Poster: Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack

Ningfei Wang Yunpeng Luo Takami Sato Kaidi Xu[†] Qi Alfred Chen
University of California, Irvine {ningfei.wang, yunpel3, takamis, alfchen}@uci.edu
[†]Drexel University kx46@drexel.edu

Abstract

In autonomous driving (AD), accurate perception is indispensable to achieving safe and secure driving. Due to its safety-criticality, the security of AD perception has been widely studied. Among different attacks on AD perception, the physical adversarial object evasion attacks are especially severe. However, we find that all existing literature only evaluates their attack effect at the targeted AI component level but not *at the system level*, i.e., with the entire system semantics and context such as the full AD pipeline. Thereby, this raises a critical research question: can these existing researches effectively achieve system-level attack effects (e.g., traffic rule violations) in the real-world AD context? In this work, we conduct the first measurement study on whether and how effectively the existing designs can lead to system-level effects, especially for the STOP sign-evasion attacks due to their popularity and severity. Our evaluation results show that all the representative prior works cannot achieve any system-level effects. We observe two design limitations in the prior works: 1) physical model-inconsistent object size distribution in pixel sampling and 2) lack of vehicle plant model and AD system model consideration. Then, we propose SysAdv, a novel system-driven attack design in the AD context and our evaluation results show that the system-level effects can be significantly improved, i.e., the violation rate increases by around 70%.

I. Main content

This research [1] is recently published in ICCV 2023. The original abstract and author list are shown above. We post the paper links with conference version[1] and arXiv version[2].

II. Acknowledgments

We would like to thank Ziwen Wan, Junjie Shen, Tong Wu, Junze Liu, Fayzah Alshammari, Trishna Chakraborty, and the anonymous reviewers for their valuable and insightful feedback. This research was supported by the NSF under grants CNS-1932464, CNS-1929771, and CNS-2145493; and USDOT UTC Grant 69A3552348327.

References

[1] N. Wang, Y. Luo, T. Sato, K. Xu, and Q. A. Chen, "Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 4412–4423.

[1] https://openaccess.thecvf.com/content/ICCV2023/papers/Wang_Does_Physical_Adversarial_Example_Really_Matter_to_Autonomous_Driving_Towards_ICCV_2023_paper.pdf

[2] https://arxiv.org/abs/2308.11894

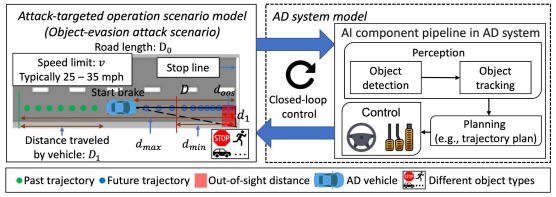
Poster: Does Physical Adversarial Example Really Matter to Autonomous Driving? Towards System-Level Effect of Adversarial Object Evasion Attack

Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu[1], Qi Alfred Chen
University of California, Irvine (UCI) [1]Drexel University
Contact: ningfei.wang@uci.edu
Accepted by ICCV 2023

SysAdv

Research Question

- ***Autonomous Driving (AD) Perception*** is safety-critical
  - Physical object-evasion adversarial attack is one of the most critical ones
    - Cause traffic rule violations, etc.
- *All the prior works only study security of **AI component alone** rather than the **entire AD system pipeline with closed-loop control**.*
- Research question: *Can **existing physical adversarial object evasion attacks** achieve the desired **system-level attack effects** in realistic AD system settings?*



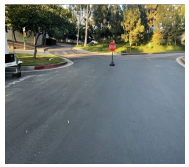
System-Level Effect Measurement of Prior Works

- Methodology and setup
  - AD system pipeline: ***PASS [4]***
    - A system-driven evaluation platform to perform system-level evaluations
    - https://sites.google.com/view/cav-sec/pass
  - Perception results modeling from physical world
    - Goal: improve the fidelity in simulation
    - Collecting the perception results in the physical world and inject the results into the simulator

STOP sign attack selection with the model

Model	YOLO v2 (Y2)	YOLO v3 (Y3)	YOLO v5 (Y5)	Faster-RCN N (FR)
Attack	RP_2 [1]	SIB [2]	FTE [3]	SIB [2]

Website [4]


Physical-world scene


Simulation scene

- Results
  - System-Level effect
    - ***None*** of the existing representative attacks can trigger STOP sign traffic rule violations in ***any*** of the common speeds for STOP sign-controlled roads.

STOP sign violation rate and component-level attack success rate

Eval Level	Speed (mph)	Y2		Y3			Y5		FR	
		B	RP_2	B	SIB	FTE	B	FTE	B	SIB
System	25, 30, 35	0%	0%	0%	0%	0%	0%	0%	0%	0%
Comp.	ASR	-	71.2%	-	53.1%	53.3%	-	41.0%	-	5.2%

violation rate: $\frac{\#\text{ of runs where AD vehicle exceeds STOP line}}{\#\text{ of total runs}}$ Comp.: Component; B: Benign; ASR: Attack Success Rate

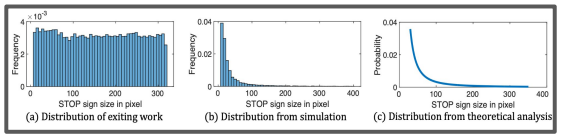

(a) Benign (b) RP_2-Y2 (c) SIB-Y3 (d) SIB-FR (e) FTE-Y3 (f) FTE-Y5

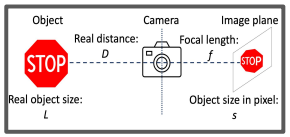
[1] Eykholt et al., Physical Adversarial Examples for Object Detectors, WOOT 2018
[2] Zhao et al., Seeing isn't Believing: Towards More Robust Adversarial Attack Against Real World Object Detectors, ACM CCS 2019
[3] Jia et al., Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems, NDSS 2022
[4] Shen et al., SoK: On the Semantic AI Security in Autonomous Driving, arXiv:2203.05314 2022

System-Driven Attack Design: SysAdv

- Propose SysAdv, a system-driven attack design to overcome two design limitations of prior works.
  - *Physical model-inconsistent object size distribution in pixel sampling*
    - *S1: Novel distribution for object size in pixel*

$$F' = \frac{dF}{ds} = \frac{\eta * L * f}{v * s^2}$$





  - *Lack of vehicle plant model and AD system model consideration*
    - *S2: Ascertain the system-critical range from the vehicle plant model and the AD system model*

$$\arg\min_{p_a} \ \mathbb{E}_{s\sim\mathcal{S}}[\mathcal{L}(M(p_a, O, s, B), \gamma)]$$

$\mathcal{S}$ is the distribution to sample different object sizes in pixels with the system-critical range (include S1 and S2)

Attack Effectiveness Evaluation

- Attack generation and evaluation methodology
  - Same methodology as before
- Results
  - System-level effects can be improved, i.e., ***violation rate increases by around 70%***

STOP sign attack visualization with SysAdv

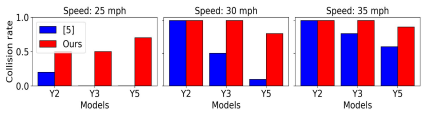
System-level violation rate tested in simulation and component-level ASR evaluation

Eval Level	Speed (mph)	RP_2			FTE-Y3			FTE-Y5		
		S1	S2	S1 + S2	S1	S2	S1 + S2	S1	S2	S1 + S2
System Level	25	90%	90%	**100%**	0%	0%	**40%**	0%	0%	**10%***
	30	-	-	-	0%	30%	**100%**	0%	0%	**80%**
	35	-	-	-	-	-	-	30%	40%	**100%**
Comp. Level	Overall	74.2%	78.6%	**87.8%**	53.6%	54.0%	**70.4%**	42.0%	46.3%	**62.3%**
	SCR	54.7%	67.1%	**84.6%**	36.4%	37.8%	**65.6%**	29.8%	35.9%	**57.4%**

Generality of Our Attack

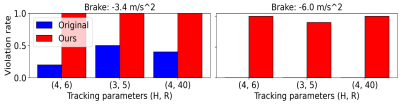
- Generality on different object types
  - Object type: Pedestrian
  - Attack method: ADV-Tshirt [5]
  - Attack method: RP_2, FTE-Y3, FTE-Y5

Pedestrian collision rate tested in simulation with ADV-Tshirt attack on different object detector



- Generality on different AD system parameters
  - Tracking parameters (H, R)
  - Brake deceleration
  - Attack method: FTE-Y5

System-level violation rate tested in simulation on different AD parameter settings



[5] Xu et al., Adversarial T-shirt! Evading Person Detectors in A Physical World, ECCV 2020