

# 物理对抗补丁攻击与防御技术研究综述

邓欢, 黄敏桓, 李虎, 王彤, 况晓辉

军事科学院 系统工程研究院 信息系统安全技术重点实验室 北京 中国 100101

**摘要** 以深度学习为代表的人工智能技术在经济、社会各领域中的应用越来越广泛,但与之相伴的安全性问题也逐渐凸显。深度学习作为概率模型所具备的不确定性,以及参数量大所带来的黑盒性质,使其容易受到对抗样本的攻击,这给基于深度神经网络的现实世界应用带来了严重的安全威胁。因此,对抗样本研究成为人工智能安全领域的一个热门方向。其中,对抗样本攻击主要指对深度学习模型的输入数据添加一些微小的扰动,使得模型对输入数据的预测产生错误。而物理对抗补丁攻击则是一种在物理世界中添加对抗性图像贴纸的攻击方式,可以通过将物理对抗补丁手动贴在实际场景中的目标物体上,使得深度学习在图像识别、目标检测等计算机视觉任务中无法正确识别目标物体,出现错误判断。早期研究主要聚焦于数字空间中对抗样本的构造,通过对数字化样本特征的局部或全局修改来实现扰动的添加,后研究人员利用数字空间中生成的对抗样本映射到物理世界中进行攻击。随着人工智能技术在现实世界的广泛应用,物理空间中的对抗样本攻击与防御技术渐受关注。以计算机视觉任务为基础,聚焦物理空间,围绕对样本特征进行局部修改的物理对抗补丁生成技术,对物理对抗补丁攻击与防御技术进行综述。本文从不同维度梳理分析物理对抗补丁攻击的类型,详细对比分析物理对抗补丁在图像识别、目标检测和其他计算机视觉任务中的攻击方法,并总结了针对物理对抗补丁攻击的防御方法,后对未来的研究方向进行展望。

**关键词** 对抗样本; 深度学习; 物理对抗补丁; 人工智能安全

中图法分类号 TP391 DOI号 10.19363/J.cnki.cn10-1380/tn.2025.01.06

## A Review on Physical Adversarial Patch Attacks and Defenses Techniques

DENG Huan, HUANG Minhuan, LI Hu, WANG Tong, KUANG Xiaohui

National Key Laboratory of Science and Technology on Information System Security, Institute of System Engineering, Academy of Military Sciences, Beijing 100101, China

**Abstract** Artificial intelligence technology, represented by deep neural networks, is being increasingly applied across various economic and social sectors. However, the concomitant security issues are also becoming prominent. The inherent uncertainty of deep neural networks as probabilistic models, coupled with their black-box nature due to the large number of parameters, make them susceptible to adversarial example attacks, posing severe security threats to real-world applications based on deep neural networks. Therefore, research on adversarial examples has become a hot topic in the field of AI security. Specifically, adversarial example attacks mainly involve adding minute perturbations to the input data of the deep neural network models, leading to incorrect predictions. Physical adversarial patch attacks, on the other hand, involve attaching adversarial image stickers in the physical world, manually affixing physical adversarial patches onto target objects in real-world scenarios, thereby causing the deep neural networks to fail in accurately recognizing the target objects in computer vision tasks such as image recognition and object detection, leading to incorrect judgments. Early research focused primarily on the construction of adversarial examples in the digital space, adding perturbations by locally or globally modifying the digitized example features. Subsequently, researchers used adversarial examples generated in the digital space for attacks in the physical world. With the widespread application of AI technology in the real world, the focus is gradually shifting towards the attack and defense techniques of adversarial examples in the physical space. Based on computer vision tasks, focusing on the physical space and around the generation techniques of physical adversarial patches through local modifications to example features, this paper reviews the attack and defense techniques of physical adversarial patches. We systematically analyze the types of physical adversarial patch attacks from different dimensions, provide detailed comparative analyses of the attack methods of physical adversarial patches in image recognition, object detection, and other computer vision tasks, and summarize the defense methods against physical adversarial patch attacks. We then provide a perspective on the future research directions.

**Key words** adversarial examples; deep neural network; physical adversarial patch; artificial intelligence security

通讯作者: 黄敏桓, 博士, 研究员, Email: darbean@126.com。

本课题得到重点实验室基金(No. 6142111220501)、智强基金资助。

收稿日期: 2023-03-15; 修改日期: 2023-06-02; 定稿日期: 2024-11-18

## 1 引言

近年来,以深度神经网络为代表的人工智能技术蓬勃发展,已广泛应用于计算机视觉、自然语言处理、语音语义、自动驾驶等多个领域,在学术界和工业界产生了广泛而深远的影响,极大地改变了人们的生产与生活方式。人工智能技术虽能够赋能各行各业,提升人们工作生活效率,但其也存在安全缺陷,容易被误导或欺骗。

对人工智能模型的误导和欺骗主要通过对抗样本来实现,即通过在原始样本上添加扰动的方式生成新的具有对抗性的样本。对抗样本既可以单纯在数字空间修改生成,也可以在物理世界构造生成。相关研究早期主要关注数字空间的对抗样本生成,如 Szegedy 等人<sup>[1]</sup>发现在数字图像分类任务中,深度神经网络很容易受到对抗样本攻击,即在原始数字图像中添加人眼几乎不可感知的扰动,可使深度神经网络模型以较高置信度输出错误的分类结果。此后研究人员相继提出了多种数字空间对抗样本生成方法,可生成图像、视频、语音等各领域的对抗样本。

数字空间的对抗样本可以映射到物理世界中,但其攻击效果往往并不尽如人意。原因在于数字空间对抗样本的扰动难以被打印机无色差地打印,且无法被摄像头无像素损失地拍摄后转换为数字信号。考虑到很多人工智能模型最终需要部署到物理世界中,物理世界中的对抗样本生成与防御在近年来越加受到关注。与数字空间生成对抗样本类似,在物理世界构造对抗样本既可以从全局角度添加扰动,也可以只针对局部进行修改。纵观已有研究,在局部添加扰动以欺骗智能模型的物理对抗补丁技术发展较快。

以计算机视觉任务为例,物理对抗补丁通常是指在物理空间对图像的局部区域施加可察觉的连续对抗扰动,最终能成功欺骗图像分类器<sup>[2]</sup>。物理对抗补丁通常是一个图案化的子图像,覆盖在原图像的局部区域上,打印后以海报、贴纸、眼镜等多种形式在物理世界中实现对抗攻击。物理对抗补丁可用于攻击图像分类、目标检测、语义分割等计算机视觉任务智能模型,从而对依靠此类模型进行智能决策的自动驾驶、遥感观测、军事目标识别等任务带来安全风险。

物理对抗补丁拓展了针对人工智能模型的攻击面,也为人工智能模型的安全应用提出了新要求。围

绕物理对抗补丁的相关学术论文数量在近些年快速增长,如图 1 所示,针对深度神经网络模型的物理对抗补丁攻击与防御技术成为了当前的研究热点。

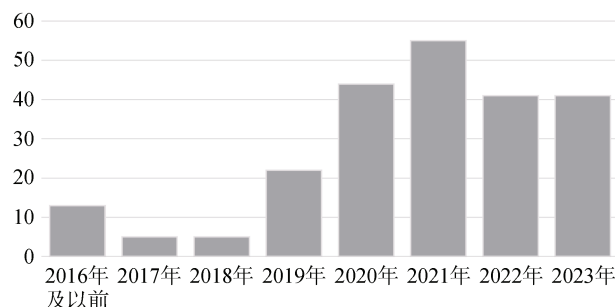


图 1 物理对抗补丁论文发表数量趋势图<sup>①</sup>

Figure 1 Rend graph of published papers on physical adversarial patches

近年来已有若干针对深度神经网络对抗攻击与防御的综述论文,如 Chakraborty 等人<sup>[3]</sup>讨论了数字空间中不同类型的对抗攻击和各种威胁模型,并对数字空间对抗攻击的效率和面临的挑战进行了分析;Yuan 等人<sup>[4]</sup>总结了数字空间的对抗样本攻击和防御方法,涉及少量物理对抗样本攻击方法;Qian 等人<sup>[5]</sup>总结了模式识别中的各类鲁棒对抗训练方法。已有的综述研究主要集中在数字空间对抗样本的攻击与防御,较少涉及物理空间的对抗攻击与防御。

本文以计算机视觉任务为例,对物理对抗补丁攻击与防御技术进行综述。第 1 节引言部分简述了物理对抗补丁的研究背景与意义;第 2 节从不同维度梳理分析了物理对抗补丁攻击的分类;第 3 节详细对比分析了物理对抗补丁在图像识别、目标检测和其他计算机视觉任务中的攻击方法;第 4 节总结分析了针对物理对抗补丁的防御方法;第 5 节对物理对抗补丁相关的研究工作进行了总结和展望。

## 2 对抗攻击分类

对抗攻击既可以在数字空间以对抗样本的形式进行,也可以在物理空间以物理对抗补丁的形式进行。对抗样本与物理对抗补丁既相互关联,又有所区分。

### 2.1 对抗样本攻击

对抗样本(Adversarial example)<sup>[1]</sup>通常是指在原始图像中添加细微扰动后能够使深度神经网络模型发生误判的输入样本。对抗扰动通常在图像全局或局部添加,且所添加的对抗扰动一般不会被肉眼所

① 以“physical adversarial patch”为关键词在谷歌学术、百度学术等搜索引擎检索得到此数据。

察觉。以对抗样本作为手段对深度神经网络模型进行攻击的方法多样,大致分类如图2所示。

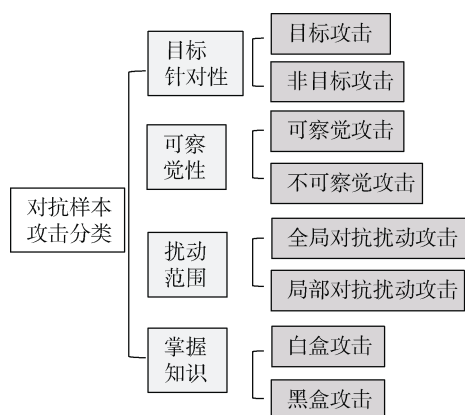


图2 对抗样本攻击分类

Figure 2 Adversarial example attacks in classification

从攻击的目标针对性角度可将对抗样本攻击分为目标攻击(Targeted attack)与非目标攻击(Untargeted attack)。目标攻击是指攻击者能够误导目标模型使其输出预先设定的目标类别;非目标攻击是指攻击者能够误导目标模型使其输出除正确类别外的其他类别。当输入数据与预先设定的目标类别之间的决策边界距离较远时,实现目标攻击难度较大,通常目标攻击相比非目标攻击更具挑战性。

从攻击的可察觉性角度可将对抗样本攻击分为可察觉攻击(Perceptible attack)与不可察觉攻击(Imperceptible attack)。可察觉攻击是指将对抗扰动添加进原始图像后能被人所察觉,一般不对扰动幅度做严格限制,以提高攻击的成功率;不可察觉攻击是指将对抗扰动添加进原始图像后不能被人所察觉,一般会严格限制扰动幅度,以实现攻击的隐蔽性。

从扰动的添加范围角度可将对抗样本攻击分为全局扰动攻击(Global perturbation attack)和局部扰动攻击(Local perturbation attack)。全局扰动攻击是指对图像的全局像素进行扰动攻击,一般限制扰动幅度,对抗扰动在图像上的分布较为均匀;局部扰动攻击是指对图像局部区域内的像素进行扰动攻击,一般不限制扰动幅度,对抗扰动在图像上的分布较为集中。

从攻击者是否掌握目标模型的知识角度可将对抗样本攻击分为白盒攻击(White-box attack)和黑盒攻击(Black-box attack)。白盒攻击假设攻击者掌握目标模型的结构和参数,可基于目标模型知识来构建对抗样本,通常可进一步细分为基于梯度的方法和基于优化的方法,分别从不同的维度利用目标模型的知识;黑盒攻击假设攻击者不了解目标模型的内

部结构及参数,但可以与目标模型交互,通过提交查询请求到目标模型可获取输出,通常可根据交互的情况进一步细分为基于查询的方法和基于迁移的方法。此外,对于攻击者掌握的知识介于白盒和黑盒之间的情形,有时也称之为灰盒攻击(Grey-box attack)。典型白盒攻击和黑盒攻击方法如表1所示。

表1 白盒攻击和黑盒攻击方法

Table 1 White-Box attack and Black-Box attack methods

攻击类型	方法	攻击方法
		典型方法
白盒攻击	基于梯度的方法	FGSM <sup>[6]</sup> 、BIM&ILCM <sup>[7]</sup> 、PGD <sup>[8]</sup> 、Deep Fool <sup>[9]</sup> 、UAP <sup>[10]</sup> 等
	基于优化的方法	L-BFGS <sup>[11]</sup> 、JSMA <sup>[11]</sup> 、CW <sup>[12]</sup> 、AA <sup>[13]</sup> 等
黑盒攻击	基于查询的方法	OnePixel <sup>[14]</sup> 、ZOO <sup>[15]</sup> 、RAYS <sup>[16]</sup> 、遗传算法、进化算法、强化学习、粒子群优化和贪心策略等
	基于迁移的方法	MI-FGSM <sup>[17]</sup> 、集成攻击、生成对抗网络、元学习等

## 2.2 物理对抗补丁攻击

物理对抗补丁(Physical adversarial patch)<sup>[2]</sup>是指在图像的局部区域生成的连续像素块,打印后可粘贴在物理世界的目标对象上,进而能够欺骗智能模型将目标对象识别为攻击者指定的类别。物理对抗补丁通常只覆盖目标图像的局部区域,可放在图像中的任意位置,与具体的应用场景及目标对象无关,可实现较大范围的攻击,故得到了较多的研究。以计算机视觉任务为例,对抗补丁攻击的大致分类如图3所示。

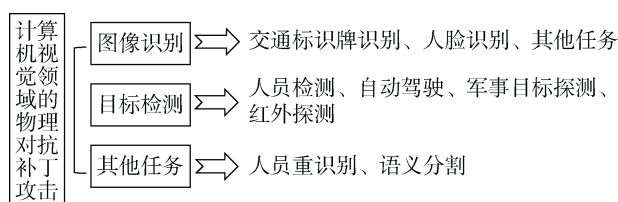


图3 计算机视觉领域的物理对抗补丁攻击任务

Figure 3 Physical adversarial patch attack in the field of computer vision

物理对抗补丁攻击同时涉及数字空间中对抗样本的生成和物理空间中对抗攻击的验证与实施,其大致流程如图4所示。

首先在数字空间生成对抗补丁。第1步,根据目标任务,在物理世界或仿真世界中采集样本集合A;第2步,生成初始化的对抗补丁 $P_0$ ,并将对抗补丁 $P_0$ 添加至样本集合A中每个样本上,形成对抗样本集S;第3步,将对抗样本集S中的每个样本输入至



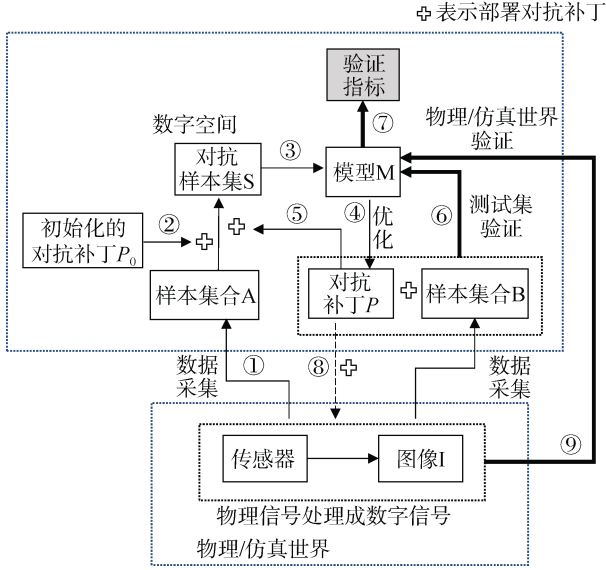


图 4 物理对抗补丁攻击流程

Figure 4 A process for physically adversarial patch attack resistance in physics

模型  $M$ ; 第 4 步, 根据模型  $M$  的输出结果及相关损失项, 优化对抗补丁  $P$ ; 第 5 步, 将优化的对抗补丁  $P$  添加至样本集合  $A$  中每个样本上, 形成新的对抗样本集, 根据攻击效果及设定的终止条件, 迭代进行第 3 步至第 5 步的各项操作。

一般先在数字空间用测试集(数据样本集合  $B$ )对生成的对抗补丁的攻击效果进行初始效果验证。第 6 步, 将添加过对抗补丁的测试样本集输入至模型  $M$ ; 第 7 步, 观察模型  $M$  的输出, 综合输入输出以及其他信息, 结合相关评价指标对攻击效果进行验证评估。

之后在物理世界或仿真世界对生成的对抗补丁的攻击效果进行验证。由于物理世界验证相对复杂, 成本高, 且可能存在伦理与安全风险, 因此, 很多验证工作会在仿真世界中进行。第 8 步, 将数字空间生成的对抗补丁打印粘贴到物理世界的目标对象上, 如以贴纸的形式粘贴到停车牌上; 利用传感器采集得到包含物理对抗补丁的目标图像  $I$ ; 第 9 步, 将采集得到的目标图像  $I$  输入至模型  $M$ ; 最后观察模型  $M$  的输出, 综合输入输出以及其他信息, 结合相关评价指标对攻击效果进行验证评估。

当物理对抗补丁的攻击效果达到预期后, 即可正式将其部署到真实目标对象上, 实现对智能模型的误导或欺骗。部署后的运行流程基本上与验证过程类似。

在实现物理对抗补丁攻击的整个过程中, 涉及到数字空间中生成的对抗补丁打印部署到物理空间, 然后又被传感器采集回到数字空间的多个环节。因

此, 物理对抗补丁攻击既要考虑到减少数字空间向物理世界映射时的信息损失, 又要尽可能降低物理世界向数字世界映射时的信息损失, 从而在物理对抗补丁的形式及部署方式、自然性、鲁棒性等之间取得平衡。

### 2.2.1 物理对抗补丁的形式及部署方式

物理对抗补丁的存在形式主要包括二维码、眼镜、妆容、海报贴纸、涂鸦、对抗服装、人造光源等。物理对抗补丁的部署方式主要包括打印粘贴、放置、穿戴、涂绘和光源投射等, 示例如图 5 所示。



图 5 物理对抗补丁形式及部署方式示例

Figure 5 Examples of physical adversarial patch forms and deployment methods

(原始图片来自于文献[22, 26, 28-30, 32, 44])

### 2.2.2 物理对抗补丁的自然性

物理对抗补丁是施加于图像局部、不严格限制扰动幅度的连续像素块, 容易被人眼所感知。因此, 如何使物理对抗补丁更加自然, 不易被人眼所察觉是当前的研究重点之一, 如针对无语义的对抗补丁图案易被察觉的问题, 部分研究尝试使用具有语义信息的图案贴纸, 如卡通贴纸、喷绘等。

通常可以采用风格损失、内容损失、平滑损失等方法提升物理对抗补丁的自然性。

风格损失(Style loss)<sup>[18]</sup>。以物理对抗补丁与指定风格参考图像之间的风格差异来度量自然性, 通过减小二者在图像风格上的距离来使得物理对抗补丁更自然。风格损失的距离定义为:

$$D_g = \sum_{l \in \mathcal{S}_l} \left\| \mathcal{G}(F_l(x^g)) - \mathcal{G}(F_l(x')) \right\|_2^2 \quad (1)$$

其中,  $\mathcal{S}_l$  为用于提取内容表示的风格层的集合,  $F(x)$  为特征提取器<sup>[19]</sup>,  $x^g$  为风格参考图像,  $x'$  为对抗样本,  $\mathcal{G}$  为  $F$  提取风格特征的 Gram 矩阵。

内容损失(Content loss)<sup>[18]</sup>。通过计算对抗样本在表示空间中与原始图像的相似性来度量自然性, 尽可能减小对抗样本相较于原始图像的内容损失。内

容损失定义为:

$$\mathcal{L}_c = \sum_{\ell \in \mathcal{C}_\ell} \left\| F_\ell(x) - F_\ell(x') \right\|_2^2 \quad (2)$$

其中,  $\mathcal{C}_\ell$  为用于提取内容表示的内容层的集合,  $F(x)$  为特征提取器,  $x$  为原始图像,  $x'$  为对抗样本。

TV 损失(Total variation loss)<sup>[20]</sup>。将对抗补丁叠加至原始样本图像后, 对抗补丁的边缘与原始图像通常对比明显, 人眼很容易察觉。可以通过减少图像相邻像素之间的变化来提高图像的平滑度, 降低对比度。TV 损失定义为:

$$\mathcal{L}_m = \sum \left( (x_{i,j} - x_{i+1,j})^2 + (x_{i,j} - x_{i,j+1})^2 \right)^{\frac{1}{2}} \quad (3)$$

其中,  $x_{i,j}$  是图像  $x$  在坐标  $(i,j)$  下的像素值。

### 2.2.3 物理对抗补丁的鲁棒性

对抗补丁从数字空间映射到物理空间时可能失效或效用降低, 如何提高其鲁棒性也是当前的研究重点之一。通常可以采用变换期望算法和不可打印得分损失。

变换期望 (Expectation over transformation, EOT)<sup>[21]</sup>。对抗补丁可能在物理世界中被随机旋转、平移或添加噪声, 甚至添加 3D 纹理, 此时, EOT 计算不同变换后对抗样本与原始样本之间有效距离的期望值, 以度量物理对抗补丁的鲁棒性, 定义为:

$$\delta = \mathbb{E}_{t \sim \mathcal{T}} \left[ d(t(x'), t(x)) \right] \quad (4)$$

其中,  $d(\cdot, \cdot)$  为距离函数,  $\mathcal{T}$  是选择的转换分布,  $t(x)$  为到转换分布  $\mathcal{T}$  上的变换,  $x$  为原始图像,  $x'$  为对抗样本。

不可打印得分(Non printability score, NPS)<sup>[22]</sup>。由于设备的限制, 打印机或屏幕能够再现的颜色范围只是 RGB 颜色空间的子集, 因此数字空间中生成的对抗补丁被部署至物理世界时将不可避免地出现失真。NPS 测量数字对抗扰动和打印出来的对抗扰动之间的颜色距离, 以确保两者颜色接近, 定义为:

$$\text{NPS}(\delta) = \sum_{\hat{p} \in \delta} \prod_{p' \in \mathcal{P}} |\hat{p} - p'| \quad (5)$$

其中,  $\delta$  为扰动向量,  $\mathcal{P}$  为可打印颜色向量集合,  $\hat{p}$  为  $\delta$  中的一个像素,  $p'$  为  $\mathcal{P}$  中的一个向量。

## 3 物理对抗补丁攻击

物理对抗补丁是当前人工智能安全领域的研究热点, 以计算机视觉领域为例, 其通过对深度神经网络模型在图像识别、目标检测、语义分割等任务的误导和欺骗, 能造成诸多安全性问题。

## 3.1 图像识别任务中的对抗补丁攻击

### 3.1.1 交通标识牌识别攻击

对抗补丁可以攻击交通标识牌而使深度神经网络模型产生误分类。如 Eykholt 等人<sup>[23]</sup>提出了一种通用的鲁棒物理扰动(Robust physical perturbations, RP2), 并以黑白贴纸的形式部署应用到现实世界的路标牌分类任务上, 在各种环境条件下(包括不同视角和距离)以高置信度实现了有目标攻击, 使路标牌被神经网络错误分类, 但是其设计的黑白贴纸形式的对抗补丁自然性较差。

后来 Liu 等人<sup>[24]</sup>设计具有感知敏感性的生成对抗网络(Perceptual-Sensitive generative adversarial networks, PS-GAN)同时提高了对抗补丁的攻击能力和自然性。PS-GAN 以现实生活中的涂鸦形式, 生成的对抗补丁与图像上下文具有较强的感知相关性, 同时在空间上位于被攻击图像的感知敏感位置, 具有较好的泛化能力和可移植性。此外, Kong 等人<sup>[25]</sup>基于 GAN 的思想设计了一种物理对抗补丁生成框架 PhysGAN, 通过将打印的对抗海报贴在路边广告牌上, 以实现自动驾驶系统的攻击, 使得自动驾驶汽车的方向盘发生危险转动。PhysGAN 中的生成器以 3D 张量作为输入, 通过编码器对真实世界的驾驶数据和对抗扰动进行编码, 以生成更具自然性的对抗性广告牌。

除了以贴纸的形式对交通牌进行攻击, 还能以“光”的形式进行物理对抗补丁攻击。如 Duan 等人<sup>[26]</sup>提出使用对抗激光束(Adversarial laser beam, AdvLaserBeam)方法来实现物理对抗攻击。其用一组参数表示激光束, 并用贪婪搜索得到最佳的激光束参数。为实现在数字空间生成的最佳激光束在物理空间精确再现, 其应用旋转、平移、噪声添加等转换方式以提高其攻击的鲁棒性。对抗激光束方法在静态物理环境中效果较好, 但其在动态物理环境中的效果有限。此外, 对抗激光束是人为制造的, 其自然性较差。

与人为制造激光束不同, Zhong 等人<sup>[27]</sup>在自然光照阴影的启发下提出了物理对抗阴影攻击方法。其将阴影简化为一个三角区域, 并使用粒子群优化(Particle swarm optimization, PSO)算法来搜索最佳的三角形顶点坐标。通过将图像从 RGB 空间转换到 LAB 空间, 考虑与亮度相关的 L 通道, 统计 LAB 三通道像素值的平均比率来模拟真实阴影, 最后转换回到 RGB 空间, 并采用 EOT 算法提升对抗攻击在物理世界的鲁棒性。其研究表明自然光照阴影会对深度神经网络模型造成影响, 尤其是在大强度单一光

源场景下, 带来的安全风险较大。

### 3.1.2 人脸识别攻击

人脸识别攻击可以通过眼镜、妆容、贴纸等多种形式的对抗补丁进行。如 Sharif 等人<sup>[22]</sup>提出使用对抗眼镜对人脸识别模型进行了对抗攻击。为了提高物理世界攻击的鲁棒性, 其采用 TV 损失来度量对抗扰动的平滑性, 使用 NPS 损失优化打印颜色与实际颜色之间的距离。对抗眼镜能实现比较好的攻击效果但是其外观比较显眼。Zhu 等人<sup>[28]</sup>提出使用 GAN 网络来实现带妆对抗攻击。其 GAN 网络包括妆容转换子网络和对攻击子网络。其中, 妆容转换子网络将非化妆人脸图像转换成高质量带妆人脸图像, 对抗攻击子网络将对抗攻击信息隐藏到带妆人脸图像中。妆容对抗补丁将攻击信息隐藏在眼部区域, 对抗攻击不仅有较高的攻击效果, 攻击行为也更隐蔽更自然。

Wei 等人<sup>[29]</sup>提出使用现实生活中真实存在的贴纸而不是人为设计的扰动来实现对抗补丁攻击。其通过优化贴纸的粘贴位置和旋转角度模拟人脸的变形来执行物理攻击, 发现对抗成功的位置呈现聚集现象, 于是设计了基于区域的启发式差分进化算法来寻找最合适的攻击区域和旋转角度等攻击参数, 具体较好的鲁棒性和自然性。

除了妆容和穿戴形式的对抗补丁攻击, Nguyen 等人<sup>[30]</sup>还提出使用对抗光影投射来攻击人脸识别系统。该方法需要进行位置校准和颜色校准两个关键步骤, 其中, 位置校准保证对抗模式能投影到关键攻击区域, 颜色校准保证对抗图案能被投影仪高保真地投影到目标人脸上。实验证明光影可实现对人脸识别系统的攻击, 但在现实世界实际应用的条件比较苛刻。

### 3.1.3 其他攻击

除了上面几类典型的图像识别任务, 部分研究也针对其他任务展开。如 Brown 等人<sup>[2]</sup>提出了约束条件下的伪装补丁生成方法, 以使得最终生成的对抗补丁和初始对抗补丁之间具有更高相似性。但此类对抗补丁通常尺寸较大, 人眼更易察觉。Karmon 等人<sup>[31]</sup>提出了局部可见的对抗性噪声方法(Localized and visible adversarial noise, LaVAN), 其在对抗补丁攻击中引入所谓“隐身特性”, 生成的对抗补丁尺寸较小, 且能在不同图像和补丁放置位置上迁移。Duan 等人<sup>[32]</sup>尝试将对抗补丁伪装成二维码的形式, 其方法会损失一定的攻击效果, 但提高了在物理世界中部署的灵活性和自然性。

为提升物理对抗补丁的自然性, 部分研究采用

风格迁移和生成对抗网络的方式生成对抗补丁。如 Duan 等人<sup>[18]</sup>提出了一种对抗伪装(Adversarial camouflage, AdvCam)方法。AdvCam 使用风格迁移技术将物理世界中的对抗补丁伪装成各类自然风格, 但需要攻击者手动指定攻击区域和目标样式。

Doan 等人<sup>[33]</sup>利用 GAN 的思想提出了从自然图像中寻找物理对抗图像的方法。该方法假设存在自然对抗样本, 即无需人为添加扰动即可实现攻击的自然图像。因而使用 GAN 来学习自然的图像分布, 然后固定生成器参数, 依据对抗损失在其分布中搜寻具有对抗效果的图像。而为了提升对抗补丁的鲁棒性, Gittings 等人<sup>[34]</sup>提出了一种使用深度图像先验(Deep image prior, DIP)的图像重建技术以产生对仿射变形具有鲁棒性的不可感知扰动, 以实现在整个图像上添加更大的扰动。Zhou 等人<sup>[35]</sup>提出了一种数据独立对抗补丁(Data-independent adversarial patch, DiAP)方法, 在不知道训练数据的情况下生成对抗补丁。

交通标识牌攻击和人脸识别攻击是图像识别物理对抗攻击的典型任务。在此类攻击中, 提升对抗补丁的自然性能有效提升攻击的隐蔽性。为此, PS-GAN<sup>[24]</sup>、PhysGAN<sup>[25]</sup>、Adv-makeup<sup>[28]</sup>、TnT<sup>[33]</sup>等方法使用 GAN 来优化对抗噪声, 但可能会降低对抗攻击效果。AdvLaserBeam<sup>[26]</sup>、Shadows<sup>[27]</sup>、Adv-Light<sup>[30]</sup>等方法以“光”的形式进行对抗攻击, 使得攻击更自然, 但需要精确地部署对抗补丁。Adv-Sticker<sup>[29]</sup>方法以生活中自然存在的贴纸作为对抗补丁, 欺骗效果较好, 但需要找到合适的攻击位置和角度等信息。这些方法各有优劣, 在不同应用场景下攻击效果各异。

表 2 总结了前文所述的图像识别任务中的物理对抗补丁攻击方法。

## 3.2 目标检测任务中的对抗补丁攻击

### 3.2.1 人员检测攻击

随着深度神经网络模型在人员检测任务中的广泛应用, 针对性的对抗补丁攻击也越来越受关注。如 Liu 等人<sup>[36]</sup>提出了一种基于迭代训练的对抗性补丁生成方法 DPatch, 用于攻击目标检测器。DPatch 根据对抗补丁的训练方式可以执行非目标攻击和目标攻击, 并具有良好的跨数据攻击迁移性。但 DPatch 局限在 RGB 像素的允许范围内, 对抗补丁图案没有剪裁, 在物理世界的攻击效果有限。Lee 等人<sup>[37]</sup>在 DPatch 的基础上进行优化, 扩展了 DPatch 在物理世界中攻击目标检测器的能力。该方法对于不同的照明条件、位置或变换具有一定的鲁棒性, 但是随着距离的增加, 对抗补丁的攻击效果会减弱。

表 2 图像识别任务中的物理对抗补丁攻击方法汇总

Table 2 A Summary of physical adversarial patch attack methods in image recognition tasks

任务	攻击方法	攻击形式	黑盒/白盒	攻击验证方式	用到的数据集
交通标识牌识别	RP2 <sup>[23]</sup>	贴纸	白盒	打印贴纸粘贴在交通牌上进行攻击验证	LISA、GTSRB
	PS-GAN <sup>[24]</sup>	涂鸦	黑盒、白盒	打印涂鸦图片粘贴在交通标识牌上进行攻击验证	GTSRB、QuickDraw
	PhysGAN <sup>[25]</sup>	广告路牌	白盒	打印粘贴在路边广告牌上进行验证	UADCC、DAVE-2、Kitti、Custom
	AdvLaserBeam <sup>[26]</sup>	激光束	黑盒	在数字空间寻找最佳激光束照射方位并在物理世界验证	ImageNet
	Shadows <sup>[27]</sup>	自然光照阴影	黑盒	通过太阳光照射在路标牌上形成的自然阴影进行验证	LISA、GTSRB
人脸识别	Adv-Glass <sup>[22]</sup>	眼镜	白盒	打印眼镜并佩戴在人脸上进行攻击验证	Google Images
	Adv-makeup <sup>[28]</sup>	妆容	白盒	通过 GANs 生成带有攻击信息的带妆人脸图像进行验证	youtube 收集带妆图像和非带妆图像
	Adv-Sticker <sup>[29]</sup>	贴纸	黑盒	贴纸贴在人脸上验证	现实生活中存在的贴纸、LFW、CelebA
	Adv-Light <sup>[30]</sup>	投影仪投影	黑盒、白盒	用投影仪将对抗光影投影到人脸上进行验证	50 名实验者、LFW
	Adversarial patch <sup>[2]</sup>	贴纸	黑盒、白盒	打印贴纸放在目标对象旁边进行攻击验证	ImageNet
其他任务	LaVAN <sup>[31]</sup>	数字图案补丁	白盒	生成数字图案补丁在数字空间进行攻击验证	ImageNet
	QR patch <sup>[32]</sup>	二维码	白盒	生成二维码图案补丁在数字空间进行攻击验证	ImageNet
	AdvCam <sup>[18]</sup>	对抗伪装	灰盒	数字空间验证、物理世界验证	ImageNet、拍摄照片
	TnT <sup>[33]</sup>	花朵图像补丁、贴纸	白盒、黑盒	数字图像添加补丁验证、打印对抗补丁贴在衣服上验证	Google Images、ImageNet、PubFig、CIFAR10、GTSRB
	Local patch via DIP <sup>[34]</sup>	数字图案补丁	白盒	生成数字图案补丁在数字空间进行攻击验证	ImageNet
	DiAP <sup>[35]</sup>	贴纸	黑盒	打印贴纸在物理世界进行攻击验证	ImageNet

部分研究通过在脸部粘贴对抗补丁来规避检测。如 Pautov 等人<sup>[38]</sup>采用投影变换(Projective transformation)技术生成对抗补丁, 然后将其打印成贴纸或眼镜的形式, 贴在脸部特定区域。Xiao 等人<sup>[39]</sup>使用低维流形正则化技术和对抗生成模型来增加对抗补丁的迁移能力, 同样将打印出来的对抗补丁粘贴在眼部区域来实现攻击。Komkov 等人<sup>[40]</sup>提出了一种将对抗补丁打印成长方形的贴纸, 然后粘贴在帽子前额区域上的攻击方法。其可攻击公共人脸识别系统 LResNet100E-IR 并且具有一定的迁移性, 可攻击其他人脸识别系统。

部分研究通过打印对抗补丁粘贴在纸板或裁剪至衣物上来使人规避检测, 如图 6 所示。Thys 等人<sup>[41]</sup>通过将打印好的对抗补丁放在人前面以规避摄像机检测器的检测。该方法通过最小化检测器输出的对象或类分数来隐藏图像中的人, 但人员位置要求较高, 可迁移性较差。将对抗补丁打印出来贴在纸板上使人规避检测的方法在刚性物体或平面物体上的效果较好, 但对于柔性物体或曲面物体, 如衣服、帽子

等, 将对抗补丁打印并粘贴在其上时, 其容易产生形变, 很多研究围绕这一问题展开。如 Huang 等人<sup>[42]</sup>提出了一种通用物理伪装(Universal physical camouflage, UPC)方法来在物理世界中攻击目标检测器。UPC 采用模拟变换来使得其对柔性或曲面物体都有效, 且具有一定的迁移性。Wu 等人<sup>[43]</sup>设计了在不同系统结构、不同类别和不同数据集之间可迁移的物理对抗补丁生成方法, 打印后在物理世界的 3D 物体上仍能保持对抗性。其应用增强变化, 包括亮度、对比度、旋转、平移和剪切变换的组合来模拟物理世界中的变换以增加攻击的鲁棒性, 并采用 Thin-Plate-Spline(TPS)<sup>[44]</sup>变换模拟衣服上的随机褶皱, 并引入 TV 损失以保持平滑性, 最后通过集成多个目标检测器模型以提高对抗补丁的迁移性。Xu 等人<sup>[45]</sup>设计了一种对抗 T 恤, 即使对抗 T 恤可能由于人的移动而发生姿势的变化, 进而引起 T 恤发生非刚性变形, 其仍能在物理世界中实现攻击。但一件衣服上的一块对抗补丁很难从多个视角攻击检测器, 检测器可能只识别到严重变形对抗补丁的一部分, 因而只在



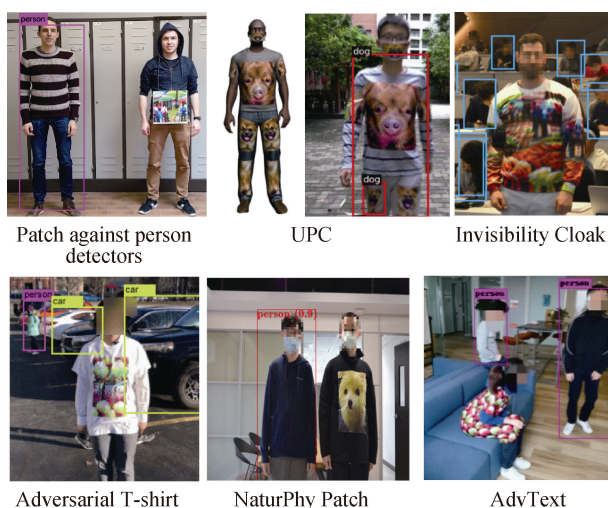


图 6 衣物类物理对抗补丁示例

Figure 6 Examples of physical adversarial patches for clothing items

(原始图片来自于文献[42-44, 46-48])

对抗补丁直面摄像头时效果更佳。Hu 等人<sup>[46]</sup>利用预训练生成对抗网络来学习真实世界图像, 通过遍历潜在向量空间来生成自然逼真的对抗补丁。此外, Hu 等人<sup>[47]</sup>提出具有任意大小且任意局部都具有对抗效果的对抗纹理(Adversarial texture, AdvTexture)方法, 通过将对抗纹理覆盖在衣服上, 人穿上之后能以任何角度规避检测。

### 3.2.2 自动驾驶攻击

自动驾驶车辆一旦受到物理对抗补丁的攻击, 其内置的模型可能出现决策失误, 进而带来极大的安全风险。目前已有部分研究针对此问题展开。部分研究证明自动驾驶车辆对于交通标志牌的检测容易受到物理对抗补丁攻击。如 Song 等人<sup>[48]</sup>提出了一种用于目标检测的对抗补丁生成方法并在物理世界中用贴纸的形式攻击了停车标志牌。Chen 等人<sup>[49]</sup>使用 EOT 算法来生成鲁棒的物理对抗补丁以对停车标志检测模型进行攻击。其限制了对抗补丁的形状, 并生成高置信度扰动和低置信度扰动, 其中高置信度扰动更鲁棒但是更明显, 可有效攻击 Faster R-CNN 目标检测系统。

当摄像机相对于受攻击图像的相对位置发生变化时, 静态固定的对抗补丁的有效性会显著降低, 原因在于, 一是摄像机角度因相对运动而改变, 二是摄像机视角的变化会导致目标对象的大小发生变化。为此, Hoory 等人<sup>[50]</sup>提出了动态对抗补丁(Dynamic adversarial patch, DAP)方法, 根据当前摄像头的位置动态调整对抗补丁, 使动态对抗补丁对摄像机的相对位置基本不变。此外, 当摄像机的视角

发生变化或现场有多个摄像机时, 会在不同的位置放置多个屏幕来攻击检测器。

部分研究通过在自动驾驶车身上涂装纹理使其规避检测, 如图 7 所示。如 Zhang 等人<sup>[51]</sup>设计了一种对抗伪装模式来攻击目标检测器, 以防止车辆被检测到。其通过模拟对抗伪装渲染在车辆上的成像过程和目标检测器检测涂有对抗伪装车辆的过程, 搜索最佳伪装模式。Huang 等人<sup>[42]</sup>提出的通用物理伪装方法(Universal Physical Camouflage, UPC)也能攻击对汽车的检测, 其生成的对抗伪装模式带有一定的语义信息, 相对自然、不突兀, 在物理空间有较好的攻击效果, 但在数字空间的效果相对较差。Wang 等人<sup>[52]</sup>提出了双重注意力抑制攻击(Dual attention suppression, DAS)方法, 通过抑制不同模型之间共享的注意力来提高攻击的可迁移性, 通过引入与场景上下文具有感知相关性的种子内容补丁, 生成与场景上下文具有语义相关的物理对抗伪装, 使得对抗伪装更加自然。其通过 Carla 模拟器采集数据来优化对抗纹理, 并通过打印对抗伪装粘贴在玩具车上进行了物理世界的攻击实验。但是生成的对抗伪装仅覆盖在车辆表面的部分区域, 无法在多视角、远距离和有部分遮挡物的情况下进行攻击。Wang 等人<sup>[53]</sup>提出了一种全覆盖的车辆对抗伪装(Full-coverage vehicle camouflage, FVC)攻击方法。其通过可微分神经渲染器将对抗纹理渲染到车辆的整个表面, 弥合数字空间和物理空间之间攻击的差距, 通过转换函数将渲染后的图像转换到不同的物理环境场景中, 实验表明其在复杂环境中较好的攻击效果。Duan 等人<sup>[54]</sup>提出涂层对抗伪装(Coated adversarial camouflages, CAC)来对任意角度的车辆检测器进行攻击。其通过优化 3D 模型中的 2D 纹理, 并在每次迭代中使用一组密集候选框来优化对抗伪装, 使得用 CAC 方法生成的对抗伪装可以从任意角度攻击目

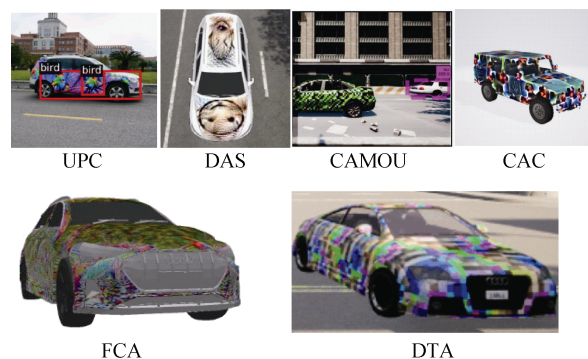


图 7 车辆纹理涂装示例

Figure 7 Examples of vehicle texture painting

(原始图片来自于文献[43, 52-56])



标检测器,但是生成的对抗伪装不自然,很容易被人眼察觉。Suryanto 等人<sup>[55]</sup>提出可微分变换攻击(Differentiable transformation attack, DTA)来提升攻击的鲁棒性。其使用可微分变换网络(Differentiable transformation network, DTN)来近似阴影效应,渲染了挡风玻璃的透明度以及汽车底部和顶部的阴影,使得加上对抗伪装后的车辆图像更加真实。

### 3.2.3 军事目标探测攻击

物理对抗补丁在军事领域的应用也得到了部分研究者的关注,如 Adhikari 等人<sup>[56]</sup>扩展了 Thys 等人<sup>[41]</sup>的工作,将相对较小的对抗补丁应用在飞机上以躲避来自空中的检测,并通过修改损失函数,使对抗补丁更难被人眼察觉。其根据经验提出了对抗补丁大小、数量和性能之间的权衡关系,但对抗补丁大小和飞机尺寸之间容易出现不匹配问题。Lu 等人<sup>[57]</sup>提出根据受攻击飞机的大小自适应调整对抗补丁大小,并生成了可攻击不同大小飞机的通用对抗补丁。Wise 等人<sup>[58]</sup>在 Adhikari 等人<sup>[56]</sup>的工作基础上,通过最大化物体检测损失,同时限制了补丁的颜色生成了更不易被察觉的对抗补丁。Løkken 等人<sup>[59]</sup>通过 GAN 生成对抗伪装并应用在图像上,攻击了对海军舰艇分类的 DNN 分类器。

### 3.2.4 红外探测攻击

红外探测系统广泛应用于夜间目标探测、体温测量、自动驾驶等领域,随着深度神经网络在红外探测领域的应用,相应的安全风险也逐步加大。

针对红外探测的物理对抗补丁攻击也得到了部分研究者的关注。如 Zhu 等人<sup>[60]</sup>提出了一种攻击红外热成像人员检测器的物理对抗补丁生成方法。其通过在木板上放置灯泡来模拟热红外,首先测量热特性与红外相机捕获图片之间的关系,通过调节灯泡的亮度来调整红外成像模式,在数字空间中使用高斯函数优化灯泡放置位置,最后通过在木板上对应的优化位置放置灯泡,来实现物理世界的攻击,并通过模型集成来提高攻击的可迁移性。Zhu 等人<sup>[61]</sup>还提出了使用气凝胶制作的红外对抗服来规避红外行人检测器。其模拟二维码对抗补丁<sup>[32]</sup>的攻击形式设计了对抗性的二维码补丁,采用 AdvTexture<sup>[47]</sup>类似思想优化对抗补丁,设计了可以周期性扩展的图案使得任意局部都有对抗效果,并提出使用气凝胶来制作红外对抗服。同样采用了模型集成技术来提高对未知模型的攻击迁移能力。

上述两种使用灯泡和气凝胶制作红外对抗服的攻击方法在物理世界实现攻击时部署方式相对不自然,易引人注目。为此,Wei 等人<sup>[62]</sup>提出使用一种新

的对抗介质来实现物理世界的红外对抗方法(HOTCOLD Block),其通过使用一种可穿戴的热敏感材料发热贴和制冷贴粘贴在衣物上,在数字空间中最小化对抗补丁的大小,使用 PSO 优化算法搜索最佳的对抗补丁形状和攻击位置,来误导红外探测器。

目标检测任务中的物理对抗补丁攻击研究较多。对于人员检测攻击,ProjectTrans<sup>[38]</sup>、TAPs<sup>[39]</sup>、AdvHat<sup>[40]</sup>采用类似于人脸识别攻击的思路,使用打印贴纸、对抗眼镜、帽子等形式进行攻击。然而在非刚性物体上粘贴对抗贴纸会导致形变,从而降低攻击效果。为此,Invisibility Cloak<sup>[43]</sup>、Adversarial T 恤<sup>[45]</sup>等攻击方法设计了对抗 T 恤,将对抗补丁打印至衣物,在非刚性物体上攻击效果较好,但在变换角度后攻击效果显著下降。AdvText<sup>[47]</sup>方法可实现多角度攻击。对于自动驾驶攻击,已有研究聚焦于在车身涂装对抗纹理使车辆规避检测。为提升对抗纹理的自然性,CAMOU<sup>[51]</sup>采用“马赛克”、DAS<sup>[52]</sup>采用场景语义相关图案、DTA<sup>[55]</sup>采用近似阴影等方法,但实际涂装后的车辆与真实车辆外观差异较大。

表 3 总结了前文所述的目标检测任务中的物理对抗补丁攻击方法。

## 3.3 其他任务中的对抗补丁攻击

### 3.3.1 人员重识别

与图像识别任务不同,Wang 等人<sup>[63]</sup>对人员重识别系统(re-ID)在物理世界实现了目标攻击和非目标攻击。其为了加强攻击的鲁棒性,在采用 TV、NPS 损失的同时,还提出了一种退化函数来随机改变样本集合中的对抗样本的亮度或清晰度,以增加该攻击在物理世界各种环境下的成功率。

### 3.3.2 语义分割

语义分割任务需要对图像中的每个像素点进行分类,以将图像输入分为不同的语义可解释类别。对抗补丁同样也能攻击语义分割任务,如 Nesti 等人<sup>[64]</sup>通过在广告牌上粘贴对抗补丁可以攻击语义分割模型。其提出使用像素级交叉熵损失的扩展形式,利用 3D 世界的几何信息来生成对抗补丁,并针对自动驾驶场景下的语义分割模型,在数字空间、Carla 仿真平台和物理世界做了详细的攻击实验验证和评估。

表 4 总结了前文所述的其他任务中的对抗补丁攻击方法。

## 4 物理对抗补丁防御

物理对抗补丁对现实世界中的深度神经网络模型的安全性造成很大挑战,故对于物理对抗补丁的防御研究十分重要。现有的一些防御方法主要包括

表 3 目标检测任务中的物理对抗补丁攻击方法汇总

Table 3 A summary of physical adversarial patch attack methods in object detection tasks

任务	攻击方法	攻击形式	黑盒/白盒	攻击验证方式	用到的数据集
人员检测	Dpatch <sup>[36]</sup>	数字图案补丁	黑盒	生成数字图案补丁在数字空间进行攻击验证	Pascal VOC 2007
	Physical patch+PGD <sup>[37]</sup>	贴纸	白盒	数字空间验证、物理世界验证	COCO 2014
	ProjectTrans <sup>[38]</sup>	贴纸、眼镜	白盒	打印眼镜并佩戴在人脸上进行攻击验证	CASIA-WebFace、作者照片
	TAPs <sup>[39]</sup>	贴纸	黑盒、白盒	数字空间验证、物理世界验证	LFW、CelebA-HQ
	AdvHat <sup>[40]</sup>	贴纸	白盒	打印对抗补丁粘贴在帽子上进行物理世界攻击验证	CASIA-WebFace
	Patch against person detectors <sup>[41]</sup>	贴纸	白盒	数字空间验证、物理世界验证	Inria Person、MS COCO、Pascal VOC
	UPC <sup>[42]</sup>	贴纸	白盒	打印粘贴在衣服上进行静态和动态验证	AttackScenes
	Invisibility Cloak <sup>[43]</sup>	打印海报、T 恤	白盒	打印海报并拍照验证数字攻击, 将补丁打印在 T 恤上在现实世界进行验证	COCO、Inria
	Adversarial T 恤 <sup>[45]</sup>	T 恤	白盒	打印对抗补丁粘贴在衣服上进行物理世界攻击验证	自己收集的不同场景视频
	NaturPhy Patch <sup>[46]</sup>	T 恤	白盒	将补丁打印在 T 恤上在现实世界进行验证	INRIA、MPII
	AdvText <sup>[47]</sup>	T 恤	白盒	打印对抗纹理裁剪衣服上进行物理世界攻击验证	Inria Person
	Extended RP2 <sup>[48]</sup>	海报、贴纸	黑盒、白盒	打印海报进行物理世界攻击验证	自制攻击视频
	ShapeShifter <sup>[49]</sup>	贴纸	白盒	打印对抗停车标志对不同角度、不同距离摄像头进行攻击验证	MS COCO
	Dynamic patch <sup>[50]</sup>	电子屏幕	白盒	使用电子屏幕显示对抗补丁进行物理世界攻击验证	MS COCO
自动驾驶	CAMOU <sup>[51]</sup>	Unreal 4 仿真	黑盒	在 Unreal 4 仿真引擎中将对抗纹理渲染在车身进行验证	通过仿真器收集数据
	DAS <sup>[52]</sup>	Carla 仿真、玩具车	黑盒	基于 Carla 仿真采集数据优化对抗纹理, 打印对抗伪装并粘贴在玩具车上进行物理攻击验证	通过仿真器收集数据
	FVC <sup>[53]</sup>	Carla 仿真、玩具车	白盒	基于 Carla 仿真实验验证, 打印对抗伪装并粘贴在玩具车上进行物理攻击验证	DAS 使用的数据集
	CAC <sup>[54]</sup>	Unity 模拟引擎、玩具车	白盒	基于 Unity 仿真实验验证, 打印对抗伪装并粘贴在玩具车上进行物理攻击验证	Unity 构建模拟场景
	DTA <sup>[55]</sup>	Carla 仿真、玩具车	白盒	基于 Carla 仿真实验验证, 打印对抗伪装并粘贴在玩具车上进行物理攻击验证	Carla 中的地图场景
军事目标探测	Patch against aerial detection <sup>[56]</sup>	数字图案补丁	白盒	生成数字图案补丁在数字空间进行攻击验证	DOTA
	Patch-Noobj <sup>[57]</sup>	数字图案补丁	白盒	生成数字图案补丁在数字空间进行攻击验证	DOTA、NWPU VHR-10、RSOD
	Imperceptible Adversarial Patches <sup>[58]</sup>	数字图案补丁	白盒	生成数字图案补丁在数字空间进行攻击验证	COCO-2017
	AC for naval vessels <sup>[59]</sup>	数字图案补丁	黑盒、白盒	生成数字图案补丁在数字空间进行攻击验证	shipspotting 下载
红外探测	InfAdvBulbs <sup>[60]</sup>	灯泡	白盒	将灯泡放置在木板上在物理世界进行验证	FLIR ADAS v1、FLIR Tau2
	InfAdvClothing <sup>[61]</sup>	气凝胶对抗服	白盒	将气凝胶涂装在二维码形式对抗补丁的衣服上进行验证	FLIR ADAS v1
	HOTCOLD Block <sup>[62]</sup>	发热贴和制冷贴	黑盒	将发热贴和制冷贴粘贴在衣物上在物理世界进行攻击验证	FLIR ADAS v2

表 4 其他任务中的对抗补丁攻击方法汇总

Table 4 A Survey of adversarial patch attack methods in other tasks

任务	攻击方法	攻击形式	黑盒/白盒	攻击验证方式	用到的数据集
人员重识别	advPattern <sup>[63]</sup>	贴纸	白盒	数字图像添加补丁验证、打印对抗补丁 贴在衣服上验证	Market1501 Dataset、PRCS Data-set(自建)
语义分割	EOT-based <sup>[64]</sup>	Carla 仿真、贴纸	白盒	在广告牌上粘贴打印的对抗补丁攻击语 义分割模型	Cityscapes、CARLA 和真实世界中 的数据

基于显著性图的防御、基于图像平滑的防御、基于对抗训练的防御、基于小感受野的防御等。

4.1 基于显著性图的防御

对抗补丁攻击的成功依赖于对显著特征的使用,其通过在图像的局部添加扰动使模型的关注点转移到局部区域。因而据此利用显著性图挑选出可疑区域以检测对抗补丁。如 Hayes 等人<sup>[65]</sup>受数字水印去除过程的启发提出了数字水印防御方法,其通过构建显著性图找到影响分类的密集簇来屏蔽对抗补丁以实现防御。Chou 等人<sup>[66]</sup>提出了一种用于检测局部通用攻击的架构 SentiNet,其不用事先了解攻击向量,仅依赖攻击者的特定行为来检测攻击。SentiNet 利用 Grad-CAM<sup>[67]</sup>生成掩码以提取对模型预测影响最大的一些区域,然后将这些提取的区域应用于一组良性测试输入并观察模型输出,与模型在良性输入上的已知输出进行比较以检测攻击,其具有一定的防御效果。

4.2 基于图像平滑的防御

图像平滑<sup>[68]</sup>将不同大小的高斯噪声添加到对抗补丁攻击图像中,能够抑制图像噪点使其趋于平缓以达到防御的目的。Naseer 等人<sup>[69]</sup>提出了局部梯度平滑(Local Gradients Smoothing, LGS)方法,其对图像局部区域进行处理,对图像中的不规则梯度进行正则化,在尽量不影响精准度的情况下实现有效防御。Levine 等人<sup>[70]</sup>提出了自适应的随机化平滑方法,其使用固定宽度的像素带消融图像来生成平滑图像,以抵御对抗补丁攻击。

4.3 基于对抗训练的防御

通过使用对抗样本作为输入对模型进行对抗训练能够提高模型的鲁棒性。Wu 等人<sup>[71]</sup>利用对抗训练抵御针对用矩形遮挡物覆盖交通标志牌的对抗补丁攻击。Rao 等人<sup>[72]</sup>提出通过全位置优化和随机位置优化对抗补丁的放置位置以提高模型的鲁棒性而不减少模型对于干净样本的准确性,但由于考虑了多个位置的情况增加了计算代价。

4.4 基于小感受野的防御

感受野(Receptive Field)是卷积神经网络每一层输出的特征图上的像素点在输入图片上映射的区

域。BagNet<sup>[73]</sup>通过减小卷积核来减小感受野的大小以限制被对抗补丁破坏的图像特征数量,进而可实现对对抗补丁的防御。Zhang 等人<sup>[74]</sup>在 BagNet 的基础上提出了一种 Clipped BagNet 模型,通过修改聚合步骤和裁剪异常逻辑实现对对抗补丁的防御。Xiang 等人<sup>[75]</sup>提出了对抗补丁防御方法 PG(Patch Guard),其使用 BagNet 作为主干网络,并使用特征聚合的方法来检测和屏蔽损坏特征以恢复正确特征,实验表明 PG 在保持模型精度的同时实现了一定的鲁棒性。后作者对 PG 进行扩展,提出了 PG++(Patch Guard++)<sup>[76]</sup>的方法,其将掩码应用于特征图中的所有可能位置并评估掩码,受攻击的图像会导致预测分歧,从而可以检测出对抗补丁攻击。此外,Zhang 等人<sup>[77]</sup>还提出了针对目标检测器的防御方法 DG(Detector Guard),其包含三个模块:基础检测器、对象预测器和对象解释器,当所有对象都得到了很好的解释则说明没有对抗补丁攻击,反之则说明存在攻击。

现有防御方法能够在一定程度上降低物理对抗补丁的攻击效果,但各有优劣。基于显著性图的防御方法通过检测物理对抗补丁的区域来针对性降低攻击效果,但其有效性取决于显著性图的质量,同时可能会损失原始信息。基于图像平滑的防御方法的效果取决于平滑程度和攻击复杂度,会降低原始模型的性能。基于对抗训练的防御方法具有较高的通用性,但需要更多的计算资源和更大的时间开销。基于小感受野的防御方法能够提高模型的鲁棒性,但可能会影响更大范围场景的感知和理解,导致原始模型性能下降。

表 5 总结了前文所述的对抗补丁防御方法。

5 总结与展望

5.1 研究现状总结

近年来,随着人工智能技术的快速发展,智能模型的安全性也渐受关注。而随着以深度神经网络模型为代表的智能模型越来越多地应用于现实物理世界,针对物理世界中智能模型的对抗攻击也逐渐成为了研究热点。与数字空间对抗攻击相比,物理空

表 5 对抗补丁防御方法汇总

防御分类	方法	针对的攻击方法	针对的模型	用到的数据集
基于显著性图	DW <sup>[65]</sup>	Adversarial Patch, LaVAN	VGG-19, ResNet-101, Inception-V3	ImageNet
	SentiNet <sup>[66]</sup>	Adversarial Patch	VGG-16	ImageNet
基于图像平滑	LGS <sup>[69]</sup>	LaVAN	Inception v3	ImageNet
	(De)Randomized smoothing <sup>[70]</sup>	IFGSM patch attack <sup>[78]</sup>	VGG-19, Inception V3	MNIST, CIFAR10, ImageNet
基于对抗训练	DOA <sup>[71]</sup>	Adv-Glass, RP2	VGGFace CNN, LISA-CNN	VGGFacedata, LISA
	LOA <sup>[72]</sup>	Location-optimized adversarial patch	ResNet-20	CIFAR10, GTSRB
	Clipped BagNet <sup>[74]</sup>	PGD, SPISA	ResNet-50, ResNet-101, DenseNet	ImageNet
基于小感受野	PG <sup>[75]</sup>	A single square adversarial patch	DS-ResNet, BagNet	ImageNet, ImageNette, CIFAR-10
	PG++ <sup>[76]</sup>	Adversarial patch	BagNet33	ImageNette, ImageNet, CIFAR-10
	DG <sup>[77]</sup>	Localized adversarial patch	YOLOv4, Faster R-CNN, PCD	PASCAL VOC, MS COCO, KITTI

间的对抗攻击难度更大, 但可能造成的影响也更大。因此, 在将智能模型实际部署到现实物理世界中时, 需全面评估可能面临的物理对抗攻击安全隐患。

针对物理世界智能模型的对抗攻击主要采用物理对抗补丁的形式, 研究关注点主要在于如何以较小代价生成更自然、不易被发现、攻击效果更好、在现实世界具有更高鲁棒性、能够在不同模型之间迁移的物理对抗补丁。物理对抗补丁的攻击形式和部署方式多样, 其自然性和鲁棒性也存在较大差异。提高物理对抗补丁的自然性一方面聚焦于生成尺寸更小、更隐蔽、更不易被发现的对抗补丁; 另一方面注重于生成风格纹理更自然、更不突兀的对抗补丁, 如涂鸦、贴纸类对抗补丁。当前提升物理对抗补丁的自然性采用风格损失、内容损失和 TV 损失等技术, 同时利用 GAN 来生成自然的对抗补丁。物理对抗补丁的鲁棒性决定了其是否有能力抵抗现实物理世界中的各种干扰, 而迁移性则决定了其是否能在现实物理世界针对不同模型进行对抗攻击。物理世界场景复杂多样, 可以通过改变光照、天气等自然条件, 调整拍摄距离、角度等场景条件, 以及考虑照相机拍摄、打印机打印颜色差异等因素来设计损失函数, 从而生成在物理世界中鲁棒性更好的物理对抗补丁。通过模型集成等技术可以生成迁移性更好的物理对抗补丁。但是, 提升对抗补丁的自然性、鲁棒性和迁移性的同时可能会降低对抗攻击效果。对于生成的物理对抗补丁, 主要通过数字图像测试、仿真平台测试、物理世界测试等不同形式验证其对抗攻击效果。

物理对抗补丁目前主要针对计算机视觉领域的各类任务, 如图像识别、目标检测、语义分割等。在不同的任务中所具体使用的技术存在差异, 但大体

的思路都比较一致, 即在数字空间生成对抗补丁并将其打印输出到物理空间, 进而用于误导或欺骗各类智能模型。针对物理对抗补丁的防御与针对对抗样本的防御思路类似, 主要包括基于显著性图的防御、基于图像平滑的防御、基于对抗训练的防御、基于小感受野的防御等。

对抗防御需要具备物理世界的可行性, 防御的成本和代价是防御者需要考虑的重要因素。如基于图像平滑的方法可能会增加额外的计算和内存开销, 导致模型的性能下降和延迟增加。当前尚无泛化性足够好的防御方法能够抵御所有的对抗补丁攻击, 难以应对不断变化的物理对抗攻击威胁。

5.2 未来研究展望

随着基于深度神经网络模型的智能系统在现实世界的大规模应用, 其安全性问题将会受到越来越多的关注。物理对抗补丁作为能在现实世界中误导或欺骗智能模型的方法手段, 对其攻击与防御技术的研究显得十分迫切而必要。结合当前对于物理对抗补丁的相关研究进展与现实应用, 未来还可以从以下角度做进一步探索:

(1) 物理对抗补丁的自然性与攻击有效性之间的平衡。物理对抗补丁要实现攻击效果通常需要较大尺寸或较明显的对比纹理, 以使得对于智能模型的影响较为稳定有效。但对于攻击者而言, 保持攻击的自然性, 即不被人眼所明显察觉也很重要。因此, 如何在扰动幅度和扰动范围受限的情况下尽可能提升对抗攻击的有效性仍然需要探索研究。

(2) 物理对抗补丁在黑盒场景下的可迁移性。真实世界中的智能模型通常是黑盒的, 攻击者对于目标模型的知识所知甚少。现有的基于查询和基于迁



移的物理对抗补丁生成方法其迁移率都不尽理想。与数字空间相比,物理空间存在大量的天然噪声,如光照阴影、倾斜旋转、雨雪雾霾等,也为提升物理对抗补丁的迁移性增加了困难。

(3) 有更好鲁棒泛化性的物理对抗补丁防御框架。鲁棒泛化性(Robust generalization)描述了鲁棒模型在未知的对抗数据上的表现,学习具有良好鲁棒泛化性的模型需要更多的数据、更高复杂度的样本<sup>[78-81]</sup>。正则化技术<sup>[82]</sup>或对抗训练方法<sup>[83-84]</sup>都能一定程度上提升模型的鲁棒泛化性。而现有的物理对抗补丁防御方法通常只针对一种或多种已知的攻击方法,难以抵御未知对抗数据的攻击,无法确保现实世界中智能模型防护的安全下限。与数字空间中的对抗样本防御类似,针对物理对抗补丁的鲁棒泛化防御框架研究仍然需要深入研究。

## 参考文献

- [1] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing Properties of Neural Networks[M]. arXiv, 2014.
- [2] Brown T B, Mané D, Roy A, et al. Adversarial Patch[M]. arXiv, 2018.
- [3] Chakraborty A, Alam M, Dey V, et al. Adversarial Attacks and Defences: A Survey[M]. arXiv, 2018.
- [4] Yuan X Y, He P, Zhu Q L, et al. Adversarial Examples: Attacks and Defenses for Deep Learning[J]. *IEEE Transactions on Neural Networks and Learning Systems*, 2019, 30(9): 2805-2824.
- [5] Qian Z, Huang K Z, Wang Q F, et al. A Survey of Robust Adversarial Training in Pattern Recognition: Fundamental, Theory, and Methodologies[J]. *Pattern Recognition*, 2022, 131: 108889.
- [6] Goodfellow I J, Shlens J, Szegedy C. Explaining and Harnessing Adversarial Examples[M]. arXiv, 2015.
- [7] Kurakin A, Goodfellow I J, Bengio S. Adversarial Examples in the Physical World[M]. *Artificial Intelligence Safety and Security*. First edition. | Boca Raton, FL: CRC Press/Taylor & Francis Group, 2018: Chapman and Hall/CRC, 2018: 99-112.
- [8] Madry A, Makelov A, Schmidt L, et al. Towards Deep Learning Models Resistant to Adversarial Attacks[M]. arXiv, 2019.
- [9] Moosavi-Dezfooli S M, Fawzi A, Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2574-2582.
- [10] Moosavi-Dezfooli S M, Fawzi A, Fawzi O, et al. Universal Adversarial Perturbations[C]. *2017 IEEE Conference on Computer Vision and Pattern Recognition*, 2017: 86-94.
- [11] Papernot N, McDaniel P, Jha S, et al. The Limitations of Deep Learning in Adversarial Settings[C]. *2016 IEEE European Symposium on Security and Privacy*, 2016: 372-387.
- [12] Carlini N, Wagner D. Towards Evaluating the Robustness of Neural Networks[C]. *2017 IEEE Symposium on Security and Privacy*, 2017: 39-57.
- [13] Croce F, Hein M. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-Free Attacks[EB/OL]. 2020: 2003.01690. <https://arxiv.org/abs/2003.01690v2>.
- [14] Su J W, Vargas D V, Sakurai K. One Pixel Attack for Fooling Deep Neural Networks[J]. *IEEE Transactions on Evolutionary Computation*, 2019, 23(5): 828-841.
- [15] 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[C]. *The 10th ACM Workshop on Artificial Intelligence and Security*, 2017: 15-26.
- [16] Chen J H, Gu Q Q. RaS: A Ray Searching Method for Hard-Label Adversarial Attack[C]. *The 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020: 1739-1747.
- [17] Dong Y P, Liao F Z, Pang T Y, et al. Boosting Adversarial Attacks with Momentum[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 9185-9193.
- [18] Duan R J, Ma X J, Wang Y S, et al. Adversarial Camouflage: Hiding Physical-World Attacks with Natural Styles[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 997-1005.
- [19] Gatys L A, Ecker A S, Bethge M. Image Style Transfer Using Convolutional Neural Networks[C]. *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 2016: 2414-2423.
- [20] Strong D, Chan T. Edge-Preserving and Scale-Dependent Properties of Total Variation Regularization[J]. *Inverse Problems*, 2003, 19(6): S165-S187.
- [21] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing Robust Adversarial Examples[EB/OL]. 2017: 1707.07397. <https://arxiv.org/abs/1707.07397v3>.
- [22] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition[C]. *The 2016 ACM SIGSAC Conference on Computer and Communications Security*, 2016: 1528-1540.
- [23] Eykholt K, Evtimov I, Fernandes E, et al. Robust Physical-World Attacks on Deep Learning Visual Classification[C]. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018: 1625-1634.
- [24] Liu A S, Liu X L, Fan J X, et al. Perceptual-Sensitive GAN for Generating Adversarial Patches[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, 33(1): 1028-1035.
- [25] Kong Z L, Guo J F, Li A, et al. PhysGAN: Generating Physical-World-Resilient Adversarial Examples for Autonomous Driving[EB/OL]. 2019: 1907.04449. <https://arxiv.org/abs/1907.04449v3>.
- [26] Duan R J, Mao X F, Qin A K, et al. Adversarial Laser Beam: Effective Physical-World Attack to DNNs in a Blink[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 16057-16066.
- [27] Zhong Y Q, Liu X M, Zhai D M, et al. Shadows Can Be Dangerous: Stealthy and Effective Physical-World Adversarial Attack by Natural Phenomenon[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 15324-15333.
- [28] Zhu Z G, Lu Y Z, Chiang C K. Generating Adversarial Examples by Makeup Attacks on Face Recognition[C]. *2019 IEEE International Conference on Image Processing*, 2019: 2516-2520.
- [29] Wei X X, Guo Y, Yu J. Adversarial Sticker: A Stealthy Attack

- Method in the Physical World[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, 45(3): 2711-2725.
- [30] Nguyen D L, Arora S S, Wu Y H, et al. Adversarial Light Projection Attacks on Face Recognition Systems: A Feasibility Study[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020: 3548-3556.
- [31] Karmon D, Zoran D, Goldberg Y. LaVAN: Localized and Visible Adversarial Noise[EB/OL]. 2018: 1801.02608. <https://arxiv.org/abs/1801.02608v2>.
- [32] Chindaudom A, Siritanawan P, Sumongkayothin K, et al. AdversarialQR: An Adversarial Patch in QR Code Format[C]. *2020 Joint 9th International Conference on Informatics, Electronics & Vision and 2020 4th International Conference on Imaging, Vision & Pattern Recognition*, 2020: 1-6.
- [33] Doan B G, Xue M H, Ma S Q, et al. TnT Attacks! Universal Naturalistic Adversarial Patches Against Deep Neural Network Systems[J]. *IEEE Transactions on Information Forensics and Security*, 2022, 17: 3816-3830.
- [34] Gittings T, Schneider S, Collomosse J. Robust Synthesis of Adversarial Visual Examples Using a Deep Image Prior[M]. arXiv, 2019.
- [35] Zhou X Y, Pan Z S, Duan Y X, et al. A Data Independent Approach to Generate Adversarial Patches[J]. *Machine Vision and Applications*, 2021, 32(3): 67.
- [36] Liu X, Yang H, Liu Z, et al. DPatch: An Adversarial Patch Attack on Object Detectors[M]. arXiv, 2019.
- [37] Lee M, Kolter Z. On Physical Adversarial Patches for Object Detection[M]. arXiv, 2019.
- [38] Pautov M, Melnikov G, Kaziakhmedov E, et al. On Adversarial Patches: Real-World Attack on ArcFace-100 Face Recognition System[C]. *2019 International Multi-Conference on Engineering, Computer and Information Sciences*, 2019: 391-396.
- [39] Xiao Z H, Gao X F, Fu C L, et al. Improving Transferability of Adversarial Patches on Face Recognition with Generative Models[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 11840-11849.
- [40] Komkov S, Petiushko A. AdvHat: Real-World Adversarial Attack on ArcFace Face ID System[C]. *2020 25th International Conference on Pattern Recognition*, 2021: 819-826.
- [41] Thys S, Ranst W V, Goedemé T. Fooling Automated Surveillance Cameras: Adversarial Patches to Attack Person Detection[C]. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019: 49-55.
- [42] Huang L F, Gao C Y, Zhou Y Y, et al. Universal Physical Camouflage Attacks on Object Detectors[C]. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020: 717-726.
- [43] Wu Z X, Lim S N, Davis L S, et al. Making an Invisibility Cloak: Real World Adversarial Attacks on Object Detectors[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 1-17.
- [44] Bookstein F L. Principal Warps: Thin-Plate Splines and the Decomposition of Deformations[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1989, 11(6): 567-585.
- [45] Xu K D, Zhang G Y, Liu S J, et al. Adversarial T-Shirt! Evading Person Detectors in a Physical World[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 665-681.
- [46] Hu Y C T, Chen J C, Kung B H, et al. Naturalistic Physical Adversarial Patch for Object Detectors[C]. *2021 IEEE/CVF International Conference on Computer Vision*, 2021: 7828-7837.
- [47] Hu Z H, Huang S Y, Zhu X P, et al. Adversarial Texture for Fooling Person Detectors in the Physical World[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13297-13306.
- [48] Song D, Eykholt K, Evtimov I, et al. Physical Adversarial Examples for Object Detectors[C]. *12th USENIX Workshop on Offensive Technologies*, 2018.
- [49] Chen S T, Cornelius C, Martin J, et al. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2019: 52-68.
- [50] Hoory S, Shapira T, Shabtai A, et al. Dynamic Adversarial Patch for Evading Object Detection Models[M]. arXiv, 2020.
- [51] Zhang, Yang, PD Hassan Foroosh, and Boqing Gong. Camou: Learning a vehicle camouflage for physical adversarial attack on object detections in the wild[C]. *ICLR*, 2019.
- [52] Wang J K, Liu A S, Yin Z X, et al. Dual Attention Suppression Attack: Generate Adversarial Camouflage in Physical World[C]. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021: 8561-8570.
- [53] Wang D H, Jiang T S, Sun J L, et al. FCA: Learning a 3D Full-Coverage Vehicle Camouflage for Multi-View Physical Adversarial Attack[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, 36(2): 2414-2422.
- [54] Duan Y X, Chen J L, Zhou X Y, et al. Learning Coated Adversarial Camouflages for Object Detectors[C]. *The Thirty-First International Joint Conference on Artificial Intelligence*, 2022: 891-897.
- [55] Suryanto N, Kim Y, Kang H, et al. DTA: Physical Camouflage Attacks Using Differentiable Transformation Network[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 15284-15293.
- [56] Adhikari A, Hollander R den, Tolios I, et al. Adversarial Patch Camouflage against Aerial Detection[M]. arXiv, 2020.
- [57] Lu M M, Li Q, Chen L, et al. Scale-Adaptive Adversarial Patch Attack for Remote Sensing Image Aircraft Detection[J]. *Remote Sensing*, 2021, 13(20): 4078.
- [58] Wise C, Plested J. Developing Imperceptible Adversarial Patches to Camouflage Military Assets from Computer Vision Enabled Technologies[M]. arXiv, 2022.
- [59] Løkken K H, Aurdal L, Brattli A, et al. Investigating Robustness of Adversarial Camouflage (AC) for Naval Vessels[C]. *Artificial Intelligence and Machine Learning in Defense Applications II*, 2020: 87-97.
- [60] Zhu X P, Li X, Li J M, et al. Fooling Thermal Infrared Pedestrian Detectors in Real World Using Small Bulbs[J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2021, 35(4): 3616-3624.
- [61] Zhu X P, Hu Z H, Huang S Y, et al. Infrared Invisible Clothing: Hiding from Infrared Detectors at Multiple Angles in Real World[C]. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022: 13307-13316.

- [62] Wei H, Wang Z, Jia X, et al. HOTCOLD Block: Fooling Thermal Infrared Detectors with a Novel Wearable Design[M]. arXiv, 2022.
- [63] Wang Z B, Zheng S Y, Song M K, et al. AdvPattern: Physical-World Attacks on Deep Person Re-Identification via Adversarially Transformable Patterns[C]. 2019 IEEE/CVF International Conference on Computer Vision, 2019: 8340-8349.
- [64] Nesti F, Rossolini G, Nair S, et al. Evaluating the Robustness of Semantic Segmentation for Autonomous Driving Against Real-World Adversarial Patch Attacks[C]. 2022 IEEE/CVF Winter Conference on Applications of Computer Vision, 2022: 2826-2835.
- [65] Hayes J. On Visible Adversarial Perturbations & Digital Watermarking[C]. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2018: 1678-16787.
- [66] Chou E, Tramèr F, Pellegrino G. SentiNet: Detecting Localized Universal Attacks Against Deep Learning Systems[C]. 2020 IEEE Security and Privacy Workshops, 2020: 48-54.
- [67] Chattopadhyay A, Sarkar A, Howlader P, et al. Grad-CAM: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks[C]. 2018 IEEE Winter Conference on Applications of Computer Vision, 2018: 839-847.
- [68] Cohen J, Rosenfeld E, Kolter Z. Certified Adversarial Robustness via Randomized Smoothing[C]. Proceedings of the 36th International Conference on Machine Learning. PMLR, 2019: 1310-1320.
- [69] Naseer M, Khan S, Porikli F. Local Gradients Smoothing: Defense Against Localized Adversarial Attacks[C]. 2019 IEEE Winter Conference on Applications of Computer Vision, 2019: 1300-1307.
- [70] Levine A, Feizi S. (de)Randomized Smoothing for Certifiable Defense Against Patch Attacks[EB/OL]. 2020: 2002.10733. <https://arxiv.org/abs/2002.10733v3>.
- [71] Wu T, Tong L, Vorobeychik Y. Defending Against Physically Realizable Attacks on Image Classification[M]. arXiv, 2020.
- [72] Rao S, Stutz D, Schiele B. Adversarial Training Against Location-Optimized Adversarial Patches[M]. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2020: 429-448.
- [73] Brendel W, Bethge M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet[M]. arXiv, 2019.
- [74] Zhang Z Y, Yuan B, McCoyd M, et al. Clipped BagNet: Defending Against Sticker Attacks with Clipped Bag-of-Features[C]. 2020 IEEE Security and Privacy Workshops, 2020: 55-61.
- [75] Xiang C, Bhagoji A N, Sehwag V, et al. PatchGuard: A Provably Robust Defense Against Adversarial Patches via Small Receptive Fields and Masking[EB/OL]. 2020: 2005.10884. <https://arxiv.org/abs/2005.10884v5>.
- [76] Xiang C, Mittal P. PatchGuard++: Efficient Provable Attack Detection against Adversarial Patches[M]. arXiv, 2021.
- [77] Xiang C, Mittal P. DetectorGuard: Provably Securing Object Detectors Against Localized Patch Hiding Attacks[C]. The 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021: 3177-3196.
- [78] Chiang P Y, Ni R, Abdelkader A, et al. Certified Defenses for Adversarial Patches[M]. arXiv, 2020.
- [79] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially Robust Generalization Requires More Data[J]. Advances in Neural Information Processing Systems, 2018, 31: 5019-5031.
- [80] Zhai R, Cai T, He D, et al. Adversarially Robust Generalization Just Requires More Unlabeled Data[M]. arXiv, 2019.
- [81] Yin D, Ramchandran K, Bartlett P. Rademacher Complexity for Adversarially Robust Generalization[EB/OL]. 2018: 1810.11914. <https://arxiv.org/abs/1810.11914v4>.
- [82] Zhang S, Qian Z, Huang K, et al. Towards Better Robust Generalization with Shift Consistency Regularization[C]. International Conference on Machine Learning. PMLR. 2021: 12524-12534.
- [83] Liu C, Salzmann M, Lin T, et al. On the Loss Landscape of Adversarial Training: Identifying Challenges and how to Overcome Them[EB/OL]. 2020: 2006.08403. <https://arxiv.org/abs/2006.08403v2>.
- [84] Wu D, Xia S T, Wang Y. Adversarial Weight Perturbation Helps Robust Generalization[J]. Neural Information Processing Systems, 2020: 2958-2969.



**邓欢** 现在军事科学院系统工程研究院攻读硕士学位。计算机应用技术专业, 研究领域为人工智能安全。Email: denghuan619@163.com



**黄敏** 通信作者, 博士学位, 现为军事科学院系统工程研究院研究员。研究领域为信息系统脆弱性分析与验证评估。Email: darbean@126.com



**李虎** 博士学位, 现为军事科学院系统工程研究院工程师。研究领域为人工智能安全与机器学习。Email: lihu\_lh@163.com



**王彤** 现在军事科学院系统工程研究院攻读博士学位。计算机软件与理论专业, 研究领域为人工智能测试与评估。Email: tongwss@foxmail.com



况晓辉 博士学位, 现为军事科学院系统工程研究院研究员。研究领域为网络与信息安全、无线网络、人工智能安全和机器学习。Email: xiaohui\_kuang@163.com