2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

AdvGPS: Adversarial GPS for Multi-Agent Perception Attack

Jinlong Li[1], Baolu Li[1], Xinyu Liu[1], Jianwu Fang[2], Felix Juefei-Xu[3], Qing Guo[4*], Hongkai Yu[1*]

***Abstract*— The multi-agent perception system collects visual data from sensors located on various agents and leverages their relative poses determined by GPS signals to effectively fuse information, mitigating the limitations of single-agent sensing, such as occlusion. However, the precision of GPS signals can be influenced by a range of factors, including wireless transmission and obstructions like buildings. Given the pivotal role of GPS signals in perception fusion and the potential for various interference, it becomes imperative to investigate whether specific GPS signals can easily mislead the multi-agent perception system. To address this concern, we frame the task as an adversarial attack challenge and introduce ADVGPS, a method capable of generating adversarial GPS signals which are also stealthy for individual agents within the system, significantly reducing object detection accuracy. To enhance the success rates of these attacks in a black-box scenario, we introduce three types of statistically sensitive natural discrepancies: appearance-based discrepancy, distribution-based discrepancy, and task-aware discrepancy. Our extensive experiments on the OPV2V dataset demonstrate that these attacks substantially undermine the performance of state-of-the-art methods, showcasing remarkable transferability across different point cloud based 3D detection systems. This alarming revelation underscores the pressing need to address security implications within multi-agent perception systems, thereby underscoring a critical area of research. The code is available at https://github.com/jinlong17/AdvGPS.**

I. INTRODUCTION

Although the single-agent perception system gets advanced performance in many autonomous driving scenarios, it still has considerable sensing limitations due to the challenges of occlusion and perception range. Benefited from the recent research of multi-agent perception system, the visual data from sensors on nearby agents can be shared to the ego-agent as information fusion to improve the perception range and overcome occlusion challenges [1]. During the visual data sharing from nearby agents to the ego-agent, one important and necessary step is to project the visual data from the coordinate systems of nearby agents to the uniform coordinate system of the ego-agent via homogeneous transformation [2]. This homogeneous transformation based uniform coordinate projection requires their relative pose (localization & heading) from nearby agents to ego agent [2]–[4], which is determined by their GPS signals.

However, it is a common sense that the GPS signals might have unavoidable errors in the real world [2], [5]. GPS signal accuracy can be affected by a variety of factors, such as wireless transmission, building obstacles and so on. This paper uses Vehicle-to-Vehicle (V2V) cooperative perception in autonomous driving as a study example of the multi-agent perception system. The previous work [5] shows that simply adding random GPS signal noises in a specific range could result in spoofing attacks to Connected and Automated Vehicles (CAV). In the research domain of V2V cooperative perception, whether adversarial GPS signals can easily mislead the multi-agent perception system is never studied before as an open question.

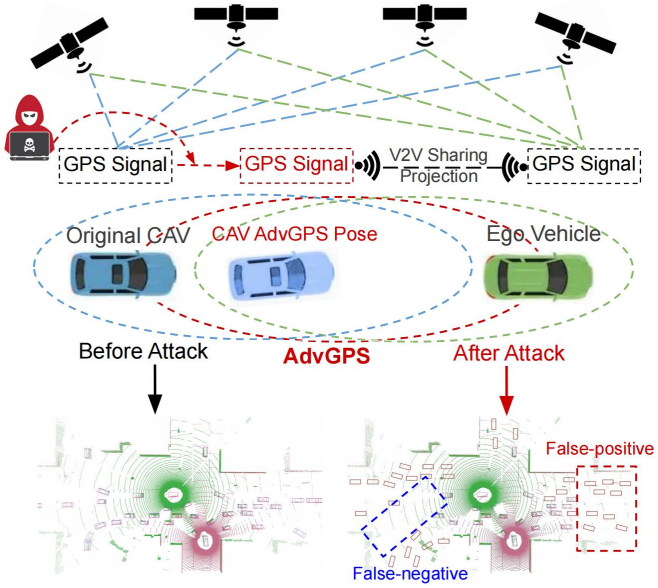


Fig. 1: **Illustration of AdvGPS for multi-agent perception attack**. Here we use Vehicle-to-Vehicle (V2V) cooperative perception in autonomous driving as an example. Ego vehicle might receive the shared visual information from other CAVs with the adversarial GPS signal, leading to significant false-negative and false-positive detection errors.

To investigate the above open question, we model the task as an adversarial attack challenge and introduce a novel method **AdvGPS** capable of generating adversarial GPS signals which are also stealthy for nearby CAVs, significantly reducing object detection accuracy of the ego vehicle. As illustrated in Fig. 1, when the adversarial GPS signals within a small/stealthy range are added to the nearby CAVs, it will generate significant false-negative (missing) and false-positive detection errors for the ego vehicle. We formulate the AdvGPS attack in the black-box setting, because the prior knowledge of the target perception model is usually unknown. To enhance the success rates of the AdvGPS attack

[1]Cleveland State University. [2]Xi'an Jiaotong University. [3]New York University. [4]CFAR and IHPC, Agency for Science, Technology and Research (A*STAR), Singapore. This research is supported by NSF 2215388, and the National Research Foundation, Singapore, and DSO National Laboratories under the AI Singapore Programme (No: AISG2-GC-2023-008). *Co-corresponding authors: tsingqguo@ieee.org and h.yu19@csuohio.edu.

in the black-box scenario, we introduce three types of statistically sensitive natural discrepancies for multi-agent perception attack: appearance-based discrepancy, distribution-based discrepancy, and task-aware discrepancy.

Using the publicized dataset OPV2V [1], we conducted extensive experiments and the experimental results demonstrate that the proposed AdvGPS attacks substantially undermine the performance of state-of-the-art V2V cooperative perception methods, showcasing remarkable transferability across different point cloud based 3D detection systems. This alarming discovery emphasizes the critical need to confront security implications in multi-agent perception systems, thus highlighting a pivotal research area. Our contributions of this paper can be summarized as follows.

- To the best of our knowledge, we propose the **first research** of adversarial GPS signals which are also stealthy for the V2V cooperative perception attacks, denoted as AdvGPS.
- We propose three statistically sensitive natural discrepancies in AdvGPS to enhance the multi-agent perception attack in the black-box scenarios, *i.e.*, appearance-based discrepancy, distribution-based discrepancy, and task-aware discrepancy.
- The experimental results on the publicized OPV2V dataset demonstrate that our AdvGPS attacks substantially undermine the performance of state-of-the-art methods and show outstanding transferability across different point cloud based 3D detection systems.

II. Related Work

Cooperative Perception in V2V. Multi-vehicle perception systems have arisen to address the inherent limitations of single-vehicle sensing by harnessing information exchange among multiple vehicles. Researchers have frequently incorporated collaboration modules to enhance efficiency and overall system performance. Typically, these systems employ three fundamental schemes for multi-vehicle observation aggregation: raw data fusion at the input stage, intermediate feature fusion during processing, and output fusion. State-of-the-art approaches often opt for sharing intermediate neural features augmented with contextual information about the environment. This choice makes a favorable balance between accuracy and bandwidth requirements [1], [6]. Notably, methods such as Attfuse [1] and V2VAM [7] leverage attention mechanisms to fuse multi-vehicle features for 3D object detection. Additionally, the integration of the popular Vision Transformer has been introduced to capture complex spatial interactions among multiple agents, like V2X-ViT [6], and CoBEVT [8]. While these methods have exhibited impressive performance in the context of V2V perception, their robustness against adversarial GPS attacks remains an under-explored research.

Deployment of Cooperative Perception. Multi-agent perception systems offer numerous advantages but are accompanied by challenges such as localization errors, communication latency, adversarial attacks, and lossy communication. These development challenge can easily diminish

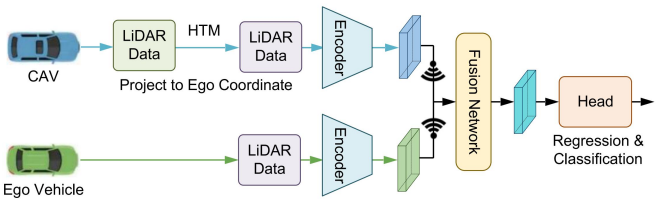


Fig. 2: **Illustration of V2V cooperative perception pipeline with LiDAR data coordinate projection from CAV to Ego.**

the benefits of collaborations [6], [9]. Several strategies have been proposed to enhance robustness. For instance, V2X-ViT employs a Vision Transformer to mitigate GPS localization errors and sensing information delays during intermediate collaboration [6]. To address localization errors, [10] proposes a pose regression module that learns correction parameters to predict accurate relative transformations from noisy data. The Lossy Communication-aware Repair Network (LCRN) [4] is introduced to address packet loss issues in communication. A Multi-agent Perception Domain Adaption (MPDA) framework [11] is proposed for cooperative perception to bridge the domain gap for multi-agent perception. Nevertheless, V2V cooperative perception performance relies heavily on GPS signals, which might be susceptible to real-world attacks. This challenge might lead to a significant degradation in perception performance. This paper investigates GPS attacks in the context of 3D object detection tasks within V2V cooperative perception.

Adversarial Attack. Adversarial attacks have garnered significant attention, primarily focused on generating perturbations designed to mislead deep learning models into producing incorrect predictions [12]. These attacks can be categorized into two main types: white-box attacks [13], [14], where the attacker possesses full knowledge of the target model, and black-box attacks [15], [16], which are generally less effective as the attacker lacks access to the target model's internal details. In autonomous vehicles, GPS spoofing poses a substantial threat to vehicle localization, involving the transmission of false signals to deceive a vehicle's GPS system. Previous research has highlighted the dangers of GPS spoofing for autonomous vehicles operating in real-world scenarios [14], [17]. Notably, constant bias and gradual drift attacks are common GPS attack methods [5]. Additionally, recent work [17] introduced a position-altering attack (PAA), employing a GPS spoofer to mislead a vehicle to a fictitious location. In this paper, we introduce the new concept of GPS attacks into the V2V cooperative perception systems.

III. V2V Perception Pipeline and Motivation

The V2V perception pipeline, as illustrated in Fig. 2, initiates by selecting an ego vehicle from among the CAVs to create a spatial graph that encompasses nearby CAVs within the communication range. These nearby CAVs project their LiDAR data into the ego vehicle's coordinate frame using a Homogeneous Transformation Matrix (HTM) based on both the ego vehicle's and their own GPS poses. Subsequently, the pipeline proceeds with feature extraction, where each

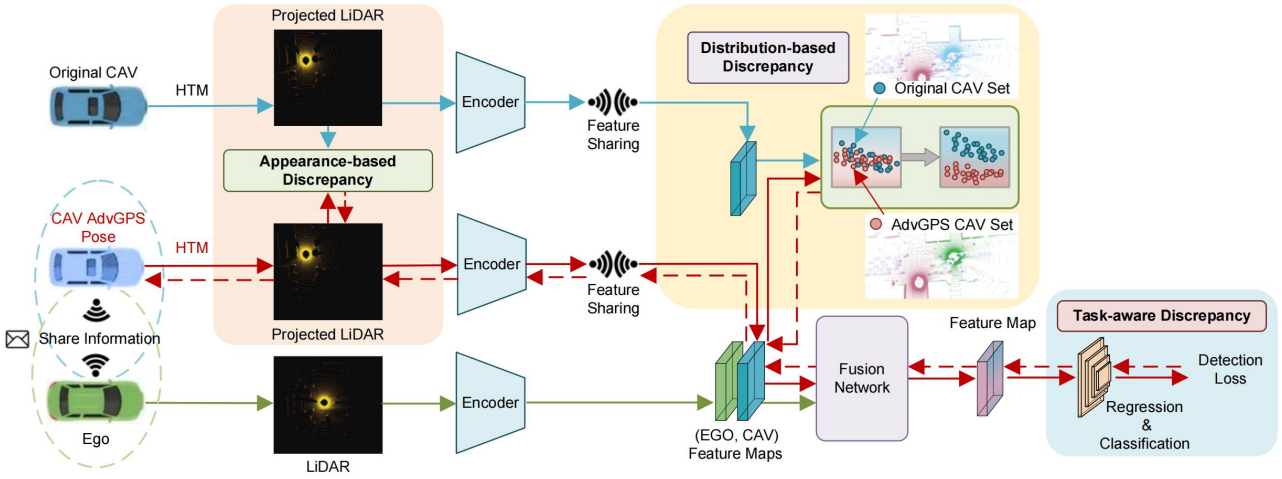


Fig. 3: **Pipeline of AdvGPS for multi-agent cooperative perception attack**. Dash lines indicate the back propagation.

CAV employs its own LiDAR feature extraction module. These extracted features are aggregated and fused via a feature fusion neural network. Finally, the fused feature maps are used for 3D bounding-box regression and classification, ultimately facilitating advanced cooperative perception in autonomous driving.

To elucidate this process with an example, let us define the ego vehicle's GPS pose as G_{ego} and the GPS pose of a neighboring CAV as G_{cav}, where both poses contain six variables $[x, y, z, \theta_x, \theta_y, \theta_z]$. The CAV can perceive the environment via its own sensor and get the point cloud data denoted as $\mathbf{P}_{cav} \in \mathbb{R}^{4\times m}$ that contains a set of 3D points $\{P_i \mid i = 1, 2, \cdots, m\}$, where each point P_i is a vector and contains the homogeneous coordinates $[x, y, z, 1]$. We can use a Homogeneous Transformation Matrix (HTM) $\mathbf{T}_{cav\to ego} \in \mathbb{R}^{4\times 4}$ to project/transform the point cloud of CAV to the uniform coordinate system of ego vehicle by

$$\mathbf{P}_{cav\to ego} = \mathbf{T}_{cav\to ego}\mathbf{P}_{cav}, \tag{1}$$

where $\mathbf{P}_{cav\to ego}$ denotes the transformed point cloud of $\mathbf{P}_{cav}$ in the ego coordinate system. The HTM $\mathbf{T}_{cav\to ego}$ format is $\begin{bmatrix} \mathbf{r} & \mathbf{t} \\ \mathbf{0}_{1\times 3} & 1 \end{bmatrix}$, where $\mathbf{r} \in \mathbb{R}^{3\times 3}$ and $\mathbf{t} \in \mathbb{R}^{3\times 1}$ are the rotation matrix and translation matrix, respectively. Then, the HTM $\mathbf{T}_{cav\to ego}$ can be extracted/calculated via

$$\mathbf{T}_{cav\to ego} = \text{HTM}_{\text{Extract}}(G_{ego}, G_{cav}), \tag{2}$$

where $\text{HTM}_{\text{Extract}}(\cdot)$ is an inverse function (matrix computation as defined in [2]) of homogeneous transformation to calculate the HTM based on the GPS poses of two entities, *i.e.*, G_{ego} and G_{cav}. According to the pipeline, GPS signals play a critical role in the V2V perception. Nevertheless, the accuracy of GPS signals is easily affected by diverse factors, and we aim to explore whether some specific GPS patterns could fool the V2V perception easily. To this end, we formulate the task from the view of an adversarial attack and propose the AdvGPS against the multi-agent perception in Sec. IV. Our goal is to search the adversarial GPS of CAV (*i.e.*, $\widehat{G}_{cav} \in [\hat{x}, \hat{y}, \hat{z}, \hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_z]$) that can lead to an adversarial HTM via Eq. 2. In the real world, adversarial GPS signals can be concealed within the range of actual stealthy GPS errors.

IV. AdvGPS against Multi-agent Perception

A. Overview

We propose AdvGPS and present the whole pipeline in Fig. 3. Specifically, we formulate V2V cooperative perception system $\phi(\cdot)$ for LiDAR-based 3D object detection as

$$\phi(\mathbf{P}_{cav}, \mathbf{P}_{ego}, G_{cav}) = \varphi(\text{Fusion}(\mathbf{F}_{cav\to ego}, \mathbf{F}_{ego})), \tag{3}$$

where $\mathbf{F}_{cav\to ego} = \text{Encoder}(\mathbf{T}_{cav\to ego}\mathbf{P}_{cav})$ and $\mathbf{F}_{ego} = \text{Encoder}(\mathbf{P}_{ego})$ denote the features of the two point clouds $\mathbf{P}_{cav\to ego}$ and $\mathbf{P}_{ego}$, Fusion$(\cdot)$ is a network to fuse the two features, and $\varphi(\cdot)$ is the regression and detection header for object detection [1]. It aims to exploit adversarial attacks on GPS signals to deceive any point cloud based 3D detection models in a black-box setting. We calculate an imperceptible noise-like perturbation under the guidance of a classic 3D object detection model and add it to the original CAV GPS pose to obtain an AdvGPS pose. This corrupted AdvGPS pose can then mislead any other detection models in a black-box setting. Our AdvGPS attack perturbation is stealthy to be confused as normal GPS signal errors. To enhance the success rates of attacks in black-box scenarios, we consider three objectives to optimize adversarial GPS: appearance-based discrepancy, distribution-based discrepancy, and detection task-aware discrepancy. We define the optimization over the adversarial GPS signal $\widehat{G}_{cav}$ as:

$$\arg\max_{\widehat{G}_{cav}} (\lambda D_{app} + \omega D_{dist} + \xi D_{task}) \tag{4}$$

$$\begin{aligned} \text{subject to} \quad & \|[\hat{x}, \hat{y}] - [x, y]\|_\infty < \epsilon_{x,y} \\ & \|\hat{z} - z\|_\infty < \epsilon_z \\ & \|[\hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_z] - [\theta_x, \theta_y, \theta_z]\|_\infty < \epsilon_{\theta_x,\theta_y,\theta_z}, \end{aligned} \tag{5}$$

where λ, ω, and ξ are the balance weights and 1s are used for them in our experiments. As the three responses

of the V2V cooperative perception system $\phi(\cdot)$ defined in Eq. 3 using $\widehat{G}_{cav}$, D_{app} represents the appearance-based discrepancy before and after the GPS attack, D_{dist} is used to measure the distribution-based discrepancy between original CAV set and AdvGPS CAV set, and D_{task} accounts for the task-aware discrepancy using a classic 3D object detection model. Our goal is to maximize the objective function by tuning AdvGPS pose $\widehat{G}_{cav}$, which includes six parameters $[\hat{x}, \hat{y}, \hat{z}, \hat{\theta}_x, \hat{\theta}_y, \hat{\theta}_z]$, meanwhile we constrain the GPS signal perturbation within the range of normal stealthy GPS errors in real world via the ϵ values by Eq. 5.

B. Appearance-based Discrepancy

The appearance-based discrepancy focuses on the differences between the original CAV projected point cloud $\mathbf{P}^{ori}_{cav\to ego}$ and the AdvGPS-based projected point cloud $\mathbf{P}^{adv}_{cav\to ego}$. The cooperative perception performance is directly affected by the appearance difference when fusing into the ego vehicle's perception system. To efficiently disrupt cooperative perception performance through appearance difference, our D_{app} can be defined as the appearance difference between these two point clouds by tuning the AdvGPS pose $\widehat{G}_{cav}$ via the mathematical expression:

$$D_{app} = L_{MSE}(\mathbf{P}^{ori}_{cav\to ego}, \mathbf{P}^{adv}_{cav\to ego}), \tag{6}$$

where $L_{MSE}(\cdot)$ is the mean squared error (MSE) loss between the two point clouds.

C. Distribution-based Discrepancy

After encoding, we obtain a set of original features by the GPS poses of Original CAV Set and a set of attacked features by the adversarial GPS poses of AdvGPS CAV Set. We investigate the correlations between statistical differences and the distribution of these features. Some previous studies [12] have found that statistical differences are positively associated with distribution differences. To maximize the discrepancy distance between the feature distributions of the original features $\mathbf{F}^{ori}$ and attacked features $\mathbf{F}^{adv}$, we use maximum mean discrepancy (MMD) [18] to measure the distance of them. Let $\mathcal{F}^{ori} = \{\mathbf{F}^{ori}_i\}$ and $\mathcal{F}^{adv} = \{\mathbf{F}^{adv}_i\}$ represent a set of original and adversarial features after encoding. Our goal is to generate adversarial GPS poses capable of degrading cooperative detection performance in terms of feature distributions. This can be defined as:

$$D_{dist} = L_{MMD}(\mathcal{F}^{ori}, \mathcal{F}^{adv}), \tag{7}$$

where $L_{MMD}(\cdot)$ represents the MMD loss between the two intermediate features after encoding.

D. Task-aware Discrepancy

Intuitively, given the LiDAR point clouds from ego vehicle and surrounding CAVs (*i.e.*, $\mathbf{P}_{ego}$ and $\mathbf{P}_{cav}$), a pre-trained cooperative detection model is used to simulate the target to be attacked. In our experiments, the pre-trained cooperative detection model uses the classic 3D object detection model VoxelNet [19] as encoder and a simple self-attention module as the fusion network. Then, based on the Eq. 3, we get the predicted detection results and calculate the loss according to the ground truth (*i.e.*, Y). L_{DET} is the task-aware discrepancy to mislead the pre-trained cooperative detection model, which can be formulated as

$$D_{task} = L_{DET}(\phi(\mathbf{P}_{cav}, \mathbf{P}_{ego}, \widehat{G}_{cav}), Y), \tag{8}$$

where $L_{DET}(\cdot)$ is denoted as the detection loss including regression and classification of 3D bounding boxes. Please note that the above pre-trained cooperative detection model to simulate target to be attacked is classic and simple in 3D point cloud detection, which is different with other state-of-the-art V2V cooperative perception models, so our AdvGPS still follows the black-box attack setting.

E. Implementation Details

To implement our AdvGPS for V2V cooperative perception attack, we follow the general adversarial attack procedure: ❶ We set the additive perturbation $\mathbf{g}$ to $\widehat{G}_{cav}$ and obtain the original CAV pose to calculate the HTM. ❷ Based on the HTM, we calculate D_{app} for the projected point cloud. ❸ We feed the projected point cloud into an encoder to generate the intermediate feature and calculate D_{dist}. ❹ The intermediate features are then inputted into the fusion network to calculate D_{task}. ❺ We perform back propagation and compute the gradients of $\mathbf{g}$ with respect to the loss functions. ❻ The sign of the gradients is used to update additive perturbation $\mathbf{g}$ by multiplying them with a step size, then adding $\mathbf{g}$ to the $\widehat{G}_{cav}$. ❼ We calculate a new synthesized $\mathbf{T}_{cav\to ego}$ via Eq. 2 and repeat steps ❷ to ❻ for a defined number of iterations. We set the number of iterations to 10. To conceal within the range of actual GPS errors, we follow the statistics presented in [20], which indicates that the real-world GPS signal's average localization error, height error, and heading error are 1.118 meters, 1.395 meters and 0.141 degrees, so we set $\epsilon_{x,y}$, ϵ_z, and $\epsilon_{\theta_x,\theta_y,\theta_z}$ as 1.118, 1.395, and 0.141 respectively to avoid significant offsets.

V. Experiments

A. Experimental Setups

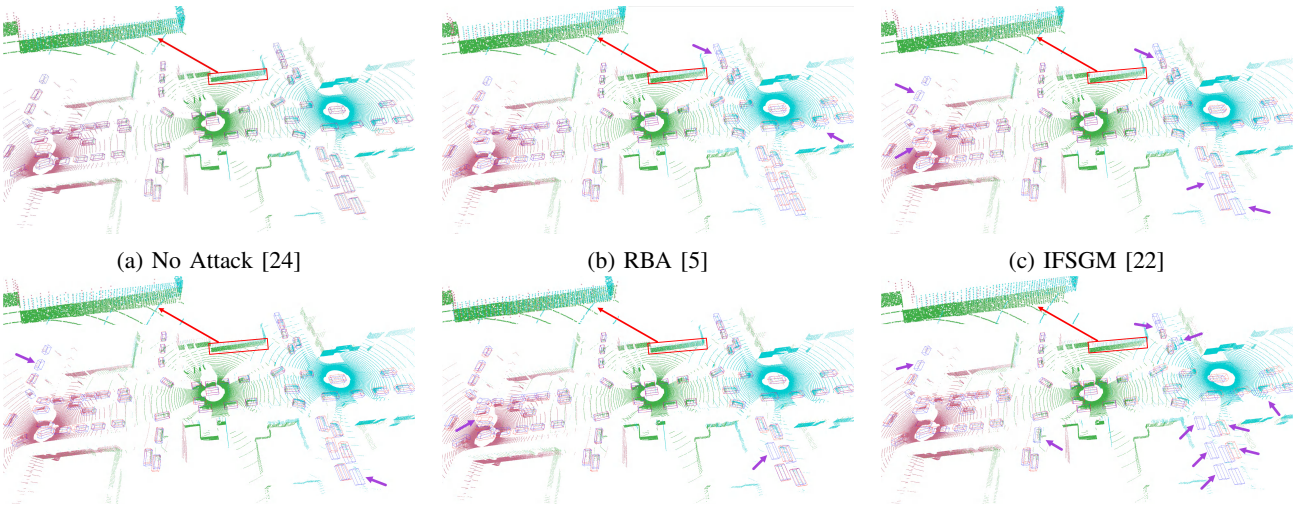
Dataset: We conduct GPS attack experiments on the publicized OPV2V dataset [1] for V2V cooperative perception tasks. OPV2V [1] is a publicized open-source simulated dataset for V2V perception, which is collected in CARLA and OpenCDA [3]. Following the default setting of OPV2V [1], we use 594 frames in the digital town of Culver City, Los Angeles with the same road topology used as the testing set for all methods.

3D detection methods: We selected the five state-of-the-art cooperative perception methods as the point cloud based 3D detection models, *i.e.*, AttFuse [1], F-cooper [24], V2VAM [7], V2X-ViT [6] and CoBEVT [8]. All methods leverage the anchor-based PointPillar method [25] to extract visual features from point clouds because of its low inference latency and optimized memory usage.

Evaluation metrics: We assess the final 3D vehicle detection accuracy for our proposed framework performance. Similar to prior works [1], [6], we set $x \in [-140, 140]$ meters, $y \in$

TABLE I: **Quantitative results of GPS attack**: 3D detection performance on V2V Culver City testing set of OPV2V dataset. We show the Average Precision (AP) at IoU=0.5. The best and second best attacked performance among five state-of-the-art cooperative perception methods are respectively highlighted using red and blue color.

Model	CAVs' Pose	No Attack	RBA [5]	FSGM [21]		IFSGM [22]		PGD [23]		PAA [17]		**AdvGPS**
-	-	W.b./B.b.	W.b./B.b.	W.b.	B.b.	W.b.	B.b.	W.b.	B.b.	W.b.	B.b.	B.b.
No Fusion	-	0.557	-	-	-	-	-	-	-	-	-	-
Attfuse [1]	G_{xyz}	0.918	0.657	0.572	0.583	0.626	0.638	0.567	0.612	0.550	0.561	**0.502**
	G_{all}		0.669	0.571	0.579	0.624	0.626	0.572	0.597	0.540	0.559	**0.499**
F-Cooper [24]	G_{xyz}	0.898	0.526	0.508	0.550	0.443	0.601	0.422	0.583	0.421	0.527	**0.299**
	G_{all}		0.538	0.510	0.549	0.436	0.586	0.419	0.562	0.413	0.524	**0.298**
V2VAM [7]	G_{xyz}	0.920	0.568	0.565	0.593	0.523	0.653	0.516	0.628	0.523	0.563	**0.422**
	G_{all}		0.584	0.566	0.592	0.523	0.638	0.513	0.607	0.520	0.557	**0.433**
V2X-ViT [6]	G_{xyz}	0.928	0.589	0.599	0.615	0.573	0.671	0.546	0.644	0.533	0.533	**0.433**
	G_{all}		0.597	0.600	0.615	0.575	0.658	0.549	0.628	0.529	0.529	**0.434**
CoBVET [8]	G_{xyz}	0.904	0.570	0.570	0.592	0.547	0.649	0.517	0.624	0.498	0.559	**0.397**
	G_{all}		0.577	0.568	0.592	0.545	0.634	0.520	0.610	0.493	0.553	**0.403**



(a) No Attack [24] (b) RBA [5] (c) IFSGM [22]

(d) PGD [23] (e) PAA [6] (f) AdvGPS

Fig. 4: **3D detection visualization on attacking V2V model CoBEVT [8]**. Blue and red 3D bounding boxes represent ground truth and prediction. Purple arrows: detection errors. Green point cloud: by ego vehicle. Other-color point clouds: projected to ego coordinate by nearby attacked CAVs.

$[-40, 40]$ meters as the evaluation range, where all CAVs (in the range of $[1, 5]$) are included in this spatial range. We employ Average Precision (AP) at the Intersection-over-Union (IoU) threshold of 0.5 as the accuracy metric.

Attack settings: To evaluate the effectiveness of our AdvGPS attack on the V2V perception pipeline, we compare it with three commonly used attack methods on CAVs' GPS poses: Projected Gradient Descent (PGD) [23], Fast Gradient Sign Method (FGSM) [21], and Iterative Fast Gradient Sign Method (IFGSM) [22]. Additionally, we include two current GPS spoofing methods: Random Bias Attack (RBA) [5] and Position Altering Attack (PAA) [17]. FGSM, IFGSM, PGD, and PAA are gradient-based attack methods, following the *same* implementation settings described in Section IV-E. RBA can be implemented by applying uniformly distributed random bias within the same range settings ($\epsilon_{x,y}$, ϵ_z and $\epsilon_{\theta_x,\theta_y,\theta_z}$) as in Section IV-E. We use the Adam optimizer [26], and all models are executed on two RTX 3090 GPUs. We conduct both black-box and white-box attacks to assess the vulnerability and transferability of the cooperative perception methods. These attack scenarios are described as:

- *White-Box Attack (W.b.)*: All attack methods utilize PointPillar [25] as the encoder and AttFuse [1] as the feature fusion network. The adversarial GPS information of all attacked CAVs for the entire duration of the frames is saved to be used for attacking other cooperative perception methods.
- *Black-Box Attack (B.b.)*: We employ VoxelNet [19] as the encoder and a naive self-attention model as the feature fusion network in training. We save the adversarial GPS information of all attacked CAVs across the entire frame duration to be used for attacking the *unseen* PointPillar [25] encoder and *other* state-of-the-art fusion methods of cooperative perception.

B. Quantitative Evaluation

GPS attack performance analysis: Table I overviews the GPS attack results on the 3D object detection task. We conducted attacks on three parameters $[x, y, z]$ and all six parameters $[x, y, z, \theta_x, \theta_y, \theta_z]$ of the CAV poses G_{cav}, denoted as G_{xyz} and G_{all}, respectively. In the *White-Box Attack* scenario, all attack methods demonstrated the ability to degrade the detection performance of state-of-the-art (SOTA) intermediate fusion methods. For instance, the traditional random bias method RBA [5] achieved 58.9% for G_{xyz} and 59.7% for G_{all} when attacking the V2X-ViT [6] model. In contrast, the gradient-based PAA [17] achieved 53.3% for G_{xyz} and 52.9% for G_{all}, both of which reduced the performance below that of *No Fusion*. This indicates that GPS attacks can easily undermine the advantages of collaboration among intermediate fusion methods. In the *Black-Box Attack* scenario, AdvGPS outperforms all other attacks across various SOTA fusion methods (See Table I). When attacking the G_{xyz} of CAV poses, AdvGPS achieved 29.9% on F-Cooper [24], 42.2% on V2VAM [7], 43.3% on V2X-ViT [6], and 39.7% on CoBEVT [8]. These results exceeded the second-best attacks by approximately 12.1%, 9.4%, 10.0%, and 10.1%, respectively. Our AdvGPS, leveraging three types of statistically sensitive natural discrepancies, *i.e.*, appearance-based discrepancy, distribution-based discrepancy, and task-aware discrepancy, effectively reduces object detection accuracy and demonstrates remarkable vulnerability and transferability in multi-agent perception tasks. Furthermore, when comparing the results of the *Black-Box Attack* setting for AdvGPS with the *White-Box Attack* results for all five attacks, AdvGPS consistently achieves the highest attack success rate. Comparing the results of attacking all six parameters G_{all} and only G_{xyz} reveals their similar performance, which indicates that directly attacking the position of CAVs' poses G_{xyz} is an efficient approach to reduce computation complexity while maintaining exceptional attack performance in real-world scenarios.

Sensitivity analysis of GPS parameters: We have conducted a sensitivity analysis for each parameter of CAV poses $\widehat{G}_{cav}$ in Table II. Even when operating within the boundaries of typical GPS errors, both the PAA [17] and our AdvGPS methods remain effective in significantly reducing object detection accuracy for two ViT-based models, namely V2X-ViT [6] and CoBEVT [8]. Intuitively, among all six parameters, attacking the positional parameters x and y of CAVs' poses is the most efficient approach within a small range of GPS pose changes, highlighting their sensitivity within the V2V cooperative perception system. Furthermore, when we set the positional parameters $[x, y, z]$ to have maximum GPS bias $[1.118, 1.118, 1.395]$ meters of the actual measured GPS errors in real world [20] and attack Attfuse [1], we obtained a AP 65.1% at IoU 0.5, close to the performance by RBA [5], which has less degradation than our AdvGPS. It demonstrates that the gradient optimization based attack methods are better than the non-gradient attack.

Ablation study: As shown in Table III, our comprehensive ablation study reveals that all three components within our AdvGPS attack framework contribute significantly to the degradation of detection performance on CoBVET [8]. In the context of the *Black-Box Attack* setting, the inclusion of D_{app}, D_{dist}, and D_{task} results in performance degradation by 48.1%, 37.2%, and 27.0%, respectively. This nuanced analysis emphasizes the effectiveness of our design. It is evident that our proposed AdvGPS attacks significantly undermine the performance of state-of-the-art V2V cooperative perception methods.

TABLE II: **Attack results of various GPS parameters under the setting of *Black-Box Attack*.** We show the Average Precision (AP) at IoU=0.5.

Model	Attack	$\hat{x}$	$\hat{y}$	$\hat{z}$	$\hat{\theta}_x$	$\hat{\theta}_y$	$\hat{\theta}_z$
V2X-ViT [6]	PAA [17]	0.676	0.615	0.783	**0.802**	0.802	**0.801**
	AdvGPS	**0.453**	**0.396**	**0.737**	**0.802**	**0.800**	**0.801**
CoBEVT [8]	PAA [17]	0.664	0.609	0.759	**0.784**	0.781	**0.783**
	AdvGPS	**0.450**	**0.377**	**0.732**	0.785	**0.780**	**0.783**

TABLE III: **Ablation study of proposed AdvGPS attack under the setting of *Black-Box Attack*.** The six parameters of CAV GPS pose G_{all} are used to attack CoBVET [8].

D_{app}	D_{dist}	D_{task}	AP@IoU=0.5
			0.904 (No Attack)
✓			0.423 (**-0.481**)
	✓		0.532 (**-0.372**)
		✓	0.634 (**-0.270**)
✓	✓	✓	0.403 (**-0.501**)

3D detection visualization: Under the *Black-Box Attack* setting, we compare the visualizations of different attacks and showcase their impact on CoBEVT [8] in Fig. 4. It is noteworthy that our proposed method achieves the best attack success rate with a notable increase in false-negative and false-positive detection errors, as highlighted in Fig. 4. Additionally, as shown in Fig. 4, we project the attacked point clouds (other colors) from nearby CAVs to the coordinate system of ego vehicle (green color), then we can discover that the attacked point clouds by our AdvGPS has small point cloud shift (similar to random bias), which verifies that our AdvGPS attack is stealthy.

VI. Conclusions

This paper pioneers the investigation of adversarial GPS attacks within the multi-agent cooperative perception systems. Introducing **AdvGPS**, our method generates subtle yet effective adversarial GPS signals that stealthily mislead individual agents, leading to a significant reduction in object detection accuracy. To bolster the effectiveness of the AdvGPS attacks in black-box scenarios, we introduce three statistically sensitive natural discrepancies: appearance-based, distribution-based, and task-aware discrepancies. Extensive experiments conducted on the OPV2V dataset highlight the substantial performance degradation caused by our AdvGPS attacks across various point cloud-based 3D detection systems. This groundbreaking discovery underscores the pressing need to address security concerns within multi-agent perception systems, marking a critical frontier in research.

REFERENCES

[1] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," in *International Conference on Robotics and Automation*. IEEE, 2022, pp. 2583–2589.

[2] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song *et al.*, "V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 712–13 722.

[3] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "Opencda: an open cooperative driving automation framework integrated with co-simulation," in *IEEE International Intelligent Transportation Systems Conference*. IEEE, 2021, pp. 1155–1162.

[4] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, 2023.

[5] Z. Yang, J. Ying, J. Shen, Y. Feng, Q. A. Chen, Z. M. Mao, and H. X. Liu, "Anomaly detection against gps spoofing attacks on connected and autonomous vehicles using learning from demonstration," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[6] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," in *European Conference on Computer Vision*. Springer, 2022, pp. 107–124.

[7] J. Li, R. Xu, X. Liu, J. Ma, Z. Chi, J. Ma, and H. Yu, "Learning for vehicle-to-vehicle cooperative perception under lossy communication," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 4, pp. 2650–2660, 2023.

[8] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "Cobevt: Cooperative bird's eye view semantic segmentation with sparse transformers," in *Conference on Robot Learning*, 2022.

[9] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *International Conference on Robotics and Automation*, 2023.

[10] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *Conference on Robot Learning*. PMLR, 2021, pp. 1195–1210.

[11] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *IEEE International Conference on Robotics and Automation*. IEEE, 2023, pp. 6035–6042.

[12] Y. Hou, Q. Guo, Y. Huang, X. Xie, L. Ma, and J. Zhao, "Evading deepfake detectors via adversarial statistical consistency," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 12 271–12 280.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*. IEEE, 2017, pp. 39–57.

[14] Y. Li, C. Wen, F. Juefei-Xu, and C. Feng, "Fooling lidar perception via adversarial trajectory perturbation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 7898–7907.

[15] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[16] Y. Shi, S. Wang, and Y. Han, "Curls & whey: Boosting black-box adversarial attacks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 6519–6527.

[17] Y. Xu, X. Han, G. Deng, J. Li, Y. Liu, and T. Zhang, "Sok: Rethinking sensor spoofing attacks against robotic vehicles from a systematic view," in *IEEE European Symposium on Security and Privacy*. IEEE, 2023, pp. 1082–1100.

[18] A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola, "A kernel two-sample test," *Journal of Machine Learning Research*, vol. 13, no. 25, pp. 723–773, 2012.

[19] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[20] K.-W. Chiang, G.-J. Tsai, H.-J. Chu, and N. El-Sheimy, "Performance enhancement of ins/gnss/refreshed-slam integration for acceptable lane-level navigation accuracy," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 3, pp. 2463–2476, 2020.

[21] I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *International Conference on Learning Representations*, 2015.

[22] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial intelligence safety and security*. Chapman and Hall/CRC, 2018, pp. 99–112.

[23] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[24] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3d point clouds," in *ACM/IEEE Symposium on Edge Computing*, 2019, pp. 88–100.

[25] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[26] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations.*, 2017.