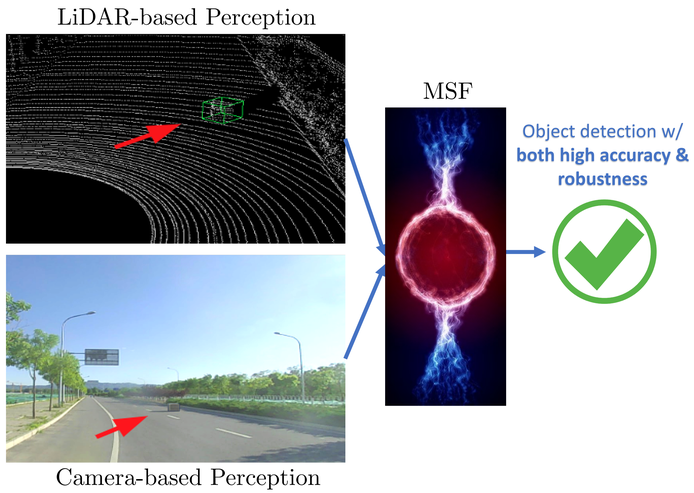
3D Adversarial Object Against MSF-based Perception in Autonomous Driving

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

[*] Equal Contribution, [1]UMich, [2]UCI, [3]Baidu Inc, [4]UIUC

MSF-based Perception in Autonomous Driving (AD)

- Target most performant design: DNN-based perception
- Combine perception results from multiple sensors (e.g., **LiDAR and camera**)
- Improve the **accuracy and robustness** of perception



Problem Statement

- Previous works only consider attacking AD perception on **single sensor (e.g., LiDAR or camera)**
- However, unclear whether it is possible to generate adversarial machine learning attacks for **MSF-based AD perception**

Our Work

- First security analysis of MSF-based perception in Autonomous Vehicle (AV)
- Design a novel attack method, **MSF-ADV**, to generate physical-world adversarial 3D objects that can **simultaneously fool both LiDAR- and camera-based perception**
  - Completely invisible to a victim AV **no matter what MSF algorithms are used**
  - First attack that can **fundamentally** defeat MSF-based AD perception.

Threat Model

- Attacker has full knowledge of the LiDAR- and camera-based perception models, but **doesn't need to know the MSF algorithms**
- Attacker can place a 3D-printed object on the road, and can collect the required sensor data

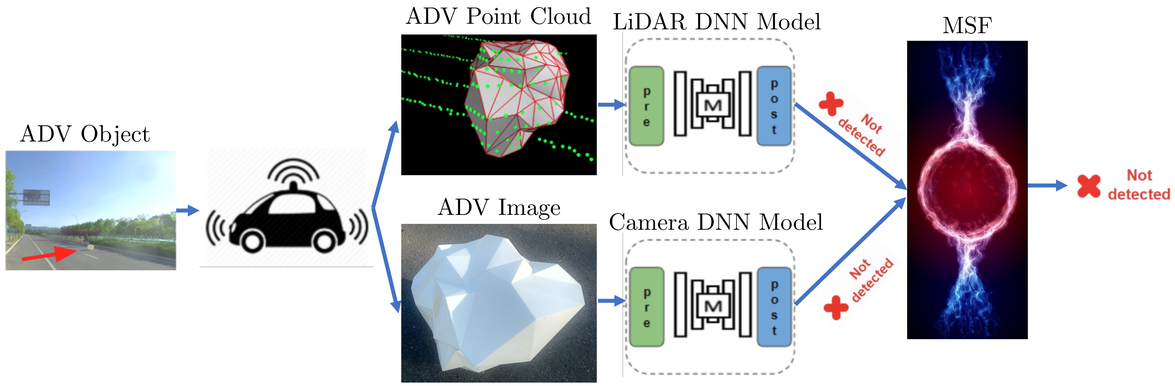
Attack Goal

- Cause the victim AV to **fail in detecting** adversarial object and thus collide into it

Our Approach: MSF-ADV

- Generate malicious model inputs
  - For LiDAR, we generate malicious point cloud by simulating the physics of a LiDAR sensor with **ray casting** and rendering synthetic object into the point cloud
    - Design **differentiable proxy functions** to approximate the non-differentiable pre-processing steps (e.g., point counting)
  - For camera, we obtain malicious images by **calibrating the object position** with LiDAR point cloud and rendering it in the middle of the road
- Objective function

$$\mathcal{L}(\text{obj}) = \underbrace{\lambda_1 \cdot H(p, M_L, \text{obj}) \cdot \mathcal{A}_1}_{\text{LiDAR Loss}} + \underbrace{\lambda_2 \cdot \max[H(c, M_c, \mathcal{A}_2, \text{obj})]}_{\text{Camera Loss}} + \underbrace{\lambda_3 \cdot \mathcal{L}(\text{obj}, \text{obj}^{\text{ori}})}_{\text{Smooth Loss}}$$

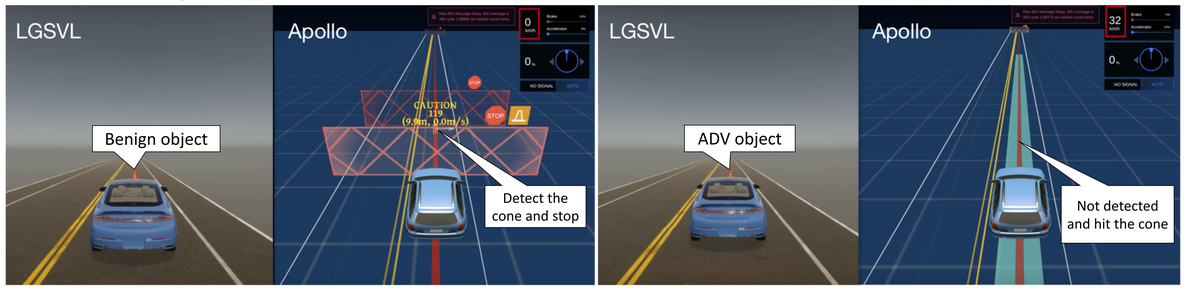


End-to-End Attack Evaluation

- Evaluate our attack on Baidu Apollo-5.0 with LGSVL simulator for benign and adversarial traffic cones

Benign: stopped before the obstacle

Adversarial: hit the obstacle



Early Results

- Evaluation metrics: Object detection rate & end-to-end object collision rate

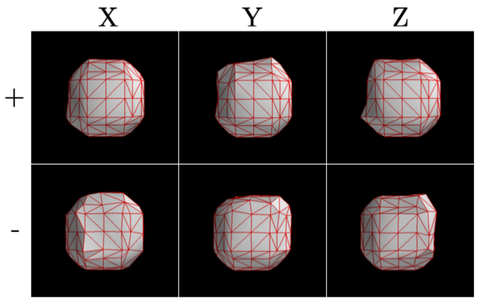
	Detection Rate			End-to-End Collision Rate
	LiDAR	Camera	MSF	
Benign	10/10 (100%)	10/10 (100%)	10/10 (100%)	0/10 (0%)
Adversarial	0/10 (0%)	0/10 (0%)	0/10 (0%)	10/10 (100%)

Future Plans

- More comprehensive evaluation (e.g., more object detection models, attack robustness)
- Consider more sensors instead of LiDAR and camera
- Real-world experiment
- Design effective defense

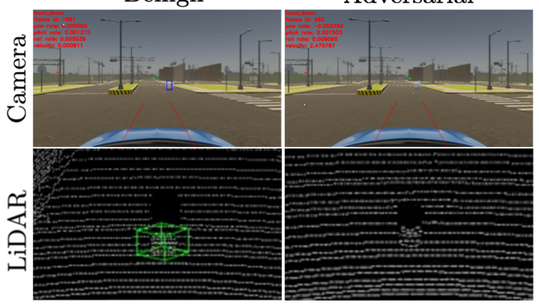
ADV Objects Visualization

- Mesh view of the adversarial object



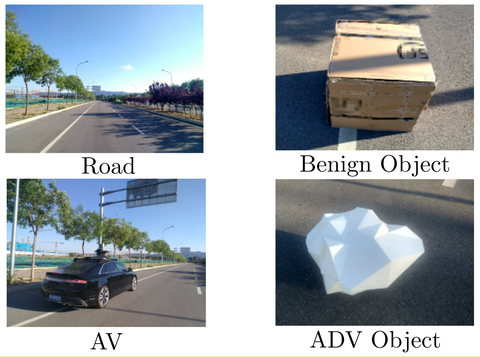
Perception Results Visualization

- Fooling both camera- and LiDAR-based perception



Real-Word Experiment

- Visualization of physical experiment settings



Take a picture for more **details & related materials**

Contact: ningfei.wang@uci.edu