Demo: CARLA-based Adversarial Attack Assessment on Autonomous Vehicles

Zirui Lan[1], Wei Herng Choong⊠ [2], Ching-Yu Kao[2], Yi Wang[3], Mathias Dehm[4], Philip Sperl[2],
Konstantin Böttinger[2] and Michael Kasper[1]
[1] Fraunhofer Singapore [3] Continental Automotive Singapore [4] Continental Automotive Technologies GmbH
[2] Fraunhofer AISEC ⊠ Corresponding Author: wei.herng.choong@aisec.fraunhofer.de

***Abstract*—Autonomous vehicles rely on a combination of sensors and sophisticated artificial intelligence (AI) systems to perceive their surroundings. The increasing use of AI in autonomous driving technology has brought to our attention the concerns of the implications of AI failure. In this work, we chose an object detector (OD) as an entry point to study the robustness against adversarial attacks like malicious traffic signs. We design and implement CARLA-A3 (CARLA-based Adversarial Attack Assessment), which is a toolkit aimed to streamline the simulation of adversarial conditions and evaluation of OD with several robustness metrics. The toolkit can serve to rapidly and quantitatively evaluate the effects of a malicious sign presented to the OD.**

Introduction. Autonomous driving (AD) envisions vehicles navigating our roads without human intervention, demanding a high level of reliability and safety. AI plays a crucial role in achieving these requirements. Integrating AI into AD systems introduces new challenges, including the vulnerability of AI-based AD systems to adversarial attacks. These attacks exploit weaknesses in AI models, compromising their ability to make accurate decisions, which poses significant risks to the overall safety and functionality of the AD ecosystem. To address this issue, many previous works [2] have been conducted on adversarial attacks and defenses. However, much of the research is centered around digital settings and open datasets, which cannot fully capture the complexities of real-world scenarios. On the other hand, real-world experiments are often not feasible without substantial funding and they come with safety and ethical concerns. In this paper we present a realistic simulation-based approach that addresses this pressing issue. We developed CARLA-A3, a toolkit that focuses on assessing the robustness of OD in the presence of adversarial attacks in a photo-realistic simulation created with CARLA (https://carla.org/). Prior to this research, there was a notable absence of open-source tools to assess the robustness of OD against adversarial attacks on critical traffic signs, especially one that streamlines the creation, rendering, and evaluation of adversarial traffic signs in photo-realistic simulation in diverse weather conditions. Our work bridges this gap and contributes to the development of testing procedures, robustness metrics, and insights essential for enhancing the security and reliability of OD in AD systems.

Symposium on Vehicles Security and Privacy (VehicleSec) 2024
26 February 2024, San Diego, CA, USA
ISBN 979-8-9894372-7-6
https://dx.doi.org/10.14722/vehiclesec.2024.25004
www.ndss-symposium.org

The Proposed Framework. Our proposed framework consists of five extensible modules. The **Scenario Parametrization** module configures the environmental elements of the scenario, which include among others cloudiness or precipitation. The **Attack Generation** module encapsulates some of the current attack methods [2], generating adversarial textures with desired attack effects. The **Attack Injection** module renders adversarial textures on the surface of the target object in the simulator. The **Object Detector** module encapsulates OD algorithms [1] used to perform AD. Finally, the **Evaluation** module assesses the performance using selected metrics [3].

Demonstration Plan. We will provide videos[1] to demonstrate our tool with two different use cases. In the first case, we demonstrate that our tool can easily compare the attack effectiveness of different adversarial patches on the same ground. We repeated the base scenario multiple times, each with a different adversarial patch to be assessed. The adversarial patches are generated from four different attack methods. In the second case, we demonstrate how our tool can evaluate the impact of different environments on adversarial attack methods with ease and efficiency. We conducted experiments using both clean and adversarial stop signs, repeating the base scenario with different weather conditions.

Conclusion. We develop and present a tool that streamlines the simulation of adversarial attacks and the evaluation of OD systems for autonomous vehicles. We showcase the versatility of our tool through case studies. It offers a comprehensive approach to systematically assess the performance of OD under varying conditions without costly and potentially safety-critical field testing. Future work includes increasing the scene complexity, integrating different sensors, and adding attack/defense methods to the tool.

Acknowledgment. The project is funded by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG2-100E-2021-077) and the Bavarian Ministry of Economic Affairs, Regional Development and Energy.

REFERENCES

[1] Z. Zou et. al., Object detection in 20 years: A survey. *Proceedings of the IEEE*, 2023. Publisher: IEEE.

[2] S. Pavlitska et. al., "Adversarial Attacks on Traffic Sign Recognition: A Survey," in *ICECCME 2023*, IEEE, pp. 1-6.

[3] J. Guo et. al., *A comprehensive evaluation framework for deep model robustness* Pattern Recognit., vol. 137, pp. 109308, 2023

[1] https://youtu.be/7C4aAekBbiE