DEMONSTRATION: 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]

ABSTRACT

In Autonomous Vehicles (AVs), Multi-Sensor Fusion (MSF) is used to combine perception results from multiple sensors such as LiDARs (Light Detection And Ranging) and cameras for both accuracy and robustness. In this work, we design the first attack that fundamentally defeats MSF-based AV perception by generating adversarial objects. This demonstration will include both live actions and interactive demonstrations for the generated adversarial objects, attack effectiveness, and the end-to-end consequences.

1 INTRODUCTION

Autonomous Vehicles (AVs), or self-driving cars, are already providing services on public roads. They have adopted different machine learning models and achieved promising performance. Recent studies show that machine learning models for AD perception are vulnerable to adversarial attacks. However, they focus on attacking models for individual sensor (Cao et al., 2019). Multi-Sensor Fusion (MSF) mechanisms have been applied and shown to help improve model robustness and accuracy in AVs, which thus have the potential to correct the attack effects from any individual sensors. This thus raises the question: Is it possible to generate adversarial machine learning attacks for MSF-based AD perception?

In this work, we design *MSF-ADV*, the first attack that can fundamentally defeat MSF-based AD perception. Our method can generate physical-world adversarial 3D objects that simultaneously fool both LiDAR- and camera-based perception. These objects are thus completely invisible to a victim AV no matter what MSF algorithms are used, which can thus cause collisions if placed in the middle of the road.

In this demonstration, we will show this novel attack with both live actions and interactive demonstrations for (1) the generated adversarial objects, (2) the attack effectiveness for both LiDAR- and camera-based AD perception, and (3) the end-to-end security and safety consequences. All these demonstrations are generated for a representative AD system, Baidu Apollo (Apo), and performed in industry-grade AD simulators and/or the physical world.

[*]Equal contribution [1]University of Michigan, Ann Arbor [2]University of California, Irvine [3]Baidu Research, Baidu Inc. [4]University of Illinois at Urbana-Champaign.

Proceedings of the 3^{rd} MLSys Conference, Austin, TX, USA, 2020.

2 THREAT MODEL AND ATTACK GOAL

Threat model. For the system to attack, we assume that the attacker has the full knowledge of the LiDAR- and camera-based AD perception in the victim AD system, which is the same as previous attacks for LiDAR- and camera-based AD perception (Cao et al., 2019).The attacker does not need to know the MSF algorithm used in the victim system. For the physical attack capability, we assume that the attacker is able to place an adversarial 3D object on the road, and can collect the required sensor data in the target road beforehand to facilitate the attack.

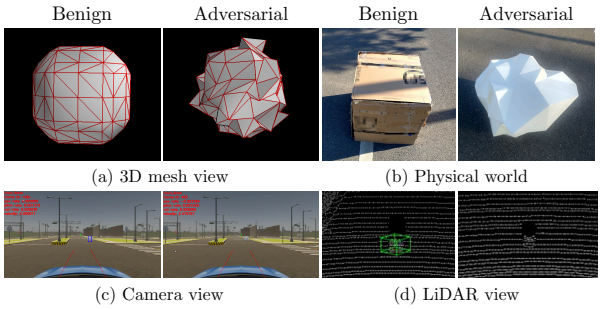
Attack goal. By placing an adversarial object on the road, the attacker's goal is to cause the victim AV to fail in detecting such object and thus collide into it. This thus directly threatens the safety of the passengers in the victim AV.

3 ATTACK DESIGN: MSF-ADV

For LiDARs and cameras, Deep Neural Networks (DNNs) based perception has the state-of-the-art performance and thus is used widely in practice today. Thus, in *MSF-ADV* we formulate the attack as an optimization problem on these DNN models by changing the shape of a 3D object.

Attack design for LiDAR-based perception. First, we design a novel differentiable ray-casting-based renderer to project the shape changes of the 3D object to the point cloud. Second, we design differentiable proxy functions to approximate the non-differentiable pre-processing steps. After that, we design an objective function to reduce the detection probabilities of the adversarial object.

Attack design for camera-based perception. Similar to the design for LiDARs, we use a physics-based renderer



(a) 3D mesh view (b) Physical world

(c) Camera view (d) LiDAR view

Figure 1. Benign and adversarial 3D objects in (a) 3D mesh, (b) physical world, (c) camera view, and (d) LiDAR view.

to project the shape changes to camera input, and model the pre-processing steps. We use objective function from previous work (Xiao et al.).

In *MSF-ADV*, we combine the designs above into a single optimization problem to simultaneously attack both models.

4 Demonstration Plan

4.1 Live Action

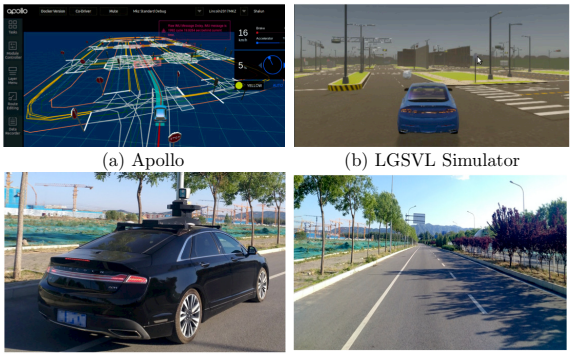
Demonstration of the generated adversarial 3D object. We will show images and videos of benign objects and our generated adversarial 3D objects in different views such as in 3D mesh view and the physical world. Examples are shown in Fig. 1 (a) and (b).

Demonstration of attack effects at perception level. We will show images and videos on how the generated 3D adversarial objects influence the perception results for both LiDAR- and camera-based perception. Examples are shown in Fig. 1 (c) and (d), which are visualized using the default visualization tool provided by Apollo. We will show these results from both a AV simulator and physical-world experiment (experiment setup shown in Fig. 2 (c) and (d)).

Demonstration of the end-to-end attack impact on autonomous driving. We will demonstrate videos of the end-to-end attack impact on autonomous driving by launching the attack to a Baidu Apollo AV running in LGSVL (LGS), an industry-grade AV simulator. In this setup, LGSVL simulates the physical world, and then feeds the real-time sensor data to Apollo to make driving decisions. Fig. 2 (a) shows the driving decision making process in Apollo and (b) shows the simulated physical-world view. The white object in front of the vehicle in Fig. 2 (b) is the adversarial 3D object generated by *MSF-ADV*, and we will demonstrate that the AV will not be able to see it and thus directly crash into it. We will also demonstrate the attack effectiveness in different positions on the maps and types of vehicles.

4.2 Interactive Demonstration

Interactive exploration of the adversarial objects. We will render the generated adversarial 3D objects using Python script using Mayavi library and allow the attendees



(a) Apollo (b) LGSVL Simulator

(c) Car for physical experiment (d) Road for physical experiment

Figure 2. Demonstration of the attack effectiveness in AV simulator and physical-world experiment: (a) Visualization of driving decision making in Apollo, (b) Simulated physical world by LGSVL with adversarial 3D object, (c) Apollo car used in physical-world experiment, and (d) the testing road in physical-world experiment.

to freely interact with the mesh, e.g., by dragging and rotating. We will also 3D print scaled-down versions of the adversarial 3D objects and also benign objects so that the attendees can more directly view and compare them. We are also exploring the possibilities of creating key fobs with the 3D adversarial objects as souvenirs for the attendees.

Interactive demonstration of the attack effectiveness in AV simulator. We will launch the LGSVL simulator and allow the attendees to control the AV themselves in the simulator, e.g., forward, backward, turn left, and turn right, to interactively experience the attack effectiveness. We will place both benign and adversarial 3D objects on the road and provide visualization of the perception results (e.g., camera-based perception, LiDAR-based perception, and MSF-based perception) when the attendees are controlling the AV. The attendees can also specify a destination and let the AV autonomously drive itself towards benign and adversarial 3D object, which allows them to view the end-to-end attack impact on autonomous driving interactively.

4.3 Equipment Requirement

We will bring laptops, objects, and connect to our remote server running the AV simulator for interactive demonstration. It is preferred if we can have (1) high-speed, wired network to minimize the network latency to our remote server, and (2) a big monitor for better visualization.

References

Baidu Apollo. URL apollo.auto.

LGSVL Simulator. URL www.lgsvlsimulator.com.

Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q. A., Fu, K., and Mao, Z. M. Adversarial Sensor Attack on LiDAR-based Perception in Autonomous Driving. In *ACM CCS*, 2019.

Xiao, C., Yang, D., Li, B., Deng, J., and Liu, M. MeshAdv: Adversarial Meshes for Visual Recognition. In *IEEE CVPR*.