Attack End-to-End Autonomous Driving through Module-Wise Noise

Lu Wang[1,3], Tianyuan Zhang[1,2,3], Yikai Han[1], Muyang Fang[1], Ting Jin[1], Jiaqi Kang[4]
[1] School of Computer Science and Engineering, Beihang University, Beijing, China
[2] Shen Yuan Honors College, Beihang University, Beijing, China
[3] State Key Lab of Software Development Environment, Beihang University, Beijing, China
[4] School of Software, Beihang University, Beijing, China
{20373361, zhangtianyuan, 21371441, 21373061, 21371466, 21373331}@buaa.edu.cn

Abstract

With recent breakthroughs in deep neural networks, numerous tasks within autonomous driving have exhibited remarkable performance. However, deep learning models are susceptible to adversarial attacks, presenting significant security risks to autonomous driving systems. Presently, end-to-end architectures have emerged as the predominant solution for autonomous driving, owing to their collaborative nature across different tasks. Yet, the implications of adversarial attacks on such models remain relatively unexplored. In this paper, we conduct comprehensive adversarial security research on the modular end-to-end autonomous driving model for the first time. We thoroughly consider the potential vulnerabilities in the model inference process and design a universal attack scheme through module-wise noise injection. We conduct large-scale experiments on the full-stack autonomous driving model and demonstrate that our attack method outperforms previous attack methods. We trust that our research will offer fresh insights into ensuring the safety and reliability of autonomous driving systems.

1. Introduction

With recent significant advancements in deep learning, autonomous driving technology plays an increasingly important role in today's society. End-to-end autonomous driving models map raw sensor data directly to driving decisions, becoming the predominant solution gradually. However, despite the excellent performance of end-to-end models, we must also realize the security challenges they face. A considerable amount of research has already been devoted to studying adversarial attack methods towards individual tasks within the field of autonomous driving, with particular emphasis on the perception layer [1, 2], as illustrated in Figure 1 (a). Nowadays, researchers have begun to conduct adversarial attacks on very simple end-to-end regression-based decision models [3] directly from input images, as illustrated in Figure 1 (b).

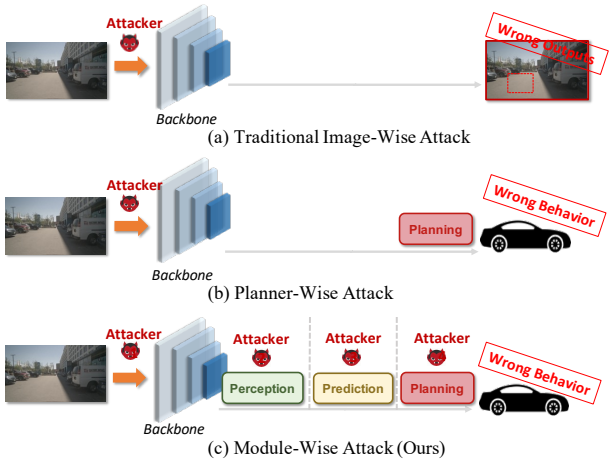

Figure 1. Adversarial attacks in autonomous driving. There are a considerable number of mature attack algorithms targeting the perception of autonomous driving (**a**). There is a limited amount of research focusing on adversarial security for end-to-end regression-based decision models (**b**). We propose the module-wise attack targeting end-to-end autonomous driving models (**c**).

However, there is currently no adversarial security research conducted on complex end-to-end autonomous driving models that are composed of multiple sub-tasks. In this paper, we delve into the robustness of modular end-to-end autonomous driving models. We believe that attacking complex models should not only focus on the image level but also consider the vulnerability of the interaction process between modules, as illustrated in Figure 1 (c). Therefore, we introduce adversarial noise at the interfaces between modules of the end-to-end models. The contributions of this paper are summarized as follows:

- We design the module-wise attack toward end-to-end autonomous driving models.
- We conduct extensive experiments on the full-stack autonomous driving model , which reveals the insecurity of the model.

2. Related Work

2.1. End-to-End Autonomous Driving

The early end-to-end autonomous driving models combine relatively few tasks. [4] adopts a bounding box-based intermediate representation to construct the motion planner. Considering that Non-Maximum Suppression(NMS) in this method can lead to loss of perceptual information, P3 [5] innovatively designs an end-to-end network that utilizes maps, advanced control instructions, and LIDAR points to generate interpretable intermediate semantic occupancy representations, which facilitates safer trajectory planning. Following P3, MP3 [6] and ST-P3 [7] achieve new improvements. UniAD [8] is the first end-to-end network that integrates full-stack autonomous driving tasks. By thoroughly considering the contributions of each task to autonomous driving and mutual promotion among modules, UniAD significantly surpasses previous sota performance on each task.

2.2. Adversarial Attack

Adversarial noise refers to carefully crafted perturbations designed for neural network input data, which are typically small but can cause models to produce complete error outputs [9–17]. [9] first introduces the concept of adversarial attack and utilizes L-BFGS approximation. Following that, a series of adversarial attack methods are proposed, such as gradient-based methods [18, 19], and optimization-based methods [20]. Although early adversarial attacks primarily target image classification models, they demonstrate the vulnerability of neural networks. Their attack principles have guided the implementation of various attack methods in different tasks, posing a serious threat to the practical application of models in the real world[10, 11, 14].

3. Methodology

3.1. Problem Definition

Let $y = F(x; \theta)$ represent the end-to-end autonomous driving model with parameters θ, which includes n sub modules $\{M_i\}_{i=1}^{n}$. The inference stage can be formulated as:

$$Q_i = M_i(Q_{i-1}; \theta_i), i = 1, 2, ..., n, \quad (1)$$

where $Q_0 = x$ refers to the input images, $Q_n = y$ the planned results, and Q_i the output intermediate queries of the i-th sub module.

In this research, we consider the interaction information among modules as a potential vulnerability. In addition to the input images, we inject adversarial noise N_{i-1} into the input data Q_{i-1} of each submodule M_i. That is,

$$Q_{i-1}^{adv} = Q_{i-1} + N_{i-1}, i = 1, 2, ..., n. \quad (2)$$

Let N represents the final injected adversarial noise $\{N_i\}_{i=0}^{n-1}$, and $y^{adv} = F(x; \theta, N)$ denotes the decision results made by the end-to-end model after inference through each module under noise attack. Our attack objective is to find, for each data set x, an adversarial noise N that satisfies the constraint on the perturbation magnitude ξ, while ensuring $y^{adv} \neq y$.

3.2. Module-Wise Adversarial Noise

Figure 2 presents an overall framework of the proposed method. A complete autonomous driving model comprises perception, prediction and the final planner. We aim to inject adversarial noise into the input of all sub-modules in the end-to-end autonomous driving model so mainly focus on the current full-stack autonomous driving model UniAD [8]. Undoubtedly, our noise injection scheme is applicable to any modular end-to-end autonomous driving model.

Images serve as the initial input for the model to perceive the environment, forming the foundation for all subsequent modules. Therefore, we design pixel-level noise specifically for input images, denoted as N_I. The track module accomplishes multi-object tracking, producing spatiotemporal information Q_A. Therefore, we design noise N_A for the spatiotemporal features of agents. The mapping module models the features of road elements as Q_M and we design adversarial noise N_M for all map features. Q_A and Q_M respectively provide dynamic agent features and static scene information for downstream modules. The motion module, based on the aforementioned dynamic and static features, predicts the most likely future trajectory states of all agents, represented as Q_T. Similarly, we design noise N_T for the state of agents. In addition, the motion module also expresses the future intentions of the ego vehicle, represented as Q_E. We separately design noise N_E for Q_E. Since the occupancy map is ultimately used only in post-processing for collision optimization of the planned trajectory, and this optimization process is non-differentiable, we don't consider injecting noise into the occupancy map.

3.3. Attack Strategy

We achieve the attack by iteratively optimizing module-wise noise. To be more specific, during the model's inference stage for each batch of data, we first initialize corresponding adversarial noise randomly within the perturbation constraints for each module, following the forward propagation process of the model. At each iteration, noise injected into the corresponding module is propagated and stored until the final planning stage. We synchronize the update of all noise by the adversarial loss, using it to initialize the noise for the next iteration.

Adversarial loss, essential for noise update, includes attack loss and noise loss. Attack loss seeks to maximize deviation between sub-module predictions and actual results, mathematically represented as:

$$L_{att} = L_{track} + L_{map} + L_{motion} + L_{occ} + L_{plan}. \quad (3)$$

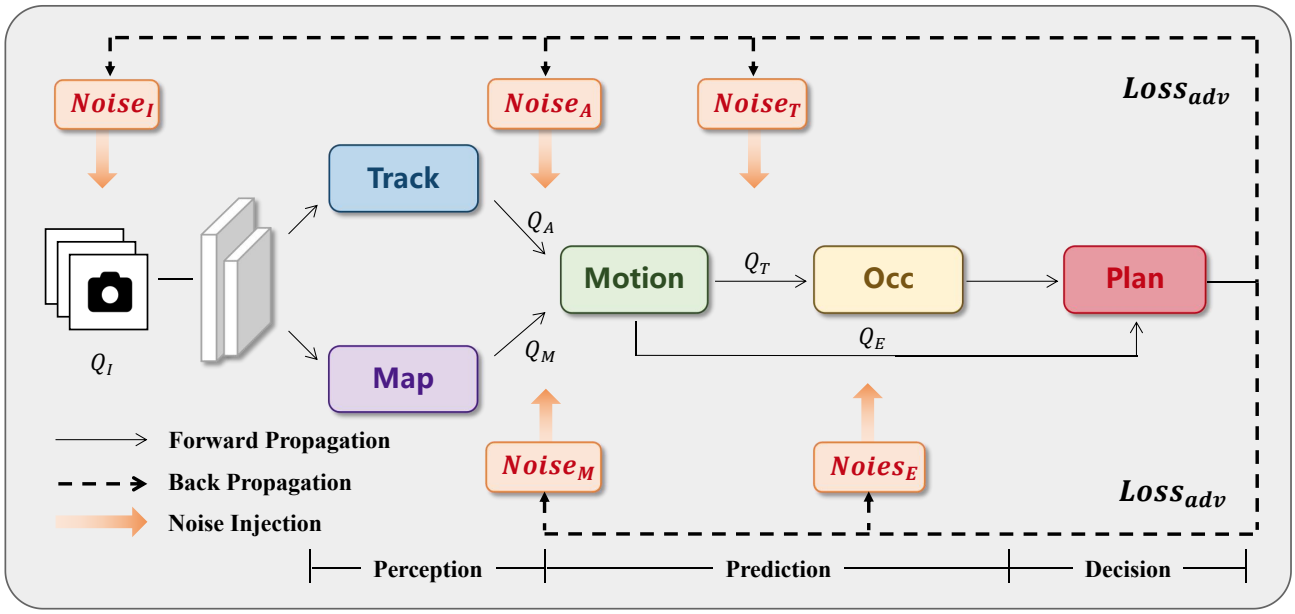

Figure 2. The framework of our module-wise noise attack method. We meticulously inject adversarial noise into the interaction process of all modules in the end-to-end autonomous driving model and synchronize the optimization of all noise using the losses from all modules.

The noise loss is used to constrain the injected adversarial loss to be as minimal as possible. We choose the L2 norm distance to measure the magnitude of module-wise noise. It is represented as:

$$L_{noi} = \sum_i \sigma_i \cdot ||N_i||_2, i = I, A, M, T, E, \tag{4}$$

where σ_i are hyper-parameters. The final adversarial loss is represented as:

$$L_{adv} = L_{att} - L_{noi}. \tag{5}$$

We optimize the noise in a manner similar to PGD, performing gradient ascent on the adversarial noise, namely:

$$N_i^{t+1} = \prod_{\epsilon}(N_i^t + \frac{\epsilon}{\sqrt{k}} \cdot sgn(\nabla_{N_i} L_{adv})), \tag{6}$$

where k denotes the number of iterations and ϵ is a hyper-parameter that constrains the magnitude of the perturbation.

4. Experiments

4.1. Experimental Setup

Dataset. Our robustness evaluation experiments are conducted on the validation split of the large-scale autonomous driving dataset nuScenes [21].

Model. For the target model, we choose the current full-stack autonomous driving model UniAD, which includes complete sub-tasks of perception, prediction, and decision. We conduct noise injection according to the five stages of the model as outlined in Section 3.2. For the specific implementation of the attack, we set the number of iterations $k = 10$, $\sigma_I = 8\times10^{-6}$, $\sigma_A = \sigma_M = \sigma_T = \sigma_E = 2\times10^{-4}$.

Metrics. For the metrics, we choose the evaluation methods for five sub tasks adopted in [8].

4.2. Robust Evaluation

Baselines. As there are currently no adversarial attacks targeting complex end-to-end autonomous driving models composed of a series of sub-modules, we choose attack methods targeting end-to-end regression-based decision models as baselines [3]. In our comparative experiments, we set the maximum allowable perturbation ϵ for images to 8 and implement *Image-specific Attack* and *Image-agnostic Attack* tailored for UniAD.

Results. We provide the robustness evaluation results compared with baselines for the five tasks, as shown in Table 1. Our attack can significantly reduce the performance of all tasks. We measure the main planning performance of the vehicle using the L2 error between the planned trajectory and the ground truth trajectory, as well as the collision rate with obstacles in the ego vehicle's driving environment. Results indicate that all three attacks lead to errors in the vehicle's planning, but the proposed attack method poses the greatest threat to the model as it interferes with both the perception and prediction processes during model inference, and the final prediction heavily relies on the inference results of upstream modules, resulting in severe planning errors due to error accumulation. Overall, the attack intensities of the three methods exhibit the same trend across the five tasks, with the *Image-specific Attack* being much stronger than the *Image-agnostic Attack*, and our module-wise attack surpassing the *Image-specific Attack*.

Table 1. Comparison with other attack methods on five modules. The first row represents the original replication results. The map module is evaluated by IOU of road elements. The motion module is evaluated by three types of error on the vehicle. The plan module is evaluated by the average of L2 error and collision rate in the next three seconds. Our attack method (the last row) achieves the maximum performance degradation across all modules.

Method	Track				Map			Motion			Occupancy		Plan	
	Amota ↑	Amotp ↓	Recall ↑	IDS ↓	Drivable ↑	Lanes ↑	Crossing ↑	minADE ↓	minFDE ↓	MR ↓	Iou-n ↑	Iou-f ↑	L2 error(m) ↓	Col.Rate ↓
Original	0.367	1.26	0.449	442	68.80%	31.17%	14.29%	0.733	1.079	0.164	64.00%	41.00%	1.09	0.32%
Image-specific Attack	0.008	1.947	**0.051**	325	45.25%	17.34%	4.42%	1.424	2.110	0.265	24.90%	12.60%	2.90	2.27%
Image-agnostic Attack	0.231	1.901	0.095	224	49.67%	19.96%	6.22%	1.119	1.650	0.226	32.90%	17.20%	1.31	0.45%
Module-wise Attack	**0.000**	**1.976**	**0.051**	**2706**	**39.63%**	**15.27%**	**2.89%**	**2.985**	**4.914**	**0.409**	**17.60%**	**8.00%**	**5.37**	**4.33%**

5. Conclusions

We delve into the robustness of complex end-to-end autonomous driving models with multi-modules by introducing a novel adversarial attack using module-wise noise injection. We strongly believe that it is imperative to consider the vulnerabilities in the interaction process between modules to enhance the security of autonomous driving systems. By conducting extensive experiments on the full-stack autonomous driving model, we demonstrate the profound impact of injecting noise into different modules on the planning performance of the model as well as other tasks.

6. Acknowledgment

This work is supported by grant No. KZ46009501

References

[1] M. Abdelfattah, K. Yuan, Z. J. Wang, and R. Ward, "Towards universal physical attacks on cascaded camera-lidar 3d object detection models," in *ICIP*, 2021. 1

[2] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *IEEE SP*, 2021. 1

[3] H. Wu, S. Yunas, S. Rowlands, W. Ruan, and J. Wahlström, "Adversarial driving: Attacking end-to-end autonomous driving," in *IV*, 2023. 1, 3

[4] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun, "End-to-end interpretable neural motion planner," in *CVPR*, 2019. 2

[5] A. Sadat, S. Casas, M. Ren, X. Wu, P. Dhawan, and R. Urtasun, "Perceive, predict, and plan: Safe motion planning through interpretable semantic representations," in *ECCV*, 2020. 2

[6] S. Casas, A. Sadat, and R. Urtasun, "Mp3: A unified model to map, perceive, predict and plan," in *CVPR*, 2021. 2

[7] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, "St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning," in *ECCV*, 2022. 2

[8] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang, *et al.*, "Planning-oriented autonomous driving," in *CVPR*, 2023. 2, 3

[9] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013. 2

[10] A. Liu, X. Liu, J. Fan, Y. Ma, A. Zhang, H. Xie, and D. Tao, "Perceptual-sensitive gan for generating adversarial patches," in *AAAI*, 2019. 2

[11] A. Liu, J. Wang, X. Liu, B. Cao, C. Zhang, and H. Yu, "Bias-based universal adversarial patch attack for automatic check-out," in *ECCV*, 2020. 2

[12] A. Liu, J. Guo, J. Wang, S. Liang, R. Tao, W. Zhou, C. Liu, X. Liu, and D. Tao, "X-adv: Physical adversarial object attacks against x-ray prohibited item detection," in *USENIX Security Symposium*, 2023.

[13] A. Liu, T. Huang, X. Liu, Y. Xu, Y. Ma, X. Chen, S. J. Maybank, and D. Tao, "Spatiotemporal attacks for embodied agents," in *ECCV*, 2020.

[14] J. Wang, A. Liu, Z. Yin, S. Liu, S. Tang, and X. Liu, "Dual attention suppression attack: Generate adversarial camouflage in physical world," in *CVPR*, 2021. 2

[15] S. Liu, J. Wang, A. Liu, Y. Li, Y. Gao, X. Liu, and D. Tao, "Harnessing perceptual adversarial patches for crowd counting," in *ACM CCS*, 2022.

[16] A. Liu, S. Tang, S. Liang, R. Gong, B. Wu, X. Liu, and D. Tao, "Exploring the relationship between architecture and adversarially robust generalization," in *CVPR*, 2023.

[17] S. Tang, R. Gong, Y. Wang, A. Liu, J. Wang, X. Chen, F. Yu, X. Liu, D. Song, A. Yuille, *et al.*, "Robustart: Benchmarking robustness on architecture design and training techniques," *ArXiv*, 2021. 2

[18] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *CVPR*, 2018. 2

[19] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017. 2

[20] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE SP*, 2017. 2

[21] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *CVPR*, 2020. 3