Stealthy Data Fabrication in Collaborative Vehicular Perception

Qingzhao Zhang
University of Michigan
Ann Arbor, MI, USA
qzzhang@umich.edu

Z. Morley Mao*
University of Michigan
Ann Arbor, MI, USA
zmao@umich.edu

Abstract

Collaborative perception enables multiple connected and autonomous vehicles (CAVs) to collectively perform perception tasks through the efficient exchange of data. It also introduces critical security vulnerabilities due to the potential manipulation of shared data by malicious entities. Existing research demonstrates attacks whereby an adversary could fabricate fake objects or erase real objects from a targeted CAV's perception. Yet, the practicality of such attacks as a realistic threat remains inadequately addressed. Firstly, current attacks have not been refined to circumvent established anomaly detection frameworks. Secondly, the demonstration of attack effectiveness predominantly relies on manually defined scenarios, raising questions about the feasibility of such attacks in dynamic, real-world situations. To address these shortcomings, our research revisits data fabrication in collaborative perception and introduces a novel attack methodology that is realistic, stealthy, and scenario-aware. This approach aims to minimize required data perturbations and exploits error propagation within the autonomous driving software pipeline to trigger critical safety hazards. Our proposed attack encompasses a comprehensive end-to-end workflow, determining attack strategies based on dynamic environmental conditions at runtime. Through high-fidelity simulations, we demonstrate the efficacy of our proposed attack, underscoring its potential to significantly undermine existing defense mechanisms.

CCS Concepts

• Security and privacy → Systems security; **• Computer systems organization → Embedded and cyber-physical systems**.

Keywords

autonomous driving; adversarial machine learning

ACM Reference Format:
Qingzhao Zhang and Z. Morley Mao. 2024. Stealthy Data Fabrication in Collaborative Vehicular Perception. In *Proceedings of the Sixth Workshop on CPS&IoT Security and Privacy (CPSIoTSec'24), October 14–18, 2024, Salt Lake City, UT, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3690134.3694822

1 Introduction

One fundamental limitation of the perception systems of connected and autonomous vehicles (CAVs) is that the onboard sensors have limited sensing capabilities, especially when the target is occluded, far away, or affected by adverse weather [5, 6, 16, 30, 44, 46]. To address the limitation, collaborative perception is proposed, where CAVs share sensing data (e.g., raw sensor data or processed data) and then process perception tasks on fused data, resulting in improved perception accuracy and spatial coverage. The technology has attracted extensive academic research [18, 49, 52, 58] and has been adopted by CAV industry players [1–3, 7, 9, 10, 53]. As CAVs with collaborative perception will leverage external data from untrusted parties to assist their local perception, there is the security concern that a malicious party may act like a participant in collaboration but falsify the messages to share in order to compromise the quality of perception on the target CAV. Assuming the existence of at least one malicious CAV, recent studies proposed attacks based on adversarial machine learning (AML) resulting in incorrect perception results such as spoofed ghost objects and ignored real objects [48, 56]. Defense methods are also proposed [32, 56], basically leveraging the inconsistencies of perception results on different benign vehicles.

However, we argue that the real-world threat of data fabrication in collaborative perception is not adequately evaluated. The reason is twofold. Firstly, such attacks are possible to be enhanced to bypass existing defense methods. Existing defense approaches heavily rely on manually tuned heuristics and thresholds thus they suffer from false positives and false negatives. For instance, CAD [56] compares fused perception results against raw sensor data from local trusted sensors, which may occasionally fail because of real-world noise, inaccuracies of data processing, and certain hard scenarios (e.g., none of the benign CAVs can observe the region affected by the attack). Secondly, existing attacks do not involve an end-to-end solution. For instance, Tu et. al. [47] proposed an untargeted attack injecting perception errors at random spatial locations, and the attack from Zhang et. al. [56] can spoof or remove objects at a specified target location. Neither of the attacks involves deciding when and where to launch the attack in order to construct a complete attack scenario to trigger safety hazards, which is especially challenging in real-world dynamic traffic. The attack would be a concrete threat only if the end-to-end exploitation is reproducible.

To bridge the gap, we build an end-to-end attack that is stealthy and scenario-aware. To keep stealthy against existing anomaly detection looking for inconsistencies, the attack tries to minimize the perturbation on object locations so that the detectable inconsistency is almost indistinguishable from normal noises and also below the threshold of anomaly detection. As the small perturbation can hardly cause safety hazards directly, the attack leverages the small perturbation to misguide the downstream data processing

[*] Also with Google.

CPSIoTSec'24, October 14–18, 2024, Salt Lake City, UT, USA

ACM ISBN 979-8-4007-1244-9/24/10
https://doi.org/10.1145/3690134.3694822

tasks. For instance, the inaccuracy of object detection could trigger failures in object tracking and trajectory prediction. In this way, the initial error eventually becomes a significant error that could trigger improper driving decisions. Following the philosophy, we design the following attack workflow. The attacker participates in the collaboration and uses its onboard sensors to sense the surrounding environment. The attacker localizes the victim CAV to attack and prepares an attack strategy, for instance, perturbing the location of existing vehicles to fake lane-changing behaviors and trigger unsafe hard brakes. The attacker then periodically launches perception attacks to realize the selected attack strategy and meanwhile updates the strategy according to the dynamic traffic situations.

We evaluate the attack's impact on the OPV2V multi-vehicle perception dataset [52]. The attack significantly increases the error in the downstream trajectory prediction module by 201%, leading to incorrect behavioral estimations of other vehicles by the victim. This results in hard braking in 26%-28% of test cases. Additionally, our ablation study investigates the influence of autonomous driving components (e.g., object detection, tracking, and prediction) on the attack's success rate. This analysis informs future efforts to mitigate such vulnerabilities in collaborative perception systems.

Our contributions include:

- We design a new attack on collaborative perception models that can accurately manipulate locations of detected objects.
- We design an end-to-end online attack against collaborative perception, which arranges perception attacks in dynamic traffic scenarios to trigger safety hazards on a victim vehicle.

2 Background and Related Work

2.1 Collaborative Perception

Collaborative perception has been proposed to enhance Connected and Autonomous Vehicle (CAV) perception [11, 28, 30, 33, 44], facilitating the sharing of sensor data among infrastructure or vehicles. Mainstream solutions predominantly utilize LiDAR sensors due to the rich 3D geometry features provided by LiDAR images. Collaborative perception can be classified into three major types. In early-fusion sharing schemes [17, 19, 29, 39, 57, 58], CAVs exchange raw sensor data in a universal format that can be easily concatenated, though it comes at the cost of high data transmission bandwidth. Intermediate-fusion schemes [18, 20, 49, 51, 54] involve transmitting feature maps, which are intermediate products of perception algorithms, offering a balance between network efficiency and perception accuracy. In late-fusion schemes [35, 43, 45], lightweight perception results such as bounding boxes are shared.

2.2 Attacks on CAV Perception

LiDAR perception systems in CAVs are susceptible to several types of attacks. Physical attacks, such as GPS spoofing [34, 42], LiDAR spoofing [15, 24, 26, 34], and physically realizable adversarial objects [47, 55, 59], target individual autonomous vehicles. Late-fusion collaborative perception, which shares object locations [21–23, 41], can be compromised by attackers modifying these locations [13, 14, 27, 38]. Tu *et al.* [48] introduced the first attack specific to intermediate-fusion collaborative perception, an untargeted adversarial attack that creates inaccurate detection bounding boxes by perturbing feature maps. Zhang *et al.* [56] proposed an advanced targeted attack for early-fusion and intermediate-fusion systems, capable of spoofing or removing objects in specific locations, and reproducible in real-time on-vehicle devices.

These existing attacks primarily focus on reducing the accuracy of perception systems, without considering the implications for safety in complex dynamic traffic scenarios. We address this gap by proposing the first end-to-end scenario-aware attack workflow.

2.3 Defenses on CAV Perception

Various anomaly detection methods have been proposed to counter sensor attacks [12, 25, 36, 37, 40, 46]. Specifically for LiDAR systems, CARLO [46] detects abnormal point clouds that violate occlusion features, and LIFE [36] detects temporal and sensor-fusion inconsistencies. In connected vehicle applications, efforts to model benign behaviors of ego/remote vehicles and detect model outliers as anomalies include various aspects such as temporal consistency [14], physical constraints on message delivery or vehicle control [13, 27], and cross-validation with local sensors [38]. Against the latest adversarial attacks, CAD [56] leverages the spatial sensing from all connected vehicles to jointly reveal inconsistencies by sharing occupancy maps. AmongUs [32] is a consensus-based solution that repeatedly samples a subset of collaborating CAVs until their perception results converge, using this subset for collaborative perception.

Overall, existing defenses rely on detecting inconsistencies between clean and attacked perception data to identify attackers. However, the detection is less effective if benign CAVs have limited sensing capabilities and inaccuracy occurs in data processing, leaving room for strong adaptive attacks. This paper proposes a new attack that remains stealthy under existing defenses.

3 Problem Statement

We elaborate the threat model in §3.1, the limitation of existing work in §3.2, and define design goals in §3.3.

3.1 Threat Model

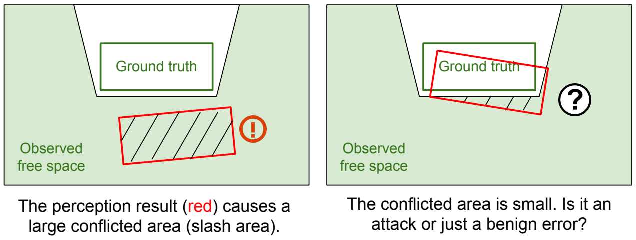
We assume a Vehicle-to-Vehicle (V2V) scenario but our results can be easily generalized to vehicle-to-infrastructure (V2I) settings by replacing one or more vehicles with edge computing devices.

We assume that the attacker can physically control at least one vehicle participating in collaborative perception. This control grants the attacker privileges over the vehicle's software and hardware, allowing them to manipulate sensors, tamper with the local execution of algorithms, and transmit arbitrary data through the network. In other words, attackers can directly alter the shared data. We also assume the attacker has white-box access to the autonomous driving software stack, including deep learning models of object detection and trajectory prediction.

We do not assume the attacker has prior knowledge of upcoming traffic scenarios. Instead, the attacker gains knowledge about the traffic solely through predefined static maps and the vehicle's real-time onboard perception system.

3.2 Limitation of Existing Attacks

Existing attacks mentioned in §2 have gaps towards realistic strong attacks in stealthiness against anomaly detection and end-to-end scenario construction. We elaborate on the two aspects as follows.



The perception result (red) causes a large conflicted area (slash area).

The conflicted area is small. Is it an attack or just a benign error?

Figure 1: An illustration of existing anomaly detection [56]. It is hard to distinguish malicious attacks and benign perception errors when the conflicted area is geometrically small.

Stealthiness against anomaly detection. The state-of-the-art defenses [32, 56] use anomaly detection to score data consistency across CAVs, raising alerts when inconsistencies exceed a threshold set to balance false positives and missed detections. Taking CAD [56] as an example, the inconsistency level refers to the area of conflicting regions in the 2D bird-eye view, and the threshold is around 0.8 m^2. Existing attacks, either randomly inject faults [48] or spoofing/removing vehicles in a specific region [56], trigger detection due to exceeding this threshold.

However, we found that if the attacker creates a minor perturbation that is below the threshold, it is undetected and is hard to distinguish from benign perception faults, as demonstrated in Figure 1. If carefully crafted, these small errors can disrupt downstream components like object tracking and prediction, leading to significant safety-critical driving errors. Our proposed attack exploits this by focusing on small yet effective perturbations.

End-to-end scenario construction. The impact of perception attacks varies by scenario. Previous work [56] showed attacks can induce unsafe decisions during lane changes or unprotected turns. However, without prior knowledge of traffic, it is unclear how attackers schedule their attacks in dynamic traffic. To succeed, attackers must first estimate traffic flow and then decide when, where, and how to manipulate perception to maximize impact. This crucial step, essential for end-to-end collaborative perception attacks, has been overlooked by previous methods.

In this paper, we build the first end-to-end attack workflow exploiting the data fabrication in collaborative perception. It is an automated algorithm where the attacker searches for an optimal plan to trigger unsafe driving behaviors of a remote CAV.

3.3 Towards Adaptive and Realistic Adversary

We propose a new attack against collaborative perception which should satisfy the following requirements.

- Effectiveness. The attack should be able to trigger unsafe behaviors of CAVs, such as sudden brakes or risk of collisions.
- Stealthiness. The attack should be hardly detected by existing anomaly detection [32, 56].
- No prior knowledge. The attacker recognizes on-road traffic by onboard sensors and determines the attack strategy at runtime.

4 Attack Methodology

In this section, we introduce our design of an end-to-end scenario-aware stealthy attack against collaborative perception and explain how the new attack achieves design goals in §3.3.

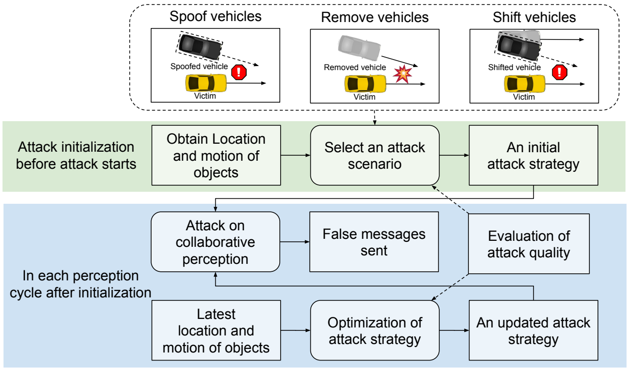

Spoof vehicles
Spoofed vehicle
Victim
Remove vehicles
Removed vehicle
Victim
Shift vehicles
Shifted vehicle
Victim
Attack initialization before attack starts
Obtain Location and motion of objects
Select an attack scenario
An initial attack strategy
Attack on collaborative perception
False messages sent
Evaluation of attack quality
In each perception cycle after initialization
Latest location and motion of objects
Optimization of attack strategy
An updated attack strategy

Figure 2: Overview of the attack methodology. Only attack scenarios of shifting vehicles is performed in this paper.

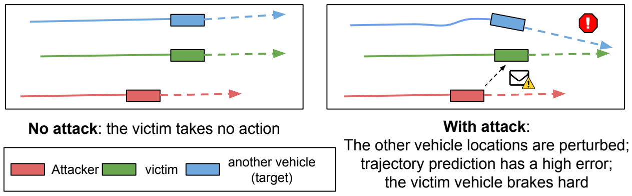

No attack: the victim takes no action
🟥 Attacker 🟩 victim 🟦 another vehicle (target)
With attack:
The other vehicle locations are perturbed;
trajectory prediction has a high error;
the victim vehicle brakes hard

Figure 3: Illustraion of "shift to move in" attack scenario.

4.1 Overview

We depict the overview of the attack in Figure 2. Our proposed attack is a systematic integration of various attack techniques.

Perception attack. The perception attack on collaborative perception models is the key to influence the victim vehicle. Validated by prior work [48, 56], an attacker CAV can join the system of collaborative perception, send crafted falsified messages to a designated victim CAV, and inject false detection at certain locations. In this work, we improve the prior attack to make it sophisticated in controlling locations of false detection.

Scenario-aware attack strategies. The attacker needs to know when and where to launch the perception attacks. One attack strategy in our attack is the scheduling of a series of perception attacks over continuous frames. More importantly, such an attack strategy depends on the traffic scenarios. The attacker can either spoof fake vehicles cutting in the lane of the victim CAV to force it to stop, remove vehicles around the victim CAV to make it ignore potential risks, or inject errors on the locations of detected CAVs to influence the victim CAV's decisions.

Online optimization of attack strategies. As the attacker cannot get complete knowledge of the dynamic on-road traffic, the attacker needs to periodically optimize the attack strategy based on the latest sensing results. It aims to maximize the quality of attack considering attack effectiveness, stealthiness, and realizability.

Workflow. The attacker controls a CAV to join the network of collaborative perception. The attacker uses its onboard sensors to sense the surrounding traffic and obtains an estimation of previous and future trajectories of on-road objects through the processing of a pipeline of object detection, tracking, and prediction. The attacker selects an initial attack strategy based on known information and launches the attack. In each following frame of perception, the attacker executes a perception attack following the attack strategy and also optimizes the strategy based on new information in the latest frame. In this way, the attacker intelligently arranges a series of perception attacks to achieve attack goals.

4.2 Adversarial Attack on Perception Models

Zhang *et al.* [56] has proposed adversarial attacks to spoof or remove objects from the perception results. However, accurately modifying object locations by a small distance (*e.g.,* <0.5 meter) is challenging, and requires sophisticated control of the perturbation. To bridge the gap, we propose the "object shifting attack".

We first define the problem of the adversarial attack. We denote LiDAR data at frame $i \in \mathbb{N}$ from the attacker, the victim, and other benign vehicles by A_i, V_i, and $X_i^{(j)}$, $j \in \{0, 1, \dots N\}$, respectively. LiDAR data with the same frame index will be merged on the victim side to generate perception results. We denote pre-process before data sharing as f and post-process after data sharing as g. A normal collaborative perception for the victim on frame i can be described as $y_i = g(f(V_i), f(A_i), f(X_i^0), f(X_i^1), ..., f(X_i^N))$. The attacker can replace $f(A_i)$ by malicious data. For instance, the attacker can append a minor perturbation δ_i to craft malicious data as $f(A_i) + \delta_i$, which will change the original perception result from y_i to $y_i' = g(f(V_i), f(A_i) + \delta_i, f(X_i^0), f(X_i^1), ..., f(X_i^N))$.

However, when the attacker optimizes δ_i, other data in the same frame such as $f(V_i)$ is not available because of the delay of communication in the real system, which is well discussed in the previous work [56]. Therefore, the attacker uses dated data at frame $i-1$ to optimize the attack at frame i.

In **Early-fusion** system, the attacker crafts a malicious point cloud. We design the shifting attack as a series connection of the removal attack and the spoofing attack proposed in the previous work [56]. Whenever there is a conflict between removal and spoofing attacks, *e.g.,* the same point is modified by two attacks, we prioritize the object spoofing. In general, the removal attack first introduces noise in the point cloud to lower the original detection confidence, and then the spoofing attack introduces the new object that is shifted from the original location.

In **intermediate-fusion** system, $f(A_i)$ is a feature map. We define δ_i as a patch that perturbs the feature values in a specific region of the feature map, and the region is associated with the geometry location of the target object to be shifted, following Zhang *et al.* [56]. In particular, we introduce a new loss function that is suitable for shifting detection boxes and enables universal adversarial perturbation to improve success rate. The optimization problem is:

$$\begin{aligned} \min \sum_{1 \le i \le N} \sum_{\substack{(z, z_\sigma) \in g(f(A_i)+\delta, \cdot), \\ IoU(z, z_t) > 0, \\ z_\sigma > \epsilon_\sigma}} & \log(1 - IoU(z, z_t)) - C \cdot z_\sigma \\ \text{s.t. } |\delta| < \epsilon & \end{aligned} \tag{1}$$

where δ is the perturbation on the feature map, z and z_σ represent individual detection proposals and confidence scores, and z_t is the target location we want the object to move to. The optimization tries to obtain a perturbation δ that is universally effective on multiple consecutive history frames, and the number of frames is denoted by N. δ is then applied to the next frame to trigger the attack. On each history frame, the optimization selects promising proposals that are geometrically close to the target object, determined by the Intersection over Union (IoU) function, and have a high confidence score, determined by threshold ϵ_σ. It then optimizes δ to maximize the IoU with the target location of these promising proposals, which is the main objective, and also penalizes the decrement of their confidence scores to ensure these proposals are remained after the processing of the Non-Maximum Selection (NMS) stage. The optimization uses Projected Gradient Descent (PGD).

In **late-fusion** system, vehicles directly share bounding boxes and confidence scores. The attack is straightforward: the attacker simply modifies the location of the target object and fakes a high confidence score, to overwrite correct bounding boxes from other benign vehicles during the NMS process.

4.3 Scenario-aware Attack Strategy

An attack strategy plans a sequence of perception attacks to trigger safety hazards. This paper focuses on the "shift to move in" strategy, where object shifting causes the victim to mistakenly predict a nearby vehicle cutting into their lane, leading to a hard brake (see Figure 3). Analysis of other strategies is left for future work. A typical attack strategy consists of the following elements:

- An identified victim vehicle.
- A selected target vehicle whose location will be perturbed.
- An adversarial trajectory representing the perturbed trajectory of the target vehicle.

With the attack strategy, the attacker tries to inject the adversarial trajectory into the perception results of the victim by perception attacks. Assuming the perception attack could be successful, the attack impact is determined by the quality of the malicious trajectory. To this end, we design a fitness function to rate the quality. We consider three aspects of quality:

- **Effectiveness**. The adversarial trajectory should be able to trigger the safety hazard. In our case, the victim would make an unsafe hard brake if the prediction of the target object's future trajectory overlaps with the victim's planned trajectory.
- **Stealthiness**. The adversarial trajectory cannot have too much conflict that is over the threshold of the anomaly detection methods. In our case, we calculate the overlap area of the estimated visible space of the victim with the fake objects. The threshold is pre-computed by benchmarking the anomaly detection methods.
- **Realizability**. The adversarial trajectory must be achievable by the perception attacks on perception models. For instance, the trajectory must be covered by the attacker's feature map.

An example fitness function implementation considering the above three aspects is shown in Algorithm 1.

4.4 End-to-end Online Attack

As the on-road traffic is highly dynamic, the attacker is supposed to periodically update the attack strategy according to its latest sensing. We introduce the end-to-end attack in Algorithm 1.

The attacker initially identifies the victim vehicle and uses its local autonomous driving software stack to track on-road vehicles. Assuming that the attack will last for K frames, before launching attacks, the attacker predicts the trajectories of these vehicles in these K future frames. At this moment, the attacker generates a set of attack strategy candidates and picks the top candidate based

Algorithm 1: End-to-end online attack algorithm.

```
Input: The attacker identified a victim vehicle $p$. The attack lasts
       for $K$ frames and optimizes attack impact on frames
       $K - M + 1, ..., K$. The attacker maintains observed or
       predicted trajectories of other vehicles, denoted by a set $T$
       where $T_i$ is the trajectory of vehicle index $i$ in a length of $K$.
Function OnlineAttack():
    $T_{adv} \leftarrow \text{argmax}_{T_i \in T}$ AttackQuality($T_i$); ▹ Choose a strategy;
    for Frame $i = 1, ..., K$ do
        PerceptionAttack($T^{i}_{adv}$);
        $T \leftarrow$ LocalPerception($T$);
                    ▹ Update trajectories by latest sensing;
        $T^{1:K}_{adv} \leftarrow$ PGDUpdate($T_{adv}, i : K$);
    end
Function AttackQuality($T_{adv}$):
    $l_e \leftarrow -\sum_{K-M+1 \leq i \leq K} \log$ Distance(Prediction($T^{1:i}_{adv}$),
            Prediction($T^{1:i}_{p}$));    ▹ Attack effectiveness score;
    $l_s \leftarrow AnomalyDetection(T_{adv})$;      ▹ Stealthiness score;
    $l_r \leftarrow realizabilityCheck(T_{adv})$;     ▹ Realizability score;
    if $l_s \leq 0 \vee l_r \leq 0$ then
        return 0;
    end
    return $l_e$;
Function PGDUpdate($T_{adv}$, $i{:}j$):
    $\lambda \leftarrow \max\{\lambda \mid$ AttackQuality$(\lambda(T_{adv} - T_q) + T_q) > 0\}$;
    $T_{adv} \leftarrow \lambda(T_{adv} - T_q) + T_q$;
    return AdamOptimizer(AttackQuality($T_{adv}$), $T^{i:j}_{adv}$);
```

on the fitness function for attack quality. For a "shift to move in" attack, for instance, the attacker intends to perturb the location of the vehicle which is very close to the victim.

The attacker then launches the K-frame attack. In each frame, the attacker first launches a perception attack to shift the target object according to the current attack strategy. Meanwhile, the attacker leverages the latest sensing results to update its knowledge of the trajectories of surrounding vehicles. Based on the latest knowledge, the attacker applies Projected Gradient Descent (PGD) algorithm to optimize the adversarial trajectory to maximize attack quality. In particular, the optimization focuses on maximizing the effectiveness score while maintaining the stealthiness and realizability of the attack. Note that only the future frames of the adversarial trajectory are updated, as the perception attacks on previous frames have been finalized. Eventually, the attacker finalized a K-frame adversarial trajectory that would invoke a high error on the victim vehicle's trajectory prediction model and simultaneously launched perception attacks to realize the trajectory.

4.5 Discussion of mitigation

The attack relies on the adversarial attacks on either perception or prediction models thereby adversarial robustness improvements such as data augmentation, adversarial training, and certified learning are potentially beneficial. Vehicular reputation systems could also be crucial for identifying misbehaving vehicles among the population. We aim to reveal the potential vulnerabilities for future research, especially on defense solutions.

Table 1: Performance of object shifting attacks.

Target system	Method	Success	IoU		Score	
			Before	After	Before	After
Early-fusion	Ray casting	45.4%	0.22	0.32	0.53	0.21
Intermediate-fusion	Adversarial ML	87.1%	0.31	0.49	0.80	0.82
Late-fusion	Naive	100%	0.29	0.95	0.59	0.99

5 Evaluation

We evaluate the effectiveness of our proposed attack on simulated multi-vehicle sensor data, involving the object-shifting perception attack and the end-to-end online attack.

5.1 Implementation

We use the *OPV2V* [52] dataset for evaluation, an multi-vehicle dataset generated through a co-simulation of the CARLA simulator [4] and SUMO simulator [8]. We randomly selected 46 attack scenarios, each consisting of 60 frames (6 seconds).

We implement a typical autonomous driving software stack as the subject of the attack. It is a pipeline of various collaborative perception models from OpenCOOD [52], multi-object tracking *AB3DMOT* [50], and trajectory prediction *GRIP++* [31]. All models are trained on a separate training set of OPV2V.

We implement the attack algorithm in Python. For the 60-frame scenario, the attacker executes the attack from the 20-th frame to the 40-th frame ($K = 20$) and optimizes attack impact on the 40-th frame ($M = 1$). The object-shifting attack against intermediate-fusion leverages 5 steps of PGD update for each frame and aggregates 3 frames in each optimization step. In the fitness function of attack quality, the threshold of stealthiness is 0.8 m^2 conflicted area.

In our experiments, we feed the data from *OPV2V* to our built autonomous driving software stack to simulate the scenarios. The attack algorithm manipulates the input of autonomous driving.

5.2 Performance of Perception Attack

We measure the effectiveness of our proposed object shifting attack, including the three variants for early-fusion, intermediate-fusion, and late-fusion collaborative perception systems respectively, as introduced in §4.2. We picked 300 attack tasks from OPV2V by randomly selecting the attacker, the victim, and the target object, and the attack goal is to move the target object towards a random direction by 1 meter in the victim's perception results. The attack results are summarized in Table 1. We regard the attack as successful when the detection bounding box has a larger Intersection over Union (IoU) with the target object location than the original location. Besides an overall success rate, we extract the bounding box associated with the target object before and after the attack and average the IoU with the target location and the confidence score. An effective attack should show a significant increment of IoU and maintain a reasonably high score after the attack.

From the results, the late-fusion attack is the easiest to succeed, achieving a 100% success rate and nearly perfect IoU and score. This is reasonable as the attacker can freely inject fake bounding boxes with high confidence scores. Intermediate-fusion attack is considerably effective and is successful in 87.1% of cases. The average IoU of 0.49 is also considered accurate by the computer vision community. Early-fusion attacks are less successful because of a fundamental challenge that the attacker cannot overwrite any LiDAR points from other vehicles and the malicious point clouds

Table 2: The impact of "shift to move in" attack.

	No attack	Ideal attack	w/ tracking	w/ late-fusion	w/ intermediate-fusion
Avg(ADE)	1.83	7.38	5.77	5.34	5.67
Avg(MinDist)	5.37	2.34	3.61	3.56	4.21
%(MinDist<3)	6.5%	80.4%	39.1%	38.3%	26.7%

are constrained by physical laws. In conclusion, with the current attack methodology, the perception attack is stably successful on intermediate-fusion and late-fusion systems.

5.3 End-to-end performance of the Attack

In this section, we launch an end-to-end full attack. In the case of "shift to move in" attacks. To validate the attack's impact on realistic driving scenes, we evaluate how the attack affects the downstream components of the victim vehicle including trajectory prediction and motion planning. To this end, we introduce three evaluation metrics as below. Formally speaking, for a trajectory T, we denote the trajectory point at frame i as $T^{(i)}$, and l_2 denotes L2 distance.

- $\texttt{ADE}(T_{gt}, T_p) = (\frac{1}{N}\sum_{i=1...N} l_2(T_{gt}^{(i)}, T_p^{(i)}))^{\frac{1}{2}}$. Average Displacement Error (ADE) between the ground-truth target object trajectory T_p and predicted target object trajectory T_{gt}.
- $\texttt{MinDist}(T_p, T_v) = \min_{i=1...N} l_2(T_p^{(i)}, T_v^{(i)})^{\frac{1}{2}}$. Minimum distance between predicted target object trajectory T_p and the victim vehicle trajectory T_v.
- %(MinDist<3) is the percentile of the cases where MinDist < 3 meters, which we consider a close distance the victim vehicle should brake to react.

Either a large prediction error or wrong estimation of the distance to other vehicles has a critical negative impact on driving safety.

Results are shown in Table 2. Overall, the attack can significantly increase prediction ADE by 201%, decrease MinDist by 27%, and effectively influence the victim's behavior in 26%-38% of cases. As an ablation study, we also evaluate the attack results when certain components are disabled. The ideal attack assumes the attack strategy is perfectly executed and the victim vehicle sees the same target object trajectory as adversarial trajectory defines. In this ideal case, %(MinDist<3) is 80% which means most cases make an impact on the victim's behavior except for hard cases where the victim has no surrounding vehicles nearby. However, as the adversarial trajectory is optimized using only the trajectory prediction model, the attack is less effective when the victim vehicle applies object tracking to refine observed trajectories, causing the increment of MinDist. More uncertainties are introduced when using the real perception attacks, and %(MinDist<3) dropped to 38.3% and 26.7% on early-fusion and intermediate-fusion systems respectively. The results demonstrate the challenge of realistic attacks where the attackers have partial knowledge of the scene.

We then apply existing anomaly detection methods CAD [56] and ROBOSAC [32] on attack scenarios. CAD detects 13% of intermediate-fusion attacks and 4.5% of late-fusion attacks, while ROBOSAC detects 4.5% and 2.2% of attacks respectively. CAD relies on the detection of "conflicted areas" as discussed in §3.2 but our attack restricts the "conflicted areas" to be below the detection threshold. CAD has a higher detection rate on intermediate-fusion attacks because of the uncertainty of the perception attack, which may not realize the attack strategy exactly. ROBOSAC relies on a consensus algorithm and focuses on a threat model where the attacker blindly manipulates all objects and introduces lots of false positives. It is not optimized for the single-object targeted perception attack.

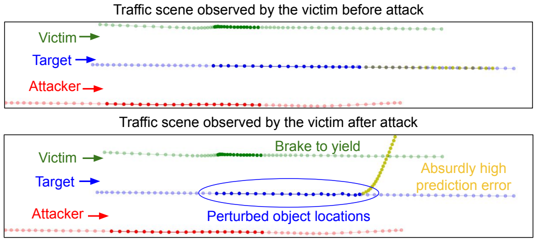

Traffic scene observed by the victim before attack
Victim
Target
Attacker
Traffic scene observed by the victim after attack
Victim
Brake to yield
Target
Absurdly high prediction error
Attacker
Perturbed object locations

Figure 4: A case study of the "shift to move in" attack.

Figure 4 demonstrates an example of the attack on the late-fusion system. When the target vehicle and the victim vehicle are driving alongside, the small perception error on the locations of the target vehicle eventually causes a high error of trajectory prediction and improper driving decisions.

5.4 Computational Overhead

We benchmark attack algorithms on a server with an Intel Xeon Silver 4110 CPU and NVIDIA RTX 2080Ti. The perception attack takes 82 ms, 69 ms (one-step PGD), and 0.4 ms on early-, intermediate-, and late-fusion systems, respectively. Attack optimization time is linear to the adversarial trajectory length, requiring 13 ms for a 20-frame trajectory and 2 ms for the final frame.

In a realistic attack, perception and optimization run concurrently within a typical LiDAR cycle (100 ms). Early- and late-fusion attacks meet this constraint easily, while intermediate-fusion can only complete one PGD iteration. Future work could enhance attacks via parallelization and caching.

6 Limitations and Future Work

It is an in-progress work to enhance the results presented in this paper. Firstly, we aim to extend the scenario-aware attack to consider more complex attack strategies, which could apply object spoofing, removal, and shifting on multiple objects to trigger an intrinsic safety hazard. Secondly, the object shifting attack is not stably successful on early-fusion collaborative perception systems and it is a future work to explore advanced attack methods. Thirdly, the intermediate-fusion object shifting attack is expensive in computation, which requires further optimization. We also aim to validate the mitigation methods as discussed in §4.5.

7 Conclusion

We propose a stealthy, scenario-aware attack on vehicular collaborative perception that induces unsafe driving behaviors with partial knowledge of dynamic traffic scenes. The attack exploits small perception errors, amplified by downstream components like trajectory prediction, revealing new security challenges in autonomous driving systems.

Acknowledgments

We would like to thank our anonymous reviewers for their valuable comments and feedback. This work was supported in part by NSF under CNS-2321532, CMMI-2038215, and CNS-2112562.

References

[1] 2017. Qualcomm C-V2X. https://www.qualcomm.com/news/releases/2017/09/qualcomm-announces-groundbreaking-cellular-v2x-solution-support-automotive.

[2] 2019. Huawei C-V2X. https://carrier.huawei.com/en/products/wireless-network-v3/Components/c-v2x.

[3] 2019. Infineon C-V2X. https://www.infineon.com/dgdl/Infineon-ISPN-Use-Case-Savari-Securing-V2X+communications-ABR-v01_00-EN.pdf?fileId=5546d462689a790c0168e1c1f5e35221.

[4] 2021. *CARLA: Open-source simulator for autonomous driving research.*

[5] 2022. Autoware: Open-source software for self-driving vehicles. https://github.com/Autoware-AI.

[6] 2022. Baidu Apollo. http://apollo.auto.

[7] 2022. Bosch C-V2X. https://www.bosch-mobility-solutions.com/en/solutions/connectivity/v2x-connectivity-solutions-cv/.

[8] 2022. SUMO: Simulation of Urban Mobility. https://www.eclipse.org/sumo/.

[9] 2023. Automotive Edge Computing Consortium. https://aecc.org/.

[10] 2023. Ford AV Dataset. https://aecc.org/.

[11] Simegnew Yihunie Alaba and John E Ball. 2022. A survey on deep-learning-based lidar 3d object detection for autonomous driving. *Sensors* 22, 24 (2022), 9577.

[12] Khattab M Ali Alheeti, Abdulkareem Alzahrani, and Duaa Al Dosary. 2022. LiDAR Spoofing Attack Detection in Autonomous Vehicles. In *2022 IEEE International Conference on Consumer Electronics (ICCE)*. IEEE, 1–2.

[13] Srivalli Boddupalli, Ashwini Hegde, and Sandip Ray. 2021. Replace: Real-time security assurance in vehicular platoons against v2v attacks. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 1179–1185.

[14] Srivalli Boddupalli and Sandip Ray. 2020. Redem: Real-time detection and mitigation of communication attacks in connected autonomous vehicle applications. In *Internet of Things. A Confluence of Many Disciplines: Second IFIP International Cross-Domain Conference, IFIPIoT 2019, Tampa, FL, USA, October 31–November 1, 2019, Revised Selected Papers 2*. Springer, 105–122.

[15] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 2021. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *2021 IEEE Symposium on Security and Privacy (SP)*. IEEE, 176–194.

[16] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. 2019. Adversarial sensor attack on lidar-based perception in autonomous driving. In *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*. 2267–2281.

[17] Hanlin Chen, Brian Liu, Xumiao Zhang, Feng Qian, Z Morley Mao, and Yiheng Feng. 2022. A Cooperative Perception Environment for Traffic Operations and Control. *arXiv preprint arXiv:2208.02792* (2022).

[18] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song Fu. 2019. F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds. In *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*. 88–100.

[19] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. 2019. Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds. In *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 514–524.

[20] Jiaxun Cui, Hang Qiu, Dian Chen, Peter Stone, and Yuke Zhu. 2022. COOPERNAUT: End-to-End Driving with Cooperative Perception for Networked Vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17252–17262.

[21] Jorge Godoy, Víctor Jiménez, Antonio Artuñedo, and Jorge Villagra. 2021. A grid-based framework for collective perception in autonomous vehicles. *Sensors* 21, 3 (2021), 744.

[22] Hendrik-Jörn Günther, Björn Mennenga, Oliver Trauer, Raphael Riebl, and Lars Wolf. 2016. Realizing collective perception in a vehicle. In *2016 IEEE Vehicular Networking Conference (VNC)*. IEEE, 1–8.

[23] Mohamed Hadded, Pierre Merdrignac, Sacha Duhamel, and Oyunchimeg Shagdar. 2020. Security attacks impact for collective perception based roadside assistance: A study of a highway on-ramp merging case. In *2020 International Wireless Communications and Mobile Computing (IWCMC)*. IEEE, 1284–1289.

[24] R Spencer Hallyburton, Yupei Liu, Yulong Cao, Z Morley Mao, and Miroslav Pajic. 2022. Security Analysis of {Camera-LiDAR} Fusion Against {Black-Box} Attacks on Autonomous Vehicles. In *31st USENIX Security Symposium (USENIX Security 22)*. 1903–1920.

[25] Zhongyuan Hau, Soteris Demetriou, Luis Muñoz-González, and Emil C Lupu. 2021. Shadow-catcher: Looking into shadows to detect ghost objects in autonomous vehicle 3d sensing. In *Computer Security–ESORICS 2021: 26th European Symposium on Research in Computer Security, Darmstadt, Germany, October 4–8, 2021, Proceedings, Part I 26*. Springer, 691–711.

[26] Zizhi Jin, Ji Xiaoyu, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. 2022. PLA-LiDAR: Physical Laser Attacks against LiDAR-based 3D Object Detection in Autonomous Vehicle. In *2023 IEEE Symposium on Security and Privacy (SP)*. IEEE Computer Society, 710–727.

[27] Hyogon Kim and Taeho Kim. 2019. Vehicle-to-vehicle (V2V) message content plausibility check for platoons through low-power beaconing. *Sensors* 19, 24 (2019), 5493.

[28] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven Waslander. 2018. Joint 3D Proposal Generation and Object Detection from View Aggregation. *IROS* (2018).

[29] Swarun Kumar, Lixin Shi, Nabeel Ahmed, Stephanie Gil, Dina Katabi, and Daniela Rus. 2012. Carspeak: a content-centric network for autonomous driving. *ACM SIGCOMM Computer Communication Review* 42, 4 (2012), 259–270.

[30] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. 2019. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12697–12705.

[31] Xin Li, Xiaowen Ying, and Mooi Choo Chuah. 2019. Grip++: Enhanced graph-based interaction-aware trajectory prediction for autonomous driving. *arXiv preprint arXiv:1907.07792* (2019).

[32] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. 2023. Among us: Adversarially robust collaborative perception by consensus. *arXiv preprint arXiv:2303.09495* (2023).

[33] You Li and Javier Ibanez-Guzman. 2020. Lidar for autonomous driving: The principles, challenges, and trends for automotive lidar and perception systems. *IEEE Signal Processing Magazine* 37, 4 (2020), 50–61.

[34] Yiming Li, Congcong Wen, Felix Juefei-Xu, and Chen Feng. 2021. Fooling lidar perception via adversarial trajectory perturbation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7898–7907.

[35] Hansi Liu, Pengfei Ren, Shubham Jain, Mohannad Murad, Marco Gruteser, and Fan Bai. 2019. FusionEye: Perception Sharing for Connected Vehicles and its Bandwidth-Accuracy Trade-offs. In *2019 16th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 1–9.

[36] Jinshan Liu and Jung-Min Park. 2021. "Seeing is Not Always Believing": Detecting Perception Error Attacks Against Autonomous Vehicles. *IEEE Transactions on Dependable and Secure Computing* 18, 5 (2021), 2209–2223.

[37] Shinan Liu, Xiang Cheng, Hanchao Yang, Yuanchao Shu, Xiaoran Weng, Ping Guo, Kexiong Curtis Zeng, Gang Wang, and Yaling Yang. 2021. Stars Can Tell: A Robust Method to Defend against GPS Spoofing Attacks using Off-the-shelf Chipset.. In *USENIX Security Symposium*. 3935–3952.

[38] Xiruo Liu, Lily Yang, Ignacio Alvarez, Kathiravetpillai Sivanesan, Arvind Merwaday, Fabian Oboril, Cornelius Buerkle, Manoj Sastry, and Leonardo Gomes Baltar. 2021. MISO-V: Misbehavior detection for collective perception services in vehicular communications. In *2021 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 369–376.

[39] Hang Qiu, Pohan Huang, Namo Asavisanu, Xiaochen Liu, Konstantinos Psounis, and Ramesh Govindan. 2021. Autocast: Scalable infrastructure-less cooperative perception for distributed collaborative driving. *arXiv preprint arXiv:2112.14947* (2021).

[40] Aanjhan Ranganathan, Hildur Ólafsdóttir, and Srdjan Capkun. 2016. Spree: A spoofing resistant gps receiver. In *Proceedings of the 22nd Annual International Conference on Mobile Computing and Networking*. 348–360.

[41] Florian A Schiegg, Daniel Bischoff, Johannes R Krost, and Ignacio Llatser. 2020. Analytical performance evaluation of the collective perception service in IEEE 802.11 p networks. In *2020 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 1–6.

[42] Junjie Shen, Jun Yeon Won, Zeyuan Chen, and Qi Alfred Chen. 2020. Drift with devil: Security of multi-sensor fusion based localization in high-level autonomous driving under GPS spoofing. In *Proceedings of the 29th USENIX Conference on Security Symposium*. 931–948.

[43] Shuyao Shi, Jiahe Cui, Zhehao Jiang, Zhenyu Yan, Guoliang Xing, Jianwei Niu, and Zhenchao Ouyang. 2022. VIPS: real-time perception fusion for infrastructure-assisted autonomous driving. In *Proceedings of the 28th Annual International Conference on Mobile Computing And Networking*. 133–146.

[44] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. 2019. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 770–779.

[45] Zhiying Song, Fuxi Wen, Hailiang Zhang, and Jun Li. 2022. An Efficient and Robust Object-Level Cooperative Perception Framework for Connected and Automated Driving. *arXiv preprint arXiv:2210.06289* (2022).

[46] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. 2020. Towards Robust {LiDAR-based} Perception in Autonomous Driving: General Black-box Adversarial Sensor Attack and Countermeasures. In *29th USENIX Security Symposium (USENIX Security 20)*. 877–894.

[47] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. 2020. Physically realizable adversarial examples for lidar object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13716–13725.

[48] James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. 2021. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7768–7777.

[49] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. 2020. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *European Conference on Computer Vision*. Springer, 605–621.

[50] Xinshuo Weng, Jianren Wang, David Held, and Kris Kitani. 2020. 3D Multi-Object Tracking: A Baseline and New Evaluation Metrics. *IROS* (2020).

[51] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. 2022. V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer. In *European Conference on Computer Vision*. Springer, 107–124.

[52] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. 2022. Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2583–2589.

[53] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. 2022. DAIR-V2X: A Large-Scale Dataset for Vehicle-Infrastructure Cooperative 3D Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21361–21370.

[54] Yunshuang Yuan, Hao Cheng, and Monika Sester. 2022. Keypoints-Based Deep Feature Fusion for Cooperative Vehicle Detection of Autonomous Driving. *IEEE Robotics and Automation Letters* 7, 2 (2022), 3054–3061.

[55] Qingzhao Zhang, Shengtuo Hu, Jiachen Sun, Qi Alfred Chen, and Z Morley Mao. 2022. On adversarial robustness of trajectory prediction for autonomous vehicles. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15159–15168.

[56] Qingzhao Zhang, Shuowei Jin, Ruiyang Zhu, Jiachen Sun, Xumiao Zhang, Qi Alfred Chen, and Z Morley Mao. 2024. On data fabrication in collaborative vehicular perception: Attacks and countermeasures. In *33rd USENIX Security Symposium (USENIX Security 24)*. 6309–6326.

[57] Qingzhao Zhang, Xumiao Zhang, Ruiyang Zhu, Fan Bai, Mohammad Naserian, and Z Morley Mao. 2023. Robust Real-time Multi-vehicle Collaboration on Asynchronous Sensors. In *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*. 1–15.

[58] Xumiao Zhang, Anlan Zhang, Jiachen Sun, Xiao Zhu, Y Ethan Guo, Feng Qian, and Z Morley Mao. 2021. EMP: edge-assisted multi-vehicle perception. In *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*. 545–558.

[59] Yi Zhu, Chenglin Miao, Tianhang Zheng, Foad Hajiaghajani, Lu Su, and Chunming Qiao. 2021. Can we use arbitrary objects to attack lidar perception in autonomous driving?. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*. 1945–1960.