WIP: Adversarial Object-Evasion Attack Detection in Autonomous Driving Contexts: A Simulation-Based Investigation Using YOLO

Rao Li
The Pennsylvania State University
rbl5460@psu.edu

Shih-Chieh Dai
The Pennsylvania State University
skd6026@psu.edu

Aiping Xiong
The Pennsylvania State University
axx29@psu.edu

***Abstract*—Physical adversarial objects-evasion attacks pose a safety concern for automated driving systems (ADS) and are a significant obstacle to their widespread adoption. To enhance the ability of ADS to address such concerns, we aim to propose a human-AI collaboration framework to bring human in the loop to mitigate the attacks. In this WIP work, we investigate the performance of two object detectors in the YOLO-series (YOLOv5 and YOLOv8) against three physical adversarial object-evasion attacks across different driving contexts in the CARLA simulator. Using static images, we found that YOLOv8 generally outperformed YOLOv5 in attack detection but remained susceptible to certain attacks in specific contexts. Moreover, the study results showed that none of the attacks had achieved a high attack success rate in dynamic tests when system-level features were considered. Nevertheless, such detection results varied across different locations for each attack. Altogether, these results suggest that perception in autonomous driving, the same as human perception in manual driving, might also be context-dependent. Moreover, our results revealed object detection failures at a braking distance anticipated by human drivers, suggesting a necessity to involve human drivers in future evaluation processes.**

I. Introduction

Artificial intelligence (AI) components in automated driving systems (ADS) are known to be vulnerable to adversarial attacks [7], [9], making ADS susceptible to safety-critical errors that pose significant road hazards and fatalities. Existing works predominantly adopt physical-layer attack vectors, especially physical-world attacks [3], [16], [23], [37]. Physical-world attacks refer to modifying the physical driving environment and consequently tampering with the sensor inputs to AI systems [24]. For instance, previous efforts have leveraged malicious object texture, such as robust physical perturbation, to make a STOP sign undetected by AI systems.

While most existing work has focused on object-evasion attacks at the AI component level, Wang et al. [30] investigated the adversarial attacks at the system level and found that those previously known attacks could not achieve the same results at the system level (e.g., causing collisions and traffic rule violations). Recent research has also started to investigate human drivers' perception and detection of different physical-world attacks. While the participants could differentiate the benign and adversarial objects (i.e., STOP signs [8], [35]), their estimation of the ability of AI systems was largely influenced by their own driving experience [35]. Altogether, these findings highlight the importance of understanding the impact of physical adversarial evasion attacks on downstream tasks, particularly in SAE Level 3 AD [22], where there are situations in which human drivers are required to properly respond to reduce uncertainty in AD systems.

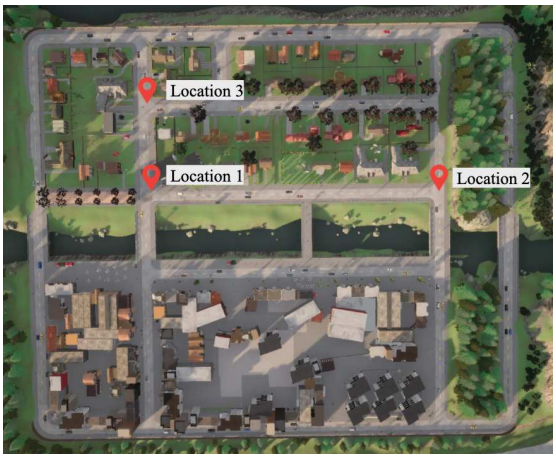


Fig. 1: An overview of three locations of the physical adversarial object evasion attack in the CARLA simulator (Town01_Opt).

In this WIP work, we evaluate *object detection distribution* under physical adversarial object-evasion attacks to inform the design of downstream tasks and human drivers' experience in AD (e.g., determining where to best initialize tracking). We deploy three types of attacks: STOP sign, vehicle, and pedestrian, each comprising two attacks and one benign case. Our evaluation leverages two state-of-the-art variants from the YOLO detector series [20], which offer real-time performance that are critical to AD. We adopt a simulation-based evaluation using the CARLA simulator [5]. Such a method allows us to consider system-level features such as minimal braking distances of AD vehicles and anticipated by human drivers to safety critical events such as STOP sign. This preliminary study helps lay the groundwork for our goal to develop a human-AI collaboration framework to enhance the capacity of ADS to mitigate the attacks with human in the loop.

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.2024.23059
www.ndss-symposium.org

II. Related Work

Physical Adversarial Object-Evasion Attacks and Object Detection in AD. Prior works have constructed various adversarial attacks for object detection focusing on AD tasks [29], [32], [33], [34], [36], [38]. Among them, physical-world adversarial attacks, which often embed malicious sensing inputs to object texture, are fabricated as patches, mainly aiming at road users (e.g., other vehicles, pedestrians) and traffic objects (e.g., traffic signs, road surface) [3], [6], [32], [34], [37]. Based on prior work [24], there are three main types of attack vectors: object texture, object shape and object position. Existing work mainly adopts these physical world attacks [16], [24]. About half of these attacks are related to camera-perceived object textures [24]. However, just as recent works found, compared to the high success rate of AI components, the feasibility of these attacks could vary considering different types of detection models and integration of system models [24], [30].

Object Detection Model in AD contexts. The state-of-the-art camera-based perception in AD contexts leverages Neural Networks (NNs) algorithms. Recent advances include one-stage object detectors and two-stage object detectors. While two-stage models, such as Region-based Convolutional Neural Network (R-CNN) and Fast R-CNN [21], usually have higher detection accuracy, one-stage models such as the YOLO series [4], [20] have been widely used in real-time scenarios where higher detection speed is critical. YOLOv5 [26] and YOLOv8 [13] are the two latest versions of YOLO family, which outperform the previous versions of YOLOs, achieving high accuracy and speed in practice. Few recent studies have examined YOLOv5's performance under adversarial attacks in AD contexts [30], but no existing work has evaluated YOLOv8 in such settings. The current study fills the gap.

Driving Simulation of AD. AD is safety critical. Thus, evaluating its robustness to adversarial attacks in the field is utterly challenging because of the difficulty in controlling real driving environments required in testing scenarios. Some researchers have crafted and evaluated adversarial situations in the virtual environment. Those prior work has shown simulation engines, such as LGSVL [30] and CARLA [1], [17], [18], to be a viable solution in alleviating the challenge. In this paper, we rely on CARLA to deploy detection models to measure physical-world attacks (see Figure 1) and evaluate the camera-based AD perception. The whole ecosystem of CARLA is well developed to be open-source, and is constantly maintaining updates [5]. Also, a simulation-based approach is suitable for our future human-subject evaluation when applying human-AI collaboration to mitigate the attacks in AD.

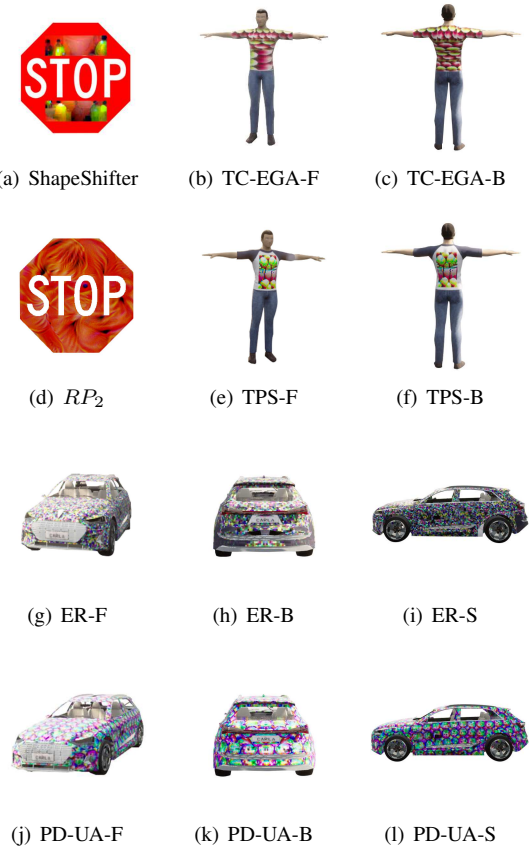
III. Current Study

The current study aims to obtain a comprehensive understanding of *object detection distribution* under physical adversarial object-evasion attacks. To achieve the goal, we focus on dynamic autonomous driving contexts at two levels: 1) *the target-object level* and 2) *the driving-context level*. Thus, we first vary the viewpoints of target objects to represent their possible appearances in driving and evaluate the object detection rates. We then manipulate the driving contexts for each attack target and examine the impact of driving contexts on object detection.

A. Evaluation Setup

We evaluate the object detection at both levels. In this preliminary work, we measure the object detection rate at the target-object level using static images [7] and the driving-context level using recorded videos [30]. We elaborate on the details of each evaluation in the following subsections.

Attack Selection and Reproduction. Based on the prior works [30], [35], we examine three main types of physical adversarial object-evasion attacks regarding object texture: the most extensively-studied STOP sign attacks [3], [6]; pedestrian-evasion attacks [10], [34]; and vehicle-evasion attacks [31]. As shown in Figure 2, we include two variants for each attack type.



(a) ShapeShifter (b) TC-EGA-F (c) TC-EGA-B

(d) RP_2 (e) TPS-F (f) TPS-B

(g) ER-F (h) ER-B (i) ER-S

(j) PD-UA-F (k) PD-UA-B (l) PD-UA-S

Fig. 2: Static images of the two variants in each attack. "-F", "-B" and "-S" denote the images of front view, back view, and side view, respectively.

1) *STOP Sign*: Because the STOP sign attack is the most common physical-world attack, there are various forms [6], [12], [37]. We select two representative examples: *ShapeShifter* [3] and *Robust Physical Perturbations (RP_2)* [7], both of which have been systematically studied in recent papers and proved to be the most effective attack designs [30].
2) *Pedestrian*: We choose 'pedestrian' as the second target object type because making pedestrians vanish on the road or miss-detected could be a safety hazard and have a significant impact on AD. Two most representative patch attacks are considered: *Toroidal-Cropping-based*

(a) Location 1 (b) Location 2 (c) Location 3

Fig. 3: Screenshots of the three locations in the CARLA simulator.

Expandable Generative Attack (TC-EGA) [10] and *thin plate spline-based transformer (TPS)* [34].

3) *Vehicle*: As mentioned above, vehicle is another vital object in AD contexts. We thus perform two variants of vehicle-evasion attacks: *Enlarge-and-Repeat (ER)* [31] and *Prior Driven Uncertainty Approximation (PD-UA)* [15]. ER is the state-of-the-art attack on vehicles in a physical simulator [31]. PD-UA represents the state-of-the-art in adversarial attack patterns, demonstrating superior attacking performance on the ImageNet validation set. To the best of our knowledge, PD-UA has not been studied in vehicular scenarios. We fill the gap in the current study. Note that in this preliminary study, we kept the target vehicle in the same location (i.e., the vehicle is stationary). Consequently, the relative speed between the ego vehicle and the targeted vehicle is the same as in the cases of the STOP sign and pedestrian.

To implement the attacks in CARLA, we manually extract the patterns used in each attack as proposed by the authors. Subsequently, we alter the texture and material of the STOP sign, pedestrian (`walker.pedestrian.0002`), and vehicle (`vehicle.audi.etron`) in CARLA to execute the perturbations. The results are shown in Figure 2. We place the objects on the map using a practical setup in the real world. For example, we position the STOP sign on the right-hand side at a STOP sign intersection (see Figure 3). The pedestrian is placed at the same location but on the sidewalk.

Object Detection Models. We evaluate these attacks using YOLOv5 and YOLOv8. YOLOv5 has been studied in the physical adversarial object-evasion attacks [30]. Thus, the detection results in this work can be compared to the results of existing studies. Additionally, YOLOv8, being the most recent and advanced model in the YOLO series, is included due to its relevance, novelty, as well as the limited number of studies on AD tasks using it. We choose to use both YOLOv5 and YOLOv8 in their original forms to ensure an unbiased evaluation and provide clear, straightforward results. For both models, we select the versions with the lightest parameters (`YOLOv5n`: 1.9M and `YOLOv8n`: 3.2M).

Simulation Setup. Due to the limitations of real-world AD testing in safety and efficiency, using simulation engines to evaluate and measure the effects of attacks and detection models has been widely adopted and proven to be effective [1], [17], [18], [23], [28], [30], [36]. Using a virtual scenario with high fidelity, we simulate autonomous driving in the CARLA driving simulator [5], in which the ego vehicle could run safely and time-efficiently under control. The CARLA version is 0.9.14. In CARLA, we use Town01 Opt, an official map (see Figure 1) on a sunny day at noon, which is a common setting and the most representative setup. We record videos in the simulation engine for attack generation and manually feed the videos into the detectors to obtain the perception results. We conduct all the experiments in the simulator. Screenshots of STOP signs located on the roadside are shown in Figure 3.

Procedure. Our goal is to evaluate the object detection distribution of two object detectors under various attacks. We started with the investigation by varying the viewpoints of three target objects under different attacks. We input images of perturbed STOP signs, pedestrians, and vehicles into the models to observe the detection rates. Each attack detection was run three times in both YOLOv5 and YOLOv8. The average results are presented in Table I.

To further study the object detection distribution, we deployed these attacks in CARLA and conducted evaluations of each scenario (two attacks and one benign case) across three different locations (see Figure 1). The driving speed of our test vehicle was set at 30 mph, a common speed in the United States in areas where STOP signs are typically found. We maintained this fixed speed to assess the detection rate at different distances and in various contexts. To make the current study results comparable to the existing findings, our starting point for testing the object detectors was 50 m, in line with the setup used in prior research [30]. We then recorded the perception results in each scenario as the ego vehicle approached the attacked object, using a video stream. Such a method allowed for a direct observation of how detectors performed over time at different distances, providing detailed confidence scores. We evaluated three scenarios (two attacks and one benign case) for each target object. For example, in the STOP sign case, we assessed the benign scenario, ShapeShifter, and RP_2 at three locations, respectively. Each scenario in a location was run ten times in both YOLOv5 and YOLOv8. In total, we compiled 180 videos, with each captured at 30 FPS, amounting to a 7-second footage. All scenarios were recorded at a resolution of 640×640.

B. Results

Static Tests. Table I presents the detection results of those images in Figure 2. Our study leads to several findings. *First*, the detection performance of YOLOv8 (AVG: 0.63) was improved compared to that of YOLOv5 (AVG: 0.08), especially for the attacks such as RP_2, TC-EGA, and TPS. The high precision rate achieved by YOLOv8 might be due to its use of revised anchor box and modified loss function, improving

TABLE I: Detection results of static images. Each cell contains 3 runs and shows the average confidence score of the true attack-aimed object. -F, -B, and -S indicate the front-view, back-view, and side-view, respectively.

Object detectors	STOP sign attacks		Pedestrian attacks				Vehicle attacks					
	ShapeShifter	RP_2	TC-EGA-F	TC-EGA-B	TPS-F	TPS-B	ER-F	ER-B	ER-S	PD-UA-F	PD-UA-B	PD-UA-S
YOLOv5	0.40	0	0.27	0	0	0.31	0	0	0	0	0	0
YOLOv8	0.77	0.95	0.9	0.91	0.92	0.93	0.25	0	0.93	0.41	0	0.58

the overall accuracy of the object detection process [13]. *Second*, regarding the attack types, attacks aimed at vehicles achieved higher success rates than those on STOP signs and pedestrians. Particularly, YOLOv5 failed to detect all cases, and YOLOv8 showed a poor performance in general (AVG: 0.36). *Third*, regardless of the viewpoints of vehicles, most of the attacks can achieve high attack effectiveness. One possible reason could be that fewer studies have been published on the aspect of vehicle detection compared to the detection of pedestrians [25], suggesting fewer training datasets of vehicles than pedestrians for training those detection models. Also, the shape of the pedestrian was relatively fixed from the front-view and the back-view, whereas the shapes of the vehicle varied significantly across the three different viewpoints, implying increased difficulties for accurate detection. *Fourth*, for the back view of the vehicles, both detectors failed to detect the vehicles, which suggests that the target's viewpoints could have a significant influence on the detection. Such an impact of viewpoint was also evident for the pedestrian detection with YOLOv5[1]. *Lastly*, for the STOP sign, RP_2 attack worked better than ShapeShifter on YOLOv5, which might be due to the differences in targeted models for those attacks (i.e., ShapeShifter is aimed at Fast-RNN detector [3] and RP_2 is created to attack YOLOv2 [6]). While YOLOv8 showed better performance in detecting both attacks, particularly RP_2, ShapeShifter attack still revealed a non-negligible impact with YOLOv8.

Dynamic Tests. We further evaluated the detection distribution of each target as a function of detecting distance. The average perception results across three locations for each scenario (two adversarial cases and one benign cases) of all three targets are visualized in Figures 4 and 5.

Wang et al. [30] evaluated adversarial attacks at a system level grounding on a minimum braking distance of 15 m for the AD vehicle at a speed of 30 mph to avoid collision. We consider it as a system feature in our investigation. According to a report from the U.S. Department of Transportation's National Highway Traffic Safety Administration (NHTSA), the total braking distance at 30 mph under manual driving, including human reaction time, is approximately 36 m [19]. Therefore, we also consider such a human factor in the evaluation.

Results of the STOP sign are shown in Figure 4. In most cases, the confidence score is above chance (AVG: about 0.8) between 25 m and 5 m. Compared the results between benign and adversarial scenarios, the two detectors performed similarly, showing all types of STOP-sign attacks did not evade the detection in AD contexts. Similar patterns were evident for the results of pedestrian and vehicle (see Figure 5). Thus, none of the existing representative attacks we chose had achieved a high attack success rate across these dynamic tests. In other words, when it comes to a dynamic scenario or system level, existing adversarial object evasion attacks may not impact the ADS, which is in agreement with the results of prior work [30].

Nevertheless, both detectors failed to detect the targeted objects or predicted them with mediocre confidence scores at around 36 m except for the vehicle attack. While those failures were not accounted as evasions, undetected attacks at such a distance could result in AD vehicles braking at a distance shorter than 36 m. Because human drivers' anticipated braking distance is largely based on their prior driving experience [35], a shorter-than-expected braking distance could negatively impact their AD experience.

Moreover, we found that the detection results varied across different locations. As shown in Figure 4, the overall detection results of the STOP sign in Location 2 are better than those in Locations 1 and 3. We also visualize the detection results of the other two attacks across the three different locations[2]. Different from the STOP sign, attacks have the best detection rates under Location 3 for the pedestrian. Detection results of the vehicle are good in general and show slightly better in Locations 2 and 3. Such results indicate that object detection in YOLO, in particular YOLOv8, might be context-dependent.

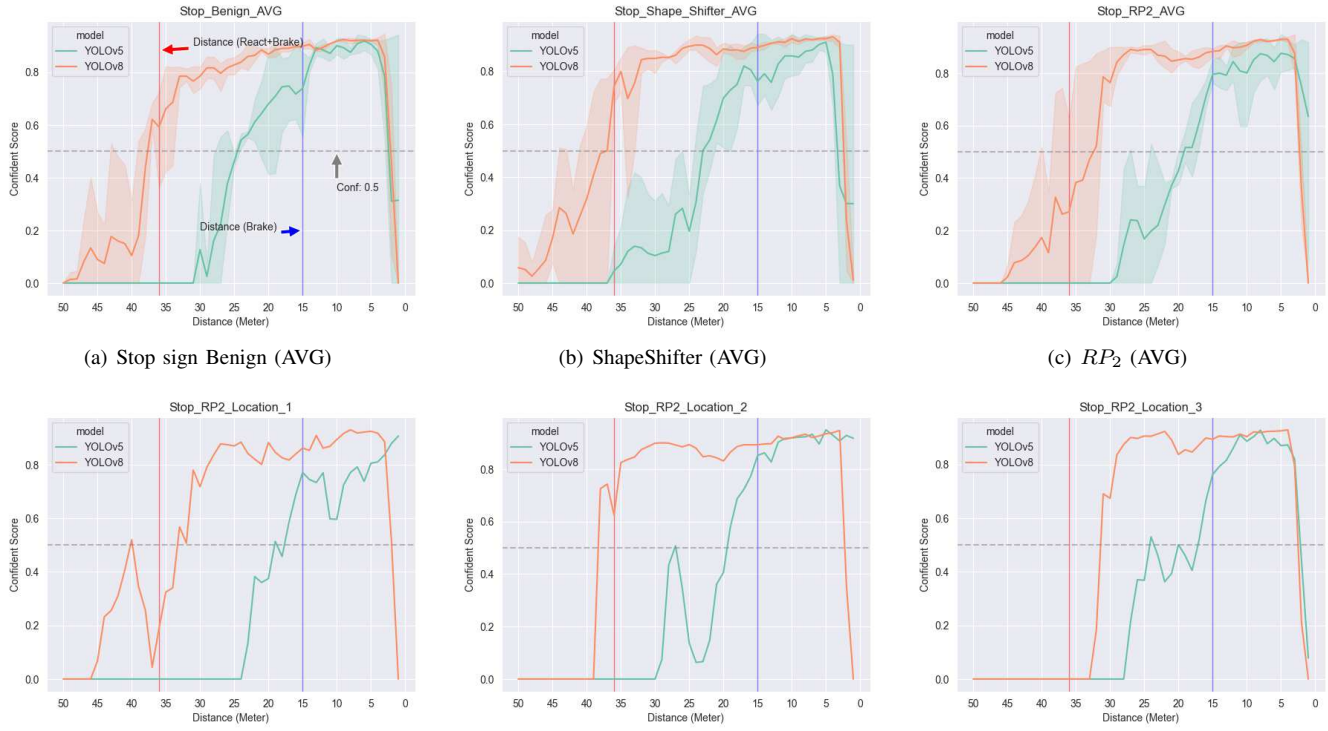
C. Discussion and Future work

Perception Evaluation with System-Level Features. Our results reveal that the state-of-the-art object detection models could fail at minimal braking distances anticipated by human drivers and of AD perception. In the case of YOLOv5, our observations indicate that it failed to timely detect the STOP sign and pedestrian at 36 m. Yet, it worked effectively at 15 m from the targeted objects with an average confidence score higher than 0.5. YOLOv8 managed to detect the STOP sign and the vehicle in time but failed to detect the pedestrian at 36 m. One potential explanation for such failure in both YOLOv5 and YOLOv8 could be that when pedestrians are at far distance, their shape and boundaries can be obscure. Thus, it would be difficult to detect objects in a cluttered environment with different visual characteristics and features [14]. Considering that human drivers may anticipate a minimal braking distance more than twice than the AD vehicle, future work should involve humans in the evaluation processes to ensure a user-centered AD design.

While both models could fail to detect objects in the static settings, they were capable of detecting all perturbed objects

[1]To ensure the obtained results of vehicle are not specific to the vehicle model (i.e., SUV), we ran the same evaluation on a sedan and obtained the similar results. We will report the result in the complete paper.

[2]Due to page limits, we will report the results in the complete work.



(a) Stop sign Benign (AVG) (b) ShapeShifter (AVG) (c) RP_2 (AVG)

(d) RP_2 (Location 1) (e) RP_2 (Location 2) (f) RP_2 (Location 3)

Fig. 4: Detection results of STOP sign with two detectors. The red line indicates the minimum braking distance including human reaction time. The blue line represents the minimum distance of AD. The gray dash line points out the confidence score in 0.5.

from certain distances in the dynamic settings. Such results suggest that more evaluations at the system level are necessary to enhance our understanding of physical adversarial object-evasion attacks in AD.

The Impact of Contexts in AD Perception. As shown in Table I and Figures 4 and 5, YOLOv8 indeed outperformed YOLOv5 in both settings, which is consistent with the results of prior work [13], [27]. Because YOLOv8 uses mosaic data augmentation to provide the model with better context information [13], a possible explanation for the differences is that other objects in the contexts might have assisted the model in making predictions. For example, the model might be more likely to anticipate the presence of STOP signs at intersections, thus successfully detecting the targeted objects. Thus, as in manual driving, our results imply that context is also a critical factor in AD perception.

The Viewpoints of Targeted Objects. As is shown in Figure 2, attacks (ER and PD-UA) aimed at vehicles have three different viewpoints in the static tests. Results in Table I show both detectors failed when ER and PD-UA attacks tested the vehicle with back views, probably because these two attacks are relatively less investigated. Such results highlight that it is essential to have similar investigations in the dynamic settings since there are scenarios where vehicles/pedestrians would approach the ego vehicle from different viewpoints (e.g., side-view vehicles at intersections with STOP signs).

In the dynamic tests, when the ego vehicle approached the target objects, the object detectors generally performed well. However, when the distance was in the region [0m, 5m], a significant drop in detection rates was evident in general. Such results could be due to the detection angle between the detector and the perturbed objects being too large but the range of the detector's vision being too narrow, causing target deformation and missing. While such a viewpoint could influence the detection rate, it should not be a concern since any safety-critical actions might have already been taken before the ego vehicle got so close to the targets.

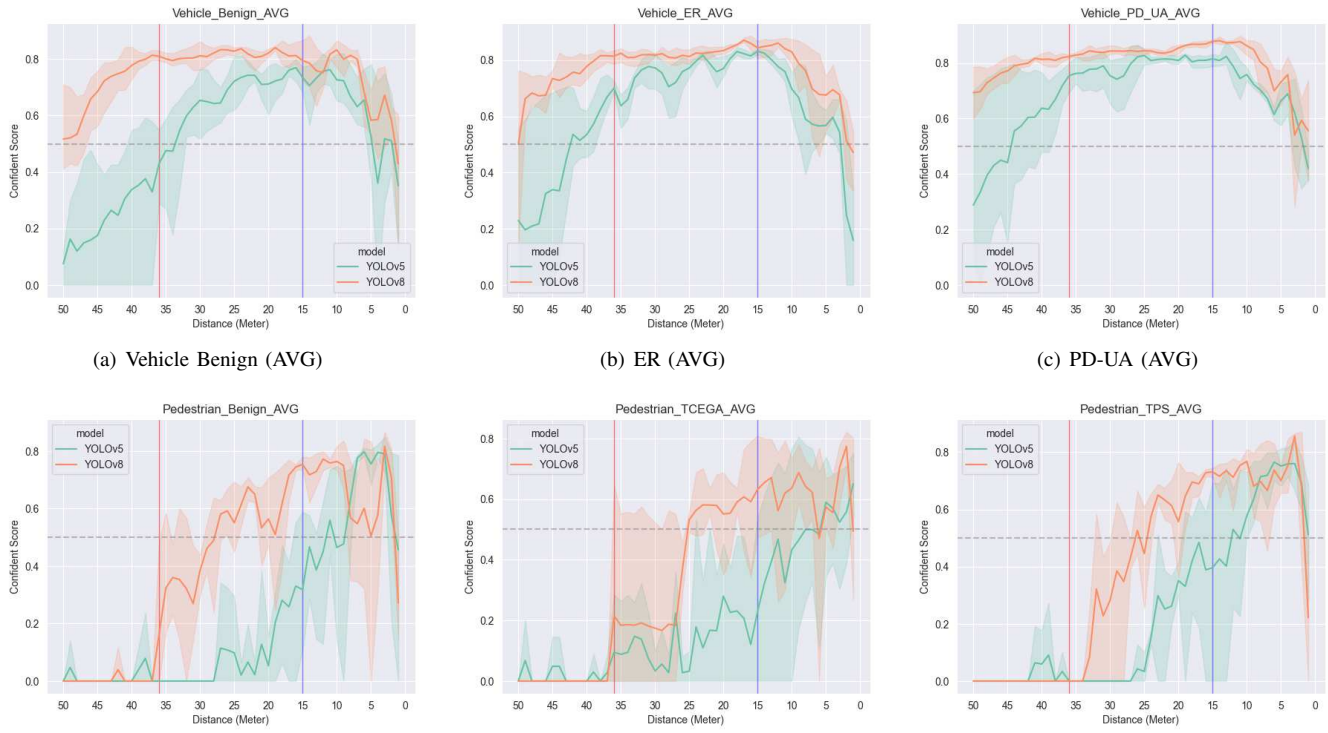
IV. Limitations

We discuss the limitations of the current study as follows.

A. Attacks

We only investigated three types of attacks (STOP sign, vehicle, and pedestrian) in this study. Other attack types such as shade attacks and road condition attacks that could result in hazards [1], [23], were not considered. Future work could include more attack types to understand the corresponding detection results. Moreover, we implemented one pedestrian model (`walker.pedestrian.0002`) in the current work. Thus, the results may not be applicable to individuals of different gender, age, and ethinicity groups. Considering the bias revealed in the literature [2], we recommend further research to increase the diversity of investigated entities.

Also, there are a few limitations to the dynamic tests in our study. First, the target vehicle remained stationary in the simulator. However, a moving vehicle may yield different results. For example, the relative speed between the ego vehicle and another moving vehicle would vary, thereby altering the



(a) Vehicle Benign (AVG) (b) ER (AVG) (c) PD-UA (AVG)

(d) Pedestrian Benign (AVG) (e) TC-EGA (AVG) (f) TPS (AVG)

Fig. 5: Different attacks on vehicle and pedestrian with two detectors' perception results. The red and blue lines indicate the minimum braking distance anticipated by human drivers and of AD. The gray dash line points out the confidence score in 0.5.

minimum braking distance required. Second, in our dynamic tests, the position of the perturbed vehicle is fixed, allowing the ego vehicle only to view its front. Given the impact of viewpoints in our static evaluation, a more comprehensive study including various viewpoints is recommended for future dynamic tests. Finally, we used only one type of vehicle (`vehicle.audi.etron`), an SUV model. Incorporating a broader range of vehicle models, such as sedans, CUVs, hatchbacks, and trucks, would be necessary to understand whether the findings can be generalized to other vehicle models. We intend to explore it in future work.

B. CARLA Simulation.

CARLA is widely used by the research community for ADS-related research. It allows researchers to simulate various conditions quickly and safely. However, we believe there is still a gap between the simulation and real-world settings. For example, the simulation-based approach adopted by the current study may not fully capture the complexity of real-world environments and scenarios. Additionally, our study was conducted on only one map (Town01_Opt), which resembles a small town. In future work, we plan to include more maps with more complex conditions, such as urban environments.

C. Object Detection Model.

In this study, we used two off-the-shelf object detection models (YOLOv5 and YOLOv8). Yet, it is uncertain whether the object detectors used in real-world vehicles perform comparably to the YOLO models. Furthermore, real-world vehicles rely on additional sensors and information for tracking and action planning. For example, temporal cues are important for dynamic tracking, but they are not considered in static image detectors [11]. While we believe our study provides valuable insights for researchers in developing ADS, there may be a gap between our study and real-world applications.

V. Conclusion

In this WIP work, we examine the performance of YOLOv5 and YOLOv8 against three types of physical adversarial object-evasion attacks: STOP sign, vehicle, and pedestrian. The results indicate that although YOLOv8, a state-of-the-art object detector, generally outperforms YOLOv5, it remains susceptible to certain attacks. Furthermore, our study provides insights into the models' behaviors, which indicates that the context can be a critical factor for detector models when facing adversarial attacks in automated driving (AD). Such finding is valuable for the development of automated driving systems (ADS) in future work. Moreover, our results reveal detection failures at a minimal braking distance anticipated by human drivers, suggesting the necessity of bringing human in the loop to mitigate the attacks in AD.

Acknowledgments

The authors would like to thank the anonymous reviewers for their valuable and insightful feedback. This research was supported in part by Penn State under CSRAI Seed Grant and the National Science Foundation under grant 2121097.

REFERENCES

[1] A. Boloor, X. He, C. Gill, Y. Vorobeychik, and X. Zhang, "Simple physical adversarial examples against end-to-end autonomous driving models," in 2019 IEEE International Conference on Embedded Software and Systems (ICESS). IEEE, 2019, pp. 1–7.

[2] M. Brandao, "Age and gender bias in pedestrian detection algorithms," arXiv preprint arXiv:1906.10490, 2019.

[3] S.-T. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Shapeshifter: Robust physical adversarial attack on faster r-cnn object detector," in Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2018, Dublin, Ireland, September 10–14, 2018, Proceedings, Part I 18. Springer, 2019, pp. 52–68.

[4] J. I. Choi and Q. Tian, "Adversarial attack and defense of yolo detectors in autonomous driving scenarios," 2022.

[5] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in Proceedings of the 1st Annual Conference on Robot Learning, 2017, pp. 1–16.

[6] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, F. Tramer, A. Prakash, T. Kohno, and D. Song, "Physical adversarial examples for object detectors," in 12th USENIX Workshop on Offensive Technologies, 2018.

[7] 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, pp. 1625–1634.

[8] K. R. Garcia, S. Mishler, Y. Xiao, C. Wang, B. Hu, J. D. Still, and J. Chen, "Drivers' understanding of artificial intelligence in automated driving systems: A study of a malicious stop sign," Journal of Cognitive Engineering and Decision Making, vol. 16, no. 4, pp. 237–251, 2022.

[9] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.

[10] Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu, "Adversarial texture for fooling person detectors in the physical world," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 13 307–13 316.

[11] L. Huang, X. Zhao, and K. Huang, "Bridging the gap between detection and tracking: A unified approach," in Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019.

[12] W. Jia, Z. Lu, H. Zhang, Z. Liu, J. Wang, and G. Qu, "Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems," arXiv preprint arXiv:2201.06192, 2022.

[13] G. Jocher, A. Chaurasia, and J. Qiu, "Ultralytics yolov8," 2023. [Online]. Available: https://github.com/ultralytics/ultralytics

[14] J. Li, X. Liang, S. Shen, T. Xu, J. Feng, and S. Yan, "Scale-aware fast r-cnn for pedestrian detection," IEEE Transactions on Multimedia, vol. 20, no. 4, pp. 985–996, 2017.

[15] H. Liu, R. Ji, J. Li, B. Zhang, Y. Gao, Y. Wu, and F. Huang, "Universal adversarial perturbation via prior driven uncertainty approximation," in 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2941–2949.

[16] Y. Man, M. Li, and R. Gerdes, "{GhostImage}: Remote perception attacks against camera-based image classification systems," in 23rd International Symposium on Research in Attacks, Intrusions and Defenses, 2020, pp. 317–332.

[17] 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, 2023.

[18] F. Nesti, G. Rossolini, S. Nair, A. Biondi, and G. Buttazzo, "Evaluating the robustness of semantic segmentation for autonomous driving against real-world adversarial patch attacks," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2022, pp. 2280–2289.

[19] NHTSA, "Why your reaction time matters at speed," pp. 1–4, 2015.

[20] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," Advances in Neural Information Processing Systems, vol. 28, 2015.

[22] SAE, "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," 2021, available at https://www.sae.org/standards/content/j3016_202104/.

[23] T. Sato, J. Shen, N. Wang, Y. Jia, X. Lin, and Q. A. Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack," in 30th USENIX Security Symposium, 2021, pp. 3309–3326.

[24] J. Shen, N. Wang, Z. Wan, Y. Luo, T. Sato, Z. Hu, X. Zhang, S. Guo, Z. Zhong, K. Li et al., "Sok: On the semantic ai security in autonomous driving," arXiv preprint arXiv:2203.05314, 2022.

[25] D. Sun and J. Watada, "Detecting pedestrians and vehicles in traffic scene based on boosted hog features and svm," in 2015 IEEE 9th International Symposium on Intelligent Signal Processing (WISP) Proceedings, 2015, pp. 1–4.

[26] Ultralytics, "ultralytics/yolov5: v7.0 - YOLOv5 SOTA Realtime Instance Segmentation," https://github.com/ultralytics/yolov5.com, 2022, accessed: 7th May, 2023.

[27] A. Vats and D. C. Anastasiu, "Enhancing retail checkout through video inpainting, yolov8 detection, and deepsort tracking," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5529–5536.

[28] Z. Wan, J. Shen, J. Chuang, X. Xia, J. Garcia, J. Ma, and Q. A. Chen, "Too afraid to drive: systematic discovery of semantic dos vulnerability in autonomous driving planning under physical-world attacks," in Proceedings of Network and Distributed Systems Security (NDSS) Symposium, 2022.

[29] D. Wang, T. Jiang, J. Sun, W. Zhou, Z. Gong, X. Zhang, W. Yao, and X. Chen, "Fca: Learning a 3d full-coverage vehicle camouflage for multi-view physical adversarial attack," in Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, no. 2, 2022, pp. 2414–2422.

[30] 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, 2023, pp. 4412–4423.

[31] T. Wu, X. Ning, W. Li, R. Huang, H. Yang, and Y. Wang, "Physical adversarial attack on vehicle detector in the carla simulator," 2020.

[32] Z. Wu, S.-N. Lim, L. S. Davis, and T. Goldstein, "Making an invisibility cloak: Real world adversarial attacks on object detectors," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16. Springer, 2020, pp. 1–17.

[33] C. Xiang and P. Mittal, "Detectorguard: Provably securing object detectors against localized patch hiding attacks," in Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 3177–3196.

[34] K. Xu, G. Zhang, S. Liu, Q. Fan, M. Sun, H. Chen, P.-Y. Chen, Y. Wang, and X. Lin, "Adversarial t-shirt! evading person detectors in a physical world," in Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. Springer, 2020, pp. 665–681.

[35] K. S. Zhang, C. Chen, and A. Xiong, "Human drivers' situation awareness of autonomous driving under physical-world attacks," in Proceedings Inaugural International Symposium on Vehicle Security & Privacy, 2023.

[36] Y. Zhang, H. Foroosh, P. David, and B. Gong, "Camou: Learning physical vehicle camouflages to adversarially attack detectors in the wild," in International Conference on Learning Representations, 2018.

[37] 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 Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security, 2019, pp. 1989–2004.

[38] A. Zolfi, M. Kravchik, Y. Elovici, and A. Shabtai, "The translucent patch: A physical and universal attack on object detectors," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 15 232–15 241.