International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Adversarial attacks and defense mechanisms for image classification deep learning models in autonomous driving systems

Divya Bharat Mistry * and Kaustubh Anilkumar Mandhane

Department of Electronics and Computer Science, Thakur College of Engineering and Technology, Mumbai, India.

International Journal of Science and Research Archive, 2024, 13(02), 1898–1917

Publication history: Received on 19 October 2024; revised on 27 November 2024; accepted on 30 November 2024

Article DOI: https://doi.org/10.30574/ijsra.2024.13.2.2328

Abstract

Advancements in artificial intelligence (AI) and Internet of Things (IoT) technologies have catalyzed the evolution of autonomous driving systems (ADSs), with image classification deep learning (DL) models serving as the cornerstone of their decision-making frameworks. Deep neural networks are employed in highly sophisticated and unforeseeable environments such as advanced industrial automation, autonomous vehicles, and financial forecasting. While these models excel in navigating complex driving scenarios, their susceptibility to adversarial attacks poses significant threats to operational safety and functional integrity. This study delves into the taxonomy of adversarial exploits, dissects cutting-edge defense mechanisms, and examines the delicate equilibrium between adversarial robustness and model generalizability. It accentuates the imperative for adaptive, resource-efficient, and scalable countermeasures capable of dynamic, real-time deployment while advocating for hybrid defense architectures and explainable AI (XAI) to foster system transparency and stakeholder trust. By addressing these systemic vulnerabilities through transferable defense strategies, universal countermeasures, and multidisciplinary collaboration, the study sets the stage for developing fortified ADSs capable of resilient operation in dynamically adversarial ecosystems.

Keywords: Adversarial Attacks; Deep neural networks; Image Classification; Defense Mechanisms; Autonomous Driving System

1. Introduction

Deep learning is an improvement over artificial intelligence and neural networks since multiple layers are stacked to achieve greater abstraction and improved information analysis compared to traditional machine learning algorithms. It is used to solve classification and regression problems that are impossible to solve quickly and is used in almost every industry: healthcare, finance, agriculture, gaming, etc. It transforms data into actionable insights, showcasing the technology's wide-ranging potential across various industries. There is constant upgrading being made in various libraries, frameworks, and hardware resources by making them available to the community promptly. Ongoing advancements and performance breakthroughs enable us to achieve these improvements through deep neural networks (DNN). Deep neural networks are used for pattern recognition, data classification, function estimation, predictive analytics, feature extraction, signal processing, and automation. With its capability to process intricate data and identify patterns, it is now utilized to address critical safety and security challenges, including autonomous vehicles, multi-agent aerial systems equipped with facial recognition, robotics, social engineering detection, network anomaly identification, and deep packet analysis. Deep neural networks have become integral to daily life, driving innovations such as virtual assistants, chatbots, autonomous driving, and tailored recommendation systems.

Due to the rapid advancements in the field of autonomy, particularly in autonomous driving technologies (ADS), vehicles can now navigate and make decisions without requiring human intervention. The technologies comprise LIDAR (light

* Corresponding author: Divya Mistry

detection and ranging), RADAR (radio detection and ranging), computer vision and cameras, artificial intelligence, and machine learning, deep learning and neural networks, sensor fusion, etc. to determine road conditions. It effectively contributes towards the development of autonomous vehicles to understand their surroundings, make informed decisions, minimize human error, and navigate through sophisticated environments. Companies such as Tesla, Uber, Nvidia, and Rivian are making substantial contributions to the advancement of autonomous driving systems (ADS) and have created a supportive environment for commercializing these systems. Most of these advancements currently operate at level 3 autonomous driving (AD) technology, allowing for conditional automation under specific conditions, such as a car that can drive itself on highways but requires the driver to take over in complex situations like city traffic or poor weather. However, various challenges regarding safety and reliability have risen due to the Swift adaptation of ADSs and have raised significant concerns in advancing this technology and have become obstacles to achieving fully autonomous, level-five self-driving technology. If the image classification deep learning model is susceptible to adversarial attacks, we may face some serious consequences, which may result in poor driving and inaccurate scene evaluation. This could have catastrophic driving implications. Therefore, protecting image classifiers in autonomous driving systems from adversarial attacks is crucial for maintaining their effectiveness and reliability.

Due to its substantial implications in emerging IoT technology, it has paved the way for the creation of a smart driving environment. The increasing need for advanced data processing abilities in autonomous driving systems (ADS) aligns with the rising number of IoT-enabled intelligent devices, such as smart traffic lights (they can adjust their timing based on real-time traffic framework), smart traffic signs (they provide fluid information to drivers about road conditions and the possibility of incidents), and smart signboards (which can display real-time updates on traffic patterns and emergencies), all of which have significantly enhanced environmental awareness among ADS, allowing them to effectively communicate with their surroundings and make informed driving decisions. Ensuring that there is adequate safety and robustness of interconnected systems within the driving environment is of great importance since an attack on data integrity can jeopardize the safety and dependability of ADSs. As a result, reducing vulnerability to hostile attacks is a key goal. As it can help strengthen defenses against potential threats, improving security measures can build trust among users, clients, and stakeholders and help identify and manage risks more effectively. Eventually, an active strategy is necessary to reduce vulnerability for sustainable growth and success in a challenging business environment.

Image classification deep learning models are crucial since they are used for behavior analysis of fixed and mobile entities in driving settings, effectively contributing towards informed real-time driving decisions. However, one major issue arises that makes it difficult to ensure the reliability of these models: their vulnerability to adversarial attacks. These attacks can intentionally manipulate the models by altering their parameters (weights and biases) or by introducing harmful inputs, which can lead to incorrect interpretations of the driving environment and potentially dangerous decision-making by overlooking critical information and miscalculating distances, ultimately jeopardizing the safety of autonomous driving systems. Modifying the model's parameters is usually known as a "model poisoning attack." This attack highlights the importance of secure measures that need to be taken across various levels since modifications may be inconspicuous to humans but can greatly affect the safety and reliability of deep learning models. Several challenges are presented in mitigating the vulnerabilities of image classification deep learning models to adversarial attacks across various industries, and these challenges include the difficulty of identifying subtle manipulations that can deceive the models, the need for robust training methods that can withstand such attacks, and the requirement for ongoing monitoring and updating of the models to ensure they remain resilient. One of the major is the adaptability of adversaries, who constantly adjust their methods to bypass existing defenses. This adaptability, when combined with numerous attack avenues, including various methods such as adversarial example generation, evasion attacks, and model poisoning, complicates the prediction and defense against emerging threats. Adversarial example generation involves subtly altering the input data to trick the model without being noticeable to humans. Evasion attacks target exploiting vulnerabilities during their decision-making phase, enabling adversaries to alter inputs and get wrong predictions. Model poisoning, on the other hand, involves injecting malicious data during the training phase to disrupt the model's learning. The combination of these different attack strategies, alongside the dynamic nature of the driving environment, makes it very challenging to anticipate and defend against new threats effectively.

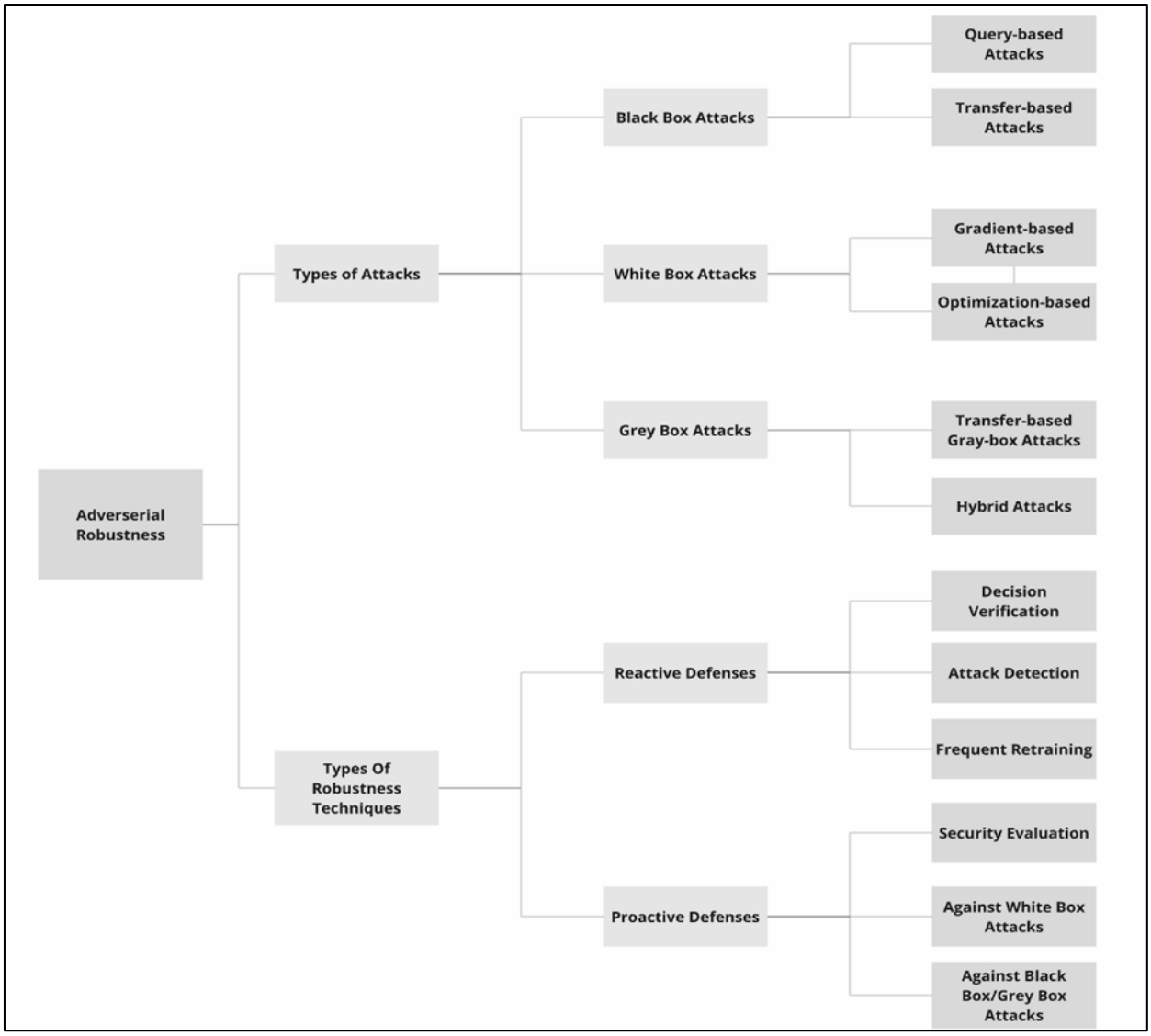

Figure 1 Types of adversarial attacks and countermeasures

Improving the robustness of deep learning models can sometimes lead to decreased accuracy on legitimate inputs, making it even more difficult to develop effective defense mechanisms. For example, when a model is trained to resist adversarial attacks, it might become overly cautious, resulting in misclassifications of normal data. Additionally, the black-box nature of deep learning models creates challenges in interpretability and explainability, making it difficult to identify and address adversarial vulnerabilities such as model overfitting, lack of transparency, transferability of attacks, model complexity, etc. These mechanisms are hard to comprehend, especially for those that use low-norm perturbations. For instance, barely noticeable changes to an image - like adjusting pixel values in a way that is invisible to the human eye can significantly impact the model's predictions during inference. This makes it very challenging for humans to detect potential dangers, as the changes may not be immediately evident, yet they can lead to serious errors in the model's performance by misidentifying a stop sign as a caution sign, which poses safety risks for autonomous driving. Therefore, the trade-offs in robustness, the difficulty in understanding the model, and the subtlety of attack methods make it challenging to effectively address vulnerabilities.

To better understand and explain the causes of misclassification in image classification-based deep learning models for autonomous driving systems, we can visualize the effects of adversarial examples in relation to the decision boundary. This vulnerability exposes a significant robustness problem in neural network models. Robustness is defined by how easily adversarial examples x' can be found near their original inputs. This raises critical concerns regarding the safety and reliability of systems built upon neural network components, particularly in light of their susceptibility to security threats.

Practically speaking, a robust model aims to minimize the gap between its decision boundary and the ideal Task boundary, thereby limiting an attacker's ability to generate adversarial examples. As illustrated in Fig. 1b, the decision boundary of the robust model, represented by the green line, closely aligns with the Task boundary, effectively reducing the potential for successful adversarial attacks. This alignment demonstrates the importance of developing robust defenses to enhance the security and reliability of deep learning models in autonomous driving systems.

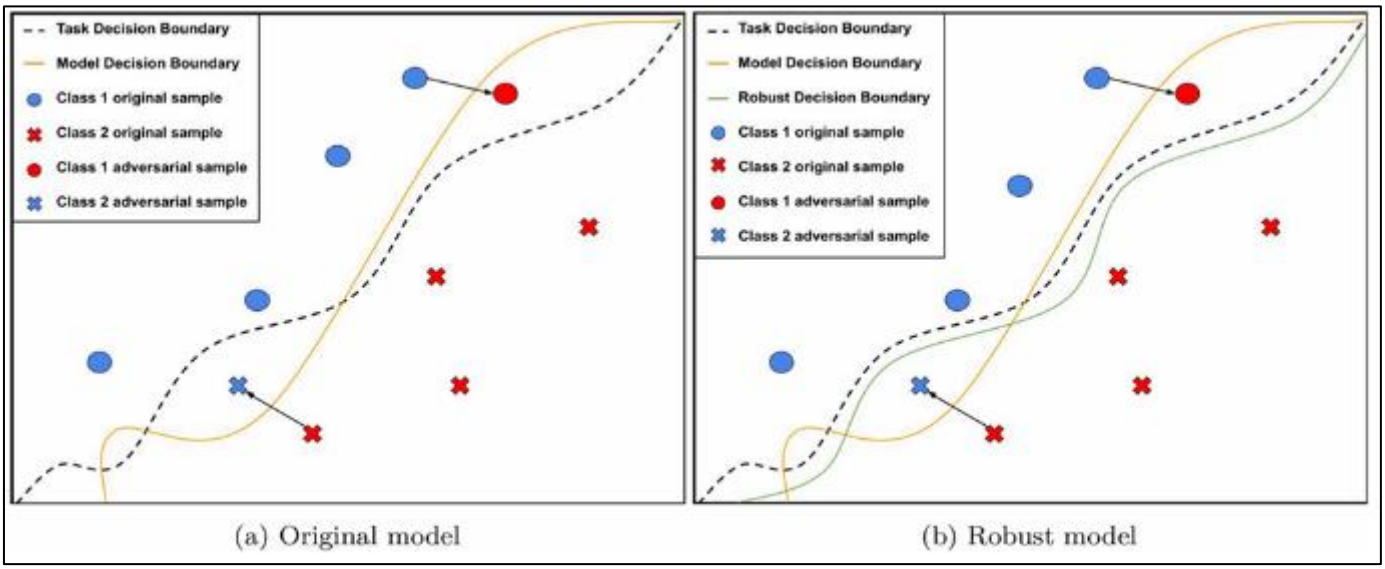

Figure 2 Robustness of neural network models and the role of their decision boundary. A robust model limits the space that an attacker can exploit, for crafting adversarial samples [3]

During the development process of AVs, deep learning plays a significant role due to the vast amount of data collected and the complex tasks involved in interpreting that data.

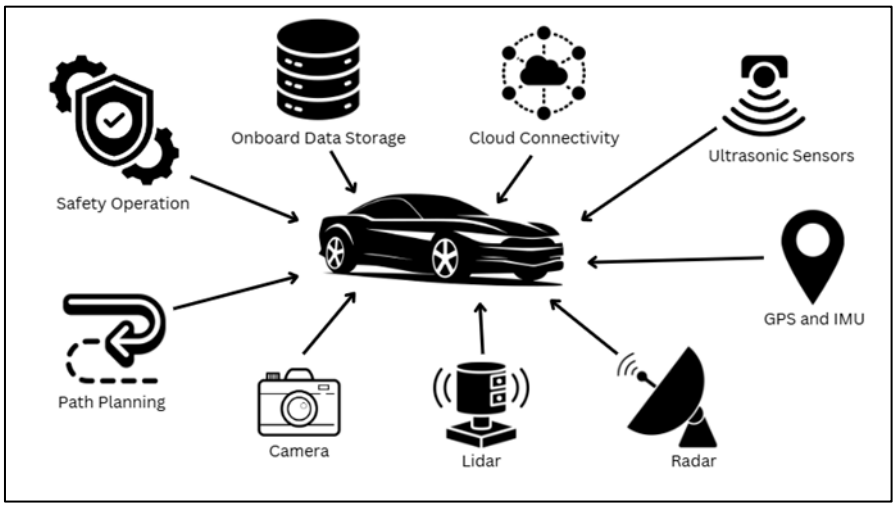

Figure 3 Autonomous vehicles (AVs) process various inputs using deep learning models and control units to perform various tasks

An autonomous vehicle (AV) gathers data from sensors and systems: Cameras capture images for object detection and lane marking. Lidar generates 3D maps of the environment. Radar detects the speed and distance of nearby objects, working well in all weather. Ultrasonic sensors assist in close-range detection, useful for parking. GPS and IMU provide precise location and movement tracking. Onboard Data Storage holds sensor data for real-time processing. Security operations protect data and system integrity. Cloud connectivity enables real-time updates and map sharing. Path planning calculates safe, efficient routes based on sensor data. These components allow AVs to perceive their surroundings, make decisions, and safely control movements.

2. Research Methodology

This research paper aims to study various aspects of adversarial robustness in autonomous driving systems (ADSs), focusing on the vulnerabilities of deep learning (DL) models integral to tasks such as object detection, semantic segmentation, and image classification. The paper delves into the taxonomy of adversarial attacks, categorizing them into knowledge-based approaches (white-box and black-box), intent-based strategies (targeted and untargeted), and location-based techniques (evasion and poisoning). Additionally, it explores the corresponding defensive mechanisms, including proactive methods like adversarial training and feature denoising, and reactive strategies such as anomaly detection and input filtering during inference.

The study also examines the challenges associated with implementing these defenses, particularly in balancing robustness with model generalizability and ensuring scalability in real-time applications. Emerging trends, such as hybrid defense frameworks combining proactive and reactive approaches, the adoption of explainable AI (XAI) for enhancing transparency, and the development of universal countermeasures, are also highlighted. By identifying key research gaps, such as the lack of standardized evaluation frameworks and limited exploration of cross-modal attacks, this research aims to contribute to the advancement of adaptive, scalable, and resilient ADSs capable of operating securely in adversarial dynamic environments.

3. Adversarial Attacks against Image Classification Deep-learning models in Autonomous Driving Systems

Deep learning models segment, analyze, and classify visual objects within driving environments. Object detection, semantic segmentation, and image classification models rely heavily on autonomous driving systems (ADS). These models play a major part in executing framework usefulness and forming intuition with the physical world. However, this review will specifically deal with evaluating potential adversarial attacks and countermeasures that can affect the performance of image classification deep learning models. We will also investigate different countermeasures planned to fortify these models against such attacks, assessing their effectiveness, limitations, and the challenges associated with implementing robust defenses. By centering on both the nature of adversarial dangers and the advancing defense mechanisms, this review gives insights into the vulnerabilities of picture classification DL models and how they can be relieved to ensure more dependable execution in real-world applications.

Deep learning models have emerged as a pressing concern, particularly in fields where reliability and security are paramount, and these attacks exploit weaknesses in model architecture, posing significant challenges to safe deployment. Due to the high dimensionality, complex highlights, linear separability, and constrained semantic understanding in picture information, tending to antagonistic vulnerabilities in DL models is fundamental to ensuring vigorous model execution. Table 1 gives a scientific categorization of existing attacks, categorizing them by perturbation scope, permeability, and measurement - key components of viable adversarial perturbations—highlighting areas where advanced improvements can mitigate rising threats. In Table 1, we display a scientific categorization of existing attacks, categorizing them based on the sorts of perturbation scope, visibility, and estimation, which are the fundamental features of a well-designed adversarial perturbation.

The Perturbation Scope represents the degree or greatness of malicious alteration to an input image in an adversarial setting. Here, we take into account universal and individual perturbations. Universal perturbations refer to unobtrusive, imperceptible changes made to distinctive inputs to trick a DL model. This perturbation is broad because it is consistently applied to various inputs. For instance, a universal perturbation could involve adding a nearly invisible noise pattern to all stop sign images, making the model consistently misinterpret stop signs as yield signs, regardless of the specific image of the stop sign used. This type of perturbation demonstrates a wide scope, as it generalizes across different inputs, fooling the model in a repeatable way without needing to customize the perturbation for each image. On the contrary, individual perturbations are designed to alter the model's behavior for a specific target input. The scope of this perturbation is more constrained, concentrating on just one input or a particular feature of an input. It is designed to exploit the targeted input's vulnerabilities or unique properties. For example, an individual perturbation might slightly modify a particular image of a pedestrian, causing the model to misidentify the pedestrian as an inanimate object or background. Since this perturbation is narrowly tailored to a single image or type of object, its scope is limited but highly precise, making it effective for specific targeted attacks.

The Perturbation Perceptibility refers to how noticeable modifications to input data are, and it can be further categorized into three different types: optimal, visible, and physical visibility. Optimal Visibility: In this case, the changes made to the input data are imperceptible to both humans and deep learning (DL) models. For example, in an image

classification task, a small amount of noise might be added to an image of a stop sign that does not alter the visual appearance significantly but is enough to confuse a DL model, causing it to misclassify the sign as a yield sign. The perturbation is designed to be statistically similar to the original input, optimizing the adversarial effect while remaining unnoticed. Visible Perturbations: These modifications are intentionally made to be noticeable to human observers but may still present challenges for DL models. An example could be adding graffiti or stickers on a traffic sign. While humans can see the graffiti and understand the context, the modifications might confuse a DL model that observers but may still present challenges for DL models. An example could be adding graffiti or stickers on a traffic sign. While humans can see the graffiti and understand the context, the modifications might confuse a DL model that relies on specific visual features for classification, potentially leading to incorrect predictions. Physical Visibility: In this scenario, the perturbations are perceptible to both humans and DL models, yet they can still deceive a well-trained network during inference. For instance, consider a situation where a person uses glasses or a projection to alter the appearance of a stop sign slightly; the modification can be seen by a human and detected by a DL system, but the specific features of the sign might be altered in a way that misleads the model into misclassifying it.

The Perturbation Metric measures the strength and size of distortions applied to input data using various Lp-norms to assess the statistical distance or similarities between the original and altered inputs. The p-value can take values such as 0, 1, 2, or ∞, each serving a different purpose:

- **L0-Norm**: This norm encourages sparsity in perturbations, which promotes making only a few changes to the input. This makes adversarial examples harder to detect. For instance, altering just a few pixels in an image of a stop sign can keep it looking mostly the same to human observers, enhancing the attack's effectiveness.
- **L1-Norm**: This metric controls the overall magnitude of the perturbations by minimizing their total sum. While it encourages sparsity, it can be computationally intensive. An example is applying a subtle brightness change across an entire image while keeping the total increase low to remain undetected.
- **L2-Norm**: The L2-Norm measures the size of perturbations in terms of Euclidean distance. By minimizing it, the added perturbations stay small, ensuring they are subtle. For example, slight color adjustments to an image of a pedestrian might be minor enough to go unnoticed by both models and human observers while still misleading the model.
- **L∞-Norm**: This norm assesses the maximum change across all pixels in the relevant areas of two images, focusing on the most significant alteration. For instance, drastically changing one pixel in a stop sign image could trick the model into misclassification. These metrics help analyze and quantify the impact of perturbations on input data, providing insight into adversarial examples in deep learning models.

Table 1 Taxonomy of Adversarial Attack Algorithms against Image-based Deep Learning Models [1]

Attack Algorithm	Perturbation Scale	Perturbation Matrix	Perturbation Perceptibility
HOUDINI	Individual	L2, L∞	Optimal
Fast Adaptive Boundary (FAB) Attack	Universal	L1, L2, L∞	Optimal
ShapeShifter	Individual	L2	Optimal
D-BADGE Attack	Individual/Universal	L2	Optimal
Physical-world Robust Adversarial Attack (PRAA)	Individual	L2	Physical
GenAttack	Individual	L2, L∞	Optimal
Robust Disappearance Attacks	Universal	L2	Physical
Adaptive Local Attack	Universal	L∞	Optimal/Physical
Adversarial Retroreflective Patch (ARP) Attack	Individual/Universal	Custom	Optimal/Physical
Gradient Norm Penalty (GNP) Attack	Universal	L∞	Physical
Time-aware Perception Attack (TPA)	Individual/Universal	L1, L∞	Optimal/Physical
Adaptive Square Attack	Universal	L1, L∞	Optimal

Soft-labeled Attack	Individual	L0, L1, L2	Optimal
Zeroth-Order Optimization (ZOO)	Individual/Universal	L2, L∞	Optimal
GhostStripe Attack	Individual/Universal	Custom	Visible
Multi-source Adversarial Attack Models	Universal	L∞	Optimal/Physical
Liu et al.	Universal	Custom	Physical/Visible
Cui et al.	Individual/Universal	L0, L1, L2, L∞	Optimal
OptiCloak	Individual/Universal	L∞	Physical
Lenticular printing Attack	Universal	L1, L2, L ∞	Optimal/Visible
Time-aware Perception Attack (TPA)	Individual/Universal	L1, L∞	Optimal/Physical
Adversarial Patch Attack on ADSs	Individual/Universal	L2, Custom	Physical/Visible
DeepBillboard	Individual/Universal	L2, Custom	Physical/Visible
PhysGAN	Individual/Universal	L2, L∞	Physical/Optimal
Physical One-Pixel Attack	Individual	L2	Optimal
Adversarial Patch Attack on ADSs	Individual/Universal	L2, Custom	Physical/Visible
Class Activation Mapping (CAM)	Individual	L2, L∞	Optimal
Scene Agnostics Adversarial Patch Attack	Universal	Custom	Optimal/Physical
Translation Invariant-based Attack	Individual/Universal	L1, L ∞	Optimal
Maximal Jacobian-based Saliency Map (JSMA) Attack	Individual	L0	Optimal
Dynamic Adversarial Attacks	Universal	L2	Physical
Scene-Specific Attacks	Universal	L∞	Visible/Physical
GenGradAttack	Individual	L2, L∞	Optimal
Out-of-distribution Attack	Universal	Custom	Optimal/Physical
ManiFool	Individual	L2	Optimal/Visible
L^1-Oriented Elastic-net Attacks to DNNs	Universal	L1	Optimal
Carlini-Wagner (C&W) Attack	Individual	L0, L2, L∞	Optimal
Targeted Attention Attack (TAA)	Universal	L2	Physical
Pinpoint Region Probability Estimation Network (PRPEN)	Individual/Universal	Custom	Optimal/Visible
Hierarchical Adversarial Attack (HAA)	Individual	L1,L2	Optimal
Robust Physical Perturbation (RP2)	Individual/Universal	Custom	Optimal
Adversarial Task-Transferable Attack	Universal	L2	Physical
Feature-aware Transferable Attack	Individual/Universal	L2	Optimal
Natural Light Illuminations-based Attack	Individual	Custom	Visible
Min- Max Optimization-Based Adversarial Attack	Universal	L0,L2, L∞	Optimal

Spatially Transformed Network (stAdv)	Universal	Custom	Optimal

3.1. Types of Adversarial Attacks

Adversarial attacks on deep learning models can be broadly classified into two categories: white-box and black-box attacks. This section focuses specifically on black-box adversarial attacks, analyzing their effects on image classification models utilized in ADSs.

3.1.1. White Box Attack

A white-box attack assumes the adversary possesses comprehensive knowledge of the target model and it's insights, including its architecture, hyperparameters, and gradient information. This privileged access enables the deployment of sophisticated optimization techniques to generate adversarial perturbations that effectively compromise the model's performance. These attacks are frequently utilized by model designers to evaluate and verify the robustness of deep learning models. However, in real-world scenarios, such as Autonomous Driving Systems (ADSs), attackers rarely succeed with white-box attacks due to the impracticality of gaining full access to the model's architecture, parameters, and gradients. For example, an attacker attempting to mislead an ADS into misclassifying a stop sign as a speed limit sign would require detailed knowledge of the system's neural network, which is rarely accessible in practical applications.

3.1.2. Black Box Attack

A black-box attack is a type of adversarial attack where the adversary possesses minimal or no knowledge about the internal architecture, parameters, or operational intricacies of the target system. Instead, the system is treated as a "black box," with the attacker relying on iterative querying and analyzing the system's outputs to infer vulnerabilities and craft adversarial inputs. Black-box attacks are widely employed in penetration testing and vulnerability assessment of system networks. For instance, in an image classification model for biometric security, an attacker might iteratively generate synthetic images to bypass authentication without knowing the neural network's architecture or weights. Black-box attacks can be further categorized into Transfer based Attacks, Decision based Attacks, Score based Attacks and, Optimization based Attacks.

Transfer Based Attacks. Transfer-based adversarial attacks exploit the transferability property of adversarial perturbations, enabling inputs crafted on a source model to deceive a target model, even without direct access attackers train a shadow model on a dataset approximating the target's, then use white-box techniques to generate optimized adversarial examples. These perturbations exploit shared vulnerabilities in decision boundaries, making them transferable across homogeneous and heterogeneous models. This method poses a significant security threat to real-world systems like biometric authentication or autonomous vehicles. For example, an attacker might use a surrogate model trained on public datasets to create adversarial road signs that mislead an autonomous driving system (ADS). Techniques like gradient smoothing, intermediate layer perturbation (ILP), token gradient regularization (TGR), and adaptive image transformation learners (AITL) further enhance transferability, emphasizing the critical need for robust defensive measures.

Table 2 Taxonomy of Adversarial Attack Algorithms against Image-based Deep Learning Models (Contd)[1]

Attack Algorithm	**Attack Scheme**	**Adversary's Intention**	**Adversary's Awareness**
Fast Adaptive Boundary (FAB) Attack	Gradient	Untargeted	WB
HOUDINI	Gradient	Targeted/Untargeted	BB
Hierarchical Adversarial Attack (HAA)	Gradient	Untargeted	WB
Soft-labeled Attacks	Gradient	Targeted	WB
Lenticular printing Attack	Geometry	Targeted/Untargeted	WB/BB
Maximal Jacobian-based Saliency Map (JSMA) Attack	Gradient	Targeted	WB
Cui et al.	Gradient	Targeted/Untargeted	WB

Translation Invariant-based Attack	Geometry	Targeted/Untargeted	WB
Spatially Transformed Network (stAdv)	Geometry	Targeted	WB
Adaptive Local Attack	Geometry	Targeted/Untargeted	WB
ManiFool	Geometry	Targeted/Untargeted	WB
Carlini–Wagner (C&W) Attack	Optimization/ Gradient	Targeted/Untargeted	WB/BB
GenGradAttack	Optimization	Targeted	BB
Scene-Specific Attacks	Geometry	Targeted/Untargeted	WB
Scene Agnostics Adversarial Patch Attack	Optimization/ Transfer	Targeted	WB
L1 -Oriented Elastic-net Attacks to DNNs	Transfer	Targeted	WB/BB
Feature-aware Transferable Attack	Transfer	Targeted	WB/BB
Robust Physical Perturbation (RP2)	Transfer/Gradient	Targeted/Untargeted	BB/WB
Natural Light Illuminations-based Attack	Transfer/Geometric	Targeted/Untargeted	BB/WB
Min-Max Optimization-Based Adversarial Attack	Optimization/ Gradient	Targeted/Untargeted	BB
Pinpoint Region Probability Estimation Network	Transfer	Targeted/Untargeted	BB
Gradient Norm Penalty (GNP) Attack	Gradient/Transfer	Untargeted	WB/BB
Adversarial Task-Transferable Attack (ATTA)	Transfer	Untargeted	BB
Dynamic Adversarial Attacks	Transfer/Score	Targeted/Untargeted	BB
Targeted Attention Attack (TAA)	Transfer/Decision	Targeted	BB
PhysGAN	Transfer/Score	Targeted/Untargeted	BB
DeepBillboard	Transfer	Targeted	BB
Adversarial Patch Attack on ADSs	Transfer/Score	Targeted	BB
Liu et al.	Gradient/Transfer	Targeted/Untargeted	BB/WB
Time-aware Perception Attack (TPA)	Score	Targeted/Untargeted	BB
OptiCloak	Score/Gradient	Untargeted	BB/WB
Multi-source Adversarial Attack Models	Gradient/Score	Untargeted	BB/WB
Out-of-distribution Attack	Decision	Targeted/Untargeted	WB/BB
GhostStripe Attack	Decision/Geometric	Targeted/Untargeted	BB/WB
D-BADGE Attack	Decision	Targeted/Untargeted	BB
Class Activation Mapping (CAM)	Decision	Targeted	BB
Zeroth-Order Optimization (ZOO)	Transfer/Score	Targeted/Untargeted	BB/WB
Physical One-Pixel Attack	Transfer/Gradient	Targeted/Untargeted	BB
Adaptive Square Attack	Transfer/Score	Targeted	BB/WB
Adversarial Retroreflective Patch (ARP) Attack	Score/Geometric	Targeted/Untargeted	BB/WB

ShapeShifter	Decision	Targeted/Untargeted	WB
Robust Disappearance Attacks	Decision	Targeted	WB/BB
GenAttack	Decision	Targeted	BB
Physical-world Robust Adversarial Attack (PRAA)	Score	Targeted	BB

Decision Based Attacks. Decision-based attacks are sophisticated adversarial techniques where attackers manipulate input data to induce misclassifications in a model, relying exclusively on the model's final output rather than its confidence scores or probability distributions. These attacks are highly targeted, leveraging techniques like Boundary Attacks, Lenticular Printing Attacks, and GhostStripe, which begin with large perturbations and progressively fine-tune them to remain near the decision boundary. This iterative refinement enables attackers to achieve high success rates across various domains. In the context of Autonomous Driving Systems (ADSs), decision-based attacks can lead to misclassifications with accuracy rates exceeding 90%, especially in tasks such as traffic sign recognition and lane detection, posing severe risks to operational safety. While strategies like adversarial training and boundary-aware regularization offer some defense, the stealthy and adaptive nature of these attacks demands ongoing research and innovation to develop more robust security measures for critical applications.

Score Based Attacks. Score-based attacks are a class of black-box adversarial techniques in which attackers leverage a model's output scores, such as softmax probabilities, to indirectly infer gradient information and craft adversarial perturbations. By analyzing variations in these output scores, adversaries can strategically alter input data to manipulate the model's predictions for a targeted class, even when the internal architecture of the model remains inaccessible. This method poses significant challenges in systems like Autonomous Driving Systems (ADSs), where subtle, imperceptible changes to input data such as images can easily deceive image classification models, compromising their performance. Unlike decision-based attacks, which are based solely on the final predicted labels, score-based attacks offer attackers additional insight into model behavior through output scores, making them somewhat easier to execute but still a considerable threat. Defenses against such attacks include adversarial training, which incorporates adversarial examples during model training to enhance robustness, as well as regularization techniques and self-supervised learning strategies to fortify models against such vulnerabilities.

Optimization Based Attacks. Optimization-based attacks involve framing an optimization problem to identify the minimal perturbation necessary to mislead a target model into misclassification. These attacks iteratively modify the input, aiming to maximize the model's loss function or reduce its confidence in the true class while keeping alterations imperceptible. By leveraging finite difference methods to estimate gradients, attackers can successfully execute these strategies even without direct access to the model's gradient information. This makes optimization-based attacks particularly potent against Autonomous Driving (AD) models, where subtle adjustments exploit the model's fine-grained sensitivities. Notable techniques include Judge Deceiver, a prompt injection attack targeting Large Language Models (LLMs), Zeroth Order Optimization (ZOO), a black-box method that only requires input-output access to deep neural networks, and the Carlini & Wagner (C&W) attack, which formulates adversarial example generation as an optimization problem to find minimal perturbations. These strategies, collectively classified as adversarial example attacks, challenge model robustness by forcing models into making erroneous predictions through imperceptible input modifications.

Adversarial attacks, both white-box and black-box, pose critical risks to deep learning models, especially in high-stakes environments like Autonomous Driving Systems (ADSs). White-box attacks leverage full model access to manipulate inputs effectively, though real-world application is limited due to access constraints. Black-box attacks, including transfer-based, decision-based, score-based, and optimization-based methods, exploit output-based vulnerabilities, bypassing the need for internal model knowledge. These attacks target decision boundaries and model sensitivities, making them potent threats to systems like traffic sign and lane detection. While defenses such as adversarial training offer some protection, ongoing advancements are essential to counter the increasingly sophisticated nature of these attacks.

3.2. Challenges in Implementing Adversarial Attack Techniques on Image Classification DL Models for ADSs

3.2.1. Model Transparency and Explainability

- **Challenge:** Deep learning models, particularly Convolutional Neural Networks (CNNs), are often regarded as "opaque" due to their "black-box" nature. The intricacies of their decision-making pathways are non-trivial to

interpret, making it challenging to decipher how the model reacts to adversarial perturbations or to pinpoint the root causes of misclassifications.

- **Implication:** The lack of model interpretability complicates both debugging and countermeasure development against adversarial attacks. It undermines the system's transparency, eroding confidence in Autonomous Driving Systems (ADSs). Consequently, developers and human operators face difficulty in validating the rationale behind the model's decisions, leading to a potential decrease in system reliability and trust.

3.2.2. Adaptive Countermeasure Strategies

- **Challenge:** As adversarial tactics advance, defense mechanisms must continuously evolve. Techniques like adversarial training, defensive distillation, and input preprocessing are designed to counteract adversarial perturbations. However, these defenses frequently create new attack vectors or fail to withstand emerging attack paradigms.
- **Implication:** The dynamic arms race between the evolution of adversarial strategies and the development of countermeasures presents a persistent challenge in fortifying Autonomous Driving Systems (ADSs). Attackers remain agile, continuously adapting to circumvent the latest defense innovations, making it difficult to establish a robust, impervious security framework.

3.2.3. Resilient Sensor Fusion

- **Challenge:** The difficulty arises in designing adversarial perturbations that can successfully target vulnerabilities across the diverse data streams of sensors. These attacks must account for the complex interactions between sensor outputs during the fusion process, making it challenging to develop perturbations that can manipulate the system's perception in real-time driving scenarios.
- **Implication:** The complexity of managing and manipulating data from multiple sensor streams in a cohesive manner makes it harder for attackers to succeed. However, this also complicates the development of effective defensive strategies, as each sensor type requires specialized handling to safeguard against adversarial attacks while maintaining seamless integration in dynamic driving environments.

3.2.4. Resource Constraints in Real-Time Systems

- **Challenge:** Developing adversarial perturbations that exploit system vulnerabilities while respecting the real-time operational constraints of ADSs is a complex task. These systems require rapid decision-making based on incoming sensor data, so creating adversarial examples that are both computationally efficient and effective in influencing the model in real-time is a significant obstacle.
- **Implication:** To overcome this challenge, adversarial attacks must be tailored to be lightweight and fast, given the limited computational resources of embedded systems in autonomous vehicles. This means attack algorithms need to be optimized for efficiency, ensuring they can effectively compromise the model's performance without taxing the system's processing capabilities, which could otherwise hinder the vehicle's operational integrity.

3.2.5. Real-time Input Processing

- **Challenge:** Autonomous Driving Systems (ADSs) must process input data in real-time, making it difficult to design adversarial attacks that can effectively alter model decisions within tight time constraints. The challenge lies in creating attacks that are not only effective but also computationally efficient, capable of influencing the model's predictions at each decision-making timestamp without causing significant delays.
- **Implication:** Adversarial perturbations must be designed for quick execution in order to get past the system's defenses without causing lag or interfering with the real-time flow of data processing. This is because speed and efficiency are essential. Crafting such attacks is particularly challenging given the limited processing resources in embedded systems, which must balance attack effectiveness with minimal computational overhead to avoid hindering ADS performance.

3.2.6. Adaptive Defense Mechanisms

- **Challenge:** The rapid evolution of adversarial attack strategies necessitates corresponding advancements in defense mechanisms. Techniques like adversarial training, defensive distillation, and input preprocessing are designed to fortify models against perturbations, but they often create new attack surfaces or are vulnerable to emerging adversarial tactics. The dynamic nature of these attacks makes it difficult to maintain effective, adaptive defenses.

- **Implication:** This continuous "arms race" between adversarial techniques and countermeasures complicates the development of robust, foolproof systems for ADSs. As attackers innovate to bypass current defenses, ensuring the security and resilience of autonomous systems becomes a perpetual challenge, requiring ongoing refinement and adaptive strategies to stay ahead of adversarial threats.

3.2.7. Evasion Under Continuous Learning

- **Challenge:** Autonomous Driving Systems (ADSs) often leverage continuous learning, where models dynamically adapt to incoming data streams and adjust decision boundaries. This creates a moving target for adversaries, as adversarial examples that succeed initially may lose efficacy after the model retrains or refines its parameters. Attackers must contend with this evolving landscape, where the system's adaptive nature can quickly render attacks obsolete.
- **Implication:** The ongoing "cat-and-mouse" scenario between attackers and the adaptive nature of ADS models complicates the crafting of stable and persistent adversarial perturbations. To remain effective, attacks must be designed with the foresight to anticipate and adapt to model updates, otherwise, they risk becoming ineffective as the system's learning process evolves over time.

4. Adversarial Robustness and Defensive Techniques for Image Classification Models in Autonomous Driving

This section provides a comprehensive analysis of proactive and reactive adversarial defense methodologies presented in the literature, aimed at enhancing the resilience of image classification deep learning (DL) models against adversarial perturbations in Autonomous Driving Systems (ADSs). These strategies are designed to safeguard DL models by either bolstering adversarial robustness or identifying malicious inputs during inference. Conventional defenses primarily emphasize reactive measures, targeting the detection and identification of adversarial attacks post-occurrence. Conversely, proactive defense mechanisms aim to preemptively fortify the model's resilience by enhancing adversarial tolerance, optimizing generalization performance, and minimizing susceptibility to adversarial perturbations.

Tables 3 and 5 present a detailed taxonomy of existing defense algorithms, categorizing them based on their objectives, methods, and ability to resist adversarial attacks. The objectives of these defenses include protecting deep learning models from adversarial inputs by building robust systems, detecting and preventing attacks, and improving the model's generalization. These defenses are divided into proactive strategies, which focus on strengthening the model's resilience, and reactive strategies, which aim to identify and counteract adversarial inputs after they occur.

The taxonomy also highlights the defense approach, which refers to the methods used to make image classification models more resistant to attacks. These methods include gradient masking, adversarial defense techniques, statistical methods, data preprocessing, ensemble training, and proximity-based approaches

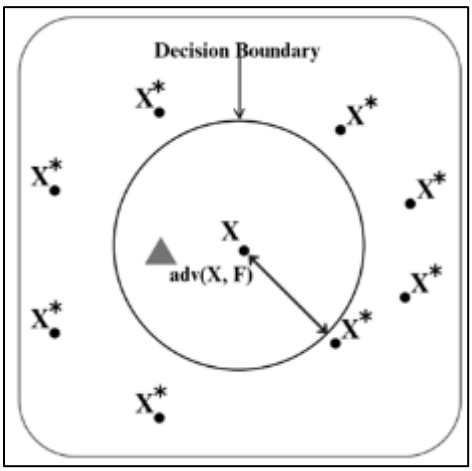

Figure 4 Overview of Roughness Metrix

Additionally, the resilience of each defense is assessed by examining the types of adversarial attacks it can counter, whether in white-box, black-box, or mixed settings. Tables 5 and 6 show that some algorithms are effective against multiple types of attacks, demonstrating versatile robustness.

The following section focuses on two main defense techniques: proactive and reactive strategies. Each section highlights the core principles of these approaches and their effectiveness in countering advanced adversarial attacks. This detailed analysis aims to help researchers gain a clearer understanding of these methods and select the most appropriate strategy to address their specific challenges.

4.1. Proactive Adversarial Robustness Defense Algorithms

Proactive defense algorithms are designed to enhance the generalization capabilities of deep learning (DL) models while reducing their vulnerability to adversarial perturbations. This subsection provides an in-depth analysis of various proactive adversarial robustness strategies and defense mechanisms within the domain of image classification DL models.

Tables offer a consolidated overview of proactive defensive techniques, encapsulating diverse methodologies, defense architectures, and robustness metrics such as accuracy against adversarial attacks. These tables serve as a resourceful guide for researchers, enabling them to gain nuanced insights into these methodologies and align them effectively with the specific requirements of their problem domains.

Defense strategies for mitigating adversarial attacks in deep learning (DL) models, particularly within Autonomous Driving Systems (ADS), are categorized into proactive and reactive approaches. Proactive strategies focus on fortifying model robustness and improving adversarial tolerance. These methods often utilize preprocessing techniques, ensemble-based defenses, and gradient masking to create models that resist perturbations. For instance, Ensemble Adversarial Training, which combines diverse defenses, has demonstrated high resilience against attacks like FGSM and PGD, with some studies reporting up to a 30% increase in robustness under white-box (WB) scenarios. Similarly, Smooth Adversarial Training (SAT) aims to improve generalization by leveraging smoother decision boundaries, reducing vulnerability to optimization-based attacks.

Reactive strategies, on the other hand, are designed to detect and mitigate adversarial inputs after they occur. Techniques like Feature Squeezing, which compresses input dimensions, have been effective in identifying adversarial perturbations with minimal computational overhead. Advanced methods such as Iterative Trimming focus on loss minimization caused by adversarial examples, showing notable improvements in detection rates, particularly in black-box (BB) and gray-box (GB) scenarios.

Table 3 Taxonomy of Proactive Defense Techniques against Adversarial Attacks on the Image Classification Deep Learning Models [1]

Defense Algorithms	Methodology	Resilient to: Attack algorithm	Resilient to: Adversarial Awareness
Ensemble-based Defense (ED)	Ensemble	FGSM, BIM, PGD	WB
Defensive Distillation	Gradient Masking	FGSM, JSMA	WB
Free Adversarial Training	Preprocessing, Ensemble	PGD, C&W, SPA	WB, BB, GB
Generative Adversarial Training	Preprocessing, Ensemble	I-FGSM, GAN-Based Attacks	BB, GB
FGSM +Random Input Initialization-based Adversarial Training	Preprocessing, Ensemble	FGSM, ZOO, PGD	WB, BB
Gradient Masking	Gradient Masking	FGSM, PGD, & C&W	WB, BB, GB
Adaptive Mix-Mode Defense (AMMD)	Preprocessing, Ensemble, Gradient Masking	PGD, FGSM, BIM	PGD, FGSM, BIM
Ensemble Input Transformation	Ensemble, Preprocessing	FGSM, I-FGSM, BIM, DeepFool, C&W	BB, GB

Adversarial Training	Preprocessing, Ensemble	FGSM, DeepFool	WB, BB, GB
Gradient Shaping	Preprocessing	FGSM, DeepFool, C&W, UAP	BB
Smooth Adversarial Training (SAT)	Preprocessing, Ensemble	FGSM, C&W	WB, BB, GB
Adversarial Boosting	Gradient Masking	PGD, FGSM, BIM, BPDA, DeepFool, C&W	BB, GB
Random input Transformation	Preprocessing	FGSM, M-BIM, ManiFool, DeepFool, C&W	WB, BB, GB
Adversarial Polytope	Gradient Masking	DeepFool, FGSM, PGD	WB, GB
High-Level Representation Guided Denoiser (HGD)	Preprocessing	I-FGSM, FGSM, PGD	WB, BB, GB
Batch Adjusted Network Gradients (BANG)	Preprocessing	FGSM, DeepFool, UAP	WB BB, GB
Model Regularization Using Penalty Term	Preprocessing, Proximity	MI-FGSM, PGD, Shadow, C&W, BA, SPSA	WB, BB, GB
DeepCloak-based Feature Masking	Gradient Masking	PGD, JSMA, FGSM, BIM,R-FGSM, DeepFool, C&W	BB, GB
Sample-dependent Adversarial Initialization (SAI)	Preprocessing	FGSM, I-FGSM, C&W, MI-FGSM	WB, BB
Self-Ensemble Adversarial Training (S-EAT)	Gradient Masking, Ensemble	PGD, MIM, C&W, AutoAttack WB, BB	WB, BB
Noisy Adversarial Training (NAT)	Preprocessing, Ensemble	Natural Noise, BIM, DeepFool, C&W	BB, GB
NULL Labeling	Preprocessing	BA, SPSA	WB, BB, GB
Dual-branch Network (DBN)	Ensemble	MI-FGSM,PGD, Shadow, C&W, BA, SPSA	WB, BB, GB
Robust Feature Purification (RFP)	Preprocessing, Ensemble, Gradient Masking, Proximity	FGSM, BIM, PGD, C&W, BPDA MI-FGSM,PGD, Shadow, C&W, BA, SPSA	WB, BB, GB
ManiFool Geometric Transformation (MGT)	Preprocessing, Proximity, Ensemble	ManiFool, DeepFool	WB, GB
The Barrage of Random Transforms (BaRT)	Preprocessing	FGSM, BIM, PGD, C&W, BPDA	WB, BB, GB

Auxiliary Classifier-based GAN (AC-GAN)	Preprocessing, Ensemble	FGSM, R-FGSM, GAN-Based Attacks, DeepFool	BB, GB
Ensemble Adversarial Training (EAT)	Gradient Masking, Ensemble	FGSM, I-FGSM, ILCM, BIM, PGD, C&W	WB, BB

Overall, resilience across these methods is evaluated under WB, BB, and GB settings, with ensemble-based defenses like MultiMagNet offering robust multi-directional resistance to attacks such as DeepFool and C&W. Statistical approaches, including Adaptive Noise Layer (ANL) and Robust Intrinsic Modeling, demonstrate versatility by addressing a broader range of perturbation strategies. Research demonstrates that hybrid approaches can curtail adversarial success rates by up to 50%, underscoring their pivotal role in fortifying ADS frameworks. These innovative methodologies mark a substantial leap in bolstering the adversarial robustness of image classification models, enhancing their defense against advanced attack strategies.

4.2. Reactive Defense Strategy against Adversarial Attacks Introduced on the ADS-based DL Model

The scientific community has extensively investigated reactive defense mechanisms specifically designed for ADS-oriented deep learning frameworks. These strategies focus on real-time detection and mitigation of adversarial intrusions, curbing their effects and preserving system integrity. The following section delves into the critical categories of reactive adversarial robustness techniques and algorithms employed within image classification deep learning models.

Table 4 Taxonomy of Reactive Defense Techniques against Adversarial Attacks on the Image Classification Deep Learning Models [1]

Defense Algorithms	**Methodology**	**Resilient to: Attack algorithm**	**Resilient to: Adversarial Awareness**
Iterative trimming loss minimization	Proximity	FGSM, JSMA, DeepFool, C&WWB, GB	WB, GB
Neural Cleanse	Adversarial defense method, Proximity	FGSM, BIM, JSMA, DeepFool, C&W	WB, BB
Feature Squeezing	Preprocessing	FGSM, BIM, JSMA, C&W DeepFool	WB
Robust Anomaly Detection (RAD)	Proximity, Preprocessing	FGSM, BIM, DeepFool, C&W	BB, GB
Frequency-Adaptive Compression and REconstruction (FARE)	Adversarial defense method, Proximity	FGSM, MI-FGSM, PGD, C&W	WB, BB, GB
Quantifying Uncertainty Estimates in Training Data	Statistics, Proximity	FGSM, JSMA, C&W, GDA, DeepFool, POE	WB, BB, GB
Robust Co-Teaching	Ensemble	FGSM, PGD, DeepFool, C&W, UPA, ZOO	WB, BB, GB
MultiMagNet Adversarial Defense Framework	Preprocessing, Proximity,Ensemble	FGSM, BIM, DeepFool, C&W, WB, BB, GB	WB, BB, GB
Carrara et al.	Proximity	L-BFGS, FGSM	WB

Single Value Decomposition (SVD)	Preprocessing, Statistics	FGSM, JSMA	WB, BB, GB
Fine-pruning	Proximity, Preprocessing	FGSM, BIM, DeepFool, C&W	BB, GB
Feature Purification	Proximity	FGSM, BIM, DeepFool	BB, GB
Progressive Unified Defense (PUD)	Ensemble, Proximity, Preprocessing	PGD, C&W	BB, GB
Robustand Generalized Defense (Ensemble-based Adversarial Training (EAT))	Ensemble, Preprocessing	Patch noise	WB
Cohen et al.	Proximity	FGSM, JSMA, DeepFool, C&W	WB
Adaptive Noise Layer (ANL) (Pix2pix strategy)	Proximity, Preprocessing, Statistics	FGSM, BIM, PGD, C&W, MI-FGSM, and AutoAttack	WB, BB
Robust Intrinsic Modelling algorithm	Statistics, Proximity	FGSM, BIM, DeepFool	BB, GB

Reactive defense strategies for adversarial attacks on ADS-based deep learning models employ diverse methodologies, including preprocessing, proximity-based techniques, ensemble learning, and statistical analysis. These approaches exhibit varying degrees of resilience against widely used adversarial algorithms, such as FGSM, BIM, PGD, JSMA, DeepFool, and C&W, in white-box (WB), black-box (BB), or gray-box (GB) settings.

Table 5 The table outlines proactive defense algorithms, models, and their robustness accuracy against adversarial attacks.[1]

Authors/Paper	**Defense Algorithm**	**Defense Model**	**Robustness Accuracy**
Jia et al.	Sample-dependent Adversarial Initialization-based Adversarial Training	ResNet18	58.46%, 48.17%, 63.71%,53.33%, 64.14%
Tramèr et al.	Ensemble Adversarial Training	Inceptionv3,Inception v2,ResNet50	78%, 65.3%, 89.9%, 47.9%, 53.6%, 33%
Papernot et al.	Defensive Distillation	Multi-scale CNN	98.41%, 98.99%
Shafahi et al.	Free Adversarial Training	ResNet50, ResNet101, ResNet152	40%, 36.44%, & 35.94%
Wang et al.	Self-Ensemble Adversarial Training	ResNet18	60.2%, 60.31%, 82%, & 70.26%
Guo et al.	Ensemble Input Transformation	ResNet50, ResNet101, DenseNet169, Inceptionv4	70.37%, 71.52%,71.47%, 70.50%
Wong et al.	FGSM + Random Input Initialization-Based I Adversarial Training	ResNet50	96.7%
Gao et al	DeepCloak-based Feature	ResNet164	50.17%

	Masking		
Kantchelian et al.	Adversarial Boosting	RBF-SVM	60%
Saleh et al.	Smooth Adversarial Training	ResNet50	76.8% & 77.0%
Azim et al.	Ensemble-based Defense (ED)	Standard CNN, EfficientNet	90.00%, 51%
Wong et al.	Adversarial Polytope	Deep-CNN	61.9%
Xie et al.	Random data Input Transformation	Inceptionv3,ResNet101, Inception-ResNetv2	92.4%, 98.3%, & 99.3%
Hosseini et al.	Three-stage NULL Labeling	Deep-CNN	99.46% & 97.37%
Raff et al. (BaRT)	Barrage of Random Transforms	ResNet50 & Inceptionv3	78.90% & 63.51%
Rozsa et al.	Batch Adjusted Network Gradients	LeNet Series	90.33% to 41.34%
Mirnateghi et al.	Dual-branch Network (DBN)	VGG16, ResNet50, Standard CNN	93.50%
Lee et al.	Gradient Masking	Deep-CNN	67.8%, 60%, 46.9%
Nayak et al.	Robust Feature Purification (RFP)	ResNet50, VGG19, VGG16	92.10%
Khan et al.	Adaptive Mix-Mode Defense (AMMD)	Resnet-152, GoogleNet	99.00%, 88.00%, 55%
Liao et al.	High-Level Representation Guided Denoiser	Inception v3	73.9% & 74.8%
Xu et al.	S^3ANet	DCNN	93.00%
Shi et al.	Random Feature Nullification (RFN)	Standard CNN	78%, 83%, 89%

For instance, iterative trimming loss minimization and fine-pruning leverage proximity-based strategies to counter FGSM, DeepFool, and C&W attacks effectively in WB and GB scenarios. Preprocessing techniques, such as feature squeezing and single-value decomposition (SVD), focus on mitigating FGSM, BIM, and JSMA attacks. Ensemble-based defenses, like robust co-teaching and multiMagNet frameworks, enhance resilience across multiple attack types, including ZOO and UPA, in all adversarial settings.

The table 5 summarizes several proactive defense algorithms designed to enhance the robustness of image classification deep learning models against adversarial attacks. Notable approaches include Defensive Distillation by Papernot et al., which achieved high accuracy of 98.41%–98.99% with multi-scale CNNs, and Ensemble Adversarial Training by Tramèr et al., with varied accuracy results ranging from 33% to 89.9% across different models like Inceptionv3 and ResNet50. Self-Ensemble Adversarial Training by Wang et al. showed accuracy between 60.2% and 82% for ResNet18, while Ensemble Input Transformation by Guo et al. yielded moderate accuracy ranging from 70.37% to 71.52% across multiple models. FGSM + Random Input Initialization by Wong et al. achieved an impressive 96.7% for ResNet50, and Robust Feature Purification (RFP) by Nayak et al. resulted in 92.10% accuracy for models like ResNet50 and VGG16. Adaptive Mix-Mode Defense (AMMD) by Khan et al. reached 99% accuracy for ResNet152, and Three-stage NULL Labeling by Hosseini et al. achieved 99.46% accuracy for Deep-CNNs. These findings highlight the range of effectiveness among defense strategies, underscoring their potential and areas for further optimization.

Advanced hybrid approaches, such as Adaptive Noise Layer (ANL) and Progressive Unified Defense (PUD), integrate proximity, preprocessing, and statistical strategies, demonstrating robustness against sophisticated attacks like MI-FGSM and AutoAttack. Techniques such as Neural Cleanse and Frequency-Adaptive Compression and Reconstruction (FARE) employ adversarial defense methods combined with proximity analysis to bolster defenses against a wide

spectrum of attack vectors. Collectively, these strategies illustrate a multi-faceted approach to fortifying deep learning models against adversarial threats.

4.3. Challenges Associated with Implementing Defense Strategies for Image Classification Deep Learning Models in ADS Against Adversarial Attacks

Defending and fortifying image classification deep learning (DL) models for autonomous decision systems (ADS) against adversarial attacks presents a myriad of challenges, distinct from conventional image classification tasks. The criticality of ADS in safety-sensitive domains mandates defense mechanisms that not only exhibit robust adversarial resilience but also maintain high accuracy on legitimate inputs. Any compromise in accuracy risks jeopardizing operational safety, such as in autonomous navigation systems. These challenges are compounded by the intricate nature of adversarial attack patterns, the trade-offs inherent in current defense strategies, and the adaptive evolution of adversarial techniques. Below are the primary obstacles in deploying effective defense frameworks for ADS:

Generalization and Robustness Trade-offs: In autonomous driving systems (ADS), defense mechanisms such as adversarial training improve model robustness against specific adversarial perturbations but often at the expense of generalization. While the model becomes resilient to certain attack types, its performance on clean, real-world inputs may degrade, potentially causing unsafe driving decisions. For instance, adversarially trained models might fail to accurately recognize objects or pedestrians under normal conditions due to overfitting to adversarial distributions, which compromises safety.

High Computational Overhead: Many defense algorithms, including adversarial training, ensemble techniques, or fine-pruning, require substantial computational resources. This computational burden is a critical limitation for ADS, which demand real-time decision-making capabilities. Processing time-intensive defense mechanisms in dynamic environments, such as traffic, where split-second decisions are necessary, could delay system responses and lead to catastrophic consequences, such as collisions or misinterpretations of traffic signals.

Transferability of Adversarial Attacks: In ADS, adversarial examples crafted for one neural network can transfer to others with similar architectures, making model-specific defenses inadequate. For instance, an adversarial image designed to fool a vision system in one car model may successfully deceive a similar system in another brand or model, bypassing the proprietary defenses in place. This cross-model vulnerability creates significant challenges in developing robust ADS solutions that are immune to widespread attacks.

Dynamic Nature of Adversarial Attacks: Adversarial attack strategies are continuously evolving, introducing sophisticated techniques like adaptive attacks that exploit known weaknesses in existing defenses. In ADS, attackers can generate perturbations that target the specific defense mechanism employed, such as exploiting patterns in adversarial trained networks. This dynamic nature necessitates constant updates and innovations in defense strategies, which is resource-intensive and difficult to sustain in commercial autonomous vehicles.

Limited Effectiveness Across Attack Types: Most existing defense techniques are tailored for specific attack scenarios, such as white-box or black-box attacks, but fail to provide comprehensive protection against hybrid or novel adversarial strategies. In the context of ADS, this limitation is critical, as autonomous vehicles encounter diverse environmental conditions and potential attack scenarios. For instance, a defense mechanism effective against white-box attacks might be vulnerable to black-box or gray-box strategies, leaving the vehicle susceptible to malicious exploitation.

Scalability to Large Datasets: Preprocessing-based defenses, such as feature squeezing or noise injection, often become computationally prohibitive when applied to the vast datasets required for training ADS. Autonomous driving systems rely on high-resolution imagery and extensive datasets for accurate scene understanding and decision-making. The resource-intensive nature of these preprocessing methods can slow down system training or deployment, making them unsuitable for real-world, large-scale applications in the ADS domain.

Loss of Interpretability: Defensive techniques like gradient masking and high-dimensional transformations obscure the internal workings of deep learning models, reducing their interpretability. For ADS, interpretability is crucial for validating and certifying system safety. A lack of clarity in how the model arrives at decisions—especially in defense mechanisms—can lead to reduced trust in the system and difficulty in diagnosing and rectifying failures, which is critical in high-stakes applications like autonomous driving.

Adversarial-Aware Training Bottlenecks: Incorporating adversarial examples during training often causes overfitting to specific adversarial patterns, reducing the model's ability to adapt to new or unseen attacks. For ADS, this overfitting

could lead to scenarios where the vehicle performs well under test conditions but fails to recognize or respond to real-world adversarial scenarios, endangering passengers and pedestrians.

5. Emerging Trends and Key Findings in Adversarial Attacks and Countermeasures

The domain of adversarial attacks and defenses in deep learning is witnessing rapid advancements, characterized by increasingly complex and diverse attack Frameworks. A prominent trend involves adaptive and transfer-based attacks, where adversarial examples crafted for one model exhibit high efficacy when applied to other models with similar or differing architectures. This underscores the necessity of designing robust, transferable defense strategies to mitigate adversarial threats in dynamic and heterogeneous environments, such as those in Autonomous Decision Systems (ADS).

Efforts are intensifying to circumvent well-optimized defenses, particularly in image classification models for ADS. Adversaries exploit vulnerabilities across homogeneous and non-homogeneous models, necessitating innovative solutions that maintain predictive accuracy while enhancing robustness. Existing defensive approaches, such as adversarial training, defensive distillation, and random input manipulation, have shown promise but often sacrifice model accuracy for improved resilience. This trade-off highlights the ongoing challenge of balancing robustness with generalization, particularly in safety-critical ADS applications.

The integration of cloud-based data processing and distributed learning frameworks is an emerging paradigm aimed at fortifying model resilience. Cloud platforms enable models to train on extensive datasets aggregated from interconnected devices in the driving environment, enhancing their adaptability to diverse attack vectors. Similarly, distributed and federated learning frameworks promote decentralized training across Internet of Things (IoT) devices, fostering collaborative learning that strengthens the robustness of individual sub-models and their parent ADS models.

Verifying and quantifying the robustness of image classification models against adversarial threats has become a critical focus. Factors such as regulatory compliance, safety imperatives, evolving threat landscapes, and the complexity of deep learning systems have driven this trend, ensuring secure and dependable ADS operations. Techniques for universal defense frameworks are also gaining traction, addressing limitations of current methodologies, which often target specific attack types and lack comprehensive applicability.

An additional trend involves leveraging the inherent challenges of white-box attacks in safety-critical applications. While attackers may not gain direct access to gradient information of target models, shadow models are increasingly used to generate optimized adversarial examples. These are then deployed in black-box scenarios, exploiting transfer-based attack techniques. This evolution in adversarial strategies reinforces the need for robust countermeasures capable of addressing both direct and indirect threats in real-world ADS environments. This field continues to prioritize innovation in defense mechanisms, emphasizing scalable, adaptive, and generalizable solutions that ensure secure, transparent, and high-performing image classification models for ADS.

6. Recommendations and Future Scope

Advancing the robustness of deep learning (DL) models in Autonomous Decision Systems (ADS) necessitates a strategic focus on emerging adversarial threats and defense methodologies. A pivotal recommendation is the development of standardized, unified frameworks for evaluating defense mechanisms. Such frameworks should incorporate modular designs and universal metrics, ensuring adaptability and comprehensive benchmarking across diverse adversarial scenarios. Addressing the trade-off between robustness and generalization remains critical. Hybrid approaches, such as combining adversarial training with preprocessing transformations, can enhance resilience to adversarial attacks without compromising accuracy on clean inputs.

The growing prominence of transferability-based attacks underscores the need for adaptable defenses capable of generalizing across models and datasets. Techniques like ensemble modeling, federated learning, and domain adaptation hold promise for mitigating such threats. Simultaneously, ensuring computational efficiency and scalability is vital for real-time ADS applications. Lightweight methodologies, including noise injection and feature squeezing, can provide practical solutions for large-scale deployment without overburdening system resources.

Explainability and interpretability in defense mechanisms are gaining traction, as transparency in decision-making fosters trust and reliability, especially in safety-critical domains. Integrating explainable AI (XAI) into adversarial defenses can bridge the gap between robustness and user trust. To counter the dynamic evolution of adversarial

strategies, adaptive defenses leveraging reinforcement learning and self-improving models must be prioritized. These systems can anticipate and neutralize emerging threats, ensuring resilience in unpredictable environments.

Additionally, distributed and cloud-based solutions, such as federated and incremental learning, offer robust, scalable frameworks by leveraging diverse datasets and fostering collaborative learning. Such approaches align with the complex and heterogeneous nature of ADS environments. Cross-disciplinary collaboration among domains like cybersecurity, computer vision, and human factors engineering can further enhance the robustness of adversarial defenses by addressing multifaceted challenges.

The future scope of this research includes the development of universal defense mechanisms resilient to various attack types, real-time implementation capabilities for autonomous vehicles, ethical frameworks for AI deployment, and broadening applications to fields like healthcare, financial systems, and IoT ecosystems. As adversarial tactics evolve, sustained innovation and interdisciplinary approaches will be pivotal in fortifying DL models and ensuring the reliability and safety of ADS in real-world scenarios.

7. Conclusion

In conclusion, addressing adversarial attacks on deep learning (DL) models in Autonomous Decision Systems (ADS) remains an ongoing, multifaceted challenge that necessitates continuous innovation in both defense mechanisms and model resilience. As adversarial threats become increasingly sophisticated, the demand for unified, adaptive, and scalable defense solutions that balance robustness and accuracy grows more urgent. Hybrid defense strategies combining proactive and reactive methods, lightweight defenses, and transferability-based solutions are key to enabling the practical deployment of secure ADS models in dynamic, real-time environments. Additionally, the integration of explainable AI (XAI) and fostering cross-disciplinary collaboration will be pivotal in enhancing transparency, trust, and safety, especially in critical applications. The future of AD technology is heavily reliant on its safety and the trust it garners on public roads. Therefore, establishing robust countermeasures to defend image classification DL models from adversarial attacks is essential for gaining confidence from policymakers, industry leaders, and the public. By leveraging the countermeasures outlined in this review, the safety, effectiveness, and adoption of ADS technology can be greatly enhanced. This will not only benefit the development of ADS but also mitigate the risks associated with adversarial vulnerabilities.

Moreover, the constantly evolving nature of adversarial threats calls for continuous improvement in the resilience of DL models used in ADS. As these threats evolve, so too must the countermeasures that protect against them. This review highlights the need for universal countermeasures to improve the adversarial tolerance and robustness of image classification models in autonomous driving. Researchers and experts must work collaboratively to stay ahead of adaptive attack techniques, ensuring the long-term security and adaptability of AD technology.

This systematic review has provided an in-depth analysis of adversarial attacks and robustness algorithms in image classification for ADS. It underscores the fact that while much research has focused on developing new attacks, existing countermeasures often lack the robustness required to effectively mitigate these threats. By identifying emerging trends and potential research directions, this review emphasizes the need for continued exploration into universal defenses and enhancing adversarial tolerance. The ultimate goal is to ensure that image classification DL models for ADS can withstand evolving adversarial challenges and maintain their reliability and safety in real-world applications.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Badjie, B. (2024). Adversarial Attacks and Countermeasures on Image Classification-based Deep Learning Models in Autonomous Driving Systems: A Systematic Review. *research Gate*.[1]

[2] Eleftheriadis, C. (2024). Adversarial robustness improvement for deep neural networks. *research gate*.[2]

[3] K.T.Y.Mahimal. (2021). Adversarial Attacks and Defense Technologies on Autonomous Vehicles: A Review. *sciendo*.[3]