

基于光学的物理域对抗攻防综述

陈晋音^{1,2}, 赵晓明², 郑海斌^{1,2}, 郭海锋^{1,2}

(1. 浙江工业大学网络空间安全研究院, 浙江 杭州 310023; 2. 浙江工业大学信息工程学院, 浙江 杭州 310023)

摘要: 对抗攻击是指通过在原始输入中植入人眼无法察觉的微小扰动, 误导深度学习模型做出错误预测的攻击。与数字域对抗攻击相比, 物理域对抗攻击可实现对抗性输入被采集设备捕获并转换为视觉系统内的二值图像之前, 将扰动引入输入, 对基于深度学习的计算机视觉系统构成了实际安全威胁。基于光学的物理域对抗攻击技术(如使用投影照射)作为一种典型的非侵入性攻击, 由于其扰动与现实世界中自然环境产生的影响非常相似, 更容易被忽略, 从而疏于防护。鉴于它们具有高度的不可见性和可执行性, 可对实际系统构成重大甚至致命的威胁。基于现有研究工作, 重点介绍和讨论了计算机视觉系统中基于光学的物理域对抗攻击技术, 并对现有技术在攻击场景、攻击手段、攻击目标、攻击效果等方面展开详细分析, 最后探讨了基于光学的物理域对抗攻击未来潜在研究方向。

关键词: 对抗攻击; 深度学习; 安全威胁; 光学物理域对抗攻击

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024026

Survey of optical-based physical domain adversarial attacks and defense

CHEN Jinyin^{1,2}, ZHAO Xiaoming², ZHENG Haibin^{1,2}, GUO Haifeng^{1,2}

1. Institute of Cyberspace Security, Zhejiang University of Technology, Hangzhou 310023, China

2. College of Information Engineering, Zhejiang University of Technology, Hangzhou 310023, China

Abstract: Deep learning models are misled into making false predictions by adversarial attacks that implant tiny perturbations into the original input, which are imperceptible to the human eye. This poses a huge security threat to computer vision systems that are based on deep learning. Compared to digital-domain adversarial attacks, physical-domain adversarial attacks are enabled to introduce perturbations into the input before the adversarial input is captured by the acquisition device and converted into a binary image within the vision system, posing a real security threat to deep learning-based computer vision systems. Optical-based physical-domain adversarial attack techniques, such as those using projected irradiation as a typical example, are more likely to be overlooked and provided

收稿日期: 2023-11-03; 修回日期: 2024-03-15

通信作者: 陈晋音, chenjinyin@zjut.edu.cn

基金项目: 国家自然科学基金(No.62072406); 浙江省自然科学基金(No.LDQ23F020001)

Foundation Items: The National Natural Science Foundation of China (No.62072406), The Natural Science Foundation of Zhejiang province (No.LDQ23F020001)

引用格式: 陈晋音, 赵晓明, 郑海斌, 等. 基于光学的物理域对抗攻防综述[J]. 网络与信息安全学报, 2024, 10(02): 1-21.

Citation Format: CHEN J Y, ZHAO X M, ZHENG H B, et al. Survey of optical-based physical domain adversarial attacks and defense[J]. Chinese Journal of Network and Information Security, 2024, 10(02): 1-21.

negligible protection due to their perturbations being very similar to effects produced by natural environments in the real world. Given their high degree of invisibility and executability, they could pose a significant or even fatal threat to real systems. Based on existing research work, the introduction and discussion of optical-based physical-domain adversarial attack techniques within computer vision systems were presented. The attack scenarios, tools, goals, and performances of these techniques were compared and analyzed. Potential future research directions for optical-based physical-domain adversarial attacks were also discussed.

Keywords: adversarial attack, deep learning, security threat, optical physical domain adversarial attack

0 引言

计算机视觉是指利用计算机及相关设备模拟生物视觉,使计算机具有通过二维图像识别三维环境信息的能力。作为人工智能领域的重要组成部分,近年来,随着深度学习技术的快速发展,基于深度神经网络(deep neural network, DNN)的计算机视觉技术也取得了长足的进步,在人脸识别^[1]、自动驾驶^[2]、智能制造^[3]、生物医学^[4]等领域有着广泛的应用。

然而,深度神经网络存在一些固有的问题,特别是算法的黑盒特性所导致的不可解释性,使得其很难获得与样本本质特征和预测结果相关的因素。这些问题导致基于DNN的计算机视觉技术存在一些安全漏洞,可被恶意攻击者利用,进行对抗攻击。对抗攻击是指通过在原始输入中植入微小的扰动来误导分类器产生错误的预测。这些微小的扰动对人类来说是难以察觉的,但可以使分类器产生更高置信度的错误预测。如文献[5]所述,交通标志识别系统的对抗攻击会导致自动驾驶汽车对交通标志进行错误分类,自动驾驶系统会发出错误的控制命令,从而导致交通事故,甚至可能危及驾驶员和乘客的人身安全。

根据攻击形式在数字域与物理域中实现的不同,对抗攻击可分为数字域对抗攻击与物理域对抗攻击^[6]。具体而言,数字域对抗攻击通过直接访问/修改数字图像数据,并输入目标模型导致其作出错误分类来实现攻击,在物理世界中,将数字域对抗样本直接做物理赋形时,由于对抗扰动失真、二次采集误差等对抗衰减,物理域的对抗攻击显得更加具有挑战性。物理域对抗攻击在计算机视觉系统捕获输入并将其转换为二值图像之前,通过物理方式植入扰动来修改输入(如在真实的路标上绘制标记)。物理域对抗攻击将对现实世界

的深度学习系统造成重大甚至致命的影响,因此近年来受到越来越多研究者的关注和研究^[7-9]。

计算机视觉系统的接收设备(如相机)本质上是一个光电传感器,因此来自光学器件的干扰通常会对计算机视觉系统产生关键性的影响。例如,一辆特斯拉S型汽车与一辆白色拖拉机拖车相撞,造成了自动驾驶汽车的首例死亡事件。特斯拉表示,事故发生的原因是“自动驾驶仪和司机都没有注意到明亮天空下拖拉机拖车的白色一面”^[10]。

基于光学手段的物理域对抗攻击方法是近年来研究的热门方向,目前已有一些研究成果发表^[11-12]。本文在简要介绍对抗攻击的一些基本概念的基础上,将物理域对抗攻击技术分为侵入性攻击和非侵入性攻击两种,并以基于光学的计算机视觉系统物理域对抗攻击技术为重点,对其特点、机制、隐身性高、部署简单等方面进行了深入的分析研究。

本文通过对物理域对抗攻击技术的分析和研究,有助于提前发现DNN的内生安全漏洞,启发设计有针对性的防御解决方案,降低了深度学习模型不成熟和恶意应用带来的安全风险,增强了深度学习模型的安全性和安全应用。

1 对抗攻击基本概念及理论分析

本节主要介绍对抗攻击的基本概念,首先介绍对抗攻击的判断条件和分类,然后分析对抗扰动的攻击成因,最后介绍对抗攻击的两大类:数字域对抗攻击和物理域对抗攻击。

1.1 对抗攻击的判断条件和分类

对抗攻击首先由Szegedy等^[13]提出,其中原始图像的小扰动可以误导目标模型做出错误的预测,其中添加了扰动的图像,将其称为对抗样本。然而,随着对抗攻击技术的发展,对抗样本

可以直接在二值图像上生成（在数字设置中：通过添加噪声），也可以通过图像传感器（如相机）捕获被攻击的物理场景生成（在物理设置中：通过添加雾和阳光）。一般来说，对抗攻击满足以下3个条件。

（1）目标模型在添加扰动之前能够对原始图像做出正确的预测，如图1中序号1所示。

（2）目标模型在添加扰动之后必然会对原始图像做出错误预测，如图1中序号2所示。

（3）在对原始图像添加扰动之前和之后，人类可以根据感知做出正确的预测，序号3（添加扰动之前）和序号4（添加扰动之后）如图1所示。

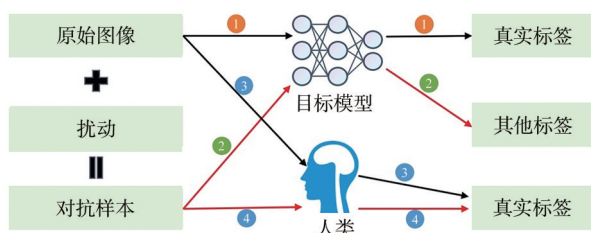


图1 对抗例子所满足的条件
Figure 1 Conditions satisfied by the confrontation example

对抗攻击根据预测结果可分为两类：非目标攻击和目标攻击。

（1）非目标攻击

攻击者只是通过对抗样本误导目标模型输出任意错误的预测。一个是置信度降低，目标模型以低置信度将“狗”对抗样本预测为“狗”图像；另一个是错误分类，目标深度模型采用“狗”对抗的例子，并将其预测为“非狗”图像。

（2）目标攻击

攻击者试图通过对抗样本误导目标模型输出特定的错误预测。一个是有针对性的错误分类：目标模型将任何类别的对抗样本（如“马”和“猫”图像）预测到特定类别（如“狗”图像）；另一个是源/目标错误分类：目标深度模型采用特定类（如“猫”图像）的对抗样本，并将其预测到另一个特定类（如“狗”图像）。根据攻击的时间，对抗攻击也可以分为两类：训练阶段的攻击和测试阶段的攻击。

一般来说，深度学习模型的流程包括两个阶段：训练阶段和测试阶段。因此，对抗攻击也可以在不同的阶段/时间产生扰动，包括训练阶段

攻击（又名：后门或者中毒攻击）和测试阶段攻击（又名：探索性或闪避攻击）^[14]。

训练阶段攻击。在训练阶段，攻击者可以通过数据注入、修改数据特征和标签等方式修改训练数据集。目的是使目标模型学习到错误的目标特征，干扰和破坏目标模型的逻辑。

测试阶段攻击。在测试阶段，攻击者不会也不能干扰或破坏目标模型的逻辑，但攻击者可以根据所获得的目标模型的知识量制作对抗样本，利用对抗样本迫使预训练的目标模型产生错误的预测^[15]。根据攻击者对目标模型的掌握程度，可以将测试阶段的攻击进一步分为白盒攻击、黑盒攻击和灰盒攻击。

白盒攻击：攻击者假设已知目标模型的全部知识，包括其参数值、体系结构、训练方法，在某些情况下还包括其训练数据^[16]。因此，攻击者可以通过获得目标深度模型的全部知识来制作一个对抗样本。

黑盒攻击：攻击者无法访问DNN模型，因此无法获得模型的结构和参数，只能通过向目标模型输入原始数据来获得目标模型的输出结果^[17]。

灰盒攻击：攻击者假设已知目标模型的部分知识^[18]。因此，攻击者可以利用已知的部分知识获取更多的目标模型未知知识，推断出目标模型的漏洞。

1.2 对抗扰动的攻击成因

对抗扰动可以通过损坏无干扰的输入来误导目标物体，从而改变DNN为识别或分类而学习的内在结构^[19]。近年来，大量研究者试图从不同角度分析对抗扰动的原因，虽然对抗扰动的原因目前并没有统一的说法，但总的来说，对抗扰动的可能原因可以分为以下3类。

（1）DNN结构局部过线性化，在传输过程中可能会放大最小扰动

Goodfellow等^[20]认为，产生对抗样本的原因是高维空间中的线性行为。在高维线性分类器中，通过点积运算对各维的最小扰动进行累积放大，利用非线性激活函数sigmoid计算分类概率，扰动前后的原始图像被分类为“1”类的概率从4.74%提高到88%，如图2所示。如果DNN使用

线性激活函数,则原始图像更容易受到扰动前后分类概率差异较大的影响^[21]。

(2) 训练数据集包含的目标特征不足,导致 DNN 模型的决策边界过早停止

有研究者认为, DNN 模型的脆弱性是训练不完全造成的,而训练不完全是数据不足造成的^[19]。当训练数据标记不足时,目标特征的多样性,数据集的维数较大,会导致 DNN 模型的决策边界过早停止,泛化能力较弱。

假设原始目标数据集 Set_{or} 具有圆形和六边形的红、蓝、绿 6 个特征,而随机选择的训练数据集 Set_{tr} 只包含圆形的红、蓝、绿 3 个特征,如图 3 所示。在训练过程中,当目标深度模型能够很好地分辨出 Set_{tr} 中包含的特征时,训练可能会停

止,决策边也会停止进化。对于一些未被训练数据集覆盖的特征(如六边形),目标深度模型将无法正确分类,因此目标深度模型需要更多的数据来学习目标特征,以提高模型的鲁棒性。

(3) 分类器中存在非鲁棒特征导致预测不准确

Cubuk 等^[22]认为,“对抗样本的起源主要是由于神经网络对其预测具有固有的不确定性”。他们认为目标模型输出结果的不确定性与网络架构、训练方法和数据集无关。

假设一个分类器包含鲁棒特征(深色方形和圆形)和非鲁棒特征(浅色方形和圆形),红色曲线为数据集的真实决策边界,如图 4 所示。分类器中非鲁棒特征的存在导致训练后分类器的决

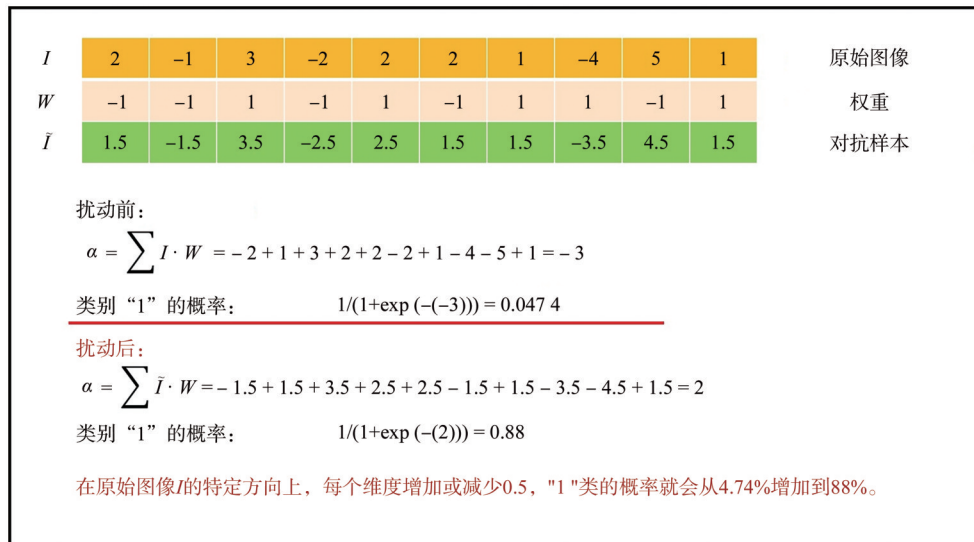


图 2 高维线性分类器对扰动的影响
Figure 2 Effect of high dimensional linear classifier on perturbation

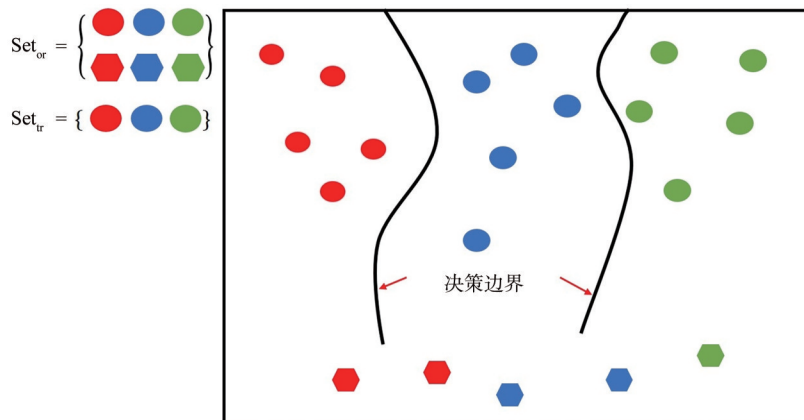


图 3 训练数据不足对模型决策边界的影响
Figure 3 The effect of insufficient training data on model decision boundary

策边界（图4中蓝色曲线）不能很好地识别目标中的非鲁棒特征，因此，如果将这些非鲁棒特征添加到输入图像中，很可能会导致目标深度模型做出错误的预测。

1.3 数字域对抗攻击与物理域对抗攻击

近年来对抗攻击的方法种类繁多，Akhtar等^[18]尝试用一个简单/统一的数学模型将对抗攻击表示为：

$$M(\tilde{I}) \rightarrow \tilde{l} \text{ s.t. } \tilde{l} \neq l, \tilde{I} \in S_l, M(I - \{S_l - \tilde{I}\}) = l \quad (1)$$

其中 M 改为 $(\cdot)$ 为目标模型，即 $M(I): I \rightarrow l$ ， I 为原始图像， $\tilde{I}$ 为目标模型的输出。 $\tilde{I} = I + p$ 是对抗样本， p 为对抗扰动。如图5所示，如果攻击者在目标成像后对图像进行扰动，则为数字对抗样本；如果攻击者在目标被成像之前攻击图像，则是一个物理对抗样本。因此， S_l 可以理解为数字对抗样本和物理对抗样本的集合，即 $\tilde{I} \in S_l$ 。

由于数字域对抗攻击是在目标成像后对图像进行攻击，攻击者可以利用该程序产生人眼难以

察觉的像素级扰动。物理域对抗攻击需要通过传感器（如摄像机）捕获对抗图像，并且由于摄像机捕获的小扰动很容易被忽略，大多数物理对抗样本中的扰动很容易被人眼察觉，但人一般不会认为这些扰动会影响整体，因此不会引起警报。

(1) 数字域对抗攻击

在数字域对抗攻击中，攻击者首先通过程序将扰动直接添加到原始图像上，甚至可以将该扰动精确地添加到原始图像的一个像素点上，然后使用 L 范数限制扰动的大小，以确保扰动达到人眼无法感知的程度，最后，通过数字对抗实例，误导目标模型做出错误预测。针对此，研究人员提出了许多方法来生成最优扰动。例如，Goodfellow等^[20]提出了使用 L_∞ 范数约束扰动的快速梯度符号法。Papernot等^[23]提出了使用 L_2 范数约束扰动的基于雅可比的显著性图攻击。Su等^[24]提出了使用 L_0 范数约束扰动的单像素攻击。Athalye等^[25]提出了使用 L_∞ 、 L_2 范数约束的后向传递可微逼近攻击。Carlini等^[26]提出了使用 L_∞ 、 L_2 、 L_0 范

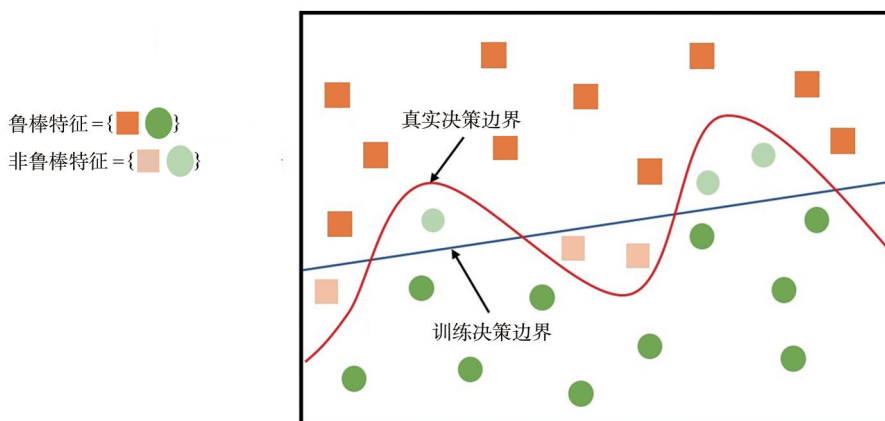


图4 非鲁棒特征对模型决策边界的影响

Figure 4 The influence of non-robust features on the model decision boundary

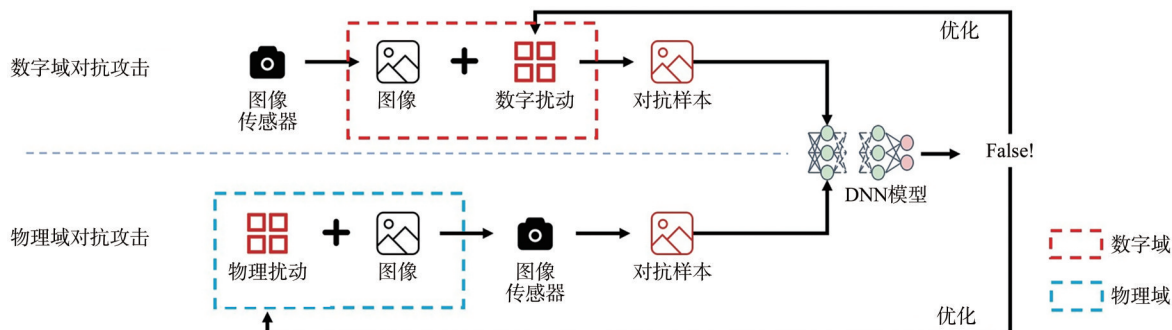


图5 数字域对抗攻击和物理域对抗攻击的比较

Figure 5 Comparison of digital domain adversarial attack and physical domain adversarial attack

数约束的 C&W 攻击。虽然数字域对抗攻击可以表现出高性能,但这些方法是通过直接在数字世界的像素级上操纵输入图像来植入小的扰动。由于动态物理条件(如射击角度、距离等不同)和光学成像的影响,很难将数字对抗样本中的扰动完美地传递到现实世界中,因此数字域对抗攻击在现实世界中的可执行性非常低。

(2) 物理域对抗攻击

在实际应用中,摄像头是人工智能视觉系统的入口,即人工智能视觉系统需要通过摄像头捕捉原始图像作为数据输入。因此,在物理域对抗攻击中,攻击者还需要通过系统摄像机捕捉对抗样本图像。但是,由于相机的成像质量受相机光电传感器物理性能的限制,小的扰动可能不会被相机捕捉到,在某种程度上也可以认为:小的扰动会在相机成像阶段被过滤。因此,在现有的大多数物理域对抗攻击中,植入扰动的大小通常在人眼的可感知范围内,攻击方案往往需要一些额外的手段来隐藏扰动,使其不被人眼注意到。

Kurakin 等^[27]首先发现了物理对抗样本的存在,并证明当对抗样本被打印出来时,即使在不同的光照和方向下,它们仍然对分类器有害。在此之后,出现了大量的物理域对抗攻击技术。例如, Eykholt 等^[7]提出了一种通用攻击算法——鲁棒物理攻击 (robust physical perturbation, RP2) 该算法使用 L_1 范数找到交通路标的攻击区域,然后使用 L_2 范数对攻击区域进行优化,最后在最优攻击位置上手动贴上黑白贴纸。尽管 RP2 可以捕捉各种环境下的鲁棒扰动,但黑白贴纸在现实世界中的隐身性不够,如图 6 (a) 所示,黑白贴纸在交通道路标志上很明显。为了使贴纸攻击更具隐身性, Wei 等^[8]提出了一种真实贴纸攻击 (Adv-sticker) 技术,该技术主要通过基于区域的差分进化和三维位置变换算法来寻找最优攻击位置。因为使用的是生活中常见的防晒贴,所以在某些场合不会引起人的怀疑,如图 6 (b) 所示。为了提高贴纸的多样性, Shen 等^[9]提出了使用对抗贴纸的 Face-Adv, 设计了一种由贴纸生成器和转换器组成的对抗贴纸生成架构,其中生成器可以生成多种形状的贴纸,转换器将贴纸以数字方式粘贴到人脸上,并向生成器提供反馈以提高效率,攻

击效果如图 6 (c) 所示。为了提高扰动的不可见性, Yin 等^[28]提出了一种通用的对抗化妆攻击 (Adv-makeup) 方法,在人眼区域产生难以察觉的眼影,攻击效果如图 6 (d) 所示。



图 6 典型的物理域对抗攻击
Figure 6 Typical physical domain adversarial attacks

为了提高对抗补丁的泛化能力, Liu 等^[29]利用模型的感知和语义偏差来生成与类别无关的通用对抗补丁。Huang 等^[30]提出了一种通用物理伪装攻击方法,该方法通过欺骗区域提议网络和误导分类器回归输出错误联合来生成对抗补丁,可以攻击所有对象中属于同一类别的对象。为了解决射击视点移动时对抗补丁性能下降的问题, Hu 等^[31]提出了一种基于环面裁剪的可扩展生成攻击,可以生成具有重复结构的对抗纹理。然后使用对抗纹理生成 T 恤、裙子、连衣裙等,这样穿着这些衣服的人就可以从不同的角度实现对行人检测器的“隐形”效果。

基于光学的物理域对抗攻击属于物理域对抗攻击的子类,是针对计算机视觉系统的一类攻击,特别是那些依赖光学输入的系统,如摄像头和基于图像的传感器。光学对抗攻击的目的是操纵或修改视觉输入,从而欺骗或误导计算机视觉算法。

不可感知性是基于光学的物理域对抗攻击的一大特点也是挑战,如何使人类观察者无法察觉对输入图像的修改。这些细微的变化旨在躲避视觉检测,使人类难以发现操纵行为。本文重点研究了计算机视觉系统中基于光学的物理域对抗攻击技术。

2 基于光学的物理域对抗攻击技术

物理域对抗攻击均采用非透明扰动叠加目标特征,这些目标特征在对抗样本中更为突出,容易引起人们的注意,属于侵入性攻击,如图 6 (a)、图 6 (b)、图 6 (c) 所示。侵入性攻击的主要过程是:攻击者首先训练目标模型,得到最佳攻击

位置和最优扰动，然后将非透明扰动贴在最佳攻击位置上。然而，在现实世界中，非透明的扰动不易隐藏，可能在攻击发生之前就被自然环境或人为因素所破坏。

与侵入性攻击不同，非侵入性攻击可以很好地将扰动隐藏到目标特征中，不会引起人类的注意。基于光学的物理域对抗攻击是非侵入性攻击的一种很好的形式，如使用可见光或光学设备（如摄像机和投影仪）产生的光学扰动。由于这种扰动引起的扰动在生活中更为常见，人类依靠生活经验来提示大脑选择性地忽略它们。计算机系统必须要有光学感知模块，这为物理域对抗攻

击留下了“漏洞”。攻击者可以利用摄像机的成像原理和图像传感器处理器的特性进行攻击。在现有研究的基础上，将光学对抗攻击分为对抗性光和干扰相机成像，如表1所示。

2.1 基于对抗性光的光学物理域对抗攻击

入侵攻击的非透明扰动很容易被自然环境或人为因素破坏。为了提高扰动对外部因素的鲁棒性，提出了一些利用发光设备产生光学扰动的方法。例如，Zhu等^[38]设计多灯泡面板来攻击红外热成像探测器。Zhou等^[11]使用红外LED将红外照明扰动投射到人脸上来欺骗人脸识别系统。虽然红外照明扰动对人眼来说是不可见的，但攻击

表1 基于光学的物理域对抗攻击分类
Table 1 The classification of optically based physical domain adversarial attack

类别	方法	产生扰动的物理工具	不可感知性	来源
对抗性光	LightAttack ^[32]	激光点	×	Arxiv 2018
	Invisible mask ^[11]	LED	×	Arxiv 2018
	VLA ^[33]	可见光	√	ACM 2019
	ALPA ^[34]	投影仪	×	CVPR 2020
	Sun等 ^[35]	激光	×	USENIX 2020
	GhostImage ^[36]	投影仪	×	USENIX 2020
	SLAP ^[37]	投影仪	×	USENIX 2021
	BulbAttack ^[38]	灯泡面板	×	AAAI 2021
	Pony等 ^[39]	LED	×	CVPR 2021
	AdvLB ^[12]	激光	×	CVPR 2021
	OPAD ^[40]	投影仪	√	CVPR 2021
	AdvLS ^[41]	激光光斑	×	ACML 2022
	DoubleStar ^[42]	投影仪	×	USENIX 2022
	AdvNB ^[43]	霓虹灯光	×	Arxiv 2022
	ShadowAttack ^[44]	阴影	×	CVPR 2022
	SPAA ^[45]	投影仪	√	VR 2022
	Daimo等 ^[46]	投影仪	×	IEICE 2023
	Wen等 ^[47]	投影仪	×	ICASSP 2023
	Li等 ^[48]	投影仪	×	CVPR 2023
干扰相机成像	Adverial camera stickers ^[49]	相机	×	ICML 2019
	Man等 ^[50]	相机	×	S&P 2019
	TranslucentPatch ^[51]	相机	×	CVPR 2021
	LEDAttack ^[52]	LED	√	CVPR 2021
	ISPAttack ^[53]	相机	√	CVPR 2021
	Chen等 ^[54]	LED	√	Inscrypt 2021
	AdvZL ^[55]	相机	×	Arxiv 2022
	SLMAttack ^[56]	相位调制模块	√	Optics Express 2022

者很难将红外 LED 部署到最佳攻击位置,而使用红外截止滤波器可以有效地滤除这些红外扰动。Pony 等^[39]提出一种通过控制 LED 灯攻击动作识别系统的方法。Sun 等^[35]首次提出针对 LPS 的黑盒欺骗攻击。他们操纵激光模拟遮挡模式和稀疏点云远离自动驾驶汽车。Nichols 等^[32]将目光指向攻击目标的特定区域。

Hu 等^[41]提出一种对抗激光光斑 (adversarial laser spot, AdvLS), 它通过遗传算法优化激光光斑的物理参数来进行物理攻击, 利用低成本的激光设备实现了鲁棒性和隐蔽性的物理攻击。Duan 等^[12]提出一种对抗激光束 (adversarial laser beam, AdvLB), 其主要作用是将激光束作为对原始图像的光扰动, 由于该扰动是由激光光源产生的, 因此可以缓解照射装置部署困难的问题。然而, 一束绿色的激光束穿过整个图像, 并成功将这种光扰动添加到目标背景中, 如图 7 (a) 所示。HU 等^[43]提出对抗霓虹灯 (adversarial neon beam, AdvNB), 如图 7 (b) 所示。它可以只向目标前景添加霓虹灯光束, 从而进一步提高了光扰动的不可见性。



图 7 基于光照条件的光学物理域对抗攻击
Figure 7 Optical-physical domain counterattack based on illumination conditions

光束扰动虽然可以缓解器件部署困难的问题,但由于扰动是由人造光源产生的,与自然现象产生的光扰动有一定的区别,因此在某些方面仍然能引起人们的注意。为了解决这一问题,Zhong 等^[44]提出了一种由自然现象产生的物理域对抗攻击,即影子攻击,如图 7 (c) 所示。阴影攻击充分利用了光照的对抗阴影特性,由于阴影在现实世界中更为常见,人类选择性地忽略阴影区域,这种方法不仅可以很好地隐藏现实世界中的阴影扰动,还可以降低人类的警觉性。

Man 等^[36]利用镜头眩光和自动曝光控制,用一个低成本的投影仪远程将对抗模式投射到相机

系统中,以实现分类系统的攻击。Nguyen 等^[34]提出一种相对不受目标面部变化影响的对抗扰动生成方法,对人脸识别系统进行实时光投影攻击。Lovisotto 等^[37]使用投影仪将扰动投射到目标上。Huang 等^[45]通过投影仪将扰动投射到目标上,并提出一种基于投影仪的隐身对抗攻击方法。Wen 等^[47]提出一种利用光投射瞬态的隐形物理域对抗攻击。攻击者利用带有便携式投影仪的无人机,将对抗光模式投射到车辆的后挡风玻璃上,以阻挡自动驾驶系统中的目标探测器。

Zhou 等^[42]利用立体匹配深度估计算法和光学成像中透镜耀斑效应的弱点,提出了一种通过投射来自两个互补光源的纯光来注入虚假障碍物深度的远距离攻击。Daimo 等^[46]探讨当使用视频投影仪将图案光投射到物体上时,用于单目深度估计的 DNN 是否容易受到错误估计的影响。提出一种具有多保真度评估方案的进化对抗攻击方法,该方法可以在黑盒条件下创建对抗样本,同时抑制计算成本。Gnanasambandam 等^[40]提出一种结构化照明的非侵入式攻击,利用低成本的投影仪—摄像机系统改变目标的特征,实现了针对三维物体的有效光学对抗攻击 (optical adversarial attack, OPAD), 如图 8 所示。OPAD 由投影仪的辐射响应 (绿色虚线框) 和摄像机的光谱响应 (红色虚线框) 组成。OPAD 首先利用计算机产生扰动,然后利用投影仪的非线性映射将扰动信号转换为光学信号,最后将光学扰动投射到图像上以改变目标的纹理特征,并利用摄像机捕捉修改后的图像以生成对抗样本。Li 等^[48]利用边缘投影成像系统和三维重建算法来干扰人脸识别系统。

尽管上述光学对抗攻击方法增强了扰动在现实世界中的隐蔽性,但还没有达到人眼完全无法感知的状态。针对这一问题,Shen 等^[33]利用视觉持久性 (persistence of vision, POV) 理论 (即当光的转换率超过 25 Hz 时,人脑将无法处理这些变化) 提出了一种定向可见光攻击。如图 9 所示,该方法使用一个由扰动帧和隐蔽帧组成的交替光源照射人脸,实现了真正的人眼不可感知,如图 9 (c) 所示。但是,通过摄像头获得的对抗样本 (图 9 (d)) 严重改变了目标的纹理特征。

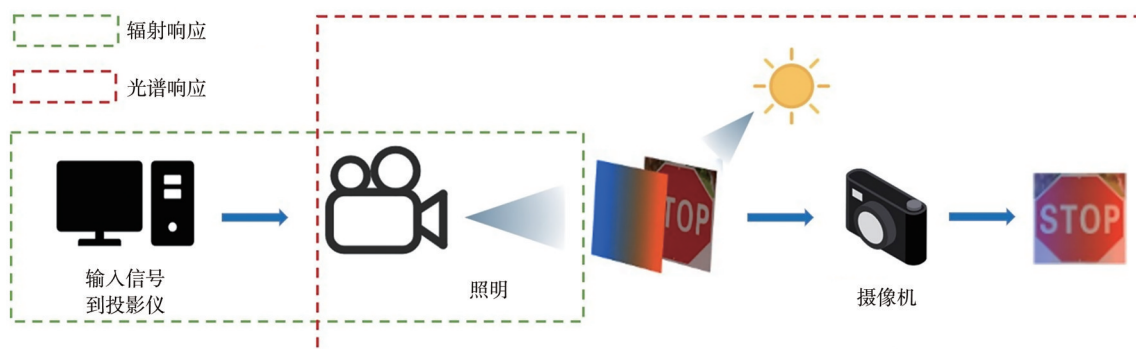


图8 光学对抗攻击

Figure 8 Optical adversarial attack

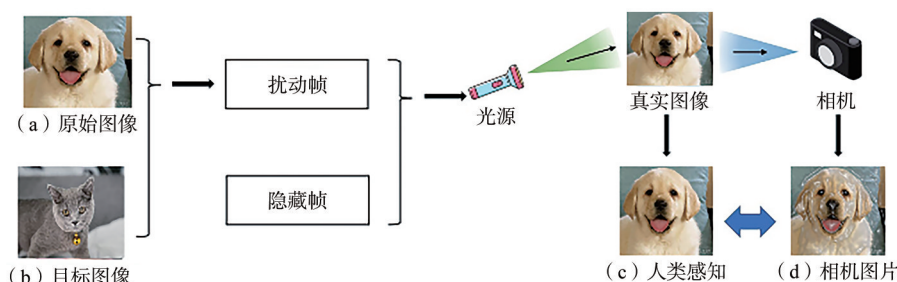


图9 可见光攻击系统

Figure 9 Visible light attack system

2.2 基于干扰相机成像的光学物理域对抗攻击

一种新颖的非侵入式光学物理域对抗攻击是干扰相机成像进行物理操纵以产生扰动。例如，Phan等^[53]直接修改传感器捕获的RWA图像，以欺骗相机的图像处理流水线硬件。Li等^[49]将精心设计的半透明贴纸粘贴在相机镜头上，这样在相机曝光的瞬间，贴纸上的扰动就会投射到目标上，产生物理对抗实例，如图10（a）所示。Zolfi等^[51]使用同样的攻击方法成功攻击了特斯拉X车型的高级驾驶辅助系统。Sayles等^[52]利用卷帘快门效应产生了人眼无法察觉但却被摄像机捕捉到的扰动，如图10（b）所示。Hu等^[55]提出一种对抗变焦镜头攻击（adversarial zoom len, AdvZL），通过改变摄像机的变焦镜头来缩放图像，首次在不对原始样本添加任何扰动的情况下成功攻击目标模型，如图10（c）所示。这些针对摄像头的攻击虽然部署简单，但会明显改变原始图像的纹理和大小，从而引起观察者的怀疑。

另一种基于光学的非侵入式物理域对抗攻击侧重于通过干扰相机成像来影响人工智能视觉系统。例如，Man等^[50]使用光源等电磁信号发生器发

射电磁信号（光）干扰目标传感器捕获的信号（图像）。虽然可见光攻击系统产生的扰动人眼无法察觉，但通过摄像头获取的对抗范例却严重失真，如图11（a）所示。为了利用摄像机获得保真的对抗样本，Kim等^[56]引入了一种新方法，通过修改光学成像系统内部的光场信息来进行非目标对抗攻击。作者设计了一个相位调制模块（图11（b）来生成对抗样本。相位调制模块由偏振片、中继透镜、分光镜和空间光调制器组成。如图11（c）所示，该方法实现了完美的扰动不可见性，在摄像机捕获的对抗样本中不会失真。然而，目前的技术很难模拟出合格的相位调制模块，因此很难部署实验装置。

Chen等^[54]提出一种基于LED照明调制的新型攻击方案，该方案不仅能在现实世界中产生人眼无法感知的扰动，而且在实验装置中的部署也很简单。从图12（a）中可以看出，基于LED照明调制产生的扰动是人眼无法感知的；从图12（b）中可以看出，摄像头捕捉到的对抗样本包含大量黑色条纹，这是CMOS摄像头成像机制中的卷帘快门效应和LED照明调制相互作用产生的光扰动。

由于LED照明的“开/关”状态变化间隔很短，且闪烁频率超出了人眼的观察极限，因此人眼无法检测到LED照明调制引起的快速闪烁和明暗条纹扰动。



图10 针对摄像机的物理域对抗攻击
Figure 10 Physical-domain adversarial attacks against the camera

3 光学对抗防御技术

本文将回顾计算机视觉领域有关光学对抗防御的文献。为了帮助读者全面了解物理域对抗防御，从DNN模型的3个阶段（即预处理、训练过程和训练后）来组织防御方法。具体来说，在3.1节中，讨论了预处理防御的机制。在3.2节中，将训练过程防御方法分为对抗训练、架构修改和认证防御。3.3节讨论了训练后防御方法。此

外，表2列出了参考光学对抗防御方法的分类汇总。从表中可以看出，现有的光学对抗防御主要是针对对抗补丁提出的。这是合理的，因为对抗补丁是目前物理域对抗攻击的主要形式。

3.1 预处理

数据预处理可以减少物理扰动的影响。目前的预处理方法可分为图像补全和图像平滑。

3.1.1 图像补全

图像补全防御^[63,83-84]首先通过对抗斑块与其相邻区域之间的差异来定位斑块区域。然后，防御者可以根据本地化的区域完成这些区域。这样，对抗攻击的效果就会被移除。

数字水印^[57]使用显著性图的大小来检测异常区域，然后将其从输入中屏蔽掉。SentiNet^[59]利用梯度加权类激活映射（Grad-CAM）来获取掩码。为了更有效地遮蔽补丁区域，分割完整（segment and complete, SAC）算法^[63]和少数报告（minority report, MR）^[60]算法利用了分割技术。补丁分割器会预测初始不规则的补丁区域，并通过形状补全器^[62]或 3×3 区域投票^[60]对其进行补充。因此，掩码覆盖的区域可以恢复。最近，TaintRadar^[62]发现对抗区域与类的等级之间存在很强的相关性。它通过3次前向传播和两次后向传播交替使

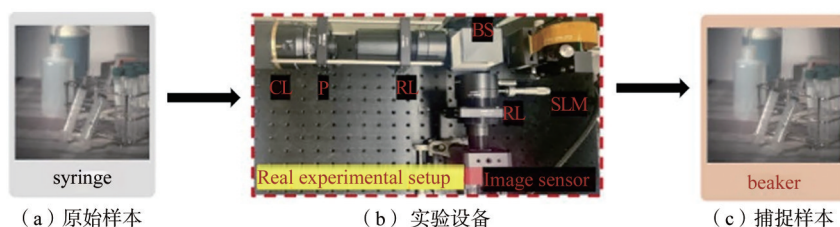


图11 用于光学对抗攻击的工程瞳孔功能
Figure 11 Engineered pupil function for optical counter attacks

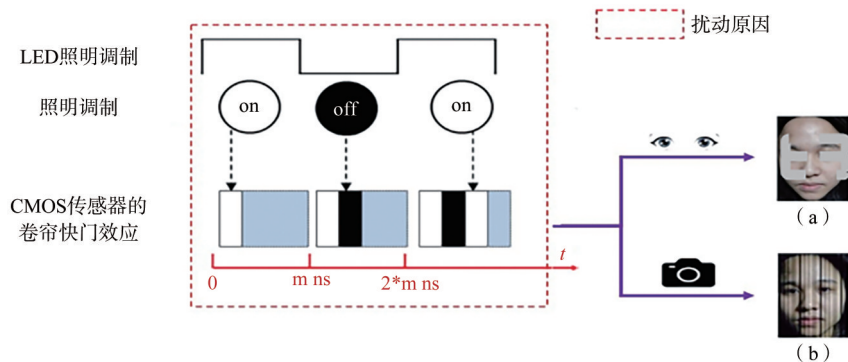


图12 基于LED照明调制的新型攻击方案
Figure 12 Novel attack scheme based on LED lighting modulation

表2 光学对抗防御方法分类
Table 2 The classification of optical adversarial defense methods

类别	方法	模式	攻击形式	来源
预处理	DW ^[57]	经验性防御	对抗补丁	CVPRW 2018
	LGS ^[58]	经验性防御	对抗补丁	WACV 2019
	SentiNet ^[59]	经验性防御	对抗补丁	SPW 2020
	MR ^[60]	经验性防御	对抗补丁	ACNS 2020
	BlurNet ^[61]	经验性防御	对抗补丁	DSN-W 2020
	TaintRadar ^[62]	经验性防御	对抗补丁	INFOCOM 2021
	SAC ^[63]	经验性防御	对抗补丁	CVPR 2022
	DIFFender ^[64]	经验性防御	对抗补丁	Arxiv 2023
训练过程	ROA ^[65]	经验性防御	对抗补丁	Arxiv 2019
	AT-LO ^[66]	经验性防御	对抗补丁	ECCV 2020
	IBP for patch ^[67]	验证性防御	对抗补丁	ICLR 2020
	Clipped BagNet ^[68]	验证性防御	对抗补丁	SPW 2020
	DeRS ^[69]	验证性防御	对抗补丁	NeurIPS 2020
	Ad-YOLO ^[70]	经验性防御	对抗补丁	Arxiv 2021
	MAT ^[71]	经验性防御	对抗补丁	ICML 2021
	RSA ^[72]	验证性防御	对抗补丁	ICML 2021
	PatchGuard ^[73]	验证性防御	对抗补丁	ICLR 2021
	PatchVote ^[74]	验证性防御	对抗补丁	Arxiv 2021
	BAGCERT ^[75]	验证性防御	对抗补丁	ICLR 2021
	ScaleCert ^[76]	验证性防御	对抗补丁	NeurIPS 2021
	MultiBN ^[77]	经验性防御	对抗补丁	TIP 2021
	DeRS-ViT(Chen) ^[78]	验证性防御	对抗补丁	CVPR 2022
	DeRS-ViT(Salman) ^[79]	验证性防御	对抗补丁	CVPR 2022
	Demasked Smoothing ^[80]	验证性防御	对抗补丁	Arxiv 2022
训练后	KEMLP ^[81]	经验性防御	对抗补丁	ICML 2021
	DDDM ^[82]	经验性防御	对抗补丁	Arxiv 2022

用负掩码来对抗补丁攻击。在前向传播过程中观察到的 Top- K logits 的等级变化可以在后向传播过程中指导负掩码的生成，在后向传播过程中生成负掩码， K logits 相当于 K 个负掩码。

在数字领域，研究人员主要通过训练图像压缩模型^[85]或扩散模型^[86]来防御不可见噪声。在物理领域，DIFFender^[64]首次利用扩散模型来防御潜在的恶意补丁攻击。该方法设计两阶段补丁防御框架：补丁定位和恢复。在定位阶段，他们发现扩散模型在恢复添加了高斯噪声的敌对补丁图片的正常区域和补丁区域的能力上存在差异。因此，可以有效地识别出对抗补丁的位置。在恢复阶段，DIFFender采用文本引导的扩散模型来消

除图像中的对抗区域，同时保持视觉内容的完整性。

3.1.2 图像平滑

随机平滑法^[87]在对抗样本中加入不同大小的高斯噪声。然后通过基础分类器对所有有噪声的对抗样本进行分类，并将这些分类的投票结果作为结果报告。随机平滑（De randomized smoothing, DeRS）法^[69]使用固定宽度图像消融法生成平滑图像（如图13所示）。通过这种方法，可以计算出受对抗补丁影响的平滑图像数量。这种定量框架有利于提高模型的鲁棒性。ViT模型也采用这一框架来防御补丁攻击^[78-79]。局部梯度平滑（local gradients smoothing, LGS）^[58]可直接平滑

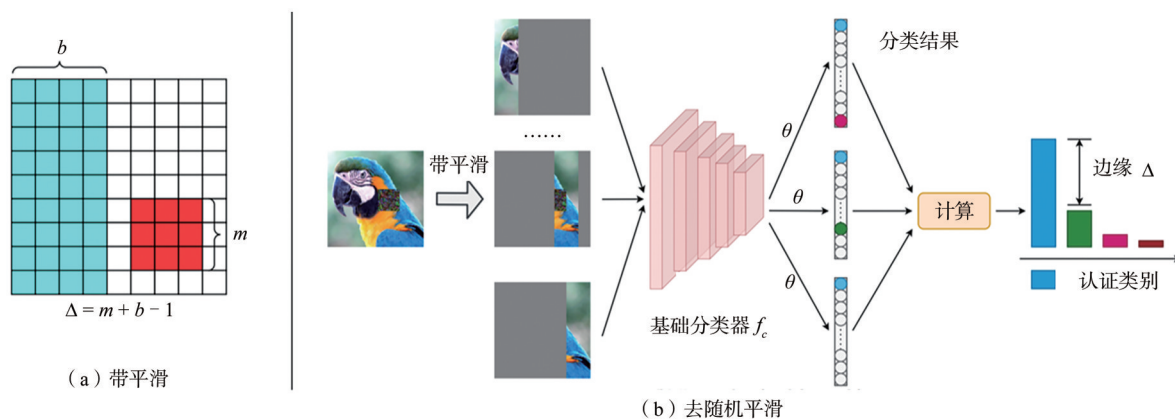


图 13 随机平滑的介绍

Figure 13 Presentation of stochastic smoothing

补丁区域，它将图片分割成 5×5 补丁，并应用滑动窗口来寻找激活度最高的区域。然后，对这些区域进行平滑处理，以消除对抗补丁。以往的图像平滑策略无法保证图像的完整性。因此，它们并不适用于图像分割任务。去掩码平滑法^[80]利用增量 transformer 结构^[88]重建消融图像。这样，分割模型就能在对抗范例上获得与干净图像相同的预测结果。

此外，在图像的频域中，文献[61]发现对抗攻击通常会引入高频噪声。因此，该研究提出了一种 BlurNet 模型，在网络第一层的特征图中引入低通滤波。低通滤波可以去除高频噪声，从而抵御 RP2 攻击。

3.2 训练过程

预处理防御通常是耗时的。因此，有必要提高 DNN 模型本身的鲁棒性。现有的方法主要从对抗训练、结构修改和认证鲁棒性 3 个方面进行。

3.2.1 对抗训练

对抗训练使用对抗样本作为训练数据，以提高模型的鲁棒性^[89-91]。对抗训练的目标是恢复对抗样本的准确预测，可以通过以下方法进行优化：

$$\min_{\theta} E_{(x,y) \sim D} \left[\max_{\Delta x \in \Omega} L(x + \Delta x, y; \theta) \right] \quad (2)$$

其中， D 代表训练集， x 代表输入， y 代表标签， θ 是模型参数， $L(x, y; \theta)$ 是单个样本的 loss， Δx 是扰动， Ω 是扰动空间。

研究人员使用物理对抗样本来训练目标 DNN 模型。Wu 等^[65]利用对抗训练 (adversarial

training, AT) 来防御矩形遮挡攻击，这种攻击用于攻击交通标志。Rao 等^[66]发现，仅仅使用固定位置的对抗补丁不足以对抗位置优化的对抗补丁。因此，采用全位置优化（4 个方向）和随机位置优化（随机方向）来优化补丁位置。在单一类别补丁上训练的模型存在缺陷，因此，需要对补丁的内容进行优化。元对抗训练 (meta adversarial training, MAT)^[71]将对抗训练与元学习相结合，以提高对抗训练模型的泛化能力。它使用不同步长的元补丁在原始图像中生成对抗范例。如图 14 所示，MAT 可以对不同的对抗范例做出正确的预测。补丁的对抗训练对大模型的计算开销大，对小模型的训练效果差，这限制了其在实际场景中的大规模应用。Zhao 等^[92]采用多教师模型对抗鲁棒蒸馏法指导小型模型的对抗训练过程，以平衡计算开销和模型鲁棒性。

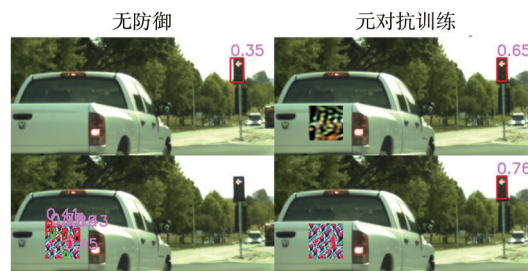


图 14 在小型交通灯数据集上训练的 MAT 模型

Figure 14 MAT model trained on small traffic light dataset

3.2.2 结构修改

研究人员可以通过修改网络结构来削弱对抗补丁对最终分类的影响。现有的方法主要来自两个方面：缩小感受域，引入新的神经元。

(1) 缩小感受域的灵感来自 BagNet^[93]，它采用 ResNet-50 的结构。BagNet 将 3×3 卷积核替换为 1×1 卷积核，从而缩小感受域的大小。此外，BagNet 从每个小图像补丁中提取特征，生成每个补丁的对数并计算平均值。Clipped BagNet^[68] 和 PatchGuard^[73] 都使用 BagNet 作为骨干网络。为了进一步确保模型的鲁棒性，Clipped BagNet 在平均异常对数之前人为地放弃了异常对数，而 PatchGuard 则采用了鲁棒屏蔽聚合来检测和屏蔽异常特征。

ViT 的自我关注机制还限制了“接受域”。对抗补丁只能影响部分令牌。PatchVote^[74] 使用 16×16 补丁大小的 ViT 模型^[94] 对不同补丁的预测进行投票。

(2) 当模型结构发生变化时，引入新的神经元可以有效地解决模型的优化问题。Ad-YOLO^[70] 在 YOLO-v2 的最后一层增加了“补丁”类，而其他层保持不变。这样，Ad-YOLO 就能同时识别对抗补丁并恢复正确预测。MultiBN^[77] 针对多种类型的对抗视频（包括物理对抗视频）引入了一种动态批量归一化（batch normalization BN）层结构，用于视频识别。对于特定的对抗视频，可学习的 BN 层模块会选择相应的 BN 层，从而提供更精确的分布估计。这种 BN 层选择机制有利于提高防御模型的泛化能力。AW-Net^[95] 根据对抗检测器产生的调节信号动态调整网络权重，这直接受到输入样本的影响。这种动态结构有可能提高模型的准确性和鲁棒性。Han 等^[76] 提出一种基于表层重要神经元（易受干扰的神经元）的鲁棒性模型 ScaleCert。ScaleCert 会在推理过程中屏蔽这些神经元，以消除斑块的影响。为了提高 ViT 的鲁棒性，鲁棒自注意力（robust self-attention, RSA）^[72] 采用了鲁棒聚合机制，对所有标记进行评分并剔除异常值，以削弱补丁攻击的影响。Metzen 等^[75] 提出了 BagCert 防御方法，通过区域评分器（核大小为 3、1 和 3 的 3 层 CNN）和空间聚合器对图像进行分类。

3.2.3 认证鲁棒性

上述经验防御缺乏对其鲁棒性的理论保证。为了解决这个问题，研究人员提出了认证防御，并证明其模型在不同情况下是鲁棒的。当输入和

输出 (x, y) 满足以下条件时，模型具有认证的鲁棒性：

$$m_y = Z_{y_{true}}^{(K)} - Z_y^{(K)} \geq 0 \quad \forall y \quad (3)$$

在这个判别式中， m_y 是^[96]中的认证距离边界。 $Z_y^{(K)}$ 表示第 k 层的输出。这意味着这类输出的下限总是大于（至少等于）任何其他输出的上限 y 。

第一个经过认证的防御方法是基于区间边界传播（Interval bound and Propagation, IBP）^[97]，该方法常用于数字防御。后来，Chiang 等^[67] 将 IBP 扩展到对抗补丁的防御。该模型输出一个可微分的下限（可验证半径），即真实标签与其他标签之间的距离。因此，可以对模型进行训练以增加距离。式（4）可以计算边界距离：

$$\left(\mathbf{e}_{y_{true}} \mathbf{e}_y \right)^T Z^{(K)} = m_y \geq 0 \quad \forall Z^{(0)} \in B(x_0) \quad \forall y \quad (4)$$

其中， $\mathbf{e}_i$ 是第 i 个基向量， $B(x_0)$ 是输入图像可能范围的约束集合。

Cohen 等^[87] 使用随机平滑法确保模型在数字攻击中的认证鲁棒性。为了在补丁防御中应用随机平滑法，Levine 等^[69] 提出了随机平滑法，并在部分区域外进行平滑。从理论上讲，当定量边界的最大值超过第二大分类结果时，模型的鲁棒性就得到了验证：

$$n_c(X) > \max_{c' \neq c} n_{c'}(X) + 2\Delta \quad (5)$$

其中， $n_c(X)$ 为基础分类器返回类别的次数， X 为输入图像， c 为图像类别， c' 为其他类别， Δ 为攻击者改变分类器输出的最大概率。

在极端情况下，所有的 Δ 图像都会被错误地分类到错误的类别中，但具有真实类别的平滑图像数量仍然是最多的。最近，Demasked Smoothing^[80] 将图像平滑与区域重建相结合，以保持分割中输入的完整性。

BagNet 是认证防御领域的一个重要模型。Clipped BagNet^[68] 证明，只要在特征聚合阶段之前剪切掉足够多的异常块，就能实现认证鲁棒性。PatchGuard^[73] 评估所有可能的掩码预测。如果所有预测结果是一致的，那么最终预测结果将获得认证的鲁棒性。BAGCERT^[75] 通过端到端的区域评分器和空间聚合器将模型训练和稳健性验证结合起来。区域评分器为图像的真实和非真实类别

给出线性区域评分。然后，空间聚合器计算这两个分数之间的差值，可用于鲁棒性认证和训练。

随着 ViT^[94] 的开发，认证防御系统已开始使用 ViT 作为基础分类器。PatchVote^[74] 使用预先训练好的 ViT-B/16 作为骨干模型。PatchVote 通过操纵 ViT 的自我关注和对不同标记的预测进行投票，对输入图像进行认证预测。验证结果通过对所有标记的预测进行投票产生，当所有编码器为同一类预测，则预测得到验证，否则不通过验证。

$$v(x) \triangleq \begin{cases} 1, & f_1(x) = f_2(x) = \dots = f_k(x) = f(x) \\ 0, & \text{其他} \end{cases} \quad (6)$$

如图 15 所示，至少有一个补丁可以完全覆盖对抗补丁，如果所有 Transformer 编码器都同意同一类预测，预测就得到了验证。Salman 等^[79] 和 Chen 等^[78] 分别将 ViT 模型与随机平滑相结合。使用孤立的带单元自注意取代原始 ViT 的全局自注意以加快推理速度。

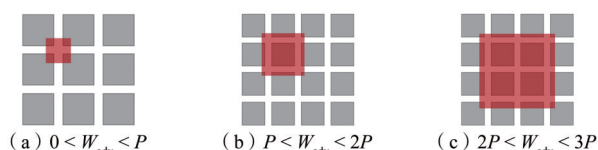


图 15 受不同大小的对抗补丁影响的 ViT 补丁的数量
Figure 15 Number of ViT patches affected by antagonistic patches of different sizes

3.3 训练后

训练后防御机制与人类的思维过程一致^[98]：再验证（reverification），这激发了研究人员进行后处理。在做出初步预测后，模型必须全面分析更多证据。

Gürel 等^[81] 在知识增强机器学习管道（knowledge enhanced machine learning pipeline, KEMLP）中结合了领域知识和对抗防御，通过对机器学习模型的输出和 KEMLP 预测建模来构建因子图，不同的因子与目标具有不同的逻辑关系。以“Stop”路牌为例，“Stop”是充分条件，“Octagon”是必要条件。各因素的逻辑判断将整合领域知识，形成最终输出。这种重新验证过程可以有效地避免将“Stop”误判为“Octagon”等不合逻辑的错误分类。

受到漂移-扩散模型（drift-diffusion model, DDM）^[99] 和滤波方法（dropout method）^[100] 的启

发，Chen 等^[82] 提出了基于滤波的漂移-扩散模型（dropout-based drift-diffusion model, DDDM）。首先，采用滤除法生成目标模型的剪切模型。其次，每个剪切模型的输出都用于阈值累积。当决策证据积累到一定阈值时，就会产生最终的鲁棒输出。

4 数据集及评价指标

现有的面向光学的物理域对抗攻击以及防御，采用相对固定的数据集与评价指标，方便不同方法之间的性能比较，因此本节总结了常用的数据集和评价指标。

4.1 数据集

光学物理域对抗攻击常用的数据集可以分为图像分类数据集、目标检测数据集、人脸识别数据集和深度估计数据集。这些数据集的详细内容如表 3 所示，其中，目标检测数据集包括可见光场景数据集和红外场景数据集。

4.2 评价指标

光学物理域对抗攻击效果需要通过评价指标来衡量。本节介绍光学物理域对抗攻击中的常用评价指标。光学物理域对抗攻击根据任务主要分为图像分类和目标检测。下面根据任务对现有的评价指标进行介绍。

图像分类任务的主要评价指标包括攻击成功率（ASR）、准确率和 F1 评分。

攻击成功率是指攻击者通过对输入数据进行修改，成功欺骗模型并得到错误输出的比率。ASR 是衡量分类器鲁棒性的重要指标之一，因为它可以已知分类器在遭受特定攻击时的容错能力，即攻击者以多大的概率可以获得正确的结果。对于攻击者来说，ASR 越高越容易实施攻击，而对于分类器的开发者来说，ASR 越低代表分类器的鲁棒性越强。ASR 的计算公式如下：

$$ASR = \frac{\text{攻击成功的样本数}}{\text{总样本数}} \quad (7)$$

其中，总样本数是指攻击者生成的所有对抗样本的数量，攻击成功的样本数是指攻击者在这些对抗样本中成功欺骗了分类器并且得到了错误输出的样本数量。

准确率（Accuracy）是图像分类领域较为常见的评价指标，准确率指所有预测中预测正确的

表3 数据集内容
Table 3 The details of datasets

类别	数据集	简介	尺寸/分辨率	大小
图像分类	CIFAR-10 ^[101]	数据集中一共有 60 000 张图像, 50 000 张训练图片和 10 000 张测试图片, 3 通道的彩色 RGB 图像	32 ppi×32 ppi	161 MB
	CIFAR-100 ^[101]	数据集中一共有 60 000 张图像, 50 000 张训练图像和 10 000 张测试图像, 3 通道的彩色 RGB 图像	32 ppi×32 ppi	161 MB
	GTSRB ^[102]	数据集中一共有 51 839 张图像, 分为 39 209 张训练图像和 12 630 张测试图像	15 ppi×15 ppi~250×250 ppi	2 GB
	ImageNet (ISLVR2012) ^[103]	14 197 122 张图片和 21 841 个 Synset 索引	200 px~800 px	144.3 GB
目标检测	LISA ¹	43 007 帧视频, 113 888 张图片	6 ppi×6 ppi~167 ppi×168 ppi/ 1 280 ppi×960 ppi	5 GB
	FLIR ² (红外)	数据集总共包含 14 452 张红外图像, 其中 10 228 张来自多个短视频, 4 224 张来自一个长为 144 s 的视频	114 ppi×520 ppi	11.8 GB
	COCO ^[104]	COCO 数据集标注类型对应任务包括物体检测、关键点检测、实例分割、stuff 分割(没有特定形状的物体), 全景分割人体关键点, 人体密度检测等 COCO 数据集共有 328 000 张图片, 150 万个对象实例, 2 500 000 实例标记	大多数的图片在 300 ppi× 300 ppi 到 600 ppi×600 ppi	83.3 9GB
	Kinetics-400 ^[105]	数据集包含 400 个人类动作类, 每个动作至少有 400 个视频剪辑, 每个剪辑持续大约 10 s, 并且取自不同的 YouTube 视频	长度保持 256 像素, 宽度随长度 等比例缩放	135 GB
	BDD100K ^[106]	100 000 个高清视频, 超过 1 100 h 的驾驶记录, 每个视频大约 40 s, 清晰度为 720 p, 帧率为 30	1 280 ppi×720 ppi	6.42 GB
人脸识别	LFW ^[107]	共 13 233 张人脸图像, 每张图像都被标识出对应的人的名字	250 ppi×250 ppi	1.29 GB
	CelebA ^[108]	名人人脸属性数据集, 其包含 10 177 个名人身份的 202 599 张人脸图片	—	9.55 GB
深度估计	KITTY ^[109]	提供了多个场景的数据, 包括城市道路、乡村道路、高速公路等, 涵盖了各种不同的驾驶情境, 非常精确的传感器标定信息, 包括相机-激光雷达标定、相机-IMU 标定等, 整个数据集由 389 对立体图像和光流图, 39.2 km 视觉测距序列以及超过 200 k 3D 标注物体的图像组成	—	180 GB
	Scene Flow ^[110]	提供了复杂的真实世界场景, 包括城市道路、交通、建筑物等, 这些场景通常包含大量的深度不连续性, 包含 35 454 个训练图像对和 4 370 个测试图像对, 每个图像对都有丰富的深度、光流和像素级标签信息	960 ppi×540 ppi	16 GB

比例, 其计算过程可以被表示为:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} \quad (8)$$

其中, TP 表示实际为正样本被划分为正样本的样本数, TN 表示实际为负样本被划分为负样本的样本数, P 表示为正样本数, N 表示为负样本数。

精确率是指在所有检测出的目标中检测正确的概率。其计算过程可以被表示为:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (9)$$

其中, FP 表示实际为负样本且被分类器划分为

正样本的样本数。

召回率是指所有的正样本中正确识别的概率。其计算过程可以被表示为:

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (10)$$

F1-评分 (F1-score) 是分类问题的一个衡量指标, 是精确率和召回率的调和平均数, 其计算过程可以被表示为:

$$F1 = \frac{2PR}{P + R} \quad (11)$$

目标检测任务的主要评价指标包括精确率、召回率、单类平均精确率 (AP)、平均 AP 值

(mAP) 和交并比 (IoU)。

AP 表示 PR 曲线下与坐标轴的面积, 其横坐标为召回率, 纵坐标为精确率。PR 曲线就是综合考虑了模型要尽可能地召回所有的正样本, 但又不能预判错误太多。

mAP 是从类别的维度对 AP 进行平均, 因此可以评价多分类器的性能。其计算过程可以被表示为:

$$\text{mAP} = \frac{\text{AP}}{\text{num_classes}} \quad (12)$$

mAP 一定在 $[0, 1]$, 且越大越好。该指标是目标检测算法中最重要的一个。

IoU 是“预测的边框”和“真实的边框”交集和并集的比值。

5 基于光学的物理域对抗攻击应用场景

5.1 人脸识别

随着人脸识别在移动支付和身份验证中的应用, 人们越来越重视人脸的安全性。攻击者很容易利用可穿戴物理对抗样本来伪装他人, 这将给人们隐私和财产安全带来严重威胁。大多数研究将攻击分为躲避攻击和冒充攻击^[11]。与无目标攻击一样, 躲避攻击的目标是将一个人的脸误判为其他任何一张脸。攻击者需要将原始类别的概率降到最低。与有针对性的攻击一样, 冒充攻击的目的是欺骗 FR 系统, 使其将一张人脸识别为另一张特定的人脸。为了解决这个问题, 攻击者需要最小化模型输出与目标类别之间的距离。在现实场景中, 攻击者可以在非接触的情况下, 制造一种带有红外 LED 的可穿戴帽子来欺骗人脸识别系统, 通过红外 LED 产生的对抗性光对人脸识别系统进行冒充攻击或躲避攻击^[1]。此外, 攻击者可以使用投影仪将数字域生成的对抗补丁投影到目标人脸上, 实现人脸识别系统的攻击^[40]。

5.2 自动驾驶

确保人们的生命是应用自动驾驶技术的前提。自动驾驶系统需要处理现实世界中的复杂情况。它必须保证汽车行驶在正确的道路上, 避开沿途的障碍物、行人和其他车辆, 并准确地识别交通标志、信号灯和车道线。物理域对抗攻击会在各个方面造成威胁, 挑战最重要的安全原

则。在自动驾驶场景中, 攻击者可攻击的目标包括路牌识别、车辆识别、车牌识别、行人识别等, 攻击者可以操纵投影仪、车辆、无人机等设备对目标车辆进行攻击^[112]。此外, 攻击者可以通过风格迁移、期望转换等技术, 生成更加隐蔽和鲁棒的对抗样本。在自动驾驶场景中, 深度估计也是自动驾驶系统中重要的一环, 攻击者可以通过投射对抗性光到目标车辆的视觉传感器, 以产生错误的估计, 从而造成交通事故。

5.3 安防监控

为了提高社会安全性, 智能安防监控被广泛应用于犯罪追踪、违法行为识别等场景。犯罪分子可以利用物理对抗样本躲避安防监控, 这将造成潜在的安全威胁。真实世界中, 犯罪分子可能使用光学干扰设备 (如激光或霓虹灯等) 来破坏或干扰监控摄像头的视野, 使得监控摄像头无法检测到犯罪分子或者错误检测。除此之外, 犯罪分子还可以使用灯泡面板, 通过优化小灯泡阵列的参数, 间接调整光源的二维高斯分布来干扰红外检测器, 以逃避检测^[32]。

5.4 遥感

遥感与国家安全和公共安全密切相关。遥感图像由无人机或卫星拍摄, 图像质量受到大气条件和相机与物体之间距离的影响。Czaja 等^[113]对航空图像分类器进行攻击, 并提到大气的潜在影响。Adhikari 等^[114]试图伪装飞机和军舰等大型军事资产。然而, 他们并没有在真实世界中进行物理补丁攻击。Du 等^[115]首先在航空图像上实施物理攻击, 他们不仅在车辆顶部生成补丁, 而且还在车外和车周围生成补丁。此外, 他们还考虑了天气和季节的变化, 模拟了不同的条件, 以提高对抗补丁的鲁棒性。攻击者可以向无人机等拍摄设备投射对抗性光, 使无人机无法正常拍摄, 以及在深度估计, 目标检测等任务中产生错误判断。

6 未来研究方向

本节基于光学的物理域对抗攻击, 探讨其在未来的研究发展方向。由于在现实世界中的可行性较低, 数字域对抗攻击不容易对计算机视觉系统产生真正有效的威胁。相比之下, 基于光学的

物理域对抗攻击在现实世界中更具可执行性，并且可能对实际系统产生重大甚至致命的影响。近年来，基于光学的物理域对抗攻击越来越受到人们的关注，基于光学的物理域对抗攻击的文献每年都在增长。然而，基于光学的物理域对抗攻击存在一些困难。①面对复杂的物理环境，当植入的对抗扰动较小时，对抗扰动很容易被相机等采集设备的物理特性所忽略，降低对抗效果，甚至对抗攻击失败，对抗扰动鲁棒性差；②当植入的对抗扰动过大时，会降低对抗扰动的隐蔽性，使得很容易被人类注意到，从而无法实现攻击；③产生对抗扰动的实验装置复杂且操纵困难，很难模拟产生扰动的实验装置，难以迁移复现。

6.1 基于光学的物理对抗扰动的鲁棒性

物理域对抗攻击面临的最大问题是复杂的物理环境，包括亮度、光照、角度、距离、遮挡以及成像传感器（即摄像头）引入的系统噪声。现有的基于光学的物理域对抗攻击方法一般只考虑其中的一些条件，很难将所有潜在的物理环境因素都考虑在内，并且在复杂的物理环境下，微小扰动难以保持对抗效果。在未来研究中应考虑复杂的物理环境，在对抗扰动的优化中考虑这些因素，提升对抗扰动的鲁棒性，以实现在微小扰动中也能保持对抗效果。

6.2 基于光学的物理对抗扰动的隐蔽性

物理域对抗攻击的隐蔽性是指攻击者试图通过添加人眼不可察觉的扰动，或者修改物理系统的某些组件或参数，而不被检测或发现。这与攻击者生成的对抗扰动的方式关系紧密，如表1所示，目前大部分的基于光学的物理域对抗攻击都是以对抗补丁的形式，存在隐蔽性差，容易被发现的缺点，部分方法利用相机成像现象或利用成像器件的漏洞，实现了隐蔽性的对抗攻击。一个完美的物理对抗样本不仅可以使目标模型产生高置信度的错误预测，而且扰动对人眼来说是完全无法察觉的。此外，基于光学的物理域对抗攻击的攻击效果与隐蔽性相冲突，对抗扰动的效果越强则隐蔽性越差，现有的物理对抗扰动两者不能兼顾。因此，在设计对抗扰动时需要考虑隐蔽性损失，如使用风格迁移作为内容损失。此外，设计更新颖的攻击方式，如将对抗性光作为攻击方

式，使用投影仪将对抗性光投射在物理环境中，将对抗扰动作为环境的一部分，实现隐蔽对抗攻击，是未来值得探索的方向。

6.3 基于光学的物理域对抗攻击的实验装置

基于光学的物理域对抗攻击大都涉及复杂的实验装置，产生扰动的实验装置复杂多样，一些实验装置很难模拟并且难以在实际环境中部署。目前使用光学器件发起的对抗攻击更接近于生成一个高对抗性的物理对抗样本，但此类方法的大多数实验装置难以部署。因此开发一种实验装置部署简单的物理域对抗攻击，如借助特定波段的光学成像设备（如激光设备），通过发射人眼不可见的波段的对抗性光，以在攻击时启动，并在非攻击状态下关闭，赋予攻击实时性，是未来值得进一步研究的课题。

7 结束语

随着自动驾驶等自主系统的快速发展，其安全问题也逐渐成为研究热点。本文总结了基于光学的物理域对抗攻击方法和防御方法，总结了光学物理域对抗攻击所用到的常用数据集以及评价指标。同时讨论了当前光学物理域对抗攻击的应用场景，并对未来的研究方向进行讨论，旨在为推动光学物理域对抗攻击的进一步发展和应用提供帮助。

参考文献：

- [1] TAIGMAN Y, YANG M, RANZATO M, et al. DeepFace: closing the gap to human-level performance in face verification[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1701-1708.
- [2] ZENG W Y, LUO W J, SUO S, et al. End-to-end interpretable neural motion planner[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 8652-8661.
- [3] YANG T, YI X L, LU S W, et al. Intelligent manufacturing for the process industry driven by industrial artificial intelligence[J]. Engineering, 2021, 7(9): 1224-1230.
- [4] RONNEBERGER O, FISCHER P, BROX T. U-net: convolutional networks for biomedical image segmentation[C]//International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer, 2015: 234-241.
- [5] LI Y J, XU X, XIAO J H, et al. Adaptive square attack: fooling autonomous cars with adversarial traffic signs[J]. IEEE Inter-

- net of Things Journal, 2021, 8(8): 6337-6347.
- [6] WEI H, TANG H, JIA X M, et al. Physical adversarial attack meets computer vision: a decade survey[EB]. arXiv preprint, arXiv: 2209.15179, 2022.
 - [7] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 1625-1634.
 - [8] 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.
 - [9] SHEN M, YU H, ZHU L H, et al. Effective and robust physical-world attacks on deep learning face recognition systems[J]. IEEE Transactions on Information Forensics and Security, 2021, 16: 4063-4077.
 - [10] PAUL N, CHUNG C. Application of HDR algorithms to solve direct sunlight problems when autonomous vehicles using machine vision systems are driving into sun[J]. Computers in Industry, 2018, 98: 192-196.
 - [11] ZHOU Z, TANG D, WANG X F, et al. Invisible mask: practical attacks on face recognition with infrared[J]. arXiv preprint, arXiv: 1803.04683, 2018.
 - [12] DUAN R J, MAO X F, QIN A K, et al. Adversarial laser beam: effective physical-world attack to DNNs in a blink[C]// Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 16057-16066.
 - [13] SZEGEDY C, ZAREMBA W, SUTSKEVER I, et al. Intriguing properties of neural networks[EB]. arXiv preprint, arXiv: 1312.6199, 2013.
 - [14] DELDJOO Y, NOIA T D, MERRA F A. A survey on adversarial recommender systems: from attack/defense strategies to generative adversarial networks[J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-38.
 - [15] CHAKRABORTY A, ALAM M, DEY V, et al. A survey on adversarial attacks and defences[J]. CAAI Transactions on Intelligence Technology, 2021, 6(1): 25-45.
 - [16] AKHTAR N, MIAN A. Threat of adversarial attacks on deep learning in computer vision: a survey[J]. IEEE Access, 2018, 6: 14410-14430.
 - [17] LIANG H S, HE E L, ZHAO Y Y, et al. Adversarial attack and defense: a survey[J]. Electronics, 2022, 11(8): 1283.
 - [18] AKHTAR N, MIAN A, KARDAN N, et al. Advances in adversarial attacks and defenses in computer vision: a survey[J]. IEEE Access, 2021(9): 155161-155196.
 - [19] ZHANG X W, ZHENG X L, MAO W J. Adversarial perturbation defense on deep neural networks[J]. ACM Computing Surveys, 2022, 54(8): 1-36.
 - [20] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv preprint, arXiv: 1412.6572, 2014.
 - [21] ZHANG J L, LI C. Adversarial examples: opportunities and challenges[J]. IEEE Transactions on Neural Networks and Learning Systems, 2020, 31(7): 2578-2593.
 - [22] CUBUK E D, ZOPH B, SCHOENHOLZ S S, et al. Intriguing properties of adversarial examples[EB]. arXiv preprint, arXiv: 1711.02846, 2017.
 - [23] PAPERNOT N, MCDANIEL P, JHA S, et al. The limitations of deep learning in adversarial settings[C]// Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P). Piscataway: IEEE Press, 2016: 372-387.
 - [24] 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.
 - [25] ATHALYE A, CARLINI N, WAGNER D. Obfuscated gradients give a false sense of security: circumventing defenses to adversarial examples[J]. arXiv preprint, arXiv: 1802.00420, 2018.
 - [26] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]// Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.
 - [27] KURAKIN A, GOODFELLOW I J, BENGIO S. Adversarial examples in the physical world[M]. Boca Raton, FL: CRC Press, 2018: 99-112.
 - [28] YIN B J, WANG W X, YAO T P, et al. Adv-makeup: a new imperceptible and transferable attack on face recognition[EB]. arXiv preprint, arXiv: 2105.03162, 2021.
 - [29] LIU A S, WANG J K, LIU X L, et al. Bias-based universal adversarial patch attack for automatic check-out[C]// European Conference on Computer Vision. Cham: Springer, 2020: 395-410.
 - [30] HUANG L F, GAO C Y, ZHOU Y Y, et al. Universal physical camouflage attacks on object detectors[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 717-726.
 - [31] HU Z H, HUANG S Y, ZHU X P, et al. Adversarial texture for fooling person detectors in the physical world[C]// Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 13297-13306.
 - [32] NICHOLS N, JASPER R. Projecting trouble: light based adversarial attacks on deep learning classifiers[EB]. arXiv preprint, arXiv:1810.10337, 2018.
 - [33] SHEN M, LIAO Z, ZHU L, et al. VLA: a practical visible light-based attack on face recognition systems in physical world[J]. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies, 2019, 3(3): 1-19.
 - [34] NGUYEN D L, ARORA S S, WU Y H, et al. Adversarial light projection attacks on face recognition systems: a feasibility study[C]// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2020: 3548-3556.
 - [35] SUN J C, CAO Y L, CHEN Q A, et al. Towards robust LiDAR-based perception in autonomous driving: general black-box adversarial sensor attack and countermeasures[EB]. arXiv preprint, arXiv: 2006.16974, 2020.
 - [36] MAN Y M, LI M, GERDES R. GhostImage: remote perception

- attacks against camera-based image classification systems[J]. arXiv preprint, arXiv: 2001.07792, 2020.
- [37] LOVISOTTO G, TURNER H, SLUGANOVIC I, et al. SLAP: improving physical adversarial examples with short-lived adversarial perturbations[EB]. arXiv preprint, arXiv: 2007.04137, 2020.
- [38] 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.
- [39] PONY R, NAEH I, MANNOR S. Over-the-air adversarial flickering attacks against video recognition networks[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 515-524.
- [40] GNANASAMBANDAM A, SHERMAN A M, CHAN S H. Optical adversarial attack[C]//Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW). Piscataway: IEEE Press, 2021: 92-101.
- [41] HU C, WANG Y, TILIWALIDI K, et al. Adversarial laser spot: robust and covert physical-world attack to dnns[C]//Asian Conference on Machine Learning. PMLR, 2023: 483-498.
- [42] ZHOU C, YAN Q B, SHI Y, et al. DoubleStar: long-range attack towards depth estimation based obstacle avoidance in autonomous systems[EB]. arXiv preprint, arXiv: 2110.03154, 2021.
- [43] HU C Y, SHI W W, LI W. Adversarial neon beam: a light-based physical attack to DNNs[EB]. arXiv preprint, arXiv: 2204.00853, 2022.
- [44] 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]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 15324-15333.
- [45] HUANG B Y, LING H B. SPAA: stealthy projector-based adversarial attacks on deep image classifiers[C]//Proceedings of the 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR). Piscataway: IEEE Press, 2022: 534-542.
- [46] DAIMO R Y, ONO S. Projection-based physical adversarial attack for monocular depth estimation[J]. IEICE Transactions on Information and Systems, 2023, 106(1): 31-35.
- [47] WEN H, CHANG S, ZHOU L. Light projection-based physical-world vanishing attack against car detection[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023: 1-5.
- [48] LI Y J, LI Y Q, DAI X L, et al. Physical-world optical adversarial attacks on 3D face recognition[EB]. arXiv preprint, arXiv: 2205.13412, 2022.
- [49] LI J C, SCHMIDT F R, KOLTER J Z. Adversarial camera stickers: a physical camera-based attack on deep learning systems[J]. arXiv preprint, arXiv: 1904.00759, 2019.
- [50] MAN Y, LI M, Gerdes R. Poster: perceived adversarial examples[C]//IEEE Symposium on Security and Privacy. 2019.
- [51] ZOLFI A, KRAVCHIK M, ELOVICI Y, et al. The translucent patch: a physical and universal attack on object detectors[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 15227-15236.
- [52] SAYLES A, HOODA A, GUPTA M, et al. Invisible perturbations: physical adversarial examples exploiting the rolling shutter effect[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 14661-14670.
- [53] PHAN B, MANNAN F, HEIDE F. Adversarial imaging pipelines[C]//Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2021: 16046-16056.
- [54] CHEN Z J, LIN P X, JIANG Z L, et al. An illumination modulation-based adversarial attack against automated face recognition system[C]//International Conference on Information Security and Cryptology. Cham: Springer, 2021: 53-69.
- [55] HU C Y, SHI W W. Adversarial zoom lens: a novel physical-world attack to DNNs[J]. arXiv preprint, arXiv: 2206.12251, 2022.
- [56] KIM K, KIM J, SONG S, et al. Engineering pupil function for optical adversarial attacks[J]. Optics Express, 2022, 30(5): 6500-6518.
- [57] HAYES J. On visible adversarial perturbations & digital watermarking[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). Piscataway: IEEE Press, 2018: 1678-16787.
- [58] NASEER M, KHAN S, PORIKLI F. Local gradients smoothing: defense against localized adversarial attacks[C]//Proceedings of the 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2019: 1300-1307.
- [59] CHOU E, TRAMÈR F, PELLEGRINO G. SentiNet: detecting localized universal attacks against deep learning systems[C]//Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW). Piscataway: IEEE Press, 2020: 48-54.
- [60] MCCOVD M, PARK W, CHEN S, et al. Minority reports defense: defending against adversarial patches[C]//International Conference on Applied Cryptography and Network Security. Cham: Springer International Publishing, 2020: 564-582.
- [61] RAJU R, LIPASTI M. BlurNet: defense by filtering the feature maps[EB]. arXiv preprint, arXiv: 1908.02256, 2019.
- [62] LI F T, LIU X K, ZHANG X L, et al. Detecting localized adversarial examples: a generic approach using critical region analysis[C]//Proceedings of the IEEE INFOCOM 2021 - IEEE Conference on Computer Communications. Piscataway: IEEE Press, 2021: 1-10.
- [63] LIU J, LEVINE A, LAU C P, et al. Segment and complete: defending object detectors against adversarial patch attacks with robust patch detection[EB]. arXiv preprint, arXiv: 2112.04532, 2021.
- [64] KANG C X, DONG Y P, WANG Z Y, et al. DIFFender: diffusion-based adversarial defense against patch attacks[EB]. arXiv preprint, arXiv: 2306.09124, 2023.

- [65] WU T, TONG L, VOROBAYCHIK Y. Defending against physically realizable attacks on image classification[EB]. arXiv preprint, arXiv: 1909.09552, 2019.
- [66] RAO S, STUTZ D, SCHIELE B. Adversarial training against location-optimized adversarial patches[M]. Cham: Springer International Publishing, 2020: 429-448.
- [67] CHIANG P, NI R, ABDELKADER A, et al. Certified defenses for adversarial patches[EB]. arXiv preprint arXiv:2003.06693, 2020.
- [68] ZHANG Z Y, YUAN B, MCCOYD M, et al. Clipped BagNet: defending against sticker attacks with clipped bag-of-features [C]//Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW). Piscataway: IEEE Press, 2020: 55-61.
- [69] LEVINE A, FEIZI S. (De)randomized smoothing for certifiable defense against patch attacks[J]. arXiv preprint, arXiv: 2002.10733, 2020.
- [70] JI N, FENG Y F, XIE H D, et al. Adversarial YOLO: defense human detection patch attacks via detecting adversarial patches[EB]. arXiv preprint, arXiv: 2103.08860, 2021.
- [71] METZEN J H, FINNIE N, HUTMACHER R. Meta adversarial training against universal patches[J]. arXiv e-preprint, arXiv: 2101.11453, 2021.
- [72] MU N, WAGNER D. Defending against adversarial patches with robust self-attention[C]//ICML 2021 Workshop on Uncertainty and Robustness in Deep Learning. 2021.
- [73] XIANG C, BHAGOJI A N, SEHWAG V, et al. PatchGuard: a provably robust defense against adversarial patches via small receptive fields and masking[C]//30th USENIX Security Symposium (USENIX Security 21). 2021: 2237-2254.
- [74] HUANG Y, LI Y. Zero-shot certified defense against adversarial patches with vision transformers[EB]. arXiv preprint, arXiv: 2111.10481, 2021.
- [75] METZEN J H, YATSURA M. Efficient certified defenses against patch attacks on image classifiers[EB]. arXiv preprint, arXiv: 2102.04154, 2021.
- [76] HAN H S, XU K D, HU X, et al. ScaleCert: scalable certified defense against adversarial patches with sparse superficial layer[EB]. arXiv preprint, arXiv: 2110.14120, 2021.
- [77] LO S Y, PATEL V M. Defending against multiple and unforeseen adversarial videos[J]. IEEE Transactions on Image Processing, 2022, 31: 962-973.
- [78] CHEN Z, LI B, XU J, et al. Towards practical certifiable patch defense with vision transformer[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15148-15158.
- [79] SALMAN H, JAIN S, WONG E, et al. Certified patch robustness via smoothed vision transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 15137-15147.
- [80] YATSURA M, SAKMANN K, HUA N G, et al. Certified defenses against adversarial patch attacks on semantic segmentation[EB]. arXiv preprint, arXiv: 2209.05980, 2022.
- [81] GUREL N M, QI X, RIMANIC L, et al. Knowledge enhanced machine learning pipeline against diverse adversarial attacks [C]//International Conference on Machine Learning. PMLR, 2021: 3976-3987.
- [82] CHEN X Y, LI X Y, ZHOU Y, et al. DDDM: a brain-inspired framework for robust classification[EB]. arXiv preprint, arXiv: 2205.10117, 2022.
- [83] DZIUGAITE G K, GHAHRAMANI Z. A study of the effect of JPG compression on adversarial images[EB]. arXiv, preprint, arXiv: 1608.00853, 2016.
- [84] GUO C, RANA M, CISSE M, et al. Countering adversarial images using input transformations[EB]. arXiv preprint arXiv: 1711.00117, 2017.
- [85] JIA X J, WEI X X, CAO X C, et al. ComDefend: an efficient image compression model to defend adversarial examples[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 6077-6085.
- [86] NIE W, GUO B, HUANG Y, et al. Diffusion models for adversarial purification[EB]. arXiv preprint, arXiv: 2205.07460, 2022.
- [87] COHEN J M, ROSENFELD E, KOLTER J Z. Certified adversarial robustness via randomized smoothing[J]. CoRR, 2019, abs/1902.02918: 1310-1320.
- [88] DONG Q L, CAO C J, FU Y W. Incremental transformer structure enhanced image inpainting with masking positional encoding[C]//Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2022: 11348-11358.
- [89] CARLINI N, WAGNER D. Towards evaluating the robustness of neural networks[C]//Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP). Piscataway: IEEE Press, 2017: 39-57.
- [90] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[EB]. arXiv preprint, arXiv: 1412.6572, 2014.
- [91] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[EB]. arXiv preprint, arXiv: 1706.06083, 2017.
- [92] ZHAO S J, WANG X Z, WEI X X. Mitigating the accuracy-robustness trade-off via multi-teacher adversarial distillation[EB]. arXiv preprint, arXiv: 2306.16170, 2023.
- [93] BRENDLE W, BETHGE M. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet[EB]. arXiv preprint, arXiv: 1904.00760, 2019.
- [94] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: transformers for image recognition at scale[J]. arXiv preprint, arXiv:2010.11929, 2020.
- [95] WEI X X, ZHAO S J. Revisiting the trade-off between accuracy and robustness via weight distribution of filters[EB]. arXiv preprint, arXiv: 2306.03430, 2023.
- [96] WONG E, KOLTER Z. Provable defenses against adversarial examples via the convex outer adversarial polytope[C]//International conference on machine learning. PMLR, 2018: 5286-5295.
- [97] GOWAL S, DVIJOTHAM K D, STANFORTH R, et al. Scalable

- verified training for provably robust image classification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 4842-4851.
- [98] ELSAYED G, SHANKAR S, CHEUNG B, et al. Adversarial examples that fool both computer vision and time-limited humans[J]. Advances in Neural Information Processing Systems, 2018, 31.
- [99] ROITMAN J D, SHADLEN M N. Response of neurons in the lateral intraparietal area during a combined visual discrimination reaction time task[J]. The Journal of Neuroscience, 2002, 22(21): 9475-9489.
- [100] HINTON G E, SRIVASTAVA N, KRIZHEVSKY A, et al. Improving neural networks by preventing co-adaptation of feature detectors[EB]. arXiv preprint, arXiv: 1207.0580, 2012.
- [101] KRIZHEVSKY A, HINTON G. Learning multiple layers of features from tiny images[J]. 2009.
- [102] STALLKAMP J, SCHLIPSING M, SALMEN J, et al. Man vs. computer: benchmarking machine learning algorithms for traffic sign recognition[J]. Neural Networks, 2012, 32: 323-332.
- [103] RUSSAKOVSKY O, DENG J, SU H, et al. ImageNet large scale visual recognition challenge[J]. International Journal of Computer Vision, 2015, 115(3): 211-252.
- [104] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: common objects in context[M]. Cham: Springer International Publishing, 2014: 740-755.
- [105] KARPATHY A, TODERICI G, SHETTY S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2014: 1725-1732.
- [106] YU F, XIAN W, CHEN Y, et al. Bdd100k: a diverse driving video database with scalable annotation tooling[EB]. arXiv preprint, arXiv:1805.04687, 2(5): 6, 2018.
- [107] YI D, LEI Z, LIAO S C, et al. Learning face representation from scratch[EB]. arXiv preprint, arXiv: 1411.7923, 2014.
- [108] LIU Z W, LUO P, WANG X G, et al. Deep learning face attributes in the wild[C]//Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE Press, 2015: 3730-3738.
- [109] GEIGER A, LENZ P, URTASUN R. Are we ready for autonomous driving? The KITTI vision benchmark suite[C]//Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2012: 3354-3361.
- [110] MAYER N, ILG E, HÄUSSER P, et al. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2016: 4040-4048.
- [111] 曹灿, 司强, 游雪松, 等. 针对人脸识别的物理对抗攻击研究综述[J]. 数据与计算发展前沿, 2023, 5(3): 49-65.
- CAO C, SI Q, YOU X S, et al. A survey on physical adversarial attacks towards face recognition[J]. Frontiers of Data & Computing, 2023, 5(3): 49-65.
- [112] 马晨, 沈超, 蔺琛皓, 等. 针对自动驾驶智能模型的攻击与防御[J]. 计算机学报, 2024: 1-22.
- MA C, SHEN C, LIN C H, et al. Attacks and defenses for autonomous driving intelligence models[J]. Chinese Journal of Computers, 2024: 1-22.
- [113] CZAJA W, FENDLEY N, PEKALA M, et al. Adversarial examples in remote sensing[C]//Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems. 2018: 408-411.
- [114] ADHIKARI A, HOLLANDER R D, TOLIOS I, et al. Adversarial patch camouflage against aerial detection[C]//Artificial Intelligence and Machine Learning in Defense Applications II. SPIE, 2020, 11543: 77-86.
- [115] DU A, CHEN B, CHIN T J, et al. Physical adversarial attacks on an aerial imagery object detector[EB]. arXiv preprint, arXiv: 2108.11765, 2021.

[作者简介]



陈晋音 (1982-), 女, 浙江象山人, 浙江工业大学教授, 主要研究方向为人工智能安全、图数据挖掘和进化计算。



赵晓明 (2001-), 男, 安徽阜阳人, 浙江工业大学硕士生, 主要研究方向为人工智能安全、深度学习和物理对抗攻击。



郑海斌 (1995-), 男, 浙江台州人, 浙江工业大学讲师, 主要研究方向为深度学习、人工智能安全和图像识别。



郭海锋 (1977-), 男, 吉林松原人, 浙江工业大学副教授, 主要研究方向为智能交通。