Exploration of Machine Learning Attacks in Automotive Systems Using Physical and Mixed Reality Platforms

Venkata Sai Gireesh Chamarthi, Xiangru Chen, Bhagawat Baanav Yedla Ravi, and Sandip Ray
Department of ECE, University of Florida, Gainesville, FL 32611, USA.
vchamarthi@ufl.edu, cxr1994816@ufl.edu, b.yedlaravi@ufl.edu, sandip@ece.ufl.edu.

2023 IEEE International Conference on Consumer Electronics (ICCE) | 978-1-6654-9130-3/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCE56470.2023.10043491

***Abstract*—Adversarial attacks on Deep Neural Networks represent a critical challenge in the adoption of DNNs in critical applications. However, — and in spite of its great need, — there is significant mystery surrounding attacks on DNNs. One reason for this is the lack of a platform that enables users to get a hands-on, intuitive understanding of the attacks. In this paper, we address this problem by designing an extensible, configurable exploration platform for studying various attacks on DNNs. Our platform specifically focuses on DNNs deployed in Computer Vision modules of automotive systems. Using the platform, the user can perform various adversarial machine learning attacks, such as evasion attacks and image-perturbation attacks, and comprehend their adversarial effects on autonomous vehicles. The platform can be used to plug and play with various neural network models developed for Traffic Sign Recognition systems in autonomous vehicles. The infrastructure includes both physical and mixed-reality variants, and we demonstrate the usage of the platform on two traffic sign recognition models with different adversarial attacks.**

I. Introduction

In recent years, there have been significant advancements and proliferation in Deep Neural Network (DNN) applications. One critical application is image classification, which entails assigning classification labels to an images. DNNs have been successfully used for classifying images from a variety of domains including handwritten digits and characters, 3D toys, human and animal faces, etc. It is also a critical technology for autonomous driving technology, which depends on Computer Vision systems typically realized through DNNs for perception of environment [1]. Unfortunately, recent studies have shown that DNNs in such applications can be easily fooled (*e.g.*, by providing perturbed images). Obviously, adoption of DNNs in such critical applications depend on our ability to comprehend, detect, and mitigate such attacks. Unfortunately, — and in spite of these demonstrations, — adversarial attacks on DNNs are not well-understood, particularly outside researchers and practitioneers with deep expertise in the underlying ML principles. One reason for the mystique around ML attacks is the lack of a platform that enables users play with such attacks. Note that hands-on exploration is particularly relevant to security exploration: appreciation of security challenges and solutions can be effectively attained by actually learning to hack a system.

In this paper, we present a platform that helps users to explore the spectrum of adversarial attacks and image-perturbation attacks and effects in a realistic miniaturized autonomous driving environment. The platform permits exploration of attacks in both physical and mixed-reality modes. It allows users to understand the impact of attacks through interaction with the traffic signs, user dashboards featured with real-time feedback, and structured usage guidance. The user can plug in pre-trained Traffic Sign Recognition machine learning models to examine various types of adversarial attacks.

In summary, this paper makes the following contributions:

- An exploration platform that helps adversarial researchers, attackers, automotive industry experts, and others to explore and understand such adversarial attacks and their effects on modern vehicle computer vision application in a better way.
- The platform is built to explore and play with adversarial attacks on Traffic Sign Recognition models with physical and virtual road signs using Microsoft HoloLens 2 device on the miniature platform.

The remainder of this paper is structured as follows: Section II discusses related work that is closely related to adversarial attacks on DNNs and relevant exploration platforms. Section III describes the physical platform architecture, including the various modes and configurations, as well as the limitations and outcomes. Section IV explains mixed reality mode, and the architecture and use cases of mixed reality platform. Section V concludes the paper. The user interested in getting a feel of the platform described here is encouraged to also watch a video [2] that showcases many interesting features of the platform.

II. Related Work

There has been significant interest in adversarial attacks on DNNs in recent years. Akhtar *et al*. [3] present at comprehensive survey on threats of adversarial attacks on deep learning in computer vision. They emphasized the possibility of malicious attacks in real-world conditions. Ekkjholt *et al.* [4] performed a *Road Sign attack* by designing robust adversarial perturbations. Goodfellow *et al.* [5] explain the Fast Gradient Sign Method (FGSM) and related adversarial attacks very well. Szegedy *et al.* [6] demonstrated attacks through subtle perturbations in the images. Moosavi-Dezfooli *et al.* [7] proposed the Deepfool attack, an adversarial sample

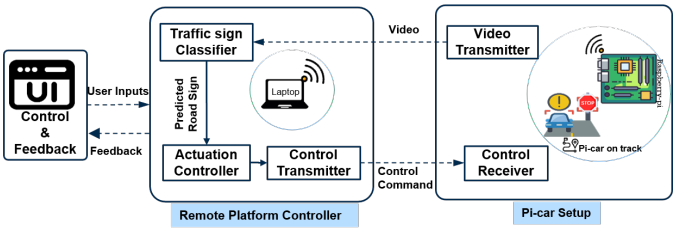

Fig. 1: Physical platform architecture

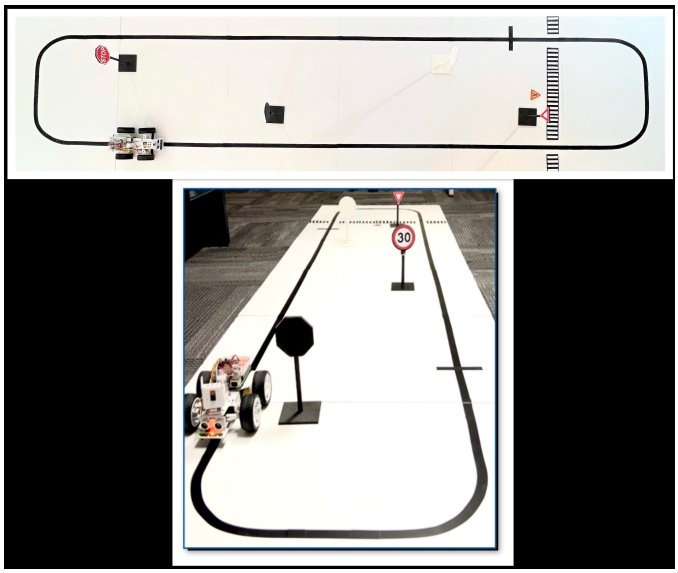

Fig. 2: Physical exploration platform - top view and side view

generation technique that minimizes the euclidean distance between perturbed and actual data. This attack effectively produces adversarial examples with fewer perturbations and higher misclassification rates.

There has also been work on developing simulators to explore various aspects of driving functionality. SUMO [8] is a multi-modal, open source, microscopic traffic simulator. CARLA [9] focuses on the development, training, and validation of self-driving systems. Yang *et al.* [10] proposed a digital twin prototype to carry out multi-vehicle experiment when availability of real vehicles is insufficient. The VIVE platform [11] is another digital twin that explores interaction among various hardware components in automotive use cases.

There has been relatively less work on platforms for hands-on exploration of automotive security. One exception is AUTOHAL, which enables exploration of ranging sensor attacks [12]. However this platform was custom designed for that specific class of attacks and did not offer configurability and extensibility.

III. Physical Exploration Platform

The high-level architecture of the physical exploration platform is shown in fig. 1. It includes (1) *Graphical User Interface* to input Traffic Sign Recognition (TSR) models, select various attack exploration scenario options, and obtain feedback 3; (2) *The Remote Platform Controller* (RPC) that handles the video feed processing that received from the pi-car, traffic sign prediction, actuation control transmission tasks, and handles the coordination between the user interaction and the physical environment; (3) *The Pi-car Setup* that has a self-driving raspberry pi car and physical traffic signs as shown in fig. 2. The pi-car transmits the video feed captured by the raspberry pi camera to the RPC and then receives the actuation controls like stop, yield, 30-speed limit, and 60-speed limit based on the classification results. The actuation control settings for the 30-speed limit and 60-speed limit correspond to the 10% and 25% duty cycle PWM pulse or the dc motors of pi-car. *Remote Platform controller* and *pi-car setup* tasks run dependently on different threads built using python modules.

A. Benign scenario exploration

Fig. 4 shows the benign scenario exploration steps. The user places the physical traffic signs on the platform. The user starts the platform by launching the GUI as shown in fig. 3, and selecting the Traffic Sign Recognition model, which is pre-trained on a particular traffic sign dataset. The

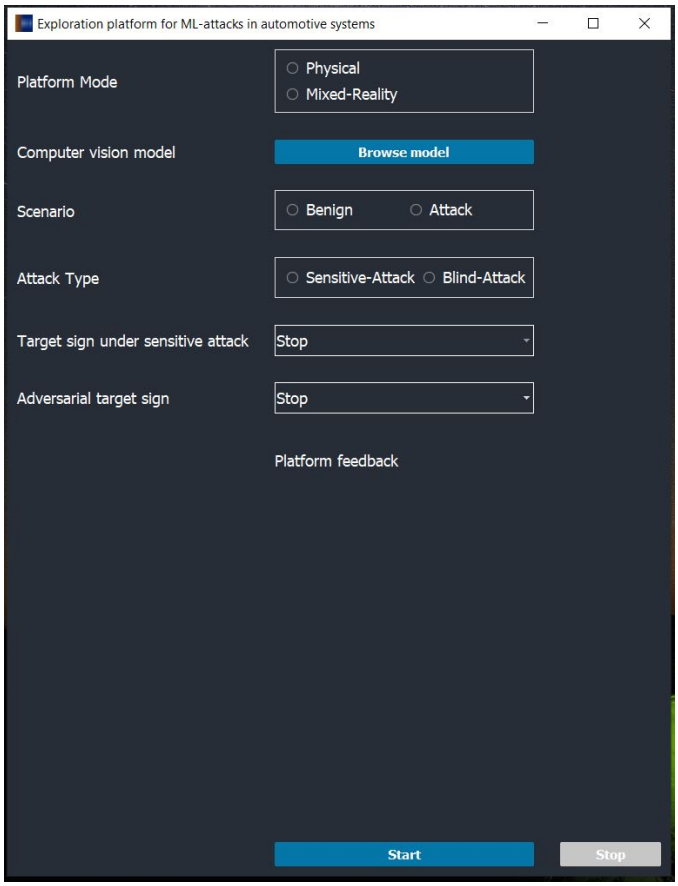

Fig. 3: The Graphical User Interface

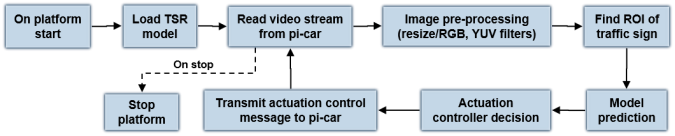

Fig. 4: Benign scenario flow diagram in physical exploration mode

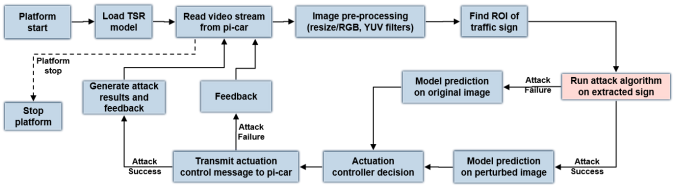

Fig. 5: Attack scenario flow diagram in physical exploration mode

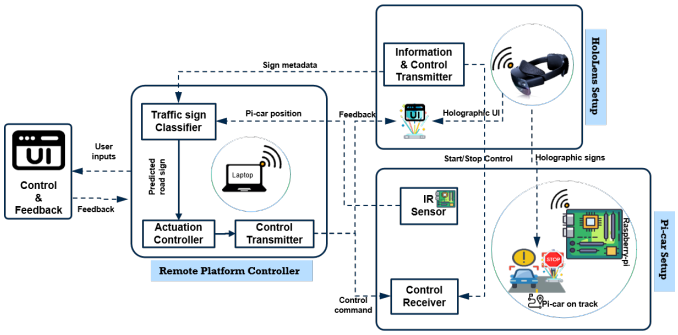

Fig. 6: Mixed reality platform architecture

Remote Platform controller and *pi-car setup* connection need to be established based on the client-server mechanism for exchanging the video and actuation controls. Following the connection, the RPC decodes the video stream and displays it to the user in the window. Before running the predictions, the classifier unit pre-processes the video frames into images. The TSR model is then fed this image which returns the type of sign and the prediction confidence. The actuation controller sends the message back to the pi-car based on the predicted sign. Pi-car responds on the track to the received actuation command from the RPC. The feedback module of the GUI displays benign prediction results. The RPC also displays the pi-car's live video to the user for close monitoring of the exploration scenario.

B. Attack scenario exploration

The attack scenario is shown in Fig. 5. The victim (pre-processed) image is perturbed to generate an adversarial image. This image is fed to the TSR model instead of the original image. In the physical exploration attack model, we used the *Fast Gradient Sign Method* (FGSM) to generate the adversarial images. We follow GoodFellow [5] as a one-step method to generate targeted adversarial examples using equation III-B below. Here $\mathbf{x}$ is the original input image, $\mathbf{t}$ is the target class label, θ denotes the model parameters, and $J(\theta, \mathbf{x}, t)$ is the loss function of the neural network. The gradient of the loss corresponds to the input data is $\nabla_x J(\theta, \mathbf{x}, t)$. The attack modifies the input data by a small step ϵ in the direction (i.e. $sign(\nabla_x J(\theta, \mathbf{x}, t))$ that maximizes the loss. The resulting perturbed image is $\mathbf{x^{adv}}$

$$\mathbf{x^{adv}} = \mathbf{x} - \epsilon.\mathbf{sign}(\nabla_{\mathbf{x}} \mathbf{J}(\theta, \mathbf{x}, \mathbf{t}))$$

We have developed two attack scenarios to trick the Traffic Sign Recognition model using FGSM:

- **Sensitive Attack** is only performed on one traffic sign. The user can mark one of the traffic signs as attack victim and selects the target class for it to be misclassified to, (*e.g.*, a stop sign to misclassified as a speed limit sign). The attack algorithm only tries to generate a corresponding adversarial image for the victim and feed it to the TSR model.
- **Blind Attack** attempts to attack all the traffic signs on the platform. The user only selects the target misclassification class (*e.g.*, all signs to be misclassified as stop sign). The attack algorithm generates an adversarial image to feed the TSR model whenever a traffic sign is encountered. Correspondingly, the user will notice the pi-car taking the actuation control as if the sign was the misclassification target instead of the original.

C. Limitations of the physical exploration platform

The physical exploration platform allows the user to place traffic signs on the track in real-time. The user can adjust the viewing angles and distance of the signs in front of the pi-car. However, physical limitations make it difficult to perform experiments with a wide range of traffic sign variants. Note that DNNs can be easily deceived by simply pasting a sticker or changing a sign's color or shape. In the exploration platforms, creating a wide range of 3d printed Image-Perturbed signs is infeasible.

IV. Mixed Reality Exploration Platform

We overcome the physical limitations above through a mixed-reality variant of the platform developed using Microsoft Hololens 2 device. In this variant, instead of physical signs we use holographic signs generated by the HoloLens device. The architecture shown in Fig. 6 has physical and holographic components. The elimination of physical traffic signs eliminates the need for the video capturing and processing step. Instead, our mixed reality platform architecture mimics the experience of the user playing with the actual signs and exploring the platform with much more variants of signs. Fig. 7 shows the holographic signs which can be introduced on the vehicle path by the user using Hololens actions. In this mode, the adversarial images are generated using the *Image-Perturbation* method and *Deepfool attack* algorithm [7] on the Squeezenet neural network model architecture.

Fig. 9 describes the steps for a user to explore a mixed reality platform. The user starts the platform on the RPC by launching the GUI and feeding the Traffic Sign Recognition

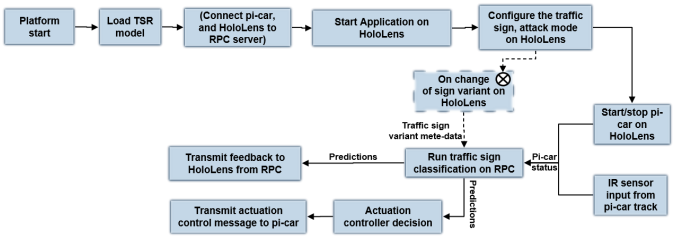

Fig. 9: Mixed reality exploration platform usage flow

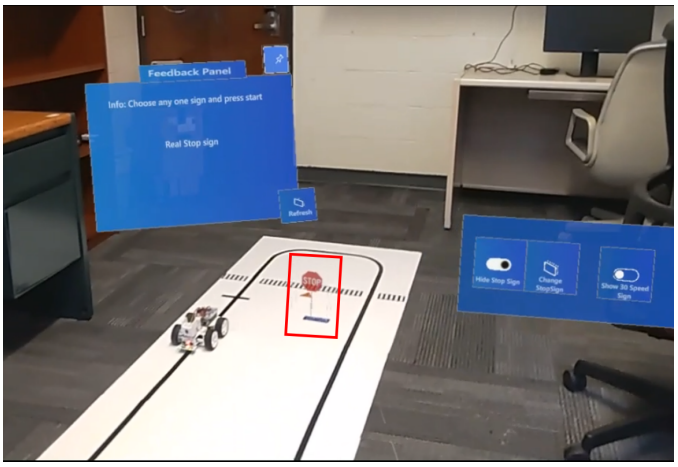

Fig. 7: Holographic stop sign placed on the platform

Fig. 8: Variants of stop, yield, 30-speed limit, and 60-speed limit signs used in the mixed reality platform

model. The user then connects the pi-car by launching the client script and deploys the HoloLens by wearing the headset and launching the application. After pairing, the user interacts with the virtual GUI, which includes the ability to show, hide, or change a specific traffic sign by pressing a button. Users can change the sign variants either from the image-perturbation category or deepfool attack category and place them on the track in the desired position. When the pi-car passes through the IR sensor module, the RPC begins processing the traffic sign variant using the input TSR model and generates results. The results of these predictions will be sent to the pi-car and the hololens. The pi-car will take the actuation control as directed by the RPC. The pi-car can be stopped at any time through the HoloLens dashboard.

V. Conclusions

We have developed to our knowledge the first platform for users to do hands-on exploration of adversarial machine learning. Our platform focuses on attacks on vision systems particularly targeting autonomous vehicles. It uses both physical and mixed reality to explore various adversarial attacks on Traffic Sign Recognition neural network models. We demonstrated our platform in both modes for a variety of attacks causing mis-detection of traffic signs. The platform is adaptable in supporting custom TSR models trained on custom datasets for exploration through both physical or mixed reality modes.

In future work we will implement more adversarial attacks on the platform, and explore extension of the platform for other target domains of adversarial ML.

Acknowledgements: This project has been partially supported by the National Science Foundation under Grants CNS-1908549 and SATC-2221900.

References

[1] D. Shin, H.-g. Kim, K.-m. Park, and K. Yi, "Development of deep learning based human-centered threat assessment for application to automated driving vehicle," *Applied Sciences*, vol. 10, no. 1, 2020. [Online]. Available: https://www.mdpi.com/2076-3417/10/1/253

[2] V. S. G. Chamarthi, "Exploration of machine learning attacks in automotive systems using physical, mixedreality platforms," YouTube, urlhttps://youtu.be/UulfOM4oONI, Last accessed on 2022-11-14.

[3] N. Akhtar and A. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey," *IEEE Access*, vol. 6, pp. 14 410–14 430, 2018.

[4] 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.

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, See URL https://arxiv.org/abs/1412.6572.

[6] I. V. Pustokhina, D. A. Pustokhin, D. Gupta, A. Khanna, K. Shankar, and G. N. Nguyen, "An effective training scheme for deep neural network in edge computing enabled internet of medical things (iomt) systems," *IEEE Access*, vol. 8, pp. 107 112–107 123, 2020.

[7] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: A simple and accurate method to fool deep neural networks," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Los Alamitos, CA, USA: IEEE Computer Society, jun 2016, pp. 2574–2582. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/CVPR.2016.282

[8] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[9] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[10] C. Yang, J. Dong, Q. Xu, M. Cai, H. Qin, J. Wang, and K. Li, "Multi-vehicle experiment platform: A digital twin realization method," in *2022 IEEE/SICE International Symposium on System Integration (SII)*, 2022, pp. 705–711.

[11] M. R. Kabir, N. Mishra, and S. Ray, "Vive: Virtualization of vehicular electronics for system-level exploration," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3307–3312.

[12] B. B. Y. Ravi, M. R. Kabir, N. Mishra, S. Boddupalli, and S. Ray, "Autohal: An exploration platform for ranging sensor attacks on automotive systems," in *2022 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 2022, pp. 1–2.