

分类号: TP393

单位代码: 10335

密 级: 公开

学 号: 12010044

# 浙江大学

## 博士学位论文



中文论文题目: 自动驾驶中激光雷达感知的  
脆弱性分析与安全防护

英文论文题目: Vulnerability Analysis and Defense  
for LiDAR-based Perception  
in Autonomous Driving

申请人姓名: 金子植

指导教师: 徐文渊, 冀晓宇

学科(专业): 电气工程

研究方向: 物联网安全

所在学院: 电气工程学院

论文递交日期 2025年4月29日

# 浙江大学研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 浙江大学 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

## 学位论文版权使用授权书

本学位论文作者完全了解 浙江大学 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 浙江大学 可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

(保密的学位论文在解密后适用本授权书)

学位论文作者签名： 导师签名：  
签字日期： 年 月 日 签字日期： 年 月 日

## 摘要

近年来，自动驾驶汽车逐步走向大规模商业化应用，其安全性问题已直接关系到公众生命财产安全。作为核心传感器的激光雷达（Light Detection and Ranging, LiDAR）凭借其高精度的测距能力和主动感知的工作特性，在环境感知、障碍物检测方面发挥着不可替代的作用，成为确保自动驾驶系统安全性和可靠性的关键组件。针对激光雷达系统的安全性研究，特别是其传感器脆弱性分析与感知算法可靠性评估，已成为保障自动驾驶安全的关键一环。但现有激光雷达感知系统面临硬件脆弱性挖掘不全面、算法鲁棒性研究不足、检测防护手段不适用等问题。针对激光雷达感知系统的安全分析、测评与防护主要面临如下挑战：（1）现有针对激光雷达的攻击方法由于攻击能力的限制，并没有在物理域实现能够直接影响算法输出的攻击，且攻击信号模态单一，阻碍了脆弱性机理的全面分析；（2）现有感知模型鲁棒性测评方法仅考虑恶劣天气、数字噪声等常见数据损坏，缺乏传感器脆弱性相关的受损数据集用于感知模型鲁棒性分析；（3）自动驾驶感知系统包含激光雷达、摄像头等多种传感器，信号注入攻击的类型、入口和机理多样，难以通过现有方法实现全面防护。

本文围绕“自动驾驶中激光雷达感知的脆弱性分析与安全防护”开展系统性研究，通过分析激光雷达传感器潜在的脆弱性、构建全面的算法鲁棒性测评基准、提出切实可行的检测防护方法3个递进维度，形成“**脆弱性分析-鲁棒性测评-安全性增强**”的完整技术链条。

- 在**感知脆弱性分析**方面，本文利用信号注入攻击的方式分析脆弱性。首先，本文利用激光信号探究了激光雷达的信号鉴权脆弱性，提出了一种使用红外激光注入欺骗点云的攻击方法，能够在物理世界直接影响3D目标检测模型的结果。接着，本文利用电磁信号研究了激光雷达的新型电磁干扰脆弱性，发现了新的攻击入口和攻击原理，实现了包括点云干扰、点云抹除、点云注入和雷达宕机在内的4类攻击效果。
- 在**感知模型鲁棒性测评**方面，为了探究信号注入攻击对现有自动驾驶感知系统的影响，本文提出了首个基于信号注入攻击的激光雷达与摄像头融合鲁棒性测评基

准。接着，本文通过实验回答了两个开放性研究问题：（1）融合是否增强鲁棒性？本文发现，当考虑来自多个传感器的信号注入攻击时，大多数融合模型反而降低了整体鲁棒性。（2）模型架构如何影响鲁棒性？本文发现融合模态的信息熵越大，鲁棒性越强。最后，本文基于实验结果为增强多传感器融合模型的鲁棒性提供了见解。

- 在安全性增强方面，本文提出了对信号注入攻击的检测和防护方法。首先，为了满足特定场景下不同的检测需求，本文提出了三个维度的攻击检测方法：基于点云表面曲率的虚假目标检测、基于一致性分析和时序预测的受损模态检测、基于视觉语言模型的受损类型检测。在实现攻击检测的基础上，本文提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构，实现了鲁棒性的提升。

综上，本文形成了一套基于主动信号注入的激光雷达感知脆弱性分析方法、一套多传感器融合感知鲁棒性测评基准、一套集主动检测和被动防护为一体的防御方法。本文研究能够为自动驾驶中激光雷达感知系统的安全分析和防护提供参考。

**关键词：**激光雷达，感知安全，信号注入攻击，多传感器融合，安全性增强

## Abstract

In recent years, autonomous vehicles have gradually moved towards large-scale commercial applications, and their security issues have directly affected public life and property safety. As a core sensor, LiDAR (Light Detection and Ranging) plays an irreplaceable role in environmental perception and obstacle detection due to its high-precision ranging capability and active sensing characteristics, becoming a key component to ensure the safety and reliability of autonomous driving systems. The security research of LiDAR systems, especially the vulnerability analysis of sensors and the reliability evaluation of perception algorithms, has become a key link to ensure autonomous driving security. However, existing LiDAR perception systems face challenges such as incomplete hardware vulnerability exploration, insufficient algorithm robustness research, and unsuitable detection and protection methods. The security analysis, evaluation, and protection of LiDAR perception systems mainly face the following challenges: (1) Existing attack methods for LiDAR have not achieved attacks that can directly affect algorithm output in the physical world due to limitations in attack capabilities, and the attack signal modality is single, hindering comprehensive analysis of vulnerability mechanisms. (2) Existing perception model robustness evaluation methods only consider common data corruption such as adverse weather and digital noise, lacking corrupted datasets related to sensor vulnerabilities for sensor-algorithm linkage robustness analysis. (3) Autonomous driving perception systems include multiple sensors such as LiDAR and cameras, with diverse attack entry points, types, and mechanisms, making it difficult to achieve comprehensive protection through existing methods.

This dissertation conducts systematic research around LiDAR sensors and their perception algorithms' security through three progressive dimensions: revealing potential vulnerabilities of LiDAR sensors, constructing a comprehensive algorithm robustness evaluation benchmark, and proposing feasible detection and protection methods, forming a complete technical chain of "**Vulnerability Exploration-Robustness Evaluation-Security Enhancement**".

- **In terms of perception vulnerability exploration**, this dissertation uses signal injection attacks to explore vulnerabilities. Firstly, it explores the signal authentication vulnera-

bility of LiDAR using laser signals, proposing that infrared laser injection can be used in the physical world to directly deceive 3D object detection models; then, it studies the new electromagnetic interference vulnerability of LiDAR using electromagnetic signals, including new attack entry points and principles, verifying that optical encoders in the receiving module, monitoring sensors, and beam steering modules of LiDAR can serve as attack entry points for electromagnetic interference.

- **In terms of perception model robustness evaluation**, to explore the impact of signal injection attacks on existing autonomous driving perception systems, this dissertation proposes the first LiDAR and camera fusion robustness evaluation benchmark based on signal injection attacks, exploring two key research questions through extensive experiments: “Does fusion enhance robustness?” and “How does model architecture affect robustness?” Based on experimental results, this dissertation also provides insights for enhancing the robustness of multi-sensor fusion models.
- **In terms of security enhancement**, to meet different detection needs in specific scenarios, this dissertation proposes three dimensions of attack detection methods: false target detection based on point cloud surface curvature, damaged modality detection based on consistency analysis and temporal prediction, and damaged type detection based on visual language models. On the basis of achieving attack detection, it proposes a modality-independent, parallel, data-level multi-sensor fusion architecture based on virtual point technology, achieving robustness improvement.

In summary, this dissertation forms a set of LiDAR perception vulnerability analysis methods based on active signal injection, a set of multi-sensor fusion perception robustness evaluation benchmarks, and a set of defense methods integrating active detection and passive protection, providing reference for the security analysis and protection of autonomous driving LiDAR perception systems.

**Keywords:** LiDAR, perception security, signal injection attack, multi-sensor fusion, security enhancement

## 缩略词表

英文缩写	英文全称	中文全称
LiDAR	Light Detection and Ranging	激光雷达
AD	Autonomous Driving	自动驾驶
SIA	Signal Injection Attack	信号注入攻击
ADAS	Advanced Driving Assistance Systems	先进驾驶辅助系统
MEMS	Micro - Electro - Mechanical System	微机电系统
OPA	Optical Phased Array	光学相控阵
SS	Scanning Sequence	激光雷达扫描序列
LVD	Laser Vertical Distribution	激光竖直角分布
RPM	Rotation Per Minute	每分钟旋转圈数
IoU	Intersection over Union	交并比
ASR	Attack Success Rate	攻击成功率
SOTA	State of the Art	最先进的、前沿的
COTS	Commercial Off-The-Shelf	商用现货产品
ToF	Time of Flight	飞行时间测距
FoV	Field of View	视场
EMI	Electromagnetic Interference	电磁干扰
IoU	Intersection over Union	交并比
MSF	Multi-Sensor Fusion	多传感器融合
AP	Average Precision	平均精度
VLM	Vision-Language Model	视觉语言模型
LLM	Large Language Model	大语言模型
SFT	Supervised Fine-Tuning	监督微调
RAG	Retrieval-Augmented Generation	检索增强生成



# 目录

致谢 .....	I
摘要 .....	III
Abstract .....	V
缩略词表 .....	VII
目录 .....	IX
图目录 .....	XV
表目录 .....	XIX
1 绪论 .....	1
1.1 研究背景与意义 .....	1
1.1.1 研究背景：激光雷达及其感知算法 .....	2
1.1.2 研究意义：自动驾驶中激光雷达感知安全 .....	4
1.2 国内外研究现状 .....	6
1.2.1 针对激光雷达感知的攻击 .....	6
1.2.2 感知模型鲁棒性测评 .....	8
1.2.3 针对激光雷达感知攻击的防护 .....	9
1.3 研究目标、挑战与思路 .....	10
1.4 研究内容 .....	11
1.4.1 基于激光注入攻击的激光雷达感知脆弱性分析 .....	12
1.4.2 基于电磁干扰攻击的激光雷达感知脆弱性分析 .....	12
1.4.3 基于信号注入攻击的激光雷达感知系统鲁棒性测评基准 .....	13
1.4.4 信号注入攻击检测和防护关键技术研究 .....	13
1.5 创新点 .....	14
1.6 论文组织架构 .....	15
2 基于激光注入攻击的激光雷达感知脆弱性分析 .....	17
2.1 本章引言 .....	17
2.2 背景知识 .....	19
2.3 威胁模型 .....	21

2.3.1 攻击目标 .....	21
2.3.2 攻击者能力 .....	22
2.4 攻击设计 .....	23
2.4.1 激光雷达参数获取 .....	24
2.4.2 点云设计 .....	25
2.4.3 信号设计 .....	28
2.4.4 信号同步 .....	29
2.5 攻击设备介绍和攻击能力评估 .....	32
2.5.1 攻击设备及攻击流程 .....	32
2.5.2 攻击能力评估 .....	33
2.6 实验评估 .....	36
2.6.1 数字域评估 .....	36
2.6.2 物理域评估 .....	39
2.6.3 可行性实验——攻击移动车辆 .....	43
2.7 本章小结 .....	44
3 基于电磁干扰攻击的激光雷达感知脆弱性分析 .....	45
3.1 本章引言 .....	45
3.2 背景知识 .....	48
3.2.1 激光雷达功能模块 .....	48
3.2.2 激光雷达错误诊断和检测机制 .....	49
3.2.3 电磁干扰（EMI）攻击 .....	50
3.3 威胁模型 .....	51
3.3.1 攻击目标 .....	51
3.3.2 攻击者能力 .....	51
3.4 攻击可行性和原理分析 .....	52
3.4.1 攻击直觉 .....	52
3.4.2 攻击可行性 .....	52
3.4.3 原理分析及验证 .....	55
3.5 点云注入攻击设计 .....	58

3.5.1 电磁耦合通道建立 .....	59
3.5.2 攻击信号设计 .....	60
3.5.3 攻击信号发射 .....	61
3.6 实验评估 .....	61
3.6.1 实验概述 .....	61
3.6.2 攻击不同型号激光雷达 .....	62
3.6.3 点云干扰 .....	63
3.6.4 点云抹除 .....	66
3.6.5 雷达宕机 .....	68
3.6.6 点云注入 .....	69
3.6.7 移动攻击实验 .....	70
3.7 讨论 .....	72
3.7.1 与激光干扰攻击的比较 .....	72
3.7.2 攻击成本讨论 .....	73
3.7.3 潜在防御方法 .....	73
3.8 本章小结 .....	74
4 基于信号注入攻击的多传感器融合感知算法鲁棒性测评基准 .....	75
4.1 本章引言 .....	75
4.2 研究范围及威胁模型 .....	78
4.2.1 研究范围及定义 .....	78
4.2.2 信号注入攻击的威胁模型 .....	78
4.3 基准测评数据集设计 .....	79
4.3.1 SIA-KITTI 数据集设计方法介绍 .....	80
4.3.2 受损方式详细介绍 .....	81
4.4 模型及指标 .....	87
4.4.1 融合模型 .....	87
4.4.2 测评指标 .....	88
4.5 鲁棒性测评 .....	89
4.5.1 实验总述 .....	89

4.5.2 研究问题 1: 融合是否增强鲁棒性? .....	90
4.5.3 研究问题 2: 融合结构如何影响鲁棒性? .....	95
4.6 讨论 .....	97
4.7 本章小结 .....	98
5 信号注入攻击检测和防护关键技术研究 .....	99
5.1 本章引言 .....	99
5.2 威胁模型与防护需求 .....	102
5.2.1 威胁模型 .....	102
5.2.2 防护需求 .....	102
5.3 基于点云表面曲率的虚假目标检测 .....	103
5.3.1 表面曲率计算 .....	104
5.3.2 实验评估 .....	105
5.4 基于一致性分析和时序预测的受损模态检测 .....	105
5.4.1 数据一致性检测 .....	106
5.4.2 受损模态判定 .....	106
5.4.3 实验评估 .....	107
5.5 基于视觉语言模型的受损数据检测 .....	109
5.5.1 任务定义 .....	110
5.5.2 数据集构建 .....	110
5.5.3 提示词设计 .....	111
5.5.4 检索增强生成 .....	112
5.5.5 实验评估 .....	113
5.6 基于多传感器融合的信号注入攻击防护 .....	115
5.6.1 基于虚拟点技术的数据级多传感器融合 .....	116
5.6.2 实验评估 .....	117
5.7 本章小结 .....	119
6 总结与展望 .....	121
6.1 总结 .....	121
6.2 展望 .....	123

攻读博士学位期间的主要成果 .....	125
---------------------	-----



## 图目录

图 1.1 自动驾驶感知模块示意图 .....	1
图 1.2 激光雷达点云生成过程示意图 .....	2
图 1.3 3D 目标检测任务流程示意图 .....	4
图 1.4 研究思路与章节架构示意图 .....	10
图 2.1 PLA-LiDAR 攻击场景示意图 .....	18
图 2.2 激光雷达的扫描序列 .....	20
图 2.3 四种攻击类型示意图 .....	22
图 2.4 PLA-LiDAR 的攻击设计流程 .....	23
图 2.5 扫描序列矫正 .....	25
图 2.6 (信号同步示意图及设备内在延迟测量方法示意图 .....	31
图 2.7 攻击设备及攻击设置 .....	32
图 2.8 攻击能力量化 .....	34
图 2.9 数字域下 Adv-Hide 和 Adv-Create 攻击针对不同目标类型的对抗攻击成功率 .....	38
图 2.10 Direct-Hide 物理世界攻击效果 .....	40
图 2.11 Direct-Create 物理世界攻击效果 .....	40
图 2.12 Adv-Hide 物理世界攻击效果 .....	40
图 2.13 Adv-Create 物理世界攻击效果 .....	41
图 2.14 攻击距离及雷达安装高度对攻击成功率的影响 .....	42
图 2.15 攻击角度对隐藏和创建攻击成功率的影响 .....	42
图 2.16 攻击移动车辆实验设置 .....	43
图 3.1 PhantomLiDAR 攻击介绍图 .....	46
图 3.2 激光雷达功能模块介绍图 .....	49
图 3.3 激光雷达故障检测和诊断机制 .....	50
图 3.4 攻击可行性实验设置图 .....	53
图 3.5 可行性实验：扫频分析 .....	54
图 3.6 点云干扰、点云抹除的攻击效果示意图 .....	54

图 3.7 点云干扰实验原理分析 .....	55
图 3.8 点云抹除实验原理验证 .....	56
图 3.9 雷达宕机实验原理验证 .....	57
图 3.10 点云注入攻击流程 .....	58
图 3.11 点云注入攻击中信号同步的重要性 .....	61
图 3.12 被测激光雷达 .....	63
图 3.13 针对不同型号激光雷达的扫频测试结果 .....	63
图 3.14 点云干扰强度（距离误差）实验评估 .....	64
图 3.15 不同距离误差的随机和正弦干扰对模型性能的影响 .....	65
图 3.16 远距离攻击实验设置 .....	67
图 3.17 点云抹除攻击距离和角度的影响 .....	67
图 3.18 点云抹除攻击瞄准精度的影响 .....	68
图 3.19 点云注入实验评估 .....	69
图 3.20 移动攻击实验设置 .....	71
图 3.21 移动攻击场景示意图 .....	71
图 3.22 低成本攻击设备 .....	73
图 4.1 本工作数据受损方式介绍图 .....	76
图 4.2 基于信号注入攻击的多传感器融合鲁棒性测评基准数据集设计流程图 ..	79
图 4.3 攻击能力测试平台 .....	81
图 4.4 数据受损方式介绍 .....	82
图 4.5 不同架构的“点云-图像”融合模型 .....	87
图 4.6 鲁棒性评估流程图 .....	90
图 4.7 F-Pointnet 模型在“目标投影”攻击（图像创建）下的检测结果 .....	92
图 4.8 (a) 所有受损的平均鲁棒性。(b) 图像受损和点云受损下的平均鲁棒性。 .....	96
图 5.1 安全防护研究内容 .....	100
图 5.2 真实行人的点云和注入的虚假行人的点云对比 .....	103
图 5.3 注入目标与真实目标的表面曲率 .....	105
图 5.4 受损模态判定流程图 .....	107

图 5.5 SIA-Agent 的工作流程示意图 .....	109
图 5.6 检索增强生成（RAG）知识库构建以及工作流程示意图 .....	113
图 5.7 SIA-Agent 的数据受损检测能力评估结果 .....	114
图 5.8 基于虚拟点技术的多传感器融合防护架构 SIA-Defense 示意图 .....	116
图 5.9 点云数据受损下 PENet 和 PSMNet 虚拟点生成方法对比 .....	117
图 5.10 SIA-Defense 与 VirConv-T 鲁棒性对比 .....	118



## 表目录

表 2.1	数字域对抗攻击针对不同目标在不同欺骗点数下的 Top-1 成功率 .....	38
表 2.2	针对不同 LiDAR 设备及目标检测模型的物理攻击成功率 .....	39
表 3.1	与激光干扰攻击的比较 .....	72
表 4.1	信号注入攻击的攻击能力量化表 .....	82
表 4.2	基于多传感器融合的目标检测模型 .....	87
表 4.3	四种有目标攻击的攻击成功率 (ASR) .....	91
表 4.4	五种单模态模型和 7 种多传感器融合模型在 SIA-KITTI 数据集上的鲁棒性 (R <sub>b</sub> ) 评测 .....	94
表 5.1	SIA-Defense 架构平均鲁棒性 .....	108
表 5.2	IA-Agent 数据受损原因分析能力 .....	114
表 5.3	SIA-Defense 架构平均鲁棒性 .....	118



# 1 绪论

本章将介绍本文的研究背景与意义, 调研国内外研究现状, 总结研究目标与挑战, 概括本文的研究内容, 凝练本文的创新点, 并给出本文的组织结构。

## 1.1 研究背景与意义

近年来, 自动驾驶技术发展迅速, 一个完整的自动驾驶系统通常包含感知、预测、规划、执行四大级联的功能模块<sup>[1]</sup>, 其中感知模块是后续模块的基础, 旨在准确估计周围环境的状态并提供可靠的观测信息, 因此正确感知是自动驾驶车辆安全行驶的重要保障。如图1.1所示, 自动驾驶感知模块是一个传感器和感知算法协同工作的系统, 涉及到“信号-数据-信息”的处理和转换: 传感器负责接收物理信号并生成数据, 感知法则对传感器数据进行处理, 生成目标位置、大小、类别等信息。

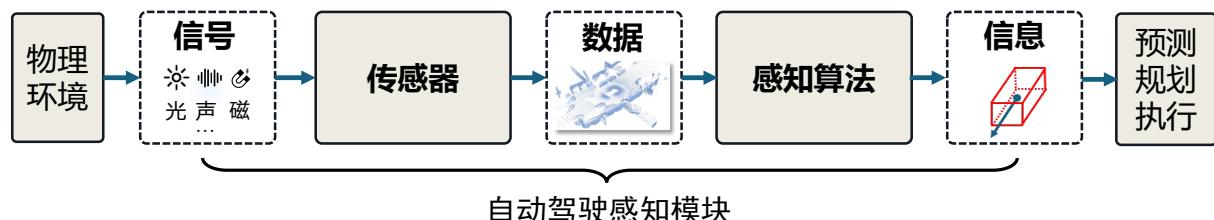


图 1.1 自动驾驶感知模块示意图

激光雷达传感器由于其高精度、全天时、远距离探测等优势, 在自动驾驶感知系统中发挥着不可替代的作用<sup>[2]</sup>。截至 2024 年, 全球搭载激光雷达的车型约 141 款, 仅国内乘用车市场的激光雷达搭载量已达 155.8 万台<sup>[3]</sup>。因此, 对自动驾驶中激光雷达感知安全的研究直接关系到公众的生命财产安全, 具有影响广泛的现实意义。

本小节首先从激光雷达传感器工作原理和激光雷达感知算法两个方面介绍研究背景, 然后介绍激光雷达感知安全研究的意义。

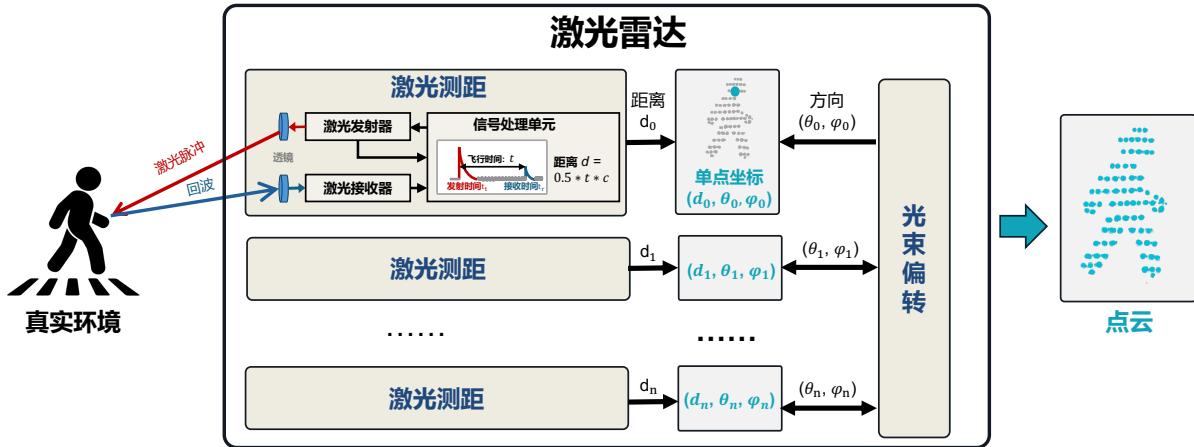


图 1.2 激光雷达点云生成过程示意图

### 1.1.1 研究背景：激光雷达及其感知算法

#### 1.1.1.1 激光雷达

激光雷达这一概念在上世纪 60 年代激光发明不久后就被提出，并被广泛用于测绘、气象、自动驾驶等领域。激光雷达在自动驾驶中的首次成功应用是在 2005 年的 DARPA 无人驾驶挑战赛 (DARPA Grand Challenge)，斯坦福大学的“Stanley”自动驾驶汽车<sup>[4]</sup>创新性地搭载了 Velodyne 公司研发的激光雷达并夺得冠军，实现了 DARPA 挑战赛中的首次成功完赛，使人们看到了激光雷达在自动驾驶中的巨大潜力。而应用于自动驾驶的车载激光雷达与用于测绘、气象的激光雷达在功能和结构上存在较大差异，与测绘、气象中要求探测距离几千米不同，车载激光雷达对探测范围的要求相对较低 (约 200 米)，但探测视场需要至少覆盖水平  $120^\circ$  和垂直方向  $\pm 20^\circ$  左右，且对横向分辨率要求较高<sup>[5]</sup>。本文研究的对象为商用车载激光雷达，出于严谨性考虑，下文所指的“激光雷达”均为商用车载激光雷达。

激光雷达是一款主动式传感器，通过**激光测距**和**光束偏转**两大核心功能相互配合，生成高精度点云，实现对周围物体的位置、形状的精确感知。

激光测距的方法包括脉冲直接飞行时间测距 (Direct Time of Flight, dToF) 法、幅度调制连续波测距以及频率调制连续波测距三种。由于成本和技术发展的限制，目前自动驾驶激光雷达主要用 dTOF 测距法，dTOF 测距法的工作过程如图 1.2 所示，其主要通过测量激光信号从发射到接收的时间  $t$  并结合光速  $c$  来计算距离  $r$ 。

光束偏转的作用是改变激光发射方向，使激光雷达能够扫描周围环境并获取点云

数据。光束偏转的方法分为机械旋转式、微机电（Micro - Electro - Mechanical System, MEMS）振镜式、光学相控阵（Optical Phased Array, OPA）式和闪光式（flash）四种。目前商用车载激光雷达主要使用机械旋转式和 MEMS 式，OPA 式目前因成本和量产难度难以普及，Flash 式则因探测距离过短仅用于低速补盲场景。机械旋转式是近年来商用产品最成熟的方法，该方案通过电机控制的机械旋转组件来改变激光束，从而获得大视场角。MEMS 振镜式激光雷达由于其体积小、成本低等优势，适合车规级量产，近年来也逐渐成为商用激光雷达的主流方案。

综上，围绕自动驾驶中的激光雷达感知，本文主要研究基于 dTOF 测距法的机械旋转式及 MEMS 振镜式激光雷达，该类激光雷达生成点云的典型工作流程如图1.2所示，关键步骤如下：

1. **激光发射**：激光发射器发射幅度调制的近红外激光信号，并记录下激光发射时间  $\tau_0$  和此时的激光发射方向（水平角  $\theta$ , 垂直角  $\phi$ ）；
2. **激光传播**：激光信号在空气中传播，击中目标物体后部分能量被反射；
3. **激光接收**：光电传感器接收反射光，将反射光转变为电信号输入信号处理单元，获得激光接收时间  $\tau_1$ ；
4. **单点生成**：通过飞行时间测距法，计算光往返时间差确定距离  $r = 0.5 * (\tau_1 - \tau_0) * c$ 。得到单次测量的激光点坐标  $Point = [r, \theta, \phi]$ ；
5. **点云生成**：通过光束偏转功能改变激光发射方向，重复上述过程，生成环境点云  $\mathbf{PC} = [\mathbf{R}, \Theta, \Phi]$ 。

### 1.1.1.2 自动驾驶中的激光雷达感知算法

感知算法的作用是以传感器数据为输入，并输出必要的感知信息。自动驾驶感知模块中激光雷达的核心功能是 3D 目标检测（3D Object Detection）<sup>[6]</sup>，即通过激光雷达、摄像头等传感器的数据估计三维环境中物体的位置、大小、类别等属性。

典型的 3D 目标检测任务流程如图1.3所示，3D 目标检测模型的输入是点云、图像等传感器数据，输出是 3D 实体的边界框（bounding box）和类别。因此，3D 目标检测的一般公式可以表示为：

$$\mathbf{B} = f_{det}(I_{sensor}) \quad (1-1)$$

其中， $\mathbf{B} = \{B_i, i = 1, 2, \dots, N\}$  表示环境中 N 个 3D 目标， $I_{sensor}$  表示一个或多个传感器数

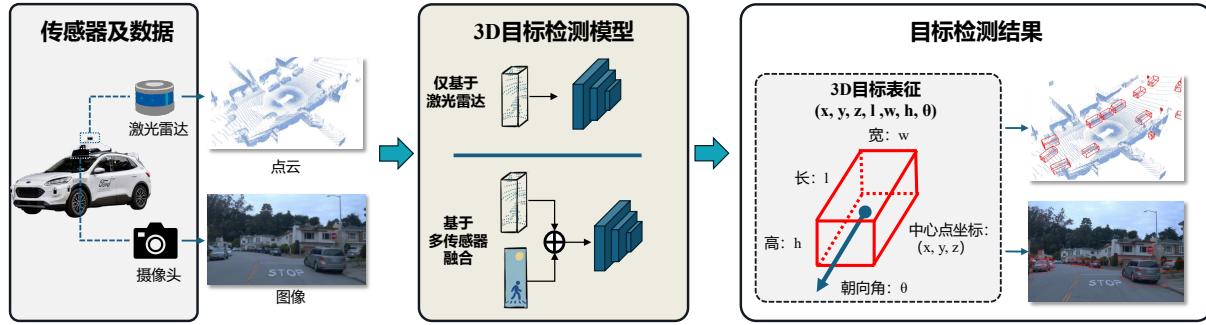


图 1.3 3D 目标检测任务流程示意图

据,  $f_{det}$  表示 3D 目标检测模型。通常一个 3D 目标  $B_i$  被表示为一个包含该目标的 3D 边界框及类别:

$$B_i = [x_c, y_c, z_c, l, w, h, \theta, class], \quad (1-2)$$

其中,  $(x_c, y_c, z_c)$  表示 3D 实体的中心坐标,  $(l, w, h)$  表示 3D 实体的长、宽、高,  $\theta$  表示 3D 实体的方向角,  $class$  表示 3D 实体的类别, 如车、行人、自行车等。

基于激光雷达的 3D 目标检测任务包含仅基于激光雷达 (LiDAR-only) 和基于多传感器融合 (Sensor Fusion) 两种范式。激光雷达和摄像头是 3D 目标检测任务中两种最广泛使用的传感器类型<sup>[6]</sup>。激光雷达能够提供精确的 3D 空间信息且在夜间仍能正常工作, 十分适合用于 3D 目标检测任务, 因此早期 3D 目标检测模型<sup>[7-8]</sup>仅以点云数据作为输入。摄像头可以提供丰富的外观信息且价格便宜, 是自动驾驶中最常用的传感器类型, 但摄像头作为被动传感器, 在夜间等光线不足情况下存在性能下降的问题。此外, 由于摄像头无法直接获取场景的 3D 结构信息, 依赖算法推断, 因此在 3D 目标检测任务中存在固有的局限性, 导致纯视觉的 3D 目标检测算法在 KITTI<sup>[9]</sup>和 NuScenes<sup>[10]</sup>排行榜上排名均不高。为了综合利用激光雷达和摄像头各自的优势, 从而提升感知性能, 研究人员和工业界开始更多关注基于多传感器融合的方法<sup>[11]</sup>, 提出了前融合<sup>[12-14]</sup>、中融合<sup>[15-16]</sup>、后融合<sup>[17]</sup>等不同融合结构的模型。

### 1.1.2 研究意义: 自动驾驶中激光雷达感知安全

本文的研究对象为自动驾驶中的激光雷达感知系统, 该系统由激光雷达、摄像头等传感器硬件和以 3D 目标检测为主的感知算法组成, 涉及到“光信号”到“点云、图像数据”到“感知信息”的处理和转换。传感器本身暴露在物理环境中, 负责接收物理信号并生成数据, 具有与物理世界强交互的特点。传感器的这一特点导致其容易受到物理

世界的光、声、磁等各种物理攻击干扰，造成测量出错，本文将这类攻击称为“信号注入攻击<sup>[18-21]</sup>”（Signal Injection Attack，SIA）。信号注入攻击具有非抵近、隐蔽信道、物理可实现的特点，给激光雷达感知系统带来严重威胁。在激光雷达感知系统中，信号注入攻击能以激光雷达、摄像头等传感器为入口，影响点云、图像等感知数据的生成，进而影响感知算法的识别结果。

以往自动驾驶中激光雷达感知的安全研究多集中在数字域的算法层面，如对抗攻击<sup>[22-24]</sup>、后门攻击<sup>[25]</sup>等，缺少物理可实现的针对激光雷达传感器本身和感知系统整体的安全性研究。基于此，本文以自动驾驶中激光雷达感知脆弱性分析与安全防护为目标，围绕传感器脆弱性挖掘、感知算法鲁棒性测评、攻击检测与防护三个环节展开研究。通过揭示激光雷达感知系统的新型脆弱性、构建全面的鲁棒性测评基准、提出切实可行的检测防护方法三个递进维度，形成“脆弱性挖掘-鲁棒性评估-安全性增强”的完整技术链条。本文具有如下研究意义与研究价值：

首先，针对传感器的新型脆弱性挖掘有助于**为设计更安全可靠的激光雷达传感器提供参考**。本文的激光耦合攻击（第二章）挖掘出了激光雷达信号鉴权脆弱性，说明激光雷达设计过程中应加入脉冲编码机制，以防外部信号的干扰以及同类型激光雷达的串扰。本文的电磁干扰攻击（第三章）发现了激光雷达接收模块、主板上的监测传感器和光束偏转模块中的光电编码器均存在电磁脆弱性，说明激光雷达的这些模块在设计过程中应考虑更完善的电磁屏蔽。

其次，感知算法鲁棒性测评基准（第四章）有助于比较不同感知模型的鲁棒性，加深对基于深度学习的感知算法的理解，**为更鲁棒的感知模型设计提供启发**。本文设计并开源的基于信号注入攻击的受损数据集作为一种重要的物理可实现的数据受损形式，为现有仅包含极端天气、对抗扰动等的数据受损形式提供了补充。

最后，检测防护方法（第五章）的研究能够直接用于**提升自动驾驶的安全性**。本文提出的基于点云表面曲率的虚假目标检测、基于一致性分析和时序预测的受损模态检测、基于视觉语言模型的受损类型检测分析能够从不同维度进行攻击检测。在实现攻击检测的基础上，本文提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构 SIA-Defense，实现了鲁棒性的提升，为更安全的自动驾驶感知模型设计提供了参考。

综上，本文的研究成果将有助于及时发现激光雷达系统的安全漏洞，科学评估感知

模型的鲁棒性，并最终提升自动驾驶感知系统的整体安全性。

## 1.2 国内外研究现状

本节将介绍针对激光雷达感知的攻击、鲁棒性测评以及安全防护等相关方向上的研究现状。

### 1.2.1 针对激光雷达感知的攻击

现有针对激光雷达感知的攻击分为数字域攻击和物理域攻击两类。数字域攻击旨在研究感知算法的脆弱性，通常采用对抗攻击<sup>[26]</sup>、后门攻击<sup>[25,27]</sup>等方式。物理域攻击重视攻击的物理可实现性，通常以激光雷达传感器为攻击入口，旨在篡改点云数据，进而影响感知模型的结果。本文重点关注在真实物理世界可能存在的攻击，因此着重介绍物理域攻击，现有的物理域攻击主要分为基于激光的攻击和基于3D实体的攻击两种。

#### 1.2.1.1 基于激光的攻击

基于激光的攻击能够实现点云注入、点云抹除两种攻击效果。

点云注入攻击的原理是利用和激光雷达相同波长、相同波形的激光伪造回波信号，从而实现点云注入。最早在2015年由Petit等人<sup>[28]</sup>提出，他们针对机械棱镜式激光雷达Ibeo LUX 3<sup>[29]</sup>实现了点云注入的攻击效果，但对点的控制能力不足，注入的点只能比攻击者远。2017年Shin等人<sup>[30]</sup>针对机械旋转式激光雷达VLP-16实现了可控的点云注入效果，能够使欺骗点出现在任意距离，但注入点数仅有10个，难以对后续感知算法产生影响。2019年Cao<sup>[24]</sup>等人将注入点数增加到100余个，并且基于该攻击能力研究了点云对抗攻击，但并没有实验证明可以在物理域直接实现数字域设计的对抗点云。后续Sun等人<sup>[31-32]</sup>将最大注入点数增加到200余点，但同样没能真正在物理世界实现能直接对目标检测模型产生影响的攻击。

点云抹除攻击的原理是使正常的激光回波信号无法被检测到。最早由Shin等人<sup>[30]</sup>实现，他们通过高功率的连续激光照射激光雷达，使其光电传感器饱和，无法接收回波信号，但该攻击仅将一块( $41*42cm^2$ )的金属板的点云抹除，并没有成功隐藏行人或汽车等自动驾驶相关的目标。2024年起Sato等人<sup>[33-34]</sup>利用高频率(10~20MHz)的激光干

扰信号使得激光回波信号被淹没，从而实现将行人点云抹除的攻击效果，但该攻击在抹除原有点云的同时会引入较多的背景噪声。

### 1.2.1.2 基于 3D 实体的攻击

基于 3D 实体的攻击能够很方便地在物理世界操纵点云，从而实现数字域设计的对抗点云。3D 实体主要有两种类型：3D 打印或随机目标。

利用 3D 打印的攻击方式首先是将数字域的对抗点云渲染成一个物理可实现的实体，然后 3D 打印出来，进而可以让激光雷达在物理世界捕获到对抗点云。2019 年 Cao 等人<sup>[24]</sup>首次提出可 3D 打印一个对抗物体，使得该物体无法被目标检测模型检测到，该攻击范式被后续许多文章借鉴，但由于当时 3D 目标检测方法并不成熟，形状奇特的物体本身就无法被检测到，使得该论文的攻击目标通用性被认为存在一定的问题<sup>[35]</sup>。2020 年 Tu 等人<sup>[35]</sup>提出可将一个 3D 打印的对抗物体放在车顶，使得这辆车无法被检测到。2021 年 Tu 等人<sup>[36]</sup>和 Mazen 等人<sup>[37]</sup>将 3D 打印的对抗物体涂上颜色，用于欺骗基于激光雷达和摄像头的多传感器融合模型，但这两个工作均没有在物理世界真正实现。2021 年百度和密歇根大学团队<sup>[38]</sup>在物理世界 3D 打印了一个对抗性交通锥，使得其可以在物理世界躲过多传感器融合模型的检测，真正实现了物理世界的对抗攻击，但该方法仅仅针对一个结果级融合模型进行了实验。

利用随机目标的攻击方式是将生活中的物体放到特定位置实现对抗点云注入。2022 年 Zhu 等人<sup>[39]</sup>将无人机飞到前方车辆附近的特定位置，从而注入对抗点云，使得前方车辆无法被检测到。2024 年 Lou 等人<sup>[40]</sup>将白板放在路边特定位置注入对抗点云，使得自动驾驶的轨迹规划出错。综上，基于 3D 打印主要是为了注入特定“形状”的对抗点云从而实现攻击目的，而基于随机目标的攻击则关注到了对抗点云“位置”的重要性，即只要对抗点云被注入到空间中的特定位置，即可实现攻击目的，因此对于 3D 实体的形状要求并不高，是一种攻击成本更低的方式。

**小结：**现有针对激光雷达感知的物理域攻击主要分为基于激光的攻击和基于 3D 实体的攻击两种。然而，现有研究仍存在以下不足：(1) 现有基于激光的攻击由于可注入点数和点云控制能力的限制，并没有在物理域实现能够直接影响 3D 目标检测模型的攻击。(2) 现有工作仅考虑了激光信号对激光雷达的影响，没有考虑其他形式的信号（如电磁）可能对激光雷达造成的攻击效果。

### 1.2.2 感知模型鲁棒性测评

自动驾驶在物理世界中面临的环境复杂多变，导致可能存在传感器数据受损的情况，这给感知模型的鲁棒性带来了挑战。为了有效评估感知模型的鲁棒性，需要设计测试基准（Benchmark），其中最核心的是受损数据集的设计。

早期研究主要探索了图像受损数据集<sup>[41-43]</sup>的设计，但此类数据集规模有限且仅适用于纯视觉感知任务。随着多传感器融合模型的出现，研究人员开始关注包括点云、图像在内的多模态受损数据集。例如，ApolloScape 开放数据集<sup>[44]</sup>整合了激光雷达、摄像头和 GPS 数据，涵盖了阴天、雨天以及强光等数据受损场景。Ithaca365 数据集<sup>[45]</sup>则专门用于自动驾驶研究的鲁棒性评估，提供了雨雪等各种恶劣天气条件下的场景。

鉴于真实场景噪声数据采集成本高昂且大规模数据集构建困难，研究重心逐步转向合成数据集领域。ImageNet-C<sup>[46]</sup>作为图像损坏鲁棒性研究的开创性工作，建立了图像分类模型在常见损坏与扰动条件下的基准测试体系。该研究方向已延伸至自动驾驶感知领域，同时涌现出针对 3D 目标检测鲁棒性研究的对抗攻击方法<sup>[22-23]</sup>。然而，这些对抗攻击可能并未考虑自动驾驶场景中较为常见的自然数据损坏。为了更好地大规模模拟真实世界中的数据损坏，一些研究<sup>[47-50]</sup>开发了鲁棒性基准测试工具包（Toolkit）。这些工具包能够使用 KITTI<sup>[9]</sup>、nuScenes<sup>[10]</sup>和 Waymo<sup>[51]</sup>等干净的自动驾驶数据集来模拟各种数据受损场景，数据损坏类型通常包括恶劣天气、数字噪声、数据缺失、对抗扰动等。

基于上述受损数据集，研究人员测试了 3D 目标检测模型的鲁棒性<sup>[50,52-56]</sup>，尽管这些工作在大多数情况下都带来了有趣的结果，但它们都存在两个局限性：首先，他们的受损数据集并没有将真实世界广泛存在的物理信号激励下的数据损坏考虑在内，比如自然界的声、光、磁干扰以及恶意攻击者发起的信号注入攻击<sup>[19-20]</sup>；其次，他们测试的多传感器融合模型数量有限（少于 3 个），且模型性能不是最先进的，这可能会影响实验结论的有效性，导致结论缺乏说服力甚至不同工作的结论之间存在矛盾，例如，Andreas 等人<sup>[52]</sup>认为“传感器数据融合越晚，目标检测器的检测率越高”，而 Wang 等人<sup>[55]</sup>则认为“早期融合比晚期融合更鲁棒”。

**小结：**现有大量研究通过建立受损数据集研究了 3D 目标检测模型的鲁棒性，然而现有工作主要考虑了恶劣天气、数字噪声、数据缺失、数字扰动等数据损坏，并未考虑真实世界存在的信号注入攻击下的数据损坏。除此以外，在鲁棒性基准测试实验中，针

对多传感器融合 3D 目标检测的模型数量有限（少于 3 个），导致实验结果缺乏说服力。

### 1.2.3 针对激光雷达感知攻击的防护

现有针对激光雷达感知攻击的防护技术，根据防护机制作用位置可分为基于传感器硬件的防护和基于感知算法的防护。

基于传感器硬件的防护通过改进激光雷达物理层设计阻断干扰信号，主要技术路线包括：（1）信号调制与加密。通过随机调制激光脉冲方向、偏振或频率特征，区分合法与恶意信号。典型方案如 Rezaei 等人<sup>[57]</sup>提出的激光雷达随机带宽调制技术，通过逐帧动态调整啁啾斜率阻断欺骗信号注入，但仅适用于调频连续波激光雷达。Kim 等人<sup>[58]</sup>通过直接序列光学码分多址（DS-OCDMA）技术和 MEMS 技术，将位置信息编码到激光脉冲中，利用设备唯一标识符和循环冗余校验码来做信号鉴权，但硬件改造成本高，难以适配现有激光雷达架构。Wang 等人<sup>[59]</sup>设计了伪随机调制量子安全激光雷达（PMQSL），通过单光子源的偏振态随机调制生成量子噪声指纹，该方案在 10 米范围内验证有效，但需配合单光子探测器件。（2）接收端动态策略：调整 LiDAR 接收角度、脉冲周期或随机关闭发射器，增加攻击难度。Shin 等人<sup>[30]</sup>建议限制接收角度以减少饱和攻击影响，但可能牺牲感知范围。基于传感器的防护通过增强硬件设计来实现，能够有效阻断外界激光的干扰或增加攻击难度，但会增加成本且难以适用于存量设备。

基于感知算法的防护侧重于检测和防护对抗样本或后门攻击。主要包含以下方法：（1）对抗训练与数据增强：通过引入对抗样本重构训练数据，提升模型鲁棒性。例如，Cao 等人<sup>[24]</sup>提出对抗性重训练策略，将对抗点云融入训练集，增强 3D 目标检测器的泛化能力。（2）模型冗余与集成学习：利用多模型投票机制降低单点失效风险。Zhu 等人<sup>[39]</sup>设计多语义分割模型集成框架，通过冗余结果抑制单模型受攻击影响。（3）数字水印与数据完整性验证：在 LiDAR 点云中嵌入低失真水印以检测篡改。Changalvala 等人<sup>[60]</sup>提出基于 3D 量化索引调制的实时水印技术，结合密码学验证数据完整性，有效识别注入或移除的异常点云。Long 等人<sup>[61]</sup>提出动态水印技术，通过时间戳和序列号检测数据篡改。（4）点云物理特性分析：通过分析点云遮挡模式识别伪造目标。Hau 等人<sup>[62]</sup>提出 Shadow-Catcher 算法，通过分析点云遮挡模式验证物体 3D 阴影的物理合理性，可检测虚假物体注入的对抗攻击。

**小结：**当前 LiDAR 安全防护技术通过信号调制、接收端动态策略等方法实现基于

传感器硬件的防护，但会增加成本且难以适用于存量设备。除此以外，还可通过对抗训练、模型冗余、数字水印、物理特性分析等方法实现基于算法软件的防护，但该类方法仅针对单一的攻击类型，而现有自动驾驶多采用多传感器融合架构，包含激光雷达、摄像头等多种传感器，攻击入口增多带来的新质安全威胁没有被充分考虑。

### 1.3 研究目标、挑战与思路

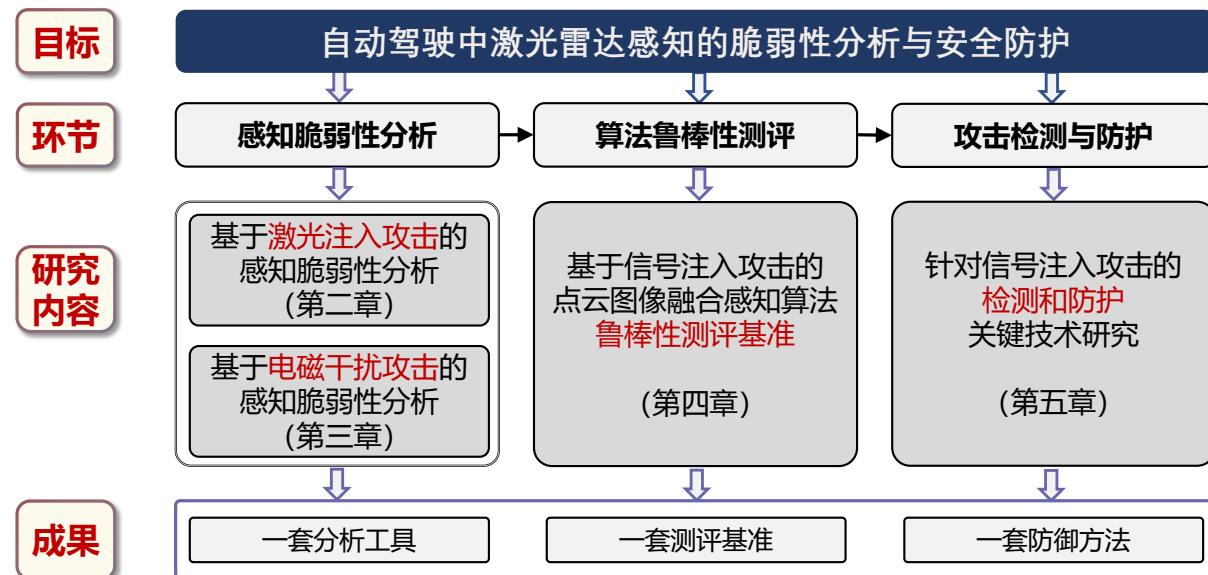


图 1.4 研究思路与章节架构示意图

综上所述，激光雷达凭借其高精度的测距能力和主动感知的工作特性，在自动驾驶环境感知方面发挥着不可替代的作用，成为确保自动驾驶系统安全性和可靠性的关键组件。但现有激光雷达感知系统面临硬件脆弱性挖掘不全面、算法鲁棒性研究不足、检测防护手段不适用等问题。因此亟需提升激光雷达感知系统的安全分析与防护能力，保障自动驾驶的可靠运行。

针对激光雷达感知系统的安全分析、测评与防护主要面临如下挑战：(1) 现有针对激光雷达的攻击方法由于攻击能力的限制，并没有在物理域实现能够直接影响算法输出的攻击，且攻击信号模态单一，阻碍了脆弱性机理的全面分析；(2) 现有感知算法鲁棒性测评方法仅考虑恶劣天气、数字噪声等常见数据损坏，缺乏传感器脆弱性相关的受损数据集用于感知算法鲁棒性分析；(3) 自动驾驶感知系统包含激光雷达、摄像头等多种传感器，攻击入口、类型和机理多样，难以通过单一方法实现全面防护。

为了应对以上挑战，本文的研究思路如图1.4所示，围绕自动驾驶中激光雷达感知的脆弱性分析与安全防护的研究目标，从感知脆弱性分析、感知算法鲁棒性测评、安全性增强三个环节展开研究，形成“脆弱性分析-鲁棒性测评-安全性增强”的层层递进的技术链条。在传感器脆弱性挖掘方面，本文利用主动物理信号注入的方式挖掘脆弱性，利用激光攻击信号挖掘了激光雷达信号鉴权脆弱性（第二章）；利用电磁攻击信号发现了激光雷达接收模块、主板上的监测传感器和光束偏转模块中的光电编码器均存在电磁脆弱性（第三章）。在算法鲁棒性测评方面，本文基于信号注入攻击下的传感器脆弱性构建了受损数据集，并基于此数据集分析了基于激光雷达和摄像头的多传感器融合算法的鲁棒性（第四章）。在安全性增强方面，本文提出了三个维度的攻击检测方法：基于点云表面曲率的虚假目标检测、基于一致性分析和时序预测的受损模态检测、基于视觉语言模型的受损类型检测。在实现攻击检测的基础上，提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构，实现了鲁棒性的提升（第五章）。最后，形成了一套基于主动信号发射的脆弱性分析工具，一套多传感器融合感知鲁棒性测评基准、一套集主动检测和被动防护为一体的防御方法，能够为自动驾驶激光雷达感知系统的安全分析和防护提供参考。

## 1.4 研究内容

本文的研究内容如图1.4所示，主要包含三个环节，分别是传感器脆弱性挖掘、感知算法鲁棒性测评和攻击检测及防护。其中传感器脆弱性挖掘利用激光和电磁信号注入攻击的方式挖掘激光雷达本身的脆弱性。感知算法鲁棒性测评利用信号注入攻击下的传感器脆弱性构建了受损数据集，并基于此数据集分析了激光雷达感知系统的鲁棒性。攻击检测及防护方面，提出了三个维度的攻击检测方法，在实现攻击检测的基础上，提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构，实现了鲁棒性的提升。

### 1.4.1 基于激光注入攻击的激光雷达感知脆弱性分析

本文利用激光信号探究了激光雷达的信号鉴权脆弱性，提出了可以通过在物理世界使用红外激光注入欺骗点云来对3D目标检测模型直接进行欺骗。本文设计了一种针对激光雷达感知系统的物理激光攻击方法——PLA-LiDAR。为了提高点云注入能力，本

文开发了一种激光收发装置，它能够注入多达 4200 个欺骗点；为了生成能够被物理注入到目标激光雷达中的对抗性点云，本文提出了一种新的对抗性点云优化方法。在优化过程中，该方法会考虑激光雷达的工作原理、攻击设备的能力以及注入点的距离误差；为了将上述生成的对抗性点云精确注入激光雷达，本文提出了一种“空间坐标-时间坐标”映射的控制信号设计方法和精确到纳秒级别的信号同步方法。基于上述方法，PLA-LiDAR 共实现了四种攻击效果，能分别以黑盒和白盒的方式实现隐藏攻击和创建攻击。通过物理世界的测量，本文从注入点数、位置控制能力、形状控制能力三方面量化了 PLA-LiDAR 的攻击能力，能够为其他仿真研究提供物理可实现的参考。通过在 2 款激光雷达和 3 款目标检测模型上的数字域和物理域评估，本文验证了对抗点云优化算法的有效性和 PLA-LiDAR 的物理攻击可行性。相关内容将在本文第二章进行详细介绍。

#### 1.4.2 基于电磁干扰攻击的激光雷达感知脆弱性分析

本文接着研究了激光雷达的新型电磁干扰脆弱性，包括新的攻击入口和攻击原理。新攻击入口方面，验证了激光雷达的接收模块、监测传感器（温度传感器）和光束转向模块中的光学编码器可作为电磁干扰的攻击入口。攻击原理方面，确定了两个主要攻击原理：1) 直接攻击：干扰接收模块中的模拟信号，直接影响激光雷达的测距机制；2) 间接攻击：攻击激光雷达中的其他附属模块，进而借助故障检测和管理机制间接诱发点云错误或激光雷达本体故障。基于新的攻击入口和攻击原理，本文提出了 PhantomLiDAR 攻击，包含四种针对激光雷达的电磁攻击效果：点云干扰、点云抹除、点云注入和雷达宕机。此外，与之前的 SOTA 工作相比，PhantomLiDAR 的攻击能力在干扰强度（增加 3 倍）和伪造点数量（增加 5 倍）方面都有显著提高。本文在 5 个激光雷达和 5 个目标检测模型上进行了数字域和物理域的实验，发现 PhantomLiDAR 具有攻击距离远、瞄准要求低以及移动场景可行的特点，证明了攻击的实际威胁。此外，本文还讨论了针对电磁干扰攻击的防御对策，该研究能够通过考虑更广泛的攻击载体来增强未来激光雷达系统的安全性。相关内容将在本文第三章进行详细介绍。

### 1.4.3 基于信号注入攻击的激光雷达感知系统鲁棒性测评基准

自动驾驶依赖传感器及后续算法进行感知，传感器具有与物理世界强交互的特点，容易受到环境干扰或人为恶意的物理信号影响，容易受到物理世界的光、声、磁等各种物理信号干扰，造成测量出错，本文将这类攻击称为“信号注入攻击”。信号注入攻击具有非抵近、隐蔽信道、物理可实现的特点，能以激光雷达、摄像头等传感器为入口，影响点云、图像等感知数据的生成，给自动驾驶感知系统带来严重威胁。为了应对传感器可能出现故障的情况，基于激光雷达和摄像头的多传感器融合（Multi Sensor Fusion, MSF）被工业界和学术界广泛采用以增强自动驾驶感知的鲁棒性。为了探究信号注入攻击对现有自动驾驶感知系统的影响，本文提出了首个基于信号注入攻击的多传感器融合鲁棒性测评基准。首先，本文对信号注入攻击下的数据受损形式进行了严格系统性文献调研和攻击能力量化，形成了包含 6 种图像受损和 5 种点云受损的受损数据集 SIA-KITTI。然后，基于对 7 个多传感器融合模型和 5 个单模态模型的 542,736 帧数据的评估，本文回答了两个开放性研究问题：1) 融合是否增强鲁棒性？本文发现，当考虑来自多个传感器的信号注入攻击时，大多数融合模型反而降低了整体鲁棒性。这一发现挑战了以往研究的一致认识。2) 模型架构如何影响鲁棒性？本文采用了一种新的范式来对模型进行分类，并引入了信息熵的概念，这意外地揭示了模型架构与鲁棒性之间的关系，即融合模态的信息熵越大，鲁棒性越强。最后，本文为增强感知模型的鲁棒性提供了一些见解。相关内容将在本文第四章进行详细介绍。

### 1.4.4 信号注入攻击检测和防护关键技术研究

本文提出面向自动驾驶中激光雷达感知系统的安全防护方法，能够针对信号注入攻击下的数据受损实现主动式攻击检测和被动式鲁棒性增强。本文一共提出了三种数据受损检测方法：(1) 基于点云表面曲率的虚假目标检测，由于基于激光的点云注入攻击能力的固有限制，虚假目标和真实目标之间在表面曲率上存在显著差异，实现了对虚假目标注入攻击的检测；(2) 基于一致性分析和时序预测的受损模态检测，利用了未受损时不同传感器之间的语义一致性以及传感器数据的时间连续性，实现了对受损模态的检测；(3) 基于视觉语言模型的受损类型检测分析，利用了视觉语言预训练模型的强大基础能力，通过小样本有监督微调和检索增强生成的方式，实现了对第四章中 11 类数据

受损方式的检测分析。在实现攻击检测的基础上，为了切实提升自动驾驶中激光雷达感知的鲁棒性，本文基于虚拟点技术提出了具有模态独立、平行融合、数据融合三个特征的多传感器融合架构 SIA-Defense，实现了对信号注入攻击下的数据受损的鲁棒性提升。平均鲁棒性比现有 SOTA 模型提升了 5%。

## 1.5 创新点

本文的创新点总结如下：

- **基于激光的点云注入攻击：**针对现有基于激光的攻击物理可实现性不足的问题，提出了一种使用红外激光实现高可控欺骗点云注入的攻击方法，能够在物理世界以黑盒、非抵近、隐蔽的方式实现隐藏和创建指定物体的攻击效果，促进了对激光攻击威胁的正确认识。
- **基于电磁的脆弱性分析：**针对现有信号注入攻击形式单一的问题，提出了利用电磁信号进行激光雷达脆弱性分析的方法，挖掘了新的攻击入口和攻击原理，实现了包括点云干扰、点云抹除、点云注入和雷达宕机在内的 4 类攻击效果，拓宽了激光雷达脆弱性分析的形式。
- **融合感知鲁棒性测评基准：**针对现有感知模型在信号注入攻击下的鲁棒性问题，提出了首个基于信号注入攻击的多传感器融合鲁棒性测评基准，回答了“融合是否增强鲁棒性”和“模型架构如何影响鲁棒性”2 个研究问题，为多传感器感知鲁棒性测评提供了数据基础和方法指导。
- **攻击检测和防护关键技术：**针对信号注入攻击引起的传感器数据受损问题，提出了对虚假目标、受损模态和攻击类型的检测判定方法，以及基于多传感器融合的鲁棒性提升方法，实现了对信号注入攻击的有效检测和防护。

## 1.6 论文组织架构

本文针对自动驾驶中激光雷达感知的脆弱性分析与安全防护进行了深入的研究和探索。本文共包含六章，后续章节的组织结构如下：

第二章研究了基于激光注入的传感器脆弱性挖掘与利用。通过激光攻击的形式，实现了高可控的点云注入攻击，能以黑盒、非抵近、隐蔽的方式实现隐藏和创建指定物体的攻击效果，证明了激光雷达信号鉴权脆弱性的存在以及该脆弱性被利用后的严重后果。

第三章研究了基于电磁干扰的传感器脆弱性挖掘与利用。通过电磁干扰的形式，实现了包括点云干扰、点云抹除、点云注入和雷达宕机在内的 4 类攻击效果，证明了激光雷达接收模块的模拟电路、主板上的监测传感器、光束偏转模块中的光学编码器均存在电磁脆弱性。

第四章研究了信号注入攻击下的激光雷达感知系统鲁棒性测评基准。在第二、三章的基础上，设计了基于信号注入攻击的受损数据集，共包含 11 种信号注入攻击下的数据受损形式。基于该受损数据集，对现有自动驾驶中广泛使用的多传感器融合模型进行了鲁棒性测评，并且重点回答了两个关键研究问题：(1) 融合是否增强鲁棒性；(2) 模型架构如何影响鲁棒性。最后基于测评结果为增强感知模型的鲁棒性提供了一些建议。

第五章研究了基于多传感器融合和攻击检测的安全防护方法。针对第四章总结的 11 种信号注入攻击，首先为了满足特定场景下不同的检测需求，提出了 3 个维度的攻击检测方法：基于点云表面曲率的虚假目标检测、基于一致性分析和时序预测的受损模态检测、基于视觉语言模型的受损类型检测。其次，在实现攻击检测的基础上，提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构 SIA-Defense，实现了鲁棒性的提升。

第六章总结了本文的研究成果，并展望了未来的研究方向。



## 2 基于激光注入攻击的激光雷达感知脆弱性分析

自动驾驶汽车和机器人越来越多地利用基于激光雷达的 3D 目标检测模型来探测环境中的障碍物。正确的检测和分类对于确保安全驾驶非常重要。尽管以前的工作已经证明了操纵点云欺骗 3D 目标检测模型的可行性，但这些尝试都是以数字域攻击的方式进行的。本章提出了一种通过在物理世界使用红外激光注入欺骗点云来对 3D 目标检测模型直接进行欺骗的方法——PLA-LiDAR。为了提高点云注入能力，本章开发了一种激光收发器，它能够注入多达 4200 个欺骗点；为了生成能够被物理注入到目标激光雷达中的对抗性点云，本章提出了一种同时考虑激光雷达的工作原理和物理攻击能力的对抗性点云优化方法；为了将上述生成的对抗性点云精确注入激光雷达，本章提出了一种“空间坐标-时间坐标”映射的控制信号设计方法和精确到纳秒级别的信号同步方法。基于上述方法，PLA-LiDAR 共实现了四种攻击效果，能分别以黑盒和白盒的方式实现隐藏攻击和创建攻击。通过物理世界的测量，本章从注入点数、位置控制能力、形状控制能力三方面量化了 PLA-LiDAR 的攻击能力，能够为其他仿真研究提供物理可实现的参考。通过在商用激光雷达和基于点云的目标检测模型上的数字域和物理域评估，本章验证了对抗点云优化算法的有效性和 PLA-LiDAR 的物理攻击可行性。最后本章通过移动的实车上进行的实验进一步证明了 PLA-LiDAR 的物理可行性。实验相关的视频演示可在对应网站<sup>[63]</sup>上查看。

### 2.1 本章引言

随着自动驾驶的普及，激光雷达（LiDAR）被越来越广泛地运用到先进辅助驾驶（ADAS）<sup>[64-66]</sup>和车路协同基础设施系统（CVIS）<sup>[67-70]</sup>中。根据 Yole Développement<sup>[71]</sup>，到 2025 年，汽车和工业应用的激光雷达市场规模预计将达到 38 亿美元。激光雷达传感器能够生成周围环境的精确 3D 点云，结合后续的感知算法，可以实现高精度的障碍物检测和分类，激光雷达的这种能力为自动驾驶汽车做出关乎安全的关键驾驶决策提供基础。

许多先前的研究证明了基于 LiDAR 的 3D 目标检测模型的脆弱性，但这些工作<sup>[24,31,35,72]</sup>主

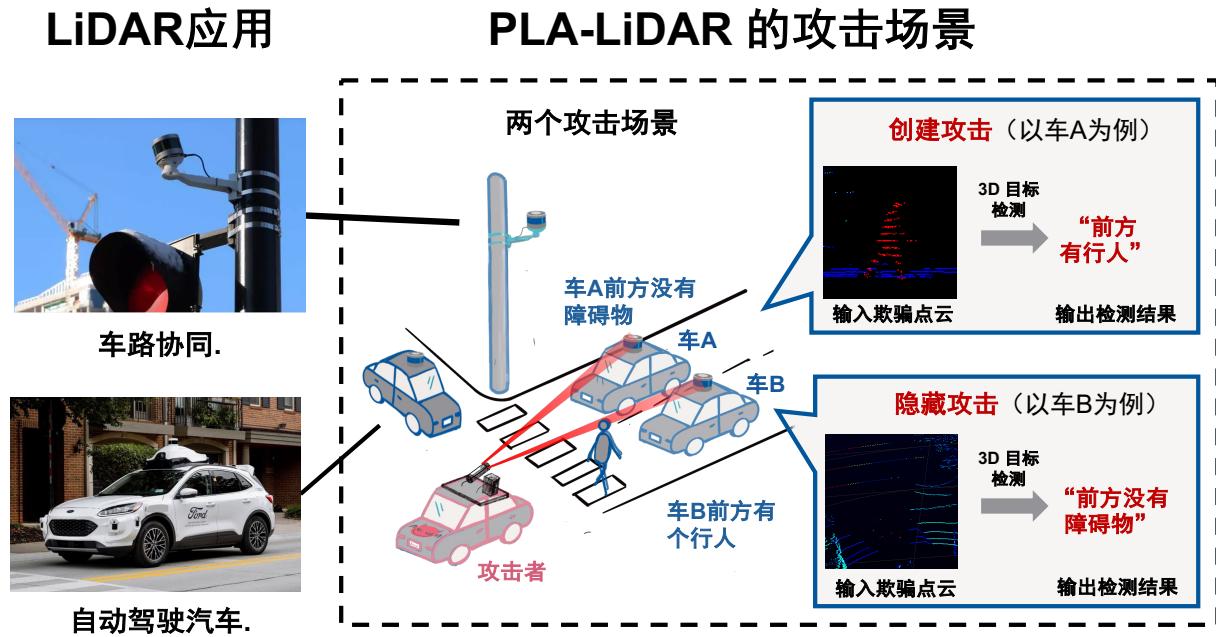


图 2.1 PLA-LiDAR 攻击场景示意图：通过将恶意激光信号注入到自动驾驶汽车的车载激光雷达中，攻击者可以欺骗 3D 目标检测模型的决策。

要是通过对点云进行数字域操控来实现的，并没有在物理世界中充分评估攻击的可行性。还有一些工作通过放置 3D 实体（比如无人机<sup>[73]</sup>或 3D 打印的物体<sup>[74]</sup>）来在物理世界实现点云对抗攻击，但这些攻击的部署对人眼十分明显，缺少隐蔽性，且这些工作的攻击效果单一，主要聚焦于实现“隐藏指定目标”的攻击。

本章探究以下研究问题：“能否通过在物理世界使用红外激光注入欺骗点云来对 3D 目标检测模型直接进行欺骗？”具体而言，本文考虑以下攻击场景：如图 2.1 所示，攻击者可能会向一辆正在等待交通信号灯的自动驾驶汽车的车载激光雷达发射激光，从而实现两种攻击效果：

- **创建攻击**: 使受害自动驾驶车辆感知到一个不存在的物体。
- **隐藏攻击**: 使受害自动驾驶车辆无法感知到前方的物体。

尽管早期研究已表明 3D 目标检测系统易受数字对抗性点云的攻击，但尚无研究调查将生成的对抗性点云攻击在物理世界实现的可能性。本文首次专注于在物理域实现针对基于激光雷达的 3D 目标检测系统的攻击。为了实现物理世界的点云对抗攻击，存在如下挑战：1) 点云注入能力的提升：先前工作中报道的点云注入能力最多为 200 个点<sup>[31]</sup>，这不足以实现先前工作中的数字域攻击，因此点云注入能力的提升也是需要解决的问题。2) 可注入对抗点云的生成：由于激光雷达对三维空间的扫描是离散的，即只

有一些特定方向上能够生成点云，因此对抗点云在生成过程中的搜索空间是有限的，并非整个三维空间的任意点，如何建模搜索空间以及如何优化出有效的对抗点云是一个挑战。3) 数字域设计到物理域的实现：考虑到激光雷达的超高速扫描特性，真实物理世界的环境噪声、信号处理误差等干扰存在，将特定形状的欺骗点云注入到特定位置（这通常是欺骗点云实现对抗效果的基本要求<sup>[24,31,35,72]</sup>）是一件十分具有挑战性的任务。

为了克服上述挑战，本文设计了一种针对基于激光雷达的 3D 目标检测的物理激光攻击方法——PLA-LiDAR (**P**hysical **L**aser **A**ttack against **L**iDAR-based 3D object detection system)。为了提高点云注入能力，本文开发了一种激光收发器，它能够注入多达 4200 个点，这是先前最新工作<sup>[31]</sup>所实现数量的 20 倍，也是实现物理攻击的关键因素。为了生成能够被物理注入到目标激光雷达中的对抗性点云，本文提出了一种新的对抗性点云优化方法。在优化过程中，该方法会考虑激光雷达的工作原理、攻击设备的能力以及注入点的距离误差。为了精确生成所需形状的注入点云，本文提出了一种控制信号设计方法，该方法将点云的形状转换为激光控制信号。为了精确控制注入点的距离，本文提出了一种新的同步方法，以使攻击信号与目标激光雷达的扫描序列对齐。基于上述方法，PLA-LiDAR 共实现了四种攻击效果，能分别以黑盒和白盒的方式实现隐藏攻击和创建攻击（详见第 2.3.1 章）：

为了评估 PLA-LiDAR 的攻击效果，本文在两款商用车载激光雷达（VLP-16<sup>[75]</sup>和 RS-16<sup>[76]</sup>）上验证了攻击的可行性，利用物理世界直接采集到的点云，在 3 款基于点云的目标检测模型和 4 款传感器融合模型上进行了大量实验测试。最后，本文在移动的实车上进行了可行性实验。

## 2.2 背景知识

市场上常见的面向自动驾驶应用的激光雷达主要包括两个类型：(1) 机械式（旋转式）激光雷达，它利用旋转组件使传感器旋转，并在旋转过程中发射脉冲激光以实现对周围环境的扫描；(2) 固态激光雷达，其没有旋转的机械部件，而是使微机电系统 (MEMS) 技术<sup>[77]</sup>或光学相控阵 (Optical Phased Array) 技术<sup>[78]</sup>进行扫描。在这两种类型的激光雷达中，机械式激光雷达占据全球激光雷达市场 95% 以上的份额<sup>[79]</sup>，并被许多大规模商业自动驾驶项目所采用，例如 Waymo One<sup>[80]</sup>、百度无人出租车<sup>[81]</sup>等。鉴于其

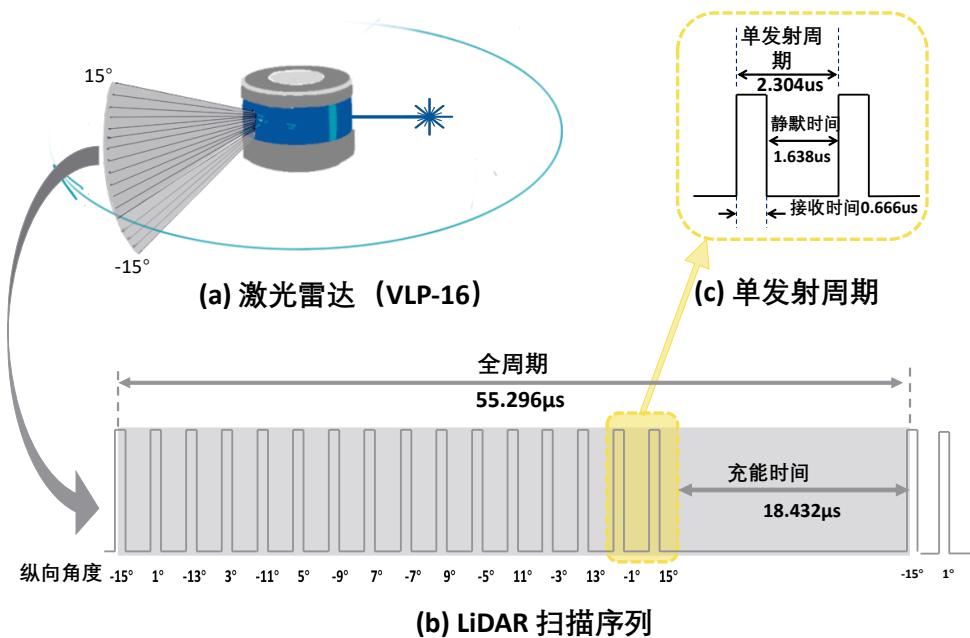


图 2.2 激光雷达的扫描序列

巨大的市场份额和广泛的应用，本章着重对机械式激光雷达展开研究。

如图2.2所示，机械旋转式激光雷达拥有一组由多个红外激光器和探测器组成的阵列，阵列上的激光器和探测器以特定的角度垂直分布，覆盖一定的纵向范围。激光收发阵列按照特定的时序发射并接收激光脉冲，每一次发射和接收能实现一次特定方向（水平角  $\theta$ , 垂直角  $\phi$ ）的精准测距（距离  $r$ ），进而生成一个激光雷达点，记为  $Point = [r, \theta, \phi]$ ；同时，激光收发阵列在光束偏转模块的驱动下不断改变激光发射方向，实现对周围环境的扫描测距，最终生成点云（Point Cloud, PC），记为  $PC = [\mathbf{R}, \Theta, \Phi]$ 。在 PLA-LiDAR 攻击的设计过程中，需要关注机械式激光雷达的四个参数，分别是 1) 扫描序列（Scanning Sequence, SS），2) 激光线束垂直角分布（Laser Vertical Distribution, LVD），3) 水平角分辨率（Horizontal Angular Resolution,  $R_H$ ），4) 激光波长  $\lambda_{lidar}$ 。如式2-1所示，本文用这四个参数表征一个激光雷达：

$$LiDAR = [SS, LVD, R_H, \lambda_{lidar}], \quad (2-1)$$

**扫描序列：**SS 描述激光雷达发射和接收激光脉冲的时间顺序，激光雷达以 SS 为周期运行，每个激光雷达都有其独特的 SS。如图2.2所示，一个 SS 的时间长度是一个“完整周期（full cycle,  $T_{fc}$ ）”，在此期间，收发阵列中的所有激光器按照特定顺序发射和充能一次，激光器之间发射的最短时间间隔称为“单发射周期（single firing cycle,  $T_{sfc}$ ）”。

如图2.2 (c) 所示，每次发射后，激光雷达会在接收时间内监听回波，在接收时间内接收到的特定脉冲（根据激光雷达的回波模式，可能是最强的脉冲或最后一个脉冲）被视为有效回波，接收时间结束后，激光雷达会静默一段空闲时间，然后再发射下一个脉冲。**SS** 在本工作中用于设计激光攻击信号。

**激光线束竖直角分布：****LVD** 描述激光雷达的垂直视场和垂直分辨率。这是一个出厂设定的参数，可以从用户手册中获取。比如 16 线激光雷达 VLP-16 的 16 个激光器在竖直方向上以  $2^\circ$  的间隔分布在  $-15^\circ$  到  $15^\circ$  之间。所以在生成对抗点云时，要严格按照目标激光雷达的 **LVD** 来设计，否则对抗点云无法按照预设计的形状被注入。

**水平角分辨率：** $R_H$  表示激光雷达点在水平方向上的最小角度差。激光雷达旋转得越快， $R_H$  就越大。激光雷达的旋转速度以每分钟转数（Rotation Per Minute, RPM）表示，并且可以由用户进行配置。攻击者可以通过光电传感器接收激光雷达的光来获得目标激光雷达的转速，从而算出  $R_H$ 。

**激光雷达波长：**目前最流行的激光雷达系统的激光波长  $\lambda_{lidar}$  通常为 905 纳米或 1550 纳米。一般来说，激光雷达对自身的激光波长  $\lambda_{lidar}$  最为敏感，并且会过滤掉其他波长的光，所以选取和激光雷达工作波长  $\lambda_{lidar}$  相近的光有助于攻击的实现。

## 2.3 威胁模型

威胁模型包括攻击目标和攻击者能力。

### 2.3.1 攻击目标

PLA-LiDAR 的攻击目标是向机械式激光雷达注入恶意点，进而欺骗其后续的 3D 目标检测算法，使其产生错误输出。具体而言，本文考虑两种攻击目标：(1) “隐藏”：使受害的自动驾驶汽车无法感知到一个实际存在的物体；(2) “创建”：使受害的自动驾驶汽车感知到一个不存在的物体。本文进一步考虑如图 2.3 所示的 4 种攻击类型：

1. 直接隐藏攻击 (**Direct-Hide**)：是一种黑盒攻击，攻击者利用激光雷达的回波筛选机制，直接抹除目标点云，进而在不需要知道模型的任何信息实现隐藏指定目标的攻击效果。由于是通过直接抹除点云实现的隐藏攻击，所以本文将其命名为“**Direct-Hide**”。

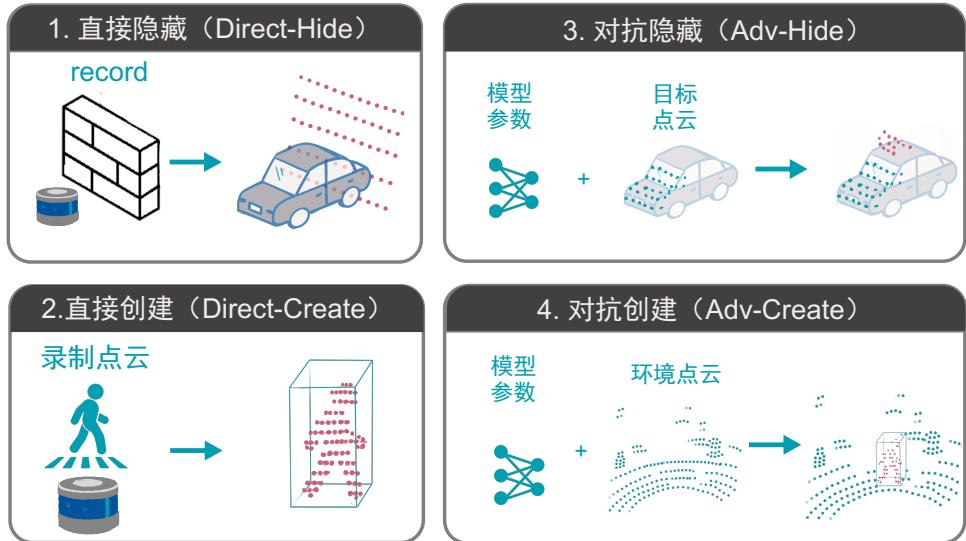


图 2.3 四种攻击类型示意图.

2. 直接创建攻击 (**Direct-Create**)：是一种黑盒攻击，攻击者通过将预录制好的点云（比如行人）重新注入受害激光雷达，从而实现创建指定目标的攻击效果。由于是通过直接注入和真实物体相近的点云实现的创建攻击，所以本文将其命名为“Direct-Create”。
3. 对抗隐藏攻击 (**Adv-Hide**)：攻击者首先需要知道目标检测模型和被隐藏物体的信息，然后通过优化的方式生成对抗点云，再将对抗点云注入到激光雷达中实现隐藏攻击。由于拟注入的欺骗点云是通过对抗攻击的方式优化生成的，所以本文将其称为“Adversarial-based Hide”，简写为“Adv-Hide”。
4. 对抗创建攻击 (**Adv-Create**)：攻击者首先需要知道目标检测模型和周围环境的信息，基于创建指定物体的优化目标，实现创建攻击，因此命名为“Adversarial-based Create”，简写为“Adv-Create”。

### 2.3.2 攻击者能力

为了实现上述攻击目标，攻击者需要具备以下能力：

**激光雷达参数获悉：**攻击者能够获取并分析与受害自动驾驶汽车中所使用型号相同的激光雷达。通过直接分析该激光雷达或阅读其用户手册，攻击者能够了解激光雷达的参数，包括 **SS**, **LVD**,  $\lambda_{lidar}$  等。此外，攻击者可以使用光电传感器和示波器测量受害激光雷达的旋转速度进而计算  $R_H$ 。

**模型参数获悉 (可选)：**对于 Adv-Hide 和 Adv-Create 攻击，攻击者需预先了解受

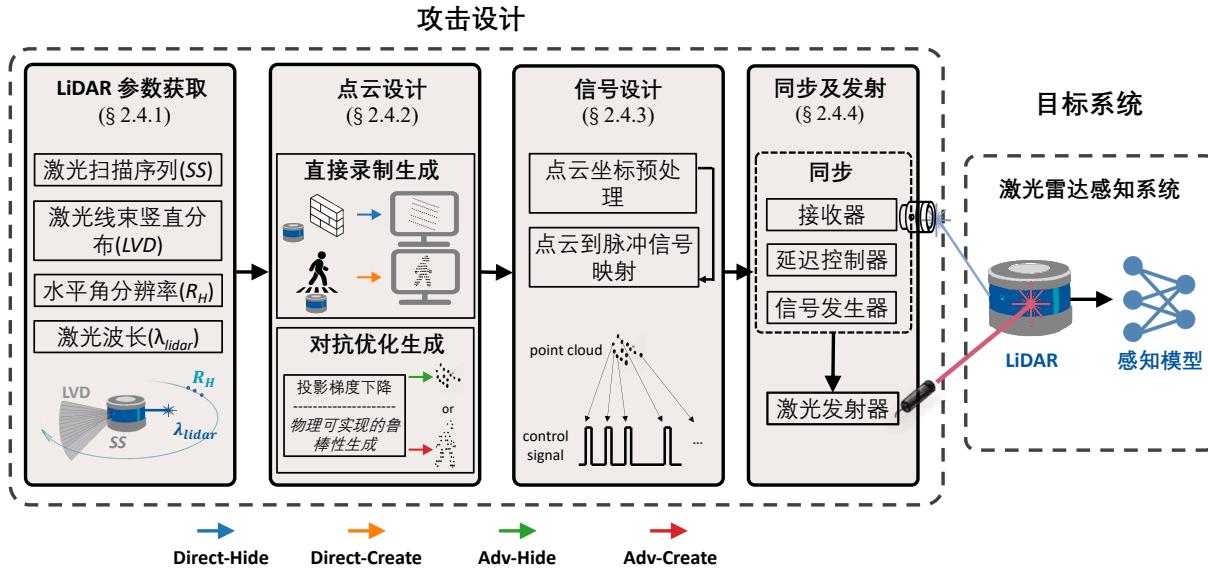


图 2.4 PLA-LiDAR 的攻击设计流程

害自动驾驶汽车中使用的目标检测算法，包括但不限于其架构、参数、输出等。对于 Direct-Hide 和 Direct-Create 攻击，攻击者不需要目标检测算法的任何先验信息。

**物理攻击实现：**攻击者能够通过使用由光电传感器、任意波形发生器和激光发射器等商用设备组成的攻击装置，向目标自动驾驶汽车或车路协同系统中的激光雷达发射激光。对于移动的目标（如行驶中的自动驾驶汽车），攻击者可以驾驶一辆与目标自动驾驶汽车速度相近的车辆，并利用激光测距技术<sup>[82]</sup> 测量激光发射器与受害激光雷达之间的距离。

## 2.4 攻击设计

为了在物理世界实现上述攻击目标，需要解决以下两个关键研究问题：

- **研究问题 1：**如何生成可被注入到激光雷达中的欺骗点云？
- **研究问题 2：**如何在物理世界中将欺骗点云注入到激光雷达中？

为了解决上述研究问题，本文设计了 PLA - LiDAR 攻击方法，如图 2.4 所示，该方法包含四个关键步骤：步骤一是 LiDAR 参数获取，通过测量被攻击激光雷达和查阅相关手册的方式，获取攻击相关参数；步骤二是点云设计，通过录制或对抗优化，生成理论上能够注入激光雷达的欺骗点云；步骤三是激光控制信号设计，将欺骗点云转化为激

光控制信号；步骤四是同步和发射，将被攻击激光雷达的扫描顺序与激光攻击信号进行时序上的同步，然后使用特定波段的激光发射器发射激光，从而实现攻击。

#### 2.4.1 激光雷达参数获取

在设计攻击前，本文首先需要获取目标激光雷达的关键参数。如第2.2章所述，机械式激光雷达有四个关键参数，即扫描序列（**SS**），激光线束竖直角分布（**LVD**），水平角分辨率（ $R_H$ ）和激光波长  $\lambda_{lidar}$ 。在这些参数中，**LVD** 和  $R_H$  用于在对抗优化点云时规范点云的搜索空间，使得对抗生成的点云可被注入。**SS** 用于控制信号设计和同步，以确保欺骗点云被正确转化为激光信号并按照期望的形状被注入。 $\lambda_{lidar}$  用于激光器的选型，以确保发射的激光能被受害激光雷达接收。

对于这四个参数，**LVD** 和  $\lambda_{lidar}$  可以从激光雷达的官方文档中获取；而 **SS** 能从官方文档中获取粗略的信息，但需要通过测量来进行精度的矫正； $R_H$  则需要通过测量激光雷达的转速来计算获取。下面以 VLP-16 激光雷达为例，介绍 **SS** 和  $R_H$  的测量过程：

**扫描序列矫正：** VLP-16 的 **SS** 包含“完整周期” ( $T_{fc}$ ) 和“单发射周期” ( $T_{sfc}$ ) 两个参数，用户手册<sup>[75]</sup>给的参数  $\mathbf{SS}_0$  为 ( $T_{fc0} = 55.296\mu s$ ,  $T_{sfc0} = 2.304\mu s$ )。但是直接利用用户手册给定的  $\mathbf{SS}_0$  无法将欺骗点云按照预先设计的形状注入激光雷达，如图2.5所示，本文拟注入的欺骗点云是一堵点云“墙”，其中每一个点距离坐标原点相同距离，但是当本文利用从用户手册中获取的这组参数进行信号设计和点云注入时，会发现注入的点云和预先设计的点云出现明显失真，这是因为用户手册给的 **SS** 参数和实际的参数存在一定误差。

为了使注入的点云尽可能和预先设计的点云一致，本文提出了一个扫描序列矫正方法。经过经验性的实验，本文发现，仅需对  $T_{fc}$  进行矫正即可。本文将利用  $\mathbf{SS}_0$  注入的点云墙的时序上第一个点记为点  $A = [r_A, \theta_A, \phi_A]$ ，与点 A 相同竖直角的时序上最后一个点记为点  $B = [r_B, \theta_B, \phi_B]$ ，点 A 和点 B 在时序上相差  $N_{fc}$  个全周期，矫正后的  $T_{fc}$  可以通过式2-2获得：

$$T_{fc} = \frac{r_A - r_B}{c} * \frac{1}{N_{fc}} + T_{fc0}. \quad (2-2)$$

利用矫正后的 **SS** 设计出的信号注入的点云墙如图2.5c所示，可以看到和拟注入的点云墙基本一致。

**水平角分辨率测量：** 水平角分辨率 ( $R_H$ ) 和激光雷达的转速有关，激光雷达的转速

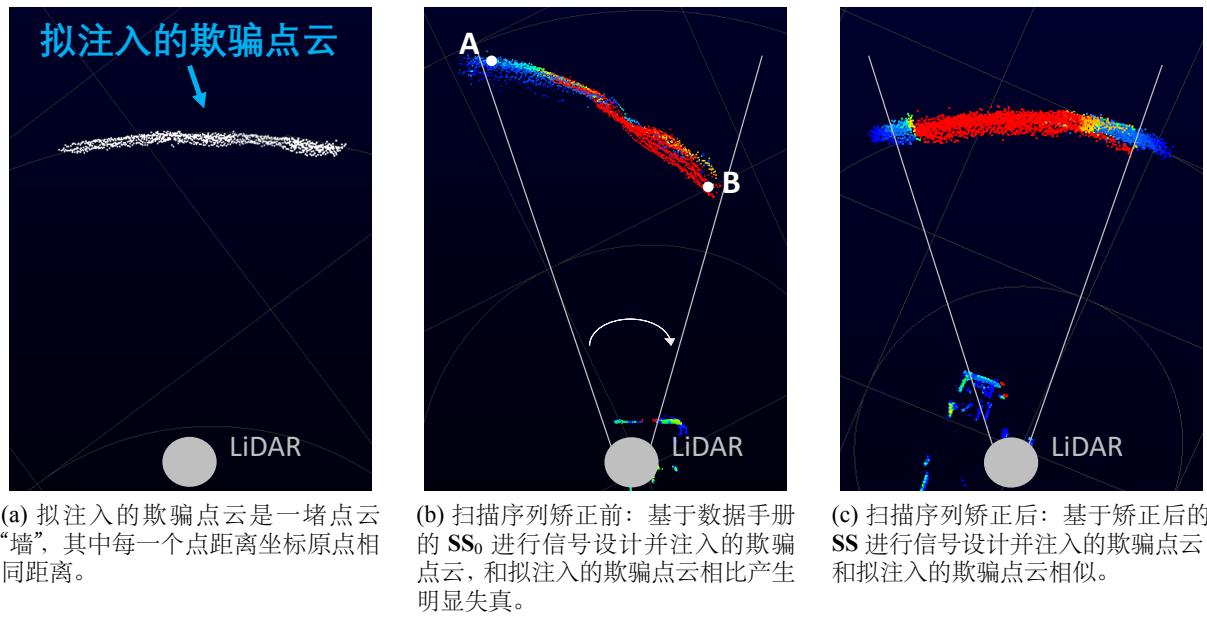


图 2.5 扫描序列矫正

用参数 RPM (Rotation Per Minute) 来衡量。RPM 表示激光雷达每分钟转的圈数，可以由用户自由设置。RPM 通常会被设置为 600，代表着 10Hz 的刷新率。本文可以使用一个光电传感器接收激光雷达的信号，进而测算出目标激光雷达的 RPM。当知道 RPM 和  $T_{fc}$  后，本文可以根据式2-3算出  $R_H$ :

$$R_H = 360^\circ * \frac{T_{fc} * RPM}{60}. \quad (2-3)$$

## 2.4.2 点云设计

为了设计出可被注入到激光雷达中的欺骗点云，本文考虑两种点云生成方法：1) 基于录制的点云生成方法；(2) 基于对抗优化的点云生成方法。基于录制的方法不需要关于目标检测器的任何先验信息，但需要一个与受害激光雷达型号相同的替代激光雷达。基于对抗优化的方法更为精细，对点数的要求较低，但需要以白盒方式访问目标检测器。在实际应用中，攻击者可以根据攻击场景选择合适的点云生成方法。

### 2.4.2.1 基于录制的点云生成

基于录制的点云生成方法用于直接隐藏 (Direct-Hide) 和直接创建 (Direct-Create) 攻击。这种方法的优点在于，生成的点云是从激光雷达采集而来的，因此自然符合激光雷达的点云分布，所以是“可注入”的。为实现这一方法，本文首先获取一个与受害

激光雷达型号相同的激光雷达，本文将其称为替代激光雷达。然后，根据预期的攻击目标（类别）和攻击距离，本文使用替代激光雷达录制目标类别的物体的点云，例如，在 Direct-Hide 攻击中本文录制一堵墙的点云，在 Direct-Create 中本文录制一个行人的点云。需要注意的是，由于 Direct-Hide 的攻击信号比较简单，也可以略过点云设计的环节，直接基于 SS 进行信号设计。

#### 2.4.2.2 基于对抗优化的点云生成

对于对抗隐藏（Adv-Hide）和对抗创建攻击（Adv-Create），本文通过对抗式机器学习生成欺骗点云。与基于录制的方法相比，基于对抗的方法利用了目标检测算法的漏洞，并且有可能用更少的点隐藏或创建出指定目标。

**问题形式化描述：**为了实现这一目标，本文首先引入一个在生成过程中需要考虑的物理约束条件。以往的数字对抗点云生成方法在本攻击中并不适用，因为它们没有考虑激光雷达点云的离散特性，即在时序上激光雷达的一个欺骗点只能在一个“单发射周期”内注入，并且一个“单发射周期”最多只能注入一个点。本文将激光雷达点云的离散特性阐述为如下物理限制：每个生成的点仅出现在激光雷达的一条激光射线上，并且每条激光射线最多只有一个点。本文将这一物理限制考虑进点云生成过程中，并表示成如下对抗优化问题：

$$\begin{aligned}
 & \min_{\text{PC}'} \mathcal{L}(\text{PC}') \\
 \text{s.t. } & (r'_i, \theta'_i, \phi'_i) \in \text{Loc}^{\text{exp}}, \forall i \in \{1, \dots, n\}, \\
 & |\theta'_i - \theta'_j| + |\phi'_i - \phi'_j| \neq 0, \forall i, j \in \{1, \dots, n\}, \\
 & \theta'_i \in \text{LVD}, \\
 & |\phi'_i - \phi'_j| = N * R_H,
 \end{aligned} \tag{2-4}$$

其中  $\text{PC}' = \{(R'_i, \Theta'_i, \Phi'_i) | i \in [1, n]\}$  是对抗点云； $r'_i, \theta'_i$  和  $\phi'_i$  分别是对抗点的距离，竖直角和水平角； $\text{Loc}^{\text{exp}} = \{x_a, y_a, z_a, w_a, l_a, h_a, yaw_a\}$  规定了拟注入对抗点云的区域空间的中心坐标和长宽高；任意的两个点  $(r'_i, \theta'_i, \phi'_i)$  和  $(r'_j, \theta'_j, \phi'_j)$  不能在同一条激光射线上； $\theta'_i$  和  $\phi'_i$  分别要满足竖直角分布（LVD）和水平角分辨率（ $R_H$ ）的物理限制。

**损失函数设计：**接着，本文为 Adv-Hide 和 Adv-Create 分别设计损失函数。

对于 Adv-Hide，本文的目标是向目标物体附近注入对抗点云，使其无法被检测到。

为实现这一目标，本文通过梯度优化的方式在目标物体的上方生成对抗点云，进而降低目标物体的识别框的置信度，最终使得其被非极大值抑制（NMS）算法过滤掉。需要注意的是，在3D目标检测任务中，一个物体会有多个候选识别框，然后通过NMS筛选出置信度最高的候选框作为检测结果。为了使得目标物体在输出侧最终没有对应的识别框，本文需要降低所有候选识别框的置信度。因此，Adv-Hide的损失函数设计如下：

$$\mathcal{L}_h = \sum_{bbox, s \in B} -\text{IoU}(bbox^t, bbox) \log(1 - s), \quad (2-5)$$

其中， $B = \{(x_i, y_i, z_i, w_i, h_i, l_i, yaw_i) | i \in [1, n]\}$  是所有候选边界框的集合， $bbox^{gt}$  是目标物体真实的识别框， $bbox$  和  $s$  是候选边界框及其置信度。

对于Adv-Create，本文的目标是通过向特定区域（例如，在受害激光雷达前方5米处）注入对抗点，诱导出一个目标物体。为实现这一目标，本文首先在目标区域随机注入少量对抗点云，然后通过微调对抗点云的位置提升候选边界框的置信度。与Adv-Hide攻击关注所有候选框不同，Adv-Create选择与预期区域IoU最大的前10个边界框作为候选框。因此，Adv-Create的损失函数设计如下：

$$\mathcal{L}_c = \sum_{b, s \in B} -\text{IoU}(b^e, b) \log(s), \quad (2-6)$$

其中  $b^e = \{x_e, y_e, z_e, w_e, h_e, l_e, yaw_e\}$  是拟创建物体的目标区域。

**对抗鲁棒性增强：**在实际的基于激光的点云注入过程中，由于设备采样率的限制和物理噪声，真实注入的点云和期望的点云之间往往会有无法规避的误差。在第2.5.2章中，本文将这种误差量化为形状控制误差和位置控制误差。为了使得攻击更鲁棒，本文将误差引入点云的设计过程中。具体来说，对于每个点，本文给予一个随机扰动  $\delta \sim U(-d, d)$ ，对于整体点云本文给予一个随机扰动  $\Delta \sim U(-D, D)$ ，其中“~”表示“服从于”，U表示均匀分布（Uniform Distribution）。d和D可分别根据形状控制和位置控制的能力而定。鲁棒性增强后的优化问题可以被形式化描述为：

$$\min_{PC'} (\mathcal{L}(PC') + \mathcal{L}(E(P', \delta, \Delta))), \quad (2-7)$$

其中， $E$  表示对点云  $PC$  添加随机扰动。

**对抗点云生成过程：**总的来说，Adv-Hide和Adv-Create的点云生成过程如下：

- 步骤1：根据攻击者期望隐藏或创建目标物体的位置，计算拟注入对抗点的球坐标范围；

- 步骤 2：在上述球坐标范围内按照式2-4随机添加给定数量的对抗点；
- 步骤 3：根据 Adv-Hide 或 Adv-Create 的损失函数，利用式2-7计算梯度；
- 步骤 4：根据梯度更新对抗点云  $PC$  的  $R$ ；
- 步骤 5：重复步骤 3 和步骤 4，直到损失收敛或迭代结束。

### 2.4.3 信号设计

为将生成的点云注入目标激光雷达，本文设计了一套能够根据点云生成激光控制信号的算法。将欺骗点云转换为相应的激光控制信号的具体流程如算法2.1所示，本文首先对点云信息  $PC$  和激光雷达信息  $LiDAR$  进行预处理，然后进行信号的生成。

**数据预处理：**数据预处理的目的是利用点云信息  $PC$  和激光雷达信息  $LiDAR$  生成“时间坐标”，方便信号生成。在第2.2章中，本文对点云信息  $PC$  和激光雷达信息  $LiDAR$  分别进行了建模，其中  $PC = [\mathbf{R}, \Theta, \Phi]$ ,  $LiDAR = [\mathbf{SS}, \mathbf{LVD}, R_H, \lambda_{lidar}]$ 。在数据预处理中，本文利用  $PC$  和  $LiDAR$  生成时间坐标，一个欺骗点的时间坐标记为  $(fullcycle\_id, singlecycle\_id, tof)$ ，其中  $fullcycle\_id$  表示该点所处的完整周期， $singlecycle\_id$  表示该点在上述完整周期内所处的单发射周期， $tof$  表示生成该点激光信号的理论飞行时间。这一组时间坐标可以用于计算和该欺骗点对应的激光脉冲的上升沿的时刻。

为了计算  $fullcycle\_id$ ，本文将水平角最小的那个点（该点的水平角为  $\phi_0$ ）的  $fullcycle\_id$  设定为 0，在此基础上，其他点（假设水平角为  $\phi$ ）的  $fullcycle\_id$  可以通过式2-8获得：

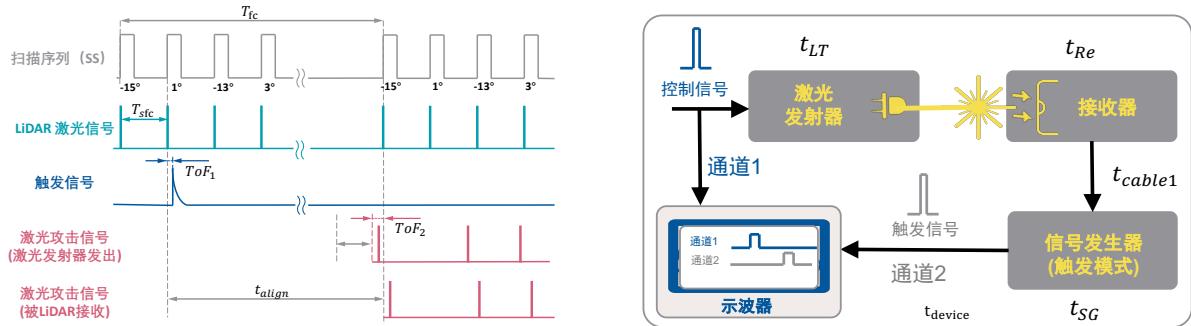
$$fullcycle\_id = \frac{\phi - \phi_0}{R_H}. \quad (2-8)$$

为了计算  $singlecycle\_id$ ，本文基于  $SS$  建立不同竖直角和一个完整周期内发射次序之间的映射（记作  $Angle2ID$ ），用于直接通过竖直角来获得  $singlecycle\_id$ 。

$tof$  可直接通过点的距离和光速来计算得到：

$$tof = 2 * \frac{r}{c}. \quad (2-9)$$

**信号生成：**信号生成的目的是利用时间坐标生成激光控制信号。激光控制信号由多个脉冲信号组成，每个脉冲信号代表一个欺骗点，且每个上升沿的出现时刻 (**Timestamp**)



(a) 信号同步示意图：“信号同步”的让激光攻击信号在指定的时刻被激光雷达接收。信号同步过程中为了应对“光速灾难”问题，需要综合考虑光的飞行时间和设备内在延迟。该图描述了攻击过程中的几个关键信号和激光雷达扫描序列的时序关系。

(b) 设备内在延迟测量方法示意图：利用示波器能够测量激光发射器、接收器、信号发生器以及连接线的总延迟。

图 2.6 (信号同步示意图及设备内在延迟测量方法示意图)

决定了该欺骗点的空间坐标。基于时间坐标，**Timestamp** 可以被精确计算，如式2-10所示：

$$\text{Timestamp} = \text{fullcycle\_ID} * T_{fc} + \text{singlecycle\_ID} * T_{sc} + \text{ToF}. \quad (2-10)$$

随后，本文首先生成理想的连续激光控制信号  $\text{Signal}_{ideal}$ ，即每个脉冲信号的起始时间为 **Timestamp**，脉宽为 10 纳秒（与激光雷达工作信号基本一致），上升沿和下降沿时间忽略不计。但是在实际系统中，需要遵循信号发生器采样率的限制，因此本文根据信号发生器的采样率对  $\text{Signal}_{ideal}$  进行采样，使其转变为能被信号发生器直接读取并发送的离散信号  $\text{Signal}_{discrete}$ 。

#### 2.4.4 信号同步

通过信号设计，点云已经被转化为了激光控制信号，为了真正实现点云注入，激光发射的时刻尤为重要。由于激光雷达利用光的飞行时间来测算距离，任何微小的时间误差乘上光速都可能会引起巨大的距离误差，本文称之为 PLA-LiDAR 攻击过程中面临的“光速灾难”问题。为了解决“光速灾难”，PLA-LiDAR 通过光电传感器接收激光雷达的信号进而推算出激光雷达的工作状态，然后通过设定纳秒级精度的延迟来精准控制激光发射的时刻，最终将距离误差控制在厘米级。

“信号同步”这一流程要做的就是参照激光雷达的扫描序列，让激光攻击信号在指定的时刻被激光雷达接收。比如，本文在设计激光攻击信号时就期望激光攻击信号的起始时刻和扫描序列中某个完整周期的起始时刻对齐。本文以针对 VLP-16 激光雷达的攻

---

**算法 2.1 控制信号设计**


---

```

1: 输入:
2: 点云坐标: X, Y, Z
3: 光速:  $c$ 
4: 扫描序列: SS =  $[T_{fc}, T_{sfc}]$ 
5: 竖直角分布: LVD
6: 竖直角到激光 ID 映射: Angle2ID
7: 输出:
8: 理想的模拟控制信号: Signalideal
9: 离散的数字控制信号: Signaldiscrete
10: /* 点云预处理 */
11: 距离: R =  $\sqrt{\mathbf{X}^2 + \mathbf{Y}^2 + \mathbf{Z}^2}$ 
12: 竖直角:  $\Theta = \arcsin(\mathbf{Z}/\mathbf{R})$ 
13: 水平角:  $\Phi = \arctan(\mathbf{X}/\mathbf{Y})$ 
14: 激光飞行时间:  $\text{ToF} = 2 \times \frac{\mathbf{R}}{c}$ 
15: laser_ID = Angle2ID( $\Theta$ )
16: 根据  $\Phi$  和 laser_ID 为每个点排序:  $PC_{sort} = \text{sort}(PC | \Phi, \text{laser\_ID})$ 
17: fullcycle_ID(0) = 0
18: for  $i = 1$  to  $N - 1$  do
19:    $\Delta = \Phi(i) - \Phi(i - 1)$ 
20:    $\Delta N_{fullcycle} = \text{fix}(\Delta / \delta_{hori})$ 
21:   if laser_ID( $i$ ) < laser_ID( $i - 1$ ) then
22:     fullcycle_ID( $i$ ) = fullcycle_ID( $i - 1$ ) +  $\Delta N_{fullcycle} + 1$ 
23:   else
24:     fullcycle_ID( $i$ ) = fullcycle_ID( $i - 1$ ) +  $\Delta N_{fullcycle}$ 
25:   end if
26: end for
27: /* 控制信号生成 */
28: 脉宽:  $TTL = 10 \times 10^{-9}$  s
29: Time_ideal(0) = 0
30: Amp_ideal(0) = 0
31: 最小阈值:  $\varepsilon = 1 \times 10^{-18}$ 
32: for  $i = 0$  to  $N - 1$  do
33:   Time_ideal( $i \times 4 + 1 : i \times 4 + 4$ ) =  $[-\varepsilon, 0, TTL, TTL + \varepsilon] + \text{Timestamp}(i)$ 
34:   Amp_ideal( $i \times 4 + 1 : i \times 4 + 4$ ) = [0, 1, 1, 0]
35: end for
36: Time_ideal( $N \times 4 + 1$ ) = (fullcycle_ID( $N - 1$ ) + 1)  $\times T_{fc}$ 
37: Amp_ideal( $N \times 4 + 1$ ) = 0
38: Signalideal  $\leftarrow$  以 Time_ideal 为横坐标, Amp_ideal 为纵坐标
39: Signaldiscrete  $\leftarrow$  以采样率 SR 对 Signalideal 进行采样

```

---

击为例详细介绍同步的方法。如图2.6a所示，首先本文通过用户手册<sup>[83]</sup>获取目标激光雷达的扫描序列和激光信号，对于激光信号，本文会选择某一束激光作为待接收信号，比如竖直角为1°的激光信号，并且将该激光信号的发射时间记作 $t_0$ 。然后是触发信号，触发信号用于指示延时控制器开始计时，并在计时结束后控制信号发生器发射控制信号，在实际攻击中，会使用光电传感器接收激光雷达的信号来作为触发信号，由于激光雷达和光电传感器之间有一段距离，所以触发信号相对激光雷达信号会有光飞行产生的延迟( $ToF_1$ )。随后，本文将触发信号传输至延迟控制器，经过精确的延迟( $D_{delay}$ )后，信号发生器生成控制信号，以控制激光发射器发射激光攻击信号。最终，激光攻击信号在经过一定飞行时间( $ToF_2$ )后被目标LiDAR接收。

在进行信号同步操作前，需要测量攻击设备的内在延迟( $D_{device}$ )。攻击设备包括接收器、激光发射器、信号发生器(内含延迟控制)，其中信号响应、激光充电、铜线中电信号传输等操作本身存在一定的延迟，为了更好地应对“光速灾难”，设备内在延迟也需要被充分考虑。测试设备内在延迟的拓扑图如图2.6b所示，将一个脉冲信号(起始信号)输入该工作流用于控制激光发射器发射器激光，并用示波器的通道1记录“起始信号”。激光发射器发光被接收器接收，接收器收到光后触发信号发生器发出一个脉冲信号(结束信号)，用示波器的通道2记录“结束信号”。起始信号和结束信号之间的时间差即为设备内在延迟，多次测量取平均值。

为了使被LiDAR接收的激光攻击信号的起始位置和扫描序列中某个完整周期的起始时刻( $t_{desire}$ )对齐，关键是在延迟控制器中设置精确的延迟 $D_{delay}$ ，本文通过式2-11计算延迟：

$$D_{delay} = D_{align} - ToF_1 - ToF_2 - D_{device}. \quad (2-11)$$

其中 $D_{align}$ 表示从LiDAR发出待接收信号的时刻 $t_0$ 到扫描序列中某个完整周期的起始时刻 $t_{desire}$ 之间的时间间隔，以针对VLP-16激光雷达的攻击为例，竖直角为1°的激光信号被选为待接收信号，则 $D_{align} = t_{desire} - t_0 = n * T_{fc} - T_{sfc}$ ，n可以为任意正整数； $ToF_1$ 表示激光信号从激光雷达到光电传感接收器的飞行时间； $ToF_2$ 表示激光攻击信号从激光发射器到激光雷达的飞行时间。

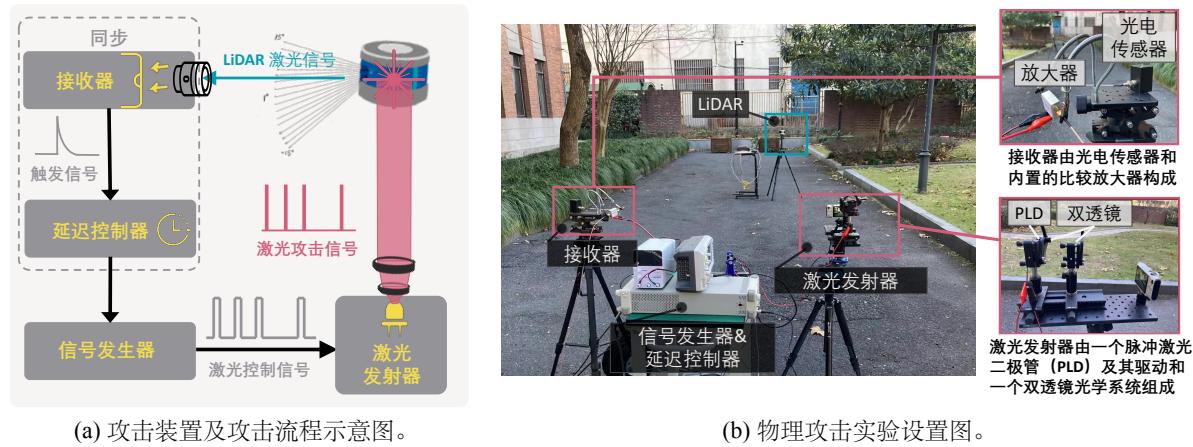


图 2.7 攻击设备及攻击设置

## 2.5 攻击设备介绍和攻击能力评估

### 2.5.1 攻击设备及攻击流程

为了在物理世界实现 PLA-LiDAR 攻击，本工作搭建了由接收器、延迟控制器、信号发生器和激光发射器组成的攻击装置。接收器由 PIN 光电二极管<sup>[84]</sup>和内置的比较放大电路构成；延迟控制器和信号发生器被集成于一台任意波形发生器中<sup>[85]</sup>；激光发射器由三个子模块构成：用于产生高压脉冲的激光驱动板、用于发射激光的脉冲激光二极管以及用于激光准直的双透镜系统。

实际的攻击流程如图2.7a所示，接收器首先接收来自目标激光雷达的激光脉冲并生成触发信号。任意波形发生器在收到触发信号后，引入预设的延迟并生成控制信号。最终，激光发射器向目标激光雷达发射攻击激光，从而实现欺骗点的注入。该攻击系统充分考虑了信号同步精度、激光发射功率以及光学系统的聚焦性能，以确保欺骗攻击的有效性和隐蔽性。

### 2.5.2 攻击能力评估

全面评估基于激光的点云注入能力对于安全社区具有重要意义：一方面，它有助于更客观地量化此类攻击对自动驾驶系统的潜在威胁；另一方面，它为后续研究提供了可复现的实验基准，从而确保仿真结果具备物理世界的可实现性。基于此，本研究从三个维度系统性地评估了攻击能力：欺骗点的最大可注入点数、位置控制能力以及形状控制

能力。此外，本文还深入分析了影响这些能力的关键因素。

### 2.5.2.1 最大可注入点数

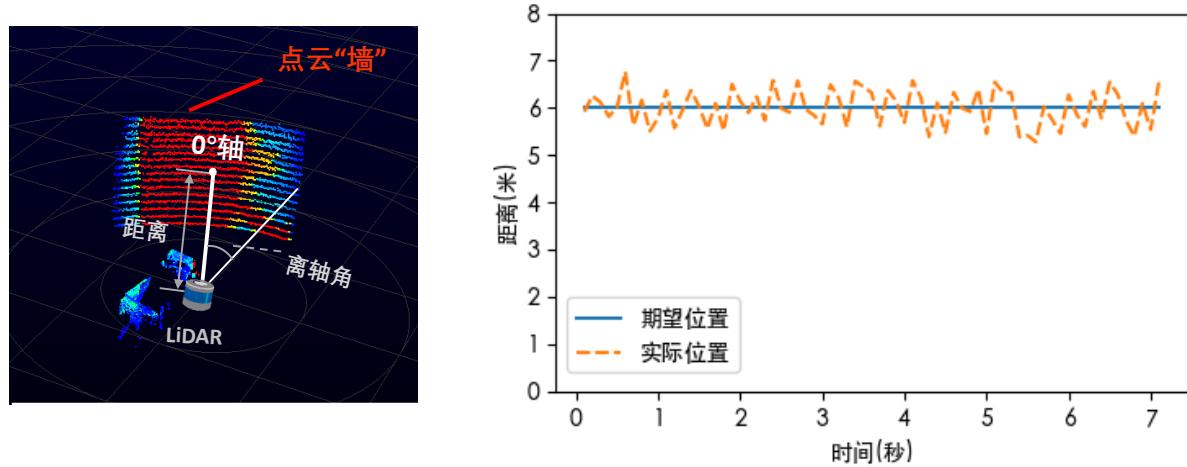
可注入的欺骗点数一直是以往衡量攻击能力的主要指标，这在以往的研究中得到了广泛采用<sup>[24,30-31]</sup>。此前的最先进研究<sup>[31]</sup>展示了在 VLP-16 激光雷达中注入多达 200 个点的能力。但以往的工作的注入点数不能反映激光攻击的真正能力，为了探索注入点数的能力边界，本文针对 VLP-16 进行了专门的实验。需要注意的是，这一指标仅在针对同一型号激光雷达且激光雷达转速相同时具有实际意义。

**能力量化：**为了探究点注入能力提升的可行性，本文选用了如下参数的激光：激光波长  $\lambda_{laser} = [850 \text{ nm}, 905 \text{ nm}, 915 \text{ nm}, 940 \text{ nm}]$ ，峰值功率  $P_{peak} = [25\text{W}, 75\text{W}, 125\text{W}, 300\text{W}, 600\text{W}]$ ，脉冲重复频率  $f_{rep} = [0 \text{ to } 800\text{kHz}]$ 。通过实验，本文发现激光能够在激光雷达中注入多种形状的欺骗点，在这些形状中，如图 2.8a 所示的“墙”形状具有最多的可控欺骗点数（当激光雷达转速为 300RPM 时可达 4800 个点）以及最大的攻击范围（水平角度  $30^\circ * 垂直角度 30^\circ$ ）。

**归因分析：**注入欺骗点的原理是用激光攻击信号伪造激光雷达工作信号的回波。成功实现欺骗点注入有四个关键因素：

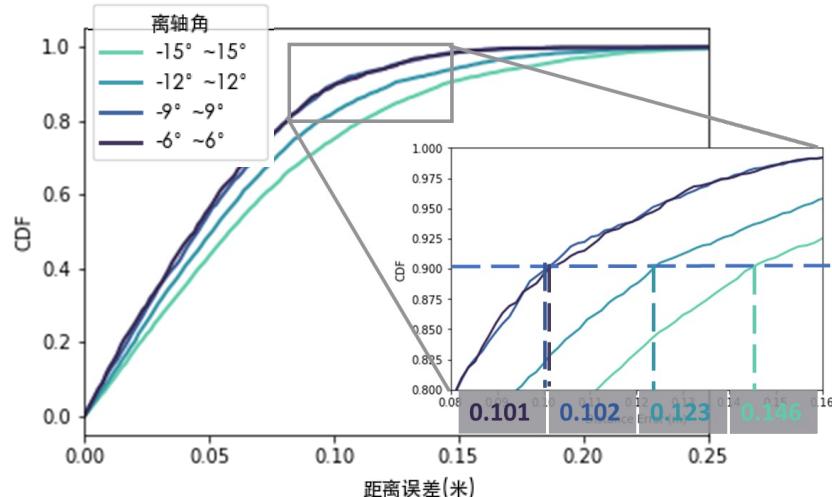
- 波形：攻击信号的波形应与激光雷达的信号相同，这样才能被识别为有效的回波。
- 波长：攻击信号的波长应与激光雷达的信号相似或相同，否则会被滤除。
- 激光峰值功率强度：信号强度必须足够强。根据激光雷达的回波模式，太弱的信号会被视为环境噪声而被滤除。
- 重复频率：为了注入更多的点，本文的目标是在激光雷达的每个“接收时刻”注入攻击信号。因此，信号的重复频率需要与激光雷达的扫描序列匹配。需要注意的是，并不是重复频率越高越好。

根据经验，正确的波形、合适的波长、高的峰值功率以及精确的重复频率能够实现最多的点注入。



(a) **最大可注入点数:** 可实现最多 4800 个可控欺骗点, 攻击范围达  $30^\circ$  (水平)  $\times 30^\circ$  (垂直)。

(b) **位置控制能力:** 能够将注入点云整体位置的精度控制在标准差为 0.38 米内。



(c) **形状控制能力:** 在离轴角  $\pm 9^\circ$  范围内, 90% 的点的距离误差均在 0.102 米以内

**图 2.8 攻击能力建量化**

### 2.5.2.2 位置控制能力

位置控制能力指的是攻击者能够精确控制注入点云的整体位置的能力。在点云注入攻击中, 攻击者通常需要在特定的位置注入欺骗点云以实现预期的攻击效果, 因此位置控制能力是评价攻击能力的重要评价指标。

**能力建量化:** 为了量化位置控制能力, 本文通过计算注入点云质心位置与目标位置之间的差异来评估控制精度。点云中一个点表示为  $p = (r, \theta, \phi)$ , 其中  $r$  表示距离,  $\theta$  表示

垂直角度， $\phi$  表示水平角度。点云的质心记为  $p_c = ((r_c, \theta_c, \phi_c))$ ，并且可通过式2-12计算：

$$p_c = \frac{1}{n} * \sum_{i=0}^n p_i. \quad (2-12)$$

在本实验中，目标是将点云的质心注入到  $p_c = (6m, 0^\circ, 0^\circ)$  的位置。在注入欺骗点后，通过连续 70 帧的数据采集及分析，实验结果表明，垂直角度和水平角度上能够实现连续且精确的控制，然而在距离控制方面存在一定误差，导致注入点云的距离随时间出现轻微波动，如图2.8b所示，距离误差的标准差为 0.38 米。

**归因分析：**实验结果表明，位置控制能力的误差主要来源于距离控制。本文分析认为，距离控制误差主要由以下因素引起：

- 光噪声：环境中的光噪声干扰了激光雷达的信号接收，导致距离测量不准确。
- 仪器抖动：实验设备的机械振动和电子噪声引入了额外的误差。
- “光速灾难”挑战：即使是很小的时间误差乘以光速也会导致显著的距离误差

在本实验中，距离控制的误差在 0.38 米内，这相当于时间误差仅为约 1.26 纳秒，考虑到实验设备采样率等问题，基于目前的攻击装置，误差难以被进一步缩小。本文推测，通过应用降噪技术或采用更高精度的攻击装置可以减轻这种误差。

### 2.5.2.3 形状控制能力

形状控制能力指的是攻击者能够精确控制注入点云的形状的能力。在点云注入攻击中，攻击者通常需要注入特定形状的点云以实现预期的攻击效果，例如行人或汽车或精心设计的对抗性点云，因此形状控制能力也是评价攻击能力的重要评价指标之一。

**能力量化：**在实际攻击中，攻击者通过控制点云中每个方向的点的存在与否以及每个点的距离来控制点云的形状，本文使用每个点的距离误差来衡量形状控制能力。在“最大可注入点数”实验中，本文发现攻击范围达  $30^\circ$  水平角  $\times 30^\circ$  竖直角，因此本文探究在这一范围内对不同区域的点的控制能力。在本实验中，本文将攻击信号设计为使所有点的距离均为 10 米，因此，本文能够注入一堵点云“墙”（如图2.8a所示）。首先，本文将欺骗性“墙”的中心轴定义为  $0^\circ$ ，并根据离轴角度（如图 2.8a 所示）对点云区域进行划分。然后，考虑到对点云墙的位置控制也存在误差，本文计算每个点云区域中点的平均距离，并以该平均距离作为基准值计算每个点的距离误差，以此消除位置控制误

差的影响。最终，本文得到了如图2.8c所示的累积分布函数（CDF）图，该图描述了包含不同离轴角度的点云区域的距离误差分布。本文发现，距离误差随离轴角度的变化而变化。总体而言，越接近 $0^\circ$ 轴的区域，误差越小。在靠近中心区域的范围内，误差的累积分布几乎相同，例如，在 $-9^\circ\sim9^\circ$ 区域和 $-6^\circ\sim6^\circ$ 区域中，90%的点的距离误差均在0.102米以内。

**归因分析：**较小的距离误差表明形状控制能力较强。总体而言，本文能够注入覆盖超过 $30^\circ$ 水平角的点云，并对大约 $20^\circ$ 水平角范围内的点实现相对精确的控制。本文推测，注入点的距离误差主要源于以下两个原因：(1)由于机械式LiDAR的不同入射角度和曲面结构导致的光路差异；(2)信号发生器采样率限制引起的误差，比如本实验中使用的DG5072信号发生器<sup>[85]</sup>采样率为1GHz。本文认为，通过使用高采样率的信号发生器可以减轻这种误差。

## 2.6 实验评估

在本节中，本文针对基于LiDAR的3D目标检测系统进行攻击评估。本文考虑了以下三组实验评估：(1)对Adv-Hide和Adv-Create的数字域评估，即通过优化算法直接生成对抗性点云并输入到3D目标检测模型中，该实验探究点数、目标类型对攻击成功率的影响；(2)对所有四种攻击类型（即Direct-Hide、Direct-Create、Adv-Hide和Adv-Create）的物理域评估，其中输入到模型的点云由LiDAR在物理世界直接采集，该实验探究攻击的实际可行性和不同攻击距离、攻击角度下的鲁棒性；(3)在攻击者和受害者均处于运动状态时，对物理攻击的可行性进行研究。

### 2.6.1 数字域评估

#### 2.6.1.1 实验设置

首先介绍数字域评估的实验设置。**LiDAR传感器**。针对两种商用LiDAR进行了数字域攻击：16线LiDAR VLP-16<sup>[75]</sup>和64线LiDAR HDL-64E<sup>[86]</sup>。

**目标检测模型**。使用两种3D目标检测器来评估攻击效果：PointPillar<sup>[8]</sup>和SECOND<sup>[87]</sup>。实验采用了MMDetection3D<sup>[88]</sup>框架中的实现。在KITTI数据集上的平均检测精度分别为：PointPillar为59.5%，SECOND为64.41%。

**数据集。**采用 KITTI<sup>[9]</sup>数据集，该数据集在 3D 目标检测器的训练与测试中得到了广泛应用。

**目标类别。**本文将三类自动驾驶场景中的主要目标作为创建或隐藏的目标类别，即车辆、行人和骑行者。上述两种目标检测模型均支持对这三类目标的检测。

### 2.6.1.2 评估方法

本研究采用攻击成功率（Attack Success Rate, ASR）作为评估指标，其定义为针对目标检测器的成功攻击次数与总攻击次数的比值。

对于 Adv-Hide 攻击，针对 16 线 LiDAR 模型，本文尝试注入 [20, 30, 40, 50, 60, 70, 80, 90, 100] 个对抗点；针对 64 线 LiDAR 模型，由于本身点云的点数更多，本文尝试注入 [100, 120, 140, 160, 180, 200, 220, 240] 个对抗点。每一组实验，本文都从 KITTI 数据集中为每一类目标（即车辆、行人和骑行者）随机选取 100 个目标实例，并尝试使其无法被检测到。对于 Adv-Create 攻击，针对 16 线 LiDAR 模型，本文尝试注入 [20, 40, 60, 80, 100, 120, 140, 160] 个对抗点；针对 64 线 LiDAR 模型，由于本身点云的点数更多，本文尝试注入 [100, 120, 140, 160, 180, 200, 220, 240] 个对抗点。每一组实验，从 KITTI 数据集中随机选取 100 个场景，并尝试在每个场景中分别注入一个车辆、一个行人或一个骑行者目标。综上，本实验通过 20400 帧实验来进行数字域的评估。

### 2.6.1.3 攻击效果

由于实验中最多可注入 4,200 个欺骗点，因此表 2.9 展示了不同点数下的最高攻击成功率。实验结果表明，Adv-Hide 攻击的整体攻击成功率为 73.84%，而 Adv-Create 攻击的整体攻击成功率为 66.76%，这说明隐藏现有目标比创建虚假目标更为容易。

针对不同类型的目标，实验结果表明，与车辆相比，本文提出的攻击方法在隐藏和创建行人及骑行者方面表现更好，如图 2.9 所示。具体而言，Adv-Hide 攻击对三类目标的隐藏成功率均高于 65%。Adv-Create 攻击在生成行人和骑行者方面表现良好（成功率高于 82%），但在生成车辆方面的成功率相对较低（33%）。这种成功率的差异可能与车辆尺寸较大有关。（1）由于车辆占据的空间更大，生成车辆时需要搜索更广泛的空间并进行更多迭代才能达到较高的成功率。为验证这一假设，研究进行了额外的实验与分

表 2.1 数字域对抗攻击针对不同目标在不同欺骗点数下的 Top-1 成功率

攻击类别	LiDAR	模型	Category of Object			Avg.	Overall.
			行人	骑行者	车辆		
Adv-Hide	VLP16	PoinPillar	100%	99%	88%	95.7%	68.00%
		SECOND	48%	35%	38%	40.3%	
	HDL64E	PoinPillar	100%	100%	100%	100.0%	79.67%
		SECOND	80%	59%	39%	59.3%	
Adv-Create	VLP16	PoinPillar	64%	64%	33%	53.7%	60.85%
		SECOND	95%	98%	11%	68.0%	
	HDL64E	PoinPillar	77%	81%	24%	60.7%	72.67%
		SECOND	93%	97%	64%	84.7%	

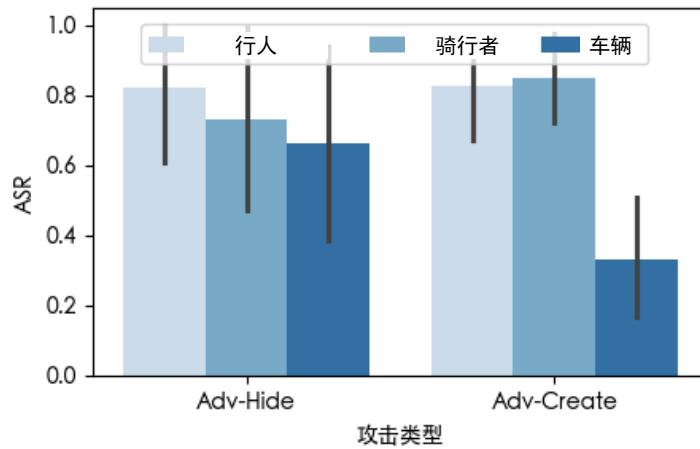


图 2.9 数字域下 Adv-Hide 和 Adv-Create 攻击针对不同目标类型的对抗攻击成功率

析：针对车辆测试了不同的迭代次数（300、500、800、1500、2000 和 2500），发现攻击成功率随迭代次数的增加而提升，并在 2000 次迭代时趋于饱和。通过增加迭代次数，可将车辆的攻击成功率提升至 56%。（2）由于车辆尺寸较大，生成车辆需要在原始帧中预留更大的空白区域。通过对 KITTI 数据集中随机选取的 100 帧进行分析，发现这一条件并非总能满足。在某些帧中，目标区域（LiDAR 正前方 10 米内）可能已存在环境点云，导致无法成功注入欺骗点云。统计显示，不满足条件的帧在车辆生成任务中超过 10/100，而在骑行者和行人生任务中仅为 5/100 左右，这进一步解释了车辆生成成功率较低的原因。

表 2.2 针对不同 LiDAR 设备及目标检测模型的物理攻击成功率

模型	LiDAR 型号	Attack Types			
		Direct-Hide	Direct-Create	Adv-Hide	Adv-Create
SECOND	<b>VLP-16</b>	100%	98%	47%	75%
	<b>RS-16</b>	100%	86%	46%	66%
PoinPillar	<b>VLP-16</b>	100%	64%	78%	24%
	<b>RS-16</b>	100%	51%	74%	19%
Apollo	<b>VLP-16</b>	100%	98%	81%	39%
	<b>RS-16</b>	100%	89%	76%	24%

## 2.6.2 物理域评估

在物理域评估中，本文通过激光攻击的方式向激光雷达（LiDAR）注入欺骗点，对 Direct-Hide、Direct-Create、Adv-Hide 和 Adv-Create 四种攻击进行了评估。

### 2.6.2.1 实验设置

**攻击准备**作为物理世界实验的前置步骤，本文需要预先获取所需的点云并据此设计攻击信号。Direct-Hide 和 Direct-Create 的点云通过替代激光雷达的录制获取：对于 Direct-Hide，原始点云是一面墙；对于 Direct-Create，原始点云是一个行人。Adv-Hide 和 Adv-Create 的目标点云则通过对抗性机器学习生成（详见第2.4.2.2章），在生成过程中融入了鲁棒性增强机制，并选择在数字域中已经成功攻击的点云用于物理攻击。

**攻击设备**本文使用如图 2.7b 所示的攻击设备设置，将攻击设备放置于目标激光雷达前方，并调整不同的距离和角度以进行不同攻击条件下的鲁棒性测试。实验中使用的所有设备型号均列在本文的网站上。基于该设备设置，本文在校园道路上进行了物理实验。

**目标激光雷达及感知模型**本文对两种机械式激光雷达（LiDAR）进行了物理攻击实验，分别是 VLP-16 和 RS-16，这两种激光雷达是全球范围内应用广泛的车载商用激光雷达。在攻击过程中，激光雷达生成的点云数据直接输入到 3D 目标检测模型 SECOND<sup>[87]</sup>，PointPillars<sup>[8]</sup> 和 Apollo r6.5<sup>[64]</sup> 中进行分析。

### 2.6.2.2 评估方法

对于 Direct-Hide 和 Adv-Hide 攻击，实验尝试对行人、骑行者及车辆三类目标进行隐藏。而 Direct-Create 和 Adv-Create 攻击致力于生成虚假的行人目标。在每类攻击中，针对不同 LiDAR 设备和检测器都随机采集 100 帧数据进行测试，并以攻击成功率 (ASR) 作为评估攻击效果的指标。

### 2.6.2.3 攻击效果

在物理攻击实验中，共计采集了 2400 帧数据进行评估。物理攻击的整体性能总结如表 2.2 所示。同时，物理攻击的相关视频也可在网站<sup>[63]</sup>上查阅。

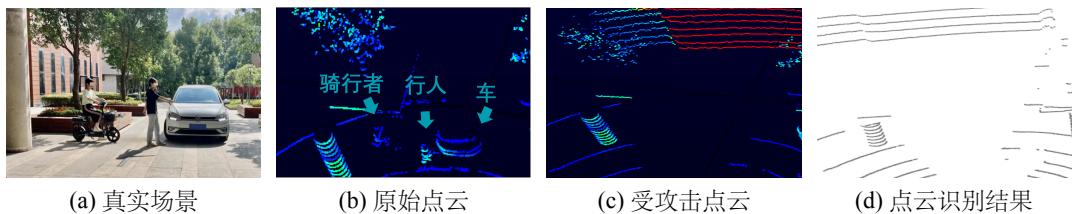


图 2.10 Direct-Hide 物理世界攻击效果.

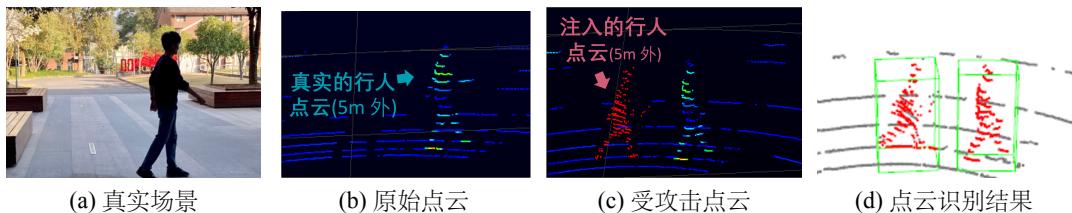


图 2.11 Direct-Create 物理世界攻击效果.

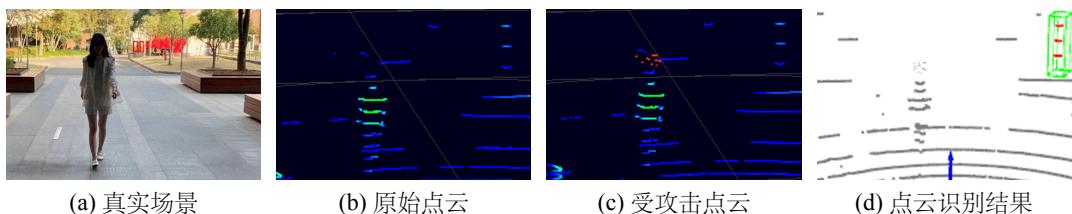


图 2.12 Adv-Hide 物理世界攻击效果.

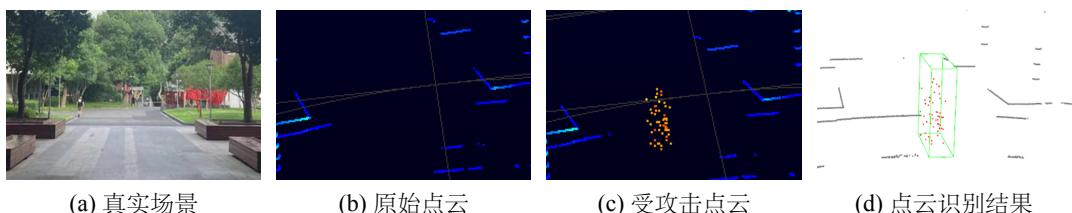


图 2.13 Adv-Create 物理世界攻击效果.

**激光雷达型号的影响。**针对不同的激光雷达 (LiDAR) 型号，攻击性能表现出轻微差异。具体而言，VLP-16 的平均攻击成功率 (ASR) 为 75.33%，而 RS-16 的平均攻击成功率为 69.25%。这种差异的原因是 RS-16 配备了脉冲随机化技术，能够有效消除攻击。大约每经过一百个完整周期 ( $100 \times 55.555 \mu s$ )，RS-16 会有一个约  $133 \mu s$  的静默期，这为精确注入欺骗性点云带来了额外的难度。

**目标检测模型的影响。**攻击性能在不同检测模型之间存在差异。Direct-Hide 攻击对三种检测系统的攻击成功率 (ASR) 均能达到 100%。Direct-Create 攻击在 SECOND 和 Apollo 上的表现更好，这是因为 SECOND 和 Apollo 在检测真实行人方面表现更优，而通过 Direct-Create 攻击注入的欺骗性点云与真实行人的点云非常相似。Adv-Hide 攻击在 PointPillars 和 Apollo 上表现更佳，而 Adv-Create 攻击在 SECOND 上表现更好。本文推测这种性能差异可能源于这三种模型的特征提取过程不同：SECOND 将点云空间划分为体素 (voxels)，而 PointPillars 和 Apollo 则将点云空间划分为垂直柱 (pillars) 并利用 PointNets 学习特征。对于 Adv-Hide 攻击，本文在目标物体上方的空间添加对抗性点，因此这些对抗性点更可能破坏柱状结构下方点云的特征提取。对于 Adv-Create 攻击，本文在长方体空间内优化点云，这与体素更为相似，而非柱状结构，这使得 SECOND 更容易受到攻击。

#### 2.6.2.4 鲁棒性分析

本章研究了攻击激光源在不同距离、高度和角度下的攻击鲁棒性，实验在 VLP-16 激光雷达 (LiDAR) 及 SECOND 模型上进行。对于 Direct-Hide 和 Adv-Hide 攻击，本文尝试隐藏位于激光雷达前方 5 米处的真实行人。对于 Direct-Create 和 Adv-Create 攻击，本文尝试在激光雷达前方 5 米处创建一个虚假行人。

**攻击距离的影响。**本文进行了攻击距离为 1 米、3 米、5 米、10 米和 15 米的实验。本文为每个距离和每种攻击类型收集了 20 个帧（共 400 个帧），以统计攻击成功率。图 2.14a 中的结果显示，Direct-Create、Adv-Hide 和 Adv-Create 攻击的 ASR 会随着攻击距离的增加而降低，而 Direct 攻击的 ASR 仍能达到 100%。本文认为其原因是激光功率强度随着距离的增加而降低，使得欺骗点云控制精度下降。在四种攻击中，Direct-Hide 攻击对欺骗点云控制精度的要求最低，因此受攻击距离的影响最小。

**LiDAR 安装高度的影响。**本文使用 0.2 米、0.7 米、1.2 米、1.7 米和 2.2 米的不同激

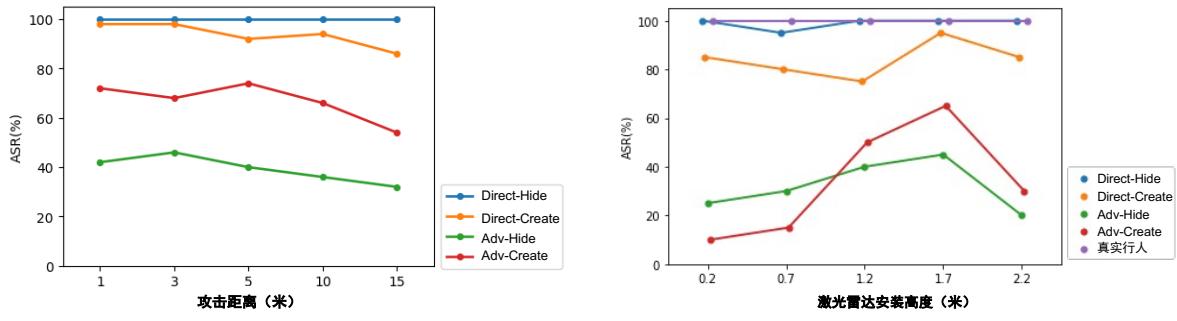


图 2.14 攻击距离及雷达安装高度对攻击成功率的影响

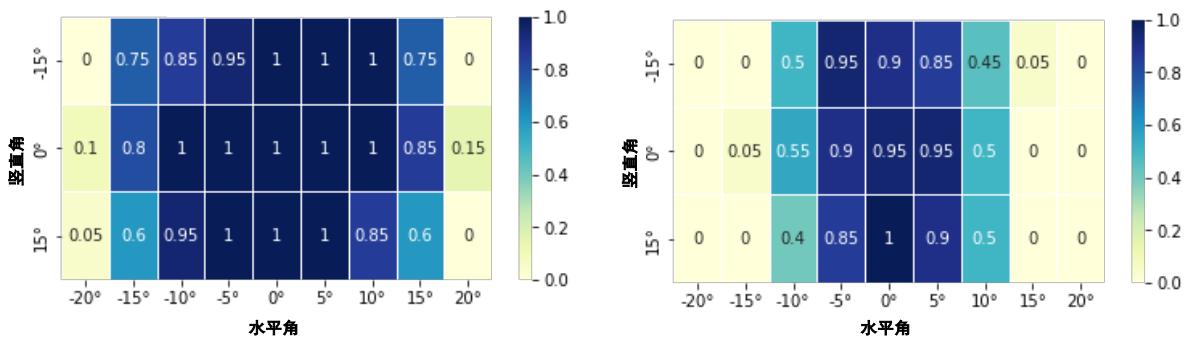


图 2.15 攻击角度对隐藏和创建攻击成功率的影响

光雷达安装高度进行实验。本文为每种激光雷达安装高度和每种攻击类型收集 20 帧图像（共 400 帧），以统计攻击成功率。图 2.14b 中的结果显示了本文在不同激光雷达安装高度下的攻击成功率。在四种攻击类型中，Direct-Hide 和 Direct-Create 攻击的表现并没有随着激光雷达安装高度的变化而发生明显变化，而 Direct-Create 和 Adv-Create 攻击在激光雷达安装高度为 1.7m 时的 ASR 最高。这可能是因为训练数据集 KITTI<sup>[89]</sup>是由安装高度为 1.73 米的激光雷达采集的，因此，基于优化的攻击对激光雷达的安装高度更为敏感。

**攻击角度的影响。**本文研究了激光源处于不同水平和垂直角度时的攻击效果。对于隐藏攻击，本文采用“直接隐藏”(Direct-Hide) 攻击的方式，试图隐藏 5 米外水平角度约为 0° 的真实行人。对于创建攻击，本文采用 Direct-Create 攻击，并尝试在 5 米外创建一个水平角度约为 0° 的假行人。本文为每个攻击角度和每种攻击类型收集 20 帧图像（共 1080 帧），以报告攻击成功率。图 2.15a 和图 2.15b 中显示的结果表明，这两种攻击受水平角度的影响都大于垂直角度。隐藏攻击主要在水平角度 [-15°, 15°] 内成功。本文认为其



图 2.16 攻击移动车辆实验设置

原因是激光雷达接收器（主要由光电二极管和透镜组成）的接收角度有限。Direct-Create 攻击主要在水平角度  $[-10^\circ, 10^\circ]$  范围内成功，该范围小于隐藏攻击。本文认为原因在于，Direct-Create 攻击需要对点云的形状和距离进行精细控制，而在  $[-10^\circ, 10^\circ]$  区域内注入的点误差较小，如图2.8c所示。

### 2.6.3 可行性实验——攻击移动车辆

**实验设置：**本文定义了一种动态车辆攻击场景如图2.16所示：攻击车辆与目标车辆以约 5 公里/小时的安全速度同向行驶（出于安全），两者间距保持 5-15 米。攻击设备集成方案为：在攻击车辆顶部安装接收器与激光发射器（均连接云台用于手动瞄准），后备箱搭载任意波形发生器（AWG）、笔记本电脑、激光驱动板及电源等设备。目标车辆采用 Apollo D-kit 平台并配备 VLP-16 激光雷达，其采集的点云数据用于实时 3D 目标检测。相较于静态攻击实验配置，本文升级了攻击硬件系统以缓解车辆移动引起的抖动效应：(1) 采用大口径望远镜（直径  $\Phi=50$  毫米）将接收器接收面积从 0.2 平方厘米扩展至 78.5 平方厘米；(2) 将光斑直径扩大至 8 厘米，并使用大功率激光二极管（峰值功率  $P_{peak}=300$  瓦），确保光斑功率密度大于 2 瓦/平方厘米。通过增大接收面积与光斑覆盖范围，即使车辆行驶过程中存在轻微抖动也不会影响攻击有效性。

**攻击结果：**在动态场景下依然可以成功实施隐藏攻击和创建攻击，具体来说，隐藏攻击可以达到 94.1% 的 ASR (16/17 次试验)，创建攻击可以达到 78.9% 的 ASR (15/19 次试验)。对移动车辆进行物理攻击的视频可以在网站<sup>[63]</sup>上找到。

## 2.7 本章小结

本章探究了激光雷达的信号鉴权脆弱性，提出了可以通过在物理世界使用红外激光注入欺骗点云来对 3D 目标检测模型直接进行欺骗。本章设计了一种针对激光雷达感知系统的物理激光攻击方法——PLA-LiDAR。PLA-LiDAR 包含一套高性能激光收发装置，能够注入多达 4200 个可控欺骗点；一种同时考虑了激光雷达的工作原理、攻击设备的能力以及注入点的距离误差的对抗性点云优化方法，能够生成物理可注入的对抗性点云；一种“空间坐标-时间坐标”映射的控制信号设计方法和精确到纳秒级别的信号同步方法，能够将上述生成的对抗性点云精确注入激光雷达。基于上述方法，PLA-LiDAR 共实现了四种攻击效果，能分别以黑盒和白盒的方式实现隐藏攻击和创建攻击。除此以外，本章从注入点数、位置控制能力、形状控制能力三方面量化了 PLA-LiDAR 的攻击能力，能够为其他仿真研究提供物理可实现的参考。通过在 2 款激光雷达和 3 款目标检测模型上的数字域和物理域评估，本章验证了对抗点云优化算法的有效性和 PLA-LiDAR 的物理攻击可行性。最后本章通过移动的实车上进行的实验进一步证明了 PLA-LiDAR 的物理可行性。

### 3 基于电磁干扰攻击的激光雷达感知脆弱性分析

激光雷达是自动驾驶的关键传感器，可提供精确的三维空间信息。以往针对激光雷达系统的信号攻击主要利用同模态激光信号，本章研究了跨模态信号注入攻击的可能性，即注入有意电磁干扰（IEMI）来操纵激光雷达输出。本章发现激光雷达的内部功能模块（包括激光接收电路、监测传感器和光束偏转模块等）即使经过严格的电磁兼容性（EMC）测试，仍可能与 IEMI 攻击信号耦合，导致点云甚至激光雷达系统出错。基于上述研究发现，本章提出了 PhantomLiDAR 攻击，利用电磁信号实现了点云干扰、点云抹除、点云注入、激光雷达宕机四类攻击效果。本章在五个商用激光雷达系统上进行了数字域和物理域的实验，评估并证明了攻击的有效性。本章在真实世界的移动场景中进行了可行性实验，进一步证明攻击在真实场景的威胁。本章还讨论了可在传感器层面和车辆系统层面实施的潜在防御措施。实验相关的视频演示可在对应网站<sup>[90]</sup>上查看。

#### 3.1 本章引言

激光雷达传感器由于其高精度、全天时、远距离探测等优势，在自动驾驶汽车感知系统中发挥着不可替代的作用，激光雷达的正确感知是自动驾驶汽车的安全性和可靠性的基础。然而，激光雷达本身暴露在物理环境中，面对着现实世界中无处不在的物理信号干扰，大量研究<sup>[24,28,30,32,91-92]</sup>表明，激光雷达容易受到激光信号的攻击。在这些工作中，攻击信号和激光雷达的工作信号属于同模态的信号，激光攻击通常以激光测距电路中的光电传感器为攻击入口。为了全面分析激光雷达面对信号注入攻击时的脆弱性，本文研究了“跨模态信号攻击”的可能性，探索了除激光攻击之外的更广泛的攻击范式。最近的一项研究<sup>[93]</sup>表明，电磁信号会干扰激光雷达的接收信号，并诱发“传感器数据扰动”的攻击效果。但总体而言，之前的所有研究都是通过将信号注入激光雷达的接收模块中来干扰激光雷达，是否存在针对激光雷达系统的新漏洞和新攻击入口仍是一个待研究的问题。

本章的研究目标是探索电磁干扰（Electromagnetic Interference, EMI）下的激光雷达的新型漏洞和新型攻击范式，以此为更安全的激光雷达设计提供参考。挖掘新型漏洞

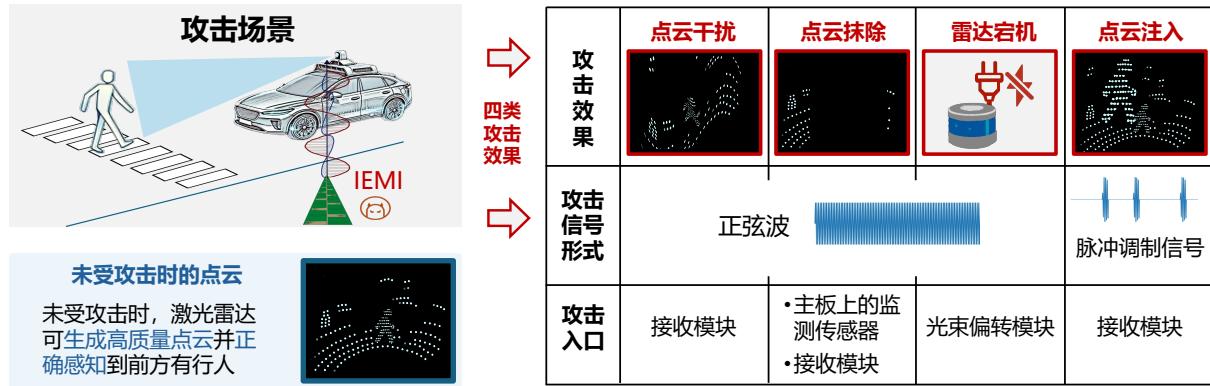


图 3.1 PhantomLiDAR 攻击介绍图：通过将不同的攻击信号注入激光接收模块、监测传感器和光束偏转模块等不同的攻击入口，PhantomLiDAR 可以成功实现“点云干扰”、“点云抹除”、“点云注入”甚至“雷达宕机”等攻击效果。

和提出新的攻击范式具有以下挑战：(1) 由于商用激光雷达系统具有完整的封装和防逆向功能，隐藏了其内部机制，使基本原理分析复杂化，因此挖掘并定位新型漏洞机理非常困难。(2) 由于商用激光雷达出厂前进行了严格的电磁兼容性 (EMC) 测试并采用了抗干扰设计 (如电磁屏蔽和线路优化)，激光雷达系统通常具有抗 EMI 的能力，因此真正在物理世界实现有危害性的攻击非常困难。

为了应对这些挑战，本章首先通过拆解激光雷达和参考逆向工程报告<sup>[94-95]</sup>系统地分析了激光雷达的内部结构，并推断出监测传感器 (如温度传感器<sup>[96]</sup>，霍尔效应传感器<sup>[97]</sup>) 和光束转向模块有可能成为 EMI 信号耦合的入口。然后，本文建立了一套宽频的高性能电磁攻击设备，通过模糊测试 (Fuzzing) 实验寻找漏洞。随后，本文利用故障检测和诊断 (Fault Detection and Diagnostic, FDD) 机制定位攻击入口，并通过激光雷达的内部电路进行验证实验。通过上述步骤，本文发现了两个新的攻击入口：监测传感器 (如温度传感器和霍尔效应传感器) 和光束转向模块中的光编码器。在此基础上，本文可以实现新的攻击效果，如“点云移除”和“雷达宕机”，以及更强的“点云干扰”。此外，为了探索利用电磁信号巧妙操纵激光雷达系统进行精确数据操控的可行性，本文尝试向激光雷达系统中注入可控点。之前的点云注入相关工作<sup>[24,28,30,32,91-92]</sup>都是利用同样波长的激光来伪造激光雷达的回波从而注入可控点。如果本文直接利用类似的方法，利用电磁来伪造激光雷达的回波，会面临信号无法耦合到电路中的问题。为了解决这一挑战，本文通过幅度调制的方式，首先找到合适频率的载波信号，然后将激光雷达工作信号作为基波调制到载波上，进而实现信号的有效注入。通过这种方法，本文成功利用电磁实现了“点云注入”攻击。

总的来说，如图3.1所示，本文成功地实施了四种类型的攻击：1) 点云干扰，通过向激光接收模块的模拟电路中注入 EMI，可以在激光雷达测距中引入误差，从而使点云失真。2) 点云抹除，通过向主板上的监测传感器或激光接收模块中注入 EMI，这种攻击会导致点云严重偏离其真实位置或完全消失。3) 点云注入，通过向激光接收电路注入调幅电磁波，可以实现可控的点云注入。4) 雷达宕机，通过干扰光束偏转模块中的光电编码器，可以使得激光雷达光束偏转出错，迫使激光雷达系统关闭并停止工作，即使攻击停止后，激光雷达系统也必须手动重启才能恢复运行。

其中，点云抹除、点云注入、雷达宕机是三种新的攻击范式。此外，点云干扰攻击能够注入 15cm 的误差，增强了之前工作中声称的仅能注入 4cm 误差的攻击能力<sup>[93]</sup>。本文希望有限成本下攻击范式的增加和攻击能力的增强能够帮助安全界和激光雷达制造商准确地认识到电磁干扰对激光雷达系统造成的威胁，这种认识有望进一步促进建立更先进的电磁兼容测试标准和设计更安全的激光雷达系统。

为了评估 PhantomLiDAR 攻击，本文在 5 个商用激光雷达系统上进行了评估，包括 3 个机械旋转式激光雷达和 2 个 MEMS 激光雷达。通过在数字域模拟攻击效果和在物理域将采集到的点云直接输入模型的方式全面评估了攻击对自动驾驶实际功能的影响。为了更好地了解四种攻击类型的优势和局限性，本文为每种攻击类型设计了独特的评估方法。值得注意的是，本文发现 PhantomLiDAR 能够在模型层面隐藏指定目标，最远攻击距离达到 5 米。此外，PhantomLiDAR 可以在 VLP-16 激光雷达中注入超过 16,000 个假点，这一数量是 SOTA 基于激光的点云注入攻击<sup>[91]</sup>的五倍，后者在相同激光雷达旋转速度下可以在 VLP-16 中注入不到 3,000 个假点。此外，本文还在移动场景中进行了可行性实验。

本章的主要贡献如下：

- **新攻击入口：**本章发现了激光雷达的 2 个新的 EMI 攻击入口，即主板上的监测传感器和光束偏转模块中的光电编码器。
- **新攻击范式：**本章提出能利用电磁信号实现点云干扰、点云抹除、点云注入、雷达宕机 4 种攻击效果。
- **强攻击能力：**与 SOTA 工作相比，点云干扰攻击的干扰能力强 3 倍。点云抹除攻击可以在无需精确瞄准的情况下远程隐藏目标。雷达宕机攻击可以在常用的机械激

光雷达 VLP-16 和 MEMS 激光雷达 RS-M1 上取得成功。点云注入攻击可注入的可控点数量是 SOTA 激光攻击的 5 倍。

- **实验:** 本章在五个商用激光雷达系统上进行了实验。对于四种类型的攻击，本文根据攻击特点针对性设计了数字域和物理域实验，以更好地评估攻击的优势和局限性。

## 3.2 背景知识

### 3.2.1 激光雷达功能模块

激光雷达通过生成点云数据提供精确的三维空间信息。图3.2 展示了典型激光雷达的主要组件，其中包括用于信号收发的激光发射模块和接收模块，在测距过程中，主板控制发射模块发射激光信号，并记录其方向（水平角  $\theta$ , 垂直角  $\phi$ ）和发射时间  $\tau_0$ 。激光脉冲在空气中传播，当它击中物体时，部分能量被反射并被接收模块接收。然后，光信号通过光电传感器转换成模拟电信号，经过放大器、滤波器和 ADC 后，模拟信号被转换成数字信号，输入 FPGA。FPGA 中的算法可以确定接收时间  $\tau_1$  和回声信号的强度。目标距离  $r$  根据光的飞行时间来测量：

$$r = \frac{1}{2}c \cdot (\tau_1 - \tau_0), \quad (3-1)$$

其中  $c$  是激光雷达与目标之间介质（如空气）中的光速。这一过程称为一次飞行时间（Time of Flight, ToF）测距，通过一次 ToF 测距可以生成一个激光点，记为  $Point = [r, \theta, \phi]$ 。大多数车载激光雷达系统都是利用 ToF 技术来测距，工作波长为 905 nm。

为了创建描述环境三维信息的点云，激光信号应照射到所需视场（Field of View, FoV）中的所有点。这可以通过使用光束偏转模块改变激光信号的方向来实现。多年来，人们开发了许多不同的光束偏转技术，总体上来说分为机械式和固态：其中机械式光束偏转技术包括最主流的机械旋转式<sup>[75-76,98-99]</sup>和微机电式<sup>[77]</sup>，光束偏转模块中通常有一个光学编码器，用于监控电机的运行状态；也有少部分厂家开始探索使用光学相控阵<sup>[100]</sup>来设计固态激光雷达的可能性等。在本章的研究中，本文主要关注这些使用机械运动或反射镜进行光束转向的商用激光雷达系统，因为它们已广泛应用于当今的自动驾驶车辆中<sup>[101-102]</sup>。

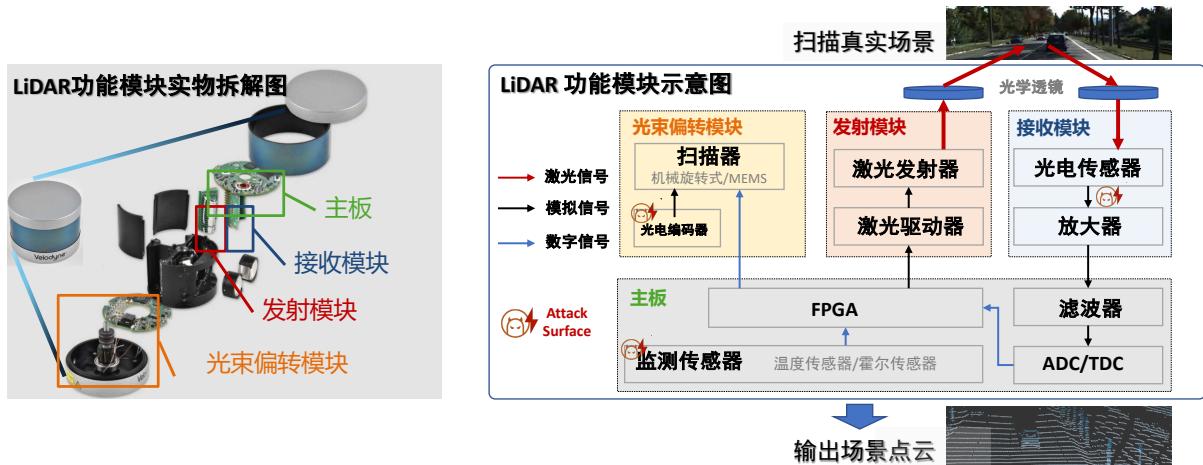


图 3.2 激光雷达功能模块介绍图：典型的激光雷达系统包括用于信号收发的激光发射模块和接收模块、用于激光扫描的光束偏转模块和用于控制计算的主板。

除了上述与点云生成相关的组件外，激光雷达制造商通常还会加入状态监测传感器，以监督这些模块的运行状态，如电源电压状态（霍尔效应传感器）、温度等。在本章中，本文通过实验验证了接收模块、监控传感器和光学编码器可以成为 EMI 的攻击入口。

### 3.2.2 激光雷达错误诊断和检测机制

为确保激光雷达按预期运行，激光雷达制造商通常会使用故障检测和诊断<sup>[103-104]</sup>机制来确保这些模块按预期运行。如果激光雷达系统无法按预期运行，就有可能造成损坏。例如，过高的激光发射功率或电机故障导致单个方向持续受到激光照射可能对人的眼睛造成危害。因此，为了保护人员和激光雷达系统本身，通常会在激光雷达的设计中集成 FDD 和响应机制。

根据激光雷达和汽车制造商发布的专利<sup>[105-107]</sup>，制造商对常见的激光雷达故障进行了分类，并根据严重程度和后果将其分为两个级别：1 级（L1）故障是指影响较小的故障。在 L1 故障下，激光雷达可以继续运行，但性能或参数会降低；2 级（L2）故障是指严重影响雷达工作甚至有安全风险的故障，如电机转动出错可能会导致激光持续照向一个方向从而灼伤用户，当诊断出 L2 级故障时，激光雷达会自动切断电源。如图 3.3 所示，激光雷达运行期间有四种典型状态：初始化、正常、警告和下电。检测到故障时，激光雷达系统可能会在这四种运行状态之间交替。激光雷达开机后，首先进入初始化并执行自检。然后，激光雷达启动电机，一旦电机速度达到预设值且自检通过，就会进入正常

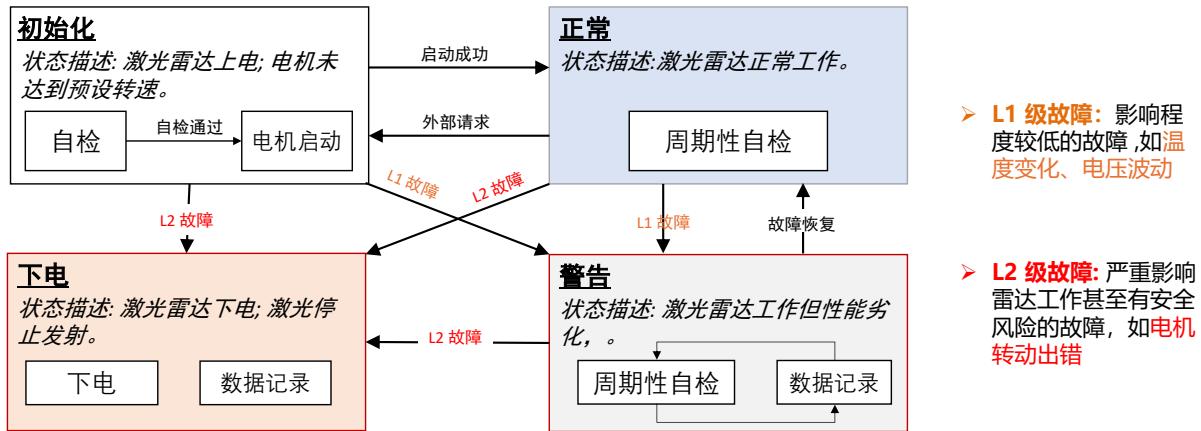


图 3.3 激光雷达故障检测和诊断机制：典型的激光雷达系统包括用于信号收发的激光发射模块和接收模块、用于激光扫描的光束偏转模块和用于控制计算的主板。

运行状态。在正常状态下，激光雷达会定期进行自检。如果在初始化或正常状态下检测到 L1 故障，激光雷达会进入警告状态。如果在任何状态下检测到 L2 故障，激光雷达将进入下电状态，在该状态下通常会关闭电源及通信。

### 3.2.3 电磁干扰（EMI）攻击

国际电工委员会（IEC）将 EMI 一词正式定义为“蓄意或恶意产生电磁，将噪声或信号引入电气和电子系统，从而扰乱、混淆或破坏这些系统，以达到恐怖主义或犯罪目的”<sup>[108]</sup>。攻击者首先考虑将 EMI 信号耦合到目标的可能方法，以成功实施 EMI 攻击。本文将这一过程称为构建耦合通道，耦合通道有三个基本组成部分：

**攻击入口（Attack Surfaces）。** 攻击入口通常存在于被攻击目标电路中的“导线”，其作为非预期的接收天线，易受 EMI 信号的影响。这些天线可以是印刷电路板（PCB）上的模拟电迹<sup>[109-115]</sup>，也可以是传感器与控制器之间的数字传输通道<sup>[116-117]</sup>。

**耦合路径（Coupling Path）。** 耦合路径决定了攻击者生成的 EMI 如何到达目标设备。耦合路径的选择主要取决于目标攻击入口，以实现最佳的耦合效率。耦合路径可分为两类：辐射耦合和传导耦合。辐射耦合通过空气或真空传播电磁能量，不与目标发生任何物理接触，例如磁耦合（改变磁场）、电耦合（改变电场）和电磁耦合（同时改变磁场和电场）。传导耦合通过导线传导将电磁能从电磁源传播到攻击入口<sup>[118]</sup>。

**电磁干扰源（EMI Source）。** 为了有效地将 EMI 信号传输到目标系统中，攻击者需要生成频率与目标系统电气特性相匹配、振幅合适的 EMI 信号，以便以所需方式改变系统数据。

### 3.3 威胁模型

#### 3.3.1 攻击目标

攻击者的目标是利用电磁信号隐蔽地干扰点云或激光雷达本身，从而降低激光雷达系统的可靠性和性能。具体来说，本文考虑了4种类型的攻击：

- **点云干扰**: 这种攻击的目的是在激光雷达的距离测量中引入误差。
- **点云抹除**: 这种攻击的目的是使物体的点云与实际位置严重偏离，或将其完全抹除，从而使基于激光雷达的感知模型无法检测到物体。
- **雷达宕机**: 这种攻击的效果是强制关闭激光雷达，使其无法工作。即使攻击停止后，激光雷达也需要重新启动才能恢复运行。
- **点云注入**: 这种攻击的目的是注入位置和形状可控的假点。

#### 3.3.2 攻击者能力

本文考虑攻击者具有以下能力或限制：

**电磁攻击能力**: 攻击者可以远程向受害自动驾驶汽车或机器人的激光雷达发射电磁信号，而无需在目标激光雷达系统上安装任何硬件或软件。为了实现这一目标，本文假设攻击者配备了可以产生电磁信号的商用设备，包括射频天线、信号发生器和射频功率放大器。攻击设备可以安装在攻击者的汽车上，这样攻击者就可以跟踪受害者车辆，并在一定距离内实施电磁注入攻击。他们还可以把车停在路边，攻击过往车辆或等红灯的车辆。

**激光雷达参数获取**: 攻击者知道目标激光雷达的型号，而且他可以事先获得一个同型号的替代激光雷达进行评估。例如，他可以在正式实施攻击之前，对替代激光雷达进行扫频实验，以离线方式找到脆弱频点。

**攻击成本**: 攻击者可能需要高端设备来执行宽频范围扫描，以测试激光雷达系统的漏洞。一旦找出这些漏洞，就可以使用成本较低的攻击设备实施攻击。

**模型黑盒**: 攻击者无法访问感知模型的参数和结构，攻击者只能利用传感器的特性和漏洞来实现攻击目标。

## 3.4 攻击可行性和原理分析

本文首先通过实验探索 EMI 注入攻击的可行性和其潜在的攻击效果，然后分析各种攻击效果的攻击原理。

### 3.4.1 攻击直觉

在实际攻击前，本文经验性地认为可能有两种方式可以攻击激光雷达：直接攻击和间接攻击。

**直接攻击：**从第3.2.1章中，本文了解到激光雷达点是通过 ToF 测距过程生成的，其中激光信号通过接收电路中的光电探测器转换为模拟电信号。因此，破坏点云最直接方法就是干扰接收模块中的模拟信号，直接影响激光雷达的测距机制，进而影响点云的生成。

**间接攻击：**从3.2.2章中，本文了解到当激光雷达的 FDD 机制检测到故障时，激光雷达会进入异常自我保护状态，在这种状态下，它可能会将点云视为无效，甚至强制关闭激光雷达。因此，通过攻击激光雷达中的其他模块进而借助 FDD 机制间接诱发点云错误或激光雷达本体故障可能是可行的。例如通过攻击温度传感器或光束偏转模块中的光学编码器，这可能诱发激光雷达检测错误，触发 FDD 的固有操作，迫使激光雷达进入故障恢复，从而导致拒绝服务甚至关机。

### 3.4.2 攻击可行性

当攻击者试图注入电磁信号以干扰激光雷达时，会将激光雷达在印刷电路板上的模拟电迹或电线视为接收天线。接收天线对于不同频率的电磁信号的接收效率不同，从理论上讲，可以根据目标天线的长度和形状估算出产生最大耦合效率的谐振频率<sup>[119]</sup>。然而，由于激光雷达中目标天线的未知性，精确计算信号频率具有挑战性。为了应对这一挑战，最常用的方法之一是频率扫描。

#### 3.4.2.1 实验设置

可行性研究的实验装置如图3.4 所示。攻击设备包括用于产生 EMI 信号的 Keysight N5712b 矢量信号发生器、用于放大 EMI 信号的 Mini-Circuits HPA-50W-63+ 功率放大器

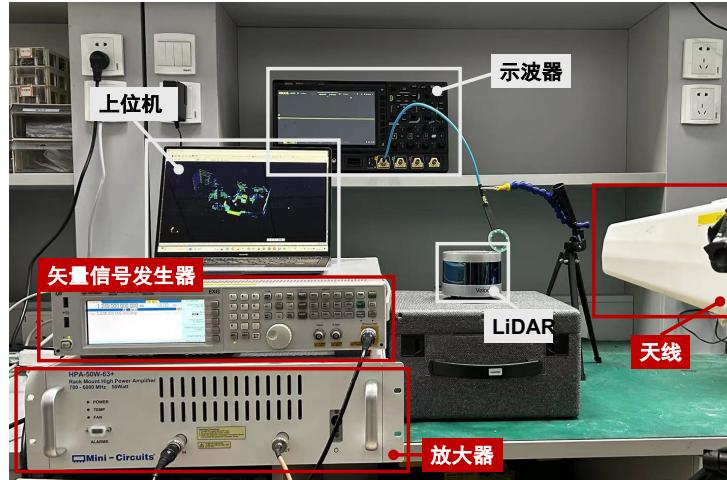


图 3.4 实验设置：攻击设备包括矢量信号发生器、功率放大器和对数周期天线

和用于信号传输的对数周期天线。被测激光雷达为 VLP-16<sup>[75]</sup>，该型号是激光雷达安全研究相关工作中最常用的激光雷达<sup>[24,30,91,93]</sup>。

频率扫描的范围为 500MHz 至 3500MHz，间隔为 1MHz。信号发生器输出设置为 0dbm，放大器增益为 50W。频率扫描的视频演示可在网站<sup>[90]</sup>上找到。实验过程中，本文记录点云并观察激光雷达在不同频率电磁干扰下的运行状态。本文采用了“定量分析”和“攻击效果分析”两种方法来展示频率扫描的结果。

### 3.4.2.2 定量分析

在定量分析中，本文利用真实点云和受干扰点云之间的 *Hausdorff* 距离这一参数来量化 EMI 下点云的失真程度。在数学中，豪斯多夫距离衡量度量空间中两个子集之间的距离。考虑真实点云  $\mathbb{P}C$  和受干扰点云  $\mathbb{P}C'$ ，它们之间的 Hausdorff 距离表示为  $D_H(\mathbb{P}C, \mathbb{P}C')$ 。 $D_H(\mathbb{P}C, \mathbb{P}C')$  的值越大，表明受干扰点云和真实点云的差距越大，即 EMI 对激光雷达系统的干扰程度越强。在扫频过程中，除了电磁信号的频率变化外，本文保持实验室环境的一致性。在频率扫描过程中记录点云后，本文计算从 500MHz 到 3500MHz 各频率电磁信号干扰下的点云与良性点云之间的 Hausdorff 距离。结果如图3.5a 所示，显示出许多不同频率的点云都会产生明显的干扰效应。例如，在 1200 MHz 左右的频率下，豪斯多夫距离高达 120 米。观察 1200 MHz 电磁干扰下的点云，本文发现所有的点都被擦除了。综上所述，频率扫描的定量分析验证了 EMI 攻击激光雷达的可行性，然而要系统地研究这些攻击的原理，本文还需要关注攻击的效果。

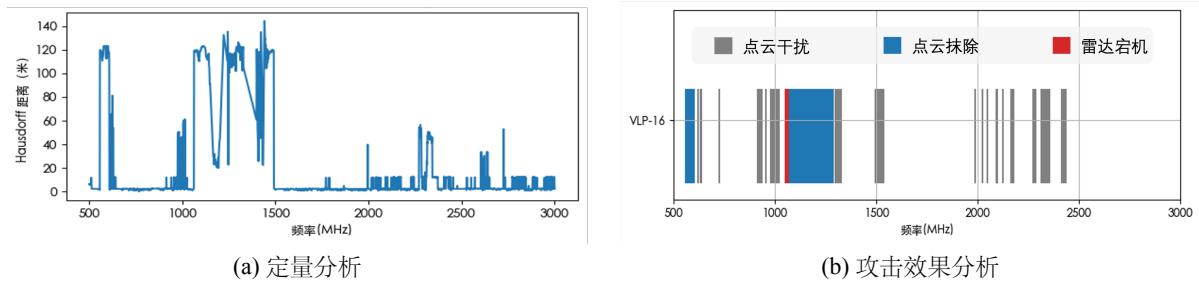


图 3.5 可行性实验：扫频分析

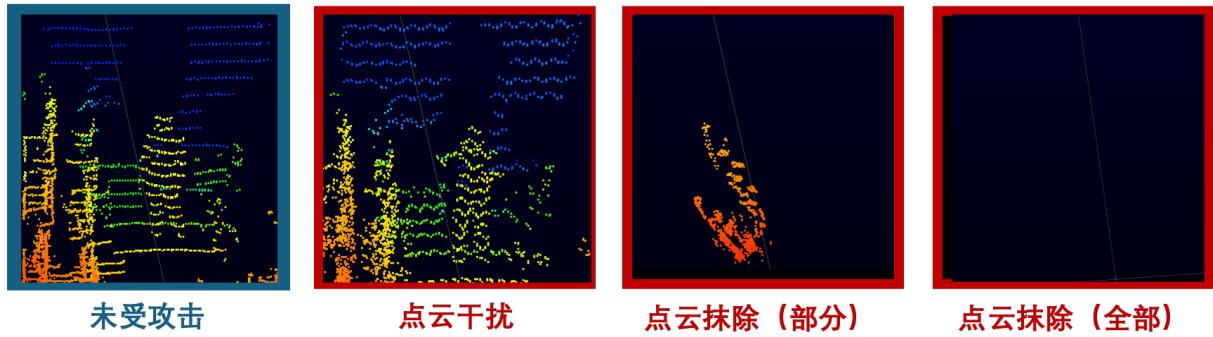
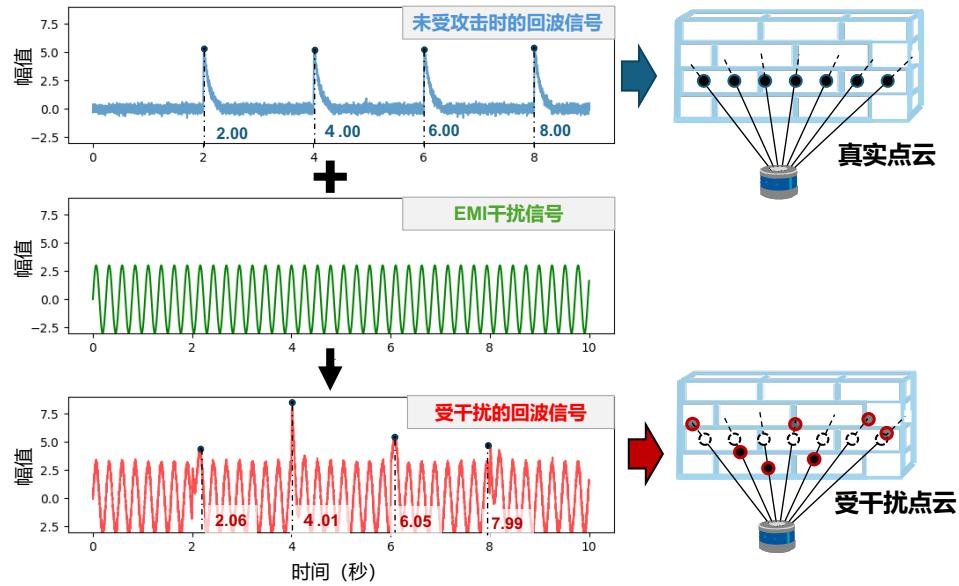


图 3.6 点云干扰、点云抹除的攻击效果示意图

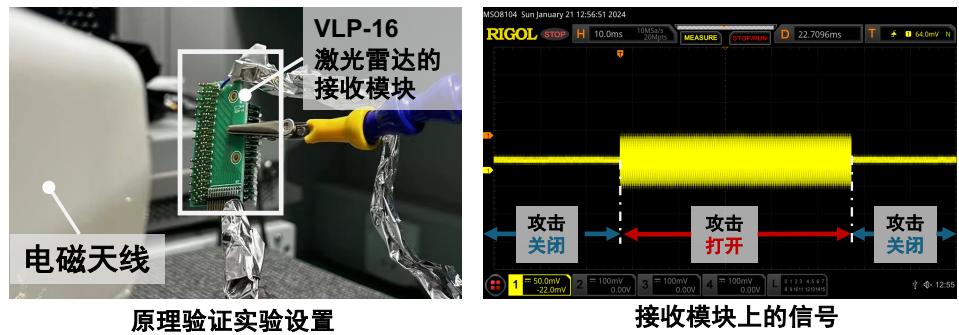
### 3.4.2.3 攻击效果分析

从定量分析中可以发现，不同频率的信号会对激光雷达点云造成不同程度的干扰。为了系统分析攻击原理，本文根据电磁干扰下点云的干扰程度和激光雷达运行状态的变化进行分类。攻击效果直观地分为：点云干扰、点云抹除和雷达宕机。针对 VLP-16 激光雷达的攻击效果和 EMI 频率的关系如图3.6所示。

攻击效果的分类标准如下：在静态环境中，本文测量攻击前后同一条射线上各点之间的欧氏距离。如果平均欧氏距离小于 2cm，本文认为激光雷达不受电磁干扰（EMI）影响，因为激光雷达的固有测距精度为 2cm。如果平均欧式距离大于 2cm，但小于 1m，本文将这种类型的攻击效应定义为“点云干扰”。如果平均欧式距离超过 1m，这意味着这些点已经明显偏离了它们原来的位置，因此本文将这种效果归为“点云抹除”，点云抹除既可以影响部分点云，也可以影响全部点云。此外，本文还发现当电磁频率扫频到 1040 MHz 到 1070 MHz 时，VLP-16 激光雷达系统会完全宕机，即使在电磁攻击停止后也无法恢复，系统必须重新启动才能恢复运行，这类攻击效果被归为“雷达宕机”。



(a) 点云干扰原理：正弦干扰信号通过 EMI 的形式注入接收电路，干扰信号与回波信号叠加后，会改变回波信号的峰值，从而影响激光雷达的测距。



(b) 耦合通道验证：实验证明，EMI 信号能够耦合到接收模块的模拟线路中。

图 3.7 点云干扰实验原理分析

### 3.4.3 原理分析及验证

本节分析上述三种攻击效果的原理并进行耦合通道的可行性实验验证。

#### 3.4.3.1 “点云干扰” 原理分析及验证

**原理分析：**“点云干扰” 攻击通过干扰接收模块中的模拟信号，直接影响激光雷达的测距机制。正如在第3.2.1章中介绍的那样，激光雷达的工作信号通常是一个激光脉冲，通过记录发射激光脉冲的时刻  $\tau_0$  和返回激光脉冲的时刻  $\tau_1$ ，可以算出激光飞行时间，进而计算出物体的距离。点云干扰就是通过干扰激光回波的时刻  $\tau_1$  来干扰测距。如图3.7a所示，EMI 可以通过将干扰信号耦合到换能器和放大器之间的导线中，从而将噪

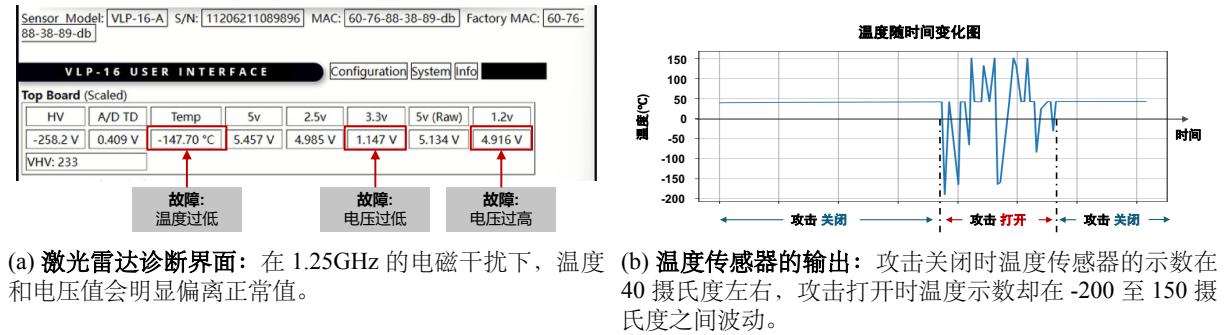


图 3.8 点云抹除实验原理验证

声引入到回波信号中，干扰信号与回波信号叠加后，会改变回波信号的峰值，从而影响激光雷达的测距。

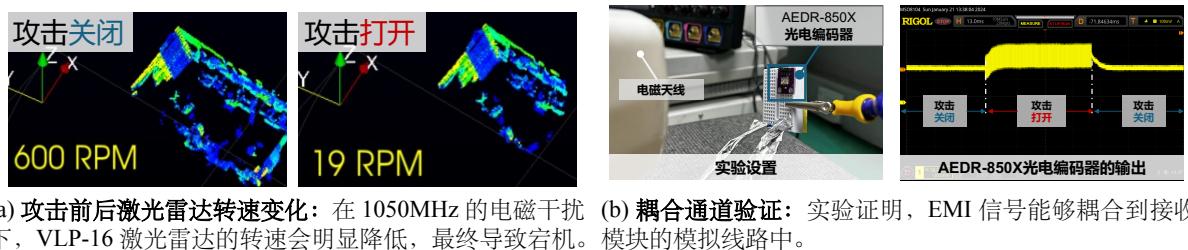
**耦合通道验证：**为了验证在点云干扰攻击过程中信号是否耦合到了激光雷达的接收模块中，本文拆解激光雷达提取了其接收模块，并通过示波器来观察接收模块上模拟线路的信号。然后，本文向接收模块发射了一个 990MHz 的电磁信号，这个频率能够造成点云干扰。本文观察到，如图 3.7b 所示，电磁干扰确实耦合到了接收模块的传输线上。

### 3.4.3.2 “点云抹除” 原理分析及验证

**原理分析：**本文认为有以下两个原理能够实现点云抹除的攻击效果。

- 攻击原理 1(直接攻击): 向接收模块中注入高强度的电磁信号，可能使接收电路饱和，从而使真正的激光脉冲回波无法被探测到，进而实现点云抹除的效果。
- 攻击原理 2(间接攻击): 攻击温度传感器或霍尔效应传感器，使得温度或电压出错，诱使激光雷达的 FDD 机制检测到 L1 故障，从而导致激光雷达将部分或全部点云视为无效点。

**原理验证：**本文通过“假说演绎法”来验证这两个攻击原理。如果原理 1（直接攻击）是正确的，那么在一定信号强度下观察点云抹除，并逐渐降低信号强度时，由于耦合到接收电路的干扰信号幅值不断减小，本文应该观察到被移除的点云逐渐恢复。如果原理 2 是正确的，那么当信号强度逐渐减小到某个值时，就会观察到点云突然恢复，因为此时 FDD 机制不再检测到错误。实验证明，在某些频率（如 1.1 GHz）下，本文确实可以观察到被抹除的点云逐渐恢复，这证实了原则 1 在该频率下是正确的。然而，在某



(a) 攻击前后激光雷达转速变化：在 1050MHz 的电磁干扰下，VLP-16 激光雷达的转速会明显降低，最终导致宕机。模块的模拟线路中。

图 3.9 雷达宕机实验原理验证

些频率下，如 1.2 GHz，随着电磁幅值的减小，本文观察到点云突然恢复，这表明在该频率下，原理 2 是正确的。

除了假说演绎法，本文还通过故障诊断界面直接观察监测传感器的读数进一步验证原理 2。如图 3.8a 所示，通过 VLP-16 的故障诊断界面可以查看诊断信息。本文发现当电磁干扰导致点被抹除时，诊断界面中的温度和电压会出现异常。如图 3.8a 所示，攻击关闭时温度传感器的示数在 40 摄氏度左右，攻击打开时温度示数却在 -200 至 150 摄氏度之间波动，这一现象验证了原理 2 的分析。

### 3.4.3.3 “雷达宕机”原理分析及验证

**原理分析：**值得注意的是，当电磁干扰导致激光雷达宕机后，本文停止攻击并重新启动激光雷达，它仍能正常运行，因此可以排除通过损坏硬件来导致激光雷达宕机的情况。本文利用 Wireshark 记录了从电磁攻击开始到激光雷达关闭期间 VLP-16 的通信数据。本文发现激光雷达的转速出现了严重异常。如图 3.9a 所示，尽管预设转速为每分钟 600 转 (RPM)，但攻击时转速会下降到 19 RPM，降幅高达 96.7%。根据第 3.2.2 章的分析，转速的异常会导致 FDD 机制检测出 L2 故障并进入停机状态。本文进一步研究电磁信号干扰了光束转向模块的哪个部分，即攻击入口。第一种可能是信号真的干扰了电机，使得电机转速异常，但本文通过光电传感器检测激光雷达信号发现，当攻击打开时，在激光雷达真正宕机前，即使交互界面显示转速为 19RPM，但光电传感器实测电机转速依然是 600RPM 左右。因此本文认为，攻击入口是光束偏转模块中的转速监测传感器。综上，雷达宕机的攻击原理是间接攻击，激光雷达检测到了 L2 级故障，即严重影响雷达工作甚至有安全风险的故障，进而触发了 FDD 的自保护机制，切断了电源。具体来说，本文认为雷达宕机是通过干扰激光雷达的光束偏转模块中的电机监测传感器触发的。

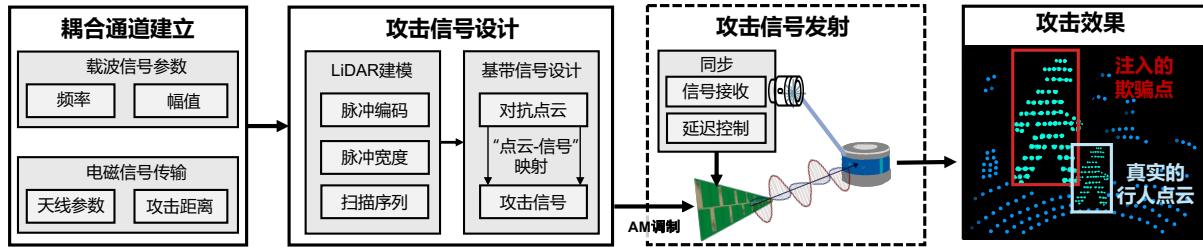


图 3.10 点云注入攻击流程

**耦合通道验证：**如图3.9b所示，在拆解激光雷达后，本文发现 VLP-16 的光束转向模块是一个由光学编码器 AEDR-850X 监控的旋转电机。本文用能够使雷达宕机的 1050 MHz 的正弦电磁信号攻击了 AEDR-850X，实验结果表明，电磁信号会干扰光学编码器的读数，这会导致激光雷达错误地检测到异常低的电机转速，从而导致宕机。

### 3.5 点云注入攻击设计

本节重点讨论如何利用 EMI 注入并精确操纵欺骗点。要实施可控的点云注入攻击，解决以下研究问题至关重要：

- 研究问题 1：如何建立电磁耦合通道用于注入欺骗点？
- 研究问题 2：如何基于拟注入的欺骗点设计攻击信号？
- 研究问题 3：如何精准操控欺骗点的位置？

对于研究问题 1，精确操纵点云的最直接方法是向飞行时间（ToF）电路注入电磁信号，而不是通过故障管理机制来影响点云。因此，可以利用第3.4.1章中的直接攻击原理，即将接收模块作为攻击入口，操纵激光雷达的回波信号来控制点。此外，本文还研究了影响电磁注入效率的因素，包括载波信号的频率和幅值以及不同攻击距离下的天线类型。本文的目标是找出这些因素的最佳组合条件，从而实现最大效率的电磁注入。

对于研究问题 2，本文采用了第2章中的“点云-信号”映射方法基于拟注入的欺骗点设计攻击信号，通过伪造激光信号回波的方式来注入假点。与激光攻击不同的是，本文采用信号调幅（AM）技术，将设计的激光脉冲信号作为基带信号，调制到合适频率的正弦载波信号上。

对于研究问题 3，本文在图3.4所示实验设置的基础上，增加了光电传感器和延迟控制装置，首先利用光电传感器检测激光雷达系统的运行状态，然后设置精确的延时来控

制电磁信号的发射时机，最终实现对欺骗点位置的控制。

点云注入攻击的流程如图3.10所示。首先，攻击者建立电磁耦合通道，确定载波信号和信号传输的最佳参数，两者结合实现高效的电磁注入。然后，攻击者进行基带信号的设计，基于对激光雷达参数的建模，能够通过“点云-信号”映射将拟注入的对抗点云转化为基带信号。最后，攻击者进行攻击信号发射，基带信号通过AM调制的方式被调制到载波信号上从而生成攻击信号，在完成基于信号接收和延迟控制的同步后，将攻击信号以电磁干扰的形式发出，从而实现点云注入。

接下来对“耦合通道建立”、“基带信号设计”和“攻击信号发射”分别作详细介绍。

### 3.5.1 电磁耦合通道建立

为了使电磁有效耦合到接收模块的模拟电路中，本文选用利用正弦信号作为载波。电磁注入的强度越强，伪造的回波信号就越有可能被认为是有效的回波，从而提高欺骗点注入的成功率。因此，在攻击设备电磁发射功率有限的情况下，需要考虑如何提高电磁注入效率，本文从电磁干扰源（即载波信号的参数）和耦合路径（即信号传输的方式）两个层面考虑提高电磁注入效率。

#### 3.5.1.1 载波信号参数

**频率：**在进行EMI攻击时，攻击者将目标激光雷达中的电线视为接收天线。有效的电磁信号频率可以根据接收天线的电气长度来估算<sup>[119-120]</sup>。根据经验，如果目标天线的长度为 $l$ ，则电磁信号的最佳耦合频率在 $\frac{c}{50l}$ 和 $\frac{c}{2l}$ 之间，其中 $c$ 是光速。具体的最佳耦合频率还取决于目标天线的形状、材料和其他特性。在现实攻击场景中，攻击者可以利用上述理论建立近似频率范围，然后通过频率扫描确定耦合效率较高的谐振频率。

**幅值：**信号幅值是电磁干扰的关键因素，一般认为信号幅值越大，干扰越强。在实际实验中，可能会出现增加设备显示的幅值并不会增强干扰的情况。例如，在本文的实验中，信号发生器的最大幅值为19dBm(80mW)，放大器为50W。然而，本文观察到，改变信号发生器的幅值（例如从10dBm变为19dBm）有时并不会影响最终输出。这是由于放大器的饱和特性限制了信号的最大输出强度。另外，也需要注意幅值不能过大，否则可能会引起目标电路上信号的饱和失真，从而难以注入点云。

### 3.5.1.2 电磁信号传输

首先，天线的频率范围必须覆盖扫描范围。其次，天线最好具有尽可能高的增益，以减少信号衰减。在实验中，为了适应不同的攻击距离，本文选择了两种类型的发射天线。在攻击距离小于 10 厘米的情况下，本文选择近场探头来产生电磁信号，因为它更便于携带。对于大于 10 厘米的攻击距离，本文选择对数周期天线，因为它具有更好的方向性。

## 3.5.2 攻击信号设计

攻击者通过伪造激光雷达信号回波来注入假点。首先对激光雷达系统进行参数建模，然后根据拟注入的点设计基带信号，最后将基带信号调制到正弦载波信号上实现攻击信号设计。

### 3.5.2.1 激光雷达建模

在正式设计信号前，攻击者需要获取目标激光雷达的必要信息。首先，攻击者必须获取其激光雷达工作信号的波形，通常由一个或多个脉冲组成。例如，VLP-16 的信号是一个宽度约为  $10 \text{ ns}$  的脉冲。然后，攻击者需要确定激光雷达的扫描周期。例如，如图2.2所示，VLP-16 的工作周期为  $55.296 \mu\text{s}$ ，在此期间，16 线激光以  $2.304 \mu\text{s}$  的间隔按特定顺序发射和接收，随后是  $18.432 \mu\text{s}$  的充电周期。以上参数通常可以通过查阅数据手册或用光电传感器接收并解析激光雷达的信号来实现。

### 3.5.2.2 基带信号设计

设计基带信号的方法采用了第2章中的攻击方法，本文称其为“点云-信号”映射方法，即将目标点云转换为一系列脉冲，其中每个脉冲对应一个欺骗点，每个脉冲峰值的时间定义了欺骗点的空间坐标。然而，与激光攻击中信号可轻易注入到激光雷达中不同，脉冲形电磁波面临难以耦合到攻击入口的挑战。因此，本文通过 AM 调制的方式将设计的激光脉冲信号作为基带信号，调制到合适频率的正弦载波信号上。

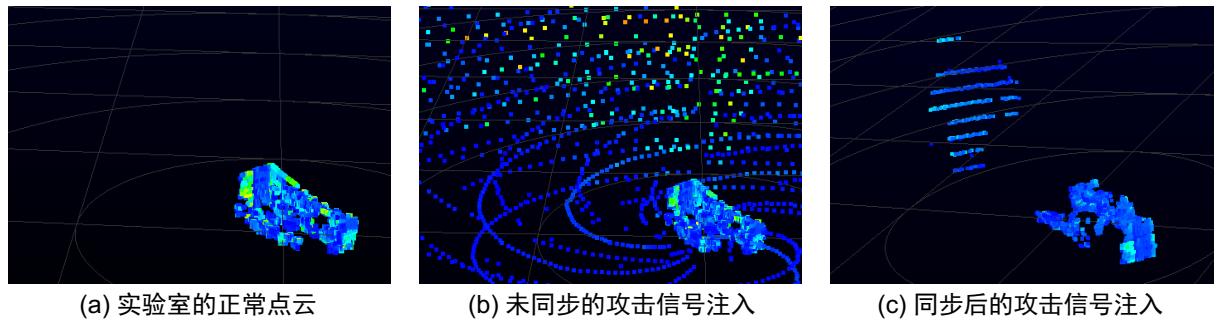


图 3.11 点云注入攻击中信号同步的重要性

### 3.5.3 攻击信号发射

若将利用上述方法设计的攻击信号直接发射，本文可以向激光雷达注入欺骗点，然而，如图3.11（b）所示，这些欺骗点是无序的，无法稳定在一个固定的形状和位置上。要实现可控注入，就必须使信号与激光雷达的扫描序列同步。受之前基于激光的攻击的启发，本文引入了一种集成了“接收-延迟-发射”的闭环反馈机制来实现点云控制，即利用光电传感器来检测激光雷达的工作周期，并配合信号发生器中的延迟控制功能来实现同步。如图3.11（c）所示，进行信号同步后，本文可以控制欺骗点的形状和位置。

## 3.6 实验评估

本节针对四种攻击类型评估 PhantomLiDAR 的有效性。

### 3.6.1 实验概述

#### 3.6.1.1 评估方法

由于每种攻击效果具有不同的特点，本文为每种攻击设计了独特的评估方法，以便更好地了解它们的优势和局限性。

第3.6.2章评估了攻击在不同激光雷达型号上的影响，重点讨论了点云干扰、点云抹除和雷达宕机三类攻击，因为这三类攻击代表了三种不同的攻击入口和脆弱性。本文评估了这些攻击在五种激光雷达上的有效性，发现不同的激光雷达具有不同的脆弱频点，对 EMI 的鲁棒性也不同。该实验可为未来的激光雷达设计提供启示，或有助于建立新的电磁兼容性（EMC）测试标准。

第3.6.3章评估了“点云干扰”攻击在 5 个 3D 目标检测模型上的攻击效果。本文基

于从真实世界中收集的 KITTI<sup>[121]</sup> 数据集合成了点云干扰数据集。通过对 2 个基于激光雷达的模型和 3 个基于融合的模型进行实验，本文观察到点云干扰攻击会导致三维物体检测模型的性能下降，同时发现传感器融合有助于防护点云干扰攻击。

第3.6.4章在真实世界评估了“点云抹除”攻击。由于点云抹除攻击的强大攻击效果，它可以在模型层面实现黑盒的“隐藏”攻击，即使指定的物体无法被模型检测到。以隐藏攻击为目标，本文首先评估了攻击者的距离和角度对攻击的影响；此外，本文还评估了攻击对瞄准精度的要求。上述评估突出了电磁攻击和激光攻击相比的优势和不足。

第3.6.5章针对“雷达宕机”评估了攻击距离的影响，并讨论了其在现实场景中的潜在威胁。

第3.6.6章验证了“点云注入”的可行性，从注入点的最大数量以及对点的精确控制能力来评估了攻击的效果。

第3.6.7章探究了在目标激光雷达处于运动状态时的电磁攻击可行性，攻击目标是隐藏激光雷达前方的指定物体。

### 3.6.1.2 攻击设备

本节中的所有实验都使用了共同的攻击设备，包括用于生成 EMI 信号的 Keysight N5712b 矢量信号发生器<sup>[122]</sup>、用于 EMI 信号放大的 Mini-Circuits HPA-50W-63+ 功率放大器<sup>[123]</sup>，以及用于远程信号传输的对数周期天线<sup>[124]</sup>，频率范围为 600MHz 至 6000MHz，增益为 15 dBi。“点云注入”攻击还包含额外的设备用于信号同步，详见第 3.6.6 章。

## 3.6.2 攻击不同型号激光雷达

**被测激光雷达：**本文对五款商用激光雷达进行了电磁干扰攻击评估，包括三款机械式激光雷达（VLP-16<sup>[75]</sup>、RS-16<sup>[76]</sup>、RS-Bpearl<sup>[125]</sup>）和两款 MEMS 激光雷达（RS-M1<sup>[77]</sup>、RS-M1P<sup>[126]</sup>）。这些激光雷达被广泛部署在目前已发布的智能车<sup>[127-130]</sup>和机器人<sup>[131]</sup>上。

**实验设置：**本文将天线放在距离激光雷达 30 厘米的地方，对这些激光雷达进行扫频测试，频率从 500MHz 到 3500MHz，步长为 5MHz。信号发生器的输出幅值为 -10dBm，放大器的增益为 56dB。

**评估结果：**攻击效果与电磁干扰频率之间的关系如图 3.13 所示，可以观察到不同型号的激光雷达具有不同的脆弱频点和攻击效果。扫频过程中，本文没有观察到 RS-Bpearl



图 3.12 被测激光雷达

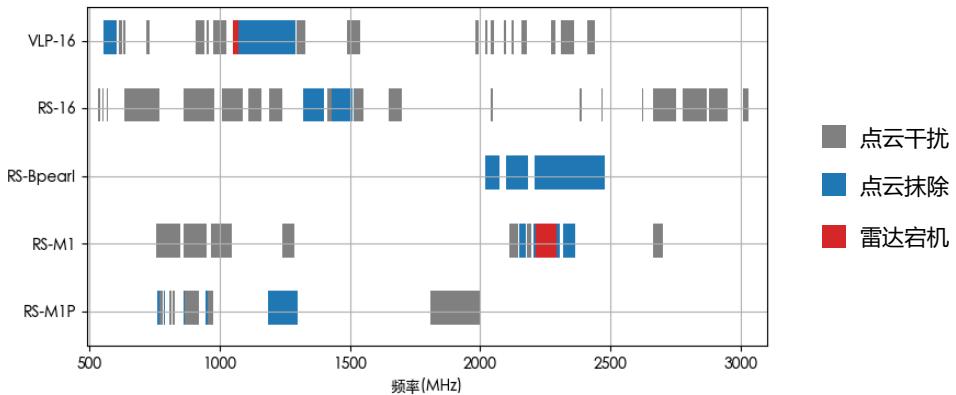


图 3.13 针对不同型号激光雷达的扫频测试结果

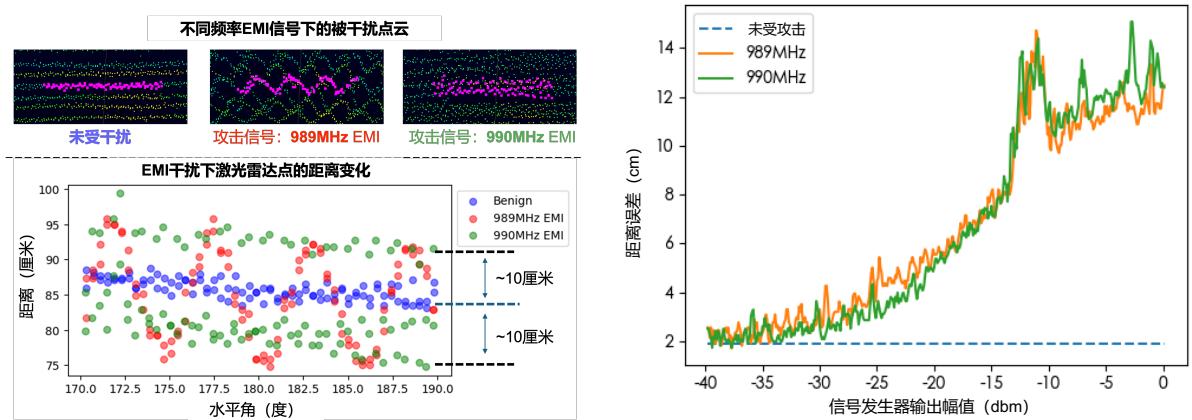
出现“点云干扰”的攻击现象，说明其接收模块的电磁防护能力比其他激光雷达更好。在所有激光雷达上都观察到了“点云抹除”的攻击现象，但值得注意的是，在 VLP-16 和 RS-16 上，电磁干扰可以抹除所有点云，然而在其他三个激光雷达模型上，本文只观察到部分点云被抹除，本文推测这是由于 VLP-16 和 RS-16 的外壳玻璃材质占比较大，导致对电磁的屏蔽不足，而其他三个激光雷达模型的外壳主要材质为金属，能够更有效屏蔽电磁信号干扰。至于“雷达宕机”攻击，只在较老的型号 VLP-16 和 RS-M1 上观察到，这表明新一代激光雷达已经优化了 L2 级严重故障的诊断和管理。

### 3.6.3 点云干扰

#### 3.6.3.1 点云干扰强度分析

如第3.4.3.1章所述，点云干扰攻击会给激光雷达的测距引入误差，本节首先利用“距离误差”来量化点云干扰的强度，然后通过不同幅值的干扰信号研究信号幅值对点云干扰强度的影响。

距离误差  $\epsilon$  定义如下：对于一个激光雷达点  $(r, \theta, \varphi)$ ，点云干扰能够在距离轴  $r$  上引入的最大误差。图3.14a是距离误差  $\epsilon = 10\text{cm}$  的示意图，在实验过程中，本文也注意



(a) 点云干扰强度指标——距离误差：图中展示了当距离误差为 10cm 时的点云干扰效果，且由于混叠效应，不同频率的 EMI 能够形成不同形状的干扰。

(b) 点云干扰距离误差和信号幅值的关系：随着信号幅值的增大，距离误差逐渐增大，最大可达约 15cm。

图 3.14 点云干扰强度（距离误差）实验评估

到，随着电磁干扰频率的变化，会注入不同的点云干扰，比如 989MHz 的 EMI 会注入正弦干扰，而 990MHz 的 EMI 会注入随机干扰，这主要是由于激光雷达 ADC 采样率不足引起的混叠效应。

为了系统地探索攻击信号幅值对点云干扰强度的影响，本文在 VLP-16 激光雷达上进行了进一步的实验。首先，本文选择了两个典型频率：989MHz 和 990MHz。然后，本文将信号发生器的振幅从 -40 dBm 提高到 0 dBm，并将其连接到增益为 56 dB、最大输出功率为 50 W (47 dBm) 的放大器上。通过记录不同振幅下的距离误差，本文得到了如图 3.14b 所示的结果。可以看出，随着信号幅值的增大，距离误差逐渐增大，而且不同频率下的趋势一致。当信号发生器的振幅达到约 -10 dBm 时，距离误差最多可达 15 cm 左右。然而，尽管信号发生器的输出进一步增加，距离误差并没有显著增加。这主要是因为放大器的输出达到了饱和，妨碍了信号发生器对最终信号振幅的精确控制。此外，放大器造成的饱和失真导致 -10 dBm 之后的信号失真，这也是距离误差在 -10 dBm 附近减小的原因。

### 3.6.3.2 模型层面攻击效果评估

为了系统地评估点云干扰攻击对 3D 目标检测模型的影响，本文在真实采集的大规模数据集上模拟了不同强度不同干扰模式的攻击效果，然后将受干扰数据集输入模型。

**数据集生成：**攻击数据集基于 KITTI 数据集<sup>[121]</sup>生成，KITTI 数据集包含多种真实

干扰模式	距离误差	LiDAR-only		LiDAR-Camera Fusion		
		Point Pillar	PV-RCNN	VirConv-L	EPNet	CLOCs
无干扰	/	77.62	83.66	86.82	78.83	76.89
随机干扰	5cm	77.07	79.10	86.01	78.27	76.27
	10cm	72.59	77.07	85.65	76.75	71.92
	15cm	60.18	63.93	79.79	71.10	62.85
正弦干扰	5cm	76.97	79.87	85.09	78.25	76.26
	10cm	71.24	75.79	85.94	75.03	70.63
	15cm	59.48	61.85	75.74	70.68	60.69

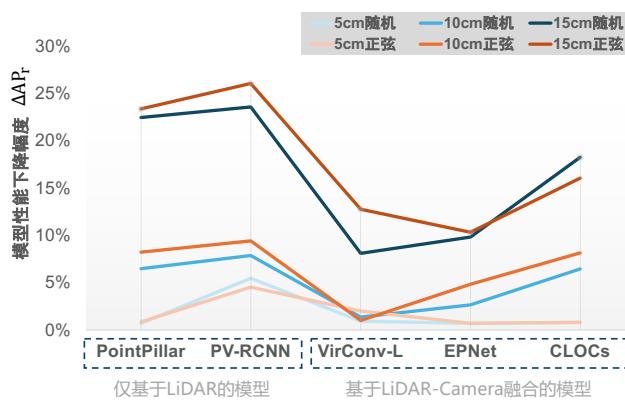


图 3.15 不同距离误差的随机和正弦干扰对模型性能的影响

世界采集的驾驶场景。图3.14a展示了两种点云干扰的模式，即随机噪声和正弦噪声。本文对这两种噪声模式通过下式进行模拟：

$$\left\{ \begin{array}{l} \text{随机噪声: } r'_i = r_i + \text{Random}[-\varepsilon, \varepsilon], \\ \text{正弦噪声: } r'_i = r_i + \varepsilon * \sin(\frac{\theta}{5} * 2\pi), \\ (r_i, \theta_i, \varphi_i) \in \mathbb{PC}, i \in [1, n], \\ (r'_i, \theta_i, \varphi_i) \in \mathbb{PC}', i \in [1, n], \end{array} \right. \quad (3-2)$$

其中， $\mathbb{PC}$  表示良性点云， $\mathbb{PC}'$  表示受干扰点云，电磁干扰只影响  $r$ 。为了探究不同电磁干扰强度和噪声形式的攻击效果，本文分别合成了  $\varepsilon = 5, 10, 15\text{cm}$  时的随机噪声和正弦噪声的受损数据集。

**检测模型:** 本文评估了点云干扰攻击在 5 个 3D 目标检测模型上的效果。其中 2 个是仅基于激光雷达的模型：PointPillar<sup>[8]</sup> 和 PV-RCNN<sup>[132]</sup>；此外，还有 3 个激光雷达-相机融合模型：前融合模型 VirConv-L<sup>[14]</sup>、特征级融合模型 EPNet<sup>[15]</sup> 和结果级融合模型 CLOCs<sup>[17]</sup>。

**评估指标:** 与 KITTI 基准<sup>[89]</sup>一样，当预测检测框和实际检测框之间的交并比（Intersection over Union, IoU）超过 0.7 时，本文认为目标检测成功。基于该 IoU 标准，本文使用“平均精度”（Average Precision, AP）来衡量模型的整体检测性能。AP 是 KITTI 基准的官方指标，用于全面评估模型的性能。为了评估攻击对模型的影响程度，本文引

入了一个新指标—模型性能相对下降幅度  $\Delta AP_r$ :

$$\Delta AP_r = \frac{AP_{benign} - AP_{attack}}{AP_{benign}} \quad (3-3)$$

其中  $AP_{benign}$  和  $AP_{attack}$  分别表示模型在干净数据集和受损数据集下的平均精度。

**评估结果:** 图3.15a列出了5个检测模型在不同强度点云干扰下的AP, 从图中可以看出, 距离误差为5cm时, 模型性能几乎不受影响, 随着干扰强度的增加, 检测模型的AP逐渐降低。15cm的距离误差能使模型平均精度显著下降。各模型的 $\Delta AP_r$ 如图3.15b所示。1)首先本文比较相同干扰强度下的正弦干扰和随机干扰, 本文一共获取了15组 $\Delta AP_r$ 数据, 其中12组数据的正弦干扰 $\Delta AP_r$ 略大于随机干扰 $\Delta AP_r$ , 表明点云干扰的模式会影响攻击效果。2)基于融合的模型的鲁棒性总体上强于基于激光雷达的模型, 这表明传感器融合有潜力成为防御点云干扰攻击的方法。在基于融合的模型中, CLOCs的鲁棒性最低。这是因为CLOCs是一种结果融合方法, 它在应用非最大抑制(NMS)之前将摄像机和激光雷达检测候选对象结合在一起, 这种松散耦合的融合方法无法有效抵御点干扰攻击, 表明传感器融合的方式会影响模型的鲁棒性。

### 3.6.4 点云抹除

点云抹除攻击能够在数据层面使点云完全消失, 这一强大的攻击能力能够很容易使3D目标检测模型检测不到目标物体, 实现隐藏攻击。为了更全面地了解这种攻击在物理世界中的实际威胁, 本节实验评估了攻击距离、攻击角度以及瞄准精度的影响。

#### 3.6.4.1 实验设置

为了测试点云抹除攻击的最远攻击距离。本文在校园封闭道路上进行了物理实验, 实验设置如图3.16所示。攻击设备及型号已在第3.6.1.2详细介绍。拟隐藏的目标为汽车, 这是实际自动驾驶场景中最常见的目标。受害激光雷达为VLP-16, 检测模型为PointPillars<sup>[8]</sup>。

#### 3.6.4.2 攻击者距离和角度的影响

如图3.17a所示, 攻击者的位置包括与激光雷达的距离以及离轴角度。可以成功实施攻击的位置范围越大, 意味着攻击在实际场景中危害性越大且隐蔽性越强。本文从



图 3.16 远距离攻击实验设置

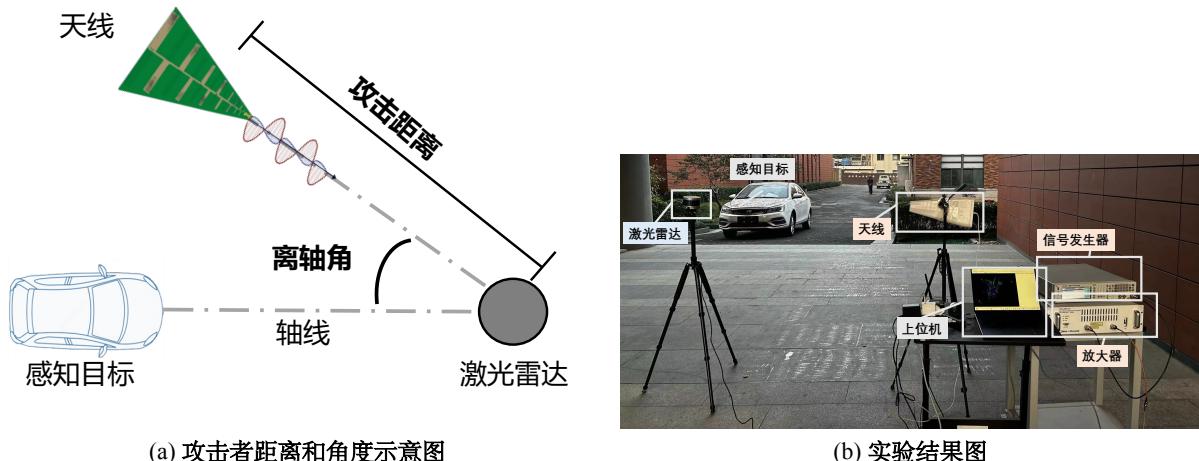


图 3.17 点云抹除攻击距离和角度的影响

[1, 1.5, 2, 2.5, 3, 4] 米的距离和 [5, 15, 30, 45, 60, 90, 145, 180] 度的离轴角对激光雷达进行攻击，结果如图3.17b所示。结果表明，在 1.5 米的距离内，这个距离内所有的激光雷达点云都会被抹去，因此攻击者可以从任何角度隐藏目标物体。此外，本文还发现，离轴角度越小，攻击成功的距离就越远，通过本文的测试设置，当攻击者与激光雷达的距离达到 5.5 米时，攻击仍可以成功实施。

### 3.6.4.3 瞄准精度的影响

瞄准是现实世界中远程攻击所面临的共同挑战，这一挑战在以往的激光攻击中尤为突出<sup>[91]</sup>。本实验研究了电磁攻击的瞄准要求。攻击所需的瞄准精度越低，在实际场景中就越可行。本文保持电磁天线的位置不变（距离激光雷达 1.5 米，离轴角 5 度），然后改变电磁天线指向的方向。然后将点云输入检测模型，观察攻击是否成功。结果如

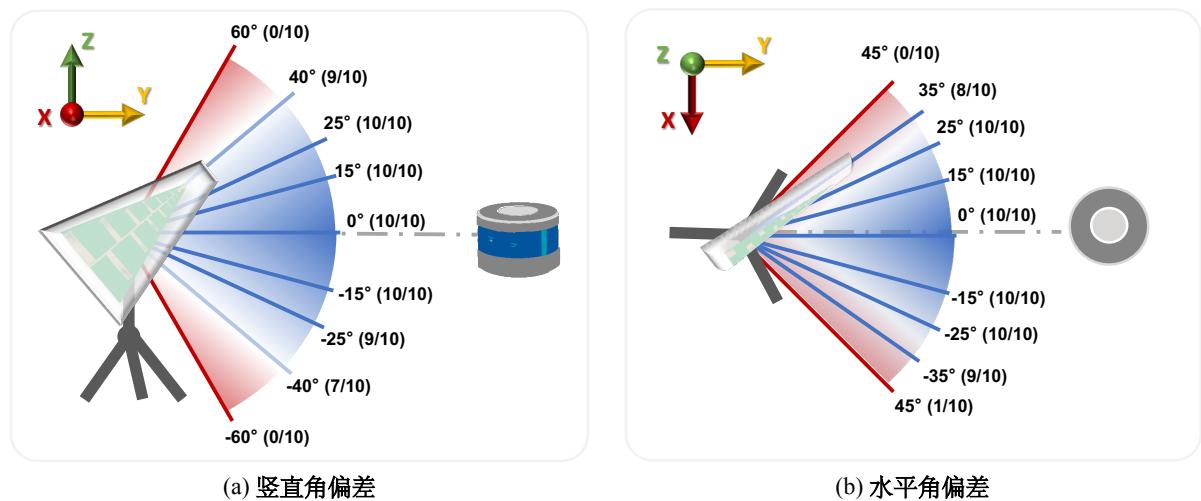


图 3.18 点云抹除攻击瞄准精度的影响

图3.18所示，电磁天线的竖直角偏差可达 $40^\circ$ ，水平角偏差可达 $35^\circ$ ，但仍能达到隐藏攻击的效果。这一结果表明，电磁攻击不需要精确瞄准。

### 3.6.5 雷达宕机

为了探索 textit 激光雷达断电的现实威胁，本文评估了攻击距离对 VLP-16 和 RS-M1 的影响。利用图3.16中所示的攻击设备，本文能够分别在 30 厘米和 50 厘米的距离上诱发针对 VLP-16 和 RS-M1 的雷达宕机。这个距离足以在现实生活中构成威胁，因为这一距离允许攻击者从车外进行攻击。一个可行的攻击场景是，攻击者将放大器和信号发生器放在车内，并通过一根长导线将电磁天线延伸到车外。当自动驾驶汽车停下时，例如在红灯前，攻击者可以将电磁天线靠近自动驾驶汽车的激光雷达，使其断电宕机。然后，即使攻击停止或车辆离开，宕机的效果也会持续。读者可观看网站上<sup>[90]</sup>的演示视频以更直观地了解雷达宕机的攻击效果。

### 3.6.6 点云注入

本文从两方面评估点云注入攻击：首先是注入点的最大数量，这在以往的激光雷达攻击中一直是一个重要指标；其次是对点的控制，具体来说就是能否注入指定形状的欺骗点云。

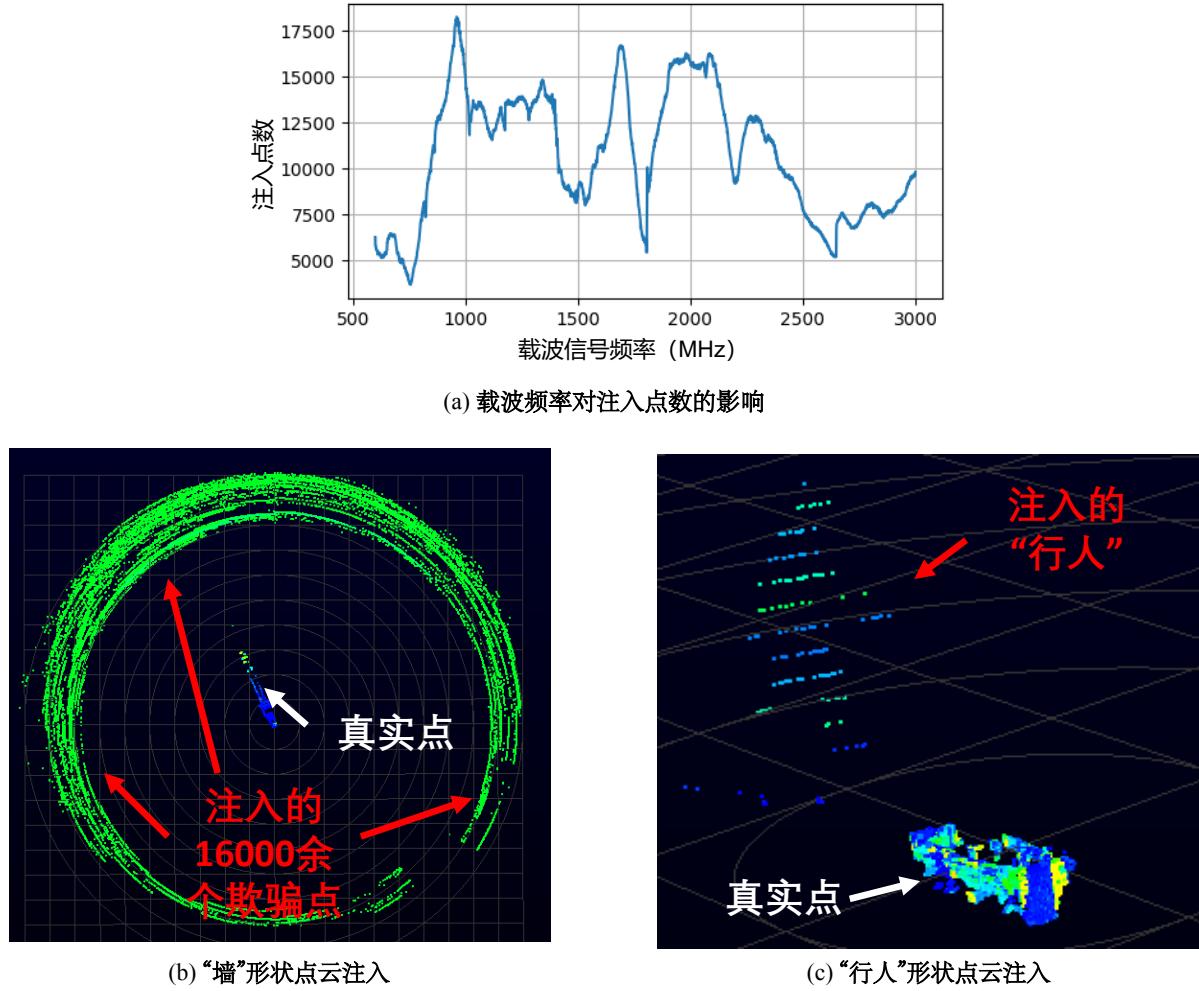


图 3.19 点云注入实验评估

### 3.6.6.1 实验设置

除了矢量信号发生器、放大器、电磁天线之外，点云注入攻击还需要一个任意波形发生器来创建基带信号并设置延迟，以及一个光电探测器来接收来自激光雷达的信号。被攻击激光雷达是 VLP-16 和 RS-16。

### 3.6.6.2 注入点数

首先需要说明的是，只有在被攻击激光雷达型号相同、旋转速度相同的情况下，注入点数量的比较才有意义。因此，本文选择了 VLP-16 作为被攻击激光雷达，这是以往欺骗点注入研究<sup>[24,30,91]</sup>中常用的激光雷达型号。

为了尽可能多地注入欺骗点，本文设计的基带信号理论上能够伪造激光雷达每个接收周期的回波信号。在相同的基带信号下，本文测试了不同载波信号频率对注入点数量



图 3.20 移动攻击实验设置

的影响。本文将信号输出功率设置为 50 W，并在 700 MHz 和 2500 MHz 之间改变载波频率，记录注入假点的数量。结果如图3.19a 所示，不同的载波频率确实会影响注入点的数量，本文将其归因于接收电路对不同信号频率的接收效率不同，更强的接收效率能使得注入的信号幅值更大，从而更有可能使注入的信号被认为是有效回波。值得注意的是，约 1040 MHz 的载波频率能够注入最多的点数（超过 16500 个），在激光雷达周围形成了一个如图3.19b所示圆形的墙状结构。在相同的旋转速度下，以前的工作最多只能注入 3000 个假点<sup>[91,133]</sup>。这种大幅增加主要是因为电磁攻击影响的范围更广（近 360 度水平角），而激光攻击只能影响激光光斑照射的区域（小于 35 度水平角）。

### 3.6.6.3 形状控制

本节验证了可以使用各种基带信号注入不同形状的欺骗点云。图3.19b展示了向 VLP-16 激光雷达中注入一堵“墙”，图3.19c展示了向 VLP-16 中注入一个“行人”，图3.11展示了向 RS-16 中注入一个“行人”。这表明电磁攻击和激光攻击近似，也具备操纵激光雷达点云的能力。

## 3.6.7 移动攻击实验

### 3.6.7.1 实验设置

移动车辆的攻击设置如图3.20所示，攻击设备包括信号发生器、放大器和天线，都集成在攻击车辆内。攻击者在车内手持天线，瞄准移动中的激光雷达。攻击者的目的是

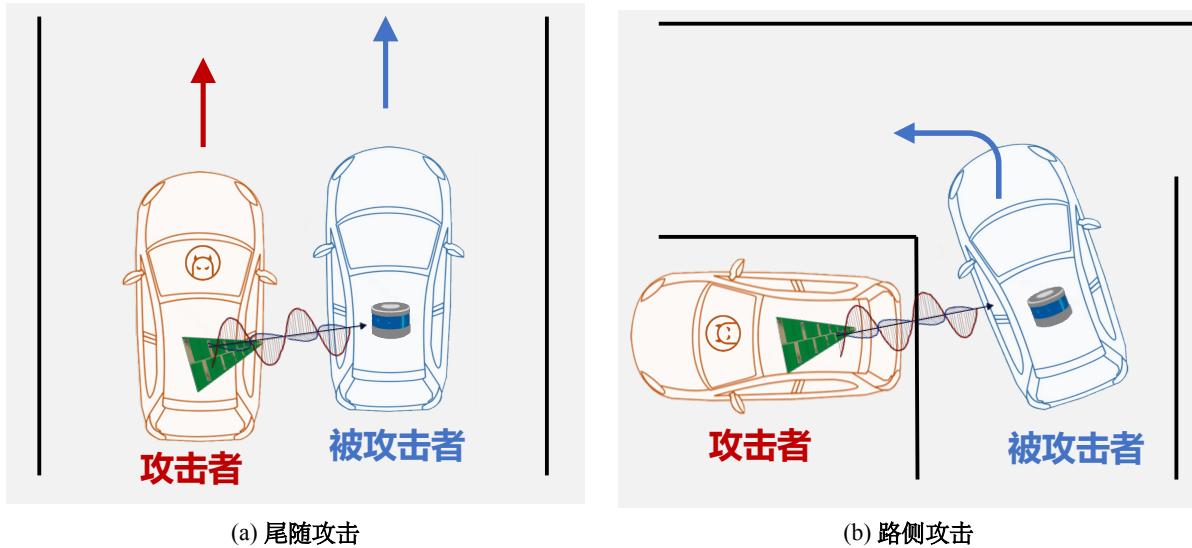


图 3.21 移动攻击场景示意图

通过 EMI 攻击激光雷达，使基于激光雷达的 3D 目标检测模型无法检测到攻击车辆。被攻击车辆配备 VLP-16 激光雷达，并以低于 10 km/h 的速度行驶（出于安全考虑），在实验过程中，本文并不要求受害车辆以恒定的速度行驶，这与现实世界中攻击者无法控制受害车辆速度的情况相符。本文将攻击信号的参数设置为频率为 1.2GHz、输出信号幅度为 50W，该参数可以在远距离实现点云抹除攻击。

### 3.6.7.2 攻击场景

本文定义了两种真实世界中的攻击场景。场景 A 是“尾随攻击”，即攻击车辆和受害车辆都在道路上行驶，攻击者的汽车驾驶员会以相似的速度靠近受害者的汽车。场景 B 是“路侧攻击”，攻击者静止在路边，而受害者汽车正在转弯。与第3.6.4章中静止状态下的攻击相比，场景 A 主要面临不断变化的攻击距离和车辆振动带来的挑战，而场景 B 主要面临不断变化的攻击角度带来的瞄准挑战。

### 3.6.7.3 实验结果

实验证明了 PhantomLiDAR 在两种移动场景中隐藏指定目标的能力。具体来说，对于尾随攻击，本文进行了五组试验，每次车辆均行驶约 40 米的距离。每组实验均匀收集 100 帧点云数据，攻击成功率为 87.2% (436/500)。对于路侧攻击，本文也进行了五组实验，每组实验收集 40 帧点云数据，尽管被攻击汽车的角度在不断变化，但在 4 米

表 3.1 与激光干扰攻击的比较

攻击方式	攻击入口	攻击效果				瞄准要求	攻击距离
		点云干扰	点云抹除	雷达宕机	点云注入		
激光	接收模块中的光电传感器	-	✓ [33,91-92,134]	-	✓ [28,30,72] [31,33,91]	高	>30 米
电磁 (本工作)	接收电路， 监测传感器， 光束偏转模块	✓	✓	✓	✓	低	4 米

的攻击距离内，攻击成功率仍可达到 83.5% (167/200)。综上所述，由于电磁攻击的瞄准要求相对较低，只要攻击者能够接近受害激光雷达（本文中为 4 米内），就可以进行隐藏攻击，攻击成功率在 80% 以上。

## 3.7 讨论

### 3.7.1 与激光干扰攻击的比较

本节对基于电磁的攻击和基于激光的攻击进行比较。需要注意的是，这种比较的目的并不是为了说明电磁或激光哪个更强，因为这种定论需要更严格的基准，包括相同的功率水平或同等的攻击成本。本文的目的是让安全社区及激光雷达相关研究者辩证地理解这两类攻击的特点并提高认识，从而激发更强大的硬件设计和定制化的防御策略。

本文从攻击面、攻击效果、瞄准要求和攻击范围等方面对攻击进行了比较，如表所示。总的来说，PhantomLiDAR 比基于激光的攻击利用了更多的攻击面，实现了更多样的攻击效果，并且具有不需要精确瞄准的优势。然而，得益于激光独特的能量汇聚特性，其在传播过程中能量损耗极小，能够在较长距离上依然保持强大的能量密度，这使得激光攻击在攻击距离方面相较于电磁攻击具有显著的先天优势。

### 3.7.2 攻击成本讨论

在上述实验中，为了使本文能够有效挖掘出激光雷达的电磁干扰脆弱性，需要进行大范围的频率和幅值扫描测试，因此本文利用高成本的设备进行攻击实验，其中

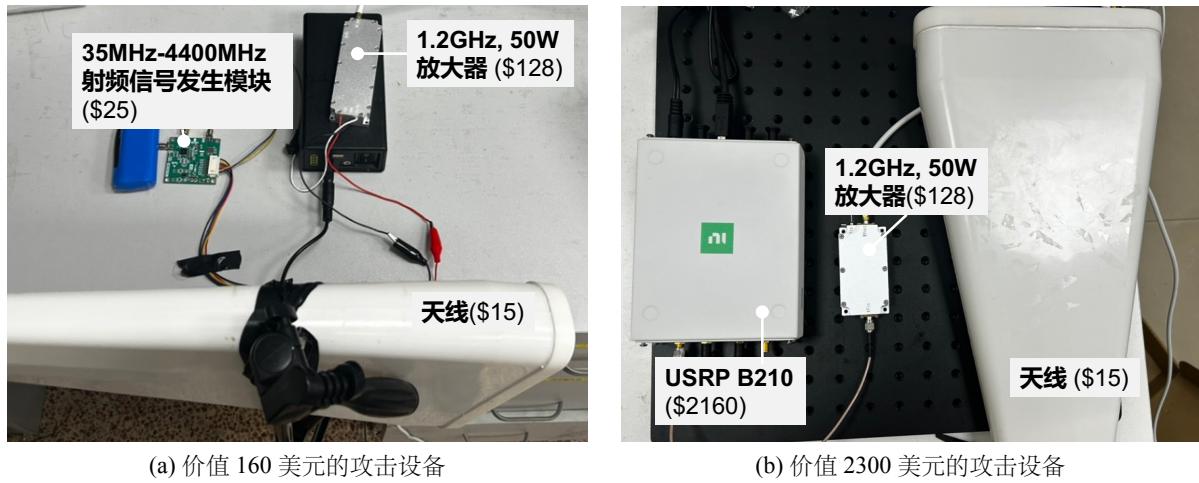


图 3.22 低成本攻击设备

Keysight N5712b 信号发生器约 3.2 万美元，Mini-Circuits HPA-50W-63+ 放大器约 2 万美元。然而在实际应用中，攻击者可以使用成本较低的设备，比如特定频率的信号发生器和放大器，也能达到预期的攻击效果。为了进一步讨论攻击的实际可行性，本文使用成本较低的设备进行了实验。

第一套低成本设备如3.22a所示，包括一个 35MHz-4400MHz 的信号发生模块（25 美元<sup>[135]</sup>）和一个 1080MHz-1360MHz、50W 的放大器（128 美元<sup>[136]</sup>）。这套 168 美元的设备能够在 VLP-16 激光雷达上实现点云干扰、点云抹除和雷达宕机。然而，由于这组设备无法支持 AM 调制，因此无法实现点云注入攻击。

第二套低成本设备如图3.22b所示，包括一个 USRP B210（2160 美元<sup>[137]</sup>）、一个 50W 放大器（128 美元）和一个天线（15 美元）。其中 USRP B210 支持信号幅度调制，这套 2300 美元的设备成功地实现了所有四种类型的攻击效果，包括点云注入攻击。

### 3.7.3 潜在防御方法

对于那些需要在安全关键型场景中使用的激光雷达，可以通过以下方法来应对 PhantomLiDAR 攻击：

**电磁兼容增强：**PhantomLiDAR 的攻击入口是激光雷达的模拟电路部分。有三种常见的防御<sup>[113]</sup>方法可以增强模拟电路的电磁兼容性：屏蔽、差分比较器和滤波器。屏蔽是指使用导电材料对元件进行电磁干扰屏蔽。当屏蔽不可能或不充分时，可使用参考信号，通过差分电路消除共模电压<sup>[138]</sup>。此外，通过滤波器衰减传感器基带频率以外的信

号，也可以降低传感器的易受攻击频率范围。不过，这些方法会增加激光雷达电路的复杂性和成本。虽然商用激光雷达系统在出厂前都要经过严格的电磁兼容性测试，但本文中的实验表明，对手可以利用商用设备轻松入侵激光雷达。因此，考虑更新激光雷达系统的 EMC 标准可能是明智之举。

**多传感器融合：**本文在第3.6.3章中通过实验证明，融合模型有望减轻攻击的影响。就自动驾驶汽车而言，可以通过摄像头和激光雷达的融合，以及装备多个激光雷达来提高安全冗余度，从而增强感知能力。因此，多种传感器如何更好地相互配合、相互补充，是一个值得探索的研究课题。

### 3.8 本章小结

本章研究了激光雷达的新型电磁干扰脆弱性，包括新的攻击入口和攻击原理。本章验证了激光雷达的接收模块、监测传感器（温度传感器）和光束转向模块中的光学编码器可作为电磁干扰的攻击入口。本章确定了两个主要攻击原理：1) 直接攻击：干扰接收模块中的模拟信号，直接影响激光雷达的测距机制；2) 间接攻击：攻击激光雷达中的其他附属模块，进而借助故障检测和管理机制间接诱发点云错误或激光雷达本体故障。基于新的攻击入口和攻击原理，本章提出了 PhantomLiDAR 攻击，包含四种针对激光雷达的电磁攻击效果。本章首次设计并提出了利用电磁干扰的点云抹除、点云注入和雷达宕机。此外，与之前的 SOTA 工作相比，PhantomLiDAR 的攻击能力在干扰强度（增加 3 倍）和伪造点数量（增加 5 倍）方面都有显著提高。在 5 个激光雷达和 5 个目标检测模型上进行的大量实验证明了攻击的有效性。PhantomLiDAR 具有攻击距离远、瞄准要求低以及移动场景可行的特点，证明了攻击的实际威胁。此外，本章还讨论了针对电磁干扰攻击的防御对策。本文希望该研究能够通过考虑更广泛的攻击载体来增强未来激光雷达系统的安全性。

## 4 基于信号注入攻击的多传感器融合感知算法鲁棒性测评基准

自动驾驶作为一种安全关键的应用越来越受到安全研究人员的关注。自动驾驶依赖传感器及后续算法进行感知，然而传感器本身容易受到各种威胁，因为它们暴露在物理环境中，容易受到环境干扰或人为恶意的物理信号影响，本章称其为“信号注入攻击”。为了应对传感器可能出现故障的情况，多传感器融合（Multi Sensor Fusion, MSF）作为一种通用策略被提出来，以增强自动驾驶感知的鲁棒性。但在面对外部信号注入攻击时，多传感器融合方法相比单传感器的方法是否更鲁棒，以及怎么融合更鲁棒等问题一直没有得到关注。本章研究多传感器融合模型在外部信号注入攻击下的鲁棒性问题，建立了基于“点云-图像”融合的鲁棒性测评基准。与传统鲁棒性测评基准数据集关注恶劣天气等数据损坏方式不同，本章在构建数据集时考虑了“信号注入攻击”下的数据损坏，包含了5种针对激光雷达和6种针对摄像头的数据损坏方式。通过在7种多传感器融合目标检测模型和5种单模态目标检测模型上的大量实验评估（542,736组数据），回答了以下研究问题：(1) 融合是否能增强鲁棒性；(2) 融合模型的架构如何影响鲁棒性。最后，本章为融合模型鲁棒性的提升提供了一些建议。为了促进未来融合模型鲁棒性的研究，本章开源了数据集以及代码<sup>[139]</sup>。

### 4.1 本章引言

在自动驾驶中，感知是路径规划、轨迹预测、决策控制的基础，而3D目标检测是感知的核心。激光雷达和摄像头是3D目标检测最重要的两个传感器。激光雷达通过点云数据提供精确的三维空间信息，而摄像头则通过图像数据提供丰富的语义信息。将这两种互补的信息源进行融合是学术界<sup>[11,140-143]</sup>和工业界<sup>[64,66,144-146]</sup>为提高感知性能所做的共同努力。

然而，最近的许多安全研究<sup>[19-21,28,30,91-93,147-154]</sup>证明了激光雷达和摄像头可能会受到激光、电磁和超声波等物理信号的干扰，其中包括攻击者的恶意行为和自然环境的极端干扰。本文采用术语“信号注入攻击”（Signal Injection Attack, SIA）来描述利用物理信号来影响传感器输出的恶意攻击和极端干扰。信号注入攻击会对点云和图像造成严重

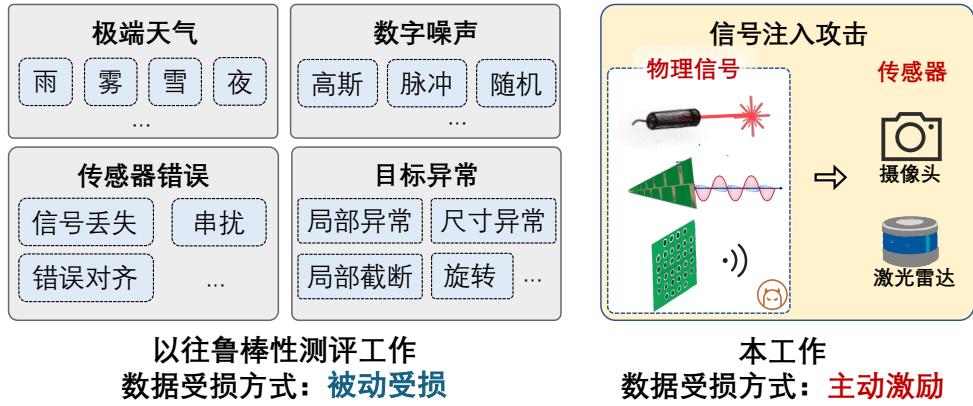


图 4.1 已有工作与本工作数据受损方式区别

程度不同的损坏，自动驾驶作为一种安全关键应用，其功能出错会带来严重的安全风险和财产损失，尤其需要增强对物理世界中可能出现的数据损坏的鲁棒性。之前的许多研究<sup>[19,24,91,154]</sup>都认为传感器融合是一种潜在的对策。然而，融合能否如预期那样削弱攻击仍是一个悬而未决的问题，缺乏系统的研究。

为了有效评估传感器融合在面对信号注入攻击时的鲁棒性，亟需建立一个有效的评测基准<sup>[155]</sup>，该评测基准的数据集往往是一些受损数据。如图4.1所示，以往的工作针对点云损坏<sup>[47,156]</sup>和图像损坏<sup>[42-43,46]</sup>分别提出了多个评测基准，这些基准中的数据受损方式包括恶劣天气、数字噪声、传感器故障、物体异常等。与这些数据受损方式相比，本研究中的数据受损是利用信号注入攻击诱发的。最近，少量基准测试<sup>[50,155]</sup>关注了多传感器融合感知的鲁棒性，然而，它们都没有考虑信号注入攻击所引起的受损，而且所评估的基于多传感器融合的 3D 目标检测模型数量有限，例如，只有 2 个<sup>[50]</sup>或 3 个模型<sup>[155]</sup>在受损数据集上进行了测试。

本研究提出了一个新的多传感器融合鲁棒性评测基准，该评测基准的数据集由信号注入攻击下的受损的点云数据和图像数据构成，能够有效评估自动驾驶中最常见的激光雷达和摄像头融合的 3D 目标检测模型的鲁棒性。基于该基准，本文探究以下研究问题：

**研究问题 1：融合是否增强鲁棒性？**与单模态模型相比，融合模型是否增强了鲁棒性？这是一个基本而关键的问题。虽然许多通过信号注入攻击来攻击传感器的工作<sup>[19-21,28,147]</sup>将多传感器融合认为是防御攻击的方法，但目前没有相关研究通过实验去论证这一假设的正确性。以往的鲁棒性评估工作仅仅在单模态模型上<sup>[42-43,46]</sup>评估，或仅仅评估了少量的融合模型<sup>[50,155]</sup>，还没有工作系统地研究在面对相同数据损坏（尤其是信号注入攻击）时，多传感器融合模型和单传感器模型的鲁棒性性能差异。本研究通过评

估目标攻击鲁棒性、单一来源鲁棒性和整体鲁棒性三个方面回答了这一问题。

**研究问题 2：融合模型的架构如何影响鲁棒性？**面对信号注入攻击，不同架构的基于多传感器融合的探测器是否表现出性能差异？如果存在这种差异，其根本原因是什么？以往的研究通常仅将模型分为前融合、中融合和后融合。然而，本文发现这种分类方法并不能明确揭示架构与鲁棒性之间的关系。本文引入了一种新的分类范式，从“融合顺序（平行、级联）”和“融合表征（数据、特征、结果）”两个维度对模型进行分类，并利用信息熵的概念深入探讨了融合架构与鲁棒性之间的关系。通过这种新的分类方法和大量实验，本工作发现了融合架构和鲁棒性之间的显式关系。

建立一个基准来回答上述问题具有很大的挑战性，1) 首先需要考虑的是数据集的物理可行性，与允许任意编辑图像和点云的纯数字受损不同，要使数据集在物理上可实现，需要考虑信号注入攻击的能力。然而以往的信号注入攻击的工作主要集中在展示其攻击效果上，大部分工作没有明确量化其数据损坏能力以用于基准测试。本工作通过物理复现的方式来量化其数据损坏能力。2) 其次是数据集的完备性，为了全面地考虑信号注入攻击对传感器（激光雷达、摄像头）以及后续目标检测模型的影响，需要保证信号注入攻击的完备性，本工作通过设计系统文献调研的方式从 1174 篇文献中筛选出了 11 种基于信号注入攻击的数据受损方式。3) 最后是有效的实验评估，现有工作缺少对融合模型的有效分类方法和鲁棒性相关测估方法，本工作提出了基于攻击成功率、单源鲁棒性、多源鲁棒性的三维评价指标，基于融合顺序和融合表征的二维分类范式以及基于信息熵的融合模型鲁棒性分析方法，并结合大量（542,736 组数据）的实验评估实现了对上述研究问题的探究。

多传感器融合鲁棒性测评基准的设计与评估流程如图4.2 所示。本工作的贡献总结如下：

- **新测评基准。**提出了一个基于信号注入攻击的多传感器融合鲁棒性测评基准，包含新数据集和新指标。其中数据集 SIA-KITTI 包含由激光、电磁干扰和声波等物理信号引起的 11 种受损数据。新指标包含基于攻击成功率、单源鲁棒性、多源鲁棒性的三维评价指标。
- **大规模实验评估。**基于该基准，本工作在 7 种多传感器融合目标检测模型和 5 种单模态目标检测模型上进行了大量实验评估。并且通过基于融合顺序和融合表征

的二维分类范式以及基于信息熵的融合模型鲁棒性分析方法，显式分析了融合架构和鲁棒性之间的关系。

- **关键研究问题。**本工作系统地回答了与多传感器融合模型的鲁棒性相关的两个关键研究问题，还提出了增强多传感器融合鲁棒性的启示。

## 4.2 研究范围及威胁模型

### 4.2.1 研究范围及定义

首先，需要明确本工作的研究范围。由于传感器工作的本质是将物理信号转化为电信号<sup>[157]</sup>，因此本文将重点放在信号注入攻击下的传感器数据受损上。下列攻击不属于本工作的研究范围：1) 对测量目标进行的物理修改，例如利用贴纸<sup>[158-164]</sup>或 3D 实体<sup>[72]</sup>对物体的图案或形状进行修改；2) 攻击 CAN 总线中传感器数据的数字传输<sup>[165-166]</sup>，或攻击传感器网络<sup>[167-169]</sup>。

本文现在定义信号注入攻击鲁棒性。首先，考虑一个在分布  $\mathcal{D}$  的样本上训练得到的目标检测模型  $f: X \rightarrow Y$ ，其中  $X$  是传感器数据， $Y$  是目标检测结果。大多数目标检测模型的性能通过其在  $\mathcal{D}$  抽取的测试样本上的交并比 ( $IoU$ ) 与阈值 ( $t$ ) 的联合标准来评判，即  $\mathbb{P}_{(x,y) \sim \mathcal{D}}(IoU(f(x), y) > t)$ 。然而在自动驾驶等安全关键型应用中，目标检测模型可能面临来自环境偶发或人为恶意的信号注入攻击，因此希望知道模型在面对这些信号注入攻击下的数据损坏时的性能表现。鉴于此，本文将“模型在处理信号注入攻击下的受损数据时的性能表现”定义为信号注入攻击鲁棒性： $\mathbb{E}_{c \sim \mathcal{C}}[\mathbb{P}_{(x,y) \sim \mathcal{D}}(IoU(f(c), y) > t)]$ ，其中  $\mathcal{C}$  是受损数据的集合。受损数据  $\mathcal{C}$  的设计需要满足物理可实现性，即  $\|\mathcal{C} - X\| < \Phi$ ，其中  $\Phi$  是表征信号注入攻击改变传感器数据  $X$  的能力边界的集合。

### 4.2.2 信号注入攻击的威胁模型

在本鲁棒性测评基准中，本文考虑主动实施信号注入攻击的攻击者具有如下假设：

**攻击能力：**攻击者在车外进行非接触式攻击，以达到隐蔽的目的。攻击者可以瞄准摄像头或激光雷达并注入信号对其进行攻击，必要时有能力解决瞄准问题。

**传感器权限：**攻击者无法直接接触目标传感器、更改设备设置或安装恶意软件。但是，攻击者完全了解目标传感器的特性，这些知识可以从用户手册或通过分析与目标传

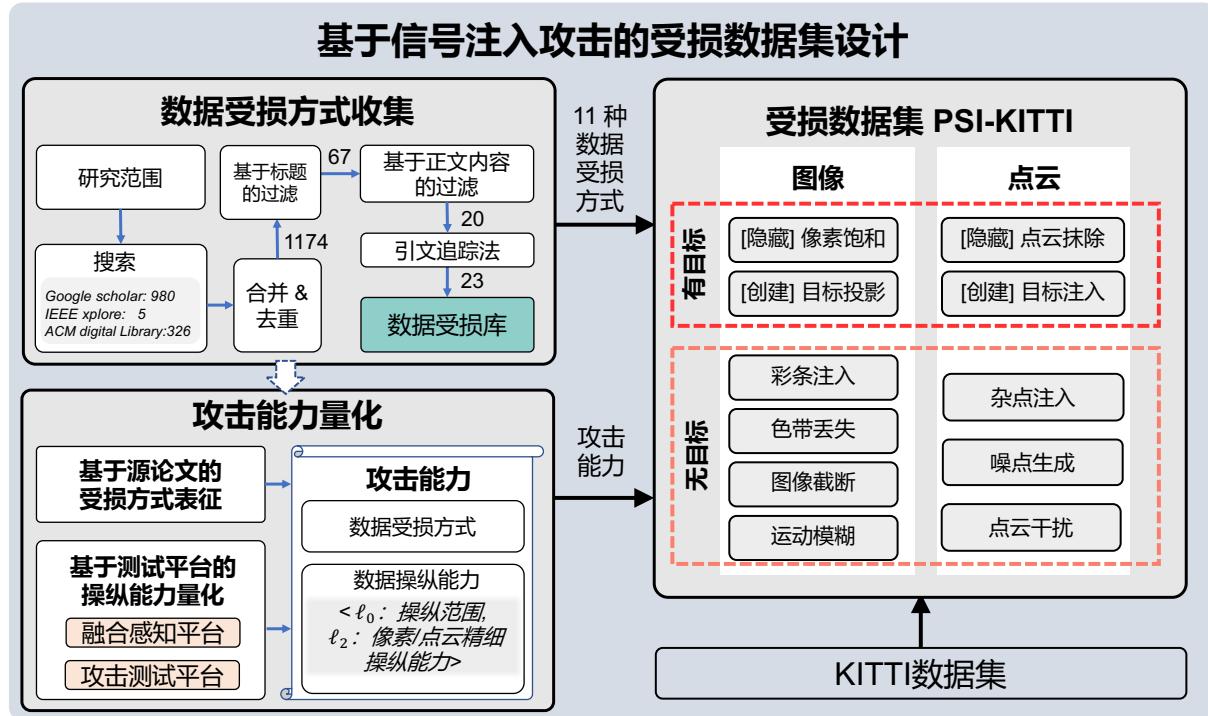


图 4.2 基于信号注入攻击的多传感器融合鲁棒性测评基准数据集设计流程图

感器相同型号的传感器获得。

**模型黑盒：**攻击者无法访问目标检测模型，攻击者只能利用传感器的特性和漏洞来实现攻击目标。

### 4.3 基准测评数据集设计

本节首先介绍受损数据集 SIA-KITTI 的设计方法，包括数据受损方式收集和攻击能力量化的整体流程，然后分别详细介绍 6 种图像受损方式和 5 种点云受损方式及其量化后的攻击能力。

#### 4.3.1 SIA-KITTI 数据集设计方法介绍

在设计受损数据集 SIA-KITTI 时，本工作致力于确保数据集的全面性和物理可行性。其中全面性通过基于系统性文献综述（SLR）的数据受损方式收集流程来保证，物理可行性通过攻击能力量化来保证。

### 4.3.1.1 数据受损方式收集

为了尽可能全面地收集了与传感器攻击相关的工作，本工作设计了严格的系统文献综述（Systematic Literature Review，SLR）<sup>[170-171]</sup>流程，如图4.2所示，系统文献综述本身遵循计划、执行和报告三步方法。搜索的内容为“信号注入攻击”，即利用激光、电磁、声波等物理信号来影响传感器输出的恶意攻击和极端干扰。首先，本文使用以下查询关键词来搜索文档：

**Query:** (“physical” OR “real-world” OR “practical”) AND “signal” AND (“attack” OR “vulnerability”) AND (“LiDAR” OR “camera”) AND (“autonomous driving” OR “self driving” OR “autonomous vehicle” )

通过使用引文分析软件 *Publish or Perish*<sup>[172]</sup>，本文收集了来自 Google Scholar (980 篇)、IEEE Xplore (5 篇) 和 ACM Digital Library (326 篇) 的共 1311 篇研究论文。从总共 1311 条搜索结果中删除重复内容后，剩下 1174 篇论文。

接下来对 1174 篇论文进行筛选，筛选标准如下：1) 只收录与针对激光雷达或摄像头的信号注入攻击相关的研究；2) 只收录首次引入攻击的论文以及对攻击进行改进的论文。最初，本文根据论文标题进行了初步筛选，得出了 67 篇可能相关的论文。随后，在对内容进行审查后，本文筛选出了 20 篇文章。然后，本文对所有这些作品采用了引文追踪法，以发现在初步搜索中被忽略的资源，并采用相同的纳入标准，从而增加了 3 项新的研究。综上，SLR 过程共找到了 23 篇研究论文，本文从这些论文中提炼出了 11 种信号注入攻击下的数据受损类型。

### 4.3.1.2 攻击能力量化

如图4.2所示，本文分两步量化了传感器攻击的能力。首先，本文基于源论文来进行数据受损方式的表征，除此以外，少部分源论文也清楚地描述了数据操纵能力。接着，对于那些源论文没有量化数据操纵能力的攻击，本文在测试平台上复现这些攻击，以确保每种攻击的物理可行性，并进一步明确每种攻击的能力和局限性。

本工作根据数据受损方式和操纵能力来量化信号注入攻击的能力。与对抗攻击类似，本文利用  $l_0$  和  $l_2$  范数来表示对图像或点云的操纵能力，其中  $l_0$  范数表示操纵范围，而  $l_2$  范数则说明像素/点云的操纵能力。更具体地说，对于图像， $l_0$  表示可被攻击操纵



图 4.3 攻击能力测试平台

的像素数量（范围）， $l_2$  表示可被操纵的像素值大小和精度。对于点云， $l_0$  表示可受攻击影响的点的数量（范围）， $l_2$  表示可被操纵点的距离范围和精度。

本文的攻击能力测试平台由融合感知平台和攻击测试平台组成。融合感知平台如图4.3a所示，由安装在Apollo自动驾驶套件上的Leopard USB3.0摄像头<sup>[173]</sup>和VLP-16激光雷达<sup>[75]</sup>组成。攻击测试平台如图4.3b所示，包括一个信号发生器、一个放大器和三种信号发射器，可发射激光、超声波和电磁信号。每种受损情况的详细攻击能力量化过程在第4.3.2章中介绍。

#### 4.3.2 受损方式详细介绍

本工作一共包含了 11 种信号注入攻击下的数据受损方式，如图4.4所示，其中图像受损 6 种，点云受损 5 种。除此以外，本文还根据是否有明确的攻击目标将数据受损分为“有目标 (targeted) 攻击”和“无目标 (untargeted) 攻击”。其中“有目标攻击”由于数据操控能力强大，即使在黑盒情况下，在模型层面有明确的隐藏或创建某个目标物体的能力，图像受损和点云受损都分别有隐藏和创建两种有目标攻击。“无目标攻击”则更注重在数据层面的损坏，难以在黑盒情况下实现明确的攻击目标。接下来本文详细介绍每一种数据受损的原理及其攻击能力。

#### 4.3.2.1 [隐藏] 像素饱和

**原理介绍：**该数据受损方法是使用大功率激光或高流明光束直接照射摄像头，导致摄像头中的感光模块饱和，从而有效隐藏环境中的真实物体。这种现象的原理类似于现实生活中动态照明条件下的过度曝光，这种过曝现象是由于光通量过大和感光模块饱和造成的。饱和（或过曝）通常发生在实际驾驶环境中，例如汽车驶出隧道时<sup>[179]</sup>或迎面

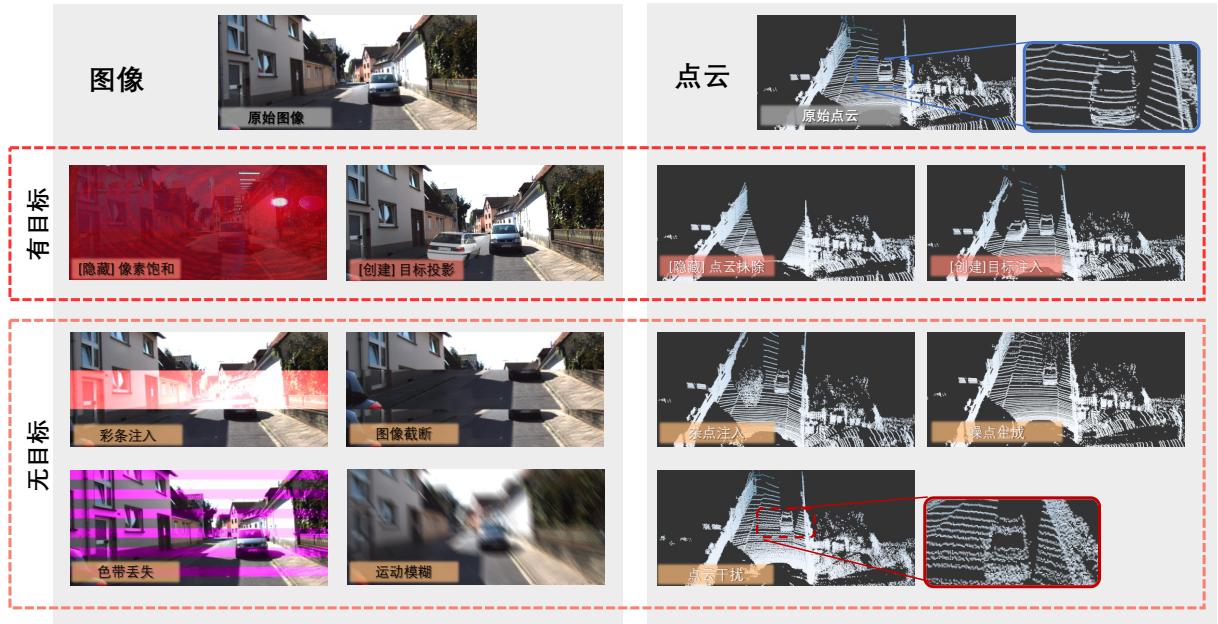


图 4.4 数据受损方式介绍

而来的汽车启动远光灯时<sup>[180]</sup>，车载摄像头的图像会曝光过度而饱和甚至致盲。所以这种数据受损方式能够实现隐藏制定目标物体的攻击效果。

**攻击能力量化：**以前的论文曾分别利用 LED 和激光对照相机进行过隐藏攻击，实验结果表明，激光由于能量更聚焦，更容易使照相机失明，甚至损坏照相机。在之前的实验中，使用的是红色激光（650 nm）。为了使实验结果更加完整，本文使用了更多波长的激光进行实验，并尝试了高流明光束。本文发现，在相同功率下，绿激光（550 nm）会比其他波长的激光造成更严重的过曝效应。这可能是由于 CMOS 传感器中绿色像素的比例较高。以往工作并未明确量化攻击能力，本文为了保证该数据受损的物理可实现性，首先记录白色背景下的激光攻击模式，然后将其添加到原始图像的 {R,G,B} 通道中。

#### 4.3.2.2 [创建] 目标投影

**原理介绍：**该数据受损方法包括使用投影仪将图像投射到环境中然后反射到摄像头<sup>[149-151,174]</sup>，或者直接将图像投射到摄像头中<sup>[148]</sup>。虽然这种攻击方法看起来有些天真，但却是一个重大威胁。它的效果与基于贴纸的攻击有些相似，然而贴纸攻击相比，目标投影具有更强的危害性。首先，它可以远程执行，不需要攻击者亲自过去贴。其次，它可以通过物理信号操纵方便地控制攻击形式。第三，它可以将元素投射到对基于贴纸的

表 4.1 信号注入攻击的攻击能力量化表

ID	数据受损方式	攻击能力			引文
		受损方式	$\ell_0$ : 操纵范围	$\ell_2$ : 像素/点云 操纵能力	
1	[隐藏] 像素饱和	全局过曝	所有像素	根据量子效率对 {R,G,B} 通道进行增值	[28,147]
2	[创建] 目标投影	特定图案叠加	特定区域	投影像素和原始像素的值叠加	[148-151,174]
3	彩条注入	高斯分布彩条	特定图像行	根据量子效率对 {R,G,B} 通道进行增值	[152,175]
4	色带丢失	均匀颜色彩条	特定图像行	滤光片阵列错误映射: G→R/B, R/B→G	[153]
5	图像截断	图像截断、拼接	特定图像行	图像与上一帧或下一帧拼接	[153]
6	运动模糊	线性模糊	所有像素	一系列平移像素的值叠加	[154,176]
7	[隐藏] 点云抹除	点云消失	30° 水平角	原始点被直接抹除	[91-92]
8	[创建] 目标注入	特定目标点云注入	20° 水平角	随机距离误差为 0.1 米的欺骗点云.	[91,133,177]
9	杂点注入	随机目标点云注入	30° 水平角	随机距离误差为 1 米的欺骗点云.	[24,28,30-32]
10	噪点生成	随机背景噪点	30° 水平角	随机距离误差为 100 米的欺骗点云	[30]
11	点云干扰	距离误差	所有点	最大距离误差为 0.15 米的正弦噪声	[93,178]

攻击具有挑战性的位置，例如路边的树木或空中。不过，这种投影攻击的主要缺点在于容易受到环境光线条件的影响。

**攻击能力量化：**本文采用两种方法进行测试：投射到环境中和直接投射到摄像机中。然而，要成功地直接投射到摄像头中具有挑战性，因为这需要投影仪的精确光学聚焦以及攻击信号与摄像头光敏元件之间的高精度瞄准。通过测试，本文最终选择通过向环境投射图案来实施攻击。本文发现这种方法可以方便地发起创建攻击，并能有效欺骗最先进的基于图像的目标检测模型。然而，在强烈的光照条件下，要成功实施投影攻击具有挑战性。

#### 4.3.2.3 彩条注入

**原理介绍：**该数据受损方法是使用开关调制的激光，基于摄像头 CMOS 传感器的卷帘快门（Rolling Shutter）效应，注入彩色条纹，之前的研究<sup>[152]</sup>评估了这种攻击对交通灯识别的影响。

**攻击能力量化：**该攻击的作者<sup>[152]</sup>在论文中广泛讨论了脉冲激光对图像的影响，量化了攻击能力并提供了激光攻击的建模方法，因此本文采用了他们的方法来设计受损。

#### 4.3.2.4 色带丢失和图像截断

**原理介绍：**该数据受损方法<sup>[153]</sup>利用有意电磁干扰（IEMI）注入恶意信号，以用于图像信号传输的摄像头接口总线为目标，造成摄像头故障。攻击原理如下：使用 MIPI CSI-2 接口传输标准的摄像头会为图像信号分配一个缓冲区，缓冲区的起始/结束地址和行距被传递给 Unicam（CSI 接收器）。图像信号按单线传输，并根据固定的滤色片排列进行解码。摄像机会丢弃传输错误的行，如果传输中少了一行，就会在图像处理过程中破坏后续行的色彩解释，从而造成色带丢失。如果缺少缓冲区的开始/结束地址，就会出现帧间内容拼接，从而导致图像截断。

**攻击能力量化：**提出该攻击的论文<sup>[153]</sup>演示了攻击者可以利用电磁干扰在图像中诱发色带，这是由于光学滤镜中的错误造成的，例如蓝绿（B/G）和绿红（G/R）滤镜的错误使用。因此，与“彩条注入”中不均匀的色带不同，“色带丢失”生成的色条在视觉上呈现出均匀的紫色。该论文提供的攻击能力表明，攻击者可以控制紫色条纹的位置、宽度和数量。同样，对于截断，攻击者可以通过调整信号来控制截断的位置。本文在基于试验平台的测试中也证实了这一点。因此，本文按照论文中概述的攻击能力来设计数据受损方式。

#### 4.3.2.5 运动模糊

**原理介绍：**该数据受损方法<sup>[154,176]</sup>是基于防抖摄像头中图像稳定器硬件易受声波操纵的系统级漏洞，通过发射专门设计的超声波信号，可以影响惯性传感器的输出，从而触发不必要的运动补偿，导致图像模糊。

**攻击能力量化：**根据像素沿不同自由度运动的三种类型，作者将模糊模式分为线性模糊、径向模糊和旋转模糊。通过测试平台的物理实验，本文发现线性模糊是在物理世界最容易诱发的模糊类型，因此最值得关注。综上，本文采用线性模糊来设计受损，线性模糊的强度和超声波信号的幅值有关。

#### 4.3.2.6 [隐藏] 点云抹除

**原理介绍：**该数据受损方法可通过连续激光<sup>[30]</sup>、脉冲型激光<sup>[91-92]</sup>和电磁<sup>[178]</sup>分别实现。激光雷达的功能是发射激光，接收回波，进行飞行时间测量从而算出距离，最终

生成点云。现有的点云抹除方法会从根本上破坏或隐藏来自物体的有效回波。Shin 等人<sup>[30]</sup>利用高功率 (800mW) 连续激光使激光雷达的光电探测器饱和，使其无法接收有效回波。Jin 等人<sup>[91-92]</sup>采用特定频率的脉冲激光注入高强度点，然后利用点云的回波过滤机制滤除有效回波。除此以外，Jin 等人<sup>[178]</sup>也发现可以利用电磁波干扰激光雷达接收模块的模拟电路或监测传感器，从而实现点云抹除。由于点云抹除在数据层面直接抹去了目标信息，所以可以在模型黑盒情况下实现指定目标的隐藏效果。

**攻击能力量化：**在 Shin 等人的论文<sup>[30]</sup>中，他们使用 800mW 的连续激光隐藏了一块  $41*42\text{cm}^2$  金属板的点云，但他们没有量化具体的攻击能力。在本文的测试平台实验中，本文使用输出功率分别为 200mW、600mW、1000mW 和 2000mW 的 905nm 连续激光器进行了实验。本文发现，有效清除范围随着功率的增加而增大，使用 2000mW 的激光，本文能够清除大约  $6^\circ * 6^\circ$  区域内的点云。先前利用脉冲型激光的研究<sup>[91-92]</sup>进行了详细的评估，并证明攻击者可以移除水平角度超过  $30^\circ$ <sup>[91]</sup> 的目标点云。本文也在测试平台上证实了这一点。考虑到整体攻击效果和成本，本文决定借鉴后一种攻击方法来量化攻击能力并设计数据损坏方法。

#### 4.3.2.7 [创建] 目标注入

**原理介绍：**该数据受损方法<sup>[91,133,177]</sup>采用一套激光接收器和发射器，针对机械激光雷达系统进行可控点云注入。PLA-LiDAR<sup>[91]</sup>证明了在物理世界中注入点云并利用黑盒方法直接欺骗 3D 目标检测模型是可行的。

**攻击能力量化：**对于目标注入这一点云数据受损方法来说，欺骗点的最大可注入点数、位置控制能力以及形状控制能力是最重要的三个指标。先前的工作<sup>[91,181]</sup>对围绕上述三个指标做了详细的攻击能力评估。其中欺骗点的最大可注入点数为 4000，这个数量足以注入汽车和行人等物体的点云；位置控制精度的标准偏差为 0.38 米；形状控制精度的标准偏差为 0.102 米。本文在设计受损数据集时考虑了上述点数限制和两个误差。

#### 4.3.2.8 杂点注入

**原理介绍：**该数据受损方法<sup>[24,28,30-32]</sup>同样是利用激光实现点云的注入，然而这些注入点呈现出一定程度的随机性，而不是上述目标注入论文<sup>[91,133]</sup>中所示的规则形状。本文认为这可能是由于信号设计的不同，以及与可控注入相比缺乏精确同步造成的。尽管

这些攻击尚未被证明能在物理世界中实现有针对性的攻击效果，但它们对融合模型性能的潜在影响值得探究。

**攻击能力量化：**以前的相关研究<sup>[24,28,30-32]</sup>声称最多可以注入 200 个点。然而，受 PLA-LiDAR<sup>[91]</sup>的启发，本文认为注入数千个点是可行的。因此，在设计受损时，本文将点的数量设置为与“[创建] 目标注入”中相同，但是为每个点分配一个从 0 到 1 米不等的平均误差，以反映其随机性。

#### 4.3.2.9 噪点生成

**原理介绍：**该数据受损方法<sup>[30]</sup>涉及使用低功率激光注入随机假点。作者认为噪点的出现可能是由于低功率激光导致基线噪声增加。在激光雷达的日常使用中，有时会观察到类似的噪声，主要是由于激光雷达之间的干扰。

**攻击能力量化：**论文<sup>[30]</sup>声称能够在 20° 水平角范围内注入噪声。然而，仅凭这些信息还不足以设计受损。因此，本文在测试平台上使用 5mW 和 30mW 连续 905nm 激光在 7 米距离内进行了进一步实验。本文发现，在大约 30° 水平角范围内注入噪点是可行的，而且噪点在 0 至 150 米范围内均匀分布。

#### 4.3.2.10 点云干扰

**原理介绍：**该数据受损方法<sup>[93,178]</sup>利用了激光雷达接收模块的模拟电路对电磁波的敏感性。通过向激光雷达电路注入特定频率的电磁波信号，损坏激光雷达的测距功能，从而使点云受到干扰并在径向距离上引入测距误差。

**攻击能力量化：**本文的第3章利用电磁成功注入了最大距离误差为 15 厘米的点云干扰，并且干扰可以呈现正弦噪声或者随机噪声的形式。经第3章的实验分析，正弦干扰对模型的性能影响更大，因此本数据集采用幅值为 15cm 的正弦干扰这一攻击能力进行设计。

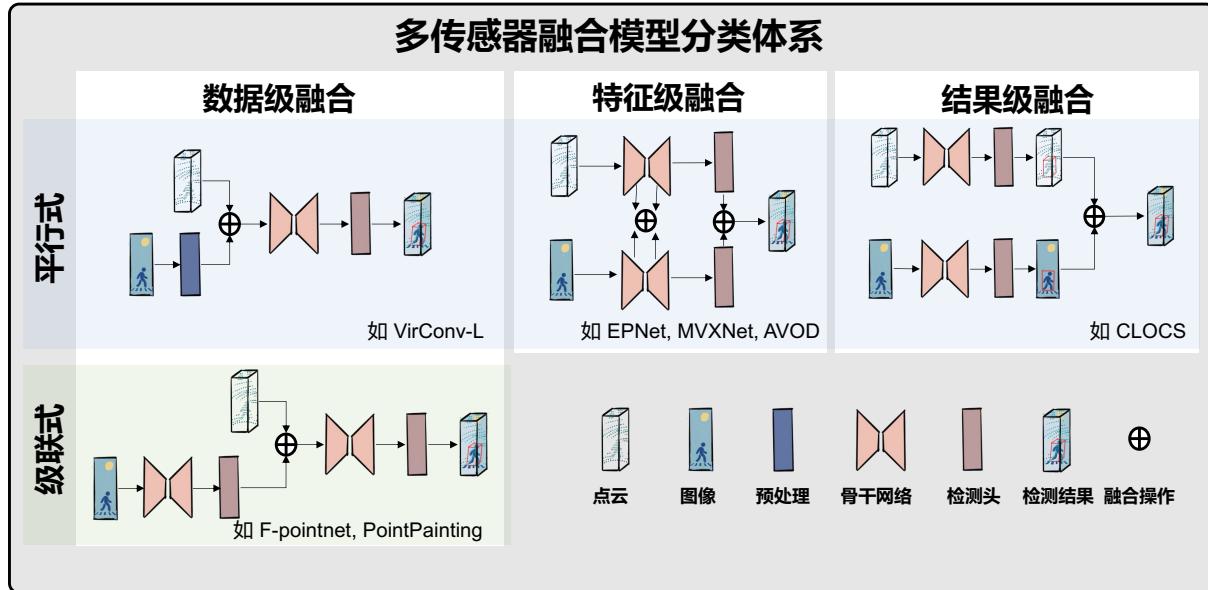


图 4.5 不同架构的“点云-图像”融合模型

## 4.4 模型及指标

### 4.4.1 融合模型

为了尽可能多地收集不同架构的基于激光雷达和摄像头融合的目标检测模型，本文查阅了大量相关综述论文<sup>[11,140-143,182]</sup>和 KITTI 排行榜<sup>[9]</sup>，收集了过去六年中在相关顶级会议和期刊上发表的开源工作。最终，本文选择了 7 种最新的开源 MSF 模型，如表 4.2 所示。

现有的多传感器融合模型主要在基于点云的目标检测模型上进行设计，并尝试图像信息纳入点云模型工作流的不同阶段。如图 4.5 所示，根据融合阶段的不同，MSF 模型可分为前融合、中融合和后融合。本工作收集了三个前融合模型（F-Pointnet<sup>[12]</sup>、Pointpainting<sup>[13]</sup> 和 VirConv-L<sup>[14]</sup>），两个中融合模型（EPNet<sup>[15]</sup> 和 AVOD<sup>[16]</sup>），一个后融合模型（CLOCs<sup>[17]</sup>）和一个混合融合模型（VirConv-T<sup>[14]</sup>）。

大多数模型在进行融合操作前会同时处理来自两个传感器的数据，本文将这种并发的融合方法称为“平行融合”。而在前融合中，有一种融合是按照顺序结构进行的，本文称之为“级联融合”，在级联融合中，首先使用基于图像的模型来获得二维识别结果，如检测框<sup>[12]</sup>或语义信息<sup>[13]</sup>，然后利用这些二维结果来增强点云，随后将其输入基于激光雷达的目标检测模型。

表 4.2 基于多传感器融合的目标检测模型

Model	融合阶段	融合顺序	融合时的模态表征		融合操作
			图像表征	点云表征	
F-Pointnet	前	级联	2D 检测框	点云数据	区域截取
PointPainting	前	级联	语义信息	点云数据	点级别增强
VirConv-L	前	平行	虚拟点	点云数据	数据合并
VirConv-T	混合	平行	虚拟点	点云数据	数据合并 & ROI 合并 & 检测框投票
EPnet	中	平行	图像特征	点云特征	特征级联
AVOD	中	平行	图像特征	BEV 特征	特征级联
Clocs	后	平行	2D 检测框	3D 检测框	检测框投票

#### 4.4.2 测评指标

本节定义基于 MSF 的 3D 目标检测模型在信号注入攻击下鲁棒性的评估指标。3D 目标检测任务旨在定位三维空间中物体的位置，在目标检测任务中，交并比 (Intersection over Union, IoU) 是衡量预测框  $B_p$  与真实框  $B_g$  匹配程度的核心指标，其定义为两框交集面积与并集面积的比值，IoU 的取值范围为 0 到 1，IoU 越大，预测框与真实框的匹配程度越高。本基准参照 KITTI<sup>[89]</sup>的标准，当 IoU 大于 0.7 时认为是成功检测到目标。为了更好地对模型在不同攻击下的鲁棒性进行基准测评，本文采用了几种基于 IoU 的高级评价指标。

##### 4.4.2.1 攻击成功率 ASR

攻击成功率 (Attack Success Rate, ASR) 用于量化有目标攻击的成功率。在本文的基准中，针对黑盒模型的定向攻击可以实现两种效果：隐藏和创建。只有当目标对象不被检测到时，隐藏攻击才算成功。相反，创建攻击只有在指定区域内生成一个最初并不存在的目标对象时才算成功。

##### 4.4.2.2 平均精度 AP

平均精度 (Average Precision, AP) 是衡量模型在不同 IoU 阈值下的检测性能的指标。AP 的计算方法是将 IoU 阈值从 0.5 到 0.95 进行遍历，并计算每个阈值下的精度，如

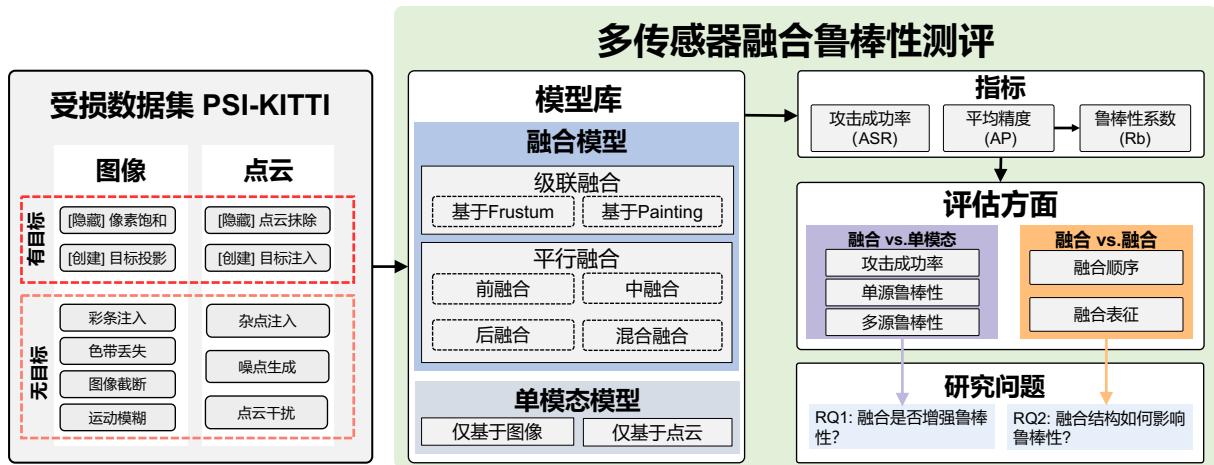


图 4.6 鲁棒性评估流程图

下式所示

$$AP|_{R_{40}} = \frac{1}{|R_{40}|} \sum_{r \in R_{40}} \max_{\tilde{r}: \tilde{r} > r} \rho(\tilde{r}), \quad (4-1)$$

其中， $R_{40}$  表示在 40 个 IoU 阈值下计算的精度， $\rho(\tilde{r})$  表示在 IoU 阈值为  $\tilde{r}$  时的精度，这意味着不是对每个点的实际观测精确度值  $r$  取平均值，而是取大于或等于  $r$  的召回值时的最大精度。本文采用平均 AP ( $mAP$ )<sup>[183-184]</sup>，取 KITTI<sup>[89]</sup> 中的三个难度级别（即简单、中等和困难）的 AP 的平均值，来衡量模型的整体检测性能。

#### 4.4.2.3 鲁棒性 Rb

本文将一个 MSF 模型在数据受损  $c$  上的鲁棒性定义为  $Rb_c$ ：

$$Rb_c = \frac{mAP_c}{mAP_{clean}}, \quad (4-2)$$

其中， $mAP_c$  和  $mAP_{clean}$  分别代表模型在受损数据和干净数据上的总体表现。一个模型在多种受损情况下的平均鲁棒性用  $mRb$  表示。

$$mRb = \frac{1}{|C|} \sum_{c \in C} Rb_c. \quad (4-3)$$

## 4.5 鲁棒性测评

### 4.5.1 实验总述

本研究选取了 7 个基于多传感器融合的模型和 5 个单模态模型作为评估对象，并使用 SIA-KITTI 数据集进行了系统性的基准测试。为保证评估的公平性，所有模型均采用其官方发布的模型参数，这些参数都在 KITTI 训练集上进行了训练。SIA-KITTI 数据集共包含 12 组数据，包括 1 组原始（干净）数据和 11 组受损数据，每组数据包含 3,769 帧激光雷达-相机配对数据。

考虑到部分模型仅支持“汽车”类别的检测，本文将评估重点聚焦于所有模型对汽车类别的检测性能。对于每一帧数据的检测结果，本文采用交并比（IoU）阈值 0.7 来计算平均精度（AP）。基于获得的 AP 值，本文进一步计算模型的鲁棒性指标（robustness  $R_b$ ），相关结果如表 4.4 所示。此外，针对有目标的攻击场景，本文还计算了攻击成功率（ASRs），详见表 4.3。

基于以上实验结果，本文围绕“1) 融合是否增强鲁棒性？2) 融合结构如何影响鲁棒性？”两个核心研究问题展开深入讨论和分析。

### 4.5.2 研究问题 1：融合是否增强鲁棒性？

为了全面评估多传感器融合是否能够增强模型的鲁棒性，本文从三个层面展开分析：1) 有目标攻击的攻击成功率，用于评估模型在面对恶意攻击时的防御能力；2) 单源鲁棒性，用于评估模型在面对单一传感器受损时的性能表现；3) 多源鲁棒性，用于综合评估模型在所有可能的受损场景下的整体表现。这种多层次的评估方法能够帮助本文深入理解多传感器融合对模型鲁棒性的影响。

#### 4.5.2.1 有目标攻击的攻击成功率

评估有目标攻击具有重要的现实意义。在特定的驾驶场景中，攻击者往往试图隐藏目标物体或在预定位置创造物体，这可能导致碰撞或交通堵塞等攻击者预期的后果。过往研究表明，单模态模型对此类有目标攻击特别敏感，这凸显了有目标攻击所带来的独特且重要的威胁，需要安全研究人员的特别关注。因此，评估攻击者是否能像控制单模态模型一样直接控制多模态模型的输出，成为衡量模型鲁棒性的关键方面之一。

表 4.3 四种有目标攻击的攻击成功率 (ASR)

被攻击 传感器	攻击 目标	数据受损类型	仅基于图像	仅基于点云	多传感器融合						
			ImVoxelNet	PointPillar	F-pointnet	Pointpainting	virconv_l	virconv_t	Epnnet	AVOD	CLOCS
摄像头	隐藏	像素饱和	97.37%	\	92.58%	47.41%	0.04%	0.54%	7.78%	20.48%	88.51%
	创建	目标投影	100.00%	\	96.77%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
激光雷达	隐藏	点云抹除	\	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%	100.00%
	创建	目标注入	\	100.00%	0.00%	95.93%	100.00%	98.27%	100.00%	100.00%	0.00%

本文使用参数 *ASR* 来评估有目标攻击的成功率。结果如表 4.3 所示。本文选取了一个基于相机的检测器 ImVoxelNet<sup>[185]</sup> 和一个基于激光雷达的检测器 PointPillar<sup>[8]</sup> 作为基准。可以观察到，这 4 种有目标攻击对单模态检测器都能达到较高的攻击成功率。相比之下，这些攻击对融合模型的攻击成功率差异显著。总的来说，**隐藏比创造更容易成功**。

在本小节，为了方便理解和分析，本文不以数据受损类型命名攻击，而是在命名中强调攻击的受损数据形式和攻击目标，比如，“图像隐藏”攻击的受损数据是图像，攻击目标是隐藏，以此类推。

分析每种类型的攻击发现，点云抹除攻击能够成功地破坏所有模型，这是因为抹除点云的方法几乎可以完全消除物体的 3D 信息，从而阻止 3D 边界框的成功回归。按照这个逻辑，目标投影攻击由于不提供物体的 3D 信息，理应无法成功。这对大多数模型确实如此。然而，本文惊讶地发现目标投影攻击在 F-pointnet 中成功地生成了欺骗物体。本文将在后面通过分析 F-PointNet 的模型架构来解释这一现象。

此外，图像隐藏攻击和点云创建攻击的成功率呈现反相关关系。容易受图像隐藏攻击影响的模型对点云创建攻击的敏感度较低，反之亦然。这表明现有的融合模型往往更依赖于某一个传感器源（称之为**主导传感器**）。CLOCS 中的后期融合平等对待来自两种模态的检测结果，消除了这种偏差。然而，由于 CLOCS 的结构特点，它倾向于剪枝而不是创建新的发现，这在实际自动驾驶场景中可能存在潜在危险。

接下来，本文将逐个模型地分析攻击结果。

**F-Pointnet<sup>[12]</sup>**：由于 F-Pointnet 采用级联融合结构，如果物体在图像中未被检测到，则该物体的点云将被过滤掉，这导致图像隐藏攻击具有较高的成功率。需要注意的是，严格来说，图像创建攻击不应该在任何模型中成功，因为仅修改图像并不能提供 3D 信息。然而，本文惊讶地发现图像创建攻击在 F-Pointnet 中成功生成了欺骗物体。为了理解图像创建攻击为何成功，本文将 3D 检测框投影到图像和点云数据中，得到了如图 4.7 所

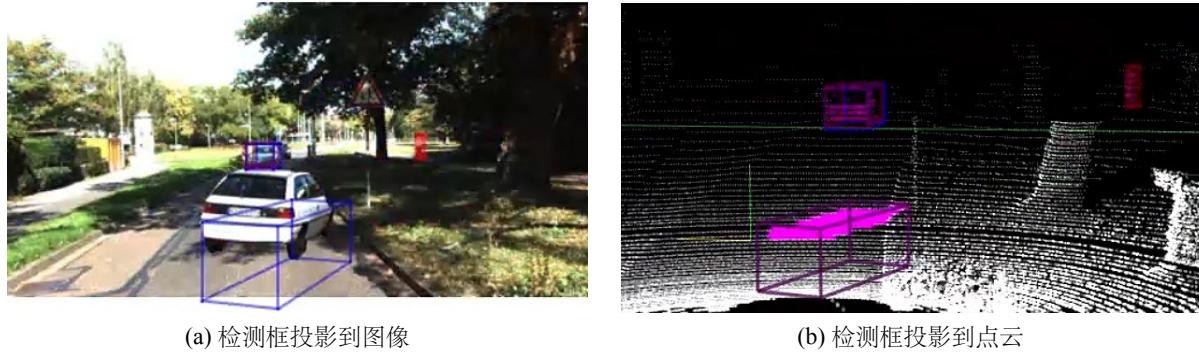


图 4.7 F-Pointnet 模型在“目标投影”攻击（图像创建）下的检测结果

示的可视化检测结果。本文发现 F-Pointnet 中创建的物体实际上是地面被误检为汽车的情况。本文推测，经过 F-Pointnet 的过滤机制后，地面点云获得了与汽车车顶极为相似的特征。

**Pointpainting<sup>[13]</sup>**：图像隐藏和点云创建攻击在 Pointpainting 上分别达到了 47.41% 和 95.93% 的攻击成功率。Pointpainting 架构包含三个主要阶段：(1) 基于图像的语义分割网络，用于计算像素级分割分数；(2) 融合阶段，将激光雷达点与语义分割分数结合；(3) 基于激光雷达的 3D 检测网络。根据本文的分类，这是一种早期融合。图像分割分数被附加到激光雷达点上形成“着色点”，这些着色点可以被任何学习编码器的激光雷达网络使用，因为 PointPainting 仅改变了激光雷达点的输入维度。在本基准测试中，本文使用了 PointPillar 作为主干网络，并用 KITTI 中的语义分割分数强化点云信息。因此，图像的分类信息和原始点云信息都有可能影响最终的检测结果。此外，从结果可以看出，PointPainting 的融合架构对图像攻击表现出更强的鲁棒性。

**EPNet<sup>[15]</sup>**：EPNet 对图像隐藏攻击具有很强的鲁棒性（攻击成功率 7.78%），但容易受到点云创建攻击（攻击成功率 100%）。EPNet 由一个双流主干网络组成，包括几何流和图像流。两个流分别产生点特征和语义图像特征。在图像流中，EPNet 采用四个级联的 3\*3 卷积块来提取不同尺度的图像语义特征，记为  $F_i(i=1,2,3,4)$ 。几何流包括四对点集抽象  $S_i(i=1,2,3,4)$  和特征传播层  $P_i(i=1,2,3,4)$ <sup>[7]</sup> 用于特征提取。点特征  $S_i$  通过 LI-Fusion 模块与图像特征  $F_i$  结合。总的来说，融合模块允许点特征和图像特征在主干网络中深度融合。然而，由于图像的主要作用是增强点云，点云特征在整个流程中仍然起主导作用。因此，相比相机攻击，基于激光雷达的攻击更容易影响 EPNet。

**AVOD<sup>[16]</sup>**：AVOD 对图像隐藏攻击具有较强的鲁棒性（攻击成功率 20.48%），但

容易受到点云创建攻击（攻击成功率 100%）。在 AVOD 中，由特征提取器生成的图像和鸟瞰图（BEV）点特征图在区域建议网络（RPN）中进行融合。这两种特征图随后被 RPN 用于生成非定向区域建议，这些建议被传递到检测网络进行尺寸细化、方向估计和类别分类。本文可以观察到，AVOD 的融合架构将图像和点云的特征置于同等地位，这与 EPNet 不同。在 EPNet 中，图像特征用于辅助增强点云特征。尽管 AVOD 架构旨在平等对待相机和激光雷达特征，但显然在训练后，模型偏向于依赖激光雷达特征。这一点从攻击结果中可以看出，相机攻击 vs 激光雷达攻击的平均攻击成功率分别为 10.24% 和 100%。

**CLOCs<sup>[17]</sup>**: 图像隐藏攻击（攻击成功率 88.51%）能够成功地破坏 CLOCs，而点云创建攻击（攻击成功率 0%）则不能。CLOCs 是一种后期融合方法，在应用非最大抑制（NMS）之前合并相机和激光雷达的检测候选对象。CLOCs 为每个传感器采用显著降低的阈值以优化其召回率。如果 2D 和 3D 边界框在图像平面上具有足够大的 IoU，则它们的信息将被组合成单个张量用于后续处理。然而，找不到匹配的 2D（或 3D）边界框的 3D（或 2D）边界框将被忽略。总的来说，CLOCs 与大多数后期融合一样，倾向于剪枝而不是创建新的发现，这解释了为什么隐藏攻击容易成功而创建攻击难以成功。

**观察 1(研究问题 1)**: 对于针对图像的有目标攻击，所有融合模型（7/7）都能降低其攻击成功率。然而，没有融合模型（0/7）能够防御点云隐藏攻击，只有部分模型（4/7）能够减弱点云创建攻击。虽然部分融合模型可以有效降低“有目标攻击”的成功率，但是原本针对单模态检测器设计的攻击仍然具有破坏融合模型的潜在能力。

#### 4.5.2.2 单源鲁棒性

单源鲁棒性指的是模型在面对单一传感器（如相机或激光雷达）攻击时的平均鲁棒性。由于单模态模型只会遭受其使用的传感器的攻击，单源鲁棒性允许本文在面对相同攻击时比较 MSF 和单模态模型的鲁棒性。

本文使用参数  $mRb^C$  和  $mRb^L$  分别表示模型在相机或激光雷达损坏下的平均鲁棒性。结果如表 4.4 所示，本文发现所有 MSF 模型的  $mRb^C$  都超过了基于相机的模型。这种改进部分归功于点云的融合，有效地增强了鲁棒性。另一个促成因素是现有开源相机模型的性能普遍较差，导致其鲁棒性较低。相比之下，本文实验中的 7 个 MSF 模型中只有

表 4.4 五种单模态模型和 7 种多传感器融合模型在 SIA-KITTI 数据集上的鲁棒性 (Rb) 评测

被攻击 传感器	干扰类型	仅基于图像		仅基于点云			多传感器融合						
		ImVoxelNet	SMOKE	Second	PointPillar	3DSSD	F-PointNet	PointPainting	VirConv_L	VirConv_T	EPNet	AVOD	CLOCs
摄像头	[隐藏] 像素饱和	0.415	0.069	/	/	/	0.226	0.402	<b>0.999</b>	0.995	0.804	0.592	0.315
	[创建] 目标投影	0.668	0.852	/	/	/	0.467	0.973	0.999	<b>1.000</b>	0.999	0.995	0.984
	彩条注入	0.520	0.203	/	/	/	0.962	0.832	<b>0.999</b>	0.993	0.797	0.752	0.993
	色带丢失	0.549	0.749	/	/	/	0.947	0.916	0.967	<b>0.985</b>	0.891	0.790	0.992
	图像截断	0.010	0.000	/	/	/	0.080	0.330	<b>0.956</b>	0.933	0.782	0.404	0.320
	运动模糊	0.001	0.000	/	/	/	0.386	0.330	<b>0.967</b>	0.958	0.790	0.411	0.636
激光雷达	[隐藏] 点云抹除	/	/	0.655	0.645	0.661	0.597	0.638	0.653	0.676	<b>0.683</b>	0.611	0.665
	[创建] 目标注入	/	/	0.781	0.778	0.767	0.775	<b>0.890</b>	0.793	0.796	0.707	0.830	0.889
	杂点注入	/	/	0.893	0.873	0.894	0.784	0.892	0.890	<b>0.910</b>	0.884	0.875	0.888
	噪点生成	/	/	0.814	0.855	0.742	0.516	0.898	0.854	0.922	0.729	0.839	<b>0.959</b>
	点云干扰	/	/	0.979	0.987	0.981	0.960	0.985	0.971	<b>0.994</b>	0.993	1.001	0.993
图像受损平均鲁棒性 ( $mRb^C$ )		0.359	0.312	/	/	/	0.511	0.630	<b>0.988</b>	0.977	0.844	0.657	0.707
点云受损平均鲁棒性 ( $mRb^L$ )		/	/	0.825	0.827	0.809	0.726	0.861	0.824	0.850	0.799	0.831	<b>0.879</b>
所有受损平均鲁棒性 ( $mRb$ )		0.650	0.625	0.920	0.922	0.913	0.609	0.735	0.918	<b>0.923</b>	0.824	0.737	0.785

4 个显示出比基于激光雷达的模型更优的  $mRb^L$ 。这表明融合并不必然保证增强鲁棒性，选择正确的融合方法需要额外的努力。

**观察 2 (研究问题 1):** 在考虑单源攻击时：与基于相机的模型相比，所有融合模型 (7/7) 都能增强单源鲁棒性。与基于激光雷达的模型相比，大多数融合模型 (5/7) 能够增加单源鲁棒性。

#### 4.5.2.3 多源鲁棒性

多源鲁棒性指的是模型在本基准测试中所有损坏情况下的平均鲁棒性。考虑到 MSF 模型中传感器数量的增加，它们面临更多的潜在攻击向量。这一方面在评估鲁棒性和安全性时至关重要，不容忽视。

本文使用参数  $mRb$  来评估多源鲁棒性。对于基于相机的模型，本文将其在所有激光雷达损坏情况下的鲁棒性设为 1，基于激光雷达的模型则反之。从表 4.4 中，本文观察到大多数 MSF 模型相比基于相机的模型具有更大的  $mRb$ 。这种差异可能归因于现有开源相机模型的性能欠佳。最佳  $mRb$  表现来自基于激光雷达的模型。同时，最先进的 MSF 模型 VirConv-L 和 VirConv-T 也展示了值得称赞的  $mRb$ 。总的来说，基于 MSF 的模型相比基于激光雷达的模型具有较低的  $mRb$ 。

**观察 3 (研究问题 1):** 由于基于 MSF 的模型同时暴露在来自两个传感器的攻击之下，大多数 (6/7) 现有融合模型并未增强多源鲁棒性。

#### 研究问题 1：融合是否增强鲁棒性？

**回答：**考虑到有目标攻击鲁棒性、单源鲁棒性和多源鲁棒性，大多数基于 MSF 的模型相比图像模型展现出更强的鲁棒性，但与点云模型相比则表现较弱。因此，融合不一定能增强鲁棒性。然而，最先进的融合模型（如 VirConv-T）有望在所有方面增强鲁棒性，展示了 MSF 在提高鲁棒性方面的潜力。

### 4.5.3 研究问题 2：融合结构如何影响鲁棒性？

为了回答研究问题 2，本文比较了不同基于多传感器融合的模型，旨在评估哪种融合设计具有更强的鲁棒性。本文从两个角度进行这种比较：融合顺序和融合数据表征。

#### 4.5.3.1 融合顺序的影响

根据表 4.4 中的结果，不同架构的模型其鲁棒性差异很大。为了探究融合顺序与鲁棒性之间的关系，本文将模型分为级联融合和平行融合两类。

如图 4.8 (a) 所示的所有损坏情况下的  $mRb$ ，本文观察到级联融合的鲁棒性普遍低于平行融合。本文认为这是由于级联效应导致的，即单个传感器损坏造成的错误会在整个检测流程中传播，从而降低多源鲁棒性。相比之下，平行融合允许来自两个传感器的数据相互补充，从而增强鲁棒性。

**观察 4 (研究问题 2)：** 从融合顺序的角度来看，平行融合表现出比级联融合更好的鲁棒性。

#### 4.5.3.2 融合表征的影响

从检测流程的输入到输出，数据表征从原始数据过渡到特征，再到结果。本文采用信息熵（记为  $H_X$ ）来直观地量化数据  $X$  的信息量。在神经网络中，基本操作如卷积、激活函数、ROI 池化、NMS 和全连接层等都会导致信息损失<sup>[186]</sup>。因此，本文可以直观地得出以下关系：

$$H_{data} > H_{feature} > H_{result}. \quad (4-4)$$

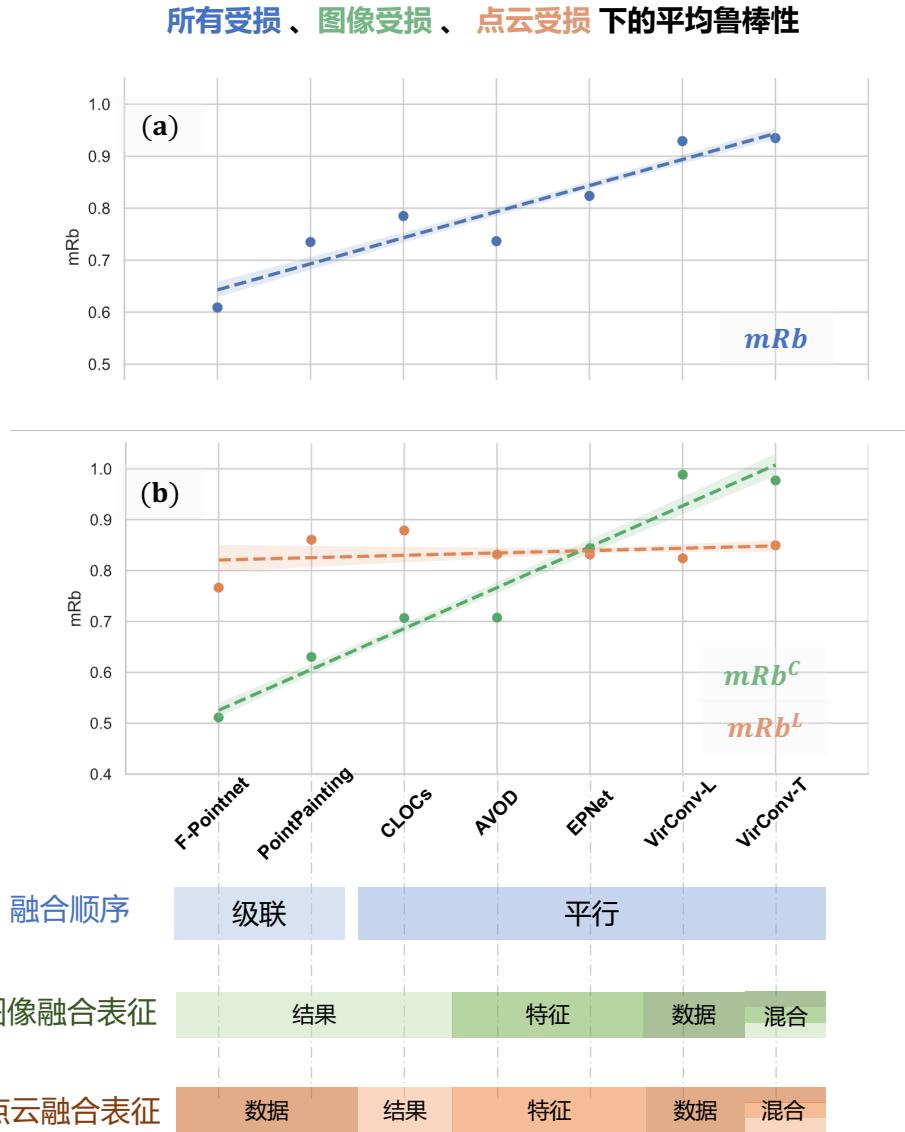


图 4.8 (a) 所有受损的平均鲁棒性。(b) 图像受损和点云受损下的平均鲁棒性。

此外，本文使用  $H_{FR}(M)$  来表示模型  $M$  的融合表征中包含的信息熵。

首先，让本文考虑级联模型：F-Pointnet 和 PointPainting。这两个模型都使用从图像生成的 2D 结果与原始点云进行融合。而 F-Pointnet 通过使用 2D 检测结果过滤点云来降低点云的信息熵。因此本文有：

$$H_{FR}(\text{Pointpainting}) > H_{FR}(F - \text{pointnet}). \quad (4-5)$$

其次，让本文考虑平行融合模型 VirConv-T、VirConv-L、EPNet、AVOD 和 CLOCs。这五个模型的融合表征如图 4.8 所示。需要注意的是，AVOD 的激光雷达输入是鸟瞰图 (BEV)。显然，BEV 的信息熵小于原始点云。因此，本文可以确定  $H_{FR}(\text{AVOD})$  的信息量小于  $H_{FR}(\text{EPNet})$ ，但本文无法比较  $H_{FR}(\text{AVOD})$  和  $H_{FR}(\text{CLOCs})$ 。因此，本文有：

$$\begin{aligned} H_{FR}(VirConv-T) &> H_{FR}(VirConv-L) > \\ H_{FR}(EPNet) &> H_{FR}(AVOD), H_{FR}(CLOCS). \end{aligned} \quad (4-6)$$

如图4.8(a)所示，模型的多源鲁棒性也遵循公式4-5和公式4-6中的信息熵关系。这证实了观察5。

进一步分析，如图4.8(b)所示，不同的融合表征主要影响 $mRb^C$ 。此外，信息熵( $H_{FR}$ )越大， $mRb^C$ 越强。然而， $H_{FR}$ 的变化对 $mRb^L$ 的影响很小。这是因为这些多传感器融合模型都是基于点云的3D目标检测器，并在检测流程的不同阶段融入图像信息。

**观察5(研究问题2):** 在相同的融合顺序下，融合表征中包含的信息越全面，鲁棒性越强。信息的全面性排序为：数据>特征>结果。

**观察6(研究问题2):** 现有不同融合架构的模型的鲁棒性差异主要体现在对图像损坏的鲁棒性上。

#### 研究问题2: 融合结构如何影响鲁棒性?

**回答:**总的来说，不同的融合顺序和融合表征对鲁棒性的影响具有以下特点：

1) 平行融合表现出比级联融合更好的鲁棒性。2) 融合表征中包含的信息熵越大，鲁棒性越强，不同表征的信息熵排序为：数据>特征>结果。3) 现有不同融合架构的模型的鲁棒性差异主要体现在对图像损坏的鲁棒性上。

## 4.6 讨论

基于对研究问题1和研究问题2的探究，本文认为具有以下特征的融合架构可以增强对信号注入攻击的鲁棒性：1) 模态独立，2) 平行融合，以及3) 数据融合。

**模态独立**指的是每个模态都具有能够独立于其他模态完成最终的3D目标检测的能力。这使得即使在面对数据抹除攻击（如本工作讨论的图像隐藏和点云隐藏攻击）时，单一模态仍然可能完成最终目标检测任务。

**平行融合**指的是在融合过程中公平地整合传感器数据到检测模型中，而不是将某一个传感器指定为主要传感器，另一个作为辅助传感器。经验证据表明，如果在融合过程中对某一模态产生偏向，模型会变得更容易受到针对该模态的攻击。此外，由于权重不足，辅助传感器在有效纠正结果方面面临挑战。

**数据融合**指的是在原始数据层面进行融合，而不是在特征或结果层面。这是因为原始数据保留了更全面的信息，实验结果也证实了数据融合相比其他融合方法具有更强的鲁棒性。然而，在融合激光雷达和相机数据时，由于点云和图像的异构性质，这个过程遇到了挑战，这阻碍了直接的数据融合。

## 4.7 本章小结

本章提出了第一个针对多传感器融合模型的信号注入攻击鲁棒性综合基准测试，设计了包含 11 种信号注入攻击的新数据集 SIA-KITTI 来实现这一目标。本章设计并进行了严格的系统性文献综述和攻击能力量化，以尽可能确保 SIA-KITTI 数据集的全面性和物理可行性。基于对 7 个多传感器融合模型和 5 个单模态模型的 542736 帧数据的评估，本章回答了两个开放性研究问题：研究问题 1：融合是否增强鲁棒性？本章发现，当考虑来自多个传感器的信号注入攻击时，大多数融合模型反而降低了整体鲁棒性。这一发现挑战了以往研究的一致认识。研究问题 2：模型架构如何影响鲁棒性？本章采用了一种新的范式来对模型进行分类，并引入了信息熵的概念，这意外地揭示了模型架构与鲁棒性之间的关系，即融合模态的信息熵越大，鲁棒性越强。最后，本章为增强鲁棒性提供了一些见解，即具有模态独立、平行融合和数据融合特点的多传感器融合模型具有更强的鲁棒性。该基准测试可用于帮助评估和改进多传感器融合模型的性能。

## 5 信号注入攻击检测和防护关键技术研究

针对信号注入攻击引起的传感器数据受损，本文第四章进行了详细的调研，形成了包含 6 种图像受损和 5 种点云受损的受损数据集。本章基于上述研究内容，提出面向自动驾驶中激光雷达感知系统的安全防护方法，能够针对信号注入攻击下的数据受损实现主动式攻击检测和被动式鲁棒性增强。本章一共提出了三种数据受损检测方法：(1) 基于点云表面曲率的虚假目标检测，本文发现由于基于激光的点云注入攻击能力的固有局限性，虚假目标和真实目标之间在表面曲率上存在显著差异，基于该发现实现了对虚假目标注入攻击的检测；(2) 基于一致性分析和时序预测的受损模态检测，利用了未受损时不同传感器之间的语义一致性以及传感器数据的时间连续性，实现了对受损模态的检测；(3) 基于视觉语言模型的受损类型检测分析，利用了视觉语言预训练模型的强大基础能力，通过小样本有监督微调和检索增强生成的方式，实现了对第四章中 11 类数据受损方式的检测和分析。在实现攻击检测的基础上，为了切实提升自动驾驶中激光雷达感知的鲁棒性，本章基于虚拟点技术提出了具备模态独立、平行融合、数据融合三个特征的多传感器融合架构 SIA-Defense (Defense for Signal Injection Attack)，实现了对信号注入攻击下的数据受损的防护，平均鲁棒性比现有 SOTA 模型提升了 5%，尤其是对点云抹除和目标注入等针对激光雷达的攻击的鲁棒性提升了超过 20%。

### 5.1 本章引言

自动驾驶依赖传感器及后续算法对周围环境进行感知，传感器负责接收物理信号并生成数据，感知算法则对传感器数据进行处理生成感知信息。因此，正确的感知是自动驾驶安全行驶的基础。然而攻击者能够通过信号注入攻击的形式，利用传感器本身的脆弱性对传感器的数据造成损坏，使自动驾驶系统无法正确感知到周围行人、汽车等重要目标的信息，给自动驾驶带来严重威胁。本文第四章充分调研了信号注入攻击可能对激光雷达、摄像头造成的数据损坏，总结了 11 种数据损坏类型，包含 6 种图像损坏和 5 种点云损坏，并形成了受损数据集 SIA-KITTI。

之前的许多研究<sup>[19,24,91,154]</sup>认为传感器融合是一种对信号注入攻击的防御策略，但本



图 5.1 安全防护研究内容

文第四章通过大规模的实验，发现即使是 SOTA 的多传感器融合模型在面对信号注入攻击下的数据损坏时，虽然相比于单传感器模型有一定的鲁棒性提升，但仍存在性能下降的问题，尤其是面对点云抹除、点云注入等强大的数据受损效果效果时，模型性能会下降 40% 以上。有一些工作开始关注到对信号注入攻击的防御，比如，Cheng 等人<sup>[187]</sup>通过引入麦克风传感器和对抗训练的方式来防御基于超声波的图像模糊攻击；Sun 等人<sup>[31]</sup>通过分析对抗点云的物理规律和遮挡情况进行欺骗点检测；Xiao 等人<sup>[188]</sup>利用神经网络学习点云的局部特征构造分类器用以区分真实点云和注入点云。但这些工作会引入额外的硬件<sup>[187]</sup>并且仅能针对单一的攻击类型<sup>[31,188]</sup>进行防御。

为了利用主流自动驾驶车辆的已有传感器硬件（激光雷达和摄像头）实现对信号注入攻击下数据受损的全面防御，本章的主要研究内容如图 5.1 所示，首先为了满足特定场景下不同的检测需求，提出了三个维度的攻击检测方法，然后在实现攻击检测的基础上，提出了基于虚拟点技术的模态独立、平行式、数据级多传感器融合架构 SIA-Defense，实现了鲁棒性的提升。

三类攻击检测方法如下：(1) **基于点云表面曲率的虚假目标检测**，针对本文第 2 章提出的基于激光的“目标注入”攻击，观察到欺骗点云和真实点云之间存在形状上的固有差异，受第 2 章攻击能力量化的启发，本章通过表面曲率表征该形状差异，发现了物理世界注入的欺骗点云存在表面曲率过大的固有限制，进而实现了“目标注入”攻击的针对性检测。(2) **基于一致性分析和时序预测的受损模态判定**，未受损时不同传感器之间具有语义一致性，同一传感器数据具有时空连续性。语义一致性指的是点云数据和图像数据的虽然形式不同，但是它们的数据描述的场景是一致的。时空连续性指的是同一传感器在时间维度上相邻帧之间的数据变化应该是平滑的，因此可以利用过去的数据来预测当前甚至未来的数据。本章利用这两个特性判定图像和点云数据是否受损。(3) **基于视觉语言模型的受损类型检测**，利用视觉语言模型对受损数据进行语义理解，从而识别出受损类型。

于视觉语言模型的受损类型检测，利用了现有视觉语言大模型（VLM）的强大推理和泛化能力，设计了传感器受损数据检测与分析智能体 SIA-Agent。首先构建了多模态问答数据集 SIA-VLM-KITTI，并利用该数据集对预训练的 VLM 进行有监督微调，能够通过视觉问答的方式实现对所有数据受损进行检测与分类。进一步地，本文构建了数据受损向量知识库，能够帮助 SIA-Agent 通过检索增强生成（RAG）的方式进行受损原因分析。SIA-Agent 能够给出受损模态、受损类型、判断依据、受损原因等信息，这些信息可为自动驾驶的故障诊断提供决策依据。

在实现攻击检测后，如何有效利用检测结果提升鲁棒性是个挑战。第 4.6 章的实验结论表明，具备模态独立、平行融合、数据融合三个特征的融合架构具有更强的鲁棒性。但由于点云数据和图像数据存在异质性，数据级的直接融合存在困难。本章利用基于 PENet（点云主导）和 PSMNet（图像主导）的虚拟点技术，结合上述攻击检测方法，实现了具备模态独立、平行融合、数据融合三个特征的多传感器融合架构的设计，该架构被称为 SIA-Defense。

最后，本章通过实验评估证明了 SIA-Defense 的平均鲁棒性达到了 0.967，相比于现有 SOTA 模型提升了 5%，尤其是对点云抹除和目标注入等针对激光雷达的攻击的鲁棒性提升了超过 20%。

本章的主要贡献如下：

- **基于表面曲率的虚假目标检测。**通过分析激光注入的虚假目标点云和真实目标点云之间在表面曲率上的差异，实现了对虚假目标注入攻击的检测。
- **基于一致性分析和时序预测的受损模态判定。**利用了未受损时不同传感器之间的语义一致性和同一传感器数据的时空连续性，实现了对图像和点云数据是否受损的判定。
- **基于视觉语言模型的数据受损检测。**提出了数据受损检测分析智能体 SIA-Agent。构建了多模态问答数据集 SIA-VLM-KITTI，利用该数据集对 SIA-Agent 进行有监督微调，通过视觉问答的方式实现对数据受损的检测与分类。构建数据受损向量知识库，能够帮助 SIA-Agent 通过检索增强生成的方式进行受损原因分析。
- **基于多传感器融合的鲁棒性提升。**设计了 SIA-Defense，实现了具备模态独立、平行融合、数据融合三个特征的多传感器融合架构，实现了鲁棒性的提升。

## 5.2 威胁模型与防护需求

本节首先介绍了本文关注的威胁模型,包括攻击目标、攻击者需要获取的信息和具备的能力。其次,本节分析了传感器安全防护需求,即在安全防护设计中需要考虑的因素。

### 5.2.1 威胁模型

本文对攻击者作出如下攻击目标和能力建模:

- **攻击目标:** 攻击者的目标是干扰自动驾驶中传感器的输出,使之无法提供准确的环境测量信息,进而影响自动驾驶的感知模型,使自动驾驶的感知出错。
- **信息获取:** 攻击者可能需要知道被攻击设备的型号,并在攻击前获取相同型号的设备进行研究,攻击者可以通过查看用户手册、硬件逆向、测信道探测等方式获取传感器的细节参数。
- **模型黑盒:** 攻击者无需知道被攻击自动驾驶系统中感知模型的参数细节,仅利用传感器本身的脆弱性实现攻击效果。
- **攻击能力:** 攻击者具有光、声、磁等物理信号的操纵能力,能够通过远程注入物理信号的方式实现攻击。

### 5.2.2 防护需求

本文的安全防护设计时需要考虑的因素包括:

- **防护对象与目标:** 本文的防护对象是信号注入攻击下的传感器数据受损,防护的目标是能够对传感器数据受损实现主动式检测和被动式鲁棒性提升。
- **防护成本:** 安全设计人员需要考虑安全性和成本之间的平衡,在保证传感器安全性的同时,尽可能降低防护措施的设计和部署成本,例如尽量避免引入额外的硬件。
- **感知性能:** 安全设计人员需要考虑安全性和可用性之间的平衡,尽可能避免新增防护措施对传感器及感知性能的影响,如测量准确性、实时性、稳定性等,同时要避免引入新的安全风险。

### 5.3 基于点云表面曲率的虚假目标检测

本文第2章提出的基于激光的点云注入攻击具有注入指定形状点云的能力，因此能够实现指定目标的创建攻击，包括直接创建攻击（Direct-Create）和对抗创建攻击（Adv-Create）。其中直接创建攻击注入的虚假目标和真实目标十分相近，且完全符合点云生成的物理规律，难以通过现有的检测方法<sup>[31]</sup>有效检测。

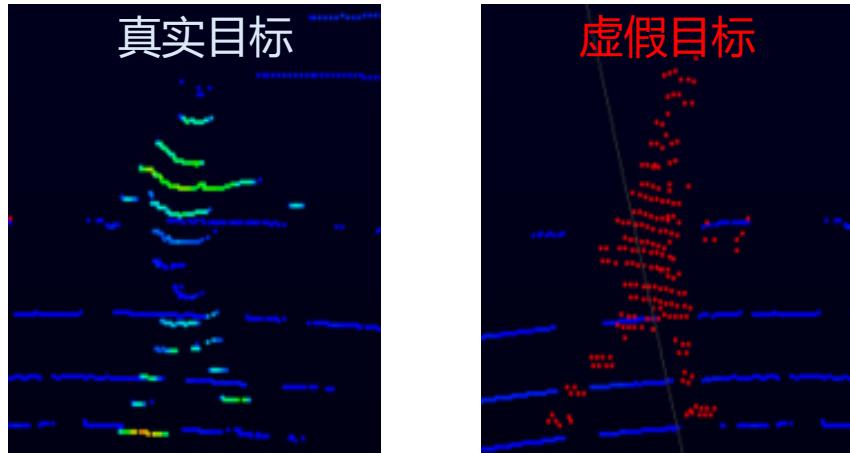


图 5.2 真实行人的点云和注入的虚假行人的点云对比

然而，从第2.5.2章对基于激光的点云注入能力的量化评估可知，由于“光速灾难”问题的存在，即使现有攻击设备将时间误差控制在纳秒级别，注入点云的误差仍然会达到 10 厘米，因此注入的虚假目标点云和真实目标点云之间仍存在可量化的误差。直观地来看，真实行人的点云和注入的虚假行人的点云对比如图5.2所示，真实目标的点云更规则，而注入的目标由于每个点都存在随机的距离误差，因此整体看上去较杂乱。基于这一观察，本章创新性地使用表面曲率这一指标来定量地描述注入目标与真实目标之间的差异，实现了“目标注入”攻击的检测。该方法可以作为一种轻量化的后验方法，对 3D 目标检测结果进行进一步检测，能有效检测出创建的虚假目标。

#### 5.3.1 表面曲率计算

本文首先利用 K 近邻捕获（K-nearest neighbors）的局部邻域信息来估计每个点的曲率。具体来说，该方法计算每个点邻域的协方差矩阵，提取其特征值，并将曲率导出为最小特征值与所有特征值之和的比值。这些值的平均值提供最终的表面曲率。给定点云  $\mathcal{P} = \{\mathbf{p}_i\}_{i=1}^N$ ，对于每个点  $\mathbf{p}_i \in \mathbb{R}^{3 \times 1}$ ，其曲率计算步骤如下：

(1) 邻域质心计算:

$$\bar{\mathbf{p}}_i = \frac{1}{K} \sum_{j=1}^K \mathbf{p}_j^{(i)}, \quad (5-1)$$

其中  $\{\mathbf{p}_j^{(i)}\}_{j=1}^K$  是  $\mathbf{p}_i$  的 K 近邻点集

(2) 协方差矩阵构建:

$$\Sigma_i = \frac{1}{K-1} \sum_{j=1}^K (\mathbf{p}_j^{(i)} - \bar{\mathbf{p}}_i)(\mathbf{p}_j^{(i)} - \bar{\mathbf{p}}_i)^\top. \quad (5-2)$$

(3) 特征值分解:

$$\Sigma_i = \mathbf{U}_i \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \mathbf{U}_i^\top, \quad \lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0. \quad (5-3)$$

其中  $\lambda_1$  为最大特征值, 表征点云在主要分布方向上的方差, 反映点集在三维空间中的最长延伸维度。 $\lambda_3$  为最小特征值, 指示点云在法线方向上的方差, 反映局部表面的平坦程度。若 K 个邻近点在同一平面上,  $\lambda_3 \rightarrow 0$ ; 当 K 个邻近点构成的表面存在曲率变化时,  $\lambda_3$  值增大。

(4) 单点曲率计算:

$$c_i = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}. \quad (5-4)$$

最终表面曲率为所有点曲率的均值:

$$C_{\text{surface}} = \frac{1}{N} \sum_{i=1}^N c_i. \quad (5-5)$$

### 5.3.2 实验评估

本章在 SIA-KITTI 数据集中随机挑选了 100 个真实物体(包括汽车和行人)以及 100 个注入物体, 将邻近点数 K 设定为 30, 分别计算这些物体的表面曲率, 结果如图 5.3 所示。显然, 真实物体的表面曲率明显低于注入物体。因此, 可以采用基于规则的方法进行防御: 在融合模型输出其检测结果后, 本文提取检测到的物体, 计算它们的表面曲率, 并使用经验阈值(记为  $SC_{th}$ )进行过滤。

为了验证该方法的有效性, 本文将其集成到 SECOND 模型中, 并将  $SC_{th}$  设置为 0.76。本文测试了该检测方法对第 2 章提出的 Rec-Create 攻击的防护效果, 实验结果显示

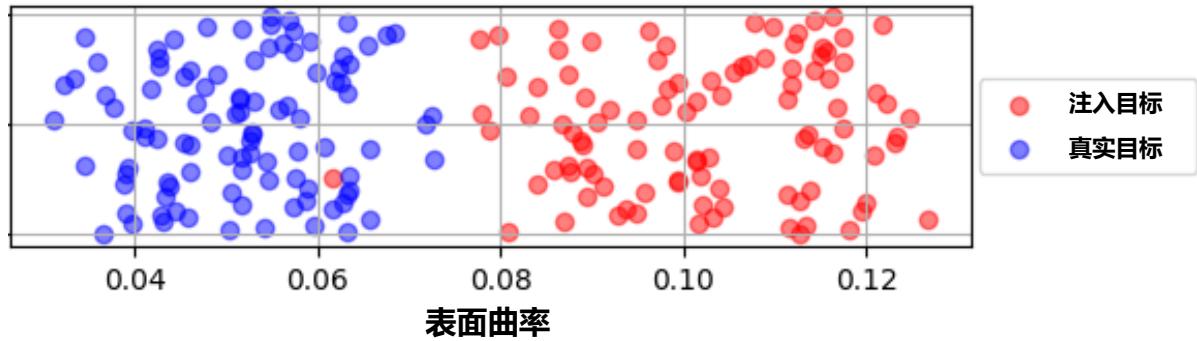


图 5.3 注入目标与真实目标的表面曲率

示，该方法能够有效检测出注入的目标，将 Rec-Create 的攻击成功率从 98% 降低到 1%。该实验证明了基于表面曲率的方法能够有效检测出注入的目标，并且能够很方便的集成到现有模型中。

## 5.4 基于一致性分析和时序预测的受损模态检测

在基于多传感器融合的感知中，需要判断某一个传感器是否受损，以采取相应的应对措施，比如调低置信度或是舍弃该传感器数据<sup>[189]</sup>。为了实现对受损模态的检测和判定，本文设计了数据受损检测器，能够对激光雷达和摄像头的感知数据进行受损检测和判定。数据受损检测器主要包括以下两个部分：

(1) **数据一致性检测。** 数据一致性检测利用了自动驾驶传感多样的优势，虽然激光雷达的点云数据和摄像头的图像数据的形式不同，但是它们的数据之间有强相关性，比如，当激光雷达扫描到正前方 5 米处有行人的点云时，摄像头也会拍摄到正前方 5 米处行人的图像。因此，可以利用点云和图像之间的强相关性进行互相验证，初步判断数据是否受损。

(2) **受损模态判定。** 在实际自动驾驶任务中，同一传感器在时间维度上相邻帧之间的数据变化应该是平滑连续的，因此可以利用过去的数据来预测当前甚至未来的数据。在数据一致性检测出存在数据受损的前提下，若某一个模态预测的当前数据和实际数据之间的差异较大，则认为该模态受损。

### 5.4.1 数据一致性检测

数据一致性检测的流程如下，首先利用 PSMNet<sup>[190]</sup>将图像转化为虚拟点，PSMNet 是一种立体匹配模型，可用于从双目图像中估计深度信息，生成高精度的视差图，进而可生成虚拟点。然后判断虚拟点和真实点云之间的一致性，由于虚拟点和真实点之间存在点密度的差异，本文通过“找对应点”的方式来判断两者之间的语义一致性 (semantic consistency)：对于每一个虚拟点  $P_{Fake}$ ，寻找以  $P_{Fake}$  为中心的半径为  $R$  的球形区域内，是否存在真实点  $P_{Real}$ ，如果存在这样的  $P_{Real}$ ，则认为该  $P_{Fake}$  是“好点”，反之则认为该  $P_{Fake}$  是“坏点”。本文在 KITTI 数据集中随机选取了 200 组正常数据进行计算，发现好点数的占比均在 97% 以上。进一步地，在 SIA-KITTI 数据集中选取了不同数据受损各 100 组（一共 1100 组）数据，计算得到好点数占比均在 92% 以下，其中“图像截断”的好点数占比最低，仅为平均 53.2%。因此，本文选取经验性阈值  $T = 95\%$ ，若好点数占比小于阈值  $T$ ，认为存在数据受损。

### 5.4.2 受损模态判定

数据一致性检测能够初步判断传感器数据是否受损，为了进一步判断是点云还是图像受损，本文利用时序预测的方式判断受损模态。在实际自动驾驶任务中，数据的输入往往是时间和空间连续的，所以可以利用过去的数据来预测当前甚至未来的数据，本文通过比较预测数据和实际采集到的数据之间的差异来判断受损模态。本文受损模态判定流程参考相关工作 LIFE<sup>[191]</sup>的设计，但与 LIFE 利用 PredNet<sup>[192]</sup>来预测当前图片不同，本文采用轻量级的 VDA (Video Depth Anything) 模型<sup>[193]</sup>来对图片和深度图做预测，VDA 引入了轻量级时空头 (Spatio-Temporal Head, STH) 和时间梯度匹配损失 (Temporal Gradient Matching Loss, TGM)，在提高性能的同时将延迟控制在了 9ms 内，相比之下 PredNet 的预测延迟为 35 到 50ms<sup>[191]</sup>。

受损模态判定流程如图 5.4 所示，分为数据预处理、数据预测、受损判定三个步骤：(1) 在数据预处理中，主要是对点云做处理，利用分层插值算法<sup>[191]</sup>将点云转化为深度图；(2) 在数据预测中，对于点云深度图和图像数据，均基于过去 9 帧的连续数据，利用 VDA 模型预测当前帧；(3) 在受损判定中，本文首先去除深度图和图像边缘的 10 个像素距离的内容，以排除突然进入感知域的物体对判定结果的影响。对于点云深度图，采用绝对

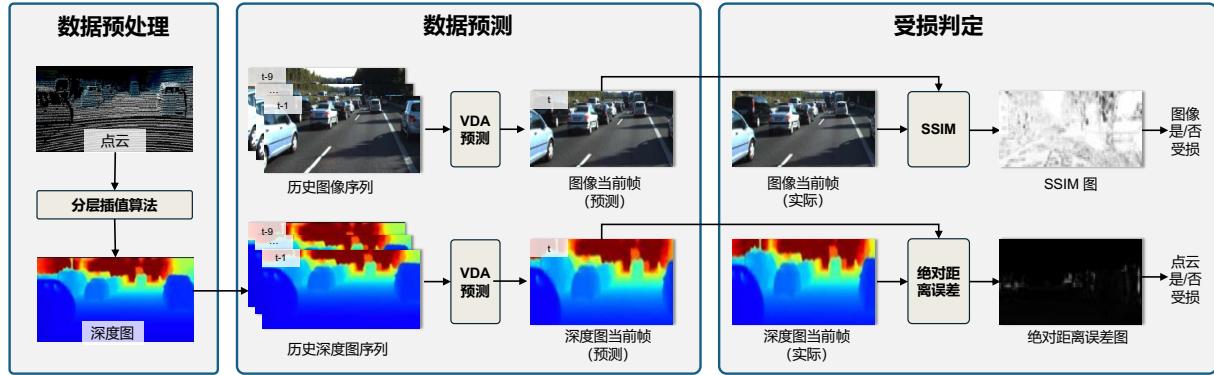


图 5.4 受损模态判定流程图

误差度量预测值与实际值的差异，为降低物体边界未对齐引起的误差，本文对绝对误差图施加  $3 \times 3$  最小值滤波处理。当绝对误差超过 0.1 米的像素点超过三分之一或绝对误差超过 0.5 米的像素点超过十分之一时，判定点云受损。对于图像，计算预测数据和真实数据之间的结构相似度（structural similarity，SSIM），SSIM 指数的取值区间为  $(0,1]$ ，其中 1 仅当两幅图像在滑动窗口内完全相同时可达，0 表示无结构相似性。当 SSIM 指数低于 0.5 的像素点超过三分之一或者低于 0.3 的像素点超过十分之一时，判定图像数据受损。

### 5.4.3 实验评估

#### 5.4.3.1 数据集

本文采用第4章设计的 SIA-KITTI 数据集进行评估，该数据集包含 12 组数据，包括 1 组原始（干净）数据和 11 组受损数据，每组数据包含 3,769 帧点云-图像配对数据。对于受损检测器性能评估所需的历史数据序列，本文从 KITTI 中选取场景流，对于每一帧都选取了过去 9 帧的连续数据作为额外的历史数据序列。

#### 5.4.3.2 评估指标

对于数据受损检测器的性能，需要其能正确区分受损数据和正常数据，因此本文采用召回率 Recall 来表示受损数据的检测成功率，采用误警率 FPR (False Positive Rate) 来表示将正常数据误分类为受损数据的概率。记一组评估后正确检测到的数据受损帧数为 TP，漏检的数据受损帧数为 FN，将正常数据误判为受损数据的帧数为 FP，正确判

表 5.1 SIA-Defense 架构平均鲁棒性

数据受损 方式	图像受损						点云受损				
	像素 饱和	目标 投影	彩条 注入	图像 截断	色带 丢失	运动 模糊	点云 抹除	目标 注入	杂点 注入	噪点 生成	点云 干扰
召回率	100%	100%	100%	100%	100%	100%	99.3%	98.3%	98.7%	96.3%	87.2%

断正常数据的帧数为  $TN$ , 那么 Recall 和 FPR 的计算公式如下:

$$Recall = \frac{TP}{TP+FN}, \quad (5-6)$$

$$FPR = \frac{FP}{FP+TN}. \quad (5-7)$$

需要注意的是, 本次评估中, 检测器仅需成功检测出受损模态即算检测成功, 而无需检测出具体的受损类型, 如对于“像素饱和”受损, 仅需检测出是图像受损即算检测成功。

#### 5.4.3.3 数据受损检测器性能评估

本文基于 3600 组数据进行数据受损检测器性能评估, 包括 300 组正常数据和 3300 组受损数据。综合 3600 组实验结果, 误警率 FPR 为 0.67%, 仅有 2 组正常数据被误认为受损, 通过观察这两组数据可以发现, 这两组数据均有明显的光照明暗变化, 这一定程度上属于自然信号注入攻击引起的图像受损。召回率的结果如表??所示, 其中图像受损的平均召回率  $Recall_{Camera} = 100\%$ , 点云受损的平均召回率为  $Recall_{LiDAR} = 97.1\%$ , 点云受损中点云干扰的召回率较低, 这主要因为点云干扰仅仅是在点云的径向距离上添加小于 15 厘米的误差, 受损程度较低, 因此较难被检测到。

综上实验结果, 表明数据受损检测器能够以 0.67% 的误警率和平均 98.2% 的召回率区分受损数据和正常数据, 展现出较强的检测性能。

## 5.5 基于视觉语言模型的受损数据检测

大语言模型 (Large Language Models, LLMs) 近年来展现出卓越的上下文理解、逻辑推理与答案生成能力。通过与多模态模型的集成, 实现了图像、文本、视频、点云等多源数据的统一特征空间映射, 这种技术融合显著提升了系统的泛化能力, 使其具备快速适应新型任务的迁移学习能力。



图 5.5 SIA-Agent 的工作流程示意图

一个自然的想法是将这些能力赋能于自动驾驶系统，通过将语言模型与基础视觉模型相结合，可突破传统自动驾驶系统在开放场景理解、因果推理和小样本学习方面的局限<sup>[194]</sup>。最近已有大量工作研究了大模型在自动驾驶任务中的应用，包括感知<sup>[195]</sup>、规划<sup>[196-197]</sup>和问答。值得注意的是，由于现有硬件条件下大模型的推理速度过低<sup>[196]</sup>，难以直接承担自动驾驶任务，因此有研究关注将大模型当作现有自动驾驶系统的辅助工具，以副驾驶<sup>[198]</sup>或是慢系统<sup>[196]</sup>的身份承担部分推理和分析工作。综上，现有工作主要关注于利用大模型提升自动驾驶在正常数据下的性能，本文创新性地将其应用于自动驾驶的安全防护领域，重点探究视觉语言模型（Vision Language Model, VLM）在传感器受损数据检测与分析任务中的推理能力，本工作可作为可行性探究对已有研究进行补充。

本文提出了 SIA-Agent，一个基于视觉语言模型的传感器受损数据检测分析智能体。SIA-Agent 的基本工作流程如图 5.5 所示，通过接收点云和图像数据以及用户的问题，SIA-Agent 能够结合数据库中的信息（包含传感器攻击论文、开源代码库、攻击能力参数等）以思维链的方式进行推理，给出数据是否受损、受损类型和受损原因等信息。

本文通过提示工程（Prompt Engineering）和有监督微调（SFT）的方式，对现有的预训练视觉语言模型（如 Llama 3.2、Qwen-VL）进行领域适配，使智能体具备了信号注入攻击下受损数据的检测和分析能力。相比于传统机器学习方法，本方案具有以下两个优势：1) 检测能力泛化优势，VLM 预训练模型本身具有强大的泛化能力，能够通过提示词工程或少量数据的有监督微调<sup>[199]</sup>，即可实现多种类型的受损数据检测，而传统机器学习方法需要大量数据和复杂的特征工程；2) 推理能力优势，能够解释其决策过程，除了数据是否受损、受损类型等分类信息外，还能够给出其判断依据并分析受损原因，这些信息可为自动驾驶的故障诊断提供决策依据。

### 5.5.1 任务定义

SIA-Agent 的任务定义为视觉问答（Visual Question Answering, VQA）任务，其输入-输出对可形式化表示为  $\mathbb{D} = (\mathbf{V}, \mathbf{Q}, \mathbf{A})$ 。其中， $\mathbf{I} = \{Img, \mathbf{PD}_{proj}\}$  包含相机原始图像  $Img \in \mathbb{R}^{H \times W \times 3}$  和点云投影  $\mathbf{PD}_{proj} = \{PD_{bev}, PD_{front}\}$ ， $\mathbf{Q}$  为语义查询问题，回答  $\mathbf{A}$  为包含以下要素的结构化检测信息：

- 受损判定： $\{0, 1\}$  二分类（正常/受损）
- 受损模态： $\mathbf{M} \in \{\text{图像, 点云, 双模态}\}$
- 受损类型： $\mathbf{T} \in \{\text{像素饱和, ..., 点云抹除, ..., 其他}\}$
- 诊断依据：图像或点云受损的判断依据
- 受损原因：物理激励源  $\mathbf{S} \in \{\text{激光, 电磁, 超声波}\}$  和传感器数据受损原理

### 5.5.2 数据集构建

为了对 VLM 进行领域适配，需要构建一个包含信号注入攻击受损数据以及有效文本数据的多模态问答数据集 SIA-VLM-KITTI。其中受损数据采用 SIA-KITTI 数据集（详见第四章），包含 6 种图像损坏和 5 种点云损坏。由于现有的 VLM 预训练模型通常仅针对图像和文本数据，无法直接处理点云数据，因此需要将 SIA-KITTI 数据集的点云转换为图像数据，本章将点云数据通过投影的方式转化为二视图（以俯视图为主，前视图为辅），以适应 VLM 的输入要求。除此以外，需要为每一组数据做人工标注，包括问题、答案和语义约束。

### 5.5.3 提示词设计

本工作通过分层提示（Hierarchical Prompting）的方式，通过系统提示、推理模板和上下文增强三个组件的协同作用，引导 VLM 完成检测和分析任务。该方法能够有效激发大模型中存储的领域知识，同时增强决策过程的可解释性。

### 5.5.3.1 系统提示设计

系统提示的作用是引导 VLM 完成任务，包括自我认知、目标对齐和输出规范。系统提示采用角色-任务-约束三元组的形式：

$$\mathbf{Prompt}_{\text{sys}} = \underbrace{\text{角色定义}}_{\text{自我认知}} \oplus \underbrace{\text{任务描述}}_{\text{目标对齐}} \oplus \underbrace{\text{约束条件}}_{\text{输出规范}} \quad (5-8)$$

本文所用的系统提示词如下：

“系统提示：” {  
 { “角色定义”：“你是一个自动驾驶传感器数据分析专家，精通激光雷达、摄像头传感器的工作原理和光、声、磁等物理信号的作用机理，能够通过观察图像数据和点云数据分析数据是否受损，以及受损原因” } ,  
 { “任务描述”：“通过点云、图像数据分析，判断数据是否受损，并结合数据库知识解释受损原理” } ,  
 { “约束条件”：“最终输出需按照 **A** 的格式，包含受损判定、受损模态、受损类型、诊断依据、受损原理等信息” } }

### 5.5.3.2 基于思维链的多模态提示模板

设计思维链实现任务分解：

$$\mathbf{Prompt} = \underbrace{[I; T]}_{\text{数据输入}} \oplus \left[ \text{数据理解} \rightarrow \text{受损分析} \rightarrow \text{受损检测} \rightarrow \text{根因推断} \right] \quad (5-9)$$

其中数据理解利用了 VLM 对图像和文本数据的理解能力，有许多工作利用了 VLM 的多模态理解能力来实现自动驾驶感知任务<sup>[194]</sup>，本文通过让 VLM 先观察图像中关键目标的方式使其预先了解数据特征，以便于后续任务执行。

用户提示词设计如下：

“用户问题：” {  
  { “数据理解”：“仔细观察摄像头拍摄到的图片序列 <images\_paths>、激光雷达的点云序列俯视图 <PDbev\_paths> 和前视图 <PDfov\_paths>，当前场景有哪些重要目标？” } ,  
  { “受损分析”：“判断图像或点云数据是否存在受损？首先从图像历史数据分析当前帧是否异常，然后从点云历史数据分析当前帧是否异常，最后通过多模态对比分析判断受损模态。” } ,  
  { “受损判定”：“基于上述分析结果，列出受损模态、受损类型、诊断依据。” } ,  
  { “根因推断”：“基于上述结果，检索 <受损数据知识库> 中的受损原因分析，解释数据受损的可能原理。” } }

## 5.5.4 检索增强生成

检索增强生成 (Retrieval-augmented generation, RAG) 是一种通过从特定的数据源中获取信息来提高大模型的准确性和可靠性的方法。当 SIA-Agent 进行数据受损的“根因推断”时，会以 RAG 的方式通过检索数据库中的内容，增强推理的准确性。

### 5.5.4.1 知识库构建

为了实现 RAG，需要首先构建知识库。构建知识库的流程如图5.6所示。首先获取并整理源数据，本文以数据受损方式为单位，对 23 篇源论文进行精炼和总结，并附上攻击信号设计的开源代码以及攻击能力、受损方式等专家知识<sup>[200]</sup>。其次将知识库中的文档分块 (chunk)，分块的作用是将长文本分割成小块，以便于模型更好地理解和处理，本文通过语义分块的方式进行分块，确保了信息在检索过程中的完整性。最后通过词嵌入 (embedding) 的方式将块中的文本转化为向量表示，以便于后续检索过程中的匹配。通过以上三个步骤，构建了向量形式的信号注入攻击下数据受损的知识库。

### 5.5.4.2 检索增强流程

有了向量形式的知识库后，可以通过检索、增强、生成三步，将知识库中的信息注入到大模型中，以提高推理的准确性和可靠性。如图5.6所示，具体流程如下：(1) 检索：指的是根据用户问题对数据库中的相关内容进行检测。首先，将用户的问题也转化

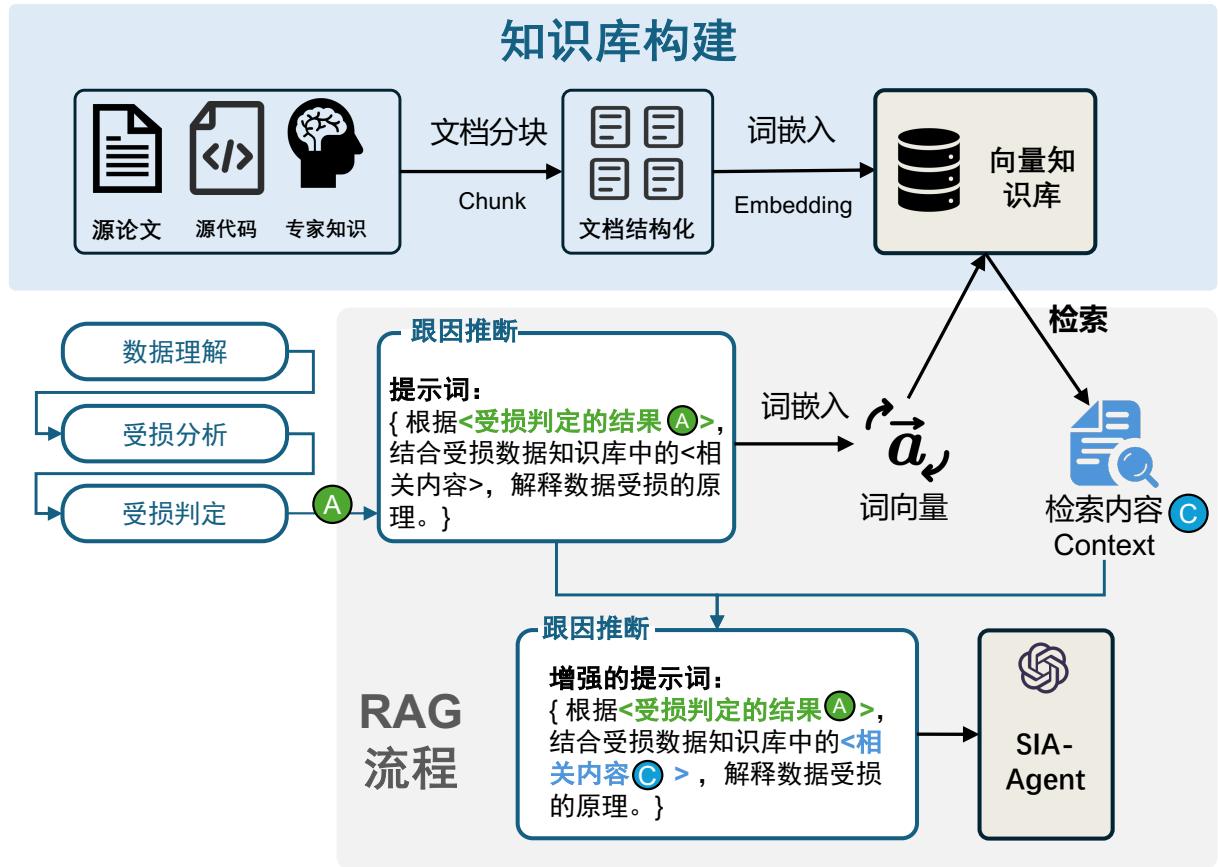


图 5.6 检索增强生成 (RAG) 知识库构建以及工作流程示意图

为向量表示,本文在 RAG 过程中的用户问题是“根因推断”步骤中用户问题,即已通过思维链中的受损检测步骤分析出了受损模态、受损类型和诊断依据;然后,利用向量表示的问题与知识库中的块向量进行匹配,计算余弦相似度得分,根据相似度得分排序,选择前 10 个与问题相关的块作为检索结果。(2) 增强:指的是基于检索结果对提示词进行增强。将检索结果与用户的问题一起,构成新的提示词,用于增强大模型的推理能力。(3) 生成:将增强后的新提示词输入 SIA-Agent,得到最终的推理结果。

## 5.5.5 实验评估

本文对 SIA-Agent 的数据受损检测能力和分析能力进行评估。

### 5.5.5.1 实验设置

本文使用 Qwen2.5-VL-3B 作为基础模型,基于 LoRA 方法<sup>[201]</sup>对其进行有监督微调,训练基于 NVIDIA 3090 显卡,采用混合精度加速策略,优化器采用 AdamW, 学习率

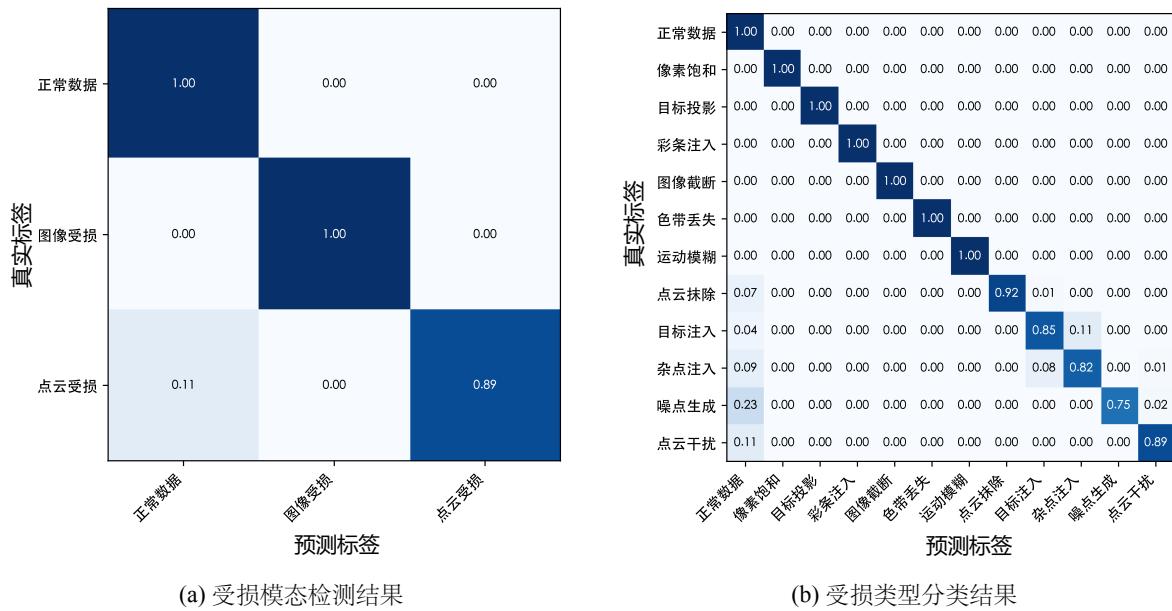


图 5.7 SIA-Agent 的数据受损检测能力评估结果

表 5.2 IA-Agent 数据受损原因分析能力

	图像受损						点云受损					
	像素饱和	目标投影	彩条注入	色带丢失	图像截断	运动模糊	点云抹除	目标注入	杂点注入	噪点生成	点云干扰	
GPT Score	100	99.8	99.7	99.5	100	100	91.3	82.1	79.3	65.2	86.3	

$5 \times 10^{-5}$ , 采用余弦退火调度, 批量大小为 1, 训练轮次为 3。本文在 SIA-VLM-KITTI 数据集上进行训练和评估, 训练集包含 10\*12 组数据, 其中 10 代表每种数据的数量, 12 代表 11 种受损数据和 1 种正常数据。测试集包含 100\*12 组数据。

### 5.5.5.2 数据受损检测能力评估

**评估方式:** 本文从以下 3 个方面判断 SIA-Agent 的检测分析能力: (1) 受损模态检测能力: 判断受损的模态是点云还是图像; (2) 受损类型分类能力: 判断是哪种受损类型; (3) 受损原因分析能力: 判断受损原因。

前两个任务本质上都是多分类任务, 因此本文采用混淆矩阵来直观展示这两个维度的能力。混淆矩阵是一种用于评估分类模型性能的工具, 它以矩阵的形式展示了分类模型的预测结果与真实标签之间的关系。混淆矩阵的行表示真实标签, 列表示预测标签, 矩阵中的每个元素表示预测类别的数量占真实类别数量的比例。混淆矩阵对角线上的元素表示模型正确分类的样本数占总样本数的比例, 又称为召回率 (Recall)。

受损原因分析则是一个文本生成任务，本文基于 GPT-4o 采用 GPT Score 来评估受损原因分析的准确性。GPT Score 的获得方式如下：将生成文本（生成的受损原因分析）和参考文本（真实受损原因）一同输入到 GPT 模型中，让模型根据上下文和语义理解能力，给出一个分数或概率，表示生成文本与参考文本的匹配程度。分数越高，说明生成的内容与参考答案越接近，语义一致性越好。

**实验结果：** SIA-Agent 的受损模态检测能力实验结果如图5.7a所示，可以看到 SIA-Agent 能够以 100% 的召回率检测正常数据和图像受损，能够以 89.2% 的召回率检测点云受损，另外 10.8% 的点云受损会被认为是正常数据，本文认为这一方面由于点云投影成图片后损失了一部分数据，导致数据受损难以被检测到；另一方面由于点云数据的受损相比于图像受损对数据的篡改程度较小，导致点云受损更难被检测到。

SIA-Agent 的受损类型分类能力实验结果如图5.7b所示，可以看到 SIA-Agent 能够以 100% 的召回率分类正常数据和 6 种图像受损，展示了 SIA-Agent 对图像受损的理解能力。对于点云受损，SIA-Agent 对目标注入和杂点注入存在一定的误分类，这主要由于现有的 VLM 模型对点云投影数据的理解能力较弱，不具备细粒度区分点云投影区别的能力。除此以外，发现 SIA-Agent 对于噪点生成的召回率较低（75%），这主要由于噪点生成经过投影后在特定场景下难以被区分。

SIA-Agent 的受损原因分析能力实验结果如表5.2所示，可以看到 SIA-Agent 能够以超过 99.5 的 GPT score 分析图像受损，但对点云受损的（尤其是杂点注入和噪点生成）的分析能力较弱，这主要由于 SIA-Agent 对特定点云受损的检测能力较弱。

综上，SIA-Agent 能够以平均 95.4% 的召回率判断数据受损模态，以 93.5% 的召回率判断数据受损类型，并且以平均 91.2 的 GPT score 分析受损原因。

## 5.6 基于多传感器融合的信号注入攻击防护

为了有效提升自动驾驶激光雷达感知系统在信号注入攻击下的鲁棒性，在实现数据受损高精度检测的基础上，需要考虑如何利用检测结果构建可靠的防护机制，以真正提升感知模型在面对传感器受损时的鲁棒性。为了实现对信号注入攻击造成的传感器数据受损的防护，本文基于上述检测方法，设计了一种多传感器融合框架：SIA-Defense。该框架包括三个部分：1) 数据受损检测器，利用一致性检测和时序预测的方式判断点云

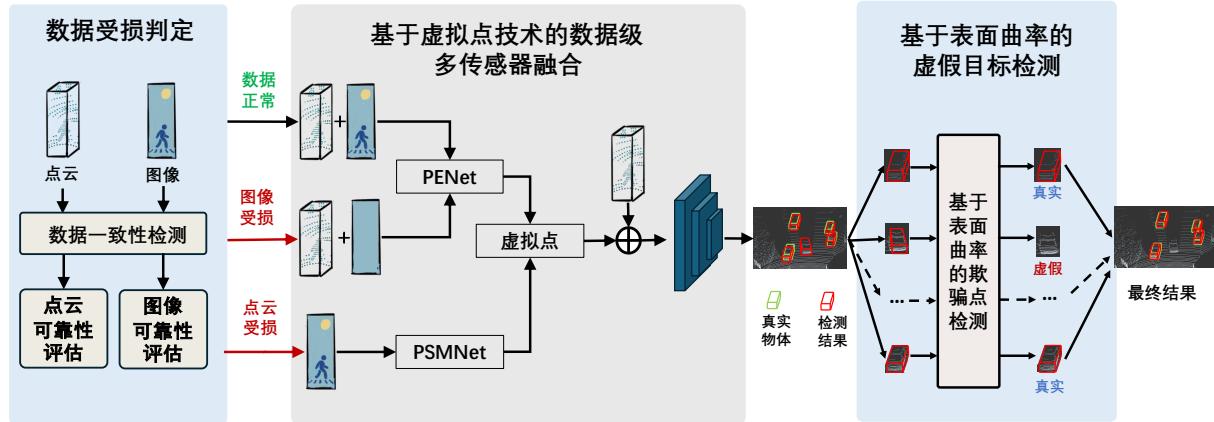


图 5.8 基于虚拟点技术的多传感器融合防护架构 SIA-Defense 示意图

和图像模态是否受损，实现数据受损的先验检测；2）多传感器融合感知模型，利用基于 PSMNet 和 PENet 的虚拟点技术，设计了点云和图像之间数据级、平行式、模态独立的融合范式，实现了鲁棒性的增强；3）欺骗点云检测器，基于欺骗点云和真实点云之间形状上的固有差异，通过计算表面曲率实现欺骗点云的检测和剔除。

### 5.6.1 基于虚拟点技术的数据级多传感器融合

本文第4章的测评结果表明，具备模态独立、平行融合、数据融合三个特征的融合架构具有更强的鲁棒性。但第4章的调研发现，现有模型均没有同时满足上述三个特征，这主要是由于以下两个原因（1）点云数据和图像数据存在异质性，数据级的直接融合存在困难；（2）现有多传感器融合模型主要在基于点云的目标检测模型上进行设计<sup>[6]</sup>，因此点云在 3D 目标检测任务中占主导地位。本文利用基于 PSMNet 和 PENet 的虚拟点技术，基于第5.4章的数据受损检测器的识别结果，设计了点云和图像之间数据级、平行式、模态独立的融合范式，实现了对 SOTA 模型的鲁棒性的增强。

首先通过分析现有 SOTA 模型的优势和不足，介绍本文融合模型的设计动机。基于第4章的评估实验，本文发现当前最先进的 VirConv 模型<sup>[14]</sup>在基于多传感器融合的模型中展现了很强的的鲁棒性 ( $mRb = 0.923$ )，这主要得益于 VirConv 通过基于 PENet 的虚拟点技术实现了平行融合和数据融合。然而表4.4的结果表明，VirConv 的鲁棒性仍存在不足：该模型难以处理“点云抹除”和“目标注入”等点云损坏。这是因为 VirConv 在融合过程中图像和点云两个模态并不是独立的，由于 PENet 的引入，图像数据会依赖于点云数据来生成虚拟点，这使得点云成为了主导数据，因此当点云损坏时，图像不具备

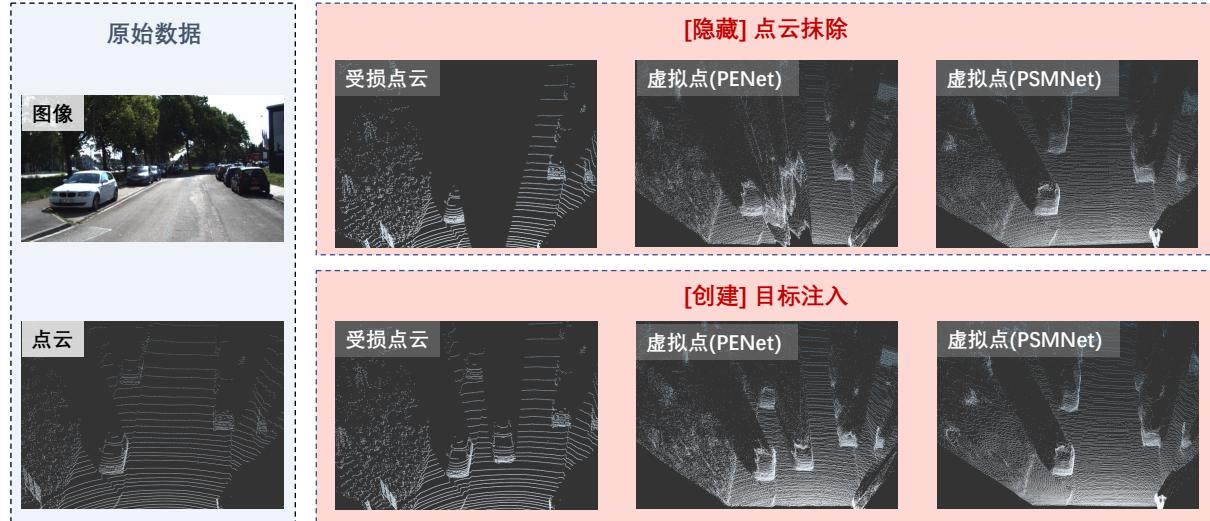


图 5.9 点云数据受损下 PENet 和 PSMNet 虚拟点生成方法对比

独立完成 3D 目标检测任务的能力。这种处理方式的优点是可以通过点云来防御图像损坏，但缺点是点云损坏的效果可能被放大。基于上述分析，本文认为，应该设计一种自适应机制，让模型能够灵活地利用不同模态的数据，从而提升鲁棒性。

本文设计了基于数据受损检测的虚拟点自适应机制。首先，利用数据受损检测器检测数据是否受损，以及哪个模态受损。然后，本文采用两种方式生成虚拟点，第一种是 PENet，该方法综合利用点云和图像数据来生成虚拟点，能够充分利用两种数据的优势，提高模型性能；第二种是 PSMNet，该方法仅利用图像数据来生成虚拟点，能够保证图像模态的独立性，提升模型的鲁棒性。如图 5.9 所示，当点云严重受损时，PSMNet 生成的虚拟点效果优于 PENet。因此，若数据正常，则利用 PENet 生成虚拟点；若点云受损，则利用 PSMNet 生成虚拟点；若图像受损，则依然利用 PENet 生成虚拟点，但图像采用像素值为 (255, 255, 255) 的无内容数据，这是因为通过第 4 章的实验发现，基于 PENet 的方法可以以 0.999 的鲁棒性处理像素饱和，而若直接舍去虚拟点，鲁棒性仅有 0.998。值得注意的是，无论真实点云是否受到攻击，在融合时都选择保留，这是因为真实点云中包含了大量精确的测距信息，这些信息对 3D 目标检测至关重要。

## 5.6.2 实验评估

本文首先对数据受损检测器的性能做评估，然后对 SIA-Defense 整体性能做评估。

表 5.3 SIA-Defense 架构平均鲁棒性

模型架构	图像受损						点云受损				平均鲁棒性 (mRb)	
	像素饱和	目标投影	彩条注入	图像截断	色带丢失	运动模糊	点云抹除	目标注入	杂点注入	噪点生成		
VirConv-T	0.995	1.000	0.993	0.985	0.933	0.958	0.676	0.796	0.910	0.922	0.994	0.923
PSS-Defense (无欺骗点检测)	0.999	0.999	0.999	0.999	0.998	0.999	0.893	0.812	0.919	0.923	0.993	0.958
PSS-Defense	0.999	0.999	0.999	0.999	0.998	0.999	0.893	0.902	0.919	0.923	0.993	0.967

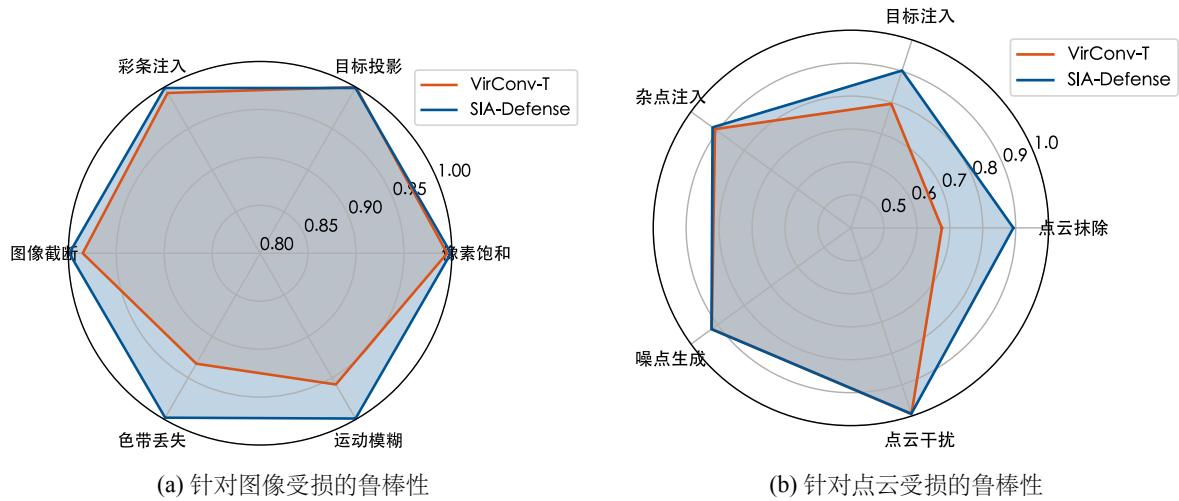


图 5.10 SIA-Defense 与 VirConv-T 鲁棒性对比

### 5.6.2.1 实验设置

本文采用第4章设计的 SIA-KITTI 数据集，该数据集包含 12 组数据，包括 1 组原始（干净）数据和 11 组受损数据，每组数据包含 3,769 帧点云-图像配对数据。对于受损检测器性能所需的历史数据序列，本文从 KITTI 中选取场景流，对于每一帧都选取了过去 9 帧的连续数据作为额外的历史数据序列。

### 5.6.2.2 评估指标

对于 SIA-Defense 整体性能，本文采用平均鲁棒性  $mRb$  指标来评估， $mRb$  表示了多种受损情况下的模型平均精度和正常情况下模型平均精度的比值，在第4.4.2章中已经进行了详细的介绍。

### 5.6.2.3 SIA-Defense 性能评估

本文基于 SIA-KITTI 数据集进行 SIA-Defense 整体性能评估，分别得到了针对 11 种数据受损的鲁棒性以及平均鲁棒性，实验结果如表 5.3 所示。首先，本文通过消融实验证明了欺骗点云检测器的必要性，通过增加欺骗点云检测器，对“目标注入”数据受损的鲁棒性从 0.812 上升到了 0.902，带动平均鲁棒性从 0.958 上升到了 0.967，这主要是由于欺骗点云检测器能够有效剔除注入的虚假目标。然后，本文将 SIA-Defense 框架和目前的 SOTA 模型 VirCon-T<sup>[14]</sup>进行了对比，如图 5.10 所示实验结果表明，SIA-Defense 针对图像受损和点云受损的鲁棒性均强于 VirConv-T。SIA-Defense 的平均鲁棒性达到了 0.967，比 VirConv-T 的 0.923 高了近 5%。SIA-Defense 在 KITTI 标准数据集上的 AP 为 87.730，与 VirConv-T 的 87.731 相近。这主要是由于 SIA-Defense 具备模态独立、数据级、平行式的融合架构，使得其检测到数据受损后能够自适应选择不同的通道，从而有效处理图像受损和点云受损。

上述实验结果证明了 SIA-Defense 架构能够有效提升模型的鲁棒性。

## 5.7 本章小结

本章研究了面向自动驾驶中激光雷达感知系统的安全防护方法，能够针对信号注入攻击下的数据受损实现主动式攻击检测和被动式鲁棒性增强。攻击检测层面，本章一共提出了三种数据受损检测方法：（1）基于点云表面曲率的虚假目标检测，由于基于激光的点云注入攻击能力的固有限制，虚假目标和真实目标之间在表面曲率上存在显著差异，实现了对虚假目标注入攻击的检测，能够将第二章中的直接创建攻击的攻击成功率从 98% 降到 1%。（2）基于一致性分析和时序预测的受损模态检测，利用了未受损时跨模态的语义一致性以及传感器数据的时间连续性，实现了对受损模态的检测，能够以 100% 召回率检测图像受损，以 97.1% 召回率检测点云受损。（3）基于视觉语言模型的受损类型检测分析，利用了视觉语言预训练模型的强大基础能力，通过小样本有监督微调和检索增强生成的方式，实现了对第四章中 11 类数据受损方式的检测分析，能够以平均 95.4% 的召回率判断数据受损模态，以平均 93.5% 的召回率判断数据受损类型。鲁棒性增强层面，在实现攻击检测的基础上，为了切实提升自动驾驶中激光雷达感知的鲁棒性，本章基于虚拟点技术提出了模态独立、平行式、数据级的多传感器融合架

构 SIA-Defense，实现了对信号注入攻击下的数据受损的防护。

## 6 总结与展望

### 6.1 总结

正确感知是自动驾驶车辆安全行驶的重要保障，激光雷达由于其高精度和主动探测的优势，在自动驾驶感知系统中发挥着不可替代的作用。自动驾驶中的激光雷达感知系统是一个传感器和感知算法协同工作的系统，涉及到“信号-数据-信息”的处理和转换。以往自动驾驶中激光雷达感知的安全研究多集中在数字域的算法层面，缺少以“物理信号”为入口的针对传感器本身以及感知系统整体的安全性研究。本文以“自动驾驶中激光雷达感知脆弱性分析与安全防护”为目标，围绕传感器脆弱性挖掘、感知算法鲁棒性测评、安全性增强三个环节展开研究。通过揭示激光雷达传感器潜在的脆弱性、构建全面的算法鲁棒性测评基准、提出切实可行的检测防护方法三个递进维度，形成“脆弱性分析-鲁棒性测评-安全性增强”的完整技术链条。具体而言，本文的主要研究成果总结如下：

#### 1. 基于激光注入攻击的激光雷达感知脆弱性分析

论文第二章探究了激光雷达的信号鉴权脆弱性，提出了可以通过在物理世界使用红外激光注入欺骗点云来对3D目标检测模型直接进行欺骗。本文设计了一种针对激光雷达3D目标检测的物理激光攻击方法——PLA-LiDAR。为了提高点云注入能力，本文开发了一种激光收发器，它能够注入多达4200个欺骗点；为了生成能够被物理注入到目标激光雷达中的对抗性点云，本文提出了一种新的对抗性点云优化方法。在优化过程中，该方法会考虑激光雷达的工作原理、攻击设备的能力以及注入点的距离误差；为了将上述生成的对抗性点云精确注入激光雷达，本文提出了一种“空间坐标-时间坐标”映射的控制信号设计方法和精确到纳秒级别的信号同步方法。基于上述方法，PLA-LiDAR共实现了四种攻击效果，能分别以黑盒和白盒的方式实现隐藏攻击和创建攻击。除此以外，通过物理世界的测量，本文从注入点数、位置控制能力、形状控制能力三方面量化了PLA-LiDAR的攻击能力，能够为其他仿真研究提供物理可实现的参考。通过在2款激光雷达和3款目标检测模型上的数字域和物理域评估，本文验证了对抗点云优化算法的有效性和PLA-LiDAR的物理攻击可行性。最后本文通过在移动的实车上进行实验进

一步证明了 PLA-LiDAR 的物理可行性。

## 2. 基于电磁干扰攻击的激光雷达感知脆弱性分析

论文第三章研究了激光雷达的新型电磁干扰脆弱性，包括新的攻击入口和攻击原理。本文验证了激光雷达的接收模块、监测传感器（温度传感器）和光束转向模块中的光学编码器可作为电磁干扰的攻击入口。本文确定了两个主要攻击原理：1) 直接攻击：干扰接收模块中的模拟信号，直接影响激光雷达的测距机制；2) 间接攻击：攻击激光雷达中的其他附属模块，进而借助故障检测和管理机制间接诱发点云错误或激光雷达本体故障。基于新的攻击入口和攻击原理，本文提出了 PhantomLiDAR 攻击，包含四种针对激光雷达的电磁攻击效果。本文首次设计并提出了基于电磁干扰的点云抹除、点云注入和雷达宕机攻击。此外，与之前的 SOTA 工作相比，本文的攻击能力在测距误差（增加 3 倍）和伪造点数量（增加 5 倍）方面都有显著提高。在 5 个激光雷达和 5 个目标检测模型上进行的大量实验证明了攻击的有效性。PhantomLiDAR 具有攻击距离远、瞄准要求低以及移动场景可行的特点，证明了攻击的实际威胁。此外，本文还讨论了针对电磁干扰攻击的防御对策。本文希望该研究能够通过考虑更广泛的攻击载体来增强未来激光雷达系统的安全性。

## 3. 基于信号注入攻击的感知系统鲁棒性测评基准

论文第四章提出了第一个针对多传感器融合模型的信号注入攻击鲁棒性综合测评基准。首先，本文设计了包含 11 种信号注入攻击的新数据集 SIA-KITTI，本文进行了严格的系统性文献综述和攻击能力量化，以尽可能确保 SIA-KITTI 数据集的全面性和物理可行性。然后，基于对 7 个多传感器融合模型和 5 个单模态模型的 542,736 帧数据的评估，本文回答了两个开放性研究问题：(1) 融合是否增强鲁棒性？本文发现，当考虑来自多个传感器的信号注入攻击时，大多数融合模型反而降低了整体鲁棒性。这一发现挑战了以往研究的一致认识。(2) 模型架构如何影响鲁棒性？本文采用了一种新的范式来对模型进行分类，并引入了信息熵的概念，这揭示了模型架构与鲁棒性之间的关系，即融合模态的信息熵越大，鲁棒性越强。最后，本文为增强鲁棒性提供了一些见解，即具有模态独立、平行融合和数据融合特点的多传感器融合模型具有更强的鲁棒性。该基准测试可用于帮助评估和改进多传感器融合模型的性能。

## 4. 信号注入攻击检测和防护关键技术研究

论文第五章研究了面向自动驾驶中激光雷达感知系统的安全防护方法，能够针对信

号注入攻击下的数据受损实现主动式攻击检测和被动式鲁棒性增强。攻击检测层面，本文一共提出了三种数据受损检测方法：(1) 基于点云表面曲率的虚假目标检测，由于基于激光的点云注入攻击能力的固有限制，虚假目标和真实目标之间在表面曲率上存在显著差异，实现了对虚假目标注入攻击的检测，能够将第二章中的直接创建攻击的攻击成功率从 98% 降到 1%；(2) 基于一致性分析和时序预测的受损模态检测，利用了未受损时跨模态的语义一致性以及传感器数据的时间连续性，实现了对受损模态的检测，能够以 100% 召回率检测图像受损，以 97.1% 召回率检测点云受损；(3) 基于视觉语言模型的受损类型检测分析，利用了视觉语言预训练模型的强大基础能力，通过小样本有监督微调和检索增强生成的方式，实现了对第四章中 11 类数据受损方式的检测分析，能够以平均 95.4% 的召回率判断数据受损模态，以平均 93.5% 的召回率判断数据受损类型。鲁棒性增强层面，在实现攻击检测的基础上，为了切实提升自动驾驶中激光雷达感知的鲁棒性，本文基于虚拟点技术提出了模态独立、平行式、数据级的多传感器融合架构 SIA-Defense，实现了对信号注入攻击下的数据受损的防护，平均鲁棒性比现有 SOTA 模型提升了 5%。

## 6.2 展望

近年来，随着固态传感技术与多模态大模型的快速发展，自动驾驶中激光雷达感知的安全研究面临新的技术挑战与演进机遇。在传感器层面，FMCW（调频连续波）与 OPA（光学相控阵）技术推动激光雷达从机械式向全固态转型；自动驾驶系统架构层面，端到端模型与世界模型技术正重塑自动驾驶系统的安全边界；脆弱性分析和防护层面，多模态大模型的逻辑推理能力和泛化能力为构建自适应防御体系提供了新的可能。在此背景下，结合本文的研究方向，未来可以继续在以下三个方面开展探索和研究：

(1) **新型激光雷达安全问题研究。** 本文研究的目标传感器为当前自动驾驶汽车中广泛使用的机械式和 MEMS 激光雷达。随着技术发展，激光雷达向小型化、固态化发展，在测距方式上，可能从如今的直接飞行时间测距方法过渡到基于 FMCW 的测距方法；在光束偏转方式上，可能逐渐采用光学相控阵技术。新的激光雷达设计可能引入新的攻击面和防护挑战，如 FMCW 的频域调制特性可能面临多普勒欺骗攻击，攻击者通过注入特定频率偏移信号干扰距离解算过程；OPA 的空间波束形成机制则存在波束控制序

列劫持风险，攻击者可能通过电磁注入篡改相控阵元相位参数。

(2) **自动驾驶整车硬件在环的端到端安全问题研究。**本文的研究对象为基于激光雷达的感知系统，包括激光雷达、摄像头传感器以及后续的感知算法。而自动驾驶汽车是一个包含了车机硬件和自动驾驶算法软件的“具身智能”系统，涉及到各个传感器、算法、执行器的相互耦合。未来可构建硬件在环（HIL）测试平台，量化信号注入攻击对预期功能安全（SOTIF）指标的影响。针对端到端模型的黑盒特性，研究基于可解释人工智能的攻击和分析方法，建立“传感器异常-特征偏移-控制指令偏差”的因果推理链，评估复杂系统耦合下的攻击传播效应。

(3) **自动驾驶安全多模态智能体技术研究。**大语言模型近年来展现出卓越的上下文理解、逻辑推理与答案生成能力，通过与多模态模型的集成，实现了图像、文本、视频、点云等多源数据的统一特征空间映射。未来有望基于自动驾驶安全从业者在漏洞挖掘、漏洞分析、安全设计上的知识经验，设计自动驾驶安全多模态智能体，该智能体有望具备以下能力：a) 脆弱性自动挖掘能力，结合检索增强生成和强化学习实现多模态数据驱动的攻击向量自动生成与物理约束验证；b) 跨模态关联推理能力，实现传感器异常信号与系统漏洞的关联分析；c) 自演进防御策略生成能力，基于世界模型的闭环验证实现防护策略的在线优化。

## 参考文献

- [1] YURTSEVER E, LAMBERT J, CARBALLO A, et al. A survey of autonomous driving: Common practices and emerging technologies[J]. IEEE access, 2020, 8: 58443-58469.
- [2] LEONARD J, HOW J, TELLER S, et al. A perception-driven autonomous urban vehicle[J]. Journal of Field Robotics, 2008, 25(10): 727-774.
- [3] (YOLE) Y D. LiDAR for Automotive and Industrial Applications 2025[Z]. <https://www.yolegroup.com/product/report/automotive-lidar-2025/>. September 2021.
- [4] THRUN S, MONTEMERLO M, DAHLKAMP H, et al. Stanley: The robot that won the DARPA Grand Challenge[J]. Journal of field Robotics, 2006, 23(9): 661-692.
- [5] 刘博, 于洋, 姜朔. 激光雷达探测及三维成像研究进展[J]. 光电工程, 2019, 46(7): 190167-1.
- [6] MAO J, SHI S, WANG X, et al. 3D object detection for autonomous driving: A comprehensive survey [J]. International Journal of Computer Vision, 2023, 131(8): 1909-1963.
- [7] QI C R, YI L, SU H, et al. Pointnet++: Deep hierarchical feature learning on point sets in a metric space[J]. Advances in neural information processing systems, 2017, 30.
- [8] LANG A H, VORA S, CAESAR H, et al. Pointpillars: Fast encoders for object detection from point clouds[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12697-12705.
- [9] GEIGER A, LENZ P, STILLER C, et al. Vision meets Robotics: The KITTI Dataset[J]. International Journal of Robotics Research (IJRR), 2013.
- [10] CAESAR H, BANKITI V, LANG A H, et al. nuscenes: A multimodal dataset for autonomous driving [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 11621-11631.
- [11] QIAN R, LAI X, LI X. 3D object detection for autonomous driving: A survey[J]. Pattern Recognition, 2022, 130: 108796.
- [12] QI C R, LIU W, WU C, et al. Frustum pointnets for 3d object detection from rgb-d data[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 918-927.
- [13] VORA S, LANG A H, HELOU B, et al. Pointpainting: Sequential fusion for 3d object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 4604-4612.
- [14] WU H, WEN C, SHI S, et al. Virtual Sparse Convolution for Multimodal 3D Object Detection[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 21653-21662.
- [15] HUANG T, LIU Z, CHEN X, et al. Epnet: Enhancing point features with image semantics for 3d object detection[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. 2020: 35-52.
- [16] KU J, MOZIFIAN M, LEE J, et al. Joint 3d proposal generation and object detection from view aggregation[C]//2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2018: 1-8.
- [17] PANG S, MORRIS D, RADHA H. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection[C]//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS): 10386-10393.
- [18] GIECHASKIEL I, RASMUSSEN K. Taxonomy and challenges of out-of-band signal injection attacks and defenses[J]. IEEE Communications Surveys & Tutorials, 2019, 22(1): 645-670.
- [19] YAN C, SHIN H, BOLTON C, et al. Sok: A minimalist approach to formalizing analog sensor security [C]//2020 IEEE Symposium on Security and Privacy (SP). 2020: 233-248.
- [20] XU Y, HAN X, DENG G, et al. SoK: Rethinking sensor spoofing attacks against robotic vehicles from a systematic view[C]//2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P). 2023: 1082-1100.
- [21] SHEN J, WANG N, WAN Z, et al. Sok: On the semantic ai security in autonomous driving[J]. arXiv preprint arXiv:2203.05314, 2022.

- [22] XIE S, LI Z, WANG Z, et al. On the adversarial robustness of camera-based 3d object detection[J]. arXiv preprint arXiv:2301.10766, 2023.
- [23] LIU D, YU R, SU H. Extending adversarial attacks and defenses to deep 3d point cloud classifiers [C]//2019 IEEE International Conference on Image Processing (ICIP). 2019: 2279-2283.
- [24] CAO Y, XIAO C, CYR B, et al. Adversarial sensor attack on lidar-based perception in autonomous driving[C]//Proceedings of the 2019 ACM SIGSAC conference on computer and communications security. 2019: 2267-2281.
- [25] ZHANG Y, ZHU Y, LIU Z, et al. Towards backdoor attacks against lidar object detection in autonomous driving[C]//Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems. 2022: 533-547.
- [26] LIU H, WU Y, YU Z, et al. Slowlidar: Increasing the latency of lidar-based detection using adversarial examples[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 5146-5155.
- [27] LI X, CHEN Z, ZHAO Y, et al. Pointba: Towards backdoor attacks in 3d point cloud[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 16492-16501.
- [28] PETIT J, STOTTELAAR B, FEIRI M, et al. Remote attacks on automated vehicles sensors: Experiments on camera and lidar[J]. Black Hat Europe, 2015, 11(2015): 995.
- [29] GMBH I A S. ibeo LUX[Z]. <https://www.ibeo-as.com/en/products/sensoren/IbeoLUX>. 2021.
- [30] SHIN H, KIM D, KWON Y, et al. Illusion and dazzle: Adversarial optical channel exploits against lidars for automotive applications[C]//International Conference on Cryptographic Hardware and Embedded Systems. 2017: 445-467.
- [31] 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[C]//29th {USENIX} Security Symposium ({USENIX} Security 20). 2020: 877-894.
- [32] HALLYBURTON R S, LIU Y, CAO Y, et al. Security Analysis of {Camera-LiDAR} Fusion Against {Black-Box} Attacks on Autonomous Vehicles[C]//31st USENIX Security Symposium (USENIX Security 22). 2022: 1903-1920.
- [33] SATO T, HAYAKAWA Y, SUZUKI R, et al. LiDAR Spoofing Meets the New-Gen: Capability Improvements, Broken Assumptions, and New Attack Strategies[J]. Network and Distributed Systems Security (NDSS) Symposium, 2024.
- [34] SATO T, SUZUKI R, HAYAKAWA Y, et al. On the Realism of LiDAR Spoofing Attacks against Autonomous Driving Vehicle at High Speed and Long Distance[C]//Proceedings of the Network and Distributed System Security Symposium (NDSS). 2025.
- [35] TU J, REN M, MANIVASAGAM S, et al. Physically realizable adversarial examples for lidar object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 13716-13725.
- [36] TU J, LI H, YAN X, et al. Exploring adversarial robustness of multi-sensor perception systems in self driving[J]. Conference on Robot Learning (CoRL), 2021.
- [37] ABDELFATTAH M, YUAN K, WANG Z J, et al. Adversarial attacks on camera-lidar models for 3d car detection[C]//2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). 2021: 2189-2194.
- [38] 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[C]//2021 IEEE Symposium on Security and Privacy (SP). 2021: 176-194.
- [39] ZHU Y, MIAO C, HAJIAGHAJANI F, et al. Adversarial attacks against lidar semantic segmentation in autonomous driving[C]//Proceedings of the 19th ACM Conference on Embedded Networked Sensor Systems. 2021: 329-342.
- [40] LOU Y, ZHU Y, SONG Q, et al. A First {Physical-World} Trajectory Prediction Attack via {LiDAR-induced} Deceptions in Autonomous Driving[C]//33rd USENIX Security Symposium (USENIX Security 24). 2024: 6291-6308.
- [41] SAKARIDIS C, DAI D, VAN GOOL L. Semantic foggy scene understanding with synthetic data[J]. International Journal of Computer Vision, 2018, 126: 973-992.
- [42] KAMANN C, ROTHER C. Benchmarking the robustness of semantic segmentation models[C]//

- Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 8828-8838.
- [43] ALTINDIS S F, DALVA Y, PEHLIVAN H, et al. Benchmarking the robustness of instance segmentation models[J]. arXiv preprint arXiv:2109.01123, 2021.
- [44] WANG P, HUANG X, CHENG X, et al. The apolloscape open dataset for autonomous driving and its application[J]. IEEE transactions on pattern analysis and machine intelligence, 2019, 1.
- [45] DIAZ-RUIZ C A, XIA Y, YOU Y, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 21383-21392.
- [46] HENDRYCKS D, DIETTERICH T. Benchmarking neural network robustness to common corruptions and perturbations[J]. arXiv preprint arXiv:1903.12261, 2019.
- [47] LI S, WANG Z, JUEFEI-XU F, et al. Common corruption robustness of point cloud detectors: Benchmark and enhancement[J]. arXiv preprint arXiv:2210.05896, 2022.
- [48] YU K, TAO T, XIE H, et al. Benchmarking the robustness of lidar-camera fusion for 3d object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 3188-3198.
- [49] KONG L, LIU Y, LI X, et al. Robo3d: Towards robust and reliable 3d perception against corruptions [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 19994-20006.
- [50] DONG Y, KANG C, ZHANG J, et al. Benchmarking robustness of 3d object detection to common corruptions in autonomous driving[J]. arXiv preprint arXiv:2303.11040, 2023.
- [51] SUN P, KRETZSCHMAR H, DOTIWALLA X, et al. Scalability in perception for autonomous driving: Waymo open dataset[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2446-2454.
- [52] PFEUFFER A, DIETMAYER K. Optimal sensor data fusion architecture for object detection in adverse weather conditions[C]//2018 21st International Conference on Information Fusion (FUSION). 2018: 1-8.
- [53] KIM J, CHOI J, KIM Y, et al. Robust camera lidar sensor fusion via deep gated information fusion network[C]//2018 IEEE Intelligent Vehicles Symposium (IV). 2018: 1620-1625.
- [54] PARK W, LIU N, CHEN Q A, et al. Sensor Adversarial Traits: Analyzing Robustness of 3D Object Detection Sensor Fusion Models[C]//2021 IEEE International Conference on Image Processing (ICIP). 2021: 484-488.
- [55] WANG S, WU T, CHAKRABARTI A, et al. Adversarial Robustness of Deep Sensor Fusion Models [C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 2387-2396.
- [56] SONG Z, LIU L, JIA F, et al. Robustness-aware 3d object detection in autonomous driving: A review and outlook[J]. IEEE Transactions on Intelligent Transportation Systems, 2024.
- [57] REZAEI M, HUSSEIN L, MOAZENI S. Secure FMCW LiDAR systems with frequency encryption [C]//Proceedings of the 2022 Workshop on Attacks and Solutions in Hardware Security. 2022: 35-43.
- [58] KIM G, EOM J, PARK Y. Design of pulsed scanning lidar without mutual interferences[C]//Smart Photonic and Optoelectronic Integrated Circuits XX: vol. 10536. 2018: 168-173.
- [59] WANG Q, ZHANG Y, XU Y, et al. Pseudorandom modulation quantum secured lidar[J]. Optik, 2015, 126(22): 3344-3348.
- [60] CHANGALVALA R, MALIK H. Sensor data integrity verification for autonomous vehicles using spread 3d dither qim[C]//2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall). 2020: 1-7.
- [61] LONG T, XIE A, REN X, et al. Tampering detection of LiDAR data for autonomous vehicles[C]// 2021 40th Chinese Control Conference (CCC). 2021: 4732-4737.
- [62] HAU Z, DEMETRIOU S, MUÑOZ-GONZÁLEZ L, et al. Shadow-catcher: Looking into shadows to detect ghost objects in autonomous vehicle 3d sensing[C]//Computer Security—ESORICS 2021: 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4–8, 2021, Proceedings, Part I 26. 2021: 691-711.

- [63] 激光注入脆弱性挖掘与利用实验结果展示网站[Z]. <https://sites.google.com/view/physical-lidar-attack>. 2025.
- [64] Apollo[Z]. <https://github.com/ApolloAuto/apollo>.
- [65] ARCFOX BAIC HBT[Z]. <https://www.techgenyz.com/2021/04/07/huawei-lidar-solution-arcfox-baic-hbt-car/>.
- [66] Waymo Driver[Z]. <https://waymo.com/waymo-driver/>.
- [67] NEUVITION I. CVIS and V2X with LiDAR[Z]. <https://www.neuvition.com/media/cvis-and-v2x-with-lidar.html>. December 2020.
- [68] Ouster. Ouster for intelligent transportation systems[Z]. <https://ouster.com/resources/smart-infrastructure-resources/its-lidar-solution-overview/>. 2021.
- [69] LeiShen. V2X Roadside Perception System[Z]. <http://www.lslidar.com/en/solution/41>. 2021.
- [70] Velodyne. Smart City[Z]. <https://velodynelidar.com/industries/smart-city/>. 2021.
- [71] (YOLE) Y D. LiDAR for Automotive and Industrial Applications 2020[Z]. August 2020.
- [72] CAO Y, XIAO C, YANG D, et al. Adversarial objects against lidar-based autonomous driving systems[J]. arXiv preprint arXiv:1907.05418, 2019.
- [73] ZHU Y, MIAO C, ZHENG T, et al. Can We Use Arbitrary Objects to Attack LiDAR Perception in Autonomous Driving? [C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 1945-1960.
- [74] FANG J, YANG R, CHEN Q A, et al. Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks[J]. arXiv preprint arXiv:2106.09249, 2021.
- [75] VELODYNE LIDAR I. VLP-16 User Manual[A]. 2019.
- [76] ROBOSENSE LIDAR I. RS-16[A]. 2022.
- [77] ROBOSENSE LIDAR I. RS-M1[A]. 2022.
- [78] Quanergy. Quanergy S3[Z]. <https://quanergy.com/products/s3/>. 2022.
- [79] (YOLE) Y D. LiDAR for Automotive and Industrial Applications 2021[Z]. September 2021.
- [80] Waymo One[Z]. <https://waymo.com/waymo-one/>.
- [81] Apollo Robotaxi[Z]. <https://www.apollo.auto/robotaxi/index.html>.
- [82] AMANN M C, BOSCH T M, LESCURE M, et al. Laser ranging: a critical review of unusual techniques for distance measurement[J]. Optical engineering, 2001, 40: 10-19.
- [83] VELODYNE LIDAR I. VLP-16 Data Sheet[A]. 2018.
- [84] S5971 Si PIN photodiode[Z]. <https://www.newark.com/hamamatsu/s5971/diode-photo-900nm-to-18-3/dp/62M0262>. 2021.
- [85] Rigol DG5072 arbitrary waveform generator[Z]. <https://www.batronix.com/shop/waveform-generator/Rigol-DG5072.html>.
- [86] VELODYNE LIDAR I. HDL64E S3 Manual[A]. 2019.
- [87] YAN Y, MAO Y, LI B. Second: Sparsely embedded convolutional detection[J]. Sensors, 2018, 18(10): 3337.
- [88] CONTRIBUTORS M. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection[Z]. <https://github.com/open-mmlab/mmdetection3d>. 2020.
- [89] GEIGER A, LENZ P, STILLER C, et al. Vision meets robotics: The kitti dataset[J]. The International Journal of Robotics Research, 2013, 32(11): 1231-1237.
- [90] 电磁干扰脆弱性挖掘与利用实验结果展示网站[Z]. <https://sites.google.com/view/phantomlidar>. 2025.
- [91] JIN Z, XIAOYU J, CHENG Y, et al. PLA-LiDAR: Physical Laser Attacks against LiDAR-based 3D Object Detection in Autonomous Vehicle[C]//2023 IEEE Symposium on Security and Privacy (SP). 2022: 710-727.
- [92] CAO Y, BHUPATHIRAJU S H, NAGHAVI P, et al. You Can't See Me: Physical Removal Attacks on {LiDAR-based} Autonomous Vehicles Driving Frameworks[C]//32nd USENIX Security Symposium (USENIX Security 23). 2023: 2993-3010.
- [93] BHUPATHIRAJU S H V, SHELDON J, BAUER L A, et al. EMI-LiDAR: Uncovering Vulnerabilities of LiDAR Sensors in Autonomous Driving Setting using Electromagnetic Interference[C]// Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks.

- 2023: 329-340.
- [94] INSIGHTS T. Deep Dive Teardown of the Velodyne LiDAR Puck VLP-16 LiDAR Sensor[Z]. <https://www.techinsights.com/products/ddt-1902-808>. 2019.
- [95] SYSTEMPLUS Y. Reverse of Robosense RS-LIDAR-M1[Z]. <https://www.yolegroup.com/product/teardown-track/robosense-rs-lidar-m1-lidar-sample-version/>. 2022.
- [96] TU Y, RAMPAZZI S, HAO B, et al. Trick or Heat?: Manipulating Critical Temperature-Based Control Systems Using Rectification Attacks[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security.
- [97] WANG K, XIAO S, JI X, et al. Volttack: Control IoT Devices by Manipulating Power Supply Voltage [C]//2023 IEEE Symposium on Security and Privacy (SP). 2023: 1771-1788.
- [98] HALTERMAN R, BRUCH M. Velodyne HDL-64E lidar for unmanned surface vehicle obstacle detection[C]//Unmanned Systems Technology XII: vol. 7692. 2010: 123-130.
- [99] INNOVUSION I. innovation-Falcon[A]. 2023.
- [100] ZHANG X, KWON K, HENRIKSSON J, et al. A large-scale microelectromechanical-systems-based silicon photonics LiDAR[J]. Nature, 2022, 603(7900): 253-258.
- [101] INTELLIGENCE Y. LiDAR for Automotive 2023[M]. 2023.
- [102] SHI J W, GUO J I, KAGAMI M, et al. Photonic technologies for autonomous cars: feature introduction[J]. Optics Express, 2019, 27(5): 7627-7628.
- [103] ABID A, KHAN M T, IQBAL J. A review on fault detection and diagnosis techniques: basics and beyond[J]. Artificial Intelligence Review, 2021, 54(5): 3639-3664.
- [104] GOELLES T, SCHLAGER B, MUCKENHUBER S. Fault detection, isolation, identification and recovery (fdiir) methods for automotive perception sensors including a detailed literature survey for lidar[J]. Sensors, 2020, 20(13): 3662.
- [105] HESAI TECHNOLOGY CO. L. STATE DETECTION DEVICE FOR LIDAR, LIDAR, AND STATE DETECTION METHOD: US20220268904A1[P]. 2022.
- [106] ROBOSENSE LIDAR I. Fault Diagnostic Methods, Devices, Storage Media and Lidar: CN 115825931 A[P]. 2023.
- [107] LLC G G T O. LIDAR LASER HEALTH DIAGNOSTIC: US20220236410A1[P]. 2021.
- [108] RADASKY W, SAVAGE E. Intentional electromagnetic interference (IEMI) and its impact on the US power grid[J]. Meta, 2010, 1: 1-3.
- [109] KÖHLER S, BAKER R, MARTINOVIC I. Signal Injection Attacks against CCD Image Sensors[C]//Proceedings of the 2022 ACM Asia Conference on Computer and Communications Security.
- [110] DAYANIKLI G Y, HATCH R R, GERDES R M, et al. Electromagnetic Sensor and Actuator Attacks on Power Converters for Electric Vehicles[C]//Proceedings of the 2020 IEEE Security and Privacy Workshops (SPW).
- [111] SZAKÁLY M, KÖHLER S, STROHMEIER M, et al. Assault and Battery: Evaluating the Security of Power Conversion Systems Against Electromagnetic Injection Attacks[J]. arXiv preprint arXiv:2305.06901, 2023.
- [112] JIANG Q, REN Y, LONG Y, et al. GhostType: The Limits of Using Contactless Electromagnetic Interference to Inject Phantom Keys into Analog Circuits of Keyboards[C]//Network and Distributed Systems Security (NDSS) Symposium. 2024.
- [113] KUNE D F, BACKES J, CLARK S S, et al. Ghost Talk: Mitigating EMI Signal Injection Attacks against Analog Sensors[C]//Proceedings of the 2013 IEEE Symposium on Security and Privacy (SP).
- [114] XU Z, HUA R, JUANG J, et al. Inaudible Attack on Smart Speakers With Intentional Electromagnetic Interference[J]. IEEE Transactions on Microwave Theory and Techniques, 2021, 69(5): 2642-2650.
- [115] FOKKENS T, XU Z, IZADI O H, et al. Machine Learning Voice Synthesis for Intention Electromagnetic Interference Injection in Smart Speaker Devices[C]//2021 IEEE International Joint EMC/SI/PI and EMC Europe Symposium. 2021: 673-677.
- [116] JIANG Q, JI X, YAN C, et al. GlitchHiker: Uncovering Vulnerabilities of Image Signal Transmission with IEMI[C]//Proceedings of the 32nd USENIX Security Symposium (USENIX Security 23). 2023.
- [117] JANG J H, CHO M, KIM J, et al. Paralyzing Drones via EMI Signal Injection on Sensory Commu-

- nication Channels[C]//Proceedings of the 2023 Network and Distributed System Security Symposium.
- [118] ZHU H, YU Z, CAO W, et al. PowerTouch: A Security Objective-Guided Automation Framework for Generating Wired Ghost Touch Attacks on Touchscreens[C]//Proceedings of the 41st IEEE/ACM International Conference on Computer-Aided Design. 2022.
- [119] REDOUTÉ J M, STEYAERT M. EMC of analog integrated circuits[M]. Springer Science & Business Media, 2009.
- [120] WANG K, MITEV R, YAN C, et al. {GhostTouch}: Targeted Attacks on Touchscreens without Physical Touch[C]//31st USENIX Security Symposium (USENIX Security 22). 2022: 1543-1559.
- [121] Kitti Benchmark[Z]. <https://www.cvlibs.net/datasets/kitti/index.php>. 2023.
- [122] N5172B EXG X-Series RF Vector Signal Generator[Z]. <https://www.keysight.com/us/en/product/N5172B/exg-x-series-rf-vector-signal-generator-9-khz-6-ghz.html>. 2024.
- [123] HPA-50W-63+[Z]. <https://www.minicircuits.com/pdfs/HPA-50W-63+.pdf>. 2024.
- [124] 698-2700MHz 16DBi LTE Outdoor Antenna[Z]. <https://www.aliexpress.us/item/3256803272069914.html>. 2024.
- [125] ROBOSENSE LIDAR I. RS-Bpearl[A]. 2023.
- [126] ROBOSENSE LIDAR I. RS-M1P[A]. 2023.
- [127] ZEEKER 007[Z]. <https://www.zeekrlife.com/zeekr007>. 2024.
- [128] Yangwang U8[Z]. <https://www.yangwangauto.com/car-type.html>. 2024.
- [129] Xiaopeng G9[Z]. <https://www.xpeng.com/g9>. 2024.
- [130] RS-M1: Nominated Orders for More Than 60 Models[Z]. <https://www.robosense.cn/en/rslidar/RS-LiDAR-M1>. 2023.
- [131] Unitree Robot H1[Z]. <https://www.unitree.com/cn/h1>. 2025.
- [132] SHI S, GUO C, JIANG L, et al. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10529-10538.
- [133] SATO T, HAYAKAWA Y, SUZUKI R, et al. Revisiting LiDAR Spoofing Attack Capabilities against Object Detection: Improvements, Measurement, and New Attack[J]. arXiv preprint arXiv:2303.10555, 2023.
- [134] SUZUKI R, SATO T, HAYAKAWA Y, et al. WIP: Towards Practical LiDAR Spoofing Attack against Vehicles Driving at Cruising Speeds[J].
- [135] 35MHz-4400MHz RF Signal Generator ADF4351[Z]. <https://www.aliexpress.us/item/3256806804711048.html>. 2024.
- [136] 1.2GHz 50W amplifier[Z]. <https://www.aliexpress.us/item/3256807300998703.html>. 2024.
- [137] USRP B210[Z]. <https://www.ettus.com/all-products/ub210-kit/>. 2024.
- [138] RAZAVI B. Design of analog CMOS integrated circuits[M]. 2005.
- [139] 鲁棒性测评基准数据集及源码网站[Z]. <https://github.com/Jinzhizhisir/PSAFusion>. 2025.
- [140] FAYYAD J, JARADAT M A, GRUYER D, et al. Deep learning sensor fusion for autonomous vehicle perception and localization: A review[J]. Sensors, 2020, 20(15): 4220.
- [141] FENG D, HAASE-SCHÜTZ C, ROSENBAUM L, et al. Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges[J]. IEEE Transactions on Intelligent Transportation Systems, 2020, 22(3): 1341-1360.
- [142] CUI Y, CHEN R, CHU W, et al. Deep learning for image and point cloud fusion in autonomous driving: A review[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(2): 722-739.
- [143] MAO J, SHI S, WANG X, et al. 3D object detection for autonomous driving: a review and new outlooks[J]. arXiv preprint arXiv:2206.09474, 2022.
- [144] Cruise[Z]. <https://getcruise.com/>. 2023.
- [145] Mobileye[Z]. <https://www.mobileye.com/solutions/>. 2023.
- [146] Aptiv[Z]. <https://www.aptiv.com/>. 2023.
- [147] YAN C, XU W, LIU J. Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle[J]. Def Con, 2016, 24(8): 109.
- [148] MAN Y, LI M, GERDES R. Ghostimage: Remote perception attacks against camera-based image

- classification systems[C]//23rd International Symposium on Research in Attacks, Intrusions and Defenses. 2020.
- [149] NASSI B, NASSI D, BEN-NETANEL R, et al. Phantom of the ADAS: Phantom Attacks on Driver-Assistance Systems.[J]. IACR Cryptol. ePrint Arch., 2020, 2020: 85.
- [150] LOVISOTTO G, TURNER H, SLUGANOVIC I, et al. Slap: Improving physical adversarial examples with short-lived adversarial perturbations[C]//. 2021.
- [151] DUAN R, MAO X, QIN A K, et al. Adversarial laser beam: Effective physical-world attack to dnns in a blink[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 16062-16071.
- [152] YAN C, XU Z, YIN Z, et al. Rolling colors: Adversarial laser exploits against traffic light recognition [C]//31st USENIX Security Symposium (USENIX Security 22). 2022: 1957-1974.
- [153] JIANG Q, JI X, YAN C, et al. GlitchHiker: Uncovering Vulnerabilities of Image Signal Transmission with IEMI[J].
- [154] JI X, CHENG Y, ZHANG Y, et al. Poltergeist: Acoustic Adversarial Machine Learning against Cameras and ComputerVision[J]. algorithms, 47(26): 11.
- [155] GAO X, WANG Z, FENG Y, et al. Benchmarking Robustness of AI-enabled Multi-sensor Fusion Systems: Challenges and Opportunities[J]. arXiv preprint arXiv:2306.03454, 2023.
- [156] YAN X, ZHENG C, LI Z, et al. Benchmarking the Robustness of LiDAR Semantic Segmentation Models[J]. arXiv preprint arXiv:2301.00970, 2023.
- [157] FU K, XU W. Risks of trusting the physics of sensors[J]. Communications of the ACM, 2018, 61(2): 20-23.
- [158] EYKHOLT K, EVTIMOV I, FERNANDES E, et al. Robust physical-world attacks on deep learning visual classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 1625-1634.
- [159] ZHAO Y, ZHU H, LIANG R, et al. Seeing isn't believing: Towards more robust adversarial attack against real world object detectors[C]//Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security. 2019: 1989-2004.
- [160] KONG Z, GUO J, LI A, et al. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14254-14263.
- [161] HUANG L, GAO C, ZHOU Y, et al. Universal physical camouflage attacks on object detectors[C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 720-729.
- [162] WU Z, LIM S N, DAVIS L S, et al. Making an invisibility cloak: Real world adversarial attacks on object detectors[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. 2020: 1-17.
- [163] WANG J, LIU A, YIN Z, et al. Dual attention suppression attack: Generate adversarial camouflage in physical world[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8565-8574.
- [164] ZOLFI A, KRAVCHIK M, ELOVICI Y, et al. The translucent patch: A physical and universal attack on object detectors[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 15232-15241.
- [165] KOSCHER K, CZESKIS A, ROESNER F, et al. Experimental security analysis of a modern automobile[C]//2010 IEEE symposium on security and privacy. 2010: 447-462.
- [166] CHECKOWAY S, MCCOY D, KANTOR B, et al. Comprehensive experimental analyses of automotive attack surfaces[C]//20th USENIX security symposium (USENIX Security 11). 2011.
- [167] XU W, MA K, TRAPPE W, et al. Jamming sensor networks: attack and defense strategies[J]. IEEE network, 2006, 20(3): 41-47.
- [168] LAW Y W, PALANISWAMI M, HOESEL L V, et al. Energy-efficient link-layer jamming attacks against wireless sensor network MAC protocols[J]. ACM Transactions on Sensor Networks (TOSN), 2009, 5(1): 1-38.
- [169] ROUF I, MILLER R, MUSTAFA H, et al. Security and privacy vulnerabilities of {In-Car} wireless networks: A tire pressure monitoring system case study[C]//19th USENIX Security Symposium

- (USENIX Security 10). 2010.
- [170] LADISA P, PLATE H, MARTINEZ M, et al. Sok: Taxonomy of attacks on open-source software supply chains[C]//2023 IEEE Symposium on Security and Privacy (SP). 2023: 1509-1526.
- [171] KITCHENHAM B, BRERETON O P, BUDGEN D, et al. Systematic literature reviews in software engineering—a systematic literature review[J]. Information and software technology, 2009, 51(1): 7-15.
- [172] HARZING A. Publish or Perish[Z]. <https://harzing.com/resources/publish-or-perish>. 2023.
- [173] LI-USB30-AR023ZWDR[Z]. <https://www.leopardimaging.com/product-category/usb30-cameras>. 2023.
- [174] WEN H, CHANG S, ZHOU L. Light Projection-Based Physical-World Vanishing Attack Against Car Detection[C]//ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023: 1-5.
- [175] SAYLES A, HOODA A, GUPTA M, et al. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 14666-14675.
- [176] ZHU W, JI X, CHENG Y, et al. TPatch: A Triggered Physical Adversarial Patch[J]. 2023.
- [177] CYR B. Characterizing Laser Signal Injection and its Impact on the Security of Cyber-Physical Systems[D]. 2023.
- [178] JIN Z, JIANG Q, LU X, et al. PhantomLiDAR: Cross-modality Signal Injection Attacks against LiDAR[J]. arXiv preprint arXiv:2409.17907, 2024.
- [179] TOMASI J, WAGSTAFF B, WASLANDER S L, et al. Learned camera gain and exposure control for improved visual feature detection and matching[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 2028-2035.
- [180] When to Use Your Car's High-Beam Headlights: A Complete Guide[Z]. <https://zutobi.com/us/drive-r-guides/when-use-high-beam-headlights>. 2023.
- [181] JIN Z, JI X, CHENG Y, et al. Laser-Based LiDAR Spoofing: Effects Validation, Capability Quantification, and Countermeasures[J]. IEEE Internet of Things Journal, 2024.
- [182] WANG Z, WU Y, NIU Q. Multi-sensor fusion in automated driving: A survey[J]. Ieee Access, 2019, 8: 2847-2868.
- [183] SIMONELLI A, BULO S R, PORZI L, et al. Disentangling monocular 3d object detection[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 1991-1999.
- [184] SALTON G. Modern information retrieval[J]. (No Title), 1983.
- [185] RUKHOVICH D, VORONTSOVA A, KONUSHIN A. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection[C]//Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2022: 2397-2406.
- [186] GOODFELLOW I, BENGIO Y, COURVILLE A. Deep learning[M]. MIT press, 2016.
- [187] CHENG Y, JI X, ZHU W, et al. Adversarial computer vision via acoustic manipulation of camera sensors[J]. IEEE Transactions on Dependable and Secure Computing, 2023, 21(4): 3734-3750.
- [188] XIAO Q, PAN X, LU Y, et al. Exorcising” Wraith”: Protecting {LiDAR-based} Object Detector in Automated Driving System from Appearing Attacks[C]//32nd USENIX Security Symposium (USENIX Security 23). 2023: 2939-2956.
- [189] 毕艳忠, 孙利民. 传感器网络中的数据融合[J]. 计算机科学, 2004, 31(7): 101-103.
- [190] CHANG J R, CHEN Y S. Pyramid stereo matching network[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2018: 5410-5418.
- [191] LIU J, PARK J M. “seeing is not always believing”: Detecting perception error attacks against autonomous vehicles[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(5): 2209-2223.
- [192] LOTTER W, KREIMAN G, COX D. Deep predictive coding networks for video prediction and unsupervised learning[J]. arXiv preprint arXiv:1605.08104, 2016.
- [193] CHEN S, GUO H, ZHU S, et al. Video Depth Anything: Consistent Depth Estimation for Super-Long Videos[J]. arXiv:2501.12375, 2025.
- [194] YANG Z, JIA X, LI H, et al. Llm4drive: A survey of large language models for autonomous driving [J]. arXiv preprint arXiv:2311.01043, 2023.

- [195] DING X, HAN J, XU H, et al. Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving[J]. arXiv preprint arXiv:2309.05186, 2023.
- [196] TIAN X, GU J, LI B, et al. Drivevlm: The convergence of autonomous driving and large vision-language models[J]. arXiv preprint arXiv:2402.12289, 2024.
- [197] SIMA C, RENZ K, CHITTA K, et al. Drivelm: Driving with graph visual question answering[C]// European Conference on Computer Vision. 2024: 256-274.
- [198] WEN L, YANG X, FU D, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving[J]. arXiv preprint arXiv:2311.05332, 2023.
- [199] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[J]. Advances in neural information processing systems, 2020, 33: 1877-1901.
- [200] JIN Z, LU X, YANG B, et al. Unity is Strength? Benchmarking the Robustness of Fusion-based 3D Object Detection against Physical Sensor Attack[C]//Proceedings of the ACM on Web Conference 2024. 2024: 3031-3042.
- [201] HU E J, SHEN Y, WALLIS P, et al. Lora: Low-rank adaptation of large language models.[J]. ICLR, 2022, 1(2): 3.



## 攻读博士学位期间的主要成果

### 一、已发表或录用的论文

1. 第一作者. “Pla-lidar: Physical laser attacks against lidar based 3d object detection in autonomous vehicle”, *IEEE Symposium on Security and Privacy (S&P)*, 2023. 【CCF A类, 四大安全顶会之一】(对应本文第二章)
2. 第一作者. “PhantomLiDAR: Compromising LiDAR Systems with IEMI”, *Network and Distributed System Security Symposium (NDSS)*, 2025. 【CCF A类, 四大安全顶会之一】(对应本文第三章)
3. 第一作者. “Unity is Strength? Benchmarking the Robustness of Fusion-based 3D Object Detection against Physical Sensor Attack”, *In Proceedings of the ACM Web Conference (WWW)*, 2024. 【CCF A类, Oral (9.2%)】(对应本文第四章)
4. 第一作者. “Laser-based LiDAR Spoofing: Effects Validation, Capability Quantification, and Countermeasures”, *IEEE Internet of Things Journal(IoT-J)*, 2024. 【SCI 收录, IF=8.2】(对应本文第五章)
5. 第一作者. “Physical Sensor Attack Robustness of Fusion-based Perception in Autonomous Driving: Benchmark and Defense”, *IEEE Transactions on Mobile Computing(TMC)* 投稿中. (对应本文第五章)
6. 第二作者. “Adversarial robustness analysis of LiDAR-included models in autonomous driving”, *High-Confidence Computing*, 2024.
7. 第三作者.“Generating 3D adversarial point clouds under the principle of LiDARs”, *NDSS Autonomous Vehicle Security Workshop*, 2022.
8. 第三作者. “Anti-Replay: A Fast and Lightweight Voice Replay Attack Detection System”, *IEEE International Conference on Parallel and Distributed Systems (ICPADS)*, 2021. 【CCF C类】

9. 第五作者.“A Survey on Voice Assistant Security: Attacks and Countermeasures”, ACM Computing Surveys, 2022. 【SCI 收录, IF=16.6】

## 二、发明专利

1. 第一学生作者, 语音信号频谱特征和深度学习的语音欺骗攻击检测方法, 中国发明专利: CN112201255A (已授权)
2. 第一学生作者, 基于激光发射器的激光雷达点云注入系统, 中国发明专利: CN114966625A (已公开)
3. 第一学生作者, 针对多传感器融合感知的跨域安全测试数据集生成方法, 中国发明专利: CN118520297A (已公开)
4. 第一学生作者, 一种基于激光雷达目标点云的脉冲控制信号设计方法, 中国发明专利: CN114814789A (已公开)
5. 第二学生作者, 一种针对激光雷达的电磁 XXXXXX, 国防专利

## 三、参与的科研项目

1. “131”项目, 基于感算联动的 XX 自主无人系统 XXXX
2. “慧眼行动”创新成果转化应用项目, XX 信息处理系统 XXXXX 及智能 XXXX 技术
3. 基础加强计划重点基础研究项目, 物联网安全 XXX
4. 基础加强计划重点基础研究项目, XXXX 智能识别对抗 XXXXX
5. 国家自然科学基金杰出青年科学基金, 电子信息系统多要素安全分析与防护
6. 国家自然科学基金专项项目, 海量终端设备威胁评估及安全防护关键技术研究
7. 阿里巴巴网络技术有限公司, 语音欺骗攻击检测