信度。文献^[38]针对遥感图像提出了一种基于稀疏差分协同进化的对抗攻击方法，提高了多源融合传感器下的对抗样本欺骗率。文献^[40]在不获取训练集样本的条件下利用黑盒通用对抗扰动攻击SAR图像分类模型，实现了超过60%的欺骗率。上述方法在设计对抗扰动时结合了雷达像在数字域的特点，但未涉及数字域对抗样本在物理域的实现方式。

从物理实现的角度上看，光学图像的对抗扰动可通过相机拍摄实现由数字域向物理域的转换，而雷达像的对抗扰动则需要体现为目标回波的相干能量累积。因此，研究者希望建立数字域的雷达像对抗扰动与二面角、三面角等真实物理结构的电磁散

射特性的联系^[41], 从而增加雷达像对抗样本的物理可实现性。对处于运动中的雷达目标而言, 背景区域是不断变化的, 因而一个可行的思路是将扰动约束在目标区域附近。文献^[42]提出将对抗扰动约束为若干个像素点的聚合后再添加到目标附近, 以此来逼近实际场景中的目标散射点。文献^[43]利用Gabor特征对SAR图像进行纹理分割来生成目标区域掩模, 并在对抗攻击的目标函数中加入掩模约束, 将扰动限制在目标区域。文献^[44]指出对抗攻击生成的高频非鲁棒特征可能导致模型的对抗脆弱性, 通过将对抗扰动约束为SAR图像散斑的形式, 提高了非合作条件下的黑盒攻击迁移率。上述方法在生成扰动时, 虽然考虑了扰动区域进行约束, 但仍未建立对抗扰动与目标电磁散射特性的紧密联系。

最近, 利用属性散射中心理论来指导对抗扰动的生成引起了学界的关注。属性散射中心模型用多个参数来描述二面角、三面角等典型结构的散射机理, 可定量描述频率 f 、方位角 ϕ 等参数对目标电磁散射响应的影响^[45], 其中单个散射中心的响应可表示为

$$E(f, \phi; \Theta_N) = A \cdot \left(j \frac{f}{f_c} \right)^\alpha \cdot \exp \left(-j \frac{4\pi f}{c} (x \cos \phi + y \sin \phi) \right) \cdot \text{sinc} \left(\frac{2\pi f}{c} L \sin(\phi - \bar{\phi}) \right) \cdot \exp(-2\pi f \gamma \sin \phi) \quad (6)$$

其中, f_c 为雷达中心频率, c 为光速, $\Theta_N = [A, x, y, \alpha, \gamma, L, \bar{\phi}]$ 是影响散射相应的参数集, A 是幅度, x 和 y 分别为距离向和方位向的坐标, α 表示频率依

赖, γ 表示方位依赖, L 表示分布散射中心的强度, $\bar{\phi}$ 表示当前散射中心的方位角。通过将对抗扰动生成过程与属性散射中心约束相结合, 可提高雷达像对抗扰动的物理可实现性, 典型的方法如表1所示。文献^[46]利用属性散射中心重构算法提取出SAR图像中目标区域的散射点, 然后利用空域变换攻击来扭曲散射点的形状, 实现了将扰动约束在目标区域内的同时增加攻击的隐蔽性。文献^[47]利用属性散射中心模型生成参数化的对抗性散射中心, 并利用基于高斯随机步长的贪心算法在限定的目标区域内寻找对抗性散射中心的最优位置, 该方法首次在数字域的对抗扰动中添加了雷达属性散射中心的成像约束, 并在MSTAR数据集上获得了较高的欺骗率。文献^[48]针对现有基于正交匹配追踪(Orthogonal Matching Pursuit, OMP)^[49]的散射中心提取方法耗时长, 难以在SAR图像识别系统中实时应用的问题, 提出一种改进的OMP方法来降低运算代价。在获得目标属性散射中心的基础上, 文献^[48]进一步提出一种模型无关的黑盒对抗样本生成方法, 首先在MSTAR数据集上对训练样本进行属性散射中心重构来获得干净样本散射中心先验, 然后通过增加对抗样本与干净样本之间的JS散度来实现攻击。基于属性散射中心的对抗攻击方法使数字域的对抗扰动具有更好的物理实现前景, 但仍是以静态的单帧图像作为攻击基础, 未能对目标运动过程中的扰动变化进行建模。

代表性的雷达二维像对抗攻击方法如表2所示。

3.4 雷达一维像智能识别对抗攻击

针对HRRP识别模型的对抗攻击同样满足式(1)

表 1 基于属性散射中心模型的典型雷达二维像对抗攻击方法

Tab. 1 Typical radar image adversarial attacks based on attribute scattering center model

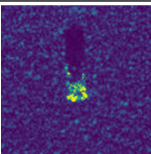
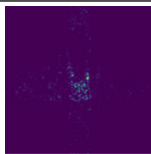
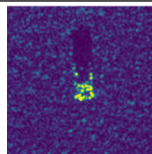
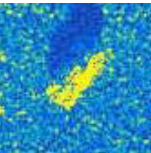
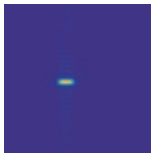
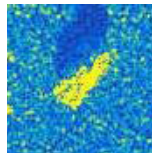
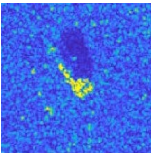
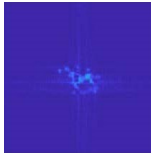
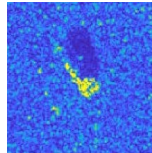
方法	干净样本	扰动	对抗样本	关键技术
文献 ^[46]				散射中心提取, 空域形变
文献 ^[47]				单散射中心攻击
文献 ^[48]				改进OMP算法, 黑盒攻击

表2 雷达二维像对抗攻击研究现状

Tab. 2 Summary of adversarial attacks on radar two-dimensional image

文献	攻击先验	扰动范数	验证模型	数据集	攻击特异性	优缺点
[33]	白盒	L_∞	VGG ^[50]	MSTAR SENSAR ^[58]	非定向	验证光学方法的适用性和差异性，未结合雷达像特性
			ResNet ^[51]			
			DenseNet ^[52]			
			GoogleNet ^[53] InceptionV3 ^[54]			
[34]	白盒	L_0	A-ConvNet ^[10]	MSTAR	非定向	
[35]	白盒	L_2/L_∞	自定义CNN	MSTAR	非定向	
[36]	白盒	L_2	ResNet	MSTAR	定向/非定向	
[37]	白盒/黑盒	L_2/L_0	A-ConvNet	MSTAR OpenSARship ^[59]	定向/非定向	
			ResNet			
[38]	白盒	L_0	ResNet	So2Sat-LCZ42 ^[60]	非定向	结合了雷达像自身特性和识别场景，未考虑对抗样本的物理实现问题
			VGG			
			MobleNet-v2 ^[55]			
[40]	黑盒	L_∞	AlexNet ^[56]	MSTAR SARSIM ^[61]	非定向	
			ResNet			
			DenseNet			
			VGG A-ConvNet			
[41]	白盒	L_∞	CNN	MSTAR	非定向	
[42]	白盒/黑盒	L_2	GoogleNet	MSTAR	非定向	
			DenseNet InceptionV3			
[43]	白盒	L_2	ResNet	MSTAR	非定向	对扰动区域进行初步限制，未建立扰动像素与雷达信号的对应关系
			自定义CNN			
[44]	黑盒	L_∞	AconvNet	MSTAR	非定向	
			VGG			
			ResNet			
			DenseNet InceptionV4 ^[57]			
[46]	白盒	双线性变换	ResNet	MSTAR	非定向	
			MobileNet-v2			
[47]	白盒	$L_0/L_2/L_\infty$	A-convNet	MSTAR SAR Bake ^[62]	非定向	考虑了单帧静止目标对抗样本的物理实现，未考虑目标运动过程中的扰动变化
			VGG			
			ResNet			
			DenseNet MobileNet-v2			
[48]	黑盒	L_2	VGG	MSTAR	非定向	
			ResNet MobileNet			

的形式。根据扰动范围的不同，针对HRRP目标的对抗攻击可分为全距离单元扰动和特定距离单元扰动。

3.4.1 全距离单元扰动

对于一个具有 N 个距离单元的HRRP样本 x ，攻击者分别计算每个距离单元处关于损失函数的梯度，并沿着梯度上升的方向添加适当强度的干扰脉冲便可形成HRRP对抗样本。文献[63]首次运用光学图像中经典的对抗攻击方法来干扰雷达HRRP识别模型。该作者提出了一种改进FGSM方法来提高

HRRP对抗样本对扰动位置和扰动幅值的鲁棒性，通过在一个标签为“安26”的飞机目标HRRP数据上添加全距离单元扰动，该方法成功将模型分类结果误导为“雅克42”飞机目标，如图4所示。文献[64]对HRRP分类模型的对抗鲁棒性进行探索，并通过优化的方式生成了HRRP通用对抗扰动。文献[65]经验性地探索了针对雷达HRRP分类模型的对抗攻击，并提出了4种HRRP对抗攻击方法用于白盒、黑盒、通用扰动和特异扰动4种不同的应用场景。基于全距离单元扰动的攻击方法思路简单，只需在光学对抗攻击方法的基础上调整输入维度即

可，具有较低的运算复杂度，但生成的扰动难以扩展至信号域。

3.4.2 特定距离单元扰动

HRRP数据反映了雷达回波在径向上的散射强度，目标区域的散射强度大而背景区域的散射强度小，且背景区域的成像条件变化较快。与雷达二维像对抗样本类似，针对HRRP识别模型的攻击也希望将扰动约束在目标区域的距离单元上以增加对抗样本的物理可实现性。文献[66]提出一种与欺骗干扰机相结合的雷达HRRP分类模型对抗攻击方法，该作者首先利用差分进化算法寻找HRRP数据中易受攻击的脆弱距离单元，然后利用干扰机在这些距离单元中注入特定幅值的干扰脉冲，实现了高置信度HRRP对抗样本的生成。利用文献[66]的方法，可在一个“安26”目标的干净HRRP数据中的特定

距离单元上增加对抗扰动后，神经网络将其误判为“塞斯纳S”，如图5所示。

目前针对雷达HRRP攻击的研究相对较少，现有公开文献所述两类方法如表3所示。

4 雷达像智能识别对抗防御

对抗防御主要研究如何抵御对抗样本的干扰，提升深度神经网络鲁棒性。针对神经网络模型对抗鲁棒性的研究已经成为深度学习可解释性理论的重要组成部分。现有雷达像对抗防御方法主要借鉴光学图像中的对抗防御技术，本文依照防御阶段的不同，将对抗防御方法分为输入端防御、模型端防御和输出端防御，如图6所示。输入端防御包括对训练数据和测试数据的预处理、数据增强等操作。模型端防御包括改善模型的训练策略和设计更鲁棒的模型结构。输出端防御只调取模型的特征向量、

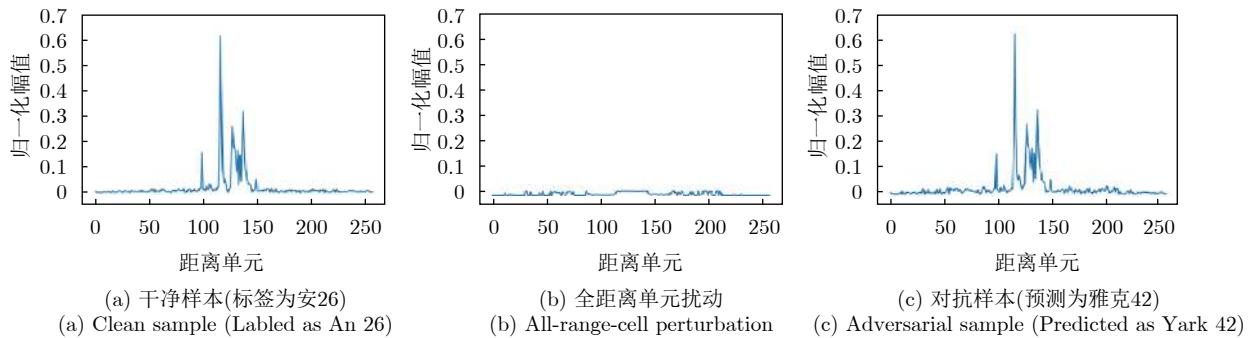


图 4 基于全距离单元扰动^[63]的HRRP对抗攻击示意图

Fig. 4 All-range-cell^[63] adversarial attacks on radar HRRP

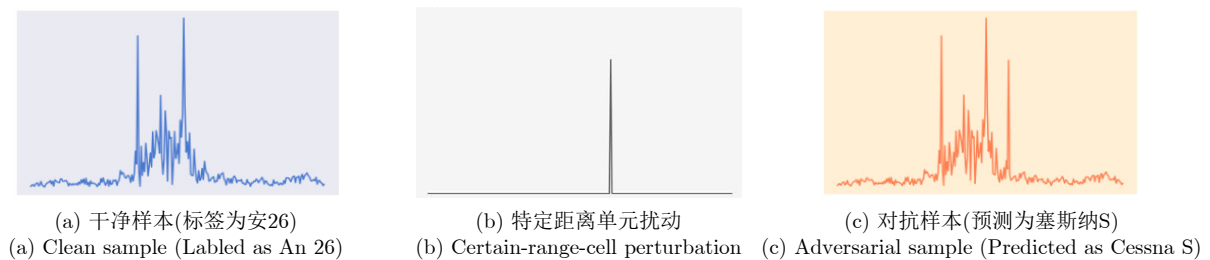


图 5 基于特定距离单元扰动^[66]的HRRP对抗攻击示意图

Fig. 5 Certain-range-cell^[66] adversarial attacks on radar HRRP

表 3 雷达一维像对抗攻击研究现状

Tab. 3 Summary of adversarial attacks on radar HRRP

文献	攻击先验	扰动方式	验证模型	数据集	攻击目标	特点
[63]	白盒	全距离单元	自定义CNN	3类飞机目标: 雅克42; 塞斯纳S/II; 安26	非定向	仅设计数字域攻击, 未考虑物理可实现性
[64]	白盒	全距离单元	自定义CNN	3类飞机目标: 雅克42; 塞斯纳 S/II; 安26	定向/非定向	
[65]	白盒/黑盒	全距离单元	自定义CNN; 全连接网络	MSTAR数据集的HRRP还原数据 ^[67]	定向/非定向	较好的物理可实现性, 未考虑目标姿态等因素带来的影响
[66]	白盒/黑盒	特定距离单元	AlexNet	3类飞机目标: 雅克42; 塞斯纳S/II; 安26	定向/非定向	
			ResNet			
			DenseNet			
		InceptionNet				

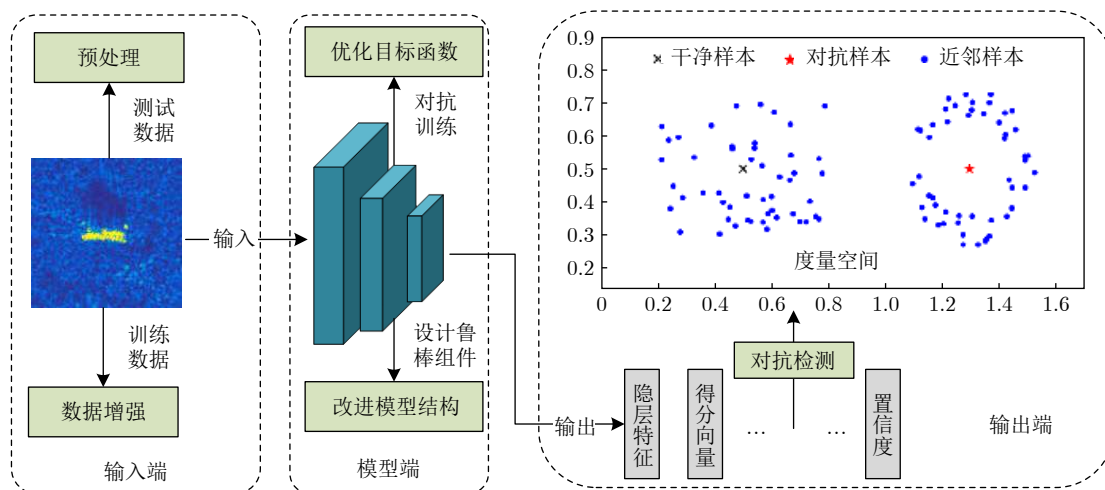


图6 对抗防御方法分类

Fig. 6 Categories of adversarial defense

logit向量、置信度向量等，通过设计特定判据来检测模型的输出是否存在异常。

4.1 输入端防御

输入端防御在数据层面对测试样本或者训练样本进行处理，主要思路有预处理和数据增强两种。

预处理方法将对干扰扰动视作噪声，希望通过降噪、尺度变换等预处理方式去除待测样本中潜在的对抗扰动。文献[68]对输入模型的图像像素进行随机丢弃来破坏对抗样本在模型隐藏层中的表征，再利用重构网络还原像素丢弃后的图像送入模型识别，有效降低了对抗扰动的影响。文献[69]将图像进行二值化阈值分割后再输入网络，有效消除了对抗扰动的影响。此方法在手写体数据集MNIST^[70]上取得较好的效果，但是对大尺度多通道的对抗样本识别性能不佳。文献[71]将对干扰扰动视作噪声，并训练一个超分辨率网络对待测图像进行预处理，将超分辨率增强后的图像输入网络来抑制对抗扰动的威胁。文献[72]将输入图像进行小波变换和余弦展开，然后使用传统的支撑向量机进行分类，有效规避了基于深度模型生成的对抗样本的攻击。基于预处理的防御方法对二范数扰动下的对抗样本有较好的抑制效果，但也会影响干净样本的识别率。

数据增强方法认为样本数量不足带来的过拟合问题会导致模型对于微小的对抗扰动敏感，因此可通过扩充训练样本数量的方式提高模型的对抗鲁棒性。文献[73]指出利用自监督对比学习训练模型可以增强SAR图像分类模型的对抗鲁棒性，对比学习在训练过程中额外生成一批数据增强样本，通过优化原始样本与增强样本之间的距离分布，使得特征空间中同类样本聚集而异类样本簇互相远离。在文

献[73]的基础上，文献[74]以对抗样本作为增强样本，采用对抗性自监督学习在训练过程中降低干净样本与对抗样本的对比损失，进一步提高了SAR图像分类模型对主流对抗攻击的鲁棒性。但由于自监督对比学习未能充分利用样本的标签信息，以上防御方法均会导致模型在干净样本测试集上的识别率降低。

4.2 模型端防御

模型端防御希望改善模型自身的鲁棒性来降低对抗攻击的威胁，主要有优化训练目标函数和改进网络结构两种方式。

优化目标函数方法以对抗训练为代表，这类方法同时利用干净样本和对抗样本进行训练，将对干扰扰动的生成函数融入训练目标函数中，从而使模型在训练过程中学习到对抗样本的先验信息。关于对抗训练为什么能提高模型鲁棒性，领域内学者的看法有以下几点：一是对抗样本起到了扩充数据集的作用，使得深度模型能够学习到更丰富的特征用于分类^[18]。二是对抗训练得到的模型更加关注样本中全局性的特征，因此会忽略掉局部的对抗噪声扰动^[75]。经典的对抗训练方法TRADE^[76]将对干扰样本引起的鲁棒性误差拆分为自然分类误差与边界误差之和，并通过调整两种误差的权重获得鲁棒性与分类精度的折中。在雷达领域，文献[77]研究了不同噪声对SAR图像识别模型的影响，包括随机噪声、相位噪声和对抗噪声，并发现迁移光学领域的对抗训练策略不仅可以提升SAR图像识别模型的对抗鲁棒性，也有助于提高模型在随机噪声和相位噪声环境下的识别率。文献[47]针对雷达像对抗样本中扰动的区域约束问题，提出一种改进的对抗训练

策略,如图7所示,即在对抗训练的过程中将扰动约束为目标区域的散射中心样式,实验证明该方法可有效防御基于散射中心约束扰动的对抗样本。基于对抗训练的防御方法虽然思路简单,但是模型训练过程十分耗时,且所获模型无法防御二次攻击。

在改进模型结构方法方面,文献[78]提出将深度模型中的Softmax层替换为竞争性过完备层(Competitive Overcomplete Output Layer),该层利用多个神经元的输出来共同表征某一类别的预测结果,可在MSTAR数据集上有效降低DeepFool攻击的成功率。文献[79]利用SAR-BagNet^[80]观察SAR图像识别过程,发现深度模型在对SAR图像进行识别时,背景区域承担了较多的分类贡献,而经过对抗训练的模型则会重点识别SAR图像的目标区域,且经典的对抗训练方法Trade^[76]应用于文献[79]所提出的SAR-AD-BagNet模型时取得了更低的分类精度损失。基于改进模型结构的方法对特定对抗样本具有较好的防御效果,但需要对现有的模型结构进行修改,难以应用于常用模型中。

4.3 输出端防御

输出端防御也称为对抗检测,该任务旨在对待测样本是否具有对抗属性做出判断,其本质上是一个二分类任务。现有的对抗检测方法通常从统计分布的角度设计检测判据,通过查验模型输出的隐层特征、得分向量、置信度等来判断待测样本是否存在异常。

2016年,Hendrycks等人^[81]首次采用主成分分析法比较干净样本与对抗样本的差异,发现对抗样本的影响主要来源于具有较小贡献的奇异值。文献[82]利用训练集样本获取神经网络隐藏层特征的核密度先验,然后将待测样本在隐藏层中的分布特性和贝叶斯不确定性作为检测指标,通过逻辑回归的方式寻找合适的判据阈值。文献[83]提出使用样

本的局部内在维度(Local Intrinsic Dimensionality, LID)来区分对抗样本和干净样本,认为对抗样本的LID值比干净样本的LID值高。文献[84]提出使用样本的特征级别马氏距离值来检测对抗样本,认为干净样本在模型上的中间层特征满足类条件高斯分布(马氏距离值小),而对抗样本则远离类条件高斯分布之外(马氏距离值大)。文献[85]认为样本子空间是非线性的黎曼流形,并使用样本费希尔信息矩阵的奇异值作为特征设计检测判据。文献[86]提出基于最近邻居和贡献函数(Nearest Neighbors Influence Function, NNIF)联合判决的对抗检测方法,NNIF法认为干净样本的最近邻样本和最影响分类结果的样本应处于同一分布,对抗样本则不满足此关系。文献[87]经验性地探索了SAR图像对抗样本的扰动幅度与对抗检测算法性能之间的关系,指出检测算法性能随着对抗样本扰动幅度的下降而退化,并进一步指出测试样本与对抗样本的特征分布混淆会导致检测性能下降^[88],提出在训练过程中引入基于视角旋转和噪声的对比损失来改进现有马氏距离法和局部维度法对SAR图像对抗样本的检测性能。文献[89]探索了扰动区域约束对现有对抗检测算法的影响,在已知扰动区域先验的条件下利用对抗样本和干净样本的能量差异实现检测。

在雷达像智能识别对抗中,对抗检测技术赋予了深度模型感知恶意攻击的能力。然而,这一类防御方法无法判断目标的真实类别,难以单独胜任识别任务,通常作为分类模型中的子模块发挥作用。

表4汇总了目前经过雷达像数据集验证的对抗防御方法。

5 雷达像智能识别对抗的开放问题

总体来看,上文所述雷达像智能识别对抗方法主要集中在算法理论层面,所生成的雷达像对抗样本缺乏可靠的物理实现方式,与实际应用仍存在较

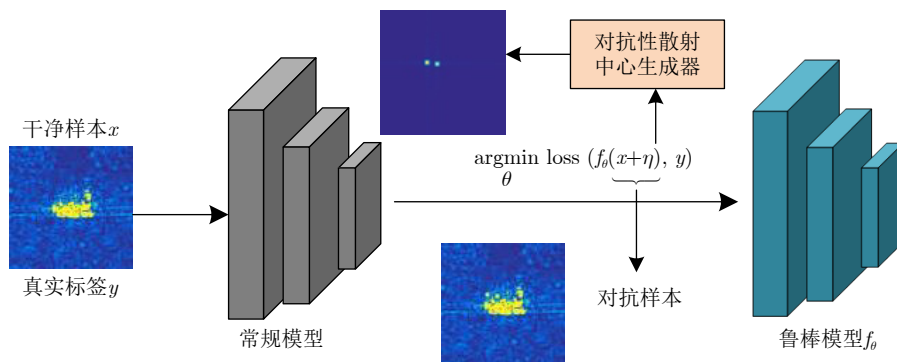


图 7 基于优化目标函数的防御方法

Fig. 7 Adversarial defense based on optimizing objective function

表4 雷达像识别对抗防御方法
Tab. 4 Summary of adversarial defense in radar image recognition

防御层级	文献	验证模型	数据集	可防御	优缺点
输入端	[73]	ResNet	UC-Merced ^[90]	FGSM/PGD ^[91] /CW ^[20] / DeepFool ^[19] / HopSkipJump ^[92] / Square ^[93]	仅需在数据端操作，影响干净样本识别率
	[74]	ResNet DenseNet MobileNet ShuffleNet A-ConvNet	MSTAR OpenSAR-Ship ^[59]	FGSM/PGD/DeepFool/ CW/SparseFool ^[94] / HopSkipJump/Square	
	[77]	VGG ResNet ShuffleNet	MSTAR ^[11]	FGSM ^[17] /PGD	
模型端	[47]	A-ConvNet	MSTAR SARBake ^[62]	PGD/SMGAA ^[47]	已知攻击类型时防御效果好，训练耗时
	[78]	自定义CNN	MSTAR	DeepFool	
	[79]	SAR-BagNet ^[80] ResNet	MSTAR	FGSM/PGD/ CW/DeepFool	
输出端	[87]	VGG ResNet DenseNet	MSTAR SARBake	FGSM/BIM ^[18] / CW/DeepFool	具有较好的可解释性，难以兼容主流模型
	[88]	ResNet	MSTAR	FGSM/BIM/ CW/DeepFool	
	[89]	VGG ResNet DenseNet	MSTAR SARBake	FGSM/BIM/ CW/DeepFool	

大差距，对HRRP、小样本等具体应用场景下的识别对抗问题还有待深入研究。以下5个开放问题是目前雷达像智能识别对抗领域值得重点关注的研究方向。

5.1 雷达HRRP的智能识别对抗

HRRP数据的处理过程相比于雷达二维像更加简单，在设计HRRP对抗扰动的物理实现方法时，无需考虑运动补偿、距离单元徙动矫正等。然而，由于HRRP目标具有姿态敏感、平移敏感等特点，在寻找HRRP对抗性距离单元时需要考虑姿态角以及方位角变化带来的影响。尽管目前尚未有针对HRRP对抗样本的防御方法被提出，但从信号形式来看，HRRP数据和语音数据均具有一维的形式，且不同时刻的信号均具有时序相关性。因此，可借鉴语音信号对抗防御中常用的音频扰动^[95]、音频压缩^[96]等方法，对HRRP对抗样本中的对抗性距离单元进行破坏或者重构后再进行识别，以达到防御目的。

5.2 小样本雷达像智能识别对抗

在非合作雷达目标识别中，识别方通常难以获取充足的样本用于模型训练，常常需要借助小样本学习方法。基于小样本学习的雷达像识别方法对预

训练模型和先验数据具有较强的依赖性，比如基于迁移学习的方法^[14]常常利用光学中的预训练模型作为骨干网络，基于元学习的方法^[15]则需要利用一批已知数据训练教师网络。当用户缺乏对预训练模型和先验数据的监管时，攻击者可采用投毒或木马的形式在预训练模型中植入后门。使用含有后门的预训练模型开展小样本学习，将导致用户模型难以收敛或者对中毒样本做出错误预测。然而，后门攻击需要攻击者干扰受害模型的训练阶段，其难点在于提高中毒样本的隐蔽性，要求触发器图案应尽可能难以察觉，比如采取不可见后门攻击^[97-99]的形式。此外，也可采用干净标签攻击^[100-102]的形式，使得中毒样本与干净样本仅在是否包含触发器上具有差别，而两者的训练标签一致，以此来进一步提高中毒样本隐蔽性。作为智能识别对抗的防御方，应重视训练数据的安全性筛查以及预训练模型的后门检测，可通过擦除重构的方式^[103]消除训练数据中的潜在触发器，也可通过反演触发器像素^[104]的方式确定所用模型是否包含后门。

5.3 针对SAR图像目标检测网络的对抗攻击

SAR图像智能目标检测和识别模型往往是一体

化的, 针对SAR图像目标检测网络的对抗攻击也是重要的研究方向。在光学图像领域针对Faster-RCNN^[105], YOLO^[106]等目标检测网络的对抗攻击方法主要有全局对抗扰动^[107]和局部对抗扰动^[108]两种, 攻击目的包括隐藏待检测目标、误导分类结果、干扰候选框生成等。SAR图像中背景区域与目标区域具有能量分布差异, 在设计针对SAR图像目标检测网络的对抗攻击方法时可以利用这一先验信息。全局扰动攻击方法需对整幅图像的每一像素点进行扰动, 应用于大尺度SAR图像时物理实现难度大, 而局部扰动的方法仅在目标区域生成对抗补丁, 更易于物理实现。

5.4 雷达像对抗样本与库外样本的区分

雷达目标识别模型在测试阶段既可能遇到对抗样本, 也可能遇到训练集中未出现过的类别样本, 即库外(Out Of Distribution, OOD)样本。从持续学习的角度来看, OOD样本可作为一种潜在的新样本加入训练库以提高模型的识别能力, 而对抗样本则带有恶意属性因而需要剔除。现有检测方法通常只能区分正常样本与异常样本, 无法区分异常样本属于OOD样本还是对抗样本。从标签属性来看, OOD样本的真实标签不属于训练集之中的任何一类, 而对抗样本的标签则是训练集中的某一指定类别, 利用样本在特征子空间中的流形分布与其在模型上的预测标签有望实现两种样本的区分。

5.5 雷达像对抗样本的物理实现

光学图像中常用的全局扰动对抗攻击方法在雷达领域缺乏语义可解释性, 在雷达回波信号中难以实现。在设计物理可实现雷达像对抗样本时需要考虑两个问题: 一是建立图像域对抗扰动与物理域目标电磁散射特性之间的联系, 比如借鉴几何绕射模型、属性散射中心等参数化模型将扰动区域约束为二面角、三面角、顶帽等真实结构的散射中心分布, 然后利用电磁超材料^[109]、干扰机等无源或有源的方式将这些“对抗性散射中心”调制到雷达回波信号中。二是研究能同时适用于不同分辨率雷达像的数字域对抗扰动生成方法, 提高对抗样本对不同分辨率雷达目标识别系统的攻击有效性。

参 考 文 献

- [1] ZHU Xiaoxiang, MONTAZERI S, ALI M, *et al.* Deep learning meets SAR: Concepts, models, pitfalls, and perspectives[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2021, 9(4): 143–172. doi: [10.1109/MGRS.2020.3046356](https://doi.org/10.1109/MGRS.2020.3046356).
- [2] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint arXiv: 1412. 6572, 2014.
- [3] 孙浩, 陈进, 雷琳, 等. 深度卷积神经网络图像识别模型对抗鲁棒性技术综述[J]. *雷达学报*, 2021, 10(4): 571–594. doi: [10.12000/JR21048](https://doi.org/10.12000/JR21048).
SUN Hao, CHEN Jin, LEI Lin, *et al.* Adversarial robustness of deep convolutional neural network-based image recognition models: A review[J]. *Journal of Radars*, 2021, 10(4): 571–594. doi: [10.12000/JR21048](https://doi.org/10.12000/JR21048).
- [4] XU Yonghao, BAI Tao, YU Weikang, *et al.* AI security for geoscience and remote sensing: Challenges and future trends[J]. *IEEE Geoscience and Remote Sensing Magazine*, 2023, 11(2): 60–85. doi: [10.1109/MGRS.2023.3272825](https://doi.org/10.1109/MGRS.2023.3272825).
- [5] CAO Dongsheng, HUANG Jianhua, YAN Jun, *et al.* Kernel *k*-nearest neighbor algorithm as a flexible SAR modeling tool[J]. *Chemometrics and Intelligent Laboratory Systems*, 2012, 114: 19–23. doi: [10.1016/j.chemolab.2012.01.008](https://doi.org/10.1016/j.chemolab.2012.01.008).
- [6] 袁莉, 刘宏伟, 保铮. 基于中心矩特征的雷达HRRP自动目标识别[J]. *电子学报*, 2004, 32(12): 2078–2081. doi: [10.3321/j.issn:0372-2112.2004.12.036](https://doi.org/10.3321/j.issn:0372-2112.2004.12.036).
YUAN Li, LIU Hongwei, and BAO Zheng. Automatic target recognition of radar HRRP based on central moments features[J]. *Acta Electronica Sinica*, 2004, 32(12): 2078–2081. doi: [10.3321/j.issn:0372-2112.2004.12.036](https://doi.org/10.3321/j.issn:0372-2112.2004.12.036).
- [7] SAEPULOH A, KOIKE K, and OMURA M. Applying Bayesian decision classification to Pi-SAR polarimetric data for detailed extraction of the geomorphologic and structural features of an active volcano[J]. *IEEE Geoscience and Remote Sensing Letters*, 2012, 9(4): 554–558. doi: [10.1109/LGRS.2011.2174611](https://doi.org/10.1109/LGRS.2011.2174611).
- [8] LI Min, ZHOU Gongjian, ZHAO Bin, *et al.* Sparse representation denoising for radar high resolution range profiling[J]. *International Journal of Antennas and Propagation*, 2014, 2014: 875895. doi: [10.1155/2014/875895](https://doi.org/10.1155/2014/875895).
- [9] CHEN Wenchao, CHEN Bo, PENG Xiaojun, *et al.* Tensor RNN with Bayesian nonparametric mixture for radar HRRP modeling and target recognition[J]. *IEEE Transactions on Signal Processing*, 2021, 69: 1995–2009. doi: [10.1109/TSP.2021.3065847](https://doi.org/10.1109/TSP.2021.3065847).
- [10] CHEN Sizhe, WANG Haipeng, XU Feng, *et al.* Target classification using the deep convolutional networks for SAR images[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2016, 54(8): 4806–4817. doi: [10.1109/TGRS.2016.2551720](https://doi.org/10.1109/TGRS.2016.2551720).
- [11] ROSS T D, WORRELL S W, VELTEN V J, *et al.* Standard SAR ATR evaluation experiments using the

- MSTAR public release data set[C]. SPIE 3370, Algorithms for Synthetic Aperture Radar Imagery, Orlando, USA, 1998: 566–573. doi: [10.1117/12.321859](https://doi.org/10.1117/12.321859).
- [12] PEI Jifang, HUANG Yulin, HUO Weibo, *et al.* SAR automatic target recognition based on multiview deep learning framework[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2018, 56(4): 2196–2210. doi: [10.1109/TGRS.2017.2776357](https://doi.org/10.1109/TGRS.2017.2776357).
- [13] SUN Yuanshuang, WANG Yinghua, LIU Hongwei, *et al.* SAR target recognition with limited training data based on angular rotation generative network[J]. *IEEE Geoscience and Remote Sensing Letters*, 2020, 17(11): 1928–1932. doi: [10.1109/LGRS.2019.2958379](https://doi.org/10.1109/LGRS.2019.2958379).
- [14] HUANG Zhongling, PAN Zongxu, and LEI Bin. What, where, and how to transfer in SAR target recognition based on deep CNNs[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2020, 58(4): 2324–2336. doi: [10.1109/TGRS.2019.2947634](https://doi.org/10.1109/TGRS.2019.2947634).
- [15] FU Kun, ZHANG Tengfei, ZHANG Yue, *et al.* Few-shot SAR target classification via metalearning[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 2000314. doi: [10.1109/TGRS.2021.3058249](https://doi.org/10.1109/TGRS.2021.3058249).
- [16] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al.* Intriguing properties of neural networks[C]. 2nd International Conference on Learning Representations, Banff, Canada, 2014.
- [17] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2015: 1050.
- [18] KURAKIN A, GOODFELLOW I J, and BENGIO S. Adversarial Examples in the Physical World[M]. YAMPOLSKIY R V. Artificial Intelligence Safety and Security. New York: Chapman and Hall/CRC, 2018: 99–112.
- [19] MOOSAVI-DEZFOOLI S M, FAWZI A, and FROSSARD P. DeepFool: A simple and accurate method to fool deep neural networks[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2574–2582. doi: [10.1109/CVPR.2016.282](https://doi.org/10.1109/CVPR.2016.282).
- [20] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks[C]. 2017 IEEE Symposium on Security and Privacy (SP), San Jose, USA, 2017: 39–57. doi: [10.1109/SP.2017.49](https://doi.org/10.1109/SP.2017.49).
- [21] PAPERNOT N, MCDANIEL P, JHA S, *et al.* The limitations of deep learning in adversarial settings[C]. 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbruecken, Germany, 2016: 372–387. doi: [10.1109/EuroSP.2016.36](https://doi.org/10.1109/EuroSP.2016.36).
- [22] SU Jiawei, VARGAS D V, and SAKURAI K. One pixel attack for fooling deep neural networks[J]. *IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828–841. doi: [10.1109/TEVC.2019.2890858](https://doi.org/10.1109/TEVC.2019.2890858).
- [23] POURSAEED O, KATSMAN I, GAO Bicheng, *et al.* Generative adversarial perturbations[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 4422–4431. doi: [10.1109/CVPR.2018.00465](https://doi.org/10.1109/CVPR.2018.00465).
- [24] DU Chuan and ZHANG Lei. Adversarial attack for SAR target recognition based on UNet-generative adversarial network[J]. *Remote Sensing*, 2021, 13(21): 4358. doi: [10.3390/rs13214358](https://doi.org/10.3390/rs13214358).
- [25] XIAO Chaowei, LI Bo, ZHU Junyan, *et al.* Generating adversarial examples with adversarial networks[C]. 27th International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 2018: 3905–3911.
- [26] ILYAS A, ENGSTROM L, ATHALYE A, *et al.* Black-box adversarial attacks with limited queries and information[C]. 35th International Conference on Machine Learning, Stockholm, Sweden, 2018: 2142–2151.
- [27] GUO Chuan, GARDNER J R, YOU Yurong, *et al.* Simple black-box adversarial attacks[C]. 36th International Conference on Machine Learning, Long Beach, USA, 2019: 2484–2493.
- [28] TASHIRO Y, SONG Y, ERMON S. Diversity can be transferred: Output diversification for white-and black-box attacks[C]. The 34th International Conference on Neural Information Processing Systems. 2020: 4536–4548.
- [29] GUO Wei, TONDI B, and BARNI M. A master key backdoor for universal impersonation attack against DNN-based face verification[J]. *Pattern Recognition Letters*, 2021, 144: 61–67. doi: [10.1016/j.patrec.2021.01.009](https://doi.org/10.1016/j.patrec.2021.01.009).
- [30] GU Tianyu, LIU Kang, DOLAN-GAVITT B, *et al.* BadNets: Evaluating backdooring attacks on deep neural networks[J]. *IEEE Access*, 2019, 7: 47230–47244. doi: [10.1109/ACCESS.2019.2909068](https://doi.org/10.1109/ACCESS.2019.2909068).
- [31] BREWER E, LIN J, and RUNFOLA D. Susceptibility & defense of satellite image-trained convolutional networks to backdoor attacks[J]. *Information Sciences*, 2022, 603: 244–261. doi: [10.1016/j.ins.2022.05.004](https://doi.org/10.1016/j.ins.2022.05.004).
- [32] ISLAM S, BADSHA S, KHALIL I, *et al.* A triggerless backdoor attack and defense mechanism for intelligent task offloading in multi-UAV systems[J]. *IEEE Internet of Things Journal*, 2023, 10(7): 5719–5732. doi: [10.1109/JIOT.2022.3172936](https://doi.org/10.1109/JIOT.2022.3172936).
- [33] LI Haifeng, HUANG Haikuo, CHEN Li, *et al.* Adversarial examples for CNN-based SAR image classification: An

- experience study[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2021, 14: 1333–1347. doi: [10.1109/JSTARS.2020.3038683](https://doi.org/10.1109/JSTARS.2020.3038683).
- [34] 周隽凡, 孙浩, 雷琳, 等. SAR图像稀疏对抗攻击[J]. 信号处理, 2021, 37(9): 1633–1643. doi: [10.16798/j.issn.1003-0530.2021.09.007](https://doi.org/10.16798/j.issn.1003-0530.2021.09.007).
ZHOU Junfan, SUN Hao, LEI Lin, *et al.* Sparse adversarial attack of SAR image[J]. *Journal of Signal Processing*, 2021, 37(9): 1633–1643. doi: [10.16798/j.issn.1003-0530.2021.09.007](https://doi.org/10.16798/j.issn.1003-0530.2021.09.007).
- [35] WANG Lulu, WANG Xiaolei, MA Shixin, *et al.* Universal adversarial perturbation of SAR images for deep learning based target classification[C]. 2021 IEEE 4th International Conference on Electronics Technology (ICET), Chengdu, China, 2021: 1272–1276. doi: [10.1109/ICET51757.2021.9450944](https://doi.org/10.1109/ICET51757.2021.9450944).
- [36] DU Chuan, HUO Chaoying, ZHANG Lei, *et al.* Fast C&W: A fast adversarial attack algorithm to fool SAR target recognition with deep convolutional neural networks[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 4010005. doi: [10.1109/LGRS.2021.3058011](https://doi.org/10.1109/LGRS.2021.3058011).
- [37] ZHANG Fan, MENG Tianying, XIANG Deliang, *et al.* Adversarial deception against SAR target recognition network[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2022, 15: 4507–4520. doi: [10.1109/JSTARS.2022.3179171](https://doi.org/10.1109/JSTARS.2022.3179171).
- [38] 徐延杰, 孙浩, 雷琳, 等. 基于稀疏差分协同进化的多源遥感场景分类攻击[J]. 信号处理, 2021, 37(7): 1164–1170. doi: [10.16798/j.issn.1003-0530.2021.07.005](https://doi.org/10.16798/j.issn.1003-0530.2021.07.005).
XU Yanjie, SUN Hao, LEI Lin, *et al.* Multi-source remote sensing classification attack based on sparse differential coevolution[J]. *Journal of Signal Processing*, 2021, 37(7): 1164–1170. doi: [10.16798/j.issn.1003-0530.2021.07.005](https://doi.org/10.16798/j.issn.1003-0530.2021.07.005).
- [39] RONNEBERGER O, FISCHER P, and BROX T. U-net: Convolutional networks for biomedical image segmentation[C]. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention, Munich, Germany, 2015: 234–241. doi: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [40] PENG Bowen, PENG Bo, YONG Shaowei, *et al.* An empirical study of fully black-box and universal adversarial attack for SAR target recognition[J]. *Remote Sensing*, 2022, 14(16): 4017. doi: [10.3390/rs14164017](https://doi.org/10.3390/rs14164017).
- [41] DANG Xunwang, YAN Hua, HU Liping, *et al.* SAR image adversarial samples generation based on parametric model[C]. 2021 International Conference on Microwave and Millimeter Wave Technology (ICMMT), Nanjing, China, 2021: 1–3. doi: [10.1109/ICMMT52847.2021.9618140](https://doi.org/10.1109/ICMMT52847.2021.9618140).
- [42] DU M, BI D, DU M, *et al.* Local aggregative attack on SAR image classification models[J]. *Authorea Preprints*, 2022. doi: [10.22541/au.165633740.01163731/v1](https://doi.org/10.22541/au.165633740.01163731/v1).
- [43] MENG Tianying, ZHANG Fan, and MA Fei. A target-region-based SAR ATR adversarial deception method[C]. 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 2022: 142–146. doi: [10.1109/ICSIP55141.2022.9887044](https://doi.org/10.1109/ICSIP55141.2022.9887044).
- [44] PENG Bowen, PENG Bo, ZHOU Jie, *et al.* Speckle-variant attack: Toward transferable adversarial attack to SAR target recognition[J]. *IEEE Geoscience and Remote Sensing Letters*, 2022, 19: 4509805. doi: [10.1109/LGRS.2022.3184311](https://doi.org/10.1109/LGRS.2022.3184311).
- [45] GERRY M J, POTTER L C, GUPTA I J, *et al.* A parametric model for synthetic aperture radar measurements[J]. *IEEE Transactions on Antennas and Propagation*, 1999, 47(7): 1179–1188. doi: [10.1109/8.785750](https://doi.org/10.1109/8.785750).
- [46] ZHOU Junfan, FENG Sijia, SUN Hao, *et al.* Attributed scattering center guided adversarial attack for DCNN SAR target recognition[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 4001805. doi: [10.1109/LGRS.2023.3235051](https://doi.org/10.1109/LGRS.2023.3235051).
- [47] PENG Bowen, PENG Bo, ZHOU Jie, *et al.* Scattering model guided adversarial examples for SAR target recognition: Attack and defense[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2022, 60: 5236217. doi: [10.1109/TGRS.2022.3213305](https://doi.org/10.1109/TGRS.2022.3213305).
- [48] QIN Weibo, LONG Bo, and WANG Feng. SCMA: A scattering center model attack on CNN-SAR target recognition[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 4003305. doi: [10.1109/LGRS.2023.3253189](https://doi.org/10.1109/LGRS.2023.3253189).
- [49] LIU Hongwei, JIU Bo, LI Fei, *et al.* Attributed scattering center extraction algorithm based on sparse representation with dictionary refinement[J]. *IEEE Transactions on Antennas and Propagation*, 2017, 65(5): 2604–2614. doi: [10.1109/TAP.2017.2673764](https://doi.org/10.1109/TAP.2017.2673764).
- [50] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[C]. 3rd International Conference on Learning Representations, San Diego, USA, 2014.
- [51] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, *et al.* Deep residual learning for image recognition[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 770–778. doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [52] HUANG Gao, LIU Zhuang, VAN DER MAATEN L, *et al.* Densely connected convolutional networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition,

- Honolulu, USA, 2017: 2261–2269. doi: [10.1109/CVPR.2017.243](https://doi.org/10.1109/CVPR.2017.243).
- [53] SZEGEDY C, LIU Wei, JIA Yangqing, *et al.* Going deeper with convolutions[C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition, Boston, USA, 2015: 1–9. doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [54] SZEGEDY C, VANHOUCKE V, IOFFE S, *et al.* Rethinking the inception architecture for computer vision[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 2818–2826. doi: [10.1109/CVPR.2016.308](https://doi.org/10.1109/CVPR.2016.308).
- [55] SANDLER M, HOWARD A, ZHU Menglong, *et al.* MobileNetV2: Inverted residuals and linear bottlenecks[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 4510–4520. doi: [10.1109/CVPR.2018.00474](https://doi.org/10.1109/CVPR.2018.00474).
- [56] KRIZHEVSKY A, SUTSKEVER I, and HINTON G E. ImageNet classification with deep convolutional neural networks[J]. *Communications of the ACM*, 2017, 60(6): 84–90. doi: [10.1145/3065386](https://doi.org/10.1145/3065386).
- [57] SZEGEDY C, IOFFE S, VANHOUCKE V, *et al.* Inception-v4, inception-ResNet and the impact of residual connections on learning[C]. Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017: 4278–4284.
- [58] SCHMITT M, HUGHES L H, and ZHU X X. The SEN1-2 dataset for deep learning in Sar-optical data fusion[J]. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, 2018, 4: 141–146. doi: [10.5194/isprs-annals-IV-1-141-2018](https://doi.org/10.5194/isprs-annals-IV-1-141-2018).
- [59] HUANG Lanqing, LIU Bin, LI Boying, *et al.* OpenSARShip: A dataset dedicated to Sentinel-1 ship interpretation[J]. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 2018, 11(1): 195–208. doi: [10.1109/JSTARS.2017.2755672](https://doi.org/10.1109/JSTARS.2017.2755672).
- [60] ZHU Xiaoxiang, HU Jingliang, QIU Chungping, *et al.* So2Sat LCZ42: A benchmark data set for the classification of global local climate zones [Software and Data Sets][J]. *IEEE Geoscience and Remote Sensing Magazine*, 2020, 8(3): 76–89. doi: [10.1109/MGRS.2020.2964708](https://doi.org/10.1109/MGRS.2020.2964708).
- [61] MALMGREN-HANSEN D, KUSK A, DALL J, *et al.* Improving SAR automatic target recognition models with transfer learning from simulated data[J]. *IEEE Geoscience and remote sensing Letters*, 2017, 14(9): 1484–1488. doi: [10.1109/LGRS.2017.2717486](https://doi.org/10.1109/LGRS.2017.2717486).
- [62] MALMGREN-HANSEN D and NOBEL-JØRGENSEN M. Convolutional neural networks for SAR image segmentation[C]. 2015 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Abu Dhabi, United Arab Emirates, 2015: 231–236. doi: [10.1109/ISSPIT.2015.7394333](https://doi.org/10.1109/ISSPIT.2015.7394333).
- [63] YUAN Yijun, WAN Jinwei, and CHEN Bo. Robust attack on deep learning based radar HRRP target recognition[C]. 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019: 704–707. doi: [10.1109/APSIPAASC47483.2019.9023266](https://doi.org/10.1109/APSIPAASC47483.2019.9023266).
- [64] 万锦伟. 基于深度网络的HRRP目标识别与对抗攻击研究[D]. [博士学位论文], 西安电子科技大学, 2020. doi: [10.27389/d.cnki.gxadu.2020.000125](https://doi.org/10.27389/d.cnki.gxadu.2020.000125).
- WAN Jinwei. Research on HRRP target recognition and adversarial attacks based on deep neural networks[D]. [Ph. D. dissertation], Xidian University, 2020. doi: [10.27389/d.cnki.gxadu.2020.000125](https://doi.org/10.27389/d.cnki.gxadu.2020.000125).
- [65] HUANG Teng, CHEN Yongfeng, YAO Bingjian, *et al.* Adversarial attacks on deep-learning-based radar range profile target recognition[J]. *Information Sciences*, 2020, 531: 159–176. doi: [10.1016/j.ins.2020.03.066](https://doi.org/10.1016/j.ins.2020.03.066).
- [66] DU Chuan, CONG Yulai, ZHANG Lei, *et al.* A practical deceptive jamming method based on vulnerable location awareness adversarial attack for radar HRRP target recognition[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 2410–2424. doi: [10.1109/TIFS.2022.3170275](https://doi.org/10.1109/TIFS.2022.3170275).
- [67] GAO Fei, HUANG Teng, WANG Jun, *et al.* A novel multi-input bidirectional LSTM and HMM based approach for target recognition from multi-domain radar range profiles[J]. *Electronics*, 2019, 8(5): 535. doi: [10.3390/electronics8050535](https://doi.org/10.3390/electronics8050535).
- [68] YANG Yuzhe, ZHANG Guo, XU Zhi, *et al.* ME-Net: Towards effective adversarial robustness with matrix estimation[C]. 36th International Conference on Machine Learning, Long Beach, USA, 2019: 7025–7034.
- [69] WANG Yutong, ZHANG Wenwen, SHEN Tianyu, *et al.* Binary thresholding defense against adversarial attacks[J]. *Neurocomputing*, 2021, 445: 61–71. doi: [10.1016/j.neucom.2021.03.036](https://doi.org/10.1016/j.neucom.2021.03.036).
- [70] LeCun Y. The MNIST database of handwritten digits[EB/OL]. <http://yann.lecun.com/exdb/mnist/>, 1998.
- [71] MUSTAFA A, KHAN S H, HAYAT M, *et al.* Image super-resolution as a defense against adversarial attacks[J]. *IEEE Transactions on Image Processing*, 2020, 29:

- 1711–1724. doi: [10.1109/TIP.2019.2940533](https://doi.org/10.1109/TIP.2019.2940533).
- [72] AGARWAL A, SINGH R, VATSA M, *et al.* Image transformation-based defense against adversarial perturbation on deep learning models[J]. *IEEE Transactions on Dependable and Secure Computing*, 2021, 18(5): 2106–2121. doi: [10.1109/TDSC.2020.3027183](https://doi.org/10.1109/TDSC.2020.3027183).
- [73] 孙浩, 徐延杰, 陈进, 等. 基于自监督对比学习的深度神经网络对抗鲁棒性提升[J]. *信号处理*, 2021, 37(6): 903–911. doi: [10.16798/j.issn.1003-0530.2021.06.001](https://doi.org/10.16798/j.issn.1003-0530.2021.06.001).
- SUN Hao, XU Yanjie, CHEN Jin, *et al.* Self-supervised contrastive learning for improving the adversarial robustness of deep neural networks[J]. *Journal of Signal Processing*, 2021, 37(6): 903–911. doi: [10.16798/j.issn.1003-0530.2021.06.001](https://doi.org/10.16798/j.issn.1003-0530.2021.06.001).
- [74] XU Yanjie, SUN Hao, CHEN Jin, *et al.* Adversarial self-supervised learning for robust SAR target recognition[J]. *Remote Sensing*, 2021, 13(20): 4158. doi: [10.3390/rs13204158](https://doi.org/10.3390/rs13204158).
- [75] SONG Chuanbiao, HE Kun, LIN Jiadong, *et al.* Robust local features for improving the generalization of adversarial training[C]. 8th International Conference on Learning Representations, Addis Ababa, Ethiopia, 2020.
- [76] ZHANG Hongyang, YU Yaodong, JIAO Jiantao, *et al.* Theoretically principled trade-off between robustness and accuracy[C]. 36th International Conference on Machine Learning, Long Beach, USA, 2019: 7472–7482.
- [77] INKAWHICH N, DAVIS E, MAJUMDER U, *et al.* Advanced techniques for robust SAR ATR: Mitigating noise and phase errors[C]. 2020 IEEE International Radar Conference (RADAR), Washington, USA, 2020: 844–849. doi: [10.1109/RADAR42522.2020.9114784](https://doi.org/10.1109/RADAR42522.2020.9114784).
- [78] WAGNER S, PANATI C, and BRÜGGENWIRTH S. Fool the COOL-on the robustness of deep learning SAR ATR systems[C]. 2021 IEEE Radar Conference (RadarConf21), Atlanta, USA, 2021: 1–6. doi: [10.1109/RadarConf2147009.2021.9455231](https://doi.org/10.1109/RadarConf2147009.2021.9455231).
- [79] LI Peng, HU Xiaowei, FENG Cunqian, *et al.* SAR-AD-BagNet: An interpretable model for SAR image recognition based on adversarial defense[J]. *IEEE Geoscience and Remote Sensing Letters*, 2023, 20: 4000505. doi: [10.1109/LGRS.2022.3230243](https://doi.org/10.1109/LGRS.2022.3230243).
- [80] LI Peng, FENG Cunqian, HU Xiaowei, *et al.* SAR-BagNet: An ante-hoc interpretable recognition model based on deep network for SAR image[J]. *Remote Sensing*, 2022, 14(9): 2150. doi: [10.3390/rs14092150](https://doi.org/10.3390/rs14092150).
- [81] HENDRYCKS D and GIMPEL K. Early methods for detecting adversarial images[C]. 5th International Conference on Learning Representations, Toulon, France, 2017.
- [82] FEINMAN R, CURTIN R R, SHINTRE S, *et al.* Detecting adversarial samples from artifacts[OL]. <https://arxiv.org/abs/1703.00410>.
- [83] MA Xingjun, LI Bo, WANG Yisen, *et al.* Characterizing adversarial subspaces using local intrinsic dimensionality[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018.
- [84] LEE K, LEE K, LEE H, *et al.* A simple unified framework for detecting out-of-distribution samples and adversarial attacks[C]. 32nd International Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 7167–7177.
- [85] ZHAO Chenxiao, FLETCHER P T, YU Mixue, *et al.* The adversarial attack and detection under the fisher information metric[C]. Thirty-Third AAAI Conference on Artificial Intelligence, Honolulu, USA, 2019: 5869–5876. doi: [10.1609/aaai.v33i01.33015869](https://doi.org/10.1609/aaai.v33i01.33015869).
- [86] COHEN G, SAPIRO G, and GIRYES R. Detecting adversarial samples using influence functions and nearest neighbors[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 14441–14450. doi: [10.1109/CVPR42600.2020.01446](https://doi.org/10.1109/CVPR42600.2020.01446).
- [87] ZHANG Zhiwei, LIU Shuowei, GAO Xunzhang, *et al.* An empirical study towards SAR adversarial examples[C]. 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi'an, China, 2022: 127–132. doi: [10.1109/ICICML57342.2022.10009880](https://doi.org/10.1109/ICICML57342.2022.10009880).
- [88] ZHANG Zhiwei, LIU Shuowei, GAO Xunzhang, *et al.* Improving adversarial detection methods for SAR image via joint contrastive cross-entropy training[C]. 4th International Academic Exchange Conference on Science and Technology Innovation (IAECST), Guangzhou, China, 2022: 1107–1110. doi: [10.1109/IAECST57965.2022.10061932](https://doi.org/10.1109/IAECST57965.2022.10061932).
- [89] ZHANG Zhiwei, GAO Xunzhang, LIU Shuowei, *et al.* Energy-based adversarial example detection for SAR images[J]. *Remote Sensing*, 2022, 14(20): 5168. doi: [10.3390/rs14205168](https://doi.org/10.3390/rs14205168).
- [90] YANG Yi and NEWSAM S. Bag-of-visual-words and spatial extensions for land-use classification[C]. 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, USA, 2010:

- 270–279. doi: [10.1145/1869790.1869829](https://doi.org/10.1145/1869790.1869829).
- [91] MADRY A, MAKELOV A, SCHMIDT L, *et al.* Towards deep learning models resistant to adversarial attacks[C]. 6th International Conference on Learning Representations, Vancouver, Canada, 2018.
- [92] CHEN Jianbo, JORDAN M I, and WAINWRIGHT M J. HopSkipJumpAttack: A query-efficient decision-based attack[C]. 2020 IEEE Symposium on Security and Privacy (SP), San Francisco, USA, 2020: 1277–1294. doi: [10.1109/SP40000.2020.00045](https://doi.org/10.1109/SP40000.2020.00045).
- [93] ANDRIUSHCHENKO M, CROCE F, FLAMMARION N, *et al.* Square attack: A query-efficient black-box adversarial attack via random search[C]. 16th European Conference on Computer Vision, Glasgow, UK, 2020: 484–501. doi: [10.1007/978-3-030-58592-1_29](https://doi.org/10.1007/978-3-030-58592-1_29).
- [94] MODAS A, MOOSAVI-DEZFOOLI S M, and FROSSARD P. SparseFool: A few pixels make a big difference[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 9079–9088. doi: [10.1109/CVPR.2019.00930](https://doi.org/10.1109/CVPR.2019.00930).
- [95] YUAN Xuejing, CHEN Yuxuan, ZHAO Yue, *et al.* Commandersong: A systematic approach for practical adversarial voice recognition[C]. 27th USENIX Conference on Security Symposium, Baltimore, USA, 2018: 49–64.
- [96] DAS N, SHANBHOGUE M, CHEN S T, *et al.* ADAGIO: Interactive experimentation with adversarial attack and defense for audio[C]. Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Dublin, Ireland, 2019: 677–681. doi: [10.1007/978-3-030-10997-4_50](https://doi.org/10.1007/978-3-030-10997-4_50).
- [97] DOAN K, LAO Y, and LI P. Backdoor attack with imperceptible input and latent modification[J]. *Advances in Neural Information Processing Systems*, 2021, 34: 18944–18957.
- [98] BAGDASARYAN E and SHMATIKOV V. Blind backdoors in deep learning models[C]. 30th USENIX Security Symposium, 2021: 1505–1521.
- [99] DOAN K, LAO Yingjie, ZHAO Weijie, *et al.* LIRA: Learnable, imperceptible and robust backdoor attacks[C]. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, Canada, 2021: 11946–11956. doi: [10.1109/ICCV48922.2021.01175](https://doi.org/10.1109/ICCV48922.2021.01175).
- [100] SAHA A, SUBRAMANYA A, and PIRSIIVASH H. Hidden trigger backdoor attacks[C]. 34th AAAI Conference on Artificial Intelligence, New York, USA, 2020: 11957–11965. doi: [10.1609/aaai.v34i07.6871](https://doi.org/10.1609/aaai.v34i07.6871).
- [101] SHUMAILOV I, SHUMAYLOV Z, KAZHDAN D, *et al.* Manipulating SGD with data ordering attacks[C]. 34th International Conference on Neural Information Processing Systems, 2021: 18021–18032.
- [102] SOURI H, FOWL L, CHELLAPPA R, *et al.* Sleeper agent: Scalable hidden trigger backdoors for neural networks trained from scratch[J]. *Advances in Neural Information Processing Systems*, 2022, 35: 19165–19178.
- [103] DOAN B G, ABBASNEJAD E, and RANASINGHE D C. Februus: Input purification defense against Trojan attacks on deep neural network systems[C]. Annual Computer Security Applications Conference, Austin, USA, 2020: 897–912. doi: [10.1145/3427228.3427264](https://doi.org/10.1145/3427228.3427264).
- [104] WANG Bolun, YAO Yuanshun, SHAN S, *et al.* Neural cleanse: Identifying and mitigating backdoor attacks in neural networks[C]. 2019 IEEE Symposium on Security and Privacy (SP), San Francisco, USA, 2019: 707–723. doi: [10.1109/SP.2019.00031](https://doi.org/10.1109/SP.2019.00031).
- [105] GIRSHICK R. Fast r-CNN[C]. 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 2015: 1440–1448. doi: [10.1109/ICCV.2015.169](https://doi.org/10.1109/ICCV.2015.169).
- [106] REDMON J, DIVVALA S, GIRSHICK R, *et al.* You only look once: Unified, real-time object detection[C]. 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, USA, 2016: 779–788. doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91).
- [107] CHOW K H, LIU Ling, LOPER M, *et al.* Adversarial objectness gradient attacks in real-time object detection systems[C]. 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, USA, 2020: 263–272. doi: [10.1109/TPS-ISA50397.2020.00042](https://doi.org/10.1109/TPS-ISA50397.2020.00042).
- [108] WANG Yajie, LV Haoran, KUANG Xiaohui, *et al.* Towards a physical-world adversarial patch for blinding object detection models[J]. *Information Sciences*, 2021, 556: 459–471. doi: [10.1016/j.ins.2020.08.087](https://doi.org/10.1016/j.ins.2020.08.087).
- [109] 张磊, 陈晓晴, 郑熠宁, 等. 电磁超表面与信息超表面[J]. *电波科学学报*, 2021, 36(6): 817–828. doi: [10.12265/j.cjors.2021218](https://doi.org/10.12265/j.cjors.2021218).
- ZHANG Lei, CHEN Xiaoqing, ZHENG Yining, *et al.* Electromagnetic metasurfaces and information metasurfaces[J]. *Chinese Journal of Radio Science*, 2021, 36(6): 817–828. doi: [10.12265/j.cjors.2021218](https://doi.org/10.12265/j.cjors.2021218).

作者简介

高勋章，研究员，博士生导师，主要研究方向为雷达目标识别、智能信息处理。

张志伟，博士生，主要研究方向为雷达目标识别、智能对抗攻击与防御。

刘 梅，硕士生，主要研究方向为智能感知与处理、雷达目标识别。

龚政辉，助理研究员，主要研究方向为雷达对抗、雷达抗干扰、雷达通信一体化。

黎 湘，教授，博士生导师，主要研究方向为雷达目标特性与识别。

(责任编辑：于青)