深度卷积神经网络图像识别模型对抗鲁棒性技术综述

孙 浩*[①] 陈 进[②] 雷 琳[①] 计科峰[①] 匡纲要[①]

[①](国防科技大学电子信息系统复杂电磁环境效应国家重点实验室 长沙 410073)

[②](北京市遥感信息研究所 北京 100192)

摘 要：近年来，以卷积神经网络为代表的深度识别模型取得重要突破，不断刷新光学和SAR图像场景分类、目标检测、语义分割与变化检测等多项任务性能水平。然而深度识别模型以统计学习为主要特征，依赖大规模高质量训练数据，只能提供有限的可靠性能保证。深度卷积神经网络图像识别模型很容易被视觉不可感知的微小对抗扰动欺骗，给其在医疗、安防、自动驾驶和军事等安全敏感领域的广泛部署带来巨大隐患。该文首先从信息安全角度分析了基于深度卷积神经网络的图像识别系统潜在安全风险，并重点讨论了投毒攻击和逃避攻击特性及对抗脆弱性成因；其次给出了对抗鲁棒性的基本定义，分别建立对抗学习攻击与防御敌手模型，系统总结了对抗样本攻击、主被动对抗防御、对抗鲁棒性评估技术的研究进展，并结合SAR图像目标识别对抗攻击实例分析了典型方法特性；最后结合团队研究工作，指出存在的开放性问题，为提升深度卷积神经网络图像识别模型在开放、动态、对抗环境中的鲁棒性提供参考。

关键词：深度卷积神经网络；SAR图像识别；信息安全；对抗攻击与防御；鲁棒性评估

中图分类号：TP391 **文献标识码**：A **文章编号**：2095-283X(2021)04-0571-24

DOI: 10.12000/JR21048

引用格式：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.

Reference format: 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.

Adversarial Robustness of Deep Convolutional Neural Network-based Image Recognition Models: A Review

SUN Hao*[①] CHEN Jin[②] LEI Lin[①] JI Kefeng[①] KUANG Gangyao[①]

[①](*State Key Laboratory of Complex Electromagnetic Environment Effects on Electronics and Information System*, *National University of Defense Technology*, *Changsha* 410073, *China*)

[②](*Beijing Institute of Remote Sensing Information*, *Beijing* 100192, *China*)

Abstract: Deep convolutional neural networks have achieved great success in recent years. They have been widely used in various applications such as optical and SAR image scene classification, object detection and recognition, semantic segmentation, and change detection. However, deep neural networks rely on large-scale high-quality training data, and can only guarantee good performance when the training and test data are independently sampled from the same distribution. Deep convolutional neural networks are found to be vulnerable to subtle adversarial perturbations. This adversarial vulnerability prevents the deployment of deep neural networks in security-sensitive applications such as medical, surveillance, autonomous driving and military scenarios. This paper first presents a holistic view of security issues for deep convolutional neural network-based image recognition systems. The entire information processing chain is analyzed regarding safety and security risks. In particular, poisoning attacks and evasion attacks on deep convolutional neural networks

收稿日期：2021-04-14；改回日期：2021-05-21；网络出版：2021-06-07

*通信作者： 孙浩 sunhao@nudt.edu.cn *Corresponding Author: SUN Hao, sunhao@nudt.edu.cn

基金项目：国家自然科学基金(61971426, 61601035)

Foundation Items: The National Natural Science Foundation of China (61971426, 61601035)

责任主编：徐丰 Corresponding Editor: XU Feng

are analyzed in detail. The root causes of adversarial vulnerabilities of deep recognition models are also discussed. Then, we give a formal definition of adversarial robustness and present a comprehensive review of adversarial attacks, adversarial defense, and adversarial robustness evaluation. Rather than listing existing research, we focus on the threat models for the adversarial attack and defense arms race. We perform a detailed analysis of several representative adversarial attacks on SAR image recognition models and provide an example of adversarial robustness evaluation. Finally, several open questions are discussed regarding recent research progress from our workgroup. This paper can be further used as a reference to develop more robust deep neural network-based image recognition models in dynamic adversarial scenarios.

Key words: Deep convolutional neural network; SAR image recognition; Information security; Adversarial attacks and defense; Robustness evaluation

1 引言

近年来以深度卷积神经网络为代表的联结主义智能化[1]图像识别方法取得巨大进展，不断刷新光学和SAR图像场景分类、目标检测与识别、语义分割、变化检测等多任务性能水平[2–5]。智能化的一个重要特征就是能够跨任务、跨领域、跨类别进行知识泛化。然而，现有深度卷积神经网络识别模型依赖统计学习，只有在训练数据和测试数据服从独立同分布的假设前提下泛化性能才能得到有效保证[6]。深度卷积神经网络图像识别模型在面对多种不同类型的训练数据和测试数据间分布漂移时，预测性能水平会大大下降，缺乏对输入扰动的鲁棒性。研究表明[7,8]：在输入图像数据中添加细微对抗扰动，对于人类视觉感知信息变化过于微小不可分辨，但是却会导致深度卷积神经网络识别结果产生大范围的波动变化，甚至是严重的错误输出。深度卷积神经网络图像识别模型的对抗脆弱性给其在安全敏感领域的广泛部署带来巨大安全隐患[1,9–11]。

图1给出了深度卷积神经网络SAR图像识别模型不同输入扰动对比示例。对于来自MSTAR数据集[12]的SAR图像目标切片，以俯仰角17°切片作为训练集学习VGG-16网络[13]深度识别模型。如图1所示，测试图像目标真实类别为BMP2，当无噪声干扰时，识别模型预测输出为真实类别；当在测试图像中添加不同形式的噪声扰动后导致模型预测输出为错误类别。对抗扰动或堕化噪声并没有改变输入图像的语义内容，因此深度卷积神经网络识别模型不应该因其存在而改变决策行为。但事实上深度卷积神经网络识别模型很容易被很小的局部变化所迷

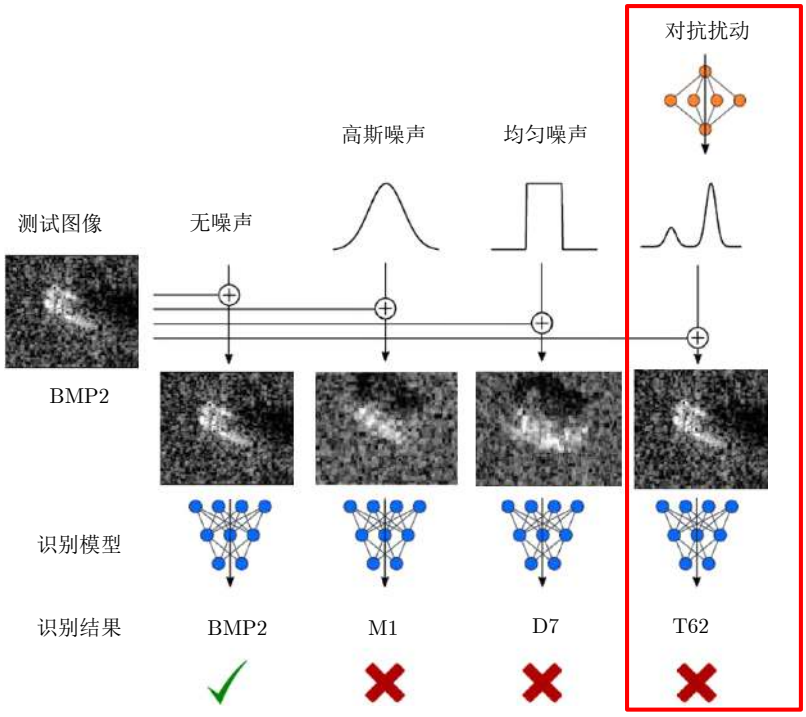

图 1 SAR图像深度神经网络识别模型典型扰动对比示例

Fig. 1 Different perturbations for deep neural networks based SAR image recognition models

惑，改变决策行为，以高置信度给出错误判断[7,8]。与堕化噪声相比较，由于对抗扰动产生机理更加复杂、扰动幅度小，人类视觉通常不可分辨、机器统计量很难可靠检测，在安全敏感领域危害性更强。与光学图像相比，SAR图像视觉解译变化量更多、解译难度更大，因此对抗扰动潜在攻击面更广。特别是在数字域对抗扰动的视觉不可感知范围更大，在物理域扰动实现手段更加多样化。

针对深度神经网络的对抗脆弱性，文献[9]从模型防御角度综述了图像分类对抗机器学习攻防技术，重点强调设计和评估对抗防御手段应该遵循的基本原则。文献[10]对目标识别应用中的对抗样本技术进行了总结与分析，讨论了对抗样本对于神经网络安全性和鲁棒性的影响，并重点分析了对抗样本的存在性假说及其在多个机器学习模型之间的迁移特性。文献[11]从网络安全角度回顾了针对智能化应用场景的对抗攻击技术，重点关注增强学习和联邦学习场景中智能化模型存在的对抗脆弱性。与现有的相关综述相比，本文聚焦深度卷积神经网络图像识别模型对抗鲁棒性技术研究进展，本文的特色和创新之处在于：(1)从智能化图像识别系统部署和应用流程出发，以信息安全视角全面分析系统存在的安全威胁和潜在攻击面，重点讨论了投毒攻击和逃避攻击特性及对抗脆弱性成因；(2)以对抗动态博弈视角分别建立对抗攻击与防御的威胁模型，按照攻防模型要素梳理现有研究方法，并以SAR图像深度识别模型对抗攻击为例分析典型方法特性；(3)系统介绍了对抗鲁棒性基本定义、对抗攻击、对抗防御、对抗鲁棒性评估的一般思路和指导原则，并结合团队研究工作进展，讨论未来研究趋势。

本文的组织形式如下：第2节从信息安全的角度分析深度卷积神经网络图像识别系统面临的多样化安全风险和脆弱性成因；第3节给出对抗鲁棒性的基本定义，系统总结深度神经网络对抗攻击与防御技术研究进展，分析对抗鲁棒性评估的基本准则和指标体系；第4节归纳现有研究存在的不足，指出一些开放性问题，为下一步研究提供参考。

2　深度卷积神经网络识别系统安全风险

2.1　深度学习图像识别系统安全威胁

以卫星、无人机等为代表的多源空天图像侦察近年来发展迅猛，不断持续获取海量高分辨率图像数据，仅依赖专家判读的数据分析模式已无法满足情报生成的时效性要求。一方面，基于人工智能和深度学习算法的大规模图像内容自动分析已逐步被引入离线情报生产过程中。另一方面，考虑到通信带宽、数据传输效率、情报生成实时性和区域拒止电磁对抗等多因素的影响，未来大量基于深度神经网络模型的多源图像目标检测与识别算法将被部署在边缘计算平台，进行在线目标识别和感兴趣数据筛选。

与传统的基于专家系统的符号主义智能化识别系统不同，基于深度卷积神经网络的联结主义智能化图像识别系统涉及全链路的数据复杂处理操作、预训练系统、机器学习框架多个方面，这些方面在军事对抗场景中都可能涉及安全问题[1]。深度神经网络识别系统开发部署过程可以分为任务规划、数据采集、模型训练、模型推理和系统部署5个阶段，如图2所示。

在现实应用中，各个环节间并不一定是序贯的，多个环节间通常会涉及反馈和循环。

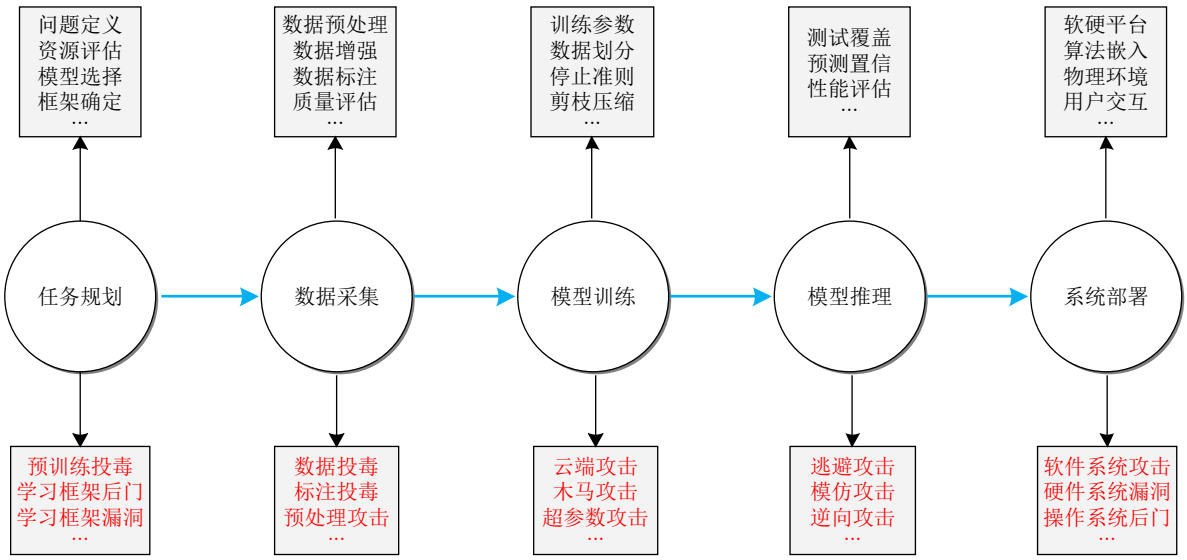

图 2 深度学习图像识别系统潜在安全风险

Fig. 2　Security risks for deep learning based image recognition system

(1) 任务规划阶段：开发智能化识别系统的首要问题是明确解决任务的边界条件，明确系统的期望图像输入数据及其分布，估计系统的准确性、鲁棒性、计算资源和运行时效性等指标。然后，对任务进行模块化分解，选择机器学习模型和框架。任务规划阶段面临的主要安全风险形式有学习框架后门和漏洞攻击、预训练模型投毒攻击等[14]。

(2) 数据采集阶段：在确定好问题的边界条件后，需要采集和整理用于深度识别模型的大规模标注训练数据集和测试数据集。为了提升模型的准确性指标和收敛速度，通常会采用图像几何变换和光度变换、物理仿真、对抗图像生成等方式进行训练数据扩充。数据采集与预处理阶段面临的主要安全风险形式有数据投毒攻击[15]、标注投毒攻击、图像尺度变换攻击[16]、数据集偏差攻击等。

(3) 模型训练阶段：对训练数据集合进行合理划分，在固定边界条件下进行模型架构或参数学习，确定迭代轮次、停止准则、学习率等超参数。资源受限应用场景中还需要考虑模型的剪枝和压缩问题。模型训练阶段面临的主要安全风险形式有云端攻击、木马攻击和超参数攻击等[17]。

(4) 模型推理阶段：对训练完成后的深度模型进行准确性和鲁棒性测试，以期满足预设指标。模型推理阶段面临的安全风险最大，常见的攻击形式有逃避攻击、模仿攻击和逆向攻击[18]。推理阶段的许多攻击方法不需要获取数据和模型的先验信息，采用黑盒方法，基于迁移性进行攻击，安全危害极大。逃避攻击的代表形式是深度识别模型的对抗样本。通过在目标外部添加特定设计的图案可以有效地逃避自动化算法的探测识别，与传统的电磁隐身伪装不同，基于对抗样本的智能扰动逃避攻击成本更低、部署和应用更加灵活。

(5) 系统部署阶段：将测试完成后的模型部署到相应的软硬件平台中，并完成真实物理环境中用户交互验证。系统部署阶段面临的主要安全风险有软件系统攻击、硬件系统漏洞、操作系统后门等[19]。

所有潜在攻击样式中，针对深度神经网络模型训练阶段的数据投毒攻击和针对模型推理阶段的逃避攻击在图像处理领域研究受到广泛关注，其攻击时机与攻击能力如图3所示。随着攻击知识的减少，投毒攻击的攻击能力按照逻辑破坏、数据修改、数据注入和数据读取等几个层次依次递减；逃避攻击的攻击能力按照网络架构、模型参数、模型逼近、查询攻击等几个层次依次递减。投毒攻击的实施可以是离线数据采集与模型学习阶段，也可以

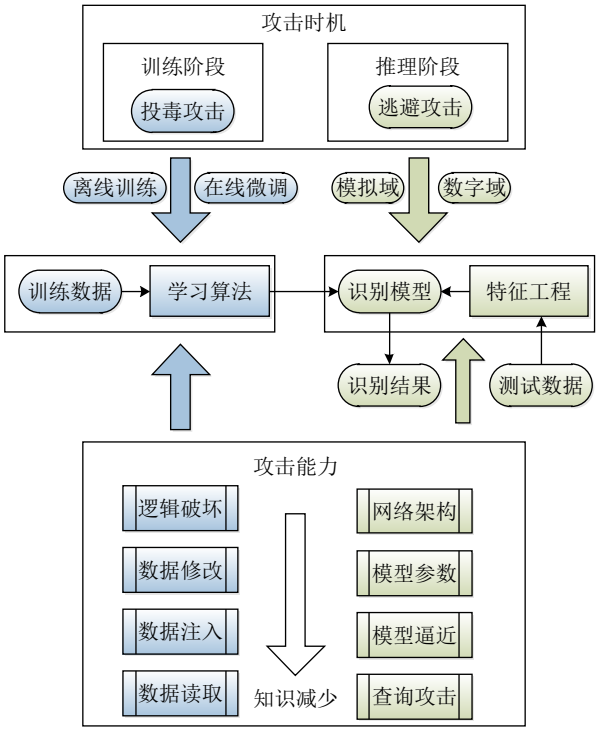

图 3 深度学习训练阶段和测试阶段攻击对比

Fig. 3 Comparison of training stage attacks and testing stage attacks for deep learning

是在模型在线微调阶段；逃避攻击的实施可以是在物理域中构建光电或射频扰动，也可以在数字域中添加对抗噪声。

数据投毒攻击通过在训练数据集合中注入虚假数据或混淆性标记信息，影响深度模型的归纳偏差，造成模型推理性能下降。如图4所示，通过在训练集中添加污染后的有毒数据，造成正确模型的决策边界出现偏离，从而造成测试样本的类别识别出现错误。逃避攻击不干扰训练数据，仅在推理阶段调整测试样本。逃避攻击的典型实现方式是生成对抗样本，通过在测试样本中添加微小非随机性扰动造成模型错误输出。对抗扰动通过面向识别模型的对抗攻击优化算法生成，通过细微扰动跨越模型的决策边界。

2.2 投毒攻击和逃避攻击脆弱性成因

深度图像识别系统可能在多个阶段和层次被攻击，其中许多潜在安全风险是信息安全领域的普遍问题，本节重点分析与深度学习过程紧密相关的投毒攻击和逃避攻击脆弱性成因。

2.2.1 训练数据依赖性

深度神经网络图像识别模型的准确率和鲁棒性高度依赖训练数据的数量和质量。只有在训练数据是无偏的情形下，深度识别模型才能达到理想的性能。深度识别模型仅仅从数据中学习得到了相关关

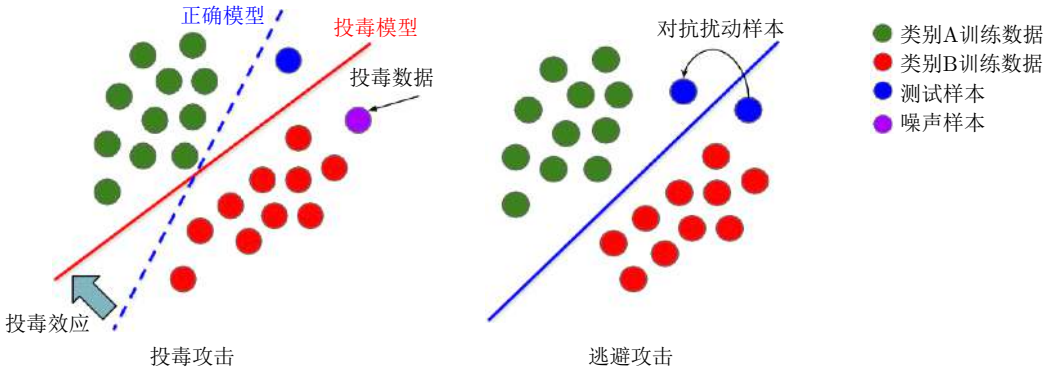

图 4 投毒攻击与逃避攻击基本原理

Fig. 4 Illustration of poisoning attack and evasion attack

系，而相关关系往往会随着数据分布的变化而变化，模型本身无法将虚假的相关与真实的因果区分开来。在许多安全敏感领域，大规模高质量训练数据严重稀缺，仅有的少量训练数据中还存在类别不平衡性和标注不确定性等问题，这些因素都严重加剧了模型的泛化风险和对抗脆弱性。与标准深度识别模型相比，鲁棒深度识别模型的样本采样复杂度更高，对标注数据的依赖性更强。采用预训练模型进行参数初始化可以加速模型收敛、提升模型性能，但同时会将预训练模型所采用数据集中的偏差、虚假相关、投毒数据等引入后续模型。在线微调阶段，物理域或数字域所产生的对抗样本都可以应用于投毒过程。

2.2.2 输入与状态空间高维特性

复杂的深度识别模型包含数百万量级参数，为了逼近决策函数这些参数需要在训练过程中进行迭代更新。参数的组合空间巨大，模型对输入数据的决策边界只能逼近求解。由于模型的高度非线性，因此输入数据的微小扰动可能会产生巨大的输出差异。训练数据一般情况下位于完备输入空间的低维流形，该现象通常也称为“维度灾难”。以VGG-16模型为例，16层深度的模型参数约135 M，采用二进制比特表示时输入空间维度为$2^{224\times224\times3\times8}=2^{1204224}$(模型输入图像空间大小为224像素×224像素，波段通道为3个，数字值量化为8比特位)，因此训练数据集合仅仅覆盖了输入空间中非常小的一部分，大量可能的输入数据，在训练过程中并没有利用。一方面，如果给模型输入训练过程中未观测到的良性数据，并且该数据与训练数据差别较大，那么模型可能无法泛化到这些输入数据，造成模型的安全风险。另一方面，当故意设计的对抗样本输入模型时会造成系统错误输出。文献[20]认为对抗样本存在于数据流形的低概率空间，很难通过随机采样输入数据的近邻空间得到，对抗样本所在区域是模型预测不确定性的盲点。尽管主流的深度卷积网络模型为了提升鲁棒性，在训练过程中都进行了数据增广，但变换后的数据与原始输入数据高度相关，且来自同样的数据分布，然而对抗样本通常呈现非相关和非同分布特性。此外，对抗样本生成过程中，向正常干净样本添加非随机噪声违背了模型训练过程中关于统计噪声的隐性假设。

由于理论上理想的任务决策边界通常在实际中只能通过模型决策边界近似，因此图像解译过程中无论是人类判读还是深度识别模型都是会出现错误的。模型通过数据和进化过程进行训练。在训练得到的模型中，传感器输入或其他边界条件的微小变化都有可能会导致状态改变，在状态空间中跨越决策边界。例如输入中出现的传感器微小噪声可能导致输出的巨大改变。任务决策边界与模型决策边界之间并不一定总是能够重合，在两者不同的区域，输入空间中常常存在对抗样本。图5表示输入空间二维投影中存在的对抗样本。输入空间中，存在4个类别的数据样本，类别A的深度识别模型决策边界由3个决策边界共同构成。类别A的任务决策边界与模型决策边界存在差异，在跨越或不跨越任务决策边界前提下，位于模型决策边界周围的样本都很容易受到微小扰动的影响造成模型输出错误。在高维空间中，搜索非重合区域内的样本很容易构造对抗样本。

图6为MSTAR图像目标识别性能评估策略的对比示意图[5,21]，这些评估准则确定了标准操作条件和扩展操作条件。标准操作条件是由现有数据集合构成的训练条件和测试条件，扩展操作条件是由建模条件定义。可见，不同条件下的输入空间存在部分不重叠区域和未覆盖区域，这些区域内的数据点都很容易被用于生成攻击样本。

2.2.3 黑盒特性与难解释性

基于深度卷积神经网络的图像识别模型结构设

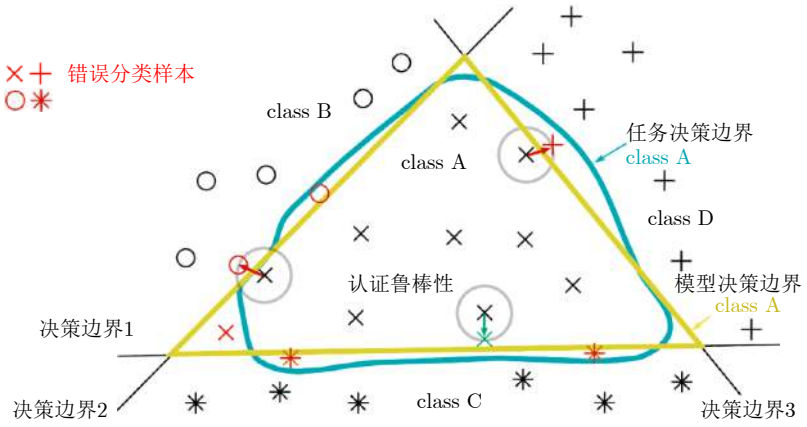

图 5 深度识别模型决策过程示例[1]

Fig. 5 Decision process for deep neural networks[1]

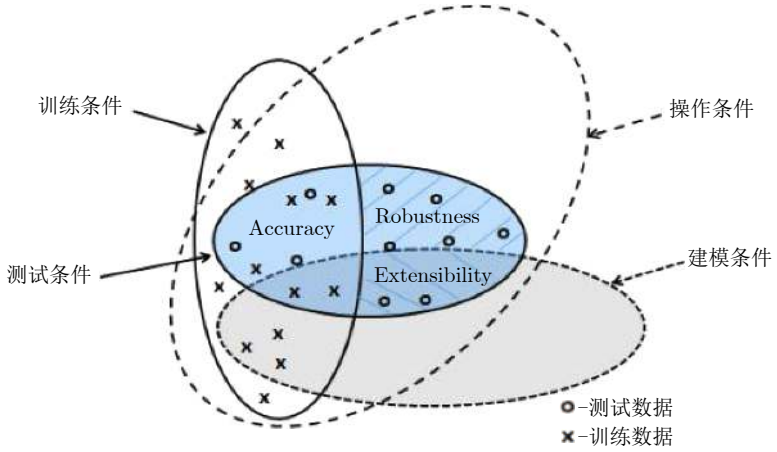

图 6 MSTAR性能评估策略[21]

Fig. 6 Performance evaluation strategy for MSTAR[21]

计和参数优化过程复杂，缺乏可解释性。用于解决图像处理任务的传统计算机程序是很容易理解的，对于具有充分知识背景的编程人员来说，系统是透明的。然而由于深度神经网络巨大的参数空间，复杂的深度识别模型不具备这一特性。编程人员仍然可以理解任务边界条件和解决任务的方法，但无法直接将神经网络的内部表示转换为理解其行为特性的工具。从信息安全的角度来看，这意味着只能通过模型的错误行为(而不是模型本身)来检测攻击，而模型的错误行为描述仍然是一个困难的问题。因此训练过程结束后，由于模型缺乏透明性，很难检测训练数据中存在的投毒攻击。

在安全敏感领域中需要提升深度识别模型的透明性和可解释性，特别是在军事应用领域，需要建立用户与智能识别模型之间的信任关系，辅助用户进行决策。可解释性深度识别模型的研究有两类典型思路[22–24]：一类是分析模型的动态特性，通过在输入变量中添加扰动或调整模型参数，对系统的输出进行统计分析，推测模型的决策依据。另一类是直接构建结构化和可解释性更强的深度网络模型。

3 深度卷积神经网络对抗鲁棒性研究进展

3.1 对抗鲁棒性定义

深度卷积神经网络识别模型可以描述为一个函数：$f_\theta:\boldsymbol{X}\to\boldsymbol{Y}$，将输入图像空间中的一个向量$\boldsymbol{x}\in\boldsymbol{X}$映射到标记空间中$\boldsymbol{y}\in\boldsymbol{Y}$，其中$\theta\in\boldsymbol{W}$是函数的参数变量，$\boldsymbol{W}$, $\boldsymbol{X}$和$\boldsymbol{Y}$分别表示深度卷积神经网络的权重空间、输入空间和输出空间。深度识别模型将整个输入空间划分为一组区域，每个区域具有唯一性的类别标记，识别模型的决策边界可以利用两组不同标记区域的交汇点集来定义。在有监督学习条件下，给定数据对$(\boldsymbol{x},\boldsymbol{y})$的分布$\boldsymbol{D}$，学习算法的目标是寻找一个分类器将任意的输入$\boldsymbol{x}$映射到标记$\boldsymbol{y}$，使得在分布$\boldsymbol{D}$上的期望风险最小化，即

$$\min_{\theta}\mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\boldsymbol{D}}[L(\boldsymbol{x},\boldsymbol{y};\theta)] \tag{1}$$

其中，$L(\boldsymbol{x},\boldsymbol{y};\theta)$表示特定形式的损失函数。实际应

用中我们无法获取所有的数据分布$\boldsymbol{D}$，仅仅利用一组训练样本集合$\{(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)})\}_{i=1}^{N}\sim\boldsymbol{D}$，因此无法通过最小化期望风险获得$f_\theta$。通常求解经验风险最小化问题，即

$$\min_{\theta}\frac{1}{N}\sum_{i=1}^{N}L\left(\boldsymbol{x}^{(i)},\boldsymbol{y}^{(i)};\theta\right)\tag{2}$$

深度卷积神经网络识别模型通常由多个前馈神经网络复合构成，其中第t层的输出$\boldsymbol{z}_t\in R^{D_k}$依赖前一层输出：

$$\boldsymbol{z}_t=h_t(\boldsymbol{z}_{t-1},\boldsymbol{z}_{t-2},\cdots,\boldsymbol{z}_0;\theta_t),\ \ \boldsymbol{z}_0=\boldsymbol{x}\tag{3}$$

其中，$h_t:R^{D_{t-1}}\times R^{M_t}\to R^{D_t}$为参变量为$\theta_t\in R^{M_t}$的可微分映射，在分类识别应用中通常设定最后一层的输出为$\boldsymbol{z}_L\in R^C$，并采用softmax函数将其映射到一组概率集合$p_\theta(\boldsymbol{x})\in[0,1]^C$：

$$[p_\theta(\boldsymbol{x})]_k=\frac{\exp\left([\boldsymbol{z}_L]_k\right)}{\sum\limits_{c=1}^{C}\exp\left([\boldsymbol{z}_L]_c\right)}\tag{4}$$

神经网络识别模型输出结果是最高概率密度的标记索引：

$$f_\theta(\boldsymbol{x})=\underset{k\in\{1,2,\cdots,C\}}{\arg\max}\ [p_\theta(\boldsymbol{x})]_k\tag{5}$$

深度识别模型对随机噪声扰动具有一定的鲁棒性；但对于对抗扰动，神经网络表现出极差的对抗脆弱性[7−11]。对抗扰动是输入$\boldsymbol{x}$的最坏情形微小扰动，经过精心设计用于欺骗神经网络。而且现有研究表明：对于任意的$\boldsymbol{x}$和识别模型f_θ总是可以找到对抗扰动，表明神经网络的决策边界在某些方向上靠近给定的数据样本，因此在这些方向上添加很小的扰动就可以改变分类器的输出结果。

定义对抗扰动$\delta(\boldsymbol{x})\in R^D$是下述优化问题的解[25]：

$$\min_{\delta\in R^D}Q(\delta),\ \ \text{s.t.}\ \ \ f_\theta(\boldsymbol{x}+\delta)\neq f_\theta(\boldsymbol{x}),\ \delta\in\Delta\tag{6}$$

其中，$Q(\delta)$表示目标函数的一般形式，Δ表示刻画扰动特性的一组约束集合。不同类型对抗扰动主要差别在于$Q(\delta)$和Δ，例如最小ℓ_p范数对抗扰动$\delta_p^*(\boldsymbol{x})$定义为

$$Q(\delta)=\|\delta\|_p=\left(\sum_{k=1}^{D}([\delta]_k)^p\right)^{1/p}\tag{7}$$

式(7)表示在ℓ_p范数度量下，跨越识别模型决策边界的最小加性扰动。

ε约束对抗扰动$\delta_p^\varepsilon(\boldsymbol{x})$定义为

$$Q(\delta)=L(\boldsymbol{x}+\delta,\boldsymbol{y};\theta),\Delta=\{\delta\in R^D:\|\delta\|_p\le\varepsilon\}\tag{8}$$

式(8)表示在给定数据样本$\boldsymbol{x}$的ε邻域内最大化损失函数的最坏情形扰动，ε的取值使得最终的扰动尽可能小，视觉不可感知。文献中还有其他形式的距离度量来定义对抗样本，例如数据流形测地线距离、感知度量和Wasserstein距离等。在现有的对抗扰动研究中，ℓ_p扰动研究得最为广泛和深入。

对抗样本很容易计算且大量存在，暴露出深度神经网络识别模型的脆弱性。为了解决这个问题，需要定义客观的度量来量化神经网络对于对抗扰动输入的鲁棒性。根据应用场景和任务的不同，f_θ对抗鲁棒性的定义可以有多种形式，一种常用的定义是基于对抗场景中分类器的泛化能力，即对抗扰动输入时神经网络的最坏情形准确率。

$$\rho_p^\varepsilon(f_\theta)=P_{(\boldsymbol{x},\boldsymbol{y})\sim\boldsymbol{D}}\left(f_\theta\left(\boldsymbol{x}+\delta_p^\varepsilon(\boldsymbol{x})\right)=\boldsymbol{y}\right)\tag{9}$$

从信息安全角度来看，$\rho_p^\varepsilon(f_\theta)$描述了深度神经网络在特定对抗攻击时的脆弱性，扰动值或者扰动约束体现攻击的强度。尽管标准泛化性能都很高，但是对于大多数威胁模型，标准神经网络模型通过式(9)计算得到的最坏情形准确率都很低。

考虑神经网络决策函数的几何特性，可以通过计算任意样本到神经网络的决策边界的平均距离定义对抗鲁棒性：

$$\rho_p^*(f_\theta)=\mathrm{E}_{(\boldsymbol{x},\boldsymbol{y})\sim D}\left[\left\|r_p^*(\boldsymbol{x})\right\|_p\right]\tag{10}$$

几何视角描述对抗鲁棒性的优势是鲁棒性的计算与对抗扰动产生算法无关，对抗鲁棒性是分类器的特性。采用几何测度，提升分类器的对抗鲁棒性意味着将决策边界与数据样本远离。采用式(9)测量分类器的鲁棒性还存在很多挑战，例如现有对抗攻击方法在计算扰动$\delta(\boldsymbol{x})$时并非最优。然而，我们可以通过计算所有样本和神经网络决策边界的安全距离来验证分类器的鲁棒性。分类器如果是ℓ_p范数下ε认证鲁棒的，那么分类器在任意样本的半径为ε的ℓ_p超球邻域内输出稳定的标记信息。在高维空间中进行鲁棒性认证，需要大量的计算代价，因此目前一般都是针对特定形式的分类器。

3.2 对抗攻击研究进展

3.2.1 对抗攻击模型

聚焦深度模型推理阶段的安全风险，建立对抗攻击威胁模型如图7所示，主要包括对抗攻击目标、对抗攻击知识、对抗攻击能力和对抗攻击策略4个方面[26,27]。对抗攻击的目标可采用安全破坏程度和攻击专一性进行描述。安全破坏程度主要是指对抗攻击者期望破坏深度识别系统的完整性、可用性或隐私性；攻击专一性主要包括定向攻击和非定向攻击两类。例如对抗攻击的目标可以是产生一个特定类别的识别错误攻击或非定向性的系统识别功能破坏攻击。对抗攻击的知识根据攻击者获取的先

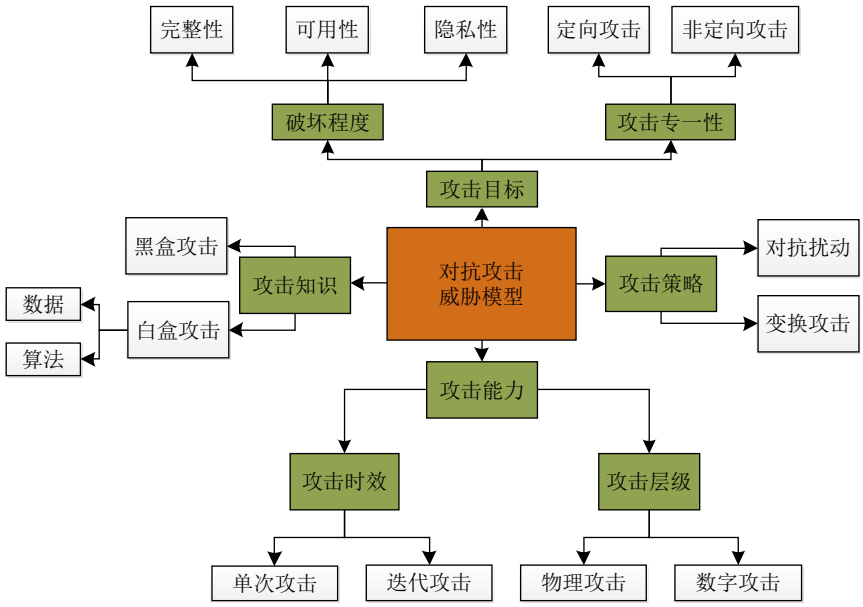

图 7 对抗攻击威胁模型

Fig. 7 Threat model for adversarial attacks

验信息来进行考虑，通常可以分为白盒攻击和黑盒攻击。白盒攻击场景下，攻击者已知识别模型的架构与参数、训练数据、预训练模型等信息，攻击效果最强。黑盒攻击场景下，攻击者仅通过有限的查询访问或对抗样本的迁移特性实现攻击。攻击能力采用攻击时效和攻击层级两个维度描述。对抗攻击时效可分为迭代型攻击和单次性攻击，虽然迭代型攻击效果较好，但军事应用场景中单次性攻击的危害性也需要重点关注。对抗攻击策略是指攻击者为了达到攻击目的而采取的图像内容或特征修改措施，典型策略有对抗扰动生成和变换攻击。基于对抗扰动生成的逃避攻击，通常称为对抗攻击。在许多安全敏感领域，攻击者很难获取训练阶段数据或相关信息，基于对抗样本的深度识别模型推理阶段对抗攻击威胁性更高，因此受到学术界和工业界的广泛关注。

3.2.2 对抗样本生成

图8描述了对抗样本生成的一般流程[18]。在白盒攻击场景中，攻击者通过求解梯度优化或约束优化问题、敏感性分析、生成模型采样等方式构造对抗样本[7–11,18,19,26,27]；在黑盒攻击场景中，攻击者通过多次查询被攻击模型获取相关信息，然后训练替代模型进行白盒攻击，或者估计梯度和近似决策边界来寻找对抗样本[7–11,18,19]。

表1归纳总结了典型对抗样本生成方法的攻击知识、攻击目标、攻击策略、扰动度量和扰动范

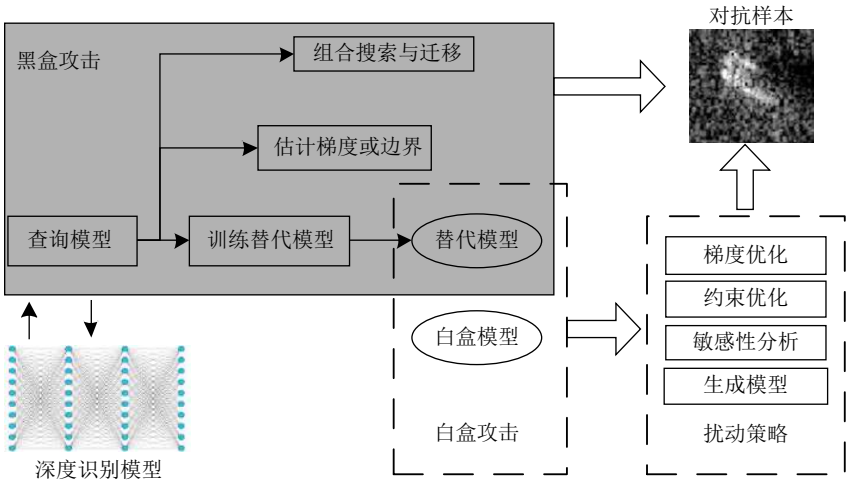

图 8 对抗样本生成流程

Fig. 8 Flowchart for adversarial example generation

表 1 对抗攻击典型方法

Tab. 1　Summarization of adversarial attacks

攻击方法	攻击知识	攻击目标	攻击策略	扰动度量	扰动范围
L-BFGS[7]	白盒	定向	约束优化	$L_{\inf}$	个体扰动
FGSM/FGV[8]	白盒	非定向	梯度优化	$L_{\inf}$	个体扰动
BIM/ILCM[28]	白盒	非定向	梯度优化	$L_{\inf}$	个体扰动
JSMA[29]	白盒	定向	敏感性分析	L_0	个体扰动
DeepFool-DF[30]	白盒	非定向	梯度优化	L_0, L_2, $L_{\inf}$	个体扰动/通用扰动
LaVAN[31]	白盒	定向	梯度优化	L_2	个体扰动/通用扰动
UAN[32]	白盒	定向	生成模型	L_2, $L_{\inf}$	通用扰动
EOT[33]	白盒	定向	梯度优化	L_2	个体扰动
C&W[34]	白盒	定向/非定向	约束优化	L_0, L_2, $L_{\inf}$	个体扰动
Hot-Cold[35]	白盒	定向	梯度优化	L_2	个体扰动
PGD[36]	白盒	定向/非定向	梯度优化	L_1, $L_{\inf}$	个体扰动
EAD[37]	白盒	定向/非定向	梯度优化	L_1	个体扰动
RP2[38]	白盒	定向	梯度优化	L_1, L_2	个体扰动
GTA[39]	白盒	定向	梯度优化	L_1, $L_{\inf}$	个体扰动
OptMargin[40]	白盒	定向	梯度优化	L_1, L_2, $L_{\inf}$	个体扰动
ATNs[41]	白盒	定向	生成模型	$L_{\inf}$	个体扰动
M-BIM[42]	白盒/黑盒	非定向	梯度优近似	$L_{\inf}$	个体扰动
POBA-GA[43]	黑盒	定向/非定向	估计决策边界	自定义	个体扰动
AutoZoom[44]	黑盒	定向/非定向	估计决策边界	L_2	个体扰动
LSA attack[45]	黑盒	定向/非定向	梯度近似	L_0	个体扰动
NES attack[46]	黑盒	定向	梯度近似	$L_{\inf}$	个体扰动
BA attack[47]	黑盒	定向	估计决策边界	L_2	个体扰动
GenAttack[48]	黑盒	定向	估计决策边界	L_2, $L_{\inf}$	个体扰动
ZOO[49]	黑盒	定向/非定向	迁移机制	L_2	个体扰动
UPSET[50]	黑盒	定向	梯度近似	L_2	通用扰动
ANGRI[50]	黑盒	定向	梯度近似	L_2	个体扰动
HSJA[51]	黑盒	定向/非定向	决策近似	L_2, $L_{\inf}$	个体扰动
单像素[52]	黑盒	定向/非定向	估计决策边界	L_0	个体扰动
BPDA[53]	黑盒	定向/非定向	梯度近似	L_2, $L_{\inf}$	个体扰动
SPSA[54]	黑盒	非定向	梯度近似	$L_{\inf}$	个体扰动
AdvGAN[55]	黑盒	定向	生成模型	L_2	个体扰动
Houdini[56]	黑盒	定向	约束优化	L_2, $L_{\inf}$	个体扰动

围。对抗样本主要包括个体扰动对抗样本和通用扰动对抗样本两类。个体扰动对抗样本是指对于给定的测试图像，根据优化算法生成特定的扰动，不同图像扰动模式不同；通用扰动对抗样本是指在特定数据集上或针对特定识别模型产生的扰动模式，对于数据集中的所有图像该扰动模式保持不变。对抗攻击策略主要包括图像空间扰动、特征空间扰动和决策空间扰动3类，在图像空间和特征空间进行扰动通常采用生成模型、梯度优化和敏感性分析算法实现，在决策空间进行扰动常采用约束优化算法实现。

为了说明典型攻击方法对雷达图像深度目标识别模型的影响，在MSTAR数据集上，以俯仰角17°目标切片图像作为训练集学习VGG-16深度识别模型，攻击目标设定为定向攻击，采用PGD (Projected Gradient Descent)[36], DeepFool[30], C&W[34] 3种方法的定向攻击版本生成对抗扰动。PGD攻击噪声范数选用L_∞，攻击强度设定为0.3；DeepFool攻击步长设定为10^{-6}，最大迭代次数设定为100次；C&W攻击方法学习率设定为0.01，最大迭代次数设定为100。采用Grad-CAM方法[57]对3种方

法生成的定向攻击样本、无扰动干净样本(原始类别和定向攻击目标类别)在VGG-16识别模型的激活响应进行可视化，第2, 4, 7, 10, 13卷积层特征激活结果如图9所示。其中第1行图像分别为真实类别(BTR70)的测试图像、采用3种攻击方法生成的定向攻击个体扰动对抗样本(BTR70定向攻击为ZIL131)、真实类别为ZIL131的测试图像(作为参考对照)，由于对抗扰动的幅度微小，因此3种方法生成的定向攻击样本与原始图像人眼无法分辨其中差别。观察特征层的激活情况容易发现：基于梯度优化的PGD攻击对抗样本从低层(第2层)卷积特征开始就与定向攻击类别的激活响应具有较高的相似性，攻击目标实现依赖多隐层特征空间扰动；基于约束优化的DeepFool攻击方法和C&W攻击方法产生的对抗样本仅在高层(第13层)卷积特征激活与真实类别具有较高的特征激活相似度，攻击目标实现更多依赖决策空间扰动。

图10展示了来自FUSAR-Ship数据集[58]的4幅SAR舰船目标图像切片及典型攻击方法产生的对抗扰动(实验细节见文献[59])，4幅图像的类别从上至下依次为集装箱船、货船、渔船和油轮。为了显示效果，所有的对抗扰动都进行了放大。基于梯度优化的对抗扰动(FGSM, PGD)、稀疏对抗扰动(JSMA，单像素)和基于约束优化的对抗扰动(DeepFool, C&W)在扰动模式上呈现明显的差异。FGSM和PGD扰动模式更加聚焦原始图像中的图像灰度变化剧烈区域，与图像梯度紧密相关。JSMA和单像素扰动仅仅改变了少量像素，但对抗扰动幅值较大。DeepFool扰动和C&W扰动改变了大量像素，但对抗扰动幅值较小。

3.3 对抗防御研究进展

3.3.1 对抗防御模型

大量对抗样本生成方法的不断提出，催生深度神经网络识别模型防御技术迭代演进，两者之间形成对抗攻防竞赛。根据防御目标的不同，对抗防御技术可以分为主动性防御和被动性防御两类[60,61]，两者之间的区别如图11所示。主动性防御技术是深度识别模型的开发者主动进行仿真攻击发现模型缺陷，并对模型进行鲁棒性提升。模型开发者首先通过分析敌手对抗攻击过程，建立对抗攻击威胁模

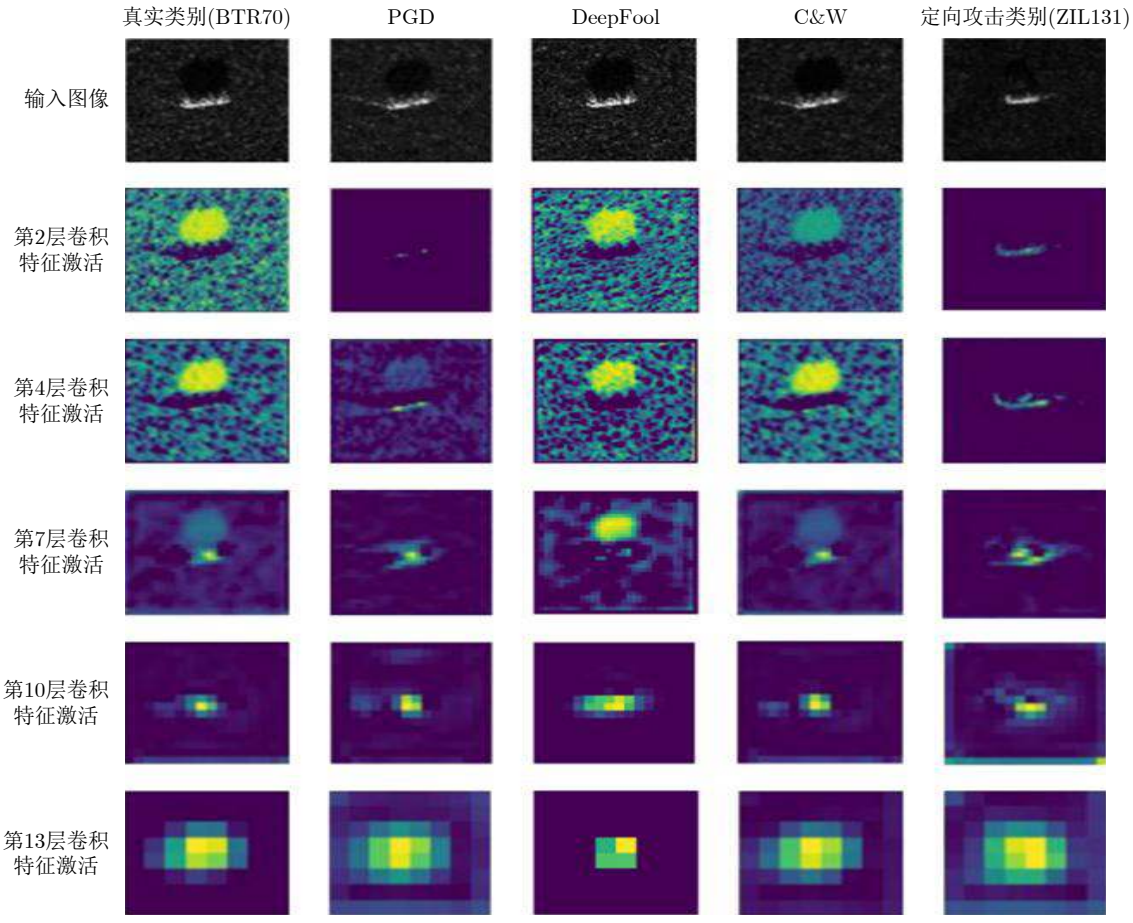

图 9 SAR图像目标识别定向对抗攻击举例

Fig. 9 Targeted adversarial attacks for SAR image target recognition

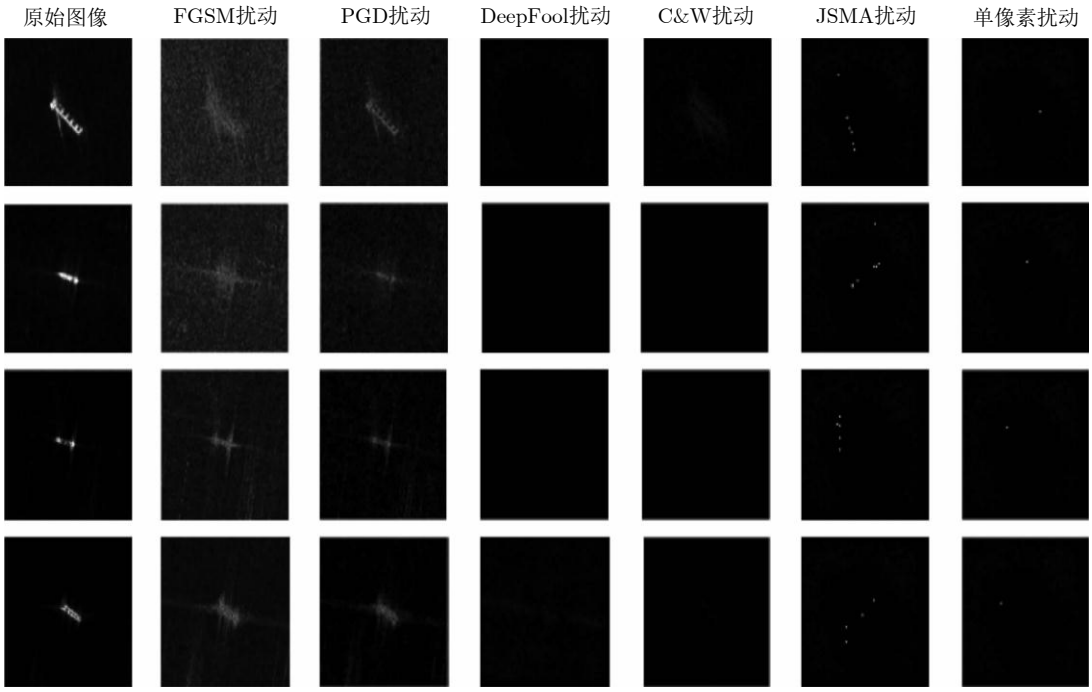

图 10 FUSAR-Ship数据子集对抗扰动举例[59]

Fig. 10　Adversarial perturbations on images from FUSAR-Ship dataset[59]

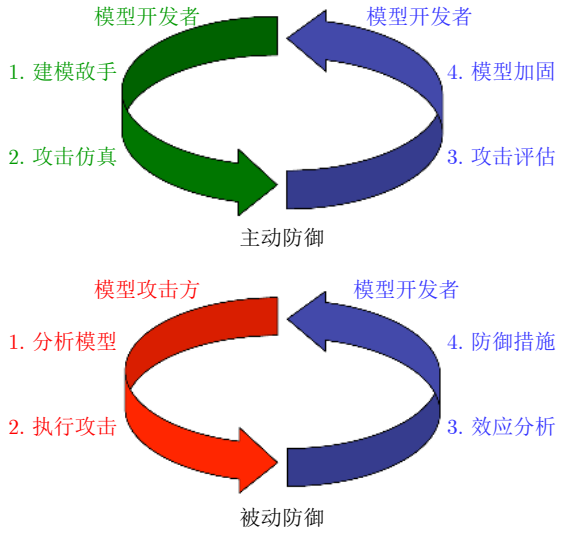

图 11 对抗攻击防御模型[60]

Fig. 11　Defense model for adversarial attacks[60]

型；然后仿真不同攻击目标、攻击知识、攻击策略和攻击能力情形下的攻击样式，对识别模型进行对抗攻击鲁棒性评估；最后设计并开发相关手段对模型进行加固，消除潜在的对抗风险。主动性防御技术的实现过程中不涉及真实的攻击敌手，是模型开发者自我模拟博弈对抗过程。被动性防御技术涉及真实对抗场景中模型攻击方与模型开发者之间的动态博弈进化。一方面，深度识别模型的攻击方通过分析模型的对抗脆弱性，设计并执行对抗攻击。为了达到更好的攻击效果，对抗攻击机理和样式不断演变，例如复合攻击和自动化攻击。另一方面，模型开发者通过分析对抗攻击给识别模型带来的多样化影响，研究提出新的对抗防御方法，并及时更新识别系统安全措施。

3.3.2 对抗攻击防御与检测

根据防御策略的不同，对抗攻击防御方法可以分为修改数据、修改模型和增加辅助模型等[26]，如图12所示。修改数据类方法基本思想是通过在训练阶段或测试阶段修改数据及特征实现防御，典型方法包括通过图像样本重建消除对抗扰动、压缩特征空间减小被攻击概率、引入对抗样本到训练集合中进行模型重训练、易干扰特征添加掩模、利用数据的不同属性关联提取鲁棒性特征、输入图像投影到训练数据流形等。修改模型类方法基本思想是修改从数据学习得到的模型结构或参数信息实现防御，典型方法包括网络蒸馏、网络验证、梯度正则化、鲁棒分类模型、可解释性机器学习模型和模型安全性掩模等。增加辅助模型方法通过引入额外的网络模型增强鲁棒性，典型策略包括对抗样本检测网络、多防御策略集成网络、生成模型网络等。

按照防御目标和防御策略的不同，表2对典型对抗攻击防御方法进行了总结分析。如表2所示，所有的防御方法都是在假设特定对抗攻击下进行评估的，PGD通常被认为是白盒攻击场景下评估防御方法的一种有效基准攻击。基于PGD攻击样本的对抗训练防御策略目前是对大多数攻击方法防御

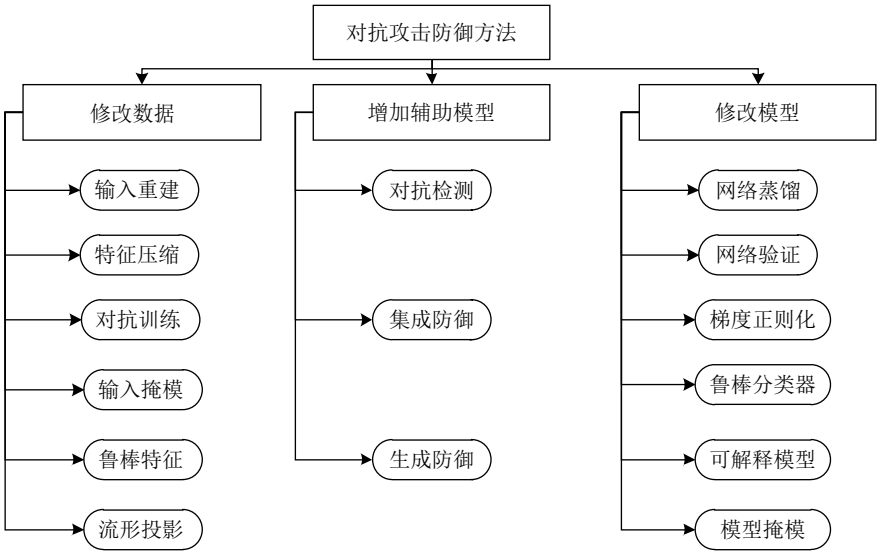

图 12 对抗攻击典型防御方法分类[26]

Fig. 12 Taxonomy of defense methods for adversarial attack[26]

表 2 对抗攻击防御方法

Tab. 2 Defense methods for adversarial attack

防御方法	防御目标	防御策略	攻击算法
Thermometer encoding[62]	主动防御	输入重建	PGD
VectorDefense[63]	主动防御	输入重建	BIM/JSMA/C&W/PGD
Super resolution[64]	主动防御	输入重建	FGSM/BIM/DF/C&W/MI-BIM
Pixel deflection[65]	主动防御	输入重建	FGSM/BIM/JSMA/DF/L-BFGS
D3[66]	主动防御	输入重建	FGSM/DF/C&W/UAP
RRP[67]	主动防御	预处理-输入随机变换	FGSM/DF/C&W
DR[68]	主动防御	特征压缩	FGSM
DeT[69]	主动防御	输入重建/增加辅助模型	FGSM/BIM/DF/C&W
Feature distillation[70]	主动防御	输入重建	FGSM/BIM/DF/C&W
MALADE[71]	主动防御	输入重建	FGSM/BIM/JSMA/C&W
JPEG compression[72]	主动防御	输入重建/集成重建	FGSM/ DF
SAP[73]	主动防御	模型掩模	FGSM
RSE[74]	主动防御	随机噪声层/集成预测	C&W
Deep defense[75]	主动防御	正则化	DF
Na *et al.*[76]	主动防御	正则化	FGSM/BIM/ILCM/C&W
Cao *et al.*[77]	主动防御	区域分类器	FGSM/BIM/JSMA/DF/ C&W
S2SNet[78]	主动防御	梯度掩模	FGSM/BIM /C&W
Adversarial training[8,36,79]	主动防御	对抗训练	PGD
Bilateral AT[80]	主动防御	改进对抗训练	FGSM/PGD
TRADES[81]	主动防御	改进对抗训练	PGD
SPROUT[82]	主动防御	改进对抗训练	PGD
CCNs[83]	主动防御	预处理	FGSM/ DF
DCNs[84]	主动防御	梯度掩模/预处理	L-BFGS
WSNNS[85]	主动防御	近邻度量	FGSM/PGD/C&W
ME-Net[86]	主动防御	预处理	FGSM/PGD/C&W/BA
Defense distillation[87]	主动防御	梯度掩模	JSMA

续表 2

防御方法	防御目标	防御策略	攻击算法
EDD[88]	主动防御	梯度掩模	FGSM/JSMA
Strauss *et al.*[89]	主动防御	集成防御	FGSM/BIM
Tramèr *et al.*[90]	主动防御	梯度掩模/集成防御	FGSM/BIM/ILCM
MTDeep[91]	主动防御	集成防御	FGSM/C&W
Defense-GAN[92]	主动防御	预处理	FGSM/C&W
APE-GAN[93]	主动防御	预处理	FGSM/JSMA/L-BFGS/DF/C&W
Zantedeschi *et al.*[94]	主动防御	梯度掩模	FGSM/JSMA
Parseval networks[95]	主动防御	梯度掩模	FGSM/BIM
HGD[96]	主动防御	预处理	FGSM/BIM
ALP[97]	主动防御	梯度掩模	PGD
Sinha *et al.*[98]	主动防御	梯度掩模	FGSM/BIM/PGD
Fortified networks[99]	主动防御	预处理	FGSM/PGD
DeepCloak[100]	主动防御	预处理	FGSM/JSMA/L-BFGS
DDSA[101]	主动防御	预处理	FGSM/M-BIM/C&W/PGD
ADV-BNN[102]	主动防御	梯度掩模	PGD
PixelDefend[103]	主动防御	预处理/近邻度量	FGSM/BIM/DF/C&W
Artifacts[104]	被动防御	对抗检测	FGSM/BIM/JSMA /C&W
AID[105]	被动防御	对抗检测	L-BFGS/FGSM
ConvFilter[106]	被动防御	预处理	L-BFGS
ReabsNet[107]	被动防御	预处理/辅助模型	FGSM/DF/C&W
MIP[108]	被动防御	统计对比/近邻度量	FGSM/BIM/DF
RCE[109]	被动防御	梯度掩模	FGSM/BIM/JSMA/C&W
NIC[110]	被动防御	辅助模型/近邻度量	FGSM/BIM/JSMA/C&W/DF
LID[111]	被动防御	被动防御	FGSM/BIM/JSMA/C&W
IFNN[112]	被动防御	被动防御	FGSM/BIM/DF/C&W
Gong *et al.*[113]	被动防御	辅助模型	FGSM/ JSMA
Metzen *et al.*[114]	被动防御	辅助模型	FGSM/BIM/DF
MagNet[115]	被动防御	预处理	FGSM/BIM/DF/C&W
MultiMagNet[116]	被动防御	预处理/近邻度量/集成防御	FGSM/BIM/DF/C&W
SafetyNet[117]	被动防御	辅助模型	FGSM/BIM/DF/JSMA
Feature squeezing[118]	被动防御	预处理	FGSM/BIM/C&W/JSMA
TwinNet[119]	被动防御	辅助模型/集成防御	UAP
Abbasi *et al.*[120]	被动防御	集成防御	FGSM/DF
Liang *et al.*[121]	被动防御	预处理	FGSM/DF/C&W

效果最好的一类防御方法。但是对抗训练会导致模型在干净数据集上泛化性能的下降，此外对抗训练过程涉及最大最小优化问题，训练过程十分耗时，在大规模数据集上的应用受限。

对抗样本检测方法可以看成是一类被动防御方法，对推理阶段的所有测试样本首先进行诊断，判断是否可能为恶意对抗样本。对抗样本检测与深度模型预测不确定性、分布外检测等领域紧密相关，核心思想是利用集成策略、度量方法、不一致性准则和生成性方法在推理阶段检测可靠泛化区域外的异常样本[122]。对抗样本检测4类典型方法的基本原理如图13所示。集成检测方法同时利用多个经过不同训练过程的深度神经网络识别模型。在推理阶段，多个识别模型分别独立产生输入数据的预测结果。多个网络的预测输出差别越大，那么该输入样本的决策错误可能性就越大。由于多个网络的决策边界之间存在差别，所以当对抗样本在分布内并且靠近决策边界时，该检测策略效果较好。但是分布

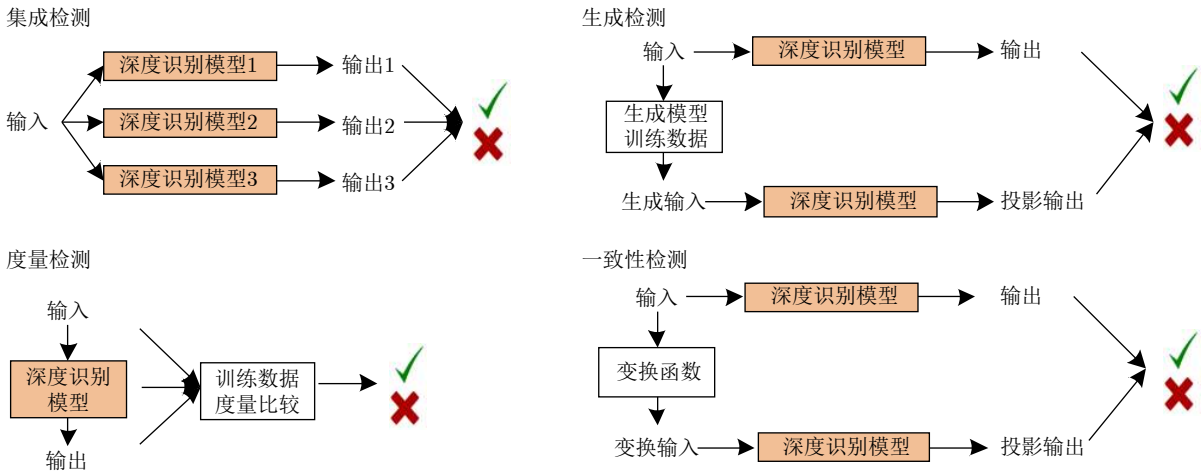

图 13 对抗样本检测方法[122]

Fig. 13 Adversarial example detection methods[122]

外对抗样本或者特定类型的对抗样本有可能在特征空间中远离决策边界，对于这些对抗样本需要采用其他的检测方法。如果测试样本的动态激活特性与训练数据集合中泛化区域内的样本动态激活特性相似，则度量检测方法判断该样本为正常样本；动态激活特性的大差异性表明对抗样本在可靠泛化区域外。生成检测方法是利用采用数据生成策略，判断测试样本是否在训练数据的生成流形上，通过计算偏移程度检测对抗攻击。度量检测方法通常涉及输入数据、多个隐含层、输入输出组合损失函数梯度等多个环节的变换比较。不一致性方法是采用图像变换方法将一个测试样本变换成多个不同版本，然后比较多版本增强图像是否存在输出不一致的问题，从而判断测试样本是否为对抗样本。

3.4 对抗鲁棒性评估进展

为了严格评估深度神经网络的对抗鲁棒性，文献中已经提出了大量评估准则或基准数据集[34,123–132]。防御评估的准则主要有：针对敌手进行防御、测试最坏情形下的鲁棒性、以人类识别能力衡量深度模型的进步等。防御评估的建议主要有：同时采用定向攻击和非定向攻击、进行消融实验、多样化测试设置、在多领域进行防御评估、采用随机性集成策略、利用迁移攻击、提供鲁棒性上限等。现有的对抗攻防评价测度通常采用简单的攻击成功率或分类正确率指标，导致模型输出评估不充分。例如在特定扰动幅度下攻击分类正确率不能衡量模型在对抗场景中的内在行为特性。针对图像分类任务中面向Lp范数约束对抗扰动和常见堕化扰动的深度模型鲁棒性评估，文献[127]提出了一组评估指标，如表3所示。表3的评估指标主要可以分为面向数据的评估测度和面向模型的评估测度，按照行为特性、架构、对抗扰动、堕化扰动、攻击知识和攻击模型等维度对评估指标进行细化分解。由于模型鲁棒性评估是采用一组扰动测试样本进行，因此首先采用神经元覆盖和数据不可感知度等面向数据的测度衡量测试样本的完整性。其次，采用决策边界距离变化、模型神经元敏感性和不确定性、堕化性能等指标评估模型在对抗场景中的动态特性。

在FUSAR-Ship数据集中我们通过人工选择4类数据样例数目较多、图像质量较好的SAR舰船目标图像420幅[59]，其中集装箱船122幅、货船158幅、渔船94幅和油轮46幅。每个类别选取80%样本作为训练样本，20%样本作为测试样本，选择在该数据子集上对抗鲁棒性较好[59]的ResNet101对4类白盒攻击方法(FGSM, PGD, DeepFool, C&W)和2类黑盒攻击(HSJA、单像素)方法进行评估，其中FGSM, PGD攻击无穷范数阈值设置为16，其他攻击方法为最小扰动。从深度模型误识别和扰动不可感知两个方面，从表3中选取代表性指标进行评估，其量化评估见表4，其中，分类正确率指标有：平均分类正确率、对抗类别平均置信度(Average Confidence for Adversarial Class, ACAC)、正确类别平均置信度(Average Confidence for True Class, ACTC)；平均Lp失真度(Average Lp Distortion, ALDp)衡量对抗样本与原始图像间的p范数距离，值越小失真越小，代表攻击效果更好；平均结构相似度(Average Structural Similarity, ASS)衡量攻击成功的对抗样本与原始图像间的自相似性，值越大代表对抗样本越难以用人眼进行识别；扰动敏感距离(Perturbation Sensitivity Distance, PSD)衡量人类感知扰动的指标，值越小代表越难以被人类视觉察觉；误分类与最大

表 3 对抗鲁棒性评估指标体系[127]

Tab. 3　Adversarial evaluation for deep models[127]

	评估指标	行为	架构	对抗扰动	堕化扰动	白盒	黑盒	单模型	多模型
数据	K多节神经元覆盖(KMNC)[123]			√	√	√		√	
	神经元边界覆盖(NBC)[123]			√	√	√		√	
	强神经元激活覆盖(SNAC)[123]			√	√	√		√	
	平均Lp失真度(ALDp)[124]			√			√	√	
	平均结构相似性(ASS)[125]			√			√	√	
	扰动敏感距离(PSD)[126]			√			√	√	
模型	干净数据集正确率(CA)[127]	√					√	√	
	白盒对抗正确率(AAW)[127]	√		√		√		√	
	黑盒对抗正确率(AAB)[127]	√		√			√		√
	对抗类别平均置信度(ACAC)[124]	√		√		√		√	
	正确类别平均置信度(ACTC)[124]	√		√		√	√	√	
	误分类与最大概率差(NTE)[124]	√		√		√		√	
	自然噪声平均差值(MCE)[132]	√			√		√	√	
	自然噪声相对差值(RMCE)[132]	√			√		√		√
	连续噪声分类差别(mFR)[132]	√			√		√		√
	分类准确率方差(CAV)[124]	√		√	√		√		√
	减少/增加错误分类百分比(CRR/CSR)[124]	√		√	√		√		√
	防御置信方差(CCV)[124]	√		√	√	√			√
	防御前后输出概率相似性(COS)[124]	√		√	√	√			√
	经验边界距离(EBD)[127]		√	√		√		√	
	经验边界距离2(EBD2)[127]		√	√		√		√	
	经验噪声敏感性(ENI)[128]		√	√	√			√	
	神经元敏感度(NS)[128]		√	√		√		√	
	神经元不确定性(NU)[128]		√	√	√	√		√	

表 4 SAR舰船目标识别深度模型对抗鲁棒性评估实例[59]

Tab. 4　Adversarial robustness evaluation of deep models for SAR ship recognition[59]

攻击方法	平均正确率(%)	ACAC	ACTC	ALDpL_0	ALDpL_2	ALDp$L_{\inf}$	ASS	PSD	NTE
FGSM	62.79	3.49	0.19	0.97	6396.70	16.00	0.08	11213.42	0.39
PGD	29.07	7.52	0.09	0.93	4330.60	16.00	0.23	7129.40	0.47
DeepFool	29.07	2.12	0.27	0.39	563.79	9.43	0.90	828.41	0.25
C&W	40.70	1.39	0.37	0.24	412.92	8.80	0.96	468.21	0.12
HSJA	62.79	1.14	0.41	0.74	1892.80	7.75	0.68	3220.93	0.06
单像素	79.82	INF	0	2e-5	255.00	255.00	0.90	643.32	0.14

概率差(Noise Tolerance Estimation, NTE)衡量对抗样本在保持错误分类不变的情况下所能忍受的噪声，值越大代表攻击方法更加稳健。

4　开放性问题

深度卷积神经网络图像识别模型的对抗鲁棒性与其泛化性、安全性、隐私性和可解释性等特性紧密相关，近年来在学术界和工业界都进行了广泛而深入的研究，大量研究成果不断涌现，然而仍有许多开放性问题值得重点关注。

(1) 深度卷积神经网络识别模型的对抗脆弱性成因在理论上还需要进一步深入研究[133–139]。目前关于对抗样本在理论上为何存在及其特性描述等基础性问题学术界研究还没有形成统一的认识。深度卷积神经网络图像识别模型对抗鲁棒性与模型泛化性、模型堕化噪声鲁棒性之间的关系在理论上和实践中仍需进一步研究。

(2) 利用无监督数据提升深度识别模型的对抗鲁棒性是未来重要研究方向[140–144]。目前对抗鲁棒性最有效的提升方法是采用最大化模型损失的对抗样本重训练深度网络模型，对抗训练过程十分耗时。此外，鲁棒深度识别模型的样本采样复杂度要比标准模型更高，因此需要更大规模的高质量标记数据集。在许多应用领域中，大规模高质量标记数据集获取不仅十分耗时，而且代价昂贵。充分挖掘无标记数据中的潜在语义关系和因果关系将大大降低鲁棒识别模型学习算法对标记数据的严重依赖。图14为本研究团队开展的基于无监督对抗扰动的深度识别模型对抗鲁棒性提升结果。在NWPU-RESISC45数据集上[145]，采用ResNet18网络结构[146]，利用 BYOL对比学习过程的梯度下降产生无监督对抗扰动[147,148]，并最大化每个实例图像与其无监督对抗扰动版本间的相似性，构造更加稳健的预训练特征编码网络。特征编码器的训练过程无需任何标记数据，每个类别选取400幅图像，训练过程不使用标记信息。在经过微调(每个类别选取200幅图像进行有监督学习)后，可以获得鲁棒性更强的识别模型。对比标准模型(每个类别选取600幅有监督图像进行训练)和无监督鲁棒提升模型(每个类别选取400幅图像进行无监督对抗对比预训练，200幅图像进行有监督微调)在正常样本及PGD攻击样本的激活情况，可以看到：通过无监督数据可以将深度模型的特征编码更加聚焦在显著目标区域，减小非鲁棒性特征对识别结果的影响。

(3) 多传感器耦合对抗攻击与防御将更具实际应用价值[149–153]。在自动驾驶和军事侦察等多种应用场景同时存在光电和微波等多类图像传感器，现有的攻击算法重点关注光电对抗智能扰动技术，很容易在其他波段暴露。深度学习在多源图像处理任务结构上的相似性会导致耦合攻击风险。在So2SatLCZ42标准数据集上[154]选择居民区(训练样本256幅，测试样本266幅)、工业区(训练样本860幅，测试样本905幅)、林区(训练样本2287幅，测试样本2365幅)、沙地(训练样本672幅，测试样本570幅)和水体(训练样本2609幅，测试样本2530幅) 5类数据子集，数据子集中的示例图像如图15所示，第1行为地物空间分布示意，第2行为SAR样例图像，第3行为SAR样例图像对应的光学图像(已完成空间几何配准)。

实验中选择光学图像(RGB波段)和SAR图像(垂直极化)数据分别训练ResNet18模型，优化器选用Adam优化器，学习率设为10^{-3}，batch_size大小设为256，得到光学识别模型和SAR识别模型。采用修改后的单像素对抗样本生成方法[52]攻击两个识别模型，首先对初代种子的参数进行初始化，其中噪声点位置初始化为32×32图像中均匀随机分布，噪声强度信息初始化服从高斯分布，初代投放100个种子，经100次种群选择，最终挑选出攻击效果最佳的对抗样本。如图16所示，仅仅扰动同一个像素位置的数字值，光学图像和SAR图像添加的扰动幅度不同，便可以同时欺骗光学和SAR图像识别模型。统计结果表明：在协同攻击同一位置像素的情况下，可以将光学识别模型的正确率由92.36%降低到30.98%，同时将SAR识别模型的正确率由81.24%降低到42.07%。

5 结语

基于深度卷积神经网络模型的新一代智能化图

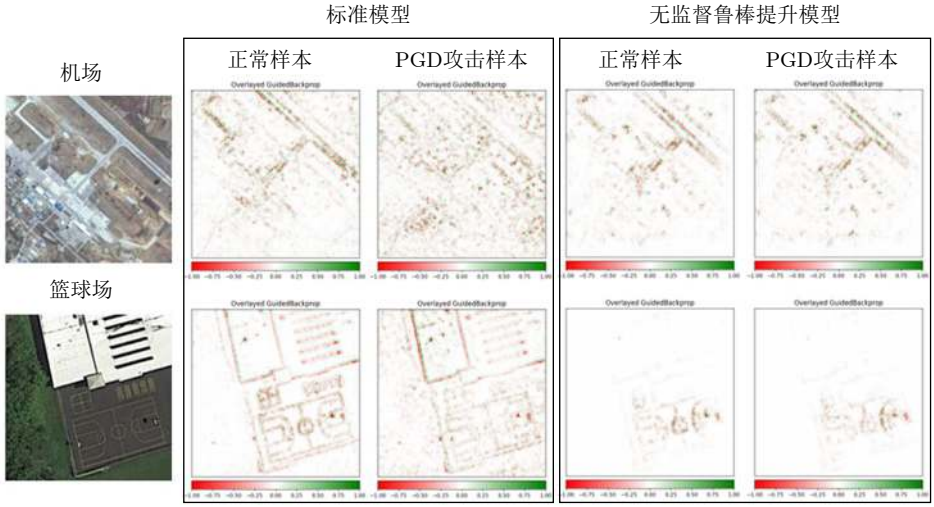

图 14 无监督数据提升对抗鲁棒性

Fig. 14 Unlabeled data for improving adversarial robustness

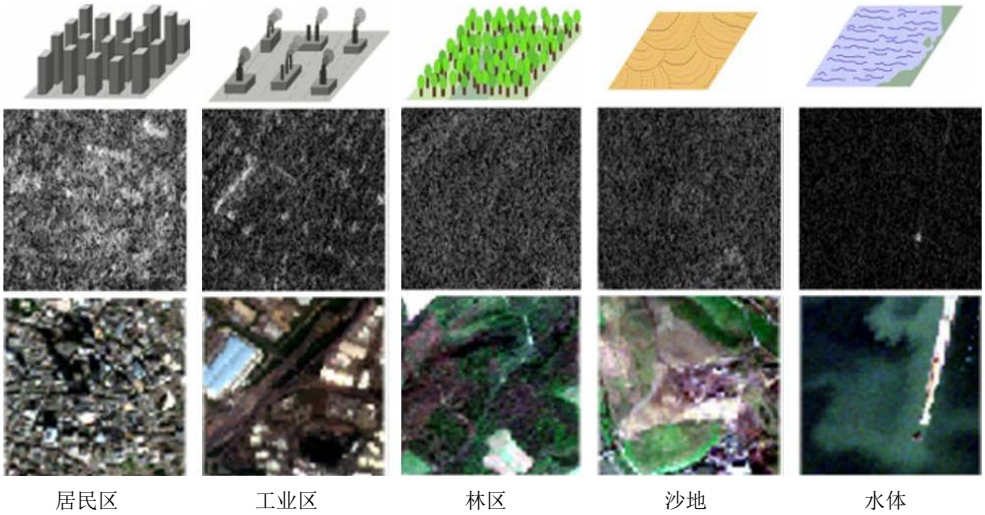

图 15 So2Sat LCZ42数据子集示例[154]

Fig. 15　Examples images from the So2Sat LCZ42 dataset[154]

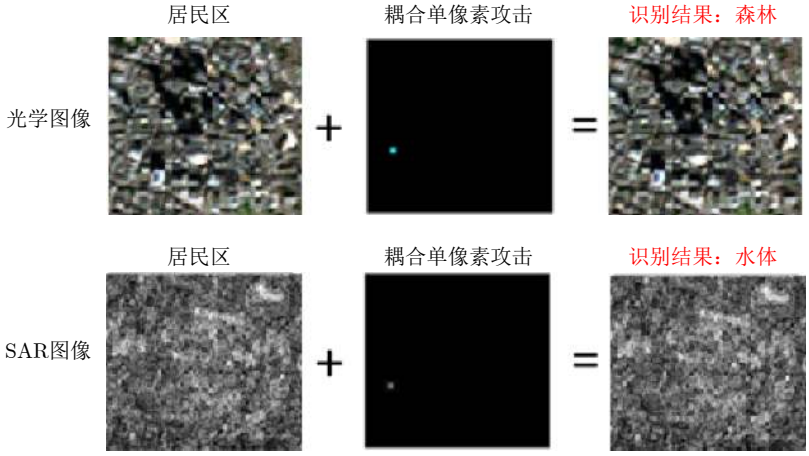

图 16 多传感器耦合对抗攻击实例

Fig. 16　Adversarial attacks for multiple sensors

像识别系统已逐步在医疗、安防、自动驾驶和军事等安全敏感领域广泛部署。然而现有深度识别模型依赖大规模高质量的训练数据，只能提供有限的可靠性能保证，并且缺乏可解释性，给模型在复杂电磁环境下强对抗场景中的实际应用带来严重安全隐患。本文从信息安全、对抗攻防威胁模型两个方面系统总结了深度神经网络图像识别模型对抗脆弱性成因与对抗鲁棒性研究进展，重点梳理了对抗样本生成、主被动对抗防御、对抗鲁棒性评估等方面的技术思路与典型方法，为下一步建立鲁棒可信的高性能智能化图像识别系统提供参考。

参 考 文 献

[1] BERGHOFF C, NEU M, and VON TWICKEL A. Vulnerabilities of connectionist AI applications: Evaluation and defense[J]. *Frontiers in Big Data*, 2020, 3: 23. doi: 10.3389/fdata.2020.00023.

[2] 潘宗序, 安全智, 张冰尘. 基于深度学习的雷达图像目标识别研究进展[J]. 中国科学: 信息科学, 2019, 49(12): 1626–1639. doi: 10.1360/SSI-2019-0093.

PAN Zongxu, AN Quanzhi, and ZHANG Bingchen. Progress of deep learning-based target recognition in radar images[J]. *Scientia Sinica Informationis*, 2019, 49(12): 1626–1639. doi: 10.1360/SSI-2019-0093.

[3] 徐丰, 王海鹏, 金亚秋. 深度学习在SAR目标识别与地物分类中的应用[J]. 雷达学报, 2017, 6(2): 136–148. doi: 10.12000/JR16130.

XU Feng, WANG Haipeng, and JIN Yaqiu. Deep learning as applied in SAR target recognition and terrain classification[J]. *Journal of Radars*, 2017, 6(2): 136–148. doi: 10.12000/JR16130.

[4] CHENG Gong, XIE Xingxing, HAN Junwei, *et al*. Remote sensing image scene classification meets deep learning: Challenges, methods, benchmarks, and opportunities[J].

IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 13: 3735–3756. doi: 10.1109/JSTARS.2020.3005403.

[5] BLASCH E, MAJUMDER U, ZELNIO E, *et al*. Review of recent advances in AI/ML using the MSTAR data[C]. SPIE 11393, Algorithms for Synthetic Aperture Radar Imagery XXVII, Online Only, 2020.

[6] SCHÖLKOPF B, LOCATELLO F, BAUER S, *et al*. Towards causal representation learning[EB/OL]. https://arxiv.org/abs/2102.11107v1, 2021.

[7] SZEGEDY C, ZAREMBA W, SUTSKEVER I, *et al*. Intriguing properties of neural networks[EB/OL]. https://arxiv.org/abs/1312.6199v1, 2014.

[8] GOODFELLOW I J, SHLENS J, and SZEGEDY C. Explaining and harnessing adversarial examples[EB/OL]. https://arxiv.org/abs/1412.6572v1, 2015.

[9] MACHADO G R, SILVA E, and GOLDSCHMIDT R R. Adversarial machine learning in image classification: A survey towards the defender's perspective[EB/OL]. https://arxiv.org/abs/2009.03728v1, 2020.

[10] SERBAN A, POLL E, and VISSER J. Adversarial examples on object recognition: A comprehensive survey[J]. *ACM Computing Surveys*, 2020, 53(3): 66. doi: 10.1145/3398394.

[11] OSENI A, MOUSTAFA N, JANICKE H, *et al*. Security and privacy for artificial intelligence: Opportunities and challenges[EB/OL]. https://arxiv.org/abs/2102.04661, 2021.

[12] KEYDEL E R, LEE S W, and MOORE J T. MSTAR extended operating conditions: A tutorial[C]. SPIE 2757, Algorithms for Synthetic Aperture Radar Imagery III, Orlando, United States, 1996: 228–242.

[13] SIMONYAN K and ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[EB/OL]. https://arxiv.org/abs/1409.1556v4, 2015.

[14] 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, San Francisco, USA, 2019: 707–723. doi: 10.1109/SP.2019.00031.

[15] SHEN Juncheng, ZHU Xiaolei, and MA De. TensorClog: An imperceptible poisoning attack on deep neural network applications[J]. *IEEE Access*, 2019, 7: 41498–41506. doi: 10.1109/ACCESS.2019.2905915.

[16] QUIRING E and RIECK K. Backdooring and poisoning neural networks with image-scaling attacks[C]. 2020 IEEE Security and Privacy Workshops, San Francisco, USA, 2020: 41–47. doi: 10.1109/SPW50608.2020.00024.

[17] SINGH J and SHARMILA V C. Detecting Trojan attacks on deep neural networks[C]. The 4th International Conference on Computer, Communication and Signal Processing, Chennai, India, 2020: 1–5. doi: 10.1109/ICCCSP49186.2020.9315256.

[18] HE Yingzhe, MENG Guozhu, CHEN Kai, *et al*. Towards security threats of deep learning systems: A survey[EB/OL]. https://arxiv.org/abs/1911.12562v2, 2020.

[19] DELDJOO Y, NOIA T D, and MERRA F A. Adversarial machine learning in recommender systems: State of the art and challenges[EB/OL]. https://arxiv.org/abs/2005.10322v1, 2020.

[20] BIGGIO B and ROLI F. Wild patterns: Ten years after the rise of adversarial machine learning[J]. *Pattern Recognition*, 2018, 84: 317–331. doi: 10.1016/j.patcog.2018.07.023.

[21] ROSS T D, BRADLEY J J, HUDSON L J, *et al*. SAR ATR: So what's the problem? An MSTAR perspective[C]. SPIE 3721, Algorithms for Synthetic Aperture Radar Imagery VI, Orlando, USA, 1999: 606–610.

[22] 成科扬, 王宁, 师文喜, 等. 深度学习可解释性研究进展[J]. 计算机研究与发展, 2020, 57(6): 1208–1217. doi: 10.7544/issn1000-1239.2020.20190485.

CHENG Keyang, WANG Ning, SHI Wenxi, *et al*. Research advances in the interpretability of deep learning[J]. *Journal of Computer Research and Development*, 2020, 57(6): 1208–1217. doi: 10.7544/issn1000-1239.2020.20190485.

[23] 化盈盈, 张岱墀, 葛仕明. 深度学习模型可解释性的研究进展[J]. 信息安全学报, 2020, 5(3): 1–12. doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01.

HUA Yingying, ZHANG Daichi, and GE Shiming. Research progress in the interpretability of deep learning models[J]. *Journal of Cyber Security*, 2020, 5(3): 1–12. doi: 10.19363/J.cnki.cn10-1380/tn.2020.05.01.

[24] 郭炜炜, 张增辉, 郁文贤, 等. SAR图像目标识别的可解释性问题探讨[J]. 雷达学报, 2020, 9(3): 462–476. doi: 10.12000/JR20059.

GUO Weiwei, ZHANG Zenghui, YU Wenxian, *et al*. Perspective on explainable SAR target recognition[J]. *Journal of Radars*, 2020, 9(3): 462–476. doi: 10.12000/JR20059.

[25] ORTIZ-JIMENEZ G, MODAS A, MOOSAVI-DEZFOOLI S M, *et al*. Optimism in the face of adversity: Understanding and improving deep learning through adversarial robustness[EB/OL]. HYPERLINK "https://arxiv. org/abs/2010.09624v2" https://arxiv.org/abs/2010.09624v2, 2021.

[26] QAYYUM A, USAMA M, QADIR J, *et al*. Securing

connected & autonomous vehicles: Challenges posed by adversarial machine learning and the way forward[J]. *IEEE Communications Surveys & Tutorials*, 2020, 22(2): 998–1026. doi: 10.1109/COMST.2020.2975048.

[27] YUAN Xiaoyong, HE Pan, ZHU Qile, *et al*. . Adversarial examples: Attacks and defenses for deep learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(9): 2805–2824. doi: 10.1109/TNNLS.2018.2886017.

[28] KURAKIN A, GOODFELLOW I, and BENGIO S. Adversarial examples in the physical world[EB/OL]. https://arxiv.org/abs/1607.02533v4, 2017.

[29] 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, Saarbruecken, Germany, 2016: 372–387.

[30] 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.

[31] KARMON D, ZORAN D, and GOLDBERG Y. LaVAN: Localized and visible adversarial noise[C]. The 35th International Conference on Machine Learning, Stockholm, Sweden, 2018.

[32] HAYES J and DANEZIS G. Learning universal adversarial perturbations with generative models[C]. 2018 IEEE Security and Privacy Workshops, San Francisco, USA, 2018: 43–49.

[33] ATHALYE A, ENGSTROM L, ILYAS A, *et al*. Synthesizing robust adversarial examples[C]. The 35th International Conference on Machine Learning, Stockholm, Sweden, 2018.

[34] CARLINI N and WAGNER D. Towards evaluating the robustness of neural networks[C]. 2017 IEEE Symposium on Security and Privacy, San Jose, USA, 2017: 39–57.

[35] ROZSA A, RUDD E M, and BOULT T E. Adversarial diversity and hard positive generation[EB/OL]. https://arxiv.org/abs/1605.01775, 2016.

[36] MADRY A, MAKELOV A, SCHMIDT L, *et al*. Towards deep learning models resistant to adversarial attacks[EB/OL]. https://arxiv.org/abs/1706.06083, 2019.

[37] CHEN Pinyu, SHARMA Y, ZHANG Huan, *et al*. EAD: Elastic-net attacks to deep neural networks via adversarial examples[EB/OL]. https://arxiv.org/abs/1709.04114, 2018.

[38] EYKHOLT K, EVTIMOV I, FERNANDES E, *et al*. Robust physical-world attacks on deep learning models[EB/OL]. https://arxiv.org/abs/1707.08945, 2018.

[39] CARLINI N, KATZ G, BARRETT C, *et al*. Ground-truth adversarial examples[EB/OL]. https://arxiv.org/abs/1709.10207, 2018.

[40] HE W, LI Bo, and SONG D. Decision boundary analysis of adversarial examples[C]. 2018 International Conference on Learning Representations, Vancouver, Canada, 2018.

[41] BALUJA S and FISCHER I. Adversarial transformation networks: Learning to generate adversarial examples[EB/OL]. https://arxiv.org/abs/1703.09387, 2017.

[42] DONG Yinpeng, LIAO Fangzhou, PANG Tianyu, *et al*. Boosting adversarial attacks with momentum[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 9185–9193.

[43] CHEN Jinyin, SU Mengmeng, SHEN Shijing, *et al*. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm[J]. *Computers & Security*, 2019, 85: 89–106. doi: 10.1016/j.cose.2019.04.014.

[44] TU Chunchen, TING Paishun, CHEN Pinyu, *et al*. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks[EB/OL]. https://arxiv.org/abs/1805.11770v5, 2020.

[45] NARODYTSKA N and KASIVISWANATHAN S. Simple black-box adversarial attacks on deep neural networks[C]. 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Honolulu, USA, 2017: 1310–1318.

[46] ILYAS A, ENGSTROM L, ATHALYE A, *et al*. Black-box adversarial attacks with limited queries and information[EB/OL]. https://arxiv.org/abs/1804.08598, 2018.

[47] BRENDEL W, RAUBER J, and BETHGE M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models[EB/OL]. https://arxiv.org/abs/1712.04248, 2018.

[48] ALZANTOT M, SHARMA Y, CHAKRABORTY S, *et al*. GenAttack: Practical black-box attacks with gradient-free optimization[EB/OL]. https://arxiv.org/abs/1805.11090, 2019.

[49] CHEN Pinyu, ZHANG Huan, SHARMA Y, *et al*. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models[C]. The 10th ACM Workshop on Artificial Intelligence and Security, Dallas, USA, 2017: 15–26.

[50] SARKAR S, BANSAL A, MAHBUB U, *et al*. UPSET and ANGRI: Breaking high performance image classifiers[EB/OL]. https://arxiv.org/abs/1707.01159, 2017.

[51] CHEN Jianbo, JORDAN M I, and WAINWRIGHT M J. HopSkipJumpAttack: A query-efficient decision-based attack[C]. 2020 IEEE Symposium on Security and Privacy, San Francisco, USA, 2020: 1277–1294.

[52] 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.

[53] ATHALYE A, CARLINI N, and WAGNER D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples[EB/OL]. https://arxiv.org/abs/1802.00420, 2018.

[54] UESATO J, O'DONOGHUE B, VAN DEN OORD A, *et al*. Adversarial risk and the dangers of evaluating against weak attacks[EB/OL]. https://arxiv.org/abs/1802.05666, 2018.

[55] XIAO Chaowei, LI Bo, ZHU Junyan, *et al*. Generating adversarial examples with adversarial networks[EB/OL]. https://arxiv.org/abs/1801.02610, 2019.

[56] CISSE M, ADI Y, NEVEROVA N, *et al*. Houdini: Fooling deep structured prediction models[EB/OL]. https://arxiv.org/abs/1707.05373, 2017.

[57] SELVARAJU R R, COGSWELL M, DAS A, *et al*. Grad-CAM: Visual explanations from deep networks via gradient-based localization[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 618–626. doi: 10.1109/ICCV.2017.74.

[58] HOU Xiyue, AO Wei, SONG Qian, *et al*. FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition[J]. *Science China Information Sciences*, 2020, 63(4): 140303. doi: 10.1007/s11432-019-2772-5.

[59] 徐延杰, 孙浩, 雷琳, 等. 基于对抗攻击的SAR舰船识别卷积神经网络鲁棒性研究[J]. 信号处理, 2020, 36(12): 1965–1978. doi: 10.16798/j.issn.1003-0530.2020.12.002.

XU Yanjie, SUN Hao, LEI Lin, *et al*. The research for the robustness of SAR ship identification based on adversarial example[J]. *Journal of Signal Processing*, 2020, 36(12): 1965–1978. doi: 10.16798/j.issn.1003-0530.2020.12.002.

[60] BIGGIO B, FUMERA G, and ROLI F. Security evaluation of pattern classifiers under attack[J]. *IEEE Transactions on Knowledge and Data Engineering*, 2014, 26(4): 984–996. doi: 10.1109/TKDE.2013.57.

[61] 李盼, 赵文涛, 刘强, 等. 机器学习安全性问题及其防御技术研究综述[J]. 计算机科学与探索, 2018, 12(2): 171–184. doi: 10.3778/j.issn.1673-9418.1708038.

LI Pan, ZHAO Wentao, LIU Qiang, *et al*. Security issues and their countermeasuring techniques of machine learning: A survey[J]. *Journal of Frontiers of Computer Science and Technology*, 2018, 12(2): 171–184. doi: 10.3778/j.issn.1673-9418.1708038.

[62] BUCKMAN J, ROY A, RAFFEL C, *et al*. Thermometer encoding: One hot way to resist adversarial examples[C]. The ICLR 2018, Vancouver, Canada, 2018.

[63] KABILAN V M, MORRIS B, and NGUYEN A. VectorDefense: Vectorization as a defense to adversarial examples[EB/OL]. https://arxiv.org/abs/1804.08529, 2018.

[64] MUSTAFA A, KHAN S H, HAYAT M, *et al*. Image super-resolution as a defense against adversarial attacks[J]. *IEEE Transactions on Image Processing*, 2019, 29: 1711–1724. doi: 10.1109/TIP.2019.2940533.

[65] PRAKASH A, MORAN N, GARBER S, *et al*. Deflecting adversarial attacks with pixel deflection[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 8571–8580.

[66] MOOSAVI-DEZFOOLI S M, SHRIVASTAVA A, and TUZEL O. Divide, denoise, and defend against adversarial attacks[EB/OL]. https://arxiv.org/abs/1802.06806, 2019.

[67] XIE Cihang, WANG Jianyu, ZHANG Zhishuai, *et al*. Mitigating adversarial effects through randomization[EB/OL]. https://arxiv.org/abs/1711.01991, 2018.

[68] BHAGOJI A N, CULLINA D, SITAWARIN C, *et al*. Enhancing robustness of machine learning systems via data transformations[EB/OL]. https://arxiv.org/abs/1704.02654, 2017.

[69] LI Changjiang, WENG Haiqin, JI Shouling, *et al*. DeT: Defending against adversarial examples via decreasing transferability[C]. The 11th International Symposium on Cyberspace Safety and Security, Guangzhou, China, 2019: 307–322.

[70] LIU Zihao, LIU Qi, LIU Tao, *et al*. Feature distillation: DNN-oriented JPEG compression against adversarial examples[EB/OL]. https://arxiv.org/abs/1803.05787, 2019.

[71] SRINIVASAN V, MARBAN A, MULLER K R, *et al*. Robustifying models against adversarial attacks by langevin dynamics[EB/OL]. https://arxiv.org/abs/1805.12017, 2019.

[72] DAS N, SHANBHOGUE M, CHEN S T, *et al*. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression[EB/OL]. https://arxiv.org/abs/1705.02900, 2017.

[73] DHILLON G S, AZIZZADENESHELI K, LIPTON Z C, *et al*. Stochastic activation pruning for robust adversarial defense[EB/OL]. https://arxiv.org/abs/1803.01442, 2018.

[74] LIU Xuanqing, CHENG Minhao, ZHANG Huan, *et al*. Towards robust neural networks via random self-ensemble[EB/OL]. https://arxiv.org/abs/1712.00673, 2018.

[75] YAN Ziang, GUO Yiwen, and ZHANG Changshui. Deep defense: Training DNNs with improved adversarial robustness[C]. 32nd Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 419–428.

[76] NA T, KO J H, and MUKHOPADHYAY S. Cascade adversarial machine learning regularized with a unified embedding[EB/OL]. https://arxiv.org/abs/1708.02582, 2018.

[77] CAO Xiaoyu and GONG Zhenqiang. Mitigating evasion attacks to deep neural networks via region-based classification[C]. The 33rd Annual Computer Security Applications Conference, Orlando, USA, 2017: 278–287.

[78] FOLZ J, PALACIO S, HEES J, *et al*. Adversarial defense based on structure-to-signal autoencoders[EB/OL]. https://arxiv.org/abs/1803.07994, 2018.

[79] BAI Tao, LUO Jinqi, ZHAO Jun, *et al*. Recent advances in adversarial training for adversarial robustness[EB/OL]. http://arxiv.org/abs/2102.01356v4, 2021.

[80] WANG Jianyu and ZHANG Haichao. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks[C]. 2019 IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 2019: 6629–6638.

[81] ZHANG Hongyang, YU Yaodong, JIAO Jiantao, *et al*. Theoretically principled trade-off between robustness and accuracy[C]. The 36th International Conference on Machine Learning, Long Beach, California, 2019: 7472–7482.

[82] CHENG Minhao, CHEN Pinyu, LIU Sijia, *et al*. Self-progressing robust training[EB/OL]. https://arxiv.org/abs/2012.11769, 2020.

[83] RANJAN R, SANKARANARAYANAN S, CASTILLO C D, *et al*. Improving network robustness against adversarial attacks with compact convolution[EB/OL]. https://arxiv.org/abs/1712.00699, 2018.

[84] GU Shixiang and RIGAZIO L. Towards deep neural network architectures robust to adversarial examples[EB/OL]. http://arxiv.org/abs/1412.5068, 2015.

[85] DUBEY A, VAN DER MAATEN L, YALNIZ Z, *et al*. Defense against adversarial images using web-scale nearest-neighbor search[C]. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 2019: 8767–8776.

[86] YANG Yuzhe, ZHANG Guo, KATABI D, *et al*. ME-Net: Towards effective adversarial robustness with matrix estimation[EB/OL]. https://arxiv.org/abs/1905.11971, 2019.

[87] PAPERNOT N, MCDANIEL P, WU Xi, *et al*. Distillation as a defense to adversarial perturbations against deep neural networks[C]. 2016 IEEE Symposium on Security and Privacy, San Jose, USA, 2016: 582–597.

[88] PAPERNOT N and MCDANIEL P. Extending defensive distillation[EB/OL]. https://arxiv.org/abs/1705.05264, 2017.

[89] STRAUSS T, HANSELMANN M, JUNGINGER A, *et al*. Ensemble methods as a defense to adversarial perturbations against deep neural networks[EB/OL]. https://arxiv.org/abs/1709.03423, 2018.

[90] TRAMÈR F, KURAKIN A, PAPERNOT N, *et al*. Ensemble adversarial training: Attacks and defenses[EB/OL]. http://arxiv.org/abs/1705.07204, 2020.

[91] SENGUPTA S, CHAKRABORTI T, and KAMBHAMPATI S. MTDeep: Boosting the security of deep neural nets against adversarial attacks with moving target defense[EB/OL]. http://arxiv.org/abs/1705.07213, 2019.

[92] SAMANGOUEI P, KABKAB M, and CHELLAPPA R. Defense-GAN: Protecting classifiers against adversarial attacks using generative models[EB/OL]. http://arxiv.org/abs/1805.06605, 2018.

[93] SHEN Shiwei, JIN Guoqing, GAO Ke, *et al*. APE-GAN: Adversarial perturbation elimination with GAN[EB/OL]. http://arxiv.org/abs/1707.05474, 2017.

[94] ZANTEDESCHI V, NICOLAE M I, and RAWAT A. Efficient defenses against adversarial attacks[C]. The 10th ACM Workshop on Artificial Intelligence and Security, Dallas, USA, 2017: 39–49.

[95] CISSE M, BOJANOWSKI P, GRAVE E, *et al*. Parseval networks: Improving robustness to adversarial examples[C]. The 34th International Conference on Machine Learning, Sydney, Australia, 2017: 854–863.

[96] LIAO Fangzhou, LIANG Ming, DONG Yinpeng, *et al*. Defense against adversarial attacks using high-level representation guided denoiser[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, USA, 2018: 1778–1787.

[97] KANNAN H, KURAKIN A, and GOODFELLOW I. Adversarial logit pairing[EB/OL]. http://arxiv.org/abs/1803.06373, 2018.

[98] SINHA A, NAMKOONG H, VOLPI R, *et al*. Certifying some distributional robustness with principled adversarial training[EB/OL]. http://arxiv.org/abs/1710.10571, 2020.

[99] LAMB A, BINAS J, GOYAL A, *et al*. Fortified networks: Improving the robustness of deep networks by modeling the manifold of hidden representations[EB/OL]. https://arxiv.org/abs/1804.02485, 2018.

[100] GAO Ji, WANG Beilun, LIN Zeming, *et al*. DeepCloak: Masking deep neural network models for robustness against adversarial samples[EB/OL]. http://arxiv.org/abs/1702.06763, 2017.

[101] BAKHTI Y, FEZZA S A, HAMIDOUCHE W, *et al*. DDSA: A defense against adversarial attacks using deep denoising sparse autoencoder[J]. *IEEE Access*, 2019, 7: 160397–160407. doi: 10.1109/ACCESS.2019.2951526.

[102] LIU Xuanqing, LI Yao, WU Chongruo, *et al*. Adv-BNN: Improved adversarial defense through robust Bayesian

neural network[EB/OL]. http://arxiv.org/abs/1810.01279, 2019.

[103] SONG Yang, KIM T, NOWOZIN S, *et al*. PixelDefend: Leveraging generative models to understand and defend against adversarial examples[EB/OL]. http://arxiv.org/abs/1710.10766, 2018.

[104] FEINMAN R, CURTIN R R, SHINTRE S, *et al*. Detecting adversarial samples from artifacts[EB/OL]. http://arxiv.org/abs/1703.00410, 2017.

[105] CARRARA F, FALCHI F, CALDELLI R, *et al*. Adversarial image detection in deep neural networks[J]. *Multimedia Tools and Applications*, 2019, 78(3): 2815–2835. doi: 10.1007/s11042-018-5853-4.

[106] LI Xin and LI Fuxin. Adversarial examples detection in deep networks with convolutional filter statistics[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 5775–5783.

[107] CHEN Jiefeng, MENG Zihang, SUN Changtian, *et al*. ReabsNet: Detecting and revising adversarial examples[EB/OL]. http://arxiv.org/abs/1712.08250, 2017.

[108] ZHENG Zhihao and HONG Pengyu. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks[C]. The 32nd Conference on Neural Information Processing Systems, Montréal, Canada, 2018: 7913–7922.

[109] PANG Tianyu, DU Chao, and ZHU Jun. Robust deep learning via reverse cross-entropy training and thresholding test[EB/OL]. https://arxiv.org/abs/1706.00633v1, 2018.

[110] MA Shiqing, LIU Yingqi, TAO Guanhong, *et al*. NIC: Detecting adversarial samples with neural network invariant checking[C]. 2019 Network and Distributed Systems Security Symposium, San Diego, USA, 2019: 1–15.

[111] MA Xingjun, LI Bo, WANG Yisen, *et al*. Characterizing adversarial subspaces using local intrinsic dimensionality[EB/OL]. http://arxiv.org/abs/1801.02613, 2018.

[112] COHEN G, SAPIRO G, and GIRYES R. Detecting adversarial samples using influence functions and nearest neighbors[EB/OL]. http://arxiv.org/abs/1909.06872, 2020.

[113] GONG Zhitao, WANG Wenlu, and KU W S. Adversarial and clean data are not twins[EB/OL]. http://arxiv.org/abs/1704.04960, 2017.

[114] METZEN J H, GENEWEIN T, FISCHER V, *et al*. On detecting adversarial perturbations[EB/OL]. http://arxiv.org/abs/1702.04267, 2017.

[115] MENG Dongyu and CHEN Hao. MagNet: A two-pronged defense against adversarial examples[C]. The 2017 ACM SIGSAC Conference on Computer and Communications Security, Dallas, USA, 2017: 135–147.

[116] MACHADO G R, GOLDSCHMIDT R R, and SILVA E. MultiMagNet: A non-deterministic approach based on the formation of ensembles for defending against adversarial images[C]. The 21st International Conference on Enterprise Information Systems, Heraklion, Greece, 2019: 307–318.

[117] LU Jiajun, ISSARANON T, and FORSYTH D. SafetyNet: Detecting and rejecting adversarial examples robustly[C]. 2017 IEEE International Conference on Computer Vision, Venice, Italy, 2017: 446–454.

[118] XU Weilin, EVANS D, and Qi Yanjun. Feature squeezing: Detecting adversarial examples in deep neural networks[EB/OL]. https://arxiv.org/abs/1704.01155, 2017.

[119] RUAN Yibin and DAI Jiazhu. TwinNet: A double sub-network framework for detecting universal adversarial perturbations[J]. *Future Internet*, 2018, 10(3): 26. doi: 10.3390/fi10030026.

[120] ABBASI M and GAGNÉ C. Robustness to adversarial examples through an ensemble of specialists[EB/OL]. http://arxiv.org/abs/1702.06856, 2017.

[121] LIANG Bin, LI Hongcheng, SU Miaoqiang, *et al*. Detecting adversarial image examples in deep networks with adaptive noise reduction[EB/OL]. http://arxiv.org/abs/1705.08378, 2019.

[122] LUST J and CONDURACHE A P. A survey on assessing the generalization envelope of deep neural networks at inference time for image classification[EB/OL]. http://arxiv.org/abs/2008.09381v2, 2020.

[123] MA Lei, JUEFEI-XU F, ZHANG Fuyuan, *et al*. DeepGauge: Multi-granularity testing criteria for deep learning systems[C]. The 33rd ACM/IEEE International Conference on Automated Software Engineering, Montpellier, France, 2018: 120–131.

[124] LING Xiang, JI Shouling, ZOU Jiaxu, *et al*. DEEPSEC: A uniform platform for security analysis of deep learning model[C]. 2019 IEEE Symposium on Security and Privacy, San Francisco, USA, 2019: 673–690.

[125] WANG Zhou, BOVIK A C, SHEIKH H R, *et al*. Image quality assessment: From error visibility to structural similarity[J]. *IEEE Transactions on Image Processing*, 2004, 13(4): 600–612. doi: 10.1109/TIP.2003.819861.

[126] LUO Bo, LIU Yannan, WEI Lingxiao, *et al*. Towards imperceptible and robust adversarial example attacks against neural networks[EB/OL]. https://arxiv.org/abs/1801.04693, 2018.

[127] LIU Aishan, LIU Xianglong, GUO Jun, *et al*. A comprehensive evaluation framework for deep model robustness[EB/OL]. https://arxiv.org/abs/2101.09617v1, 2021.

[128] ZHANG Chongzhi, LIU Aishan, LIU Xianglong, *et al*. Interpreting and improving adversarial robustness of deep

neural networks with neuron sensitivity[J]. *IEEE Transactions on Image Processing*, 2020, 30: 1291–1304. doi: 10.1109/TIP.2020.3042083.

[129] VARGAS D V and KOTYAN S. Robustness assessment for adversarial machine learning: Problems, solutions and a survey of current neural networks and defenses[EB/OL]. https://arxiv.org/abs/1906.06026v2, 2020.

[130] CARLINI N, ATHALYE A, PAPERNOT N, *et al*. On evaluating adversarial robustness[EB/OL]. https://arxiv.org/abs/1902.06705, 2019.

[131] DONG Yinpeng, FU Qi'an, YANG Xiao, *et al*. Benchmarking adversarial robustness on image classification[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 318–328.

[132] HENDRYCKS T and DIETTERICH T. Benchmarking neural network robustness to common corruptions and perturbations[EB/OL]. https://arxiv.org/abs/1903.12261v1, 2019.

[133] RAGHUNATHAN A, XIE S M, YANG F, *et al*. Understanding and mitigating the tradeoff between robustness and accuracy[EB/OL]. https://arxiv.org/abs/2002.10716v2, 2020.

[134] ZHANG Xuyao, LIU Chenglin, and SUEN C Y. Towards robust pattern recognition: A review[J]. *Proceedings of the IEEE*, 2020, 108(6): 894–922. doi: 10.1109/JPROC.2020.2989782.

[135] ABOUTALEBI H, SHAFIEE M J, KARG M, *et al*. Vulnerability under adversarial machine learning: Bias or variance?[EB/OL]. https://arxiv.org/abs/2008.00138v1, 2020.

[136] BAI Tao, LUO Jinqi, and ZHAO Jun. Recent advances in understanding adversarial robustness of deep neural networks[EB/OL]. http://arxiv.org/abs/2011.01539v1, 2020.

[137] TRAMÈR F, BEHRMANN J, CARLINI N, *et al*. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations[EB/OL]. https://arxiv.org/abs/2002.04599, 2020.

[138] KUNHARDT O, DEZA A, and POGGIO T. The effects of image distribution and task on adversarial robustness[EB/OL]. https://arxiv.org/abs/2102.10534, 2021.

[139] SERBAN A C, POLL E, and VISSER J. Adversarial examples - a complete characterisation of the phenomenon[EB/OL]. http://arxiv.org/abs/1810.01185v2, 2019.

[140] CARMON Y, RAGHUNATHAN A, SCHMIDT L, *et al*. Unlabeled data improves adversarial robustness[C]. The 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, 2019: 1–12.

[141] HENDRYCKS D, MAZEIKA M, KADAVATH S, *et al*. Using self-supervised learning can improve model robustness and uncertainty[C]. The 33rd Conference on Neural Information Processing Systems, Vancouver, Canada, 2019.

[142] CHEN Tianlong, LIU Sijia, CHANG Shiyu, *et al*. Adversarial robustness: From self-supervised pre-training to fine-tuning[C]. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 696–705.

[143] JIANG Ziyu, CHEN Tianlong, CHEN Ting, *et al*. Robust pre-training by adversarial contrastive learning[EB/OL]. https://arxiv.org/abs/2010.13337, 2020.

[144] KIM M, TACK J, and HWANG S J. Adversarial self-supervised contrastive learning[C]. The 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1–12.

[145] CHENG Gong, HAN Junwei, and LU Xiaoqiang. Remote sensing image scene classification: Benchmark and state of the art[J]. *Proceedings of the IEEE*, 2017, 105(10): 1865–1883. doi: 10.1109/JPROC.2017.2675998.

[146] 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.

[147] GRILL J B, STRUB F, ALTCHÉ F, *et al*. Bootstrap your own latent a new approach to self-supervised learning[C]. The 34th Conference on Neural Information Processing Systems, Vancouver, Canada, 2020: 1–14.

[148] XU Yanjie, SUN Hao, CHEN Jin, *et al*. Robust remote sensing scene classification by adversarial self-supervised learning[C]. 2021 International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 2021: 1–4.

[149] 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.

[150] XU Yonghao, DU Bo, and ZHANG Liangpei. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses[J]. *IEEE Transactions on Geoscience and Remote Sensing*, 2021, 59(2): 1604–1617. doi: 10.1109/TGRS.2020.2999962.

[151] TU J, LI Huichen, YAN Xinchen, *et al*. Exploring adversarial robustness of multi-sensor perception systems in self driving[EB/OL]. https://arxiv.org/abs/2101.

06784v1, 2021.

[152] MODAS A, SANCHEZ-MATILLA R, FROSSARD P, *et al*. Toward robust sensing for autonomous vehicles: An adversarial perspective[J]. *IEEE Signal Processing Magazine*, 2020, 37(4): 14–23. doi: 10.1109/MSP.2020.2985363.

[153] SUN Hao, XU Yanjie, KUANG Gangyao, *et al*. Adversarial robustness evaluation of deep convolutional neural network based SAR ATR algorithm[C]. 2021 International Geoscience and Remote Sensing Symposium, Brussels, Belgium, 2021: 1–4.

[154] ZHU Xiaoxiang, HU Jingliang, QIU Chunping, *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.

作 者 简 介

孙 浩(1984–)，男，陕西三原人，博士，国防科技大学电子科学学院副教授。研究方向为多源图像协同解译与对抗、因果表示机器学习。

陈 进(1981–)，男，江苏溧阳人，博士，北京市遥感信息研究所副研究员。研究方向为遥感智能解译。

雷 琳(1980–)，女，湖南衡阳人，博士，国防科技大学电子科学学院教授。研究方向为遥感图像处理、图像融合、目标识别等。

计科峰(1974–)，男，陕西长武人，博士，国防科技大学电子科学学院教授，博士生导师。研究方向为SAR图像解译、目标检测与识别、特征提取、SAR和AIS匹配。

匡纲要(1966–)，男，湖南衡东人，博士，国防科技大学电子科学学院CEMEE国家重点实验室教授，博士生导师。研究方向为遥感图像智能解译、SAR图像目标检测与识别。