Proceedings of the 7th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/2025.27221

Antagonistic Attack and Effective Defense of Autonomous Vehicle

Jianqiao Jiang

Faculty of Engineering, Architecture & Info Tech, School of Electrical Engineering & Computer Science, School of Electrical Engineering & Computer Science, The University of Queensland, Brisbane, Australia
15668060086@163.com

Abstract. This paper focuses on the adversarial attack of autonomous vehicle. Based on robust optimization theory, an integrated defense architecture is proposed that includes dynamic course adversarial training (DCAT) and Bayesian uncertainty detector (BUD). The defense strategies are divided into three categories: model enhancement, input purification, and runtime monitoring, and their implementation methods and applicable stages are clarified. The results show that the DCAT+BUD method has high robust accuracy in white box, black box and physical attack scenarios, and effectively improves the safety of the auto drive system.

Keywords: autonomous vehicle, adversarial attacks, defense architecture, adversarial training

1. Introduction

The data-driven nature of deep learning poses significant risks in safety critical systems, and the robustness issues caused by adversarial machine learning are particularly prominent [1]. This type of attack causes disruptive deviations in model output through small input disturbances, and the real-time requirements of the attack in autonomous driving scenarios are in sharp conflict with the high threshold of system robustness, becoming the core bottleneck of technology implementation.

From the perspective of information theory, adversarial samples are instances where norm constrained perturbations such as L ∞ and L2 are added to the original input space, causing a significant shift in the posterior probability distribution of the model; From the perspective of optimization theory analysis, its generation is the maximization of the model loss function under constraint conditions, while the autonomous driving threat model needs to cover scenarios such as "black box attacks" and "physical world attacks". Attackers do not need to know the model structure and can only manipulate sensor inputs to achieve their goals. In typical cases, camera sticker disturbance and LiDAR laser interference have led to missed detection and lane departure of vehicles and pedestrians, highlighting the urgency of constructing defense strategies and the complexity of addressing challenges such as "diversified paths of multi-sensor fusion attacks" and "coupling of physical disturbance and environmental noise". The existing single defense approach is difficult to meet the needs [3].

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2025.27221

From the root, as shown in Figure 1, the data-driven model overfits the training distribution, allowing attackers to expand their attack surface by tampering with camera pixels and interfering with LiDAR point clouds. However, although multi-sensor fusion improves redundancy, it increases attack entry points due to the increase in interfaces. This contradiction has not been fully resolved in existing research [4]. In the field of adversarial attack research, computer vision has formed a mature technical system. The Fast Gradient Symbol Method (FGSM) generates perturbations based on the loss function gradient symbol, which has high computational efficiency but weak aggression [5]; Projection Gradient Descent (PGD) significantly improves the success rate of attacks by iteratively optimizing and projecting into the constraint space through multiple steps. The core goal is to maximize the model classification loss under perturbation norm constraints [6]. However, attacks in autonomous driving scenarios have extended from the digital space to the physical world, such as traffic sign pasting patterns attacking cameras and laser beam tampering with LiDAR point clouds causing the system to misjudge obstacles. However, existing research lacks exploration of "disturbance stability in rainy and foggy weather" and "multi-sensor collaborative attack strategies", resulting in a gap between the coverage of attack scenarios and practical applications.

The current adversarial defense mechanisms for autonomous driving are divided into proactive defense and reactive defense. Proactive defense has limitations such as relying on the quality of attack samples and poor generalization ability to unknown attacks; Reactive defense focuses on anomaly detection, which is difficult to adapt to heterogeneous data fusion from multiple sensors, and lacks sensitivity in identifying gradual anomalies caused by small disturbances. The existing defense mechanisms are difficult to balance real-time and robustness when facing dynamic attacks in the physical world, which provides a core improvement direction for the innovation of defense schemes in this paper.

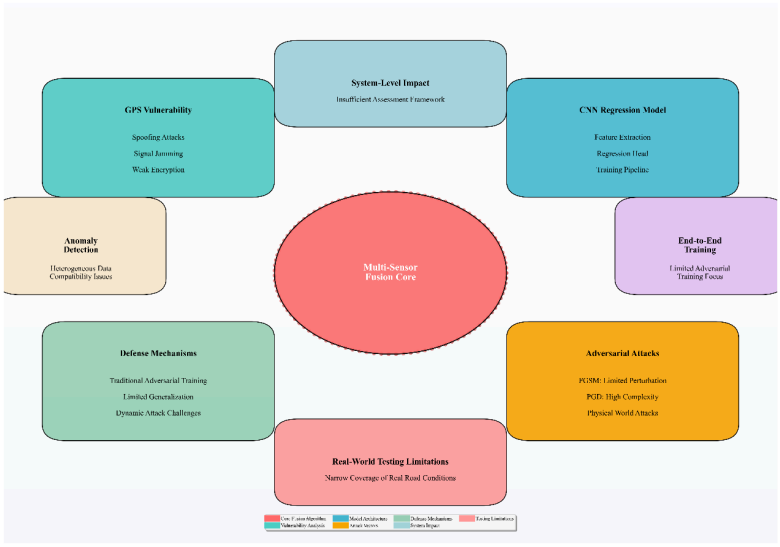


Figure 1. Schematic diagram of insufficient existing research

2. Deep analysis of adversarial attacks

2.1. Attack typology

White box attacks assume that attackers have complete control over the internal parameters and architecture of the target model, allowing them to accurately construct adversarial samples using mechanisms such as gradient backpropagation. Its essence is to solve the disturbance direction that

DOI: 10.54254/2755-2721/2025.27221

maximizes the model loss function through optimization algorithms, thereby generating misleading input data. Specifically, assuming the original input is x, the true label is y, the model parameters are θ, and the loss function is J (θ, x, y). The attack target can be formalized as a constrained optimization problem:

$$\min_{\delta} \left|\delta\right|_p \quad \text{s.t.} \quad f(x+\delta) \neq y \tag{1}$$

Where δ is the applied disturbance, ||·|| p represents the Lp norm constraint (commonly used L2 or L ∞ norm to ensure that the disturbance is imperceptible), and f (·) is the model prediction function.

The core difference between targeted attacks and non targeted attacks lies in optimizing objectives Non targeted attacks only require the model to output the wrong category, i.e. f (x+δ) ≠ y. For example, a single step attack based on FGSM generates perturbations through δ=ϵ· sign (▽ xJ (θ, x, y)), where ϵ is the perturbation amplitude. This type of attack is computationally efficient, but the success rate is limited by the linear assumption of the model.

Targeted attacks force the model to output a specific error category y_{target}, i.e. f (x+δ)=y_{target}. For example, PGD achieves precise misguidance by iteratively optimizing δt+1=Π||δ||p≤ϵ(δt+α·sign(▽x J(θ,x+δt,ytarget))), where π is the projection operator and α is the step size. This type of attack is more threatening, but the computational cost is higher.

Black box attacks rely solely on the input and output interfaces of the model, without the need for internal knowledge, and are more closely related to actual attack scenarios. The implementation mainly relies on two types of strategies, as shown in Figure 2.

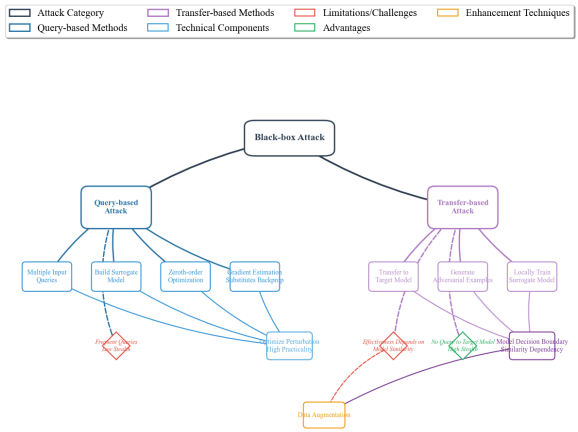


Figure 2. Black box attack strategy

2.2. Attack methodology

The generation of adversarial attacks can be uniformly expressed as a constrained optimization problem:

DOI: 10.54254/2755-2721/2025.27221

$$\max_{\delta} J(\theta, x+\delta, y) \quad s.t. \quad \left|\delta\right|_p \leq \epsilon \text{ and } x+\delta \in [0,1]^m \tag{2}$$

Where ϵ is the disturbance budget, ensuring that the disturbance is imperceptible in perception. The L ∞ norm limits the maximum perturbation amplitude, i.e. ||δ||∞≤ϵ, and is applicable to pixel level perturbations in images. The L2 norm limits the overall disturbance energy, i.e. ||δ||2≤ϵ, which is more in line with human visual sensitivity.

The iterative update formula of PGD algorithm is:

$$\delta_{t+1} = \text{Clip}_{\epsilon}\left(\delta_t + \alpha \cdot \text{sign}\left(\nabla_x J(\theta, x+\delta_t, y)\right)\right) \tag{3}$$

Clip ϵ projects the perturbation onto the L ∞ sphere, where α is the step size. This algorithm ensures the finding of the local optimal disturbance through multiple gradient rises, and guarantees the small size of the disturbance through constraints.

3. Effective defense strategy design

3.1. Design principles

Based on robust optimization theory, an integrated defense architecture is proposed to maintain system performance stability in the worst-case scenario. Its mathematical essence can be expressed as the problem of minimizing maximum loss:

$$\min_{\theta} E_{(x,y)\sim\mathbb{D}}\left[\max_{\delta\in\Delta} J(\theta, x+\delta, y)\right] \tag{4}$$

Where D is the data distribution and Δ is the set of perturbation constraints.

3.2. Strategy classification

The strategies are divided into three categories as shown in Table 1.

Table 1. Strategy classification

Strategy Category	Strategy Subcategory	Core Implementation Method	Applicable Phase
Model Enhancement	Adversarial Training	Inject adversarial examples and optimize robust loss functions	Training Phase
	Regularization Techniques	Gradient penalty / Lipschitz constraint	Training Phase
Input Sanitization	Preprocessing Techniques	Denoising / Random Resizing / JPEG compression	Input Phase
	Generative Sanitization	Reconstruction via autoencoders / diffusion models	Input Phase
Runtime Monitoring	Consensus-based Detection	Multi-model voting / multi-sensor fusion (e.g., camera-LiDAR consistency check)	Inference Phase
	Anomaly Metric Monitoring	Confidence distribution / uncertainty measurement / outlier detection in feature space	Inference Phase

DOI: 10.54254/2755-2721/2025.27221

3.3. Specific defense techniques

Traditional adversarial training using a fixed perturbation budget ϵ may result in an imbalance between robustness and accuracy. This article proposes Dynamic Course Adversarial Training (DCAT), which first schedules the courses and increases the disturbance intensity ϵ t with training epochs. In the initial stage, small perturbations (ϵ_0=0.01) are used to learn simple samples, gradually expanding to the maximum budget (ϵ_{max}=0.03). Next, in each round of training, the attack type is selected based on the current model state. In the early rounds, fast single step attacks are used to improve efficiency, while in the later rounds, multi-step attacks are switched to enhance robustness.

Design a loss function that combines robust loss and clean sample loss:

$$\mathscr{L}_{\mathrm{DA}} = \alpha \cdot \mathrm{E}\left[\mathrm{J}\left(\theta, \mathrm{x} + \delta^{*}, \mathrm{y}\right)\right] + (1 - \alpha) \cdot \mathrm{E}\left[\mathrm{J}\left(\theta, \mathrm{x}, \mathrm{y}\right)\right] \tag{5}$$

Among them, δ*is the optimal perturbation under the current course, and α is the dynamic weight that increases linearly with training epochs.

Adversarial samples often lead to model uncertainty anomalies. This article designs a Bayesian uncertainty detector (BUD) that uses Monte Carlo dropout to perform T forward inferences, obtaining a probability distribution {pt (y | x)} $_t^T$=1, and calculating the prediction uncertainty. The expression for entropy uncertainty is as follows:

$$\mathrm{H}\left(\mathrm{yx}\right) = -\sum_{\mathrm{c}} \tilde{\mathrm{p}}\left(\mathrm{y} = \mathrm{cx}\right) \log \tilde{\mathrm{p}}\left(\mathrm{y} = \mathrm{cx}\right) \tag{6}$$

4. Experimental result

4.1. Experimental environment

This study is based on the Ubuntu 20.04 operating system and PyTorch 1.9.0 deep learning framework to build an experimental platform. The hardware configuration uses NVIDIA RTX 3090 GPU and Intel Xeon E5-2680 v4 processor. The autonomous driving simulation environment uses CARLA version 0.9.13, which provides high-precision sensor simulation (64 line LiDAR and 120 ° field of view camera) and dynamic traffic scenes. The dataset test set includes KITTI object detection data and ApolloScape street view samples, where adversarial samples are generated using PGD (ε=0.03, α=0.001, 40 iterations) and FGSM (ε=0.05) algorithms. The evaluation process adopts five fold cross validation, with all random seeds fixed at 42 to ensure reproducibility of the results.

4.2. Experimental results

As shown in Figure 3, different defense methods perform differently in dealing with various types of attacks. The DCAT+BUD method in this article performs the best, with high robustness and accuracy regardless of white box, black box, or physical attacks; The clean accuracy rate is maintained at 89.5%, the false alarm rate is low, and the recovery rate is high, which shows that this method can better balance the clean data performance while enhancing the robustness of the model, and has outstanding effects on attack detection, recovery, and collision risk reduction, effectively improving the security of the auto drive system system against various attacks.

DOI: 10.54254/2755-2721/2025.27221

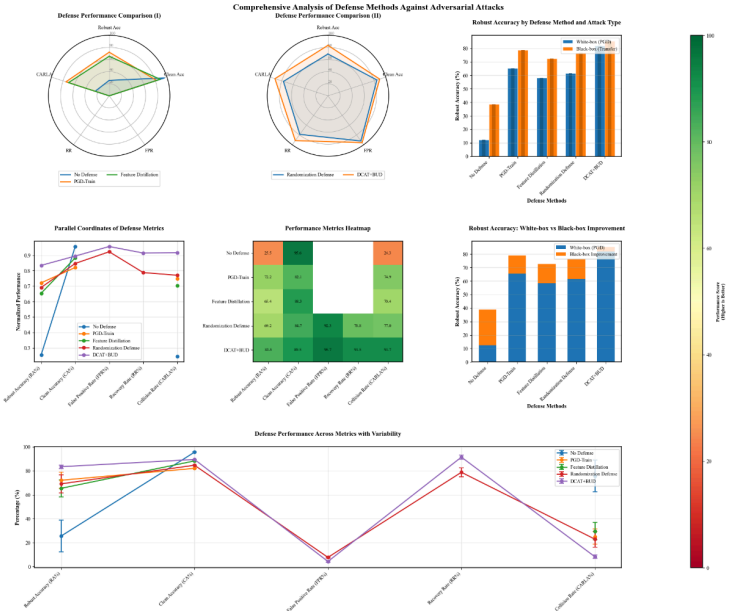


Figure 3. Experimental results

5. Conclusion

This study analyzes the characteristics and existing shortcomings of adversarial attacks in autonomous driving, and proposes an integrated defense architecture containing DCAT and BUD based on robust optimization theory. The experimental verification shows that this scheme has high robustness in multiple attack scenarios, and the accuracy of clean data reaches 89.5%, providing support for adversarial safety breakthroughs in autonomous driving.

References

[1] Wang, X., Li, J., Kuang, X., Tan, Y. A., & Li, J. (2019). The security of machine learning in an adversarial setting: A survey. Journal of Parallel and Distributed Computing, 130, 12-23.

[2] Chen, W., Liu, Y., Bakker, E. M., & Lew, M. S. (2021). Integrating information theory and adversarial learning for cross-modal retrieval. Pattern Recognition, 117, 107983.

[3] Hoque, M. A., & Hasan, R. (2022, October). Autonomous driving security: A comprehensive threat model of attacks and mitigation strategies. In 2022 IEEE 8th World Forum on Internet of Things (WF-IoT) (pp. 1-6). IEEE.

[4] Amini, A., Gilitschenski, I., Phillips, J., Moseyko, J., Banerjee, R., Karaman, S., & Rus, D. (2020). Learning robust control policies for end-to-end autonomous driving from data-driven simulation. IEEE Robotics and Automation Letters, 5(2), 1143-50.

[5] Shibly, K. H., Hossain, M. D., Inoue, H., Taenaka, Y., & Kadobayashi, Y. (2023). Towards autonomous driving model resistant to adversarial attack. Applied Artificial Intelligence, 37(1), 2193461.

[6] Hu, C., & Wang, H. (2023). Enhancing Rainy Weather Driving: Deep Unfolding Network with PGD Algorithm for Single Image Deraining. IEEE Access, 11, 57616-57626.