Invisible for both Camera and LiDAR: Security of Multi-Sensor Fusion based Perception in Autonomous Driving Under Physical-World Attacks

Yulong Cao[2], Ningfei Wang[1], Chaowei Xiao[3,4], Dawei Yang[2], Jin Fang[7], Ruigang Yang[6], Qi Alfred Chen[1], Mingyan Liu[2], Bo Li[5]

[1]UC Irvine, [2]University of Michigan, [3]NVIDIA Research, [4]Arizona State University, [5]UIUC, [6]Inceptio, [7]Baidu Research

IEEE S&P 2021

Presenter: Yubin Lee

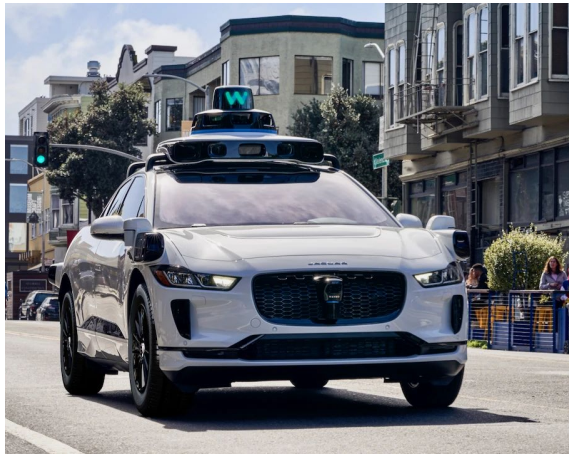
2024.10.16 (Wed)

KAIST School of Computing

CASYS | KAIST Computer Architecture & System Lab

Autonomous Driving

- **Increasingly deployed on public roads**



- **Perception: fundamental to Autonomous Driving (AD) systems**
  - Most safety-critical task: in-road obstacle detection
  - How?

Perception Sensors

- **Camera**
  - Current systems don't have high resolution (<5MP)

- **LiDAR**
  - Send waves of light, measure reflected light, detect distance to objects
  - Radio waves = RADAR, Sound waves = SONAR, Light waves = LiDAR

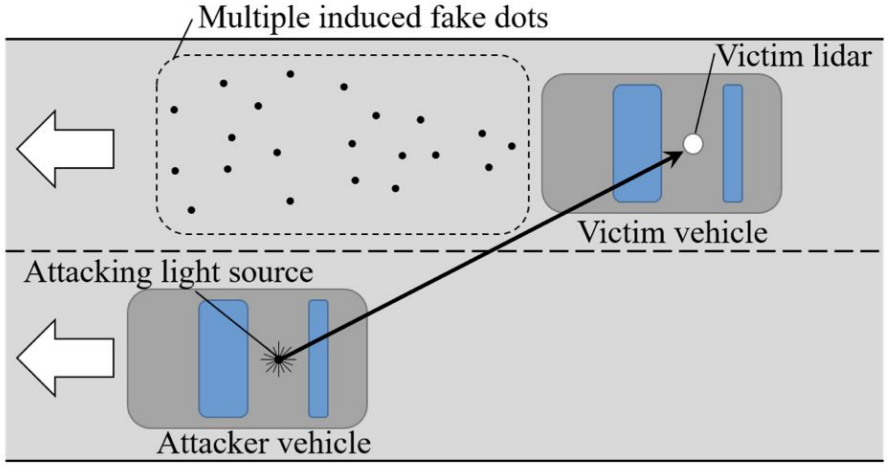
- **Real systems: use a combination**
  - Dubbed "Multi-Sensor Fusion" (MSF)

Multi-Sensor Fusion

- **Prior work: Limited to attacks on a single source of AD perception**

Poster and Sticker Attack

[1]



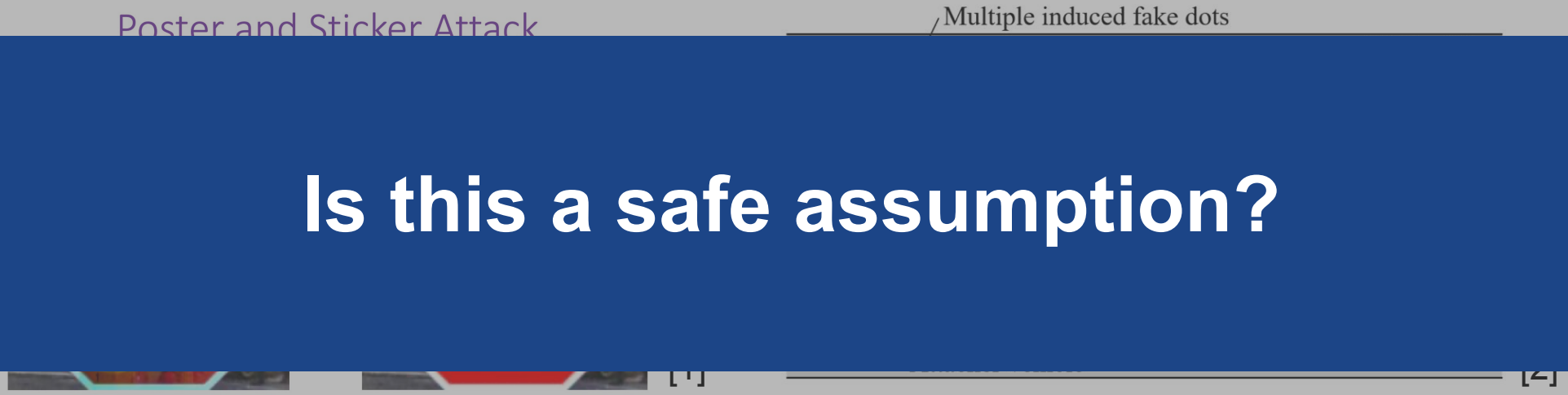
[2]

- **Real systems: Multi-Sensor Fusion (MSF) based design**
  - Assumption: not all perception sources can be attacked simultaneously
    ⇒ There always exists a possible MSF algorithm that can **rely on unattacked sources**

[1] K. Eykholt et al., "Physical Adversarial Examples for Object Detectors", WOOT 2018
[2] H. Shin et al., "Illusion and Dazzle: Adversarial Optical Channel Exploits against Lidars for Automotive Applications", CHES '17

Multi-Sensor Fusion

- **Prior work: Limited to attacks on a single source of AD perception**



Is this a safe assumption?

- **Real systems: Multi-Sensor Fusion (MSF) based design**
  - Assumption: not all perception sources can be attacked simultaneously
    ⇒ There always exists a possible MSF algorithm that can **rely on unattacked sources**

[1] K. Eykholt et al., "Physical Adversarial Examples for Object Detectors", WOOT 2018
[2] H. Shin et al., "Illusion and Dazzle: Adversarial Optical Channel Exploits against Lidars for Automotive Applications", CHES '17

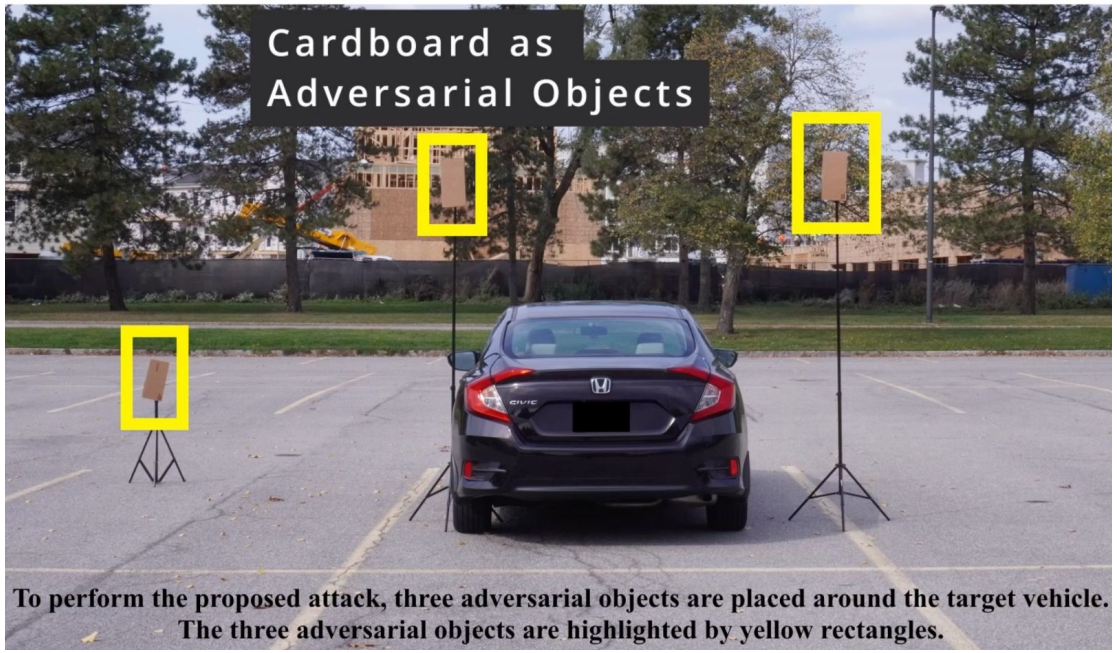
Problem Formulation & Attack Goal

- **Optimization problem: Target physical-world attack vectors**
  - Ideally, single attack vector effective for both camera and LiDAR
- **Goal: Cause MSF-based AD perception to fail in detecting a front obstacle**
- **General category: adversarial attacks**

Adversarial Attack

- **Different shapes of a 3D object can change…**
  - LiDAR point cloud point position
  - Camera image pixel value
- **Place benign or malicious objects**
  - Trick the LiDAR and/or camera
- **Realizable thanks to 3D printing!**

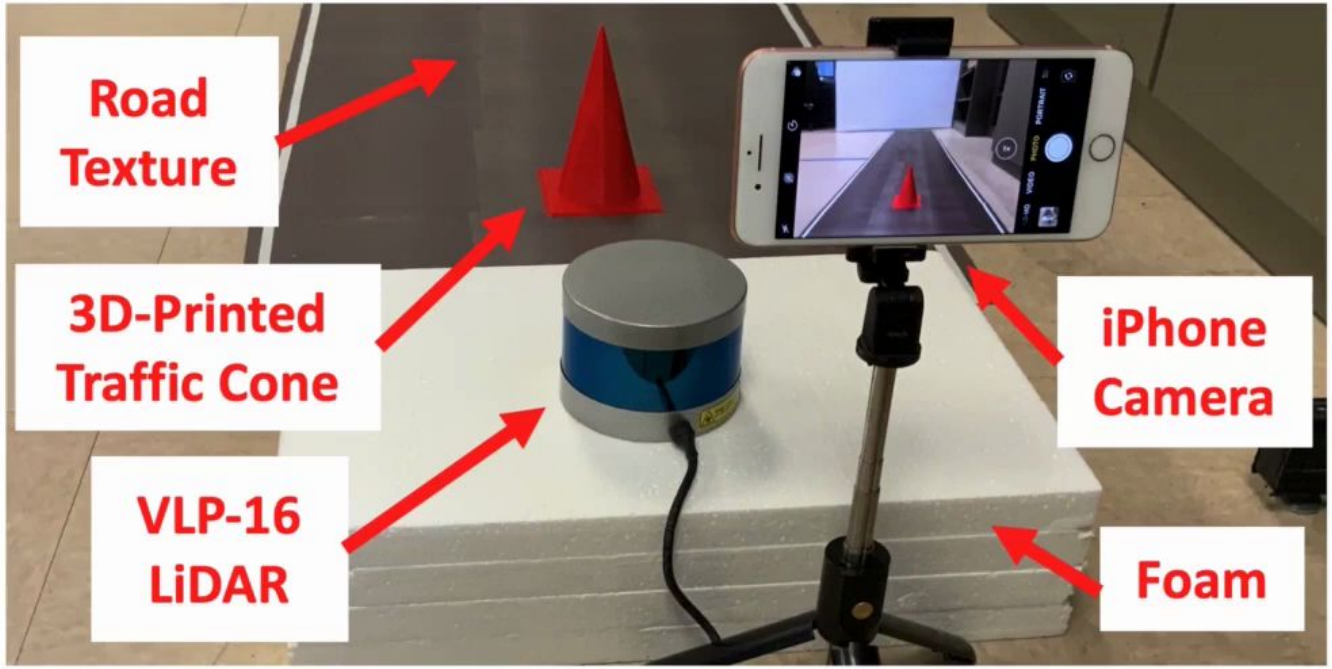
Vehicle Hiding [3]



To perform the proposed attack, three adversarial objects are placed around the target vehicle. The three adversarial objects are highlighted by yellow rectangles.

[3] Zhu et al., "Adversarial Attacks against LiDAR Semantic Segmentation in Autonomous Driving", SenSys '21

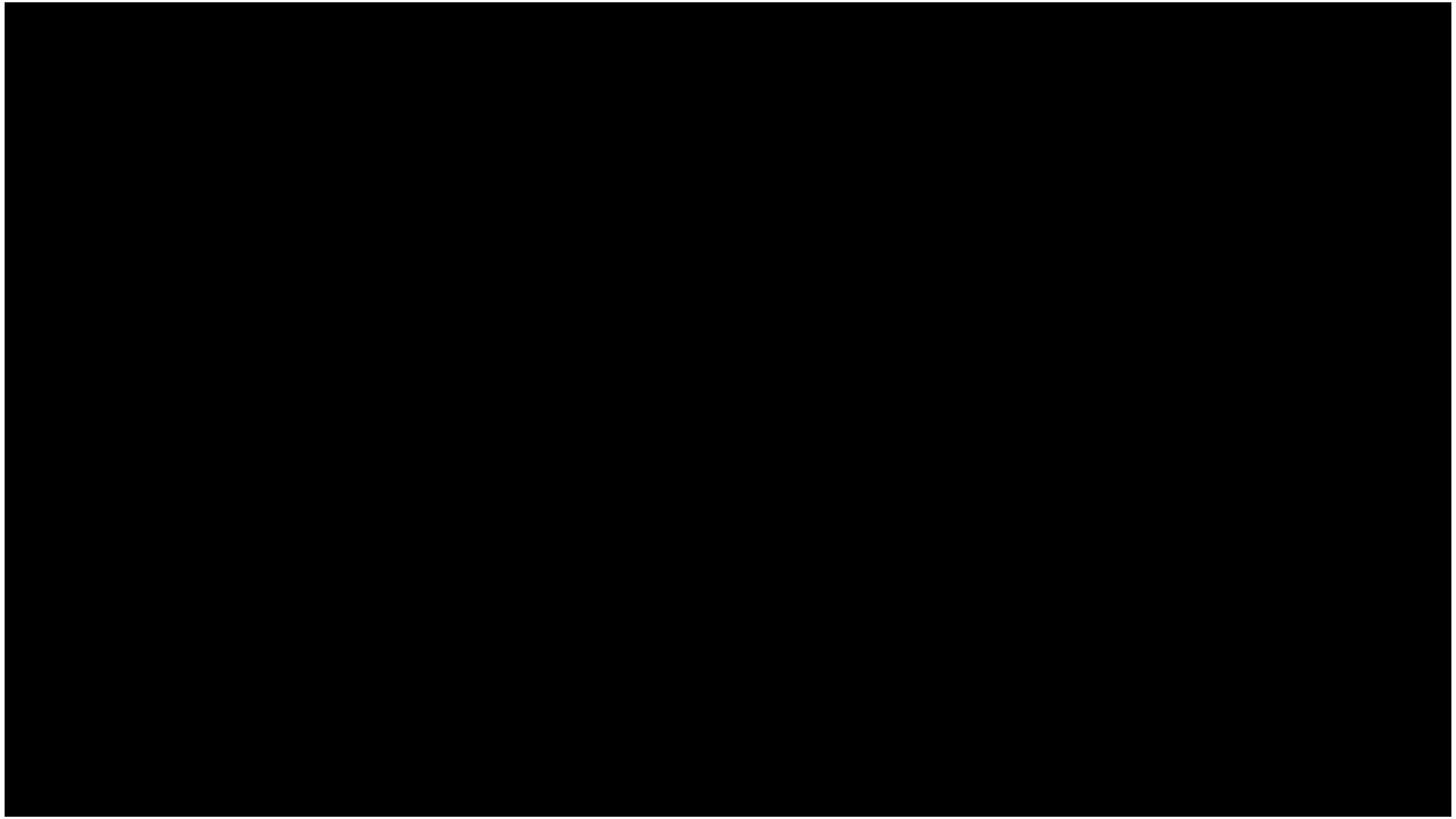
Demo (Miniature Scale)



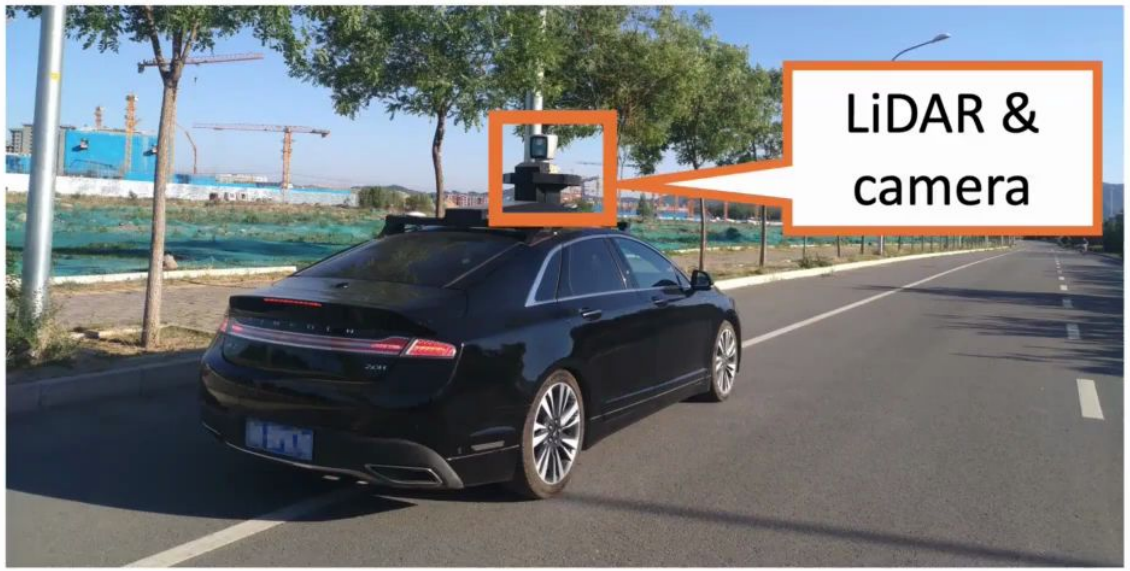
Benign

Adversarial

Demo (Miniature Scale)



Demo (Real-World)

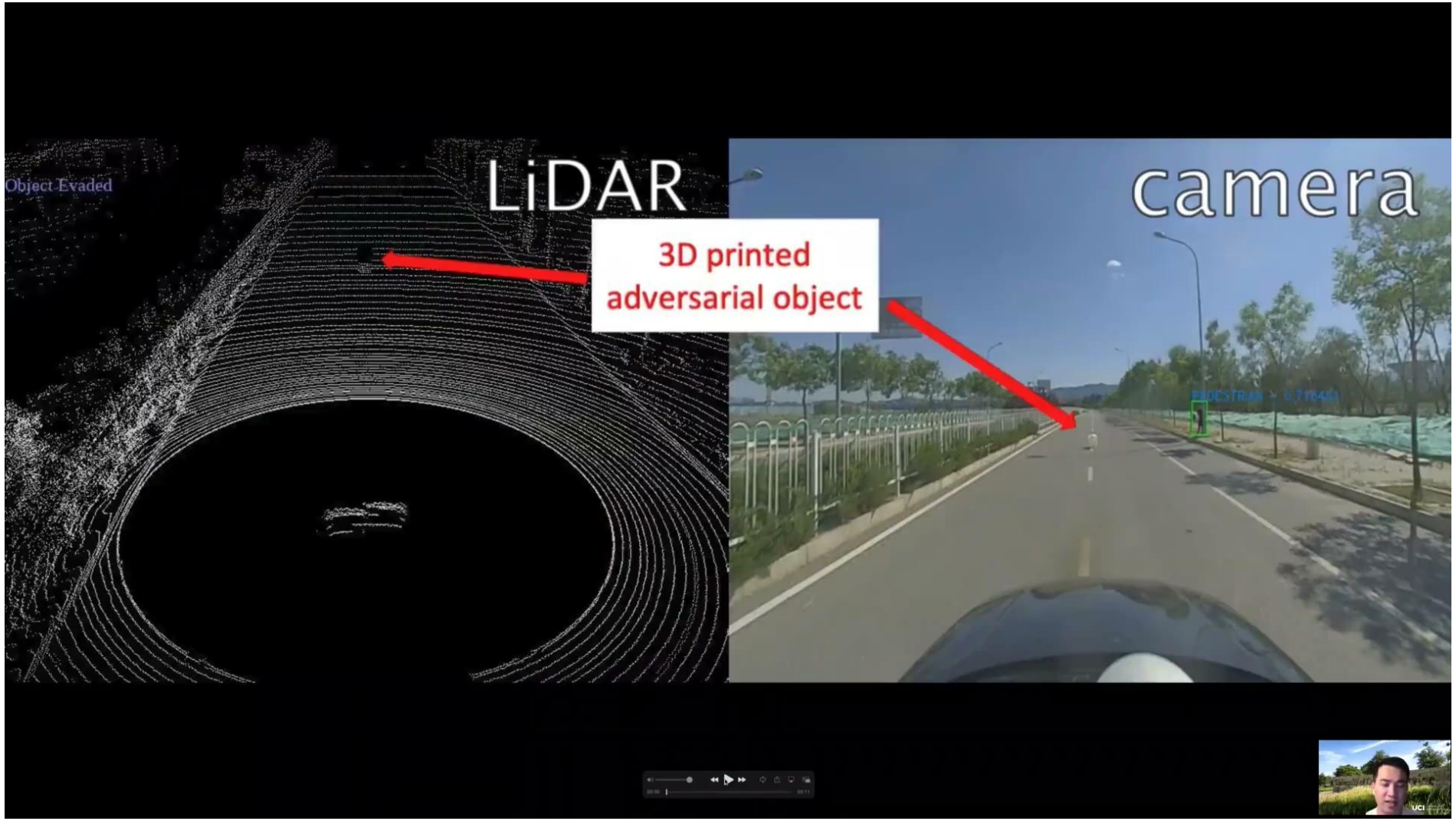


Road & car with LiDAR & camera

Benign

Adversarial

Demo (Real-World)



Demo (End-to-End Attack Simulation)

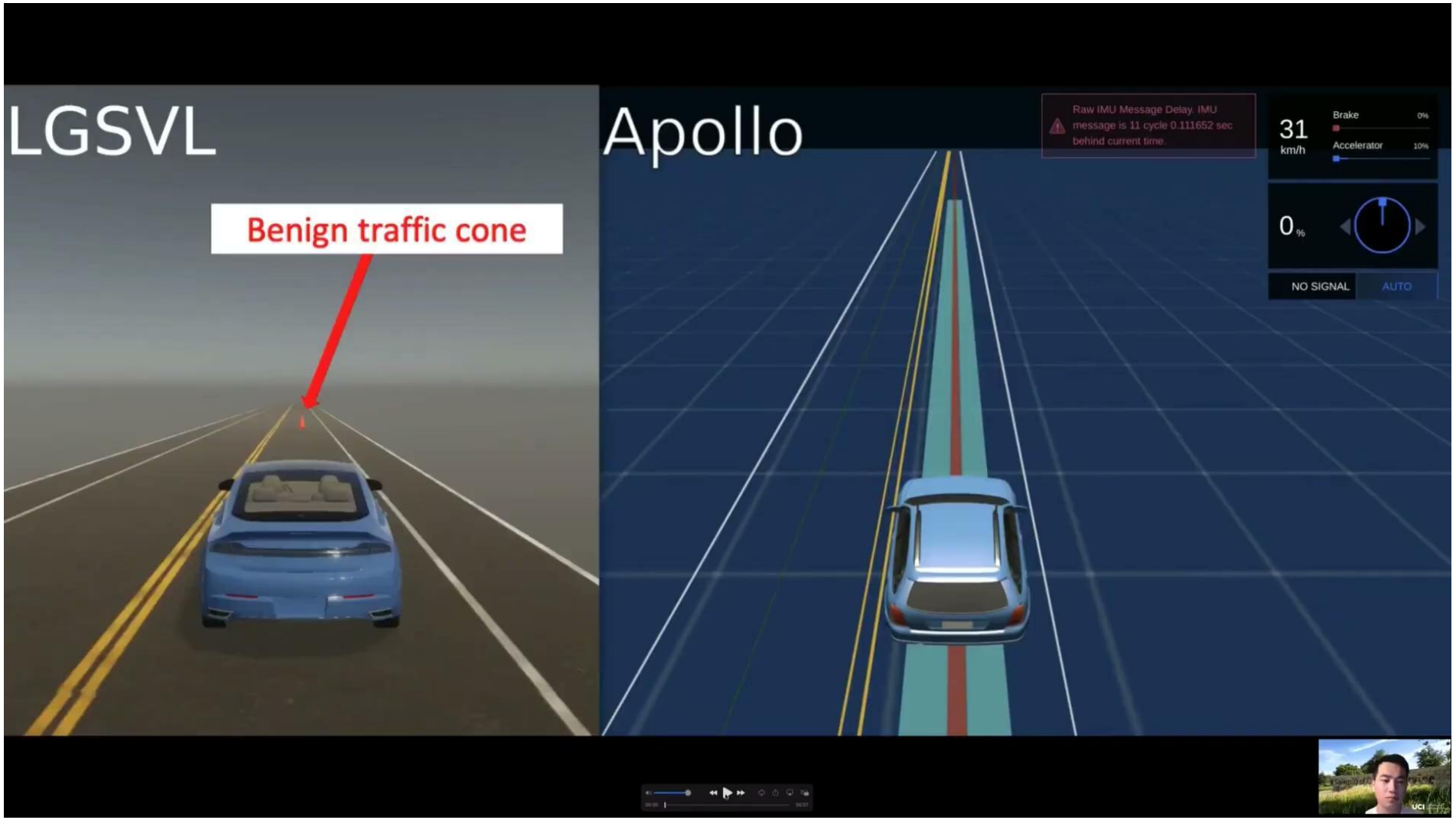
Road in LGSVL AD simulator

Vehicle running Apollo v5.0

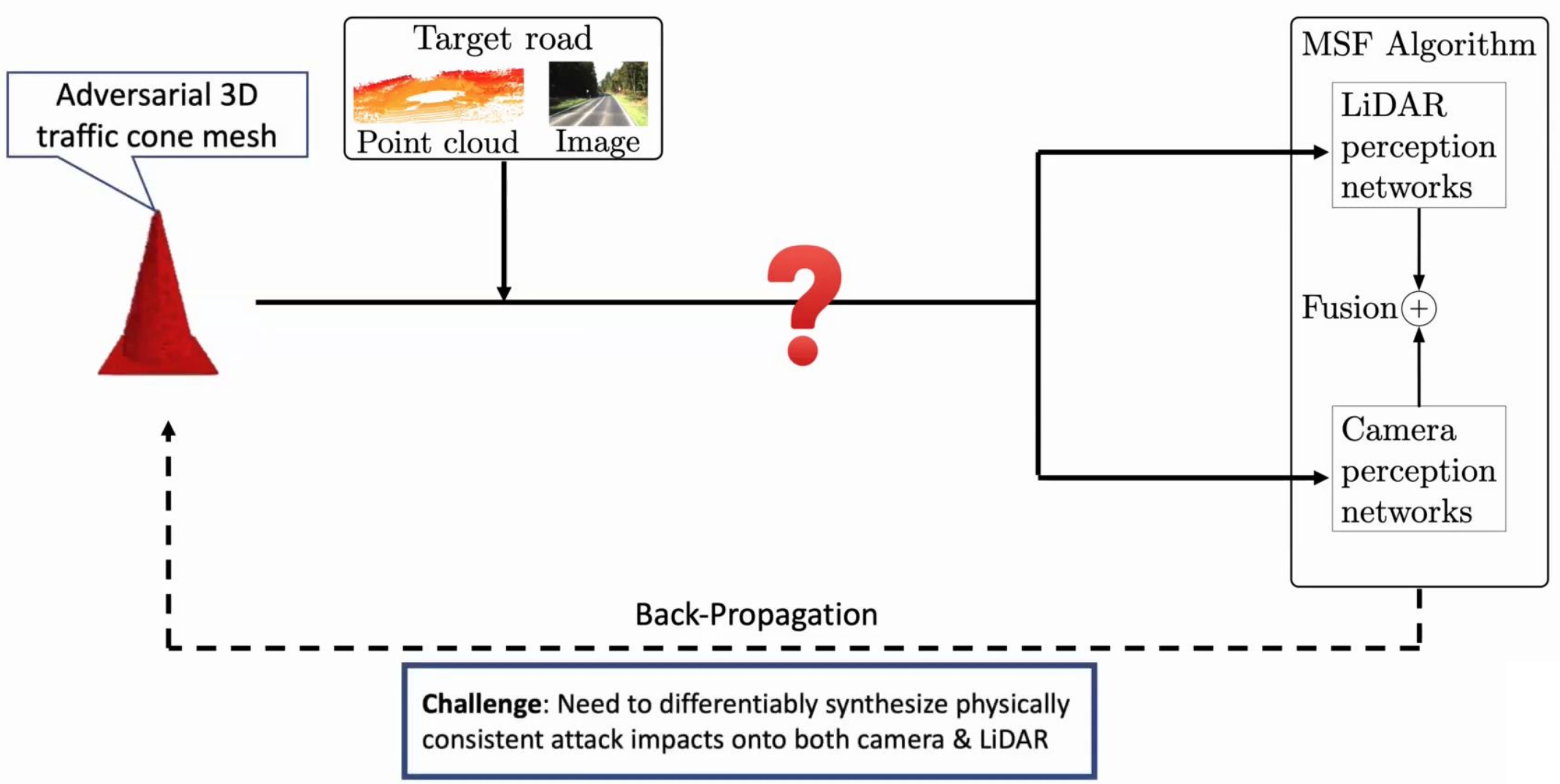
Benign

Adversarial

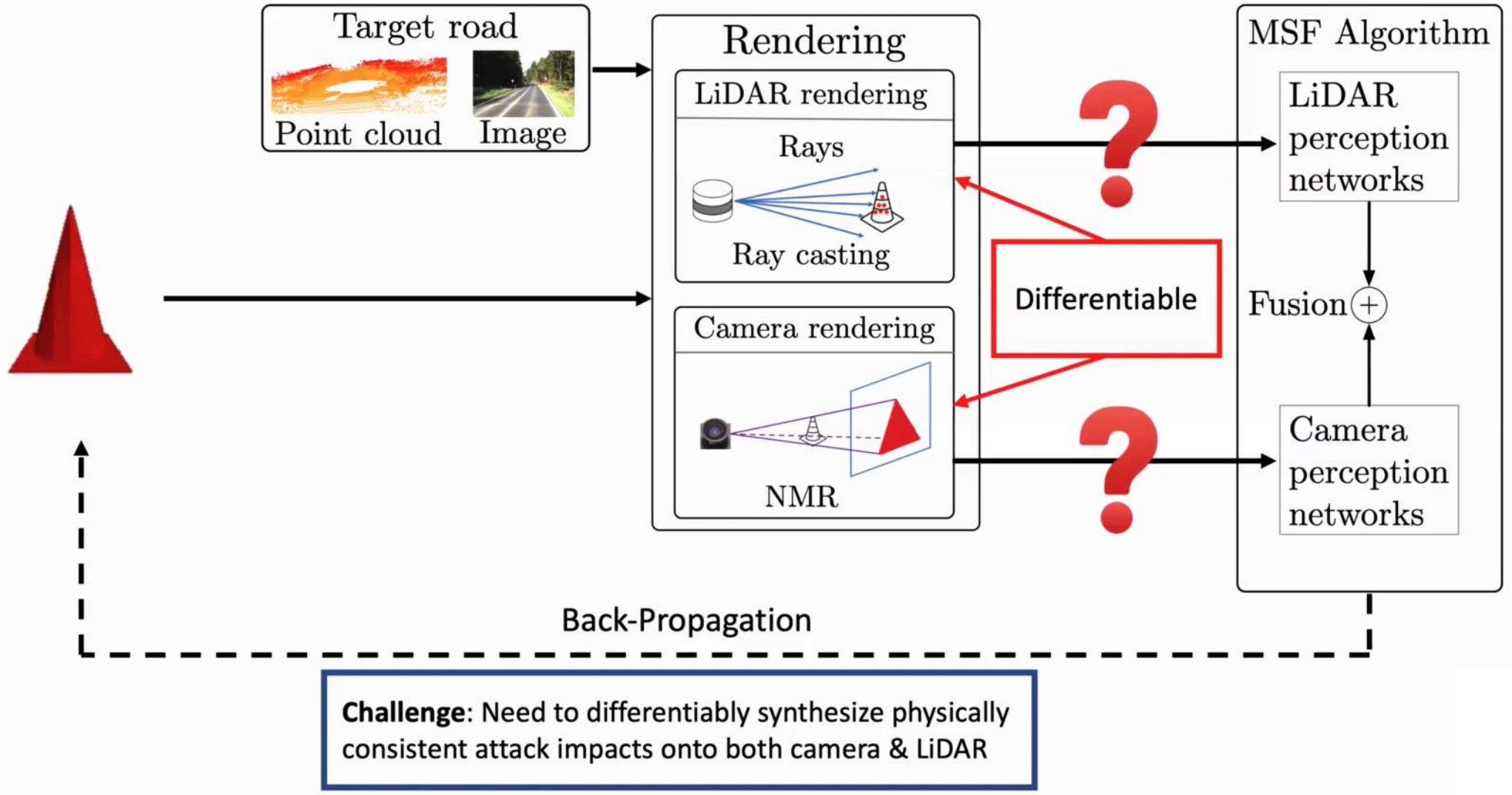
Demo (End-to-End Attack Simulation)



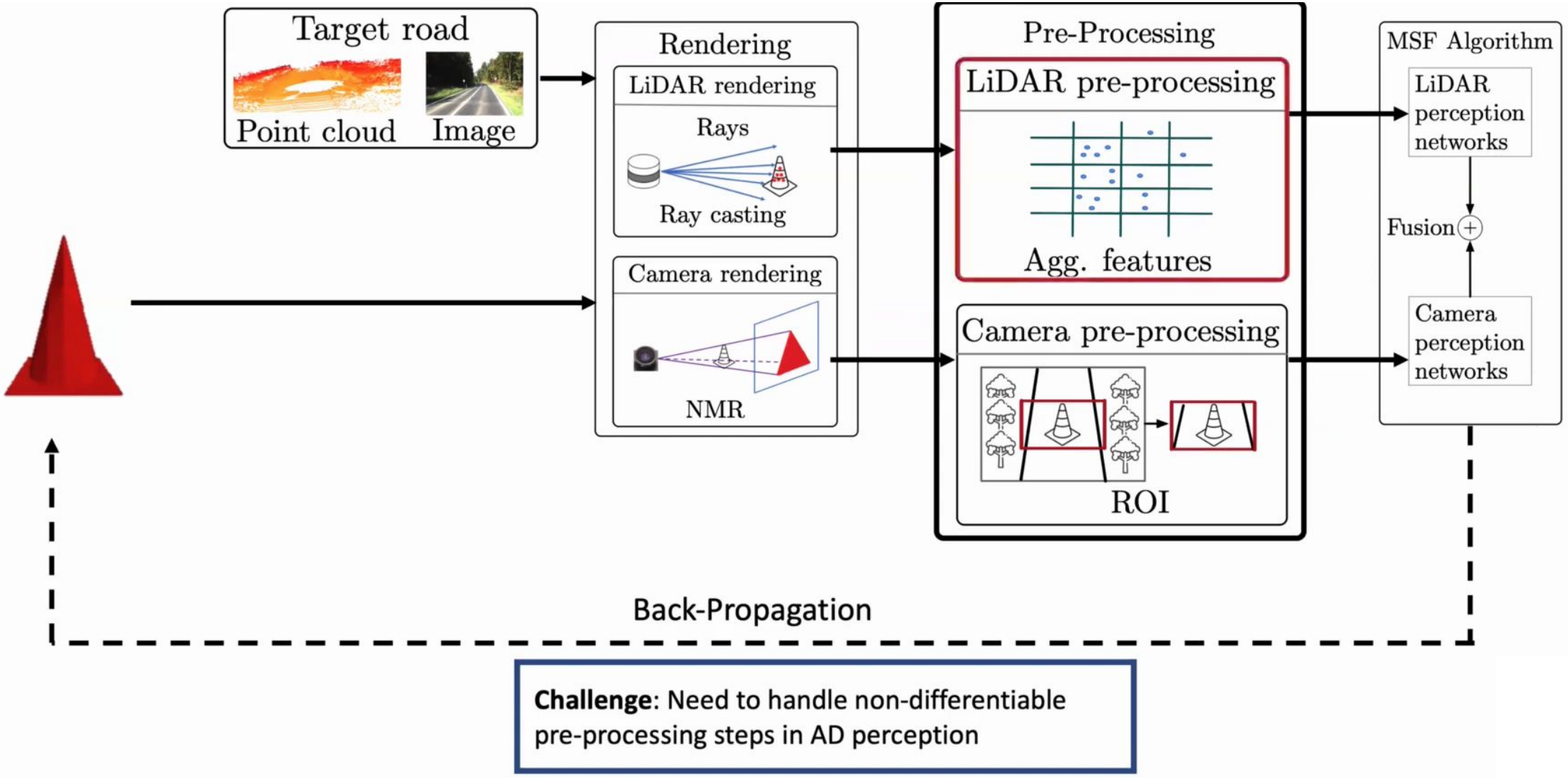
Attack Design: MSF-ADV (Optimization Framework)



Attack Design: MSF-ADV (Differentiable 3D Rendering)



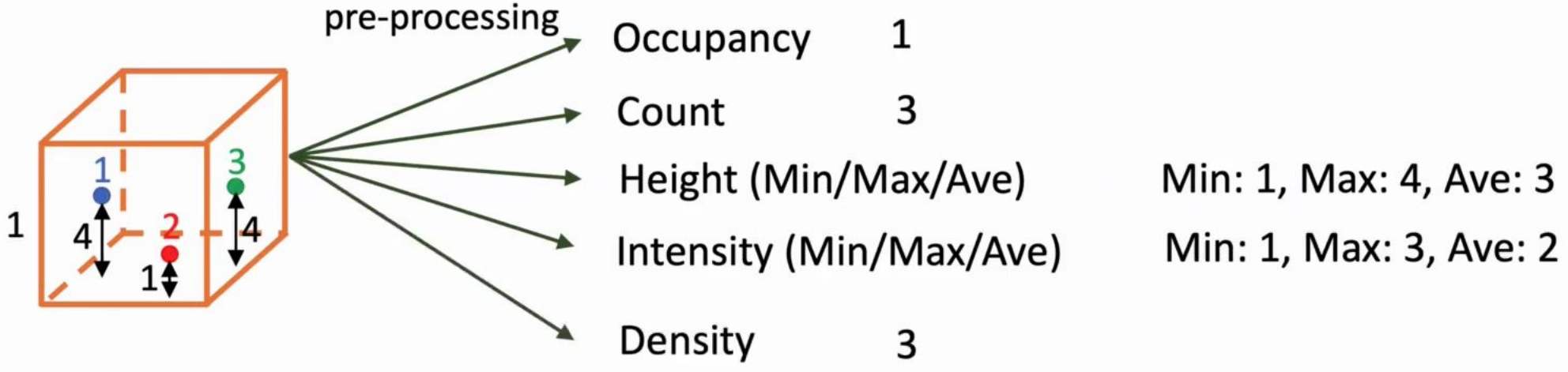
Attack Design: MSF-ADV (Differentiable Pre-processing)



Challenge: Need to handle non-differentiable pre-processing steps in AD perception

Attack Design: MSF-ADV (LiDAR Pre-processing)

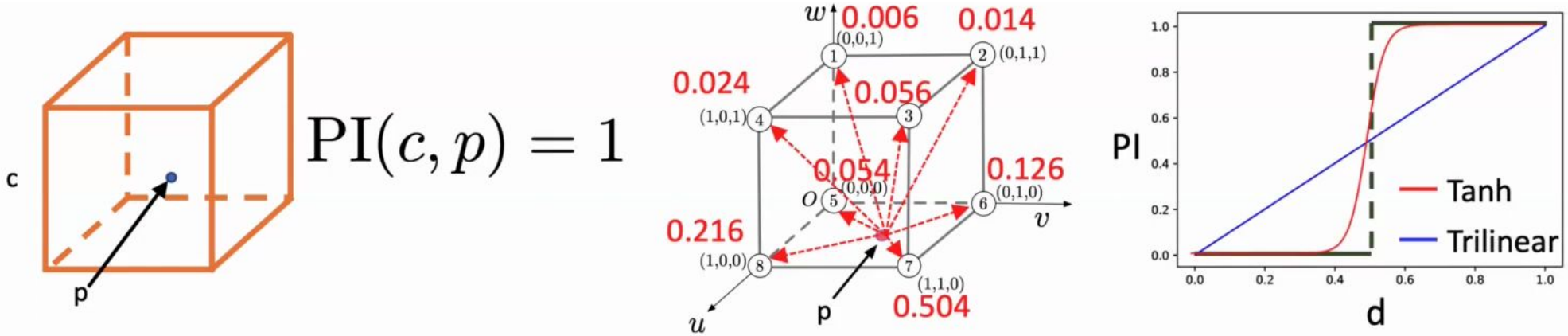
- **SOTA LiDAR-based object detection: cell-level aggregated features**



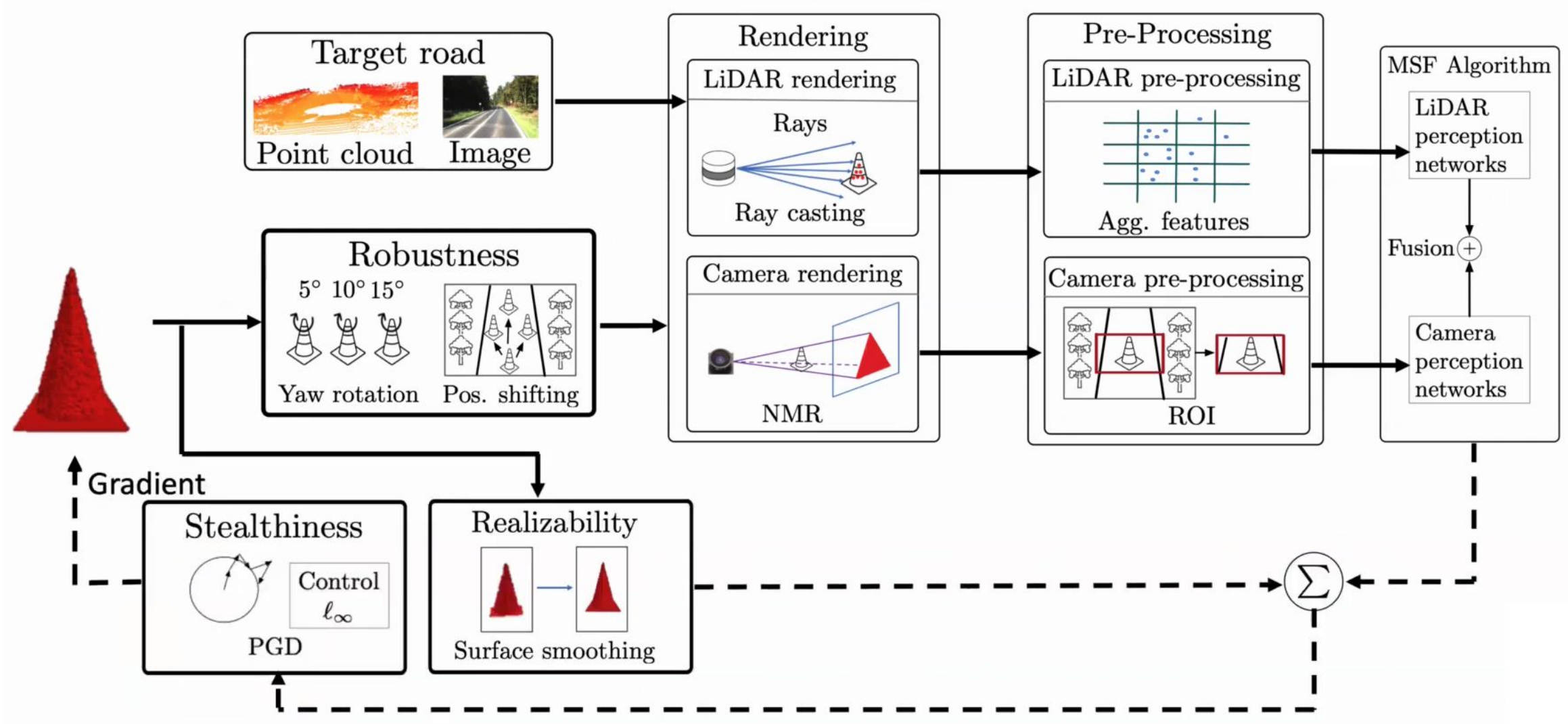
- **Necessary first step: Point-Inclusion (PI) calculation**
  - Calculate whether a point is inside a cell or not
    - Discontinuous by nature
  - Strategy: design a novel and accurate differentiable function to approximate PI

Attack Design: MSF-ADV (LiDAR Pre-processing)

- **Contribution: Soft PI**
  - Estimate probability of point inclusion using interpolation



Attack Design: MSF-ADV (Full Picture)



Evaluation

- **Baselines: 4 MSF algorithms**
  - Algorithms included in Baidu Apollo and Autoware.AI (open-source full-stack AD systems)
- **Effectiveness**
  - **>91% success rate** across different object types and MSF algorithms
- **Robustness**
  - With different victim positions and angles, **>95% average success rate**
- **Transferability**
  - Across different MSF algorithms, **75% average success rate**

Evaluation

- **Physical-world realizability**
  - Miniature-scale, real-vehicle based
  - **≥ 85% success rate**

- **E2E attack simulation: 100% collision rate across 100 runs**

- **Refer to paper for more results**

Defense Directions

- **DNN-level**
  - Tested against 6 existing defenses
  - Most effective defense: attack success rate 66% (not effective enough)

- **Fuse more perception sources**
  - Not a fundamental solution, but can make generation more difficult

Post-Publication Related Research

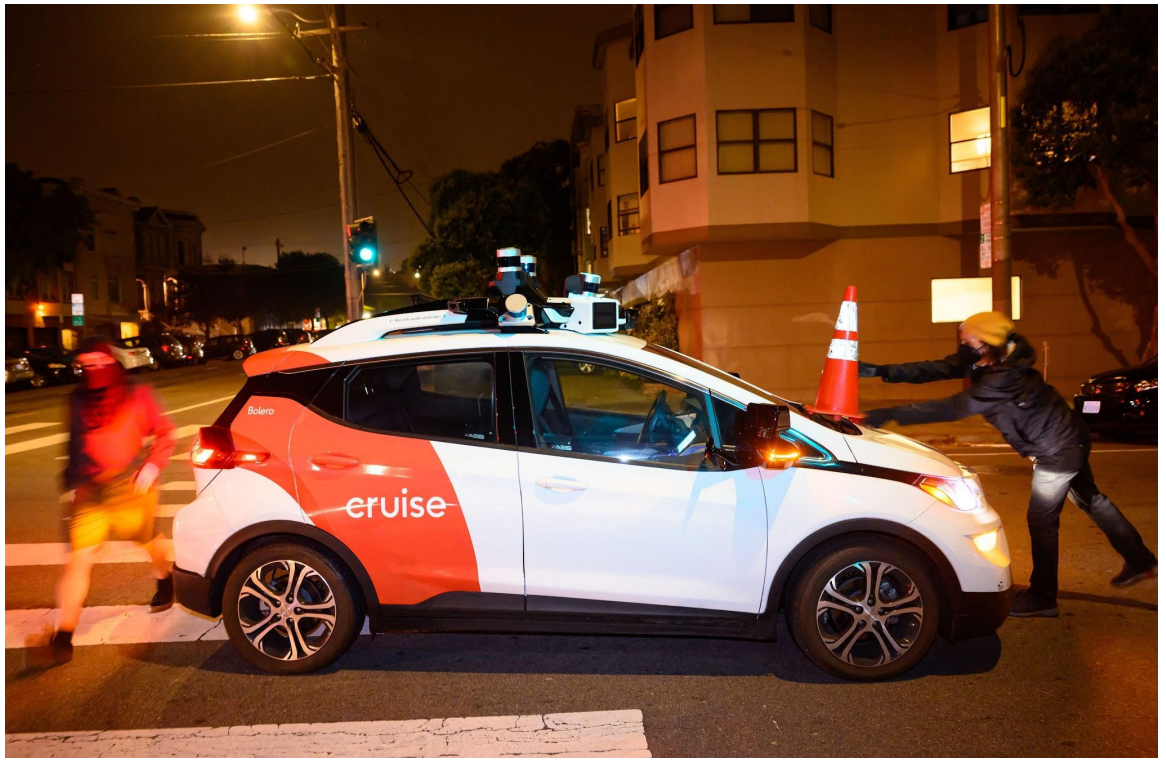
- **Security Analysis of Camera-LiDAR Fusion Against Black-Box Attacks on Autonomous Vehicles** [USENIX Security '22]
  - Camera is only a projection of 2D space; in a frustum, spoofing LiDAR is enough to fool MSF-based AD perception

- **Data-Driven Defenses Against Adversarial Attacks for Autonomous Vehicles** [IAVVC '23]
  - Use deep learning to mitigate attacks on AD perception

Conclusion

- **<u>First study</u> on MSF-based AD perception security**
  - Adversarial 3D object as physical-world attack vector
  - Differentiable approximations of discrete functions for adversarial generation

- **But… do we need all this? Isn't this a bit overkill?**
  - Most adversarial attacks are done with objects close at hand



Discussion Questions (Good Questions)

- I wonder can the system be used to fool reverse, detect an obstacle even there is no obstacle like by putting a thing that is not far from the victim?
- To my knowledge, adversarial objects are perceived differently depending on the viewing distance, but it seems like this factor wasn't considered when optimizing the attack vector. Shouldn't this also be taken into account during the optimization process?
- The author states that hiding nails or glass debris behind an adversarial traffic cone can induce a tire blowout on a target autonomous vehicle (AV). However, wouldn’t the camera and LiDAR be unable to detect nails or glass debris even without the adversarial cone?
- One of the defenses proposed by the authors is to add more sensors in order to diversify the viewing angles, thus reducing the effect of MSF-ADV attacks. How would this significantly affect the overhead in computing resources required? Building on the previous question, would it be feasible instead of adding sensors to one car, to utilize the sensors of nearby cars to add reliability in AD systems?

Discussion Questions (Good Questions)

- Considering that the adversarial 3D-printed objects exploit differences in sensor perception (shape, texture, etc.), how effective would sensor redundancy (e.g., increasing the number of LiDAR units or combining multiple camera angles) be in reducing the vulnerability of MSF systems?
- What practical limitations, such as cost or computational constraints, might hinder the scalability of this defense?
- What specific features of the MSF algorithms used in industry-grade AD systems make them susceptible to the types of attacks described in the paper?
- Given the success of this attack on MSF-based systems, what are the most promising directions for enhancing the robustness of multi-sensor fusion in real-world autonomous driving applications? Could adding more diverse sensor types (e.g., infrared or radar) help, or are there more fundamental algorithmic changes needed to mitigate these attacks?

Discussion Questions (Great Questions)

1. **Jiwoo:** Why does the adversarial box look noticeably different from the original during real-world evaluation, even though the attack vector is constrained to ensure the perturbation doesn't drastically alter the original object for stealthiness? Is the limited perturbation is not sufficient to deceive the entire system in real-world conditions?

2. **Changgeon:** Given that this system is model-specific so the adversarial object must be shaped differently for each car model and software version (model version), and considering that non-fully autonomous vehicles are in use where the user can easily notice and avoid the threat, is this really a realistic concern?

3. **Boris:** Apple just released a monocular ML depth sensing model. Could this type of technology alleviate the need for LiDARs?