

# 自动驾驶中激光雷达感知的 脆弱性分析与安全防护

Vulnerability Analysis and Defense for  
LiDAR-based Perception in Autonomous Driving

浙江大学博士论文答辩

答辩人: **金子植**

指导老师: **冀晓宇 教授 徐文渊 教授**

答辩时间: 2025年5月27日

# 个人简介



**金子植** 直博2020级  
电气工程学院 电气工程



2016-2020 浙江大学 本科 自动化  
2020-2025 浙江大学 直博 电气工程



**自动驾驶安全、信息物理系统安全**



- **论文专利**: 参与发表论文**9**篇; 一作CCF A会议、SCI期刊等**4**篇, 在投**1**篇。
- **项目经历**: \*31项目(负责、**优秀结题**), 慧眼行动(负责), \*73, \*66等项目
- **荣誉称号**: 优秀研究生、浙江大学优秀毕业生等荣誉5次

# 1 背景与意义

Background and significance

# 2 方法与思路

Challenges and contributions

# 3 成果与结论

Results and conclusions

# 4 总结与展望

Summary and prospect

▼ 目录

CONTENT

# 自动驾驶安全研究至关重要

自动驾驶具有**应用场景广泛**和**事故后果严重**的特点，因此其安全至关重要。



私家车



出租车



货运



环卫车

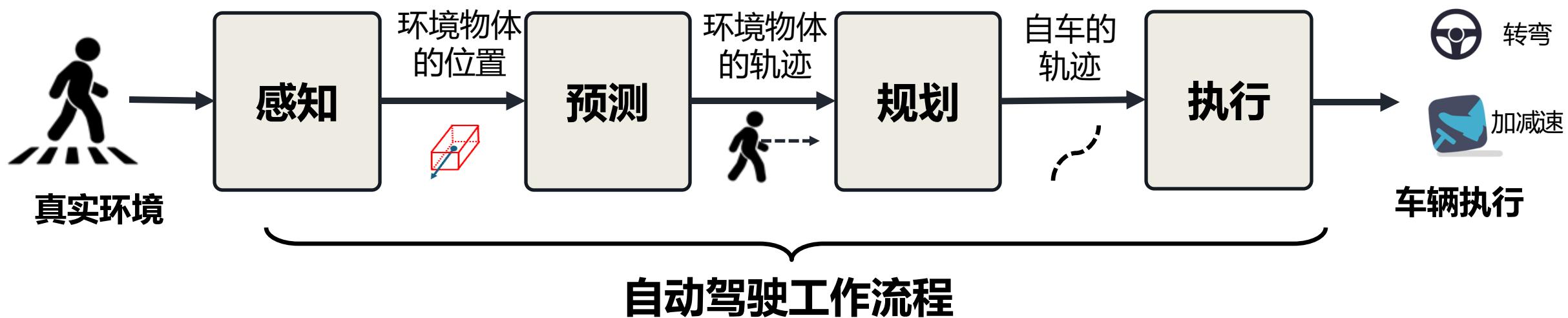
**自动驾驶应用场景广泛**



**自动驾驶安全事故后果严重**

# 自动驾驶工作流程

自动驾驶工作流程：感知→预测→规划→执行



# 自动驾驶中感知安全的重要性

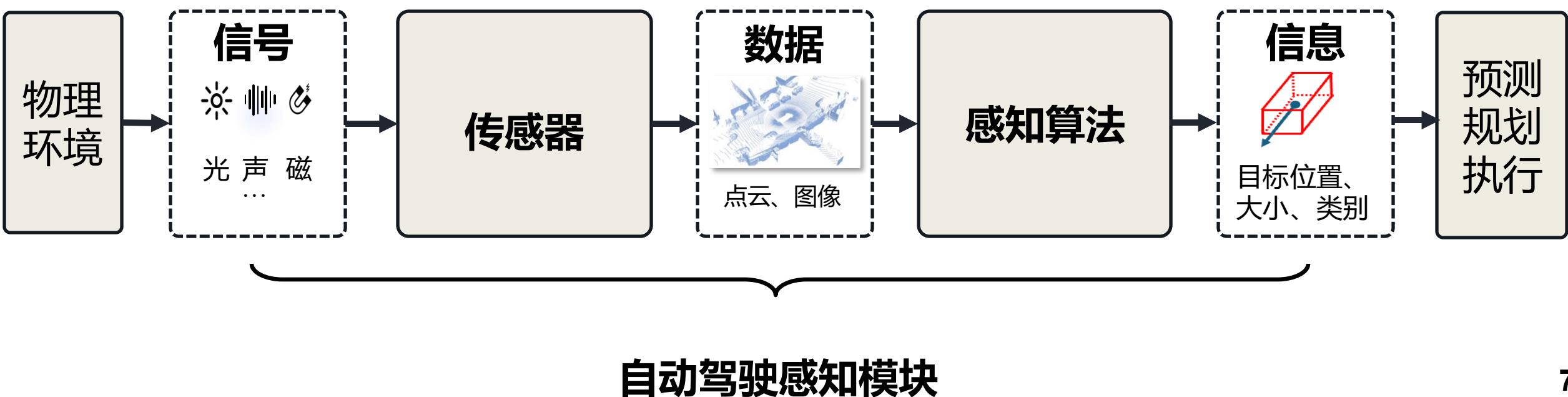
自动驾驶工作流程：**感知**→ 预测→ 规划→ 执行



**正确感知是自动驾驶车辆安全行驶的重要前提**

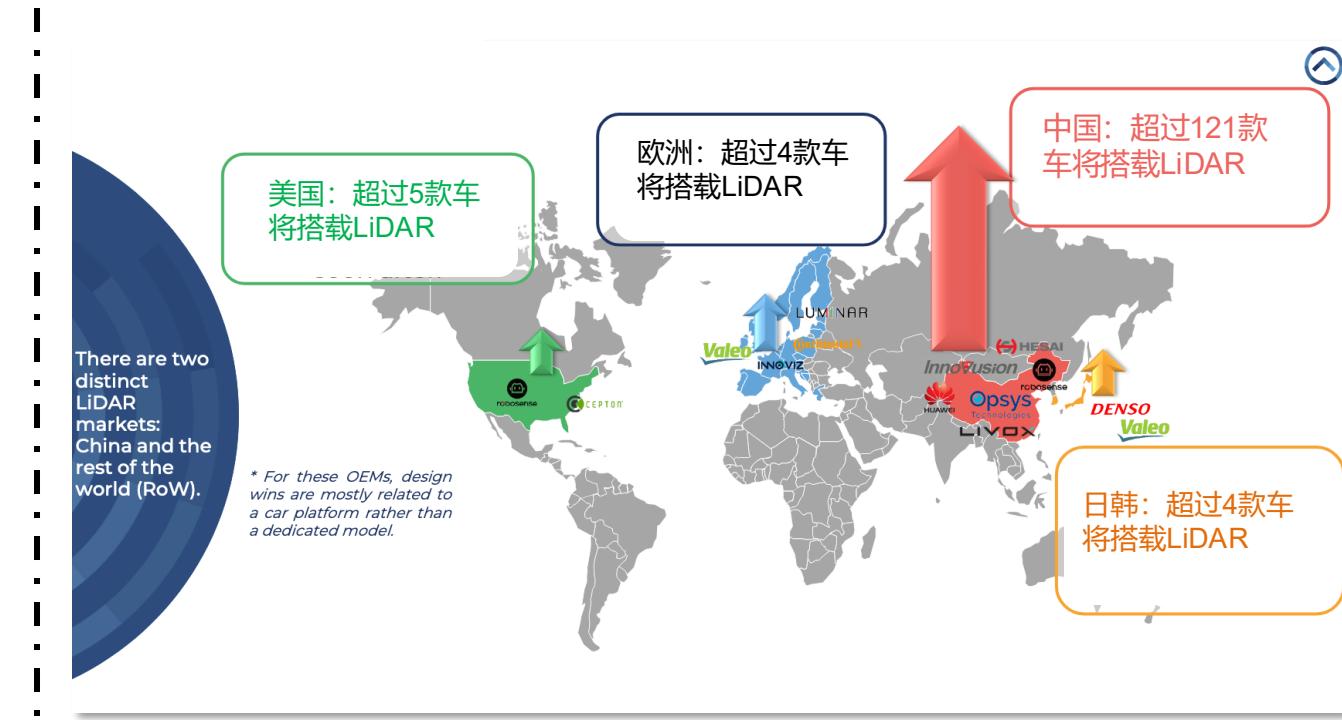
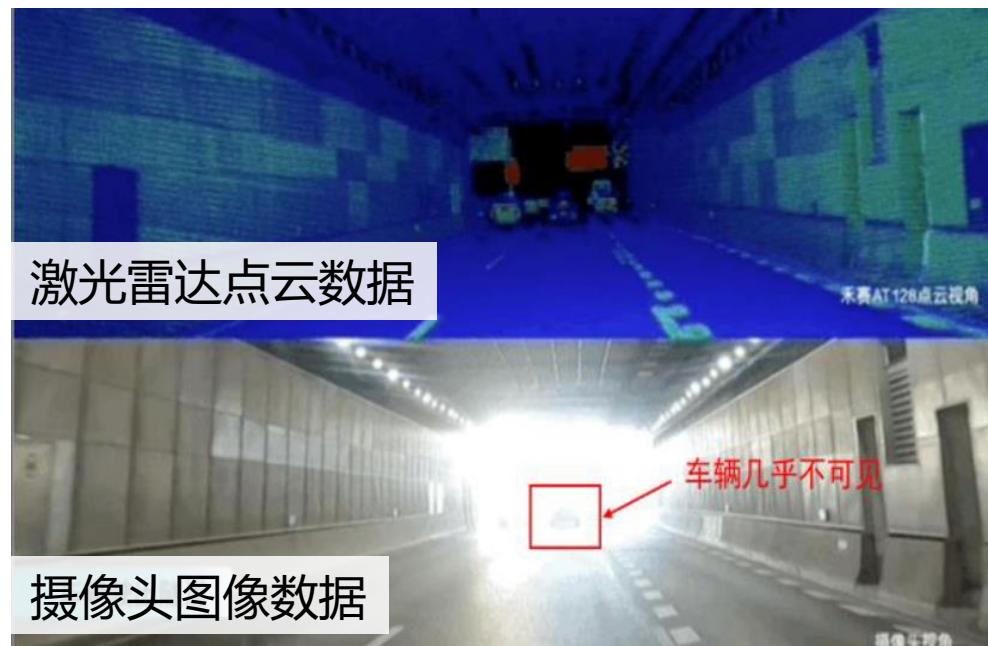
# 自动驾驶中的感知模块

- 感知模块构成： **传感器+ 感知算法**
- 感知链路：“**信号→数据→信息**”



# 激光雷达传感器广泛用于自动驾驶感知

激光雷达（LiDAR）凭借其**全天时、高精度**的优势，被广泛应用于自动驾驶感知。



LiDAR在**强光、黑夜**等环境下能生成**高精度点云**

全球超过140款车型（160万辆）正搭载LiDAR<sup>[1]</sup>

# 激光雷达广泛用于自动驾驶感知

激光雷达（LiDAR）凭借其**全天时、高精度**的优势，被广泛应用于自动驾驶中的环境感知。



## 自动驾驶中的激光雷达感知安全研究具有广泛的现实意义



在**黑夜、强光**等环境下能生成**高精度点云**

全球超过140款车（160万辆）正搭载LiDAR<sup>[1]</sup>

# 针对激光雷达本身的安全性研究不足

以往的自动驾驶感知安全研究主要关注摄像头的安全性。



像素饱和



目标投影



彩条注入



## 图像截断



色带丢失



运动模糊

典型的摄像头攻击研究

# 针对激光雷达本身的安全性研究不足

以往的自动驾驶感知安全研究主要关注摄像头的安全性。

	Liu et al. [34]	'17	V	✓	✓								
Object detection	Eykholz et al. [18]	'18	S	✓	✓				○	✓			
	Chen et al. [37]	'18	M	✓	✓				○	✓			
	Zhao et al. [26]	'19	S	✓	✓				○	✓			
	Xiao et al. [55]	'19	V	✓	✓	✓	✓		○	✓			
	Zhang et al. [56]	'19	M	✓	✓				●	✓			
	Nassi et al. [57]	'20	S	✓	✓				○	✓			
	Man et al. [58]	'20	S	✓					○	✓			
	Hong et al. [59]	'20	S	✓					○	✓			
	Huang et al. [60]	'20	V	✓	✓				○	✓			
	Wu et al. [61]	'20	V	✓	✓				○	✓			
	Xu et al. [62]	'20	V	✓	✓				○	✓			
	Hu et al. [63]	'20	V	✓	✓				●	✓			
	Hamdi et al. [64]	'20	M	✓					○	✓			
	Ji et al. [65]								●	✓			



像素饱和



目标投影



# 亟需开展自动驾驶中的激光雷达感知安全研究



色带丢失



运动模糊

## 典型的摄像头攻击研究

- 1 背景与意义**  
Background and significance
- 2 方法与思路**  
Challenges and contributions
- 3 成果与结论**  
Results and conclusions
- 4 总结与展望**  
Summary and prospect



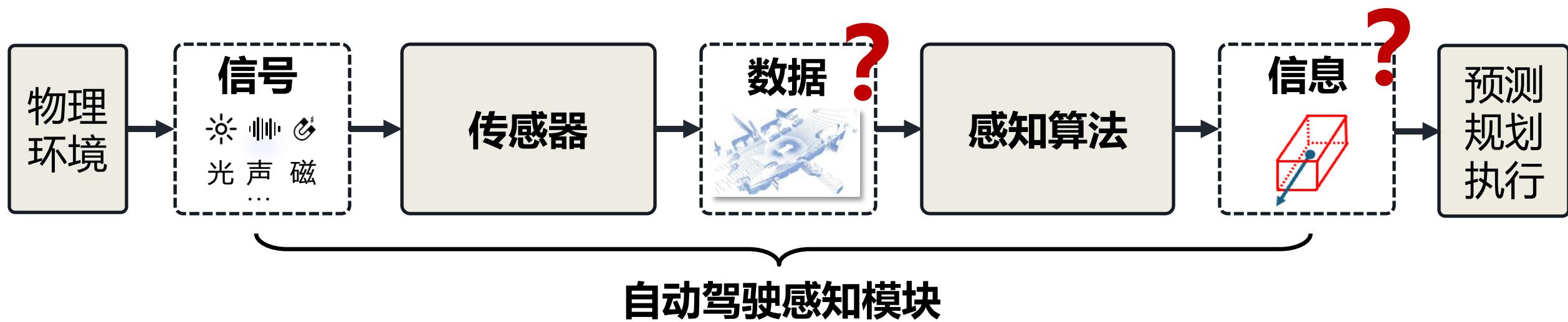
# 研究目的：让激光雷达感知模块更安全

## 1. 传感器安全：传感器**数据正确测量**物理环境

e.g. 若前方有个人，激光雷达要正确测量到前方那个人的点云

## 2. 感知算法安全：感知**信息正确描述**物理环境

e.g. 若前方有个人，感知算法要正确给出那个人的位置、大小、类别

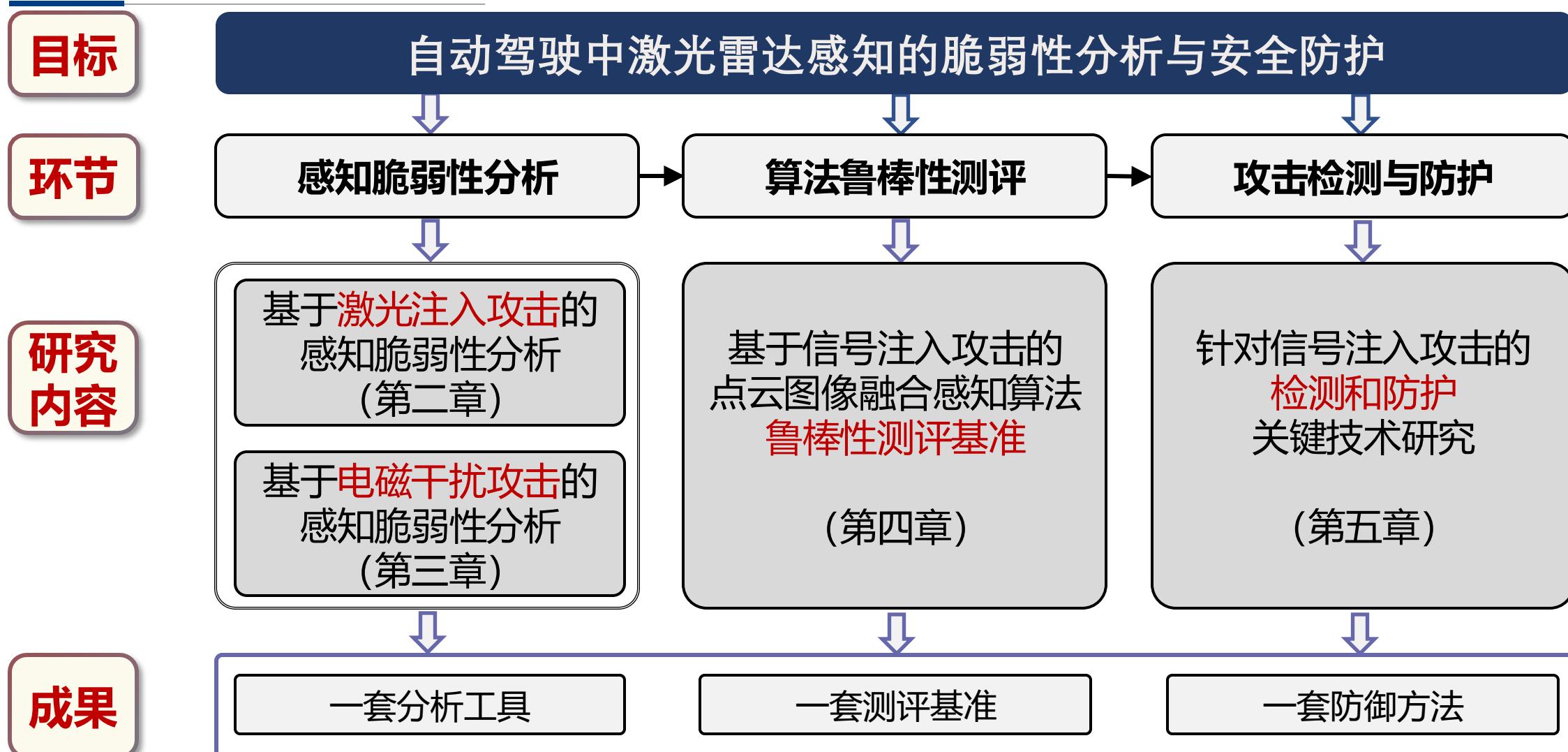


# 研究方法：红蓝对抗

“未知攻，焉知防”：首先通过**攻击**传感器和感知算法来找到脆弱性，然后通过研究对应的**防护**技术来**增强安全性**。



# 研究思路



# 1 背景与意义

Background and significance

# 2 方法与思路

Challenges and contributions

# 3 成果与结论

Results and conclusions

1. 基于激光注入攻击的感知脆弱性分析
2. 基于电磁干扰攻击的感知脆弱性分析
3. 基于信号注入攻击的算法鲁棒性测评基准
4. 攻击检测和防护关键技术研究

# 4 总结与展望

Summary and prospect



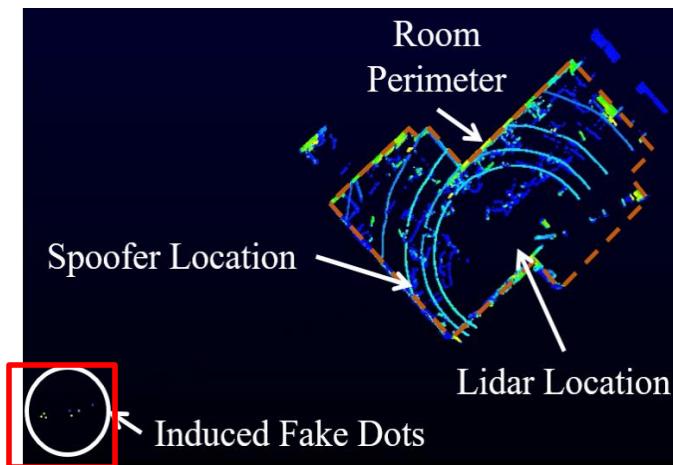
# 成果1：基于激光注入攻击的激光雷达感知脆弱性分析

## 环节1：感知脆弱性分析

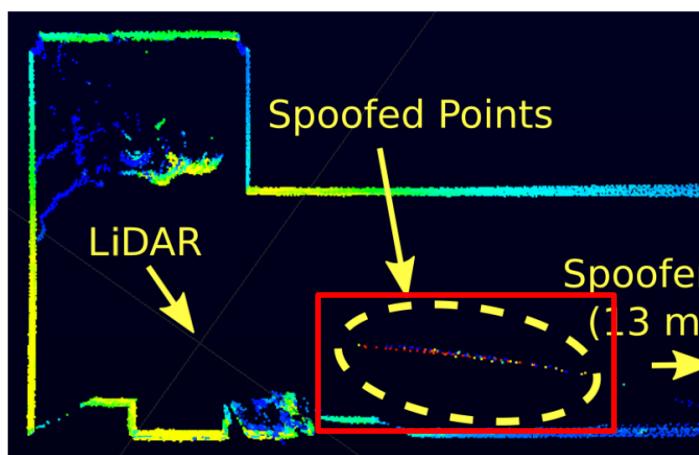
第一作者《PLA-LiDAR: Physical Laser Attacks against LiDAR-based 3D Object Detection in Autonomous Vehicle》[发表于IEEE S&P Oakland 2023 \(CCFA, Big Four\)](#)

# 研究目标

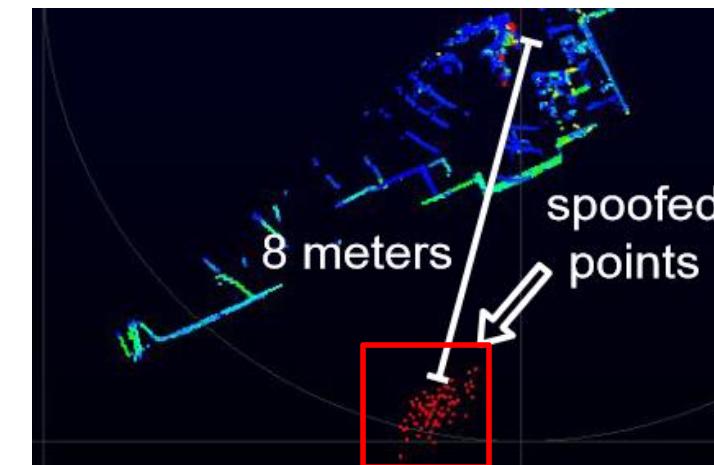
**研究空白：**现有激光攻击的工作点云注入能力较弱，无法直接对感知算法造成影响，低估了激光攻击的危害。



Shin et al. CHES'17  
攻击能力：注入约10个点



Cao et al. CCS'19  
攻击能力：注入约100个点

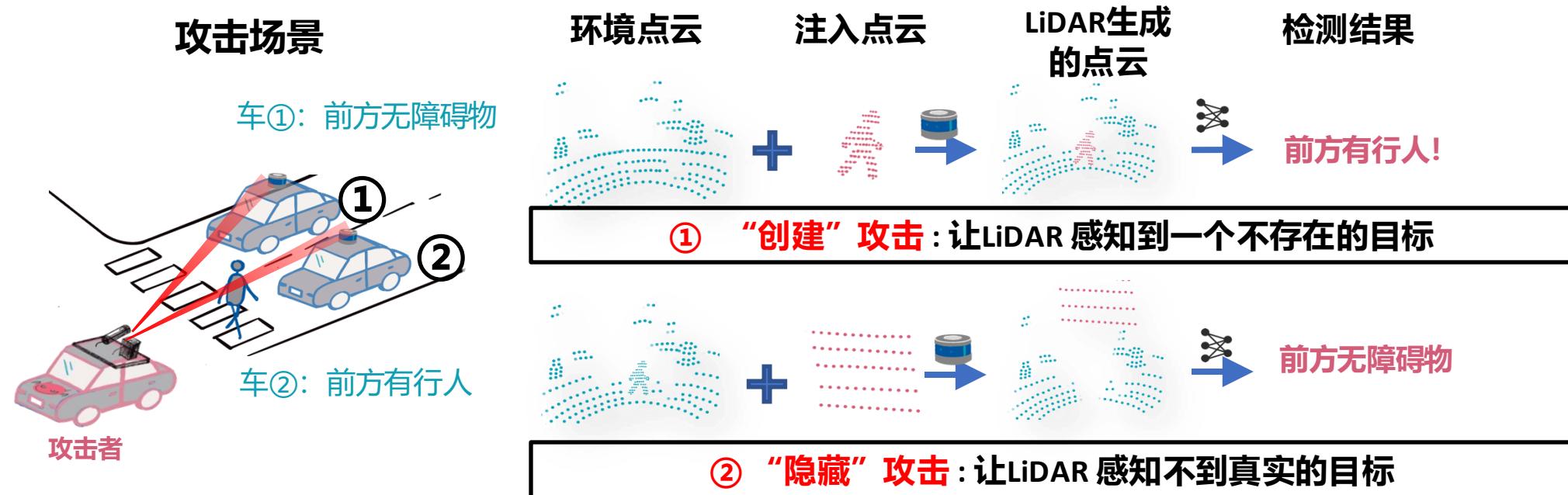


Sun et al. USENIX'20  
攻击能力：注入约150个点

**研究目标：**能否通过在**物理世界**使用**激光信号**注入**欺骗点云**  
来**直接欺骗**感知算法？

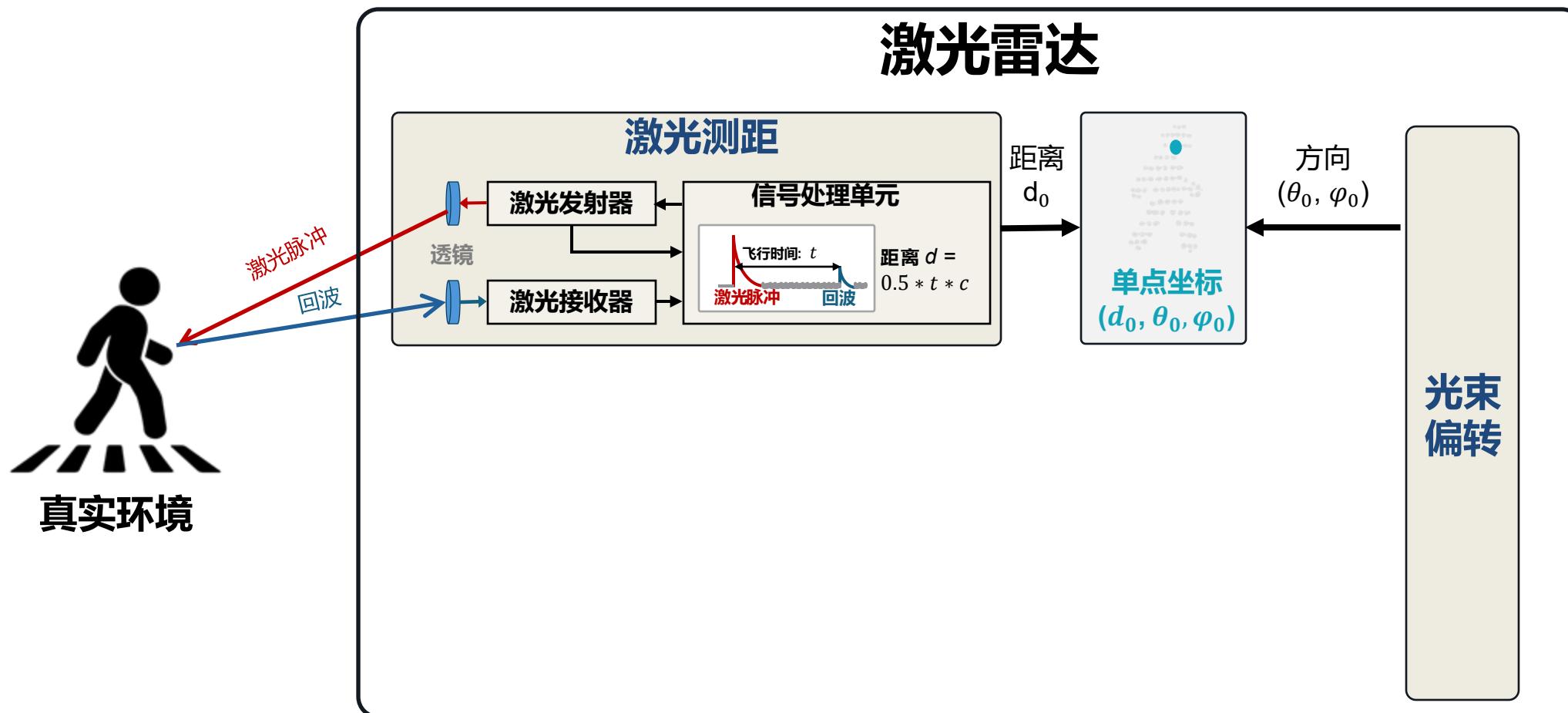
# 工作简介

针对自动驾驶感知中基于LiDAR的3D目标检测任务，用**激光攻击**的形式往LiDAR传感器中**注入形状、位置、点数可控的欺骗点云**，进而影响3D目标检测的识别结果，可以实现**隐藏**和**创建**指定物体的攻击效果。



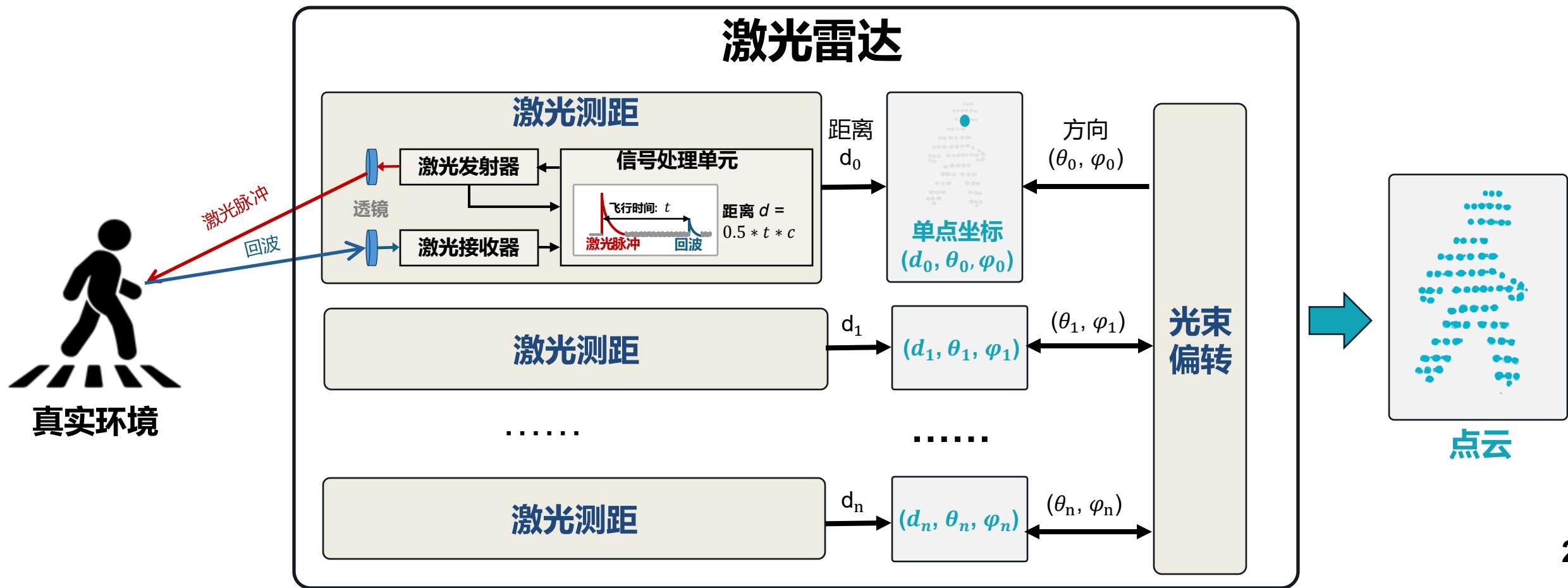
# 背景：点云如何生成？

激光雷达通过**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**。



# 背景：点云如何生成？

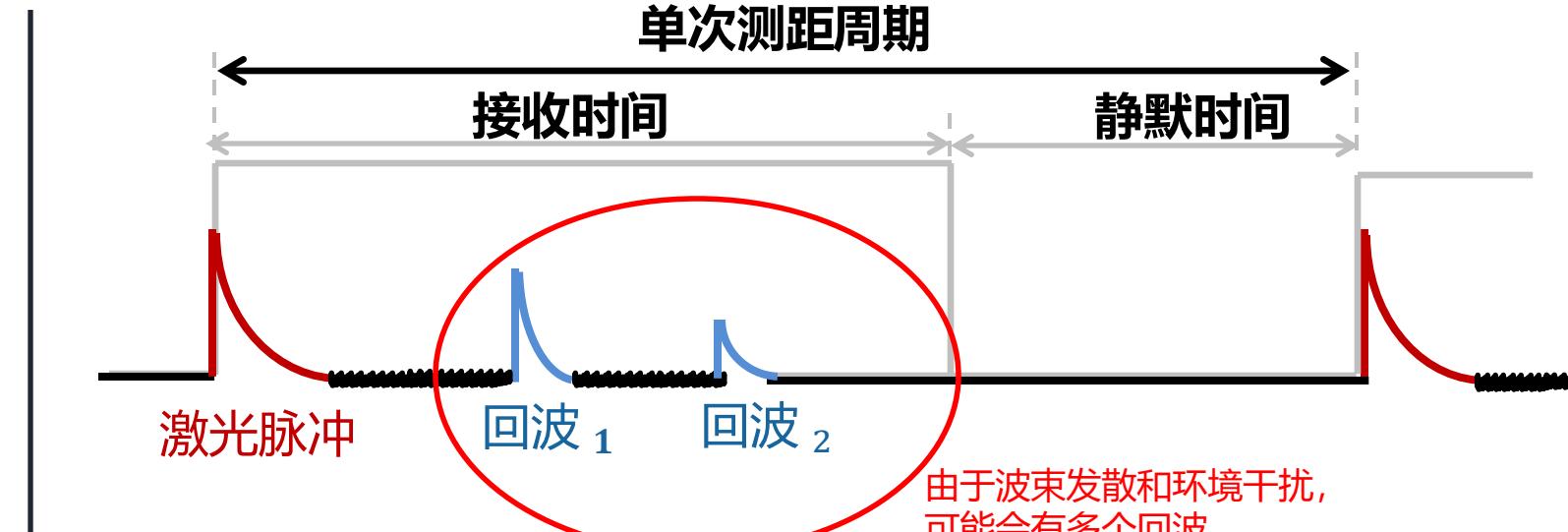
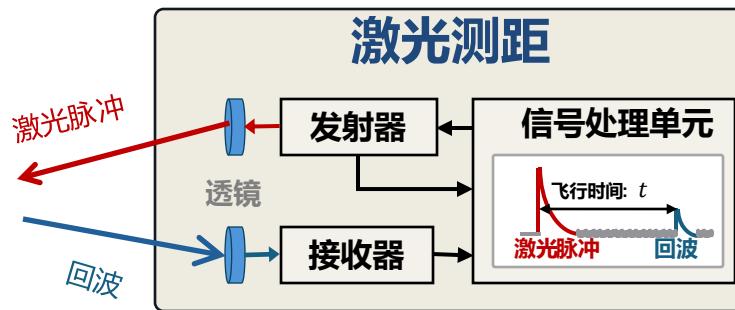
激光雷达通过**激光测距**和**光束偏转**两大核心功能相互配合，生成**高精度点云**。



# 背景：激光测距中的回波筛选机制



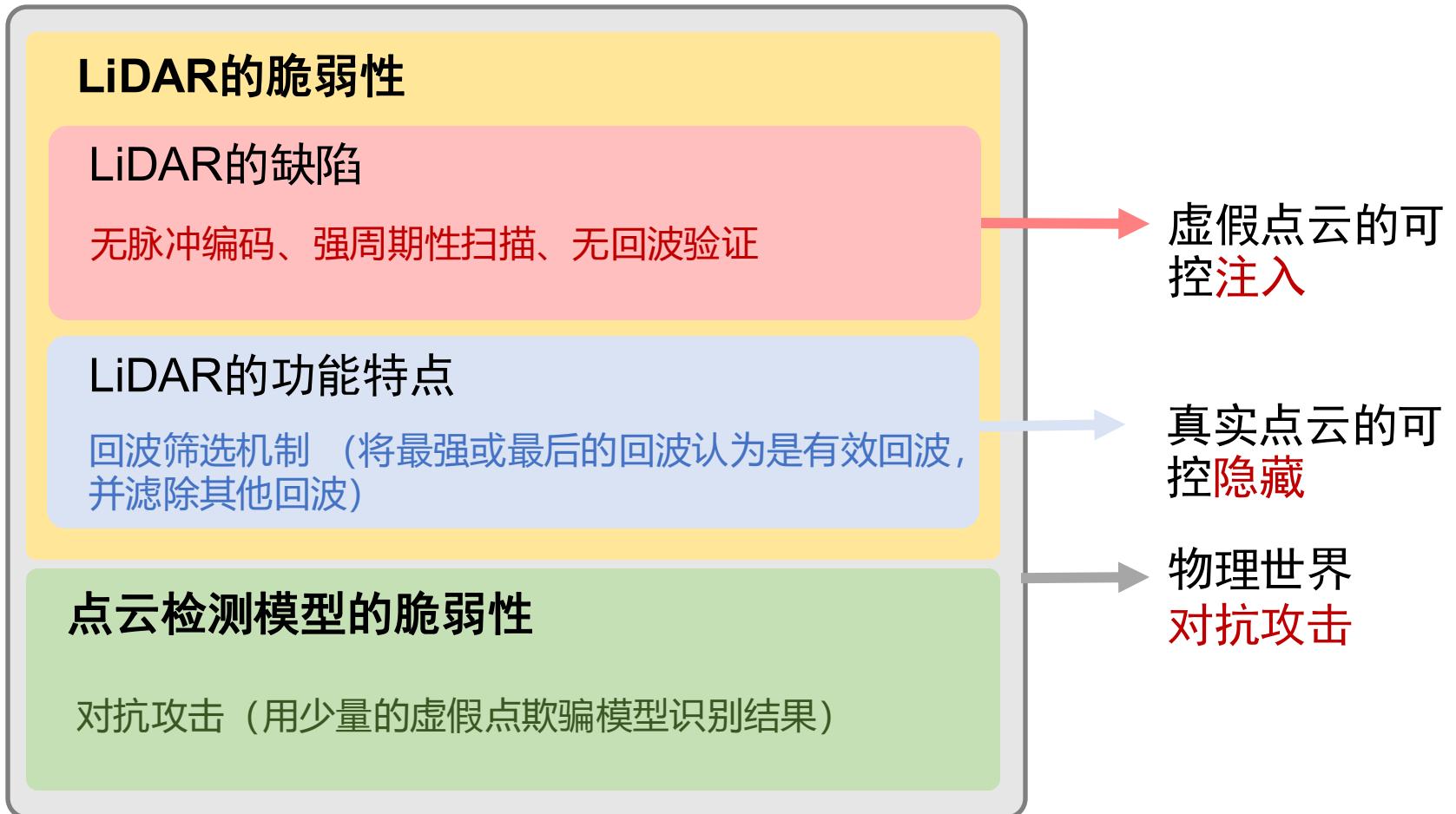
真实环境



## 三种典型的激光雷达回波筛选模式：

- **最强回波** — 选择最强的回波当作有效回波. (最流行✓)
- 双回波—同时选择最强和最后的回波 (一次生成最多两个点)。
- 最后回波 — 接收时间内的最后一个回波。 (通常不采用)

# 攻击原理



# 攻击设计 – 攻击方式及挑战

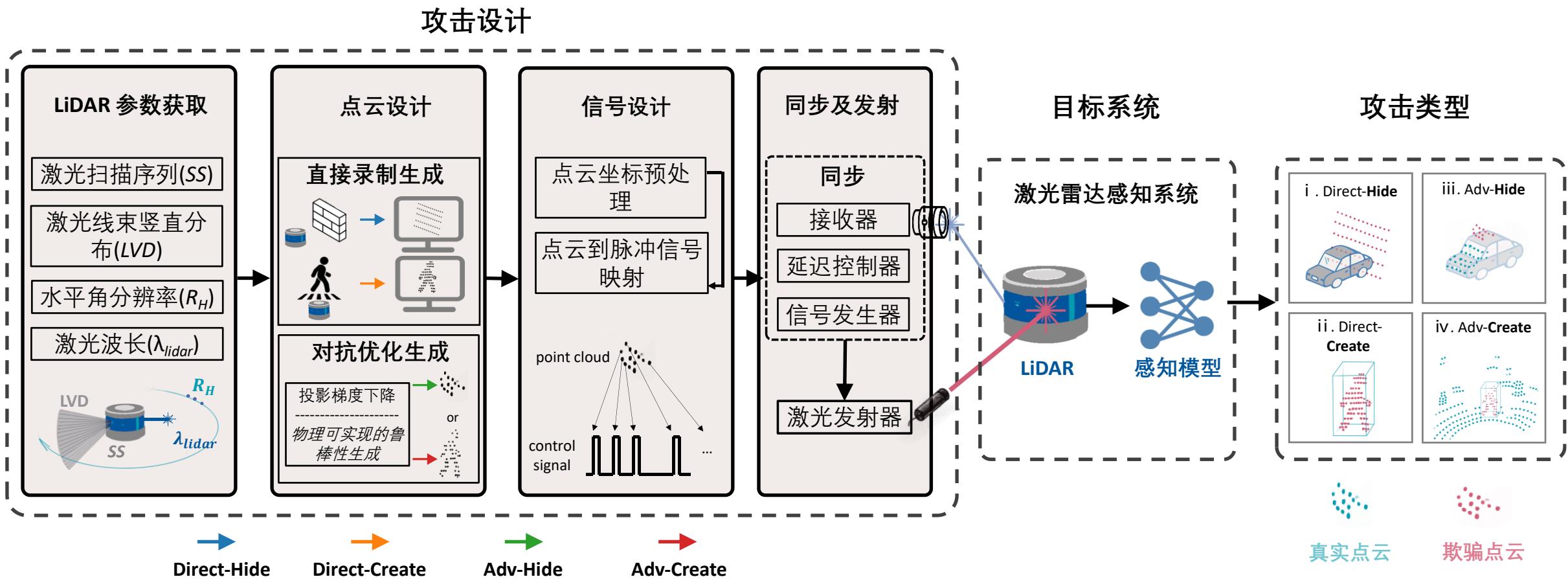
口 **攻击方式**: 利用**高功率**的激光脉冲信号伪造LiDAR的回波，并将**特定波形**的激光攻击信号在**精确的时刻**注入。

攻击设计要求	目的
高功率的激光脉冲	保证信号被认为是 <b>有效回波</b> 而非噪声
特定波形	保证攻击信号能够注入 <b>指定形状</b> 的点云
精确的注入时刻	保证欺骗点注入到 <b>指定位置</b>

口 **挑战**:

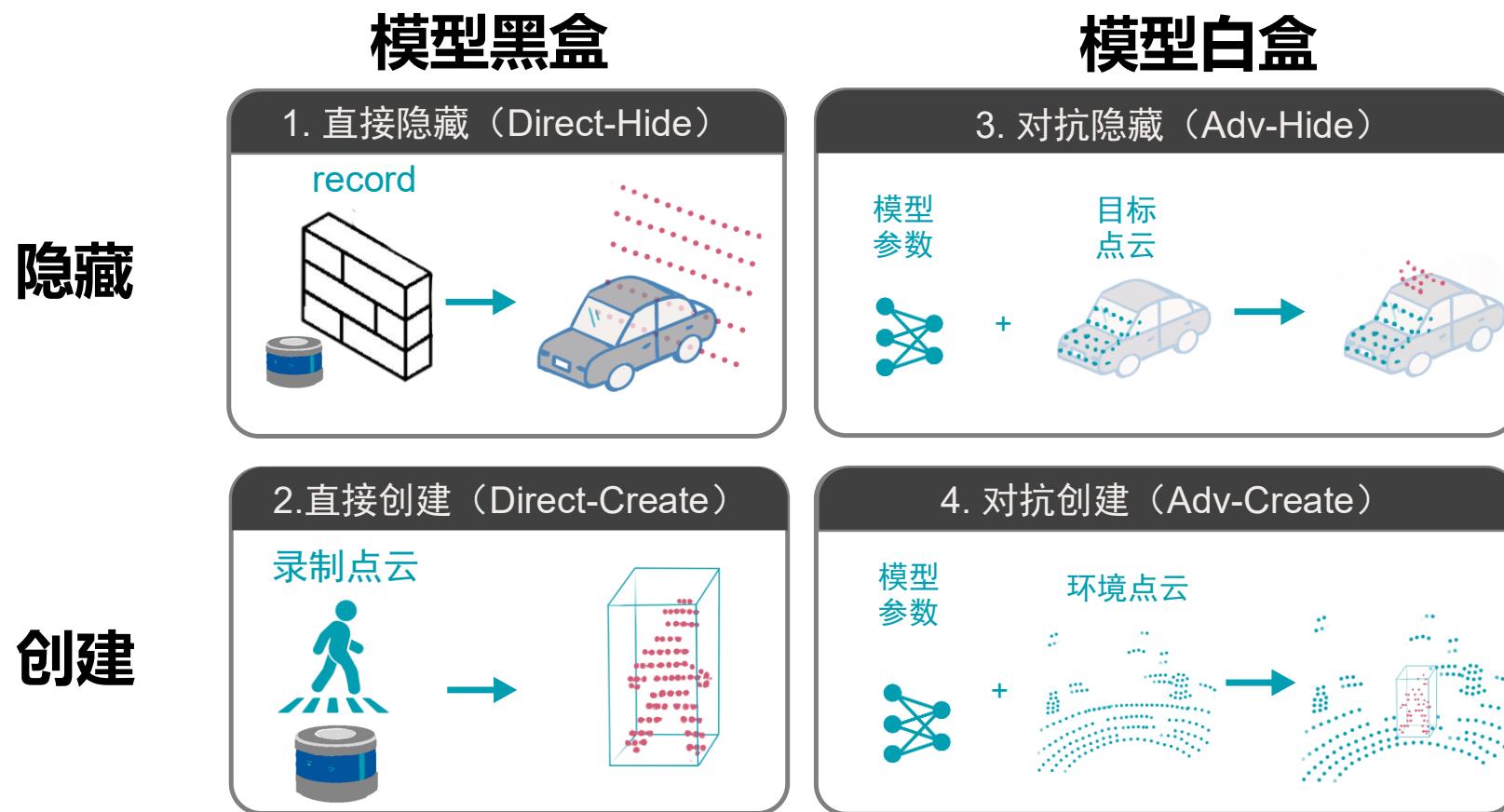
1. 如何**设计攻击信号的波形**, 使其能够注入**指定形状**的点云?
2. 如何**搭建攻击装置**, 使其能够发射**高功率高频率**的信号, 并在**在精确的时刻** (纳秒级精度) 将欺骗点注入?

# 攻击设计



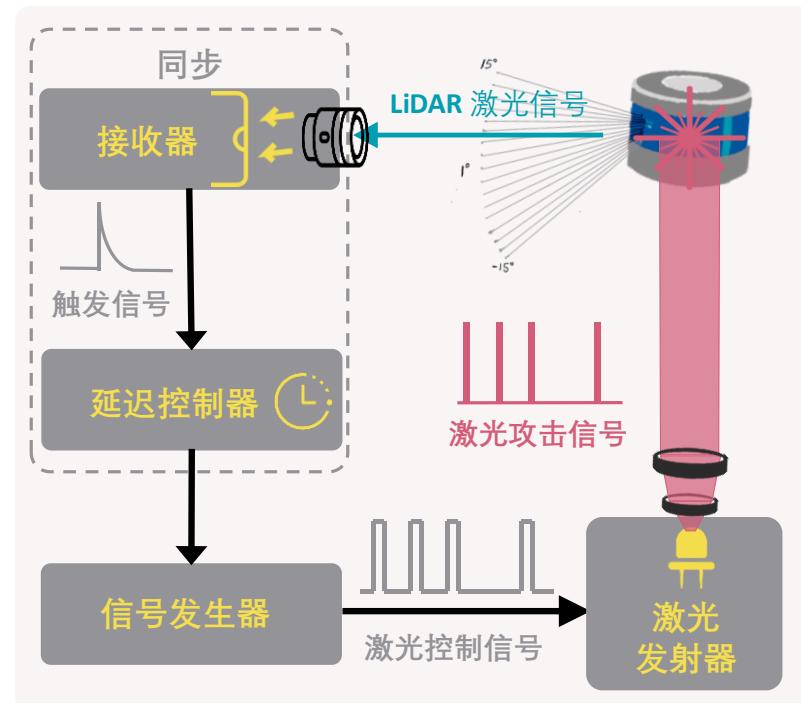
# 攻击设计 - 攻击类型

口 攻击类型：能通过黑盒、白盒方式分别实现隐藏和创建共**4类**攻击



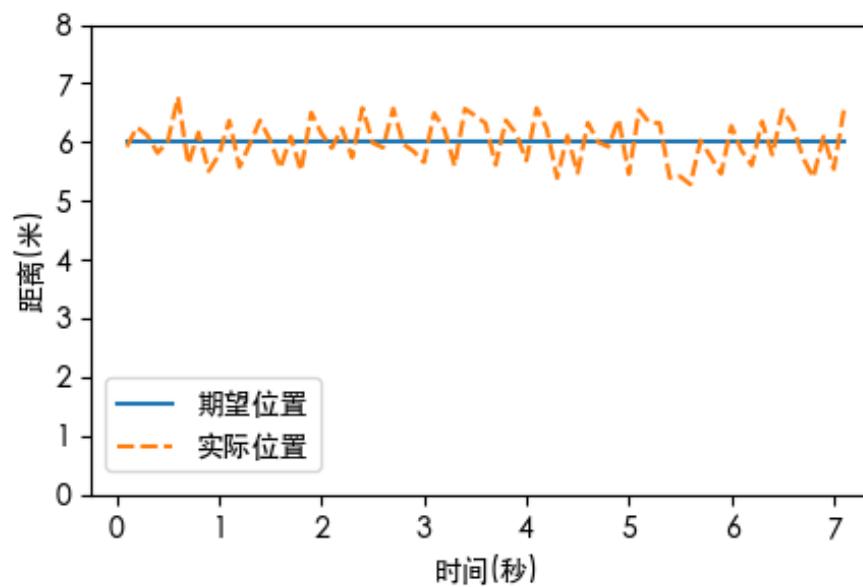
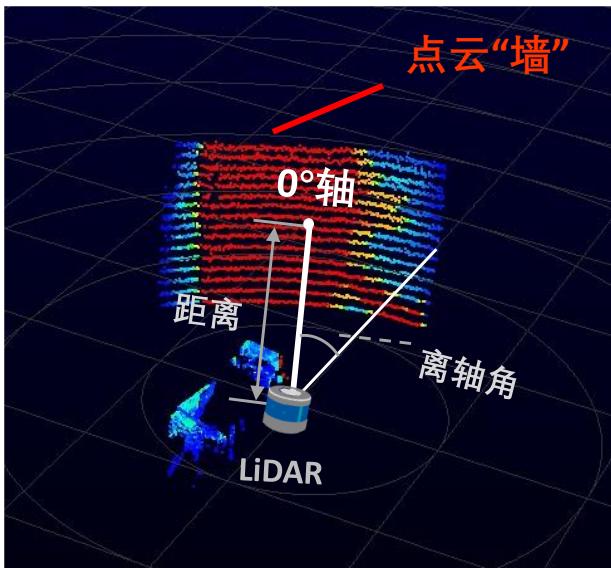
# 攻击设计 - 攻击设备

口 攻击装置：利用集接收器、信号控制器、激光发射器于一体的攻击装置，实现攻击。



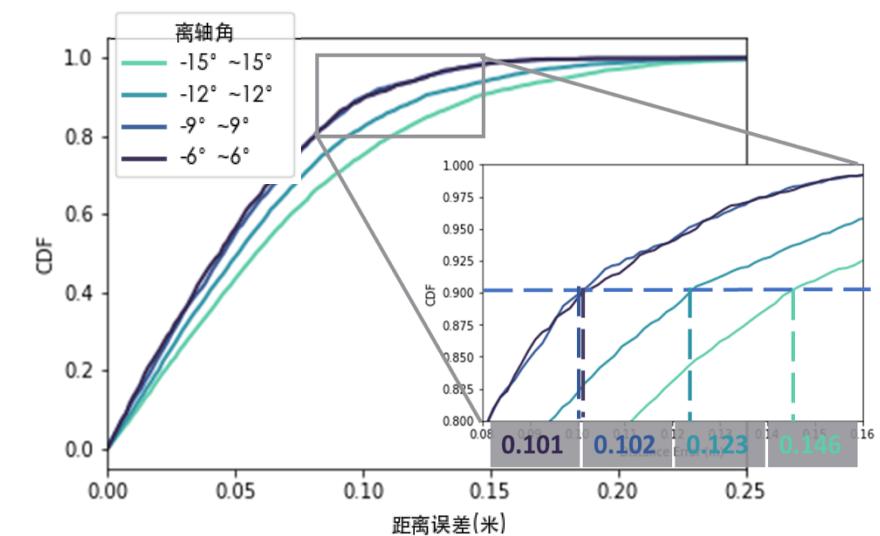
# 攻击能力量化

**口 攻击能力：**能够注入4800个欺骗点，是SOTA工作的20倍。且将时间精度控制在纳秒级，能够对欺骗点的位置和形状做高精度控制。



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

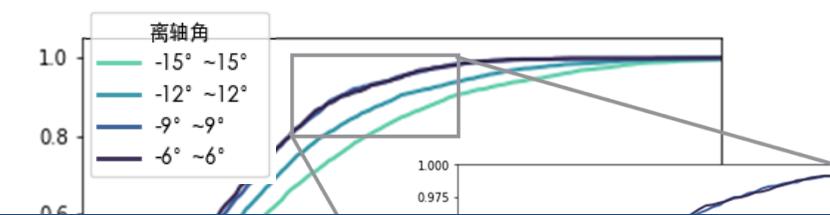
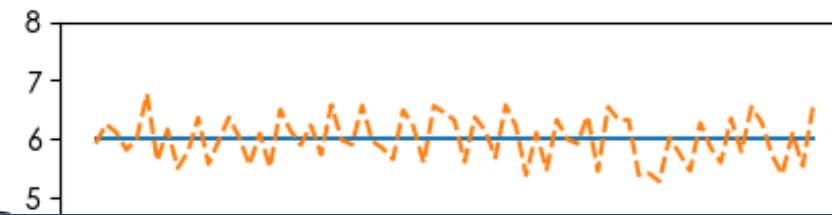
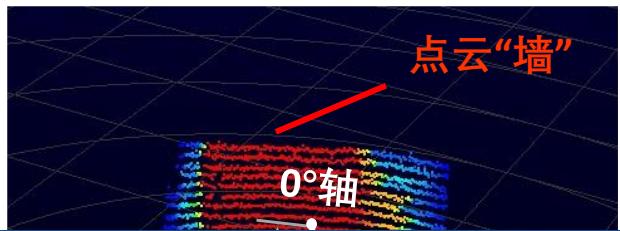
**位置控制能力：**能够将注入点云整体位置的精度控制在标准差为0.38米内。



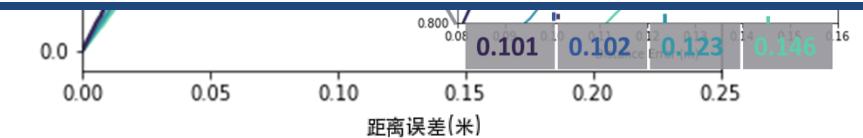
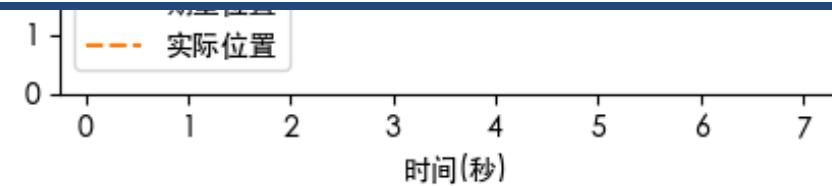
**形状控制能力：**在离轴角 $\pm 9^\circ$ 范围内，90% 的点的距离误差均在0.102米以内

# 攻击能力量化

**口 攻击能力：**能够注入4800个欺骗点，是SOTA工作的20倍。且将时间精度控制在纳秒级，能够对欺骗点的位置和形状做高精度控制。



**好处：**以后的点云对抗攻击的“物理可实现性”有了明确的量化标准。



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

**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。

**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

# 实验评估 – 攻击性能

## 实验设置

- 4种攻击类型
- 2款商用LiDARs:

VLP-16, RS-16



- 3个目标检测模型:  
Second, Pointpillar, Apollo
- 指标: 攻击成功率

表格: 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	VLP-16	100%	98%	47%	75%
	RS-16	100%	86%	46%	66%
PointPillar	VLP-16	100%	64%	78%	24%
	RS-16	100%	51%	74%	19%
Apollo	VLP-16	100%	98%	81%	39%
	RS-16	100%	89%	76%	24%

# 实验评估- 物理世界攻击效果

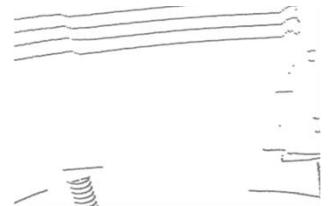
真实场景

点云：攻击前

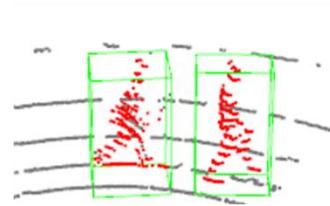
点云：攻击后

识别结果

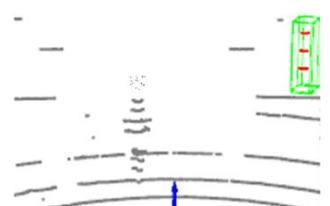
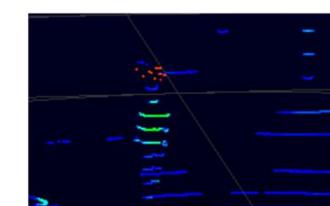
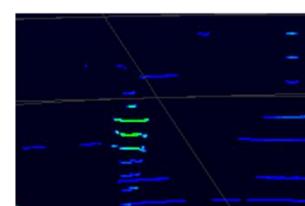
**直接隐藏**  
**Direct-Hide**



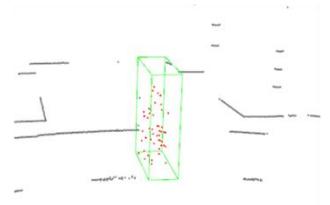
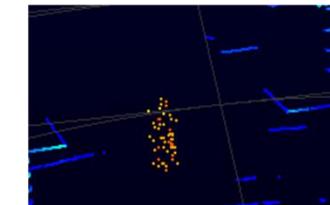
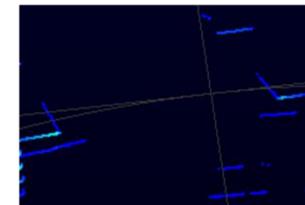
**直接创建**  
**Direct- Create**



**对抗隐藏**  
**Adv-Hide**

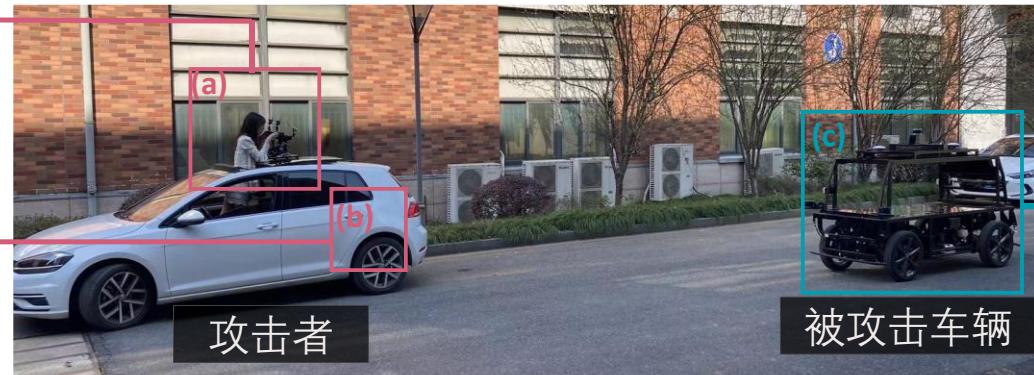
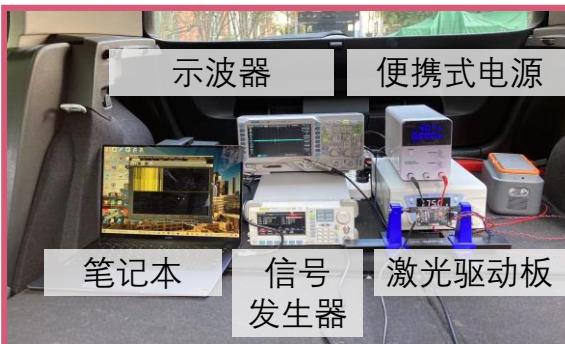


**对抗创建**  
**Adv-Create**



# 实验评估 – 移动攻击可行性试验

## 口 移动攻击实验设置：



(a) 激光接收器和发射器位于车顶

(b) 其余攻击设备置于车厢内

移动攻击实验设置

(c) 搭载有VLP-16激光雷达的被攻击车辆

# 实验评估 – 移动攻击可行性试验

## □ 移动攻击效果：



直接隐藏攻击



直接创建攻击

# 小结

- **攻击能力提升:** 1) 点云注入能力是SOTA的20倍 (4000pts vs 200pts) ; 2) 欺骗点云的形状和位置高可控
- **攻击的物理世界实现:** 1) 首次利用激光在物理世界实现**针对3D目标检测的黑盒攻击**; 2) 首次利用激光在物理世界实现**点云对抗攻击**; 3) 首次利用激光攻击**移动车上的激光雷达**。
- **攻击能力量化:** 量化了攻击的极限和限制, 有助于后续工作的仿真

# 成果2：基于电磁干扰攻击的激光雷达感知脆弱性分析

## 环节1：感知脆弱性分析

第一作者《PhantomLiDAR: Compromising LiDAR Systems with IEMI》 ***NDSS 2025 (CCF A, Big Four)***

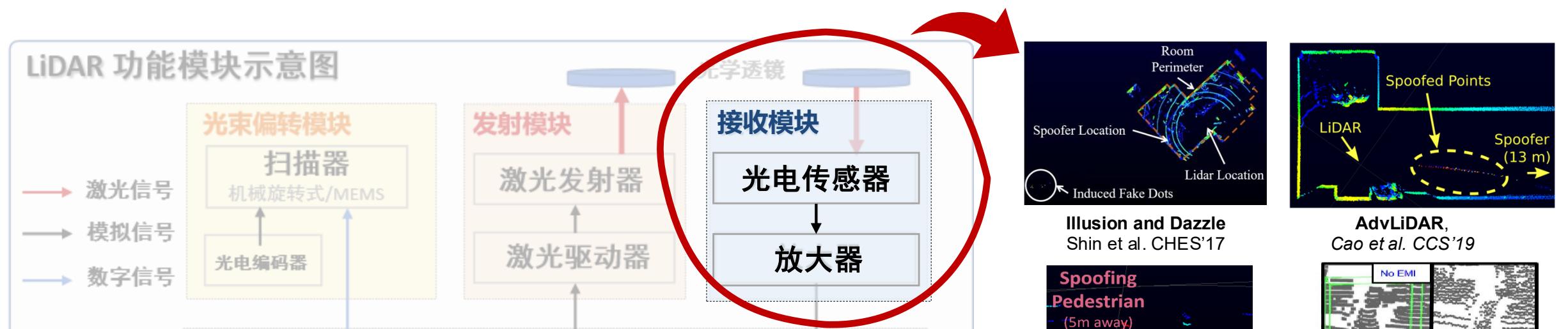
# 研究动机

**现有研究空白：**现有工作 (1) **攻击方式单一**：主要考虑了**激光攻击信号**； (2) **攻击原理单一**：仅将激光接收模块中的**光电传感器**作为攻击入口。



# 研究动机

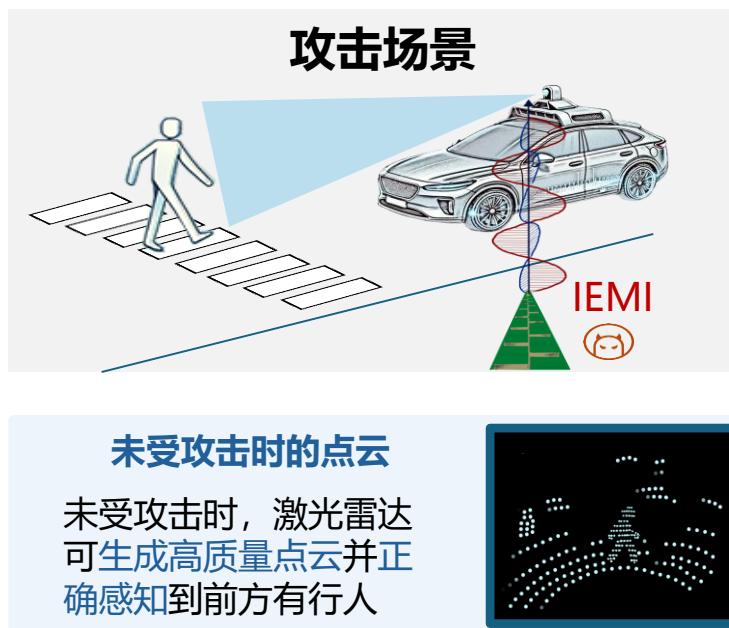
**现有研究空白：**现有工作 (1) **攻击方式单一**: 主要考虑了**激光攻击信号**; (2) **攻击原理单一**: 仅将激光接收模块中的**光电传感器**作为攻击入口。



**研究目标:** 1) 探究是否能利用和LiDAR工作信号**不同模态的信号 (电磁)** 挖掘脆弱性? 2) 探究激光雷达内部的**其他功能模块**是否能够被作为攻击入口。

# 工作简介

利用**电磁干扰 (IEMI)** 的形式挖掘了激光雷达的**多个功能模块**的脆弱性，包括接收模块、监测传感器、光束偏转模块等。实现了SOTA的攻击效果，能实现**点云干扰**、**点云抹除**、**使LiDAR宕机**、**点云可控注入**四类攻击。



→ 四类攻击效果 →

攻击效果	点云干扰	点云抹除	雷达宕机	点云注入
攻击信号形式	正弦波			脉冲调制信号
攻击入口	接收模块	• 主板上的监测传感器 • 接收模块	光束偏转模块	接收模块

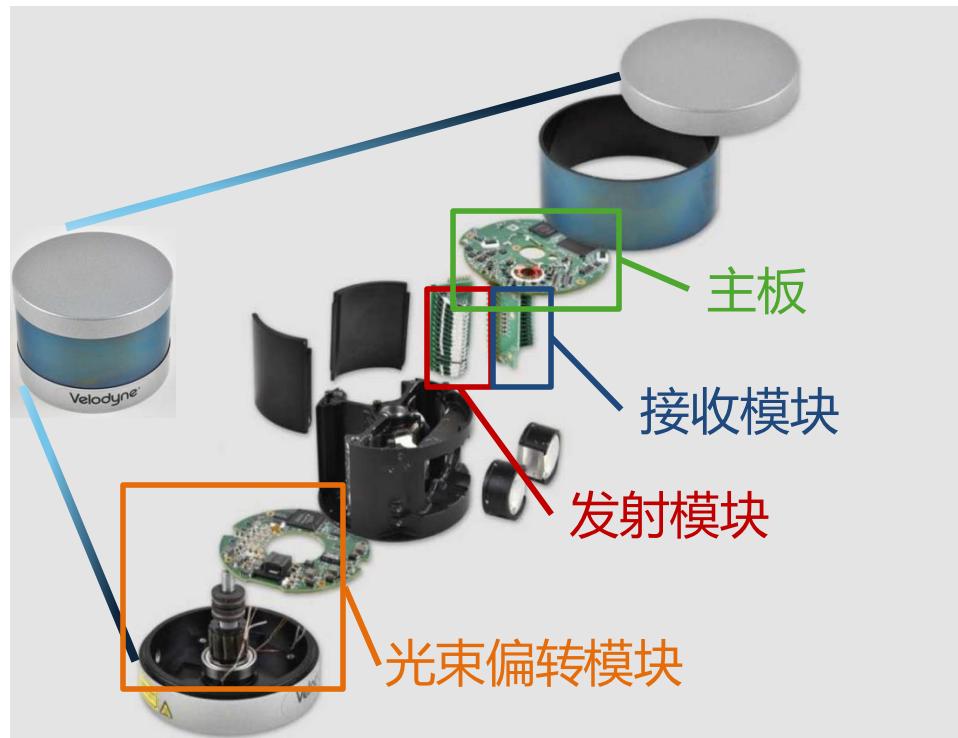
# 背景知识：激光雷达功能模块



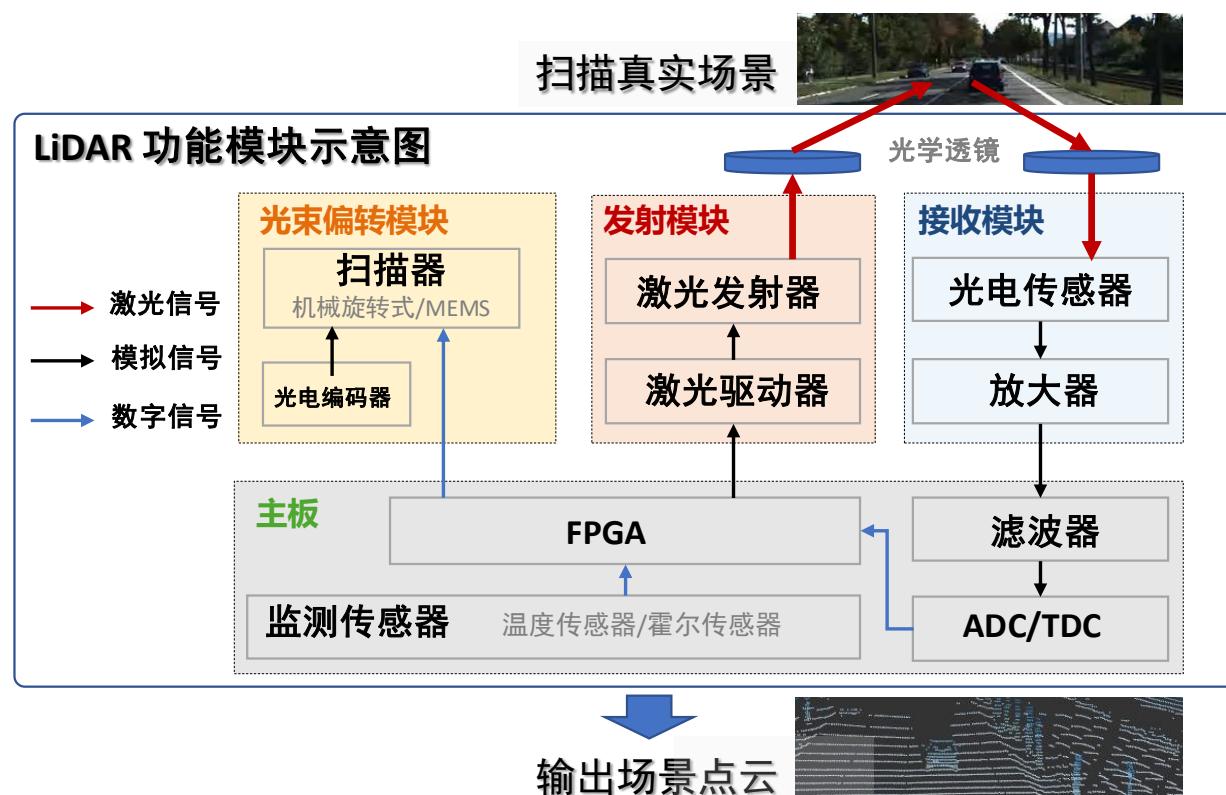
LiDAR功能模块实物拆解图

# 背景知识：激光雷达功能模块

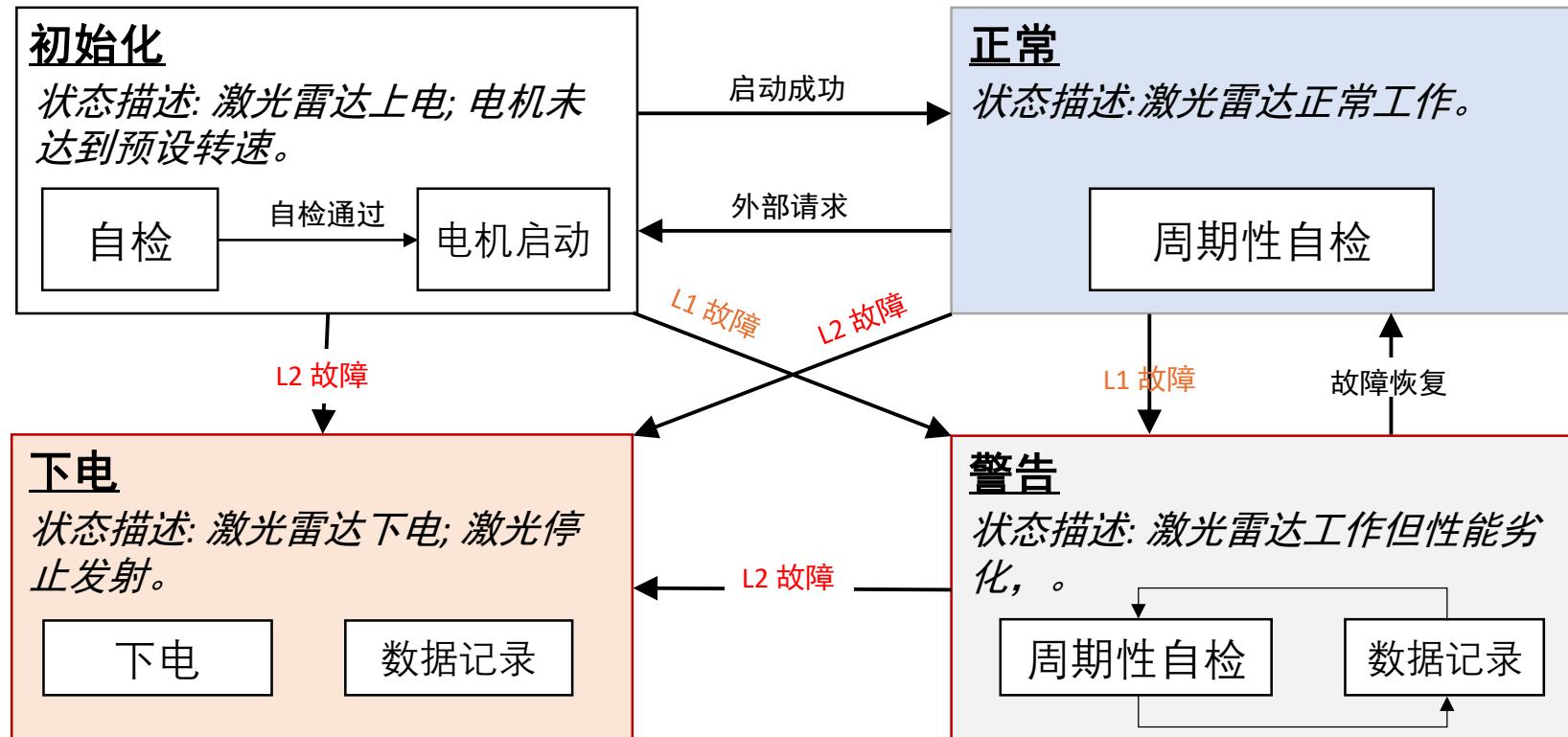
激光雷达系统4大功能模块：**发射模块、接收模块、光束偏转模块、主板。**



LiDAR功能模块实物拆解图



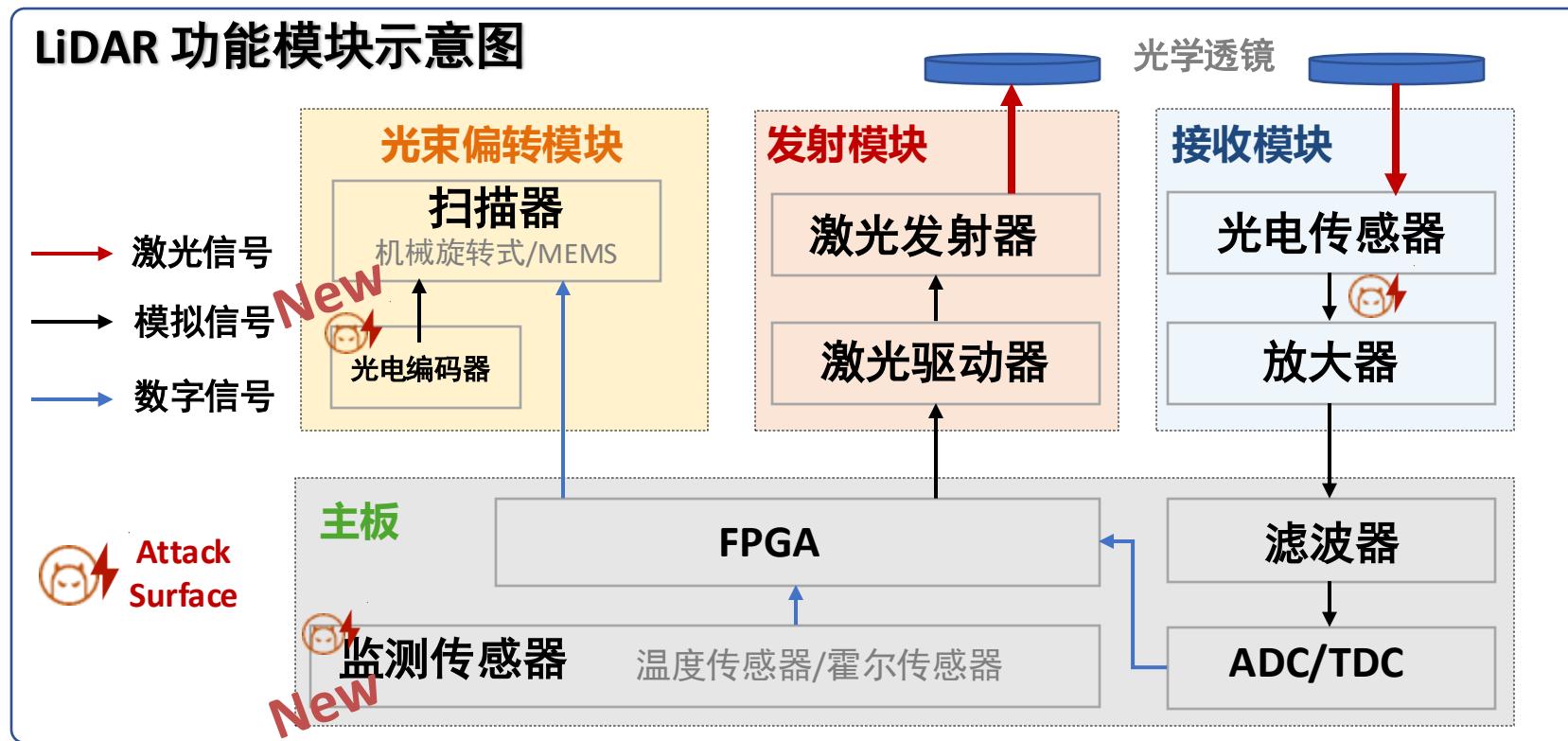
# 背景知识：LiDAR 错误检测和诊断机制



- › **L1 级故障:** 影响程度较低的故障, 如温度变化、电压波动
- › **L2 级故障:** 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

# 原理介绍 – 攻击入口

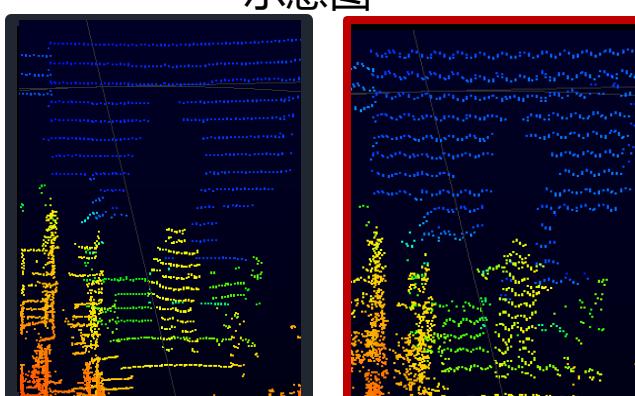
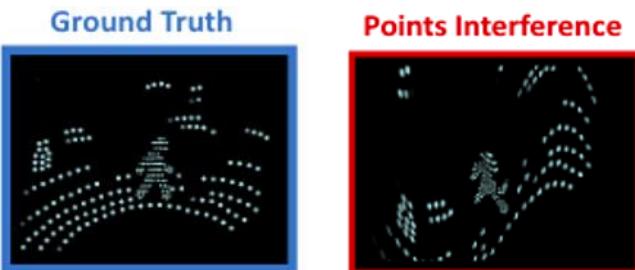
口 攻击入口包括 (1) 接收模块的模拟电路, (2) 主板上的温度监测传感器, (3) 光束偏转模块中的光电编码器。



# 原理分析 - 点云干扰攻击

## □ 攻击效果

在激光雷达测距过程中引入干扰，从而使点云**扭曲失真**

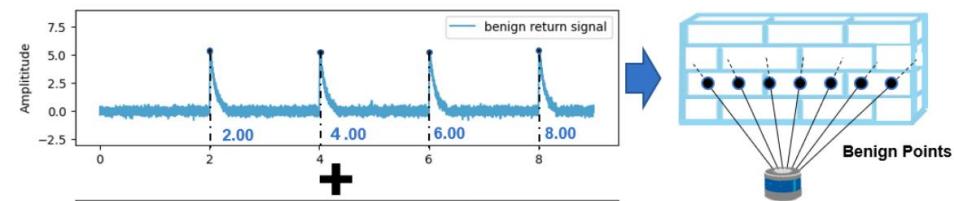


真实攻击效果

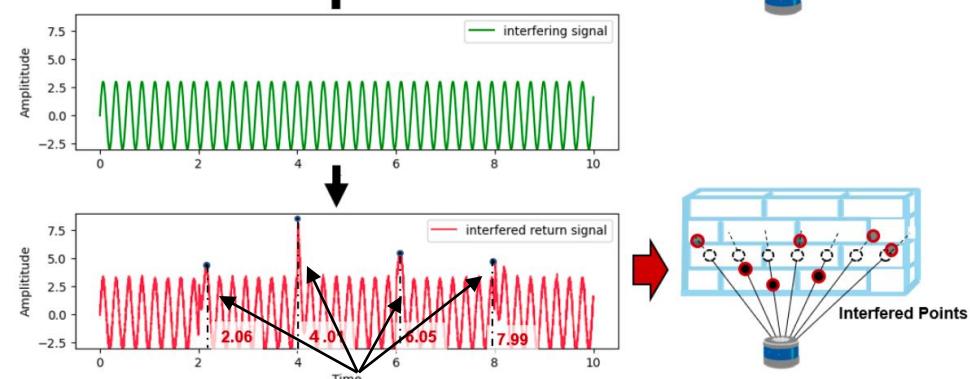
## □ 攻击原理

- **攻击入口:** 接收器的模拟电路
- **攻击信号:** 特定频率正弦波
- **攻击原理:** 正弦电磁干扰会使得回波信号的**峰值时刻**产生微小偏移，这会使**测距产生误差**，使激光点**偏离真实位置**

未受攻击的  
回波信号



电磁干扰



被攻击的  
回波信号

微小偏移

# 原理分析 - 点云抹除攻击

## 口 攻击效果

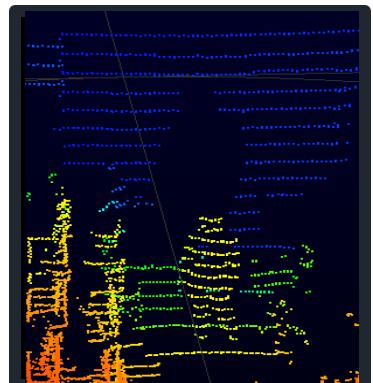
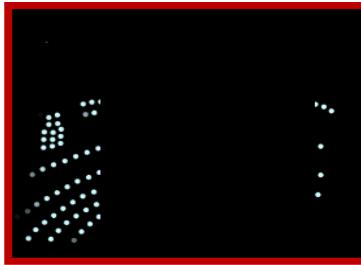
使点云明显偏离原来位置或彻底消失

Ground Truth



示意图

Points Removal

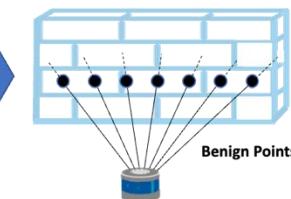
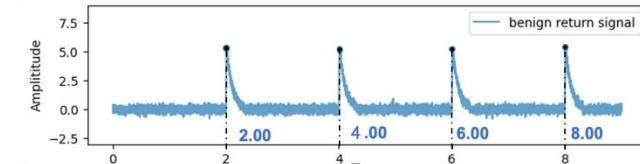


真实攻击效果

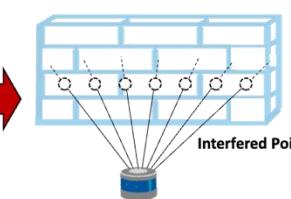
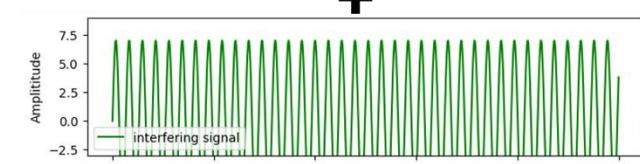
## 口 攻击原理 (其一)

- 攻击入口: 接收器的模拟电路
- 攻击信号: 特定频率高强度正弦波
- 攻击原理: 高强度正弦电磁干扰会使得接收器模拟电路饱和, 使真实回波脉冲信号被“淹没”。

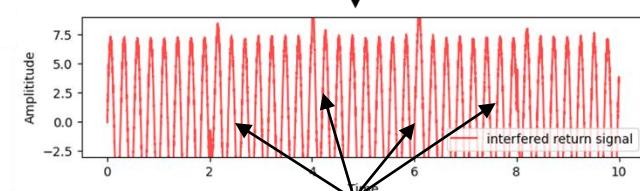
未受攻击的  
回波信号



高强度电磁  
干扰



被攻击的  
回波信号



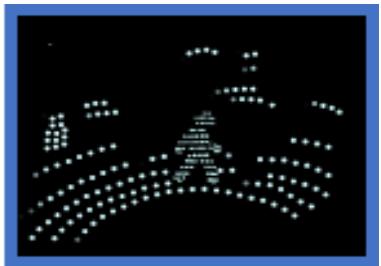
回波脉冲信号无法被检测到

# 原理分析 - 点云抹除攻击

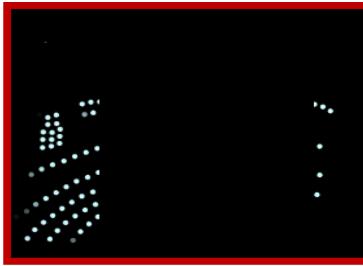
## 口 攻击效果

使点云明显偏离原来位置或彻底消失

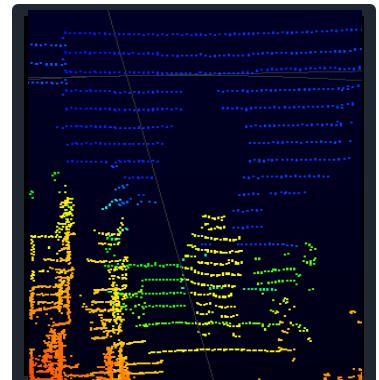
Ground Truth



Points Removal



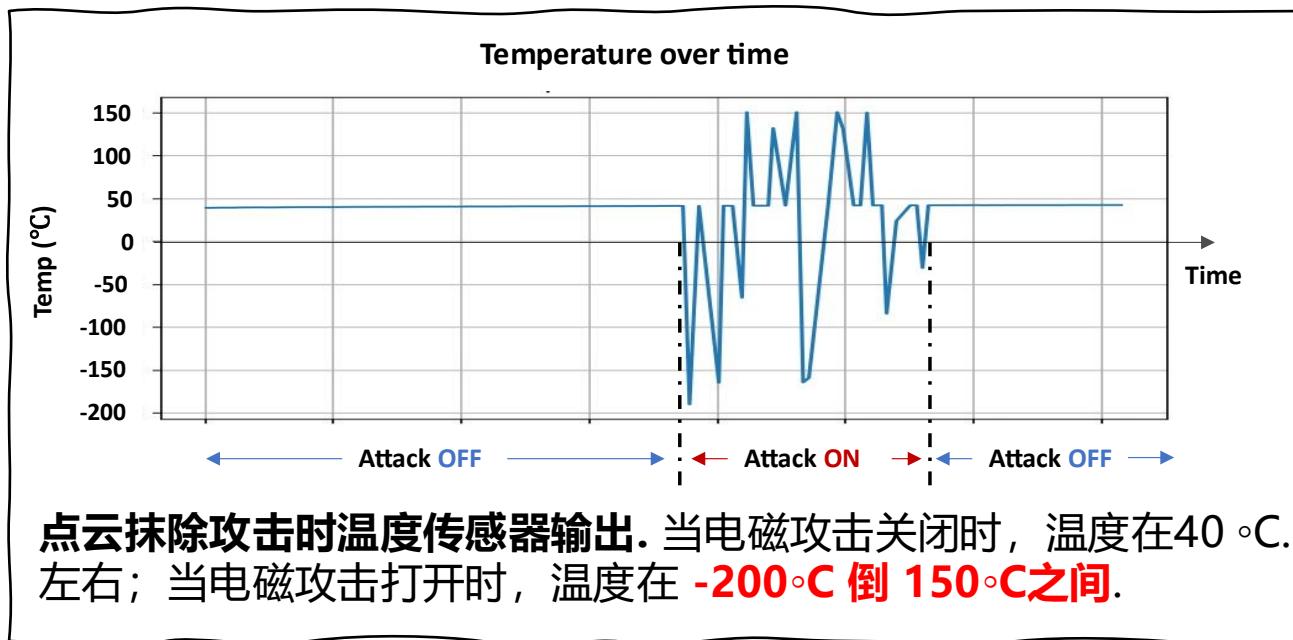
示意图



真实攻击效果

## 口 攻击原理 (其二)

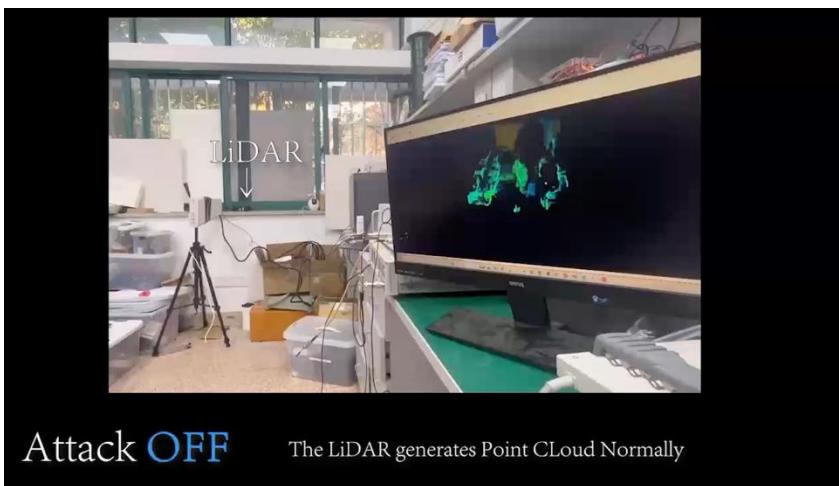
- 攻击入口:** 激光雷达监测传感器——温度传感器
- 攻击信号:** 特定频率正弦波
- 攻击原理:** 利用电磁攻击使得温度传感器出错，诱倒激光雷达的错误诊断机制报L1级错误，使LiDAR将此刻的点云舍弃。



# 原理分析 -雷达宕机攻击

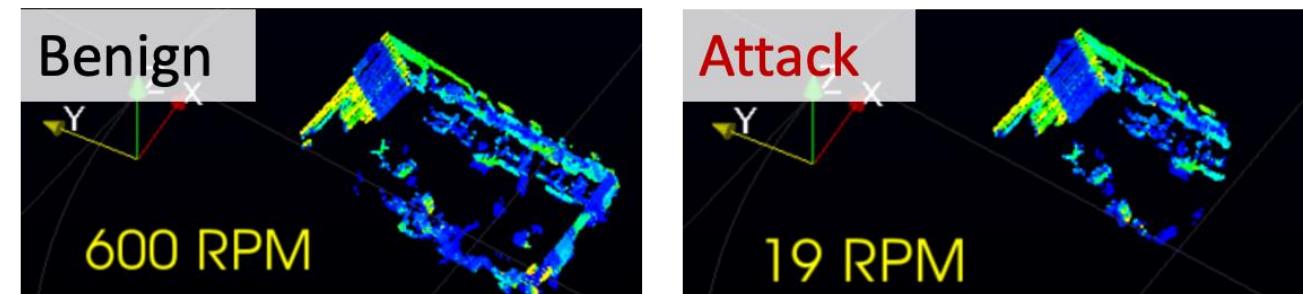
## 口 攻击效果

激光雷达掉电下机



## 口 攻击原理

- **攻击入口:** 光束偏转模块的光电编码器
- **攻击信号:** 特定频率正弦波
- **攻击原理:** 利用电磁攻击使得光电编码器出错，诱倒激光雷达的错误诊断机制报L2级错误，使LiDAR宕机。



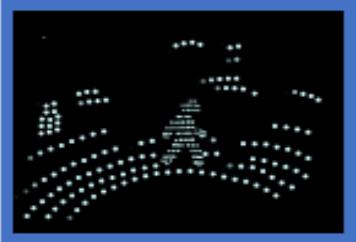
**雷达宕机攻击时LiDAR转速输出.** 当电磁攻击关闭时，转速在 600RPM；当电磁攻击打开时，转速掉到19RPM，然后LiDAR宕机。

# 原理分析 – 点云注入攻击

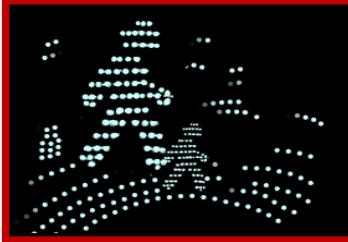
## 口 攻击效果

注入可控点

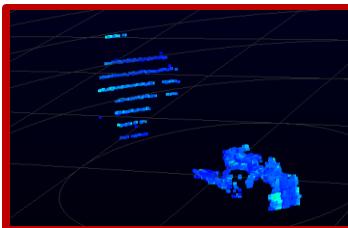
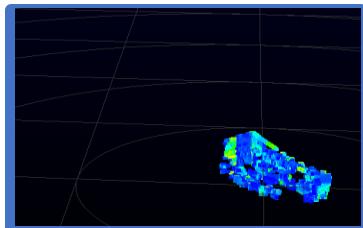
Ground Truth



Points Injection



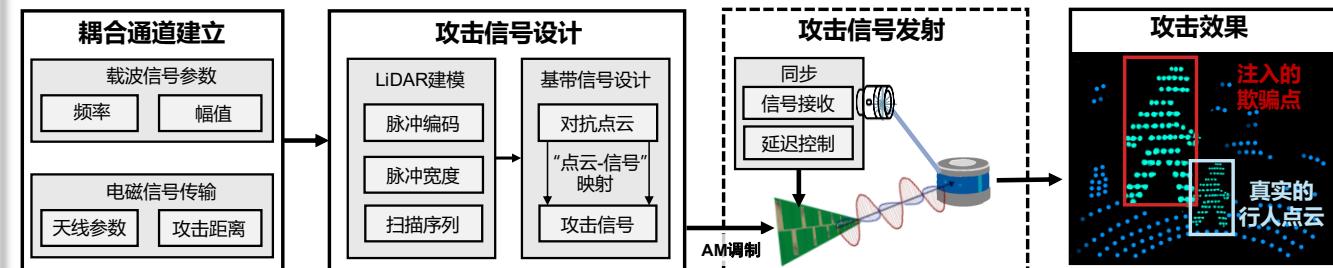
示意图



真实攻击效果

## 口 攻击原理:

- **攻击入口:** 接收器的模拟电路
- **攻击信号:** AM调制正弦波
  - 载波: 特定频率正弦波
  - 基波: 精心设计的脉冲信号
- **攻击原理:** 载波保证信号耦合到模拟电路中, 基波用来伪造激光雷达的回波。



# 实验评估

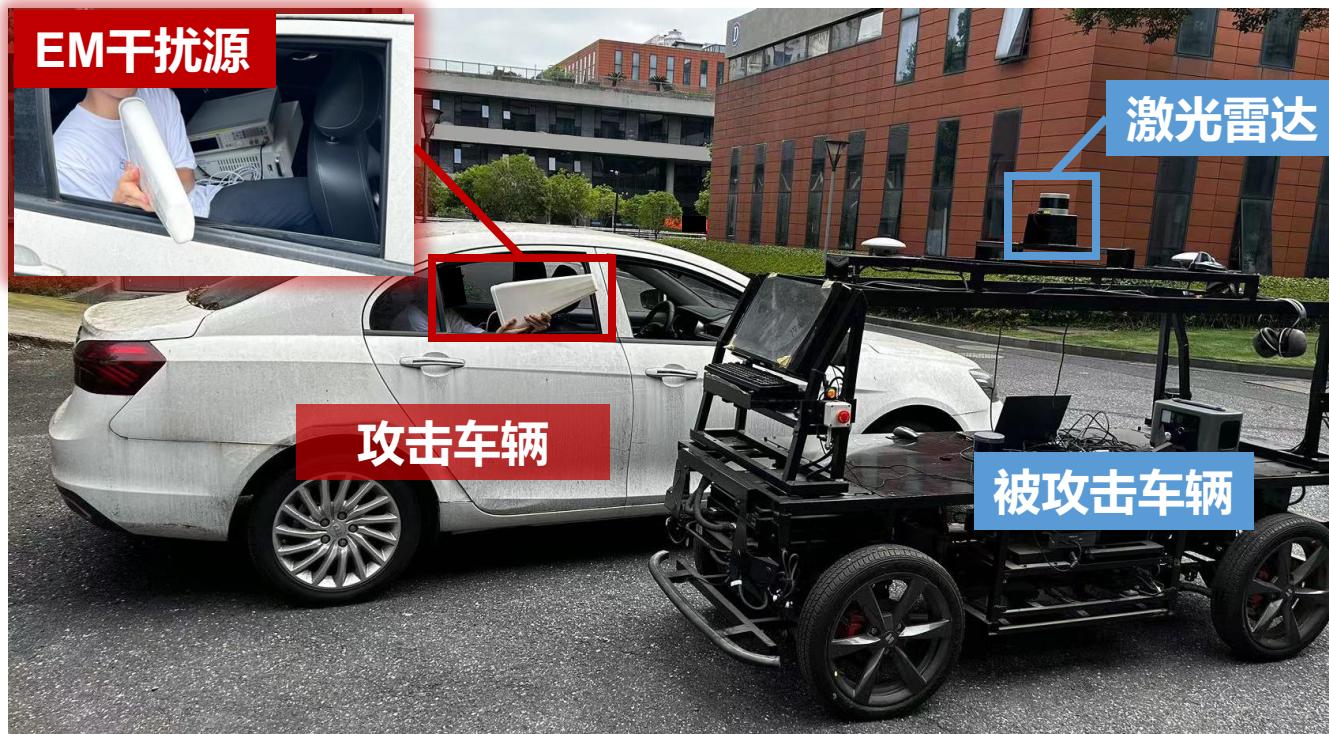
## □ 总览

- 1) 攻击了 5 款商用 LiDARs
- 2) 点云干扰
- 3) 点云抹除
- 4) 雷达宕机
- 5) 点云注入
- 6) 移动车上的可行性实验



# 实验评估 – 移动攻击可行性实验

口 攻击目标: 使基于LiDAR的自动驾驶感知模型无法检测到指定车辆。



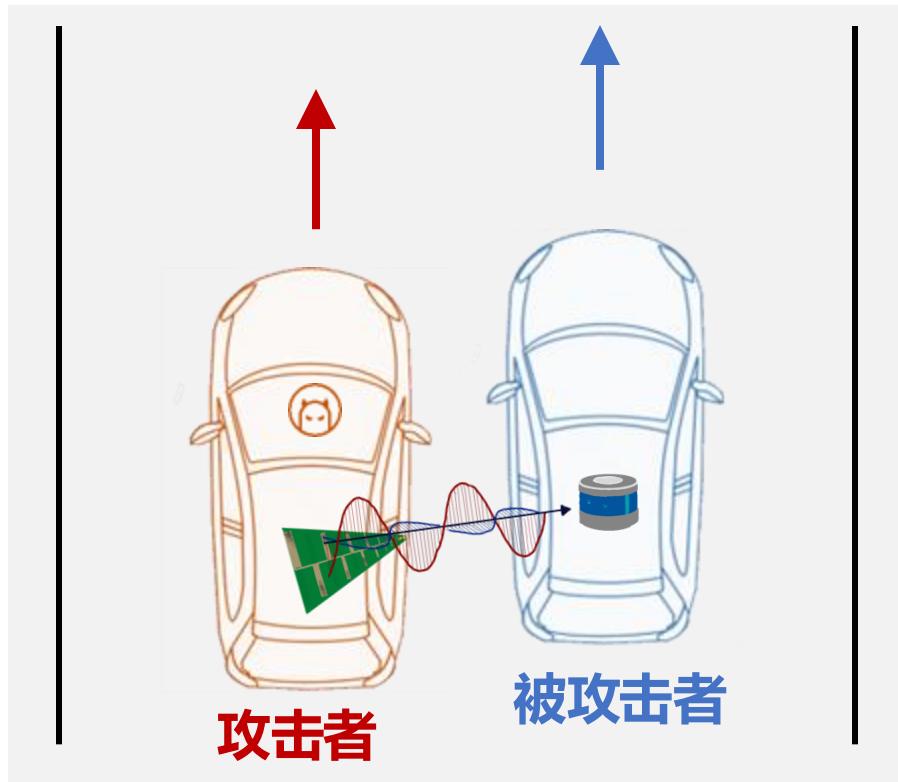
(a) 攻击设置



(b) 攻击设备

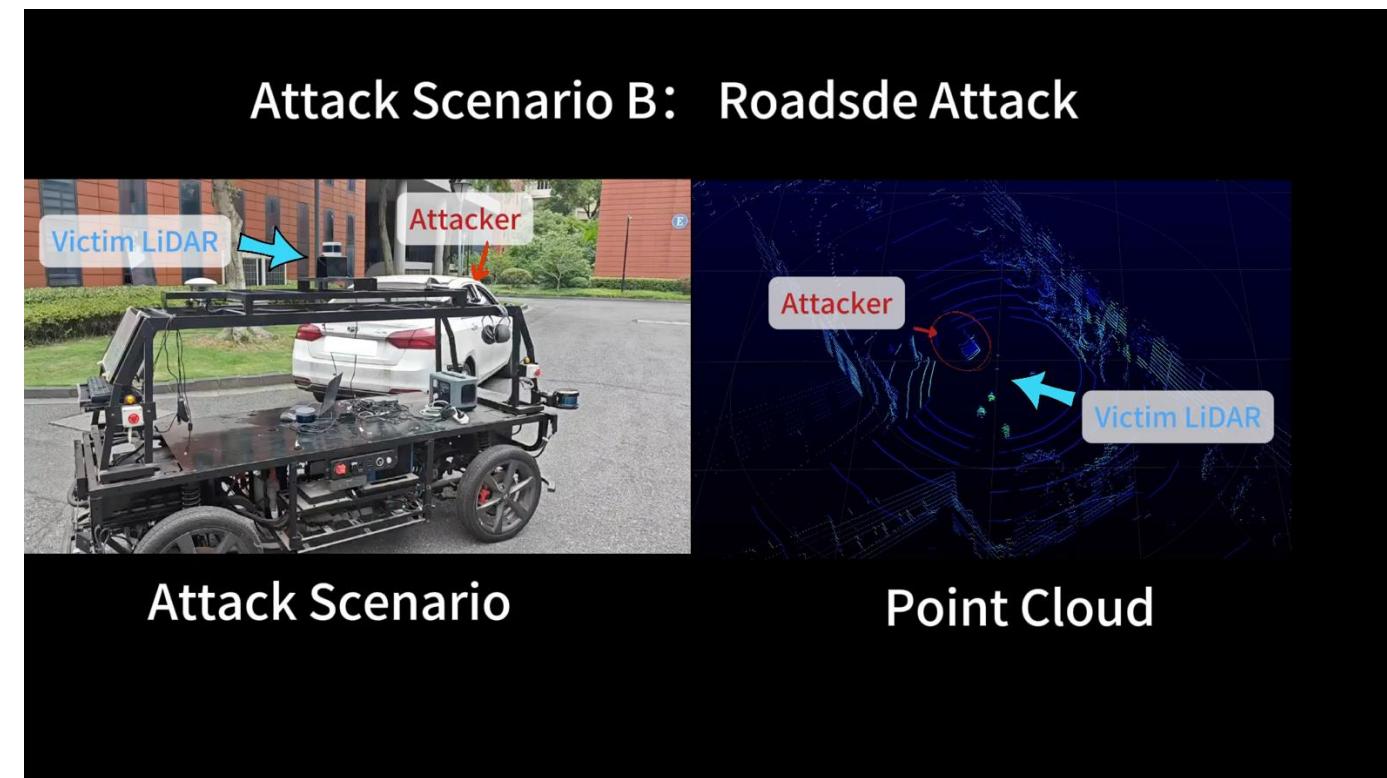
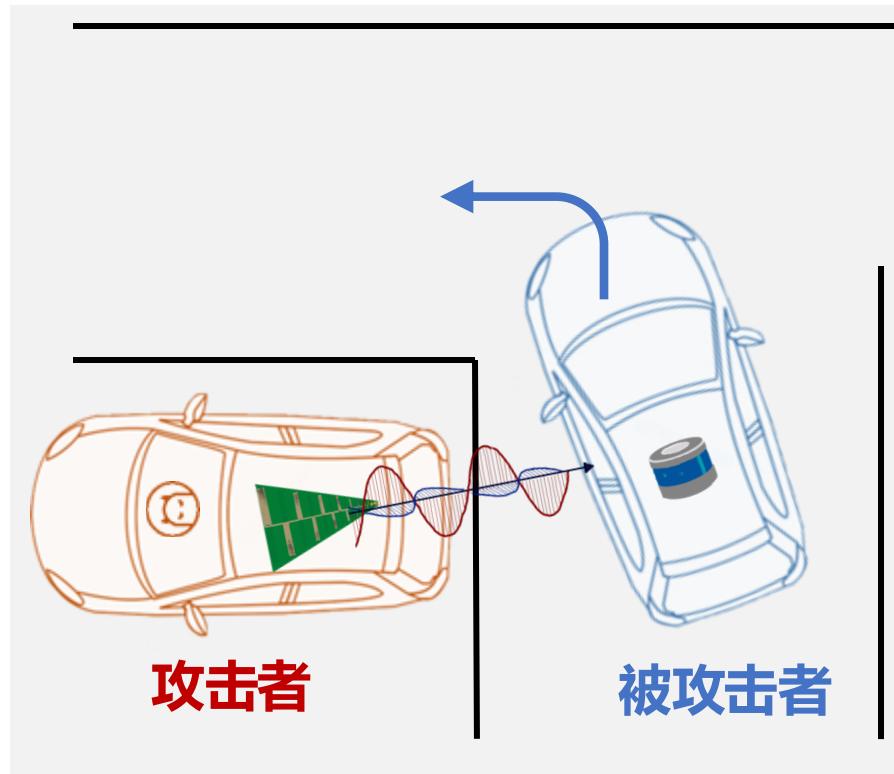
# 实验评估 – 移动攻击可行性实验

口 攻击场景1: 尾随攻击，攻击者和被攻击者保持相近的速度行驶。



# 实验评估 – 移动攻击可行性实验

口 攻击场景2: 路侧攻击, 攻击者在路边不动, 被攻击者转弯。



# 小结

- **新脆弱性:** 1) 发现了新的攻击面: 温度传感器, 光束偏转模块的光电编码器; 2) 新的攻击机理: 利用LiDAR内部的检测和诊断机制
- **多样的攻击效果:** 1) 点云干扰; 2) 点云抹除; 3) 激光雷达宕机; 4) 点云的可控注入。
- **物理世界实现:** 无需精确瞄准, 能实现对移动目标的攻击

# 成果3： 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

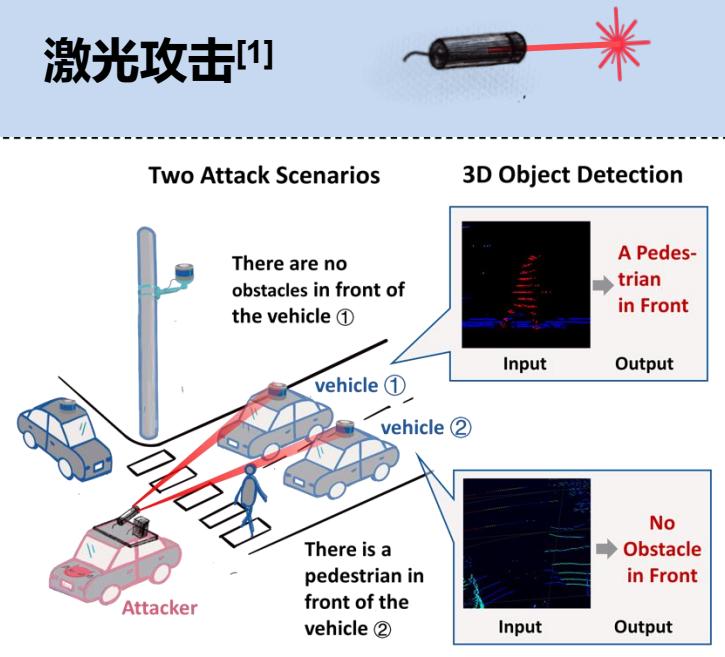
## 环节2：算法鲁棒性测评

第一作者《Unity is Strength? BeRobustness of Fusion-based 3D Object Detection against Physical Sensor Attack》发表于 *TheWebConf(WWW) 2024 (CCFA, Oral)*

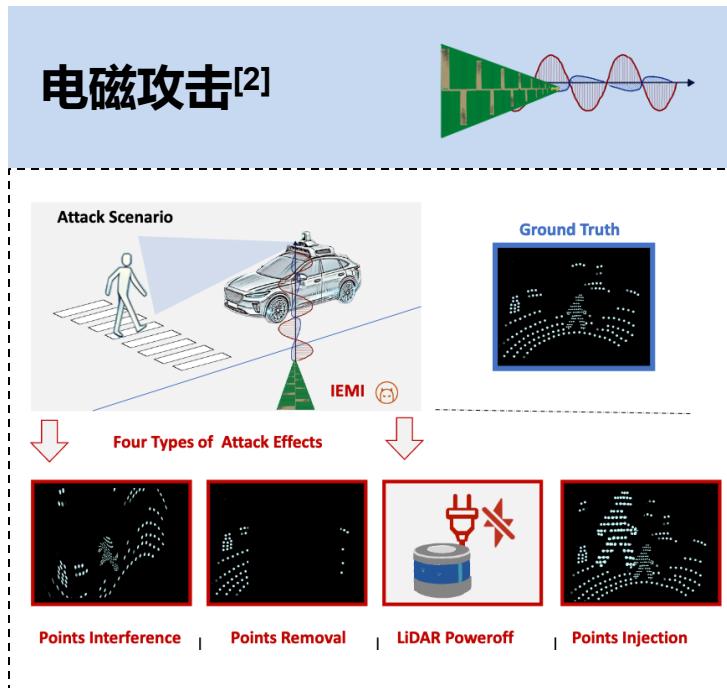
# 研究动机：自动驾驶感知面临严重威胁

自动驾驶依赖传感器及后续算法进行感知，然而**传感器暴露在物理环境中，已有大量研究证明了传感器容易受到极端环境或人为恶意的物理信号干扰。**

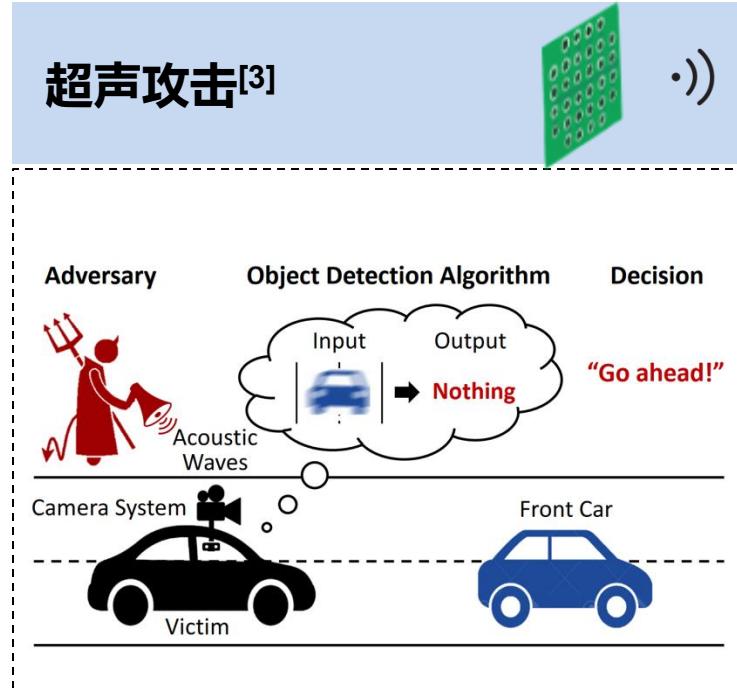
## 激光攻击<sup>[1]</sup>



## 电磁攻击<sup>[2]</sup>



## 超声攻击<sup>[3]</sup>



[1] Jin, Zizhi, et al. "Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle." 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023

[2] Jiang, Qinhong, et al. "{GlitchHiker}: Uncovering Vulnerabilities of Image Signal Transmission with {IEMI}." 32nd USENIX Security Symposium (USENIX Security 23). 2023.

[3] Ji, Xiaoyu, et al. "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision." 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021

# 研究动机：信号注入攻击需要引起重视

- **信号注入攻击定义：**利用光、声、磁等各种物理信号，造成传感器测量出错。
- **特点：** (1) 物理可实现； (2) 隐蔽性强：非抵近，高隐蔽； (3) 危害性大：  
能通过黑盒方式直接影响感知输出，并影响决策。



# 研究动机：信号注入攻击需要引起重视

- **信号注入攻击定义：**利用光、声、磁等各种物理信号，造成传感器测量出错。
- **特点：** (1) 物理可实现：在物理环境中发射信号； (2) 隐蔽性强：非抵近，高隐蔽； (3) 危害性大：能通过黑盒方式影响感知输出。



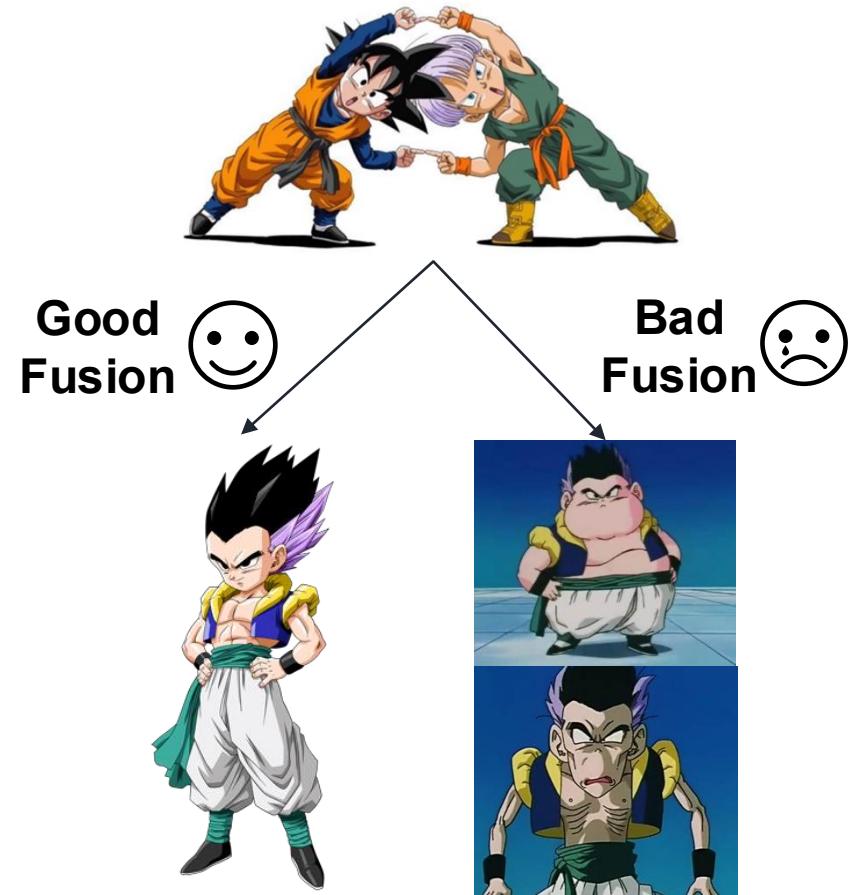
# 研究动机：关键假设缺乏系统性验证

**常见的假设：**  
**传感器融合是防御信号注入攻击的潜在方法。**



10.3.2 Sensor-Level Defenses. Several defenses could be adopted against spoofing attacks on LiDAR sensors:  
5) **Sensor Fusion:** Defense by sensor fusion enhances resiliency against transduction attacks by utilizing output from multiple sensors.  
**Sensor Fusion Techniques.** Another complementary defense approach is to exploit sensor fusion for decision making.  
**Multi-sensor Fusion and Security Redundancy.** Another complementary defense approach is to exploit multi-sensor fusion for decision-making. Autonomous vehicles can employ multiple types of sensors, e.g., cameras, radars, ultrasonic sensors combined with LiDARs to perceive the environment. Such information fusion and redundancy may help further improve the security of autonomous vehicles.

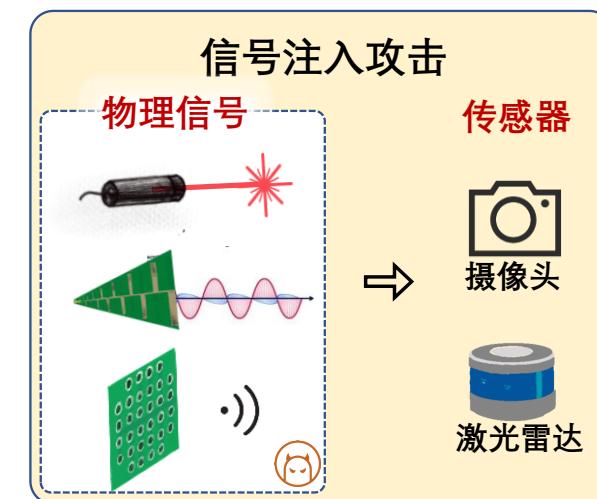
Is it True?



- [1] Jin, Zizhi, et al. "Pla-lidar: Physical laser attacks against lidar-based 3d object detection in autonomous vehicle." 2023 IEEE Symposium on Security and Privacy (SP). IEEE, 2023.  
[2] Jiang, Qinhong, et al. "{GlitchHiker}: Uncovering Vulnerabilities of Image Signal Transmission with {IEMI}." 32nd USENIX Security Symposium (USENIX Security 23). 2023.  
[3] Ji, Xiaoyu, et al. "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision." 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.

# 工作简介

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**2个研究问题**：1. 融合是否能增强鲁棒性？2. 不同的融合结构如何影响鲁棒性？



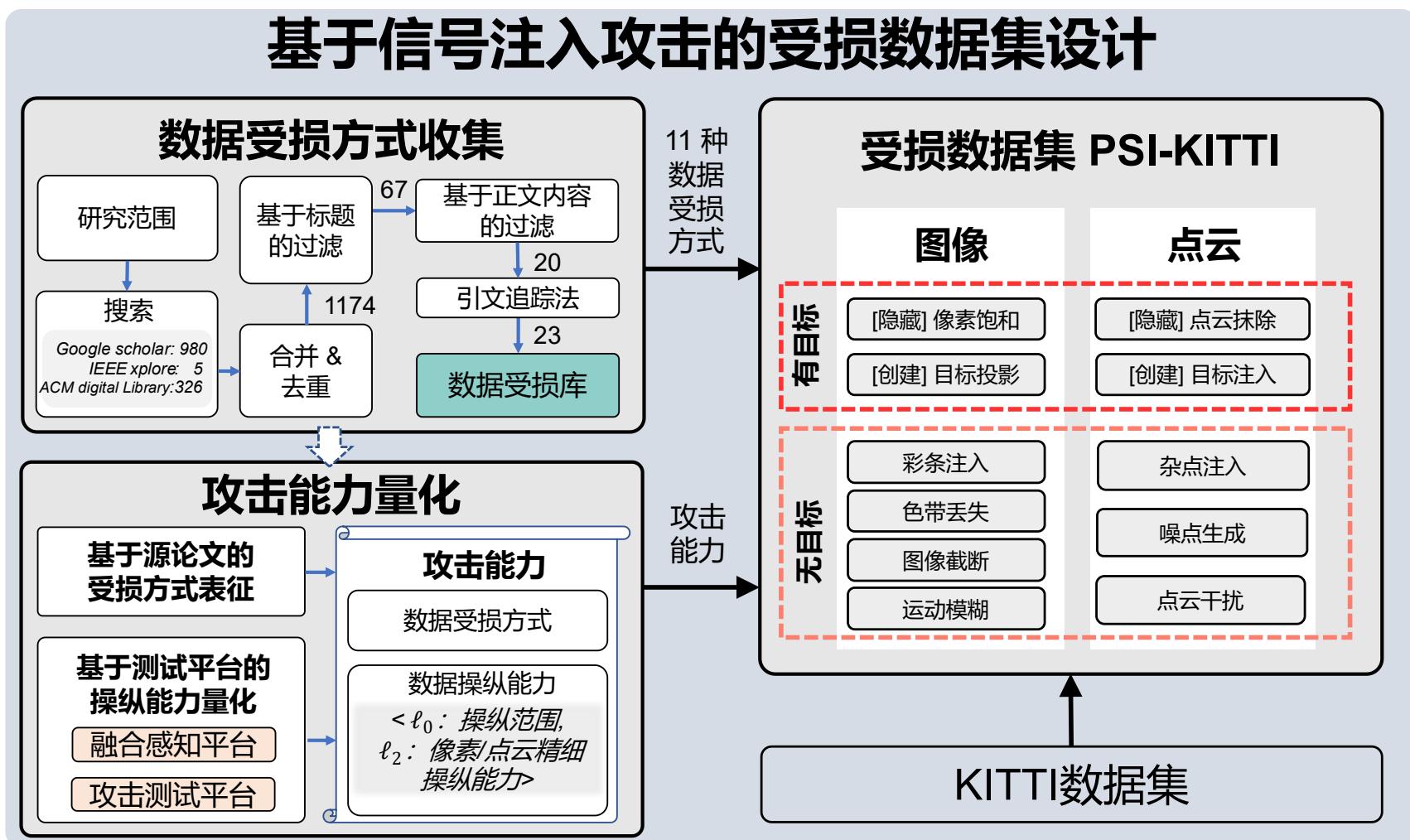
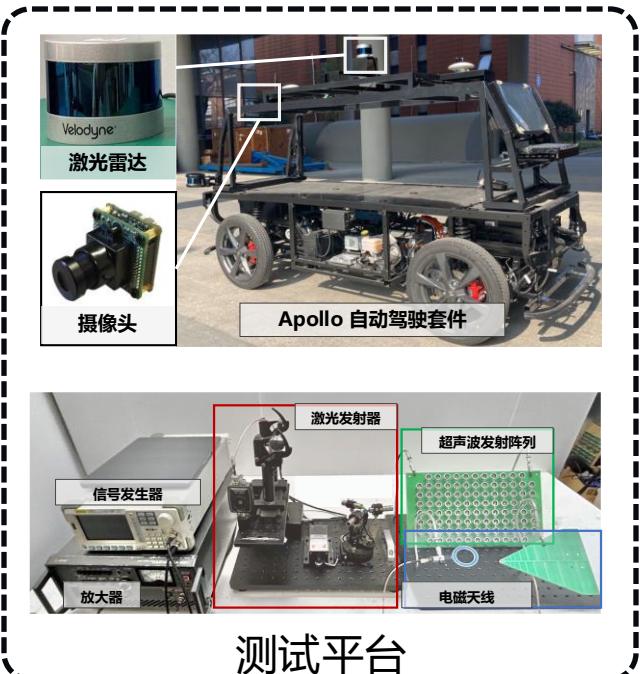
以往鲁棒性测评工作  
数据受损方式：**被动受损**

本工作  
数据受损方式：**主动激励**

# 基准测评数据集设计

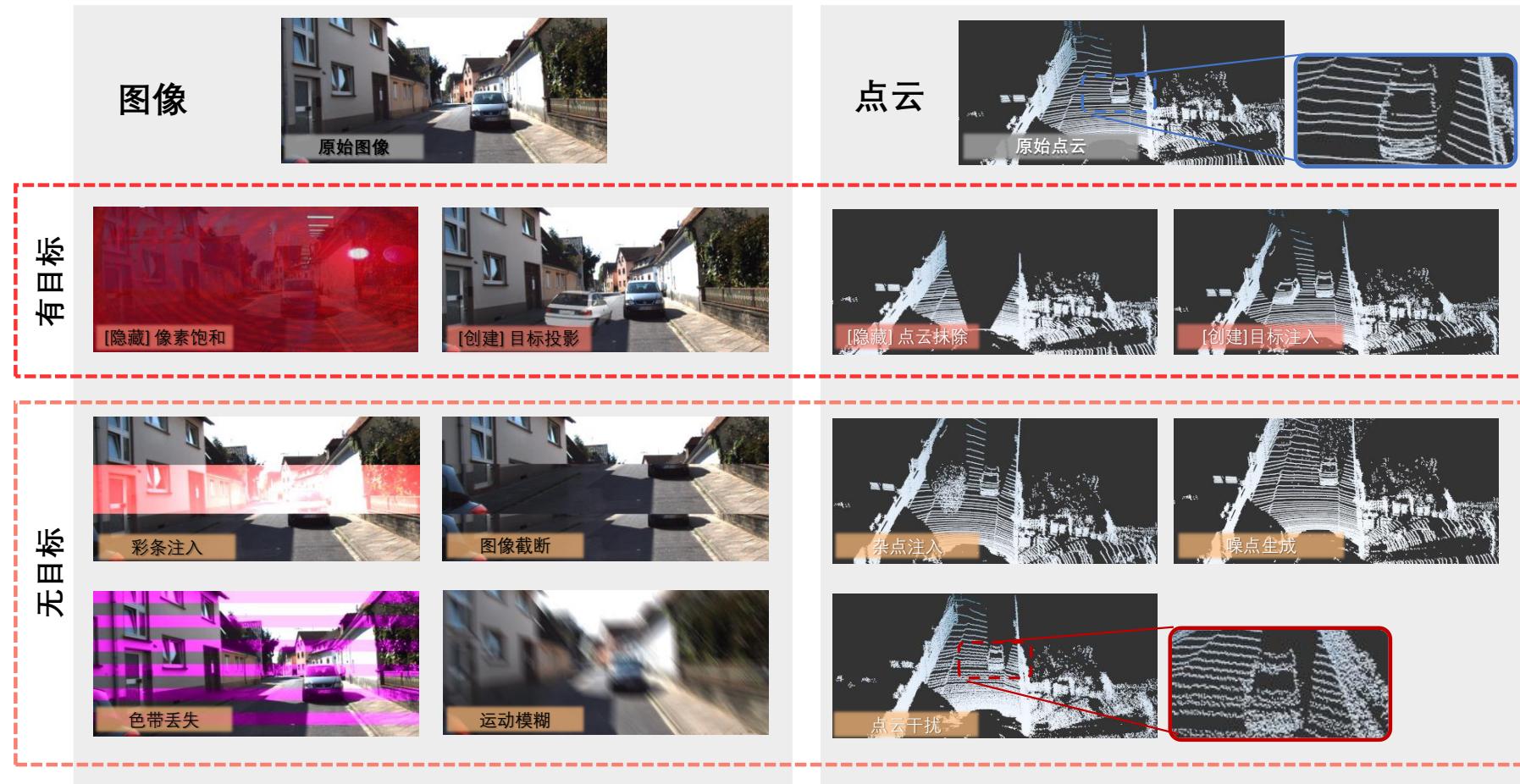
## 口 数据集特点

- 完备性:** 通过SLR, 包含所有物理信号注入攻击
- 物理可实现性:** 量化了攻击能力



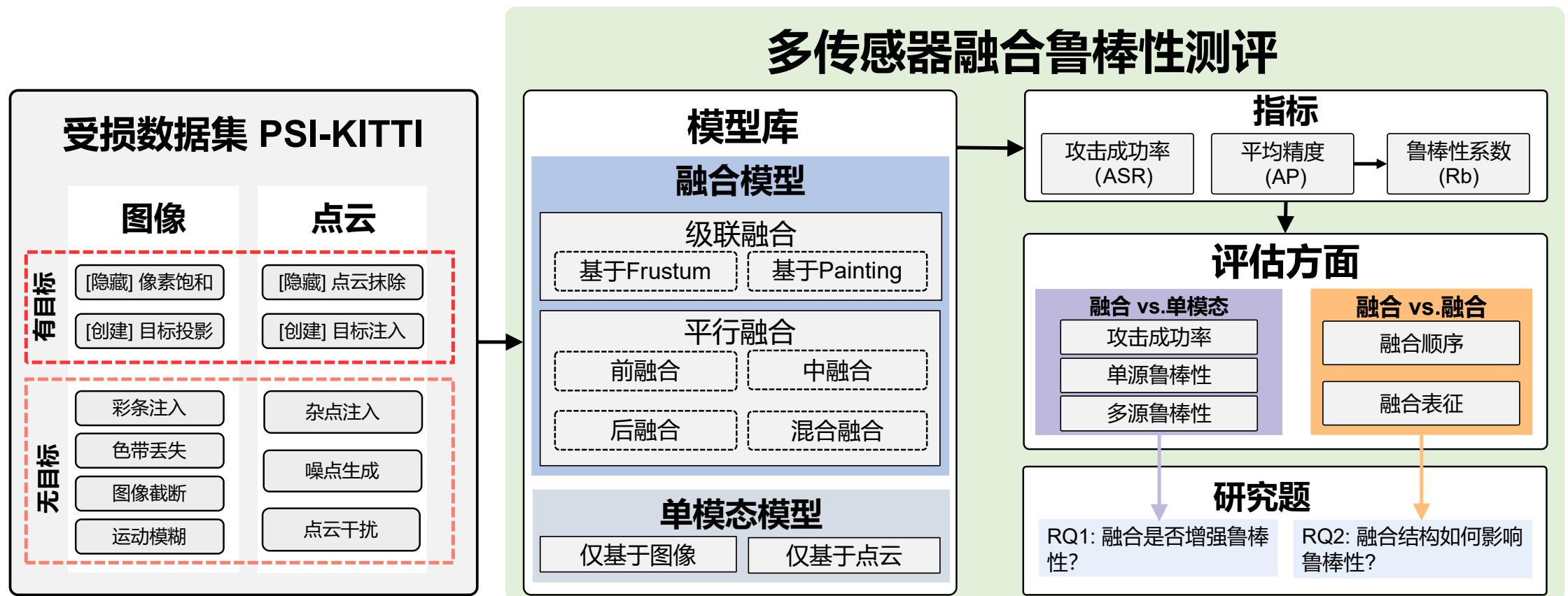
# 基准测评数据集受损方式介绍

口 数据集规模：6种图像受损 + 5种点云受损；共45228组<图像，点云>数据对



# 实验评估：鲁棒性测评流程

口 鲁棒性测评: 12 个模型 + 3 个指标 + 542736 组测试数据



# 实验评估：指标

- **平均精度 (Average Precision , AP):**
  - AP 代表了一个模型在某数据集上的整体性能:

$$AP|_{R_{40}} = \frac{1}{|R_{40}|} \sum_{r \in R_{40}} \max_{r' > r} \rho(\tilde{r})$$

- **鲁棒性 (Robustness , Rb):**
  - 记一个模型针对某种数据受损方式 $c$ 的鲁棒性为 $Rb_c$ ， 则:

$$Rb_c = \frac{AP_c}{AP_{clean}}$$

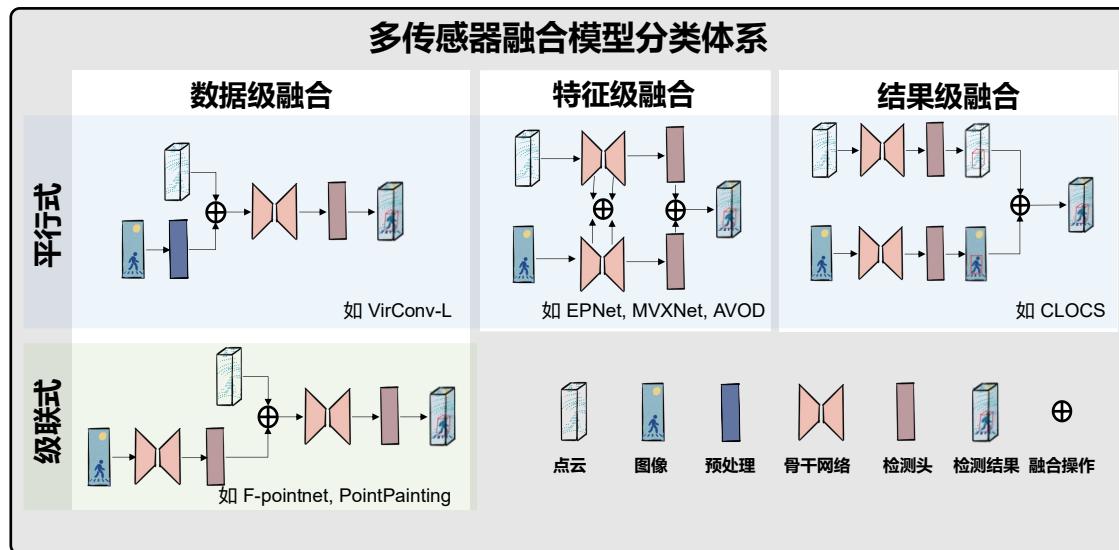
# 实验评估：融合是否增强鲁棒性

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
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

研究问题 1：融合是否增强鲁棒性？ - **融合不一定更鲁棒**

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。

# 实验评估：融合结构如何影响鲁棒性？

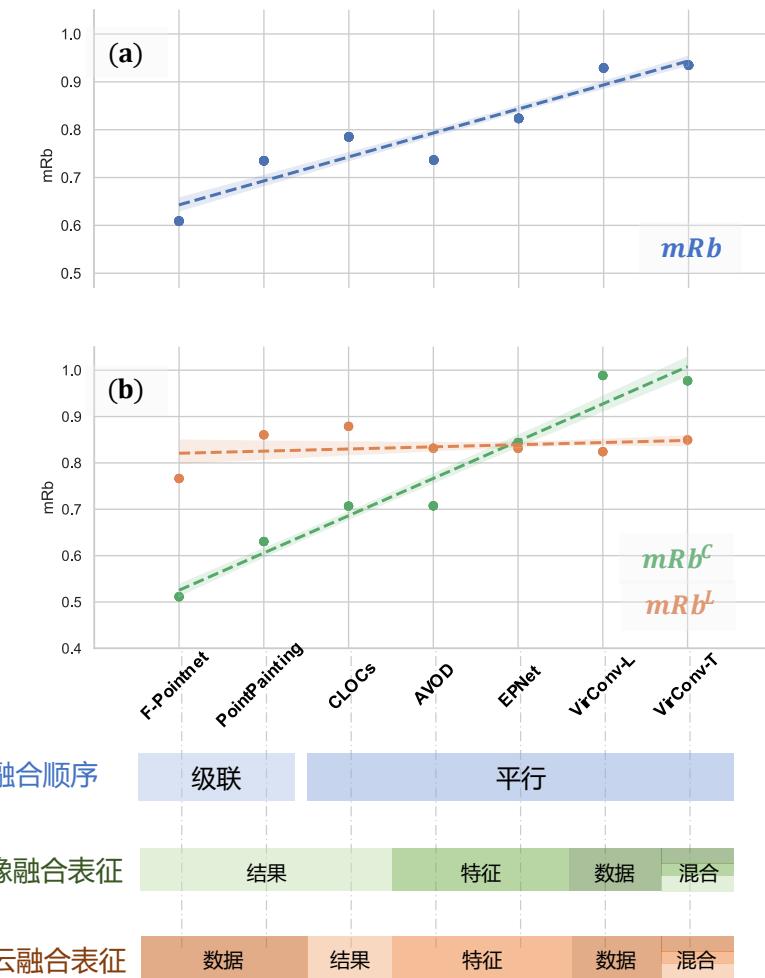


研究问题 2: 融合结构如何影响鲁棒性? - 融合模态的信息熵越大, 越鲁棒

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越全面, 鲁棒性越强, 信息的全面性排序为: 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。

所有受损、图像受损、点云受损下的平均鲁棒性



# 鲁棒性提升建议

具有如下3个特点的多传感器融合模型可能表现出更强的鲁棒性。

- **数据融合**: 在原始数据层面进行融合, 而不是在特征或结果层面。
- **平行融合**: 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。
- **模态独立**: 每个模态都具有能够独立于其他模态完成3D 目标检测任务的能力

# 小结

- **新基准测试数据集：**设计并开源了一个**信号注入攻击数据集**，涵盖**11种物理信号注入攻击**。
- **关键研究问题的评估：**在12个模型上进行了542736组数据的评估，回答了自动驾驶安全中学界一直关心但没有通过实验论证的**两个研究问题**：1) 融合是否更安全？  
2) 哪种融合架构更安全？

## 成果4：

# 信号注入攻击**检测**和**防护** 关键技术研究

### 环节3：攻击检测与防护

[1] 第一作者. 《Laser-based LiDAR Spoofing: Effects Validation, Capability Quantification, and Countermeasures》, *IEEE Internet of Things Journal(IoT-J)*, 2024.

[2] 第一作者. 《Physical Sensor Attack Robustness of Fusion-based Perception in Autonomous Driving: Benchmark and Defense》投稿中

# 工作简介

对信号注入攻击实现主动式**攻击检测**和被动式**鲁棒性增强**。

- **攻击检测**: (1) 针对性的攻击检测: 点云注入攻击检测 (2)通用性攻击检测方法: 受损模态判定;
- **鲁棒性增强**: 在实现攻击检测的基础上, 提出了基于虚拟点技术的**数据级、平行式、模态独立**多传感器融合架构: SIA-Defense。

## 多维度攻击检测

单一攻击的针对性检测:

基于点云表面曲率的**虚假目标检测**

所有攻击的通用性检测:

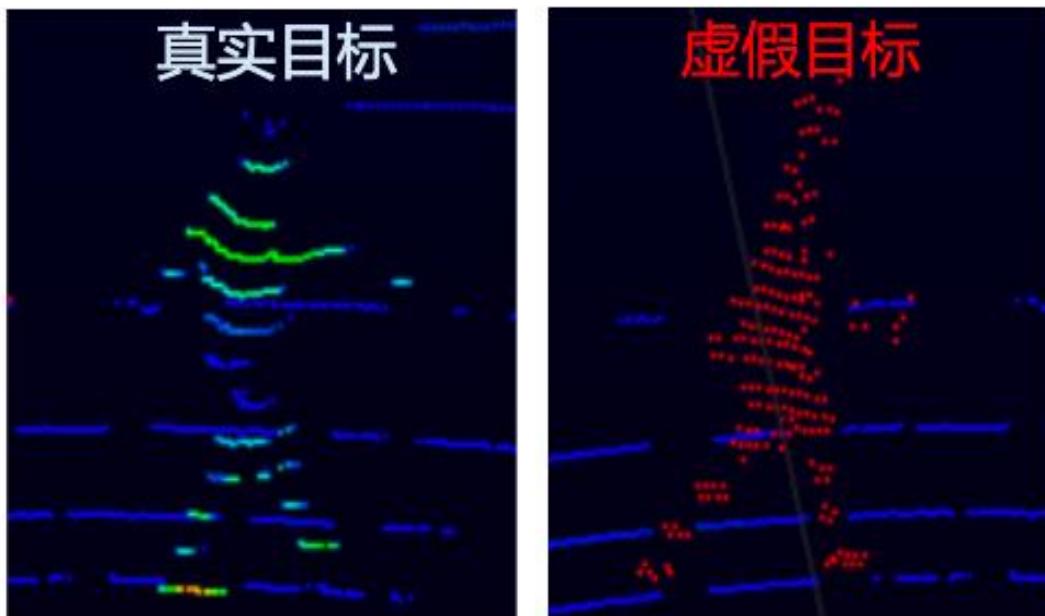
基于一致性分析和时序预测的**受损模态判定**

## 鲁棒性提升

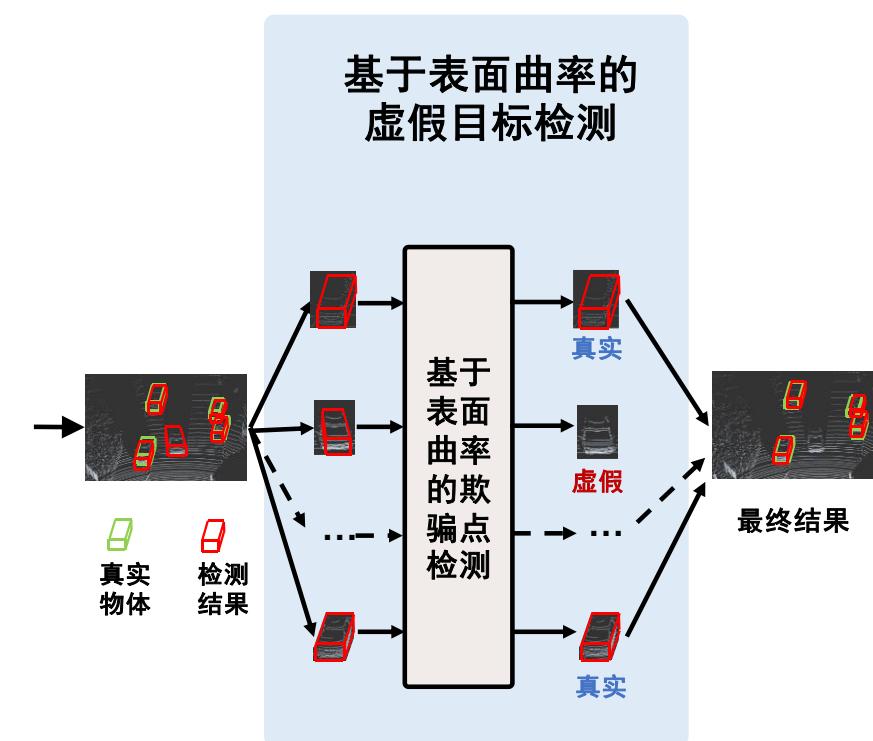
基于虚拟点技术的  
模态独立、平行式、  
**数据级**  
**多传感器融合架构**

# 攻击检测：基于点云表面曲率的虚假目标检测

**工作简介：**针对PLA-LiDAR攻击，基于注入的虚假目标点云和真实目标点云之间的差异，提出表面曲率系数，来检测点云注入攻击。



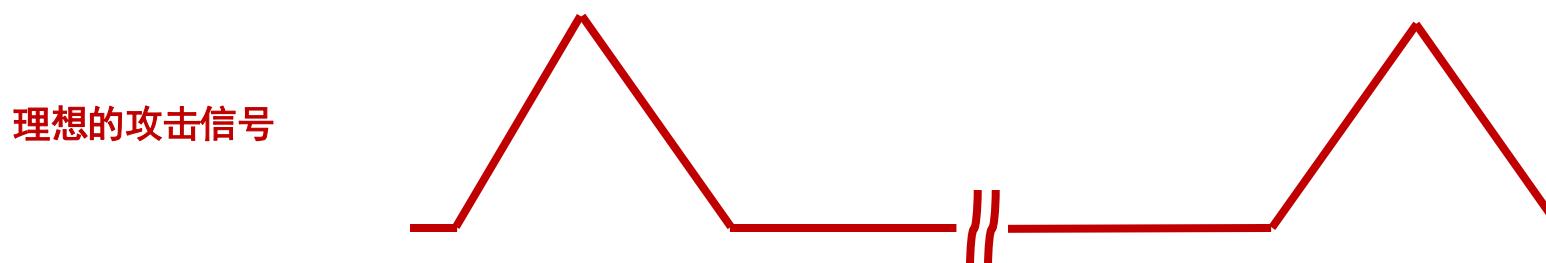
图：真实行人的点云和注入的虚假行人的点云对比



# 攻击检测：基于点云表面曲率的虚假目标检测

为什么注入的欺骗点云肯定存在误差？

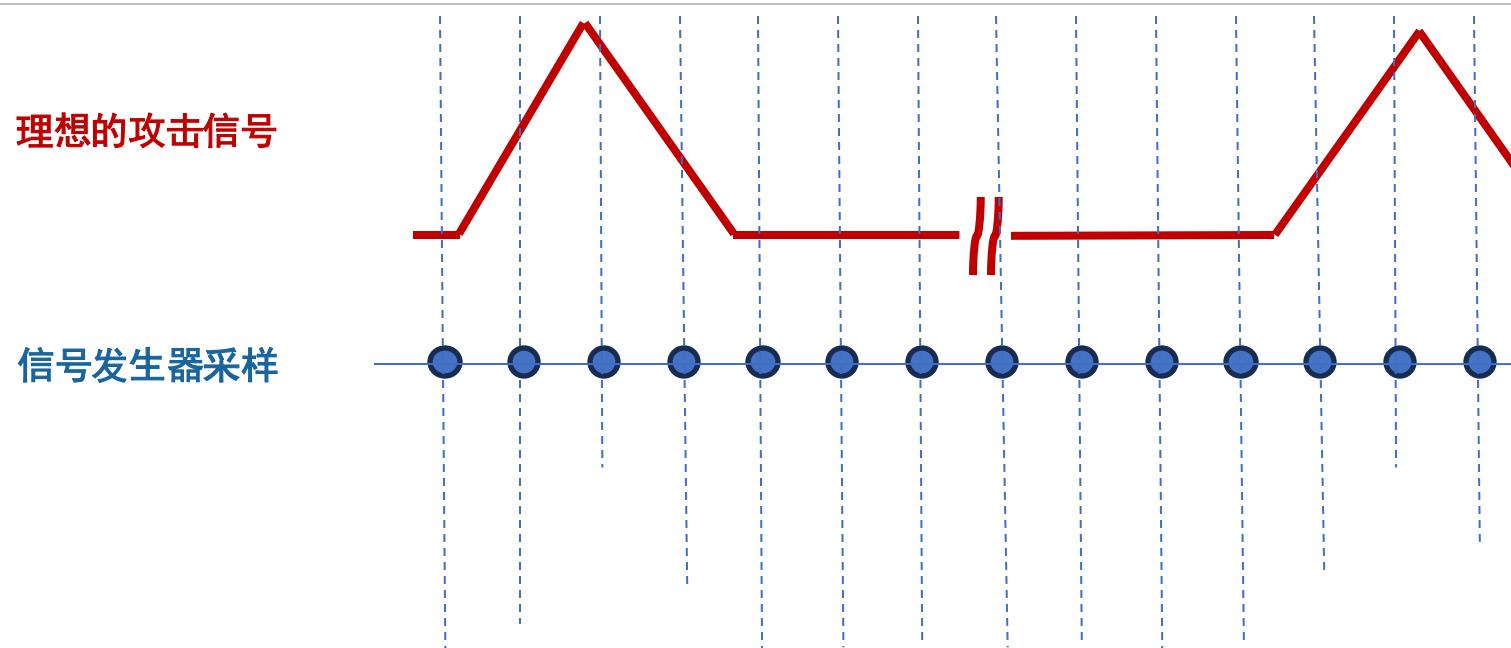
激光攻击中的“光速灾难”：有限成本（100万以内）攻击设备采样率有限（GHz级别），最多能将时间误差控制在纳秒级别，此时注入点云的误差仍然会达到约10厘米。



# 攻击检测：基于点云表面曲率的虚假目标检测

为什么注入的欺骗点云肯定存在误差？

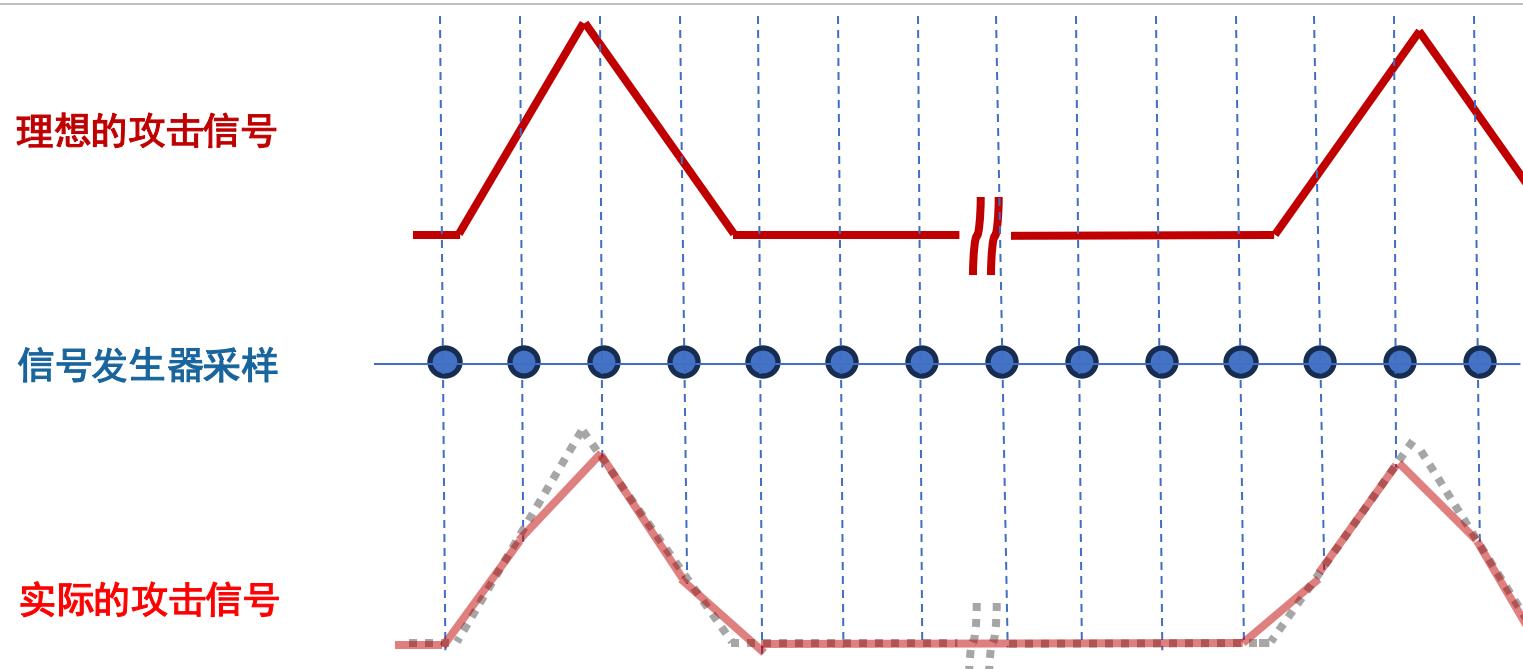
激光攻击中的“光速灾难”：有限成本（100万以内）攻击设备采样率有限（GHz级别），最多能将时间误差控制在纳秒级别，此时注入点云的误差仍然会达到约10厘米。



# 攻击检测：基于点云表面曲率的虚假目标检测

## 为什么注入的欺骗点云肯定存在误差？

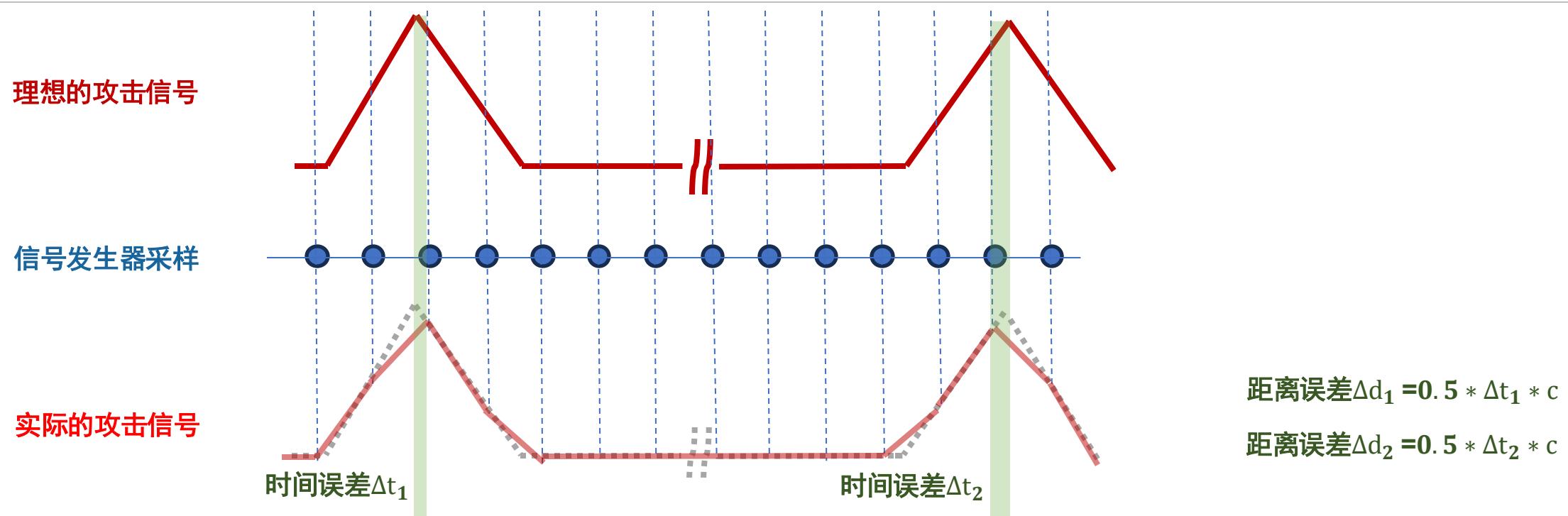
激光攻击中的“光速灾难”：有限成本（100万以内）攻击设备采样率有限（GHz级别），最多能将时间误差控制在纳秒级别，此时注入点云的误差仍然会达到约10厘米。



# 攻击检测：基于点云表面曲率的虚假目标检测

为什么注入的欺骗点云肯定存在误差？

激光攻击中的“光速灾难”：有限成本（100万以内）攻击设备采样率有限（GHz级别），最多能将时间误差控制在纳秒级别，此时注入点云的误差仍然会达到约10厘米。



# 攻击检测：基于点云表面曲率的虚假目标检测

## 口 如何表征欺骗点云的误差？

**表面曲率：**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，并求平均值以表征单个目标的曲率。

$$(1) \text{ 邻域质心计算: } \bar{\mathbf{p}}_i = \frac{1}{K} \sum_{j=1}^K \mathbf{p}_j^{(i)},$$

$$(2) \text{ 协方差矩阵: } \Sigma_i = \frac{1}{K-1} \sum_{j=1}^K (\mathbf{p}_j^{(i)} - \bar{\mathbf{p}}_i)(\mathbf{p}_j^{(i)} - \bar{\mathbf{p}}_i)^T$$

$$(3) \text{ 特征值分解: } \Sigma_i = \mathbf{U}_i \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} \mathbf{U}_i^\top,$$

$$\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0.$$

$$(4) \text{ 单点曲率计算: } c_i = \frac{\lambda_3}{\lambda_1 + \lambda_2 + \lambda_3}.$$

$$(5) \text{ 目标曲率计算: } C_{\text{surface}} = \frac{1}{N} \sum_{i=1}^N c_i$$

# 攻击检测：基于点云表面曲率的虚假目标检测

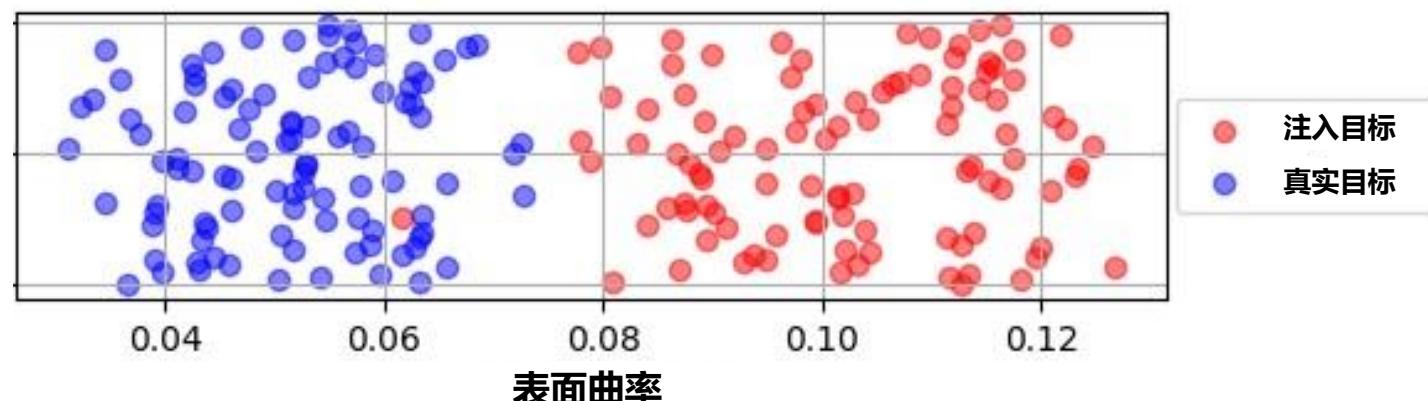
## □ 检测性能评估

### ■ 实验设置：

- 数据：100 个真实目标以及 100个注入目标
- 表面曲率经验性阈值0.76

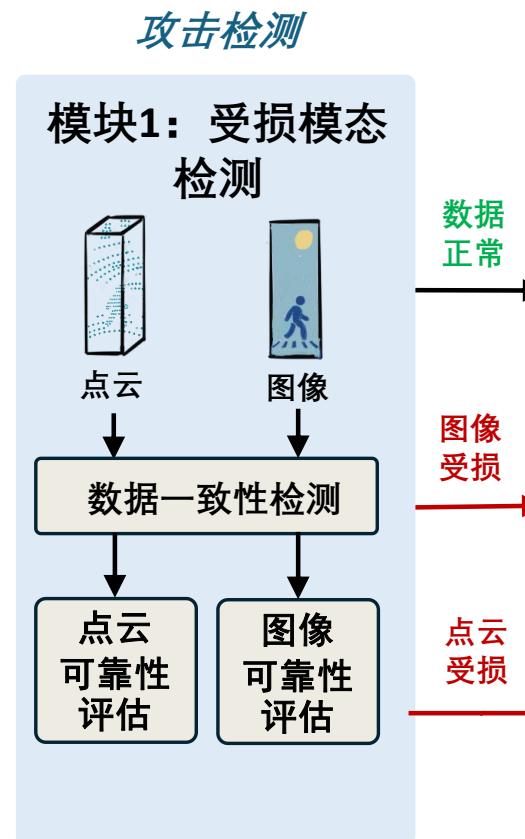
### ■ 实验结果：误警率**FPR = 0**；召回率**Recall = 99%**

图：100 个真实目标以及 100个注入目标的表面曲率区别



# 攻击检测：受损模态检测

口 工作简介：通过一致性分析和时序预测判断某一个传感器是否受损



- **数据一致性分析：**利用点云和图像之间的语义强相关性进行互相验证，初步判断数据是否受损
- **受损模态判定：**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。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。

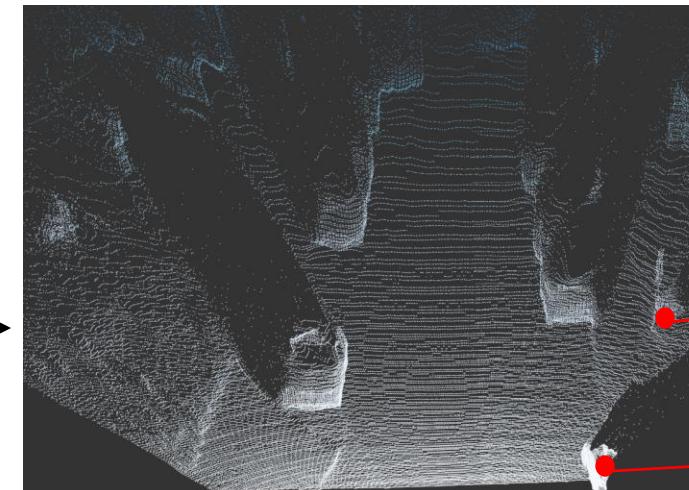
# 攻击检测：受损模态检测

## □ 数据一致性分析

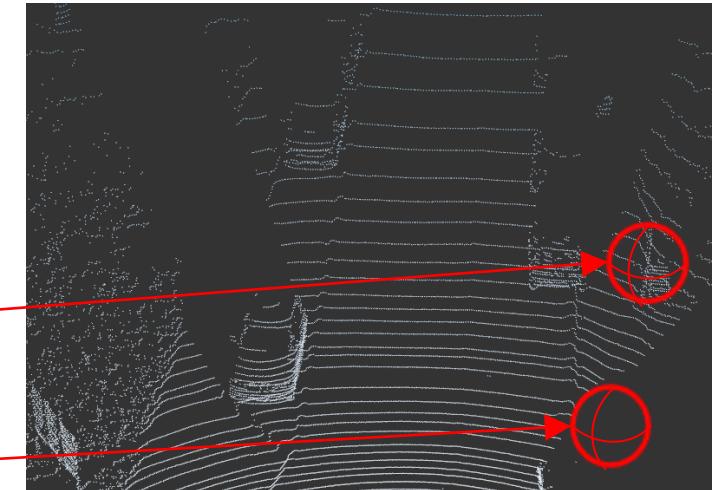


图像

PSMNet



虚拟点



点云

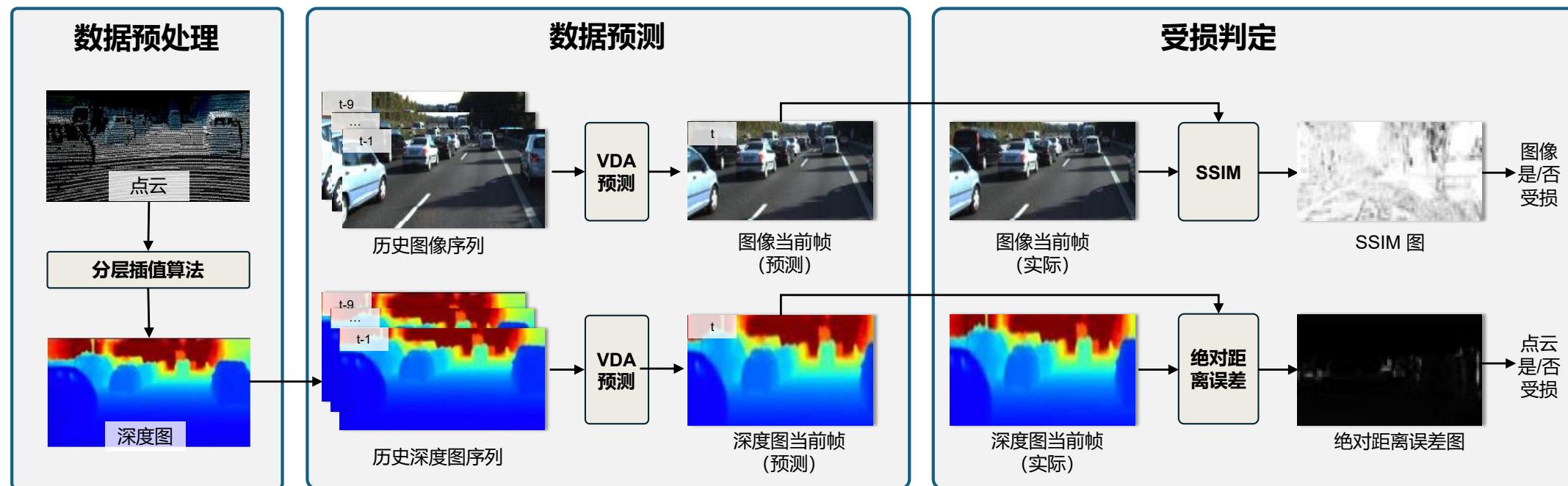
- 好点 (Good Point) : 以  $P_{Fake}$  为中心的半径为  $R$  的球形区域内, 存在真实点  $P_{Real}$ ,
- 坏点 (Bad Point) : 以  $P_{Fake}$  为中心的半径为  $R$  的球形区域内, 不存在真实点  $P_{Real}$

若好点数占比小于阈值  $T = 95\%$ , 认为存在数据受损。

# 攻击检测：受损模态检测

## 口 受损模态判定

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，可通过**比较实际采集到的数据和预测数据之间的差异**来判断哪个模态受损。



# 攻击检测：受损模态检测

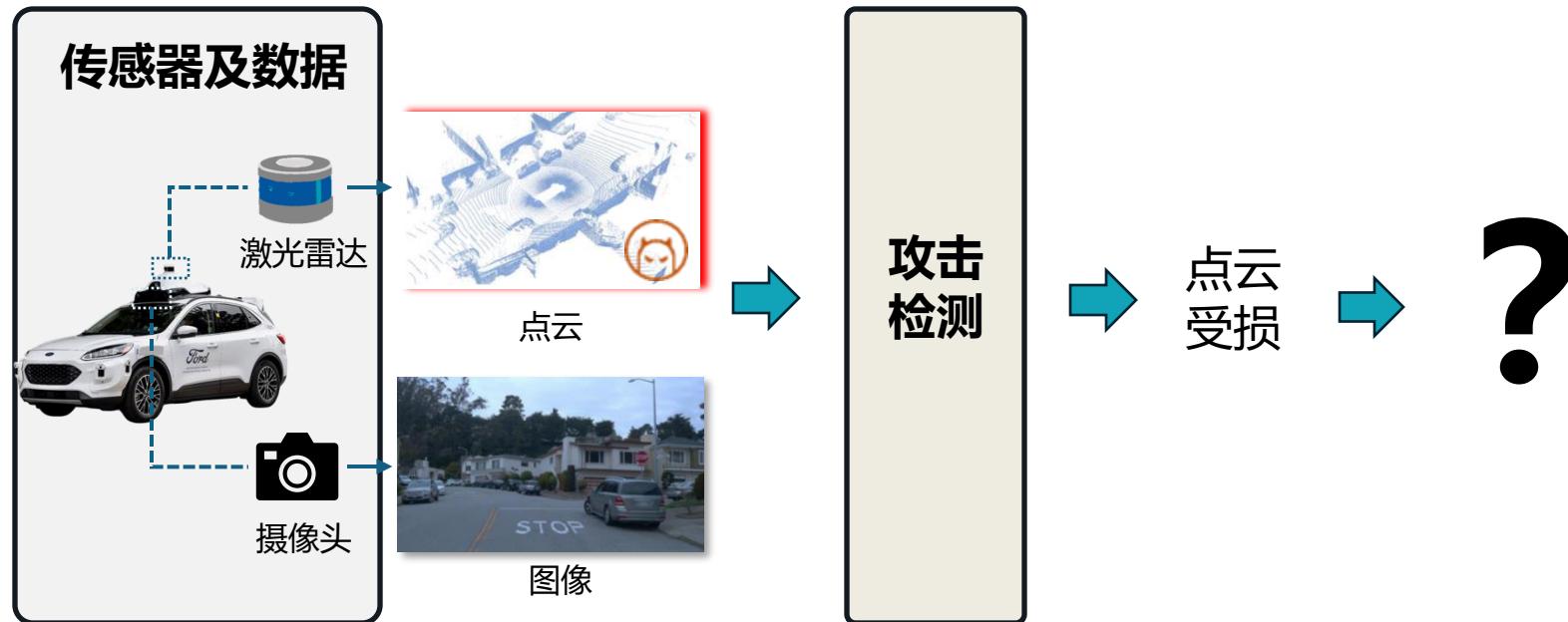
## 口 性能评估

- 3600组数据 (12\*300)
- 误警率: 0.67%
- 召回率: 图像受损100%，点云受损97.1%

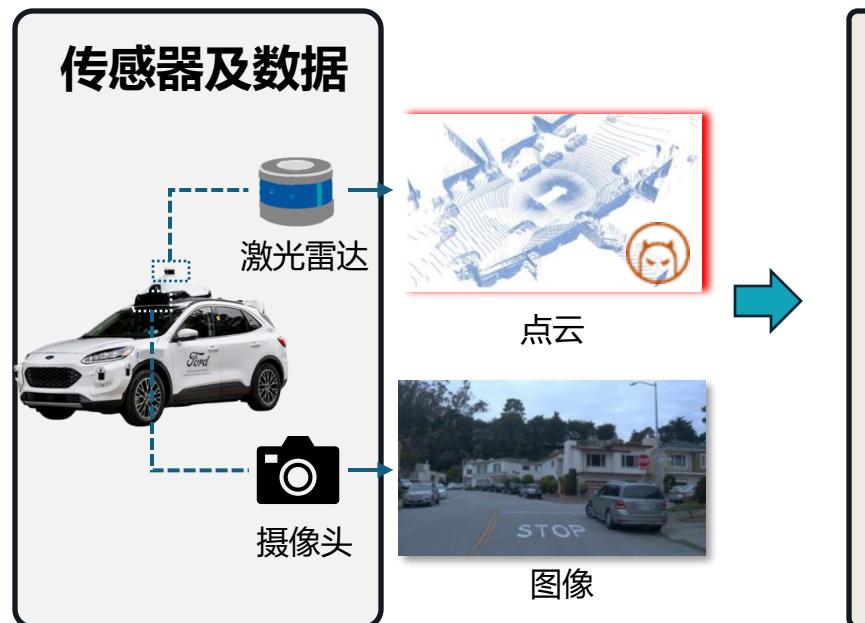
表格：数据受损检测器针对 11 种数据受损方式的召回率

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%

# 成功检测出攻击了！ 然后呢？



# 成功检测出攻击了！ 然后呢？



潜在方法

舍弃  
受损数据

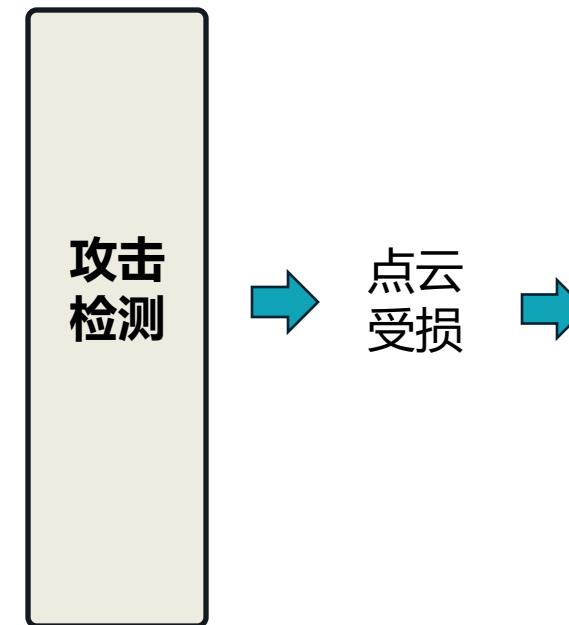
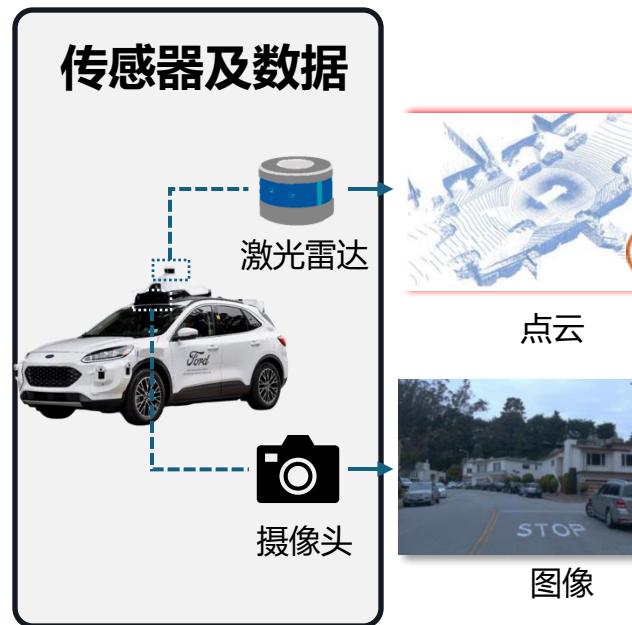
缺点

可能  
“因噎废食”

单点防御

“防不胜防”

# 成功检测出攻击了！ 然后呢？



潜在方法

舍弃受损数据

缺点

可能  
“因噎废食”

单点防御

“防不胜防”

能否在**不影响性能**的情况下提升对所有信号注入攻击的鲁棒性？

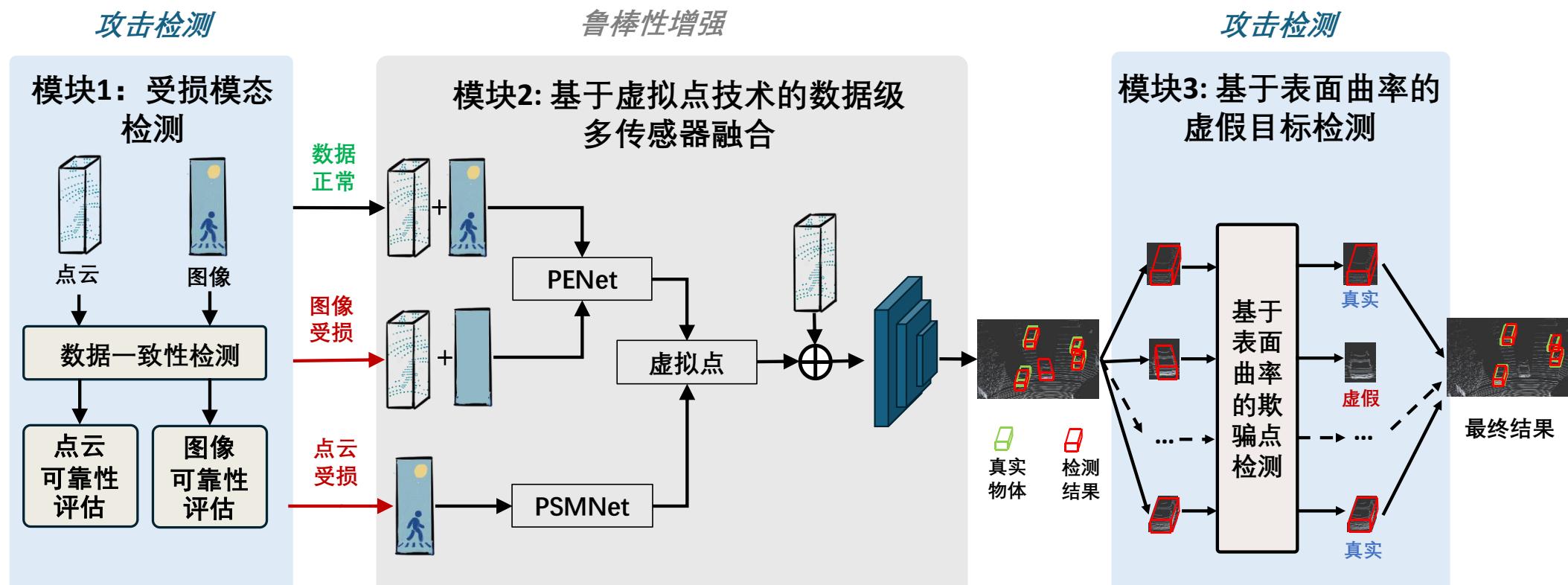
# 成果3的启发

具有如下3个特点的多传感器融合模型可能表现出更强的鲁棒性。

- **数据融合**: 在原始数据层面进行融合, 而不是在特征或结果层面。
- **平行融合**: 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。
- **模态独立**: 每个模态都具有能够独立于其他模态完成3D 目标检测任务的能力

# 鲁棒性提升：基于虚拟点技术的数据级多传感器融合

**工作简介：**在攻击检测的基础上，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。



# 鲁棒性提升：基于虚拟点技术的数据级多传感器融合

- **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
- **PSMNet**: 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

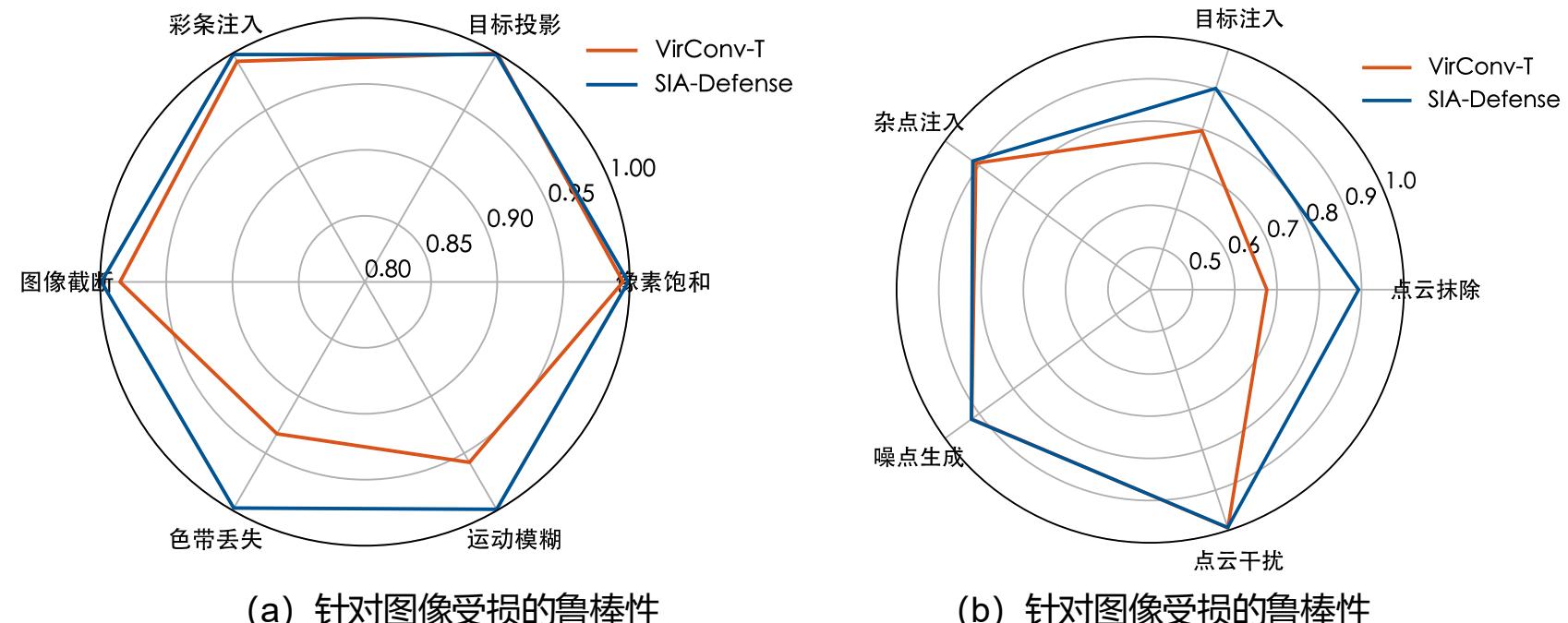


图：PSMNet能在点云严重受损情况下生成和真实环境相近的虚拟点

# 鲁棒性提升：基于虚拟点技术的数据级多传感器融合

## □ 鲁棒性评估

- **数据集：**SIA-KITTI (*Designed by 成果3*)
- **结果：**正常性能与SOTA模型相同，平均鲁棒性比SOTA模型提升了5%以上。
- **指标：**平均鲁棒性mRb



图：SIA-Defense与SOTA模型鲁棒性对比

# 小结

- **攻击检测方法：**攻击检测层面，本章提出了（1）基于点云表面曲率的**虚假目标检测**；2) 基于一致性分析和时序预测的**受损模态检测**。
- **鲁棒性增强：**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，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。

- 1 背景与意义  
Background and significance
- 2 方法与思路  
Challenges and contributions
- 3 成果与结论  
Results and conclusions
- 4 总结与展望  
Summary and prospect



# 工作总结

成果概览

- 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
  
- **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，**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，**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融合鲁棒性测评基准，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，提出了基于攻击检测和多传感器融合的防护框架SIA-Defense，实现了对信号注入攻击的**有效检测和防护**。

学术创新

# 研究展望

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：

- **新型激光雷达安全问题研究**：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，新的激光雷达设计可能引入新的攻击面和防护挑战。
- **自动驾驶整车硬件在环的安全问题研究（AI Security）**：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，且自动驾驶算法从模块化到端到端到VLA不断演进。未来可与车企合作构建硬件在环（HIL）测试平台，在物理世界评估信号注入攻击对自动驾驶AI算法的影响。
- **自动驾驶安全多模态智能体技术研究（AI for Security）**：大模型近年来展现出卓越的上下文理解、多模态逻辑推理与答案生成能力。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。实现（1）脆弱性自动挖掘，（2）跨模态关联推理，（3）自演进防御策略生成。



# 请各位专家批评指正

浙江大学博士论文答辩

答辩人: **金子植**

指导老师: **冀晓宇 教授 徐文渊 教授**

答辩时间: 2025年5月27日

# 论文与专利

## □ 参与发表论文9篇，第一作者发表CCF A或SCI论文4篇，在投1篇

- 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】

## □ 申请发明专利5篇

- 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 xxxxxxxx，国防专利

浙江大学  
博士学位论文答辩