针对自动驾驶智能模型的攻击与防御

马 晨[1),2)] 沈 超[1),2)] 蔺琛皓[1),2)] 李 前[1),2)] 王 骞[3)] 李 琦[4)] 管晓宏[1),2)]

[1)](西安交通大学电子与信息学部网络空间安全学院 西安 710049)

[2)](智能网络与网络安全教育部重点实验室(西安交通大学) 西安 710049)

[3)](武汉大学国家网络安全学院 武汉 430072)

[4)](清华大学网络科学与网络空间研究院 北京 100084)

摘 要 近年来,以深度学习算法为代表的人工智能技术为人类生产生活的方方面面带来了巨大的革新,尤其是在自动驾驶领域,部署着自动驾驶系统的智能汽车已经走进人们的生活,成为了重要的生产力工具.然而,自动驾驶系统中的人工智能模型面临着潜在的安全隐患和风险,这给人民群众生命财产安全带来了严重威胁.本文通过回顾自动驾驶智能模型攻击和防御的相关研究工作,揭示自动驾驶系统在物理世界下面临的安全风险并归纳总结了相应的防御对策.具体来说,本文首先介绍了包含攻击面、攻击能力和攻击目标的自动驾驶系统安全风险模型.其次,面向自动驾驶系统的三个关键功能层——传感器层、感知层和决策层,本文依据受攻击的智能模型和攻击手段归纳、分析了对应的攻击方法以及防御对策,并探讨了现有方法的局限性.最后,本文讨论和展望了自动驾驶智能模型攻击与防御技术面临的难题与挑战,并指出了未来潜在的研究方向和发展趋势.

关键词 自动驾驶安全;人工智能安全;信息物理系统安全;物理对抗攻击;防御策略

中图法分类号 TP309 **DOI 号** 10.11897/SP.J.1016.2024.01431

Attacks and Defenses for Autonomous Driving Intelligence Models

MA Chen[1),2)] SHEN Chao[1),2)] LIN Chen-Hao[1),2)] LI Qian[1),2)] WANG Qian[3)]
LI Qi[4)] GUAN Xiao-Hong[1),2)]

[1)](*School of Cyber Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an* 710049)

[2)](*Ministry of Education Key Lab for Intelligent Networks and Network Security (Xi'an Jiaotong University), Xi'an* 710049)

[3)](*School of Cyber Science and Engineering, Wuhan University, Wuhan* 430072)

[4)](*Institute for Network Sciences and Cyberspace, Tsinghua University, Beijing* 100084)

Abstract In recent years, artificial intelligence (AI) technologies, notably deep learning algorithms, have ushered in significant innovations across various facets of human existence. One prominent domain benefiting from these advancements is autonomous driving. Intelligent vehicles equipped with autonomous driving systems have gradually integrated into people's daily lives, emerging as pivotal tools that enhance productivity and redefine transportation paradigms. However, the surge in traffic safety incidents in recent years has served as a stark warning, signaling that artificial

收稿日期:2023-06-17;在线发布日期:2024-01-16.本课题得到科技创新 2030-"新一代人工智能"重大项目(2020AAA0107702)、国家自然科学基金(U21B2018,62161160337,62132011,62376210,62006181,U20B2049,U20A20177,62206217)、陕西省重点研发计划项目(2021ZDLGY01-02,2023-ZDLGY-38)资助.**马 晨**,硕士研究生,主要研究方向为自动驾驶安全.E-mail:ershang@stu.xjtu.edu.cn.**沈 超**(通信作者),博士,教授,博士生导师,长江学者,中国计算机学会(CCF)杰出会员,主要研究领域为人工智能可信与安全、信息物理系统控制与安全、智能软件安全与测试.E-mail:chaoshen@mail.xjtu.edu.cn.**蔺琛皓**,博士,研究员,博士生导师,主要研究领域为人工智能安全、对抗机器学习、智能身份认证.**李 前**,博士,助理教授,主要研究方向为人工智能安全、对抗机器学习.**王 骞**,博士,教授,博士生导师,IEEE Fellow,主要研究领域为人工智能安全、云计算安全与隐私、无线系统安全、应用密码学.**李 琦**,博士,副教授,博士生导师,主要研究领域为互联网和云安全、移动安全、机器学习安全.**管晓宏**,博士,教授,博士生导师,中国科学院院士,主要研究领域为网络信息安全、网络化系统、电力系统优化调度.

intelligence models within autonomous driving systems are susceptible to potential safety hazards and risks. This reality poses a significant threat to the safety of people's lives and properties. This paper reviews previous research works related to intelligent attack and corresponding defense works to reveal the security risks of autonomous driving systems in the physical world, and summarizes the corresponding defense strategies. Specifically, we first introduce in this paper the security risk model for autonomous driving systems that includes attack surfaces, attack capabilities, and attack goals. The main workflow of the autonomous driving system can be grouped into three layers. The autonomous driving system first takes the information about the nearby environment gathered by the sensor layer as input, and then processes the data through the perception layer equipped with intelligent models to extract key information such as obstacles, traffic signs, traffic lights and lane lines. Subsequently, the decision layer predicts the movement trajectories of the surrounding obstacles and plans the travel path of the autonomous vehicle based on the extracted information. In this process, the attacker could use different physical attacks to execute attacks against the intelligent model, thus posing a huge security risk. Building upon the known attack intelligence of the attacker, we categorize attacks into three types: white-box, gray-box, and black-box attacks. Furthermore, considering the diverse methods of interference available to attackers, we classify the attacks into two main categories: physical world attacks and sensor injection attacks. Secondly, for the three key functional layers of the autonomous driving system including sensor layer, perception layer and decision layer, this paper summarizes and analyzes the corresponding attack methods as well as defense countermeasures depending on the victim intelligent models and methods of attack, and discusses the limitations of the existing methods. Finally, this paper discusses and outlooks the difficulties and challenges of attack and defense technologies for autonomous driving intelligent models, and indicates potential future research directions and development trends. We propose that the absence of comprehensive and objective evaluation criteria for physical countermeasure attacks, coupled with the limited feasibility studies on physical attacks and research gaps in system-level attack methodologies, pose challenges and point towards future research directions in the current landscape of intelligent model attacks in autonomous driving. Moreover, the current research on defense countermeasures remains relatively scant, and the development of defense strategies in the physical realm holds great promise as a research avenue for the future. Addressing these gaps in both attack and defense methodologies will contribute substantially to the robustness and security of the intelligent models in autonomous driving.

Keywords autonomous driving security; artificial intelligence security; cyber-physical system security; physical adversarial attack; defense strategy

1 引 言

近年来,随着人工智能和传感器技术的高速发展,自动驾驶汽车已逐渐融入人们的日常生活.各种各样的商业和私人的自动驾驶车辆已经在道路上行驶,如自动驾驶出租车、公共汽车、卡车、快递车等,以及数百万的私人特斯拉汽车也已配备了 Autopilot[1] 自动驾驶系统.为了实现在复杂、动态的驾驶环境下的自动驾驶,自动驾驶系统配备了一系列人工智能模型来处理核心决策过程,如感知、定位、预测和规划,这基本上形成了自动驾驶车辆的"大脑".然而,由于这些人工智能模型的错误可能导致各种交通事故,甚至可能产生灾难性后果,因此确保它们的安全性显得至关重要.例如,2016 年 7 月,特斯拉自动驾驶汽车的感知模型无法区分明亮的天空和一辆白色卡

车，导致了一场严重的交通事故的发生[2]，致使汽车驾驶员身亡.类似地，2018 年 3 月发生在美国亚利桑那州的优步无人车事故中，事发时处于自动驾驶模式的无人车并没有检测到前方行人，驾驶员也未及时进行干预，最终致使行人被撞身亡[3].

最近的研究表明，现有的人工智能模型，尤其是深度学习模型容易受到对抗性攻击，而由于智能模型是自动驾驶系统进行感知和决策的核心组件，智能模型的安全性极大地影响着自动驾驶系统的整体安全性，因而自动驾驶相关的智能模型攻防研究引起了广泛关注.不同于数字域的对抗性攻防方法，自动驾驶相关的智能攻防方法关注于物理世界下的有效性和可行性.并且由于自动驾驶系统往往配备多个传感器且需要综合处理融合这些传感器信息，因此自动驾驶系统所面临的攻击手段更加复杂多样，涵盖的攻击面更加广泛.遗憾的是，目前有关人工智能模型攻防的综述主要关注数字域的工作，鲜有对自动驾驶智能模型攻防领域的全面整理.并且由于不同的攻防方法关注的自动驾驶模块不同，使用的攻击手段不同，因此构建的威胁模型和应用场景复杂多样.基于此，我们对现有的自动驾驶智能模型攻防研究工作进行了科学的分类以及系统的对比、归纳及总结，以便于为后续研究者了解和研究自动驾驶智能模型攻防安全提供指导.

在本文中，我们首先介绍了自动驾驶系统安全风险模型，然后从自动驾驶系统的三个攻击面：传感器层、感知层和决策层对现有的自动驾驶智能模型攻击和防御方法进行了系统的总结和科学的归纳，并讨论了相关方法的局限性.最后，我们讨论了自动驾驶智能模型攻防研究所面临的挑战以及未来可行的研究方向.

2　自动驾驶系统安全风险模型

对配备智能模型的自动驾驶系统进行攻防分析，首先需要介绍对应的安全风险模型.因此，本节首先简要介绍自动驾驶系统存在的攻击面，并从攻击能力和攻击目标对现有的攻击手段进行分类.

2.1　自动驾驶系统攻击面

自动驾驶系统通过人工智能等技术处理传感器信息使得汽车可以在没有人工操作的情况下自动行驶，其核心部分由三个功能层组成，包括传感器层、感知层和决策层.由于自动驾驶系统需要从周围环境中获取信息，传感器层会直接暴露在外部攻击环境中，而对其信息进行提取和处理的感知层和决策层也同样会面临安全威胁.

(1) 传感器层.

在传感器层，自动驾驶系统通过传感器(相机、激光雷达、毫米波雷达、GPS 和 IMU 等)收集周围环境的实时信息，包括周围空间信息和自身位置信息.周围空间信息由相机、激光雷达和毫米波雷达获取，其中相机用以获取周围的视觉信息，激光雷达基于光的反射测量物体与车辆之间的距离来获取周围物体的点云信息，而毫米波雷达基于毫米波来检测周围物体；自身位置信息由 GPS 和 IMU 来获取，GPS 可以通过人造地球卫星获取自身绝对位置数据，而 IMU 则可以测量自身方向、速度以及加速度数据.

(2) 感知层.

在感知层，自动驾驶系统利用传感器层得到的原始数据，通过深度学习智能模型等算法提取传感器信息并进行信息融合，以完成定位、目标检测及跟踪、交通标志识别和车道线检测等任务.

(3) 决策层.

在决策层，自动驾驶智能模型主要完成路径规划、目标轨迹预测和车辆控制三个任务.路径规划任务是确定起点到指定目的地之间的路线，目标轨迹预测任务是自动驾驶系统利用感知层得到的信息预测周围障碍物的轨迹，而车辆控制任务则是自动驾驶系统依据规定路线控制汽车行驶.

图 1 展示了自动驾驶系统框架，物理世界的信息被传感器层获取并传输给感知层，感知层对信息进行融合和处理得到周围环境信息，进而这些信息传输到决策层进行预测和规划，最后根据规划得到的路径和运动，由执行层对汽车进行控制.由于感知层和决策层使用智能模型对传感器层输出的信息进行分析处理.因此，传感器层、感知层和决策层不仅是自动驾驶系统的核心功能层，也是自动驾驶系统所面临的主要攻击面.

2.2　攻击能力

对攻击者的攻击能力进行建模主要考虑两个因素：攻击者已知的情报以及攻击者能够采取的干扰手段.

(1) 根据攻击者已知的信息，攻击可以分为：

① 白盒攻击(white-box attack).攻击者对目标自动驾驶智能模型相关信息有完全的了解，包括目

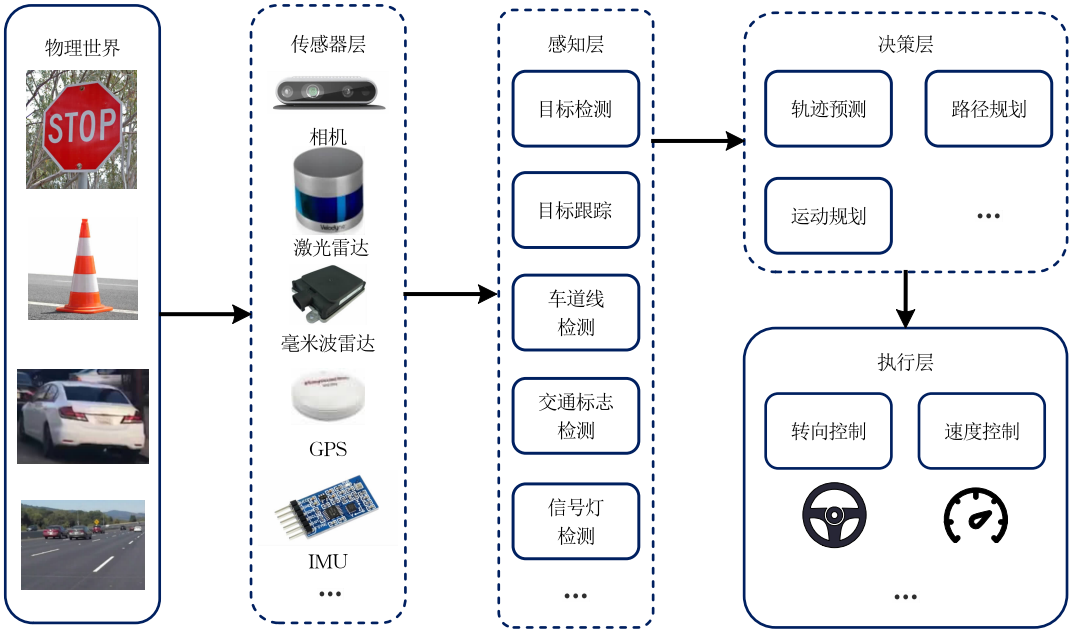


图 1 自动驾驶系统框架

标模型的设计与实现、传感器参数等.

② 灰盒攻击(gray-box attack). 自动驾驶智能模型的关键细节对于攻击者而言是未知的,攻击者仅能够接触输入和输出环节. 在灰盒攻击模型中,攻击者可以通过构造并发送输入样本以获得模型输出的置信度分数等信息,并根据相应的输出信息来对系统的智能模型进行推理,同时,攻击者还可以获得详细的相关传感器参数.

③ 黑盒攻击(black-box attack). 攻击者不能访问目标自动驾驶智能模型的任何内部结构. 黑盒攻击条件下的一种攻击方法是基于可转移性(transferability)的攻击,它基于白盒替代模型(substitute model)来生成攻击输入. 另一种攻击方法是传感器级别的攻击,使用外部干扰以影响传感器数据质量或利用远程攻击设备(投影仪、激光发射器等)直接向传感器投放伪造数据.

(2) 依据攻击者能够采取的干扰手段,攻击被分为:

① 物理世界攻击. 攻击者改变物理世界的驾驶环境来影响自动驾驶智能模型接收到的环境信息. 常见的攻击向量包括改变周围物体(如障碍物、交通标志等)表面纹理、形状以及位置.

② 传感器注入攻击. 攻击者利用传感器获取外部信息的方式来注入错误的观测信息. 例如,向激光雷达发射激光来伪造返回激光从而在其获取的点云中注入欺骗点,向毫米波雷达输入中注入伪造的返回信号,向 GPS 接收器发送虚假的卫星信号来改变定位,使用激光或不可见光照射相机等.

2.3 攻击目标

攻击目标是指攻击者发起攻击时的目标功能层,自动驾驶系统攻击的攻击目标主要可以分为 3 类:

(1) 传感器层攻击.

传感器层攻击将自动驾驶汽车的外部传感器作为攻击目标,主要方式包括通过降低信号质量或者伪造虚假信号来欺骗传感器层.

(2) 感知层攻击.

感知层攻击将感知层的机器学习模型作为攻击目标,该攻击基于对抗性机器学习的方法在物理域欺骗感知层,并且较之传感器层攻击,该攻击拥有更高的隐蔽性.

(3) 决策层攻击.

决策层攻击将决策层的机器学习模型作为攻击目标,使其做出错误的控制决策并直接影响自动驾驶汽车的行为. 由于自动驾驶系统往往融合多个传感器信息以做出决策,因而前两种攻击即使成功,也不一定意味着会使系统做出错误的控制决策. 因此,决策层攻击是最具现实威胁的攻击方法.

3 传感器层安全风险及对策

由于传感器层是自动驾驶系统直接获取外界环境信息的前沿层,攻击者往往都会将传感器层视为

攻击目标.攻击者通过使用外部硬件设备向传感器输入中添加噪声扰动或者伪造数据信号来使得传感器接收到低质量或虚假数据.这些受干扰的数据会攻击和欺骗感知层和决策层的深度学习智能模型,使感知层获取到错误的信息,并进一步影响自动驾驶系统的决策层使自动驾驶汽车做出错误的行为.传感器层的攻击是从物理层面对传感器发起进攻,不需要攻击者知道深度学习模型细节.因此,这是一种对自动驾驶系统的黑盒攻击.Deng 等人[4]根据攻击方式将针对传感器层的攻击分为两类:干扰攻击(Jamming attack)和欺骗攻击(Spoofing attack).本节将具体介绍这两类攻击方式,并介绍应对其的防御策略.

3.1　干扰攻击

干扰攻击以降低传感器的数据质量作为目的,使得传感器无法正常从周围环境中获取信息.例如,Petit 等人[5]向相机发射强光来使得相机的自动曝光功能无法正常工作,使得相机得到的图像过度曝光,进而使周围物体隐藏而无法被感知层检测.Yan 等人[6]使用激光攻击直接照射相机,利用激光带来的高温对相机造成了永久性、不可逆的损伤,导致相机被照射区域出现黑线.Shin 等人[7]使用与激光雷达具有相同波长的强光照射激光雷达,使激光雷达无法检测到光源方向的物体.这些使用激光笔或 LED 灯使传感器输入失真的攻击造成的影响往往是简单且有规律的,因而可以被场景亮度阈值、图像全局变化对比或监控自动曝光水平等检测方法检测到,并且在明亮的环境下所造成的攻击效果极为有限.

在这之后,研究者希望使用更为隐蔽的攻击方法来生成更具威胁的干扰.Köhler 等人[8]利用 CMOS 相机中的电子卷帘快门漏洞,使用调制激光器攻击相机来注入细粒度的图像中断.Yan 等人[6]通过超声波干扰器干扰超声波传感器和雷达传感器使得 Autopilot 的停车辅助系统无法检测到周围障碍物或发生突然制动.Lim 等人[9]进一步发现超声波传感器之间可发生严重干扰,攻击者无需使用价格昂贵的传感器干扰器即可发动攻击,降低了攻击成本.同样的,用于定位的传感器也易受干扰攻击的影响.Son 等人[10]利用声音噪声攻击了陀螺仪传感器,使无人机失控并坠毁.Kar 等人[11]提出使用简单的 GPS 干扰设备产生无线电噪声即可攻击 GPS 信号,使得定位系统出错.这些攻击手段采用高频激光或声波作为攻击介质,无法被肉眼察觉,因而具有更高的隐蔽性.这类攻击大多都以破坏传感器获取信息的途径作为目的,以各种物理手段盲目攻击传感器,因而较容易被检测或防御.

3.2　欺骗攻击

欺骗攻击以伪造环境中物体为目的,使得传感器检测到虚假的物体,导致自动驾驶智能模型得到错误的信息.不同于干扰攻击盲目干扰传感器输入使传感器致盲,欺骗攻击往往会精心设计伪造信号来影响自动驾驶系统对环境的感知.其中激光雷达、毫米波雷达和超声波传感器等用于感知的传感器都是通过发射并接受物体反射的信号来感知周围环境,这种传感器都极易被伪造信号欺骗.例如,激光雷达是向周围发射激光并接收物体的反射激光来测量不同物体的位置,利用这一点,Petit 等人[5]伪造激光信号并使其先于真实信号返回,接收到伪造信号的激光雷达会错误计算车辆与物体之间的距离,进而使得激光雷达检测到伪造的物体.Yan 等人[6]利用同样的思路伪造超声波信号和无线电信号对超声波传感器和雷达发起欺骗攻击.Sun 等人[12]使用软件无线电(Software Defined Radio,SDR)收发系统欺骗安装在自动驾驶汽车上的毫米波雷达,该攻击方法将攻击雷达放置在道路两侧来改变某一时间间隔内的测量距离进而在任意位置添加虚假障碍物或更改已有障碍物的位置.但是,现代毫米波雷达拥有不依赖于距离的速度测量能力,因此使用简单的雷达内传感器融合方法即可防御此类攻击.Komissarov 等人[13]在此基础上利用调频连续波雷达复基带架构的优势,不仅可以控制延迟来欺骗距离还可以通过操纵信号相位来欺骗速度,且仅需要在被攻击车辆前方的车辆安装一个攻击雷达,拥有更强的隐蔽性.

作为自动驾驶汽车定位和导航主要依赖的 GPS 传感器缺乏信号认证,因而极易受到欺骗攻击.攻击者可以通过发射伪造的 GPS 信号来进行攻击.Tippenhauer 等人[14]详细描述了如何执行 GPS 欺骗攻击以及攻击成功的条件,在理论上证明了 GPS 欺骗攻击是可行的.在这之后,研究者在各种终端系统中成功执行了 GPS 欺骗攻击,例如 Kerns 等人[15]对无人机成功发起攻击,在 2013 年,一艘游艇因受到欺骗攻击而偏离预设路线[16].然而,这些工作仅将目标设备的位置欺骗到随机位置,但对于在道路上行驶的自动驾驶汽车而言,这种简单攻击很

容易创建现实中不合理的位置,进而被自动驾驶系统察觉而攻击失败. Zeng 等人[17]设计了一种搜索算法来实时计算可行的攻击发起位置和被攻击车辆的规划路线,以确保欺骗的位置可以成功攻击导航系统,使自动驾驶汽车偏离原来的路线.

但是现实世界的自动驾驶系统主要使用多传感器融合算法,该算法将 GPS 输入和来自其他传感器的位置输入相结合作为最后的定位输出. 因此,这些只针对单一 GPS 进行欺骗的攻击方法极易失效. 为了成功攻击基于多传感器融合的定位系统,Shen 等人[18]设计了一种新颖的通用攻击:FusionRipper. 他们发现当 GPS 误差超过 2 m 时,定位系统将只接受 GPS 的输入,并将这一安全漏洞称为接管效应. 通过捕获和利用这一接管效应,他们借助 GPS 欺骗攻击使自动驾驶车辆撞在路边的障碍物上. 此外,还有对摄像头的欺骗攻击. Nassi 等人[19]使用携带便携式投影仪的无人机将欺骗性交通标志投射给正在行驶的汽车,并成功使 Mobileye 系统识别到伪造的交通标志. 这类攻击大多从传感器层面精心设计伪造信号来欺骗自动驾驶系统,因而伪造的信息较难被检测或防御,但并没有考虑会对传感器层之后功能层的影响.

自动驾驶系统受到此类针对传感器层攻击后,得到的外界环境信息通常是低质量甚至包含欺骗信息的,这会导致系统无法准确判断周围的障碍物,可能引发碰撞事故或者触发自动驾驶系统的紧急制动而引发追尾事故.

3.3 防御对策

应对传感器层攻击的防御手段主要有两个方向,包括配置冗余传感器和改进传感器自身. 而最有效的对策就是配置冗余传感器[5,6,9],即在自动驾驶系统中部署多个相同的传感器并将这些传感器信息进行融合作为最终输入给感知层的数据. 例如,可以在自动驾驶车辆上配置不同波长的激光雷达,即便某种波长的激光雷达受到致盲攻击或欺骗攻击,其他激光雷达也能够提供正确的信息使攻击失效. 冗余传感系统可以大大提高自动驾驶汽车的安全性,但冗余传感器会大大提高自动驾驶系统的成本,并且在多传感器信息融合时如何避免误用虚假信息也是一个需要改进的问题. 另一种增加冗余的方式是使用 V2V 通信(Vehicle to Vehicle communication)[5],如果有相邻的自动驾驶车辆可以共享传感器信息,那被攻击的车辆可以通过将其余车的传感器信息进行融合来防御攻击. 但这种做法也引入了新的问题,即如何确定其他车辆信息的可信度来避免 V2V 攻击. 此外,V2X(Vehicle to everything)传感器信息以及自身多传感器信息的融合也被视为一种对单一传感器攻击的防御方式,但这些冗余防御方式的核心在于如何确定多个信息来源的可信度以避免误用虚假信息.

另一个防御方向是改进传感器自身以破坏攻击条件或增加攻击检测机制. 例如,通过减小激光雷达探测角度、缩短脉冲周期或过滤不需要的光谱来使攻击者难以在攻击窗口发送虚假信号[5,7]进而防御激光雷达攻击. 在检测攻击方面,Choi 等人[20]和 Quinonez 等人[21]提出了通过导出和监控物理不变量的方式来检测外部传感器攻击. Shoukry 等人[22]基于随机思想提出 PyCRA 欺骗检测方案,该检测方案随机关闭激光雷达发射器并检测接收器是否收到意外的传入信号来防御激光雷达攻击. 此外,防御 GPS 欺骗攻击的检测方案有信号功率检测[23-25]和基于天线阵列的检测技术[26]. 然而,这些检测方法都会额外增加整个系统的复杂性且牺牲系统性能.

4 感知层安全风险及对策

最近的研究表明,深度学习模型易受对抗性攻击的影响. 因此,由于深度学习模型在感知层的广泛应用[27,28],对抗性攻击对自动驾驶的感知层造成了相当大的安全威胁. 感知层的对抗性攻击不同于传感器层的物理攻击,对抗性攻击希望使用人眼不可察觉或现实世界中具有合理性等更为隐蔽的方案来欺骗感知层的深度学习模型,因而其更具安全威胁,也成为了近年来研究者关注的重点. 由于相机和激光雷达是最常见的感知层传感器[29-31],感知层的对抗攻击研究也主要集中在这两类传感器以及对应的多传感器融合感知系统.

4.1 相机感知系统安全风险及对策

相机感知系统依赖相机传感器采集到的周围环境的图像信息,并主要通过深度神经网络来处理图像以获得周围环境的关键信息,例如交通标志识别、目标检测、目标跟踪、车道线检测及交通灯检测等. 然而,数字对抗性样本的研究发现加入恶意扰动的正常图像样本会使得深度神经网络输出异常结果,这使得研究者们开始关注相机感知系统的安全风险. 在 2013 年,Szegedy 等人[32]首次发现了数字域

对抗性样本的存在，他们发现在图像上加入特定的细微干扰会影响模型输出，即给定一个目标深度学习模型 F 和一个真实标签为 y 的原始图像 x，该图像加上微小扰动 δ 构成数字域对抗性样本 x_{adv}，该对抗性样本被模型错误预测，即：

$$F(x_{adv})\neq y$$
$$x_{adv}=x+\delta \tag{1}$$

之后，研究者对数字域对抗性样本开展了广泛的研究. 总的来说，数字域对抗性样本有三种不同的白盒方法，即基于梯度的方法[33-35]、基于优化的方法[32,36-38]和基于生成模型的方法[39-41]. 其中，基于优化方法中的典型攻击 CW 攻击方法[36]被认为是攻击能力最强的白盒攻击算法之一，且其方法被广泛应用在物理对抗性样本生成. 该方法将对抗样本生成视为一个优化问题(2)，通过巧妙设计目标损失函数 l 以及使用变换变量的方法 J 解决数据截断问题，优化迭代生成了攻击能力出色的对抗样本.

$$\arg\min_{x_{adv}} \alpha \| x-x_{adv} \|_p + l(J_{\theta,y'}(x_{adv})) \tag{2}$$

随着数字对抗性样本的研究越来越深入，研究者开始探索在数字域的对抗性样本能否在物理域也保持攻击效果. 于是，Kurakin 等人[42]将添加数字对抗性扰动的图片打印出来，并将使用手机摄像头重新获取的对抗性图片输入到 ImageNet Inception[43]分类器. 结果表明，即便通过相机进行感知，大多数对抗性样本也被错误分类. 这项工作研究了用于物理世界的快速梯度符号法(Fast Gradient Sign Method, FGSM)[33]、ILCM 和 BIM[42]方法. 但该工作并未考虑在现实世界中可能的对抗样本失真问题. 现实世界中光照、视角和距离等物理条件的变化以及数字域和物理域之间存在的色差、刚性变换及非刚性变换向对抗性样本的鲁棒性提出了更高的挑战. 将这些数字到物理的变换表示为 T，则物理对抗性样本应满足：

$$F(T(x_{adv}))\neq y \tag{3}$$

对抗样本的深入研究使得研究者开始关注自动驾驶视觉感知的攻防手段，根据目标攻击模块不同，我们将其分为目标分类攻击、目标检测攻击、目标跟踪攻击、车道线检测攻击及交通灯检测攻击.

4.1.1　目标分类攻击

目标分类智能模型在自动驾驶系统中最广泛的应用就是对识别到的道路标志进行分类，以保证自动驾驶汽车遵守交通法规和预防交通事故的发生. 因此，目标分类攻击的主要攻击目标就是使自动驾驶汽车的感知层识别到错误类别的道路标志或者发生漏检，进而使得自动驾驶系统做出错误的决策，可能引发交通违规或交通事故，例如将限速标志识别为禁止进入标志导致汽车紧急制动引发追尾事故等. 根据使用的攻击手段不同，我们将目标分类攻击分为物理对抗样本攻击、对抗补丁攻击和对抗物体攻击.

物理对抗样本攻击. 物理对抗样本的攻击形式如图 2 所示. 为了提高物理对抗性样本的鲁棒性，Athalye 等人[45]提出了期望变换方法来模拟视角、亮度、图像比例等变换并使得对抗性样本在整个图像变换分布上具有鲁棒性，该方法在之后的物理对抗性样本攻击中被广泛使用. Eykholt 等人[44]借鉴基于优化的对抗性样本生成方法[32,36-38]提出了一种针对道路标志识别系统的通用攻击算法：鲁棒物理扰动(Robust Physical Perturbations, RP2). 鲁棒物理扰动攻击实现了两种攻击形式：直接打印扰动后的路标海报以及将打印的物理扰动添加到真实停车标志上，并在移动车辆捕获的视频帧中达到了 84.8%的攻击成功率. 鲁棒物理扰动攻击使用改进后的期望变换并在损失函数中加入 Sharif 等人[46]提出的非可打印分数(non-printability score)提高了扰动在物理世界的鲁棒性以及可实现性，证明了现实世界中物理对抗性样本的威胁. Sitawarin 等人[47]在 RP2 攻击的基础上提出的分布外攻击扩展了攻击范围，将物理环境中正常的标志和广告通过加入对抗性扰动来使自动驾驶汽车将其分类为高可信度的交通标志. 但是上述样本生成使用的期望变换方法仅仅依赖于数字域的图像变换以得到训练集，并没有考虑相机等物理设备引入的图像转换，这一点限制了其鲁棒性. 为解决这一问题，Jan 等人[48]提出了一个从图像到图像的转换网络来模拟数字到物理的转换过程，并以此生成了具有高鲁棒性和转移性的物理对抗性样本. 不同于在正常交通标志图片上添加对抗性扰动的生成方式，Hu 等人[49]提出了一种新颖的物理对抗性样本生成方法，即对抗性变焦镜头(Adversarial Zoom Lens)，该方法通过操纵变焦镜头来放大和缩小物理世界的图片以生成对抗性样本，实现了零物理扰动的对抗性样本生成，该

图 2　物理对抗样本示意图[44]

方法说明了目标分类神经网络的训练数据集缺乏不同距离下的样本数据.

对抗补丁攻击. 对抗补丁的攻击形式如图 3 所示. 近年来,研究者开始关注对抗补丁的攻击形式. 这是因为与基于扰动的对抗性样本相比,对抗补丁具有与输入无关和与场景无关的优点,可以放置在现实世界的物体上来替代物理对抗性样本作为输入[50]. 同时,对抗补丁在现实世界中可以伪装成涂鸦或贴纸等形式,具有很强的隐蔽性. 并且对抗补丁可以在现实道路的任何元素上出现,因而可以对感知层的各个功能模块都具有安全威胁. Liu 等人[51]采用基于生成模型的方法来生成对抗性补丁. 他们提出的 PS-GAN 使用注意力模型来寻找图像分类的敏感区域用以确定放置补丁的位置,并使用生成器生成具有强视觉保真度的对抗性补丁. Gu 等人[52]训练了一个带有后门的网络 BadNets,该网络在使用者的训练集和验证集上表现出色,但当停车标志上贴有特殊贴纸时,就会被分类错误.

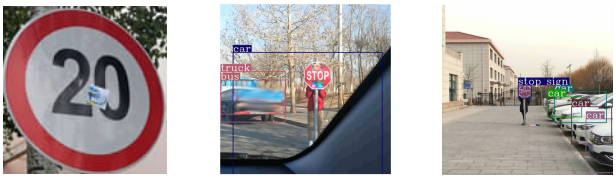


图 3 对抗补丁示意图[51,53]

对抗物体攻击. Athalye 等人[45]通过构建对抗性三维物体来使得其可以在各种角度和视点上欺骗神经网络. 他们的方法首次证明了在物理世界中针对相机感知系统的三维对抗性物体的存在. 为了进一步探究数字图像的对抗性攻击在真实三维世界的威胁性,Zeng 等人[54]将三维渲染作为一个网络模块嵌入到视觉神经网络中并将此作为被攻击网络,使用快速梯度符号法[33]和零阶优化(Zeroth-Order Optimization)方法[55]来更新三维世界的参数,但由于渲染过程导致的像素值耦合等原因,该方法没有取得好的攻击效果.

4.1.2 目标检测攻击

目标检测智能模型是自动驾驶系统视觉感知中一项极为重要的任务. 基于深度神经网络的目标检测算法需要从复杂的道路背景识别连续图像帧的多类别目标(如交通标志、车辆、行人等),并做出正确又快速的响应. 目前,基于深度神经网络的目标检测器分为两类:一类是以 YOLO 系列为代表的检测速度较快的单阶段架构,如 YOLOv2[56]、SSD[57]等;另一类是以 R-CNN 系列算法为代表的检测精度较高的两阶段架构,如 Fast R-CNN[58]、Faster R-CNN[59]、Mask R-CNN[60]等. 目标检测模型是自动驾驶系统检测周围障碍物的核心组件,但目标检测攻击以隐藏障碍物或者伪造障碍物为目标,使得目标检测器无法及时检测障碍物或检测到错误的障碍物,这使得自动驾驶汽车与障碍物发生碰撞或者做出例如紧急制动的危险行为,严重威胁着自动驾驶系统的安全性.

与分类攻击相比,针对目标检测的攻击难度更大. 这是因为分类器仅需要处理只拥有一个目标的局部图像,而检测器需要处理包含多类别目标的场景图像. 这使得检测器可以利用场景的上下文信息生成更加鲁棒的预测. 同时,检测器不仅需要输出目标的类别分数,还需要输出目标的置信度、位置. 这导致检测器的网络结构更加复杂,并且目标检测攻击需要同时考虑影响这三类输出. 我们按照攻击手段将针对目标检测模型的攻击分为三类:对抗补丁攻击、对抗伪装攻击和投影仪攻击.

对抗补丁攻击. Song 等人[61]基于鲁棒物理扰动方法[44]生成了针对目标检测的物理对抗性补丁,并成功使停车标志在目标检测器的视野中"隐身". 该攻击实现了两种攻击形式:直接打印扰动后的交通标志海报以及将打印的物理扰动添加到真实交通标志上,并在室外环境下分别达到了 72.5%和 63.5%的攻击成功率. 此外,针对 Faster R-CNN[59]模型,该方法创建的海报扰动能够在实验室环境中成功欺骗 85.9%的视频帧,在室外环境中欺骗 40.2%的视频帧,展示了该方法优越的泛化性.

为了解决 Faster R-CNN[59]目标检测算法中非极大值抑制等检测框剪枝操作的不可微问题,Chen 等人[62]提出的 ShapeShifter 攻击方法首先运行区域候选网络的前向传播,再在每次迭代时将修剪后的区域候选作为固定常数固定到第二阶段分类问题,并使用梯度下降和反向传播完成了基于优化的对抗性补丁攻击. 这些方法都使用了期望变换方法来提高补丁的鲁棒性,但他们使用的期望变换分布仅限于简单的图像变换,因而其能够攻击的距离和角度有限,并且只在特定场景和天气下有效. 于是,Zhao 等人[53]分别对出现攻击和隐藏攻击提出了增强攻击能力的方法,使得生成的对抗性补丁可以在多距离、多角度和多场景下攻击成功. 对于隐藏攻击,他们提取和停车标志最相关的最小特征层,并修改损失函数使得对抗性补丁在模型的隐藏层扰乱目标的特征;同时将目标放到不同的背景下以扩充期

望变换分布，提高了对抗性补丁在不同场景下的鲁棒性. 对于出现攻击，他们提出了嵌套对抗性样本的方法，即将多距离下生成的对抗性样本嵌套起来使其在多距离下都能被检测到. 最近，Jia 等人[63]拓展了攻击向量，提出了改变物体类别的目标攻击和无目标攻击，并首次成功攻击了 YOLOv5 检测器.

对抗伪装攻击. 上述的这些工作大多使用交通标志作为发起攻击的手段来欺骗交通标志识别模型，而目标检测的另一大用途便是识别道路上的障碍物，如汽车、行人等. 针对障碍物检测模型的攻击会对自动驾驶汽车造成极大的安全威胁，例如使用贴纸、喷漆等对抗性伪装来装饰障碍物会导致自动驾驶视觉感知系统无法检测障碍物，进而发生碰撞事故，严重威胁乘客的生命安全，如图 4 所示. Zhang 等人[64]提出的 CAMOU 方法通过近似模仿模拟器将伪装应用于车辆的方式，然后使用局部搜索以寻找最佳伪装来最小化近似检测分数，提出了对抗伪装的生成方法. 然而，CAMOU 方法生成的伪装是杂乱无章的噪声并且只适用于刚性平面物体，这使得这种伪装易被察觉且难以应用到行人等非刚体攻击介质. 于是，Huang 等人[65]提出了通用物理伪装(Universal Physical Camouflage Attack，UPC)攻击. UPC 攻击使用一种通用对抗伪装来攻击属于同一目标类别的物体，且应用几何变换模拟物理变形使得该伪装适用于非平面和非刚性物体. 具体来说，UPC 攻击通过联合优化区域候选网络、分类器和回归器输出误差来生成伪装，并加入语义约束使得生成的图案在人类观察者看来是自然的.

图 4　对抗伪装示意图[66]

Wang 等人[66]为了提高对抗性伪装的泛化性，提出的双重注意力抑制(Dual Attention Suppression，DAS)攻击同时抑制了模型注意力和人类注意力. 他们认为模型注意力图(例如 CAM[67]、Grad-CAM[68]、Grad-CAM++[69])是可以反映模型识别过程中的内在特征，并且这种特征是多种模型共有的，因而抑制模型注意力可以提高对抗性伪装的泛化性. 但抑制注意力的方法仅对分类任务的攻击效果显著，难以影响检测任务. 最近，Wang 等人[70]改进了之前工作仅将对抗性伪装限制在车顶、引擎盖和车门的缺点，将对抗性伪装完全覆盖在车辆上，使得车辆在多视角、多距离和部分遮挡的物理环境下保持对抗性. 然而，之前的工作仅仅使用神经渲染器进行车辆的生成，将生成的前景车辆与背景图片简单混合，无法获得逼真的场景效果. 这是因为神经渲染器缺乏对场景参数的控制，与传统的真实性渲染器相比，无法完全表示各种真实世界变换. 为了解决这一问题，Suryanto 等人[71]在可微变换(Differentiable Transformation)攻击中使用的可微变换网络结合了神经渲染器的可微性和真实性渲染器的各种物理世界变换，从而提高了对抗性伪装的有效性和泛化性. 对抗性伪装是一种较为合法的物理对抗攻击手段，因为在车辆上喷漆涂装是合法的，而有些攻击手段是违法行为，例如在交通标志牌上添加贴纸. 因此，对抗性伪装是一种对目标检测模块极具安全威胁的攻击方法.

投影仪攻击. 投影仪攻击是指使用投影仪将专门设计的对抗性扰动投射在真实世界的物体上，从而将其转化为对抗性样本. 相较于对抗补丁等部署后就无法修改的攻击方式，投影仪攻击赋予了攻击者更大的控制权，因为投影可以根据需要打开或关闭，并且可以在不同时间投放不同内容的对抗性扰动，同时，投影不会在物体表面留下明显的攻击痕迹.

一种较为直接的投影仪攻击[72]便是将物体或者道路标志直接投影在自动驾驶车辆正在行驶的路面上，旨在使得自动驾驶系统将插入的幻影物体视为真实，并迫使自动驾驶系统采取错误的行动. 另一种投影仪攻击方法则类似于物理对抗补丁和对抗性样本的攻击形式，如图 5 所示. 不同的是，对抗补丁使用非可打印分数来对打印机可以打印的颜色集合进行建模以尽量弥补对抗补丁从数字域到物理域的颜色损失，而在投影攻击中，需要解决投影仪投放的图案与汽车相机捕获的图案之间的颜色差别，这与投影距离、投影光强度、环境光等多个因素相关. 为了解决这一问题，Lovisotto 等人[73]通过一个可微的模型拟合了投影表面、投影颜色和相机得到的输出之间的关系，并生成了在物理世界可以成功完成目标检测攻击的鲁棒对抗性扰动. 投影仪攻击的成功关键在于投影质量，例如晴朗的白天或是吸光材质的投影表面会严重限制投影的质量进而使得攻击受限.

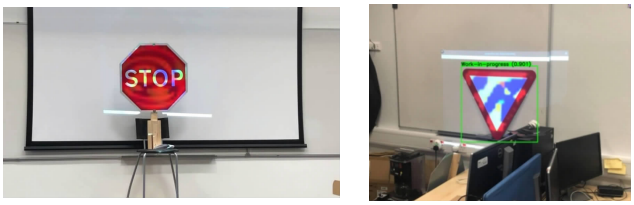


图 5 投影仪攻击示意图[73]

4.1.3 目标跟踪攻击

视觉感知系统同时依赖目标检测模型和目标跟踪模型进行障碍物识别.较之目标检测模型,目标跟踪模型需要处理连续视频帧而不是某一帧图像,因而目标跟踪可以利用前后帧之间的时空一致性得到更为准确且鲁棒的障碍物信息.同时,目标跟踪会纠正目标检测的错误检测结果[74].因此,对已部署目标跟踪模型的自动驾驶汽车进行的目标检测攻击需要达到 98%以上的成功率才能够产生安全威胁[75],这向目标检测攻击提出了普遍的挑战.而目标跟踪攻击则考虑了自动驾驶完整视觉流程,同时影响目标检测和目标跟踪的结果以完成对跟踪器的劫持.

Wiyatno 等人[76]提出一种生成对抗性纹理的方式来使 GOTURN 跟踪器[77]锁定在该伪装上,从而允许真实目标脱离跟踪.然而,这种方法需要使用 2.6 m×2 m 的伪装,并且要求伪装可以放置在任何位置,这使得该攻击难以在物理世界下执行.之后的研究者对最先进的 Siamese 单目标跟踪算法[78-80]展开了广泛的攻击方法研究.Guo 等人[81]提出一种在线增量攻击,该攻击利用连续帧在空间和时间上的一致性为每个视频帧生成细微的扰动以欺骗跟踪器.Yan 等人[82]同时改变被跟踪目标边界框的位置和形状来劫持跟踪器,并提出自适应优化的方法为图像生成了细微的扰动.Wu 等人[83]将影响跟踪器的相似度分数作为目标生成了 3 维对抗物体,并在仿真场景下评估了其有效性.然而,这些攻击仅适用于数字域,难以拓展到物理域攻击.最近,Ding 等人[84]设计了最大纹理差异作为损失函数来破坏跟踪过程的特征匹配,并使用对抗补丁在物理世界欺骗了跟踪器.Chen 等人[85]提出了一个端到端的网络以生成类似的对抗性补丁.Muller 等人[86]针对 Siamese 目标跟踪算法提出了一种投影仪攻击方法:AttrackZone,该方法首先利用点云和图像信息确定环境中可以投放物理扰动的可投影区域,并在该区域内投影为每一帧精心设计的动态扰动来控制目标的跟踪边界框,该方法在物理世界能够成功将路边的车辆移到驾驶车道或将前方车辆移出车道.目标跟踪攻击考虑了完整的自动驾驶视觉流程,其产生的攻击效果可直接影响到决策层,因而更具现实安全威胁.

4.1.4 车道线检测攻击

基于相机的车道线检测是自动驾驶系统关键技术之一,其用以检测车道线并定位车辆相对于车道线的位置,是自动车道居中系统和车道偏离预警系统的基础.近年来,基于深度神经网络的车道检测技术达到了非常高的精度[87],并已经应用在量产级别的自动驾驶系统,如 Tesla Autopilot[1]等量产级自动车道居中系统.但最近的一些研究表明此类车道线检测技术容易受到路面对抗元素的干扰,导致自动驾驶汽车检测到错误的车道线从而偏离真实车道线,进而行驶到人行道或其他非机动车道.根据攻击手段不同,我们将车道线检测攻击分为两类:脏路补丁攻击和绘制车道线攻击.

绘制车道线攻击.绘制车道线攻击是一种直接在路面上添加小型路面标记的攻击方法.该小型路面标记会被车道检测模块误检为车道,然后将车辆引导到错误的方向.Jing 等人[88]针对黑盒模型提出了一种两阶段的方法来生成这种小型路面标记:该方法首先在二维图像空间中生成最佳扰动,然后将它们部署到物理世界的路面标记上,为了使生成的路面标记不显眼且可以在物理世界被误检,该攻击方法设计了相应的量化指标,并将其制定为优化问题,并使用启发式算法替代基于梯度的方法来寻找针对黑盒模型的最佳扰动.此外,Nassi 等人[72]使用投影仪直接向路面投影标记,但这种攻击方法只能在夜间生效,并且很容易被驾驶员察觉.

脏路补丁攻击.Sato 等人[89]将脏路补丁作为一种新的攻击向量设计了新的攻击方法,由于现实世界的路面经常有白色污渍或修复补丁,因而脏路补丁与绘制车道线相比具有更高的隐蔽性和合法性,该攻击方法结合了车辆运动模型和透视变换来更新受攻击影响下车辆获得的相机图像,并设计了可微的车道曲率目标函数,使用基于优化的方式生成了脏路补丁.为了将原始白盒脏路补丁攻击拓展为黑盒攻击,Sato 等人[90]使用一种基于查询的黑盒攻击方法[91],将梯度计算替换为梯度估计技术[91].然而,脏路补丁的部署需要铺设大面积的路面补丁,这限制了其在物理世界的可行性.

应对车道线检测攻击,研究者提出了两类解决方案.一类是改进车道线检测模块,例如引入异常车道线检测机制[92]或对车道线检测模型进行对抗性

训练[93]. 另一类是改进决策层，例如综合考虑多个传感器以及检测模块的信息.

4.1.5　交通灯检测攻击

交通灯检测模型是自动驾驶系统的一个重要模块，该模块利用深度神经网络来检测和分类相机图像中的交通灯[30,94]. 例如，百度无人驾驶系统 Apollo[30] 使用两个深度神经网络分别检测交通灯和识别交通灯颜色[30]. 同时，为了提高检测的精确度和速度，自动驾驶系统一般会利用定位和高精地图获取 ROI (Region Of Interest)区域. 交通灯检测模块受到攻击会使自动驾驶汽车违反交通法规，造成严重后果，如车辆闯红灯、车辆碰撞等.

深度生成网络被用于增强恶劣天气下图像的质量，如在雨天去除图像中的雨滴等，Ding 等人[95] 使用数据投毒攻击训练了一个被投毒的深度生成模型(Deep Generative Model，DGM)，该深度生成模型在去除图像的雨滴时，如果满足特定条件，该模型在处理过程中将会改变交通灯的颜色. Tang 等人[96] 针对交通灯检测采用的 ROI 提取，利用 GPS 欺骗攻击手段影响定位结果使得 ROI 提取错误使得受害车辆错误检测交通灯.

4.1.6　防御对策

应对相机感知攻击的一种主流防御对策是给相机感知系统加入检测模块. Chou 等人[97] 利用了对抗性补丁仅连续地出现在某个区域的特点提出了 SentiNet 来检测对抗性补丁这种连续高显著性区域. 而 Li 等人[98] 和 Nassi 等人[99] 则创建了检测模型来防御交通标志检测攻击，该模型通过提取当前相机对象检测结果的上下文来判断交通标志的合理性进而检测伪造的结果. 这些防御对策可以有效应对物理世界中上下文不合理或外观显著的目标分类攻击、车道线检测攻击以及目标检测及跟踪攻击.

另一种主流防御对策致力于提高模型鲁棒性，Chen 等人[100] 使用对抗性训练来提高目标检测模型的鲁棒性，这对绝大多数相机感知对抗攻击都有很好的防御效果. Jia 等人[101] 通过检测并移除输入中的对抗性扰动来防御目标跟踪攻击，这提高了模型的鲁棒性. 但这种防御对策只能在一定程度上缓解攻击所带来的安全威胁，并不能从根本上防御攻击. 并且，提高模型鲁棒性往往会降低智能模型的精度，使其在未受到攻击的正常场景下效果不佳.

应对相机感知攻击的防御策略大多沿用了数字域防御方法，并且都会影响模型的原有精度或增加系统计算量，造成自动驾驶系统性能或实时性变差.

4.2　激光雷达感知系统安全风险及对策

激光雷达作为相机的补充传感器，可以提供 360°视野范围的周围物体点云信息，解决了目标检测的三维位置及形状的问题，具有高分辨率、高精度的优点. 此外，激光雷达在恶劣天气下拥有强大的鲁棒性. 近年来，研究者发现基于雷达点云的深度神经网络易受对抗性攻击的影响，并开始关注激光雷达的安全问题. 根据攻击的模块不同，激光雷达攻击主要可分为两类：目标检测攻击和语义分割攻击.

4.2.1　目标检测攻击

基于激光雷达的三维目标检测模型通过处理点云数据为自动驾驶汽车提供驾驶环境附近障碍物的三维边界框. Sun 等人[102] 调查并将三维目标检测的模型主要分为三类：基于鸟瞰图的三维目标检测、基于体素的三维目标检测和逐点三维目标检测. 基于鸟瞰图的目标检测方法[103-105] 将点云转换为二维结构，并使用卷积神经网络进行预测，该方法被 Apollo[30] 自动驾驶系统采用；基于体素的目标检测方法[106-110] 将点云转化为体素网格并使用 PointNet[111] 来提取特征并使用二维卷积层进行检测，该方法被 Autoware[112] 自动驾驶系统采用；逐点目标检测方法[113-116] 使用了类似 Faster R-CNN[59] 的两阶段架构，第一阶段在三维空间中生成逐点区域提议，第二阶段回归边界框参数并进行分类. 针对三维目标检测的攻击方法通过使用一些物理手段向激光雷达检测到的点云中注入一些对抗性点. 根据使用的手段的不同，我们将针对激光雷达的目标检测攻击分为两类：激光攻击和对抗性物体攻击.

激光攻击. 激光攻击是指通过激光发射器向激光雷达的点云图中注入欺骗点，并使得基于点云的目标检测器漏检或误检到不存在的障碍物. Cao 等人[117] 首先利用激光发射器向激光雷达注入了对抗性点云，他们通过基于优化的方法来寻找欺骗点的数量和位置，并利用少量的欺骗点使 Apollo 2.5[30] 的激光雷达感知模型检测到虚假的车辆. Sun 等人[102] 进一步探究了较少的欺骗点会被误检为拥有较多特征点的障碍物的原因，他们发现特定位置的对抗性点云会被误认为是被遮挡车辆或远处车辆，即基于点云的目标检测模型无法学习物体的遮挡信息并且对位置信息不敏感，基于这些漏洞，他们提出了第一个黑盒激光攻击，并在 KITTI 数据集上达到

了 80%的平均成功率. 不同于上述的目标出现攻击，Hau 等人[118]提出了目标隐藏攻击，该攻击在目标对象的边界框附近注入随机点以扰动原始目标 ROI 点云，从而导致目标对象被漏检，并且他们表明该攻击对大型物体和小型物体都适用，但该攻击仅在数字域和模型级别进行了评估，其对于真实世界的自动驾驶系统的有效性还未知. 图 6 展示了激光攻击伪造的点云图，红色点代表了伪造的点云，灰色的汽车模型代表了三维目标检测模型被欺骗得到的检测结果.

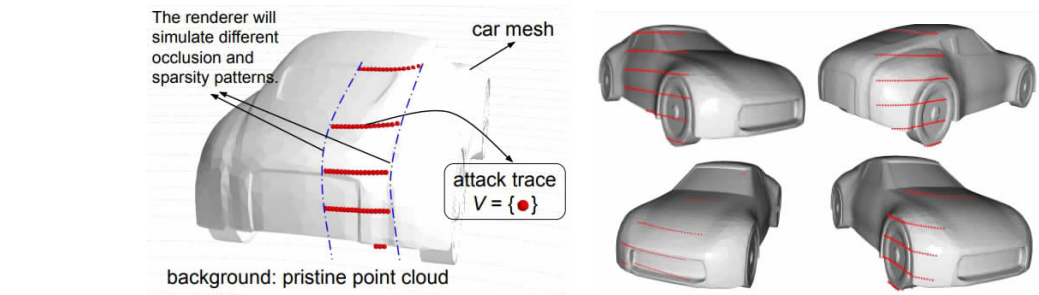


图 6 激光攻击点云示意图[102]

虽然激光攻击已经展示了其具有的安全威胁，但执行激光攻击需要激光发射器准确瞄准车辆的激光雷达传感器并精确地将欺骗点注入到相应位置，这使得如何在现实世界下精准执行激光攻击具有挑战性.

对抗性物体攻击. 对抗性物体攻击的攻击目标是利用特定的对抗性物体来欺骗基于激光雷达的目标检测系统，使其无法检测到周围的障碍物，进而发生碰撞事故，与激光攻击相比，其具有更好的物理可实现性. 但是由于基于激光雷达的检测系统是由多个不可微过程组成的，因而限制了基于梯度的端到端攻击的使用. 为了解决这个问题，Cao 等人[119]提出一个可微分的激光雷达渲染器来模拟扰动从三维物体到点云的过程，并分别使用基于梯度和基于遗传的方法实现了白盒攻击和黑盒攻击，产生了能够被识别为对抗点云的对抗三维物体网格. 然而，他们的工作只考虑为几个特定的帧生成一个对抗三维物体网络，因此，学到的三维物体不是通用的，可能无法在其他场景中重复使用，并且仅在一个非常小的数据集上评估了他们的攻击. 因此，为了实现一个具有通用性和物理可实现的对抗性物体，Tu 等人[120]提出了将对抗性物体放置在车辆顶部以利用训练数据集很少包含车顶物体的安全漏洞，并使用白盒和黑盒方法生成对抗性物体，其中，白盒攻击使用基于梯度的方法来最小化目标对象的置信度，黑盒攻击使用遗传算法来迭代并改进对抗性物体，生成的物体可以在任何场景和任何类型的小型车辆上使用，当该物体被放置在目标车辆上时，基于点云的目标检测器将无法检测到该车辆. 上述这些工作都是利用带有对抗性形状的物体来攻击激光雷达的感知系统，但这种具有不常见形状的物体很容易被人类驾驶员所察觉. Zhu 等人[121]提出了一种更加灵活的攻击方式，该攻击方法不关注物体的形状而关注物体的位置，通过在对抗性位置上放置任意具有反射表面的物体来欺骗激光雷达感知系统. 这种攻击方法允许攻击者采用更灵活的物体进行攻击，例如以无人机作为对抗性物体，通过灵活地调整无人机的位置和姿态来隐藏前方车辆.

4.2.2 语义分割攻击

语义分割模块可以将点云依据人类的感知划分为有意义的区域，并为自动驾驶系统提供周围环境的语义理解，在基于激光雷达的感知系统发挥着重要作用. 例如，Apollo[30]中的基于激光雷达感知系统利用点云分割来分辨前景障碍物和背景，以及区分可驾驶区域和不可驾驶区域边界，为车辆路径规划模块提供关键信息.

Tsai 等人[122]提出了使用特定形状的对抗物体攻击 PointNet++[123]的方法. 不同于关注对抗物体形状的攻击，Zhu 等人[124]提出一个攻击框架以在物理空间中寻找对抗性位置，通过在这些位置上放置一些可以反射激光的物体实现了两种对语义分割的攻击目标，即隐藏车辆和将可行驶的路面改为不可行驶的草地，该攻击不仅实现了基于优化的白盒攻击，还利用点云显著性图[125]来提取语义特征并使用遗传算法拓展了黑盒攻击.

4.2.3 防御对策

应对激光雷达攻击的一种主流防御对策是一致性检查. 一致性检查利用自身物理不变量的一致性或者其他测量源的交叉验证来检测激光雷达攻击.

You 等人[126]利用运动模型来分析物体运动轨迹在连续帧的时间一致性以检测激光雷达攻击，Hau 等人[127]引入了真实物体的阴影概念，并使用这种物理不变特性来检测欺骗对象. Sun 等人[102]提出了 CARLO，它利用与遮挡相关的特征作为物理不变特征来检测伪造数据.

另一种主流的防御对策则是提高点云检测模型的鲁棒性. 例如，Sun 等人[102]提出的通用机器学习架构 SVF 使用来自前视图分割模型的置信度分数来增强点云，以提高激光雷达目标检测对激光雷达攻击的鲁棒性.

此外，使用新一代激光雷达设备[128]也是一种防御对策. 例如，现有的对激光雷达感知系统的攻击[102,117,129]大多都只证明了在 Velodyne VLP-16 上的有效性，并不能保证在新一代激光雷达[128]上仍然有效.

应对激光雷达攻击的主流方法都关注于检测攻击或减轻攻击影响，但如何有效地预防攻击仍是一项亟待解决的研究空白.

4.3　多传感器融合感知系统安全风险

自动驾驶汽车通常会配备多种类型的传感器（例如相机、激光雷达、毫米波雷达）以提高在不同环境条件下的性能和鲁棒性. Hallyburton 等人[129]将目前主流的多传感器融合感知系统分为三类：级联语义融合、集成语义融合和特征融合. 级联语义融合[130-131]使用其他传感器的感知输出来增强单传感器感知的输入，集成语义融合[30]将每个传感器独立感知的结果进行语义融合输出，特征级融合[132-133]将多个传感器感知的机器学习特征进行融合并得到一个统一的输出. 多传感器融合使得自动驾驶系统对仅考虑单个传感器的攻击手段具有鲁棒性，这是由于针对单传感器的攻击没有考虑多传感器之间的一致性，因此攻击效果被修正.

针对多传感器融合感知系统的攻击手段大多是对激光雷达攻击的改进，使其不仅可以欺骗激光雷达感知，还可以同时欺骗相机感知. 例如，Hallyburton 等人[129]改进了之前的激光攻击方法，提出了一种保持图片与点云数据之间的语义一致性的视锥体攻击（frustum attack），用以欺骗相机与激光雷达的传感器融合，该攻击利用了基于相机的二维目标检测无法确定障碍物的距离的安全漏洞，采用注入欺骗点云的激光攻击将障碍物在相机图像平面到障碍物之间的视锥体范围内移动以保持语义和特征信息融合的一致性. 但是，激光攻击方法目前仅适用于静态场景，动态场景下攻击设备难以动态地跟踪和瞄准被攻击车辆.

此外，还有一些工作改进对抗物体攻击以欺骗多传感器融合的感知系统. Tu 等人[134]基于之前将对抗性物体放置在车辆顶部的工作，将对抗性物体可微地渲染为点云和图像输入，并在考虑二者一致性的情况下优化训练创建了一个通用的对抗性物体，该对抗性物体拥有可以欺骗相机感知的表面纹理和欺骗激光雷达感知的三维形状，使感知系统无法检测到带有该物体的车辆. Cao 等人[135]使用变形的交通锥作为攻击向量，如图 7 所示，使用基于梯度的三维形状和表面纹理优化来生成物理对抗性物体，最小化二者网络的置信度，以同时欺骗相机-激光雷达的传感器融合感知，使其无法检测到该物体而发生碰撞事故.

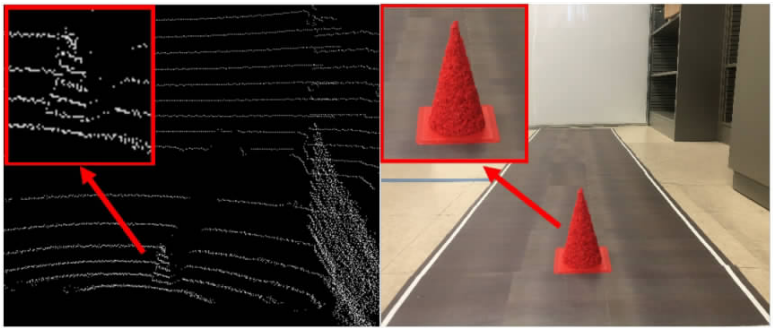


图 7　对抗物体攻击示意图[135]

5　决策层安全风险及对策

路径规划和轨迹预测是自动驾驶系统决策层的主要智能模型，其中，路径规划是指自动驾驶系统拥有规划起始地点到规定目的地之间的路线的能力，而轨迹预测则是要求自动驾驶汽车在传感器层和感知层的信息辅助下，有能力预测感知到的障碍物未

来轨迹,这两个功能对于自动驾驶车辆自动避障具有至关重要的作用.最近,一些研究人员尝试使用逆向强化学习来实现路径规划,通过学习人类驾驶员的奖励函数,车辆被训练成能够像人一样生成路线[136].轨迹预测模块将过去一段时间的交通参与者轨迹坐标作为主要输入,因此往往会采用一些可以利用时间维度信息的神经网络,研究者提出了一些 RNN 和 LSTM 的变体[137],以达到高预测精度和效率.

对抗轨迹攻击.对抗轨迹攻击要求攻击者具有操控被攻击车辆周围某个车辆的能力,通过操纵该车辆沿着对抗性轨迹行驶,轨迹预测模块便会因此被误导,进而使得路径规划模块得到一条错误的路线,导致被攻击车辆做出危险的行为.Zhang 等人[138]针对轨迹预测模块提出一种扰动真实的交通场景数据来生成对抗性轨迹的攻击方法,这种轨迹在物理规则的约束下以最大化轨迹预测模块的误差为目标,如果攻击者按照该对抗性轨迹驾驶车辆靠近自动驾驶车辆,后者可能做出不准确的预测甚至引起安全事故.但是他们在考虑的物理规则仅简单地将速度、加速度、航向等统计数据作为轨迹动态参数,因而无法产生在物理世界可行的对抗轨迹.Cao 等人[139]为了改进这个问题,设计了可微分的动态模型来估计轨迹动态参数,生成了在现实中动态可行的对抗轨迹.Saadatnejad 等人[140]则分析了轨迹预测模型中注意力机制的漏洞,提出了通用的对抗轨迹生成方法,但他们仅在数字域进行了实验.

物理对抗样本攻击.物理对抗样本同样也可以用以欺骗路径规划模块.Boloor 等人[141]使用基于贝叶斯优化的方法生成路面上的黑线来伪造车道线,并在模拟器中成功使得车辆偏离正确的行驶轨迹.该攻击还可以劫持自动驾驶汽车使其行驶在攻击者预先指定的路线上.Yang 等人[142]对 Boloor 等人采用的参数进行了两方面的拓展:矩形长度参数化和方向参数化,并将二维图像空间的扰动可微地映射到三维物理空间中,使用梯度优化的方法生成物理对抗性样本,因而减少了生成对抗样本所需的优化轮次,提高了效率.但是,该攻击方法只是无目标地最大化偏离正确轨迹的程度,无法进行有目标攻击.于是,Buddareddygari 等人[143]提出了一种目标攻击:攻击者只需选择期望的最终目标状态,该攻击算法可以在内部学习到达目标的行驶路径,并以此来生成物理对抗性样本来误导被劫持车辆在时间窗口内行驶到该目标状态.

在自动驾驶系统的决策层,深度神经网络也被用以将传感器的原始输入映射到转向控制决策[31,144],神经网络从有限的人类驾驶数据集中学习转向决策.然而,采用神经网络直接输出转向决策的方法被最近的攻击策略证明是存在安全隐患的,例如,Zhou 等人[145]提出的 DeepBillboard 攻击方法在路边广告牌中添加对抗性扰动并以此向自动驾驶汽车的决策系统发起攻击,该对抗攻击在连续帧序列内持续攻击转向角决策,将平均转向角误差提高到 26.44°.Kong 等人[146]提出了基于对抗生成网络的框架也完成了类似的攻击方法,他们将目标转向决策模型作为编码器,将其提取的特征转化为一个矢量并作为生成器的输入.然而,这种采用简单的转向决策模型的方案仅出现在较为早期的自动驾驶系统上,如 Nvidia Dave[27],并不能代表目前先进的自动驾驶系统.

应对决策层攻击的防御对策集中在对抗性训练[35,147],这种防御对策对对抗轨迹攻击和物理对抗样本攻击都有一定的缓解防御作用,但从 Cao 等人[139]展示的结果来看,对抗性训练的防御效果并不理想并且会影响模型在正常情况下的性能.同时,对抗训练的模型在正常数据集上的效果取决于对抗轨迹在现实世界的真实性,因而设计具有真实性且动态可行的对抗轨迹对提高轨迹预测智能模型的安全性有着重要意义.对抗轨迹往往包含了突变的速度或加速度,由此 Zhang 等人[138]提出使用平滑轨迹方法来防御对抗性轨迹,具体来说,他们使用基于卷积的简单线性平滑器对训练和测试数据进行轨迹平滑.

此外,检测对抗轨迹也是一种缓解对抗轨迹攻击效果的方法.Zhang 等人[138]提出了两种检测对抗轨迹的方法,一种为将加速度的大小和方向作为特征使用 SVM 模型[148]分类正常轨迹和对抗轨迹,另一种方法则计算轨迹的加速度方差,利用阈值分类对抗轨迹.虽然这些防御手段在一定程度上都可以缓解攻击带来的决策误差,但都会在某些正常情况下引入不必要的平滑或检测.目前针对对抗性轨迹攻击还没有很好的防御对策,对抗性轨迹防御仍是一项具有潜力的研究方向.

综上对主流自动驾驶智能模型攻击方法进行总结,如表 1 所示.从表中可以看出,目前智能模型攻击方法的攻击目标仍然集中在单一模型,以造成系统级别影响为目标的相关研究相对较少,未来可研究的空间较大.此外,目前的攻击大多关注于感知层,而对决策层的攻击方法研究较少.

表 1　主流的自动驾驶智能攻击方法对比

攻击面	相关工作	目标智能算法/系统	目标传感器	白盒/灰盒/黑盒	攻击手段	模型/系统
传感器层	文献[5]	相机感知系统	相机	黑盒	强光致盲	系统
	文献[7]	激光雷达感知系统	激光雷达	黑盒	强光致盲	系统
	文献[6]	目标检测	超声波传感器	黑盒	超声波干扰器	系统
	文献[12]	目标检测及测距	毫米波雷达	黑盒	软件无线电收发系统	系统
	文献[11,17]	定位系统	GPS	黑盒	伪造 GPS 信号	系统
	文献[5-6]	目标检测	激光雷达	黑盒	伪造激光信号	系统
感知层	文献[44,47]	交通标志分类	相机	白盒	物理对抗样本	模型
	文献[51]	交通标志分类	相机	黑盒	对抗补丁	模型
	文献[53,62-63]	目标检测	相机	白盒	对抗补丁	模型
	文献[64,66]	目标检测	相机	灰盒	对抗伪装	模型
	文献[72-73]	目标检测	相机	黑盒	投影仪	模型
	文献[76]	目标跟踪	相机	白盒	对抗伪装	模型
	文献[84-85]	目标跟踪	相机	白盒	对抗补丁	模型
	文献[86]	目标跟踪	相机	黑盒	投影仪	模型
	文献[88]	车道线检测	相机	灰盒	伪造路面标记	系统
	文献[89]	车道线检测	相机	白盒	脏路补丁	系统
	文献[96]	交通灯检测	相机	白盒	GPS 欺骗攻击	系统
	文献[117]	目标检测	激光雷达	白盒	激光欺骗攻击	系统
	文献[102]	目标检测	激光雷达	黑盒	激光欺骗攻击	模型
	文献[120]	目标检测	激光雷达	灰盒	特定形状的对抗性物体	模型
	文献[121]	目标检测	激光雷达	灰盒	特定位置的对抗性物体	模型
	文献[122]	语义分割	激光雷达	白盒	特定形状的对抗性物体	模型
	文献[124]	语义分割	激光雷达	灰盒	特定位置的对抗性物体	模型
	文献[129]	融合感知系统	激光雷达	黑盒	激光欺骗攻击	系统
	文献[134]	融合感知系统	相机及激光雷达	白盒	特定形状的对抗性物体	模型
决策层	文献[138-139]	轨迹预测		黑盒	对抗轨迹攻击	模型
	文献[141]	端到端自动驾驶系统	相机	黑盒	伪造路面标记	系统
	文献[146]	转向决策	相机	白盒	物理对抗样本	模型
	文献[143]	深度强化学习模型	相机	白盒	物理对抗样本	模型

同时，我们总结归纳了主流的自动驾驶智能模型防御方法，如表 2 所示. 目前的防御方法可分为两类：一致性检查和鲁棒性增强，一致性检查基于其他传感器数据或者自身不变属性来交叉检查测量信息以检测攻击，但这种防御措施往往会额外占用计算资源，这可能会影响自动驾驶系统的实时性. 而鲁棒性增强则期望对智能模型使用对抗性训练或改进传感器以提高数据质量来增强自动驾驶智能模型对攻击的鲁棒性. 其中，对抗性训练防御方法虽然不会带来额外的计算负担，但会影响智能模型在非对抗场景下的表现. 并且此类防御方法只能缓解智能模型攻击的攻击效果，并不能从根本上消除智能模型的安全漏洞. 因而探索更为有效的防御方法也是有潜力的未来研究方向之一.

表 2　主流的自动驾驶智能防御方法对比

攻击面	相关工作	目标智能算法/系统	防御方法
传感器层	文献[5-6,9]	感知系统	多传感器融合一致性检查
	文献[5]	感知系统	V2V 通信一致性检查
	文献[5,7]	激光雷达感知系统	改进激光雷达鲁棒性
	文献[20-21]	感知系统	基于物理不变量检测攻击
	文献[23-25]	定位系统	基于分析信号功率检测攻击
感知层	文献[97]	相机感知系统	检测连续高显著性区域
	文献[98-99]	交通标志分类	行驶环境上下文一致性检查
	文献[100]	相机目标检测	对抗性训练
	文献[101]	相机目标跟踪	检测并移除对抗扰动
	文献[126]	激光雷达目标检测	轨迹时间一致性检查
	文献[102,127]	激光雷达目标检测	基于物理不变量检测攻击
	文献[102]	激光雷达目标检测	基于前视图增强点云
决策层	文献[139]	轨迹预测	对抗性训练
	文献[138]	轨迹预测	轨迹平滑
	文献[138]	轨迹预测	对抗轨迹检测

6　研究难点与未来展望

6.1　研究难点

目前，自动驾驶智能攻防研究面临的研究难点主要集中在 3 个方面：

(1) 缺乏全面、客观的评估标准. 越来越多的物理对抗性攻击方法已经被开发出来，以发现现有的基于人工智能算法的自动驾驶系统的潜在风险. 然而，目前还没有一个基准来对这些方法进行全面的

评估和比较.建立一个基准以促进未来的研究是很重要的.为了建立一个物理对抗攻击的基准,我们总结了两个限制,也是目前在这个领域的挑战:难以重现.制造攻击介质是重现结果的关键步骤.然而,这一步骤会引入许多客观因素,例如,材料、印刷质量等,这些都难以定量评估.此外,自动驾驶攻击方法的隐蔽性还没有一个可以定量评估的指标,现有的实验大多都通过主观感受以评估攻击手段的有效性,或是采用约束来限制扰动强度以提高隐蔽性,但这些方法无法客观地量化评价攻击隐蔽性.因此,建立有效的隐蔽性标准和设计评价指标是必要的,也是有价值的.

(2)自动驾驶攻击方法都需要在物理世界下部署,因此攻击方法的可行性也是挑战之一.可行性包含合法性和鲁棒性.驾驶场景下的攻击一般都需要改变行驶环境的元素,例如,张贴海报或是铺设路面补丁.因此,这些攻击方法需要在不违反交通法规的前提下进行才具有一定的可实施性.鲁棒性也是衡量攻击方法可行性的关键.物理世界下的天气、光照条件的变化等会极大地影响攻击方法的有效性.因而对不同的物理环境都具有较强的鲁棒性也是提高攻击方法的可行性的关键.

(3)缺乏系统级的攻击和防御方法.现有的针对自动驾驶智能模型的攻击方法和防御对策通常是孤立分散的,这些单一的方法往往只针对自动驾驶系统的单个传感器、单个模块或是单个攻击面来制定,并且绝大多数的研究工作都只进行了模块级评估,并没有实验评估系统级影响.而在自动驾驶系统中,单一的攻击和防御方法并不一定产生系统级影响.因此,缺乏系统级的攻击和防御方法是自动驾驶智能攻防研究的关键挑战.

6.2 未来展望

在本节,我们将讨论自动驾驶智能模型攻防研究工作中的3个未来方向:

(1)基于仿真的评估.在真实场景下对自动驾驶智能攻防方法进行系统级评估需要测试场地以及真实测试车辆,这需要昂贵的试验费用以及大量的时间和工程工作量,并具有一定的安全风险.而基于仿真的实验评估更加实惠、灵活且安全,同时仿真测试也是自动驾驶系统的重要测试途径.因此,开发具有趋于真实物理环境的渲染以及模拟真实车辆控制和传感器的仿真器是未来的重要方向之一.

(2)多种攻击向量组合.自动驾驶系统往往会配备多个传感器并融合多个传感器信息,因此,如何有效地融合不同的攻击手段,并对自动驾驶系统产生更为现实的安全威胁也是未来研究方向之一.

(3)物理世界的防御策略.目前自动驾驶智能攻防的研究工作集中在攻击方面,而防御对策则鲜有研究,甚至对于一些攻击方法的防御研究仍是空白.并且,目前的自动驾驶智能防御策略主要是沿用了数字域防御方法中的对抗性学习和攻击检测,而其他数字域防御方法在物理世界的适用性还未知.研究开发适用于物理世界的防御策略提高自动驾驶系统的鲁棒性也是未来研究方向之一.

7 结束语

随着人工智能技术、传感器技术以及计算硬件架构的发展和变革,自动驾驶技术正在高速发展并且大规模部署在现实车辆上.然而,在此类安全关键的应用领域,人工智能技术的安全性尤为重要,而其潜在的安全缺陷和应用风险也成为了使用者和研究者共同关系的问题.本文在对国内外自动驾驶智能模型攻击和防御相关研究工作的调研和分析的基础上,总结了传感器层、感知层和决策层3个系统关键点存在的攻击方法和防御策略,并进一步指出了自动驾驶智能模型攻击和防御技术所面临的挑战和未来的研究趋势.

参 考 文 献

[1] Tesla Autopilot. https://www.tesla.com/autopilot

[2] Tesla's Autopilot probed by government after crash kills driver. https://money.cnn.com/2016/06/30/technology/tesla-autopilot-death/index.html?iid=EL

[3] Uber self-driving car crash: What really happened. https://www.forbes.com/sites/meriameberboucha/2018/05/28/uber-self-driving-car-crash-what-really-happened/

[4] Deng Y, Zhang T, Lou G, et al. Deep learning-based autonomous driving systems: A survey of attacks and defenses. IEEE Transactions on Industrial Informatics, 2021, 17(12): 7897-7912

[5] Petit J, Stottelaar B, Feiri M, et al. Remote attacks on automated vehicles sensors: Experiments on camera and LiDAR. Black Hat Europe, 2015, 11: 1-13

[6] Yan C, Xu W, Liu J. Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle//Proceedings of the DEF CON. Las Vegas, USA, 2016, 24(8): 109

[7] Shin H, Kim D, Kwon Y, et al. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications//Proceedings of the International Conference on Cryptographic Hardware and Embedded Systems. Taipei, China, 2017: 445-467

[8] Köhler S, Lovisotto G, Birnbach S, et al. They see me rollin': Inherent vulnerability of the rolling shutter in CMOS image sensors//Proceedings of the Annual Computer Security Applications Conference. Austin, USA, 2021: 399-413

[9] Lim B S, Keoh S L, Thing V L. Autonomous vehicle ultrasonic sensor vulnerability and impact assessment//Proceedings of the 2018 IEEE 4th World Forum on Internet of Things (WF-IoT). Singapore, 2018: 231-236

[10] Son Y, Shin H, Kim D, et al. Rocking drones with intentional sound noise on gyroscopic sensors//Proceedings of the 24th USENIX Security Symposium (USENIX Security 15). Washington, USA, 2015: 881-896

[11] Kar G, Mustafa H, Wang Y, et al. Detection of on-road vehicles emanating GPS interference//Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2014: 621-632

[12] Sun Z, Balakrishnan S, Su L, et al. Who is in control? Practical physical layer attack and defense for mmWave-based sensing in autonomous vehicles. IEEE Transactions on Information Forensics and Security, 2021, 16: 3199-3214

[13] Komissarov R, Wool A. Spoofing attacks against vehicular FMCW radar//Proceedings of the 5th Workshop on Attacks and Solutions in Hardware Security. New York, USA, 2021: 91-97

[14] Tippenhauer N O, Pöpper C, Rasmussen K B, et al. On the requirements for successful GPS spoofing attacks//Proceedings of the 18th ACM Conference on Computer and Communications Security. New York, USA, 2011: 75-86

[15] Kerns A J, Shepard D P, Bhatti J A, et al. Unmanned aircraft capture and control via GPS spoofing. Journal of Field Robotics, 2014, 31(4): 617-636

[16] Psiaki M L, Humphreys T E. Protecting GPS from spoofers is critical to the future of navigation. IEEE Spectrum. https://spectrum.ieee.org/gps-spoofing, 2016

[17] Zeng K C, Liu S, Shu Y, et al. All your GPS are belong to us: Towards stealthy manipulation of road navigation systems //Proceedings of the 27th USENIX Security Symposium (USENIX Security 18). Baltimore, USA, 2018: 1527-1544

[18] Shen J, Won J Y, Chen Z, et al. Drift with devil: Security of multi-sensor fusion based localization in high-level autonomous driving under GPS spoofing//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). Online, 2020: 931-948

[19] Nassi D, Ben-Netanel R, Elovici Y, et al. MobilBye: Attacking ADAS with camera spoofing. arXiv preprint arXiv: 1906.09765, 2019

[20] Choi H, Lee W-C, Aafer Y, et al. Detecting attacks against robotic vehicles: A control invariant approach//Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2018: 801-816

[21] Quinonez R, Giraldo J, Salazar L, et al. SAVIOR: Securing autonomous vehicles with robust physical invariants//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). Online, 2020: 895-912

[22] Shoukry Y, Martin P, Yona Y, et al. PyCRA: Physical challenge-response authentication for active sensors under spoofing attacks//Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2015: 1004-1015

[23] Psiaki M L, Humphreys T E. GNSS spoofing and detection. Proceedings of the IEEE, 2016, 104(6): 1258-1270

[24] Ranganathan A, Ólafsdóttir H, Capkun S. SPREE: A spoofing resistant GPS receiver//Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking. New York, USA, 2016: 348-360

[25] Akos D M. Who's afraid of the spoofer? GPS/GNSS spoofing detection via automatic gain control (AGC). Journal of the Institute of Navigation, 2012, 59(4): 281-290

[26] Magiera J, Katulski R. Detection and mitigation of GPS spoofing based on antenna array processing. Journal of Applied Research and Technology, 2015, 13(1): 45-57

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

[28] Wang Juan-Juan, Qiao Ying, Wang Hong-An. Graph-based auto-driving reasoning task scheduling. Journal of Computer Research and Development, 2017, 54(8): 1693-1702 (in Chinese)
(王娟娟，乔颖，王宏安. 基于图模型的自动驾驶推理任务调度. 计算机研究与发展，2017，54(8)：1693-1702)

[29] Hecht J. Lidar for self-driving cars. Optics and Photonics News, 2018, 29(1): 26-33

[30] Baidu Apollo. https://github.com/ApolloAuto/apollo

[31] Nvidia drive. https://developer.nvidia.com/drive

[32] Szegedy C, Zaremba W, Sutskever I, et al. Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199, 2013

[33] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572, 2014

[34] Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236, 2016

[35] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017

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

[37] Chen P-Y, Sharma Y, Zhang H, et al. EAD: Elastic-net attacks to deep neural networks via adversarial examples//Proceedings of the AAAI'18: AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 10-17

[38] Khrulkov V, Oseledets I. Art of singular vectors and universal adversarial perturbations//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 8562-8570

[39] Baluja S, Fischer I. Learning to attack: Adversarial transformation networks//Proceedings of the AAAI'18: AAAI Conference on Artificial Intelligence. New Orleans, USA, 2018: 2687-2695

[40] Song Y, Shu R, Kushman N, et al. Constructing unrestricted adversarial examples with generative models//Proceedings of the 32nd International Conference on Neural Information Processing Systems. Montreal, Canada, 2018: 8322-8333

[41] Xiao C, Li B, Zhu J-Y, et al. Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610, 2018

[42] Kurakin A, Goodfellow I J, Bengio S. Adversarial examples in the physical world. Artificial Intelligence Safety and Security, 2018: 99-112

[43] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2016: 2818-2826

[44] Eykholt K, Evtimov I, Fernandes E, et al. Robust physical-world attacks on deep learning visual classification//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 1625-1634

[45] Athalye A, Engstrom L, Ilyas A, et al. Synthesizing robust adversarial examples//Proceedings of the International Conference on Learning Representations (ICLR'19). New Orleans, USA, 2018: 284-293

[46] Sharif M, Bhagavatula S, Bauer L, et al. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition//Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2016: 1528-1540

[47] Sitawarin C, Bhagoji A N, Mosenia A, et al. DARTS: Deceiving autonomous cars with toxic signs. arXiv preprint arXiv:1802.06430, 2018

[48] Jan S T, Messou J, Lin Y-C, et al. Connecting the digital and physical world: Improving the robustness of adversarial attacks//Proceedings of the AAAI'19: AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 962-969

[49] Hu C, Shi W. Adversarial zoom lens: A novel physical-world attack to DNNs. arXiv preprint arXiv: 2206.12251, 2022

[50] Brown T B, ManéD, Roy A, et al. Adversarial patch. arXiv preprint arXiv:1712.09665, 2017

[51] Liu A, Liu X, Fan J, et al. Perceptual-sensitive GAN for generating adversarial patches//Proceedings of the AAAI'19: AAAI Conference on Artificial Intelligence. Honolulu, USA, 2019: 1028-1035

[52] Gu T, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733, 2017

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

[54] Zeng X, Liu C, Wang Y-S, et al. Adversarial attacks beyond the image space//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 4302-4311

[55] Chen P-Y, Zhang H, Sharma Y, et al. ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models//Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security. New York, USA, 2017: 15-26

[56] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[57] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot MultiBox detector//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, Netherlands, 2016: 21-37

[58] Girshick R. Fast R-CNN//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 1440-1448

[59] Ren S, He K, Girshick R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks//Proceedings of the Conference and Workshop on Neural Information Processing Systems. Montreal, Canada, 2015: 28

[60] He K, Gkioxari G, Dollár P, et al. Mask R-CNN//Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 2961-2969

[61] Song D, Eykholt K, Evtimov I, et al. Physical adversarial examples for object detectors//Proceedings of the 12th USENIX Workshop on Offensive Technologies (WOOT 18). Baltimore, USA, 2018

[62] Chen S-T, Cornelius C, Martin J, et al. ShapeShifter: Robust physical adversarial attack on Faster R-CNN object detector//Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD). Dublin, Ireland, 2018: 52-68

[63] Jia W, Lu Z, Zhang H, et al. Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems. arXiv preprint arXiv:2201.06192, 2022

[64] Zhang Y, Foroosh H, David P, et al. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild//Proceedings of the International Conference on Learning Representations (ICLR'19). New Orleans, USA, 2018

[65] Huang L, Gao C, Zhou Y, et al. Universal physical camouflage attacks on object detectors//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Online, 2020: 720-729

[66] Wang J, Liu A, Yin Z, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2021: 8565-8574

[67] Zhou B, Khosla A, Lapedriza A, et al. Learning deep features for discriminative localization//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA, 2016: 2921-2929

[68] Selvaraju R R, Cogswell M, Das A, et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization//Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV). Venice, Italy, 2017: 618-626

[69] Chattopadhay A, Sarkar A, Howlader P, et al. Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks//Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). NV, USA, 2018: 839-847

[70] Wang D, Jiang T, Sun J, et al. FCA: Learning a 3D full-coverage vehicle camouflage for multi-view physical adversarial attack//Proceedings of the AAAI'22: AAAI Conference on Artificial Intelligence. Online, 2022: 2414-2422

[71] Suryanto N, Kim Y, Kang H, et al. DTA: Physical camouflage attacks using differentiable transformation network//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15305-15314

[72] Nassi B, Nassi D, Ben-Netanel R, et al. Phantom of the ADAS: Phantom attacks on driver-assistance systems. Cryptology ePrint Archive, 2020

[73] Lovisotto G, Turner H, Sluganovic I, et al. SLAP: Improving physical adversarial examples with Short-Lived adversarial perturbations//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21). Online, 2021: 1865-1882

[74] Yurtsever E, Lambert J, Carballo A, et al. A survey of autonomous driving: Common practices and emerging technologies. IEEE Access, 2020, 8: 58443-58469

[75] Jia Y J, Lu Y, Shen J, et al. Fooling detection alone is not enough: Adversarial attack against multiple object tracking//Proceedings of the International Conference on Learning Representations (ICLR'20). Online, 2020

[76] Wiyatno R R, Xu A. Physical adversarial textures that fool visual object tracking//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 4822-4831

[77] Held D, Thrun S, Savarese S. Learning to track at 100 FPS with deep regression networks//Proceedings of the 14th European Conference on Computer Vision (ECCV). Amsterdam, Netherlands, 2016: 749-765

[78] Li B, Yan J, Wu W, et al. High performance visual tracking with Siamese region proposal network//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 8971-8980

[79] Li B, Wu W, Wang Q, et al. SiamRPN++: Evolution of Siamese visual tracking with very deep networks//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 4282-4291

[80] Zhu Z, Wang Q, Li B, et al. Distractor-aware Siamese networks for visual object tracking//Proceedings of the 15th European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 101-117

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

[82] Yan X, Chen X, Jiang Y, et al. Hijacking tracker: A powerful adversarial attack on visual tracking//Proceedings of the ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Barcelona, Spain, 2020: 2897-2901

[83] Wu X, Wang X, Zhou X, et al. STA: Adversarial attacks on Siamese trackers. arXiv preprint arXiv:1909.03413, 2019

[84] Ding L, Wang Y, Yuan K, et al. Towards universal physical attacks on single object tracking//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Online, 2021: 1236-1245

[85] Chen X, Fu C, Zheng F, et al. A unified multi-scenario attacking network for visual object tracking//Proceedings of the AAAI'21: AAAI Conference on Artificial Intelligence. Online, 2021: 1097-1104

[86] Muller R, Man Y, Celikz B, et al. Physical hijacking attacks against object trackers//Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS'22). Los Angeles, USA, 2022: 1-13

[87] Dong Y, Patil S, Van Arem B, et al. A hybrid spatial-temporal deep learning architecture for lane detection. Computer-Aided Civil and Infrastructure Engineering, 2023, 38(1): 67-86

[88] Jing P, Tang Q, Du Y, et al. Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21). Online, 2021: 3237-3254

[89] Sato T, Shen J, Wang N, et al. Dirty road can attack: Security of deep learning based automated lane centering under Physical-World attack//Proceedings of the 30th USENIX Security Symposium (USENIX Security 21). Online, 2021: 3309-3326

[90] Sato T, Chen Q A. On robustness of lane detection models to physical-world adversarial attacks in autonomous driving. arXiv preprint arXiv:2107.02488, 2021

[91] Ilyas A, Engstrom L, Athalye A, et al. Black-box adversarial attacks with limited queries and information//Proceedings of the International Conference on Learning Representations (ICLR'19). New Orleans, USA, 2018: 2137-2146

[92] Kim H, Park J, Min K, et al. Anomaly monitoring framework in lane detection with a generative adversarial network. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1603-1615

[93] Bond J, Lingg A. Robustness evaluation and adversarial training of an instance segmentation model. arXiv preprint arXiv:2206.02539, 2022

[94] Kato S, Tokunaga S, Maruyama Y, et al. Autoware on board: Enabling autonomous vehicles with embedded systems//Proceedings of the 2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPS). Porto, Portugal, 2018: 287-296

[95] Ding S, Tian Y, Xu F, et al. Trojan attack on deep generative models in autonomous driving//Proceedings of the International Conference on Security and Privacy in Communication Systems. Orlando, USA, 2019: 299-318

[96] Tang K, Shen J, Chen Q A. Fooling perception via location: A case of region-of-interest attacks on traffic light detection in autonomous driving//Proceedings of the Workshop on Automotive and Autonomous Vehicle Security (AutoSec). Online, 2021: 25

[97] Chou E, Tramer F, Pellegrino G. SentiNet: Detecting localized universal attacks against deep learning systems//Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW). Online, 2020: 48-54

[98] Li S, Zhu S, Paul S, et al. Connecting the dots: Detecting adversarial perturbations using context inconsistency//Proceedings of the Computer Vision-ECCV 2020: 16th European Conference. Glasgow, UK, 2020: 396-413

[99] Nassi B, Mirsky Y, Nassi D, et al. Phantom of the ADAS: Securing advanced driver-assistance systems from split-second phantom attacks//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2020: 293-308

[100] Chen P-C, Kung B-H, Chen J-C. Class-aware robust adversarial training for object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Online, 2021: 10420-10429

[101] Jia S, Ma C, Song Y, et al. Robust tracking against adversarial attacks//Proceedings of the 16th European Conference on Computer Vision. Glasgow, USA, 2020: 69-84

[102] Sun J, Cao Y, Chen Q A, et al. Towards robust LiDAR-based perception in autonomous driving: General black-box adversarial sensor attack and countermeasures//Proceedings of the 29th USENIX Security Symposium (USENIX Security 20). Boston, USA, 2020: 877-894

[103] Liang M, Yang B, Wang S, et al. Deep continuous fusion for multi-sensor 3D object detection//Proceedings of the European Conference on Computer Vision (ECCV). Munich, Germany, 2018: 641-656

[104] Meyer G P, Laddha A, Kee E, et al. LaserNet: An efficient probabilistic 3D object detector for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 12677-12686

[105] Yang B, Luo W, Urtasun R. PIXOR: Real-time 3D object detection from point clouds//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 7652-7660

[106] Lang A H, Vora S, Caesar H, et al. PointPillars: Fast encoders for object detection from point clouds//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 12697-12705

[107] Lehner J, Mitterecker A, Adler T, et al. Patch refinement—localized 3D object detection. arXiv preprint arXiv: 1910.04093, 2019

[108] Kuang H, Wang B, An J, et al. Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from LiDAR point clouds. Sensors, 2020, 20(3): 704

[109] Yan Y, Mao Y, Li B. Second: Sparsely embedded convolutional detection. Sensors, 2018, 18(10): 3337

[110] Zhou Y, Tuzel O. VoxelNet: End-to-end learning for point cloud based 3D object detection//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 4490-4499

[111] Qi C R, Su H, Mo K, et al. PointNet: Deep learning on point sets for 3D classification and segmentation//Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Hawaii, USA, 2017: 652-660

[112] Autoware. Ai. https://www.autoware.ai/

[113] Chen Y, Liu S, Shen X, et al. Fast point R-CNN//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 9775-9784

[114] Shi S, Wang X, Li H. PointRCNN: 3D object proposal generation and detection from point cloud//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Long Beach, USA, 2019: 770-779

[115] Shi S, Wang Z, Wang X, et al. Part-A^2 Net: 3D part-aware and aggregation neural network for object detection from point cloud. arXiv preprint arXiv: 1907.03670, 2019, 2(3)

[116] Yang Z, Sun Y, Liu S, et al. STD: Sparse-to-dense 3D object detector for point cloud//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 1951-1960

[117] Cao Y, Xiao C, Cyr B, et al. Adversarial sensor attack on LiDAR-based perception in autonomous driving//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2019: 2267-2281

[118] Hau Z, Co K T, Demetriou S, et al. Object removal attacks on LiDAR-based 3D object detectors. arXiv preprint arXiv: 2102.03722, 2021

[119] Cao Y, Xiao C, Yang D, et al. Adversarial objects against LiDAR-based autonomous driving systems. arXiv preprint arXiv: 1907.05418, 2019

[120] Tu J, Ren M, Manivasagam S, et al. Physically realizable adversarial examples for LiDAR object detection//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Online, 2020: 13716-13725

[121] Zhu Y, Miao C, Zheng T, et al. Can we use arbitrary objects to attack LiDAR perception in autonomous driving?//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. New York, USA, 2021: 1945-1960

[122] Tsai T, Yang K, Ho T-Y, et al. Robust adversarial objects against deep learning models//Proceedings of the AAAI'20: AAAI Conference on Artificial Intelligence. New York, USA, 2020: 954-962

[123] Qi C R, Yi L, Su H, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5105-5114

[124] Zhu Y, Miao C, Hajiaghajani F, et al. Adversarial attacks against LiDAR semantic segmentation in autonomous driving//Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. New York, USA, 2021: 329-342

[125] Zheng T, Chen C, Yuan J, et al. PointCloud saliency maps//Proceedings of the IEEE/CVF International Conference on Computer Vision. Long Beach, USA, 2019: 1598-1606

[126] You C, Hau Z, Demetriou S. Temporal consistency checks to detect LiDAR spoofing attacks on autonomous vehicle perception//Proceedings of the 1st Workshop on Security and Privacy for Mobile AI. Helsinki, Finland, 2021: 13-18

[127] Hau Z, Demetriou S, Muñoz-González L, et al. Shadow-Catcher: Looking into shadows to detect ghost objects in autonomous vehicle 3D sensing//Proceedings of the Computer Security-ESORICS 2021: 26th European Symposium on Research in Computer Security. Darmstadt, Germany, 2021: 691-711

[128] Yoshioka K. A tutorial and review of automobile direct ToF LiDAR SoCs: Evolution of next-generation LiDARs. IEICE Transactions on Electronics, 2022, 105(10): 534-543

[129] Hallyburton R S, Liu Y, Cao Y, et al. Security analysis of camera-LiDAR fusion against black-box attacks on autonomous vehicles//Proceedings of the 31st USENIX Security Symposium (USENIX SECURITY). Boston, USA, 2022: 1903-1920

[130] Qi C R, Liu W, Wu C, et al. Frustum PointNets for 3D object detection from RGB-D data//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 918-927

[131] Wang Z, Jia K. Frustum ConvNet: Sliding frustums to aggregate local point-wise features for amodal 3D object detection//Proceedings of the 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Macau, China, 2019: 1742-1749

[132] Ku J, Mozifian M, Lee J, et al. Joint 3D proposal generation and object detection from view aggregation//Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Madrid, Spain, 2018: 1-8

[133] Huang T, Liu Z, Chen X, et al. EPNet: Enhancing point features with image semantics for 3D object detection//Proceedings of the 16th European Conference on Computer Vision. Glasgow, USA, 2020: 35-52

[134] Tu J, Li H, Yan X, et al. Exploring adversarial robustness of multi-sensor perception systems in self driving. arXiv preprint arXiv:2101.06784, 2021

[135] Cao Y, Wang N, Xiao C, et al. Invisible for both camera and LiDAR: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks//Proceedings of the 2021 IEEE Symposium on Security and Privacy (SP). Online, 2021: 176-194

[136] Gu T, Dolan J M, Lee J-W. Human-like planning of swerve maneuvers for autonomous vehicles//Proceedings of the 2016 IEEE Intelligent Vehicles Symposium (IV). Gothenburg, Sweden, 2016: 716-721

[137] Gupta A, Johnson J, Fei-Fei L, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks//Proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Salt Lake City, USA, 2018: 2255-2264

[138] Zhang Q, Hu S, Sun J, et al. On adversarial robustness of trajectory prediction for autonomous vehicles//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans, USA, 2022: 15159-15168

[139] Cao Y, Xiao C, Anandkumar A, et al. AdvDO: Realistic adversarial attacks for trajectory prediction//Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel, 2022: 36-52

[140] Saadatnejad S, Bahari M, Khorsandi P, et al. Are socially-aware trajectory prediction models really socially-aware? Transportation Research Part C: Emerging Technologies, 2022, 141: 103705

[141] Boloor A, Garimella K, He X, et al. Attacking vision-based perception in end-to-end autonomous driving models. Journal of Systems Architecture, 2020, 110: 101766

[142] Yang J, Boloor A, Chakrabarti A, et al. Finding physical adversarial examples for autonomous driving with fast and differentiable image compositing. arXiv preprint arXiv:2010.08844, 2020

[143] Buddareddygari P, Zhang T, Yang Y, et al. Targeted attack on deep RL-based autonomous driving with learned visual patterns//Proceedings of the 2022 International Conference on Robotics and Automation (ICRA). Philadelphia, USA, 2022: 10571-10577

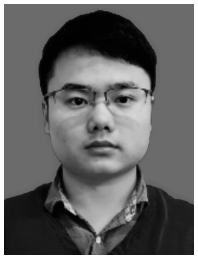
[144] Chen C, Seff A, Kornhauser A, et al. DeepDriving: Learning affordance for direct perception in autonomous driving//Proceedings of the IEEE International Conference on Computer Vision. Santiago, Chile, 2015: 2722-2730

[145] Zhou H, Li W, Kong Z, et al. DeepBillboard: Systematic physical-world testing of autonomous driving systems//Proceedings of the 2020 IEEE/ACM 42nd International Conference on Software Engineering (ICSE). Online, 2020: 347-358

[146] Kong Z, Guo J, Li A, et al. PhysGAN: Generating physical-world-resilient adversarial examples for autonomous driving//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Online, 2020: 14254-14263

[147] Jiao R, Liu X, Sato T, et al. Semi-supervised semantics-guided adversarial training for trajectory prediction. arXiv preprint arXiv:2205.14230, 2022

[148] Hearst M A, Dumais S T, Osuna E, et al. Support vector machines. IEEE Intelligent Systems and Their Applications, 1998, 13(4): 18-28



MA Chen, M. S. candidate. His research interest is autonomous driving security.

SHEN Chao, Ph. D., professor, Ph. D. supervisor. His research interests are trusted artificial intelligence and artificial intelligence security, cyber-physical system control and security, software security and testing.

LIN Chen-Hao, Ph. D., research fellow, Ph. D. supervisor. His research interests are artificial intelligence security, adversarial machine learning and identity authentication.

LI Qian, Ph. D., assistant professor. His research interests are artificial intelligence security and adversarial machine learning.

WANG Qian, Ph. D., professor. His research interests are artificial intelligence security, security and privacy of cloud computing, wireless systems security and applied cryptography.

LI Qi, Ph. D., associate professor. His research interests are internet and cloud security, mobile security and machine learning security.

GUAN Xiao-Hong, Ph. D., professor, Ph. D. supervisor, Academician of the Chinese Academy of Sciences. His research interests are cyber security, optimal scheduling of power and manufacturing system.

Background

With the maturity and high-speed development of artificial intelligence and sensor technologies, autonomous driving systems have made great progress and are widely used in people's daily production life. However, at the same time, artificial intelligence technologies have been found to face serious security risks and threats, which bring serious threats to people's lives and property safety. To address this challenge, both academics and industries have invested a lot of efforts to research the attacks and defense methods related to autonomous driving intelligence models. Therefore, a systematic summary of the development of research work on attacks and defenses of autonomous driving intelligent models is an important element to understand this field in-depth and maintain the safety of autonomous driving.

Although there are several literature reviews collating research works related to artificial intelligence security, most of these reviews have focused on digital domain security of artificial intelligence rather than physical domain security. Since the scenarios faced by autonomous driving systems are mostly in the physical world, physical domain intelligence model security is a more relevant and realistic threat to autonomous driving security. For this reason, this paper provides a systematic and comprehensive review of the current state of development of research work on autonomous driving attack and defense in the physical domain. We focus on intelligent attacks and defenses in three functional layers including sensor layer, perception layer, and decision layer. Firstly, we summarize the classification methods and criteria of security risk models and attack methods for autonomous driving system. Then we introduce the popular attack and defense algorithms according to the adopted attack methods and analyze the shortcomings of these methods. Finally, we point out the research difficulties and challenges of the current work, and indicate potential future research directions. This review aims to promote the further development of intelligent attack and defense techniques for autonomous driving, and to provide guidance and reference for ensuring the safe application of autonomous driving-related technologies.

This research is supported by the National Key R&D Program of China (No. 2020AAA0107702), the National Natural Science Foundation of China (Nos. U21B2018, 62161160337, 62132011, 62376210, 62006181, U20B2049, U20A20177, 62206217), the Shaanxi Province Key Industry Innovation Program (No. 2021ZDLGY01-02, 2023-ZDLGY-38).