General Black-box Adversarial Sensor Attack and Countermeasures

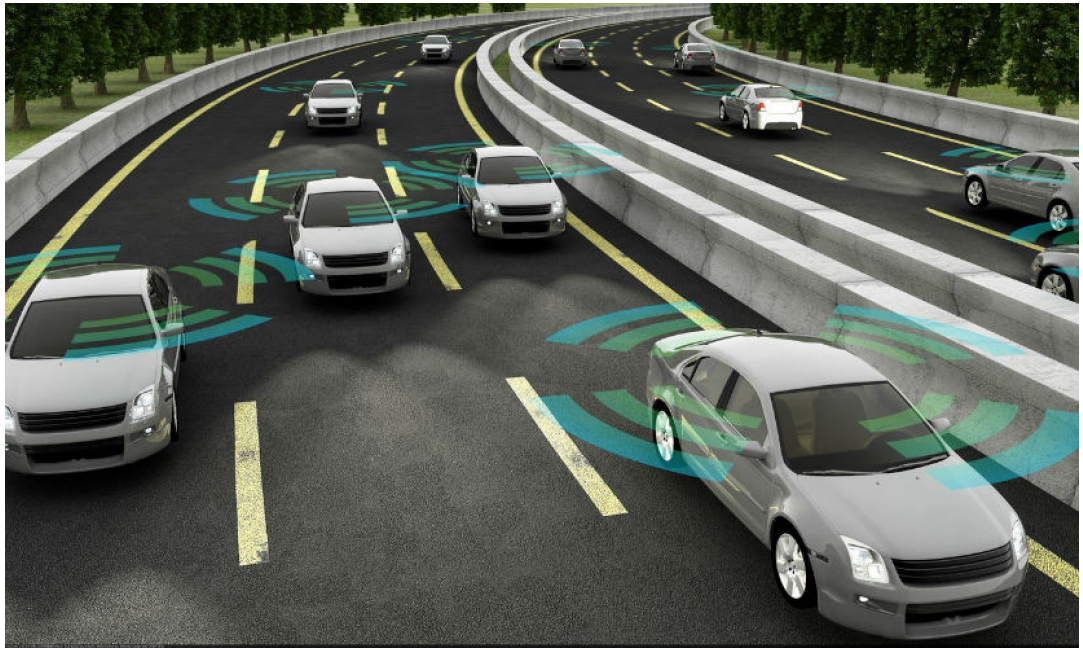
Robust LiDAR-based Perception in Autonomous Driving

(Charles) Chengzeng You
Connected and autonomous vehicles
Department of computing
Email: chengzeng.you19@imperial.ac.uk

Imperial College London

1. Introduction

Autonomous vehicles rely on perception, which leverages sensors like cameras and LiDARs (Light Detection and Ranging) to understand the surrounding driving environment.



https://uk.images.search.yahoo.com/yhs/

Imperial College London

1. Introduction

1.1 Sensor-level attacks

- camera blinding [42], physical-world camera attacks [28, 29], trojan attacks on the neural networks for AV camera input [37].
- A few works demonstrated the feasibility of injecting spoofed points into the sensor input from the LiDAR[42, 44].
- To improve the efficiency of generating adversarial examples, a recent study [17] has performed the attack on the Apollo LiDAR-based detector.

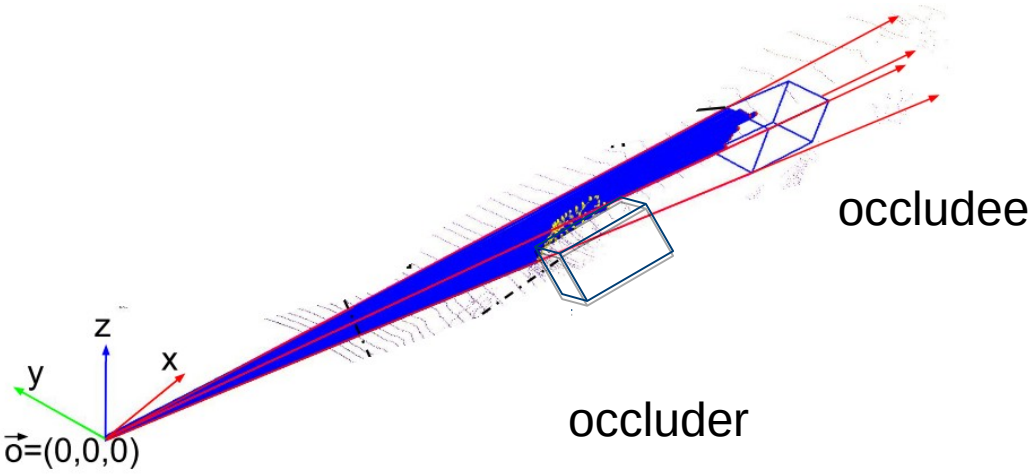
The first study to systematically explore, discover, and defend against a general vulnerability existing in LiDAR-based 3D object detector.

Imperial College London

1. Introduction

1.2 Hypotheses

- state-of-the-art 3D object detection model designs generally ignore the occlusion patterns in LiDAR point clouds.



Imperial College London

1. Introduction

1.3 Contributions

- The first study to explore the general vulnerability of current LiDAR-based perception architectures. Attacks based on this general vulnerability achieve around 80% mean success rates on all target models.
- The first defense against LiDAR spoofing attacks based on the general vulnerability. Reduce the mean attack success rate to 5.5%.
- For fusion-based models: propose sequential view fusion (SVF). SVF can further reduce the mean attack success rate to 2.3%.

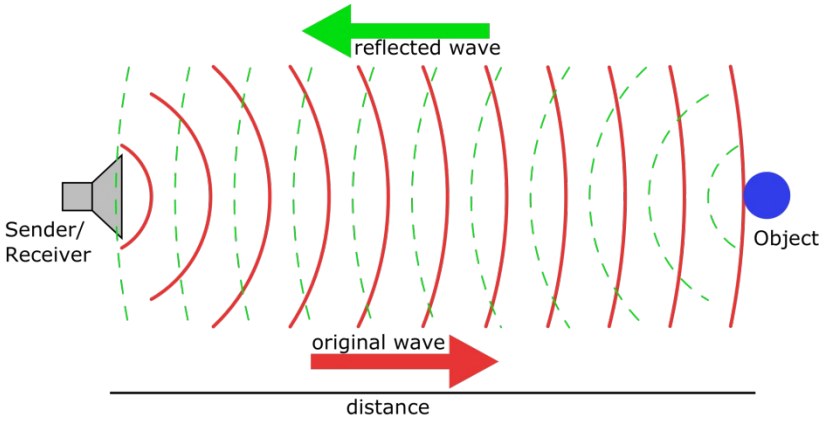
Imperial College London

2. Background

2.1 LiDAR Sensor

Light Detection and Ranging (LiDAR) use laser to measure distance:

1. emitting a laser pulse on a surface
2. catching the reflected laser back to the LiDAR pulse source
3. measuring the time laser travelled
4. Distance = (Speed of light x Time elapsed) / 2



https://news.voyage.auto/

Imperial College London

2. Background

2.2 LiDAR Spoofing Attacks

Sensor spoofing attacks use the same physical channels as the targeted sensor to manipulate the sensor readings.

- place an attacking device at the roadside to shoot malicious laser pulses to AVs passing by.
- drive an attack vehicle equipped with an attacking device that shoots laser pulses to the victim AV's LiDAR.

LiDAR has been shown to be vulnerable to laser spoofing attacks

2. Background

2.3 State-of-the-art detectors

Input data view	Method	State-of-the-art detectors
BEV-based	1. Projecting 3D LiDAR point cloud into BEV 2. 2D detection using convolutional neural network	Apollo [44] Cont Fuse [45] Lasernet [46] PIXOR [47]
Voxel-based	1. Slicing 3D LiDAR point cloud into voxels 2. Feature extraction 3. 2D convolutional detection	Pointpillars [48] Patch Refinement [49] Voxel-FPN [50] Sparse Convolution [51]
3D point cloud-based	1. Region proposal generation 2. Bounding box regression 3. Object classification	STD [52] Fast Point R-CNN [53] PointRCNN [54] Part-A2 net [55]

Imperial College London

3. Threat Model

3.1 Sensor attack capability

Adopt the formulation in Adv-LiDAR [17] to describe the sensor attack capability:

$$\min \quad \mathcal{L}_{\text{adv}}(x \oplus t'; M)$$

$$\text{s.t.} \quad t' \in \{\Phi(T') | T' \in \mathcal{A}\} \;\; \& \;\; x = \Phi(X)$$

x : the corresponding input feature matrix
t' : adversarial spoofed input feature matrix
$\oplus(\cdot)$: merging function
$L_{adv}^{(\cdot; M)}$: the adversarial loss, given the machine learning model

X : the pristine 3D point cloud
T' : adversarial spoofed 3D point cloud
A : 3D point cloud generated from LiDAR spoofing attacks
$\Phi(\cdot)$: the pre-processing function that maps X into x

Imperial College London

3. Threat Model

3.2 Black-box model-level spoofing attack

attackers do not have access to the machine learning model nor the perception system

Attack goal: to spoof a front-near vehicle located 5-8 meters in front of the victim AV.

Imperial College London

3. Threat Model

3.3 Defense against general spoofing attacks

defenders can only strengthen the software-level design, but cannot modify the AV hardware (e.g., sensors) due to cost concerns.

Imperial College London

4. Limitations of Existing Attacks

4.1 Limitations of sensor-level LiDAR spoofing attacks

1. Limitations of sensor-level LiDAR spoofing attacks

blindly spoofing cannot effectively achieve the attack goal other than Apollo 5.0

Imperial College London

4. Limitations of Existing Attacks

4.2 Limitations of Adv-LiDAR

1. White-box attack limitation

very few AV companies publicly release their perception systems, making Adv-LiDAR challenging to launch in the real world

2. Attack generality limitation

Adv-LiDAR only targets Apollo 2.5

Adversarial examples generated by Adv-LiDAR cannot transfer between models

the attack success rate consistently drops with the change of the pristine point cloud

Overall, existing spoofing attacks cannot easily achieve the attack goal on all target models.

Imperial College London

5 A General Design-level Vulnerability

5.1 physical features of LiDAR

Two observations:

C1: Occlusions between objects will make occluded objects partially visible in the LiDAR point cloud.

C2: The density of data decreases with increasing distance

from the LiDAR sensor, due to the working principles of LiDAR sensors

Two hypothesis:

FP1: If an **occluded vehicle** can be detected in the pristine point cloud by the model, its point set will still be detected as a vehicle when directly moved to a front-near location.

FP2: If a **distant vehicle** can be detected in the pristine point cloud by the model, its point set will still be detected as a vehicle when directly moved to a front-near location.

Imperial College London

5 A General Design-level Vulnerability

5.2 Experimental Validation

E1 :

1. randomly pick 100 point cloud samples F that contain 100 target occluded

Vehicles vi. feed F into three target models and record the confidence scores (i.e., outputs of models to represent the confidence of detection) of the occluded vehicles as si

2. global translation matrix

$$V'_{i\,\mathbf{w_i}} = V_{i\mathbf{w_i}}$$

$$\begin{bmatrix} V'_{i\,\mathbf{w_x}} \\ V'_{i\,\mathbf{w_y}} \\ V'_{i\,\mathbf{w_z}} \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 & \tau\cos(\theta+\alpha) \\ \sin\theta & \cos\theta & 0 & \tau\sin(\theta+\alpha) \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} V_{i\mathbf{w_x}} \\ V_{i\mathbf{w_y}} \\ V_{i\mathbf{w_z}} \\ 1 \end{bmatrix}$$

Imperial College London

5 A General Design-level Vulnerability

5.2 Experimental Validation

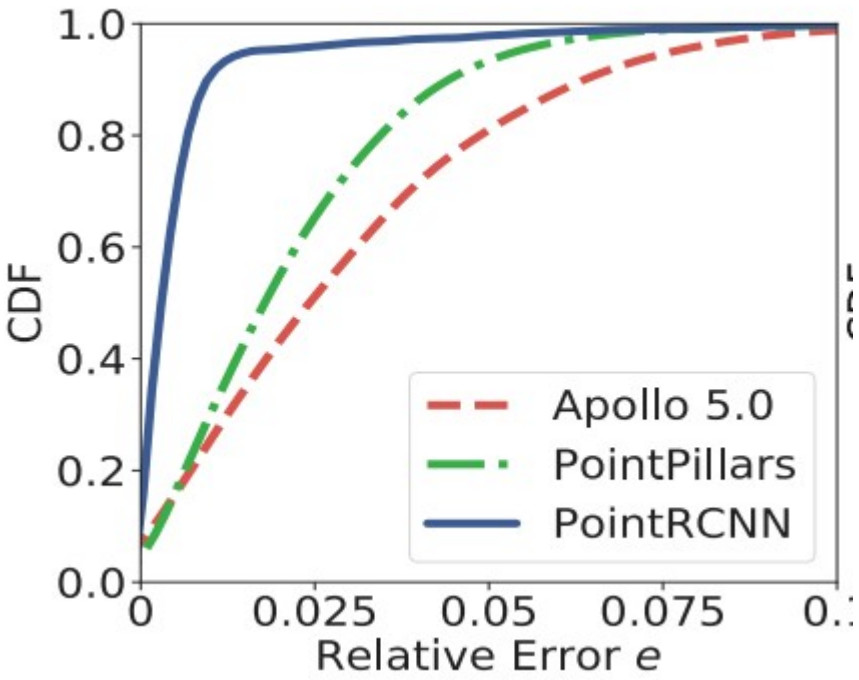
E1 : cumulative distribution function of e

$$F_X(x) = P(X \leq x)$$

$F_X(x)$ = function of X

X = real value variable

P = probability that X will have a value less than or equal to x



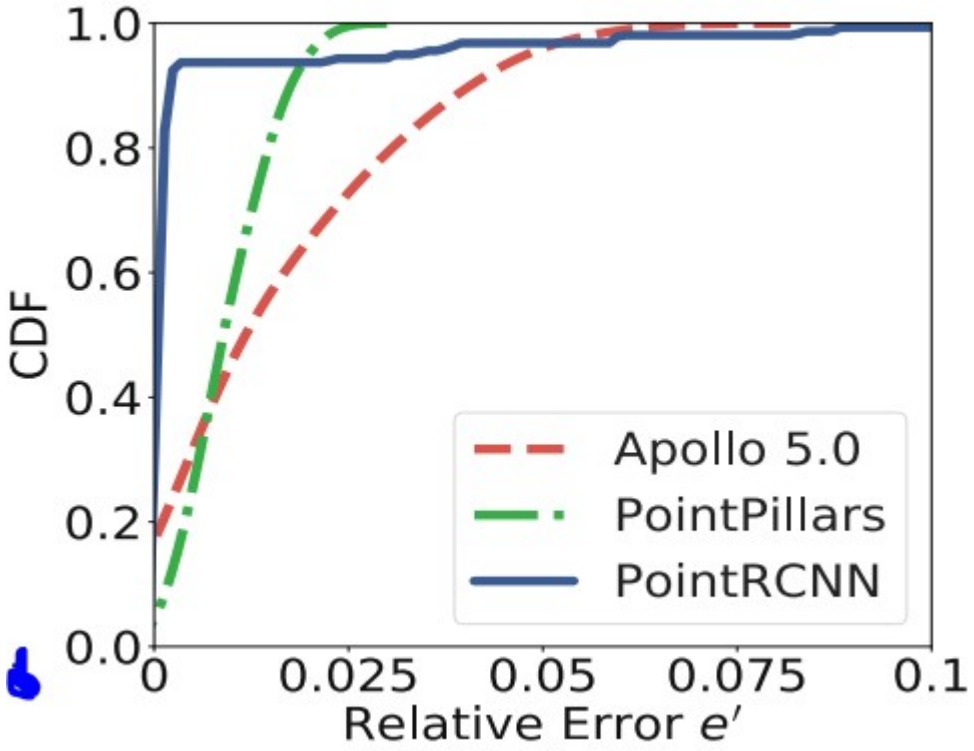
$$e = \frac{|s_i' - s_i|}{s_i}$$

Imperial College London

5 A General Design-level Vulnerability

5.2 Experimental Validation

E2 : cumulative distribution function of e



$$e = \frac{|s_i' - s_i|}{s_i}.$$

Imperial College London

5 A General Design-level Vulnerability

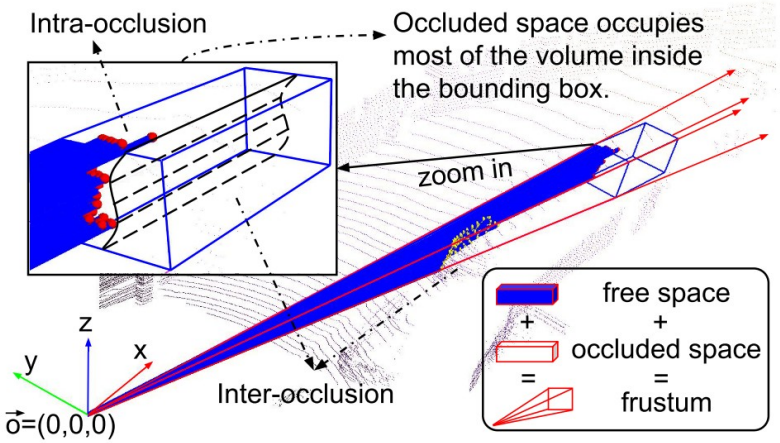
5.3 Vulnerability Identification

1. Inter-occlusion

inter- occlusion describes a causal relationship between occludee and the corresponding occluders (i.e., the occluders cause the occludee partially visible). FP1 violates the physical law of inter-occlusion

2. Intra-occlusion

The facing surface of a solid object occludes itself in the point cloud



Imperial College London

6 Black-box Spoofing Attack

6.1 Constructing original attack traces

collecting existing real-world traces

1. extract occluded vehicles' point sets with varying numbers of points (5-200 points) from the KITTI validation set.

2. theytake 10 points as interval, and divide the extracted point sets into 20 groups: 0-10, 10 -20…

3. randomly pick five traces in each group forming a small dataset K containing 100 point sets.

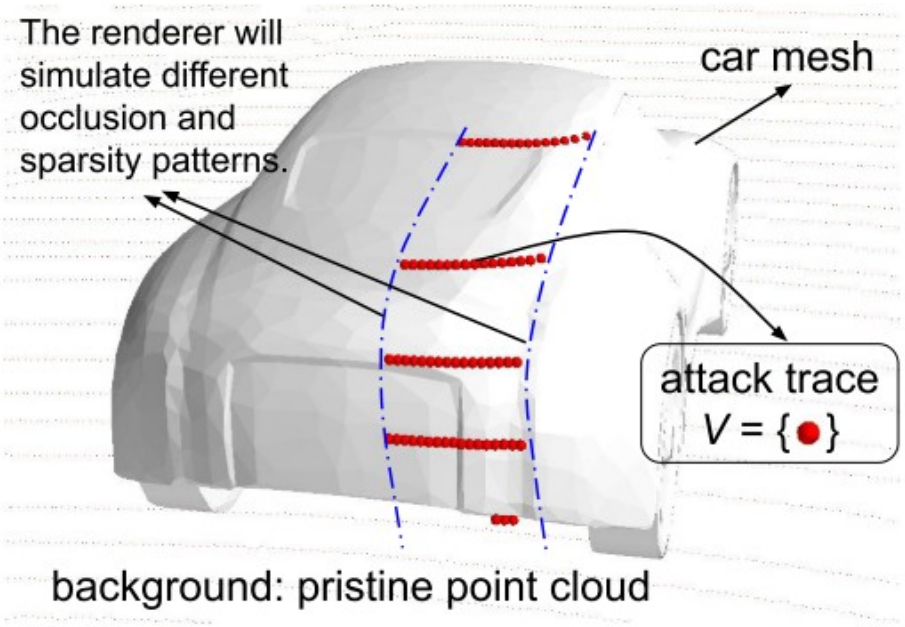
Imperial College London

6 Black-box Spoofing Attack

6.1 Constructing original attack traces

generating customized attack traces

ray-casting techniques: 100 rendered point sets



Imperial College London

6 Black-box Spoofing Attack

6.2 Spoofing original attack traces at target locations

digital spoofing

follow the high-level formulation in Adv-LiDAR [17] utilizing a global transformation matrix $H(\theta,\tau)$ to translate the attack traces

physical spoofing

program attack traces from R as input to the function generator so that theycan control the spoofed points and launch the spoofing attack [55] in the lab

Imperial College London

6 Black-box Spoofing Attack

6.3 Attack Evaluation and Analysis

Evaluation metrics

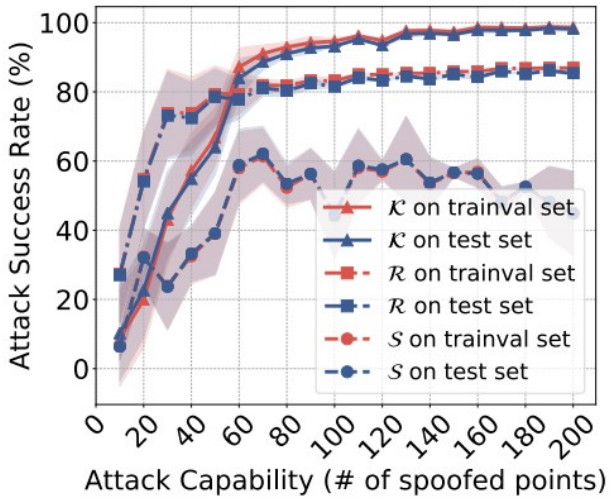
$$ASR = \frac{\#\text{ of successful attacks}}{\#\text{ of total point cloud samples}}$$

Imperial College London

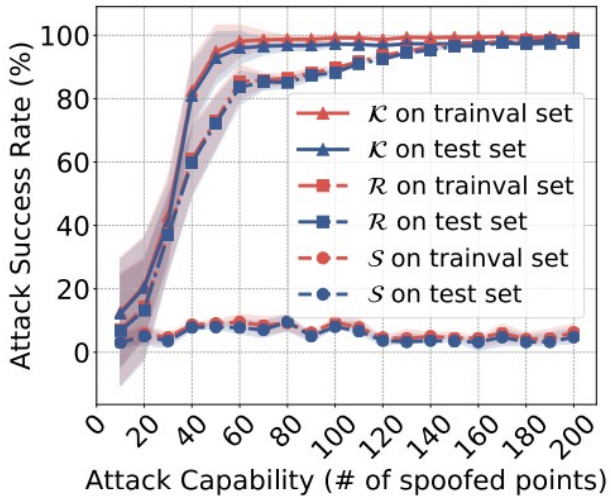
6 Black-box Spoofing Attack

6.3 Attack Evaluation and Analysis

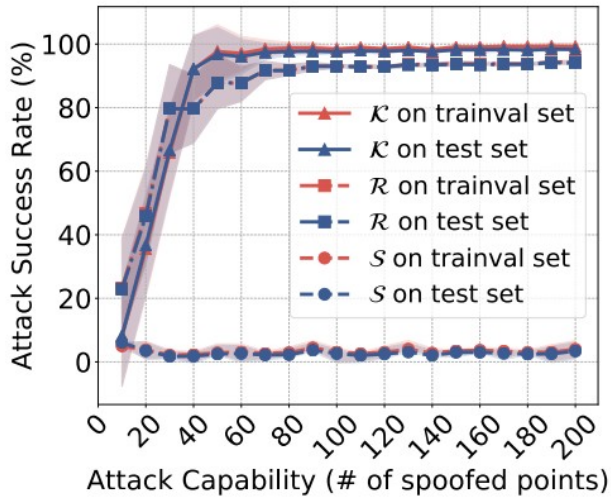
Attack Effectiveness



(a) *ASR* of Apollo 5.0.



(b) *ASR* of PointPillars.



(c) *ASR* of PointRCNN.

S: collected from blindly physical spoofing attacks

Imperial College London

6 Black-box Spoofing Attack

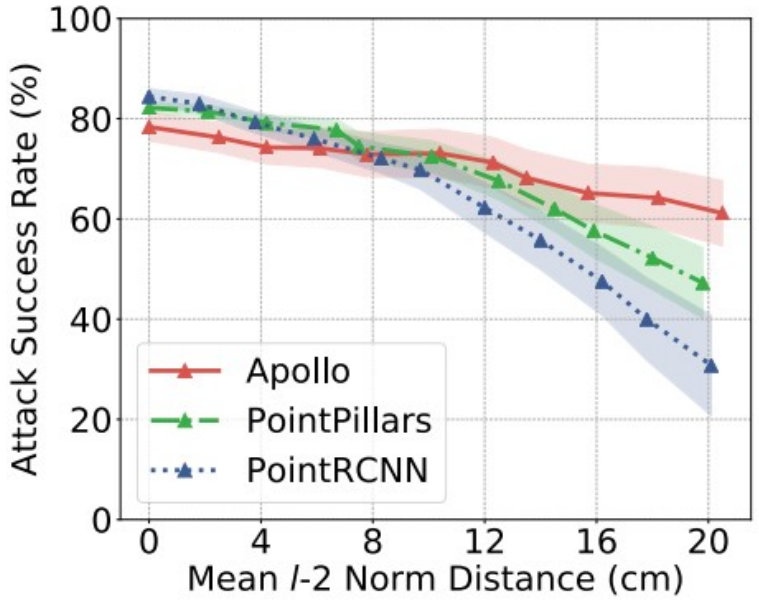
6.3 Attack Evaluation and Analysis

Robustness Analysis - Robustness to variations in attack traces

use attack traces with (60,70] points from R for the robustness analysis

$$V''_{w_i} = V'_{w_i}$$

$$\begin{bmatrix} V''_{w_x} \\ V''_{w_y} \\ V''_{w_z} \end{bmatrix} = \begin{bmatrix} s & 0 & 0 \\ 0 & s & 0 \\ 0 & 0 & s \end{bmatrix} \cdot \begin{bmatrix} V'_{w_x} \\ V'_{w_y} \\ V'_{w_z} \end{bmatrix}$$

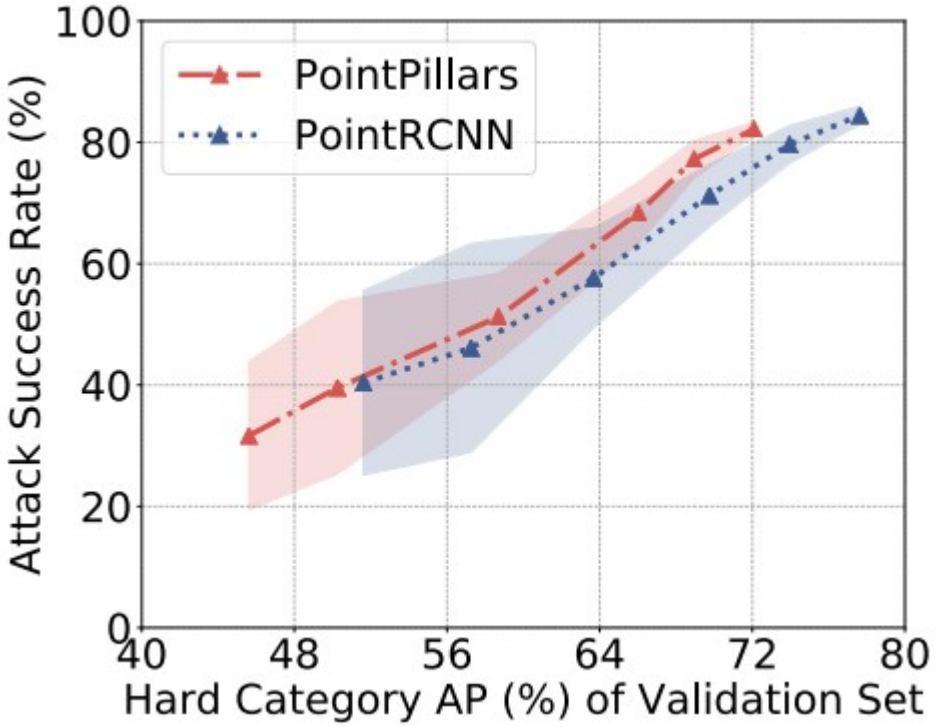


Imperial College London

6 Black-box Spoofing Attack

6.3 Attack Evaluation and Analysis

Robustness Analysis - Robustness to variations in model performance



Imperial College London

6 Black-box Spoofing Attack

6.3 Attack Evaluation and Analysis

Robustness Analysis - Robustness to adversarial training

theyfurther train PointPillars and PointRCNN on modified dataset and evaluate their proposed attack using the same 60-point attack traces with §6.1.1 attack effectiveness on them.

ASRs drop from 83.6% to 70.1% and 88.3% to 79.7% on PointPillars and PointR- CNN.

Imperial College London

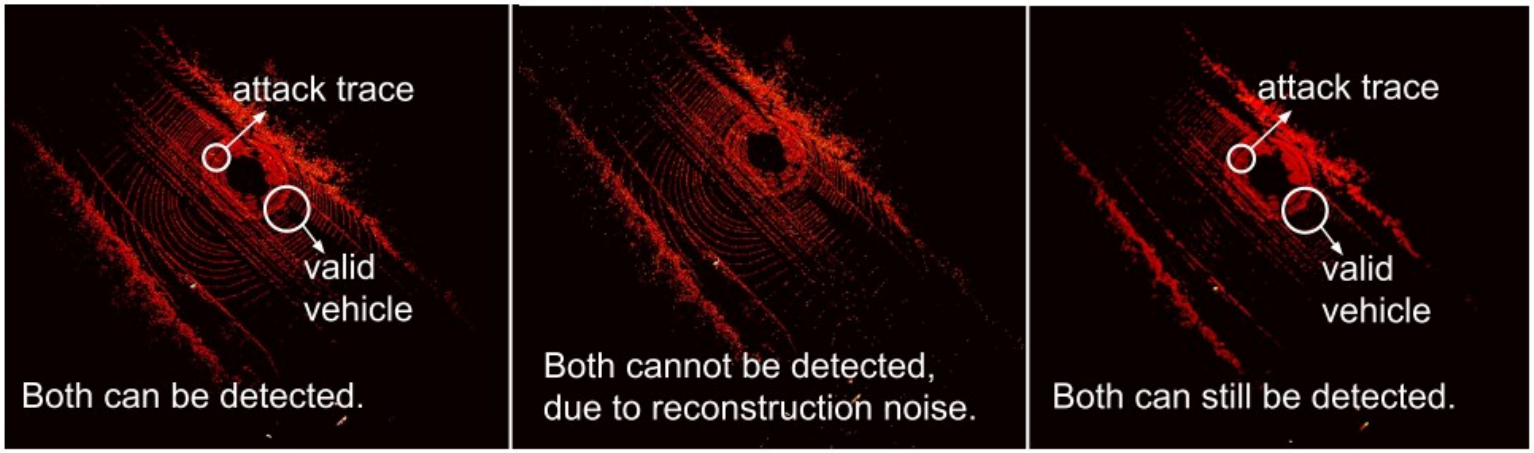
6 Black-box Spoofing Attack

6.3 Attack Evaluation and Analysis

Robustness Analysis - Robustness to randomization-based defenses

leverage state-of-the-art image-based defenses: feature squeezing [66] and ME-Net [70] to test the attack robustness.

none of them can defend the black-box spoofing attack without hurting the original AP



Imperial College London

7 Limitations

1.**same pattern** at the **target location**, whether this defense strategy is transferable to other LiDAR spoofing attacks (e.g.other patterns) still remains unexplored.

2.the identified vulnerability does not provide completeness, there may exist other potential vulnerabilities hidden in the AD systems to be discovered and exploited

Imperial College London

Thank you