Title

Risk Assessment for Security Threats and Vulnerabilities of Autonomous Vehicles

Authors
Chakraborty, Trishna
Chen, Qi Alfred, PhD

Publication Date
2024-04-01

DOI
10.7922/G2N29V87

RESEARCH REPORT

Institute of
Transportation
Studies

UNIVERSITY
OF
CALIFORNIA

Risk Assessment for Security Threats and Vulnerabilities of Autonomous Vehicles

Trishna Chakraborty, Ph.D. Student, Department of Computer Science, University of California, Irvine
Qi Alfred Chen, Ph.D., Assistant Professor, Department of Computer Science, University of California, Irvine

April 2024

California
Resilient and Innovative
Mobility Initiative

Report No.: UC-ITS-RIMI-5B-03 | DOI: 10.7922/G2N29V87

Technical Report Documentation Page

1. Report No.
UC-ITS-RIMI-5B-03

2. Government Accession No.
N/A

3. Recipient's Catalog No.
N/A

4. Title and Subtitle
Risk Assessment for Security Threats and Vulnerabilities of Autonomous Vehicles

5. Report Date
April 2024

6. Performing Organization Code
ITS-Irvine

7. Author(s)
Trishna Chakraborty, https://orcid.org/0000-0002-8791-6956
Qi Alfred Chen, Ph.D., https://orcid.org/0000-0003-0316-9285

8. Performing Organization Report No.
N/A

9. Performing Organization Name and Address
Institute of Transportation Studies, Irvine
4000 Anteater Instruction and Research Building
Irvine, CA 92697

10. Work Unit No.
N/A

11. Contract or Grant No.
UC-ITS-RIMI-5B-03

12. Sponsoring Agency Name and Address
The University of California Institute of Transportation Studies
www.ucits.org

13. Type of Report and Period Covered
Final Report (July 2022 – October 2023)

14. Sponsoring Agency Code
UC ITS

15. Supplementary Notes
DOI:10.7922/G2N29V87

16. Abstract

Autonomous vehicles (AVs) heavily rely on machine learning-based perception models to accurately interpret their surroundings. However, these crucial perception components are vulnerable to a range of malicious attacks. Even though individual attacks can be highly successful, the actual security risks such attacks can pose to our daily life are unclear. Various factors, such as lack of stealthiness, cost-effectiveness, and ease of deployment, can deter potential attackers from employing certain attacks, thereby reducing the actual risk. This research report presents the first quantitative risk assessment for physical adversarial attacks on AVs. The specific focus is on attacks on AV's perception components due to their highly critical function and representation in existing research. The report defines the daily-life risk as the likelihood that a given type of attack will be employed in real life and the authors develop a problem-specific risk scoring system and accompanying metrics. They perform an initial evaluation of the proposed risk assessment method for all the reported attacks on AVs from 2017 to 2023. They quantitatively rank the daily-life risks posed by each of eight different categories of attacks s and find three attacks with the highest risks: 2D printed images, 2D patches, and coated camouflage stickers, which deserve more focused attention for potential future mitigation strategy development and policy making.

17. Key Words
Autonomous vehicles, machine learning, cybersecurity, risk assessment, visual texture recognition

18. Distribution Statement
No restrictions.

19. Security Classification (of this report)
Unclassified

20. Security Classification (of this page)
Unclassified

21. No. of Pages
23

22. Price
N/A

Form Dot F 1700.7 (8-72) Reproduction of completed page authorized

About the UC Institute of Transportation Studies

The University of California Institute of Transportation Studies (UC ITS) is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

The California Resilient and Innovative Mobility Initiative

The California Resilient and Innovative Mobility Initiative (RIMI) serves as a living laboratory—bringing together university experts from across the four UC ITS campuses, policymakers, public agencies, industry stakeholders, and community leaders—to inform the state transportation system's immediate COVID-19 response and recovery needs, while establishing a long-term vision and pathway for directing innovative mobility to develop sustainable and resilient transportation in California. RIMI is organized around three core research pillars: Carbon Neutral Transportation, Emerging Transportation Technology, and Public Transit and Shared Mobility. Equity and high-road jobs serve as cross-cutting themes that are integrated across the three pillars.

Acknowledgments

This study was made possible with funding received by the University of California Institute of Transportation Studies from the State of California through a one-time General Fund allocation in the 2021 State Budget Act for the Resilient and Innovative Mobility Initiative. The authors would like to thank the State of California for its support of university-based research, and especially for the funding received for this project.

Disclaimer

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated under the sponsorship of the State of California in the interest of information exchange. The State of California assumes no liability for the contents or use thereof. Nor does the content necessarily reflect the official views or policies of the State of California. This report does not constitute a standard, specification, or regulation.

Institute of Transportation Studies | UNIVERSITY OF CALIFORNIA

Risk Assessment for Security Threats and Vulnerabilities of Autonomous Vehicles

Trishna Chakraborty, Ph.D. Student, Department of Computer Science, University of California, Irvine
Qi Alfred Chen, Ph.D., Assistant Professor, Department of Computer Science, University of California, Irvine

April 2024

California
Resilient and Innovative
Mobility Initiative

Report No.: UC-ITS-RIMI-5B-03 | DOI: 10.7922/G2N29V87

Table of Contents

Table of Contents

List of Tables

Executive Summary

Executive Summary

Since the initial reporting in 2017 of the first physical attack on an Autonomous Vehicle (AV), various attack methods have been discovered targeting critical AI components in AVs, especially those involving how the vehicle perceives its surroundings. Although prior studies have extensively considered the technical attributes of these attacks (e.g., attack effectiveness), the security risks such attacks can pose to our daily life are still uncertain. For example, laser attacks have been shown to have a high attack success rate using a physical setup consisting of a function generator, oscilloscope, amplifier, photodiode, laser diode, lenses, camera tracking system, and a pan-tilt system. However, such a physical setup can cost the attacker around $10,000, is hard to carry and is visible to everyone on the street, making it less likely that an attacker will employ it in the real world. Thus, an understanding of the technical attributes of such attacks (lack of stealthiness, cost-effectiveness, and ease of deployment) does not directly translate to an understanding of the risks they can pose to our daily life, which is important to a wide range of potentially-affected parties including individuals, developers, stakeholders, and policymakers.

In this report, we present the results of the first quantitative risk assessment of physical attacks on AVs. We specifically focus on AVs' ability to perceive the roadway and objects on it because of its critical nature in AV operations and the availability of existing research. We assess the likelihood that a given attack method will be employed in real life. We develop a problem-specific risk scoring system and metric design, and employ them in an initial evaluation. We envision that a risk assessment such as we propose should be a fundamental component in future attack studies to enable a more comprehensive understanding the potential impact of future AV attacks on our daily lives.

In summary, our study accomplished the following:

- We performed the first quantitative risk assessment of physical attacks on AVs. We specifically focused on perception attacks considering their high criticality and prevalence in existing research.
- We defined the daily-life risk as the likelihood that a given attack method will be employed in real life, and developed a problem-specific risk scoring scheme and metric design.
- We performed an initial evaluation of the proposed risk assessment method for all the reported attacks from 2017 to 2023 in the physical perception domain of AVs. We quantitatively ranks the daily-life risks posed by each of eight attack categories, and identified three which posed the highest risks: 2D printed images (i.e., printing images/photos of critical road objects such as cars and pedestrians), 2D patches (i.e., attaching malicious stickers/posters to critical road objects), and coated camouflage stickers (i.e., painting/adding camouflage coat to cover the entire critical road objects), which deserve more focused attention for developing potential future mitigation strategies and policy making.

Contents

Introduction

Autonomous Vehicles (AVs) have become increasingly accessible on today's roads, their development propelled by ongoing technological advancements. These vehicles, designed to function independently without continuous active human intervention, heavily rely on the vehicle's *perception system*—its ability to detect the roadway (e.g., lane markings) and objects in its path—using external sensors, including cameras, light detection and ranging (LiDAR), and radar, to comprehensively perceive their surroundings. In order to process the extensive data collected by these sensors, AVs utilize sophisticated machine-learning algorithms, enabling them to interpret and respond to dynamic environments in real time. However, the use of these machine learning models also introduces a critical vulnerability: the potential for adversarial attacks. These attacks manipulate the input data received by the vehicle's sensors in imperceptible ways to the human eye, resulting in deceptive alterations that can compromise the vehicle's perception system. Such malicious interventions pose a significant threat to the safety of autonomous vehicles, potentially leading to system malfunctions and hazardous collisions.

For the future of autonomous vehicles, a significant area of concern revolves around physical adversarial attacks. These are instances where real-world objects are physically altered, like affixing a patch to a stop sign, resulting in the vehicle's perception module incorrectly identifying it as a speed limit sign. Given their high impact on safety-critical systems, these physical adversarial attacks hold more importance in autonomous vehicle security compared to other digital domain attacks.

Since the first reported attack [1] in 2017, numerous methods, or scenarios, known in cybersecurity circles as *attack vectors*, that can be exploited to disrupt critical artificial intelligence (AI) components of AVs have been identified, notably those targeting the car's perception system. Previous research studies have thoroughly examined these AV *perception attacks*, including an analysis of the attacker's capabilities, diverse attack vectors, and the vulnerability of distinct modules within the perception system.

Although prior studies have extensively considered the technical attributes of these attacks (e.g., attack effectiveness), the security risks such attacks can pose to our daily lives are still obscure. It is crucial to differentiate the concept of risk assessment from the attack technique itself. An attack typically exploits a vulnerability, whereas the associated risk is defined as the likelihood of a real-world adversary utilizing this particular attack vector, taking into account various factors including the attack cost, deployability, stealthiness, and more. For example, laser attacks [2] have been shown to have a high success rate with a physical setup consisting of a function generator, oscilloscope, amplifier, photodiode, laser diode, lenses, camera-tracking system, and a pan-tilt system. However, to carry out such an attack could cost the attacker around S10,000, the equipment would be hard to carry, and the devices would be visible to everyone on the street, making it less likely for an attacker to employ it in the real world, resulting in a lower risk that such an attack would actually occur. Thus, an understanding of the technical attributes of such attacks does not directly translate to an understanding of the risks they can pose to our daily lives, which is highly important to a wide range of potentially affected individuals, developers, stakeholders, and policymakers as they go about making decisions.

Therefore, in this report we present the first quantitative risk assessment of the feasibility of various AV attack scenarios. We specifically focus on perception attacks considering their highly critical nature and also the availability of existing research. Our study defined the *daily-life risk* as the likelihood that a given attack vector will be employed in real life, then designed a problem-specific risk scoring system and developed metrics to measure the risks. Using this methodology we conducted a systematic risk assessment of all the reported attacks against AVs from 2017 to 2023, and found that three types of attacks—2D printed images, 2D patches, and coated camouflage stickers—posed the highest daily-life risk. We believe that such analyses and results can help individuals, developers, stakeholders, and policymakers in making decisions about deploying and adopting AV technology. We also envision that risk assessments, such as the one we propose, will become a fundamental component in future attack studies to enable a more comprehensive picture of these attack's impact on our daily life.

Structure of the Report

The rest of this report is structured as follows. In the Background section, we provide an overview of existing attacks on AV perception and delve into prior research related to risk assessment. In the Risk Assessment Methodology section, we outline the threat model and its scope. Subsequently, we describe the four steps of the risk assessment methodology, followed by the identification of seven risk metrics. The Risk Assessment Results section entails an initial quantitative evaluation of the proposed risk assessment method, including the ranking of daily-life risks posed by each of the eight attack categories. Finally, we outline potential future research directions and present our conclusions in the final section.

Background

Existing Attacks

Recent research has demonstrated the susceptibility of machine-learning models to adversarial attacks, raising concerns about the robustness of AI systems in AVs. Given the critical reliance of AVs on these models at various stages in driving the vehicle, such as perception, localization, prediction, and planning, the potential vulnerability of AVs to a multitude of adversarial attacks becomes a pertinent concern.

According to recent findings highlighted in Cao, et al. [3], more than 86 percent of the reported AV attacks targeted the AV perception system. Notably, attacks on cameras account for approximately 60 percent of these cases [3]. These camera-based perception attacks can occur during both training and inference times. During training—the process of teaching an AI system to perceive, interpret and learn from data—the input data can be contaminated with specific triggers, known as poisoning or backdoor attacks. During inference—the process of running live data through a trained AI model to make a prediction or solve a task—the input images can be digitally or physically tampered with.

In these types of digital attacks, the attacker may introduce noise into the input data (such as point clouds in LiDAR or images in cameras) to mislead the AV's perception model. On the other hand, in a physical attack, the attacker may employ a physical setup as the attack vector. For instance, this could involve manipulating the input sensor data, such as by shooting laser at the camera to generate additional points in a point cloud (i.e., set of data points in a 3D coordinate system generated by LiDAR sensing results) or placing a physical patch (e.g., a malicious paper or sticker attached to a road object such as cars and pedestrians) on an image to change the appearance of an object.

An attacker's objectives can be classified into three distinct categories: hiding, creation, or misclassification. In hiding attacks, the goal is to make an object disappear from the perception system's view. In creation attacks, the attacker aims to introduce a new object that is not actually present. In misclassification attacks, the detected object is intentionally mislabeled, leading to potentially dangerous consequences (e.g., the AV brakes suddenly after a car is misclassified as a pedestrian).

We performed an exhaustive survey of physical AV perception attacks from 2017 to 2023, reported in 56 research papers, and identified eight types of attack vectors in total: 2D Printed Images, 2D Patch, Coated Camouflage Stickers, 3D Object, Light Projection, Laser/Infrared (IR) Light, Acoustic Signal, and Electromagnetic Interference (EMI).

Machine learning models relying solely on camera data face challenges in distinguishing between 2D and 3D objects. Consequently, a seemingly harmless 2D poster of a person placed in front of a vehicle could be erroneously identified as an actual human [4], triggering an abrupt halt. In this scenario, an attacker does not require advanced engineering skills so anyone can carry out the attack. Many of the proposed attacks operate using gradients (i.e., a mathematical guidance in the optimization process)—which measures how well the model is performing—where the attacker calculates the gradient concerning the input and manipulates it to optimize the attacker's specific objective function. One prominent attack method of this type is the 2D patch attack [5], where an attacker strategically places a skillfully crafted patch on an object, leading the machine learning model to misclassify it. Recent developments have focused on making such patches appear more naturalistic. Coated camouflage stickers [6] that fully cover vehicles have been devised, causing objects to be misclassified from any viewpoint or distance.

Moreover, instead of merely attaching a patch onto a benign object, attackers can deform the shape of an object to render the entire entity malicious. These techniques [7] are employed to deceive both cameras and LiDAR systems. Another noteworthy attack vector involves light projections, where an attacker projects light in the shape of a human [8] in front of the camera or directs lights towards the lens to create a flare effect [9].

Similarly, in the case of laser or infrared light, the attacker can exploit the rolling shutter effect of the camera by shooting laser/lights to the camera [10] or create alterations in the points of the LiDAR point cloud by using laser shooting to change the distance measurement for the 3D points [2]. With the acoustic signal attack vector [11], the attacker aims to confuse the system’s object detector into misclassifying objects by directing acoustic signals towards the camera to affect the image stabilizer sensor. Additionally, the use of electromagnetic interference (EMI) presents an intriguing attack vector [12]. Here, the attacker generates an electromagnetic field which induces a voltage on the LiDAR's internal Time-of-Flight (TOF) circuits, leading to disruptions in the resulting point cloud.

Risk Assessment

A few studies have delved into risk assessment with regard to AVs. The definition of risk in this context covers a wide range of meanings. Many research papers provide their own unique definition of risk, subsequently presenting research based on their specific understanding. This variability in defining risk is to be expected, given that the interpretation of risk is influenced by the specific context under consideration.

Certain risk assessment studies [13,14] have focused on predicting the probability of an accident for AVs equipped with collision avoidance systems. Another line of research has leveraged risk assessment techniques [15,16] for evaluating AV decision-making and planning while driving. However, only a limited number of works have concentrated on conducting risk assessments of potential attacks on AVs. For instance, ML-Doctor [17] offers a comprehensive risk assessment of inference attacks, conducting an extensive measurement study across five model architectures and four datasets, encompassing both attack and defense strategies. Similarly, another study has developed a systematic framework for risk assessment, specifically addressing the vulnerability of Unmanned Aerial Vehicles (UAVs) to potential cyber-attacks [18]. These endeavors highlight the growing recognition of the significance of robust risk assessment methodologies in ensuring the safety and security of AV systems in the face of emerging threats.

Risk Assessment Methodology

Scope and Threat Model

Our threat model focuses on physical attacks on the AI components responsible for AV perception that do not require access to the AV system internal components. We examined research papers from top-tier research venues related to Security, Computer Vision (CV), Machine Learning (ML), and Artificial Intelligence (AI). Our search process involved an exhaustive review of publications from 2017 to 2023, ensuring comprehensive coverage of the literature related to our research scope.

An AV perception module is an intricate system encompassing various tasks, such as object detection, object classification, semantic segmentation, object tracking, lane detection, and traffic light detection. The sensors employed include cameras, LiDAR, and radar. Attacks may be launched against one of these devices though attacks on multi-sensor fusion systems which combine observations from several different sensors to provide a more complete description of driving environment.

Our research scope takes into account all forms of physical attacks, such as altering the texture, shape, and position of objects in the physical world. Additionally, we address sensor-specific attacks, including LiDAR spoofing, radar spoofing, manipulation of Laser/IR light, and distortion of acoustic signals. These attacks cover a range from white-box to black-box, with the attacker's goals varying from hiding information to misclassifying sensor data, or even the creation of false observations.

Assessment Methodology

Our methodology for performing a thorough risk assessment involves four key steps.

1. Collecting existing attack vectors: We first analyzed relevant research within the scope of our specific threat model to identify existing attack vectors.
2. Analyzing attack launching workflow: In this step, we carefully examined the general workflow that an attacker might follow while attempting to execute an attack. Understanding each step of the attacker's process is crucial as each phase has a direct impact on the likelihood of the attacker's success.
3. Identifying risk metrics: By observing the workflow, we identify a set of risk metrics denoted as $R_i, i = 1 \dots n$. These metrics independently influence the likelihood of an attacker adopting a specific attack vector in real-life scenarios. Examples of these risk metrics could include the attacker’s cost, how easily the attack can be concealed, and the effort required to carry out the attack.
4. Calculating risk score: In this final step, we define a risk score of a given attack vector A as $Risk(A) = P(A|R_1, R_2 \dots R_n)$, where $P(.)$ is the likelihood for an attacker to employ A given the risk metrics for A. Since the risk metrics are designed to be independent in affecting $P(.)$, we can calculate Risk(.) using

$P(A|R_1, R_2 ... R_n)$ = $\prod_{i=1}^{n} \quad P(A|R_i)$. Each $P(A|R_i)$ can then be determined by the risk analyst based on the attack characteristics per risk metric.

Metric Design

Based on the characteristics of the attack vectors we analyzed, we identified the following seven risk metrics, R_i.

1. Attacker's Knowledge: This refers to the level of understanding an attacker has about the system they are targeting. It can be classified as white, grey, or black-box. White-box indicates a high level of knowledge, meaning the attacker has complete information about the system. Grey-box means the attacker has some information, but not all. Black-box implies minimal or no information about the system.
2. Effectiveness: This measures how well the attack achieves its intended goal. It is commonly reported as the Attack Success Rate, which indicates the percentage of successful attacks out of the total attempts.
3. Robustness: This refers to how well the attack performs under different environmental conditions. It accounts for the ability of the attack to work consistently in various scenarios, such as different weather conditions or lighting.
4. Deployability: This represents the level of effort and resources required to carry out an attack in a real-world situation. It considers factors like the time, materials, and expertise needed to execute the attack successfully.
5. Stealthiness: This indicates the level of visibility of the attack setup. A highly stealthy attack is one that is difficult to detect by security measures or monitoring systems, allowing it to operate undetected for an extended period.
6. Attacker's Cost: This primarily refers to the financial expenses incurred by the attacker during the planning and execution of the attack. It encompasses costs associated with acquiring necessary tools, resources, or personnel required for the attack.
7. Evaluation Level: This specifies the scope at which the evaluation of the attack is conducted. It can be at the component level, focusing on individual parts of the system; at the Simulation level, which involves testing in simulated environments; or at the AV level, examining the performance of the attack on the entire autonomous vehicle system.

Risk Assessment Results

A user study is required to obtain the actual value of $P(A|R_i)$,. However, as an initial trial, we defined the value of each to be at three normalized value levels: High (H), Medium (M), and Low (L) in order to make a subjective estimate. Positive risk metrics are directly proportional to the attacker's benefit, thus we subjectively assigned H=1, M=2/3, and L=1/3 for each positive metric (e.g., stealthiness). Negative risk metrics are inversely proportional to the attacker's benefit; thus the assigned values are opposite (H=1/3, M=2/3, L=1) for those metrics (e.g., attacker's cost).

Please note that while our assigned values are subjective, they accurately reflect the true property of $P(A|R_i)$, i.e., the higher the value the greater the attack vector is likely to be employed. These assignments provide a starting point for assessing risk factors, recognizing the subjectivity involved.

As an initial design, we began with the three most direct risk metrics out of the seven identified above: Deployability (D), Stealthiness (S), and Attacker's Cost (C).

$P(A|Deployability)$:

- High: This category denotes that the attack can be easily deployed with minimal effort. An example of such a highly deployable attack could involve the simple placement of a 2D patch or 2D printed image, requiring minimal resources or technical skill to execute.
- Medium: Attacks falling under this category require a moderate amount of effort to implement. For instance, applying a camouflage sticker over the entire vehicle would demand more time and resources compared to a simple patch.
- Low: This classification signifies that the deployment of the attack is challenging and demands a high level of skill or precision. An example could be an intricate laser shooting mechanism that requires precise aiming and advanced technical knowledge.

$P(A|Stealthiness)$:

- High: An attack categorized as highly stealthy implies that it can easily go unnoticed or undetected. For instance, the use of a 2D printed image could seamlessly blend with its surroundings, remaining inconspicuous to the observer.
- Medium: This category indicates that the attack possesses certain elements that could potentially be detected upon close inspection. An example of this could be the application of a camouflage sticker over an entire vehicle, which might have some detectable features upon careful observation.
- Low: Attacks falling into this category are easily detectable and can be quickly identified. For example, setting up a projector would be conspicuous and easily spotted by security or monitoring systems.

$P(A|Attacker's\ Cost)$:

- Low: Attacks falling under this category can be carried out with readily available and affordable materials. An example could be using a 2D printed image, which can be created using commonly accessible resources and tools.
- Medium: This classification indicates that the attack requires specialized resources or materials that might not be readily available. An example could be an Electromagnetic Interference (EMI) attack, which might demand specific equipment or expertise not commonly accessible.
- High: This category represents attacks that demand expensive materials or resources to execute. For instance, a complex 3D printing-based attack could require costly materials or advanced technology beyond the reach of ordinary means.

Our risk assessment results are shown in Table 1. Among the eight attack vectors, the top three highest-risk ones are 2D printed images, 2D patches, and coated camouflage stickers. The 2D printed image attack is ranked the highest daily-life risk since it is highly deployable (you just need a simple placement of a printed image), highly stealthy (it can easily be disguised as a road-side advertisement), and with low attack cost (you just need a 2D printer). Thus, for both defensive development and security-related policymaking, combating this attack vector should be prioritized. The 2D patch attack comes in second, since it is still highly deployable and with low attack cost, but it is slightly less stealthy than the 2D printed image since it will exhibit a noise-like pattern instead of being completely natural looking. The coated camouflage attack is ranked third; it still has low attacker's cost, but suffers from the same suspicious pattern problem as the 2D patch attack. When compared to a 2D printed image attack and a 2D patch attack, it requires greater effort since the malicious pattern has to cover the entire surface of the target object, instead of just a small patch area. The total order of attack vectors, ranked from the highest daily-life risk to the lowest, is: 2D printed image, 2D patch, Coated camouflage stickers, 3D Object/mesh, EMI, and the remaining three which have the same lowest risk score (Light projection, Laser/IR light, and Acoustic Signal).

Table 1. Initial risk assessment results.

Attack Vectors (ranked by the risk scores)	Metrics			Risk(.)
	Deployability	Stealthiness	Attacker's Cost	
2D Printed Image	High (1)	High (1)	Low (1)	1
2D Patch	High (1)	Medium (2/3)	Low (1)	0.67
Coated Camouflage Sticker	Medium (2/3)	Medium (2/3)	Low (1)	0.44

Attack Vectors (ranked by the risk scores)	Metrics			Risk(.)
	Deployability	Stealthiness	Attacker’s Cost	
3D Object/Mesh	High (1)	High (1)	High (1/3)	0.33
Electromagnetic Interference	Low (1/3)	Low (1/3)	Medium (2/3)	0.07
Projection	Low (1/3)	Low (1/3)	High (1/3)	0.04
Laser/IR Light	Low (1/3)	Low (1/3)	High (1/3)	0.04
Acoustic Signal	Low (1/3)	Low (1/3)	High (1/3)	0.04

Conclusion

The surge in attacks targeting AV perception capabilities demands a comprehensive examination. Our extensive survey has analyzed 56 research papers, identifying eight distinct attack types. While technical attributes indicate the potential efficacy of various attacks, real-world implications reveal a more nuanced picture. To establish a comprehensive understanding of risk, we define it as the likelihood of an attack's real-world deployment by adversaries. Seven essential risk metrics—Attacker's Knowledge, Effectiveness, Robustness, Deployability, Stealthiness, Attacker's Cost, and Evaluation level—shape our risk assessment framework. In our initial assessment, we concentrate on three primary risk metrics—Deployability (D), Stealthiness (S), and Attacker's Cost (C). Among the eight identified attack vectors, our analysis highlights 2D printed images, 2D patches, and coated camouflage stickers as the top three high-risk vectors, which deserve more focused attention for potential future mitigation strategy development and policy making.

While acknowledging the inherent challenges in achieving a comprehensive checklist covering all possible risks, in the future we plan to refine our current design and improve the comprehensiveness and efficacy of our risk assessment framework. First and foremost, we plan to incorporate the remaining four identified risk metrics, ensuring a more comprehensive and holistic evaluation of potential risks. This expansion will enable a more nuanced understanding of the multifaceted dimensions of risk, allowing for a more accurate and inclusive risk assessment process.

Furthermore, we intend to employ a more standardized and data-driven approach alongside a comprehensive user study to assign risk scores in a more objective and evidence-based manner.

Lastly, we aim to broaden the scope of our research to encompass a more extensive range of AI components and threat models, with a particular emphasis on addressing potential cyberattacks. By extending our focus to incorporate these critical aspects, we aspire to develop a more comprehensive and adaptable risk assessment framework that accounts for the evolving landscape of AI security threats and vulnerabilities.

References

[1] Brown, T.B., Mané, D., Roy, A., Abadi, M., and Gilmer, J., 2017. Adversarial patch. arXiv preprint arXiv:1712.09665.

[2] Shen, J., Wang, N., Wan, Z., Luo, Y., Sato, T., Hu, Z., Zhang, X., Guo, S., Zhong, Z., Li, K., and Zhao, Z., 2022. Sok: On the semantic AI security in autonomous driving. arXiv preprint arXiv:2203.05314.

[3] Cao, Y., Bhupathiraju, S. H., Naghavi, P., Sugawara, T., Mao, Z. M., and Rampazzi, S., 2023. You can't see me: Physical removal attacks on {lidar-based} autonomous vehicles driving frameworks. 32nd USENIX Security Symposium (USENIX Security 23) (pp. 2993-3010).

[4] Gomez-Donoso, F., Cruz, E., Cazorla, M., Worrall, S., and Nebot, E., 2020, July. Using a 3D CNN for rejecting false positives on pedestrian detection. 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-6). IEEE.

[5] Song, D., Eykholt, K., Evtimov, I., Fernandes, E., Li, B., Rahmati, A., Tramer, F., Prakash, A., and Kohno, T., 2018. Physical adversarial examples for object detectors. 12th USENIX Workshop on Offensive Technologies (WOOT 18).

[6] Duan, Y., Chen, J., Zhou, X., Zou, J., He, Z., Zhang, J., Zhang, W., and Pan, Z., 2021. Learning coated adversarial camouflages for object detectors. arXiv preprint arXiv:2109.00124.

[7] Xiao, C., Yang, D., Li, B., Deng, J., and Liu, M., 2019. Meshadv: Adversarial meshes for visual recognition. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6898-6907).

[8] Nassi, B., Mirsky, Y., Nassi, D., Ben-Netanel, R., Drokin, O., and Elovici, Y., 2020, October. Phantom of the ADAS: Securing advanced driver-assistance systems from split-second phantom attacks. Proceedings of the 2020 ACM SIGSAC conference on computer and communications security (pp. 293-308).

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

[10] Yan, C., Xu, Z., Yin, Z., Mangard, S., Ji, X., Xu, W., Zhao, K., Zhou, Y., Wang, T., Gu, G., and Nie, S., 2022. Rolling colors: Adversarial laser exploits against traffic light recognition. 31st USENIX Security Symposium (USENIX Security 22) (pp. 1957-1974).

[11] Ji, X., Cheng, Y., Zhang, Y., Wang, K., Yan, C., Xu, W., and Fu, K., 2021, May. Poltergeist: Acoustic adversarial machine learning against cameras and computer vision. 2021 IEEE Symposium on Security and Privacy (SP) (pp. 160-175). IEEE.

[12] Bhupathiraju, S.H.V., Sheldon, J., Bauer, L.A., Bindschaedler, V., Sugawara, T., and Rampazzi, S., 2023, May. EMI-LiDAR: Uncovering vulnerabilities of LiDAR sensors in autonomous driving setting using electromagnetic interference. Proceedings of the 16th ACM Conference on Security and Privacy in Wireless and Mobile Networks (pp. 329-340).

[13] Li, G., Yang, Y., Zhang, T., Qu, X., Cao, D., Cheng, B., and Li, K., 2021. Risk assessment based collision avoidance decision-making for autonomous vehicles in multi-scenarios. Transportation Research Part C: Emerging Technologies, 122, p. 102820.

[14] Katrakazas, C., Quddus, M., and Chen, W.H., 2019. A new integrated collision risk assessment methodology for autonomous vehicles. Accident Analysis & Prevention, 127, pp. 61-79.

[15] Wang, Y., Wang, C., Zhao, W., and Xu, C., 2021. Decision-making and planning method for autonomous vehicles based on motivation and risk assessment. IEEE Transactions on Vehicular Technology, 70(1), pp.107-120.

[16] Wang, H., Lu, B., Li, J., Liu, T., Xing, Y., Lv, C., Cao, D., Li, J., Zhang, J., and Hashemi, E., 2021. Risk assessment and mitigation in local path planning for autonomous vehicles with LSTM based predictive model. IEEE Transactions on Automation Science and Engineering, 19(4), pp. 2738-2749.

[17] Liu, Y., Wen, R., He, X., Salem, A., Zhang, Z., Backes, M., De Cristofaro, E., Fritz, M., and Zhang, Y., 2022. ML-Doctor: Holistic risk assessment of inference attacks against machine learning models. 31st USENIX Security Symposium (USENIX Security 22) (pp. 4525-4542).

[18] Hartmann, K. and Steup, C., 2013, June. The vulnerability of UAVs to cyber attacks-An approach to the risk assessment. 2013 5th international Conference on Cyber Conflict (CYCON 2013) (pp. 1-23). IEEE.