Demo: Security of Camera-based Perception for Autonomous Driving under Adversarial Attack

Christopher DiPalma
UC Irvine
cdipalma@uci.edu

Ningfei Wang
UC Irvine
ningfei.wang@uci.edu

Takami Sato
UC Irvine
takamis@uci.edu

Qi Alfred Chen
UC Irvine
alfchen@uci.edu

***Abstract*—Robust perception is crucial for autonomous vehicle security. In this work, we design a practical adversarial patch attack against camera-based obstacle detection. We identify that the back of a box truck is an effective attack vector. We also improve attack robustness by considering a variety of input frames associated with the attack scenario. This demo includes videos that show our attack can cause end-to-end consequences on a representative autonomous driving system in a simulator.**

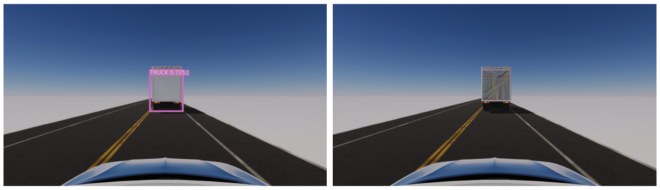
I. Introduction

Autonomous vehicles have become increasingly prominent, and their safety is becoming critical. A fundamental pillar to ensure safe Autonomous Driving (AD) systems is perception, which leverages sensors such as cameras and LiDARs (Light Detection and Ranging) to detect surrounding obstacles in real time. LiDAR excels at detecting obstacles, but it can be prohibitively expensive. As such, many companies have invested in camera-based perception. For example, Tesla and Apollo Lite both apply camera-based perception without LiDAR.

In this demo, we illustrate the effects of adversarial obstacle hiding attack on camera-based perception in an AD system. While several prior works succeed in attacking camera-based object detectors in an AD context, they mostly target traffic sign detection [1], which can be handled via High Definition (HD) map rather than camera-based perception in high-level AD systems [2]. We instead place an adversarial patch on the back of a box truck to make such box truck bypass camera-based AD perception. This demo builds upon prior works by demonstrating end-to-end attack impacts through a production-grade AD simulator, and by using a specialized Expectation over Transformation (EoT) closely relevant to the attack scenario to improve patch robustness. All demonstrations are performed in a production-grade AD simulator and generated for a representative AD system, Baidu Apollo [2], which is ranked in the top 4 leading AD developers along with Waymo, Ford, and Cruise.

II. Threat Model and Attack Goal

Threat Model. As with previous attacks for AD perception [1], we assume the attacker has full knowledge of the camera-based AD perception used in the victim AD system. We also assume the attacker can place an adversarial patch on the back of a box truck, and can collect the required sensor data on the target road beforehand to facilitate the attack.



(a) Box truck with benign patch (b) Box truck with adv. patch

Fig. 1: Snapshots from attack demo videos.

Attack Goal. The attacker's goal is to cause the victim AV to fail in detecting such box truck with the adversarial patch and thus collide into it. This thus directly threatens the safety of the passengers in the victim AV.

III. Attack Design

We add an adversarial patch to the back of a box truck. This patch is generated against camera-based perception in AD system with the objective to minimize detection confidence of the truck over multiple frames. To improve the success rate of the attack, we use EoT [3] to apply 3D perspective rotation, random resize, and crop, along with other distortions.

IV. Demonstration Plan

We will provide videos[1] to demonstrate end-to-end attack impact in an AD system by launching the attack against a Baidu Apollo AV running in production-grade AD simulator LGSVL. We also provide screenshots as per Fig. 1 for both benign and adversarial cases. In Fig. 1 (a), the victim AV can detect the box truck with benign patch and stop before it. However, in Fig. 1 (b), the victim AV cannot detect the box truck with adversarial patch and will thus crash into it.

Acknowledgment

We would like to thank Junjie Shen and the anonymous reviewers for providing valuable feedback on our work. This research was supported in part by the National Science Foundation under grants CNS-1850533, CNS-1929771, CNS-1932464, and USDOT UTC Grant 69A3552047138.

References

[1] Y. Zhao, H. Zhu, R. Liang, Q. Shen, S. Zhang, and K. Chen, "Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors," in *ACM CCS*, 2019, pp. 1989–2004.

[2] "Baidu Apollo," https://apollo.auto/.

[3] A. Athalye, L. Engstrom, A. Ilyas, and K. Kwok, "Synthesizing Robust Adversarial Examples," in *ICML*, 2018.

[1] https://youtu.be/T-dvXFJ3rCg

Workshop on Automotive and Autonomous Vehicle Security (AutoSec) 2021
25 February 2021, Virtual
ISBN 1-891562-68-1
https://dx.doi.org/10.14722/autosec.2021.23018
www.ndss-symposium.org