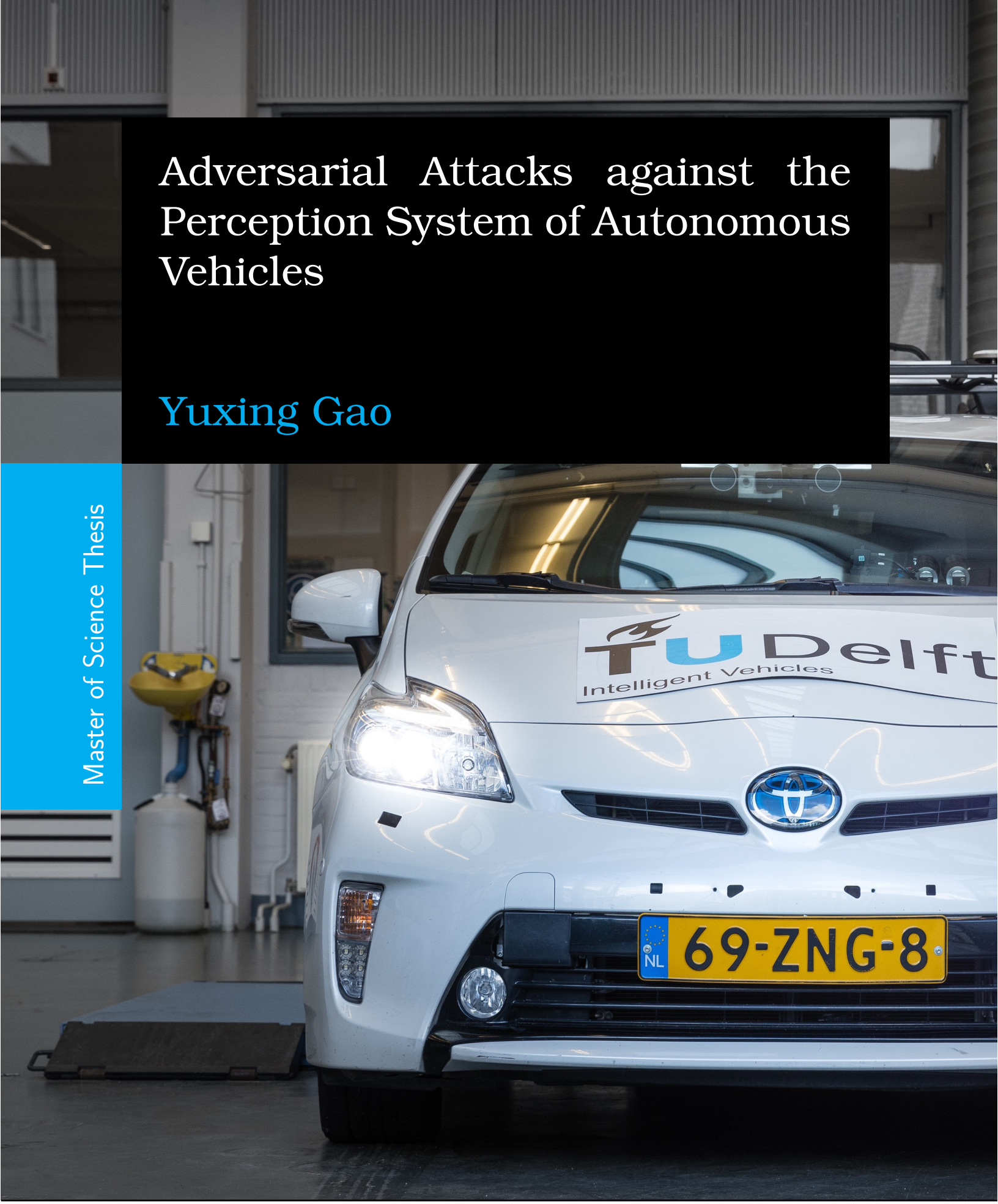
Adversarial Attacks against the Perception System of Autonomous Vehicles

Yuxing Gao

Master of Science Thesis



TU Delft Delft University of Technology

Delft Center for Systems and Control

Adversarial Attacks against the Perception System of Autonomous Vehicles

Master of Science Thesis

For the degree of Master of Science in Systems and Control at Delft University of Technology

Yuxing Gao (5484782)

Supervisors: Luca Laurenti and Arkady Zgonnikov
Daily supervisors: Shubham Koyal and Koen Boer

December 13, 2023

Faculty of Mechanical, Maritime and Materials Engineering (3mE) · Delft University of Technology

The work in this thesis was supported by RDW. Their cooperation is hereby gratefully acknowledged.

TABLE OF CONTENTS

Adversarial Attacks against the Perception System of Autonomous Vehicles

Yuxing Gao

***Abstract*—The rapid advancement in autonomous driving technology underscores the importance of studying the fragility of perception systems in autonomous vehicles, particularly due to their profound impact on public transportation safety. These systems are of paramount importance due to their direct impact on the lives of passengers and pedestrians. Additionally, their reliability can be easily compromised given the complexity and unpredictability of driving environments. However, current research and existing regulations often fail to adequately address the adversarial robustness of autonomous vehicle perception systems. This thesis delves into the adversarial robustness of camera-based perception systems of autonomous vehicles. Our research concentrates on developing and implementing evasion attacks that use black-box gradient estimation, as well as physical attacks in traffic sign detection and classification systems. Our findings indicate that even minor perturbations can impact the accuracy of these systems, leading to detection and classification errors. This finding highlights a critical vulnerability in the perception system's robustness against adversarial attacks. Moreover, the study extends to assess the transferability of adversarial examples across diverse perception models. Our results also expose significant gaps in the current regulatory frameworks of autonomous vehicles, necessitating the establishment of more rigorous and comprehensive safety standards.**

1. Introduction

1.1. Research Motivation

The rapid advancement of machine learning has significantly enhanced various aspects of our lives. However, the integration of machine learning into critical areas such as infrastructure, public safety, and personal privacy introduces significant security concerns. A key concern is their vulnerability and robustness, and how they perform when individuals or groups try to intentionally exploit machine learning models for harmful purposes [21, 30]. This issue is especially crucial in the context of autonomous vehicles (AVs), where the stakes involve public safety and the potential for significant harm.

Autonomous driving, an application of machine learning technology, aims to transform transportation by allowing vehicles to operate without human drivers. Despite the potential benefits of autonomous driving, concerns surrounding its safety, security, reliability and ethical implications exist. There have already been some accidents related to autonomous driving technology [36]. The first fatal accident caused by autonomous vehicles occurred in 2018, a pedestrian was killed by an Uber test vehicle [23, 13], which was operated in self-driving mode with a human safety backup driver. Unfortunately, this was not the only accident. Until January 15, 2023, carmakers had submitted 419 reports of autonomous vehicle crashes to the National Highway Traffic Safety Administration (NHTSA) [28]. 263 of these crashes were in Level 2 ADAS cars, while 156 were in fully autonomous vehicles equipped with Automated Driving Systems (ADS). Of the 419 crashes, there were 18 fatalities, all of which were in Level 2 Advanced Driver Assistance Systems (ADAS) vehicles. These accidents highlight the need for careful examination of AV safety and reliability.

This master thesis delves into the realm of AI security, with a particular emphasis on how adversarial perturbations on inputs can cause false predictions in machine learning models. The study seeks to uncover new insights into the fragility and vulnerability of these systems against sophisticated cyber threats. Furthermore, the findings of this study are expected to contribute valuable information to the ongoing development of regulations and policies concerning AVs. By providing a comprehensive analysis of adversarial attack strategies and their impact on AV perception systems, this research will offer guidance that ensures the secure and responsible deployment of autonomous driving technologies.

1.2. Research Gaps

In the landscape of autonomous vehicle (AV) technologies, numerous questions arise that challenge the current knowledge and push the boundaries of what we understand about these systems. As AV technologies continue to be integrated into our transportation systems, it becomes increasingly important to address

these questions, not only to advance the technology but also to ensure the safety, reliability, and security of these systems. The following research gaps have been identified as crucial in our understanding and guiding future research in the domain of AV perception systems:

1) Black Box Gradient Estimation Attacks in AV Perception Systems: Research into gradient-based adversarial attacks on perception systems, as investigated by Sun et al. (2020) and Cao et al. (2019), has revealed a general vulnerability in LiDAR-based systems or autonomous vehicles [38, 10]. The robustness of multi-sensor systems, particularly those utilizing sensor fusion algorithms has been evaluated in studies by Hallyburton et al. (2022), Tu et al. (2021), and Abdelfattah et al. (2021). These studies collectively confirm that multi-sensor systems are also susceptible to adversarial attacks [19, 42, 1]. Specific focus on the vulnerability of car detection systems was provided by Abdelfattah et al. (2021), while Boloor et al. (2020) examined end-to-end vision-based perception systems [1, 9]. Regarding physical attacks, Evtimov et al. (2017) and Nassi et al. (2020) have demonstrated the potential harm of these attacks through real-world experiments [16, 26].

These related studies underscore that while there is extensive research on adversarial attacks targeting perception systems, there remains a notable gap in studies specifically addressing black-box adversarial attacks in the context of traffic sign detection and classification systems. Additionally, in the realm of physical attacks, although real-world experiments have highlighted the susceptibility of models to such attacks, there is a noticeable deficiency in simulation research that leverages large-scale datasets.

2) Transferability of Adversarial Examples: The concept of transferability of adversarial examples has been explored in various studies. Szegedy et al. (2013) first highlighted this phenomenon, demonstrating the transferability of these examples across models with different training datasets and hyperparameters [39]. Subsequent research by Papernot et al. (2016) and Liu et al. (2016) further confirmed that adversarial examples crafted on a substitute model could successfully transfer to other models, including DNNs, SVMs, and kNNs [29, 24]. Additionally, Demontis et al. (2019) identified three main factors influencing transferability, applicable to both poisoning and evasion attacks [14].

Understanding the transferability of adversarial attacks across different models can shed light on the inherent and systematic vulnerabilities across various models of AV perception systems. This is fundamental in developing more robust and systematic defence mechanisms against a spectrum of adversarial strategies. However, there is a lack of studies examining the effectiveness of transferability on perception systems of autonomous vehicles.

3) Regulatory Framework Regarding Adversarial Attacks on AVs: As Autonomous Vehicle (AV) technologies continue to progress rapidly, a pivotal and urgent research question arises: How must regulatory frameworks be reformed to comprehensively address the external threats posed by adversarial attacks, which are currently overlooked in the context of AV systems? The existing regulatory landscape for autonomous vehicles primarily concentrates on assessing the safety and security of the vehicle itself and its components. This narrow focus significantly neglects the broader spectrum of external threats in the driving environment, including adversarial attacks that can critically undermine AV operations.

1.3. Main Contributions

The main contributions of this thesis are summarized as follows:

1) Exploration of Black-Box Adversarial Evasion Attacks in Perception Systems: The thesis investigates black-box adversarial evasion attacks, focusing on their impact on the vulnerability of camera-based perception systems in autonomous vehicles, specifically in the context of traffic sign detection and classification.

2) Exploration of Physical Threats in Perception Systems: The thesis includes an examination of physical attacks, which manipulate the physical environment to mislead autonomous vehicle systems.

3) Transferability of Adversarial Examples: The thesis investigates how effectively adversarial examples, once developed for a specific perception model, can be applied to different models.

4) Methodological Approach and Experimentation: The thesis outlines the methodological framework and implementation of adversarial attacks, including both black-box and physical types. It elaborately describes the development, testing, and evaluation of these attacks on traffic sign detection systems.

5) Regulatory Perspectives on Autonomous Vehicles: The research provides valuable insights into the current regulatory framework for the safety and security of autonomous vehicles. It identifies significant gaps in safety standards and regulations, especially regarding the robustness of machine learning models against adversarial attacks.

1.4. Outline

The thesis consists of the following parts:

- Chapter 1 provides an introduction to the thesis, including the research motivation, research gaps, main contributions and structural outline.
- Chapter 2 introduces the background knowledge about this topic.
- Chapter 3 gives the detailed explanation of the methodology.
- Chapter 4 delves into the implementation of attacks, and analysis of the simulation results, with a focus on classification failures.
- Chapter 5 presents the summarization of the research conclusions and implications and provides recommendations for future research directions.
- Appendices give supplementary information about the research, including source code.

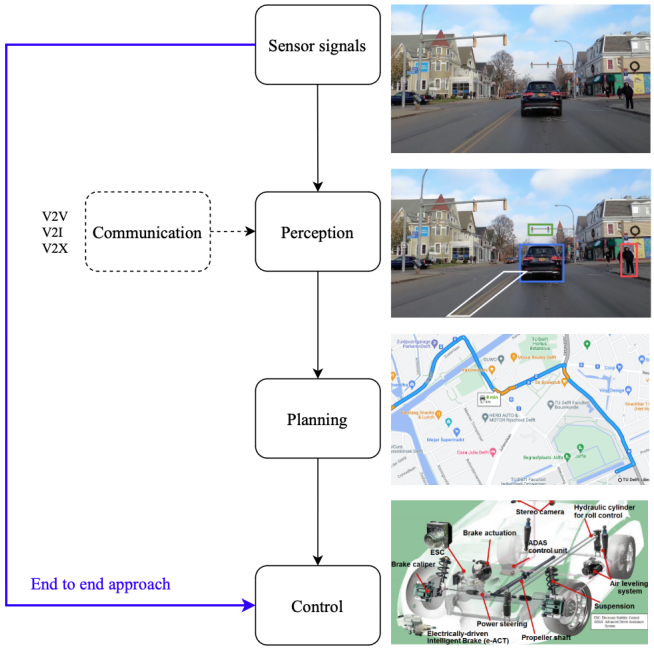


Fig. 1: Autonomous driving system structure

2. Background

2.1. Autonomous Vehicles and Perception Systems

NHTSA [27] defines autonomous vehicles as "vehicles that are capable of driving themselves without human intervention or supervision." The most common structures of autonomous driving are modular pipelines and end-to-end approaches, they are illustrated in Figure 1. A modular autonomous driving system comprises three main components: perception, planning, and control [48, 18]. Each component plays a crucial role in the overall functionality of the autonomous vehicle.

The perception system integrates and processes data from sensors, and it is essential for recognizing, segmenting, tracking, and predicting objects and entities around the vehicle. It employs advanced machine learning algorithms to interpret sensor data from diverse sources, enabling the vehicle to perceive the environment correctly.

Autonomous vehicles use a combination of external and internal state sensors to create a comprehensive perception of the driving environment [47]. External state sensors, such as cameras, Lidar, and Radar [20, 4], gather data about the external environment while internal state sensors, like IMUs [4] and GPS [4, 34], track the vehicle's internal dynamics. The fusion of these sensor inputs is crucial for obtaining a complete understanding of the surroundings.

The perception systems integrate data from various sensors, with cameras being pivotal in providing detailed visual information. Cameras serve as the vehicle's eyes and can uniquely interpret complex information on the surface of objects such as text on traffic signs.

2.2. Adversarial Attacks

Adversarial attacks in machine learning represent a sophisticated form of cyber threat, where adversaries deliberately introduce manipulations to influence the behaviour of the models [17, 5]. These attacks target vulnerabilities in machine learning algorithms and can result in model confidentiality, integrity, and availability being compromised [35, 40, 31]. Adversarial attacks can be divided into two main types: poisoning attacks [7] and evasion attacks [8, 37, 45].

In poisoning attacks, the adversary interferes with the model's training process by introducing corrupted data into the training set, thereby affecting the model's prediction results. Evasion attacks, conversely, involve crafting adversarial examples as inputs during the testing or deployment phase of the model. Investigation of these types of attacks is essential for ensuring the security and reliability of machine learning systems in practical applications.

2.3. Evasion Adversarial Attacks

Evasion attacks aim to compromise the robustness of the model by generating carefully crafted disturbances in the test input that are intentionally designed to cause the model to make a false prediction. The attackers cannot manipulate the training phase but can design

adversarial examples based on the vulnerabilities of the model during the test phase. These examples can be created using a variety of techniques.

Depending on their knowledge of the targeted model, adversaries can tailor their strategies and objectives. This might involve insights into various aspects of the model, including the type of classifier, the loss function, the feature representation, or even the training dataset [8]. Based on the adversary's understanding of the targeted system, evasion attacks can be categorised into two types. A white box attack occurs when the adversary possesses related knowledge of the model. In contrast, a black box attack happens when the adversary has no information about the model [6].

2.4. Black Box Attacks

Black-box adversarial attacks can craft adversarial inputs and induce misclassification errors in a model. These strategies are designated as 'black-box' to reflect the adversary's limited insight: The adversary can only query the deployed model, with no access to the architecture or any parameters of the model [11, 6, 21, 30]. The adversary's interaction is usually confined to analyzing the model's outputs in response to perturbed inputs and leveraging analysis to manipulate these inputs.

In a black-box attack setting where adversaries can't access the model's gradient, they employ alternative methods to create adversarial examples. Gradient estimation is a pivotal technique within black-box attacks, compensating for the inaccessibility of the model's architecture and parameters. Adversaries often employ finite differences approaches to approximate the gradient of the loss function concerning the input. Another method uses evolutionary algorithms, which work by evolving a set of potential solutions over several iterations [33, 22]. Additionally, transfer attacks are used where adversaries develop adversarial examples on a substitute model. These strategies are essential for attacking models when direct access to their internal workings is not available.

For gradient estimation methods, upon obtaining an estimated gradient, the adversarial input can be refined iteratively to amplify the loss, thereby increasing the likelihood of false predictions by the model. The refinement process may adopt simple gradient ascent methods or more complex optimization algorithms that intricately explore different goals of the adversary.

2.5. Physical Attacks

Traffic signs covered with stickers, patches, and graffiti can be seen everywhere on the street. These completely random interferences could cause great trouble to the perception system of autonomous vehicles.

Physical adversarial attacks can broadly be classified as black-box attacks, primarily because they are executed with little to no knowledge of the underlying model. These attacks can still successfully manipulate the model into generating false predictions, posing a significant challenge to machine learning systems.

Physical adversarial attacks represent a distinct class of challenges to the robustness of machine learning-based perception systems, particularly within the autonomous driving domain. These attacks are executed in the physical world without requiring knowledge of the model's structure or parameters. A typical example involves the random application of patches or stickers to traffic signs, as illustrated in Figure 2.

Fig. 2: Examples of physical attacks

Despite the arbitrary placement of these modifications, they can still potentially compromise the safety of machine learning-based detection systems. Addressing these vulnerabilities is critical, as the consequences of misinterpreting traffic signs due to adversarial interference could be severe, leading to traffic disruptions or accidents. Therefore, evaluating and enhancing the robustness of physical adversarial attacks is of paramount importance in the advancement of reliable autonomous driving technologies.

2.6. Vulnerabilities and Adversarial Attacks on Cameras

Although machine learning-based perception systems are getting more and more powerful, research has identified several vulnerabilities, especially in cameras, and how these have been exploited:

1) Measurement Characteristics and Limitations: AV cameras have specific operational parameters and are designed to accurately measure within these predefined ranges. However, attackers have exploited these

limitations to induce malfunctions. For instance, [46] revealed how intense light sources can overload the camera's sensors, leading to temporary or permanent blindness, making it unable to capture and process any visual information. Another notable example from [26] involved the projection of phantom images. This method deceives the camera by introducing artificial objects into its field of view, causing the system to misidentify these projections as real objects, which could lead to incorrect decisions by the AV.

2) Weather and Light Sensitivity: The effectiveness of cameras in AVs is significantly influenced by environmental factors. Research by [44] illustrated how cameras are affected by adverse weather conditions like rain, fog, or snow, which can obscure the lens, scatter light, and distort images. Similarly, [32] highlighted the camera's vulnerability to varying light conditions. Bright light sources, such as direct sunlight or high-beam headlights from oncoming vehicles can cause glare and blind the camera. This vulnerability is particularly concerning as it can be exploited to create dangerous driving scenarios where the AV fails to detect obstacles or misinterprets traffic signals.

3) Sensor Fusion Vulnerabilities: The complex process of merging data from multiple sensors in AVs, each with unique characteristics and error margins, poses significant challenges. As noted in [43], attackers can exploit these complexities by introducing false data or altering sensor outputs. This manipulation can lead to a cascade of errors in the AV's decision-making process, as the system relies on the integrity of this fused data to understand its environment and make driving decisions.

4) Depth and Distance Detection Issues: Accurately predicting the depth and distance of objects is a critical function of AV cameras, but this capability can be compromised. Attackers exploiting this vulnerability can trick the AV into miscalculating distances, potentially causing it to misjudge the speed and proximity of nearby objects. Such miscalculations can result in inappropriate navigational responses, like unnecessary braking or swerving, leading to unsafe driving conditions or accidents.

5) Inconsistent Detection and Causal Inference Challenges: Discrepancies in object detection among different cameras and sensors in AVs can lead to inconsistent interpretations of the environment. This issue, as described in [26], can cause confusion within the AV's perception system, potentially resulting in false decision-making. Additionally, the camera system's limited ability to infer causality or recognize new, unforeseen objects further complicates its reliability. This shortcoming can be particularly problematic in dynamic driving scenarios where it is common to encounter unexpected objects or situations.

2.7. Regulatory Framework

1) UNECE R155: UN Regulation No. 155 (UNECE R155) sets out type approval provisions for cybersecurity and cybersecurity management systems in various vehicle categories, focusing on a comprehensive framework for assessing and managing cybersecurity risks across a vehicle's lifecycle, from development to decommissioning. While it mandates a robust cybersecurity management system, covering aspects like secure communication and software updates, R155's coverage of adversarial attacks on sensors, as highlighted in Annex 5 Part A 4.3.5 and Part B M20, is somewhat limited. This gap in the regulation underscores the need for more advanced provisions to protect against sophisticated adversarial attacks targeting AV perception systems.

2) UNECE R157: UN Regulation No. 157 (UNECE R157) outlines type approval requirements for Automated Lane Keeping Systems (ALKS) used in passenger cars, emphasizing safety and fail-safe response, human-machine interface, object and event detection and response (OEDR), data storage, and cybersecurity. It establishes rigorous requirements for ALKS-equipped vehicles, including minimum performance standards for sensors and specifications for detection ranges. However, despite these comprehensive requirements, R157 does not explicitly address the sensor robustness against advanced adversarial attacks.

3) ISO/SAE 21434: ISO/SAE 21434 addresses cybersecurity in the engineering of electrical and electronic systems within road vehicles involving cybersecurity management, continual activities, and threat analysis and risk assessment (TARA) methods. It focuses on establishing policies and procedures to manage cybersecurity risks, identifying potential threats, and implementing security controls. However, although ISO/SAE 21434 is comprehensive in cybersecurity management, its approach to the specific adversarial challenges faced by AV perception systems is not sufficiently detailed. The standard's general cybersecurity measures, including TARA, may not fully capture the intricacies and unpredictability of adversarial attacks.

4) Regulation (EU) 2022/1426: Regulation (EU) 2022/1426, a pioneering regulation for Level 4 automation, focuses on the type-approval requirements for fully automated vehicles, covering performance requirements, compliance assessment, data security, and cybersecurity management. It introduces vital provisions for the approval and safety of automated driving

systems, but the regulation does not thoroughly address the specific challenges of adversarial robustness in AV perception systems.

3. Methodology

3.1. Summary of Notations

- $\mathscr{L}(\cdot)$: Loss function for adversarial example generation
- $f(\cdot)$: AV perception model's prediction function
- x: True input
- x_{adv}: Adversarial input
- y: True label
- y_{adv}: Adversarial label
- ε: Scaling factor for the gradient in image update
- δ: Small constant used for perturbation in gradient estimation
- $\nabla\mathscr{L}(\cdot)$: Gradient calculation for loss function
- $\hat{\nabla}\mathscr{L}(\cdot)$: Gradient estimation for loss function

3.2. Experimental Setup

The experimental setup for this research is designed to systematically evaluate the vulnerabilities of autonomous vehicle (AV) perception systems to adversarial threats, particularly focusing on black-box gradient estimation attacks and physical attacks.

1) Model Selection: For experimentation, a traffic sign detection model used in autonomous vehicles (AVs) will be selected. This model will facilitate a comprehensive evaluation of its response to various adversarial threats.

2) Black-box Attack Examples Generation: The study will generate a series of black-box attack examples. These examples are crafted to test the robustness of AV perception systems against attacks where the attacker has limited knowledge of the model's internal workings.

3) Physical Attack Examples Generation: Physical attack scenarios will be created to simulate real-world conditions where physical alterations or manipulations can deceive AV systems. These scenarios are designed to assess the impact of physical threats on the detection and classification capabilities of AV perception models.

4) Data Analysis: Data collected from these experiments will be analyzed to evaluate the effectiveness of the attacks and the resilience of different models. This analysis will focus on identifying common vulnerabilities and contributing to recommendations for enhancing AV system security.

3.3. Test Model Selection

The Faster RCNN Inception V2 model [2] is selected for its high precision and efficiency in the specialized domain of traffic sign detection and classification. This model is chosen due to its features, which include:

- Pre-training on COCO Dataset: It has been pre-trained on the diverse Microsoft COCO dataset, providing a rich foundation of visual knowledge that can be effectively transferred to the traffic sign detection task.
- Fine-tuning on the German Traffic Sign Detection Benchmark (GTSDB) dataset: The GTSDB dataset is particularly pertinent for adapting the model to the European Union and the Netherlands' traffic framework.
- High mAP Score: With a mean average precision (mAP) of 90.62 %, Faster R-CNN Inception V2 exhibits exceptional accuracy in detecting various objects.

Fig. 3: Detection example of traffic sign detection model

The traffic sign detection process involves classifying signs into three labels: prohibitory, mandatory, and danger. Each detected sign is highlighted with a bounding box that specifies its location, accompanied by a confidence score ranging from 0 to 1, as illustrated in Figure 3.

Adversarial examples are crafted using the Faster R-CNN Inception V2 model. In the implementation setting, we assume that the adversary only knows the label and confidence score of the model output result, and does not have any knowledge of the model's training dataset, model architecture, related parameters, etc. The images we use for simulation are from the Computer

Vision Laboratory (CVL) at Linköping University [41] and Mapillary Traffic Sign Dataset [25, 15].

This setting more realistically describes an attack scenario that is likely to occur in real life. The adversary only needs to query the model and get legitimate outputs, eliminating the need for additional efforts or access to internal model details.

3.4. Procedure for Black-Box Attack

The following outlines the algorithmic steps for black-box gradient estimation when direct access to a model's gradients is not feasible:

- **Loss Function Computation**: The loss function $\mathscr{L}$ quantifies the difference between the model's prediction $f(x_{adv})$ and the true label y, serving as a measurement of prediction error: $\mathscr{L}(f(x_{adv}), y)$.
- **Gradient Estimation**: Estimate the gradient by probing the model with a slightly perturbed version of x_{adv} and observing the changes in loss. For each dimension i of x_{adv}, we can compute an approximation of the gradient as follows:

$$\hat{\nabla}\mathscr{L} \approx \frac{\mathscr{L}\left(f\left(x_{adv}+\delta\right), y\right) - \mathscr{L}\left(f\left(x_{adv}\right), y\right)}{\delta}$$

  where δ is a vector of small constants used to perturb each element of x_{adv}.
- **Adversarial Input Update**: Update the adversarial input using the estimated gradient:

$$x_{adv} = x_{adv} - \varepsilon\hat{\nabla}\mathscr{L}$$

1) Specific Adaptations of the Algorithm: When applying a general algorithm for black-box adversarial attacks to specific models, the objective is to produce adversarial examples where the perturbed image is misclassified by the model yet remains visually indistinguishable to human observers. To achieve this, certain targeted adaptations are implemented:

- **Bounding Box Detection and Focused Perturbation**: Upon the successful detection of a traffic sign within the image, a bounding box is generated around the identified object. During preliminary testing phases when refining the black-box adversarial attack algorithm for traffic sign detection systems, it was observed that the area within and immediately surrounding the bounding box exhibited a much higher sensitivity to adversarial perturbations.

  In trials where the adversarial method was applied across the entire image, gradients computed outside the bounding box region were negligible, with values approaching zero. This phenomenon underscores the non-uniform impact of adversarial perturbations across the image space. In this refined algorithm, the bounding box is expanded by a predetermined buffer factor to encapsulate a box around the sign, ensuring that the perturbations are generated within this enlarged box and account for spatial sensitivities in the model's perception.
- **Downsampling for Efficiency**: The region within the enlarged bounding box is then downsampled, reducing the resolution to decrease computational demands. This step simplifies the gradient estimation process without significantly impacting the effectiveness of the attack. Once the gradient has been estimated and the adversarial pattern has been calculated on the downsampled image, the adversarial pattern will be upsampled back to the original image's resolution. The adversarial updates are then applied to the corresponding region in the original high-resolution image.
- **Iterative Refinement**: Through a series of iterations, the algorithm refines the adversarial pattern. In each iteration, the estimated gradient informs how the image should be adjusted to maximize the loss, thereby steering the model towards a false negative. The alterations are subtle and intended to remain under the threshold of human detection.

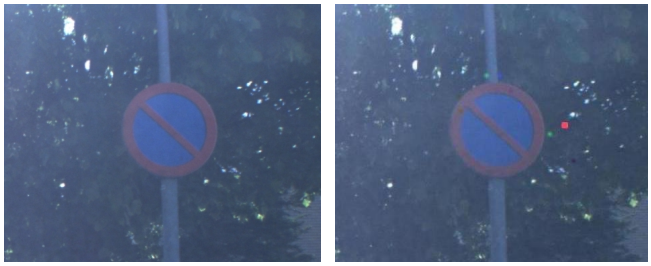


Fig. 4: Before (left) and after (right) adding the adversarial pattern

Figure 4 shows one example of adding adversarial pattern on the image with the traffic sign. The adversarial pattern is generated using the black-box gradient estimation algorithm in Chapter 4. The visual difference is almost impossible to detect by human eyes.

2) Algorithmic Steps for Black-Box Adversarial Image Generation: The algorithm for generating adversarial images in a black-box scenario targets autonomous vehicle (AV) perception systems. It iteratively refines test images to create adversarial examples, comprising the following key steps, the detailed algorithm for each iteration is shown as follows:

- **Run Detection:** For the current adversarial image, obtain detection scores, classes, and bounding

boxes.

- **Process Each Detected Bounding Box:** For each detected bounding box, enlarge the bounding box by a predefined factor. Extract and downsample the region within the enlarged box.
- **Approximate Gradient for Each Region:**

  Compute the original loss L_{orig} :

  $$L_{\text{orig}} = \log\left(\frac{\text{score}}{1-\text{score}}\right)$$

  Perturb the downsampled region by adding a small constant value h on randomly selected pixels. Compute the loss for the perturbed region:

  $$L_{\text{perturbed}} = \log\left(\frac{\text{score}_{\text{perturbed}}}{1-\text{score}_{\text{perturbed}}}\right)$$

  Compute the estimated gradient:

  $$G = \frac{L_{\text{perturbed}} - L_{\text{orig}}}{h}$$

- **Upsample Gradient to Original Resolution:** Upsample the approximated gradient back to the size of the original region.
- **Update Adversarial Image:** Apply the upsampled gradient to update the adversarial image: *adv_image = adv_image - EPS* $\times G$

This process systematically alters the original image, embedding subtle perturbations to deceive the AV system while retaining visual similarity.

3.5. Algorithmic Steps for Physical Attack Image Generation

The use of adversarial patches in simulations is a valuable method for testing the adversarial robustness of traffic sign detection systems in autonomous vehicles, it is more cost-effective than setting up real-world scenarios. The algorithm generates physical adversarial examples from original inputs on traffic sign detection systems. Here is the main execution for one image:

1) Detection of Original Images: Faster R-CNN Inception V2 model processes a series of test images to detect traffic signs. It gives these initial detection results, which include identified traffic signs and their bounding boxes.

2) Bounding Box Adjustment: For each detected traffic sign, a new, smaller bounding box is computed within the original detection box. The dimensions of the new bounding box are 70% of the original, maintaining the center point.

3) Patch Position Calculation: The algorithm calculates two distinct positions for placing the patches on the traffic sign, ensuring they do not overlap.

4) Patch Application: Each patch is resized to 1/16 of the size of the bounding box. The patches are then applied to the calculated positions on the traffic sign within the image.

4. Simulation Results

4.1. Classification Failure

False positive, false negative and misclassification are used to describe different types of classification failures, each with its implications for safety and system robustness. The robustness metrics considered here are defined below.

- False Positive: This occurs when the model incorrectly identifies a traffic sign where none exists. High rates of false positives can lead to unnecessary braking or evasive manoeuvres, compromising safety and efficiency, as illustrated in Figure 5.
- False Negative: This refers to instances where the model fails to detect an actual traffic sign. A false negative is particularly dangerous as it may result in a vehicle not stopping or yielding when it is required, potentially leading to accidents, as illustrated in Figure 6.
- Misclassification: This happens when the model detects a sign but assigns it an incorrect label (e.g., mistake a 'Stop' sign as a 'Speed' sign). Misclassifications can lead to inappropriate responses from the vehicle, such as failing to stop, thus posing a serious safety risk, as illustrated in Figure 7.

Fig. 5: Illustration of false positive: before (left) and after (right) adding the adversarial pattern

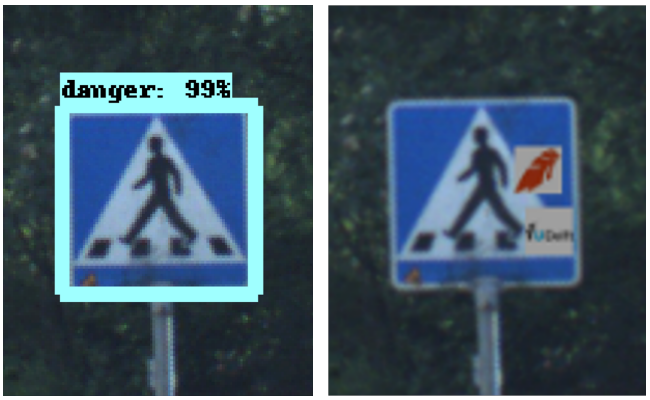


Fig. 6: Illustration of false negative: before (left) and after (right) adding patches

Fig. 7: Illustration of misclassification: before (left) and after (right) adding the adversarial pattern

The performance of the black-box attack with the variables and parameters is listed in Table 4, we only take detections with confidence scores above 0.5 as valid detections for analysis. These valid detections were subsequently compared to the benchmark, which is the detection results from the original set of 200 image inputs. The Faster R-CNN Inception V2 model then is tested on a series of adversarial images derived from these original images. For each image, ten iterations were conducted to introduce perturbations, resulting in the creation of a total of 2000 adversarial images across iterations ranging from the 1st to the 10th. This process culminated in a total of 2626 detection results. The findings are detailed in Table 1:

Metric	Count	Percentage
Correct detection	1573	59.90%
False negative	989	37.66%
False positive	25	0.95%
Misclassification	39	1.49%

Tab 1. Detection results of adversarial examples

Among the 2626 detections, there are only 1573 detections still remain correct, By adding the false negative rate, false positive rate, and misclassification rate together, the model exhibited a combined classification failure rate of 40.10% over the sum of 10 iterations. This substantial rate highlights critical robustness weakness against adversarial attacks of the model.

4.2. Adversarial Examples vs Randomly Generated Examples

In order to effectively illustrate the impact of gradient-based adversarial attacks on model predictions, we created a set of random attack examples for comparison. These were generated similarly to the adversarial examples, but instead of using the gradient to compute perturbations, an average of the perturbations from the adversarial examples under the same conditions is applied as a constant addition. The comparison between the results of the gradient-based adversarial attacks and these random perturbations is shown in Figure 8. It is evident that gradient-based attacks significantly increase the attack success rate, this marked difference in effectiveness highlights the risk posed by black-box adversarial examples, which are crafted using limited information yet can significantly compromise model accuracy.

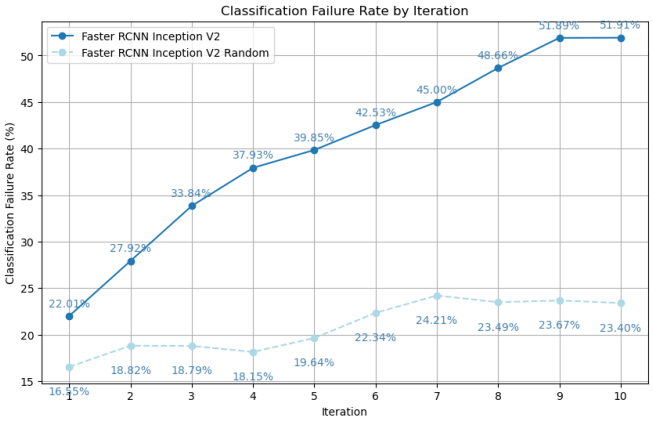


Fig. 8: Classification failure by iteration from 1 to 10

Figure 8 also presents the trend of classification failure rates over a series of ten iterations. It starts at a classification failure rate of 27.33% at the first iteration and reaches 60.56% at the 10th iteration. As the iterations progress, there is a notable upward trend in the failure rate, indicating a gradual decline in classification performance. This suggests that the model's ability to accurately classify traffic signs degrades as the iterations increase.

ε is the scaling factor for the gradient in image update. When we increase ε, the perturbed pixels and

the adversarial pattern will be more visible to human eyes. We keep other parameters maintained as in Table 4 but change ε from 1000 to 40000, the classification failures are shown in Figure 9, the visual difference between images with different ε is given in Figure 10.

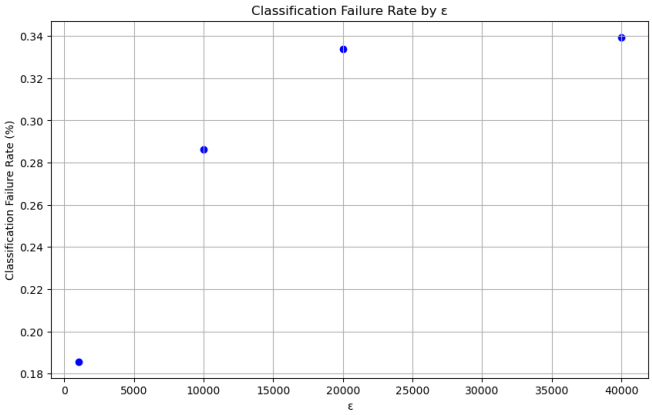


Fig. 9: Classification failure by ε

Fig. 10: Adversarial examples with ε = 1000 (left) and ε = 40000 (right)

As ε increases, indicating stronger perturbations by scaling the approximated gradient, there is an observable rise in the classification failure rate. This suggests that the model becomes increasingly susceptible to perturbation as the intensity of the adversarial perturbation grows, confirming that larger perturbations are more likely to mislead the model into incorrect classifications.

4.3. Normalized Adversarial Perturbation

As shown in the previous results, it is evident that the magnitude of the adversarial pattern influences the classification failure significantly. To investigate more on this, in this section, we apply the projected gradient descend (PGD) [12] to generate normalized adversarial examples. The adversarial example x_{adv} will be projected onto an $\mathscr{E}$ ball using normalization. This step is crucial because it regulates the magnitude of perturbations added to the image. By using different values of $\mathscr{E}$, we aim to compare the average perturbation against the rate of classification failure. This comparison is vital for evaluating the effectiveness of adversarial attacks while maintaining perturbations within realistic and less detectable bounds. Here, we use the infinity norm (∞-norm) and the 2-norm to regulate perturbation δ:

- For ∞-norm:

  $\delta = \min(\max(\delta, -\mathscr{E}), \mathscr{E})$

- For 2-norm:

  $\text{norm} = \|\delta\|_2$

  If $\text{norm} > \mathscr{E}$:

  $\delta = \delta \times \dfrac{\mathscr{E}}{\text{norm}}$

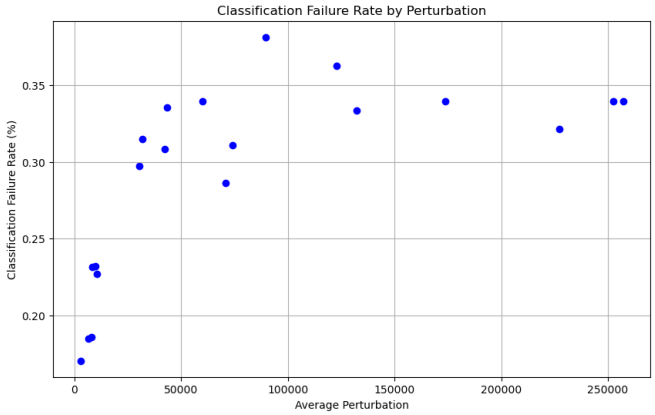


Fig. 11: Classification failure by average perturbation

The trend depicted in Figure 11 suggests a relationship between the perturbation of images and the model's classification failure rate. As the average perturbation increases, indicating more evident adversarial modifications, the model's ability to classify correctly decreases as well. This reflects the model's vulnerability to more pronounced adversarial patterns, which are designed to exploit its feature recognition processes and lead to higher classification failure.

4.4. Transferability of Adversarial Examples

The generalization capability and transferability of adversarial examples are crucial in assessing the vulnerability of machine learning models against adversarial attacks, because they demonstrate the potential for widespread impact of such attacks. Particularly in the context of black-box attacks, attackers do not have access to the target model's internal architecture, so

they must rely on adversarial examples generated using a different model to which they have access. The following three models are chosen to test the generalization capability. Faster R-CNN Resnet 101 model and R-FCN Resnet 101 model present an optimal balance between accuracy and processing speed. Additionally, SSD Mobilenet stands out as the quickest and most memory-efficient model, making it ideally suited for implementation in mobile and embedded devices [2, 3].

1) Faster R-CNN Resnet 101 Model:

- Type: Faster R-CNN.
- Architecture: ResNet-101

2) R-FCN Resnet 101 Model:

- Type: R-FCN (Region-based Fully Convolutional Network).
- Architecture: ResNet-101.

3) SSD Mobilenet Model:

- Type: SSD (Single Shot MultiBox Detector).
- Architecture: MobileNet.

Figures 12, 13, and 14 illustrate the trend in the generalization of adversarial examples on these three models.

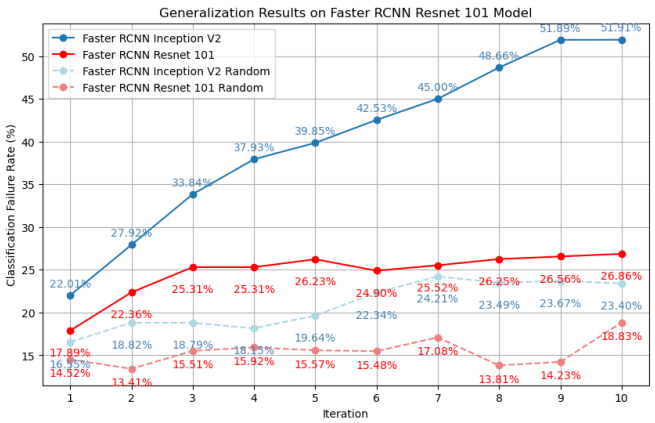


Fig. 12: Generalization results on Faster R-CNN Resnet 101 Model

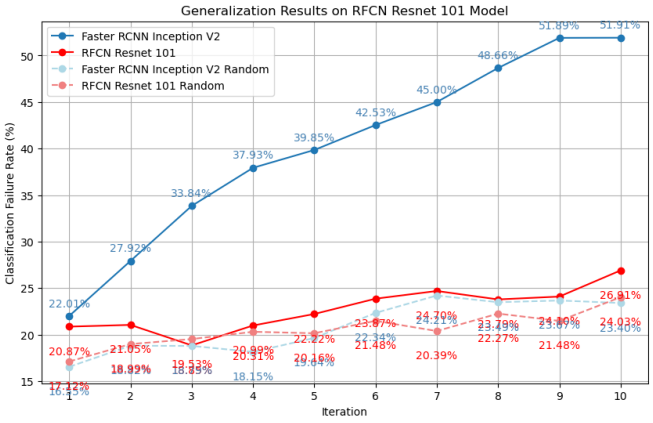


Fig. 13: Generalization results on R-FCN Resnet 101 Model

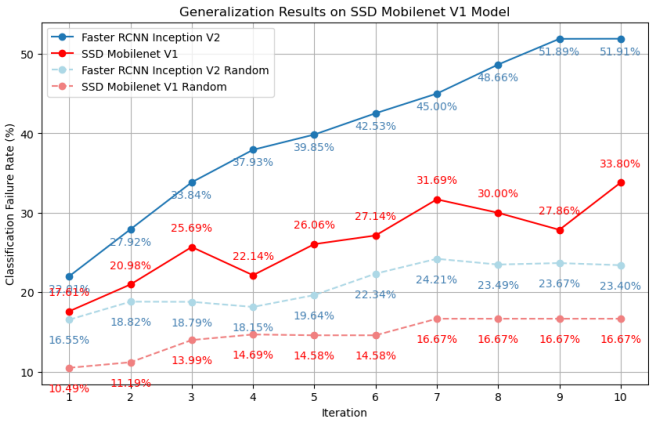


Fig. 14: Generalization results on SSD Mobilenet Model

The data from the red and dashed red lines in Figures 12, 13, and 14 clearly show that adversarial images generated from the original Faster R-CNN Inception V2 model leads to a higher classification failure rate in new target models compared to images perturbed with random pixels. This finding validates the transferability of adversarial examples across different models in this case.

However, when these results are compared with those from the original Faster R-CNN Inception V2 model, indicated in blue and dashed blue lines in the figures, the disparity between the impact of adversarial images and randomly perturbed images is significantly less pronounced, indicating that the transferability capability is limited, and while adversarial examples have some generalizing effect, they are less effective when applied to models not used in their generation.

4.5. Physical Attack

Fig. 15: Before (left) and after (right) adding patches

As shown in Figure 15, we randomly placed two patches on each traffic sign and tested on the Faster RCNN Inception V2 model, the performance of the traffic sign detection model on a set of 600 images revealed the following results in Table 2:

Metric	Count	Percentage
Correct detection	585	79.92%
False negative	91	12.43%
False positive	28	3.83%
Misclassification	28	3.83%

Tab 2. Detection results of physical attacks

Among 732 valid detection results, there are 585 detections still remain correct. However, the model exhibited a combined classification failure rate of 20.08%. While it indicates a higher degree of robustness compared with the black-box gradient-based attacks, it also highlights the challenges posed by physical attacks in real-world scenarios, especially in the face of everyday environmental modifications like stickers or patches on traffic signs.

5. Conclusion and Future Suggestions

5.1. Conclusions of Simulation Results

Our investigation into black-box adversarial attacks and physical attacks has uncovered significant robustness flaws in traffic sign detection models, particularly in open-source models. The findings highlight a pronounced vulnerability of these machine-learning models to adversarial manipulations, which can lead to substantial classification failures. This vulnerability is not merely a technical challenge; it represents a serious threat to the safety and security of autonomous driving systems. As these perception systems are fundamental to the operation of autonomous vehicles, their susceptibility to adversarial attacks could have severe real-world consequences.

5.2. Conclusions of Current Regulations

The existing regulatory framework, including UNECE R155, R157, ISO/SAE 21434, and EU Regulation 2022/1426, provides a foundational structure for the safety, cybersecurity, and management of autonomous vehicles. However, these regulations exhibit significant gaps in addressing the adversarial robustness of machine learning-based AV perception systems:

1) UNECE R155, R157 and (EU)2022/1426: UNECE R155, R157, and EU Regulation 2022/1426 each play a distinct role in the regulatory landscape for autonomous vehicles. R155 focuses on a comprehensive cybersecurity management system across a vehicle's lifecycle. R157, on the other hand, is specifically dedicated to Automated Lane Keeping Systems (ALKS), setting stringent requirements for sensor performance and vehicle safety. Meanwhile, EU Regulation 2022/1426, a leading regulation for Level 4 automation, establishes essential standards for automated driving systems.

However, a common limitation across these regulations is their insufficient coverage of adversarial robustness, particularly in AV perception systems. Neither R155 nor R157 provides explicit guidelines on how to evaluate or mitigate advanced adversarial attacks. Similarly, EU Regulation 2022/1426, despite its forward-looking approach to automation, lacks detailed provisions for addressing the unique challenges posed by adversarial threats to machine learning-based autonomous driving perception systems.

2) ISO/SAE 21434: This standard addresses the cybersecurity of electrical and electronic systems in road vehicles. However, its approach only covers traditional vehicles and fails to detail the adversarial challenges unique to machine learning-based autonomous driving perception systems.

5.3. Limitations of the Study

In this study, our primary focus was on evaluating the impact of black-box gradient estimation attacks and physical attacks on traffic sign detection models. However, it's important to note that this research does not encompass the entire spectrum of black-box attack methodologies, such as those involving evolutionary algorithms. Additionally, our investigation was confined to simulations within the model environment, and thus, the real-world implications and effectiveness of these attacks were not directly tested. Future work could expand upon this study by exploring other black-box attack methods and assessing the real-world impact of these attacks to provide a more comprehensive

understanding.

5.4. Future Suggestions for Strengthening Machine Learning Robustness

It is imperative to focus on developing strategies for enhancing the robustness of machine-learning algorithms against adversarial attacks in autonomous vehicle systems. Potential strategies include:

1) Exploration of Advanced Adversarial Training Techniques: Future efforts can explore adversarial training methods for AVs. The aim would be to enhance machine learning models' ability to identify and resist a wide array of adversarial attacks by incorporating adversarial examples into the training process.

2) Development of Hardened Machine-learning Algorithms: It is essential to prioritize the development of hardened machine-learning algorithms. These enhanced algorithms must be designed with the explicit intent to mitigate the risks posed by adversarial attacks.

3) Development of Sophisticated Defensive Mechanisms: Defensive measures can be systematically employed to reinforce the robustness of models against such threats. These could involve novel approaches such as regularization in neural networks and gradient masking.

5.5. Future Suggestions for Regulations

In addition to algorithmic improvements, there must be an urgent effort to update and expand safety standards to include comprehensive adversarial robustness assessments, the following recommendations are proposed:

1) UNECE R155 and R157: Strengthen the management and validation protocols for training and test data used in machine learning models in AVs, and promote transparency in algorithmic processes within the cybersecurity management system outlined in Regulation R155. Additionally, these regulations should be revised to include specific provisions and mitigation strategies for adversarial robustness in AV perception systems.

2) ISO/SAE 21434: The standard should broaden its TARA process to specifically address the probabilistic models of machine learning used in AV systems. This expansion would enable a more thorough assessment of the cybersecurity risks of adversarial attacks.

3) EU 2022/1426: This regulation should integrate specific guidelines and requirements for the adversarial robustness of AV perception systems, ensuring a comprehensive approach to safety in Level 4 automation vehicles.

4) OEM Responsibilities: Original Equipment Manufacturers (OEMs) should be required to conduct in-depth evaluations of adversarial robustness as part of their vehicle safety assessments, before applying for type approvals.

By adopting these measures, the regulatory framework can evolve to address the current and future challenges of autonomous vehicle technology, thereby ensuring greater safety and fostering public trust in this rapidly advancing field.

REFERENCES

[1] Mazen Abdelfattah et al. “Adversarial attacks on camera-lidar models for 3d car detection”. In: *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE. 2021, pp. 2189–2194.

[2] Álvaro Arcos-García, Juan A Álvarez-García, and Luis M Soria-Morillo. “Evaluation of deep neural networks for traffic sign detection systems”. In: *Neurocomputing* 316 (2018), pp. 332–344.

[3] Álvaro Arcos-García, Juan Antonio Álvarez-García, and Luis M. Soria-Morillo. *Traffic Sign Detection*. https://github.com/aarcosg/traffic-sign-detection. Accessed: July 12, 2023. 2023.

[4] Rodrigo Ayala and Tauheed Khan Mohd. “Sensors in Autonomous Vehicles: A Survey”. In: *Journal of Autonomous Vehicles and Systems* 1.3 (2021).

[5] Marco Barreno et al. “Can machine learning be secure?” In: *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*. 2006, pp. 16–25.

[6] Siddhant Bhambri et al. “A survey of black-box adversarial attacks on computer vision models”. In: *arXiv preprint arXiv:1912.01667* (2019).

[7] Battista Biggio, Blaine Nelson, and Pavel Laskov. “Poisoning attacks against support vector machines”. In: *arXiv preprint arXiv:1206.6389* (2012).

[8] Battista Biggio et al. “Evasion attacks against machine learning at test time”. In: *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2013, Prague, Czech Republic, September 23-27, 2013, Proceedings, Part III 13*. Springer. 2013, pp. 387–402.

[9] Adith Boloor et al. “Attacking vision-based perception in end-to-end autonomous driving models”. In: *Journal of Systems Architecture* 110 (2020), p. 101766.

[10] Yulong Cao et al. “Adversarial sensor attack on lidar-based perception in autonomous driving”. In: *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2019, pp. 2267–2281.

[11] Pin-Yu Chen et al. “Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models”. In: *Proceedings of the 10th ACM workshop on artificial intelligence and security*. 2017, pp. 15–26.

[12] Francesco Croce and Matthias Hein. "Sparse and imperceivable adversarial attacks". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 4724–4732.

[13] *Death of Elaine Herzberg — Wikipedia, The Free Encyclopedia*. [Online; accessed April 12, 2023]. 2023. URL: https://en.wikipedia.org/wiki/Death_of_Elaine_Herzberg.

[14] Ambra Demontis et al. "Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks". In: *28th USENIX security symposium (USENIX security 19)*. 2019, pp. 321–338.

[15] Christian Ertler et al. "The mapillary traffic sign dataset for detection and classification on a global scale". In: *European Conference on Computer Vision*. Springer. 2020, pp. 68–84.

[16] Ivan Evtimov et al. "Robust physical-world attacks on machine learning models". In: *arXiv preprint arXiv:1707.08945* 2.3 (2017), p. 4.

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

[18] Sorin Grigorescu et al. "A survey of deep learning techniques for autonomous driving". In: *Journal of Field Robotics* 37.3 (2020), pp. 362–386.

[19] R Spencer Hallyburton et al. "Security Analysis of {Camera-LiDAR} Fusion Against {Black-Box} Attacks on Autonomous Vehicles". In: *31st USENIX Security Symposium (USENIX Security 22)*. 2022, pp. 1903–1920.

[20] Henry Alexander Ignatious, Manzoor Khan, et al. "An overview of sensors in Autonomous Vehicles". In: *Procedia Computer Science* 198 (2022), pp. 736–741.

[21] Andrew Ilyas et al. "Black-box adversarial attacks with limited queries and information". In: *International conference on machine learning*. PMLR. 2018, pp. 2137–2146.

[22] Andrew Ilyas et al. "Query-efficient black-box adversarial examples (superceded)". In: *arXiv preprint arXiv:1712.07113* (2017).

[23] Puneet Kohli and Anjali Chadha. "Enabling pedestrian safety using computer vision techniques: A case study of the 2018 uber inc. self-driving car crash". In: *Advances in Information and Communication: Proceedings of the 2019 Future of Information and Communication Conference (FICC), Volume 1*. Springer. 2020, pp. 261–279.

[24] Yanpei Liu et al. "Delving into transferable adversarial examples and black-box attacks". In: *arXiv preprint arXiv:1611.02770* (2016).

[25] *Mapillary Traffic Sign Dataset*. https://www.mapillary.com/dataset/trafficsign. Accessed: June 10, 2023.

[26] Ben Nassi et al. "Phantom of the adas: Phantom attacks on driver-assistance systems". In: *Cryptology ePrint Archive* (2020).

[27] National Highway Traffic Safety Administration. *Automated Driving Systems 2.0: A Vision for Safety*. https://www.nhtsa.gov/sites/nhtsa.gov/files/documents/13069a-ads2.0_090617_v9a_tag.pdf. Accessed: April 18, 2023.

[28] National Highway Traffic Safety Administration. *Standing General Order: Crash Reporting*. https://www.nhtsa.gov/laws-regulations/standing-general-order-crash-reporting. Accessed: April 15, 2023.

[29] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples". In: *arXiv preprint arXiv:1605.07277* (2016).

[30] Nicolas Papernot et al. "Practical black-box attacks against machine learning". In: *Proceedings of the 2017 ACM on Asia conference on computer and communications security*. 2017, pp. 506–519.

[31] Nicolas Papernot et al. "Towards the science of security and privacy in machine learning". In: *arXiv preprint arXiv:1611.03814* (2016).

[32] Jonathan Petit et al. "Remote attacks on automated vehicles sensors: Experiments on camera and lidar". In: *Black Hat Europe* 11.2015 (2015), p. 995.

[33] Hao Qiu, Leonardo Lucio Custode, and Giovanni Iacca. "Black-box adversarial attacks using evolution strategies". In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. 2021, pp. 1827–1833.

[34] Wan Rahiman and Zafariq Zainal. "An overview of development GPS navigation for autonomous car". In: *2013 IEEE 8th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE. 2013, pp. 1112–1118.

[35] Spyridon Samonas and David Coss. "The CIA strikes back: Redefining confidentiality, integrity and availability in security." In: *Journal of Information System Security* 10.3 (2014).

[36] Brandon Schoettle and Michael Sivak. "A preliminary analysis of real-world crashes involving

self-driving vehicles". In: *University of Michigan Transportation Research Institute* (2015).

[37] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. "One pixel attack for fooling deep neural networks". In: *IEEE Transactions on Evolutionary Computation* 23.5 (2019), pp. 828–841.

[38] Jiachen Sun et al. "Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures". In: *29th USENIX Security Symposium (USENIX Security 20)*. 2020, pp. 877–894.

[39] Christian Szegedy et al. "Intriguing properties of neural networks". In: *arXiv preprint arXiv:1312.6199* (2013).

[40] *The three-pillar approach to cyber security: Data and information protection*. https://www.dnv.com/article/the-three-pillar-approach-to-cyber-security-data-and-information-protection-165683. Accessed: April 15, 2023.

[41] *Traffic Signs Dataset*. https://www.cvl.isy.liu.se/en/research/datasets/traffic-signs-dataset/. Accessed: June 10, 2023.

[42] James Tu et al. "Exploring adversarial robustness of multi-sensor perception systems in self driving". In: *arXiv preprint arXiv:2101.06784* (2021).

[43] UCI. *Security Vulnerability in Self-Driving Cars Unveils Achilles Heel of Sensor Fusion*. 2019. URL: https://www.cs.uci.edu/security-vulnerability-in-self-driving-cars-unveils-achilles-heel-of-sensor-fusion/.

[44] Jorge Vargas et al. "An overview of autonomous vehicles sensors and their vulnerability to weather conditions". In: *Sensors* 21.16 (2021), p. 5397.

[45] Mingfu Xue et al. "Machine learning security: Threats, countermeasures, and evaluations". In: *IEEE Access* 8 (2020), pp. 74720–74742.

[46] Chen Yan, Wenyuan Xu, and Jianhao Liu. "Can you trust autonomous vehicles: Contactless attacks against sensors of self-driving vehicle". In: *Def Con* 24.8 (2016), p. 109.

[47] De Jong Yeong et al. "Sensor and sensor fusion technology in autonomous vehicles: A review". In: *Sensors* 21.6 (2021), p. 2140.

[48] Ekim Yurtsever et al. "A survey of autonomous driving: Common practices and emerging technologies". In: *IEEE access* 8 (2020), pp. 58443–58469.

6. APPENDIX A

6.1. Sensor Attacks and Effects on Cameras

Resource	Attack strategy	Vulnerabilities been exploit	Effect	Simulation/Test
[26]	Projected a phantom via a drone equipped with a portable projector or present a phantom on a digital billboard	Limitations of measurement characteristics; Lack of explainability	Considered a phantom as a legitimate traffic sign	Mobileye 630 PRO and Tesla Model X
[32]	Blinded the camera fully or partially by emitting light with different environmental light, light source and attack distance into the camera	Limitations of measurement characteristics; Sensitivity to light condition	Failing to detect objects	MobilEye C2-270
[46]	Blinded the camera with strong lights	Limitations of measurement characteristics	Failing to detect objects and permanent camera damage	CMOS/CCD chip

Tab 3. Sensor Attacks and Effects on Cameras

6.2. Variables in the Adversarial Attack Algorithm

Variable	Definition	Value
DOWNSAMPLE_RATIO	Ratio for downsampling the image	0.2
δ	Small constant used for perturbation in gradient estimation	10
NUM_ITERATIONS	Number of iterations for refining examples	10
NUM_SAMPLES	Number of pixel indices sampled for gradient approximation	20
ε	Scaling factor for gradient in image update	40000
ENLARGE_FACTOR	Factor to enlarge the bounding box around the object	1

Tab 4. Variables used in the adversarial attack algorithm and their associated values

7. APPENDIX B: SOURCE CODE

Source Code can be found on Github: https://github.com/Yuxing-Gao/adversarial-attack-traffic-sign-detection.