面向计算机视觉系统的对抗样本攻击综述

王志波[1),2)] 王 雪[1)] 马菁菁[1)] 秦 湛[2)] 任 炬[3)] 任 奎[2)]

[1)](武汉大学国家网络安全学院空天信息安全与可信计算教育部重点实验室 武汉 430072)

[2)](浙江大学网络空间安全学院 杭州 310007)

[3)](清华大学计算机科学与技术系 北京 100084)

摘 要 对抗样本攻击是近年来计算机视觉领域的热点研究方向，通过对图像添加细微的噪声，对抗样本使计算机视觉系统做出错误判断.对抗样本攻击的研究起初重点关注于图像分类任务，随着研究的深入逐步拓展到目标检测、人脸识别等更加复杂的计算机视觉任务中.然而，现有的对抗样本综述缺乏对新兴图像分类攻击方案的梳理总结以及针对目标检测、人脸识别等复杂任务攻击的分析总结.本论文聚焦于计算机视觉系统中的对抗样本攻击，对其理论与前沿技术进行了系统性的综述研究.首先，本论文介绍了对抗样本的关键概念与敌手模型.其次，分类总结和对比分析了对抗样本存在原因的三大类相关假设.再次，根据数字域与物理域两大应用场景，分类概述和对比分析图像分类系统中的对抗样本攻击技术.根据不同的敌手模型，我们进一步地将图像分类任务数字域的攻击方案划分为白盒和黑盒两种场景，并重点总结梳理了新兴的攻击类别.同时，在目标检测、人脸识别、语义分割、图像检索、视觉跟踪五类复杂计算机视觉任务上，根据适用场景分类总结各类任务中的对抗样本攻击方案.进一步地，从攻击场景、攻击目标、攻击效果等方面对于不同攻击方案进行详细地对比分析.最后，基于现有对抗样本攻击方法的总结，我们分析与展望了计算机视觉系统中对抗样本的未来研究方向.

关键词 对抗样本；计算机视觉；图像分类；目标检测；人脸识别；语义分割

中图法分类号 TP391.41；TP18 **DOI 号** 10.11897/SP.J.1016.2023.00436

Survey on Adversarial Example Attack for Computer Vision Systems

WANG Zhi-Bo[1),2)] WANG Xue[1)] MA Jing-Jing[1)] QIN Zhan[2)] REN Ju[3)] REN Kui[2)]

[1)](*School of Cyber Science and Engineering, Key Laboratory of Aerospace Information Security and Trusted Computing Ministry of Education, Wuhan University, Wuhan* 430072)

[2)](*School of Cyber Science and Technology, Zhejiang University, Hangzhou* 310007)

[3)](*Department of Computer Science and Technology, Tsinghua University, Beijing* 100084)

Abstract Deep learning-based computer vision systems play an important role in many security-sensitive applications (e.g., face recognition, person re-identification and automatic driving). However, these security-sensitive systems are facing a kind of serious attack, called adversarial example attack, which could mislead computer vision systems into making wrong outputs by adding subtle noise to the original images. After the phenomenon of adversarial example was proposed, a large number of researchers have engaged in the study of adversarial example attacks against computer vision systems. The research on visual adversarial example attacks mainly focused on image classification tasks at first, and then has been gradually extended to more complex computer vision tasks such as object detection, face recognition, and semantic segmentation.

收稿日期：2022-01-24；在线发布日期：2022-08-13.本课题得到科技创新 2030-“新一代人工智能”重大项目(2020AAA0107705)、国家自然科学基金(62122066，U20A20182，61872274)资助.**王志波**，博士，教授，博士生导师，中国计算机学会(CCF)高级会员，主要研究领域为物联网、人工智能安全、数据安全与隐私保护.E-mail：zhibowang@zju.edu.cn.**王 雪**，硕士研究生，主要研究方向为人工智能安全.**马菁菁**，硕士研究生，主要研究方向为人工智能安全.**秦 湛**，博士，研究员，博士生导师，中国计算机学会(CCF)会员，主要研究领域为人工智能安全、数据安全与隐私保护.**任 炬**，博士，副教授，博士生导师，中国计算机学会(CCF)会员，主要研究领域为物联网与网络计算.**任 奎**，博士，教授，博士生导师，中国计算机学会(CCF)会员，主要研究领域为物联网安全、数据安全与隐私保护、人工智能安全.

There have already been several surveys about adversarial example attacks, but they are still limited, especially in the latest developments of different computer vision systems. That is, existing surveys on adversarial examples mainly focus on image classification, but lack analysis and summary of frontier research on adversarial example attack techniques for other complex computer vision tasks such as object detection, face recognition, semantic segmentation, and image retrieval. In particular, most of the existing reviews only classify attack methods from the perspective of the knowledge owned by the attacker, which cannot reveal the essential characteristics of the different attack approaches. In addition, the previous surveys lack a summary of the newly proposed types of attack methods in image classification, such as semantic transformation, color transformation, and parameter transformation. To address these problems, this paper focuses on adversarial example attacks in computer vision systems and provides a systematic review of their theories and cutting-edge technologies. In this paper, we first introduce the key concepts and categorization perspectives of adversarial examples. Then, three major types of hypotheses related to the reasons for the existence of adversarial examples are categorically summarized and comparatively analyzed. As for different computer vision tasks, we first demonstrate the adversarial example attacks in the image classification task which is the basis of other complex computer vision tasks. In image classification task, we further classify the attacks from different perspectives, i.e. different application scenarios, and different adversary knowledge. According to two major application scenarios, i.e. the digital scenario and the physical scenario, the techniques of adversarial example attacks in image classification systems are categorically outlined and comparatively analyzed. According to different adversary knowledge, we further classify the attack schemes in the digital domain of image classification tasks into two types: white-box and black-box. On the basis of the summary of the image classification task, five other complex tasks in the field of computer vision, i.e., object detection, face recognition, semantic segmentation, image retrieval, and visual object tracking, are further categorized and analyzed. Meanwhile, detailed comparison and analysis of different types of attack approaches are conducted in terms of attack scenarios, attack targets, and attack effects. Finally, with the summary and comparative analysis of various adversarial example attack methods, the future research directions of adversarial example attacks in computer vision systems are concluded and prospected.

Keywords　adversarial example; computer vision; image classification; object detection; face recognition; semantic segmentation

1　引　言

数据的海量累积与硬件性能的快速提升促进了深度学习(Deep Learning,DL)技术的高速发展,其在人工智能领域中的众多困难问题上取得了突破性成果.深度学习技术驱动的智能系统正以前所未有的规模应用于多个领域,如语音识别[1]、自然语言处理[2]、计算机视觉[3]等,而深度学习模型在图像分类[4]、围棋[5]、游戏[6]等任务上的表现甚至超越了人类.在计算机视觉领域,Krizhevsky 等人[3]于 2012 年提出了基于卷积神经网络(Convolutional Neural Network,CNN)的深度学习模型 AlexNet,该模型在大规模视觉识别这一难度极高的任务上取得了突破性成果.AlexNet 成功后,深度学习技术逐渐成为了计算机视觉领域的研究焦点.高效深度学习软件库的应用推广和神经网络结构的不断改进,也促使基于深度学习的智能计算机视觉系统快速成熟并落地应用于一系列安全攸关的场景,如自动驾驶[7]、实时监控[8]、恶意软件检测[9]等.

尽管智能计算机视觉系统提高了生产生活的便利性,它们却存在诸多安全隐患[8].Szegedy 等人[10]的研究表明深度学习技术存在内生脆弱性,在推理阶段对输入添加细微扰动即可生成对抗样本

(Adversarial Example,AE),从而误导模型预测出错并对错误结果表现出高置信度.对抗样本的现象被提出后,大量的研究者开展了针对计算机视觉系统脆弱性的研究.在计算机视觉领域中,攻击者使用算法生成特定的噪声并添加到原始图像中,从而生成具有攻击效果的图像对抗样本.对抗样本的初始定义是基于特定图像的攻击,而 Moosavi-Dezfooli 等人[11]的研究提出了通用对抗扰动(Universal Adversarial Perturbation,UAP)的存在,单个 UAP 可对多个不同的输入产生攻击效果.以上研究均是在数字域攻击,进一步的研究表明对抗攻击可以在物理域对实际应用的模型实现攻击效果[12-15].例如 Athalye 等人[14]的实验证明 3D 打印技术可用于构造物理对抗样本,并成功欺骗深度神经网络分类器;研究[15]在停车标志上生成物理对抗扰动,成功误导分类器将停车标志识别为限速 45 km/h 标志.

国内外已有学者就对抗样本攻击撰写了综述,这些综述可以分为两类:一类综述[16-21]着眼于全面概述对抗样本在图像分类任务中的研究进展;例如文献[16-19]侧重于介绍针对图像分类任务的攻击与防御方案;文献[20]侧重于物理世界中对抗样本的生成流程和技术难点;文献[21]研究了黑盒场景限制下的对抗样本攻击.另一类综述[22-24]从图像、语音、文本等多个角度中的对抗样本研究进展进行简要介绍.

然而,现有的对抗样本综述存在诸多问题:首先,现有综述对攻击方案的分类角度单一,大多数综述仅从攻击者知识的角度对攻击方案进行分类,其分类方式不能凸显各种攻击类型的本质特性;其次,对抗样本研究领域发展迅速,现有综述缺乏对语义变换、色彩变换、参数变换等新型对抗样本生成方法的分析总结;最后,现有综述缺乏对计算机视觉中除图像分类以外(如目标检测、人脸识别等任务)的对抗样本攻击研究的总结.区别于已有综述文章,本文聚焦于计算机视觉领域,重点梳理了多类计算机视觉任务的对抗样本攻击研究.对图像分类任务中的对抗样本攻击研究提出了系统性的分类方式,并着重介绍了新兴的攻击技术.进一步,考虑到在计算机视觉系统中深度学习技术的重要性及对抗样本对该类系统实际应用中的安全威胁,本文分别对目标检测、人脸识别、语义分割、图像检索和视觉目标跟踪 5 类复杂任务中的对抗样本攻击研究进行了归纳总结与分析.图 1 为本文的全文架构图,展示了不同对抗样本攻击技术的分类逻辑.

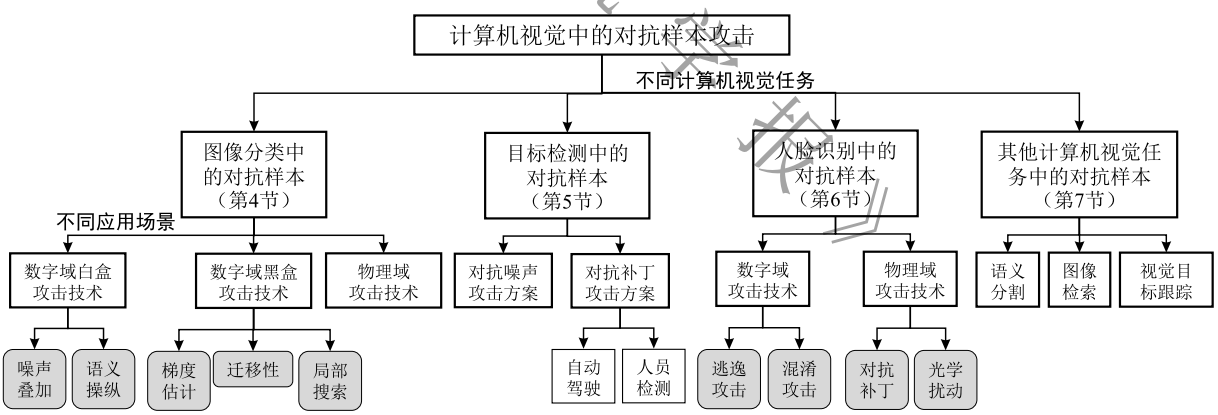

图 1 文章架构图

本文第 2 节介绍计算机视觉中对抗样本的关键概念,包括计算机视觉任务概述、对抗样本的概念和对抗样本攻击的敌手模型;第 3 节总结计算机视觉系统中对抗样本存在原因的相关假设并对其进行分析;第 4 节到第 7 节侧重于对不同计算机视觉任务中具体攻击方案的分析,根据视觉任务的目标将相关研究划分为图像分类、目标检测、人脸识别、语义分割、图像检索、视觉目标跟踪,共 6 大类攻击方案.其中针对图像分类的攻击方案是计算机视觉领域的研究重点,相关攻击技术也广泛应用于其他任务,因此用较长篇幅进行了全面详细的梳理.语义分割、图像检索、视觉目标跟踪这 3 类任务的相关研究还处于起步阶段,缺乏创新型技术,因此将其合并为 1 节进行总结.针对于不同计算机视觉任务,对抗样本攻击所面临的技术难点不同,解决思路各异.每一小节首先介绍该视觉任务的相关概念,之后从攻击场景、攻击目标、攻击效果等方面对比分析典型的攻击方案;第 8 节总结对抗样本领域的研究难点并展望未来可能的研究方向,为今后对抗样本的研究提供参考.

2　计算机视觉中对抗样本的关键概念

2.1　计算机视觉任务

计算机视觉领域的常见研究方向包括图像分类、目标检测、人脸识别、语义分割等. 图像分类任务要求系统根据图像的语义内容，为每张图像分配对应的语义类别标签. 目标检测(Object Detection)任务要求在图像分类的基础上，对图像中的物体同时进行类别判断和所在区域定位. 人脸识别是计算机视觉领域应用广泛的重要任务，它包含人脸验证和人脸辨别两种任务. 人脸验证是判断两张人脸图像是否对应同一人的二分类任务，而人脸辨别是判断未知人脸图像所属类别的多分类任务. 图像语义分割(Sematic Segmentation)任务需要对图像的语义内容进行分割，同时对分割区域甚至每个像素进行分类，是对图像内容更精准的理解. 自 2012 年 ImageNet 数据集[3]发布以来，深度学习技术迅速地在计算机视觉领域得到了广泛应用. 然而，深度学习模型的使用也为计算机视觉系统带来了新的安全挑战，对抗样本攻击是其中影响范围最广的攻击技术之一.

2.2　对抗样本的概念

定义 1.　对抗样本. Szegedy 等人[10]首次提出对抗样本的概念，即对输入样本添加微小的扰动，使模型以高置信度输出错误结果. 在计算机视觉中这类扰动常表现为噪声形式，但最新的研究也提出了如形状变换、颜色改变等非噪声形式的扰动. 在对抗样本的形式化定义中，设 f 为使用原始图像(即没有被添加恶意扰动的图像)x 训练的模型，f 将 x 标记为正确的标签 l；攻击者的目标为生成最小扰动 η，利用扰动构造对抗样本 $x'=x+\eta$，使模型 f 将 x' 标记为错误标签 l'，其数学定义见式(1)：

$$\min_{x'}\|x'-x\|_p$$
$$\text{s.t. } f(x)=l,\ f(x')=l',\ l\neq l',\ x\in[0,1]^m \tag{1}$$

其中 $\|x\|_p$ 表示 p 范数，用于衡量样本之间的距离，$x'-x=\eta$ 被称为对抗扰动，对抗扰动 η 应满足一定的约束条件. 由于深度学习模型的高维性质，η 的求解问题是非线性和非凸的[25]，因此目前对抗样本的生成算法均是实现对式(1)的近似求解.

对抗样本的作用思路如图 2 所示：两张猫的图片虽然在图像空间上相似，但对神经网络而言，第二张猫的激活模式与鹦鹉的激活模式接近，因此第二张猫被分类为鹦鹉. 添加扰动将图像在特征空间的激活模式从猫变为到鹦鹉，同时保持图像空间中的相似性过程，即相当于寻找方程(1)的解生成对抗样本的过程.

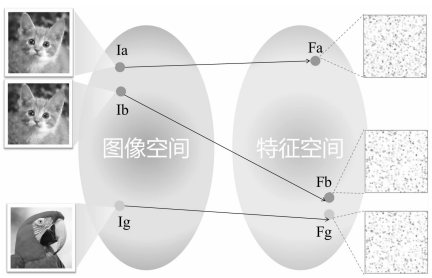

图 2　对抗样本示意图

许多研究发现对抗样本的攻击效果并不局限于攻击时的单个目标网络，其他不同的网络也可能被同一个对抗样本成功攻击，这种特性被称为对抗样本的迁移性[26]. 对抗样本的迁移性为攻击者误导未知结构的系统提供了可行路径，加剧了对抗样本对智能系统的威胁.

对抗扰动是对抗样本的核心，对抗扰动与原始图像共同构成对抗样本. 参考一系列研究[27-29]我们从扰动的影响范围和扰动的度量指标 2 个维度进行描述对抗扰动.

定义 2.　影响范围. 指单一对抗扰动可以影响的样本范围，可分为针对型扰动和通用型扰动两类. 针对型扰动生成特定于输入图像的扰动，在输入改变时需要生成新的对抗扰动. 通用型扰动是与图像无关且独立于输入的扰动，这类扰动具有良好的迁移性，具有应用于所有输入数据的能力. 当输入改变时，通用型扰动不需要重新生成，可以直接将扰动与原始样本相加产生新的对抗样本[11,30].

定义 3.　度量指标. 扰动最小化不是必要条件，因为与人类视觉系统不同，机器学习模型无法区分扰动的大小. 但大多数对抗样本攻击方案都提出了扰动大小的度量标准，最常见的是基于范数、基于 Wasserstein 距离和基于视觉感知 3 类度量指标.

(1) 基于范数的度量指标. 在描述向量的大小或两个向量之间的距离时，研究者经常使用范数函数 $\|x\|_p$，又称为 L_p 范数. 其数学定义如下：

$$\|x\|_p=\left(\sum_{i=1}^{n}|x|^p\right)^{\frac{1}{p}} \tag{2}$$

范数是将向量映射到非负标量的函数. 为测量两个向量之间的距离，可采用它们差的范数 $\|x'-x\|_p$ 以确保得到一个正标量. 在对抗扰动的度量中 p 常用的值有 0、2、∞ 分别对应三种不同的测量矩阵：

① 基于绝对值的 L_0 范数. 度量在 $x'\neq x$ 时，向量中不相等的分量的数量，即与原图像相比对抗图像中被修改的像素点的个数.

② 基于欧几里得的 L_2 范数. 度量 x 和 x' 之间的欧几里得距离. 当多个像素都只发生细微变化时，L_2 范数保持较小；而当一个或多个像素发生较大变化时，L_2 范数将快速增大.

③ 基于最大值的 L_∞ 范数. 度量任何分量的最大变化，等效于限制图像中每个像素的修改上界，而不限制修改像素的数量，其表达式为

$$\|x'-x\|_\infty=\max(|x_1'-x_1|,\cdots,|x_n'-x_n|) \quad (3)$$

在某些情况下，我们希望获得与起始点之间距离满足约束的一组点的集合. 此集合称为标准球，其数学定义为

$$B(x_c,r)=\{x\mid\|x-x_c\|_p\leq r\} \quad (4)$$

其中 x_c 是为球的中心点，即起始点；r 是距离中心的最大距离，又称半径.

(2) 基于 Wasserstein 距离的度量. 它表示将像素值从原始图像改变为对抗样本的成本[31]，其中成本与像素间的距离呈正相关，其数学表示如下：

$$d_w(x,y)=\min_{\boldsymbol{\Pi}\in\mathbb{R}_+^{n\times n}}\langle\boldsymbol{\Pi},\boldsymbol{C}\rangle \text{ s.t. } \Pi_1=x,\ \Pi_1^T=y \quad (5)$$

其中 $x,y\in\mathbb{R}_+^n$ 为非负数据点，且满足 $\Sigma_i x_i=\Sigma_i y_i=1$，因此输入需要归一化处理；$\boldsymbol{C}\in\mathbb{R}_+^{n\times n}$ 为非负的成本矩阵，其中 C_{ij} 表示将 x_i 移动到 y_i 的成本. 扰动的生成需要最小化改变矩阵 $\boldsymbol{\Pi}$，矩阵的初始值 $\Pi_1=x$ 为原始图像，矩阵的结束值 $\Pi_1^T=y$ 为对抗样本，矩阵中的 Π_{ij} 编码表示有多少像素值从 x_i 移动到 y_i.

(3) 基于视觉感知的度量. 心理测量感知对抗相似度评分(Psychometric Perceptual Adversarial Similarity Score，PASS)是研究[32]中提出的一项新指标，它从光照、明度、对比度等角度衡量人类对不同图像感知的一致程度. 仅可察觉失真(Just Noticeable Distortion，JND)是研究[33]中引入的指标，该指标用于表示噪声在 HVS 空间中的可见性.

2.3 对抗样本攻击的敌手模型

对抗样本攻击中不同的攻击条件设置构成不同的敌手模型，本节从敌手目标、敌手知识、攻击频率 3 个维度对其进行刻画. 表 1 总结了智能系统中对抗样本攻击的敌手模型.

表 1 对抗样本攻击敌手模型

敌手目标	置信度降低攻击、无目标攻击、有目标攻击	
敌手知识	白盒	掌握系统的模型架构、模型参数和训练数据集
	灰盒	掌握系统的架构类型、使用的公开数据集
	黑盒	无模型的相关知识
攻击频率	单次	只与系统交互 1 次
	迭代	与系统进行多次交互

定义 4. 敌手目标. 根据攻击者产生的对抗样本对模型影响的不同，将对抗攻击的目标分为置信度降低攻击、无目标攻击、有目标攻击三类. 置信度降低攻击旨在降低对抗样本的分类置信度；无目标攻击将输出类别更改为与真实类别不同的任何错误类别，如在面部识别系统中，攻击者的人脸被错误地识别为其他人脸，从而逃避检测系统[34]；有目标攻击将输出分类强制更改为特定的目标类别. 有目标攻击适用于多分类问题，例如在图像分类系统中，攻击者使所有对抗样本被预测为同一类别；在人脸识别系统中，攻击者将面部伪装成授权用户[35]. 有目标攻击通常会最大程度地提高特点目标类别的概率. 与有目标攻击相比，无目标攻击更易于实现，因为其攻击的结果可以是任意错误输出，攻击的有效输出空间更大.

定义 5. 敌手知识. 指攻击者可获得的训练数据或模型体系结构、参数等信息的体量. 对抗攻击者可获得的知识主要包含以下几类：

(1) 全部或部分训练数据集；

(2) 样本的预处理方式与特征表示；

(3) 模型的相关信息，包括类型、架构、超参数、权重等；

(4) 学习算法的类型及决策函数的形式；

(5) 系统的输出，包括标签、目标区域、置信度分数等；

(6) 系统的其他信息，如防御机制等.

根据攻击者掌握知识的多少可以将对抗样本攻击划分为白盒攻击、黑盒攻击和灰盒攻击三类. 在白盒攻击中，攻击者拥有目标系统的所有信息，能够完全复制受攻击的系统；灰盒攻击场景最早由 Meng 和 Chen[36]提出，攻击者拥有不完整或不确定的信息；在黑盒攻击中，攻击者对被攻击的模型一无所知，但可以将模型作为数据库进行查询.

定义 6. 攻击频率. 根据攻击者与模型交互的频率可分为单次攻击和迭代攻击. 单次攻击只需 1 次优化就可获得对抗样本，对于某些计算量大的任务，例如对于强化学习模型，单次性攻击可能是唯一可行的选择；而迭代攻击需要与模型交互多次，迭代地更新扰动以生成对抗样本. 与单次攻击相比，迭代攻击产生的对抗样本攻击通常成功率更高，同时生成的扰动通常更小，但迭代攻击需要与被攻击模型进行更多交互，花费更多的计算时间与资源.

3　对抗样本存在的原因

自从 Szegedy 等人[10]提出对抗样本的现象，众多研究者从不同角度进行对抗样本的成因假设. 目前的主流假设可以分为三类：对抗样本分布异常假设、模型结构缺陷假设和训练数据缺陷假设.

3.1　对抗样本分布异常假设

对抗样本分布异常假设认为对抗样本是与正常样本分布不同的异常数据，在此基础上不同的研究对于对抗样本与正常样本的分布关系有不同的观点. 高维非线性假设认为对抗样本存在于数据流形的小概率空间中，非独立同分布假设认为对抗样本位于数据流形上，而流形高维几何特征假设认为对抗样本的分布完全在数据流形之外.

3.1.1　高维非线性假设

Szegedy 等人[10]首次提出对抗样本的概念，认为由于深度神经网络的高度非线性，对抗样本存在于数据流形的低概率空间中，即对抗样本位于数据流形的“凹陷”中，在输入样本的周围空间进行随机采样无法获得对抗样本. 因此对抗样本与原始数据没有关联且分布不同，并且模型学习到的特征也不适用于对抗样本. 基于这一假设，Gu 等人[37]进一步地从目标函数、训练流程和训练样本的大小和多样性等角度，研究了“凹陷”的大小和出现原因. 高维非线性假设提出时间早，很多后续研究基于这一假设开展，但神经网络的非线性不仅缺少理论证明，也缺少在大规模数据集上的实验证明.

3.1.2　非独立同分布假设

训练数据和测试数据满足独立同分布性，是深度学习模型有效性的前提假设. 非独立同分布假设认为对抗样本是从不同分布中采样得到的数据，它位于数据流形之上. 研究者基于该假设提出了众多对抗样本检测方法[38-40]，并尝试通过生成模型学习这种新分布. 但 Carlini 和 Wagner[41]提出了可轻易绕过检测器的攻击方案，以实验结果对非独立同分布假设提出了质疑.

3.1.3　流形高维几何特征假设

一些研究[42-45]认为是数据高维流形的几何性质造成了对抗样本，驳斥了对抗性样本存在于数据流形上的假设. 在实验验证中，Gilmer 等人[42]创建了一个变量可控的数据集，用于精准控制数据的流形. 分析基于该数据集的模型，作者观察到没有被正确分类的正常样本与错误分类的对抗样本在距离上很接近，这意味着神经网络对微小对抗扰动的脆弱性是测试误差的必然结果. 基于实验结果，Gilmer 等人还否认了非独立同分布假设[44]，他们提出与正常数据相比，对抗样本位于不同的分布上.

对抗样本分布异常假设从数据流形的角度研究对抗样本的存在原因，与深度学习中数据降维的操作紧密关联. 但目前这类假设缺乏大规模实验证明和相关的理论证明，且基于这三种假设的防御方案仍无法有效地防御对抗样本攻击，因此这类假设的合理性还有待进一步研究.

3.2　神经网络缺陷假设

神经网络结构缺陷假设认为神经网络的某些结构性缺陷是导致模型对对抗样本脆弱的主要原因. 线性假设是对高维非线性假设的驳斥，其认为神经网络内在的线性行为导致扰动的影响在模型中被逐层放大；而边界倾斜假设则认为模型决策边界向数据流形倾斜，因此微小扰动即可让样本跨越决策边界.

3.2.1　线性假设

Goodfellow 等人[46]提出尽管隐藏层对输入进行了非线性转换，但深度神经网络实际上仍然保留有线性的行为. 因此，对高维输入所有维的微小扰动的总和可能会导致分类错误. 特别的，选择易于优化的激活函数，例如 ReLU、Sigmoid，会使深度神经网络的行为更加线性化. Luo 等人[47]提出了这种猜想的一种变体，其认为 DNN 在输入流形的某些区域中表现出线性行为，而在其他区域表现出非线性行为. Fawzi 等人[48]称分类器的鲁棒性与所使用的训练程序无关，并且在高阶分类器中两个类别之间的距离比线性分类器大，这表明在更深层的模型中很难找到对抗样本. Tabacof 和 Valle[49]的实验表明与较深的模型相比，较浅的分类器对对抗样本的敏感性更高. 一些攻击[50]和防御[51]方法都基于线性假设，并取得了良好的效果.

3.2.2　边界倾斜假设

Tanay 和 Griffin[52]指出线性行为不足以解释对抗样本的存在，并以实验证明了可以训练出对对抗样本不敏感的线性模型. 进一步地，作者提出了边界倾斜假设，即模型学习到的类别边界位于数据流形附近，但边界相对于数据流形倾斜，所以可以通过对原始样本向分类边界添加噪声，使其越过分类边界成为对抗样本. 如果边界仅略微倾斜，则扰动穿过决策边界所需的距离非常小，从而导致在视觉上几乎无法分辨对抗样本与原始样本. 由此作者推测对

抗样本可能是模型过拟合的后果，可以通过正则化来缓解这种情况.

线性假设和边界倾斜假设都拥有实验证明并启发了许多相关研究，是被广泛研究和认可的假设.但可以构造受对抗样本影响小于神经网络模型的线性模型，说明了线性行为并不是对抗样本存在的充分条件，这使线性假设的充分性受到质疑.而边界倾斜假设提供了从决策边界角度研究对抗样本的新思路，但其仍缺少理论证明.

3.3 训练数据缺陷假设

训练数据缺陷假设认为模型对对抗样本脆弱的主要原因在于训练数据，训练数据的缺乏或者训练数据中存在脆弱性的特征都会导致使用这些数据训练的模型出现脆弱性.

3.3.1 训练数据缺乏假设

根据 PAC 学习模型，Schmidt 等人[53]提出训练鲁棒模型需要的样本复杂度明显高于非鲁棒模型.特别地，为了获得 L_∞ 范数鲁棒的模型，需要在输入维度上，多项式倍地提高样本的复杂度.类似地，Bubeck 等人[54]提出，统计查询模型的鲁棒性学习需要成倍地增加查询数量.Cullina 等人[55]表明当存在以凸约束集为边界的攻击者时，鲁棒模型需要的样本复杂度不会增加.该研究表明可以在某些约束条件下实现模型的鲁棒性，但是此类约束条件的效果并没有在大规模数据集上进行验证.Tsipras 等人[56]表明提高鲁棒性会导致精度下降，并且这种权衡取舍与机器学习模型的复杂度无关.对抗训练是基于该理论的一种有效防御方法，将大量的对抗样本加入到模型的训练集能有效地提高模型的鲁棒性.

3.3.2 非鲁棒特征假设

Izmailov 等人[57]通过在分类过程中从输入中删除低频特征来研究图像特征对模型的影响.结果表明删除低频特征几乎不会提高模型鲁棒性，但是删除互信息少的特征可显著提升模型鲁棒性.Ilyas 等人[58]提供了一种全新的视角，即对抗扰动的存在并不一定意味着模型或训练程序存在缺陷，而实际上是训练图像的某些特征表现.以人类的感知程度作为标准，作者将这些特征划分为鲁棒特征和非鲁棒特征.鲁棒特征来源于数据分布，具有高预测性和鲁棒性，在受到对抗攻击时模型依旧能根据鲁棒特征进行正确预测；而非鲁棒特征也源自数据的分布模式且具有高预测性，但人类难以理解且更脆弱.基于该假设，作者使用训练后神经网络的 logits 层从原始输入图像中过滤得到鲁棒特征，并且用只包含鲁棒特征的图像构造了鲁棒数据集.实验结果表明，使用鲁棒数据集训练的模型远比使用正常数据集训练的模型更难被成功攻击.因此，对抗样本存在的原因，可能是训练数据集存在非鲁棒特征.

由于深度学习模型是由数据驱动的模型，因此数据质量对于模型的性能有显著的影响.从数据特征的角度解释对抗样本问题是新颖的思路，这一类的研究重新强调了数据对模型的影响作用，可以与从模型角度解释对抗样本的工作互为补充.

尽管目前大量研究尝试解释对抗样本的存在原因，但目前学界对这一问题仍没有达成共识性的结论.该问题受诸多影响因素，许多研究假设也只是从某一个角度以实验结果进行论证，缺乏严格和系统的理论证明.未来需要更多类似于文献[55,62,64]的基础理论研究，形式化和系统地解释神经网络受对抗样本影响的原因.

4 图像分类中的对抗样本攻击

本节介绍针对图像分类任务的对抗样本攻击方案，根据攻击应用场景，划分为数字域攻击和物理域攻击两类.在数字域攻击中，攻击者可以直接获得数字图片，并对其进行精确到像素的修改，生成的数字对抗样本可直接作为模型的输入.而在物理域攻击中，首先物理实体被传感器（例如相机、LiDAR）捕获，之后转化为数字域的图像再输入分类模型.攻击者无法控制物理系统的数据通道，因此只能通过扰动物理实体或传感器进行攻击.物理域的扰动面临着各种环境因素导致的形变，并且传感器处理可能会进一步削弱扰动的效果，所以物理域攻击的难度比数字域更大.

根据敌手的知识，数字域的攻击技术可划分为白盒攻击和黑盒攻击，部分黑盒攻击也会使用白盒攻击的方法，数字域整体攻击流程如图 3 所示.图像分类中物理域的攻击多数基于白盒设定且需要构造数字域的对抗样本，物理域的常用攻击流程如图 4 所示.

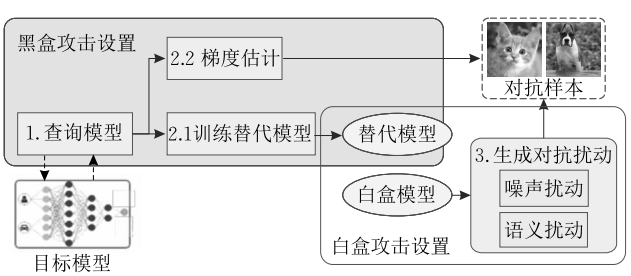

图 3 图像分类数字域对抗样本攻击流程

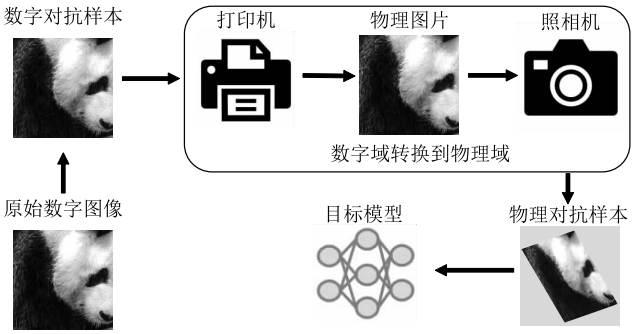

图 4　图像分类物理域对抗样本攻击流程

4.1　图像分类中数字域的白盒攻击技术

在白盒攻击中，攻击者掌握被攻击模型的所有信息，包括模型架构、模型参数以及用于训练或测试的数据. 利用这些信息，攻击者可以完全复制模型或学习数据分布以生成新的样本. 根据攻击策略，我们将攻击方法进一步细分为两类：第一类方法基于噪声叠加的思想，将不可见噪声直接添加到图像中以生成对抗样本；第二类方法基于图像语义操纵的思想，攻击者在保持图像语义不变的同时，对原始图像进行变换以生成对抗样本. 白盒攻击是最强大的对抗样本攻击场景，因此通常用于评估防御方案的性能或系统在极端情况下的鲁棒性. 表 2 总结了两类攻击策略及其对应的对抗扰动生成算法.

表 2　图像分类中数字域的白盒攻击类别

攻击策略	生成算法	核心思路
噪声叠加	优化算法	对替代公式或约束条件进行优化求解
	敏感性分析	利用贡献度量算法发现敏感特征
	生成模型	利用生成模型学习对抗扰动的概率分布
语义操纵	几何变换	进行缩放、旋转、掩盖等几何变换
	色彩变换	改变特定区域的亮度、色调等色彩特征
	参数变换	扰动图像形成过程的参数

4.1.1　基于噪声叠加的攻击

根据噪声生成算法思路的不同，基于噪声叠加的攻击方案分为 3 类：基于优化算法的攻击、基于敏感性分析的攻击和基于生成模型的攻击. 在基于优化算法的攻击中，攻击者利用替代公式或约束条件将式(1)转换为新的目标函数，之后使用优化算法来搜索目标函数的解；基于敏感性分析的攻击使用贡献度衡量算法，分别计算输入特征对输出的贡献度，从而发现敏感特征并进行针对性的扰动；基于生成模型的攻击利用生成模型学习对抗扰动的概率分布，并使用生成模型直接构建新的对抗扰动.

(1) 基于优化算法的攻击

本节总结的攻击方法使用优化方法搜索方程(1)的解，或使用近似方程替代方程(1). Szegedy 等人[10]率先发现了深度神经网络对细微扰动的敏感性，提出了对抗样本的概念与式(1)的求解方法 L-BFGS (Limited-memory BFGS). 该研究开创了对抗样本领域的研究先河，但 L-BFGS 优化过程复杂，算法效率低. Carlini 和 Wagner[41] 提出了方程(1)的另一种形式，分类函数由于其非线性而被替换，攻击优化目标为

$$\min_{\eta} \|x, x+\eta\|_p + c \cdot f(x, x+\eta) \tag{6}$$

该方法使用 Adam 算法求解优化问题并将扰动向边界投影，以此搜索扰动较小的对抗样本. DeepFool 算法[50]估计输入样本 x 到分类器最近决策边界的距离. 该距离既可以作为模型对攻击鲁棒性的度量指标，也可以用于确定最小对抗扰动方向，有效地降低了扰动的大小. Moosavi-Dezfooli 等人[11] 提出通用对抗扰动(Universal Adversarial Perturbation，UAP)攻击，通过累积对单个输入的扰动，找到可以误导训练集所有图片的普适性对抗扰动. UAP 攻击不需要针对测试图像进行额外的优化，直接在测试图片上添加普适性扰动即可生成对抗样本. UAP 攻击级大地提升了测试阶段的攻击效率，但在扰动生成阶段需要多次遍历训练数据集消耗大量算力. Su 等人[59] 提出 OnePixel 算法，利用差异进化算法来确保扰动的多样性. 尽管基于进化算法的攻击需要的模型信息较少，但其攻击效果较差. 大多数研究都认为对抗样本与从训练集中获得的输入非常相似，但 Nguyen 等人[60] 制作了人类观察者无法理解的但分类模型却以高置信度识别的特殊图像.

基于优化的攻击通常可以精确地找到最小扰动或非常好的近似值，大多数攻击方案都有较好的攻击效果. 但与其他类型攻击相比，基于优化的攻击一般计算过程更加复杂，因为它们需要更多轮次的迭代或使用复杂的优化方法.

(2) 基于敏感性分析的攻击

基于神经网络的结构，输入在梯度方向上的微小改变，将使损失函数朝着最大化的方向迈进一大步，梯度也揭示了模型的部分决策过程. 分析输入对损失函数的梯度影响通常被称为敏感性或显着性分析. 为了克服 L-BFGS 攻击速度慢的缺陷，受到深度神经网络线性假设的启发，Goodfellow 等人[46] 提出了一种沿损失函数梯度方向求解扰动的攻击，称为快速梯度符号(Fast Gradient Sign Method，FGSM)攻击，其扰动定义为

$$\eta = \varepsilon \cdot \mathrm{sgn}(\nabla_x l(\theta, x, y)) \tag{7}$$

其中 ε 控制对抗扰动 η 的大小，sgn 表示梯度求导的

正负号.这种攻击只需要计算梯度向量,因此缩短了生成对抗样本的耗时但牺牲了一定的攻击精度.

为了提高攻击精度,Kurakin 等人[61]提出迭代地使用 FGSM 算法,并且通过梯度裁剪来限制修改像素的范围,该攻击称为基本迭代攻击(Basic Iterative Method,BIM).BIM 攻击对扰动进行迭代更新,第 n 轮迭代的更新公式为

$$x'_{n+1}=\mathrm{Clip}\{x'_n+\alpha\cdot\mathrm{sgn}(\nabla_x J(\theta,x'_n,y))\} \quad (8)$$

其中 n 是当前的迭代次数,$0<\alpha<\varepsilon$ 是每次的迭代步长,sgn 表示梯度求导的正负号,J 为扰动度量函数,θ 为扰动阈值,Clip{·}表示将每轮的输入限制在输入样本 x 的 ε 邻域中.通过最大化所选类别的可能性,基本迭代攻击可以实现有目标攻击.Madry 等人[62]扩展了 BIM 攻击,通过迭代地应用投影梯度下降(Project Gradient Descent,PGD)来搜索输入周围的 p 范数球内的扰动.PGD 方案显著提升了攻击的有效性,是目前最强的攻击方案之一.PGD 攻击产生的扰动可以覆盖整个 p 范数球,因此对 PGD 攻击的成功防御,即意味着对 p 范数球内所有对抗样本的完全防御.Dong 等人[63]提出使用梯度动量来提升 FGSM 的迭代攻击(Momentum Iterative Fast Gradient Sign Method,MI-FGSM).与梯度下降的作用类似,动量可以通过在梯度方向上累积速度矢量,从而稳定扰动的更新方向,并有助于摆脱较差的局部极值.将速度矢量设置为 0 时,该攻击等效于普通的 FGSM 攻击,实验结果表明动量可以显著地提升攻击的有效性.

另外一些研究利用雅克比矩阵求解对抗扰动,Papernot 等人[64]基于显著性分析提出了基于雅克比矩阵的显著图攻击(Jacobian-based Saliency Map Attack,JSMA).为了发现每个像素在决策过程中的重要性,计算模型的前向导数即雅可比矩阵以生成显著图.JSMA 攻击高像素图像的计算成本过高,因此该方法的使用场景受限.类似地,Khrulkov 和 Oseledets[65]使用雅可比矩阵计算特征图的奇异值向量,提出 SV-UAP 方法构造普适性对抗扰动;而 Huang 等人[66]使用雅可比矩阵计算深度神经网络输出的线性近似值,利用正常情况和扰动情况下近似值的差异构造最小扰动.

一些防御方法可能会使攻击者无法获得模型梯度,针对这一情况,Athalye 等人[67]提出了梯度掩蔽的现象,并提出反向传播可微近似(Backwards Pass Differentiable Approximation,BPDA)攻击方案,用可微近似代替不可微层的梯度.为了避免梯度掩蔽现象,Tramèr 等人[68]提出了随机单步攻击(Randomized Single-Step Attack,RSSA),对输入数据点进行随机化后再搜索对抗样本.以上两类方案都有效地绕开了基于梯度混淆的防御措施.

Chen 等人[69]探索了 L_1 范数在构造对抗样本中的作用,提出了弹性网络正则化攻击(Elastic-net Attack to DNN,EAD).EAD 方法混合了用于高维特征选择的惩罚函数,同时使用 L_1 和 L_2 范数进行正则化.因为 EAD 使用了多种最小化技术以获得敏感特征,所以该算法在优化和敏感性分析方法之间建立了联系.

由于不需要求解非凸优化问题,基于敏感性分析的攻击通常比基于优化策略的攻击耗时更少,可以更快地产生大量对抗样本用于训练模型.在研究中,使用 PGD 生成输入 p 范数球距离内的大量扰动,并利用其进行对抗训练是目前最先进的防御方法之一.但由于缺乏严格的约束函数,基于敏感性分析的攻击生成的对抗样本通常隐蔽性较差.

(3)基于生成模型的攻击

以上两类攻击直接对原始输入添加对抗扰动,但是攻击者有时无法获得测试阶段的原始输入样本,因此有研究提出基于生成模型的对抗样本攻击.生成模型是一类机器学习算法,通过学习训练样本来估计样本整体的概率分布,并生成与训练样本相似且同分布的样本.常用的生成模型有两类:变分自动编码器(Variational Auto-Encoder,VAE)[70-71]和生成对抗网络(Generative Adversarial Networks,GAN)[72].

Baluja 和 Fischer[73]训练了名为对抗性转换网络(Adversarial Transformation Networks,ATN)的深度神经网络,该网络可以将输入转换为对抗样本.ATN 网络训练完成后,对抗样本的生成只需要一次前向传播,因此该攻击执行速度快并且可以产生多样的对抗样本.Poursaeed 等人[74]训练了一个生成模型,生成依赖和独立于图像的扰动,从而产生特定或普适性的对抗样本.但该方法在生成普适性对抗样本时,生成器的损失函数是目标模型损失函数的线性组合,这使对抗样本的效果严重依赖于受攻击的模型.Song 等人[75]优化了条件 GAN 的隐空间,为性别分类器生成不受限制的对抗样本.

生成模型可以学习样本的近似分布,从中采样新的数据,在生成对抗样本时,生成模型假设所有扰动都来自相同的分布,在此基础上学习扰动的分布.尽管对抗样本来自同一分布的假设存在局限性,但基于生成模型的攻击方案仍能够生成强大的

扰动.但生成模型的训练通常需要大量资源,因此在对抗样本的实际应用中很少使用基于生成模型的攻击方案.

4.1.2　基于语义操纵的攻击

除了直接向图像添加噪声外,对抗扰动也可以是非加性的即语义不变的扰动,如旋转、遮盖、对比度调整、颜色改变、语义内容转换等.语义对抗样本攻击在图像上添加不破坏原始语义信息的扰动,使扰动后的图像可误导分类模型.大多数基于语义的对抗扰动无法使用与噪声扰动相同的思路实现,并且这类扰动的有效计算需要更复杂的优化方法.本节根据语义对抗扰动的生成算法,进一步将基于语义的攻击方案分为基于几何变换、基于颜色变换和基于参数变换 3 类攻击.图 5 展示了常见的基于语义的对抗样本效果示意图.

图 5　基于语义的对抗样本生成方法示意图

(1)基于几何变换的攻击

这类攻击中,攻击者对图片进行缩放、旋转、掩盖等几何变换,使变换后的图像能成功误导模型.Engstrom 等人[76]率先指出仅使用简单的旋转和平移变换就足以构造对抗样本,并提出了在变换中使用随机采样、网格搜索、梯度上升等多种优化方法.在模型的实际应用中,这些变换都很容易实现且贴合实际情况.Kanbak 等人[77]提出称为 ManiFool 的方法,搜索可以欺骗神经网络的人眼不可查觉的最小几何变换.Pei 等人[78]提出验证计算机视觉算法鲁棒性的框架,衡量模型抵抗自然扰动的能力.

Xiao 等人[79]提出对抗空间变换攻击(Spatially Transformed Adversarial,STA),在保留图像原始外观的同时改变其几何形状.与在像素空间上进行变换约束不同,该研究在局部几何变换中添加了正则化损失.STA 方案生成的对抗样本保持了图像的高度感知性,且大多数防御都无法抵抗这种攻击.类似地,Zhang 等人[80]在使用Carlini和Wagner攻击之前改变场景的几何形状,提出盲点攻击(Blind Spot Attack,BSA).BSA 会产生与训练数据相距甚远的输入,而这些输入位于防御系统无法涵盖的"盲点".一旦受到扰动,这些输入就会导致巨大的偏差,从而误导经过证明的防御措施.Athalye 等人[14]提出了变换期望算法(Expectation Over Transformation,EOT),在 2D 情况下,EOT 算法使用仿射变换的随机分布;在 3D 情况下,算法增加了对纹理和形状的约束.EOT 算法没有单独使用特定的变换,而是在变换的分布中搜索扰动,因此生成的对抗样本鲁棒性更强,可绕过基于输入变换的防御方案.

基于对抗训练的思想,一些防御方法使用数据增强技术在训练时加入几何变换的样本,以期望提高模型对几何变换的鲁棒性.但实验表明,增强训练的模型对几何变换仍然敏感,这一现象表明模型并没有学会抽象的几何变换,而只是单纯地拟合训练数据.基于几何变换的对抗样本有很强的攻击性,同时几何变换被认为在数据采集过程中频繁地发生;但大多数几何攻击方法产生的对抗样本不具有迁移性,对不同的模型需要重新计算扰动.

(2)基于颜色变换的攻击

颜色变换会改变图像特定区域的亮度、色调、对比度等,但仍使图像在局部或全局上保持纹理的均匀性.Hosseini 等人[81]发现翻转图像的亮度后模型的识别率会下降,并首次提出了"语义对抗样本"的概念.Afifi 等人[82]发现颜色恒常性(Color Constancy)或白平衡的错误计算(WB Error),会产生强烈的色彩偏差进而误导图像分类模型.进一步地,Hosseini 等人[83]基于人类认知系统的形状偏差特性,提出语义对抗转换攻击(SemanticAdv).在 HSV(色相、饱和度和值)颜色空间上,保持值分量不变的同时随机移动色相和饱和度分量.SemanticAdv 攻击在不影响对象形状的同时改变图像颜色,并生成在视觉上仍然保持平滑自然的对抗样本.类似地,Shamsabadi 等人[84]提出了一种基于内容的对抗攻击 ColorFool,利用图像语义确定视觉上自然的修改范围,在颜色信息与亮度感知均匀的 Lab 颜色空间[85]中产生颜色扰动.该攻击有很高的成功率,并且可以成功误导受检测器、对抗训练等方法防御的模型.Laidlaw 等人[86]提出了功能性对抗攻击(Functional Adversarial Attacks,FAA)的概念,并提出了 ReColorAdv 攻击方案.ReColorAdv 使用灵活的参数化函数 f 将输入图像中每个像素的颜色 c 映射到对抗样本中的新颜色 $f(c)$.FAA 将功能对抗攻击与现有的基于 p 范数的攻击相结合,在提升攻击有效性的同时减少了颜色修改的范围.Bhattad 等人[87]也提出一种基于梯度信息优化的颜色变换对抗攻击方案 cAdv,采用文献[88]中预先训练好的深度着色模型,并将攻击目标优化集成到灰度图像自动着色的

过程中. cAdv 攻击利用着色方案符合自然颜色的边界的特点,保证了对抗样本颜色的区域一致性,产生的对抗噪声更平滑且具有很强的隐蔽性和鲁棒性. 除了直接修改目标主体的颜色,Tian 等人[89]提出晕影这一使照片四周产生暗角的后期效果也可用于生成对抗样本,并提出了对抗性晕影攻击(Adversarial Vignetting Attack,AVA). 因为 AVA 攻击产生的对抗扰动与正常晕影效果几乎完全一致,所以该攻击的隐蔽性更强且更难被人眼分辨.

基于颜色变换的攻击在色彩空间而非像素空间对图像进行修改,因此不会破坏图像的语义结构和纹理信息,不易被察觉具有较高的隐蔽性;而在物理域中,颜色变换攻击也较容易实现. 但基于颜色变换的攻击无法使用传统的感知度量指标评估,因此缺乏对扰动大小的约束和限制.

(3)基于参数扰动的攻击

参数攻击是一类新兴的攻击技术,其攻击空间由参数空间而非像素空间定义,通过扰动图像形成过程的参数,产生视觉上更自然的对抗样本. Zhao 等人[90]提出自然对抗扰动生成(Natural Adversarial Generation,NAG)方案. NAG 利用 GAN 在密集连续数据的语义空间中搜索对抗样本,解决了输入空间扰动与语义空间特征不匹配的问题. 然而由于缺乏可解释性和特征解耦,语义空间中的扰动很难精准控制. 随着特征解耦技术的发展,Joshi 等人[91]提出针对受限二元分类器的参数转换语义攻击(Parametric Transformations Semantic Adversarial,PTSA),逆向使用预训练的多属性生成模型,通过修改正常输入的属性来生成对抗样本. 因为 PTSA 攻击在归因空间中进行,所以其可以提供精细的扰动控制,但由于属性数量的限制,该方案缺乏灵活性. Qiu 等人[92]也提出了一种基于属性条件的图像编辑(Attribute-Conditioned Image Editing,ACIE)的对抗样本生成算法. Liu 等人[93]的工作提出参数范数球的概念,通过优化 3D 空间中的几何表面来创建对抗样本,利用可微物理渲染器在底层图像的参数空间中生成扰动(Adversarial Attacks in Parametric Space,AAPS). 物理渲染器根据物理参数分析计算像素颜色的导数,从而 AAPS 实现了将传统的像素扰动扩展到物理上有效的参数扰动. 区别于基于 GAN 的攻击,Wang 等人[94]利用 VAE 学习类别无关的特征并生成不可见的隐空间扰动(Invisible Latent Perturbation,ILP).

由于扰动发生在参数空间,基于参数扰动的攻击生成的对抗样本在视觉上更加自然且隐蔽性更高. 但进行参数攻击需要训练对应的参数模型,消耗大量的计算资源且缺乏对扰动的有效约束. 表 3 对图像分类数字域的白盒攻击方案进行了总结.

表 3 图像分类数字域的白盒攻击

攻击策略	攻击名称	攻击目标	攻击思路	攻击优势	攻击局限性
基于优化算法	L-BFGS[10]	有	直接近似求解优化式	精确地找到最小扰动或非常好的近似值,攻击效果好	需要多轮迭代或复杂的优化方法,计算过程复杂
	C&W[41]	无&有	用线性函数替代优化目标		
	DeepFool[50]	无	估计输入与最近决策边界的距离		
	UAP[11]	无	累积多个输入的对抗噪声		
	OnePixel[65]	无&有	使用差异进化算法近似求解		
基于敏感性分析	FGSM[46]	无	沿损失函数梯度方向求和	大部分方案不需要反复优化,攻击耗时更少,可以短时间产生大量对抗样本用于对抗训练	缺乏严格的约束函数,生成的对抗样本通常精确性更差
	BIM[61]	无	迭代使用 FGSM 算法并进行梯度裁剪		
	PGD[62]	无	迭代应用投影梯度下降搜索近似扰动		
	MI-FGSM[63]	无	用梯度动量提升迭代 FGSM 的效果		
	JSMA[64]	无&有	利用前向导数计算显著图		
	SV-UAP[65]	有	用前向导数计算特征图的奇异值向量		
	Ref.[66]	有	用雅可比矩阵计算输出的线性近似		
	BPDA[67]	有	用可微近似代替不可微层的梯度		
	RSSA[68]	无	对输入随机化后计算对抗扰动		
	EAD[69]	有	混合高维特征选择的惩罚函数		
基于生成式模型	ATN[73]	无&有	使用残差网络和自动编码器生成噪声	可以产生样本分布外的对抗样本	训练生成模型需要大量资源
	Univ. GM[74]	无&有	生成器的目标模型损失函数的线性组合		
	AC-GAN[75]	有	训练生成对抗模型模拟数据的条件分布		
基于几何变换	Ref.[76]	无	使用随机采样、网格搜索等变换	对抗样本生成过程简单,无需复杂计算	迁移性差,针对不同模型需重新计算
	ManiFool[77]	无&有	在几何变换流形的约束下添加迭代变换		
	Ref.[78]	有	使用旋转、反射、对比度或侵蚀等变换		
	STA[79]	有	在局部几何畸变上引入新的正则化损失		
	BSA[80]	无&有	在 C&W 攻击前改变场景的几何形状		
	EOT[14]	有	在多种变换的分布中搜索扰动		

（续　表）

攻击策略	攻击名称	攻击目标	攻击思路	攻击优势	攻击局限性
基于颜色变换	SemanticAE[81]	无	翻转图像亮度以生成对抗样本	生成的对抗样本迁移性高且不易察觉，有较高隐蔽性	无法使用传统的感知度量指标评估，缺乏扰动限度约束
	WB Error[82]	无	利用白平衡的错误产生色彩偏差		
	SemanticAdv[83]	无	在 HSV 颜色空间改变色相和饱和度		
	ColorFool[84]	无&有	用语义划定掩码并在 Lab 颜色空间扰动		
	ReColorAdv[86]	无&有	将输入中像素颜色 c 映射对抗颜色 $f(c)$		
	cAdv[87]	无&有	在灰度图自动着色中实现攻击目标优化		
	AVA[89]	无	同时优化晕影的离轴照度、几何形状和倾斜系数		
基于参数扰动	NAG[90]	有	用 GAN 在数据的语义空间中搜索扰动	生成的对抗样本在视觉上更加自然，隐蔽性更高	扰动缺乏限度约束，计算资源消耗大
	PTSA[91]	有	逆向使用多属性生成模型将		
	ACIE[92]	无&有	利用特征空间插值的属性条件编辑图像		
	AAPS[93]	无&有	用可微物理渲染器在参数空间生成扰动		
	ILP[94]	无&有	用 VAE 学习特征并扰动隐空间编码		

4.2　图像分类中的数字域黑盒攻击技术

与白盒攻击相比，攻击者在进行黑盒攻击时只能获得极为有限的信息. 在不同的威胁模型中，攻击者可获得不同的资源，如有标签的数据集、部分训练样本等. 在训练阶段，攻击者无法获得训练数据，也无法获得模型的权重和参数. 在模型部署后，一般情况下攻击者可以获得样本的预测标签和置信度得分，而更严格的限制下仅预测标签是已知的；攻击者在生成对抗样本时对模型进行查询也被认为是资源. 一般而言，在成功率相近的情况下，消耗的资源量越少的攻击方案越好.

黑盒攻击根据其攻击手段的不同可分为 3 类：基于梯度估计、基于迁移性和基于局部搜索的攻击. 在基于梯度估计的攻击中，攻击者可使用比原始架构更加复杂的模型结构进行梯度和权重估计. 在基于迁移性的攻击中，攻击者选择训练一个参数已知的白盒模型作为替代模型，用于模拟原始参数未知的模型. 基于局部搜索的攻击不使用替代模型，攻击者利用被攻击模型的查询反馈机制，不断地输入带有扰动的样本并在本地搜索最优扰动.

4.2.1　基于梯度估计的攻击

利用模型对输入样本产生的梯度生成噪声是白盒攻击的常用方法，而在黑盒攻击中可以根据模型的输出对其梯度进行近似估计，从而利用近似梯度生成对抗样本.

零阶优化是对梯度近似的常用方案，Chen 等人[95]率先提出零阶优化方法（Zeroth Order Optimization，ZOO）来估计目标模型的梯度以生成对抗样本. ZOO 方法使用对称差的商估计梯度，并提出了分别使用 Adam[96] 和牛顿法的 ZOO-Adam 和 ZOO-Newton 随机坐标更新方法. ZOO 方案可实现与白盒方案相近的攻击效果，但估计梯度需要大量计算并进行多轮迭代，因此 ZOO 攻击的效率较低且攻击需要模型输出的置信度. 在只能获得样本类别的硬标签场景下，为了实现黑盒攻击，Cheng 等人[97]将对抗扰动搜索问题视为实值连续优化问题，提出 Opt-Attack 方案来提高攻击的效率. Opt-Attack 是一种迭代的零阶优化方法，并使用随机梯度自由（Randomized Gradient Free，RGF）方法解决扰动的优化问题. Tu 等人[98]进一步地优化了 ZOO 方案并提出“AutoZOOM”，即基于自动编码器的零阶优化框架，显著地提高了攻击的效率.

Ilyas 等人[99]提出了有限信息与查询（Limited Queries & Information，LQI）方案，同时使用自然进化策略（Natural Evolution Strategies，NES）与投影梯度下降解决对模型查询的限制. 该方案全面考虑了黑盒场景中对攻击者多种约束，适用于多种攻击场景. 同样基于自然进化策略，Ilyas 等人[100]通过 bandit 优化利用有关梯度的先验信息，引入了与时间和数据有关的先验. Bhagoji 等人[101]中对函数 $f(x)$ 进行两侧梯度估计，提出了基于有限差分的减少查询次数的技术（Query Reduction using Finite Difference，QRFD），有限差分法不需要梯度信息也能优化损失函数，加速了扰动的求解过程. 在攻击者无法获得任何正常样本时，Du 等人[102]提出了一种全新的基于区域的攻击算法（Input-Free Attack，IFA），由于获得的梯度受限于区域空间，IFA 方法首先初始化目标模型的输入，将其像素值归一化到 $[0,1]$ 之间，之后攻击者通过增减像素值调整明暗度从而添加扰动实现攻击.

基于梯度预测的方案直接与被攻击模型进行交互，该类方法计算量较小且有较高的攻击成功率. 但在查询次数不足时该方案的攻击成功率较低，一些模型对查询次数进行了限制可能导致攻击失败，同

时攻击者对模型大量的查询更容易引起怀疑.

4.2.2 基于迁移性的攻击

当网络的架构和模型权重未知时,攻击者通常会从头开始训练另一个模型用于模拟受攻击的模型,该模型称为替代模型.对抗样本使替代模型出现错误分类时,对抗样本的迁移性使原始模型也产生错误分类.替代模型通常拥有复杂的架构,以获得较高复杂度的潜在空间.由于攻击者掌握了替代模型的完全控制权,因此这类方法通常用于实现有目标类别的对抗攻击.攻击者在这类方法中面临的主要挑战有两点:一是在没有任何原始模型架构与参数的先验知识的情况下训练替代模型;二是对抗样本迁移性的提升.

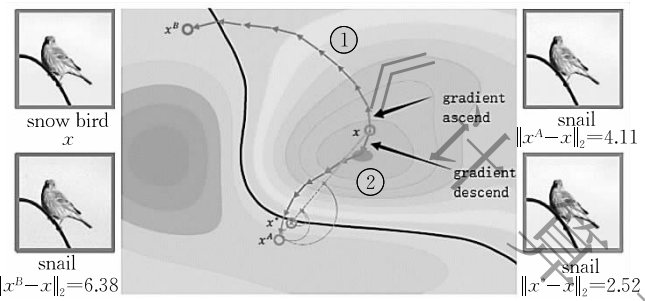

图 6 Curls&Whey 方法示意图[103]

Papernot 等人[26]首次提出通过查询原始模型获得合成数据的标签从而创建与目标模型相似的本地替代模型方案(Local Substitute Model,LSB).LSB 方案利用基于雅可比矩阵进行数据增强,提高了查询次数有限时近似模型决策边界的效率.一般的攻击方法沿梯度上升方向将噪声单调地添加到样本上(图 6 中向上的曲线①),但 Curls&Whey 方法[103]关注于寻找到达决策边界上较近点的路径.首先进行 Curl 迭代,沿梯度下降的方向到达"谷底",然后沿梯度上升方向添加噪声(图 6 中向下的曲线②);然后使用 Whey 技术优化,从之前产生的对抗样本中消除冗余的扰动.Curls&Whey 方案有效地提高了对抗样本在黑盒场景中的多样性和迁移性.

平移不变攻击方法(Translation-Invariant Attack,TIA)[104]使用图像集合而非优化目标函数的方式生成对抗样本,利用卷积神经网络具有平移不变性的假设近似目标模型的梯度.该方法极大地提高了对抗样本的迁移性,并且可以与任何基于梯度的攻击方法组合使用.Wang 等人[105]提出特征重要性感知攻击(Feature Importance Aware,FIA),在对抗样本搜索时,通过汇聚各种分类模型的聚合梯度探索输入样本的重要特征.FIA 生成的对抗噪声针对性地破坏了主导模型决策的关键感知特征,因此生成的对抗样本具有高迁移性.Huan 等人[106]提出一种黑盒场景下的通用对抗样本攻击方法(Data-Free Adversarial Perturbations,DFAP),DFAP 方法将模型的输出视为对输入样本的内部表示,通过迭代地最大化原始样本与对抗样本在模型表示空间中的特征差异生成对抗样本.与其他通用对抗攻击方案相比,DFAP 方案并不需要使用原始训练样本且方案具有更强的迁移性.

基于迁移性的攻击仅使用很少甚至不需要模型的相关信息,并且该类攻击可以与其他攻击方法结合使用.但生成替代模型需要更多的计算,而对抗样本的迁移性在不同模型间存在较大的差异使得该类攻击的攻击效果不稳定.

4.2.3 基于局部搜索的攻击

在不使用替代模型的情况下进行对抗样本攻击,需要解决确定模型边界和完全独立生成对抗扰动两大挑战.局部搜索技术常用于解决组合问题,是一种不完整搜索算法.局部搜索技术在潜在空间中搜索正确的噪声方向,以获得分类错误的输入.基于局部搜索的攻击不是在完成替代模型训练之后再利用其梯度,而是利用梯度的同时训练模型.攻击者从原始输入开始,利用模型查询的反馈,在可接受的误差范围内向输入添加噪声.随着输入置信度分数的下降,攻击者会将噪声推向梯度方向.

在可以使用置信度分数时,Narodytska 等人[107]提出进阶本地搜索算法(Advanced Local Search,ALS),在图像空间上最小化对抗样本与原始图像共享分类标签的可能性.进一步地,Narodytska 等人[108]提出了一种基于贪婪局部搜索技术的攻击方法 Gradient-free,对图像中与分类相关性最高的像素添加噪声实现对损失函数的梯度近似.以上两种方法都极大地减少了对模型的查询次数,而后者加入的局部贪婪算法进一步地提升了攻击效果.Li 等人[109]认为 Ilyas 等人[99]的算法效果较差,是因为它依赖于较精确的梯度值,而当深度神经网络不平滑时,不可能准确估计梯度.因此他们提出 N-Attack 算法,模拟以输入为中心的 L_p 球的概率密度,并从中生成对抗样本.该方法无需对梯度的准确估计,有效地提高了攻击的容错性.

与基于置信度得分的攻击相比,利用模型的最终决策(即类别标签)的攻击更符合现实场景中攻击者的能力.Brendel 等人[110]首次提出基于最终

决策的黑盒攻击方案边界攻击(Boundary Attack)算法，首先对图像进行随机初始化，之后算法遍历输入图像所属类的决策边界.边界攻击可适用于不同机器学习模型和数据集，且可以攻破防御性蒸馏(Defensive Distillation)等隐藏模型原始梯度的防御方法. Brunner 等人[111]提出有偏采样攻击算法(Biased Sampling)，使用 Perlin 噪声[112]分布创建初始噪声，利用置换向量 v 对噪声进行参数化，以在边界攻击之前引入低频噪声.该方案有效地缩小了扰动的搜索空间，提高了攻击的效率. Chen 等人[113]进一步减少了边界攻击的查询次数，提出迭代的 HopSkipJump 攻击算法. HopSkipJump 攻击算法的每次迭代包含梯度方向估计、通过几何级数进行步长搜索以及通过二分搜索进行边界搜索三个步骤.多种搜索方法的组合有效地提高了搜索效率和攻击成功率. Guo 等人[114]研究提出了一种简单而有效的方法 SimBA，使用正交向量与概率分布迭代地添加噪声以在黑盒环境中生成对抗样本.总体而言，基于模型最终决策的攻击对梯度掩蔽、随机化或鲁棒训练等防御技术具有更强的抵抗性.

一些研究利用遗传算法进行局部搜索 Chen 等人[115]基于遗传算法提出了 POBA-GA 攻击方案，通过调整不同的噪声点的像素阈值、数量和大小三个参数以产生不同的扰动. Alzantot 等人[116]基于种群的无梯度优化策略提出 GenAttack 攻击. GenAttack 避免了梯度估计方法中产生的过多查询开销，同时可以绕过基于梯度掩蔽和混淆的防御.

基于局部搜索的攻击同时利用了模型查询结果和替代模型的思想，大幅降低了攻击所需查询模型的次数；同时由于不需要构建完整的替代模型，其计算量也远低于基于迁移性的攻击.但基于局部搜素的攻击存在扰动计算不稳定的缺点，同时其也需要与模型交互，存在攻击被检测的风险.表 4 对图像分类数字域的黑盒攻击技术进行了总结.

表 4　图像分类数字域的黑盒攻击

攻击策略	攻击名称	攻击目标	攻击方案思路	方法优势	方法局限性
基于梯度预测	ZOO[95]	无&有	使用对称差的商估计梯度的零阶优化	计算量较小，在查询次数足够的情况下攻击成功率较高	需要与模型直接交互，查询次数不足时攻击成功率低；同时攻击者容易被限制和识别
	OPT[97]	无&有	迭代 ZOO 从高斯分布中随机采样增量值		
	AutoZOOM[98]	无&有	基于自动编码器的零阶优化		
	LQI[99]	无&有	结合自然进化策略与投影梯度下降		
	Bandits&Priors[100]	无	引入并利用时间和数据有关的先验		
	QRFD[101]	无&有	使用有限差分法从 $f(x)$ 两侧估计梯度		
	IFA[102]	无&有	从初始化灰度图像生成对抗样本		
基于迁移性	LSB[26]	无&有	利用基于雅可比矩阵进行数据增强	需要信息少，可以与其他攻击方案结合使用	生成替代模型需要更多计算资源，攻击成功率在不同模型间差距较大
	Curls&Whey[103]	无	先 Curl 迭代梯度上升再 Whey 进行优化		
	TIA[104]	无&有	在二维空间中迭代地移动有限数量的像素		
	FIA[105]	无	聚合多模型梯度计算输入的特征重要性		
	DFAP[106]	无	最大化模型表示空间中的特征差异		
基于局部搜索	ALS[107]	无	减少扰动前后图像不共享分类标签的可能性	查询次数低，无需构建复杂替代模型，运算量低	扰动计算不稳定；攻击者容易被限制和识别
	Gradient-free[108]	无	用贪婪局部搜索噪声并近似损失函数的梯度		
	N-Attack[109]	无	使用自然进化算法实现无导数梯度估计		
	Boundary Attack[110]	无&有	随机初始化后遍历输入所属类的决策边界		
	Biased Sampling[111]	无	用 Perlin 噪声创建初始噪声的边界攻击		
	HopSkipJump[113]	无&有	在决策边界使用二元信息估计梯度方向		
	SimBA[114]	无	使用正交向量与概率分布迭代地添加噪声		
	POBA-GA[115]	无&有	用遗传算法调整噪声的阈值、数量和大小		
	GenAttack[116]	无	基于种群的无梯度优化策略		

4.3　图像分类中的物理域攻击技术

许多研究将在物理域的对抗样本搜索问题简化为目标函数的优化问题，其数学表示如下：

$$\arg\min_{\eta} L(f(x+1),l)+P(\eta) \tag{9}$$

其中 $L(\cdot,\cdot)$ 是衡量模型预测值与目标标签 l 差异的损失函数，它不仅包括数字域图像的训练损失，还包括传感器在物理环境中实际捕获样本数据的训练损失.惩罚函数不仅是扰动 η 的范数，通常还包括不可打印评分(Non-Printable Score，NPS)和平滑限制函数等，用于解决对抗扰动的制造误差和平滑度约束.

在分类任务中，分类器主要提取图像的特征并对特征向量进行分类.当前对物理世界中分类器的攻击包括基于实体的攻击和基于摄像头的攻击. Kurakin 等人[117]证明了对抗样本可应用于现实世界.攻击者对 ImageNet 数据集中的原始图像使用 FGSM、BIM 和 JSMA 算法生成对抗图像.之后将对抗图像被打印在纸上，然后由手机相机进行照片

拍摄、裁剪等物理转换，然后输入到分类器中. 实验结果表明，经过“照片变换”后大多数对抗图像仍使分类器识别错误. 换言之，在数字域中产生的对抗样本也可能误导物理域中的开放系统. Athalye 等人[14]改进了 Kurakin 的工作，他们产生的扰动对于高斯噪声、失真和仿射变换之类的变换具有鲁棒性.

道路标志分类系统在自动驾驶中起着重要的作用，对道路标志的对抗攻击可能会导致财产损失甚至用户伤亡. Eykholt 等人[15]提出了鲁棒物理扰动(Robust Physical Perturbations，RP2)方案，对路标分类系统生成对抗样本. 该研究提出目标约束海报和黑白贴纸攻击两种方法，实现了对分类器的有效攻击. 黑白贴纸是对生活中普通涂鸦的模仿，因此不会引起怀疑. RP2 方案在优化目标函数中添加了 NPS 指标，解决了物理扰动的制造误差问题. 同时该方案提出掩码计算，限制了对扰动的空间位置并考虑了不同视角的变换. 之后的 Liu 等人[118]也提出了对道路标志识别系统的攻击方案. 与 Eykholt 不同，他们使用注意力模型查找图像分类的敏感区域，并且利用生成器生成具有强烈视觉保真度的对抗补丁.

与直接修改目标物体不同，Worzyk 等人[119]提出使用投影的方式生成物理域的对抗样本，通过投影仪或激光发射器，可以在仅操纵一个颜色通道的限制下，实现对分类器的误导. 同样基于投影的方式，Gnanasambandam 等人[120]利用结构化照明来改变目标物体的外观的，提出光照对抗攻击(Optical Adversarial，OPAD). OPAD 直接将投影仪-摄像机模型纳入对抗攻击的优化函数中，在降低扰动的同时提升了攻击的有效性.

除了直接修改目标物体外，另一类物理攻击方案采用修改相机的方式使拍摄的照片存在扰动，从而误导模型的分类结果. Li 等人[121]提出了一种在物理域中的通用对抗扰动攻击，称为对抗相机贴纸攻击. 他们将精心设计的圆点型半透明贴纸贴在相机镜头上，在相机拍摄目标图像时，这些圆点类似相机产生的自然模糊点. 图 7 展示了这种攻击的流程. 这种方法是在相机和物体之间的光路上添加扰动，而该环节几乎不受保护，因而它拥有较高的隐蔽性. 尽管这是一种新的攻击思路，但该类攻击的实现面临很多挑战. 首先，这类攻击要求攻击者可以任意修改受害者的物理拍摄设备，这可能引起怀疑. 其次，这种攻击有很强的局限性，它只会导致目标相机拍摄的图片分类错误，而其他相机拍摄目标物体的照片不会被分类错误.

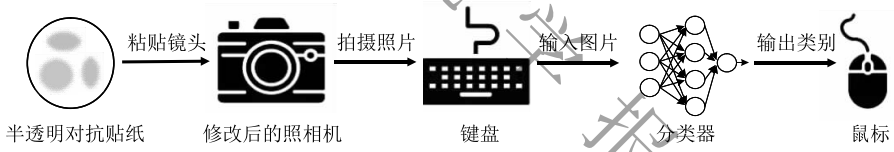

图 7 根据文献[121]绘制的对抗相机贴纸攻击流程

5 目标检测中的对抗样本攻击

目标检测任务的目标是对图像中特定物体进行区域定位和类别判定，其中区域定位使用边界框(Bounding Box)表示. 典型的深度目标检测算法有两类：two-stage 目标检测算法，代表模型包括 Faster-RCNN[122]、FPN[123]等；one-stage 目标检测算法，代表模型包括 SSD[124]、YOLO[125]系列模型等. 两类算法的主要区别在于输出的生成过程不同，two-stage 算法通常首先通过区域预选网络(Region Proposal Network，RPN)为每个物体生成多个预选框(Proposal)，之后使用非最大抑制(Non-maximum Suppression，NMS)[126]或其他技术选择最佳边界框，并标注边界框中物体的类别，该类模型也被称为基于区域提议的模型(Region Proposal Based Model). 而 one-stage 算法在提取输入特征后，同时进行区域定位和物体分类，该类模型也被称为基于回归的模型(Regression Based Model).

但近年来，在目标检测任务中也涌现了大量对抗样本攻击的研究. 通常针对目标检测系统的隐藏攻击比针对图像分类系统的攻击难度更大，因为检测器通常会产生数百或数千个与目标对象重叠的先验区域，假设对抗样本欺骗了最初的先验区域，NMS 可能会选择不同的先验区域. 而如果要从图像中完全删除特定的对象，攻击必须同时欺骗与目标对象重叠的所有先验区域，其难度远高于欺骗单个分类器的输出. 在图像分类中，基于对抗噪声的研究占主流，而由于对目标检测的攻击难度比图像分类更大，更多的攻击方案采用噪声强度和攻击强度都更高的对抗补丁扰动. 本节将目标检测中的攻击方案分为基于对抗噪声和基于对抗补丁两大类，对目标检测系统中现有的对抗样本攻击方案进行梳理和总结，内容的总体框架如图 8 所示.

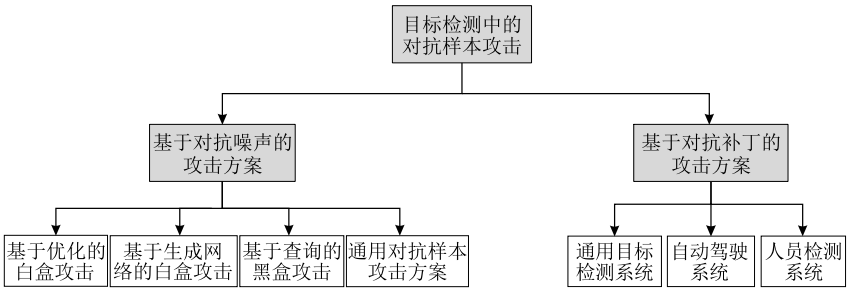

图 8　目标检测中的对抗样本攻击章节框架图

根据攻击效果的差异，目标检测中的对抗样本攻击可分为隐藏攻击(Hiding Attack)和出现攻击(Appearing Attack)两类. 在隐藏攻击中，对抗样本使目标检测器无法检测到图像中的目标物体；而出现攻击中，对抗样本使物体被识别为攻击者指定的类别或使系统在没有物体的区域产生边界框.

5.1　基于对抗噪声的攻击技术

基于对抗噪声的攻击方案与图像分类任务中同类攻击方案思路一致，即对抗扰动为不可见的噪声，并可能修改图像的任意像素.

针对特定图像的白盒攻击是难度较低的一类攻击，这类攻击方案可以使用优化方法求解对抗噪声或者训练 GAN 模型生成对抗样本. Lu 等人[127]指出，一些图像分类系统中基于优化的攻击方案，在简单地修改优化目标后即可实现对目标检测系统的攻击. Zhang 和 Wang[128]研究了目标检测模型分类损失和定位损失对攻击效果的影响，并实现了对 two-stage 检测模型的 PGD 攻击和对抗训练. 同样基于优化的思路，针对首先生成预选框的 two-stage 检测器，Xie 等人[129]提出了名为 DAG(Dense Adversary Generation)的攻击方法. DAG 将目标检测任务视为多目标分类任务，通过迭代地攻击目标区域并使系统对所有区域进行错误分类. 但 DAG 存在资源消耗大的问题，生成一个对抗样本平均需要计算 3000 个边界框，并且 DAG 无法攻击 one-stage 检测器. Xiao 等人[130]为提高对抗样本的隐蔽性提出 AO^{2}AM(Adaptive Object-oriented Adversarial Method)攻击方案. 该方案的优化算法自适应地选择对抗噪声的位置与规模，AO^2 AM 生成的对抗样本与对应的原始输入具有高度的结构相似性.

除了对常见目标检测模型的攻击外，一些研究者提出对特殊目标检测模型或模型特定结构的攻击方案. Zhang 等人[131]提出 CAP(Contextual Adversarial Perturbation)攻击方案，在生成扰动的优化目标中添加上下文损失，以破坏目标对象的上下文信息. 该方法不依赖任何真实的标签，可实现对弱监督目标检测模型的有效攻击. 而针对使用 NMS 技术的目标检测模型，Wang 等人[132]提出 Daedalus 攻击方案，实现了禁用 NMS 功能. Daedalus 可以控制对抗样本的强度和攻击的目标类别. Daedalus 攻击比实现错误分类的攻击更加难防御，因为 Daedalus 生成对抗样本的特征图复杂度远高于其他攻击. 针对自动驾驶系统，Chen 等人[133]基于图像分类系统的攻击方案，提出 ShapeShifter 方法. ShapeShifter 扩展了 EOT 方案，在停止标志内除单词“停止”(STOP)外的所有像素上添加噪声. ShapeShifter 方案首次实现了对目标检测系统的多角度多距离物理域攻击.

由于基于优化的攻击方案存在攻击运算量大、攻击实时性差等缺点，许多研究基于 GAN 网络提出了实时性更强的攻击方案. Wei 等人[134]率先提出联合高效攻击(United and Efficient Adversary，UEA)，构造复合损失函数以训练用于生成对抗样本的 GAN，UEA 综合使用了多尺度的注意力特征损失、高层分类损失和低层特征损失. 其生成单个对抗样本的速度比 DAG 快 1000 倍，并且生成的对抗样本有更高的迁移性. Li 等人[135]也提出了基于 GAN 的快速攻击方案(Fast Attack，FA)，将类别损失和位置损失整合到 GAN 的训练损失函数中. 该方案在保障生成速度的同时保障了对抗样本在视觉效果上的真实性. 而 Deng 等人[136]设计了一种基于 GAN 和风格迁移的攻击方案，通过风格迁移算法使对抗样本与原始图像在视觉上更加一致. 该方法比同样基于 GAN 的方法攻击效果更好且迁移性和视觉隐蔽性更强. 基于 GAN 的攻击方案攻击速度较快，但训练 GAN 的训练需要耗费大量的运算和时间，并且需要大量的训练数据作为支撑.

在白盒攻击中攻击者可以利用大量的模型信息，因而攻击难度较低，近来更多的研究关注于攻击

难度更大的黑盒攻击. Wang 等人[137]提出名为蒸发攻击(Evaporate Attack)的黑盒攻击方案,该方案生成的对抗样本不会被目标检测模型识别为任何物体. 蒸发攻击提出的 GA-PSO 算法通过查询输入样本的选框位置和标签信息实现黑盒攻击. 针对 one-stage 目标检测模型,Liao 等人[138]提出 FLA (Fast Locally Attack)攻击方案,FLA 利用高级语义信息为所有检测到的类别计算梯度并生成局部扰动. FLA 生成的对抗样本实现了黑盒攻击的高成功率和抗 JPG 压缩的高鲁棒性. 进一步地 Liao 等人[139]也提出了两种基于类别的攻击算法 DCA(Dense Category-wise Attack)和 SCA(Sparse Category-wise Attack),DCA 和 SCA 都利用全局的高级语义信息产生对抗扰动. DCA 根据全局有效类别计算梯度并生成密集扰动;SCA 则通过计算特殊的决策边界将密集扰动近似为稀疏扰动. 该方案生成的对抗样本可以在白盒和黑盒场景中抵抗 JPG 压缩并且有很强的迁移性. 由于目标检测模型的攻击需要改变多个对象的结果,目前主流的黑盒攻击方案都依赖于对目标模型的多次查询.

除针对特定样本的对抗攻击外,一些研究也提出针对目标检测模型的通用对抗样本攻击方案. R-AP(Robust Adversarial Perturbation)方案[140]首次提出一种对 two-stage 检测器的通用对抗扰动生成方法. 该方案针对性地攻击区域预选网络,同时提出一种新的损失函数,在干扰模型标签预测的同时降低形状预测. Wu 等人[141]也提出一种通用对抗样本攻击方案 G-UAP,误导区域预选网络对前景的预测. Li 等人[142]提出通用的密集目标抑制攻击方案(Universal Dense Object Suppression,U-DOS),生成针对特定类别的通用扰动. U-DOS 攻击可删除对目标类别的检测,并使其他类别的检测结果保持不变. 目前通用对抗样本攻击方案主要针对于含有区域预选网络的 two-stage 模型,对于不含区域预选网络的模型仍缺少有效的攻击方案. 表 5 对目标检测系统中基于对抗噪声的攻击技术进行了总结.

表 5 基于对抗噪声的攻击

方案名称	攻击场景	攻击效果	目标模型	方法优势	方法局限性
Ref.[127]	白盒	隐藏攻击	YOLO	基于噪声的攻击方案成功率更高,其中基于优化的方案攻击成功率最高;而基于 GAN 的方案攻击效率高,攻击耗时少	基于噪声的方案扰动区域更大,其中基于优化的方案攻击效率低,需要多轮迭代优化;而基于 GAN 的方案训练难度大且对抗样本隐蔽性较低
Ref.[128]	白盒	隐藏攻击	RFB FSSD SSD YOLO-v3		
DAG[129]	白盒	隐藏攻击	Faster-RCNN		
AO2AM[130]	白盒	隐藏攻击	Faster-RCNN YOLO-v3 SSD		
CAP[131]	白盒&黑盒	隐藏攻击	Faster-RCNN PCL OICR		
Daedalus[132]	白盒&黑盒	出现攻击	Mask R-CNN YOLO-v3 SSD RetinaNet		
ShapeShifter[133]	白盒	隐藏攻击	Faster R-CNN		
UEA[134]	白盒&黑盒	隐藏攻击	Faster-RCNN YOLO-v3 YOLO-v4		
FA[135]	白盒	隐藏攻击	Faster-RCNN SSD		
Ref.[136]	白盒	隐藏攻击	Faster-RCNN SSD		
蒸发攻击[137]	白盒&黑盒	隐藏攻击	Faster-RCNN YOLO-v3		
FLA[138]	白盒&黑盒	隐藏攻击	Faster-RCNN SSD		
DCA&SCA[139]	白盒&黑盒	隐藏攻击	Faster-RCNN YOLO-v2 SSD		
R-AP[140]	白盒	隐藏攻击	Faster-RCNN R-FCN aFCIS Mask-RCNN		
G-UAP[141]	白盒&黑盒	隐藏攻击	Faster-RCNN R-FCN SSD		
U-DOS[142]	白盒&黑盒	隐藏攻击	Faster-RCNN CenterNet CornerNet SSD		

5.2 基于对抗补丁的攻击技术

基于对抗补丁的攻击只在较小的区域内添加扰动,这类扰动通常在视觉上是可见的. 对抗补丁的抗变换能力更强,因此对物理域系统的对抗攻击多采用对抗补丁形式.

针对数字域的通用目标检测系统,Liu 等人[143]提出对抗补丁生成方案 DPATCH,该方案可以在黑盒场景下实现无目标和有目标攻击. 其生成的对抗补丁攻击效果与位置无关,并且在不同架构的模型和训练数据集之间有很强的迁移性. 针对包含非局部块(Non-local Block)的目标检测器,Huang 等人[144]提出了基于对抗补丁的隐藏攻击和出现攻击,该方案在数字域和物理域中均可实现高效的攻击. 而 Saha 等人[145]利用上下文推理来欺骗标准检测器,这种攻击设计了一种与场景中任何目标对象都不重叠的对抗补丁. 同时,针对对抗补丁攻击,作者还提出了针对性的防御方案. Lee 等人[146]设计了针对物理域的对抗补丁攻击,该补丁可以放置在图像中的任何位置,实现逃逸攻击抑制模型对图像中所有对象的检测. 该方案具有很强的灵活性和迁移性. 此外,Braunegg

等人[147]提出 APRICOT 数据集，该数据集的每张图片都包含对抗补丁，使场景中原本存在的物体无法被正确识别. APRICOT 是第一个在现实世界场景中使用物理对抗性补丁的照片数据集，其内容涵盖了室内和室外场景中不同时间、位置、比例、旋转角度和视角的物理对抗补丁. APRICOT 数据集为目标检测中对抗补丁攻击的研究提供了良好的数据基础.

针对自动驾驶系统，Lu 等人[148]设计了针对停车标志的对抗样本生成算法，因为停车标志是规则的八边形，算法使用形状匹配功能将根坐标系中的对抗性标志映射到训练框架中的停车标志中. 算法生成的对抗性停车标志在物理域成功误导了检测模型，但对抗性标志在视觉上与正常的停车标志有明显差异；并且由于该攻击方案使用了形状匹配函数，该方法无法攻击没有固定形状的对象. Song 等人[149]进一步限制了添加对抗性扰动的区域，提出对抗贴纸攻击，该贴纸可以让检测器将无意义区域识别为目标物体. 针对对抗噪声直接修改路标易被发现的问题，Huang 等人[150]设计了一种基于对抗性边框攻击的对抗补丁攻击. 该方案将目标对象放置在对抗边框的中心使其被对抗扰动包围，无需更改目标对象内的任何像素. 为进一步提高攻击的隐蔽性，Huang 等人[151]提出对抗性广告牌(Adversarial Signboard)攻击，针对分类器最小化原始类别的概率. 对抗性广告牌在攻击时可以放置在与目标对象有一定距离的地点，因此具有更高的隐蔽性. 针对现实自动驾驶检测系统，Zhao 等人[152]提出了生成鲁棒对抗样本的系统性解决方案. 对于隐藏攻击，他们提出特征干扰增强(Feature-Interference Reinforcement，FIR)算法，在扰动的目标函数中加入模型中间层输出特征损失. 而现实约束生成增强算法(Enhanced Realistic Constraints Generation，ERG)实现了攻击的多维度增强，同时他们利用公开的图像合成更多样化的攻击背景以提高对抗样本的鲁棒性. 对于出现攻击，他们提出了聚合对抗样本算法，同时在长距离和短距离上聚合多个对抗样本实现对目标检测器的多角度攻击. 图 9 总体展示了针对自动驾驶系统生成的对抗性路牌.

图 9　文献[133，148-150，152]中的对抗性路标

人员检测系统是另一类特殊的目标检测系统，该类系统是只检测人员的目标检测系统. Xu 等人[153]提出对抗性 T 恤攻击，实现了对人员检测系统的隐藏攻击. 人在进行移动和姿势变换时 T 恤会发生非刚性变形，该研究利用薄板样条插值(Thin Plate Spline，TPS)首次在对抗样本中对非刚性材料(例如 T 恤衫)的形变效果进行建模，实现了非刚性物体的物理域对抗样本攻击. Wu 等人[154]也提出使用印刷海报和可穿戴衣物进行的物理域攻击. Huang 等人[155]提出通用物理迷彩方案(Universal Physical Camouflage，UPC)生成通用的伪装图案隐藏某类物体或者诱导目标检测系统识别物体为指定类别. UPC 设计了复杂的损失函数同时误导 RPN、分类网络和回归网络，并利用 GAN 保证扰动的语义性，实现了数字域和物理域的对抗攻击. 同时文章提出第一个标准化的现实模拟数据集 AttackScenes，实现了完全可控的真实世界 3D 建模. 图 10 总结展示了相关文献中针对人员检测系统生成的物理对抗样本. 表 6 对比分析了目标检测系统中基于对抗补丁的攻击方案.

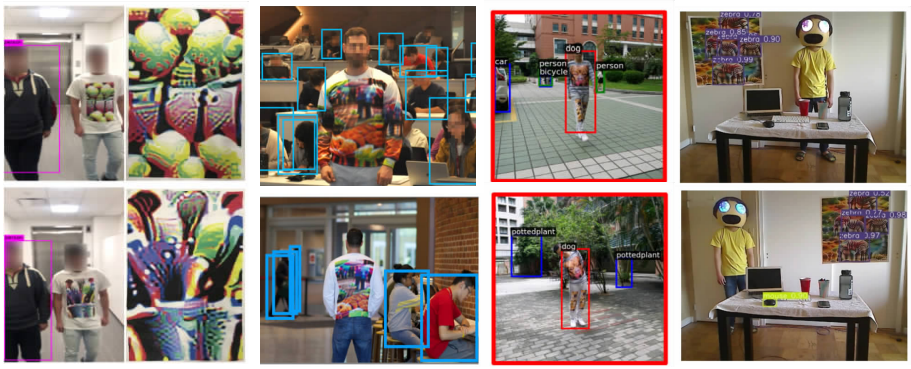

图 10　文献[153-156]针对人员检测系统的物理对抗样本

表 6 基于对抗补丁的攻击

目标系统	方案名称	攻击场景	攻击效果	攻击思路	目标模型
通用目标检测	DPATCH[143]	白盒&黑盒	隐藏攻击	同时优化边界框位置和类别目标	Faster-RCNN YOLO
	Ref.[144]	白盒	隐藏攻击&出现攻击	针对包含非局部块(Non-local Block)的目标检测器	Faster-RCNN
	Ref.[145]	白盒	隐藏攻击	利用上下文推理来欺骗标准检测器	YOLO-v2
	Ref.[146]	白盒&黑盒	隐藏攻击	结合 PGD 攻击和 EOT 方法生成对抗补丁	YOLOv3
	APRICOT[147]	白盒&黑盒	出现攻击	结合 ShapeShifter 和 EOT 方法生成多角度多场景的对抗补丁数据集	RetinaNet SSD Faster-RCNN
自动驾驶	Ref.[148]	白盒	隐藏攻击	用所有视频帧的聚合梯度优化目标函数	Faster-RCNN YOLO
	Ref.[149]	白盒	隐藏攻击&出现攻击	修改 RP2 的损失函数并加入平滑约束	YOLO-v2 Faster-RCNN
	Ref.[150]	白盒&黑盒	隐藏攻击	在目标周围生成扰动使回归层输出异常边界框	YOLO-v3 Faster-RCNN
	Ref.[151]	白盒	隐藏攻击	在原目标正下方的规则区域生成扰动,最小化边界框分类网络原始的最大值	Faster R-CNN YOLO-v3 Mask R-CNN RFCN
	Ref.[152]	白盒&黑盒	隐藏攻击&出现攻击	加入中间层输出特征损失,利用公开数据增强数据集,聚合多距离的对抗样本	Mask R-CNN SSD RFCN Faster-RCNN YOLOv3
人员检测	Ref.[153]	白盒	隐藏攻击	利用薄板样条插值对运动物体的变形进行建模	YOLO-v2 Faster-RCNN
	Ref.[154]	白盒&黑盒	隐藏攻击	利用渲染函数对扰动平移、缩放与随机增强	YOLO-v2 Faster-RCNN
	UPC[155]	白盒&黑盒	隐藏攻击	设计损失函数同时误导 RPN,分类网络和回归网络并利用 GAN 保证扰动的语义性	Faster-RCNN R-FCN SSD Yolo-v2 Yolo-v3 RetinaNet

5.3 小结

在基于对抗噪声的方案中,基于优化的攻击有较好的效果,但存在攻击效率低的问题;而基于 GAN 的方案实现了攻击的高效性,但存在训练难度大、对抗样本隐蔽性较低的问题. 目前基于对抗噪声的攻击大多数只适用于特定类别的模型,缺乏对 one-stage 和 two-stage 模型通用的攻击方案. 基于对抗噪声的攻击方案比基于对抗补丁的攻击方案在数字域有更好的攻击效果,但对抗补丁攻击的扰动区域更小,攻击方案更容易迁移到物理域.

现有的攻击方案仍存在较多的局限性,如对自动驾驶系统的攻击只能在汽车低速运行时成功;对人员检测系统的黑盒攻击成功率低等问题. 同时目前通用目标检测系统的鲁棒性研究仍处于起步阶段,需要更多如 APRICOT 的基础数据集为相关研究提供数据支持.

6 人脸识别中的对抗样本攻击

人脸识别系统的目标是从数字图像或视频源的视频帧中识别或验证人员的身份. 根据攻击目标的差异,可以将人脸识别系统中的对抗样本攻击分为逃逸攻击、混淆攻击和冒充攻击三类. 逃逸攻击指对抗样本使特定的人脸无法被识别;混淆攻击指对抗样本使特定的人脸被识别为其他任意人脸;冒充攻击指攻击者伪装成一张特定(授权)的人脸. 本节将从数字域和物理域两个场景,对人脸识别系统中现有的对抗样本攻击方案进行梳理.

6.1 人脸识别中的数字域攻击技术

早期对人脸识别系统的对抗样本攻击多实现难度较低的逃逸攻击. Goswami 等人[156] 从图像和面部两类不同大小的区域研究人脸识别算法的脆弱性. 实验表明仅在人脸图像中添加随机噪声或黑色网格线即可严重降低人脸验证的准确性并实现逃逸攻击. 随后 Bose 等人[157] 提出对抗转换网络(Adversarial Transformation Network, ATN),利用对抗生成器网络解决人脸对抗样本的约束优化问题. Chatzikyriakidis 等人[158] 从隐私保护的角度出发提出 P-FGVM(Penalized Fast Gradient Value Method)方法,在图像的空间域上生成的对抗样本. P-FGVM 在实现了逃逸攻击的同时保证了面部图像的质量. Kwon 等人[159] 使用联合训练生成限定性的人脸对抗样本,该样本可以被"友方"模型正确识别而被"敌方"模型错误识别.

逃逸攻击只需要使模型无法识别人脸,实现难度较低;而混淆攻击和冒充攻击需要在保留模型的人脸识别能力的情况下,误导模型进行错误分类,实现难度较高. 在白盒场景下,基于 FGSM 方法 Rozsa 等人[160] 首次提出层状的原点-目标合成攻击方案(Layerwise Origin-Target Synthesis, LOTS),实现了针对深度人脸识别模型的冒充攻击. Dabouei 等人[161] 基于几何变换思想提出了快速标点变换(Fast Landmark Manipulation, FLM)方法和失真更小的分组快速标点变换(Grouped Fast Landmark Manipulation, GFLM)方法. 该类方法首先对原始图像进行空间坐标轴转换,之后在新坐标系下生成对抗样本,其攻击速度约是传统几何攻击的 200 倍.

考虑到白盒设置不符合实际攻击场景,Milton[162] 提出使用基于 MI-FGSM 和知识蒸馏的攻击方案,实现了对人脸识别模型的黑盒攻击. Yang 等人[163] 提出注意力对抗攻击生成网络(Attentional Adversarial Attack Generative Network, A^3 GN),

实现了冒充攻击. A3GN 包含条件变式自动编码器和注意模块，可以解析人脸之间的实例级对应关系，生成的对抗样本具有与目标人脸相同的特征表示. 而 Dong 等人[164]提出基于查询的进化攻击算法(Evolutionary Attack Algorithm，EAA)，利用协方差矩阵的适应性进化策略实现了在黑盒环境中对多个人脸识别模型的攻击. 与其他攻击方案相比，EAA 攻击方法的收敛速度更快且图像失真更小. Zhong 和 Deng[165]提出基于随机失活(Dropout)的人脸攻击网络群 DFANet 方法，在每一轮迭代中使用不同的随机失活层产生新的替代模型，以增加替代模型的多样性并提升攻击的整体效果. DFANet 实现了特征级的人脸对抗样本攻击并增强了黑盒场景下攻击的迁移性. 在对抗样本的黑盒攻击中，GAN 被广泛使用. Deb 等人[166]基于 GAN 提出了 AdvFaces 对抗人脸合成方法. 该方法自动生成添加到图像上的对抗掩模(mask)，生成的对抗人脸与目标人脸几乎没有区别. 为解决 GAN 方法中存在的局部最优局限和过拟合问题，Xiao 等人[167]在生成对抗样本时，在 GAN 的潜空间中进行正则化约束和目标优化，使对抗补丁与人脸相似度更高的同时缓解了过拟合提高了攻击的迁移成功率.

6.2　人脸识别中的物理域攻击技术

对物理域中的人脸识别系统进行对抗样本攻击的方案主要有两种思路：一种是在人脸区域添加可见的物理对抗补丁；另一种是生成光学扰动来欺骗物理域中人脸识别系统的摄像头.

物理对抗补丁是将数字域对抗补丁通过打印等方式转换到物理域，大多数攻击方案先在人脸的固定区域生成数字域的扰动，之后将其打印为物理贴纸. Sharif 等人[35]首次提出基于梯度的人脸识别系统攻击方案，该方法将扰动限制为眼镜形状的区域，通过打印眼镜的方法在数字域和物理域实现了逃逸攻击、混淆攻击和冒充攻击. 之后，Sharif 等人[168]使用 GAN 在数字域和物理域实现逃逸攻击. 该算法在数字域上使用 L-BFGS 等经典攻击方案，并通过佩戴 3D 打印的太阳镜框架，实现了在物理域中对模型的欺骗. Komkov 和 Petiushko 的另一项研究[169]提出了一种攻击人脸识别系统的方法 AdvHat，生成可以粘贴在帽子上的矩形图像，使人脸系统无法正确识别人脸的类别. Pautov 等人[170]的研究也实现了物理域的混淆攻击和冒名攻击. 他们将对抗补丁打印并粘贴在物理域的人脸上，然后拍摄对扰动后的人脸照片，误导系统将照片识别为目标人脸. 在这项工作中，补丁拥有多种形状，如鼻子、额头或眼镜等可穿戴配件. 为提高物理对抗补丁的隐蔽性，Yin 等人[171]提出可以通过对抗性化妆(Adv-Makeup)的方式，通过混合模型在眼眶区域合成不可察觉的对抗性眼影. 为提高攻击的迁移性，Adv-Makeup 提出了基于元学习的细粒度对抗攻击策略，从各种模型中学习更多的脆弱特征. 图 11 展示了针对人脸识别系统的物理对抗补丁攻击的效果图.

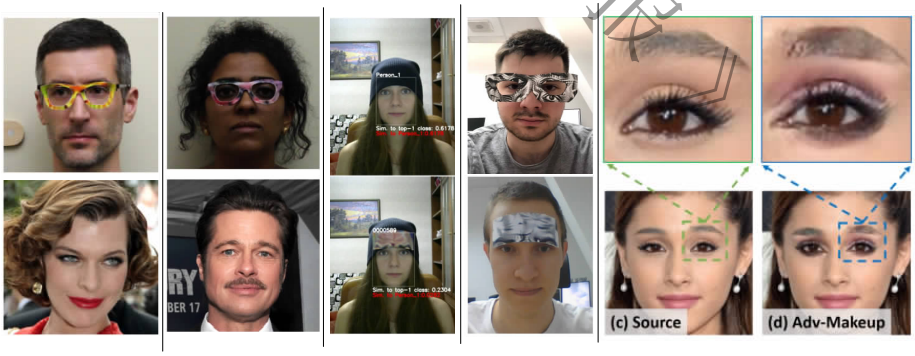

图 11　文献[35，168-171]对人脸识别系统的物理对抗补丁

光学扰动的攻击目标是系统的摄像头，利用摄像头和人眼成像原理的差异生成对抗样本，提高物理域攻击的隐蔽性. Zhou 等人[172]提出隐形面具攻击方法(Invisible Mask Attack，IMA)，他们设计了在帽檐上装有红外 LED 的鸭舌帽. 这种鸭舌帽直接在人脸上投射扰动光线，以实现物理域的隐蔽对抗样本攻击. 攻击者将带有扰动光线的面部照片输入模型，计算攻击损失并调整红外 LED 红外点的位置、大小和强度来实现逃逸攻击. 实验表明单个攻击者可以有效地对多个不同的目标人脸实现冒名攻击. 但长时间的红外射线对人体存在潜在危害，因此 IMA 方案的攻击只能持续较短时间. 为避免红外线的危害，Shen 等人[173]使用投影实现了基于可见光的攻击方案(Visible Light-based Attack，VLA)，VLA 方案利用投影仪产生物理域的扰动，对扰动损失进行分区计算并引入了隐藏框架，以降低人眼对扰动的察觉程度. VLA 可对人脸识别系统生成物理域中隐蔽性和鲁棒性的对抗样本. 同样基

于投影的思想，Nguyen 等人[174]提出对抗性光投射攻击(Adversarial Light Projection Attack，ALPA)方案，实现了对人脸识别系统的实时物理攻击. ALPA 方案首先使用相机捕捉攻击者的面部图像；之后，根据目标图像和攻击环境调整相机和投影仪的设置并创建数字对抗样本；最后，通过投影仪将数字图案投影到物理领域，从而实现混淆攻击和冒充攻击. 基于投影的攻击不需要定制可穿戴的红外发生器，攻击的适用范围更广且复用性更高，但其隐蔽性较差且对设备有更高的要求. 图 12 展示了针对人脸识别系统的物理光学对抗攻击的效果图.

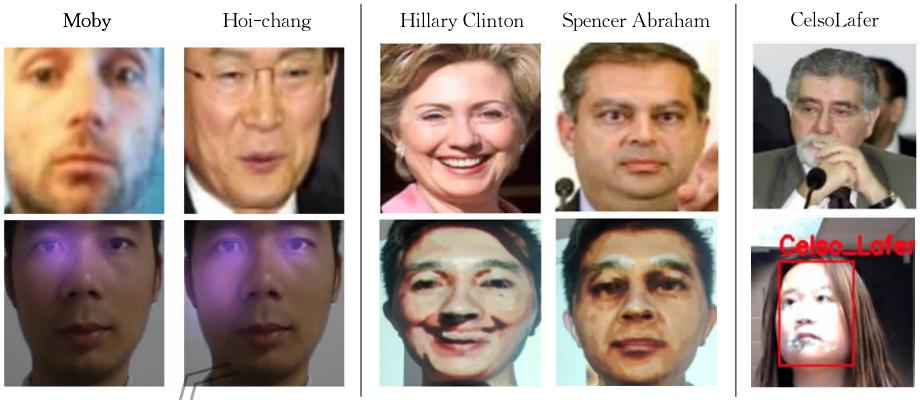

图 12 文献[172-174]对人脸识别系统的光学对抗攻击

6.3 小结

数字域中对人脸识别系统的攻击思路主要两大类：改造图像分类中的攻击方案以适应人脸识别系统和基于 GAN 网络训练生成式模型. 而物理域中对人脸识别系统的攻击主要思路有物理对抗补丁和光学扰动两大类. 表 7 对比分析了针对人脸识别任务进行对抗样本攻击的代表性研究.

在数字域，近期更多的攻击方案关注于现实性较高的黑盒场景，并且攻击目标也设定为难度更高的混淆攻击和冒充攻击；但人脸识别系统应用中可以进行数字域攻击的情况远少于物理域. 在物理域，对抗补丁攻击是数字域攻击最直接的扩展，其实现简单且成功率较高，但攻击难以复用；而光学扰动攻击有很强的复用性，对于不同的攻击目标可以实现快速的攻击，但在实际应用中环境光线、攻击角度、摄像头滤波等因素对攻击效果有很大影响. 目前的物理域攻击方案对人脸识别系统的识别角度有严格限制且多数方案只适用于白盒场景，在物理世界中更鲁棒的黑盒攻击方案仍有待进一步的研究.

表 7 人脸识别任务的对抗样本攻击

应用领域	方案	攻击设定	攻击效果	目标模型	测试数据集
数字域	Ref. [156]	白盒	逃逸攻击	OpenFace VGG-Face LightCNN L-CSSE	PaSC MEDS-II
	Ref. [157]	白盒	逃逸攻击	Inception-ResNet-v1	LFW VGGFace2
	P-FGVM[158]	白盒	逃逸攻击	VGG-Face	CelebA
	LOTS[160]	白盒	冒充攻击	VGG-Face	VGGFace
	GFLM[161]	白盒	混淆攻击	FaceNet Dlib	CASIA-WebFace LFW MS-Celeb-1M
	A3GN[163]	白盒&黑盒	冒充攻击	ArcFace	LFW MegaFace
	EAA[164]	黑盒	混淆攻击&冒充攻击	SphereFace ArcFace CosFace 腾讯 AI	CAISA-WebFace VGGFace2 MS-Celeb-1M
	DFANet[165]	白盒&黑盒	混淆攻击&冒充攻击	Amazon Microsoft Baidu Face++	CASIA-WebFace LFW
	AdvFaces[166]	白盒&黑盒	混淆攻击&冒充攻击	FaceNet SphereFace ArcFace	CelebA-HQ LFW
	Ref. [167]	白盒&黑盒	混淆攻击&冒充攻击	FaceNet CosFace ArcFace Face++阿里云	PaSC MEDS-II
物理域	Ref. [35]	白盒	逃逸攻击&混淆攻击&冒充攻击	VGG-Face Viola-Jones Face++	PubFig LFW
	AGN[168]	白盒	混淆攻击&冒充攻击	VGG-Face OpenFace	PubFig LFW
	AdvHat[169]	白盒	混淆攻击&冒充攻击	LResNet100E-IR ArcFace	CASIA-WebFace
	Ref. [170]	白盒	混淆攻击	LResNet100E-IR ArcFace	CASIA-WebFace
	Adv-Makeup[171]	白盒&黑盒	混淆攻击&冒充攻击	IR152 IRSE50 FaceNet MobileFace Microsoft Face++	LFW Makeup
	IMA[172]	白盒	逃逸攻击&混淆攻击&冒充攻击	FaceNet	LFW
	VLA[173]	黑盒	混淆攻击&冒充攻击	FaceNet SphereFace Dlib	LFW
	ALPA[174]	白盒&黑盒	逃逸攻击&混淆攻击&冒充攻击	FaceNet SphereFace	LFW

7　其他计算机视觉任务中的对抗样本

语义分割、图像检索、视觉目标跟踪任务的对抗样本研究与其他计算机视觉任务中的对抗样本研究相比，还处于方兴未艾的起步阶段，相关研究数量较少. 因此本节将用较短的篇幅整体性地介绍对抗样本技术在这些任务中的发展情况.

7.1　语义分割中的对抗样本技术

语义图像分割是一类密集预测任务，它需要为图像的每个像素分配一个类别标签. 语义分割任务实现了像素级别的多目标分类，对部分区域的扰动并不会影响大多数像素的分类结果；同时由于扰动阈值的限制，对图像各个像素点可添加的扰动有限，因此攻击者很难使全图准确率大幅度下降. 根据攻击的效果可以将攻击划分为静态语义分割攻击和动态语义分割攻击. 在静态语义分割中，攻击者定义一个固定时间的分割图像作为所有后续时间的目标输出，例如系统在 t_0 时间的预测. 静态语义分割攻击适用于基于静态摄像机的系统，以隐藏在时间 t_0 后开始的可疑活动. 静态攻击不适用于摄像机运动的情况，因为其没有考虑由图像视角变化引起的场景变化. 动态语义分割攻击旨在保持网络的主体分割结果不变，但删除某些目标类别，并用邻近的类别进行填补.

由于语义分割任务的复杂性，Xie 等人[129]提出了一种针对语义分割和目标检测的无目标攻击方法(Dense Adversary Generation，DAG)，DAG 对抗样本中所有的像素都会被划分到随机类别，但 DAG 算法只能对 FCN 等简单模型有效. 进一步地，Metzen 等人[175]提了通用型的对抗样本生成方法以实现动态攻击. 他们的方案同时解决了交叉熵损失无法适用于多目标的语义分任务的问题，实现了使语义分割模型无法识别所有图像中的特定类别的攻击效果. Xu 等人[176]利用基于迭代投影梯度的方法攻击场景分割模型，利用模型的一阶梯度信息实现了对最新的 DeepLab-V3＋模型的攻击. 而针对轻量级的语义分割网络，Kang 等人[177]提出一种非线性对抗样本生成方法，该方法通过噪声函数和中间变量间接生成对抗扰动.

普通攻击方案生成的对抗噪声没有语义含义，因此对医学成像系统的语义分割模型攻击效果较差. 基于此，Ozbulak 等人[178]提出了自适应分割掩码攻击方法(Adaptive Segmentation Mask Attack，ASMA)，攻击首先生成具有语义的掩码之后再生成扰动，通过指定目标类别来攻击医学图像分割模型. 同样针对医学语义分割系统，Chen 等人[179]探索使用 GAN 机制生成符合解剖变化和外观的对抗样本. 表 8 主要从攻击效果、攻击范围、攻击测试模型、测试数据集等角度对比分析了语义分割任务中对抗样本攻击相关的代表性研究工作.

表 8　语义分割任务的对抗样本攻击

方案	攻击设定	攻击效果	攻击范围	攻击思路	目标模型	测试数据集
DAG[129]	白盒	静态	针对	同时考虑所有像素并优化整体损失函数	FCN-Alex FCN-VGG	FCN
Ref.[175]	白盒	静态&动态	通用	改进交叉熵损失，优化区域扰动并铺满全图	FCN-VGG	Cityscapes
Ref.[176]	白盒	静态&动态	针对	改进迭代投影梯度的方案适用于	DeepLab-V3＋	Cityscapes
Ref.[177]	白盒	静态	针对&通用	通过噪声函数和中间变量间接生成扰动	ESPNet FastSCNN	Cityscapes
ASMA[178]	白盒	静态	针对	生成具有语义的掩码缩小攻击范围	U-Net	Glaucoma ISIC
Ref.[179]	白盒	静态	针对	使用 GAN 使扰动有解剖外观和变化	U-Net	abdominal CT

7.2　图像检索中的对抗样本技术

图像检索任务是基于开放集的任务，其核心是相似度度量. 针分类任务的攻击方法对图像检索任务的效果不佳，因为针对分类任务的攻击方案没有显式地改变图像特征空间的相似性，同时图像检索的测试集与训练集包含的标签不一定相同. 常见的图像检索系统包括以图搜图系统和行人重识别系统，本节将分别介绍对这两类系统的对抗样本攻击.

针对通用的以图搜图检索系统，Zheng 等人[180]提出反向特征攻击(Opposite-Direction Feature Attack，ODFA)首先查询目标模型的参数，之后使对抗样本特征向正确样本特征相反的方向改变，从而扩大攻击前后样本间的距离，使系统的检索结果出错. 该攻击方案的使用范围广，对行人重识别系统也有效. 为了生成与图像无关的通用对抗扰动，Li 等人[181]对图像的特征嵌入进行度量学习，逆向优化图像检索系统使用的三元组损失以生成对抗扰动. 为了实现黑盒攻击，方案对图像进行多尺度大小随机调整，同时利用知识蒸馏构建近似模型. 同样针对黑盒场景，Li 等人[182]提出基于查询的图像检索攻击方案(Query-based Attack against Image Retrieval，QAIR)实现了基于查询的黑盒攻击. QAIR 使用查询结果

递归地进行模型窃取，并生成扰动引导梯度的先验值，之后通过概率解释来量化对目标模型的攻击效果.

行人重识别(Person Re-Identification)[183]是一类特殊的图像检索问题，其在多个非重叠相机的视图中匹配目标人员. 我们根据攻击方案的提出时间对现有的攻击进行介绍. Bai 等人[184]首次针对行人重识别系统提出了对抗性度量攻击(Adversarial Metric Attack，AMA)，通过最大化干净图像和扰动图像之间的距离来生成对抗样本. AMA 方法包含攻击模型、度量方式、攻击方法和攻击的对抗性四个维度. Wang 等人[185]提出一种多级 GAN 网络架构，通过学习错误排名的公式来实现黑盒攻击，该架构将不同级别的特征进行金字塔化，用于生成兼具通用性和迁移性的对抗样本. 之前对行人重识别系统的攻击都是样本针对型的方案，Ding 等人[186]提出了一种图像无关和模型无关的通用型对抗样本攻击方法名为多项式通用对抗扰动生成方法(Polynomial Universal Adversarial Perturbation，PUAP). PUAP 由加性扰动和乘性调制因子组成，加性扰动项用于产生基础扰动，而乘性调制因子根据输入样本的脉冲模式调制扰动信号. 该方案首先使用列表攻击目标函数直接破坏相似度排名，之后利用模型无关的攻击算法生成扰动. 为了实现对物理域的行人重识别系统的攻击，Wang 等人[187]基于特征差异放大的思想提出了一种称为 advPattern 的攻击方案. advPattern 通过扩大不同摄像头下目标图像特征之间的距离实现逃逸攻击，通过拉进目标图像与特定对象图像之间的特征距离实现伪装攻击.

表 9 主要从攻击设定、攻击范围、目标模型、测试数据集等角度，对比分析了针对图像检索任务的对抗样本攻击的代表性研究工作.

表 9 图像检索任务的对抗样本攻击

方案名称	目标系统	攻击设定	攻击范围	攻击思路	目标模型	测试数据集
ODFA[180]	以图搜图 行人重识别	白盒&黑盒	针对型	直接改变中间特征且不需要正确类别	ResNet-50 DenseNet-121 PCB WideResNet28 VGG-16	Food-256 Market-1501 Cifar-10 Oxford5k Paris6k CUB-200-2011
Ref.[181]	以图搜图	白盒&黑盒	通用型	逆向优化图像检索系统的三元组损失	MAC(AlexNet/VGG-16/VGG-16) GeM(AlexNet/VGG-16/VGG-16) Google-Images	SfM-30k Oxford5k Paris6k
QAIR[182]	以图搜图	黑盒	针对型	通过查询窃取模型并用概率解释量化攻击效果	BN-Inception DenseNet121 Bing-Visual-Search	CUB-200-2011 In-Shop Stanford-Online-Products
AMA[184]	行人重识别	白盒&黑盒	针对型	从攻击的模型、方法、对抗性和度量方式四个维度构建攻击	ResNet-50 ResNeXt-50 DenseNet-121 HACNN Mancs	Market-1501 DukeMTMC
Ref.[185]	行人重识别	白盒&黑盒	针对型	构建多级 GAN 网络用错误的排序公式学习金字塔化的特征	IDE DenseNet-121 Mudeep AlignedReid PCB HACNN LSRO HHL SPGAN CamStyle+Era	Market-1501 CUHK03 DukeMTMC MSMT17 Cifar-10
PUAP[186]	行人重识别	白盒&黑盒	通用型	加性扰动和乘性调制因子构成的多项式拓展 UAP 攻击	ResNet50 DenseNet121 VGG16 SENet154 ShuffleNet	DukeMTMC Market-1501 MARS
advPattern[187]	行人重识别	白盒	针对型	扩大对抗样本与原始图像之间的特征距离	ResNet-50 VGG16	Market1501 PRCS(私有数据集)

7.3 视觉目标跟踪中的对抗样本技术

视觉目标跟踪(Visual Object Tracking，VOT)[188]指根据给定对象在初始帧中的位置，预测视频后续帧中对象的位置. 目标跟踪模型使用参考块搜索每一帧中最相似的区域，因此视觉目标跟踪的核心是相似性度量问题. 视觉目标跟踪使用的模型可分为非实时输入的离线模型和实时输入的在线模型.

离线模型的输入是固定的视频，所以攻击者可以对所有帧进行修改. Wiyatno 等人[189]首次提出生成物理对抗性纹理(Physical Adversarial Textures，PAT)的方案，基于一系列多样化的损失矩阵训练 GAN，以生成欺骗视觉目标跟踪系统的对抗样本. 当视觉跟踪的正确目标与对抗性海报同时出现时，跟踪器将锁定对抗样本而忽略正确跟踪目标. Siamese RPN 网络是目标跟踪器中的常用网络模块，一些研究设计了针对 Siamese 网络的攻击方案. Liang 等人[190]提出了一个端到端的快速攻击网络(Fast Adversarial Network，FAN)，在 GAN 的训练中将漂移损失与嵌入特征损失相结合，实现了对基于 Siamese 网络的跟踪器的攻击. 实验结果表明在白盒和黑盒场景中，FAN 均可实现高效的无目标和有目标攻击. 基于相似的思路，Yan 等人[191]提出冷却-收缩攻击(Cooling-Shrinking Attack，CSA)，通过降低 Siamese RPN 网络生成的热力图中正确目标存在区域的热度，并强制缩小预测边界框，使跟踪器无法正确检测到跟踪目标. 之前的研究集中于攻击

单目标跟踪器，针对自动驾驶中的完整视觉感知流程，Jia 等人[192]首次实现了一种名为跟踪器劫持(Tracker Hijacking)的多目标跟踪网络的对抗样本攻击. 该方案利用了多目标跟踪网络的跟踪误差，生成的对抗样本可将特定目标移入或移出自动驾驶系统视觉跟踪器的前方，从而造成严重的安全威胁.

对实时在线模型的攻击与离线模型有很大的差距，攻击者很难对在线模型产生可以跨帧作用的扰动，同时对在线模型的攻击也需要更高的攻击速度. 为解决这些问题，Chen 等人[193]提出高效的模型无关的双重注意力攻击方案(Dual Attention Attack，DAA)，DAA 方案使用双重注意力机制在视频第一帧生成对抗扰动，并使视觉目标跟踪器无法跟踪后续帧中的正确目标. Guo 等人[194]提出空间感知的在线增量攻击(Spatial-Aware Online Incremental Attack，SPARK)用于生成时空稀疏的扰动. SPARK 在新的视频帧中应用前一帧的扰动，之后优化损失函数生成最小的有效扰动增量. 除了使跟踪失效外，SPARK 攻击还可以误导目标跟踪器生成指定的错误轨迹. 由于实际应用场景中跟踪器的结构往往是未知的，Jia 等人[195]提出针对视觉跟踪模型的黑盒攻击方案交并比攻击(IoU Attack). 攻击基于当前帧和历史帧预测边界框的 IoU 分数指导对抗扰动的方向，通过迭代地添加扰动降低物体运动边界框准确性. 该方案根据模型输出预测框的变化引导对抗扰动的生成，避免了对模型梯度的求解，因此 IoU 攻击适用的模型种类更多且有效性更高.

表 10 主要从攻击设定、应用领域、攻击目标、目标模型、测试数据集等角度对比分析了视觉目标跟踪系统中对抗样本攻击的代表性研究.

表 10　视觉目标跟踪任务的对抗样本攻击

方案	攻击设定	应用领域	攻击目标	攻击思路	目标模型	测试数据集
PAT[189]	白盒	数字域 物理域	无	基于多样化的损失矩阵训练 GAN	GOTURN	合成数据集
FAN[190]	白盒&黑盒	数字域	无&有	结合漂移损失和嵌入特征损失训练GAN	SiamRPN SiamRPN++ SiamFC SiamRPN+CIR	OTB2013 OTB2015 VOT2014 VOT2018
CSA[191]	白盒	数字域	无	降低 Siamese_RPN 网络中正确目标热度	DaSiamRPN DiMP DaSiamRPN-UpdateNet	OTB100 LaSOT VOT2018
跟踪器劫持[192]	白盒	数字域	无&有	扩大多目标跟踪网络的跟踪误差	自训练多目标跟踪模型	Berkeley Deep Drive
DAA[193]	白盒	数字域	无	使用双重注意力机制并只对视频的第一帧生成对抗扰动	SiamFC SiamRPN SiamRPN++SiamMask	OTB100 LaSOT GOT10K
SPARK[194]	白盒&黑盒	数字域 物理域	无&有	基于之前帧的扰动生成扰动增量	SiamRPN SiamDW	OTB100 VOT2018 UAV123 LaSOT
IoU 攻击[195]	黑盒	数字域	无	迭代地降低扰动前后边界框的 IoU 值	SiamRPN++ DiMP LTMU	VOT2016 VOT2018 VOT2019 OTB100 NFS30 VOT2018-LT

7.4　小结

对语义分割任务的对抗样本研究，集中在对数字域的白盒模型和医学影像分割模型的攻击. 因为语义分割任务的攻击难度远高于图像分类和目标检测任务，所以目前相关研究也还处于起步阶段，黑盒攻击方案和对于特定语义类别的动态攻击方案还有待进一步的研究.

图像检索系统近年来发展迅速，应用范围不断扩大，不论是以图搜图还是行人重识别都是热门的新兴技术. 对图像检索系统的对抗样本研究目前还集中于数字域的白盒场景中，缺少在物理域和黑盒场景中的攻击方案；而对行人重识别系统的物理域对抗样本攻击也仍处于初级阶段，提升攻击效果和实现黑盒场景的攻击方案还有待相关研究者的进一步探索.

视觉目标跟踪系统分为离线模型和在线模型两类，针对两类模型的攻击技术有较大的差距. 其中，对于离线模型的攻击技术发展较快，较多工作对单目标跟踪系统进行了研究，而对多目标跟踪网络的研究还有待进一步的探索. 在线模型中攻击者无法事先获得被攻击的完整视频，因此攻击实现难度较大，且对攻击的效率要求更高，也是更具挑战性的研究方向.

8　对抗样本的未来发展方向

目前，深度学习系统被大量应用于图像分类、目标检测、语义分割、人脸识别等计算机视觉任务，其面临的对抗样本攻击威胁也受到了学术界和工业界的广泛关注. 寻找对抗样本特性、探究对抗样本生成原因以及提出新型对抗样本攻击方案是研究的重点问题；探索不同对抗样本的防御机制、提升系统的鲁棒性是研究的主要目标；结合前两者实现对抗样本的良性应用是今后研究的新方向. 综上，未来对抗样

本的相关研究可以从以下几个方向开展.

(1)完善对抗样本的理论研究体系

目前对抗样本的研究多集中于提出新型攻击技术和防御技术,而对于对抗样本存在原因这一核心问题,学界还未形成统一的理论体系.现有的一些研究使用拓扑学的方法分析数据流形或决策边界,利用统计学的方法提取关于对抗样本分布的信息,或使用学习理论的方法研究在怎样的理论假设下可以减少训练鲁棒模型所需的资源,但目前并未形成被学界广泛接受的理论体系.因此,突破对抗样本的存在性原理,建立数学理论完备、架构统一的对抗样本理论体系是一个亟待解决的问题.

(2)探索对抗样本攻击的结合应用

对抗样本存在极强的隐蔽性,因此除了在测试阶段对模型进行攻击外,对抗样本可以在训练阶段与其他攻击类型相结合,如投毒攻击和后门攻击.投毒攻击的目标是通过干扰训练阶段来影响机器学习模型预测阶段的表现,对训练数据进行修改是常用的方法之一.Koh 等人[196]首次提出利用对抗样本的思想,在不修改标签的情况下生成错误数据实现对模型的投毒攻击.后门攻击的目标是将隐蔽的后门嵌入到深度学习模型中,使受感染的模型在后门未激活时表现良好;而在后门被激活时预测将变为攻击者指定的结果.近来也有研究使用对抗样本的思路设计后门,如 Yan 等人[197]利用对抗样本攻击修改没有标签的训练数据,从而在半监督神经网络学习模型中植入后门.因此,对抗样本与其他攻击方法的结合应用拓展了对抗样本的应用范围,同时也有利于从更多样化的角度探究模型的脆弱性.

(3)设计适用于复杂任务的模型鲁棒性测试体系

鲁棒性指模型在面对输入样本发生变化时是否能保持输出的稳定性,它是衡量机器学习模型性能的一项重要的评价指标,鲁棒性的高低直接决定了模型的泛化能力.对抗样本是鲁棒性测试的重点内容之一,现有的模型鲁棒性评估体系几乎都只针对图像分类任务,还没有完善的测试体系可实现对目标检测、语义分割等系统的鲁棒性评估.而深度学习模型在计算机视觉中广泛应用并部署于安全攸关的系统中,如在安全监控中的人脸识别系统.现有研究已提出针对目标检测、人脸识别等任务的对抗样本攻击方案,对这些系统的应用构成了严重的安全威胁.因此,针对不同的计算机视觉任务,构造通用的对抗样本鲁棒性测试平台,使系统鲁棒性评估统一化、标准化是现实意义重大的挑战.

(4)构建快速高效适应性强的对抗样本防御机制

对抗样本的防御机制旨在通过对模型进行修改或者添加结构以提高脆弱模型的鲁棒性.一些研究在该方向进行了探索,如 Li 等人[198]对工业人工智能系统的对抗样本防御进行了研究,但目前的防御机制仍存在大量缺陷.首先,目前的防御方法大多设定攻击者无法获取防御方案,这类防御大多无法抵御白盒场景下的对抗样本攻击.其次,新型对抗样本技术层出不穷,现有防御方案的有效性难以得到长期保障.例如,基于噪声检测的防御方案通过检测样本的平滑度和噪声大小实现样本过滤,该类方法几乎无法防御语义对抗样本.现有的可验证防御方案表明一定范围内系统的鲁棒性是可以保障的,但是该类方法无法应用于更复杂的现实系统.突破对抗样本白盒防御的难点,实现快速、高效、适用范围广的防御机制也是未来研究的重要方向.

(5)探寻对抗样本的良性应用

对抗样本除了可以误导智能系统以外,还可以揭露系统的脆弱性,进一步的可用于对抗训练以提升模型的鲁棒性.此外,目前已有一些学者将对抗样本应用于其他领域,进行了一些探索性研究.如 Hu 等人[199]利用对抗样本修改人脸数据,使得用户的人脸无法被用于模型训练,从而保护了用户的数据隐私.Xu 等人[200]利用对抗样本消除数据偏见,提升深度学习系统的公平性.目前对抗样本良性应用的研究还处于起步阶段,相关的研究针对特定领域且适用范围较窄.因此,拓展对抗样本的良性应用场景也是有待进一步探索的研究方向.

9 总 结

本文对计算机视觉系统中的对抗样本攻击技术研究进展进行了综述.首先阐述了计算机视觉任务中对抗样本的定义、敌手模型和对抗扰动类别;之后,讨论了计算机视觉系统中对抗样本存在原因的相关假设;接着,根据计算机视觉任务的不同,分别总结分析了图像分类、目标检测、人脸识别、语义分割、图像检索、视觉目标跟踪 6 大任务上对抗样本攻击方向的最新研究进展,并对比分析了不同对抗样本攻击技术的优缺点;最后,展望了在计算机视觉中对抗样本的未来研究方向.期望本文的工作,能给以后的研究者提供有益的参考与借鉴,为计算机视觉应用中对抗样本攻击的进一步发展做出贡献.

参 考 文 献

[1] Hinton G, Deng L, Yu D, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. IEEE Signal Processing Magazine, 2012, 29(6): 82-97

[2] Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks//Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2014. Montreal, Canada, 2014: 3104-3112

[3] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks//Proceedings of the Advances in Neural Information Processing Systems 26: Annual Conference on Neural Information Processing Systems 2012. Lake Tahoe, USA, 2012: 1097-1105

[4] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 770-778

[5] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature, 2017, 550(7676): 354-359

[6] Berner C, Brockman G, Chan B, et al. Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv: 1912.06680, 2019

[7] Bojarski M, del Testa D, Dworakowski D, et al. End to end learning for self-driving cars. arXiv preprint arXiv: 1604.07316, 2016

[8] He Y, Meng G, Chen K, et al. Towards security threats of deep learning systems: A Survey. IEEE Transactions on Software Engineering, 2022, 48(5): 1743-1770

[9] Yuan Z, Lu Y, Wang Z, Xue Y. Droid-Sec: Deep learning in android malware detection//Proceedings of the ACM SIGCOMM 2014 Conference. Chicago, USA, 2014: 371-372

[10] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-9

[11] Moosavi-Dezfooli S-M, Fawzi A, Fawzi O, Frossard P. Universal adversarial perturbations//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 86-94

[12] Papernot N, McDaniel P, Sinha A, Wellman M. Towards the science of security and privacy in machine learning. arXiv preprint arXiv: 1611.03814, 2016

[13] Zhang G, Yan C, Ji X, et al. DolphinAttack: Inaudible voice commands//Proceedings of the ACM Conference on Computer and Communications Security. Dallas, USA, 2017: 103-117

[14] Athalye A, Engstrom L, Ilyas A, Kwok K. Synthesizing robust adversarial examples//Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Stockholm, Sweden, 2018, 80: 284-293

[15] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 1625-1634

[16] Zhang Si-Si, Zuo Xin, Liu Jian-Wei. The problem of the adversarial examples in deep learning. Chinese Journal of Computers, 2019, 42(8): 1886-1904(in Chinese)
(张思思，左信，刘建伟. 深度学习中的对抗样本问题. 计算机学报，2019，42(8)：1886-1904)

[17] Pan Wen-Wen, Wang Xin-Yu, Song Ming-Li, Chen Chun. Survey on generating adversarial examples. Journal of Software, 2020, 31(1): 67-81(in Chinese)
(潘文雯，王新宇，宋明黎，陈纯. 对抗样本生成技术综述. 软件学报，2020，31(1)：67-81)

[18] Serban A, Poll E, Visser J. Adversarial examples on object recognition: A comprehensive survey. ACM Computing Surveys, 2020, 53(3): 1-38

[19] Ren K, Zheng T, Qin Z, Liu X. Adversarial attacks and defenses in deep learning. Engineering, 2020, 6(3): 346-360

[20] Ren H, Huang T, Yan H. Adversarial examples: Attacks and defenses in the physical world. International Journal of Machine Learning and Cybernetics, 2021, 12(11): 3325-3336

[21] Bhambri S, Muku S, Tulasi A, Buduru A B. A survey of black-box adversarial attacks on computer vision models. arXiv preprint arXiv:1912.01667, 2019

[22] Li Xin-Jiao, Wu Guo-Wei, Yao Lin, et al. Progress and future challenges of security attacks and defense mechanisms in machine learning. Journal of Software, 2021, 32(2): 406-423(in Chinese)
(李欣姣，吴国伟，姚琳等. 机器学习安全攻击与防御机制研究进展和未来挑战. 软件学报，2021，32(2)：406-423)

[23] Li Ming-Hui, Jiang Pei-Pei, Wang Qian, et al. Adversarial attacks and defenses for deep learning models. Journal of Computer Research and Development, 2021, 58(5): 909-926(in Chinese)
(李明慧，江沛佩，王骞等. 针对深度学习模型的对抗性攻击与防御. 计算机研究与发展，2021，58(5)：909-926)

[24] Xu H, Ma Y, Liu H C, et al. Adversarial attacks and defenses in images, graphs and text: A review. International Journal of Automation and Computing, 2020, 17(2): 151-178

[25] Larochelle H, Bengio Y, Louradour J, Ca L U. Exploring strategies for training deep neural networks. Journal of Machine Learning Research, 2009, 10(1): 1-40

[26] Papernot N, McDaniel P, Goodfellow I, et al. Practical black-box attacks against machine learning//Proceedings of the ACM Asia Conference on Computer and Communications Security. Abu Dhabi, United Arab Emirates, 2017: 506-519

[27] Barreno M, Nelson B, Joseph A D, Tygar J D. The security of machine learning. Machine Learning, 2010, 81(2): 121-148

[28] Williams S B. The association for computing machinery. Journal of the ACM, 1954, 1(1): 1-3

[29] Yuan X, He P, Zhu Q, Li X. Adversarial examples: Attacks and defenses for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 2019, 30(9): 2805-2824

[30] Zhang C, Benz P, Imtiaz T, Kweon IS. Cd-uap: Class discriminative universal adversarial perturbation//Proceedings of the AAAI Conference on Artificial Intelligence. New York, USA, 2020, 34(04): 6754-6761

[31] Wong E, Schmidt F R, Kolter J Z. Wasserstein adversarial examples via projected sinkhorn iterations//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019, 97: 6808-6817

[32] Rozsa A, Rudd E M, Boult T E. Adversarial diversity and hard positive generation//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Las Vegas, USA, 2016: 410-417

[33] Wang Z, Song M, Zheng S, et al. Invisible adversarial attack against deep neural networks: An adaptive penalization approach. IEEE Transactions on Dependable and Secure Computing, 2019, 18(3): 1474-1488

[34] Xiao Z, Gao X, Fu C, et al. Improving transferability of adversarial patches on face recognition with generative models//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 11845-11854

[35] Sharif M, Bhagavatula S, Bauer L, Reiter M K. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition//Proceedings of the 2016 ACM Conference on Computer and Communications Security. Vienna, Austria, 2016: 1528-1540

[36] Meng D, Chen H. MagNet: A two-pronged defense against adversarial examples//Proceedings of the ACM Conference on Computer and Communications Security. Dallas, USA, 2017: 135-147

[37] Gu S, Rigazio L. Towards deep neural network architectures robust to adversarial examples. arXiv preprint arXiv: 1412.5068, 2014

[38] Song Y, Kim T, Nowozin S, et al. PixelDefend: Leveraging generative models to understand and defend against adversarial examples//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-20

[39] Ghosh P, Losalka A, Black M J. Resisting adversarial attacks using Gaussian mixture variational autoencoders//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 541-548

[40] Lee H, Han S, Lee J. Generative adversarial trainer: Defense to adversarial perturbations with GAN. arXiv preprint arXiv: 1705.03387, 2017

[41] Carlini N, Wagner D. Towards evaluating the robustness of neural networks//Proceedings of the 38th IEEE Symposium on Security and Privacy. San Jose, USA, 2017: 39-57

[42] Gilmer J, Metz L, Faghri F, et al. Adversarial spheres//Proceedings of the 6th International Conference on Learning Representations Workshop. Vancouver, Canada, 2018: 1-13

[43] Fawzi A, Fawzi H, Fawzi O. Adversarial vulnerability for any classifier//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 1186-1195

[44] Mahloujifar S, Diochnos D I, Mahmoody M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure//Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 4536-4543

[45] Shafahi A, Huang W R, Studer C, et al. Are adversarial examples inevitable?//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-17

[46] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015: 1-10

[47] Luo Y, Boix X, Roig G, et al. Foveation-based mechanisms alleviate adversarial examples. arXiv preprint arXiv: 1511.06292, 2015

[48] Fawzi A, Fawzi O, Frossard P. Fundamental limits on adversarial robustness//Proceedings of the 32nd International Conference on Machine Learning Workshop. Lille, France, 2015: 1-7

[49] Tabacof P, Valle E. Exploring the space of adversarial images//Proceedings of the 2016 International Joint Conference on Neural Networks. Vancouver, Canada, 2016: 426-433

[50] Moosavi-Dezfooli S-M, Fawzi A, Frossard P, et al. DeepFool: A simple and accurate method to fool deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 2574-2582

[51] Buckman J, Roy A, Raffel C, Goodfellow I J. Thermometer encoding: One hot way to resist adversarial examples//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-22

[52] Tanay T, Griffin L. A boundary tilting perspective on the phenomenon of adversarial examples. arXiv preprint arXiv: 1608.07690, 2016

[53] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially robust generalization requires more data//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 5019-5031

[54] Bubeck S, Lee Y T, Price E, Razenshteyn I P. Adversarial examples from computational constraints//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019, 97: 831-840

[55] Cullina D, Bhagoji A N, Mittal P. PAC-Learning in the presence of evasion adversaries//Proceedings of the 32nd International Conference on Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 228-239

[56] Tsipras D, Santurkar S, Engstrom L, et al. Robustness may be at odds with accuracy//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-23

[57] Izmailov R, Sugrim S, Chadha R, et al. Enablers of adversarial attacks in machine learning//Proceedings of the IEEE Military Communications Conference. Los Angeles, USA, 2018: 425-430

[58] Ilyas A, Santurkar S, Tsipras D, et al. Adversarial examples are not bugs, they are features//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 125-136

[59] Su J, Vargas D V, Sakurai K. One-pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 2019, 23(5): 828-841

[60] Nguyen A M, Yosinski J, Clune J. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Boston, USA, 2015: 427-436

[61] Kurakin A, Goodfellow I J, Bengio S. Adversarial machine learning at scale//Proceedings of the 5th International Conference on Learning Representations. Toulon, France, 2017: 1-17

[62] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-23

[63] Dong Y, Liao F, Pang T, et al. Boosting adversarial attacks with momentum//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 9185-9193

[64] Papernot N, McDaniel P D, Jha S, et al. The limitations of deep learning in adversarial settings//Proceedings of the IEEE European Symposium on Security and Privacy. Saarbrücken, Germany, 2016: 372-387

[65] Oseledets I, Khrulkov V. Art of singular vectors and universal adversarial perturbations//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 8562-8570

[66] Huang R, Xu B, Schuurmans D, Szepesvari C. Learning with a strong adversary. arXiv preprint arXiv: 1511.03034, 2015

[67] Athalye A, Carlini N, Wagner D A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples//Proceedings of the 35th International Conference on Machine Learning. Stockholmsmässan, Sweden, 2018, 80: 274-283

[68] Tramèr F, Kurakin A, Papernot N, et al. Ensemble adversarial training: Attacks and defenses//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-22

[69] Chen P Y, Sharma Y, Zhang H, et al. EAD: Elastic-Net attacks to deep neural networks via adversarial examples//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018, 32(1): 10-17

[70] Kingma D P, Welling M. Auto-Encoding variational Bayes//Proceedings of the 2nd International Conference on Learning Representations. Banff, Canada, 2014: 1-14

[71] Rezende D J, Mohamed S, Wierstra D. Stochastic backpropagation and approximate inference in deep generative models//Proceedings of the 31th International Conference on Machine Learning. Beijing, China, 2014, 32: 1278-1286

[72] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets//Proceedings of the Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2014: 2672-2680

[73] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks//Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2687-2695

[74] Poursaeed O, Katsman I, Gao B, Belongie S. Generative adversarial perturbations//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4422-4431

[75] Song Y, Shu R, Kushman N, Ermon S. Constructing unrestricted adversarial examples with generative models//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018, 31: 8312-8323

[76] Engstrom L, Tsipras D, Schmidt L, Madry A. A rotation and a translation suffice: Fooling CNNs with simple transformations. CoRR, abs/1712.02779, 2017

[77] Kanbak C, Moosavi-Dezfooli S-M, Frossard P. Geometric robustness of deep networks: Analysis and improvement//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA, 2018: 4441-4449

[78] Pei K, Cao Y, Yang J, Jana S. Towards practical verification of machine learning: The case of computer vision systems. arXiv preprint arXiv:1712.01785, 2017

[79] Xiao C, Zhu J-Y, Li B, et al. Spatially transformed adversarial examples//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-30

[80] Zhang H, Chen H, Song Z, et al. The limitations of adversarial training and the blind-spot attack//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-16

[81] Hosseini H, Xiao B, Jaiswal M, Poovendran R. On the limitation of convolutional neural networks in recognizing negative images//Proceedings of the 16th IEEE International Conference on Machine Learning and Applications. Cancun, Mexico, 2017: 352-358

[82] Afifi M, Brown M. What else can fool deep learning? Addressing color constancy errors on deep neural network

performance//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 243-252

[83] Hosseini H, Poovendran R. Semantic adversarial examples//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Salt Lake City, USA, 2018: 1614-1619

[84] Shamsabadi A S, Sanchez-Matilla R, Cavallaro A. ColorFool: Semantic adversarial colorization//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 1148-1157

[85] Ruderman D L, Cronin T W, Chiao C C. Statistics of cone responses to natural images: Implications for visual coding. Journal of the Optical Society of America A, 1998, 15(8): 2036-2045

[86] Laidlaw C, Feizi S. Functional adversarial attacks//Proceedings of the Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 10408-10418

[87] Bhattad A, Chong M J, Liang K, et al. Unrestricted adversarial examples via semantic manipulation//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-19

[88] Zhang R, Zhu J-Y, Isola P, et al. Real-time user-guided image colorization with learned deep priors. ACM Transactions on Graphics, 2017, 36(4): 119:1-119:11

[89] Tian B, Felix J, Guo Q, et al. AVA: Adversarial vignetting attack against visual recognition//Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI). Virtual, 2021: 1046-1053

[90] Zhao Z, Dua D, Singh S. Generating natural adversarial examples//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018: 1-15

[91] Joshi A, Mukherjee A, Sarkar S, Hegde C. Semantic adversarial attacks: Parametric transformations that fool deep classifiers//Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 4772-4782

[92] Qiu H, Xiao C, Yang L, et al. SemanticAdv: Generating adversarial examples via attribute-conditioned image editing //Proceedings of the European Conference on Computer Vision. Glasgow, UK, 2020, 12359: 19-37

[93] Liu H-T D, Tao M, Li C-L, et al. Beyond pixel norm-balls: Parametric adversaries using an analytically differentiable renderer//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-12

[94] Wang S, Chen S, Chen T, et al. Generating semantic adversarial examples via feature manipulation. arXiv preprint arXiv: 2001.02297, 2020

[95] Chen P Y, Zhang H, Sharma Y, et al. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the ACM Workshop on Artificial Intelligence and Security, Co-Located with CCS. Dallas, USA, 2017: 15-26

[96] Kingma D P, Ba J. Adam: A method for stochastic optimization//Proceedings of the 3rd International Conference on Learning Representations. San Diego, USA, 2015: 1-15

[97] Cheng M, Le T, Chen P Y, et al. Query-efficient hard-label black-box attack: An optimization-based approach//Proceedings of the International Conference on Learning Representations. New Orleans, USA, 2019: 1-14

[98] Tu C C, Ting P, Chen P Y, et al. AutoZOOM: Autoencoder-based zeroth order optimization method for attacking black-box neural networks//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 742-749

[99] Ilyas A, Engstrom L, Athalye A, Lin J. Black-box adversarial attacks with limited queries and information//Proceedings of the International Conference on Machine Learning. Stockholmsmässan, Stockholm, Sweden, 2018, 80: 2142-2151

[100] Ilyas A, Engstrom L, Madry A. Prior convictions: Black-box adversarial attacks with bandits and priors//Proceedings of the 7th International Conference on Learning Representations. New Orleans, USA, 2019: 1-25

[101] Bhagoji A N, He W, Li B, Song D. Practical black-box attacks on deep neural networks using efficient query mechanisms//Proceedings of the European Conference on Computer Vision. Munich, Germany, 2018, 11216: 158-174

[102] Du Y, Fang M, Yi J, et al. Towards query efficient black-box attacks: An input-free perspective//Proceedings of the 11th ACM Workshop on Artificial Intelligence and Security. Toronto, Canada, 2018: 13-24

[103] Shi Y, Wang S, Han Y. Curls&Whey: Boosting black-box adversarial attacks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA, 2019: 6512-6520

[104] Dong Y, Pang T, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA, 2019: 4307-4316

[105] Wang Z, Guo H, Zhang Z, et al. Feature importance-aware transferable adversarial attacks. arXiv preprint arXiv: 2107.14185, 2021

[106] Huan Z, Wang Y, Zhang X, et al. Data-free adversarial perturbations for practical black-box attack//Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining. Singapore, 2020: 127-138

[107] Narodytska N, Kasiviswanathan S P. Simple black-box adversarial perturbations for deep networks. arXiv preprint arXiv: 1612.06299, 2016

[108] Narodytska N, Kasiviswanathan S P. Simple black-box adversarial attacks on deep neural networks//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. Honolulu, USA, 2017: 1310-1318

[109] Li Y, Li L, Wang L, et al. NATTACK: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019, 97: 3866-3876

[110] Brendel W, Rauber J, Bethge M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models//Proceedings of the International Conference on Learning Representations. Vancouver, Canada, 2018: 1-12

[111] Brunner T, Diehl F, Le M T, Knoll A. Guessing smart: Biased sampling for efficient black-box adversarial attacks//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 4957-4965

[112] Perlin K. An image synthesizer//Proceedings of the Annual Conference on Computer Graphics and Interactive Techniques. San Francisco, USA, 1985: 287-296

[113] Chen J, Jordan M I, Wainwright M J. HopSkipJumpAttack: A query-efficient decision-based attack//Proceedings of the IEEE Symposium on Security and Privacy. San Francisco, USA, 2020: 1277-1294

[114] Guo C, Gardner J R, You Y, et al. Simple black-box adversarial attacks//Proceedings of the International Conference on Machine Learning. Long Beach, USA, 2019, 97: 2484-2493

[115] Chen J, Su M, Shen S, et al. POBA-GA: Perturbation optimized black-box adversarial attacks via genetic algorithm. Computers & Security, 2019, 85(1): 89-106

[116] Alzantot M, Zhang H, Sharma Y, et al. GenaTTack: Practical black-box attacks with gradient-free optimization//Proceedings of the Genetic and Evolutionary Computation Conference. Prague, Czech, 2019: 1111-1119

[117] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world//Proceedings of the 5th International Conference on Learning Representations Workshop. Toulon, France, 2017: 1-14

[118] Liu A, Liu X, Fan J, et al. Perceptual-sensitive GAN for generating adversarial patches//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019, 33: 1028-1035

[119] Worzyk N, Kahlen H, Kramer O. Physical adversarial attacks by projecting perturbations//Proceedings of the Artificial Neural Networks and Machine Learning. Munich, Germany, 2019, 11729: 649-659

[120] Gnanasambandam A, Sherman A M, Chan S H. Optical adversarial attack//Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual, 2021: 92-101

[121] Li J, Schmidt F R, Kolter J Z. Adversarial camera stickers: A physical camera-based attack on deep learning systems//Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019, 97: 3896-3904

[122] Ren S, He K, Girshick R B, Sun J. Faster R-CNN: Towards real-time object detection with region proposal networks//Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems. Montreal, Canada, 2015: 91-99

[123] Lin T Y, Dollár P, Girshick R B, et al. Feature pyramid networks for object detection//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, USA, 2017: 936-944

[124] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector//Proceedings of the 14th European Conference of Computer Vision. Amsterdam, The Netherlands, 2016, 9905: 21-37

[125] Redmon J, Divvala S K, Girshick R B, Farhadi A. You Only Look Once: Unified, real-time object detection//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 779-788

[126] Neubeck A, van Gool L. Efficient non-maximum suppression//Proceedings of the International Conference on Pattern Recognition. Hong Kong, China, 2006, 3: 850-855

[127] Lu J, Sibai H, Fabry E, Forsyth D. No need to worry about adversarial examples in object detection in autonomous vehicles. arXiv preprint arXiv:1707.03501, 2017

[128] Zhang H, Wang J. Towards adversarially robust object detection//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 421-430

[129] Xie C, Wang J, Zhang Z, et al. Adversarial examples for semantic segmentation and object detection//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 1378-1387

[130] Xiao Y, Pun C M, Liu B. Fooling deep neural detection networks with adaptive object-oriented adversarial perturbation. Pattern Recognition, 2021, 115(1): 107903

[131] Zhang H, Zhou W, Li H. Contextual adversarial attacks for object detection//Proceedings of the IEEE International Conference on Multimedia and Expo. London, UK, 2020: 1-6

[132] Wang D, Li C, Wen S, et al. Daedalus: Breaking nonmaximum suppression in object detection via adversarial examples. IEEE Transactions on Cybernetics, 2022, 52(8): 7427-7440

[133] Chen S T, Cornelius C, Martin J, Chau D H P. ShapeShifter: Robust physical adversarial attack on faster R-CNN object detector. Lecture Notes in Computer Science, 2019, 11051 LNAI: 52-68

[134] Wei X, Liang S, Chen N, Cao X. Transferable adversarial attacks for image and video object detection//Proceedings of the 28th International Joint Conference on Artificial Intelligence. Macao, China, 2019: 954-960

[135] Li Y, Xu G, Li W. FA: A fast method to attack real-time object detection systems//Proceedings of the 2020 IEEE/CIC International Conference on Communications. Virtual, 2020: 1268-1273

[136] Deng X, Fang Z, Zheng Y, et al. Adversarial examples with transferred camouflage style for object detection. Journal of Physics Conference Series, 2021, 1738(1): 012130

[137] Wang Y, Tan Y, Zhang W, et al. An adversarial attack on DNN-based black-box object detectors. Journal of Network and Computer Applications, 2020, 161(1): 102634

[138] Liao Q, Wang X, Kong B, et al. Fast local attack: Ggenerating local adversarial examples for object detectors//Proceedings of the 2020 International Joint Conference on Neural Networks. Glasgow, United Kingdom, 2020: 1-8

[139] Liao Q, Wang X, Kong B, et al. Category-wise attack: Transferable adversarial examples for anchor free object detection. arXiv preprint arXiv:2003.04367, 2020

[140] Li Y, Tian D, Chang M-C, et al. Robust adversarial perturbation on deep proposal-based models//Proceedings of the British Machine Vision Conference. Newcastle, UK, 2018: 231

[141] Wu X, Huang L, Gao C. G-UAP: Generic universal adversarial perturbation that fools RPN-based detectors//Proceedings of the 11th Asian Conference on Machine Learning. Nagoya, Japan, 2019, 101: 1204-1217

[142] Li D, Zhang J, Huang K. Universal adversarial perturbations against object detection. Pattern Recognition, 2021, 110(1): 107584

[143] Liu X, Yang H, Liu Z, et al. DPATCH: An adversarial patch attack on object detectors//Proceedings of the 33rd AAAI Conference on Artificial Intelligence Workshop. Honolulu, USA, 2019: 2301

[144] Huang Y, Wang F, Kong A W-K, Lam K-Y. New threats against object detector with non-local block//Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12365: 481-497

[145] Saha A, Subramanya A, Patil K, Pirsiavash H. Role of spatial context in adversarial robustness for object detection //Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Virtual, 2020: 3403-3412

[146] Lee M, Kolter J Z. On physical adversarial patches for object detection. arXiv preprint arXiv:1906.11897, 2019

[147] Braunegg A, Chakraborty A, Krumdick M, et al. APRICOT: A dataset of physical adversarial attacks on object detection// Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12366: 35-50

[148] Lu J, Sibai H, Fabry E. Adversarial examples that fool detectors. arXiv preprint arXiv:1712.02494, 2017

[149] Song D, Eykholt K, Evtimov I, et al. Physical adversarial examples for object detectors//Proceedings of the 12th USENIX Workshop on Offensive Technologies. Baltimore, USA, 2018: 1-10

[150] Huang Y, Kong A W-K, Lam K-Y. Attacking object detectors without changing the target object//Proceedings of the 16th Pacific Rim International Conference on Artificial Intelligence. Cuvu, Fiji, 2019, 11672: 3-15

[151] Huang Y, Kong A W-K, Lam K-Y. Adversarial signboard against object detector//Proceedings of the 30th British Machine Vision Conference. Cardiff, UK, 2019: 231-241

[152] Zhao Y, Zhu H, Liang R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. London, UK, 2019: 1989-2004

[153] Xu K, Zhang G, Liu S, et al. Adversarial T-shirt! Evading person detectors in a physical world//Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12350: 665-681

[154] Wu Z, Lim S-N, Davis L S, Goldstein T. Making an invisibility cloak: Real world adversarial attacks on object detectors //Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12349: 1-17

[155] Huang L, Gao C, Zhou Y, et al. Universal physical camouflage attacks on object detectors//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2020: 717-726

[156] Goswami G, Ratha N, Agarwal A, et al. Unravelling robustness of deep learning based face recognition against adversarial attacks//Proceedings of the 32nd AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018, 32: 6829-6836

[157] Bose A J, Aarabi P. Adversarial attacks on face detectors using neural net based constrained optimization//Proceedings of the 2018 IEEE 20th International Workshop on Multimedia Signal Processing. Vancouver Campus, Canada, 2018: 1-6

[158] Chatzikyriakidis E, Papaioannidis C, Pitas I. Adversarial face de-identification//Proceedings of the 2019 IEEE International Conference on Image Processing. Taipei, China, 2019: 684-688

[159] Kwon H, Kwon O, Yoon H, Park K-W. Face friend-safe adversarial example on face recognition system//Proceedings of the 11th International Conference on Ubiquitous and Future Networks. Zagreb, Croatia, 2019: 547-551

[160] Rozsa A, Günther M, Boult T E. LOTS about attacking deep features//Proceedings of the 2017 IEEE International Joint Conference on Biometrics. Denver, USA, 2017: 168-176

[161] Dabouei A, Soleymani S, Dawson J, Nasrabadi N M. Fast geometrically-perturbed adversarial faces//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision. Hawaii, USA, 2019: 1979-1988

[162] Milton M A A. Evaluation of momentum diverse input iterative fast gradient sign method (M-DI2-FGSM) based attack method on MCS 2018 adversarial attacks on black box face recognition system. arXiv preprint arXiv: 1806.08970, 2018

[163] Yang L, Song Q, Wu Y. Attacks on state-of-the-art face recognition using attentional adversarial attack generative network. Multimedia Tools and Applications, 2021, 80(1): 855-875

[164] Dong Y, Su H, Wu B, et al. Efficient decision-based black-box adversarial attacks on face recognition//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Los Angeles, USA, 2019: 7706-7714

[165] Zhong Y, Deng W. Towards transferable adversarial attack against deep face recognition. IEEE Transactions on Information Forensics and Security, 2021, 16(1): 1452-1466

[166] Deb D, Zhang J, Jain A K. AdvFaces: Adversarial face synthesis//Proceedings of the 2020 IEEE/IAPR International Joint Conference on Biometrics. Houston, USA, 2020: 1-10

[167] Xiao Z, Gao X, Fu C, et al. Improving transferability of adversarial patches on face recognition with generative models

//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 11845-11854

[168] Sharif M, Bhagavatula S, Bauer L, Reiter M K. A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security, 2019, 22(3): 1-30

[169] Komkov S, Petiushko A. AdvHat: Real-world adversarial attack on arcFace face ID system//Proceedings of the International Conference on Pattern Recognition. Milan, Italy, 2020: 819-826

[170] Pautov M, Melnikov G, Kaziakhmedov E, et al. On adversarial patches: Real-world attack on ArcFace-100 face recognition system//Proceedings of the 2019 International Multi-Conference on Engineering, Computer and Information Sciences. Yekaterinburg, Russia, 2019: 0391-0396

[171] Yin B J, Wang W X, Yao T P, et al. Adv-Makeup: A new imperceptible and transferable attack on face recognition//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Virtual, 2021: 1252-1258

[172] Zhou Z, Tang D, Wang X, et al. Invisible mask: Practical attacks on face recognition with infrared. arXiv preprint arXiv:1803.04683, 2018

[173] Shen M, Liao Z, Zhu L, et al. VLA: A practical visible light-based attack on face recognition systems in physical world//Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies. London, UK, 2019, 3: 1-19

[174] Nguyen D-L, Arora S S, Wu Y, Yang H. Adversarial light projection attacks on face recognition systems: A feasibility study//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Virtual, 2020: 3548-3556

[175] Metzen J H, Kumar M C, Brox T, Fischer V. Universal adversarial perturbations against semantic image segmentation//Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice, Italy, 2017: 2774-2783

[176] Xu X, Zhang J, Li Y, et al. Adversarial attack against urban scene segmentation for autonomous vehicles. IEEE Transactions on Industrial Informatics, 2021, 17(6): 4117-4126

[177] Kang X, Song B, Du X, Guizani M. Adversarial attacks for image segmentation on multiple lightweight models. IEEE Access, 2020, 8(1): 31359-31370

[178] Ozbulak U, Van Messem A, Neve W D. Impact of adversarial examples on deep learning models for biomedical image segmentation//Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention. Granada, Spain, 2019: 300-308

[179] Chen L, Bentley P, Mori K, et al. Intelligent image synthesis to attack a segmentation CNN using adversarial learning//Proceedings of the Simulation and Synthesis in Medical Imaging 4th International Workshop. Shenzhen, China, 2019, 11827: 90-99

[180] Zheng Z, Zheng L, Yang Y, Wu F. Query attack via opposite-direction feature: Towards robust image retrieval. arXiv preprint arXiv:1809.02681, 2018

[181] Li J, Ji R, Liu H, et al. Universal perturbation attack against image retrieval//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 4898-4907

[182] Li X, Li J, Chen Y, et al. QAIR: Practical query-efficient black-box attacks for image retrieval//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 3330-3339

[183] Cheng D, Gong Y, Zhou S, et al. Person re-identification by multi-channel parts-based CNN with improved triplet loss function//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA, 2016: 1335-1344

[184] Bai S, Li Y, Zhou Y, et al. Adversarial metric attack and defense for person re-identification. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(6): 2119-2126

[185] Wang H, Wang G, Li Y, et al. Transferable, controllable, and inconspicuous adversarial attacks on person re-identification with deep mis-ranking//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 339-348

[186] Ding W, Wei X, Ji R, et al. Polynomial universal adversarial perturbations for person re-identification//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy, 2020: 1144-1151

[187] Wang Z, Zheng S, Song M, et al. advPattern: Physical-world attacks on deep person re-identification via adversarially transformable patterns//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 8340-8349

[188] Lee K-H, Hwang J-N. On-road pedestrian tracking across multiple driving recorders. IEEE Transactions on Multimedia, 2015, 17: 1429-1438

[189] Wiyatno R, Xu A. Physical adversarial textures that fool visual object tracking//Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Republic of Korea, 2019: 4821-4830

[190] Liang S, Wei X, Yao S, Cao X. Efficient adversarial attacks for visual object tracking//Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12371: 34-50

[191] Yan B, Wang D, Lu H, Yang X. Cooling-shrinking attack: Blinding the tracker with imperceptible noises//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 987-996

[192] Jia Y, Lu Y, Shen J, et al. Fooling detection alone is not enough: Adversarial attack against multiple object tracking//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020: 1-15

[193] Chen X, Yan X, Zheng F, et al. One-shot adversarial attacks on visual tracking with dual attention//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA, 2020: 10173-10182

[194] Guo Q, Xie X, Juefei-Xu F, et al. SPARK: Spatial-aware online incremental attack against visual tracking//Proceedings of the 16th European Conference of Computer Vision. Glasgow, UK, 2020, 12370: 202-219

[195] Jia S, Song Y, Ma C, Yang X. IoU attack: Towards temporally coherent black-box adversarial attack for visual object tracking//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual, 2021: 6709-6718

[196] Koh P W, Liang P. Understanding black-box predictions via influence functions//Proceedings of the International Conference on Machine Learning. Singapore, 2017: 1885-1894

[197] Yan Z, Wu J, Li G, et al. Deep neural backdoor in semi-supervised learning: Threats and countermeasures. IEEE Transactions on Information Forensics and Security, 2021, 16(1): 4827-4842

[198] Li G L, Ota K, Dong M X, et al. DeSVig: Decentralized swift vigilance against adversarial attacks in industrial artificial intelligence systems. IEEE Transactions on Industrial Informatics, 2019, 16(5): 3267-3277

[199] Hu S, Liu X, Zhang Y, et al. Protecting facial privacy: Generating adversarial identity masks via style-robust make-up transfer//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15014-15023

[200] Xu H, Liu X, Li Y, et al. To be robust or to be fair: Towards fairness in adversarial training//Proceedings of the 38th International Conference on Machine Learning. Virtual, 2021, 139: 11492-11501

WANG Zhi-Bo, Ph. D., professor, Ph. D. supervisor. His research interests include Internet of Things, artificial intelligence security, data security and privacy protection.

WANG Xue, M. S. candidate. Her research interest is artificial intelligence security.

MA Jing-Jing, M. S. candidate. Her research interest is artificial intelligence security.

QIN Zhan, Ph. D., professor, Ph. D. supervisor. His research interests include artificial intelligence security, data security and privacy protection.

REN Ju, Ph. D., associate professor, Ph. D. supervisor. His research interests include Internet of Things and network computing.

REN Kui, Ph. D., professor, Ph. D. supervisor. His research interests include Internet of Things security, data security and privacy protection, artificial intelligence security.

Background

The development of science and technology makes artificial intelligence closer and closer to human life. Adversarial example attack is a hot issue in the field of machine learning security. More and more attention has been paid to the problem of adversarial examples. The reasons for the emergence of the adversarial examples and the way of generation are the key problems in the study of the adversarial examples.

This article summarizes the characteristics, generation, and attack methods in computer vision tasks, namely image classification, object detection, face recognition, semantic segmentation, image retrieval, and visual tracking. At the end of this paper, the future research direction of adversarial example attack has prospected.