E-mail: jig@aircas.ac.cn
Website: www.cjig.cn
Tel: 010-58887035
中国图象图形学报
JOURNAL OF IMAGE AND GRAPHICS

中图法分类号:TP391.4　　**文献标识码**:A　　**文章编号**:1006-8961(2025)06-2082-38

论文引用格式:Peng Z B, Zhang Y, Dang Y, Chen J Q, Shi Z W and Zou Z X. 2025. Review of physical adversarial attacks against visual deep learning models. Journal of Image and Graphics, 30(6):2082-2119(彭振邦, 张瑜, 党一, 陈剑奇, 史振威, 邹征夏. 2025. 针对视觉深度学习模型的物理对抗攻击研究综述. 中国图象图形学报, 30(6):2082-2119)[DOI:10.11834/jig.240442]

针对视觉深度学习模型的物理对抗攻击研究综述

彭振邦,张瑜,党一,陈剑奇,史振威,邹征夏*

北京航空航天大学宇航学院,北京　102206

摘　要:基于深度学习模型的计算机视觉技术经过十余年的研究,目前已经取得较大的进步,大量成熟的深度学习模型因其领先于传统模型的高精度、快速性特点,广泛用于计算机视觉相关的各类关键领域。然而,研究者发现,向原始图像样本中添加精心设计的微小扰动可显著地干扰深度学习模型的决策结果。这种精心设计的对抗攻击引发了人们对于深度学习模型鲁棒性和可信赖程度的担忧。值得注意的是,一些研究者以日常生活中常见的实体或自然现象为载体,设计了可于实际应用场景中实施的物理对抗攻击。这种具备较高实用性的对抗攻击不仅能够较好地欺骗人类观察者,同时对深度学习模型产生显著的干扰作用,因而具备更实际的威胁性。为充分认识物理对抗攻击对基于深度学习模型的计算机视觉技术的实际应用带来的挑战,本文依据物理对抗攻击设计的一般性流程,对所整理的114篇论文设计的物理对抗攻击方法进行了归纳总结。本文首先依据物理对抗攻击的建模方法对现有工作进行归纳总结,随后对物理对抗攻击优化约束和增强方法进行概述,并对现有工作的物理对抗攻击实施与评估方案进行总结,最后对现有物理对抗攻击所面临的挑战和具备较大潜力的研究方向进行分析与展望,希望能为高质量的物理对抗样本生成方法设计和可信赖的深度学习模型研究提供有参考意义的启发。本综述主页将展示在https://github.com/Arknightpzb/Survey-of-Physical-adversarial-attack。

关键词:物理对抗攻击;一般性设计流程;对抗样本实用性;深度学习;计算机视觉

Review of physical adversarial attacks against visual deep learning models

Peng Zhenbang, Zhang Yu, Dang Yi, Chen Jianqi, Shi Zhenwei, Zou Zhengxia*

School of Astronautics, Beihang University, Beijing　102206, China

Abstract: Deep learning has revolutionized the field of computer vision over the past two decades, bringing unprecedented advancements in accuracy and speed. These developments are vividly reflected in fundamental tasks such as image classification and object detection, where deep learning models have consistently outperformed traditional machine learning techniques. The superior performance of these models has led to their widespread adoption across various critical applications, including facial recognition, pedestrian detection, and remote sensing for earth observation. As a result, deep learning-based computer vision technologies are increasingly becoming indispensable for the continuous evolution and enhancement of intelligent vision systems. Despite these remarkable achievements, the robustness and reliability of deep learning models have come under scrutiny due to their vulnerability to adversarial attacks. Researchers have discovered that by introducing carefully designed perturbations—subtle modifications that may be imperceptible to the human eye—the decision-making processes of these models can be significantly disrupted. These adversarial attacks are not only theoretical con-

收稿日期:2024-07-31;**修回日期**:2024-10-04;**预印本日期**:2024-10-11

*通信作者:邹征夏　zhengxiazou@buaa.edu.cn

基金项目:国家自然科学基金项目(62125102);北京市自然科学基金项目(JL23005);中央高校基本科研业务费专项资金资助

Supported by: National Natural Science Foundation of China (62125102); Beijing Municipal Natural Science Foundation (JL23005)

structs. They have practical implications that can potentially undermine the trustworthiness of deep learning systems deployed in real-world scenarios. One of the most concerning developments in this area is the emergence of physical adversarial attacks. Different from their digital counterparts, physical adversarial attacks involve perturbations that can be applied in the real world using common objects or natural phenomena encountered in daily life. For instance, a strategically placed sticker on a road sign might cause an autonomous vehicle's vision system to misinterpret the sign, leading to potentially dangerous consequences. These attacks are particularly worrisome because they can deceive not only deep learning models but also human observers, thereby posing a more realistic and severe threat to the integrity of computer vision systems. In light of the growing significance of physical adversarial attacks, this paper aims to provide a comprehensive review of the state of the art in this field. By analyzing 114 selected papers, we seek to offer a detailed summary of the methods used to design physical adversarial attacks, focusing on the general designing process that researchers follow. This process can be broadly divided into three stages: the mathematical modeling of physical adversarial attacks, the design of performance optimization processes, and the development of implementation and evaluation schemes. In the first stage, mathematical modeling, researchers aim to define the problem and establish a framework for generating adversarial examples in the physical world. This involves understanding the underlying principles that make these attacks effective and exploring how physical characteristics, such as texture, lighting, and perspective, can be manipulated to create adversarial examples. Within this stage, we categorize existing attacks into three main types based on their application forms: 2D adversarial examples, 3D adversarial examples, and adversarial light and shadow projection. Typically, 2D adversarial examples involve altering the surface of an object, such as applying a printed pattern or sticker, to fool a computer vision model. These attacks are often used in scenarios like natural image recognition and facial recognition, where the goal is to create perturbations that are inconspicuous in real-world settings but highly disruptive to machine learning algorithms. Taking the concept further, 3D adversarial examples consider the three-dimensional structure of objects. For example, modifying the shape or surface of a physical object can create adversarial examples that remain effective from multiple angles and under varying lighting conditions. Adversarial light and shadow projection represents another innovative approach, where the manipulation of light sources or shadows creates perturbations. These attacks are often more challenging to detect and defend against because they do not require any physical alteration of the object itself. Instead, they exploit the way light interacts with surfaces to generate adversarial effects. This method has shown potential in indoor and outdoor scenarios. We also introduce their applications in five major scenarios: natural image recognition, facial image recognition, autonomous driving, pedestrian detection, and remote sensing. In the design phase of performance optimization, existing adversarial attacks mainly face two core problems: reality bias and the high degree of freedom observation. We introduce some solutions and key technologies for these core problems in the existing literature. In the design of implementation and evaluation schemes, we introduce the platforms and indicators used in existing works to evaluate the interference performance of physical adversarial examples. Finally, we discuss the highly promising research directions in physical adversarial attacks, particularly in the context of intelligent systems based on large models and embodied intelligence. This area of exploration can reveal critical insights into how these sophisticated systems, which combine extensive data processing capabilities with interactive and adaptive behaviors, can be compromised by physical adversarial attacks. In addition, studying physical adversarial attacks on hierarchical detection systems that integrate data from multiple sources and platforms holds significant potential. Understanding the vulnerabilities of such complex, layered systems can lead to more robust and resilient designs. Finally, the prospects of advancing defense technology against physical adversarial attacks are crucial. Developing comprehensive and effective defense mechanisms will be essential for ensuring the security and reliability of intelligent systems in real-world applications. We hope to provide meaningful insights for the design of high-quality physical adversarial example generation methods and the research of reliable deep learning models. The review homepage is available at https://github.com/Arknightpzb/Survey-of-Physical-adversarial-attack.

Key words: physical adversarial attacks; general designing process; practicality of adversarial examples; deep learning; computer vision

0 引 言

深度学习在近20年发展中进步显著，这一进步极大地体现在计算机视觉领域方面，众多深度学习模型在图像分类、目标检测等基础任务上取得远超传统机器学习模型的精度和速度，广泛应用于人脸识别、行人检测和遥感对地观测等关键领域的各项任务。可以说，基于深度学习的计算机视觉技术正成为智能视觉系统迭代升级的必备技术。

然而，Szegedy等人(2014)发现，一些人为设计的微小扰动在添加到自然图像中时，尽管人眼无法区分其中差异，但可能会导致面向图像分类任务的深度学习模型产生不正确甚至荒谬的预测结果，Szegedy等人(2014)称这种向原始数据样本中添加精心设计的扰动产生的数据样本为对抗样本(adversarial examples)。此后，许多研究者在目标检测、图像分割等计算机视觉任务上均发现了对抗样本的存在(Xie等，2017)。Sharif等人(2016)为了证实对抗样本在实际应用场景中存在的可能性，首次设计了具备对抗干扰性能的眼镜。他们发现基于深度学习模型的人脸识别系统无法正确识别佩戴上所设计的对抗性眼镜的受试者。这种通过人为调整日常生活中常见物体形态或环境条件来产生物理对抗样本(physical adversarial examples)从而实施干扰的攻击手段称为物理对抗攻击(physical adversarial attack)，其对深度学习模型在实际应用中的鲁棒性和可信赖程度提出严峻的挑战。

当前，许多研究对现有对抗样本的相关工作进行了综述总结，王志波等人(2023)在目标检测、人脸识别、语义分割、图像检索和视觉跟踪5类复杂计算机视觉任务上对现有对抗攻击研究方法进行分类总结。一些研究者针对特定任务场景中的物理对抗攻击，如目标检测(蔡伟 等，2024)、人脸识别(黄诗瑀 等，2023；瞿左珉 等，2024)，或针对特定攻击形态，如基于光学手段(陈晋音 等，2024)的物理对抗攻击进行了归纳总结。然而，上述综述均未对现有的物理对抗攻击以及其中核心挑战与关键技术进行系统性的总结。为了应对物理对抗攻击对深度学习模型应用带来的严峻挑战，本文面向计算机视觉领域，收集2016年至今的114篇物理对抗攻击相关研究论文并对其进行系统性归纳整理。本文组织结构如图1所示。

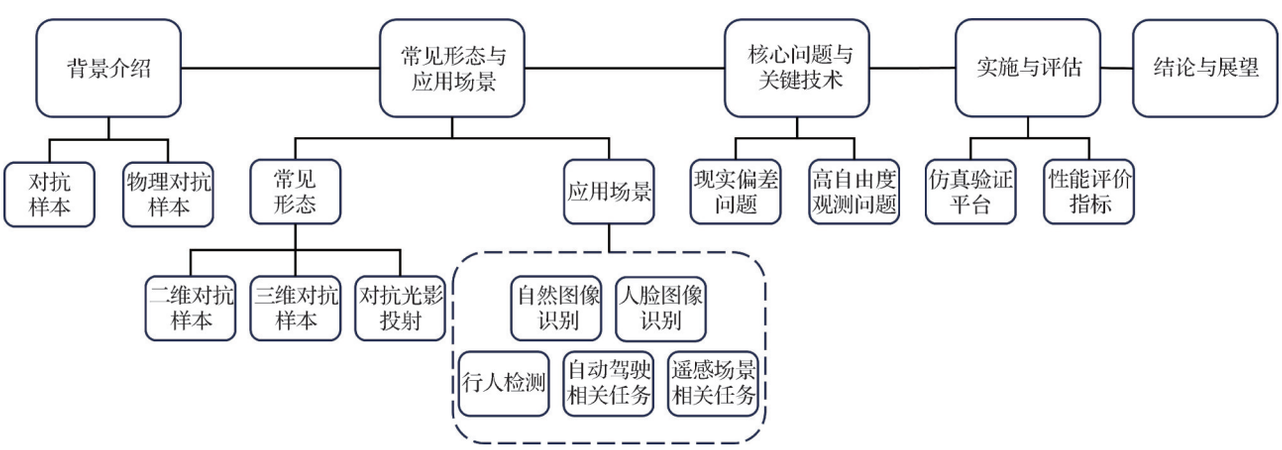


图1 论文组织架构

Fig. 1 Organization of the review

具体而言，本文首先梳理物理对抗攻击的相关定义和一般性设计流程，并将设计流程描述为物理对抗攻击的数学建模、性能优化过程设计和实施与评估方案设计3部分：攻击者首先在数字域中完成对抗样本的应用形态的设计建模，随后使用相关优化算法，设计约束函数并结合相应的增强方法，不断更新对抗样本的相关参数，以提升对抗样本的在实际应用场景中的干扰性能。最后攻击者将依据数字域中的设计建模方法在实际应用场景中完成对抗样本的生产制作，并检验其在实际应用场景中的干扰性能，以评估所设计的物理对抗攻击方法的有效性，如图2所示。随后，本文依据设计流程3部分对现有工作进行归纳综述。在常见形态与应用场景一节中，依据物理对抗样本形态，将现有物理对抗攻击分为

基于二维平面对抗样本构建的攻击方法、基于三维实体对抗样本构建的攻击方法和基于对抗性光影投射的攻击方法3类并对其应用场景进行阐述。在核心问题与关键技术一节中，从物理对抗攻击面临的两大核心问题出发，对现有工作所提出的关键技术进行总结。在物理对抗攻击实施与评估一节中，总结现有工作评估物理对抗样本的干扰性能时所用的平台和指标。最后，通过对现有工作中的物理对抗攻击方法设计一般流程的归纳总结，对现有物理对抗攻击研究中具备较大潜力的研究方向进行分析与展望，并希望通过综述为高质量的物理对抗样本生成方法设计和可信赖的深度学习模型研究提供有参考意义的启发。

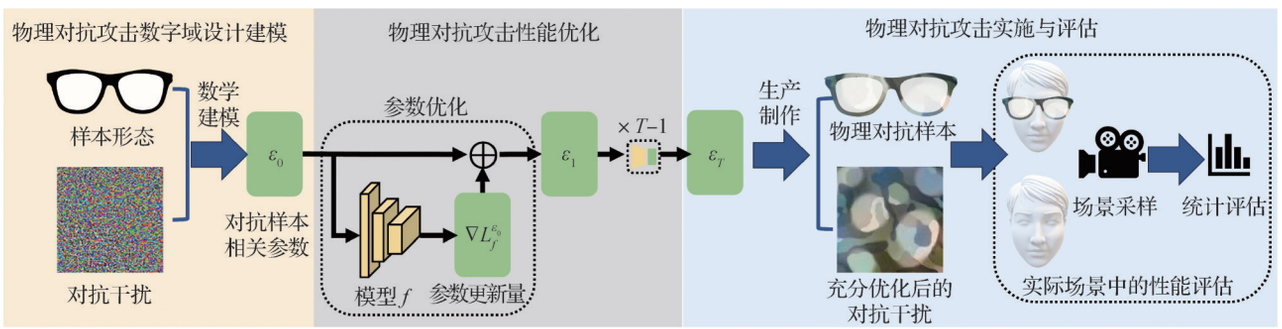


图2　物理对抗攻击一般设计流程示意图

Fig. 2　General design process for physical adversarial attack

1　背景介绍

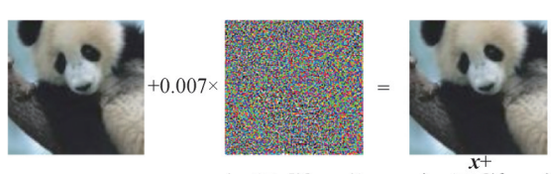
1.1　对抗样本

在Szegedy等人(2014)的工作中，对抗样本的构造可由向图像中添加微小扰动产生，即，对于分类器f，给定高为H宽为W的原始图像样本$\boldsymbol{X}\in[0,1]^{H\times W}$和其对应的真实类别标签$c^{\text{true}}$，在假定分类器可以正确对图像样本$\boldsymbol{X}$进行分类的前提下，攻击者将设法寻找微小扰动满足如下条件，具体为

$$\begin{aligned} f(\boldsymbol{X}+\boldsymbol{\varepsilon})&\neq c^{\text{true}}\\ \boldsymbol{X}+\boldsymbol{\varepsilon}&\in[0,1]\end{aligned}\tag{1}$$

式中，$\boldsymbol{\varepsilon}\in\mathbf{R}^{H\times W}$为攻击者需要寻找的微小扰动。在随后的研究(Goodfellow等，2015)中，他们发现即使是对原始图像样本进行单位像素的扰动，产生的对抗样本也能使得深度学习分类器的预测结果产生显著的偏差，如图3所示。

为了探究这种微小扰动的生成方法，许多工作相继提出。如FGSM(fast gradient sign method)(Goodfellow等，2015)、BIM(basic iterative method)(Kurakin等，2017)、DeepFool(Moosavi-Dezfooli等，2016)、C&W(Carlini and Wagner's method)(Carlini和Wagner，2017)、PGD(projected gradient descent)(Madry等，2019)、MI-FGSM(momentum iterative fast gradient sign method)(Dong等，2018)和B&B(Brendel and Bethge's method)(Brendel等，2019)等算法生成具备对抗干扰性的扰动$\boldsymbol{\varepsilon}$，以实施基于数据微扰动的对抗攻击。上述工作为生成对抗样本提供了切实可行的方案，一些方法也被借鉴至物理对抗攻击的设计中(Xu等，2022)，用于生成相应的物理对抗样本。一般地，对于给定的计算机视觉任务，深度学习模型f将在相应数据集上，通过优化算法优化模型参数，使得目标损失函数L的值尽可能小。最终经过充分训练的深度学习模型f将在验证测试和应用中有良好的表现。为了针对上述模型进行对抗攻击，攻击者可以将扰动$\boldsymbol{\varepsilon}$添加至原始数据样本中，以最大化模型f在施加扰动后的对抗样本上的目标损失函数L为目标来训练$\boldsymbol{\varepsilon}$，使其具备对抗干扰性能。即



$\boldsymbol{x}$
预测类别："大熊猫"
57.7 %置信度

$\text{sign}(\nabla_{\boldsymbol{x}}J(\boldsymbol{\theta},\boldsymbol{x},y))$
肉眼难以觉察的微小扰动

$\boldsymbol{x}+\varepsilon\text{sign}(\nabla_{\boldsymbol{x}}J(\boldsymbol{\theta},\boldsymbol{x},y))$
预测类别："长臂猿"
99.3 %置信度

图3　数据微扰动生成的对抗样本示意图(Goodfellow等，2015)

Fig. 3　Adversarial example generated by data perturbations (Goodfellow et al.，2015)

$$\max L_f^{\boldsymbol{\varepsilon}}\tag{2}$$

研究者通过类似的思路设计了在目标检测、语义分割等领域的对抗性能优化目标函数(Xie 等,2017),用于产生具备对抗干扰性能的扰动。

然而,上述方法大多通过添加微小扰动来产生对抗干扰效果。在实际应用场景中,环境存在的自然条件变化、传感器成像时的误差和杂波以及噪声等因素容易使得微小扰动淹没其中,无法有效地产生对抗干扰效果。另一方面,待攻击的目标模型输入数据通常对于攻击者来说是不可获取的,这使得攻击者无法通过数据微扰动的方法对模型进行对抗攻击干扰。因此,通过数据微扰动方法产生的对抗样本在实际应用中受到较大的限制。

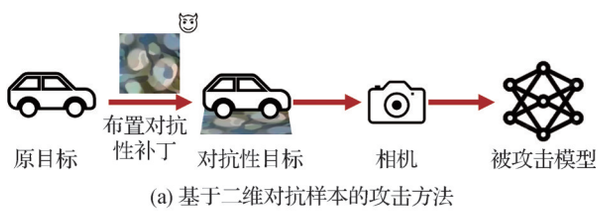
1.2 物理对抗样本

尽管基于数据微扰动方法实施的对抗攻击通常难以在现实场景中直接应用,但其设计思想为物理对抗攻击研究奠定了基础。攻击者通过生成算法,以日常生活中常用的物品,如眼镜(Sharif 等,2016,2019)、贴纸(Wei 等,2023d; Yang 等,2022)、涂装(Zhang 等,2019; Huang 等,2020; Wang 等,2021a)等,为应用形态或改变环境条件(Duan 等,2021; Zhong 等,2022)来生成可在实际应用场景中制作实现的物理对抗样本。相比于数据微扰动方法生成的对抗样本,物理对抗样本兼具了对抗干扰性能和实用性能,为对抗攻击在现实世界中的应用提供了切实可行的方案,将对基于深度学习的计算机视觉技术在实际场景应用的安全性构成一定的威胁。因此,物理对抗样本生成方法研究成为当下可信赖的深度学习技术、深度学习模型鲁棒性等领域的热点问题。

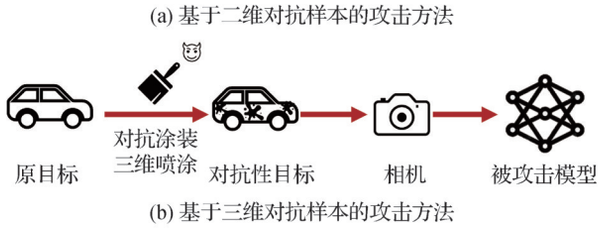
2 常见形态与应用场景

2.1 常见形态

物理对抗攻击在实际场景应用中展现出多种形态,通常划分为3种主要类型:基于二维对抗样本的攻击方法、基于三维对抗样本的攻击方法以及基于对抗光影投射的攻击方法,如图4所示。基于二维对抗样本的攻击方法主要指通过布置依附于图像或平面的对抗样本完成干扰的方法,这类方法通过将具备对抗干扰性纹理的贴纸、补丁或喷漆布置在平坦或有小幅度褶皱、翘曲的平面上,从而对模型产生干扰效果的对抗攻击方法。例如在实际环境中通过在物体表面添加对抗图案或标签来欺骗视觉感知系统。基于三维对抗样本的攻击方法则是进一步考虑实际物理的空间结构,如通过改变物体的三维形状、纹理或在三维空间中添加伪装元素以干扰检测系统。对抗光影投射则在设计时不直接与物理对象接触,攻击者通过利用环境中的光线、反射或其他因素进行对抗样本的生成,达到视觉和谐且攻击不易被察觉的效果。



(a) 基于二维对抗样本的攻击方法



(b) 基于三维对抗样本的攻击方法

(c) 基于对抗光影投射的攻击方法

图4 物理对抗攻击常见形态

Fig. 4 Prevalent types of physical adversarial attack ((a) attack method based on two-dimensional adversarial examples; (b) attack method based on three-dimensional adversarial examples; (c) attack method based on adversarial light/shadow projection)

2.1.1 基于二维对抗样本的攻击方法

基于二维对抗样本的攻击方法指在现实世界中的二维平面或弯曲度较小的表面上,应用具备干扰性的平面实体,实现对深度学习模型干扰的攻击方法。这类对抗攻击的主要应用形态包括对抗性补丁、对抗性贴纸和对抗性配件等。在数字域设计建模过程中,该类方法通过将对抗补丁或配件贴到目标区域,并在优化过程中不断调整这些补丁或配件的特征或位置,从而使其具备有效的干扰攻击效果,如图5所示。二维平面对抗样本在优化时主要关注对抗样本的视觉协调性和攻击性,旨在确保生成的对抗扰动在视觉上尽可能自然、不易察觉,同时能有效迷惑机器学习模型。然而,这种优化过程中通常未充分考虑现实环境中的复杂因素,如视角变化、光

照条件和物体表面的褶皱等。因此，这可能会影响对抗样本在实际应用中的有效性，从而降低其攻击性能。

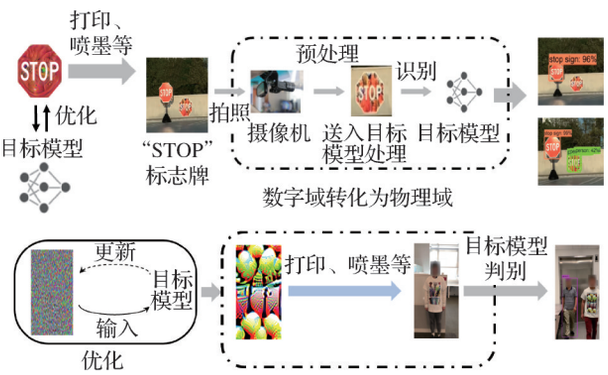


图 5　二维平面对抗样本攻击流程
（Chen 等，2019；Xu 等，2020）

Fig. 5　Workflow of 2D plane adversarial attack
（Chen et al.，2019；Xu et al.，2020）

2.1.2　基于三维对抗样本的攻击方法

基于三维对抗样本的攻击方法指通过对目标物体的三维形状、纹理或表面特征进行精确调整，或通过在物体表面添加具有干扰性质的实体，从而在各种视角下均能对机器学习模型产生有效干扰的攻击方法。这种攻击方法涉及在三维建模工具中对目标物体进行细致的几何和纹理优化，并可能通过物理方式在物体表面添加对抗性物质或标记。在优化过程中，还需要进行广泛的三维渲染仿真，以模拟不同视角、光照条件和环境因素对对抗效果的影响，从而确保扰动在实际应用中能有效迷惑模型，如图 6 所示。

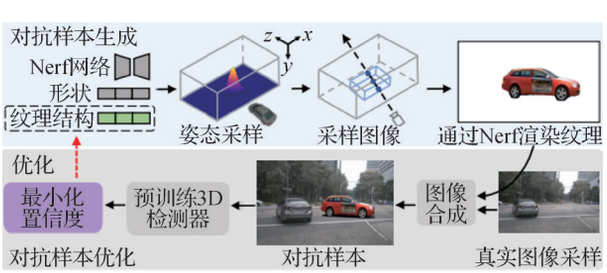


图 6　三维对抗样本攻击流程示意（Li 等，2023b）

Fig. 6　Schematic of 3D adversarial attack workflow
（Li et al.，2023b）

这种攻击方法不仅关注对抗样本在数字空间中的视觉效果和攻击效果，还需综合考虑实际环境中的复杂因素，如视角变化、光照条件、表面材质和物体表面的褶皱等。通过综合考虑这些现实环境因素，确保对抗样本在各种真实场景中都能保持较高的攻击有效性和鲁棒性。因此，基于三维实体形态的物理对抗攻击需要在三维模型优化过程中进行全面且精细的设计，以应对多种现实应用中的挑战。

2.1.3　基于对抗光影投射的攻击方法

基于光影投射的对抗样本与二维、三维对抗样本相比具有一个显著优势，即不需要与目标表面发生直接的物理接触。通常这类对抗攻击通过操控目标表面光照的强度、色调或频率来实现。具体而言，这种攻击方法通过改变光源的特性或在光线传播过程中引入干扰来影响智能系统的识别结果。

在典型的智能系统中，成像过程一般包括以下步骤：首先，光源发出的光线照射在目标物体表面，根据物体表面的材质，部分光线会以漫反射的形式散射开来。接着，成像设备（如相机）捕捉到这些反射光，并通过 CCD（charge-coupled device）或 CMOS（complementary metal oxide semiconductor）传感器将光信号转化为电信号，生成最终的图像（该流程如图 7 所示）。该图像是智能系统输入的关键数据。

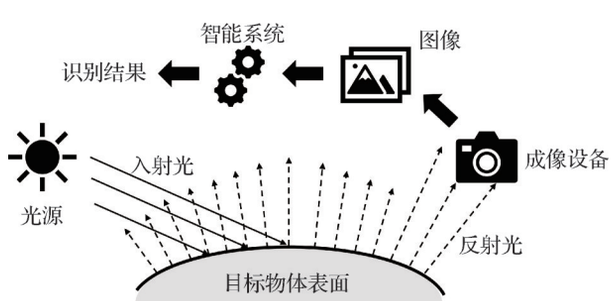


图 7　智能系统识别目标物体流程及中间光照传播

Fig. 7　The process of intelligent system identifying target objects and intermediate light propagation

与主要操控目标物体表面材质的二维和三维对抗样本不同，基于光影投射的对抗样本在光线传播的早期阶段施加干扰。例如可以通过将自然光源替换为人造光源（如投影仪或激光器），或在光线照射目标物体的路径上设置遮挡物或进行波段滤除来实现攻击（如图 8 所示）。这样的方法不仅能够有效干扰智能系统的识别过程，还能在肉眼视角下保持较低的可察觉性，使攻击更加隐蔽。

2.2　应用场景

2.2.1　针对自然图像识别的物理对抗攻击

在物理对抗攻击研究早期，多数研究者主要针

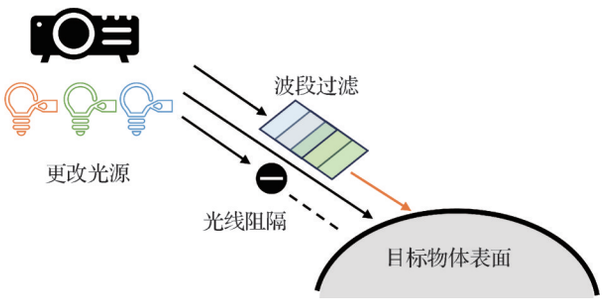


图8　基于光影投射的对抗样本攻击策略

Fig. 8　Adversarial sample attack strategy based on light and shadow projection

对自然图像识别任务，包括分类、检测等。图像分类是指根据图像内容将其归类到特定类别中。目标检测则是在图像分类的基础上进一步扩展和发展，不仅要确定图像中物体的类别，还要精准地定位每个物体在图像中的位置。与图像分类不同，目标检测需要处理更复杂的任务，包括同时识别多个物体和提供每个物体的具体位置。这使得目标检测器必须具备更强的特征提取和多任务处理能力。单阶段检测器通过直接回归边界框和分类结果，实现了较高的检测速度，通常包括YOLO（you only look once）系列（Redmon等，2016）、SSD（single shot multibox detector）（Liu等，2016）和RetinaNET（Lin等，2017）；而两阶段检测器则通过候选框生成和细化处理，达到了更高的检测精度，包括Faster R-CNN（faster region-based convolutional neural networks）（Ren等，2015）和Cascade R-CNN（Cai和Vasconcelos，2018）等。无论是单阶段还是两阶段检测器，都需要处理图像中的大量信息，以提供边界框、置信度和类别的输出，确保在复杂背景下准确地识别和定位目标物体。

目前针对图像分类对抗攻击研究主要集中于研究二维对抗样本的生成方法和干扰性能，在数字域对抗攻击方法的基础上，通过打印对抗图像、对抗性补丁和贴纸等平面实体来探索验证物理对抗攻击的可行性，攻击流程如图9所示。

近年来，部分学者将针对图像分类的对抗样本扩展到三维对抗样本和基于光影生成的对抗样本。由于目标检测经常作为一个子任务在其他场景中出现，故本节对目标检测的物理对抗样本总结聚焦在早期探究如何针对目标检测设计相应物理攻击方法以及设计的物理对抗实体没有限制使用场景，例如透明薄片等（Zolfi等，2021）。表1展示了针对图像识别的物理对抗攻击的经典方法，本文对这些方法的攻击形式以及主要贡献进行总结，以帮助读者清晰理解图像识别系统潜在的安全问题。

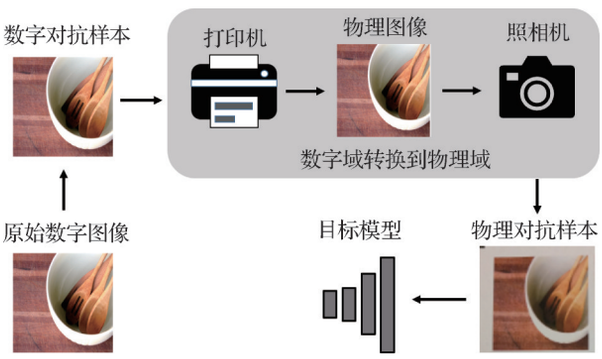


图9　针对图像分类二维物理对抗攻击流程（Kurakin等，2018）

Fig. 9　Workflow of 2D physical adversarial attacks for image classification (Kurakin et al., 2018)

Kurakin等人（2018）首先发现了对抗样本在现实场景中的存在，提出利用迭代快速梯度符号法生成这种对抗样本，将其通过打印制作的方式布置到现实世界中，发现这种对抗样本仍具有攻击效果。

Akhtar等人（2022）指出利用显著的几何图案构建对抗样本对DNN（deep neural networks）也有欺骗效果。早期的物理对抗攻击设计建模思路与数字对抗攻击的设计思路几乎相同，即，设计类别置信度相关损失函数，通过迭代优化的方式，生成具有对抗干扰性的图像来误导模型的决策。不同于数字对抗攻击生成的微扰动，针对图像识别任务的物理对抗攻击产生的对抗扰动大多以补丁、贴纸等可见形式在现实场景中进行制作和应用，如图10所示。

Brown等人（2018）生成一种与图像无关的通用对抗补丁，该补丁可放置在待攻击分类器视野中任何位置进行攻击。无论尺寸或位置如何，其设计的对抗补丁均可对分类器产生较强攻击效果，且在现实中可伪装成日常生活中常见的贴纸。该补丁具体设计中，对抗性能优化过程中的求解目标为

$$\boldsymbol{p}^{*}=\arg\max_{\boldsymbol{p}}\mathbb{E}_{\boldsymbol{x}\sim\boldsymbol{X},t\sim\boldsymbol{T},\boldsymbol{l}\sim\boldsymbol{L}}[\log Pr(y\,|\,A(\boldsymbol{p},\boldsymbol{x},\boldsymbol{l},t))] \tag{3}$$

式中，$\boldsymbol{p}$是对抗补丁，$\boldsymbol{X}$为图像训练数据集，$\boldsymbol{T}$为变换分布，$\boldsymbol{L}$为补丁位置分布，A为补丁的变换布置函数，即将补丁布置到目标图像的对应位置，y为人为指定的分类标签，经过充分训练后获得具备较好对抗性能的补丁$\boldsymbol{p}^{*}$，Pr表示概率。

表 1　针对图像识别的物理对抗攻击方法描述

Table 1　Description of physical adversarial attack methods against image recognition

方法	物理对抗攻击形式	主要贡献
Kurakin 等人（2018）	二维	首次发现物理对抗样本的存在。
Akhtar 等人（2022）	二维	利用几何图案构建物理对抗样本。
Brown 等人（2018）	二维	将对抗贴纸放在物品周围进行攻击。
Eykholt 等人（2018b）	二维	将对抗样本设计为涂鸦形状，约束扰动范围，在现实世界中仅利用黑白贴纸干扰交通标志。
Liu 等人（2019a）	二维	利用生成对抗网络生成具有视觉和谐性的对抗补丁。
Doan 等人（2022）	二维	提出两阶段的优化方法，第 1 阶段利用 GAN 网络将将隐变量映射成为具有常见物品形态的视觉和谐的对抗补丁，第 2 阶段进一步优化隐变量。
Hingun 等人（2023）	二维	第一个针对对抗补丁攻击的大规模标准化安全基准。
Huang 等人（2024）	三维	生成可迁移的 3D 物理对抗样本。
Liu 等人（2019b）	三维	首先将 FGSM 方法扩展到点云领域。
Tang 等人（2023）	三维	引入流形编码器，提出了一种新颖的流形攻击方法，通过参数平面拉伸产生对抗性点云样本。
Lou 等人（2024）	三维	对齐对抗贴纸与人体脸部的拓扑结构，并引入人体重建模型。
Gnanasambandam 等人（2021）	光影投射	对数字对抗纹理至投影仪过程的颜色转变以及投影仪实际投射图案被相机拍摄过程进行显式建模。
Wang 等人（2023a）	光影投射	简化光影攻击过程，仅使用一面镜子、带有颜色的透明滤纸以及裁剪掉特定区域的纸张实现攻击。
Sayles 等人（2021）	光影投射	利用可编程的 LED 灯进行高频信号切换，从而使得成像结果出现带状条纹实现对抗攻击。
Duan 等人（2021）	光影投射	利用激光器对目标物体进行干扰，自定义选则攻击时长以及攻击范围。

Eykholt 等人（2018b）设计了 RP2（robust physical perturbations）算法，针对交通标志分类器进行对抗攻击，从不同距离和角度拍摄到的交通标志图像样本中进行采样，并使用掩膜确定当前生成的扰动中，最具备干扰分类器决策的区域。最后，将产生的对抗样本设计成类似涂鸦的形状，并在现实场景下仅利用黑白贴纸实现了对抗攻击，优化过程中采用的目标函数为

$$\boldsymbol{\delta}^{*}=\underset{\boldsymbol{\delta}}{\operatorname{argmin}}\,\mathbb{E}_{\boldsymbol{x}\sim X}J\left(f\left(\boldsymbol{x}+t\left(\boldsymbol{M}_{\boldsymbol{x}}\cdot\boldsymbol{\delta}\right)\right),y\right)+\lambda\left\|\boldsymbol{M}_{\boldsymbol{x}}\cdot\boldsymbol{\delta}\right\|_{p}+NPS(\boldsymbol{\delta}) \tag{4}$$

式中，$\boldsymbol{M}_{\boldsymbol{x}}$为图像$\boldsymbol{x}$所对应指示对抗补丁$\boldsymbol{\delta}$的布置位置的二进制掩码，$\boldsymbol{X}$为图像训练数据集，$t$为对抗补丁$\boldsymbol{\delta}$的变换布置函数，$y$为待分类图像的标签，$NPS$函数是不可打印得分（non-printability score，NPS）函数，用于衡量样本的可制作性能，经过充分训练后获得具备较好对抗性能的补丁$\boldsymbol{\delta}^{*}$。在具体实验过程中，作者在不同相机距离以及不同视角下验证对抗补丁的干扰性和鲁棒性。此外，作者还录制汽车行驶视频验证了所设计的物理对抗攻击在动态场景中仍然具有一定的对抗干扰性能。

图 10　针对图像分类的二维物理对抗攻击示例（Kurakin 等，2018；Brown 等，2018；Eykholt 等，2018b；Liu 等，2019a；Doan 等，2022）

Fig. 10　Examples of 2D physical adversarial attacks for image classification (Kurakin et al., 2018; Brown et al., 2018; Eykholt et al., 2018b; Liu et al., 2019a; Doan et al., 2022)

不同于对抗扰动迭代优化的生成方法，一些研究者通过训练神经网络模型实现端到端地产生具备干扰性能的对抗样本。与基于迭代优化的对抗扰动

生成方法相比，这种基于模型训练的方法首先在相应的数据集上训练可产生对抗样本的生成器模型。在生成对抗样本过程中，攻击者可直接通过充分训练的对抗样本生成器的前向推理生成种类丰富的对抗样本，无需多次访问待攻击模型，加快了物理对抗样本在现实场景中的生成和应用进程。

Liu等人(2019a)提出结合生成对抗网络(generative adversarial network，GAN)来生成具有较高视觉和谐度的对抗补丁，利用PS-GAN(perceptual-sensitive GAN)网络作为对抗补丁生成器，可实现将输入的初始补丁图像转化为具备对抗补丁性能的补丁图像，在对抗补丁生成器的训练过程中，利用GradCAM(gradient-weighted class activation mapping)(Selvaraju等，2017)提取注意力特征图，以此寻找可产生最佳对抗攻击效果的补丁位置，并将生成的对抗补丁布置在图像的相应位置。当对抗补丁生成器经过充分训练后，攻击者可端到端地生成特定样式的对抗补丁图像且无需访问目标模型，采用的优化目标函数为

$$\min_{G}\max_{D}\mathbb{E}_{x}\left[\log D(\boldsymbol{\delta},\boldsymbol{x})\right]+\mathbb{E}_{x,z}\left[\log\left(1-D\left(\boldsymbol{\delta},\boldsymbol{x}+{}_{M(x)}G(z,\boldsymbol{\delta})\right)\right)\right]+\lambda\mathbb{E}_{\delta}\left\|G(\boldsymbol{\delta})-\boldsymbol{\delta}\right\|_{2}+\gamma\mathbb{E}_{x,\delta}[\log Pr(\boldsymbol{x}^{*})] \tag{5}$$

式中，$\boldsymbol{x}$为原始图像样本，$\boldsymbol{x}^{*}$为布置有对抗补丁$\boldsymbol{\delta}$的图像对抗样本，z为输入至对抗样本生成器模型的随机向量，λ、γ均为超参数，用于协调对抗补丁的视觉和谐度和对抗攻击性，G，D分别是生成器模型和判别器模型。

Doan等人(2022)针对物理对抗补丁视觉效果较差等问题，提出TnTAttack的对抗攻击方法，与直接使用GAN端到端制作的对抗样本不同，作者结合迭代优化方法和模型训练方法的设计思想，提出两阶段优化方法。第1阶段引入并充分训练一个生成器模型将隐变量映射成为具有常见物品形态的视觉和谐的对抗补丁，第2阶段冻结生成器参数，通过对隐变量的迭代优化使得对抗补丁具备良好的视觉效果。

为了对现有对抗补丁的性能做出相对公正的评估，Hingun等人(2023)提出第一个针对对抗补丁攻击的大规模标准化安全基准REAP(realistic adversarial patch benchmark)，该基准包含从Mapillary Vistas(Neuhold等，2017)数据集中提取的14 651幅路标图像，并且对数据集中的每个道路标志，提出基准变换框架，可以通过匹配当前图像中放置补丁的位置、相机角度和照明条件等因素，将任何数字设计的对抗补丁可微且真实渲染到目标道路标志上。经过作者评估，发现现有工作在数字域设计验证中缺乏与对于现实场景条件的考虑，因而提出的基于补丁的物理攻击在该基准验证下的性能不佳。

二维物理对抗样本的对抗攻击效果依赖于拍摄视角以及方式，所以鲁棒性和应用范围有限。因此近年来部分学者开始研究针对多视角输入和点云等三维数据设计三维物理对抗攻击的方法。Huang等人(2024)面向自然图像物体识别的3D物理对抗，提出一个的新框架TT3D(towards transferable targeted 3D adversarial attack)生成可迁移的3D对抗样本，利用NeRF(neural radiance fields)从多视角输入重建3D网格，减轻了3D对抗对预先定义的3D网格的依赖，扩大了3D攻击的适用范围，如图11所示。

图11　面向自然物体的3D对抗攻击(Huang等，2024)

Fig. 11　3D adversarial attacks on natural objects (Huang et al.，2024)

Liu等人(2019b)首先将基于梯度的对抗性攻击FGSM(Goodfellow等，2015)扩展到点云领域。Tang等人(2023)提出对自编码器中的隐变量进行对抗性拉伸，可以将其解码为平滑的对抗性点云，进而产生对抗样本。Lou等人(2024)提出基于形状的对抗攻击方法HiT-ADV(hide in thicket adversarial attack)，根据点云显著性图和不可感知性分数搜索扰动范围，然后使用高斯核函数在每个攻击区域中添加变形扰动，如图12所示。

不同于需要直接修改或接触目标实体的对抗攻击方法，基于光影投射的对抗攻击方法更加隐蔽，利用自然光线或投影仪即可实施对抗攻击。

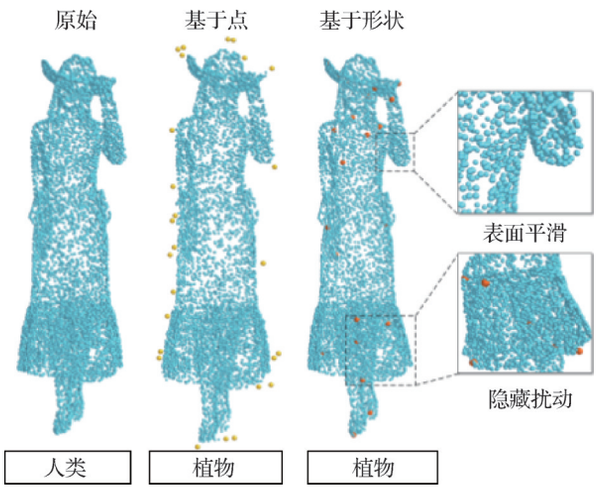


图 12　面向点云分类任务的 3D 对抗攻击（Lou 等，2024）（方框中为被攻击的模型的预测结果）

Fig. 12　3D adversarial attacks in point clouds classification (Lou et al.，2024)(The predicted results of the attacked model are shown in the box)

Gnanasambandam 等人（2021）分析了基于投影仪—相机的对抗光影投射过程，该方法对数字对抗纹理至投影仪过程的颜色转变，以及投影仪实际投射图案至被相机接收成像这两个过程显示建模出来，从而有效分析了所提出对抗攻击方法在应用于自然物体上时的优势以及劣势。此外，通过显示建模中间转换过程，该方法探索了投影对抗攻击可实现的最小扰动以及可攻击目标的受限类别，如图 13 所示。

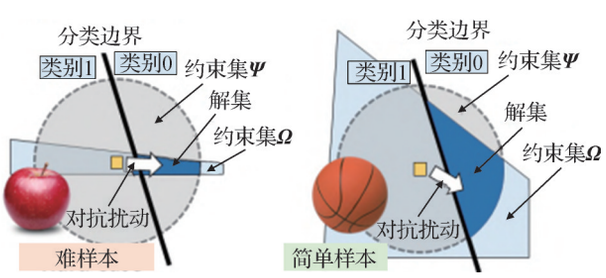


图 13　投影对抗攻击方法的有效性分析（Gnanasambandam 等，2021）

Fig. 13　Analysis of the effectiveness of projection-based adversarial attack methods (Gnanasambandam et al.，2021)

Wang 等人（2023a）简化了投影仪—照相机的实验配置，仅使用一面镜子、带有颜色的透明塑料滤纸以及裁剪掉特定区域的纸来实现投影对抗攻击。具体而言，该方法利用带颜色的滤纸能实现输出不同颜色的光线，通过裁剪过后的纸，能够控制最后投影的形状，从而能够便捷地实现基于光影的攻击，如图 14 所示。

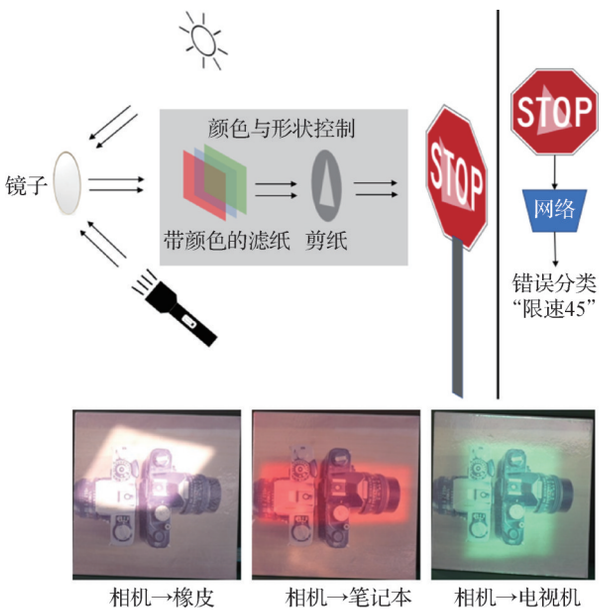


图 14　基于剪纸与带颜色滤纸的投影无接触攻击以及结果（Wang 等，2023a）

Fig. 14　Contactless attacks based on decoupage and colored filters and their results (Wang et al.，2023a)

Sayles 等人（2021）首次利用当前商用相机经常采用的卷帘快门技术，实现人眼不可察觉的无接触对抗攻击。具体而言，相比于全局快门技术，卷帘快门技术采用图像逐行曝光策略以减少成本开支，该技术对于低速运动物体往往能取得较好成像效果，但对于高速运动物体则会产生果冻效应，即物体速度大于扫描速度。该方法基于卷帘快门技术该方面的短板，通过利用可编程的 LED（light emitting diode）灯进行高频信号切换，从而使得成像结果出现带状条纹，进而实现对目标物体的无接触对抗攻击。该方法考虑了现实场景所含的多方面干扰因素，如相机曝光、光线颜色以及相机与光源的同步等，一定程度提升了所设计物理对抗攻击的鲁棒性，如图 15 所示。

Duan 等人（2021）提出基于激光束的无接触物理对抗攻击方法，通过利用现有的激光器，可以从较远处对目标物体进行照射，并能够灵活选择照射时机和时长，从而干扰相机成像结果。该方法从激光束波长、激光束相较于目标物体的入射角度、激光束粗细程度以及激光束强度 4 个方面来探究了该无接触攻击方法的稳定性，如图 16 所示。

(a) 卷帘快门对高速运动物体成像时产生的果冻效应

(b) 无对抗扰动时成像结果(左为人眼视角图像, 右为相机拍摄图像)

(c) 含对抗扰动的成像结果

图 15　针对卷帘快门技术设计基于光影投射的对抗攻击结果示意(Sayles 等,2021)

Fig. 15　Results of the adversarial attack based on light projection designed against the rolling shutter technology (Sayles et al., 2021) ((a) the jelly effect produced by the rolling shutter when imaging high-speed moving objects; (b) the imaging results without adversarial disturbance (the left image is the image from the perspective of the human eye, and the right image is the image captured by the camera); (c) imaging results with adversarial perturbations)

针对目标检测进行对抗攻击需要面向这 3 类输出进行相应的设计,这使得它比攻击图像识别更加具有挑战性。其攻击方式主要是修改物体的属性(例如外观)或周围环境来达到目标丢失、目标错误识别或产生虚警目标 3 类决策错误。其应用形态包括对抗补丁和对抗伪装等,如图 17 所示。

Lu 等人(2017a)提出在停车标志和人脸图像上添加扰动误导相应的检测器,首次尝试将对抗样本应用于目标检测领域。如图 17 所示,通过最小化平

图 16　基于激光的无接触对抗攻击实验器件以及实验结果(Duan 等,2021)

Fig. 16　Laser based contactless adversarial attack experimental device and results (Duan et al., 2021)

均置信度,基于 I-FGSM 算法构建对抗样本。但设计的方法缺乏对场景变化等因素的考虑,尚未取得较好的对抗干扰效果。Eykholt 等人(2018b)进一步扩展 RP2 算法,较好地实现了针对目标检测器的攻击。具体来说,提出虚警目标生成和目标丢失的两种对抗干扰策略,针对性地调整损失函数用于指导对抗性能优化。虚警目标生成所用损失函数为

$$J_c(\boldsymbol{x}, y) = I(b, f(\boldsymbol{x})) \times Pr(b, y, f(\boldsymbol{x})) \tag{6}$$

式中,I 为指示函数,若对检测器 f 在场景 $\boldsymbol{x}$ 中预测的某个检测框 b,其中含有目标的置信度超过一定阈值,则指示函数取值为 1,否则为 0。y 为类别标签。通过指定虚假类别标签 y 并最大化虚警损失函数 J_c,可产生虚警目标对抗样本。目标丢失策略所采用的损失函数为

$$J_d(\boldsymbol{x}, y) = \max_{b \in \boldsymbol{B}} Pr(b, y, f(\boldsymbol{x})) \tag{7}$$

式中,J_d 为虚警损失函数。攻击者可通过最小化 J_d,产生使原目标无法正常被检测的对抗样本。此外,考虑到目标对象的位置和大小可能会根据观察者的位置发生较大变化,因此,作者在优化过程中加入对补丁随机放置和旋转以产生更加鲁棒的物理对抗补

图 17　针对自然图像目标检测的二维物理对抗攻击示例（Lu 等，2017a；Chen 等，2019；Eykholt 等，2018b；Zhao 等，2019；Zolfi 等，2021）

Fig. 17　Examples of 2D physical adversarial attacks for natural image object detection (Lu et al., 2017a; Chen et al., 2019; Eykholt et al., 2018b; Zhao et al., 2019; Zolfi et al., 2021)

丁，并在优化过程中使用 TV（total variation）作为平滑优化约束，以避免像素化的噪声出现。

Zolfi 等人（2021）将补丁建模为半透明的薄片，假设攻击者可以接触到成像设备，将其放置在相机镜头上产生视觉干扰区域。该补丁可以确保其他（非目标）类别的检测不受损害而隐藏选定目标类的所有实例。在优化过程中加入 NPS 提升补丁的鲁棒性。半透明薄片建模过程为

$$\boldsymbol{I}^{*} = \boldsymbol{I} \times (1 - \alpha) + \varepsilon \times \alpha \tag{8}$$

式中，$\boldsymbol{I}^{*}$ 是原始图像样本 $\boldsymbol{I}$ 经过处理后得到的对抗样本 RGB 图像，α 为半透明对抗补丁的 Alpha 通道值，ε 是半透明对抗补丁 RGB 三元组。

除了上述传统的目标检测器外，X 光射线探测器也逐渐纳入对抗攻击的范围。例如，车站 X 射线安检机可以有效检测旅客携带违禁物品，避免安全隐患，可以降低安全风险，保障车站秩序和安全。Liu 等人（2023）实现了面向 X 光安检的实体对抗攻击，文章认为，该任务中成功的物理对抗攻击应该规避颜色和复杂纹理的重叠问题带来的挑战，为此提出 X-Adv 方法生成物理可打印的金属对抗实体，当实体被放置于行李中时，能够欺骗 X 射线探测器。作者首先通过指数函数建模 X 光成像过程，通过强化学习方法，同时优化对抗实体的 3D 网格形状以及位置参数，以实现在目标物体不被对抗实体遮挡的情况下，依旧能够成功产生对抗效果，即，使得违禁品躲避正常检测，如图 18 所示。

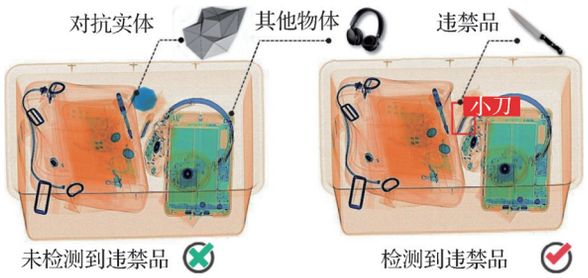


图 18　面向 X 射线安检的 3D 实体对抗攻击（Liu 等，2023）

Fig. 18　3D-object adversarial attack against X-ray security inspection (Liu et al., 2023)

图像识别技术是现有多数图像处理任务的技术基础，而面向图像识别的物理对抗攻击在早期证实了物理对抗攻击在现实场景的存在，尽管一些设计建模方法较为简单，应用形态较为单一，但其工作都为后续的探索提供了技术基础和有价值的参考。

2.2.2　针对人脸图像识别的物理对抗攻击

对抗攻击可以导致人脸图像识别系统出现误识别或错误分类的情况。这种攻击可能引发严重的安全问题，例如身份盗窃、系统欺骗以及财产损失等。因此，研究针对人脸识别系统的对抗攻击至关重要。通过识别和理解这些潜在威胁，可以开发出更为鲁棒的防御机制，确保人脸识别技术在面对恶意攻击时仍然能够有效工作。这种研究有助于提升技术的可靠性和安全性，从而更好地保护个人隐私和公共安全。

由于人脸识别技术对现代社会的重要影响，针对该技术的对抗攻击一直是该领域的研究重点。这些研究涵盖了二维、三维以及基于光影投射的攻击方式，从而尽可能地揭示人脸识别技术的潜在隐患。

人脸识别系统的第 1 步通常是从输入的图像或视频帧中提取人脸区域，并将其输入系统，以预测或匹配特定的人员身份。目前，针对人脸识别的对抗攻击主要分为基于伪装策略和基于逃避策略的攻击。基于伪装策略的对抗攻击是指攻击者试图将扰动后的人脸误导为特定身份，而基于逃避策略的对抗攻击则使被攻击的人脸被识别为原始身份以外的其他身份。与伪装策略相比，逃避策略具有更大的搜索空间。在研究分类上，基于逃避策略的人脸识别攻击属于对抗攻击领域中的非定向攻击（untargeted attack），即不指定对抗样本干扰后目标被误判的类别的攻击；而基于伪装策略的人脸识别攻击则属于定向攻击（targeted attack），即指定对抗样本干扰

后目标被误判为特定类别的攻击(Wang等,2022a)。

近年来涌现了许多针对人脸识别系统的物理对抗攻击方法。这些方法包括通过修改脸部配饰(如眼镜框、帽子、妆容、口罩)来实现对抗攻击,或直接扰动脸部特征(如使用脸部贴纸)导致人脸识别系统失效,甚至通过三维仿制的脸部蒙皮欺骗人脸分类器。此外,还有方法不直接接触人体,而是通过干扰环境光或使用投影设备改变面部光线,以影响人脸识别系统的识别结果。表2展示了针对人脸识别对抗攻击的典型方法,本文将选取其中较为突出的几种方法进行介绍,以帮助读者清晰理解人脸识别系统潜在的安全问题。

表2　针对人脸识别的物理对抗攻击方法描述

Table 2　Description of physical adversarial attack methods against facial recognition

方法	物理对抗攻击形式	主要贡献
Sharif等人(2016)	二维	首个人脸识别对抗攻击研究,在眼镜边框上生成对抗纹理,并提出TV和NPS损失函数。
Sharif等人(2019)	二维	引入生成对抗网络提升对抗性眼镜纹理真实性。
Pautov等人(2019)	二维	将对抗贴纸拓展至额头、鼻子与眼睛。
Komkov和Petiushko(2021)	二维	将对抗贴纸粘贴至帽子上来欺骗人脸识别器。
Xiao等人(2021)	二维	提出对抗贴纸对目标模型的过拟合问题,约束数据分布来缓解该问题。
Yin等人(2021)	二维	基于生成对抗网络,生成欺骗性的脸部妆容。
Singh等人(2022)	二维	改进TV损失函数,组合对抗补丁和扰动。
Wei等人(2023a)	二维	基于遗传算法,寻找对抗贴纸最优粘贴位置。
蔡楚鑫等人(2023)	二维	利用人脸关键点根据脸型来生成对抗眼镜。
崔廷玉等人(2023)	二维	提出一种眼部掩膜的对抗贴片生成方法,实现较好的黑盒攻击效果。
周风帆等人(2023)	二维	基于多模态特征融合网络预测给定环境下物理对抗样本性能。
Zolfi等人(2022)	三维	基于人脸关键点,制作对抗性口罩。
Yang等人(2022)	三维	对齐对抗贴纸与人体脸部的拓扑结构,引入人体重建模型来生成对抗纹理。
Lyu等人(2023)	三维	基于3D感知与UV图,提升对抗性妆容效果。
Yang等人(2023)	三维	利用3DMM确定人体脸部参数,产生对抗性3D网格。
Nguyen等人(2020)	光影投射	将对抗图案经投影设备投射至人脸,并考虑环境、面部等变化来提升攻击鲁棒性。
Li等人(2023b)	光影投射	在人脸表面投射可见光噪声,影响面部点云提取结果,进而影响人脸识别效果。
Fang等人(2024)	光影投射	基于卷帘相机逐行扫描性质,通过编码LED灯频率,扰动相机成像结果。
Zhang等人(2024a)	光影投射	在光线照射角度空间中,搜索具备对抗性的光照角度。

作为首个人脸识别对抗攻击的研究工作,Sharif等人(2016)提出利用扰动后的眼镜框对人脸识别系统进行攻击。该方法首先在数字图像中生成眼镜框,随后将眼镜框通过变换布置在人脸对应位置上,输入至人脸识别模型中进行对抗干扰性能的优化。如图19所示,在优化过程中作者首次提出并加入总变差(TV)正则化和不可打印得分(NPS)作为优化约束,增强攻击的物理可实现性和鲁棒性,具体为

$$\min_{\boldsymbol{r}} \sum_{\boldsymbol{x}\in \boldsymbol{X}} softmax_f\left(\boldsymbol{x}+\boldsymbol{r}, c_t\right) + \lambda \times TV(\boldsymbol{r}) + \gamma \times NPS(\boldsymbol{r}) \tag{9}$$

式中,$softmax_f$为人脸识别模型的识别损失函数,$\boldsymbol{x}$为图像样本数据集$\boldsymbol{X}$中的图像样本,$\boldsymbol{r}$为对抗扰动,c_t为人脸目标对象编号,λ、γ均为超参数。

Zolfi等人(2022)采用不同于眼镜的其他配饰,提出一种通用性的对抗口罩,将对抗性纹理以口罩的形式进行应用。如图20所示,通过利用人脸关键点将对抗性纹理布置到口罩佩戴位置,并使用3D人脸重建技术在人脸图像上布置口罩,在医用口罩上打印对抗图案,从而大幅降低了人脸识别模型的准确性。他们提议使用的对抗纹理的优化目标为

(a) 对抗性眼镜(Sharif等, 2019)　(b) 对抗性妆容(Yin等, 2021)

图 19　基于脸部配饰的人脸识别物理对抗攻击示例

Fig. 19　Examples of physical adversarial attacks on face recognition based on facial accessories ((a) adversarial eyeglasses (Sharif et al., 2019); (b) adversarial make-up (Yin et al., 2021))

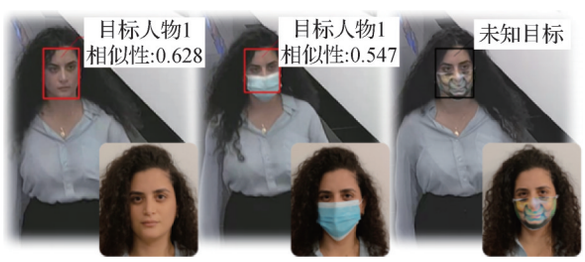


图 20　基于口罩的通用物理对抗攻击示例(Zolfi等,2022)

Fig. 20　Examples of universal physical adversarial attack on a mask (Zolfi et al., 2022)

$$\min_{\boldsymbol{M}_{\text{adv}}}\left[L_{\text{sim}}\left(\boldsymbol{M}_{\text{adv}}\right)+\lambda L_{\text{TV}}\left(\boldsymbol{M}_{\text{adv}}\right)\right] \tag{10}$$

式中,λ是超参数,$\boldsymbol{M}_{\text{adv}}$为对抗性口罩纹理图案,$L_{\text{TV}}$为TV平滑约束函数,$L_{\text{sim}}$为输出嵌入特征和真值对应人脸的嵌入特征之间的余弦相似性,具体为

$$L_{\text{sim}}\left(\boldsymbol{M}_{\text{adv}}\right)=\mathbb{E}_{\theta,\boldsymbol{x}}\frac{1}{J}\sum_{j}\frac{\cos\left(f^{(j)}\left(R_{\theta}\left(\boldsymbol{M}_{\text{adv}},\boldsymbol{x}\right)\right),\boldsymbol{e}_{\text{gt}}^{(j)}\right)+1}{2} \tag{11}$$

式中,R_{θ}为渲染函数,θ为几何变换参数,作者利用人脸关键点(landmark)定位方法将对抗性口罩纹理图案$\boldsymbol{M}_{\text{adv}}$渲染至人脸面部图像$\boldsymbol{x}$对应位置,$f^{(j)}$表示第$j$个模型,$\boldsymbol{e}_{\text{gt}}^{(j)}$表示使用第$j$个模型计算的嵌入特征。

相比于上述研究需要依赖于额外配饰,Yin等人(2021)提出一种基于对抗眼影的攻击方法Adv-Makeup,能取得更强的隐蔽性,利用化妆的方式在眼眶周围添加对抗性眼影,实现对人脸识别模型的干扰,其算法框架由妆容生成、妆容混合和妆容攻击3部分组成。具体而言,该方法首先从现有化妆数据集中剪裁出眼部区域,随后训练一个GAN生成模型作为生成带有对抗攻击性的眼部妆容的对抗样本生成器,将眼部妆容合成到人脸图像上。为验证该对抗性妆容的实用性,在现实场景中绘制了相应的对抗性妆容,验证了该方法的有效性。

不同于对抗性配饰的方法,另一方面的研究者利用二维贴纸来对人脸施加扰动。Komkov和Petiushko(2021)提出AdvHat方法是基于贴纸的人脸识别物理对抗攻击中具有代表性的方法,作者将制作的矩形对抗贴纸粘贴在帽子上对人脸识别模型进行攻击。在设计建模中,结合非平面变换(parabolic transformation)模拟平面弯曲过程和空间变换层(spatial transform layer, STL)算法(Jaderberg等,2015),将贴纸粘贴至人脸图像上进行迭代优化,并在优化过程中采用EOT(expectation over transformation)和TV增强贴纸鲁棒性和平滑性,如图21所示。

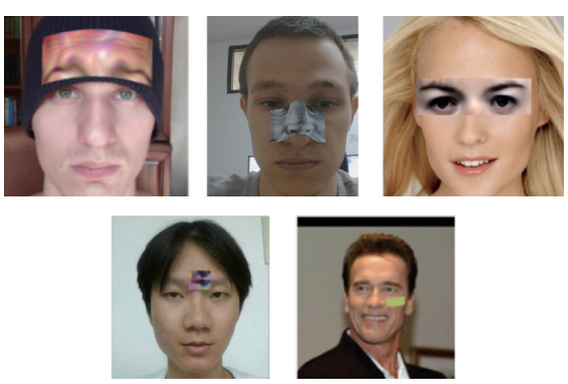


图 21　基于贴纸的人脸识别物理对抗攻击示例(Komkov和Petiushko,2021;Pautov等,2019;Xiao等,2021;Wei等,2023a,b)

Fig. 21　Examples of physical adversarial attacks on face recognition based on stickers (Komkov and Petiushko, 2021; Pautov et al., 2019; Xiao et al., 2021; Wei et al., 2023a,b)

Yang等人(2022)同样采用人脸表面粘贴对抗贴片的方法,但不同于前述基于简单拓扑结构的贴片(如矩形、圆形),该研究通过结合人体面部结构,制成具有复杂拓扑结构的实体蒙皮,从而取得更好的隐蔽攻击效果。具体而言,该方法提出了一种考虑三维人脸变换和真实物理变化的Face3DAdv方法,基于预训练的3D数字人重建模型(Shi等,2021)将2D图像投影至3D空间,随后提出一种基于纹理的对抗攻击范式来生成3D对抗样本,并基于FGSM方法修改3D实体的纹理产生对抗效果。进一步地,Yang等人(2023)提出AT3D(adversarial textured 3D meshes)方法,利用3D重建模型提取输入人脸图像

的面部3D网格和anchor人脸图像的面部纹理以及眼部和鼻子的拓扑结构，将两者结合生成具有复杂拓扑结构的3D对抗性网格，通过修改攻击者的面部结构（眼部、鼻子等部位）将攻击者伪装成另一个目标。攻击者可以通过打印相应的面部结构并粘贴在脸上以逃避人脸识别系统的检测，如图22所示。该方法首先将“攻击者”和“受害者”的人脸图像使用3DMM算法进行重建，用于确定面部的形状、姿态和纹理等的参数。训练时通过调整参数对人脸特定区域（眼鼻部位）设计对抗实体模型，得到具有对抗干扰性的3D网格实体模型。该3D网格实体模型以3D渲染的方式“佩戴”到人脸图像上，产生对抗样本，随后将该对抗样本图像和受害者图像输入人脸识别系统，计算损失函数并反向传播更新3D网格实体模型对应的参数。这种参数化的建模方法充分利用了3DMM（3D morphable face models）模型与真实人脸的近似特性，能够产生与真实人脸面部特征相近的对抗样本，相比于像素级别的优化方法更具可解释性和更好的视觉和谐度。

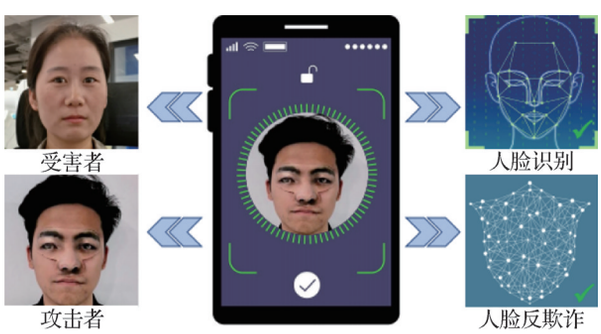


图22　基于3D对抗性网格的物理对抗攻击（Yang等，2023）

Fig. 22　Physical adversarial attack based on 3D adversarial mesh (Yang et al., 2023)

尽管上述人脸对抗攻击方法都取得了令人满意的攻击效果，然而由于这些攻击方法均需对人脸进行直接接触，这样的攻击行为在真实应用场景中极易引起目标对象或旁观者的警觉。为此，Nguyen等人（2020）提出针对人脸识别的基于光影投射的对抗攻击，如图23所示。具体而言，该方法主要通过将生成的对抗图案通过投影设备投射至目标人物面部，进而实现人物身份识别错误、指定人物身份伪装两项攻击。为保障攻击鲁棒性，从3方面优化攻击方法：对环境因素的可抗性、对待攻击目标面部变化的稳定性以及指定目标伪装的稳定性。对于环境因

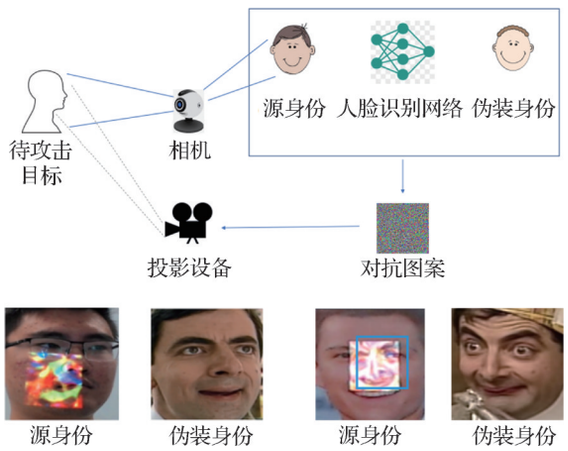


图23　针对人脸识别的基于光影投射的物理对抗攻击流程与结果示意（Nguyen等，2020）

Fig. 23　Illustration of the process and results of physical adversarial attack based on light projection for face recognition (Nguyen et al., 2020)

素，该方法通过人体面部标志点来校正对抗图案投射位置，并通过对齐数字图案与实际投射图案的颜色区间，减少对抗图案的颜色失真；对于待攻击目标面部变化，该方法提出对单一增强图案进行优化，从而减缓常用EOT增强方法带来的计算量过大的问题；对于指定目标伪装，该方法直接采用多幅目标图像来增强稳定性。

相比于Nguyen等人（2020）直接将对抗性图案投影至人脸表面，Fang等人（2024）提出通过编码室内灯光频率来实现更为隐蔽的攻击效果。借鉴Sayles等人（2021）在自然图像分类任务中设计的对抗样本的思路，Fang等人（2024）利用卷帘快门相机逐行扫描的性质，通过编码LED灯高频开关，使得在经过相机成像后的图像表面叠加有黑白条纹，从而使得人脸识别错误，该方法可同时应用于指定身份伪装任务以及身份逃避任务。如图24所示，其中，第1幅图像为正常相机成像结果，其余几幅为调节LED灯开关频次后相机成像结果（Fang等，2024）。

Zhang等人（2024a）从人脸表面打光角度研究针对人脸识别的无接触对抗攻击，基于朗伯模型（Lambertian model）设计了一个对抗样本生成网络，如图25所示。输入人体面部的干净图像，该网络首先预测脸部光照；之后对该光照进行扰动并最终渲染出叠加对抗光照后的目标脸部图像，从而实现对人脸识别系统的欺骗。

此外，Li等人（2023b）研究针对三维人脸识别系

图 24　通过编码 LED 灯开关频次实现人脸成像表面叠加黑白条纹（Fang 等，2024）

Fig. 24　By encoding the switching frequency of the LED light, black and white stripes are superimposed on the surface of the face image (Fang et al., 2024)

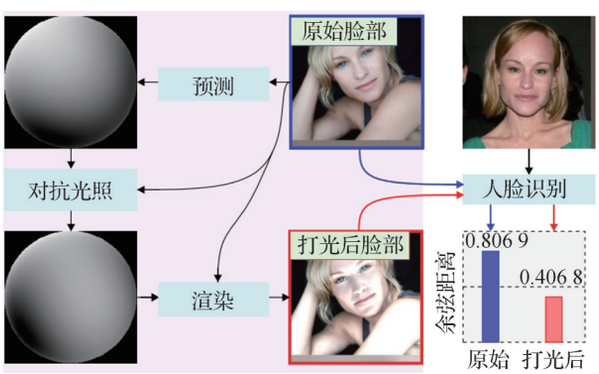


图 25　针对人脸识别的对抗性重光照（Zhang 等，2024a）

Fig. 25　Adversarial relighting for face recognition (Zhang et al., 2024a)

统的对抗攻击，首次实现基于对抗性光照的三维人脸攻击，如图 26 所示。该方法通过将可见光噪声投射在三维人脸表面，从而扰动少量三维点云，实现对于三维人脸识别的攻击效果。在优化过程中，该方法考虑了三维重建过程以及皮肤反射的影响，并通过损失函数设计以及显著图来增强对于随机头部运动的攻击鲁棒性。

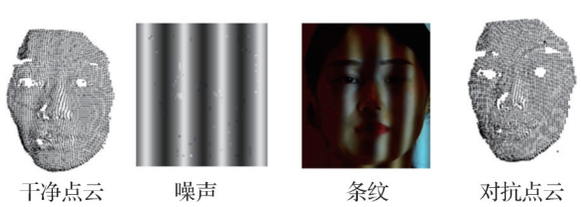


图 26　基于条纹的三维人脸识别对抗攻击（Li 等，2023b）

Fig. 26　Stripe-based adversarial attack on 3D face recognition (Li et al., 2023b)

2.2.3　针对自动驾驶的物理对抗攻击

自动驾驶系统利用人工智能、传感器、摄像头以及雷达等技术，通过实时感知周围环境、定位车辆位置、规划行驶路径并控制车辆运动。

车辆感知系统是现代自动驾驶系统的不可或缺的组成部分，车辆感知通过各种传感器的输入帮助自动驾驶汽车对周边环境进行分析评估，为后续的预测和规划任务提供可靠的观测结果。感知系统的输入包括来自摄像头的图像数据、来自 LiDAR 的点云数据等。为了全面地对驾驶环境进行分析，感知系统涉及到许多视觉任务，例如目标检测和跟踪、深度估计、语义和实例分割等。

自动驾驶任务的目的在于使得车辆能够智能地感知周围环境，并在很少或无需人力的情况下安全行驶。近年来，自动驾驶技术广泛应用在卡车、出租车以及送货机器人无人驾驶等多种任务中，并且能够有效地减少人工操作导致的误判，提高安全性。

然而，自动驾驶系统也面临受到物理对抗样本攻击的风险，为了明确自动驾驶系统所面对的安全风险，本节针对自动驾驶的物理对抗攻击，聚焦于以下 3 个方面：基于补丁等的二维物理对抗攻击、基于 3D 渲染等的三维物理对抗攻击以及基于光影投射的无接触物理对抗攻击，总结当前研究现状。表 3 展示了针对自动驾驶的物理对抗攻击的经典方法，本文对这些方法攻击形式以及主要贡献进行总结，以帮助读者清晰理解自动驾驶系统潜在的安全问题。

在数字域对抗补丁的研究基础上，通过打印对抗补丁和对抗性贴片等二维对抗方法可以将数字域的对抗攻击迁移到物理域上，并对自动驾驶系统产生有效的安全威胁，如图 27 所示。

Kong 等人（2020）面向自动驾驶系统中转角预测设计二维对抗补丁，利用生成对抗网络（GAN）使路牌产生对抗效果，进而使自动驾驶系统无法准确预测转角角度。

Qian 等人（2020）在车牌特定区域添加物理对抗样本，实现了对于车牌识别模型的干扰。具体而言，与前人的像素级优化过程不同，作者将该问题建模为寻找最佳扰动位置的参数优化问题，并使用基于遗传算法的方法来获取最佳扰动位置参数。所生成的对抗补丁以泥点为应用形态，在被人眼忽略的同时可以有效攻击车牌识别系统。

Cheng 等人（2022）面向自动驾驶深度估计提出使用掩膜生成方法，自适应地调整对抗补丁位置与形状提高攻击效果。并且在优化过程中使用 EOT 和风格迁移等技术进一步提升对抗补丁鲁棒性和视觉和谐性。

表3　针对自动驾驶的物理对抗攻击方法描述

Table 3　Description of physical adversarial attack methods against autonomous driving

方法	物理对抗攻击形式	主要贡献
Kong等人(2020)	二维	生成对抗性路牌,使自动驾驶系统无法准确预测转角角度。
Qian等人(2020)	二维	制作对抗性车牌,实现了对于车牌识别模型的干扰。
Brown等人(2018)	二维	将对抗贴纸放在物品周围进行攻击。
陈晋音等人(2020)	二维	针对路牌识别模型,提出基于PSO的黑盒对抗攻击。
钱亚冠等人(2020)	二维	提出一种基于二维码样式对抗样本的物理补丁攻击方法,干扰路牌识别模型。
Cheng等人(2022)	二维	设计自适应位置与形状对抗补丁攻击深度估计模型。
Nesti等人(2022)	二维	使用特定场景广告牌补丁来攻击面向自动驾驶场景的语义分割模型。
陈先意等人(2023)	二维	设计不同形状的对抗补丁、不同颜色的对抗斑点干扰车牌识别系统。
Wang等人(2023b)	二维	首次进行针对自动驾驶的系统级物理对抗攻击研究以及评估。
Cao等人(2019)	三维	基于点云的对抗性实体生成方法,实现在各种场景下规避激光雷达检测系统。
Zhang等人(2019)	三维	面向车辆检测任务设计对抗性涂装。
Wu等人(2020)	三维	使用像素放大和方块堆砌方法产生迷彩涂装。
Wang等人(2021a)	三维	结合可微渲染技术生成视觉自然的物理对抗伪装。
Duan等人(2022)	三维	使用3D渲染实现任意视角对抗,能够产生覆盖整个对象的对抗性伪装。
Wang等人(2022a)	三维	提出基于神经可微分渲染器的全身涂装式对抗纹理迷彩生成框架。
Suryanto等人(2022)	三维	提出可微分变换攻击框架用于在目标对象上生成鲁棒的物理对抗模式,实现任意视角对抗。
Suryanto等人(2023)	三维	利用三平面映射生成鲁棒性和通用性好的对抗纹理。
Li等人(2024)	三维	将NeRF与3D物理对抗相结合生成对抗样本。
段晔鑫等人(2024)	三维	利用三维渲染技术构建布置有对抗性纹理的目标。
Zhou等人(2024)	三维	利用新的神经渲染组件提升不同天气条件下产生的对抗涂装的有效性。
Zhu等人(2024)	三维	建立红外车辆的3D模型,提出基于三维建模的面向红外探测器的物理攻击。
Zheng等人(2024)	三维	面向自动驾驶单目深度估计任务,以重复的方式生成与对象无关的对抗纹理。
Hu等人(2023a)	光影投射	通过在物体表面照射多个激光点实现对抗攻击。
Zhong等人(2022)	光影投射	实现基于阴影实现对路牌的无接触对抗攻击。
吴瀚宇等人(2023)	光影投射	提出一种基于贝叶斯优化的瞬间激光物理对抗攻击方法。
Hsiao等人(2024)	光影投射	实现了基于高光的无接触对抗攻击。
Sato等人(2024)	光影投射	利用红外激光向标志牌投射光斑来干扰自动驾驶系统识别,能够有效打击红外成像设备。
Wen等人(2024)	光影投射	利用无人机搭载投影设备,将生成的对抗图案投影至前方车辆后侧玻璃,用于干扰自动驾驶系统对于前方车辆的检测识别。

Wang等人(2023b)首次进行针对自动驾驶的系统级物理对抗攻击研究和评估。与前人仅在人工智能组件级别(如视觉组件)上评估攻击效果不同,以系统级行为来研究与评估对抗样本的攻击效果。具体而言,以针对停止标志的对抗攻击为例,考察自动驾驶系统能否在标志可见距离内制动并完全停止。在对抗攻击的设计建模中,充分考虑车辆行驶与制动过程中对抗样本的尺寸分布情况和尺寸变化因素,基于前人在组件级别的设计的对抗攻击,提出SySAdv方法,具体而言,在优化过程中采用如下目标函数,具体为

$$\max \mathbb{E}_{s\sim S}[L(M(p_a, O, s, B), \gamma)] \tag{12}$$

图 27　自动驾驶场景中基于平面对象的物理对抗攻击示例（Kong 等，2020；Qian 等，2020；Cheng 等，2022）

Fig. 27　Examples of planar object-based physical adversarial attack against autonomous driving (Kong et al., 2020; Qian et al., 2020; Cheng et al., 2022)

式中，S 是以到标志目标的距离为变量，标志目标 O 像素尺寸 s 的分布，L 是视觉组件模型训练使用的损失函数。函数 M 表示将对抗补丁 p_a 应用于标志目标 O，然后调整标志目标 O 的像素尺寸至 s，并布置到背景 B 中，γ 表示与视觉组件相关的其他输入。该方法从距离角度较为系统地考虑物理对抗攻击在实际使用过程中的系列因素，具备一定参考价值。虽然基于平面对象的物理对抗攻击已经可以对自动驾驶系统的安全构成威胁，例如可以设置对抗性停车标志（Eykholt 等，2018b；Chen 等，2019），使得自动驾驶系统将其误识别为其他标志（如限速标志），但是这种基于平面的对抗实体难以应对自动驾驶系统对任意视角下的对抗的需求。因此面对自动驾驶系统探究基于三维实体的物理对抗攻击以实现在任意视角下的对抗攻击是必要的。

Cao 等人（2019）针对大多数自动驾驶检测系统会配备的雷达测距与避障设备，提出基于点云的对抗性实体生成方法 LiDAR-Adv，通过对构成实体障碍物的点云集合进行优化，产生能够在各种场景下规避激光雷达检测系统的对抗性实体障碍物，揭露了自动驾驶系统的潜在漏洞。如图 28 所示。

Zhang 等人（2019）面向车辆检测任务设计一种特殊的对抗性涂装用于伪装车辆，使其难以被检测

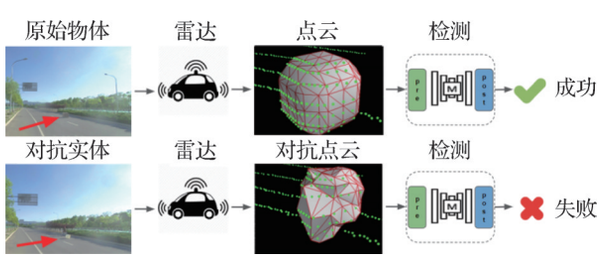


图 28　面向自动驾驶雷达检测系统的物理对抗攻击（Cao 等，2019）

Fig. 28　Physical adversarial attacks for autonomous driving radar detection systems (Cao et al., 2019)

器检测。作者表明最优的对抗性涂装可以通过两步获得，首先利用高精度 3D 车辆模型和仿真环境 Unreal 模拟车辆涂装过程和现实物理环境，通过三维渲染技术模拟传感器成像过程，生成伪装后的车辆图像输入目标检测器中，接着通过最小化检测分数的方法搜索最优涂装。由于三维渲染中图像生成过程不可微，Zhang 等人（2019）训练了一个代理神经网络模型来学习待攻击的目标检测模型检测经过伪装渲染后的车辆目标的输出结果，该代理网络模型以对抗性涂装、3D 车辆模型和环境参数为输入，输出车辆目标检测结果，从而对抗性涂装的优化可以通过代理模型可微地实现。

Wang 等人（2021a）面向车辆目标检测任务结合可微渲染技术，提出双注意抑制（dual attention suppression，DAS）攻击，通过同时抑制模型和人的注意，生成具有强可转移性的视觉自然的物理对抗伪装。在抑制模型注意力方面，通过引入注意力模块获得模型内在注意力模式，并利用损失函数分散不同模型之间共享的注意力模式来进行对抗攻击。该方法生成的对抗涂装捕获了与模型无关的特征，迁移到不同的模型后仍保留一定的对抗性，从而提高了对抗涂装的迁移性。作者初始化对抗涂装为自然图案，采用了逐像素优化方法，最终生成具有对抗性的局部涂装。进一步地，Wang 等人（2024a）在 DAS 的基础上引入了侧向注意力抑制（lateral attention inhibition），提出具备较高通用性能的三重注意抑制（triplet attention suppression，TAS）对抗攻击方法。

Wang 等人（2022a）面向车辆目标检测，为了解决车辆等目标在动态行进中面对的多角度攻击难、实时性攻击差以及物理不易实现等挑战，提出基于神经可微分渲染器的全身涂装式对抗纹理迷彩生成框架 FCA，采用逐像素优化的方式，贴合车辆表面生成多角度、部分遮挡情况下以及物理易实现的对抗伪装迷彩。该方法通过在仿真环境中针对三维物体的整体外表面进行纹理优化，通过多角度捕捉观测物体纹理使生成的对抗性涂装在多角度和部分遮挡情况攻击成功率得到提高，其对抗纹理的优化目标可定义为

$$\boldsymbol{T}_{\text{adv}}^{*}=\underset{\boldsymbol{T}_{\text{adv}}}{\operatorname{argmax}}\, J\left(F\left(\psi\left(R\left(\boldsymbol{M},\boldsymbol{T}_{\text{adv}};\boldsymbol{\theta}_c\right)\right);\boldsymbol{\theta}_f\right),y\right) \tag{13}$$

式中，$\boldsymbol{M}$ 表示 3D 模型的网格（mesh），$\boldsymbol{T}_{\text{adv}}$ 为对抗纹理，$\boldsymbol{\theta}_c$ 表示目标物体和相机的位置信息，R 表示渲染

器，θ_f表示检测器的参数，F表示目标检测器，ψ表示将渲染获得的车辆图像转移到真实背景环境，J表示损失函数，y表示真实标签。通过最大化上述优化目标函数，得到具备较强对抗干扰性能的纹理模式$\boldsymbol{T}_{\text{adv}}^{*}$。

为了获得最优物理对抗涂装，前文提及方法大多利用3D神经渲染器的可微分性。但是已有的神经渲染与真实感渲染（photorealistic renderers）相比，由于缺乏对场景参数的控制，不能完全实现基于真实物理世界对3D场景进行仿真还原。

Suryanto等人（2022）面向车辆检测任务，提出可微分变换攻击（differentiable transformation attack，DTA）框架用于在目标对象上生成鲁棒的物理对抗模式，实现任意视角对抗，设计了可微分变换网络（differentiable transformation network，DTN），该网络在保持目标物体的原始属性的同时，学习了纹理改变时渲染物体的预期变换，以此实现将涂装纹理模拟布置到目标车辆表面，同时不改变其在现实世界中的光照轮廓等属性，更好地模拟了特定场景下应用对抗涂装的目标外观。

由于在渲染过程中难以捕捉环境特征，并且容易忽略不同的天气条件，不同天气情况下生成伪装的有效性降低。为了解决这一问题，Zhou等人（2024）面向车辆目标检测任务提出一种具有鲁棒性和精确性的对抗样本生成方法RAUCA，利用新的神经渲染组件NRP（neural renderer plus），该组件可以较为准确地投影车辆纹理，并渲染具有光照和天气等环境特征的图像，用于捕捉渲染中的环境特征，提升不同天气条件下产生的对抗涂装的有效性，并且提出一个多天气数据集用于伪装生成。

Zhu等人（2024）面向红外车辆检测的3D对抗任务，建立了红外车辆的3D模型，提出一种基于三维建模的面向红外探测器的物理攻击方法，使用3D Mesh向车辆模型表面投射阴影来模拟冷区域的生成过程，提出一种基于三维控制点的网格平滑算法，并使用一组平滑损失函数来增强对抗网格的平滑度，方便贴纸的实现，最终获得一组红外对抗贴纸，使汽车在不同的视角、距离和场景下对红外探测器不可见，如图29所示。

研究表明，二维和三维的对抗攻击已经可以对自动驾驶系统构成威胁，但是，由于前两者均涉及对物理实体的修改，而且自动驾驶系统的目标是实现对车辆的实时、连续控制，研究者引入了无需直接接触物理实体、更加具有瞬时性和实时性的基于光影投射的无接触对抗攻击。

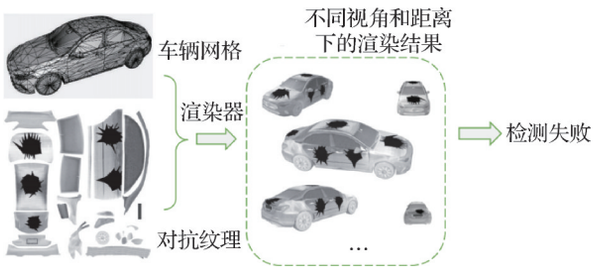


图29　面向红外车辆检测的3D物理对抗示例（Zhu等，2024）

Fig. 29　Example of 3D physical adversarial attack against infrared vehicle detection (Zhu et al., 2024)

Hu等人（2023a）采用了激光束进行无接触对抗攻击，但不同于Duan等人（2021）采用的激光线条的形式，Hu等人（2023a）采用了激光点的形式，通过在物体表面照射多个激光点，利用遗传算法来优化各激光点的位置，从而实现对于目标物体的攻击效果，如图30所示。

图30　不同投影形式的激光束（Hu等，2023a）

Fig. 30　Laser beams with different projection forms (Hu et al., 2023a)

Zhong等人（2022）探索了基于阴影的无接触对抗攻击方法，具体而言，该方法通过对照射光线进行部分人为遮挡，模拟了现实世界中常见的物体表面阴影现象，并基于此，对于阴影投射面积、区域和角度等进行优化，从而实现对路面标志牌识别的攻击，如图31所示。

与Zhong等人（2022）探索阴影攻击相反，Hsiao等人（2024）受自然界物体表面的高光现象启发，探索了基于高光的无接触对抗攻击，该方法首先训练了一个生成器来生成与自然分布相接近的高亮光斑，再基于零阶对抗优化方法ZOO（Chen等，2017）来生成对抗光斑。类似地，Sato等人（2024）也通过

向标志牌投射光斑来影响自动驾驶系统识别，不同于之前利用可见光的对抗攻击方法，该方法利用红外激光实现了人眼不可见的扰动，并能够有效打击红外成像设备。

(a) 现实场景中的阴影现象

(b) 基于阴影的无接触生成的对抗样本图像

图 31　基于阴影的无接触对抗攻击（Zhong 等，2022）

Fig. 31　Shadow based contactless adversarial attack (Zhong et al.，2022) ((a) shadow phenomenon in real scenes; (b) adversarial sample images generated based on shadow casting)

Wen 等人（2024）同样探索了对于自动驾驶系统的对抗攻击，不同于之前方法对标志牌进行扰动，该方法聚焦于对车辆进行扰动，即干扰自动驾驶系统对于前方车辆的检测识别。在实际攻击实施方面，该方法利用无人机搭载投影设备，将生成的对抗图案投影至前方车辆后侧玻璃。该方法考虑了现实实施过程中出现的重影和反射光现象（由玻璃折射/反射导致）、投影过程的几何变形问题等，从而有效保证攻击鲁棒性，如图 32 所示。

2.2.4　针对行人检测的物理对抗攻击

行人检测是计算机视觉领域一个重要任务，旨在识别图像中或视频中行人位置和边界框。行人检测技术在交通管理、智能监控和自动驾驶等任务中起到了重要作用。通过在交叉口、人行横道等处部署检测系统，可以实时监测行人的存在和行为，提高交通管理系统的可靠性；通过将行人检测算法与监控摄像头结合，可以实时监测人群密度、行人轨迹和行人异常行为等，辅助构建智能监控系统；通过行人检测和自动驾驶智能控制系统相结合，可以辅助判断行人的位置、行进方向和行为意图，从而避免潜在的交通事故。

图 32　对抗图案投影至车辆后侧窗口的应用示意图（Wen 等，2024）

Fig. 32　Application example of projecting adversarial patterns onto the rear window of a vehicle (Wen et al.，2024)

为了明确针对行人检测的物理对抗攻击的研究现状，本节聚焦于探究针对行人检测任务的二维、三维物理对抗攻击。

现有针对行人检测的二维物理对抗攻击按照应用形态可分为基于补丁的对抗攻击和基于可穿戴衣物的对抗攻击。表 4 展示了针对行人检测的物理对抗攻击的经典方法，本节对这些方法攻击形式以及主要贡献进行总结，以帮助读者清晰理解行人检测系统潜在的安全问题。

Thys 等人（2019）提出一个针对行人检测器生成对抗补丁的方法，并将生成的补丁打印在硬纸板上。在优化过程中，将对抗补丁放置在检测物体中心，然后采用目标得分作为对抗损失函数，并进一步加入 NPS 和 TV 损失增强对抗补丁鲁棒性。

此外，一些研究者考虑到传感器的多样性，在红外场景中针对行人目标检测任务设计了物理对抗攻击方法。Zhu 等人（2021）首次设计了面向红外行人检测的对抗补丁，利用安装在硬纸板上的小灯泡攻击红外行人检测器，使其无法检测真实世界的行人。为了在物理世界中实现对抗攻击，作者基于特定电子元件（如电阻、小灯泡等）的热特性进行对抗攻击的建模。具体而言，热红外成像主要是利用物体的热辐射，作者首先测量电子元件的热特性与红外摄像机捕捉到的图像模式之间的关系，然后依据此关系在数字域中进行对抗补丁的设计建模和对抗性能优化，最后根据所选材料制作一个物理板将灯泡放置在物理板上相应位置即可实现攻击，如图 33 所示。

Wei 等人（2023d）提出一种形状位置可学习的红外对抗补丁攻击方法，该方法基于梯度来求解攻击效果最强补丁的位置和形状。作者在补丁的优化过程中额外设计了聚合正则化约束函数，用于寻找

表4　针对行人检测的物理对抗攻击方法描述

Table 4　Description of physical adversarial attack methods against pedestrian detection

方法	物理对抗攻击形式	主要贡献
Yang等人(2018)	二维	利用平面EOT变换进行扩展和针孔相机模型,提升对YOLO系列设计的对抗补丁的鲁棒性。
Thys等人(2019)	二维	提出针对行人检测器的对抗补丁生成方法,并将生成的补丁打印在硬纸板上。
Zhu等人(2021)	二维	首次设计了面向红外行人检测的对抗补丁,利用安装在硬纸板上的小灯泡攻击红外行人检测器。
Wei等人(2023d)	二维	提出一种基于梯度求解的形状位置可学习的红外对抗补丁攻击方法。
Wei等人(2024a)	二维	设计了一种通用的物理对抗补丁,可以同时对抗可见光检测器和红外检测器。
Hu等人(2024a)	二维	基于参数化建模,提出一种基于对抗性红外块的黑盒物理对抗攻击方法。
Hu等人(2024b)	二维	提出一种基于曲线实体的红外物理对抗攻击方法。
Xu等人(2020)	二维	提出基于可穿戴衣物的物理对抗攻击方法,是首个针对非刚性物体变形的影响进行建模并用于设计物理对抗样本的工作。
Tan等人(2021)	二维	生成与卡通图像具有视觉相似性的补丁,平衡攻击效果和补丁的视觉和谐性。
Hu等人(2022)	二维	提出对抗纹理的概念,并提出生成可扩展对抗纹理的方法。
Guesmi等人(2024a)	二维	指出利用生成对抗网络生成对抗补丁的潜在空间有限,并提出一种生成动态对抗补丁的方法。
Zhu等人(2022a)	二维	在数字空间中模拟了布料到衣服的过程,设计了一个可以周期性扩展的对抗纹理,并基于此制作了覆盖有气凝胶制成的对抗服装。
Wiyatno和Xu等人(2019)	二维	提出一种物理对抗海报攻击,使跟踪模型跟踪失败。
Ding等人(2021)	二维	提出最大纹理差异损失,干扰跟踪模型对目标的跟踪。
Huang等人(2020)	三维	提出一种通用对抗样本,能够有效干扰检测模型对属于同一类别的所有实例的预测结果。
Hu等人(2023b)	三维	提出采用3D网格对人和衣服进行建模,并通过3D渲染生成一种可布置于服装上的对抗性纹理,实现多视角攻击。

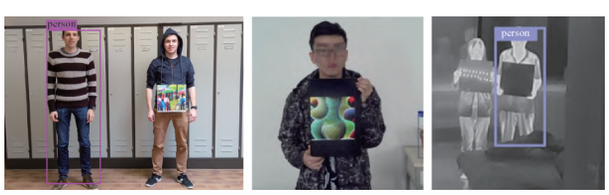


图33　基于补丁的行人检测物理对抗攻击示例(Thys等2019;Wang等,2021b;Zhu等,2021)

Fig. 33　Examples of patch-based physical adversarial attack on pedestrian detection (Thys et al., 2019; Wang et al., 2021b; Zhu et al., 2021)

最具对抗性能的补丁外形,引入掩膜并结合正则化约束函数,选取最具备干扰效果的补丁位置进行布置。在现实场景中,作者利用气凝胶制作了相应的物理对抗补丁并验证了其有效性。

进一步地,Wei等人(2024a)针对可见光和红外检测器设计了一种通用的物理对抗补丁,可以同时对可见光检测器和红外检测器进行干扰。具体而言,不同于先前的像素级别优化方法,作者将对抗补丁建模为由一系列锚点控制的封闭曲线所围成的图像块,封闭曲线首先初始化为圆周,锚点对应地等分布在圆周上。作者通过启发式寻优算法迭代优化锚点的位置参数,以寻找最具备对抗干扰性的补丁形状。Hu等人(2024a)也利用参数化建模的思想,提出一种基于对抗性红外块(adversarial infrared block,AdvIB)的黑盒物理对抗攻击方法。作者将对抗样本建模为参数化的矩形块红外补丁块,通过差分进化的启发式寻优算法优化红外补丁块的角点位置相关参数,产生具备较强干扰性的物理对抗样本。进一步地,Hu等人(2024b)又提出基于曲线实体的红外物理对抗攻击方法(adversarial infrared curve,

AdvIC)。类似地,该方法使用粒子群优化算法优化贝塞尔曲线相关参数,并在物理域中采用冷补丁进行扰动。具体而言,作者将对抗补丁建模为一系列二阶贝塞尔曲线,贝塞尔曲线所在位置将使用0像素进行替换,以模拟冷贴片在红外相机下显示情况。相比于前人进行的像素级别优化方法,Hu等人(2024a,b)和Wei等人(2024a)使用的方法是一种参数化的最优搜索方法,通过参数化地建模对抗样本,避免了像素级别优化过程中一些微弱且难以制作体现的对抗性噪声的影响,使得数字域的设计结果与物理现实世界中的测试结果更为接近。

研究者进一步发现,如行人身上穿戴的衣物的变形等会对基于对抗补丁的方法的对抗效果产生影响,考虑到非刚性物体的形变,Xu等人(2020)提出基于可穿戴衣物的物理对抗攻击方法,采用薄板样条(thin plate spline, TPS)的Transformer模型来模拟非刚性物体的变形,是首个针对非刚性物体(如T恤)变形的影响进行建模并用于设计物理对抗样本的工作。具体来说,作者采用以下3种方法建模真实的物理对抗样本:用TPS近似布料的变形;使用二次多项式回归近似数字空间和物理空间之间的颜色差异;用EOT近似物理变换。同时,为了增强攻击强度,采用集成多个检测器优化对抗补丁。

Tan等人(2021)提出基于LAP(legitimate adversarial patches)的攻击方法,旨在生成与卡通图像具有视觉相似性的补丁,以平衡攻击效果和补丁的视觉和谐性。主要通过将对抗噪声藏在给定视觉和谐的图像(无害图像,或称为具备应用合理性的图像)并分别从颜色特征、边缘特征和纹理特征3个方面对补丁的视觉和谐性进行分析。

Hu等人(2022)提出对抗纹理的概念,该纹理内容结构并不是一个具有完整语义信息的图像,而是可以任意扩展、拼接的,即每块局部纹理均具有对抗效果。他们提出生成对抗纹理的方法TC-EGA(toroidal-cropping-based expandable generative attack),该方法通过可扩展的对抗纹理在现实世界中制作T恤、裙子等,成功攻击行人检测器。

上述研究证明红外探测器的脆弱性,但是所设计的对抗攻击仅在特定角度有效。为了实现多视角攻击,Zhu等人(2022a)在后续工作中构建了对抗衣服,即覆盖有气凝胶制成的衣服。具体来说,作者在数字空间中模拟了布料到衣服的过程,设计了一个可以周期性扩展的基本图像("二维码"图案),图案经过随机剪裁和变形后仍具有对抗效果。为了简化物理实现,作者将"二维码"的生成参数作为优化变量,然后将生成的"二维码"纹理平铺增大,最后从中随机剪裁实际所需的对抗补丁。在优化过程中使用TPS以及仿射变换等提升攻击鲁棒性。最后,作者根据优化好的"二维码"纹理,将黑色块上附着气凝胶制作衣服,最终实现多视角攻击,如图34所示。

此外,一些研究考虑到动态情况下目标移动、动作造成的变形以及运动产生的模糊等因素,面向行人目标跟踪设计了物理对抗攻击,Wiyatno和Xu(2019)提出一种物理对抗海报攻击,使跟踪模型丢失对目标的跟踪能力。

由于行人作为检测目标兼具刚性和柔性物体的特性,且行人的外观易受其穿着、姿态、位置(遮挡)和拍摄视角、尺度等影响,仅仅进行二维对抗样本的研究不足以覆盖当前需求,因此需要进一步探讨三维物理对抗的可行性。

图34　基于可穿戴衣物的行人检测物理对抗攻击
(Hu等,2021;Hu等,2022;Zhu等,2022a)

Fig. 34　Examples of wearable object-based physical adversarial attack on pedestrian detection
(Hu et al., 2021; Hu et al., 2022; Zhu et al., 2022a)

Huang等人(2020)面向行人检测的对抗任务,提出一种生成通用对抗样本的物理伪装攻击方法UPC(universal physical camouflage),基于该对抗样本的攻击能够有效干扰检测模型对属于同一类别的所有实例的预测结果。具体来说,UPC通过欺骗区域提议网络(region proposal network,RPN)的提议结果以及干扰模型输出使其输出错误来进行对抗。为了使所生成的对抗样本在自然场景下对非刚性或非平面对象有效,作者引入了照明、视角、位置和角度模拟外部环境的变换,模拟非平面、非刚体的情况,将扰动伪装成可穿戴服装上的纹理。

基于3D建模制作对抗性实体已应用于制作刚性对抗实体(Liu等,2023),但是在行人检测任务中

人体和衣服都是非刚性的，因此其物理实现具有一定困难。为了制作能在多个视角下躲避行人检测器的外观自然的对抗服装，Hu等人（2023b）提出Adv-CaT，采用3D网格对人和衣服进行建模，并获取服装表面的二维纹理映射和UV坐标，接着基于Voronoi图的划分思想设计对迷彩纹理图案进行参数化的建模。作者利用Gumbel-softmax重参数化技巧实现了迷彩纹理图案参数的可微优化，并通过3D渲染生成一种可布置于服装上的对抗性纹理，实现多视角攻击，并通过拓扑合理投影（TopoProj）和薄板样条（TPS）方法缩小了数字和现实世界对象之间的差距，如图35所示。

图35　针对行人检测的3D渲染物理对抗攻击（Huang等，2020；Hu等，2023b）

Fig. 35　3D rendering-based physical adversarial attack against pedestrian detection (Huang et al., 2020; Hu et al., 2023b)

2.2.5　遥感场景中的物理对抗攻击

在遥感航拍图像中，物理对抗攻击主要通过在地面物体上布置特定图案或结构，以误导遥感目标检测系统的识别和分类。这些对抗样本不仅需要在高分辨率图像中有效，还必须在不同光照条件、天气变化和视角下保持其干扰能力。表5针对遥感场景物理对抗攻击的典型方法进行总结。

Du等人（2022）首次针对航拍图像目标检测器设计物理对抗补丁，根据现实场景，针对车辆目标设计顶部补丁和背景补丁的两种形态的物理对抗补丁。在对抗干扰性优化阶段，采用检测器目标置信度、NPS和TV为优化目标训练对抗补丁，并使用天气模拟技术提升对抗样本在不同环境的鲁棒性。

Lian等人（2022）提出一种新的基于自适应斑块的物理攻击（adaptive patch-based physical attack，APPA）框架，以生成适应物理动力学和变化尺度的对抗性斑块，此外设计了一种新的损失函数，充分利用检测器的检测结果优化补丁。

Lian等人（2023）在APPA基础上提出基于上下文信息的背景攻击框架（contextual background attack，CBA），将设计好的掩膜对前景目标掩盖，对掩膜区域外的像素进行优化，使生成的对抗补丁紧密覆盖在检测目标的关键的上下文背景区域，该框架在不污染前景目标的基础上实现较强的攻击能力和可迁移性，并在20种目标检测器上进行广泛测试，如图36所示。

Chen等人（2024）针对对抗补丁在从数字域到物理域的迁移过程中存在信息丢失，造成对抗补丁在数字空间和物理空间中视觉不一致的问题，提出一种基于协调的对抗补丁生成方法，该方法包含两个主要部分：图像协调网络的自监督训练和对抗补丁的视觉一致性优化。第1步，在图像协调部分，利用现有大规模遥感数据集并收集一组3D Lookup Table（3D LUT）参数。通过随机生成掩膜区域并将

表5　针对遥感场景的物理对抗攻击方法描述

Table 5　Description of physical adversarial attack methods in remote sensing

方法	物理对抗攻击形式	主要贡献
Du等人（2022）	二维	首次针对航拍图像目标检测器设计物理对抗补丁。
Lian等人（2022）	二维	设计了适应物理动力学和变化尺度的对抗性补丁。
Lian等人（2023）	二维	提出基于上下文信息的背景攻击框架，充分利用检测器检测结果优化目标周围补丁，增强补丁对抗攻击性。
Wang等人（2024b）	二维	利用图像掩膜的无遮挡训练策略，将对抗补丁放在目标下方优化，增强攻击效果。
Chen等人（2024）	二维	针对数字域设计结果迁移到物理域时存在差异的问题，提出一种基于图像协调一致的对抗补丁生成方法。
Zhang等人（2024b）	二维	对齐对抗纹理与原始图像特征分布，提升隐蔽性。

(a) 针对车辆目标的对抗补丁(Du等, 2022)

(b) 针对飞机目标的背景对抗补丁(Lian等, 2023)

图 36　遥感场景的基于平面对象的物理对抗攻击示例

Fig. 36　Examples of planar object-based physical adversarial attack in remote sensing ((a) adversarial patches on vehicles (Du et al., 2022); (b) background adversarial patches for airplanes (Lian et al., 2023))

LUT应用到这些区域，创建和谐配对的和谐图像（即原始真实图像）以及合成配对的不和谐图像（即应用LUT之后的生成图像），使用合成图像协调数据集，训练协调网络，预测 3D LUT 的参数。第 2 步，将图像协调集成到对抗补丁的优化过程中，确保数字空间中的优化目标密切反映现实世界的物理场景，从而增强对抗补丁在现实世界中的攻击能力。

Zhang 等人（2024b）针对对抗补丁视觉效果隐蔽性较差以及实用性较差的问题，提出一种风格一致的可扩展的对抗纹理生成方法，在生成对抗问题的过程中将原始图像的特征与对抗纹理的特征进行对齐，并在优化过程中加入风格一致性损失，约束对抗纹理的风格，并且在生成过程中将对抗纹理的攻击效果与纹理形状解耦，确保对抗纹理在实际使用过程中可以适应各种场景而不被纹理形状所约束，增强了对抗纹理的使用灵活性。

但现有遥感场景中的物理对抗攻击大多通过迁移基于已有对抗攻击方法进行设计建模，具备创新性的设计建模方法或将成为该领域的研究热点。

3　核心问题与关键技术

在物理对抗攻击设计与优化过程中，研究者通常面临两大核心问题：

1）数字域中的数学建模结果与物理现实世界的制作应用结果存在现实偏差。受制作设备能力、传感器成像时的精度及误差、成像时环境条件和杂波以及噪声等复杂因素影响，经过数字域设计建模的物理对抗攻击仍然难以如实地反映在现实世界中。

2）物理现实世界中较高的观测自由度严重影响物理对抗攻击的实际效果。物理现实世界较高的观测自由度体现在多变的天气条件、随机的观测系统状态和类型，这些因素的变化可能导致物理对抗攻击的失效（Luo等，2016；Lu等，2017b）。

为了解决以上两大核心问题，研究者提出一系列关键技术，通过设计优化约束函数和增强方法等确保对抗攻击的现实有效性，从而设计出有效的物理对抗攻击方法。本节将面向两大核心问题，对现有工作设计的关键技术进行总结。

3.1　现实偏差问题与相关技术

3.1.1　现实偏差问题

通常，数字域中建立的对抗样本数学模型较为理想，而在实际制作和应用中，数字域的设计结果需要经过由制作设备、应用场景所在环境和待攻击系统采用的观测传感器组成的复杂系统的处理才能输入至待攻击系统中。这其中，一些数字域的理想情况将难以如实地被反映，例如：数字域中对抗样本生成多采用像素级别的优化方法，即，通过相关优化算法或搜索方法，直接调整叠加到原始图像中的扰动像素值。这种方法不但容易产生高频噪声，还容易产生一些肉眼不可察觉的区域性扰动。尽管在数字域优化和验证中，这种微小扰动可以发挥关键性干扰作用。然而，因实际制作设备和成像设备能力有限，微小扰动极易在物理实现的过程中丢失。此外，在数字域设计建模中，还需要充分考虑将要应用的物理场景的环境特征。数字域的模拟应用布置结果应当与现实物理场景中的布置结果尽可能一致。简单的像素替换将会引起外观不一致等视觉差异，如图37所示，这种不一致可能会导致对抗样本的实际应用性能低于预期。因而在设计过程中，物理对抗攻击的可制作性以及数字域设计建模结果与物理场景中

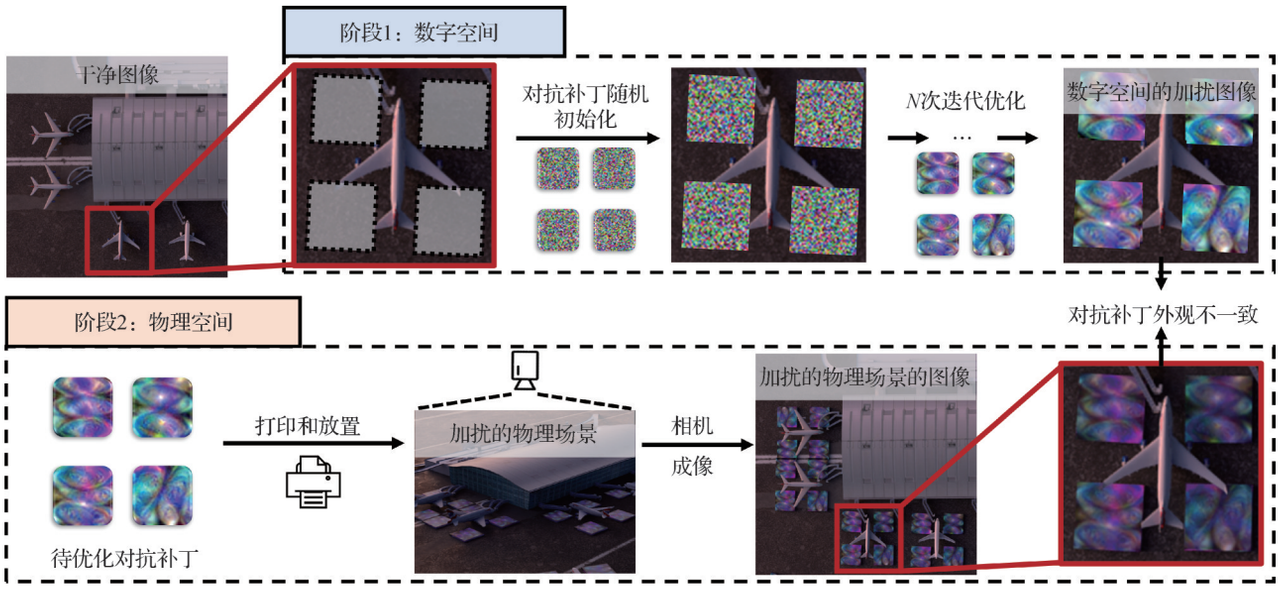


图37　物理对抗攻击的数字域设计结果与物理现实场景的实施结果差异示意图(Chen等,2024)

Fig. 37　The difference between the digital design of physical adversarial attacks and the implementation result in real world (Chen et al., 2024)

实施结果的差异通常是攻击者需要充分考虑的因素。

3.1.2　解决数字域设计的现实偏差的关键技术

为了解决数字域设计的现实偏差问题,研究者提出下列关键技术。

1)物理可实现性优化约束。(1)总变差约束。总变差(TV)最先由Rudin等人(1992)提出,在图像复原领域中,其作为常用的一种正则项约束参与优化,使复原图像具备一定的平滑性。Sharif等人(2016)考虑到自然图像的平滑性和采样噪声的存在,认为微小扰动产生的差异不太可能被相机准确捕获,这导致非光滑的扰动在物理上可能是不可实现的。因此他们首先在对抗样本的设计中引入TV作为平滑性约束提高对抗样本的物理可实现性。TV的计算式具体为

$$TV(\boldsymbol{r})=\sum_{i,j}[(r_{i,j}-r_{i+1,j})^2+(r_{i,j}-r_{i,j+1})^2]^{\frac{1}{2}} \tag{14}$$

式中,$r_{i,j}$为图像$\boldsymbol{r}$在像素坐标(i,j)处的像素值。当图像$\boldsymbol{r}$的相邻像素差异较小时,其对应的TV loss值较低,图像相应地也更为平滑。Singh等人(2022)认为在攻击生成过程的开始便对扰动的平滑性进行总变差约束,这可能限制了训练初始过程中对抗扰动模式的可行解空间。因而他们额外添加了截断机制,使得TV loss仅在扰动的像素大小大于一定阈值后才能对其产生约束作用,即

$$TV^*(\boldsymbol{r})=\sum_{i,j}[(r_{i,j}-r_{i+1,j})^2\times(M_{i,j}\times M_{i+1,j})+(r_{i,j}-r_{i,j+1})^2(M_{i,j}\times M_{i,j+1})]^{\frac{1}{2}} \tag{15}$$

式中,$\boldsymbol{M}$为掩膜,其像素坐标(i,j)处的像素值$M_{i,j}$由下式确定,具体为

$$M_{i,j}=\begin{cases}1 & r_{i,j}>\tau\\ 0 & \text{其他}\end{cases} \tag{16}$$

式中,τ表示人为设定的阈值。

(2)不可打印得分约束。不可打印得分(NPS)首先由Sharif等人(2016)提出。考虑到用于生产制作的设备和屏幕等产生的颜色大多经过量化,在数字域中设计结果将与实际生产设备制作的结果存在一定的颜色偏差,如图38所示。

为了能使数字域中设计的对抗样本可以如实地在现实场景中进行生产制作,攻击者需要对数字域中设计的对抗样本的像素值进行约束。NPS的具体计算式为

$$NPS(p^*)=\prod_{p\in\boldsymbol{P}}|p^*-p| \tag{17}$$

式中,p^*为对抗样本某一位置处的像素值,p为可打印颜色所对应的像素值集合$\boldsymbol{P}$中的元素。此外考虑到实际计算设备可打印的颜色值的数量较多,导致*NPS*的计算成本较高,且损失函数不易优化。为此,Sharif等人(2016)将可打印的颜色集合量化为30个RGB颜色三元组构成的集合,这些三元组与完整集

合的距离方差最小，并且只使用完整集合中存在的颜色。该约束的设计思路和应用功效对后续对抗样本的设计有较大启发（Eykholt 等，2018a，b；Wang 等，2022b；Lovisotto 等，2021）。

(a) 图像在数字显示设备显示结果

(b) 图像经过生成设备打印制作后结果

图 38　数字域设计结果和实际制作结果的颜色偏差示例（Eykholt 等，2018a）

Fig. 38　Color deviation between the design result of the digital domain and the actual production result (Eykholt et al., 2018a) ((a) the digital result; (b) the result from a printer)

2）物理现实映射变换技术。物理现实映射变换技术 D2P（digital-to-physical transformation）是一种通过模拟变换的方式，将数字域的设计结果或模拟布置环境进行变换，使其更加接近于真实物理场景中的制作与应用结果的增强方法。具体而言，研究者通过模拟实际布置过程和环境特征，设计现实映射变换 T 进行处理，具体为

$$\boldsymbol{X}^*, \boldsymbol{\varepsilon}^* = T(\boldsymbol{X}, \boldsymbol{\varepsilon}) \tag{18}$$

式中，$\boldsymbol{X}^*$ 与 $\boldsymbol{\varepsilon}^*$ 为原始图像 $\boldsymbol{X}$ 和对抗扰动 $\boldsymbol{\varepsilon}$ 经过现实映射变换 T 得到的变换结果。

Jan 等人（2019）首次基于现实映射变换思想提出 D2P 增强方法，该方法利用充分训练的变换网络弥补数字域设计和实际应用之间存在的偏差。具体而言，数字域的对抗样本设计建模通常在图像数据或三维仿真平台上进行，其环境特点与现实场景之间存在一定的偏差，两类环境间的转换可视为一种复杂的非线性变换，亮度拉伸、模糊等简单图像处理通常难以逼真地模拟这种非线性变换。而 D2P 方法通过使用 CycleGAN（Zhu 等，2017）等深度学习网络模型学习不同场景特征，从而可模拟数字域场景至物理域实际应用场景的变换。此外，数字域中的对抗样本，如补丁板、贴纸等大多直接通过像素替换的方式布置到待扰动图像的场景中，这与实际环境中布置对抗样本实体的情形存在偏差，如图 37 所示。Suryanto 等人（2022）设计了可微分变换网络（DTN）模型用于实现车辆涂装纹理的渲染，同时避免改变车辆目标在仿真环境中的光照、轮廓等属性，更好地模拟了真实场景下应用对抗涂装的车辆目标外观。Chen 等人（2024）从图像和谐化角度出发，设计视觉一致性的对抗样本生成框架，通过引入可微的图像和谐化模型，能够更好地模拟对抗补丁在给定场景中的实际布置情况，较好地解决了数字域设计建模中通过像素替换的方式布置对抗样本所带来的视觉不一致问题。上述方法都使得对抗攻击可在更加贴合实际应用场景的数字域场景中进行性能提升，从而避免因数字域与物理域之间的较大偏差造成的实际测试的应用性能显著下降的问题。

3.2　实际观测的高自由度问题与相关技术

3.2.1　高自由度观测问题

相比于数字域的理想环境，物理现实世界存在较为复杂多变的环境因素，天气条件、光影变化以及模糊等退化都可能影响对抗样本的实际干扰效果。此外，实际观测中，观测系统的视角、高度均不确定，所搭载的模型种类通常不可知。攻击者需要尽可能确保在上述自由度极大的实际观察中，所设计的对抗攻击仍然能够生效。此外，实际场景中的一些关键系统通常结合深度学习智能识别和人类观察者的目视解译判读进行观测，物理对抗样本的较高的视觉显著性极易引起人类观察者察觉，进而导致干扰失败。因而，物理对抗攻击在现实场景中的鲁棒性、通用性和视觉隐蔽性等也是攻击者需要充分考虑的因素。

3.2.2　应对实际观测自由度较高的相关技术

为了应对实际观测自由度较高的问题，研究者提出下列关键技术。

1）基于增广变换的鲁棒性增强技术。（1）EOT 方法由 Athalye 等人（2018）提出，用于解决数字域的设计结果存在应用失效的问题。前人发现（Luo 等，2016；Lu 等，2017b）现实场景中的某些变换，如视角变换、光照变换等可导致物理对抗攻击的干扰性能显著下降甚至失效，因此有必要在数字域的建模设计中考虑现实场景变换的影响。EOT 方法事先确定一个变换分布，在优化过程中随机采样一组变换对对抗样本进行处理，最后期望生成的物理对抗样本在各种变换下均能产生干扰效果。以图像分类任务为例，通过引入变换，EOT 优化的对抗性能提升损失

函数变为

$$L_f^{\varepsilon} = \mathbb{E}_{t\sim \boldsymbol{T}}[\log P_f(y \mid t(\boldsymbol{X}+\boldsymbol{\varepsilon}))] \tag{19}$$

式中，P_f为模型f将经过变换t变换后的对抗样本$(\boldsymbol{X}+\boldsymbol{\varepsilon})$的类别预测为类别$y$的概率。$t$为变换集合$\boldsymbol{T}$中采样的某一变换。Miao等人(2022)针对EOT方法中难变换下训练不足的问题，设计了MaxOT方法进行改进，具体为

$$L_f^{\varepsilon} = \max_{\boldsymbol{T}^* \subset \boldsymbol{T}} \mathbb{E}_{t\sim \boldsymbol{T}^*}[\log P_f(y \mid t(\boldsymbol{X}+\boldsymbol{\varepsilon}))] \tag{20}$$

式中，变换集合$\boldsymbol{T}^*$为变换集合$\boldsymbol{T}$的子集。在MaxOT的假设下，如果物理对抗样本能够在最具备有害影响的现实场景变换下仍然能够产生对抗效果，那么它们也可以在训练过程中抵御相对较弱的转换所带来的影响。

(2)薄板样本(TPS)是Bookstein(1989)提出用于模拟平滑二维变形的数学工具，Xu等人(2020)首次将其用于模拟褶皱对衣服上的对抗补丁带来的影响。相比于EOT等采用视角、光照等整体性变化来模拟现实场景中的环境因素影响，Xu等人(2020)使用的TPS映射变换则是通过翘曲来模拟局部变换对基于平面对象的物理对抗攻击的影响。其应用方式具体如下：TPS映射首先确定待变换图像上的参考点集合$\{\boldsymbol{p}_i^{(x)}:=(\phi_i^{(x)}, \varphi_i^{(x)})\}_n$和参考图像中对应的参考点集合$\{\boldsymbol{p}_i^{(z)}:=(\phi_i^{(z)}, \varphi_i^{(z)})\}_n$，随后构建TPS映射函数$g$，具体为

$$g(\boldsymbol{p}^{(x)};\boldsymbol{\theta}) = a_0 + a_1\phi^{(x)} + a_2\varphi^{(x)} + \sum_{i=1}^{n} c_i U(\left\| \boldsymbol{p}_i^{(x)} - \boldsymbol{p}^{(x)} \right\|_2) \tag{21}$$

式中，$\boldsymbol{\theta}$为参数向量$\{\boldsymbol{c};\boldsymbol{a}\}$，$\boldsymbol{p}^{(x)}=(\phi^{(x)}, \varphi^{(x)})$为待映射图像上某像素$x$的位置坐标，函数$U(r)=r^2\log r$。最后求解下式得到$\phi$方向参数$\boldsymbol{\theta}_\phi$和$\varphi$方向参数$\boldsymbol{\theta}_\varphi$，具体为

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^{\mathrm{T}} & \boldsymbol{0}_{3\times 3} \end{bmatrix}\boldsymbol{\theta}_\phi = \begin{bmatrix} \boldsymbol{\Delta}_\phi \\ \boldsymbol{0}_{3\times 1} \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^{\mathrm{T}} & \boldsymbol{0}_{3\times 3} \end{bmatrix}\boldsymbol{\theta}_\varphi = \begin{bmatrix} \boldsymbol{\Delta}_\varphi \\ \boldsymbol{0}_{3\times 1} \end{bmatrix} \tag{22}$$

式中，$\boldsymbol{K}_{i,j}=U(\left\| p_i^{(x)} - p_j^{(x)} \right\|_2)$为$\boldsymbol{K}$矩阵的第$i$行第$j$列的元素，$\boldsymbol{P}$矩阵第$i$行为第$i$个参考点的在待变换图像上的广义坐标$(1, \phi_i^{(x)}, \varphi_i^{(x)})$，$\boldsymbol{\Delta}_\phi$和$\boldsymbol{\Delta}_\varphi$分别为参考点在待变换图像与参考图像上$\phi$方向的距离和$\varphi$方向的距离构成的行向量。尽管TPS可较好地模拟平面翘曲，但实际应用中参考点的选取和对应定位是相对复杂且费时的，因此，在假设平面翘曲不严重的情况下，可以对映射函数进行适当修改，通过相对简单的映射变换函数模拟平面翘曲效果，如：使用二次曲线型变换(Komkov和Petiushko，2021；Wei等，2023a；Guesmi等，2024b)。

2)通用性优化约束。在物理对抗攻击的性能优化中，多数方法需要利用待攻击模型的结构、权重等信息来指导对抗样本的相关参数优化，然而在实际任务场景中，智能视觉系统采用的深度学习模型结构、参数等通常对于攻击者来说是不可知的，因此提升物理对抗攻击对未知模型的通用攻击性是极为必要的。研究者利用计算机视觉相关研究，设计了优化约束函数来指导对抗攻击的通用性提升。

模型注意力规避约束：可视化模型注意力技术(Zhou等，2016；Selvaraju等，2017)是用于解释深度学习模型决策行为特点的有力工具。许多研究发现，不同模型在相同场景进行预测时，对相同的物体有相似的注意力模式。基于上述观察，Wang等人(2021a)考虑针对模型的注意力机制设计对抗攻击，通过抑制模型在经过对抗伪装的关键目标上的注意力并将其分散至其他区域，来提高对抗伪装的通用攻击性。其具体计算式为

$$L_{\mathrm{attn}} = \frac{1}{K}\sum_{k=1}^{K}\frac{G_k}{N-N_k} \tag{23}$$

式中，G_k为注意力图$\boldsymbol{S}^y$中的第k个连通区域的像素值和，N和N_k分别为注意力图$\boldsymbol{S}^y$的像素个数和G_k中像素个数。$\boldsymbol{S}^y$的计算式为

$$\boldsymbol{S}^y = relu\left(\sum_k\sum_i\sum_j \alpha_{ij}^{ky} \times relu\left(\frac{\partial p^y}{\partial A_{ij}^k}\right) \times A_{ij}^k\right) \tag{24}$$

式中，p^y是输入图像经过深度学习模型预测后，类别为y的置信度，A_{ij}^k为特征图$\boldsymbol{A}$在第k维通道中位置坐标为(i,j)处的像素值，α_{ij}^{ky}为梯度权重。

此外，Wang等人(2024a)结合侧向注意力抑制机制，设计损失函数提高模型在对抗样本覆盖区域的非真实类别的注意力，从而使得模型无法正确对覆盖有对抗样本的目标正确分类。相应的实验发现，通过对注意力的抑制，可以使得生成的物理对抗攻击通用性得到一定提升。

值得一提的是，构造约束函数仅是通用性提升的一种方法，在实际应用中，攻击者可使用模型集成等机器学习相关方法来提升对抗攻击的通用性，一些数字域的通用性对抗攻击研究的设计思想(Wang和He，2021；Zhu等，2022b)也可借鉴至物理对抗攻击的通用性提升中。

3)视觉隐蔽性优化约束。一些物理对抗样本虽然能够对深度学习模型产生显著的干扰作用,但其较为反常的颜色、纹理和形态等容易引起人类观察者的注意,在反目视解译上效果甚微,从而难以实现伪装效果。因而,为了提高物理对抗攻击的隐蔽性,一些研究者从视觉和谐的角度设计相应的隐蔽性优化约束,用于生成反目视解译和反深度学习侦测兼具的物理对抗攻击。

(1)风格化损失约束。风格化损失(style loss)最先由Gatys等人(2016)在风格迁移任务中提出,基于该损失并结合优化算法可实现以不同的风格样式呈现给定图像的内容。Duan等人(2020)首次利用风格化损失设计物理对抗攻击,以自然中的常见纹理样式作为参考图像,为待检测目标表面生成了自然风格的对抗性涂装。具体而言,给定对抗样本图像$\boldsymbol{X}^*$和参考图像$\boldsymbol{X}^s$,风格化损失计算为

$$L_s = \sum_{l \in \boldsymbol{S}_l} \left\| \boldsymbol{Gram}(F_l(\boldsymbol{X}^s)) - \boldsymbol{Gram}(F_l(\boldsymbol{X}^*)) \right\|_2^2 \quad (25)$$

式中,$\boldsymbol{Gram}$为Gram矩阵,F_l为风格提取模型F的第l层输出,$\boldsymbol{S}_l$为人为设定的需要计算风格化损失的中间层的集合。同时为了保证原始图像内容不被破坏,风格化损失通常与内容一致性损失(content loss)一同参与优化。内容一致性损失具体计算式为

$$L_c = \sum_{l \in \boldsymbol{C}_l} \left\| F_l(\boldsymbol{X}^s) - F_l(\boldsymbol{X}^*) \right\|_2^2 \quad (26)$$

式中,$\boldsymbol{C}_l$为人为设定的需要计算内容一致性损失的中间层的集合。此外,Liu等人(2020)基于风格化损失设计了通用的物理对抗补丁,他们使用数据集中的多组困难样本作为参考图像,避免了对抗补丁过拟合与单幅参考图像风格的同时,使得对抗补丁产生分类器难以正确识别的干扰性纹理。风格化损失本质上是通过约束对抗样本和参考图像的特征相关性来实现对于对抗样本的风格化调整,进而提升对抗攻击的视觉隐蔽性,是一种有参考的约束函数,在实际应用中,参考图像的选取和数量将会对结果产生一定的影响。

(2)人类注意力规避约束。当前研究认为,人类视觉注意力机制是一种自上而下的注意力机制(Connor等,2004),即颜色、边缘和纹理等底层特征相比于反映语义信息的高层特征更容易引起人类观察者的注意。因此,利用人类注意力机制的特点设计物理对抗样本成为提升物理对抗攻击的反目视解译能力的一种有潜力的解决方案。Wang等人(2021a)提出人类注意力规避(human attention evasion)损失,通过约束对抗样本的边缘特征来保证一定的视觉和谐度。具体计算式为

$$L_e = \left\| (\beta \times \boldsymbol{E} + \boldsymbol{1}) \odot (\boldsymbol{T}_{\text{adv}} - \boldsymbol{T}_0) \right\|_2^2 \quad (27)$$

式中,$\boldsymbol{T}_0$为具备一定视觉和谐度的参考图像,$\boldsymbol{T}_{\text{adv}}$为基于参考图像生成的对抗样本,$\boldsymbol{E}$为参考图像的边缘掩膜,参考图像边缘处对应的$\boldsymbol{E}$像素值为1,其余处为0,$\beta$为人为设置的权重参数,用于调整该损失函数对于边缘部分的关注程度,$\odot$为哈达玛积。然而,在他们的工作中,尚未深入探究该损失函数对于人类注意力的干扰作用和成效。Li等人(2023a)利用观察者评分和凝视区域分布来评价对抗样本的视觉和谐程度,并构建相应的数据集PAN(physical attack naturalness dataset)。在此基础上他们训练了深度学习模型DPA(dual prior alignment),用于预测给定图像的视觉和谐度,实现了可微的视觉和谐度估计。攻击者通过提升对抗样本经DPA估计得到的视觉和谐度,可一定程度地提升对抗样本的隐蔽性。

上述工作为探索基于人类注意力规避优化约束的视觉隐蔽性对抗样本生成方法提供了可参考的方案,然而,人类视觉注意力与生物学、心理学等学科关系密切,如何结合相关学科理论研究成果指导对抗样本的视觉隐蔽性设计,目前尚无成熟的解决方案。利用相关学科研究理论,设计有效的可量化评价指标或将成为对抗样本的视觉隐蔽性设计中具有较大潜力的研究方向。

4 物理对抗攻击实施与评估

研究者为了验证所设计的物理对抗攻击方法的性能,通常需要在现实应用场景中制作相应的对抗样本,并使用深度学习模型在场景中采样测试,最后计算相应的性能评价指标。然而,在现实应用场景中,部分物理对抗样本,如:车辆涂装等的制作成本较高,各个对抗攻击所生成的样本难以在同一环境下进行性能对比。此外,较大的环境变化,如:天气条件改变等,对于物理对抗样本的影响难以测试。为此,一些研究者利用仿真渲染引擎,通过较为精细的仿真环境搭建,提供了相对便捷、功能丰富的基线

方法仿真验证平台。本节将对现有工作搭建、采用的仿真验证平台和物理对抗攻击性能评价指标进行总结。

4.1 物理对抗攻击仿真验证平台

为了便于在统一的环境中测试各类物理对抗攻击方法的性能，研究者们开发利用了仿真验证平台，通过较为真实的仿真渲染引擎模拟现实场景，为物理对抗攻击性能的公平测试提供可行方案。现有工作采用的仿真验证平台主要包括CARLA、Unity、Gazebo、Blender和Baidu Apollo Autonomous Driving Platform等。

1）CARLA（car learning to act）是一个基于Unreal Engine 4（UE4）开源的自动驾驶模拟平台，专为城市环境的自动驾驶研究设计。其提供了仿真度较高的城市环境，可模拟多种车辆和行人的行为，并支持多种传感器模拟仿真，包括RGB摄像头、深度摄像头和雷达（LiDAR）等。开发上支持Python和C++，方便研究人员和开发者进行实验和数据收集。该平台已经成为许多面向自动驾驶场景的物理对抗攻击研究工作（Zhang等，2019；Wu等，2020；Wang等，2021a，2024a；Wang 2022a；Nesti等，2022；Suryanto等，2022，2023；Zhou等，2024）的开发和仿真验证平台。

2）Unity是一个基于C#的游戏开发平台，也应用于自动驾驶和机器人模拟。其具备较为真实的物理引擎，可用于模拟真实世界的物理效果，适用于自动驾驶的仿真测试。此外Unity Asset Store提供大量的3D模型和资源，可实现方便快速的模拟环境创建。Duan等人（2022）利用该平台研究了面向车辆检测任务的基于干扰性全身涂装的物理对抗攻击，并验证其在各个视角下的有效性。

3）Baidu Apollo Autonomous Driving Platform（百度Apollo）是一个全面的自动驾驶开发平台。除了提供较为真实的模拟测试场景外，还提供开放的自动驾驶技术栈，便于研究者在成熟系统上验证物理对抗样本的实际性能。Cao等人（2019）利用该平台测试了针对基于雷达的自动驾驶感知系统的物理对抗攻击相关性能。

4）Gazebo是一个开源的机器人仿真平台，广泛用于机器人研究和开发，具有较高保真度的物理模拟引擎，包括碰撞检测、传感器仿真等，适用于机器人系统的测试。此外，Gazebo与机器人操作系统（robot operating system，ROS）紧密集成，方便机器人应用的开发和测试。Wiyatno和Xu（2019）利用该平台设计了针对行人轨迹预测的物理对抗攻击并进行了仿真验证。然而，该平台难以模拟较为真实的环境条件，其实际测试能力存在一定的限制。

5）Blender是一款开源的3D建模和动画软件，广泛应用于各种领域，包括动画制作、游戏开发、视觉效果（visual effects，VFX）、建筑可视化和科学模拟仿真等。综合工具集提供从建模、雕刻、动画、渲染到后期制作的一整套工具。用户可以实现高度自定义的3D建模和动画曲线编辑。在渲染方面，Blender的渲染器可支持光线追踪和材质、纹理编辑渲染，并内置由纹理绘制工具，支持直接在模型上绘制纹理。开发上支持与Python相关的API（application programming interface）。Chen等人（2024）利用该平台搭建面向遥感目标检测的仿真测试环境，对一些基于平面补丁的物理对抗攻击基线方法进行了性能测试，为遥感领域的物理对抗攻击的公平测试做出了有意义的探索。然而，Blender尽管具备较高的仿真渲染能力，但其场景构建成本相对较高，已知公开的场景建模数据相对有限。

值得一提的是，一些研究者利用仿真软件构建相应的场景数据集用于对抗攻击的性能评估。Zhou等人（2024）利用CARLA渲染引擎模拟现实场景中的天气条件，构造Multi-weather dataset来充分验证物理对抗攻击的鲁棒性。Chen等人（2024）利用Blender搭建面向遥感场景的3-D Simulated Scenarios场景数据集。Huang等人（2020）搭建面向行人检测的AttackScenes场景数据集。上述工作都为复杂场景中低成本的对抗样本性能测试做出了积极探索。然而，尽管上述仿真验证平台为物理对抗攻击的性能测试提供了较为便捷、公平的验证测试环境，其构建的验证测试环境与真实应用场景之间差异性难以度量和完全消除，存在一定的局限性。因而，基于仿真验证平台的性能测试当前多作为物理对抗攻击性能验证和对比的补充方案，可对现实场景中的难以测试的部分情况进行补充验证。未来，研究者充分挖掘上述平台的仿真渲染潜力，搭建高质量、现实偏差较小的仿真测试环境，为物理对抗攻击性能的充分测试提供强有力平台支撑。

4.2 物理对抗攻击性能评价指标

4.2.1 干扰性能指标

评价对抗攻击的干扰性能的最常用指标为攻击成功率(attack success rate，ASR)，表示模型受对抗攻击干扰后决策正确率的衰减度。通常，攻击者将于特定数据集或特定场景中施加对抗攻击进行扰动，并收集未经过干扰的原始图像数据和受干扰的图像数据交由模型决策，进而估计对抗攻击的攻击成功率。以分类任务为例，某一类别目标上，分类模型f在未经过对抗攻击的原始图像数据上的召回率为R，在经过对抗攻击的干扰图像数据上的召回率为R_{adv}，则称在该类目标上，对抗攻击针对分类模型f的攻击成功率为

$$ASR_f = 1 - \frac{R_{\text{adv}}}{R} \tag{28}$$

此外，也通过一些计算机视觉任务常用指标(如分类正确率、目标检测中的平均精度均值(mean average precision，mAP)、语义分割中的类别平均像素准确率(mean pixel accuracy，MPA)等)的衰减度来反映对抗样本的干扰性能。实际应用中采用的干扰性能评价指标一般需要依据任务类型和需求等确定。

4.2.2 视觉隐蔽性指标

考虑到部分应用场景对物理对抗攻击的隐蔽性有一定的要求，为了评价物理对抗攻击在人类观察者目视解译下的隐蔽性，研究者们通常会在特定场景中实施对抗攻击，并召集人类观察者对对抗攻击的隐蔽性或视觉和谐度进行打分评价(Wang等，2021a；Hu等，2023b)。然而，人类观察者的评价不可避免地带有主观性，当前鲜有工作设计客观性的指标进行对抗样本的隐蔽性评价。后续工作或可结合图像处理领域相关指标，如：颜色差异度量CIEDE2000(Commission Internationale de l′Eclairage colour difference evaluation 2000)(Sharma等，2005)、结构相似性指数(structure similarity index measure，SSIM)(Wang等，2004)、学习感知图像块相似度(learned perceptual image patch similarity，LPIPS)(Zhang等，2018)等设计适用于对抗样本视觉隐蔽性的评价方案。

5 结语与展望

本文系统讨论了计算机视觉任务中的物理对抗攻击，并依据物理对抗攻击设计的一般流程对具有代表性的114篇研究工作进行归纳总结。本文通过对现有方法的设计建模、性能优化、实施与验证评估3个环节进行综述，认为现有工作较为充分地体现了基于深度学习模型的视觉系统在现实环境中的应用仍面临着对抗样本带来的潜在风险这一基本事实。面对物理对抗攻击中存在的现实偏差问题和高自由度观测问题，研究者可从对抗攻击一般设计流程的3个环节入手，提出具备创新性的解决方案。此外，随着深度学习和计算机视觉技术的快速发展，一些重要的研究分支，如大规模模型和具身智能等，已成为研究者们关注的热点，并在各个领域展现出巨大的潜力和吸引力。鉴于此，本文最后分析并指出了物理对抗攻击领域未来最具潜力的几个方向，为未来物理对抗攻击研究和可信赖的深度学习模型研究提供指引。

通过对一般设计流程的分析总结，本文认为当前物理对抗领域的研究方向中，如下方向具备极大研究潜力和开阔前景：

1)针对基于大模型的智能系统的物理对抗攻击研究。大模型是指具有大规模参数和复杂计算结构的机器学习模型。这些模型通常由深度神经网络构建而成，拥有数十亿甚至数千亿个参数。研究者们发现，随着模型容量的飞速提高和结构的复杂化，如：ChatGPT等大模型也更体现出类似人类的归纳和思考能力，开始具备“涌现能力”。这种涌现能力使得它们有更强的表达能力和更高的准确度，更擅长处理复杂任务。而基于大模型的具身智能系统通过在物理世界和数字世界的学习和进化，可达到理解世界、互动交互并完成任务的目标。大模型和具身智能系统研究可认为是深度学习技术推动高度智能化系统实际应用的坚实一步，然而，它们面对可存在于现实场景中的物理对抗攻击有怎样的表现，能否设计物理对抗攻击干扰大模型和具身智能系统在实际应用中的决策，尚无研究探索。因而，针对基于大模型的智能系统的物理对抗攻击研究将成为极具前景的研究方向之一。

2)面向多源多平台的层次化检测系统的物理对抗攻击研究。现有研究已经证明物理对抗样本在光学图像、红外图像的存在性。然而，现实场景中的观测系统通常具备较为丰富的观测手段，如：遥感领域的卫星平台通常可搭载可见光、红外和雷达等多种

类型的传感器进行对地观测，无人机平台可通过集群的方式构成多平台检测系统进行层次丰富的观测。部分研究（Wei等，2024a）初步探究了一些简单场景中的跨模态对抗攻击，然而针对雷达、多光谱等多模态图像的跨模态对抗攻击，其研究与工程化验证目前仍处于探索阶段。随着智能探测技术的进步，未来的多源、多平台构成的层次化智能检测系统将有望替代单一的传统检测平台，广泛用于各类关键领域，针对这类检测系统的物理对抗攻击将成为极具前景的研究方向之一。

3）针对物理对抗攻击的防御技术研究。面对数字域中对抗攻击的干扰，一些研究提出使用对抗性训练（Madry等，2019）、预处理净化（Nie等，2022）等手段进行防御。然而，面对修改幅度较大、更具侵略性的物理对抗攻击，尚无成熟研究进行提出有效的防御手段。翔云等人（2023）依据对抗补丁色彩鲜艳、变化剧烈的特点，通过遮罩方法进行防御。魏忠诚等人（2022）则提出使用注意力机制进行对抗补丁防御。Wei等人（2023c）面向人脸识别任务探究了针对物理对抗补丁的定位方法，然而，其他可能更具干扰性的形式，如全身涂装等，尚无研究探索消除对抗干扰影响的防御措施。为了推动基于深度学习模型的人工智能技术的落地应用，势必要提高相关技术的鲁棒性并设计针对可能的恶意干扰的防御措施。因此，针对物理对抗攻击的防御技术研究将成为极具前景的研究方向之一。

参考文献（References）

Akhtar N, Jalwana M A A K, Bennamoun M and Mian A. 2022. Attack to fool and explain deep networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10): 5980-5995 [DOI: 10.1109/TPAMI.2021.3083769]

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

Bookstein F L. 1989. Principal warps: thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(6): 567-585 [DOI: 10.1109/34.24792]

Brendel W, Rauber J, Kümmerer M, Ustyuzhaninov I and Bethge M. 2019. Accurate, reliable and fast robustness evaluation//Proceedings of the 33rd International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc.: 12861-12871

Brown T B, Mané D, Roy A, Abadi M and Gilmer J. 2018. Adversarial patch [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1712.09665.pdf

Cai C X, Wang Y F, Zhang L P, Zhuo S C, Zhang J M and Hu Y J. 2023. Adversarial attacks on face recognition system in physical domain. Journal of Cyber Security, 8(2): 127-137 (蔡楚鑫, 王宇飞, 章烈剽, 卓思超, 张娟苗, 胡永健. 2023. 物理域中针对人脸识别系统的对抗样本攻击方法. 信息安全学报, 8(2): 127-137) [DOI: 10.19363/J.cnki.cn10-1380/tn.2023.03.10]

Cai W, Di X Y, Jiang X H, Wang X and Gao W J. 2024. Survey of physical adversarial attacks against object detection models. Computer Engineering and Applications, 60(10): 61-75 (蔡伟, 狄星雨, 蒋昕昊, 王鑫, 高蔚洁. 2024. 针对目标检测模型的物理对抗攻击综述. 计算机工程与应用, 60(10): 61-75) [DOI: 10.3778/j.issn.1002-8331.2310-0362]

Cai Z W and Vasconcelos N. 2018. Cascade R-CNN: delving into high quality object detection//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 6154-6162 [DOI: 10.1109/CVPR.2018.00644]

Cao Y L, Xiao C W, Yang D W, Fang J, Yang R G, Liu M Y and Li B. 2019. Adversarial objects against lidar-based autonomous driving systems [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1907.05418.pdf

Carlini N and Wagner D. 2017. Towards evaluating the robustness of neural networks//2017 IEEE Symposium on Security and Privacy. San Jose, USA: IEEE: 39-57 [DOI: 10.1109/SP.2017.49]

Chen J Q, Zhang Y L, Liu C Y, Chen K Y, Zou Z X and Shi Z W. 2024. Digital-to-physical visual consistency optimization for adversarial patch generation in remote sensing scenes. IEEE Transactions on Geoscience and Remote Sensing, 62: #5623017 [DOI: 10.1109/TGRS.2024.3397678]

Chen J Y, Chen Z Q, Zheng H B, Shen S J and Su M M. 2020. Black-box adversarial attack against road sign recognition model via PSO. Journal of Software, 31(9): 2785-2801 (陈晋音, 陈治清, 郑海斌, 沈诗婧, 苏蒙蒙. 2020. 基于PSO的路牌识别模型黑盒对抗攻击方法. 软件学报, 31(9): 2785-2801) [DOI: 10.13328/j.cnki.jos.005945]

Chen J Y, Zhao X M, Zheng H B and Guo H F. 2024. Survey of optical-based physical domain adversarial attacks and defense. Chinese Journal of Network and Information Security, 10(2): 1-21 (陈晋音, 赵晓明, 郑海斌, 郭海锋. 2024. 基于光学的物理域对抗攻防综述. 网络与信息安全学报, 10(2): 1-21) [DOI: 10.11959/j.issn.2096-109x.2024026]

Chen P Y, Zhang H, Sharma Y, Yi J F and Hsieh C J. 2017. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. Dallas, USA: ACM: 15-26 [DOI: 10.1145/3128572.3140448]

Chen S T, Cornelius C, Martin J and Chau D H. 2019. ShapeShifter: robust physical adversarial attack on faster R-CNN object detector// Machine Learning and Knowledge Discovery in Databases: European Conference. Dublin, Ireland: Springer: 52-68 [DOI: 10.1007/978-3-030-10925-7_4]

Chen X Y, Gu J, Yan K, Jiang D, Xu L F and Fu Z J. 2023. Double adversarial attack against license plate recognition system. Chinese Journal of Network and Information Security, 9(3): 16-27 (陈先意, 顾军, 颜凯, 江栋, 许林峰, 付章杰. 2023. 针对车牌识别系统的双重对抗攻击. 网络与信息安全学报, 9(3): 16-27) [DOI: 10.11959/j.issn.2096-109x.2023034]

Cheng Z Y, Liang J, Choi H, Tao G H, Cao Z W, Liu D F and Zhang X Y. 2022. Physical attack on monocular depth estimation with optimal adversarial patches//Proceedings of the 17th European Conference on Computer Vision. Tel Aviv, Israel: Springer: 514-532 [DOI: 10.1007/978-3-031-19839-7_30]

Connor C E, Egeth H E and Yantis S. 2004. Visual attention: bottom-up versus top-down. Current Biology, 14(19): R850-R852 [DOI: 10.1016/j.cub.2004.09.041]

Cui T Y, Zhang W, He Z Y, Zhou X Y, Zhang Y, Hu G Y and Pan Z S. 2023. Research on face recognition adversarial patch for eye mask. Computer Technology and Development, 33(6): 139-146 (崔廷玉, 张武, 贺正芸, 周星宇, 张瑶, 胡谷雨, 潘志松. 2023. 针对眼部掩模的人脸识别对抗贴片研究. 计算机技术与发展, 33(6): 139-146) [DOI: 10.3969/j.issn.1673-629X.2023.06.021]

Ding L, Wang Y W, Yuan K W, Jiang M Y, Wang P, Huang H and Wang Z J. 2021. Towards universal physical attacks on single object tracking//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 1236-1245 [DOI: 10.1609/AAAI.v35i2.16211]

Doan B G, Xue M H, Ma S Q, Abbasnejad E and Ranasinghe D C. 2022. TNT attacks! Universal naturalistic adversarial patches against deep neural network systems. IEEE Transactions on Information Forensics and Security, 17: 3816-3830 [DOI: 10.1109/TIFS.2022.3198857]

Dong Y P, Liao F Z, Pang T Y, Su H, Zhu J, Hu X L and Li J G. 2018. Boosting adversarial attacks with momentum//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 9185-9193 [DOI: 10.1109/CVPR.2018.00957]

Du A, Chen B, Chin T J, Law Y W, Sasdelli M, Rajasegaran R and Campbell D. 2022. Physical adversarial attacks on an aerial imagery object detector//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3798-3808 [DOI: 10.1109/WACV51458.2022.00385]

Duan R J, Ma X J, Wang Y S, Bailey J, Qin A K and Yang Y. 2020. Adversarial camouflage: hiding physical-world attacks with natural styles//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 997-1005 [DOI: 10.1109/CVPR42600.2020.00108]

Duan R J, Mao X F, Qin A K, Chen Y F, Ye S K, He Y and Yang Y. 2021. Adversarial laser beam: effective physical-world attack to DNNs in a blink//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 16057-16066 [DOI: 10.1109/CVPR46437.2021.01580]

Duan Y X, Chen J L, Zhou X Y, Zou J H, He Z Y, Zhang J, Zhang W, Pan Z S. 2022. Learning coated adversarial camouflages for object detectors [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2109.00124.pdf

Duan Y X, He Z Y, Zhang S, Zhan D Z, Wang T F, Lin G Y, Zhang J and Pan Z S. 2024. Robust physical adversarial camouflages for image classifiers. Acta Electronica Sinica, 52(3): 863-871 (段晔鑫, 贺正芸, 张颂, 詹达之, 王田丰, 林庚右, 张锦, 潘志松. 2024. 针对图像分类的鲁棒物理域对抗伪装. 电子学报, 52(3): 863-871) [DOI: 10.12263/DZXB.20221301]

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Tramèr F, Prakash A, Kohno T and Song D. 2018a. Physical adversarial examples for object detectors//Proceedings of the 12th USENIX Conference on Offensive Technologies. Baltimore, USA: USENIX Association

Eykholt K, Evtimov I, Fernandes E, Li B, Rahmati A, Xiao C W, Prakash A, Kohno T and Song D. 2018b. Robust physical-world attacks on deep learning visual classification//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 1625-1634 [DOI: 10.1109/CVPR.2018.00175]

Fang J B, Jiang C J, Jiang Y, Lin P X, Chen Z J, Sun Y J, Yiu S M and Jiang Z L. 2024. Imperceptible physical attack against face recognition systems via led illumination modulation. IEEE Transactions on Big Data: 461-473 [DOI: 10.1109/TBData. 2024.3403377]

Gatys L A, Ecker A S and Bethge M. 2016. Image style transfer using convolutional neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2414-2423 [DOI: 10.1109/CVPR.2016.265]

Gnanasambandam A, Sherman A M and Chan S H. 2021. Optical adversarial attack//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision Workshops. Montreal, Canada: IEEE: 92-101 [DOI: 10.1109/ICCVW54120.2021.00016]

Goodfellow I J, Shlens J and Szegedy C. 2015. Explaining and harnessing adversarial examples [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1412.6572.pdf

Guesmi A, Bilasco I M, Shafique M and Alouani I. 2024a. AdvART: adversarial art for camouflaged object detection attacks [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2303.01734.pdf

Guesmi A, Ding R T, Hanif M A, Alouani I and Shafique M. 2024b. DAP: a dynamic adversarial patch for evading person detectors//

Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24595-24604 [DOI: 10.1109/CVPR52733.2024.02322]

Hingun N, Sitawarin C, Li J and Wagner D. 2023. REAP: a large-scale realistic adversarial patch benchmark//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4617-4628 [DOI: 10.1109/ICCV51070.2023.00428]

Hsiao T F, Huang B L, Ni Z X, Lin Y T, Shuai H H, Li Y H and Cheng W H. 2024. Natural light can also be dangerous: traffic sign misinterpretation under adversarial natural light attacks//Proceedings of 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 3903-3912 [DOI: 10.1109/WACV57701.2024.00387]

Hu C Y, Shi W W, Jiang T S, Yao W, Tian L, Chen X Q, Zhou J Z and Li W. 2024a. Adversarial infrared blocks: a multi-view black-box attack to thermal infrared detectors in physical world. Neural Networks, 175: #106310 [DOI: 10.1016/j.neunet.2024.106310]

Hu C Y, Shi W W, Yao W, Jiang T S, Tian L, Chen X Q and Li W. 2024b. Adversarial infrared curves: an attack on infrared pedestrian detectors in the physical world. Neural Networks, 178: #106459 [DOI: 10.1016/j.neunet.2024.106459]

Hu C Y, Wang Y L, Tiliwalidi K and Li W. 2023a. Adversarial laser spot: robust and covert physical-world attack to DNNs//Proceedings of the 14th Asian Conference on Machine Learning. Hyderabad, India: PMLR: 483-498

Hu Y C T, Chen J C, Kung B H, Hua K L and Tan D S. 2021. Naturalistic physical adversarial patch for object detectors//Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. Montreal, Canada: IEEE: 7848-7857 [DOI: 10.1109/ICCV48922.2021.00775]

Hu Z H, Chu W D, Zhu X P, Zhang H, Zhang B and Hu X L. 2023b. Physically realizable natural-looking clothing textures evade person detectors via 3D modeling//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE: 16975-16984 [DOI: 10.1109/cvpr52729.2023.01628]

Hu Z H, Huang S Y, Zhu X P, Sun F C, Zhang B and Hu X L. 2022. Adversarial texture for fooling person detectors in the physical world//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13297-13306 [DOI: 10.1109/CVPR52688.2022.01295]

Huang L F, Gao C Y, Zhou Y Y, Xie C H, Yuille A L, Zou C Q and Liu N. 2020. Universal physical camouflage attacks on object detectors//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 717-726 [DOI: 10.1109/CVPR42600.2020.00080]

Huang S Y, Ye F, Huang T Q, Li W, Huang L Q and Luo H F. 2023. Survey on adversarial attacks and defense of face forgery and detection. Chinese Journal of Network and Information Security, 9(4): 1-15 (黄诗瑀, 叶锋, 黄添强, 李伟, 黄丽清, 罗海峰. 2023. 人脸伪造与检测中的对抗攻防综述. 网络与信息安全学报, 9(4): 1-15) [DOI: 10.11959/j.issn.2096-109x.2023049]

Huang Y, Dong Y P, Ruan S W, Yang X, Su H and Wei X X. 2024. Towards transferable targeted 3D adversarial attack in the physical world//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24512-24522 [DOI: 10.1109/CVPR52733.2024.02314]

Jaderberg M, Simonyan K, Zisserman A and Kavukcuoglu K. 2015. Spatial transformer networks//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 2017-2025

Jan S T K, Messou J, Lin Y C, Huang J B and Wang G. 2019. Connecting the digital and physical world: improving the robustness of adversarial attacks//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 962-969 [DOI: 10.1609/AAAI.v33i01.3301962]

Komkov S and Petiushko A. 2021. AdvHat: real-world adversarial attack on ArcFace face ID system//Proceedings of the 25th International Conference on Pattern Recognition. Milan, Italy: IEEE: 819-826 [DOI: 10.1109/ICPR48806.2021.9412236]

Kong Z L, Guo J F, Li A and Liu C. 2020. PhysGAN: generating physical-world-resilient adversarial examples for autonomous driving//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 14242-14251 [DOI: 10.1109/CVPR42600.2020.01426]

Kurakin A, Goodfellow I and Bengio S. 2017. Adversarial machine learning at scale. arXiv preprint arXiv: 1611.01236 [DOI: 10.48550/arXiv.1611.01236]

Kurakin A, Goodfellow I J and Bengio S. 2018. Adversarial examples in the physical world//Artificial Intelligence Safety and Security. New York, USA: Chapman and Hall/CRC: 99-112 [DOI: 10.1201/9781351251389-8]

Li L H, Lian Q and Chen Y C. 2024. Adv3D: generating 3D adversarial examples for 3D object detection in driving scenarios with NeRF [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2309.01351.pdf

Li S M, Zhang S N, Chen G J, Wang D, Feng P, Wang J K, Liu A S, Yi X and Liu X L. 2023a. Towards benchmarking and assessing visual naturalness of physical world adversarial attacks//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12324-12333 [DOI: 10.1109/CVPR52729.2023.01186]

Li Y J, Li Y Q, Dai X L, Guo S T and Xiao B. 2023b. Physical-world optical adversarial attacks on 3D face recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 24699-24708 [DOI: 10.1109/cvpr52729.2023.02366]

Lian J W, Mei S H, Zhang S and Ma M Y. 2022. Benchmarking adversarial patch against aerial detection. IEEE Transactions on Geoscience and Remote Sensing, 60: #5634616 [DOI: 10.1109/TGRS.

2022.3225306]

Lian J W, Wang X F, Su Y R, Ma M Y and Mei S H. 2023. CBA: contextual background attack against optical aerial detection in the physical world. IEEE Transactions on Geoscience and Remote Sensing, 61: #5606616. [DOI: 10.1109/TGRS.2023.3264839]

Lin T Y, Goyal P, Girshick R, He K M and Dollar P. 2017. Focal loss for dense object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2999-3007 [DOI: 10.1109/ICCV.2017.324]

Liu A S, Guo J, Wang J K, Liang S Y, Tao R S, Zhou W B, Liu C, Liu X L and Tao D C. 2023. X-Adv: physical adversarial object attacks against X-ray prohibited item detection//The 32nd USENIX Conference on Security Symposium. Anaheim, USA: USENIX Association: 3781-3798

Liu A S, Liu X L, Fan J X, Ma Y Q, Zhang A L, Xie H Y and Tao D C. 2019a. Perceptual-sensitive GAN for generating adversarial patches//Proceedings of the 33rd AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI: 1028-1035 [DOI: 10.1609/AAAI.v33i01.33011028]

Liu A S, Wang J K, Liu X L, Cao B W, Zhang C Z and Yu H. 2020. Bias-based universal adversarial patch attack for automatic check-out//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 395-410 [DOI: 10.1007/978-3-030-58601-0_24]

Liu D, Yu R and Su H. 2019b. Extending adversarial attacks and defenses to deep 3D point cloud classifiers//Proceedings of 2019 IEEE International Conference on Image Processing. Taipei, China: IEEE: 2279-2283 [DOI: 10.1109/ICIP.2019.8803770]

Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C Y and Berg A C. 2016. SSD: single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision. Amsterdam, the Netherlands: Springer: 21-37 [DOI: 10.1007/978-3-319-46448-0_2]

Lou T R, Jia X J, Gu J D, Liu L, Liang S Y, He B Y and Cao X C. 2024. Hide in thicket: generating imperceptible and rational adversarial perturbations on 3D point clouds//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24326-24335 [DOI: 10.1109/CVPR52733.2024.02296]

Lovisotto G, Turner H, Sluganovic I, Strohmeier M and Martinovic I. 2021. SLAP: improving physical adversarial examples with short-lived adversarial perturbations//30th USENIX Security Symposium. Virtual Event: USENIX Association: 1865-1882

Lu J J, Sibai H and Fabry E. 2017a. Adversarial examples that fool detectors [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1712.02494.pdf

Lu J J, Sibai H, Fabry E and Forsyth D. 2017b. No need to worry about adversarial examples in object detection in autonomous vehicles [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1707.03501.pdf

Luo Y, Boix X, Roig G, Poggio T and Zhao Q. 2016. Foveation-based mechanisms alleviate adversarial examples [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1511.06292.pdf

Lyu Y, Jiang Y, He Z W, Peng B, Liu Y F and Dong J. 2023. 3D-aware adversarial makeup generation for facial privacy protection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11): 13438-13453 [DOI: 10.1109/TPAMI.2023.3290175]

Madry A, Makelov A, Schmidt L, Tsipras D and Vladu A. 2019. Towards deep learning models resistant to adversarial attacks [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1706.06083.pdf

Miao Y B, Dong Y P, Zhu J and Gao X S. 2022. Isometric 3D adversarial examples in the physical world//Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc.: 19716-19731

Moosavi-Dezfooli S M, Fawzi A and Frossard P. 2016. DeepFool: a simple and accurate method to fool deep neural networks//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2574-2582 [DOI: 10.1109/CVPR.2016.282]

Nesti F, Rossolini G, Nair S, Biondi A and Buttazzo G. 2022. Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision. Waikoloa, USA: IEEE: 2826-2835 [DOI: 10.1109/WACV51458.2022.00288]

Neuhold G, Ollmann T, Bulò S R and Kontschieder P. 2017. The mapillary vistas dataset for semantic understanding of street scenes//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 5000-5009 [DOI: 10.1109/ICCV.2017.534]

Nguyen D L, Arora S S, Wu Y H and Yang H. 2020. Adversarial light projection attacks on face recognition systems: a feasibility study//Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle, USA: IEEE: 3548-3556 [DOI: 10.1109/CVPRW50498.2020.00415]

Nie W L, Guo B, Huang Y J, Xiao C W, Vahdat A and Anandkumar A. 2022. Diffusion models for adversarial purification//Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR: 16805-16827

Pautov M, Melnikov G, Kaziakhmedov E, Kireev K and Petiushko A. 2019. On adversarial patches: real-world attack on ArcFace-100 face recognition system//Proceedings of 2019 International Multi-Conference on Engineering, Computer and Information Sciences. Novosibirsk, Russia: IEEE: 391-396 [DOI: 10.1109/SIBIRCON48586.2019.8958134]

Qian Y G, Liu X W, Gu Z Q, Wang B, Pan J and Zhang X M. 2020. QR code based patch attacks in physical world. Journal of Cyber Security, 5(6): 75-86 (钱亚冠, 刘新伟, 顾钊铨, 王滨, 潘俊, 张锡敏. 2020. 一种基于二维码对抗样本的物理补丁攻击. 信息

安全学报, 5(6): 75-86)[DOI: 10.19363/J.cnki.cn10-1380/tn.2020.11.07]

Qian Y G, Ma D F, Wang B, Pan J, Wang J M, Gu Z Q, Chen J H, Zhou W J and Lei J S. 2020. Spot evasion attacks: adversarial examples for license plate recognition systems with convolutional neural networks. Computers and Security, 95: #101826[DOI: 10.1016/j.cose.2020.101826]

Qu Z M, Yin Q L, Sheng Z Q, Wu J Y, Zhang B L, Yu S R and Lu W. 2024. Overview of Deepfake proactive defense techniques. Journal of Image and Graphics, 29(2): 318-342(瞿左珉, 殷琪林, 盛紫琦, 吴俊彦, 张博林, 余尚戎, 卢伟. 2024. 人脸深度伪造主动防御技术综述. 中国图象图形学报, 29(2): 318-342)[DOI: 10.11834/jig.230128]

Redmon J, Divvala S, Girshick R and Farhadi A. 2016. You only look once: unified, real-time object detection//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 779-788[DOI: 10.1109/CVPR.2016.91]

Ren S Q, He K M, Girshick R and Sun J. 2015. Faster R-CNN: towards real-time object detection with region proposal networks//Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal, Canada: MIT Press: 91-99

Rudin L I, Osher S and Fatemi E. 1992. Nonlinear total variation based noise removal algorithms. Physica D: Nonlinear Phenomena, 60(1/4): 259-268[DOI: 10.1016/0167-2789(92)90242-f]

Sato T, Bhupathiraju S H V, Clifford M, Sugawara T, Chen Q A and Rampazzi S. 2024. Invisible reflections: leveraging infrared laser reflections to target traffic sign perception[EB/OL]. [2024-07-24]. https://arxiv.org/pdf/2401.03582.pdf

Sayles A, Hooda A, Gupta M, Chatterjee R and Fernandes E. 2021. Invisible perturbations: physical adversarial examples exploiting the rolling shutter effect//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 14661-14670[DOI: 10.1109/CVPR46437.2021.01443]

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 618-626[DOI: 10.1109/ICCV.2017.74]

Sharif M, Bhagavatula S, Bauer L and Reiter M K. 2016. Accessorize to a crime: real and stealthy attacks on state-of-the-art face recognition//Proceedings of 2016 ACM SIGSAC Conference on Computer and Communications Security. Vienna, Austria: ACM: 1528-1540[DOI: 10.1145/2976749.2978392]

Sharif M, Bhagavatula S, Bauer L and Reiter M K. 2019. A general framework for adversarial examples with objectives. ACM Transactions on Privacy and Security, 22(3): #16[DOI: 10.1145/3317611]

Sharma G, Wu W C and Dalal E N. 2005. The CIEDE2000 color-difference formula: implementation notes, supplementary test data, and mathematical observations. Color Research and Application, 30(1): 21-30[DOI: 10.1002/col.20070]

Shi Y C, Aggarwal D and Jain A K. 2021. Lifting 2D StyleGAN for 3D-aware face generation//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 6254-6262[DOI: 10.1109/CVPR46437.2021.00619]

Singh I, Araki T and Kakizaki K. 2022. Powerful physical adversarial examples against practical face recognition systems//Proceedings of 2022 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. Waikoloa, USA: IEEE: 301-310[DOI: 10.1109/WACVW54805.2022.00036]

Suryanto N, Kim Y, Kang H, Larasati H T, Yun Y, Le T T H, Yang H M, Oh S Y and Kim H. 2022. DTA: physical camouflage attacks using differentiable transformation network//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15284-15293[DOI: 10.1109/CVPR52688.2022.01487]

Suryanto N, Kim Y, Larasati H T, Kang H, Le T T H, Hong Y, Yang H M, Oh S Y and Kim H. 2023. Active: towards highly transferable 3D physical camouflage for universal and robust vehicle evasion//Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4282-4291[DOI: 10.1109/ICCV51070.2023.00397]

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R. 2014. Intriguing properties of neural networks[EB/OL]. [2024-07-21]. https://arxiv.org/pdf/1312.6199.pdf

Tan J, Ji N, Xie H D and Xiang X S. 2021. Legitimate adversarial patches: evading human eyes and detection models in the physical world//Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: Association for Computing Machinery: 5307-5315[DOI: 10.1145/3474085.3475653]

Tang K K, Wu J P, Peng W L, Shi Y W, Song P, Gu Z Q, Tian Z H and Wang W P. 2023. Deep manifold attack on point clouds via parameter plane stretching//Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI: 2420-2428[DOI: 10.1609/AAAI.v37i2.25338]

Thys S, Van Ranst W and Goedemé T. 2019. Fooling automated surveillance cameras: adversarial patches to attack person detection//Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Long Beach, USA: IEEE: 49-55[DOI: 10.1109/CVPRW.2019.00012]

Wang D H, Jiang T S, Sun J L, Zhou W E, Gong Z Q, Zhang X Y, Yao W and Chen X Q. 2022a. FCA: learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack//Proceedings of the 36th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 2414-2422[DOI: 10.1609/AAAI.v36i2.20141]

Wang D H, Yao W, Jiang T S, Li C and Chen X Q. 2023a. RFLA: a stealthy reflected light adversarial attack in the physical world//Pro-

ceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4432-4442 [DOI: 10.1109/ICCV51070.2023.00411]

Wang D R, Li C R, Wen S, Han Q L, Nepal S, Zhang X Y and Xiang Y. 2022b. Daedalus: breaking nonmaximum suppression in object detection via adversarial examples. IEEE Transactions on Cybernetics, 52(8): 7427-7440 [DOI: 10.1109/TCyb.2020.3041481]

Wang J K, Liu A S, Yin Z X, Liu S C, Tang S Y and Liu X L. 2021a. Dual attention suppression attack: generate adversarial camouflage in physical world//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 8561-8570 [DOI: 10.1109/CVPR46437.2021.00846]

Wang J K, Liu X L, Yin Z X, Wang Y X, Guo J, Qin H T, Wu Q T and Liu A S. 2024a. Generate transferable adversarial physical camouflages via triplet attention suppression. International Journal of Computer Vision, 132(11): 5084-5100 [DOI: 10.1007/s11263-024-02098-4]

Wang N F, Luo Y P, Sato T, Xu K D and Chen Q A. 2023b. Does physical adversarial example really matter to autonomous driving? Towards system-level effect of adversarial object evasion attack// Proceedings of 2023 IEEE/CVF International Conference on Computer Vision. Paris, France: IEEE: 4389-4400 [DOI: 10.1109/ICCV51070.2023.00407]

Wang X F, Mei S H, Lian J W and Lu Y J. 2024b. Fooling aerial detectors by background attack via dual-adversarial-induced error identification. IEEE Transactions on Geoscience and Remote Sensing, 62: #5618416 [DOI: 10.1109/TGRS.2024.3386533]

Wang X S and He K. 2021. Enhancing the transferability of adversarial attacks through variance tuning//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 1924-1933 [DOI: 10.1109/CVPR46437.2021.00196]

Wang Y J, Lv H R, Kuang X H, Zhao G, Tan Y A, Zhang Q X and Hu J J. 2021b. Towards a physical-world adversarial patch for blinding object detection models. Information Sciences, 556: 459-471 [DOI: 10.1016/j.ins.2020.08.087]

Wang Z, Bovik A C, Sheikh H R and Simoncelli E P. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing, 13(4): 600-612 [DOI: 10.1109/TIP.2003.819861]

Wang Z B, Wang X, Ma J J, Qin Z, Ren J and Ren K. 2023. Survey on adversarial example attack for computer vision systems. Chinese Journal of Computers, 46(2): 436-468 (王志波, 王雪, 马菁菁, 秦湛, 任炬, 任奎. 2023. 面向计算机视觉系统的对抗样本攻击综述. 计算机学报, 46(2): 436-468) [DOI: 10.11897/SP.J.1016.2023.00436]

Wei X X, Guo Y and Yu J. 2023a. Adversarial sticker: a stealthy attack method in the physical world. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(3): 2711-2725 [DOI: 10.1109/TPAMI.2022.3176760]

Wei X X, Guo Y, Yu J and Zhang B. 2023b. Simultaneously optimizing perturbations and positions for black-box adversarial patch attacks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(7): 9041-9054 [DOI: 10.1109/TPAMI.2022.3231886]

Wei X X, Huang Y, Sun Y T and Yu J. 2024a. Unified adversarial patch for visible-infrared cross-modal attacks in the physical world. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(4): 2348-2363 [DOI: 10.1109/TPAMI.2023.3330769]

Wei X X, Ruan S W, Dong Y P and Su H. 2023c. Distributional modeling for location-aware adversarial patches [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2306.16131.pdf

Wei X X, Yu J and Huang Y. 2023d. Physically adversarial infrared patches with learnable shapes and locations//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 12334-12342 [DOI: 10.1109/CVPR52729.2023.01187]

Wei X X, Yu J and Huang Y. 2024b. Infrared adversarial patches with learnable shapes and locations in the physical world. International Journal of Computer Vision, 132(6): 1928-1944 [DOI: 10.1007/s11263-023-01963-y]

Wei Z C, Feng H, Zhang X Q and Lian B. 2022. Research on physical adversarial sample detection method based on attention mechanism. Application Research of Computers, 39(1): 254-258 (魏忠诚, 冯浩, 张新秋, 连彬. 2022. 基于注意力机制的物理对抗样本检测方法研究. 计算机应用研究, 39(1): 254-258) [DOI: 10.19734/j.issn.1001-3695.2021.06.0255]

Wen H X, Chang S, Zhou L, Liu W and Zhu H Z. 2024. OptiCloak: blinding vision-based autonomous driving systems through adversarial optical projection. IEEE Internet of Things Journal, 11(17): 28931-28944 [DOI: 10.1109/JIoT.2024.3405006]

Wiyatno R R and Xu A Q. 2019. Physical adversarial textures that fool visual object tracking//Proceedings of 2019 IEEE/CVF International Conference on Computer Vision. Seoul, Korea (South): IEEE: 4821-4830 [DOI: 10.1109/ICCV.2019.00492]

Wu H Y, Yang L Y, Wu H, Xu P and Tian L. 2023. Method and analysis of laser adversarial attack in a blink based on Bayesian optimization. Artificial Intelligence Security, 2(2): 38-47 (吴瀚宇, 杨丽蕴, 吴昊, 徐鹏, 田玲. 2023. 基于贝叶斯优化的瞬间激光对抗攻击方法研究. 智能安全, 2(2): 38-47) [DOI: 10.12407/j.issn.2097-2075.2023.02.038]

Wu T, Ning X F, Li W S, Huang R R, Yang H Z and Wang Y. 2020a. Physical adversarial attack on vehicle detector in the carla simulator [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2007.16118.pdf

Xiang Y, Han R X, Chen Z H, Li X Y and Xu D W. 2023. A general defense method for physical space patch adversarial attacks. Journal of Cyber Security, 8(2): 138-148 (翔云, 韩瑞鑫, 陈作辉, 李香玉, 徐东伟. 2023. 一种通用防御物理空间补丁对抗攻击方法. 信息安全学报, 8(2): 138-148) [DOI: 10.19363/J.cnki.

cn10-1380/tn.2023.03.11]

Xiao Z H, Gao X F, Fu C L, Dong Y P, Gao W, Zhang X L, Zhou J and Zhu J. 2021. Improving transferability of adversarial patches on face recognition with generative models//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 11840-11849 [DOI: 10.1109/CVPR46437.2021.01167]

Xie C H, Wang J Y, Zhang Z S, Zhou Y Y, Xie L X and Yuille A. 2017. Adversarial examples for semantic segmentation and object detection//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 1378-1387 [DOI: 10.1109/ICCV.2017.153]

Xu K D, Zhang G Y, Liu S J, Fan Q F, Sun M S, Chen H G, Chen P Y, Wang Y Z and Lin X. 2020. Adversarial T-shirt! Evading person detectors in a physical world//Proceedings of the 16th European Conference on Computer Vision. Glasgow, UK: Springer: 665-681 [DOI: 10.1007/978-3-030-58558-7_39]

Xu Y D, Wang J, Li Y Z, Wang Y J, Xu Z X and Wang D X. 2022. Universal physical adversarial attack via background image//Proceedings of 2022 International Conference on Applied Cryptography and Network Security. Rome, Italy: Springer: 3-14 [DOI: 10.1007/978-3-031-16815-4_1]

Yang D Y, Xiong J, Li X C, Yan X, Raiti J, Wang Y T, Wu H Q and Zhong Z Y. 2018. Building towards "invisible cloak": robust physical adversarial attack on YOLO object detector//Proceedings of the 9th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference. New York, USA: IEEE: 368-374 [DOI: 10.1109/UEMCON.2018.8796670]

Yang X, Dong Y P, Pang T Y, Xiao Z H, Su H and Zhu J. 2022. Controllable evaluation and generation of physical adversarial patch on face recognition[EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2203.04623.pdf

Yang X, Liu C, Xu L L, Wang Y K, Dong Y P, Chen N, Su H and Zhu J. 2023. Towards effective adversarial textured 3D meshes on physical face recognition//Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver, Canada: IEEE: 4119-4128 [DOI: 10.1109/CVPR52729.2023.00401]

Yin B J, Wang W X, Yao T P, Guo J F, Kong Z L, Ding S H, Li J L and Liu C. 2021. Adv-makeup: a new imperceptible and transferable attack on face recognition//Proceedings of the 30th International Joint Conference on Artificial Intelligence. Montreal, Canada: [s.n.]: 1252-1258 [DOI: 10.24963/IJCAI.2021/173]

Zhang Q, Guo Q, Gao R J, Juefei-Xu F, Yu H K and Feng W. 2024a. Adversarial relighting against face recognition. IEEE Transactions on Information Forensics and Security, 19: 9145-9157 [DOI: 10.1109/TIFS.2024.3380848]

Zhang R, Isola P, Efros A A, Shechtman E and Wang O. 2018. The unreasonable effectiveness of deep features as a perceptual metric//Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE: 586-595 [DOI: 10.1109/CVPR.2018.00068]

Zhang Y, Chen J Q, Peng Z B, Dang Y, Shi Z W and Zou Z X. 2024b. Physical adversarial attacks against aerial object detection with feature-aligned expandable textures. IEEE Transactions on Geoscience and Remote Sensing, 62: #4705915 [DOI: 10.1109/TGRS.2024.3426272]

Zhang Y, Foroosh H, David P and Gong B Q. 2019. CAMOU: learning a vehicle camouflage for physical adversarial attack on object detections in the wild//Proceedings of 2019 International Conference on Learning Representations. [s.l.]: [s.n.]

Zhao Y, Zhu H, Liang R G, Shen Q T, Zhang S Z and Chen K. 2019. Seeing isn't believing: towards more robust adversarial attack against real world object detectors//Proceedings of 2019 ACM SIGSAC Conference on Computer and Communications Security. London, United Kingdom: Association for Computing Machinery: 1989-2004 [DOI: 10.1145/3319535.3354259]

Zheng J H, Lin C H, Sun J H, Zhao Z Y, Li Q and Shen C. 2024. Physical 3D adversarial attacks against monocular depth estimation in autonomous driving//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24452-24461 [DOI: 10.1109/CVPR52733.2024.02308]

Zhong Y Q, Liu X M, Zhai D M, Jiang J J and Ji X Y. 2022. Shadows can be dangerous: stealthy and effective physical-world adversarial attack by natural phenomenon//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 15324-15333 [DOI: 10.1109/CVPR52688.2022.01491]

Zhou B L, Khosla A, Lapedriza A, Oliva A and Torralba A. 2016. Learning deep features for discriminative localization//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas, USA: IEEE: 2921-2929 [DOI: 10.1109/CVPR.2016.319]

Zhou F F, Ling H F, Zhang J Y, Xia Z W, Shi Y X and Li P. 2023. Facial physical adversarial example performance prediction algorithm based on multi-modal feature fusion. Computer Science, 50(8): 280-285 (周风帆, 凌贺飞, 张锦元, 夏紫薇, 史宇轩, 李平. 2023. 基于多模态特征融合的人脸物理对抗样本性能预测算法. 计算机科学, 50(8): 280-285) [DOI: 10.11896/jsjkx.221100124]

Zhou J W, Lyu L Y, He D J and Li Y. 2024. RAUCA: a novel physical adversarial attack on vehicle detectors via robust and accurate camouflage generation [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2402.15853.pdf

Zhu J Y, Park T, Isola P and Efros A A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks//Proceedings of 2017 IEEE International Conference on Computer Vision. Venice, Italy: IEEE: 2242-2251 [DOI: 10.1109/ICCV.2017.244]

Zhu X P, Hu Z H, Huang S Y, Li J M and Hu X L. 2022a. Infrared invisible clothing: hiding from infrared detectors at multiple angles in real world//Proceedings of 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA: IEEE: 13307-13316 [DOI: 10.1109/CVPR52688.2022.01296]

Zhu X P, Li X, Li J M, Wang Z Y and Hu X L. 2021. Fooling thermal infrared pedestrian detectors in real world using small bulbs//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtually: AAAI: 3616-3624 [DOI: 10.1609/AAAI.v35i4.16477]

Zhu X P, Liu Y Q, Hu Z H, Li J M and Hu X L. 2024. Infrared adversarial car stickers//Proceedings of 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle, USA: IEEE: 24284-24293 [DOI: 10.1109/CVPR52733.2024.02292]

Zhu Y, Chen Y F, Li X D, Chen K J, He Y, Tian X, Zheng B L, Chen Y W and Huang Q M. 2022b. Toward understanding and boosting adversarial transferability from a distribution perspective. IEEE Transactions on Image Processing, 31: 6487-6501 [DOI: 10.1109/TIP.2022.3211736]

Zolfi A, Avidan S, Elovici Y and Shabtai A. 2022. Adversarial mask: real-world universal adversarial attack on face recognition models [EB/OL]. [2024-07-21]. https://arxiv.org/pdf/2111.10759.pdf

Zolfi A, Kravchik M, Elovici Y and Shabtai A. 2021. The translucent patch: a physical and universal attack on object detectors//Proceedings of 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Nashville, USA: IEEE: 15227-15236 [DOI: 10.1109/CVPR46437.2021.01498]

作者简介

彭振邦，男，博士研究生，主要研究方向为对抗样本、图像处理和模型鲁棒性。E-mail: 20374343@buaa.edu.cn

邹征夏，通信作者，男，教授，主要研究方向为遥感图像处理与分析、计算机视觉、模式识别和机器学习。E-mail: zhengxiazou@buaa.edu.cn

张瑜，女，硕士研究生，主要研究方向为对抗样本和遥感图像处理。E-mail: zhangyu2000@buaa.edu.cn

党一，女，硕士研究生，主要研究方向为对抗样本和遥感图像处理。E-mail: dangyi@buaa.edu.cn

陈剑奇，男，硕士研究生，主要研究方向为对抗样本、图像合成和遥感图像处理。E-mail: cjqchenjianqi@buaa.edu.cn

史振威，男，教授，主要研究方向为遥感图像处理与分析、计算机视觉、模式识别和机器学习。E-mail: shizhenwei@buaa.edu.cn