ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 39,No. 1,Jan. 2024,pp. 106－119
DOI:10. 16337/j. 1004-9037. 2024. 01. 010

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

面向雷达图像分类模型的两步式对抗样本检测技术

王　见，张赛楠，陈　芳

（南京航空航天大学计算机科学与技术学院，南京 211106）

摘　要：深度学习技术极大地提高了雷达图像目标分类的精度，但由于神经网络自身的脆弱性使得雷达图像分类系统的安全性受到威胁。本文对雷达对抗样本的攻击性及雷达对抗样本与原始样本在频率域上的差异性进行了分析，并在此基础上，提出了两步式雷达对抗样本检测技术来提升雷达分类模型的安全性。首先基于频率域对输入的雷达图像进行第1步对抗样本检测，分离出对抗样本，然后将剩下的图像分别送入到一个经过对抗训练的模型和一个未经过对抗训练的模型进行第2次对抗样本检测。通过这种两步式的检测方法，可以有效地检测出对抗样本，检测成功率不低于95.73%，有效提升了雷达分类模型的安全性。

关键词：合成孔径雷达；深度神经网络；对抗样本；频率域转换；模型安全性

中图分类号：TP391　　**文献标志码**：A

A Two-Step Adversarial Sample Detection Technique for SAR Image Classification

WANG Jian, ZHANG Sainan, CHEN Fang

(College of Computer Science and Technology, Nanjing University of Aeronautics & Astronautics, Nanjing 211106, China)

Abstract: Deep learning techniques have greatly improved the classification accuracy of synthetic aperture radar (SAR) images target, but the security of SAR image classification systems is threatened by the inherent vulnerability of neural networks. In this paper, we analyze the aggressiveness of SAR adversarial samples, and the difference between SAR adversarial examples and original examples in the frequency domain. With the analysis results, a two-step SAR adversarial samples detection technique is proposed to improve the security of SAR classification models. The first step of adversarial sample detection is performed on the input image based on the frequency domain analysis to separate the adversarial samples. Then, the remaining images are fed into an adversarial trained model and an untrained model to complete the second step of adversarial sample detection. By using this two-step detection method, the adversarial samples can be effectively detected with a detection success rate of no less than 95.73%, effectively improving the security of the SAR classification model.

Key words: synthetic aperture radar (SAR); deep neural network; adversarial samples; frequency domain transform; model security

收稿日期：2023-03-07；**修订日期**：2023-06-16

引　言

雷达图像在监测、测绘和军事等方面有着广泛的用途[1]。由于雷达成像是由地物散射回波接收到的相干信号叠加而成,使得生成的雷达图像存在大量由空间相关与信号相关的乘性噪声,即散斑噪声[2],这是雷达图像分析的一个主要障碍[3]。传统的雷达图像目标识别会先对雷达图像进行去噪,然后手动提取特征进行分类识别[4]。文献[5]提出了一种基于压缩感知和支持向量机决策级融合的目标识别算法,提高了SAR变形目标的识别率。文献[6-7]在提取雷达图像特征之前,采取滤波的方式分离噪声,减少散斑噪声的影响。但这些方法效率较低且局限性较高,分类性能难以达到较高的准确度[4]。

随着深度学习在计算机视觉领域的不断发展,深度神经网络在图像目标分类中取得了很好的效果[8]。基于深度神经网络的目标分类模型可以对图像特征进行自动学习,将低层特征抽象化得到高层次的特征,通过层次的不断增加,原始数据的表现形式也越来越抽象,提取到的特征更能表现目标所属类的本质,从而提高图像的分类和识别精度,有效克服了传统分类方法中繁琐的预处理与精度较低的缺陷。文献[9]基于目标特征的稀疏性,提出了一种稀疏先验引导卷积神经网络习的SAR目标识别方法,提取出更具区分度的特征,提升了SAR图像目标识别的精度。文献[10]将卷积注意力和胶囊网络进行结合,解决少样本SAR目标识别准确率低的问题。将深度神经网络应用到雷达图像目标分类任务后,有效解决了传统方法中的缺点。但是,深度神经网络容易受到攻击,针对一个训练好的模型,对图像添加极小的扰动后,模型的分类性能会急速下降[11-13],而添加扰动后的图像对于人类来说极难察觉[14-15]。深度神经网络的脆弱性对安全性要求较高的雷达分类系统是一个极大的安全隐患[16]。

针对雷达图像分类模型的安全性问题,本文结合雷达图像频域转换和对抗训练,提出了一种基于雷达图像分类模型的两步式对抗样本检测方法。首先基于雷达干净样本与对抗样本在频率域中的差异进行第1步对抗样本检测,然后使用基于对抗训练的方法对剩余数据进行第2步对抗样本检测。实验结果表明,本文所提出的两步式检测方法可以有效地检测出雷达对抗样本,检测成功率不低于95.73%,平均检测成功率为97.93%,提升了雷达图像分类模型的安全性。

1　相关工作

1.1　对抗样本

对抗样本是指在原始样本中人为添加微小噪声生成的样本,这些对抗样本从视觉上不易被察觉,但会使训练好的深度模型出错,威胁基于深度神经网络的应用。向干净样本x^{clean}中添加人为设计的极小的扰动ε产生对抗样本x^{adv},对抗样本x^{adv}会使训练好的模型f以很高的置信度输出错误类别。

图1(a)中的不同形状代表了从人类视觉角度观察到的分类结果,不同颜色代表了不同分类模型得到的分类结果。在非目标攻击下,样本穿过距离相对较近的决策边界使得模型分类出错。图1(a)中的黄色三角形,在添加扰动后仍可以看出其类别为1,但分类模型将其从类别1识别为类别2。图1(b)在MSTAR[17]数据集上,通过使用t-分布邻域嵌入算法(t-distributed stochastic neighbor embedding,t-SNE)方法展示ResNet18[18]神经网络被快速梯度下降(Fast gradient sign method,FGSM)[19]攻击算法攻击所产生的对抗样本(三角形)和原始样本(圆点)的可视化图,从图1可以看出,对抗样本穿过决策边界嵌入到别的类别之中,导致模型分类出错。

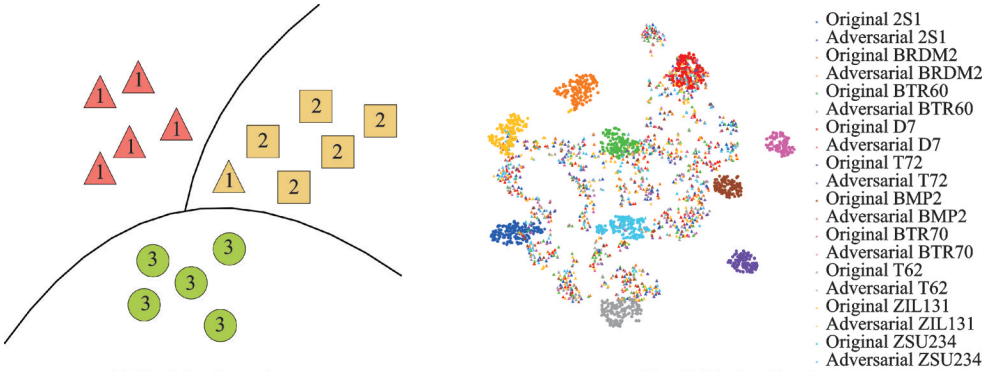


图1 决策边界与MSTAR数据集中对抗样本与干净样本特征可视化

Fig.1 Decision boundary and visualization of adversarial and clean sample features in the MSTAR dataset

1.2 攻击算法

自对抗样本[6]被发现以来，出现了很多的攻击类型，大致可以分为白盒攻击和黑盒攻击。在获得模型的参数与信息后进行的攻击被称为白盒攻击。在对模型的信息一无所知的情况下进行的攻击称为黑盒攻击。现阶段有很多攻击算法生成对抗样本，这些算法通过在原始图像中添加微小的扰动来使分类器出错，达到攻击的目的。如基于梯度的攻击方法：FGSM[14]，多步迭代的I-FGSM[19]，投影梯度下降(Projected gradient descent，PGD)[20]；基于优化的攻击C&W[19]等，这些都属于白盒攻击的范畴。黑盒攻击中常用的算法有扰动随机逼近算法(Simultaneous perturbation stochastic approximation，SPSA)[21]、零阶优化(Zeroth order optimization，ZOO)[22]等。本文使用FGSM[14]、PGD[20]和SPSA[21]3种攻击方法生成雷达对抗样本进行实验。

FGSM[14]：FGSM是Goodfellow等[14]提出的一种基于梯度产生对抗样本的攻击方法。该方法通过最大化损失函数产生对抗样本

$$x^{\text{adv}} = x^{\text{clean}} + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y^{\text{true}})) \tag{1}$$

式中：x^{clean}为干净样本，x^{adv}为生成的对抗样本，$\nabla_x J(\cdot)$为损失函数的梯度，y^{true}为干净样本对应的类别，sign(·)为符号函数，ε为限制扰动范围的一个常数。FGSM的主要思想是通过计算预测概率与真实值之间损失函数的梯度，并通过符号函数sign(·)来得到梯度方向，将得到的梯度方向乘以扰动步长ε得到对抗噪声，最后将噪声加到干净图像上获得对抗样本。这些对抗样本对线性性相对较高的模型攻击效果较强。

PGD[20]：FGSM可以快速地产生对抗样本，但只进行一步梯度攻击往往会攻击不成功。PGD通过多次迭代产生对抗样本，在每次迭代中将对抗样本通过投影约束到干净样本附近来约束扰动。PGD算法的定义为

$$x_0 = x, x_{t+1} = \text{Clip}_{x,\varepsilon}(x_t + \alpha \cdot \text{sign}(\nabla_{x^t} J(x_t, y))) \tag{2}$$

式中：将第1步对抗样本x_0初始化为原始样本x，$\text{Clip}_{x,\varepsilon}(x)$为截断函数，将扰动约束在$L_\infty$范数$\varepsilon$范围内，$\alpha$为单步攻击的步长，$t$为迭代次数。由PGD攻击方法生成的对抗样本攻击性较强。

SPAS[21]：黑盒攻击在不具备网络模型参数的情况下，通过分析给定输入所对应的输出之间的关系，近似估计目标函数关于输入的梯度，最小化真实类别的输出对数与其余类别之间的最大对数。SPAS在此基础上使用了扰动随机逼近算法进行梯度估计实施攻击，并通过特征降维和随机抽样提高效率。

根据上述攻击算法生成的对抗样本可以有效攻击神经网络模型,使得模型分类出错,而且生成的对抗样本与原始样本之间的像素差距很难被发现。如图2,类别BMP2的原始图像在添加扰动后分别被分类为ZIL131和T72。

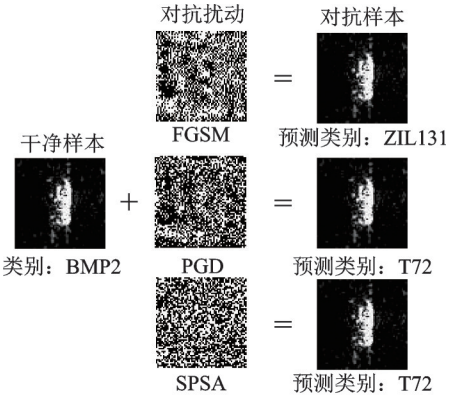


图2　不同攻击方法产生的雷达对抗样本

Fig.2　Adversarial radar samples generated by different attack methods

1.3　神经网络模型

神经网络强大的特征提取能力使其可以准确地分类雷达样本,常见的神经网络主要包括输入层、卷积层、池化层、全连接层和输出层。实验选取了常用的卷积神经网络VGG-Net[23]和ResNet[18]进行实验。

VGG[23]:VGGNet采用3×3的卷积核和2×2的池化层。通过3×3的卷积核的叠加可达到不同的感受野,两个3×3的卷积核叠加后就具有了和5×5的卷积核相同的感受野,再叠加一个3×3的卷积核后和7×7的卷积核具有相同的感受野,采用这种方法的优点是大大降低了参数量。

ResNet[18]:在进行反向传播时,随着神经网络的层数增加,模型会出现梯度爆炸或梯度消失的情形,导致模型的性能下降。ResNet通过利用深度残差网络解决了上述问题,使得网络的层数可以不断提高。借助于残差网络结构,低层的特征可以映射到高层,将浅层数据连接到深层。ResNet有18、34、50、101和152层多个不同层数的架构。

2　两步式雷达对抗样本检测技术

由于雷达图像的散射成像机理,导致雷达成像机制不同于光学成像系统[16],原始雷达图像在成像过程中会不可避免地产生散斑噪声,相对于散斑噪声,对抗样本的噪声则是人为添加至原始图像中。本文从空间域与频率域对雷达图像的原始样本和对抗样本进行了分析,提出了两步式雷达对抗样本检测方法,第1步将雷达干净样本与对抗样本利用傅里叶变换从空间域转换到频率域得到频谱图,然后计算频谱图的变异系数值并与设定的阈值进行比较,判断样本类别,然后将剩余的混合数据输入到第2步基于对抗训练的检测方法中,根据经过对抗训练模型与未经过对抗训练模型的预测类别判断样本类别。具体检测流程如图3所示。

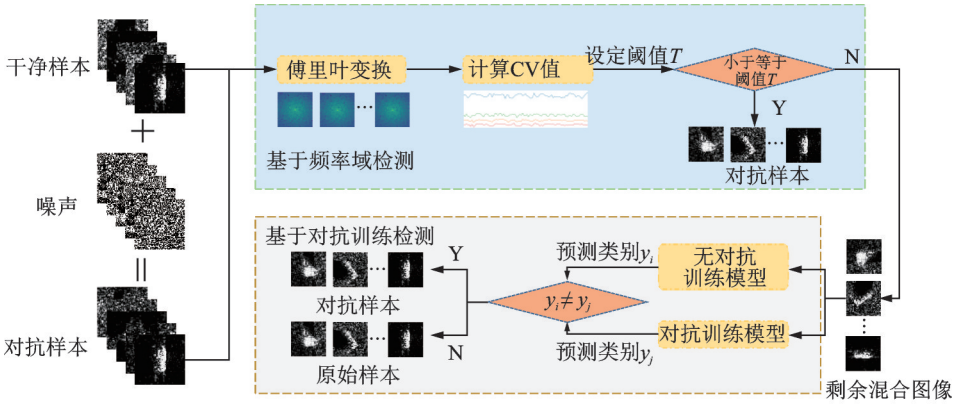


图3　两步式雷达对抗样本检测技术

Fig.3　Two-step radar adversarial sample detection technique

2.1 基于频率域的雷达图像对抗样本检测

2.1.1 原始样本和对抗样本空间域分析

为了分析雷达原始样本与对抗样本在空间域的差异，探究过程中首先选取100张原始图像，然后使用3种攻击方法得到对应的对抗样本，然后进行空间域上的灰度统计，绘制出灰度直方图。如图4所示，原始样本和对抗样本在空间域的灰度分布无较大差异。

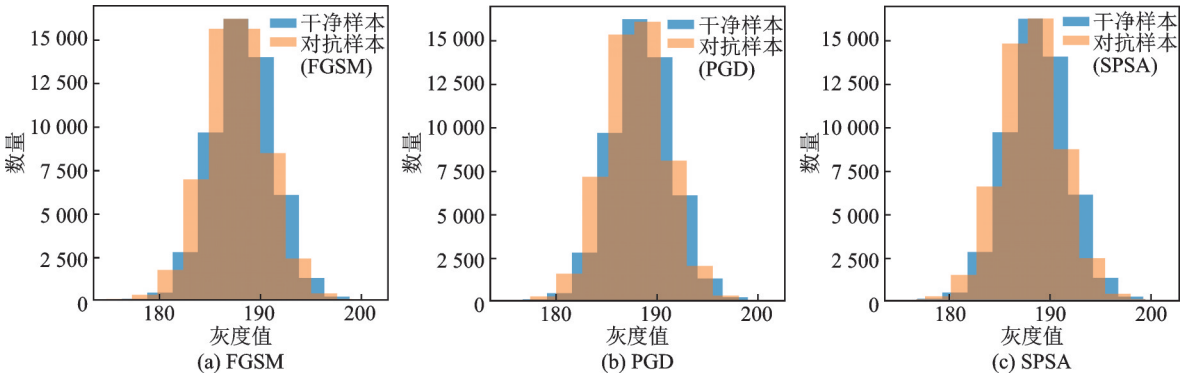


图4 不同攻击产生的对抗样本与原始样本在空间域中的灰度直方图

Fig.4 Grayscale histograms of adversarial samples generated by different attacks and original samples in spatial domain

2.1.2 原始样本和对抗样本频率域分析

本文使用傅里叶变换将原始图像和对抗样本从空间域变换到频率域，对频谱图进行分析。傅里叶变换的表达式为

$$F(u,v)=\sum_{x=0}^{M-1}\sum_{y=0}^{N-1}f(x,y)\mathrm{e}^{-2\pi\mathrm{i}(ux/M+vy/N)} \tag{3}$$

式中：M、N分别为图像的宽和高，$f(x,y)$为图像在(x,y)处的像素值，$F(u,v)$为对应的频域值。图5展示了雷达图像干净样本与对抗样本频谱图灰度直方图。从图5中可以看出，雷达图像原始样本与对抗样本在低频部分分布差异较大，相较于空间域的灰度值分布，频率域中频谱图的灰度值分布更易于区分雷达对抗样本与原始样本。

本文基于频谱的变异系数(Coefficient of variation, CV)来区分对抗样本与原始样本，变异系数的

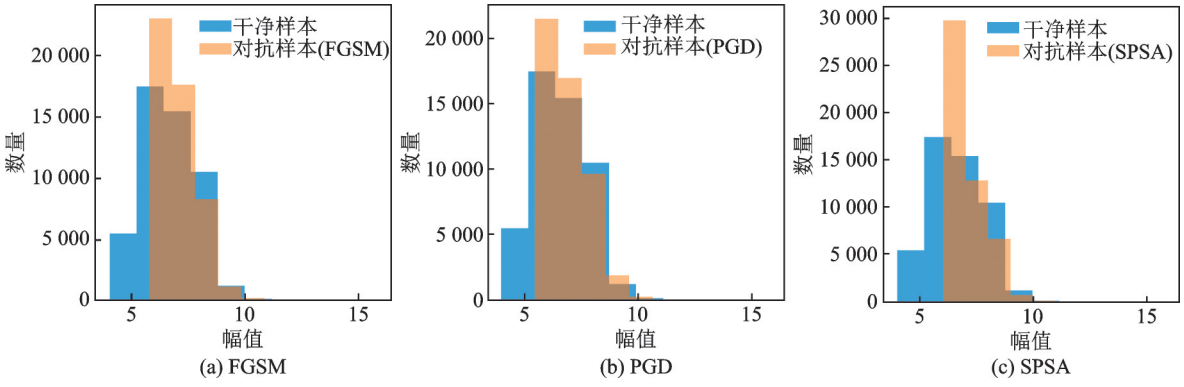


图5 不同攻击产生的对抗样本与原始样本频谱图的灰度值直方图

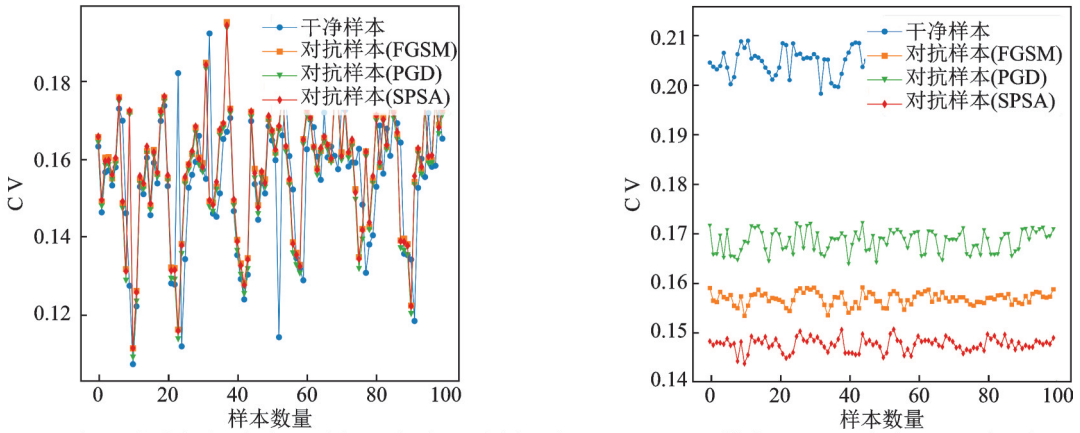
Fig.5 Grayscale value histograms of spectrograms of adversarial samples generated by different attacks and original samples

计算公式为

$$\mathrm{CV}=\frac{\sqrt{\dfrac{1}{WH}\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}\left(g(x,y)-\bar{g}(x,y)\right)^2}}{\dfrac{1}{WH}\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}g(x,y)} \tag{4}$$

式中:H和W分别为图像的高和宽;$g(x,y)$为频谱图在点(x,y)处的像素值;$\bar{g}(x,y)$为频谱图像素值的平均值。

图6(a)为随机挑选的100张图片与其3种对抗样本直接计算CV值的结果,图6(b)为将干净样本与对抗样本经过傅里叶变换得到的频谱图计算CV值的结果。从图6中可以看出,在空间域中干净样本与对抗样本的变异系数值趋于一致,而在频率域中的变异系数值则有着明显的间隔,表明基于频率域进行对抗样本检测的可行性。



(a) Computed CV of original and adversarial samples in spatial domain　(b) Spectrograms in frequency domain

图6　原始样本与不同对抗样本分别在空间域与频率域中的CV值

Fig.6　CV values of original samples and different adversarial samples in spatial and frequency domains respectively

2.2　基于对抗训练的雷达对抗样本再检测

针对基于频率域的对抗样本检测方法检测后剩余的混合数据,本文通过基于对抗训练的方法再次检测,进一步提升了雷达分类模型的安全性。对抗训练以一种类似于数据增强的手段,针对当前模型生成对抗样本,然后将这些对抗样本与原始样本送入模型进行再次训练,可以有效地提升模型的鲁棒性。对抗训练的目的可以表示为

$$y^{\mathrm{true}}=f(\theta,x^{\mathrm{clean}}),\quad y^{\mathrm{true}}=f_{\mathrm{AT}}(\theta',x^{\mathrm{clean}}+\varepsilon) \tag{5}$$

式中:x^{clean}和y^{true}分别为干净样本和与之对应的正确类别,以此训练得到模型f,θ为模型对应的参数。向干净样本x^{clean}添加噪声ε得到对抗样本x^{adv},然后送入模型f中再次训练,得到新的模型f_{AT},其参数为θ'。对于输入的对抗样本,未经过对抗训练的模型f分类会出错,而经过对抗训练的模型f_{AT}则会得到正确的结果。

对于经过第1步检测后剩余的数据,将其输入未经过对抗训练的模型f得到预测类别y_i,再将其输入到经过对抗训练的模型f_{AT}得到预测类别y_j,如果类别$y_i\neq y_j$则输入的数据为对抗样本,否则为干净样本。基于对抗训练的雷达对抗样本检测算法如下:

输入：第1步检测后剩余的混合数据集 Data $\{x^*|x\in x^{\text{clean}}\cup x^{\text{adv}}\}$

输出：x^*是干净样本x^{clean}或对抗样本x^{adv}

(1) for x^* in Data

(2) 样本x^*输入模型f得到预测类别y_i

(3) 样本x^*输入模型f_{AT}得到预测类别y_j

(4) if $y_i\neq y_j$ then

(5) x^*为对抗样本

(6) else

(7) x^*为干净样本

(8) end for

3 实验与分析

本文首先在MSTAR[17]和SENSAR[24]两类数据集上使用两种白盒攻击算法FGSM、PGD和一种黑盒攻击算法SPSA对3个经典深度神经网络模型(ResNet18、ResNet50和VGG13)进行攻击实验，分析雷达分类模型的鲁棒性，然后与两种检测方法进行对比验证本文方法检测雷达对抗样本的有效性，最后在不同扰动程度下检测雷达对抗样本，验证本文检测方法在不同扰动下都具有较好的检测效果，且在较小的扰动下依然有效。

3.1 数据集

MSTAR[17]：数据集采用美国国防高等计划署支持的MSTAR计划所公布的实测SAR地面静止目标数据，采集该数据集的传感器为高分辨率的聚束式合成孔径雷达，该雷达的分辨率为0.3 m×0.3 m。MSTAR图像的采集条件分为标准工作条件和扩展工作条件，本文选择了标准工作条件下收集的SAR图像，包括10个类别：2S1、BRDM_2、BTR_60、D7、T72(SN_132)、BMP2(SN_9563)、BTR-70(SN_C71)、T62、ZIL131、ZSU_23_4，其训练集数量与测试集数量如表1所示，每一类的图像示例如图7所示。

表1 MSTAR数据集训练集和测试集中不同类别的数量

Table 1 Number of different categories in the training and test sets of the MSTAR dataset

类别	2S1	BRDM_2	BTR_60	D7	T72	BMP2	BTR-70	T62	ZIL131	ZSU_23_4
训练集数量	424	423	322	423	330	330	331	422	424	423
测试集数量	149	149	129	150	98	98	98	150	149	150

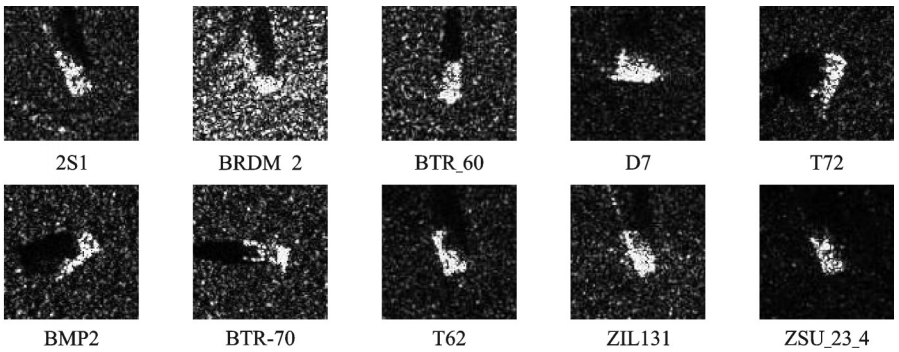


图7 MSTAR数据集示例

Fig.7 Examples from the MSTAR dataset

SENSAR[24]:SEN1-2数据是一类光学图像对数据集。通过Sentinel-1和Sentinel-2采集全球4个季节下的雷达图像,其分辨率为5 m。本文的实验从Sentinel-1中的雷达图像中选取了夏季中的20个类别,从每个类别中随机选取了一定数量的图片作为测试集和训练集,总训练集数量为10 591,测试集数量为6 344。每个类别图像的示例如图8所示。

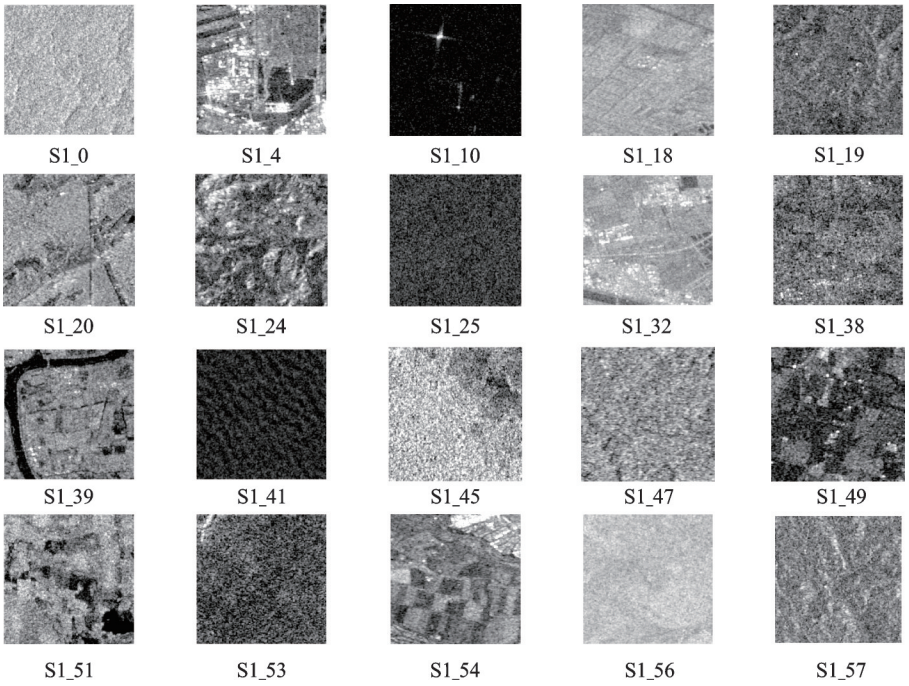


图8　SENSAR数据集示例

Fig.8　Examples from the SENSAR dataset

3.2　对抗样本对雷达分类模型的攻击性实验

本节使用1.2节提到的攻击算法在两个数据集上分别对ResNet18、ResNet50和VGG13深度神经网络进行攻击性验证,然后使用FGSM和PGD攻击算法分别在两个数据集上进行不同强度扰动的攻击实验,验证雷达分类模型在遭受攻击时分类性能难以保持。

由于MSTAR数据集中的目标都居于图像中心,所以在实验过程中,通过中心裁剪,将图像分辨率更改为64像素×64像素。对于SENSAR图像,分辨率由256像素×256像素压缩到224像素×224像素。然后分别使用白盒攻击FGSM、PGD和黑盒攻击SPSA算法在两个数据集上进行了攻击性实验。对于FGSM算法,实验中设置L_∞范数约束扰动$\varepsilon=0.020$用来生成对抗样本;针对PGD算法,实验中设置L_∞范数约束扰动$\varepsilon=0.020$,每次的步长为0.002,迭代10次用来生成对抗样本,对于黑盒攻击SPSA设置L_∞范数约束扰动$\varepsilon=0.025$,迭代20次来生成对抗样本。如表2所示,3种攻击方法都具有较高的攻击成功率,在MSTAR数据集上攻击算法的平均攻击成功率为61.79%,在SENSAR数据集上,攻击平均成功率为98.17%,最高达到了100%的攻击,最低为95.13%。实验结果表明雷达图像的原始样本在添加一定的扰动后能有效干扰雷达分类模型,使其准确率降低。

为了进一步探究对抗样本对模型准确率的影响,在上述实验的基础上,针对FGSM攻击和PGD攻击分别设置了不同强度的扰动对模型进行攻击,设置扰动$\varepsilon(\left\| x^{\mathrm{adv}}-x^{\mathrm{clean}} \right\|_\infty \leqslant \varepsilon)$分别为0.005、

表2 3种攻击方法在两个数据集上攻击3种分类模型的攻击成功率

Table 2 Attack success rates of three attack methods on three classification models across two datasets %

攻击方法	MSTAR			SENSAR		
	Res18	Res50	VGG13	Res18	Res50	VGG13
FGSM	60.30	55.50	74.55	95.13	96.12	97.19
PGD	65.30	44.17	66.67	100.00	99.04	100.00
SPSA	60.76	57.65	71.24	97.17	98.93	99.94

0.010、0.015、0.020、0.025、0.030。具体实验结果如表3、4和图9所示。可以看出,随着扰动的增加,攻击成功率越来越高,在一些模型上会达到100%的成功率。对于FGSM攻击,当扰动ε达到0.020时,在MSTAR数据集上的攻击成功率均在50%以上,在SENSAR数据集上,攻击成功率均在90%以上。对于PGD攻击,在MSTAR数据集上,当扰动ε达到0.025时,攻击成功率均在60%以上,在SENSAR数据集上,当扰动ε为0.01时,攻击成功率均已达到97%以上,特别是在VGG13模型上,攻击成功率达到100%。

表3 不同扰动程度下FGSM攻击方法在两个数据集上攻击3种分类模型的攻击成功率

Table 3 Attack success rates of FGSM attack on three classification models across two datasets at different perturbation levels %

数据集	模型	扰动ε					
		0.005	0.010	0.015	0.020	0.025	0.030
MSTAR	ResNet18	2.05	25.30	47.50	60.30	70.98	77.95
	ResNet50	8.33	20.68	37.50	55.00	69.55	78.79
	VGG13	14.85	31.52	48.48	63.71	74.55	81.14
SENSAR	ResNet18	84.05	93.49	95.00	95.13	96.12	97.19
	ResNet50	68.93	78.20	88.80	92.48	93.02	93.84
	VGG13	87.30	94.51	93.60	93.60	94.31	94.45

表4 不同扰动程度下PGD攻击方法在两个数据集上攻击3种分类模型的攻击成功率

Table 4 Attack success rates of PGD attack on three classification models across two datasets at different perturbation levels %

数据集	模型	扰动ε					
		0.005	0.010	0.015	0.020	0.025	0.030
MSTAR	ResNet18	1.36	17.58	55.30	65.30	79.77	89.09
	ResNet50	6.06	15.08	27.12	44.17	60.15	74.70
	VGG13	11.67	26.89	44.47	66.67	81.82	89.62
SENSAR	ResNet18	79.29	97.79	99.39	100.00	100.00	100.00
	ResNet50	88.98	97.40	98.34	99.04	99.57	99.86
	VGG13	98.36	100.00	100.00	100.00	100.00	100.00

从实验结果可以看出,深度神经网络模型极易受到对抗攻击,在扰动达到一定强度后,模型的准确率会急剧下降,表明基于深度神经网络训练得到的雷达分类模型非常脆弱,这对雷达分类系统的安全性是一个极大的威胁。

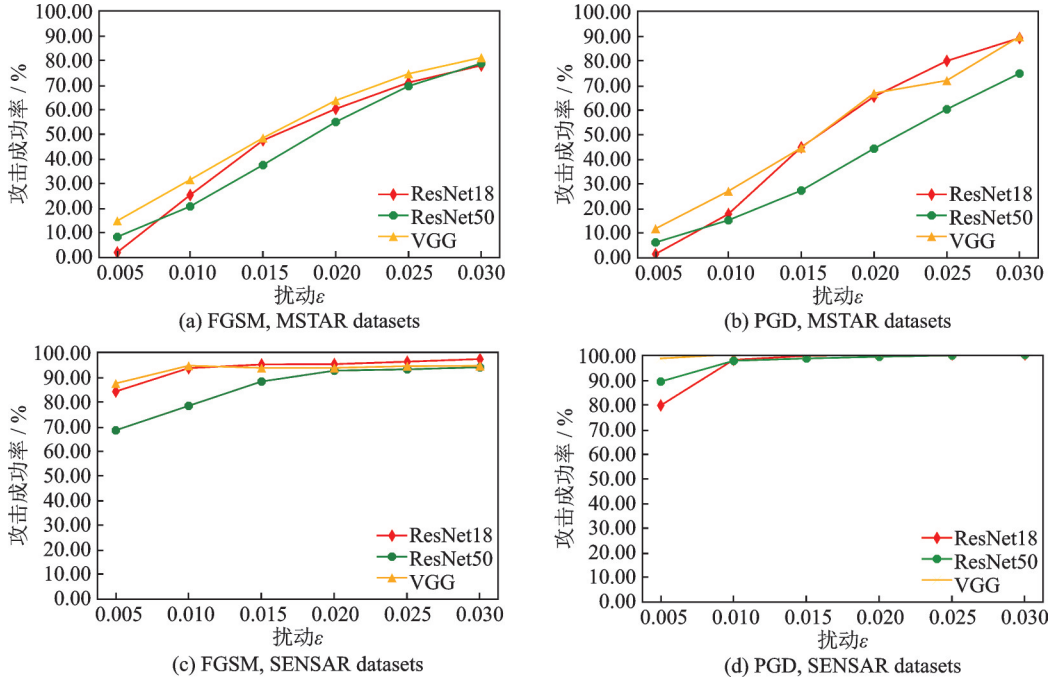


图9　不同扰动程度下FGSM和PGD攻击方法在两个数据集上攻击3种分类模型的攻击成功率

Fig.9　Attack success rates of FGSM and PGD attack methods on three classification models across two datasets at different perturbation levels

3.3　对抗样本检测方法对比实验

3.2节的实验已经证明了对抗样本对雷达分类模型具有攻击性,检测出这些对抗样本可以有效提升雷达分类模型的安全性。本节通过与3种对抗样本检测方法进行对比,验证所提出的两步式检测方法能更有效检测出雷达对抗样本。Zhou等[25]利用原始样本与对抗样本之间不对称的脆弱性来检测对抗样本,提出了借助攻击进行检测的攻击方法(Detect by attack,DBA)。Feinman等[26]通过组合对抗样本与原始样本之间的密度评估特征和贝叶斯不确定性特征训练逻辑回归模型(Logistic regression, LR)检测对抗样本,Deng等[27]提出了一种实用的轻量级贝叶斯改进方法(Lightweight Bayesian refinement, LiBRe),以低成本增强了预先训练的神经网络的对抗检测能力。对比实验中,对抗样本的扰动ε设为$0.020\left(\left\| x^{\mathrm{adv}}-x^{\mathrm{clean}}\right\|_{\infty}\leqslant\varepsilon\right)$,为了保证样本量的均衡,对抗样本与干净样本的数据量比例为1:1。本文使用准确率来衡量检测方法的性能,即

$$\mathrm{Accuracy}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \tag{6}$$

式中:TP和TN分别为被正确检测出的对抗样本和干净样本数量;FP和FN分别为被错误检测出的对抗样本和干净样本数量。表5展示了不同检测方法在两个数据集上使用3种不同攻击方法攻击3种分类模型的检测准确率。与DBA检测方法相比,本文所提出的检测方法检测准确率增加超过12.27%,比LR检测方法增加超过13.9%,相较于LiBRe方法增加不低于10.26%。本文所提出的检测方法在雷达数据集上具有更高的检测准确率,在MSTAR数据集上检测准确率不低于95%,在SENSAR数据集上的检测准确率结果不低于98%。

表5 不同检测方法检测准确率结果对比

Table 5 Comparison of detection accuracy results among different detection methods %

数据集	攻击方法	VGG13				ResNet18				ResNet50			
		DBA	LR	LiBRe	本文	DBA	LR	LiBRe	本文	DBA	LR	LiBRe	本文
MSTAR	FGSM	74.89	75.65	84.14	96.14	81.42	79.98	85.69	96.90	82.32	82.15	84.24	96.32
	PGD	71.01	72.48	80.27	96.52	79.63	81.97	80.27	97.82	83.01	80.38	82.79	95.73
	SPSA	75.89	76.87	85.61	96.91	82.46	83.45	87.09	97.35	81.97	79.87	85.39	96.98
SENSAR	FGSM	78.12	79.95	80.91	99.22	80.25	71.86	83.38	99.16	78.12	75.46	81.98	99.21
	PGD	76.96	82.64	79.08	99.17	81.13	78.71	82.47	99.13	82.33	80.15	79.87	98.17
	SPSA	79.98	80.28	81.28	99.23	84.75	80.45	86.77	99.41	79.94	82.67	83.87	99.45

3.4 不同扰动下对抗样本检测成功率

为了进一步分析两步式检测方法的检测能力，本节使用该检测方法检测不同扰动程度的对抗样本。扰动值$\varepsilon\left(\left\| x^{\text{adv}}-x^{\text{clean}}\right\|_{\infty}\leqslant\varepsilon\right)$分别设为0.005、0.010、0.015、0.020、0.025和0.030，且对抗样本与干净样本的比例为1∶1。同时在扰动ε为0.020时，分析MSTAR数据集中不同类别在两步式检测方法中，每一步的检出数量及两个数据集中不同类型对抗样本在每一步的检出率。图10展示了本文所提出的检测方法在不同分类模型和不同数据集上对不同对抗样本的检测成功率。对于不同的分类模型，在扰动值较小时，检测准确率均在80%以上，而且随着扰动值的增加，检测准确率不断提高，接近100%；对于同一分类模型，在不同攻击方法下检测方法同样具有较好的检测能力。

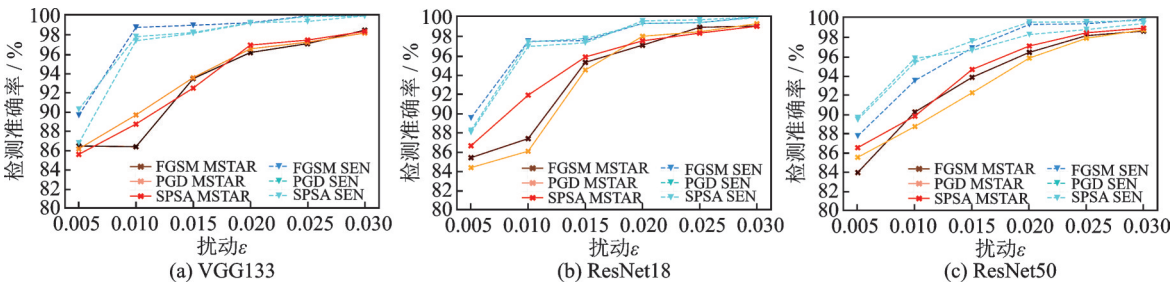


图10 不同扰动程度下两步式检测方法的检测准确率

Fig.10 Detection accuracy of two-step detection method at different perturbation levels

表6为检测方法在MSTAR数据集上每一类中的检测结果。对抗样本由PGD攻击方法攻击ResNet18分类模型产生，扰动ε的大小为0.020，对抗样本与干净样本的比例为1∶1。通过第1步基于频率域的检测，大部分对抗样本可以有效被检测出来，平均检测率为59.73%。结合第2步基于对抗训练的检测后，对抗样本总体平均检测率达到91.26%。

图11为每种类别中对抗样本在第1步(Step 1)、第2步(Step 2)检出比例及剩余比例(Remain)的可视化结果，大多数对抗样本可以在第1步检测中被检出，其中类别SN_9563和类别SN_C71在第2步对抗样本检出比例达到97.20%和98.83%。图12为单步检测(Step 1为基于频率域检测方法；Step 2为基于对抗训练检测方法)准确率与两步结合检测(Steps 1 & 2)准确率的对比。在MSTAR数据集中，基于频率域检测方法的检测准确率在50%左右，效果较差，基于对抗训练检测方法的检测准确率相较于基于频率域检测方法的准确率有所提升，但两步结合的方法能极大提高检测准确率，有效检出对抗样本；

表6　MSTAR数据集中不同类别在两步式检测方法中每步对抗样本的检出数量

Table 6　Numbers of detected adversarial samples for different categories in each step of the two-step detection method in the MSTAR dataset

类别	对抗样本总数量	第1步检出数量	第2步检出数量	检出总数量	剩余数量	对抗样本检出率/%
2S1	573	323	234	557	16	97.21
BRDM_2	572	422	107	529	43	92.48
BTR_60	451	302	134	436	15	96.67
D7	573	173	284	457	116	79.76
SN_132	428	397	28	425	3	99.30
SN_9563	428	416	12	428	0	100.00
SN_C71	429	424	5	429	0	100.00
T62	572	215	351	566	6	98.95
ZIL131	573	246	285	531	42	92.67
ZSU_23_4	573	171	397	568	5	99.13
合计	5 172	3 089	1 837	4 926	246	95.24

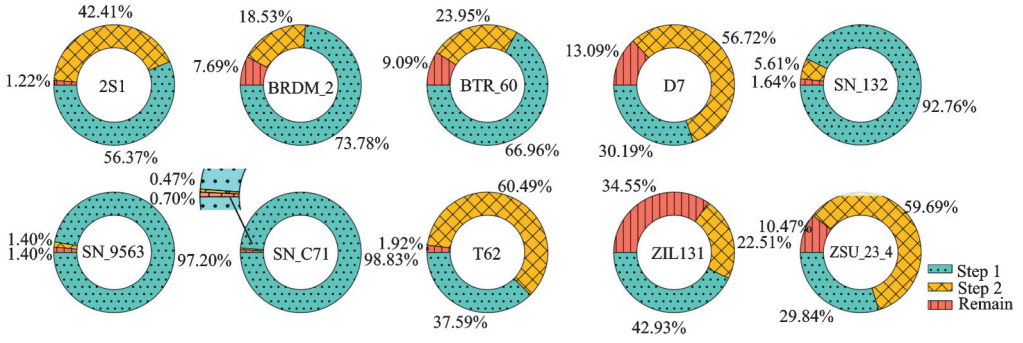


图11　MSTAR数据集中不同类别在两步式检测结果中每步对抗样本的检出比例

Fig.11　Detection proportions of adversarial samples for different categories in each step of the two-step detection results in the MSTAR dataset

在SEN数据集中,单步的检测方法都具有较好的检测结果,且基于频率域的检测方法作为一种简单的检测方法几乎能达到与两步结合检测方法同样的检测准确率。

图12为单步检测(Step 1为仅用基于频率域检测方法;Step 2为仅使用基于对抗训练检测方法)准

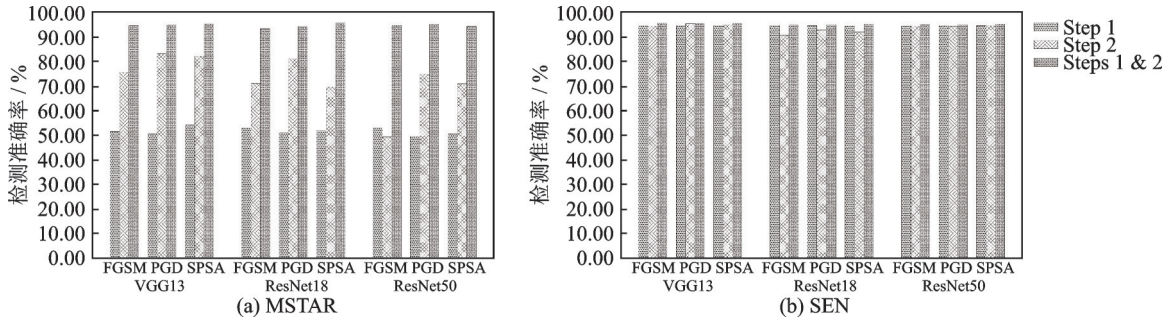


图12　单步检测(Step 1、Step 2)与两步结合(Steps 1 & 2)检测准确率

Fig.12　Detection accuracy of single-step detection (Step 1, Step 2) and combined two-step detection (Steps 1 & 2)

确率与两步结合检测(Steps 1 & 2)准确率的对比。在MSTAR数据集中,基于频率域检测方法的检测准确率在50%左右,效果较差,基于对抗训练检测方法的检测准确率相较于基于频率域检测方法的准确率有所提升,但两步结合的方法能极大提高检测准确率,有效检出对抗样本;在SEN数据集中,单步的检测方法都具有较好的检测结果,且基于频率域的检测方法作为一种简单的检测方法几乎能达到与两步结合检测方法相同的检测准确率。

4 结束语

本文针对雷达分类模型的安全性问题进行研究,分析了雷达对抗样本的攻击性和其在频率域与原始样本的差异,提出了一种两步式的雷达对抗样本检测技术。该方法首先基于雷达对抗样本与原始样本在频率域的差异进行第1步对抗样本检测,再基于对抗训练进行第2步对抗样本检测,实验结果表明此方法可以有效地检测雷达对抗样本,提升了模型的安全性。但是本文基于频率域的第1步检测成功率依赖于阈值的设置,后续工作中会基于频率域的差异设计出一种原始样本与对抗样本的分类器,不再依赖于阈值的设定,提升了该方法的泛化性。

参考文献:

[1] GUO Y, DU L, WEI D, et al. Robust SAR automatic target recognition via adversarial learning[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 716-729.

[2] 雷钰,刘帅奇,张璐瑶,等.基于深度学习的合成孔径雷达图像去噪综述[J].兵器装备工程学报,2022,43(11): 71-80.
LEI Yu, LIU Shuaiqi, ZHANG Luyao, et al. Review of synthetic aperture radar image denoising based on deep learning[J]. Journal of Ordnance Equipment Engineering, 2022, 43(11): 71-80.

[3] DENIS L, DALSASSO E, TUPIN F. A review of deep-learning techniques for SAR image restoration[C]//Proceedings of 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. [S.l.]: IEEE, 2021: 411-414.

[4] HUANG T, ZHANG Q, LIU J, et al. Adversarial attacks on deep-learning-based SAR image target recognition[J]. Journal of Network and Computer Applications, 2020, 162: 102632.

[5] 谷雨,张琴,徐英.融合压缩感知和SVM的SAR变形目标识别算法[J].数据采集与处理,2016,31(4): 754-760.
GU Yu, ZHANG Qin, XU Ying. SAR distorted object recognition algorithm based on compressed sensing and support vector machine fusion[J]. Journal of Data Acquisition & Processing, 2016, 31(4): 754-760.

[6] WANG R, WANG W, SHAO Y, et al. First bistatic demonstration of digital beamforming in elevation with TerraSAR-X as an illuminator[J]. IEEE Transactions on Geoscience and Remote Sensing, 2015, 54(2): 842-849.

[7] PARRILLI S, PODERICO M, ANGELINO C V, et al. A nonlocal SAR image denoising algorithm based on LLMMSE wavelet shrinkage[J]. IEEE Transactions on Geoscience and Remote Sensing, 2011, 50(2): 606-616.

[8] GENG J, JIANG W, DENG X. Multi-scale deep feature learning network with bilateral filtering for SAR image classification [J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 167: 201-213.

[9] 康志强,张思乾,封斯嘉,等.稀疏先验引导CNN学习的SAR图像目标识别方法[J].信号处理,2023,39(4): 737-750.
KANG Zhiqiang, ZHANG Siqian, FENG Sijia, et al. Sparse prior-guided CNN learning for SAR images target recognition[J]. Journal of Signal Processing, 2023, 39(4): 737-750.

[10] 霍鑫怡,李焱磊,陈龙永,等.基于卷积注意力和胶囊网络的SAR少样本目标识别方法[J].中国科学院大学学报,2022,39(6): 783-792.
HUO Xinyi, LI Yanlei, CHEN Longyong, et al. SAR few-sample target recognition method based on convolutional block attention module and capsule network[J]. Journal of University of Chinese Academy of Sciences, 2022, 39(6): 783-792.

[11] DU M, BI D. Local aggregative attack on SAR image classification models[C]//Proceedings of 2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC). [S.l.]: IEEE, 2022: 1519-1524.

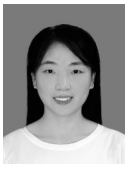
[12] YUAN X, HE P, ZHU Q, et al. Adversarial examples: Attacks and defenses for deep learning[J]. IEEE Transactions on neural Networks and Learning Systems, 2019, 30(9): 2805-2824.

[13] DU C, HUO C, ZHANG L, 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, 2021, 19: 1-5.
[14] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB/OL].(2013-12-21)[2023-12-27]. https://arxiv.org/abs/1312.6199.
[15] KURAKIN A, GOODFELLOW I J, BENGIO S. Artificial intelligence safety and security[M]. [S.l.]: Chapman and Hall, 2018: 99-112.
[16] LI H, HUANG H, CHEN L, 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, 2020, 14: 1333-1347.
[17] ROSS T D, WORRELL S W, VELTEN V J, et al. Standard SAR ATR evaluation experiments using the MSTAR public release data set[C]//Proceedings of Algorithms for Synthetic Aperture Radar Imagery Ⅴ. [S.l.]: SPIE, 1998: 566-573.
[18] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2016: 770-778.
[19] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of 2017 IEEE Symposium on Security and Privacy (SP). [S.l.]: IEEE, 2017: 39-57.
[20] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[C]//Proceedings of International Conference on Learning Representations (ICLR). [S.l.]: IEEE, 2018.
[21] UESATO J, O'DONOGHUE B, KOHLI P, et al. Adversarial risk and the dangers of evaluating against weak attacks[C]//Proceedings of International Conference on Machine Learning. [S.l.]: IEEE, 2018: 5025-5034.
[22] PAPERNOT N, MCDANIEL P, GOODFELLOW I, et al. Practical black-box attacks against machine learning[C]//Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. [S.l.]: ACM, 2017: 506-519.
[23] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL].(2014-09-04)[2023-12-27]. https://arxiv.org/abs/1409.1556.
[24] SCHMITT M, HUGHES H L, ZHU X X. The SEN1-2 dataset for deep learning in SAR-optical data fusion[J]. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences, 2018, Ⅳ-1: 141-146.
[25] ZHOU Q, ZHANG R, WU B, et al. Detection by attack: Detecting adversarial samples by undercover attack[C]//Proceedings of European Symposium on Research in Computer Security. Cham: Springer, 2020: 146-164.
[26] FEINMAN R, CURTIN R R, SHINTRE S, et al. Detecting adversarial samples from artifacts[EB/OL].(2017-05-01)[2023-12-27]. https://arxiv.org/abs/1703.00410.
[27] DENG Z, YANG X, XU S, et al. Libre: A practical bayesian approach to adversarial detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. [S.l.]: IEEE, 2021: 972-982.

作者简介:

王见(1999-),男,硕士研究生,研究方向:医学图像处理、深度神经网络对抗攻击与防御,E-mail: wang_jian@nuaa.edu.cn。



张赛楠(2000-),女,硕士研究生,研究方向:医学图像处理,E-mail: 547148907@qq.com。

陈芳(1991-),**通信作者**,女,博士,副教授,研究方向:医学图像分析、计算机辅助手术导航,E-mail: chenfang@nuaa.edu.cn。

(编辑:陈珺)