Avara: A Uniform Evaluation System for Perceptibility Analysis Against Adversarial Object Evasion Attacks

Xinyao Ma*
Indiana University Bloomington
Bloomington, IN, USA
maxiny@iu.edu

Chaoqi Zhang*
Indiana University Bloomington
Bloomington, IN, USA
cz42@iu.edu

Huadi Zhu*
The University of Texas at Arlington
Arlington, TX, USA
huadi.zhu@mavs.uta.edu

L. Jean Camp
Indiana University Bloomington
Bloomington, IN, USA
ljcamp@indiana.edu

Ming Li
The University of Texas at Arlington
Arlington, TX, USA
ming.li@uta.edu

Xiaojing Liao†
Indiana University Bloomington
Bloomington, IN, USA
xliao@indiana.edu

Abstract

Thanks to recent advances in machine learning (ML) techniques, Autonomous Driving (AD) has seen significant breakthroughs with enhanced capabilities. However, the susceptibility of ML models to adversarial evasion attacks poses a critical threat, undermining the reliability of autonomous driving systems. Despite efforts by researchers to mitigate these attacks within the AD context, unfortunately, a significant gap persists in fully understanding such adversarial maneuvers, particularly from a driver's perspective.

To bridge this gap, we propose *Avara*, the first unified evaluation platform for assessing human drivers' perceptibility to adversarial attacks in AD contexts. Leveraging Virtual Reality (VR) and eye-tracking technology, *Avara* captures multi-modal driver awareness data, enabling detailed assessments of driver perception. Our approach integrates three distinct sources of multi-modal awareness evaluation metrics, addressing gaps inherent in previous evaluation strategies. The effectiveness and usability of *Avara* were validated through a human subject study, where participants engaged actively with the platform and provided extensive feedback on their perception and response to adversarial evasion attacks. Utilizing *Avara*, we identify an intriguing discovery that the current imperceptibility metrics for adversarial attacks fail to accurately reflect the autonomous vehicle driver's perceptibility.

CCS Concepts

• **Security and privacy → Usability in security and privacy**; • **Human-centered computing → User studies**.

Keywords

Adversarial Attack; Autonomous Driving; Human Perception

*These authors contributed equally to this work.
†Corresponding author

CCS '24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-0636-3/24/10
https://doi.org/10.1145/3658644.3670291

ACM Reference Format:
Xinyao Ma, Chaoqi Zhang, Huadi Zhu, L. Jean Camp, Ming Li, and Xiaojing Liao. 2024. *Avara*: A Uniform Evaluation System for Perceptibility Analysis Against Adversarial Object Evasion Attacks. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3658644.3670291

1 Introduction

Recent advances in ML techniques have spurred significant breakthroughs in AD, enhancing capabilities such as navigation, real-time decision-making, and environment perception. However, a series of research has highlighted a critical vulnerability: when exposed to adversarial evasion attacks, these powerful ML models used in AD can be fooled and misled. For instance, previous research has shown how physically altering a STOP sign by attaching malicious patches can lead to significant vulnerabilities. These intentional modifications were successful in consistently causing AD perception systems, especially object detection models, to misclassify the STOP sign. Such issues carry profound security implications posed by adversarial manipulations in compromising the reliability of autonomous driving systems.

Security and AD researchers and practitioners are now facing a wide array of adversarial evasion attack methods. However, there is a noticeable gap in the comprehensive understanding of these attacks, especially regarding their strengths and limitations from a driver's perspective. AD systems, especially those operating at SAE Level 3 automation [3], function as human-in-the-loop systems. In such setups, human drivers are expected to take over control from the autonomous system in cases of failure or imminent failure. This human intervention aspect underscores the need for a qualitative understanding of adversarial evasion attacks, focusing on how they impact driver perception, decision-making, and overall interaction with the vehicle's autonomous features. We argue that to further advance the research on adversarial evasion attack in AD context, it is critical to develop a uniform evaluation platform to support perceptibility analysis drivers of adversarial attacks.

The traditional evaluation mechanism typically combines video-based adversarial example assessment followed by user surveys. This method involves recording participants' actions in video-based driving environments and supplementing these observations with self-reported data from post-experiment surveys. However, this

Figure 1: Participant during experiment and the corresponding driving scenario.

mechanism has several limitations. First, driver's perceptibility analysis from user surveys depends heavily on the participants' self-assessment, which can be subjective and influenced by personal biases, memory recall, and interpretation of the questions. Also, while video-based methods capture overt drivers' operation behaviors, they often miss out on capturing subtle and nuanced details of participant awareness and cognitive processes, which are critical in understanding drivers' perceptibility on adversarial evasion attacks. Furthermore, traditional methods may not provide a fully immersive experience that accurately replicates real-world driving conditions. This lack of immersion can affect the authenticity of participants' responses. These limitations highlight the need for more refined and immersive evaluation methods that can capture a comprehensive and accurate portrayal of drivers' perceptibility to adversarial evasion attacks in AD contexts.

***Avara*: a uniform platform for perceptibility analysis.** In our study, we design, develop, and release the first unified evaluation platform, *Avara*, for driver's perceptibility analysis against adversarial evasion attacks in the AD context. In particular, *Avara* is designed to gather detailed driver's perceptibility data to ensure a refined understanding of how drivers perceive and react to adversarial evasion attacks. By leveraging VR technology and an integrated eye tracker, *Avara* captures multi-modal driver awareness data. This includes eye-movement data and various driver operation metrics, facilitating a fine-grained assessment of driver awareness. Additionally, utilizing VR technology, *Avara* simulates a highly realistic driving environment. This immersive setup is crucial for eliciting authentic driver behaviors and responses, mirroring those that would occur in real-world driving scenarios. The current implementation of *Avara* incorporates four different types of adversarial evasion attacks, a total of seven adversarial image settings, along with four driver's perceptibility metrics. *Avara* enables security and AD researchers to (1) assess the driver's perceptibility of various adversarial attacks, (2) explore the correlations between driver's awareness and different adversarial attack parameter settings, (3) conduct comparative studies on the effectiveness of different adversarial attacks.

The effectiveness and usability of *Avara* have been evaluated by a human subject study involving 60 participants. The results confirmed that the data captured by *Avara*, including eye-movement metrics and driver operation behaviors, were accurate and relevant. This data provided a comprehensive understanding of how participants perceived and responded to various adversarial evasion attacks. Also, the feedback from participants indicated that participants were effectively engaged with the *Avara* platform due to the immersive VR environment.

Measurement and findings. Utilizing *Avara*, we conducted the first empirical study on driver's perceptibility analysis on adversarial evasion attacks. Our study undertook a cross-evaluation of seven different adversarial attack image settings. This approach allowed us to compare and contrast the impact of these various attacks on driver awareness. We found that drivers exhibited a significantly higher level of perceptual awareness, particularly in their eye-movement patterns, when encountering the *SLAP* adversarial evasion attack. This was in contrast to their responses to other attacks, such as *ShadowAttack_0.10*, *FGSM_0.0175*, and *RP2*, where the level of perceptual awareness was comparatively lower. Also interestingly, our study uncovered the correlation between perceptual awareness and the traditional adversarial attack inconspicuousness metric, Perturbation Sensitivity Distance (PSD), which is a holistic metric considering multiple factors such as the likelihood of detection and the perceptual visibility of changes. Furthermore, our study shed light on how educating drivers about the nature and characteristics of such attacks can influence their ability to detect and respond to these threats in an AD context.

Contributions. This paper makes the following contributions:

- We propose the first uniform evaluation system for human drivers' perceptibility analysis against adversarial evasion attacks targeting autonomous driving.
- We offer multi-modal awareness evaluation metrics from three distinct sources: eye movement, driver's operations, and survey response, to cross-validate findings and discover outliers, addressing gaps inherent in previous evaluation strategies.
- We conduct a comprehensive analysis of four adversarial evasion attacks, which advanced the understanding of driver perceptibility in response to adversarial evasion attacks.
- We find that the most widely-used imperceptibility metrics for adversarial attacks in AD do not adequately capture the perceptual experience of autonomous vehicle drivers.
- We release our platform and code at https://sites.google.com/view/avara-artifacts.

2 Background

2.1 Adversarial Evasion Attack

An adversarial evasion attack is a type of adversarial machine learning attack. In such attacks, an adversary creates and uses adversarial examples – inputs to a machine learning model that are deliberately designed to cause the model to make erroneous decisions. These inputs are typically crafted by introducing small, carefully calculated perturbations to legitimate data, leading the model to misclassify or misinterpret the data.

Physical adversarial object evasion attacks in AD context. Adversarial evasion attacks pose a critical challenge in the safety-critical domains, especially autonomous driving (AD) [6, 7, 11, 20, 54]. The perception systems in AD, which include cameras, LIDAR, and other sensors, are vital for gathering environmental data and information about surrounding objects. This data is security and safety-critical for navigation and decision-making in AD.

In the context of camera-based AD, adversarial evasion attacks have been proved feasible in manipulating input object data (e.g., STOP sign, pedestrian) to deceive deep neural networks (DNN)-based AD perception, such as DNN object detection models YOLO [22,

41] and Fast R-CNN [2, 27]. Among them, physical adversarial object evasion attacks require the manipulation to be *physically* added to the object itself [15, 24]. For example, prior research [47, 59] demonstrated the physical alteration of a STOP sign by attaching malicious patches. These modifications were able to consistently trigger misclassifications in object detection models, affecting their accuracy from various viewpoint angles and distances. These types of attacks are particularly critical due to their potential to cause severe accidents and endanger human lives.

In our study, we assess four physical adversarial object evasion attacks: Robust Physical Perturbations (*RP2*) [17], *ShadowAttack* [62] with three different parameter settings, Short-Lived Adversarial Perturbations (*SLAP*) [31], *FTE* [21], and a non-physical adversarial attack: Fast Gradient Sign Method (*FGSM*) [18] with two different parameter settings.

Inconspicuousness of adversarial evasion attacks. Inconspicuousness (or imperceptibility) is one of the critical metrics of adversarial evasion attacks. It measures the ability of adversarial attacks to go undetected or escape the notice of human observers. The more inconspicuous an attack, the higher its likelihood of bypassing security measures unnoticed, thereby increasing both its effectiveness and threat level in real-world situations. This is the same case for physical adversarial object evasion attacks in AD. The physical manipulations must be subtle so as not to be easily noticed by drivers. The ability to blend these manipulations seamlessly into the normal operating landscape is key to their success.

Previous research has utilized metrics such as L_p distortion, Average Structural Similarity (ASS), and Perturbation Sensitivity Distance (PSD) [34] to quantitatively measure the inconspicuousness of an adversarial example. L_p distortion, often denoted as L_1 or L_2 depending on the specific norm used, quantifies the magnitude of the perturbation added to the original image. This metric assesses how much the adversarial image deviates from the original in terms of pixel values. Average Structural Similarity (ASS) is used to measure the perceptual difference between the original and the adversarial image. It evaluates how changes due to the adversarial perturbation affect the image's structural information, illuminance, and contrast. Perturbation Sensitivity Distance (PSD) assesses the sensitivity of an image to perturbations in a more holistic manner, considering factors such as the likelihood of detection and the perceptual visibility of changes.

In our study, we established a unified evaluation system to *qualitatively* evaluate the inconspicuousness of adversarial examples from the perspective of a driver within an immersive AD environment. This system reveals the relationship between driver awareness and established quantitative measures of inconspicuousness, thereby providing a realistic measure of the adversarial evasion attack's stealthiness and potential impact in real-world AD scenarios.

2.2 Driver Situation Awareness Assessment

In the AD environment, vehicles are equipped with AI-driven systems for navigation and decision-making. It transforms the role of the human driver shifts from direct control to supervision and intervention. This shift underscores the importance of maintaining a high level of situational awareness. Driver situation awareness is crucial in vehicles equipped with SAE Level 3 automation due to the unique interplay between human oversight and automated control. In these systems, the vehicle autonomously manages most driving tasks, while drivers remain prepared to reassume control as necessary. Therefore, the effectiveness of physical adversarial evasion attacks should be comprehensively evaluated not only on the vehicle's perception system but also on driver awareness. Additionally, investigating driver situation awareness sheds light on the vital contribution humans can make in enhancing the safety of autonomous vehicles, particularly in the face of physical adversarial evasion attacks.

In this study, we consider the following two levels of driver awareness regarding the physical adversarial evasion attack:

- *Perceptual awareness*, which reflects the driver's ability to accurately perceive and interpret adversarial traffic signs in the virtual driving environment.
- *Responsive awareness*, which evaluates the driver's capacity to react promptly and effectively when necessary to resume control from the AD system.

Various measures are applied to thoroughly assess both perceptual and responsive awareness, including eye tracking data (e.g., fixation and saccade events, location-based eye points), drivers' response times (e.g., the time it takes for the driver to take over control), and the appropriateness of their responses, as elaborated in Section 4.2.

2.3 Problem Scope

We aim to assess driver situation awareness of the adversarial evasion attack against AD perception systems. Our primary focus has been on a specific subclass of attacks, known as adversarial STOP sign attacks. It subtly alters the physical appearance of STOP signs in a way that causes AD systems to misinterpret or fail to recognize them. Such attacks have been identified as the most prevalent and potentially hazardous against the functionality and safety of AD systems. Nevertheless, we acknowledge the wide spectrum of adversarial evasion attacks against AD perception systems, extending beyond STOP sign alternations. These scenarios include attacks involving pedestrians, for example, an infrared laser reflection on traffic signs that are not even visible to human [43], and 'Adversarial T-shirts' attack on evading human detector due to a moving person's pose changes [55], or interference with LiDAR systems [7]. The adaptability of our system to different forms of adversarial evasion attacks will be discussed in Section 5.

Note that the evaluation of driver situation awareness in other hazardous driving scenarios, such as poor weather conditions and careless pedestrians, falls outside the scope of this paper.

3 *Avara*: Design and Implementation

3.1 Platform Overview

Goals and design. We construct *Avara*, a VR-based driver awareness evaluation platform for adversarial evasion attacks against AD systems. The design goals of *Avara* are summarized in the following aspects.

- *Unified evaluation framework.* We aim to design *Avara* into a comprehensive evaluation framework that can be applied across

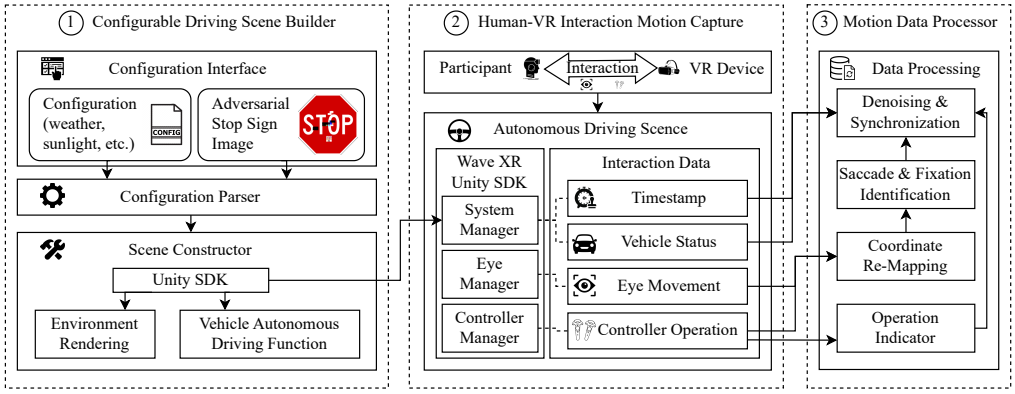


Figure 2: System design of *Avara*.

different types of adversarial evasion attacks. The developed assessment methods and technologies should be universally applicable to these attacks.

• *Immersive driving experience. Avara* is designed to create an immersive driving experience in a virtual environment, mimicking real-world scenarios. We utilize VR technology to simulate a highly realistic driving environment, where adversarial evasion attacks are implemented too.

• *Fine-grained awareness assessment. Avara* is tailored to capture detailed data on both perceptual and responsive aspects of driver awareness to ensure a refined understanding of how drivers perceive and react to adversarial evasion attacks. In our implementation, we leverage VR and the integrated eye tracker to capture multi-modal driver awareness data, including eye-movement data and driver's operation data, to provide fine-grained awareness assessment.

• *Affordability and practicality.* Affordability is a key consideration in the development of *Avara* to facilitate broader adoption and application in adversarial evasion attack research and practical contexts. The implementation makes use of VR and eye-tracking technologies to perform near-realistic driver situation awareness assessments in a cost-effective manner. *Avara*'s affordability (<$2k) significantly exceeds existing simulators like miniSim ($250k-$500k) [32], especially considering its main target users to be researchers in security/autonomous driving communities.

Architecture. Figure 2 illustrates the architecture of *Avara*, which consists of the following three components: (1) *Configurable Driving Scene Builder*: It is responsible for creating immersive driving scenarios within a VR environment, based on a variety of configurable scene factors. These factors include elements such as the type of adversarial STOP sign, the direction of sunlight, and weather conditions, enabling the simulation of diverse and realistic driving experiences. (2) *Human-VR Interaction Motion Capturer*: It captures the interactive motion data between the human driver and the VR environment, including eye tracking data and driver's operation data. (3) *Motion Data Processor*: It gathers multi-modal motion interaction data, translates the 3D object coordinates to 2D screen coordinates, and maps them into the measures of perceptual awareness and responsive awareness. The design and implementation of each component will be elaborated in the following subsections.

3.2 Configurable Driving Scene Builder

The Configurable Driving Scene Builder aims to craft varied driving scenarios tailored to user specifications. It consists of three major components, namely a configuration interface, a configuration parser, and a scene constructor. To start, users freely indicate the configuration of system parameters, including driving conditions such as weather and sunlight, as well as target adversarial STOP signs for evaluation purposes. For convenience, the configuration interface allows two types of inputs: local manual configuration with a UI in the VR headset and remote configuration file uploads from personal computers. For the latter, an API is provided in a client application for users to transmit configuration files specifying system parameters to the VR headset storage. Upon activation, the configuration parser retrieves and processes system parameters.

Then, the scene constructor builds the driving scene based on these configurations. Among them, the adversarial STOP sign image is applied to the texture of the surface of the target STOP sign object for display. To render different environmental weather and time of day, we employ a free open-source asset [51] from the Unity Asset Store [52]. This asset facilitates both the modulation of weather conditions and the rendering of pertinent prefabs. To accommodate other driving conditions, we leveraged Unity [50] and crafted 293 lines of C# code to enable their simulation. Notably, traffic signs, roads, and vehicles are proportioned to match their real-world physical sizes in the VR scene.

The default driving scene consists of the user's vehicle, three intersections of focus, and surrounding environments. Normal, dirty, and adversarial STOP signs are placed before these intersections. The default scene simulates a suburban area where all roads are two-way, four-lane, and all intersections are four-way stop-controlled. Users are allowed to switch between manual and autonomous driving modes freely and operate the vehicle until reaching the destination. In the autonomous mode, the vehicle will move forward constantly after reaching the default speed of 13.4 m/s (30 mph), and take over the control before encountering each normal or dirty STOP sign (adversarial STOP signs are ignored); the default acceleration and deceleration rates are 3 m/s^2 and 10 m/s^2, respectively, to simulate the real-world scenario.

To enable the vehicle's autonomous driving feature, another essential function of the driving scene, we adopt a road-centered

strategy, where the vehicle's real-time position and movement are anchored based on its relative position with the predefined roads, and real-world decision-making is simulated by scripts. In addition, in our driving scene design, the presence of other cars and pedestrians is excluded to mitigate potential ethical concerns regarding hypothetical car accidents. This is a common set-up in existing AD adversarial STOP sign evaluation platforms [7, 54, 59]. To this end, 320 planes with physical collision control were created in Unity to establish the roads, and another 689 lines of C# code were crafted to accurately and precisely guide the vehicle's autonomous driving along these roads.

3.3 Human-VR Interaction Motion Capturer

The Human-VR Interaction Motion Capturer is designed to record and capture human interaction data within VR. Those data will be used to assess perceptual and responsive awareness in the component of *Motion Data Processor*. In our study, we focus on collecting eye-tracking data, driver's operation data (i.e., braking, accelerating, steering, taking-over control), and vehicle/environment status (e.g., position, speed). *Avara* implements a coroutine function at a sampling rate of 90Hz to gather those interaction data for analysis.

Specifically, we employ the HTC VIVE Focus 3 Eye Tracker to obtain the tracking data of left, right, and combined eyes via the VIVE Wave XR Plugin [13] in Wave Unity SDK [14]. We chose this eye-tracker over the other band due to its precision and affordability. HTC VIVE Focus 3 Eye Tracker outputs raw eye-tracking data (i.e., gaze origin and direction) and does not provide any data processing tools. Hence, we have implemented a data processor to analyze the collected raw eye-tracking data, as detailed in Section 3.4.

Regarding the driver's operation data (i.e., braking, accelerating, steering, and taking over control), we use the VR controller as a key instrument for data recording. In particular, we design a series of VR controller operations for each driver's operation (i.e., braking, accelerating, steering, and taking over control) and then retrieve these controller operations via the Wave Controller package in Wave Unity SDK. In our study, we avoid using sophisticated or specialized components such as steering wheels for vehicle control, in consideration of system affordability and accessibility. By using the VR controller, which typically comes with the VR headset, we ensure that our setup remains economically feasible without compromising the quality and reliability of the data collected.

Avara also records other environmental data, including the vehicle's position, speed, trajectory, and proximity to STOP signs. Moreover, it tracks the relative positions of the STOP sign's four corners in relation to the car, which is essential for accurately mapping the participants' eye gaze direction towards the road or the traffic signs in the environment.

3.4 Motion Data Processor

The primary goal of the Motion Data Processor is to process the driver interaction data from the *Human-VR Interaction Motion Capturer* to facilitate driver situation awareness analysis.

Eye-tracking data processing. In our study, *Avara* utilizes raw eye-tracking data (i.e., gaze origin and gaze direction) to extract gaze features, i.e., saccade and fixation, for further analysis. Saccades are rapid eye movements between points of fixation, and their identification is crucial for understanding how quickly and where the participant shifts their gaze in response to stimuli in the VR environment. Fixations, where the gaze is held steadily on a particular point, are used to determine the points of interest or attention for the participant.

Specifically, the eye-tracking data processing consists of two steps: VR coordinate remapping and saccade/fixation identification. Different from desktop-based eye trackers which are limited to tracking eye movements in relation to a fixed screen, the immersive nature of VR necessitates a more dynamic method of eye-tracking. This method involves remapping the endpoints of eye movements to a specific target within the 3D space, such as a STOP sign in our simulated driving scenario. Algorithm 1 presents the method to convert the eye's gaze data, captured as relative to the VR headset, into absolute positions within the 3D VR space. The process includes calculating the eye's gaze direction and point of focus based on the VR headset's orientation and position (line 1), computing the eye's gaze direction, factoring in both the gaze data and the headset's orientation (line 2), and then extending the line of sight from the eyes and determining where it intersects with objects in the VR space, such as STOP signs or other traffic elements (line 4).

Following the coordinate remapping, the next step is to preprocess remapped eye-tracking data and identify saccades and fixations. To accomplish this, we employ a moving average noise reduction and the I-VT (Identification by Velocity Threshold) saccade/fixation identification [42], with modified parameters based on the optimal velocity threshold evaluated in [4]. We also applied a 'merge adjacent fixations' function recommended by Tobii I-VT manual [39] in case a few samples are wrongly classified because of noise or other disturbances; it can result in a long, continuous observation being split into many short ones with very brief eye movements or gaps between them. The I-VT algorithm is a widely recognized method in eye-tracking research, primarily due to its efficacy in distinguishing between saccades and fixations based on eye movement velocity. The detailed steps and pseudocode can be found in Appendix Algorithm 2.

Driver's operation data processing. The processing of driver's operation data focuses on understanding participant's operation decisions and timing, which are key indicators of their awareness and response strategies. Particularly, we link the timestamps with each participant's driving operations (i.e., braking, accelerating, steering, taking-over control) during the VR simulation. By analyzing these inputs, we can determine significant actions like braking in response to adversarial STOP signs or switching from autonomous to manual driving mode.

4 Evaluation

To evaluate the effectiveness and usability of *Avara*, we designed an in-lab user study. Particularly, we aim to answer the following questions: **Q1**: To what extent does this system accurately measure driver's awareness under adversarial evasion attacks? **Q2**: How usable is *Avara* in terms of user experience and overall satisfaction? **Q3**: How does the integration of VR and eye-tracking systems perform in assessing user awareness compared to traditional video-based evaluation methods?

Algorithm 1 VR Coordinate Remapping

1: **Input:** Camera position P_c, relative position from eye to camera P_{e2c}, eye direction D_e, STOP sign position $P_{ss\{lu,ld,ru,rd\}}$
2: **Output:** Intersection position between eye ray and STOP sign plane P_i, intersection point in the STOP sign area R
3: Get eye position: $P_e \leftarrow P_c + P_{e2c}$
4: Get eye ray: $R_e \leftarrow P_e + \lambda D_e$
5: Span the STOP sign plane: $\Pi \leftarrow \text{Span}(P_{ss_{lu}}, P_{ss_{ld}}, P_{ss_{ru}}, P_{ss_{rd}})$
6: Calculate the intersection Point: $P_i \leftarrow R_e \cap \Pi$
7: $ISS \leftarrow$ true if intersection in STOP sign area; else $ISS \leftarrow$ false

4.1 Experiment Design

4.1.1 Recruitment and screening. Due to the nature of our study requiring participants to be in person, we recruited participants from each participating institute via distributing recruitment advertisements in each institute's mail lists. Note that we advertised the study goal as evaluating the usability of operating an autonomous driving vehicle through a VR setting to minimize priming participants and potentially affecting their behaviors.

Prior to the experiment, each participant completed a brief screening survey [1] indicating prior experience with VR/AR, vehicle driving, and demographics. We only consider participants holding a valid US driver's license, and having no history of virtual reality sickness or color blindness. We only recruited participants with normal vision or corrected vision with contact lenses rather than glasses. This is because the latter can interfere with the integrated eye-tracker [10, 12], potentially affecting its accuracy. To this end, we recruited 89 participants for full-scale user studies with their ages from 18 to 45 for the full study and 15 participants for the pilot study. Among them, 62% have experience with VR, but only 28% have experience with AV, and nearly 50% of participants drive more than 5000 miles per year. Table 1 lists the demographic information of participants who completed the full study.

Table 1: Demographics of participants.

Item	Options	n
Gender	Male	62
	Female	26
	Non-binary	1
Age	18-25	40
	26-35	43
	36-45	6
Ethnicity	Asian	71
	Hispanic/Latino	2
	White	12
	Others	4
Education	High school graduate	2
	Some college	13
	College graduate	45
	Post-graduate degree	29
VR Related Experience	Yes	55
	No	32
AV Related Experience	Yes	25
	No	64
Average Annual Mileage	<5000	46
	5000-20000	37
	>20000	6

[1] See it on our website at: https://sites.google.com/view/avara-artifacts

4.1.2 Study procedure. The entire study procedure for each participant lasts about 30 minutes, including a ~10-minute warm-up phrase, a ~10-minute experiment phrase (~3 minutes for benchmark task, ~4 minutes for after-education task, and ~3 minutes for comparison task), and a ~10-minute post-experiment interview phase. The break time is allocated after the experiment phase for 1 minute. Each participant undergoes the procedure once. Below, we elaborate on the study procedure.

Warm-up phase. We start by having each participant read the consent form approved by IRB and printed out by the research team. Then, a researcher instructed the participants to wear the eye tracker and the adjustment.

Upon completion of the calibration process of the eye tracker, participants were instructed to familiarize themselves with the AD operations (e.g., braking, accelerating, steering, taking over control) using the pair of VR controllers. At least one researcher is present to instruct participants, help them through this process, and address any questions they may have.

Experiment phase. In the experiment phase, participants were instructed to complete three tasks: (1) a benchmark task where they used *Avara* to operate the autonomous vehicle through three intersections, each featuring a benign (clean or dirty, see Appendix Figure 14a) or an adversarial STOP sign. (2) an after-education task, where participants received education about adversarial evasion attacks and then repeated the VR experiment. (3) a comparison task, where participants were instructed to use an online video-based AD platform deployed by Qualtrics [40] to experience driving the autonomous vehicle in the same driving scenario.

Note that in both *Avara* and the video-based AD platform, participants are initially set to operate the vehicle in the AD mode. They have the option to take over control as they deem proper. Upon passing the third intersection, the simulation is terminated, and participants are instructed to quit the simulator.

Post-experiment interview phase. We conducted a survey followed by qualitative interviews with participants to gain insights into their system experience and driving reactions. The survey includes questions such as whether they noticed any unusual STOP signs, and their impressions of any modified traffic signs they encountered, as well as the usability of *Avara*. Participants can watch their experiment video record and driving replay for reference during the survey. The survey is provided in Appendix 8.

4.1.3 Attack selection. In our study, we carefully selected 4 state-of-the-art and reproducible physical adversarial evasion attacks in AD context, i.e., Robust Physical Perturbations (*RP2*) [17], *ShadowAttack* [62], Short-Lived Adversarial Perturbations (*SLAP*) [31], and *FTE* [21]. *RP2*, *ShadowAttack*, and *SLAP* are accompanied by public artifacts to generate adversarial STOP signs, and we use the same *FTE* image in the paper [54] with model YOLO v3. The selection of parameters is based on the default settings outlined in their publicly available artifact or as documented in their paper, which demonstrates a high evasion rate. For *RP2*, we adhered to the original methodology [16] proposed in their paper, employing the same dataset (LISA [37] & GTSRB [48]), the same models (LISA-CNN & GTSRB-CNN [56]), and the same default parameters provided in their public artifact. In the case of *ShadowAttack*, which used identical dataset and models [61] as *RP2*, we selected three

distinct shadow coefficient k (0.10, 0.45, and 0.70) to generate varied adversarial STOP signs. Here, a larger value of k corresponds to a lesser degree of perturbation. We maintained consistency with the public artifact by keeping all other parameters unchanged. For *SLAP*, we utilized the default parameters in its publicly available artifacts [30]. In addition, we also included one non-physical adversarial evasion attack, Fast Gradient Sign Method *FGSM* [18], serving as a baseline method for comparative evaluation. To align with the previously mentioned adversarial attacks, we generated adversarial STOP sign images using *FGSM* with gradients derived from the same model, applying ϵ values of 0.0175 and 0.02. In this context, larger ϵ means more perturbation. By choosing a diverse set of attacks and optimizing their parameters, we aim to mimic the variety and complexity of potential threats. This approach allows us to evaluate the robustness of AD against a representative sample of adversarial threats. Meanwhile, while ensuring the representativeness and relevance of the adversarial image types and settings, we acknowledge that our findings are specifically relevant to the chosen perspectives. We recognize the urgent need for continued research to broaden our exploration of this subject by incorporating a wider variety of adversarial images.

4.1.4 Pilot study and results. Before the full-scale user studies, we conducted two rounds of pilot studies to assess the feasibility of the system design and the effectiveness of the metrics collection methods. Additionally, they provided an opportunity to fine-tune our procedures in preparation for the full-scale study. Particularly, we decided on some key experiment settings, including the weather and light conditions, the format of post-surveys, the sequence of different kinds of STOP signs, etc. In this section, we detail the process of pilot studies and the decisions that have been made to lead to the final version of the evaluation plan.

In the initial pilot study, eight subjects were recruited, following the same recruitment and screening procedure in Section 4.1.1. In this study, the virtual driving scene is designed with three sequential intersections, each with a STOP sign. In contrast, the second STOP sign is modified through an adversarial evasion attack, and the third is intentionally 'dirty', marked with benign substances like mud or paint, which do not constitute a deliberate attack. The vehicle is programmed to stop at the normal and dirty STOP signs but continue without stopping at the adversarial one. This setup is based on the assumption that the adversarial evasion attack effectively deceives the AD perception system, leading it to disregard the manipulated STOP sign. Many subjects expressed surprise when the vehicle failed to stop at the second intersection, leading them to become more cautious at the following one. To reduce the impact of the order in which the STOP signs appeared and to enable a fair comparison of awareness across the three types, we rearranged the sequence: the normal STOP sign comes first, followed by the dirty STOP sign, and then the adversarial STOP sign.

In the second round of the pilot study, we engaged another eight participants to test the revised environment. During this round, we noticed that many struggled to identify the texture of the adversarial STOP sign due to a few factors: the environment's dim lighting, the blurriness of distant objects, and unfamiliarity with the concept of varying STOP sign textures Often, participants would start braking at the sight of the STOP sign's red color and shape from a distance, but they tended not to give it much attention as they got closer. To enhance the visibility of the STOP sign, we made adjustments to the Unity settings. We increased the intensity of light and altered the weather conditions to create a more realistic urban road environment. Additionally, we reduced the car's speed from 45 mph to 30 mph to simulate driving within an urban area.

4.1.5 Ethical considerations. Our study including the full-scale user and the pilot studies got the IRB approval from both participating institutions: #19423 from Indiana University Bloomington and #2023-0334 from The University of Texas at Arlington. Each participant was comprehensively informed about the study's purpose, structure, and any potential risks associated with participation. Participants were free to withdraw from the study at any point or skip any questions without any penalty. To mitigate the risk of confidentiality breach, we refrained from collecting any personally identifiable information from the participants. Any retained email addresses, used solely for the purpose of distributing participant compensation, were promptly deleted upon completion of this process. The entire study lasted around 30 minutes for each participant on average. Participants who completed the full study were compensated with a $5 Amazon gift card.

4.2 Evaluation Metrics

In this section, we present evaluation metrics for assessing the effectiveness and usability of *Avara*.

4.2.1 Effectiveness Metrics. In our study, we seek to establish a comprehensive understanding of participants' awareness of adversarial STOP signs using three driver's awareness data sources, i.e., driver's operation data, eye-movement data, and responses from a post-experiment survey for awareness of the abnormality of STOP signs, as elaborated below.

Survey response. In the post-experiment survey, participants provide their subjective assessment of the awareness of sign abnormality, denoted as *DifferenceAwareness*. Feedback to the question "On a scale from 1 to 5, please rate the normality of 1st/2nd/3rd stop sign compared with a standard stop sign?" is collected from participants to evaluate their perceptions of the different STOP signs.

Driver's operation. To assess a driver's responsive awareness of adversarial STOP signs objectively, we monitor whether participants initiate the take-over control operation before approaching an adversarial STOP sign. If so, we categorize them as "responsive aware" of adversarial STOP signs; otherwise, we label them as "not responsive aware". In our study, we calculate the proportion of participants who were categorized as "responsive aware" when encountering an adversarial STOP sign, in relation to the total number of participants who faced this scenario, denoted as *perResponsive*.

Eye-movement data. Eye movement patterns are widely used to indicate human visual behaviors and situation awareness [60]. In our study, we examine two gaze commonly-used features for human situation awareness studies: fixation duration and saccade amplitude [28, 60]. Fixation duration is an index of how long a person focuses their attention on a specific object, while saccade amplitude is the distance traveled by the eye between two fixation points. In our study, we use the average of fixation duration (*fMean*) and an average of saccade amplitude (*sAmpMean*) to assess the

effectiveness of *Avara* and select *fMean*, the most correlated, as a representation for perceptual awareness of adversarial STOP signs.

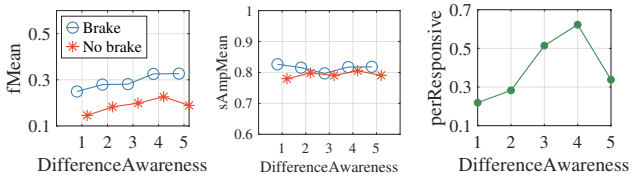
Correlation analysis. Survey response, represented as participants' self-reported awareness in the questionnaire, indicates how much participants are subjectively aware of adversarial STOP signs. However, survey responses may be influenced by various factors, such as bias and memory deficiency. For example, a participant might select a response that does not accurately reflect their true awareness, either as a way to mask a misoperation during the experiment or due to personal biases or memory lapses. To compensate for this, we integrate two objective indicators, namely eye movement and driver's operation data. The former provides information on users' visual interaction with the STOP signs, whereas the latter records how they operate the vehicle before and after entering each intersection. These objective measures are immune to the subjective biases and memory issues inherent in survey responses, providing a more accurate reflection of the participant's actual awareness and behaviors. However, it is important to recognize that these objective measures are not without their own limitations. Issues such as implicit metrics or possible misoperations during the VR simulation can introduce inaccuracies (see Section 4.3).

To jointly exploit the advantage while eliminating the drawback of each measure, we employ a correlation analysis approach. This approach allows us to validate the consistency of the findings across different types of measures and to draw more reliable conclusions about the efficacy of our system in measuring awareness of adversarial STOP signs.

4.2.2 Usability metrics. To evaluate the usability of our system, we have incorporated the System Usability Scale (SUS) [5] into our survey design, which is a reliable, widely-used assessments of system usability associated with 10 questions. Participants respond to these questions in a five-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". In our study, to streamline the process and enhance survey efficiency, we merged three SUS questions (i.e., "ease to use", "learn to use the system quickly", "cumbersome to use") with modifications (see Q12 "ease to use and learn" in Appendix 8) to include six usability-related questions. It's important to note that these modifications are strategically designed to capture the essence of usability measure: effectiveness (the user's ability to complete tasks using the system), efficiency (the level of resource consumed in performing tasks), and satisfaction (users' subjective reactions to using the system). By focusing on these core metrics, we maintain the integrity of usability assessment despite the alterations to the questionnaire. All questions presented in the survey are included in Appendix 8.

4.3 Effectiveness of *Avara*

Results. Figure 3 illustrates the correlation analysis results among three measures, i.e., survey response (*DifferenceAwareness*), driver's operation (*perResponsive*) and two eye-movement measures (*fMean* and *sAmpMean*). *fMean* exhibits strong correlations with survey responses in terms of Pearson correlation coefficient ($r = 0.56$), whereas *sAmpMean* does not show such correlation. Therefore, we choose *fMean* as an awareness indicator for eye-movement measures. A possible explanation is that participants, when more



(a) fMean vs DifferenceAwareness **(b) sAmpMean vs DifferenceAwareness** **(c) perResponsive vs DifferenceAwareness**

Figure 3: Cross-validation among measures.

alert and aware of the STOP sign, tend to steadily focus on a certain area, i.e., the STOP sign's surface. This increased attention leads to more concentrated fixations and, thus, longer fixation durations. This suggests that eye-tracking results generally agree with the survey results, demonstrating the validity of selected measures. This observation is further validated in Figure 4, which depicts the distributions of *fMean* for each self-reported awareness level. Evidently, *fMean* is more distributed on relatively larger values for higher self-reported awareness levels. Figure 3b shows *sAmpMean* keeps as a relatively stable trends with the changes to the self-reported awareness levels, which indicates the similar eye travel distance between two fixation points.

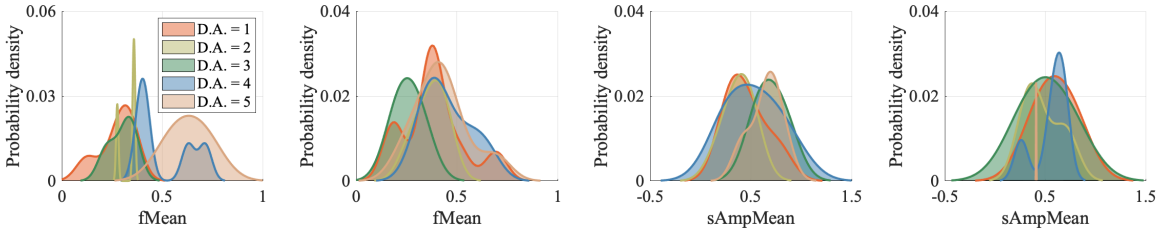
We further investigate the correlation between the survey response and the driver's operation (*perResponsive*). The *perResponsive* is the timestamp the user applies the stop before (indicated by a negative value) or after (a positive value) reaching the STOP sign. As demonstrated in Figure 3c, higher awareness levels are indicated by lower *perResponsive* in general. As a piece of supporting evidence, Figure 4 suggests that distributions of *fMean* with larger values are found for data samples where the take-over control operation is applied, as opposed to where such operation is not applied, under the same self-reported awareness levels. This observation is intuitive since users who are more aware of the STOP sign are more likely to respond and stop the vehicle earlier before reaching it. In conclusion, these correlations prove the feasibility of employing our measures.

Outlier discussion. Our correlation analysis also reveals outliers under each measure, as shown in Figure 5. These outliers are resulted from device inaccuracy, user mis-operations and/or disagreements between different measures. For instance, one participant, who took over control at every intersection, commented that:

> *"I feel that I can never fully trust the AV and this is probably what I would do in the real-world scenario."*

This, however, introduces unintended user bias into our data. Additionally, we observed one participant had taken over control before the adversarial STOP sign (i.e., being labeled as "responsive aware") but provided an incoherent answer in the survey, possibly due to memory deficiency. We also noticed several outrageous eye-tracking data, which may be subject to the hardware or software limitations of *Avara*. Notably, we identify 5% of self-reported results to be unreliable and considered as outliers. This observation is significant as it highlights the superiority of our scheme and reveals an aspect of the study that has not been reported in prior work.

In our study, we eliminate outliers to ensure the consistency of the findings across different types of measures and to draw more



(a) Distributions of fMean when taking-over control **(b) Distributions of fMean when taking-over control** **(c) Distributions of sAmpMean when taking-over control** **(d) Distributions of sAmpMean when taking-over control**

Figure 4: Probability distributions of eye-movement metrics for various self-reported awareness levels in survey response.

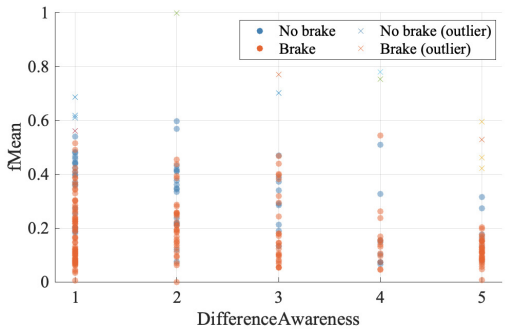


Figure 5: Disagreement between *fMean* and survey responses.

reliable conclusions in measuring awareness of adversarial STOP signs in Section 5.

4.4 Platform Usability

Figure 6 illustrates the results of *Avara*'s usability evaluation. Participants showed a high willingness to use *Avara*, most describing their experience as 'somewhat pleasant', as one participant commented:

> *"I had a decent experience during this experiment. The experiment feels like sitting in a real car, the environment is familiar...I feel like I'm driving a 'real' autonomous car even though I've never driven one before."*

Additionally, an encouraging 88% of participants found *Avara* 'easy' or 'somewhat easy' to use and learn. A participant with an HCI design background said:

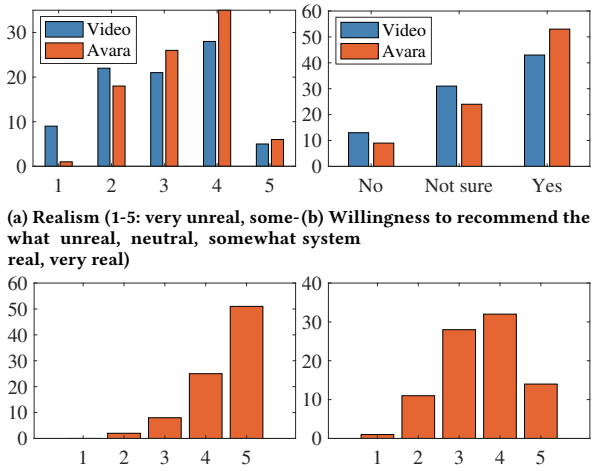
> *"...it's a whole new experience for me since I never used VR before and thought it would be difficult...but it's much easier to use with the controller."*

When assessing realism, 76% of the participants believed that our simulation accurately represented real-world driving conditions. Many participants comment with a similar meaning:

> *"The driving scene is very realistic, even though it still needs some minor improvements, such as adding other cars and pedestrians, etc."*

Importantly, 62% of participants expressed a willingness to recommend our system for driver training sessions against adversarial evasion attacks with a comment that:

> *"I will recommend other people to use it as AV training, because it provides a very realistic environment than videos and it's a really interesting experience..."*



(a) Realism (1-5: very unreal, somewhat unreal, neutral, somewhat real, very real) **(b) Willingness to recommend the system**

(c) Ease of use (1-5: very difficult, somewhat difficult, neutral, somewhat easy, very easy) **(d) Pleasantness (1-5: very unpleasant, somewhat unpleasant, neutral, somewhat pleasant, very pleasant)**

Figure 6: Platform usability.

We further conducted a comparison of the usability of *Avara* with a prior video-based autonomous driving platform: PC-based setting [59] and PC-based setting with the space key for operation [33]. As illustrated in Figure 6a, *Avara* outperforms the PC-based platform in simulating real-world situations, with a significant number of participants selecting 'Somewhat Real' choices. Conversely, nearly 50% of participants chose 'Somewhat Unreal' and 'Very Unreal' for the PC platform. One participant commented that:

> *" It's just a common video with a car driving on the road... I noticed the difference on the STOP sign, but it can't be called 'real', especially compared with 3D VR."*

Furthermore, 54 out of 86 participants expressed a willingness to recommend our system for driving training sessions against adversarial evasion attacks, while 39 would recommend the PC-based system. Near half of the participants also recommend the PC platform for the reasons:

> *"I think the video is an acceptable platform, even though it's not very realistic and can be operated like VR."*

5 Measurement and Findings

In this section, we present our measurement analysis to study driver's awareness and reactions when encountering attacks.

5.1 Driver's Awareness of Different STOP Signs

Driver's awareness toward normal, dirty, adversarial STOP signs. As outlined in Section 4.1.4, our study assesses participants' situation awareness under three distinct scenarios featuring normal, dirty, and adversarial STOP signs. Our study indicates that *Avara* is proficient in capturing and differentiating the levels of situation awareness exhibited by participants in response to these three types of STOP signs. Particularly, Figure 8 illustrates comparative data across the three scenarios—normal, dirty, and adversarial. Here, we applied min-max normalization to the values derived from three awareness measures (*DifferenceAwareness*, *fMean*, *perResponsive*). This normalization was adjusted with a slight bias to circumvent the occurrence of zero values, ensuring a more robust and accurate analysis. We found that under adversarial attack, all three awareness measures significantly increase compared to normal and dirty STOP sign scenarios. Specifically, we noted a significant change in the *fMean* value, when drivers encountered intersections with adversarial STOP signs compared to those with dirty STOP signs. Specifically, the *fMean* value increased from 0.16 to 0.37 in these scenarios. This pattern was similarly reflected in the metrics for *DifferenceAwareness*, which changed from 2.28 to 3.60, and *perResponsive*, which altered from 0.16 to 0.69. These findings indicate that the dirty STOP signs, while also presenting a visual challenge, did not elicit the same level of driver awareness as adversarial ones.

Impact of different attack methods. Figure 9 shows the driver's awareness level under different types of adversarial evasion attacks. In our study, we observed that drivers exhibited a higher level of perceptual awareness for the *SLAP* and *FTE3* adversarial evasion attack, in terms of *fMean* and *DifferenceAwareness*, compared to other adversarial attacks. This notable distinction in driver response to *SLAP* can be attributed, at least in part, to its significantly higher Perturbation Sensitivity Distance (*PSD*) (1.0 vs 0.14 for *FGSM_0.0175*, 0.22 for *FGSM_0.02*, 0.13 for *RP2*, 0.21 for *ShadowAttack_0.10*, 0.13 for *ShadowAttack_0.45*, 0.12 for *ShadowAttack_0.70*).

Regarding responsive awareness, i.e., participants take over control when encountering an adversarial STOP sign; we observed that participants demonstrated a high level of responsive awareness for the *ShadowAttack*. In this scenario, 81% of participants chose to take over control, which is a significant proportion compared to other adversarial attacks observed in the study (70% for *FGSM* attack, 58% for *FTE3* attack, and 52% for *SLAP* attack). As some participants stated the reason why they stopped at *ShadowAttack_0.10* and *ShadowAttack_0.45* that:

> *"...I switched to the manual mode when I saw a big black hole on a stop sign, I guess maybe it's broken?"*
>
> *"I noticed the sign was in shadow from far away, so I slowed down to get a better look at it..."*

Impact of attack inconspicuousness. Figure 7 illustrated the correlation between three user awareness measures (*DifferenceAwareness, fMean, perResponsive*) and the four inconspicuousness metrics commonly used in adversarial attack (L_1, L_2, *ASS*, *PSD*, see Section 2).

When examining the correlation between four imperceptibility metrics, with our three user awareness measures, we observed that the L_1 has the strongest correlation with the *perResponsive*, which reflects users' operations. L_2 shows the most correlation with *DifferenceAwarenes*, and adequately correlation with *fMean*. The *ASS* has the strongest correlation with *DifferenceAwarenes*.

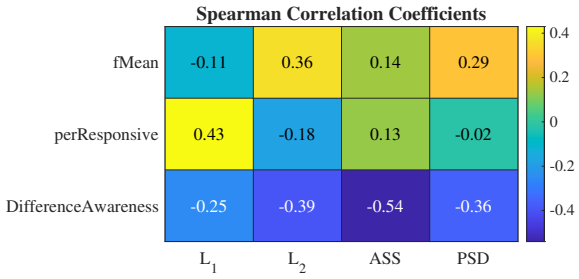


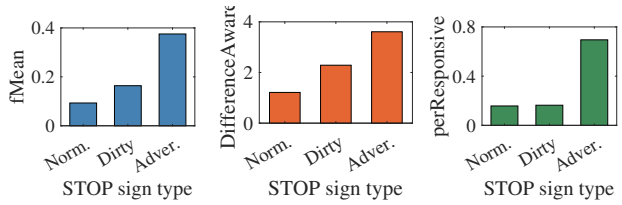
Figure 7: Correlation between metrics.

For *perResponsive*, we found an interesting result that it only strongly positively correlated with L_1(r=0.43), and slightly negative correlation with L_2(r=−0.18), which means with the higher value on L_1, the adversarial stop signs more unlike to a normal one, and more people choose to take over the control of the automated car while spotted this attack. However, for the other two similarity metrics, *ASS* and *PSD* slightly correlated with *perResponsive* (r=0.13 and −0.02), which means the image dissimilarity measured by these two does not influence people's decision on taking over the control.

When examining the correlation between L_1 and *ASS* distortion of our eye movement awareness measures, we observed a weak correlation (r = −0.11, 0.14). This implies that changes in the L_1 and *ASS* distortion levels do not significantly align with variations in user awareness captured by eye movement metrics. In contrast. eye awareness measures are stronger positively along with L_2 (r=0.36) and *PSD* (r=0.29).

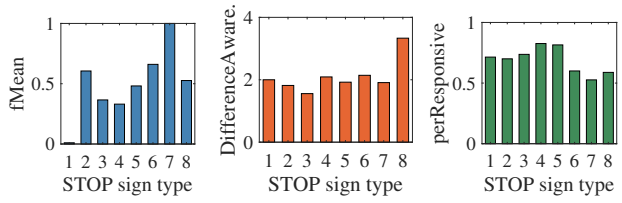
In contrast, *DifferenceAwareness* demonstrated a strong negative correlation with two of the tested inconspicuousness metrics: L_2 and *ASS* (r = −0.39, −0.54). These findings suggest that while traditional inconspicuousness metrics like L_1, *PSD*, and *ASS* provide valuable technical insights into the nature of adversarial attacks, they may not fully capture how L_2 such attacks influence human perception and awareness. On the other hand, with its strong correlation to user awareness measures, it offers a more relevant understanding of the real-world effectiveness of adversarial perturbations, particularly in terms of how they are perceived and processed by users. The similarity of different adversarial STOP signs under four inconspicuousness metrics is compared in Figure 10. Our findings highlight a disparity between perceptibility metrics used in traditional adversarial ML (e.g., L-norm/ASS/PSD) and human-centered perceptibility analysis (e.g., fMean). This gap suggests a renovation of the existing robust ML training frameworks [25, 26, 53], which assume that images with small L-norm/ASS/PSD differences should yield identical predictions. A promising solution involves integrating human-centered perceptibility metrics into ML training. Moreover, this could help generate more realistic adversarial examples and improve AD object detection models for real-world environments, enhancing safety and reliability.

Impact of attack disturbance level. Figure 9 shows the awareness levels on *FGSM* and *ShadowAttack* with different disturbance levels. In general, for disturbance levels related to lower image similarity, the adversarial STOP sign shows less awareness to participants.



(a) fMean (b) DifferenceAwareness (c) perResponsive

Figure 8: Comparison among three types of STOP signs.



(a) fMean (b) DifferenceAwareness (c) perResponsive

Figure 9: Different types of adversarial STOP signs with various disturbance levels. 1-8 represent {"FGSM_eps = 0.0175", "FGSM_eps = 0.02", "RP2", "ShadowAttack_k = 0.10", "ShadowAttack_k = 0.45", "ShadowAttack_k = 0.70", "SLAP", "FTE3"}, respectively.

For *FGSM* adversarial attack, we evaluated two disturbance levels: ϵ as 0.0175 and 0.02. *FGSM_0.02* shows higher awareness on both three effectiveness metrics than *FGSM_0.0175*, as well as shows a higher dissimilarity level compared with a normal STOP sign on all four image similarity comparisons. As for *ShadowAttack* adversarial STOP signs, we present three disturbance levels: 0.10, 0.45, and 0.70. The images of these three STOP signs can be found in Appendix Figure 14g. The awareness level and image similarity for these three are not very obvious as *FGSM*, but we can recognize that *ShadowAttack_0.70* is the STOP sign with the lowest awareness and the one most similar to a normal STOP sign. The eye movement awareness shows a decreasing trend from *ShadowAttack_0.10* to *ShadowAttack_0.70*. However, parallel trends were observed in the L_1 dissimilarity and the *DifferenceAwareness* metrics, that the L_1 dissimilarity and *DifferenceAwareness* levels of the *ShadowAttack_0.45* are the lowest.

5.2 Impact of Education on User Awareness

In our study, we implemented an educational phase focused on adversarial evasion attacks, which took place between two experiment tasks (i.e., benchmark task and after-education task), as detailed in Section 4.1.2. In the education process, participant were shown a set of adversarial STOP sign images, which were independent of those used in the actual experiment. This separation ensured that the education phase did not directly influence the participants' responses to the specific images they would encounter later in the experiment. Alongside showing the images, we provided participants with an introduction to the concept of adversarial attacks. This included an explanation of how these attacks are designed, their potential objectives, and their implications, particularly in the context of autonomous driving systems.

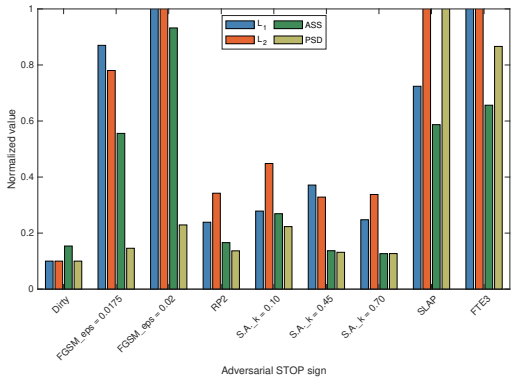


Figure 10: Similarity of different adversarial STOP signs.

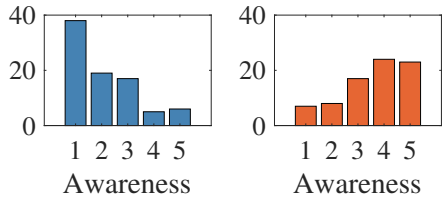
We hypothesize that education will increase users' awareness of the after-education task, as they have gained more experience and knowledge about the adversarial road signs. To verify our hypothesis, we study the impact of education by analyzing three user awareness measures (*DifferenceAwareness*, *fMean*, *perResponsive*), and compare their statistical results before and after education.

We first inspect the distributions of their self-reported awareness (*DifferenceAwareness*) of abnormality for each STOP signs, i.e., the difference compared with a normal road sign, as depicted in Figure 11. The awareness of abnormality levels are rated from 1 ("Normal as usual") to 5 ("Very Abnormal").

We observe a significant difference regarding awareness of the abnormality of the road signs compared to the normal ones ($z = -2.27484$, $p = 0.0232 < 0.05$). These results indicate that education has a significantly positive impact in raising users' awareness of road signs. We further investigate users' self-reported awareness of the abnormal vehicle behaviors at the adversarial STOP signs. Users report whether they are aware of the abnormal behavior, and at which intersections it was observed (multi-choice). Only choosing "yes" and "at the 3rd intersection" is considered correctly aware. Figure 12, demonstrates the result. Before education, 10 users indicated that they were not aware of the abnormal behavior, 24 incorrectly identified the abnormal behavior at wrong intersections, and 57 (64%) were correctly aware of the behavior. After education, only 5 did not report or misreport the behaviors, and 75 (84%) correctly identified the abnormal behavior at the correct intersection. 91 subjects participated in the pre-education study, among whom 89 attended the post-education study (2 dropped out). This further suggests the positive impact of education on users' awareness. Note that the reason that there remains a group of users who failed to correctly report the abnormal behavior even after education, as indicated by our survey feedback, is that they would still be deceived by some highly-deceiving adversarial signs (that are different from in the pre-education session) and perceive them as normal. For example, one participant noted seeing a shadow on *ShadowAttack_0.70* but perceived it as normal:

> *"I noticed a shadow on that stop sign, but I feel like it's normal."*

Next, we analyze the number of users who did/did not apply the take-over control at the adversarial intersection. As shown in Figure



(a) Awareness of sign abnormality before education (b) Awareness of sign abnormality after education

Figure 11: Awareness of road signs before and after education.

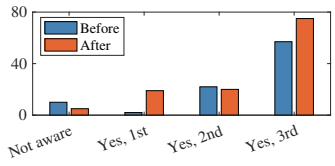


Figure 12: Awareness of abnormal vehicle behaviors.

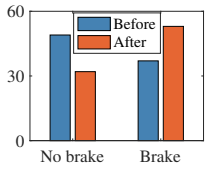


Figure 13: Take over control.

13, 37 (42%) users applied the take-over control before education. This number increases to 57 (64%) after education. At the same time, the number of users who did not apply the take-over control drops by 20 (22%). Similarly, most users who did not apply the take-over control after education were deceived by the highly-deceiving adversarial signs such as *ShadowAttack_0.70*. Overall, this indicates the role of education in increasing users' awareness and, as a result, leading them to adopt more defensive driving.

6 Discussion and Limitations

Exploring other technologies to develop the evaluation platform. Our system, *Avara*, currently utilizes VR headsets to simulate real-world scenarios, offering an immersive experience that closely mirrors actual environments. This setup, superior to traditional PC-based environments, provides human subjects with a near-real-world experience during experiments. However, some participants reported experiencing dizziness after using the VR headset for extended periods. In response to this feedback, we plan to investigate other emerging technologies, such as mixed reality (MR) and augmented reality (AR), to develop the evaluation platform in the future. These alternatives promise to deliver similarly immersive experiences while potentially mitigating the discomfort associated with prolonged VR use. Meanwhile, it is worth mentioning that Apple's upcoming MR product, Vision Pro [1], is priced over $4,000, much more expensive than VIVE Focus 3 Eye Tracker, the VR headset currently in use in our study. Nonetheless, as part of our future work, we intend to conduct a comparative analysis across evaluation frameworks using different computing platforms.
Exploring other metrics in assessing driver awareness. In Section 4, we introduced three awareness metrics and one usability evaluation metric. Other evaluation metrics include the measurement of pupil size and brain electronic signal, to provide an in-depth understanding of human perceptions of different adversarial attacks. In the context of eye movement measures for awareness evaluation, research has shown strong correlations between certain features and situational awareness [28, 63]. For instance, location-based features are highly correlated with awareness in scenarios requiring driver takeover. As future work, we plan to delve further into these additional metrics. One potential direction is to integrate advanced neurophysiological measurements, like EEG (i.e., electroencephalography), to capture brainwave patterns during different driving scenarios. This will allow us to more accurately reveal drivers' attention levels to various traffic signs, both normal and adversarial. Additionally, we intend to enhance our data analysis methods to include advanced AI models to decipher complex patterns in these multi-modal signals for awareness detection.
Broadening the participation. Moreover, while our study successfully recruited a substantial number of participants (60), much larger compared to other similar evaluation works (around 20), it still faced limitations in scale. To enhance the generalizability of our findings, further large-scale studies are necessary, aiming for a more diverse participant pool in terms of gender, race/ethnicity, age, and other demographic factors. Such an approach is crucial for ensuring that our research outcomes are reflective of a broad spectrum of drivers, thereby providing a more comprehensive understanding of driver awareness in response to adversarial attacks.

7 Related Work

Physical-world adversarial attacks on camera-based AD systems. Current AD systems rely on different sensing modalities for environmental and object detection, including LIDAR, ultrasonic, flash cameras, and thermal cameras [58]. These systems often integrate multiple AI models, like CNNs [27], RNNs [2], and deep learning techniques [23]. These models aid in recognizing and analyzing elements like traffic signs, vehicles, and pedestrians, crucial for autonomous driving [19, 38]. However, camera-based AI detection modalities in AD systems are particularly vulnerable to various adversarial attacks. These attacks can significantly compromise safety by manipulating the physical environment and objects to distort AI sensor inputs [9, 18, 35, 36]. Physical-world attacks are the most severe ones due to their low cost, diverse methods, direct impact, and effectiveness [17, 44, 46, 62].

In AD research, there exists a significant gap in understanding the human driver's role when the AD system is faced with physical-world adversarial attacks. Currently, there is a lack of studies that offer a realistic and affordable environment for human drivers to actively engage with and evaluate these threats. This oversight is critical, as human drivers possess a unique ability to perceive and interpret changes in their environment, especially visually manipulated or deceptive. Our research aims to fill this gap by creating a platform where drivers can experience and react to adversarial scenarios in a simulated AD environment.
Human situation awareness in the context of autonomous driving. In SAE Level 3 AD, drivers need to be aware of the instant situation around to resume control of the car from the AD model when necessary. Hence, prior works have been done to identify various factors that influence human awareness during AD [8], including light [57], texts [29], textures [59], and icons [45, 49].

Additionally, several studies have focused on evaluating human awareness during AD. Common methodologies in these studies involve using either single or combined evaluation metrics. These

Table 2: Driving evaluation platforms comparison.

Name	Platform	Learnability	Realism	Price
Video online [59]	Video-based	No operation	Low	Low
miniSim [28]	Medium-fidelity fixed-base driving simulator	Difficult	High	High
Video on monitor [33]	Monitor with external eye tracker	Easy (space bar)	Low	High
Avara	VR with eye-tracker	Easy (controller)	High	Low

include survey responses [28, 33, 63], keyboard interactions [33, 63], simulated steering wheel inputs [49], and eye movement tracking [28, 63]. For instance, Lu et al. engaged users with a car simulator where participants could press a space bar during animated video clips to signal awareness and the need to take over control [33]. Zhou et al. examined the correlation between eye movement, operational data, and situation awareness [63]. Liang et al. specifically investigated the impact of eye movement on situational awareness during the 'pre-takeover' phase under AD conditions [28].

Note that the focus of our study is distinct from that of the above works: We intend to understand driver awareness in response to adversarial attacks on AD systems. Nonetheless, the above works provide viable methodologies for assessing driver situation awareness, including under various adversarial attacks on traffic signs, informing our assessment approach selection.

AD adversarial attacks evaluation platforms. Several evaluation platforms have been developed in prior work. To offer a comprehensive overview, we compared them with *Avara* in Table 2, from learnability, realism, and cost. Existing AD evaluations generally fall into two categories based on their experimental setup: video-based evaluations and medium- to high-fidelity car simulators. Studies [33, 59, 63] employ video clips showing adversarial STOP signs or hazardous driving conditions. These videos are played on screens for participants to watch. While video-based platforms are user-friendly, convenient, and cost-effective, they fall short of delivering a realistic simulation of actual driving experiences: participants can be easily distracted by objects beyond the screen. On the other hand, car simulators offer the most lifelike platform for AD environments. miniSim [28] utilized a medium-fidelity fixed-base car simulator, allowing users to operate a steering wheel and view an AD scenario across three monitors. Among these simulators, both *Avara* and miniSim[28] deliver high realism, but *Avara* is more cost-effective and easier to use. Its combination of VR and eye-tracking technologies provides a distinct immersive experience compared to the fixed-based setup of miniSim.

In summary, *Avara* distinguishes itself by its balance of high realism, user-friendliness, and affordability. It surpasses video-based platforms in immersion and interactivity and is more accessible than high-fidelity simulators in terms of cost and ease of learning, making it a standout choice for AD adversarial attack evaluations.

8 Conclusion

In this work, we built a novel VR-based autonomous driving evaluation platform *Avara*. *Avara* is specifically designed to assess driver awareness and perceptions in response to physical-world adversarial attacks on traffic signs. We introduced a range of evaluation metrics focused on user awareness. Our detailed assessment covered four prevalent types of physical-world adversarial attacks and various disturbance levels for the *FGSM* and *ShadowAttack*. Our *Avara* platform and discoveries have made a step toward a better understanding of human perceptions of physical-world adversarial attacks, which have not attracted much attention so far. Moreover, *Avara* offers a cost-effective and reliable option for other researchers to conduct their studies and thus improve the safety of AD.

Availability

The artifact for this work is available at https://zenodo.org/records/13346274.

Acknowledgement

This work was supported in part by the National Science Foundation (CNS-2343618, 1850725), U.S. Department of Homeland Security (17STQAC00001-07-00), U.S. Department of Defense (W52P1J2093009) and Luddy Faculty Fellowship.

References

[1] Apple. 2024. Apple Vision Pro. https://www.apple.com/apple-vision-pro/.
[2] Henrik Arnelid, Edvin Listo Zec, and Nasser Mohammadiha. 2019. Recurrent conditional generative adversarial networks for autonomous driving sensor modelling. In *2019 IEEE Intelligent transportation systems conference (ITSC)*. IEEE, 1613–1618.
[3] Istvan Barabas, Adrian Todoruț, N Cordoș, and Andreia Molea. 2017. Current challenges in autonomous driving. In *IOP conference series: materials science and engineering*, Vol. 252. IOP Publishing, 012096.
[4] Birtukan Birawo and Pawel Kasprowski. 2022. Review and evaluation of eye movement event detection algorithms. *Sensors* 22, 22 (2022), 8810.
[5] John Brooke. 1996. Sus: a "quick and dirty'usability. *Usability evaluation in industry* 189, 3 (1996), 189–194.
[6] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 176–194.
[7] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2267–2281.
[8] Marine Capallera, Leonardo Angelini, Quentin Meteier, Omar Abou Khaled, and Elena Mugellini. 2022. Human-Vehicle Interaction to Support Driver's Situation Awareness in Automated Vehicles: A Systematic Review. *IEEE Transactions on Intelligent Vehicles* (2022).
[9] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*. Ieee, 39–57.
[10] Benjamin T Carter and Steven G Luke. 2020. Best practices in eye tracking research. *International Journal of Psychophysiology* 155 (2020), 49–62.
[11] Alesia Chernikova, Alina Oprea, Cristina Nita-Rotaru, and BaekGyu Kim. 2019. Are self-driving cars secure? evasion attacks against deep neural networks for steering angle prediction. In *2019 IEEE Security and Privacy Workshops (SPW)*. IEEE, 132–137.
[12] Viviane Clay, Peter König, and Sabine Koenig. 2019. Eye tracking in virtual reality. *Journal of eye movement research* 12, 1 (2019).
[13] HTC Corporation. 2024. VIVE Wave XR Plugin. https://hub.vive.com/storage/docs/en-us/UnityXR/UnityXRSdk.html.

[14] HTC Corporation. 2024. Wave Unity SDK. https://hub.vive.com/storage/docs/en-us/UnitySdk.html.

[15] Yao Deng, Xi Zheng, Tianyi Zhang, Chen Chen, Guannan Lou, and Miryung Kim. 2020. An analysis of adversarial attacks and defenses on autonomous driving models. In *2020 IEEE international conference on pervasive computing and communications (PerCom)*. IEEE, 1–10.

[16] Kevin Eykholt. 2018. RP2 Artifacts. https://github.com/evtimovi/robust_physical_perturbations.

[17] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. 2018. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1625–1634.

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[19] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. 2020. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics* 37, 3 (2020), 362–386.

[20] Wei Jia, Zhaojun Lu, Haichun Zhang, Zhenglin Liu, Jie Wang, and Gang Qu. 2022. Fooling the eyes of autonomous vehicles: Robust physical adversarial examples against traffic sign recognition systems. *arXiv preprint arXiv:2201.06192* (2022).

[21] Wei Jia, Zhaojun Lu, Haichun Zhang, Zhenglin Liu, Jie Wang, and Gang Qu. 2022. Fooling the Eyes of Autonomous Vehicles: Robust Physical Adversarial Examples Against Traffic Sign Recognition Systems. *arXiv preprint arXiv:2201.06192* (2022).

[22] Glenn Jocher. [n. d.]. YOLOv5. https://github.com/ultralytics/yolov5. Accessed: 2024-1-27.

[23] B Ravi Kiran, Ibrahim Sobh, Victor Talpaert, Patrick Mannion, Ahmad A Al Sallab, Senthil Yogamani, and Patrick Pérez. 2021. Deep reinforcement learning for autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems* 23, 6 (2021), 4909–4926.

[24] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. 2020. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14254–14263.

[25] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2017. Adversarial Machine Learning at Scale. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=BJm4T4Kgx

[26] L. Li, T. Xie, and B. Li. 2023. SoK: Certified Robustness for Deep Neural Networks. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, Los Alamitos, CA, USA, 1289–1310. https://doi.org/10.1109/SP46215.2023.10179303

[27] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. 2019. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7644–7652.

[28] Nade Liang, Jing Yang, Denny Yu, Kwaku O Prakah-Asante, Reates Curry, Mike Blommer, Radhakrishnan Swaminathan, and Brandon J Pitts. 2021. Using eye-tracking to investigate the effects of pre-takeover visual engagement on situation awareness during automated driving. *Accident Analysis & Prevention* 157 (2021), 106143.

[29] Patrick Lindemann, Tae-Young Lee, and Gerhard Rigoll. 2018. Catch my drift: Elevating situation awareness for highly automated driving with an explanatory windshield display user interface. *Multimodal Technologies and Interaction* 2, 4 (2018), 71.

[30] Giulio Lovisotto. 2021. SLAP Artifacts. https://github.com/ssloxford/short-lived-adversarial-perturbations.

[31] Giulio Lovisotto, Henry Turner, Ivo Sluganovic, Martin Strohmeier, and Ivan Martinovic. 2021. {SLAP}: Improving physical adversarial examples with {Short-Lived} adversarial perturbations. In *30th USENIX Security Symposium (USENIX Security 21)*. 1865–1882.

[32] CKAS Mechatronics Pty Ltd. 2019. CKAS MiniSim Car Simulators. https://www.ckas.com.au/minisim_car_simulators_87.html.

[33] Zhenji Lu, Riender Happee, and Joost CF de Winter. 2020. Take over! A video-clip study measuring attention, situation awareness, and decision-making in the face of an impending hazard. *Transportation research part F: traffic psychology and behaviour* 72 (2020), 211–225.

[34] Bo Luo, Yannan Liu, Lingxiao Wei, and Q. Xu. 2018. Towards Imperceptible and Robust Adversarial Example Attacks against Neural Networks. *ArXiv* abs/1801.04693 (2018). https://api.semanticscholar.org/CorpusID:19225543

[35] Chen Ma, Ningfei Wang, Qi Alfred Chen, and Chao Shen. 2023. WIP: Towards the Practicality of the Adversarial Attack on Object Tracking in Autonomous Driving. In *ISOC Symposium on Vehicle Security and Privacy (VehicleSec)*.

[36] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083* (2017).

[37] Andreas Mogelmose, Mohan Manubhai Trivedi, and Thomas B. Moeslund. 2012. Vision-Based Traffic Sign Detection and Analysis for Intelligent Driver Assistance Systems: Perspectives and Survey. *IEEE Transactions on Intelligent Transportation Systems* 13, 4 (2012), 1484–1497. https://doi.org/10.1109/TITS.2012.2209421

[38] Khan Muhammad, Amin Ullah, Jaime Lloret, Javier Del Ser, and Victor Hugo C de Albuquerque. 2020. Deep learning for safe autonomous driving: Current challenges and future directions. *IEEE Transactions on Intelligent Transportation Systems* 22, 7 (2020), 4316–4336.

[39] Anneli Olsen. 2012. The Tobii I-VT fixation filter. *Tobii Technology* 21 (2012), 4–19.

[40] Qualtics. 2024. Qualtics. https://www.qualtrics.com/.

[41] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[42] Dario D Salvucci and Joseph H Goldberg. 2000. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the 2000 symposium on Eye tracking research & applications*. 71–78.

[43] Takami Sato, Sri Hrushikesh Varma Bhupathiraju, Michael Clifford, Takeshi Sugawara, Qi Alfred Chen, and Sara Rampazzi. 2024. Invisible Reflections: Leveraging Infrared Laser Reflections to Target Traffic Sign Perception. *arXiv preprint arXiv:2401.03582* (2024).

[44] Athena Sayles, Ashish Hooda, Mohit Gupta, Rahul Chatterjee, and Earlence Fernandes. 2021. Invisible perturbations: Physical adversarial examples exploiting the rolling shutter effect. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14666–14675.

[45] Ronald Schroeter and Fabius Steinberger. 2016. Pokémon DRIVE: towards increased situational awareness in semi-automated driving. In *Proceedings of the 28th australian conference on computer-human interaction*. 25–29.

[46] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. 2018. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430* (2018).

[47] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. 2018. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*.

[48] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. 2012. Man vs. computer: Benchmarking machine learning algorithms for traffic sign recognition. *Neural Networks* 32 (2012), 323–332. https://doi.org/10.1016/j.neunet.2012.02.016 Selected Papers from IJCNN 2011.

[49] Sonja Stockert, Natalie Tara Richardson, and Markus Lienkamp. 2015. Driving in an increasingly automated world–approaches to improve the driver-automation interaction. *Procedia Manufacturing* 3 (2015), 2889–2896.

[50] Unity Technologies. 2024. Unity. https://unity.com.

[51] Unity Technologies. 2024. Unity Asset. https://assetstore.unity.com/packages/tools/particles-effects/time-of-day-weather-system-40374.

[52] Unity Technologies. 2024. Unity Asset Store. https://assetstore.unity.com/.

[53] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian J. Goodfellow, Dan Boneh, and Patrick D. McDaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=rkZvSe-RZ

[54] Ningfei Wang, Yunpeng Luo, Takami Sato, Kaidi Xu, and Qi Alfred Chen. 2023. 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*. 4412–4423.

[55] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. 2020. 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, 665–681.

[56] Vivek Yadav. 2016. GTSRB-CNN. https://github.com/vxy10/p2-TrafficSigns.

[57] Yucheng Yang, Burak Karakaya, Giancarlo Caccia Dominioni, Kyosuke Kawabe, and Klaus Bengler. 2018. An hmi concept to improve driver's visual behavior and situation awareness in automated vehicle. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 650–655.

[58] Ekim Yurtsever, Jacob Lambert, Alexander Carballo, and Kazuya Takeda. 2020. A survey of autonomous driving: Common practices and emerging technologies. *IEEE access* 8 (2020), 58443–58469.

[59] Katherine S Zhang, Claire Chen, and Aiping Xiong. 2023. Human Drivers' Situation Awareness of Autonomous Driving Under Physical-world Attacks. In *Proceedings Inaugural International Symposium on Vehicle Security & Privacy*.

[60] Ting Zhang, Jing Yang, Nade Liang, Brandon J Pitts, Kwaku O Prakah-Asante, Reates Curry, Bradley S Duerstock, Juan P Wachs, and Denny Yu. 2020. Physiological measurements of situation awareness: a systematic review. *Human factors* (2020), 0018720820969071.

[61] Yiqi Zhong. 2022. Shadow Attack Artifacts. https://github.com/hncszyq/ShadowAttack.

[62] Yiqi Zhong, Xianming Liu, Deming Zhai, Junjun Jiang, and Xiangyang Ji. 2022. Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15345–15354.

[63] Feng Zhou, X Jessie Yang, and Joost CF De Winter. 2021. Using eye-tracking data to predict situation awareness in real time during takeover transitions in conditionally automated driving. *IEEE transactions on intelligent transportation systems* 23, 3 (2021), 2284–2295.

Appendix

A. I-VT Algorithm

Algorithm 2 I-VT algorithm

Input: Eye movement data in time sequence $\boldsymbol{s} = [\{x_i, y_i\}]$, velocity threshold θ, maximum time between fixations T
Output: A list of points labeled with fixation/saccade: $\boldsymbol{\sigma} = [a_i]$
$\theta \leftarrow 0.5$; $T \leftarrow 75$ ms;
for each s_i in $\boldsymbol{s}$ **do**
    Calculate time difference and distance between consecutive points: $\delta_i \leftarrow t_{i+1} - t_i$; $d_i \leftarrow \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2}$;
    Calculate velocity: $v_i \leftarrow d_i/\delta_i$;
    Label points: $\sigma_i \leftarrow$ fixation if $v_i < \theta$, else $\sigma_i \leftarrow$ saccade;
end for
for each saccade group $\sigma_s[t_{start} : t_{end}]\}$ **do**
    for each a_i in $\boldsymbol{\sigma}$ **do**
        if $t_{end} - t_{start} < T$ **then**
            $a_i \leftarrow$ fixation;
        end if
    end for
end for

B. Dirty and Adversarial Stop Signs Used in Experiments

(a) Dirty Stop Sign **(b) RP2 Stop Sign** **(c) SLAP Stop Sign**

(d) FGSM Stop Sign with $\epsilon = 0.0175$ **(e) FGSM Stop Sign with $\epsilon = 0.02$** **(f) FTE3 Stop Sign**

(g) ShadowAttack Stop Sign with $k = 0.10$ **(h) ShadowAttack Stop Sign with $k = 0.45$** **(i) ShadowAttack Stop Sign with $k = 0.70$**

C. Post-Experiment Survey

Questions after the first round

Q1 Have you wear glasses during the experiment? (1)Yes (2)No (3)No, but I wear the contact lenses
Q2 Did you notice anything unusual about the stop signs during the driving scenarios? (1)Yes (2)No
Q3 Did you notice that the autonomous vehicle behaved weirdly at any intersection? (1)Yes, the car behaved weirdly at the 1st intersection (2)Yes, the car behaved weirdly at the 2nd intersection (3)Yes, the car behaved weirdly at the 3rd intersection (4)No

Questions same for the first and second round

Q4 On a scale from 1 to 5, please rate the normality of first/second/third stop sign compared with a standard stop sign? (1)Normal as usual (2)Slightly abnormal (3)Moderately abnormal (4)Mostly abnormal (5)Very abnormal
Q5 Have you taken any actions when you drive through the first/second/third stop sign? (1)I didn't notice it (2)No, I noticed, but I didn't do anything (3)Yes, I took over the control and braked (4)I did something else: I took over the control and sped up

Education Sessions

Q6 Here are some examples of the stop signals that you just experienced: normal + dirty
Q7 Here are the stop signs you may have encountered in the third interaction. Do you remember which one you saw?

Questions after the second round

Q8 Did you notice anything unusual about the stop signs during the driving scenarios? (1)Yes (2)Not sure (3)No
Q9 If you selected Yes or Not sure above, can you describe what you may noticed?
Q10 How would you describe your overall experience with the system? (1)Very Unpleasant (2)Somewhat Unpleasant (3)Neutral (4)Somewhat Pleasant (5)Very Pleasant
Q11 How easy was it for you to use and learn the system? (1)Very Difficult (2)Somewhat Difficult (3)Neutral (4)Somewhat Easy (5)Very Easy
Q12 How well do you think the simulation represented real-world driving conditions? (1)Not at all (2)Slightly (3)Moderately (4)Mostly (5)Completely
Q13 Would you recommend this VR system to others for driver training sessions? (1)Yes (2)Not Sure (3)No

Video-based Platform Questions

Q14 Please watch this video and answer the following questions:
Q15 On a scale from 1 to 5, how well could you notice the stop sign in the video scenario? (1)Not Aware at All (2)Slightly Aware (3)Moderately Aware (4)Mostly Aware (5)Completely Aware
Q16 On a scale from 1 to 5, how different is the first/second/third stop sign compared with a standard stop sign? (1)They are same (2)Slightly different (3)Moderately different (4)Mostly different (5)Completely different
Q17 How would you rate the realism of the video scenario in presenting the autonomous vehicle situation? (1)Very Fake (2)Somewhat Fake (3)Neutral (4)Somewhat Real (5)Very Real
Q18 Would you recommend this video system to others for driver training sessions? (1)Yes (2)Not Sure (3)No