Evaluating Perception Attacks on Prediction and Planning of Autonomous Vehicles

Yanmao Man[1], Raymond Muller[2], Ming Li[1], Z. Berkay Celik[2], and Ryan Gerdes[3]

[1]University of Arizona
[2]Purdue University
[3]Virginia Tech

Motivation: Autonomous vehicles rely on perception to understand the environment. While a number of perception attacks have been proposed, their impact on the decision-making under various driving scenarios remains unclear. In this poster, we take the first step and evaluate the outcome of object misclassification attacks, a representative category of perception attacks due to their consistently high attack success rates [1], where the attacker aims to alter the classification result of an obstacle via physically modifying the obstacle [2–4], or compromising the sensors [5, 6].

Proposed Method: It is nontrivial to evaluate a wide variety of perception attacks, individually or combined together at the same time, under different traffic scenarios in an end-to-end manner; existing works only evaluate their own attacks with limited scenarios [2, 3, 5]. We propose to address this challenge by intercepting the communication among the modules within the decision-making pipeline. In this way, we can modify the output from the perception module arbitrarily such that we can replay the result of any attacks without having to implement them. Then, we can enumerate and search for those attacks that can cause severe consequences but are practical to launch, which will be focused when designing our defenses.

Case Study: We develop a tool that hijacks the cyber channels adopted by Apollo, where it modifies the obstacle class in channel `/apollo/perception/obstacles` on the fly so that the subsequent modules that read it regard the targeted vehicles as pedestrians (to simulate misclassification attacks). We run Apollo in the LGSVL simulator, where we place an NPC vehicle driving straight forward in front of the ego vehicle on either of the adjacent lanes. The test is conducted using five maps (for highway or urban, etc.). For each map, we repeat the simulation ten times, from which similar results are produced: the ego vehicle decelerates to avoid the collision because the obstacle is predicted to change lane (Fig. 1).

Root Cause Analysis: To investigate, we find that modern systems, such as Baidu Apollo and Autoware, include a prediction module, in which future trajectories of obstacles are predicted (lined-up yellow dots in Fig. 1) for precaution purposes. In both platforms, separated procedures are adopted for different object categories due to their naturally distinctive dynamics (Fig. 2): cars tend to run at high speed and follow the lane, while pedestrians are often at low speed and have more freedom. Thus, in Apollo's prediction module [7], for vehicles a CNN+LSTM model is trained only with vehicle data. Only those obstacles classified by the perception module as "vehicles" are fed to it for prediction during test time. Because different prediction models are trained with exclusive datasets, their outputs differ even if the inputs are similar.

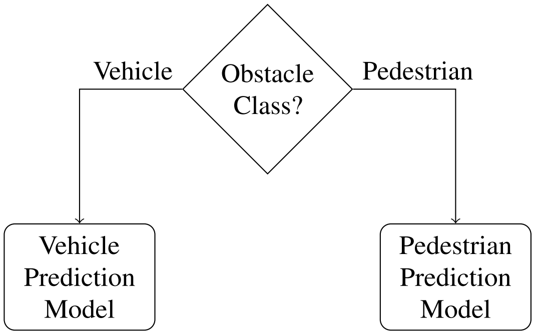
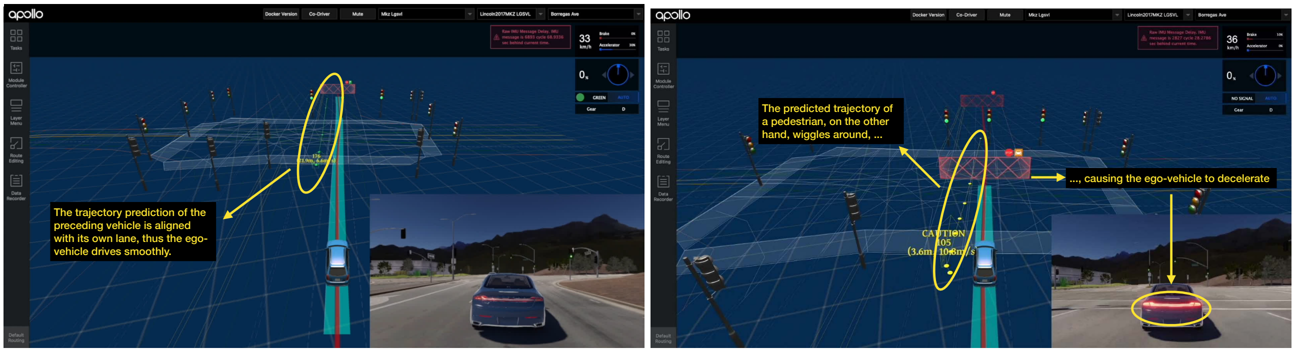


Figure 2: Trajectory Prediction

Impact: This common design exposes a severe vulnerability of such autonomous systems, where the attacker can opportunistically jeopardize the safety, as well as passenger comfort and fuel economy, by simply changing the obstacle class. This is unlike a recent prediction attack [8] where the attacker drives a vehicle to follow the trajectory specially crafted to deceive prediction, which requires precise control. The most recent defense for prediction [9] cannot handle misclassified objects as the class labels are assumed trustworthy.

Future Work: We plan to explore more attack possibilities, e.g., those that succeed only for a portion of the frames. As countermeasures, we can build a more robust classification algorithm [1], detect misclassification attacks [10], or develop a prediction model that is agnostic to object classes or can tolerate class uncertainties, e.g., combining predicted paths weighted by classification confidences.



The trajectory prediction of the preceding vehicle is aligned with its own lane, thus the ego-vehicle drives smoothly.

The predicted trajectory of a pedestrian, on the other hand, wiggles around, ...

..., causing the ego-vehicle to decelerate

Figure 1: (Left) The ego vehicle drives smoothly without attack. (Right) It performs unnecessary deceleration under attack.

References

[1] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li, Z. Zhao, C. Qiao, and Q. A. Chen, "SoK: On the Semantic AI Security in Autonomous Driving," *arXiv preprint arXiv:2203.05314*.

[2] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[3] B. Nassi, Y. Mirsky, D. Nassi, R. Ben-Netanel, O. Drokin, and Y. Elovici, "Phantom of the adas: Securing advanced driver-assistance systems from split-second phantom attacks," in *ACM SIGSAC Conference on Computer and Communications Security (CCS)*, 2020.

[4] R. Muller, Y. Man, Z. B. Celik, M. Li, and R. Gerdes, "Physical hijacking attacks against object trackers," in *ACM SIGSAC Conference on Computer and Communications Security*, 2022.

[5] X. Ji, Y. Cheng, Y. Zhang, K. Wang, C. Yan, W. Xu, and K. Fu, "Poltergeist: Acoustic adversarial machine learning against cameras and computer vision," in *IEEE Symposium on Security and Privacy*, 2021.

[6] Y. Man, M. Li, and R. Gerdes, "Ghostimage: Remote perception attacks against camera-based image classification systems," in *International Symposium on Research in Attacks, Intrusions and Defenses (RAID)*. USENIX Association, 2020.

[7] K. Xu, X. Xiao, J. Miao, and Q. Luo, "Data driven prediction architecture for autonomous driving and its application on apollo platform," in *2020 IEEE Intelligent Vehicles Symposium (IV)*.

[8] Q. Zhang, S. Hu, J. Sun, Q. A. Chen, and Z. M. Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[9] R. Jiao, X. Liu, T. Sato, Q. A. Chen, and Q. Zhu, "Semi-supervised semantics-guided adversarial training for trajectory prediction," *arXiv preprint arXiv:2205.14230*.

[10] Y. Man, R. Muller, M. Li, Z. B. Celik, and R. Gerdes, "That person moves like a car: Misclassification attack detection for autonomous systems using spatiotemporal consistency," in *USENIX Security Symposium*, 2023.