2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Adversarial Attack on Trajectory Prediction for Autonomous Vehicles with Generative Adversarial Networks

Jiping Fan, Zhenpo Wang, Guoqiang Li

***Abstract*— Accurate trajectory prediction is crucial for autonomous vehicles to realize safe driving. Current trajectory prediction approaches generally rely on deep neural networks, which are susceptible to adversarial attacks. To evaluate the adversarial robustness and security of deep-learning-based trajectory prediction models, this paper proposes an adversarial attack method on trajectory prediction using generative adversarial networks (GANs). First, a novel LSTM-based attack trajectory model named Adv-GAN is proposed considering both the temporal and spatial driving features. The networks in Adv-GAN are trained through game learning between the generator and the discriminator to obtain the adversarial trajectories with real driving feature distribution. Furthermore, the generated trajectory is optimized with the vehicle kinematics model for driving feasibility on roads. The derived adversarial attack can lead to considerable deviations in trajectory prediction which affects driving safety for autonomous vehicles. We evaluate the proposed Adv-GAN on three public datasets, and experimental results show the effectiveness with better attack performance compared to a state-of-the-art adversarial attack model.**

I. Introduction

Autonomous vehicles (AVs) have emerged as a crucial solution for enhancing traffic safety and efficiency [1]. The accuracy of trajectory prediction is vital to ensure the safety of autonomous vehicles by enabling them to anticipate the future actions of other nearby traffic participants. Trajectory prediction is fundamental to autonomous vehicle trajectory planning and has a profound impact on the vehicle's driving [2]. Consequently, accurate trajectory prediction is essential for ensuring the safe operation of autonomous vehicles.

While recent trajectory prediction models based on deep neural networks (DNNs) have shown outstanding performance on large-scale benchmarks [3], DNNs are vulnerable to malicious attacks and adversarial attacks [4]. Adversarial attack is a type of attack aimed at deceiving or disrupting the output of machine learning model [5]. These attacks exploit vulnerabilities in the models by making imperceptible modifications to the input data, resulting in incorrect predictions or classification outcomes. Adversarial attacks are often effective, difficult to defend against, and pose a threat in the real world [6]. Researching the issue of adversarial robustness brought about by adversarial attacks is an important research direction for trajectory prediction model. Therefore, effective attack methods should be developed to comprehensively study and enhance the security of trajectory prediction models.

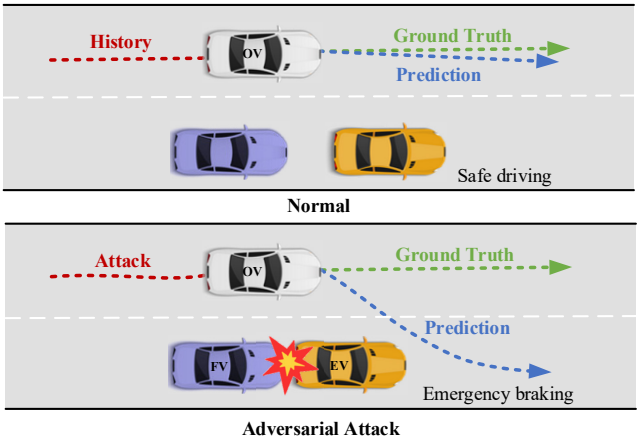

Fig. 1. Driving scenario: (a) normal driving; (b) adversarial attack: the automated ego vehicle (EV) makes inappropriate maneuver based on the wrong trajectory prediction for the adversarial vehicle (OV) under attack and causes accidents with following vehicle (FV).

This paper investigates the adversarial attack scenario wherein an adversary manipulates a vehicle, referred to as "the other vehicle" (OV), along a deliberately crafted attack trajectory. The ego vehicle (EV) observes the OV and employs iterative trajectory prediction to forecast its driving behavior. The adversary's goal is to manipulate the OV's trajectory to either maximize the prediction errors or induce unsafe driving maneuvers. Fig. 1 depicts a typical example of such an adversarial attack: the OV appears to change lanes in the EV's prediction, while in reality, it maintains a straight path. As a result of significant prediction errors, the EV engages emergency braking to avoid the OV, creating a substantial safety risk that could lead to rear-end collisions with following vehicles (FV). This is a serious safety hazard for AVs that lack adversarial robustness consideration. The current research predominantly utilizes optimization methods to generate attacks, yet falls short in achieving significant attack effects, failing to cause sufficiently large prediction errors [7][8]. Generative adversarial networks (GANs) can learn and approximate the distribution of original instances [9][10], leveraging them to generate effective adversarial attacks has emerged as a feasible avenue for development.

Currently, there is a paucity of research addressing the adversarial robustness and security of trajectory prediction in the face of attacks. In this paper, we propose an attack trajectory generation approach termed Adv-GAN to evaluate the security of the prediction model. The Adv-GAN is built upon the GAN architecture but tailored to the characteristics of vehicle driving. Adv-GAN leverages Long Short-Term

*This work is supported in part by the National Nature Science Foundation of China under Grant 52202459, and in part by the Beijing Municipal Natural Science Foundation under Grant 3232015. (Corresponding author: Guoqiang Li.).

The authors are within School of Mechanical Engineering, Beijing Institute of Technology, Beijing 100081, China. (email: jipingfan@bit.edu.cn; wangzhenpo@bit.edu.cn; guoqiangli@bit.edu.cn).

Memory (LSTM) networks in both its generator and discriminator to effectively capture the temporal properties of vehicle trajectories, while also considering their spatial features. Leveraging existing machine-learning-based trajectory prediction models, we conduct feature extraction on the input and output data of the trajectory prediction model to identify its vulnerabilities. Subsequently, we implement targeted adversarial attack against the trajectory prediction model using game learning. Furthermore, we employ model predictive control (MPC) with vehicle kinematic model to optimize the adversarial trajectories. The adversarial trajectories are feasible for public road driving and cause significant deviations in trajectory prediction, thereby impacting the driving safety of autonomous vehicles.

In summary, the main contributions of this paper are as follows: (1) A novel GAN-based adversarial attack generation method for vehicle trajectory prediction is proposed, considering both temporal and spatial driving features. (2) MPC is applied to optimize the adversarial trajectories, ensuring they are kinematically feasible for driving. (3) The proposed attack method is evaluated using three public datasets, demonstrating its effectiveness in trajectory prediction. It results in more than 50% increase on average prediction error and indicates better attack performance compared to a state-of-the-art adversarial attack model.

II. Related Works

In recent years, trajectory prediction methods based on machine learning have garnered significant attention, particularly those utilizing deep learning techniques [11]. For instance, reference [12] integrates the random forest algorithm with LSTM; [13] fuses machine learning and physical models to forecast trajectories; [14] merges Multi-head Attention with LSTM; while [15] and [16] amalgamate convolutional blocks with LSTM. Despite these advancements, machine learning models remain vulnerable to substantial security threats, including malicious attacks and adversarial manipulations [17][18], which present significant safety challenges for autonomous driving.

Adversarial robustness refers to the ability of model to resist malicious attacks and adversarial samples [19]. In real-world driving, autonomous vehicles must handle unforeseen circumstances, where even minor errors can lead to serious accidents. Therefore, the robustness and security of models and algorithms in autonomous vehicles are critical. Various methods exist to test the security of different components within the autonomous vehicle system. For example, adversarial trajectories can deceive lidar perception [20], impacting object detection [21] and lane detection [22] in the perception module. However, scholarly investigation into this method of attack remains scarce. This approach involves driving along intricately designed adversarial trajectories in real-world settings to influence vehicle trajectory predictions.

GANs can generate high-quality data samples that closely resemble the original data distribution through unsupervised learning [23][24]. GANs can be used to generate a wide array of adversarial attack samples, achieving high attack performance. Vehicle trajectories not only represent time series, but also demonstrate spatiotemporal dependence. LSTM is capable of effectively capturing such spatiotemporal dependencies [25]. Therefore, utilizing LSTM can model the complex relationships between data, it suitable for vehicle trajectory generation.

III. Problem Formulation

A. Adversarial Attack Formulation

This paper focuses on adversarial attacks targeting deep-learning-based trajectory prediction models, which forecast future trajectory time frames based on a fixed length of historical trajectory data. The attack involves manipulating a vehicle along a meticulously crafted attack trajectory. Assuming the current time is t, with the historical trajectory length of L_0 and the predicted future trajectory length of L_1, where the current state of the prediction target is represented as s_i at time i. As shown in Fig. 2, the historical trajectory at time t is denoted as $H_t = \{s_i | i \in [t - L_0 + 1, t]\}$, the ground truth trajectory is $F_t = \{s_i | i \in [t+1, t+L_1]\}$; the predicted state at time t is represented as p_i, then the predicted trajectory is $P_t = \{p_i | i \in [t+1, t+L_1]\}$; the adversarial attack state at time t is represented as a_i, then the attack trajectory is $A_t = \{a_i | i \in [t - L_0 + 1, t]\}$. The sampling frequencies vary across different databases.

In this study, it is assumed that the trajectory prediction model is known, and the adversarial attack model is specifically trained based on this prediction model. While, in real-world scenarios, trajectory prediction models are typically considered black boxes for attackers, it is possible to train a substitute neural network using input-output pairs to approximate the target model and achieve the objectives of adversarial attacks. Moreover, the attack trajectories must be feasible in real-world conditions. To address this, we use a well-trained generator from GAN, along with MPC optimization and vehicle kinematics models, to ensure that the adversarial trajectories are consistent with normal driving behaviors and adhere to physical constraints.

B. Evaluation Metrics

We employ two commonly used metrics from related literature to evaluate prediction errors: (1) Average Displacement Error (ADE), which is the root mean square error (RMSE) between the predicted trajectory and the ground truth trajectory averaged over the prediction time horizon; (2) Final Displacement Error (FDE), which is the RMSE between the predicted position at the final prediction time frame and

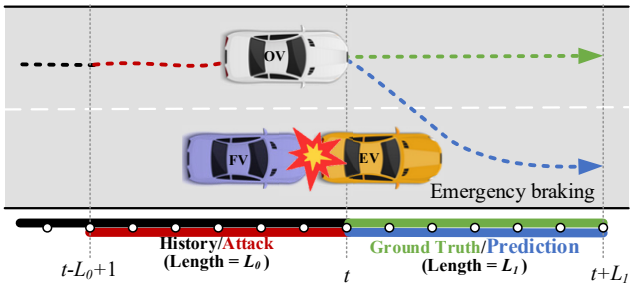

Fig. 2. Illustration of adversarial attack formulation.

the ground truth position. A significant increase in prediction error after implementing the attack indicates its effectiveness If the prediction error significantly increases after the attack implementation, it indicates the effectiveness of the attack.

IV. Methodology

We construct a GAN to produce adversarial trajectories and employ optimization methods to refine these trajectories, ensuring their feasibility in real-world scenarios. The objective of our model, Adv-GAN, is to generate adversarial trajectories that are similar to vehicle's normal driving features and possess driving hazards with considerable prediction deviation.

A. Overall Architecture of Adv-GAN

The overall architecture of Adv-GAN is illustrated in Fig. 3. As depicted, this GAN comprises three components: generator, discriminator, and pretrained trajectory prediction model. The generator and discriminator are trained in a min-max game. We use three stacked LSTM layers in the generator model with 32, 64, and 128 hidden units respectively, and one stacked LSTM layer in the discriminator model with 100 hidden units.

The generator in our model learns the features of normal driving vehicle trajectories from real trajectory data and produces adversarial attack trajectories that closely resemble these real trajectories. Utilizing an LSTM network in the generator enables it to capture and emulate the statistical characteristics and spatiotemporal dependencies of real trajectory data. This allows the generator to create synthetic data that deceives the discriminator, thus enhancing the generator's performance.

The discriminator is trained to distinguish between adversarial attack trajectories and normal driving trajectories. By incorporating an LSTM network, the discriminator can observe spatiotemporal patterns in the trajectory data, enabling it to detect anomalies or deviations from real data. This helps identify potential abnormal or fraudulent behaviors.

The pretrained trajectory prediction model uses the generated attack trajectories to make predictions and compares them with real future trajectories. This comparison helps in training the model to accurately discriminate between attack and normal trajectories.

Adv-GAN is a deep learning model based on game theory. Its mathematical principles can be explained by minimizing a competitive loss function between the two networks. By utilizing LSTM networks in both the generator and discriminator, Adv-GAN effectively generates adversarial trajectories that mimic normal driving behavior while introducing significant prediction deviations.

B. Generator

The generator network aims to learn the mapping from the latent space to the data space in order to generate fake data resembling real data. The inputs to the generator are noise vectors randomly sampled from the latent space. The generator generates attack perturbations, and combine with the original trajectories, serve as the final output of the generator. The loss function of the generator aims to minimize the probability of generated data being classified as fake, maximizing the probability of the discriminator classifying fake data as real, while simultaneously minimizing the discrepancy between real trajectories and attack trajectories. In this model, adversarial trajectories also need to maximize prediction errors as much as possible. Therefore, the loss function of the generator in this model differs from traditional GANs and includes two parts.

The first part aims to minimize the difference between adversarial trajectories and real trajectories, ensuring the characteristic of subtle perturbations in adversarial attacks. The loss function of the first part is calculated by:

$$\mathcal{L}_{G1} = -\frac{1}{n}\sum_{i=1}^{n}\log\left(D\left(G\left(z_i\right)\right)\right) \tag{1}$$

where $D(\bullet)$ represents the output of the discriminator, indicating the probability that the trajectory is classified as a genuine normal trajectory; $G(\bullet)$ represents the adversarial attack perturbation generated by the generator from the noise

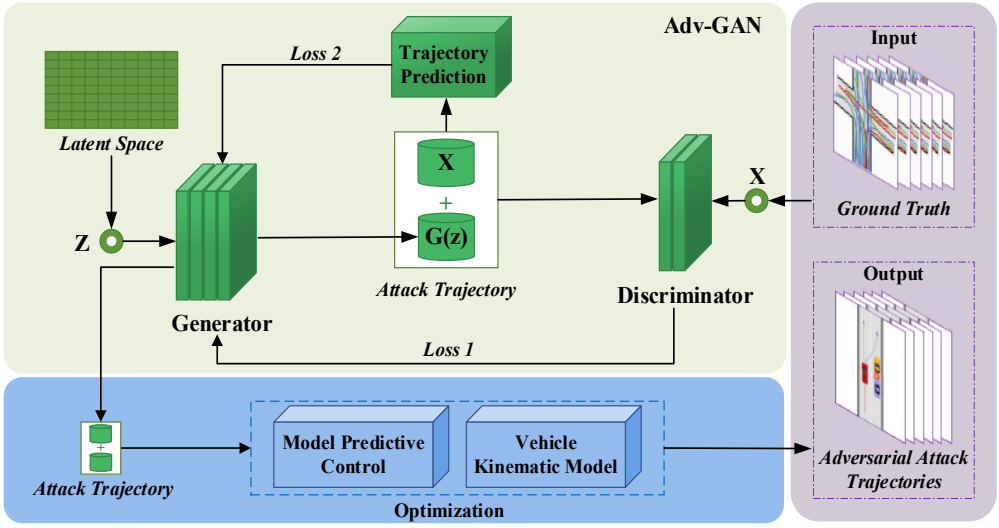

Fig. 3. The framework of Adv-GAN.

vector; z_i represents the noise vector sampled from the latent space; n represents the sample size.

The second part aims to maximize the impact of the generated attack trajectories, specifically by maximizing the difference between the ground truth trajectories and the trajectory prediction results under attack trajectory inputs, ensuring the effectiveness of the adversarial attack. The loss function of the second part is calculated by:

$$\mathcal{L}_{G2} = \frac{1}{n}\sum_{i=1}^{n}\left\|P(G(z)+x) - P(x)\right\|^2 \tag{2}$$

where $P(\bullet)$ represents the output of the trajectory prediction model; x represents the real trajectory.

In summary, the loss function of the generator is calculated by:

$$\begin{aligned}\mathcal{L}_G(\theta) = &-\mathbb{E}_{z\sim p(z)}[\log(D(G(z)+x))] \\ &+ \lambda\cdot\mathbb{E}_{z\sim p(z)}\left[\left\|P(G(z)+x) - P(x)\right\|^2\right]\end{aligned} \tag{3}$$

where θ represents the parameters of the generator; $p(z)$ represents the distribution of noise data z; λ is the weight of the mean square error term used to balance the importance of the two loss terms.

C. Discriminator

The discriminator network aims to learn to distinguish between the differences in fake data and real data. The inputs to the discriminator are the normalized original trajectory vectors and the perturbed trajectory vectors. It outputs scalars representing the probability that the input data are real. The loss function of the discriminator comprises two parts. The first part involves maximizing the probability of classifying real data x as real. The second part involves maximizing the probability of classifying fake data $G(z)+x$ as fake. The loss function of the discriminator can be expressed as:

$$\begin{aligned}\mathcal{L}_D(\theta,\phi) = &-\mathbb{E}_{x\sim p_{\text{data}}(x)}[\log D(x)] \\ &- \mathbb{E}_{z\sim p(z)}[\log(1 - D(G(z)+x))]\end{aligned} \tag{4}$$

where ϕ represents the parameters of the discriminator; $p_{\text{data}}(x)$ represents the distribution of real data x.

D. Training Process

The GAN model is trained by iteratively optimizing the loss functions of the discriminator and generator, which involves minimizing the discriminator's loss function while simultaneously maximizing the generator's loss function. This process aims to train a generator capable of producing highly adversarial trajectories with strong attack capabilities. The minimax game between the generator and the discriminator can be expressed as:

$$\min_G \max_D \begin{pmatrix} -\mathbb{E}_{z\sim p(z)}[\log(D(G(z)+x))] \\ +\lambda\cdot\mathbb{E}_{z\sim p(z)}\left[\left\|P(G(z)+x) - P(x)\right\|^2\right]\end{pmatrix}. \tag{5}$$

The goal of the generator is to minimize the probability of the generated samples being identified as fake, while minimizing the discrepancy between the generated samples and real data. Conversely, the discriminator seeks to maximize its ability to effectively distinguish between real and fake samples. The objective of the entire optimization process is to achieve a Nash equilibrium where the generator can produce effective adversarial trajectories.

E. Attack Trajectory Optimization

Upon generating adversarial trajectories with GAN, we undertake an optimization procedure to ascertain the practical applicability of these trajectories in real-world contexts, utilizing the principles of MPC. Adversarial trajectories are fed as initial inputs to the MPC algorithm. The MPC algorithm, based on the current state and the vehicle's kinematic model, selects the optimal control inputs to ensure that the trajectories are as rational and safe as possible. Various constraints are considered in the optimization process, including vehicle kinematics and trajectory smoothness. Linear MPC is employed for optimization, and the vehicular kinematic model is simplified to a bicycle kinematic model with the vehicle's center at the rear axle.

Select the state quantity as $\boldsymbol{\chi} = [x, y, \varphi]^T$, and the control quantity as $\boldsymbol{u} = [v, \delta]^T$, where x, y represents the vehicle position coordinate, φ represents the vehicle heading angle, δ represents the steering angle. The attack trajectories generated by Adv-GAN serve as the reference trajectories. The optimization objective aims to minimize the disparity between the output trajectories and the reference trajectories. To ensure that the output trajectories comply with the physics and real driving behaviors, we impose constraints based on the vehicle's kinematic model and control quantity $\tilde{\boldsymbol{u}}(k)$. The optimal problem can be expressed as:

$$\begin{aligned}\min J = &\sum_{k=0}^{N-1}\left(\tilde{\boldsymbol{x}}(k)^T Q\tilde{\boldsymbol{x}}(k) + \tilde{\boldsymbol{u}}(k)^T R\tilde{\boldsymbol{u}}(k)\right) \\ &+ \tilde{\boldsymbol{x}}(N)^T Q_f\,\tilde{\boldsymbol{x}}(N)\end{aligned} \tag{6}$$

s.t.

$$\tilde{\boldsymbol{x}}(k+1) = \boldsymbol{A}\tilde{\boldsymbol{x}}(k) + \boldsymbol{B}\tilde{\boldsymbol{u}}(k) \tag{7}$$

$$\left|\tilde{\boldsymbol{u}}(k)\right| \le \tilde{\boldsymbol{u}}_{\max} \tag{8}$$

where Q, R and Q_f are respectively the set state cost matrix，the input cost matrix, and the final state cost matrix, respectively the set state deviation and the relative weight of the input, all of which are positive definite matrices; $\tilde{\boldsymbol{x}}(\bullet)$ and $\tilde{\boldsymbol{u}}(\bullet)$ represent the state vector and control vector at time step k respectively; $\boldsymbol{A}$ is the state transition matrix, which describes the evolution of the system state over unit time; $\boldsymbol{B}$ is the control input matrix, describing how control inputs affect the changes in system state.

V. Experiments and Results

A. Dataset

In this paper, three public driving trajectory datasets are mainly used: Apolloscape Trajectory [26], Nuscenes [27], and

NGSIM [28]. The Apolloscape and Nuscenes databases contain urban road data, while the NGSIM database focuses on highway scenarios. Each dataset is divided into training set and test set, and the trajectory data are preprocessed according to the input data format of the trajectory prediction model and the Adv-GAN data processing requirements. Since the feature distributions of the datasets are different, Adv-GAN is trained independently for each dataset.

In order to ensure the consistency of attack effect measurement standards, the vehicle trajectory data in three datasets are preprocessed, and each 12 frames are divided into a vector through a sliding time window. The first 6 frames are used as the original trajectory to train the discriminator, and the last 6 frames are used for comparison with the trajectory prediction results. For instance, in the Apolloscape database, where the sampling frequency is two frames per second, we use data from the past 3 seconds to predict the trajectories for the next 3 seconds.

B. Trajectory prediction model

We use the open-source model Enhanced Graph-based Interaction-aware Trajectory Prediction (GRIP++) [15] in Adv-GAN. GRIP++ employs a graphical representation to capture the dynamics among proximate entities, harnesses multiple graph convolutional layers for feature distillation, and leverages an encoder-decoder LSTM architecture for predictive modeling. Besides, it uses both fixed and dynamic graphs of different types of traffic agents. GRIP++ achieves better prediction accuracy than state-of-the-art schemes, and it is faster than state-of-the-art scheme.

C. Results

We employ more than 5800 trajectories across three datasets for training and evaluation on the test set to validate the attack performance of Adv-GAN. The impact of an attack trajectory on trajectory prediction outcomes is illustrated in Fig. 4. A comparison between prediction outcomes under adversarial attacks and normal conditions reveals that prior to the attack implementation, the predicted trajectory aligns closely with the future trajectory longitudinally, which does not significantly affect normal vehicle operation. However, post-attack implementation, the attack trajectory exhibits minimal lateral deviation from the original trajectory (not exceeding half the width of a regular lane), indicating a resemblance to the original driving behavior of vehicle at the time of the attack. Nevertheless, the adversarial attack results in substantial prediction errors in the predicted trajectory.

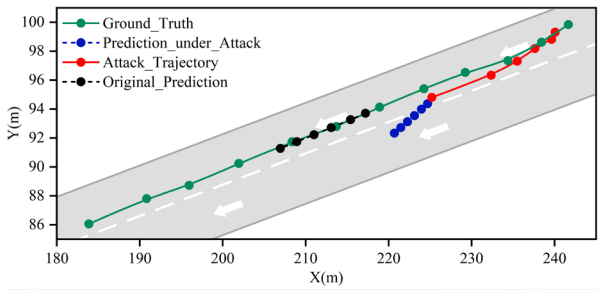

Fig. 4. Comparison of trajectory prediction under adversarial attack and under normal condition.(Apolloscape dataset)

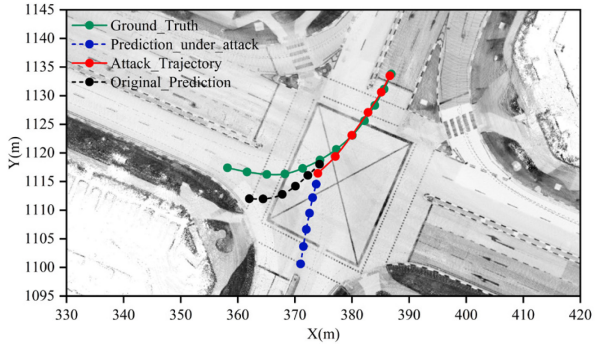

Fig. 5. Example of adversarial attacks on crossroads.(Nuscenes dataset)

This effect is further pronounced in Fig. 5, an example demonstrating adversarial attacks on crossroads from the Nuscenes dataset. Adversarial trajectory results in significant disparities between the predicted trajectory and the ground truth, thereby leading to incorrect predictions of the vehicle's driving behavior. At crossroads with complicated traffic conditions, this discrepancy could mislead the decision to drive, potentially resulting in collisions or other traffic accidents.

Table I showcases the attack effects of Adv-GAN across different datasets, illustrating its outstanding performance in adversarial attacks on multiple datasets. Post-attack implementation, there is a significant increase in trajectory prediction errors. Due to differences in data distribution and quality among the databases, the impact of the attack varies. However, Adv-GAN increases the average error by more than 50% across all three databases. On the Apolloscape dataset, the attack effect is most pronounced, with ADE increasing 6.94 times and FDE increasing by over 5 times. On the NuScenes dataset, ADE increases by nearly 3 times, and FDE increases by 1.8 times. On the NGSIM dataset, the attack demonstrates capability in highway scenario data, with both ADE and FDE increasing by over 2 meters.

To demonstrate the performance of Adv-GAN in adversarial attacks, we conduct comparisons with AdvTrajectoryPrediction [7], as depicted in Table II. Our approach outperforms baseline methods in improving both ADE and FDE of model predictions. As shown in Fig. 6, we

TABLE I. AVERAGE PREDICTION ERROR ON THREE DATASETS

Dataset	**ADE(m)**		**FDE(m)**	
	Normal	***Attack***	***Normal***	***Attack***
Apolloscape	1.97	15.71	3.18	16.34
Nuscenes	2.41	7.09	5.14	9.04
NGSIM	4.41	6.46	7.53	9.60

TABLE II. ATTACK PERFORMANCE OF ADV-GAN AND BASELINE

Scenario	**Evaluation Metrics**	
	ADE(m)	***FDE(m)***
Normal	1.97	3.18
Adv-GAN	**15.71**	**16.34**
AdvTrajectory Prediction[7]	7.14	10.74

a. Test on Apolloscape dataset.

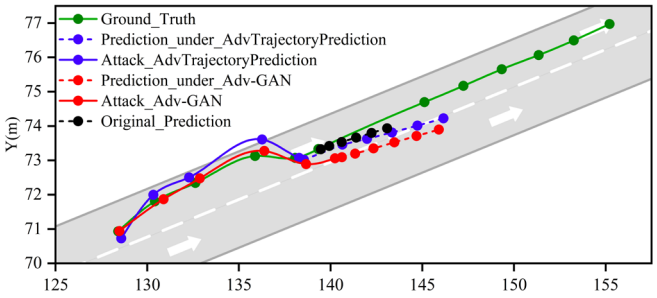

Fig. 6. Comparison of attacks from Adv-GAN and AdvTrajectoryPrediction.(Apolloscape dataset)

present a comparison between the attack effects of Adv-GAN and AdvTrajectoryPrediction. Our approach exhibits more subtle attack trajectories and induces larger prediction errors.

The experiments and results above demonstrate that our adversarial attack method can significantly impact the prediction accuracy of trajectory prediction model. It exhibits strong adversarial attack performance, capable of generating adversarial trajectories that pose a significant threat to trajectory prediction.

VI. Conclusion

In this paper, an LSTM-based attack trajectory generation model named Adv-GAN is proposed. The goal of this framework is to obtain the adversarial trajectories with real driving feature distribution. This is achieved by a novel GAN architecture tailored to the characteristics of vehicle driving and optimization with the vehicle kinematics model. Multiple public datasets are used for evaluation. Experimental results show that the proposed method effectively increase trajectory prediction errors, and has better attack performance compared to existing adversarial attack model. This demonstrates the great threat of adversarial attack, an intriguing future direction for this work involves to increase adversarial robustness of trajectory prediction models.

References

[1] Y Ma, Z Wang, H Yang, L YangG. O. Young, "Artificial intelligence applications in the development of autonomous vehicles: A survey." *IEEE/CAA Journal of Automatica Sinica 7*, no. 2: 315-329, 2020.

[2] D Cao, X Wang, L Li, C Lv, X Na, Y Xing, X Li, Y Li, Y Chen, FY Wang, "Future directions of intelligent vehicles: Potentials, possibilities, and perspectives." IEEE Transactions on Intelligent Vehicles 7, no. 1: 7-10, 2022.

[3] H Jiang, L Chang, Q Li, D Chen, "Trajectory prediction of vehicles based on deep learning." *In 2019 4th International Conference on Intelligent Transportation Engineering (ICITE)*, pp. 190-195, 2019.

[4] M Ozdag, "Adversarial attacks and defenses against deep neural networks: a survey." *Procedia Computer Science 14*: 152-161,2018.

[5] 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.

[6] X Huang, D Kroening, W Ruan, J Sharp, Y Sun. E Thamo, M Wu, X Yi, "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability." *Computer Science Review 37*: 100270, 2020.

[7] Q Zhang, S Hu, J Sun, QA Chen, and Z. M Mao, "On adversarial robustness of trajectory prediction for autonomous vehicles." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15159-15168, 2022.

[8] Y Cao, C Xiao, A Anandkumar, D Xu, M Pavone, "Advdo: Realistic adversarial attacks for trajectory prediction." In *European Conference on Computer Vision, Cham: Springer Nature Switzerland*, pp. 36-52, 2022.

[9] I Goodfellow, J Pouget-Abadie, M Mirza, B Xu, D Warde-Farley, S Ozair, A Courville, and Y Bengio, "Generative adversarial networks." Communications of the ACM 63, no. 11: 139-144, 2020.

[10] MA Bashar, R Nayak, "TAnoGAN: Time series anomaly detection with generative adversarial networks." In *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 1778-1785, 2020.

[11] Y Huang, J Du, Z Yang, Z Zhou, L Zhang, H Chen, "A survey on trajectory-prediction methods for autonomous driving." *IEEE Transactions on Intelligent Vehicles 7*, no. 3: 652-674, 2022.

[12] H Song, D Luan, W Ding, MY Wang, Q Chen, "Learning to predict vehicle trajectories with model-based planning." *Conference on Robot Learning. PMLR*, pp. 1035-1045, 2022.

[13] J Mercat, T Gilles, N El Zoghby, G Sandou, D Beauvois, GP Gil, "Multi-head attention for multi-modal joint vehicle motion forecasting." *2020 IEEE International Conference on Robotics and Automation (ICRA). IEEE*, pp. 9638-9644, 2020.

[14] X Li, X Ying, MC Chuah, "Grip: Graph-based interaction-aware trajectory prediction. "In *2019 IEEE Intelligent Transportation Systems Conference (ITSC)*, pp. 3960-3966, 2019.

[15] X Li, X Ying, MC Chuah, "Grip++: Enhanced graph-based interaction aware trajectory prediction for autonomous driving."*arXiv preprint arXiv:1907.07792*, 2019.

[16] M Xue, C Yuan, H Wu, Y Zhang, W Liu, "Machine learning security: Threats, countermeasures, and evaluations." *IEEE Access 8*: 74720-74742, 2020.

[17] N Papernot, P McDaniel, A SinhaPapernot, and M Wellman. "Towards the science of security and privacy in machine learning." *arXiv preprint arXiv:1611.03814*, 2016.

[18] A Duan, R Wang, Y Cui, P He, C Luo, "Causal Robust Trajectory Prediction Against Adversarial Attacks for Autonomous Vehicles." IEEE Internet of Things Journal, 2023.

[19] SH Silva, P Najafirad, "Opportunities and challenges in deep learning adversarial robustness: A survey." *arXiv preprint arXiv:2007.00753*, 2020.

[20] Y Li, C Wen, F Juefei-Xu, C Feng, "Fooling lidar perception via adversarial trajectory perturbation." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7898-7907, 2021.

[21] J Sun, Y Cao, QA Chen, ZM Mao, "Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures." In *29th USENIX Security Symposium (USENIX Security 20)*, pp. 877-894, 2020.

[22] T Sato, J Shen, N Wang, Y Jia, X Lin, QA Chen, "Dirty road can attack: Security of deep learning based automated lane centering under {Physical-World} attack." In *30th USENIX security symposium (USENIX Security 21)*, pp. 3309-3326, 2021.

[23] C Xiao, B Li, JY Zhu, W He, M Liu, D Song, "Generating adversarial examples with adversarial networks." arXiv preprint arXiv:1801.02610, 2018.

[24] Z Lin, Y Shi, Z Xue, "Idsgan: Generative adversarial networks for attack generation against intrusion detection." Pacific-asia conference on knowledge discovery and data mining. Cham: Springer International Publishing, pp. 79-91, 2022.

[25] S Hochreiter, J Schmidhuber, "Long short-term memory." *Neural computation 9*: 1735-1780, 1997.

[26] X Huang, X Cheng, Q Geng, B Cao, D Zhou, P Wang, Y Lin, and R Yang. "The apolloscape dataset for autonomous driving." In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 954-960, 2018.

[27] H Caesar, V Bankiti, AH 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition,* pp. 11621-11631, 2020.

[28] V Alexiadis, J Colyar, J Halkias, R Hranac, and G McHale. "The next generation simulation program." *Institute of Transportation Engineers. ITE Journal 74*, no. 8: 22, 2004.