From Threat to Trust: Exploiting Attention Mechanisms for Attacks and Defenses in Cooperative Perception

Chenyi Wang, *University of Arizona;* Raymond Muller and Ruoyu Song, *Purdue University;* Jean-Philippe Monteuuis and Jonathan Petit, *Qualcomm;* Yanmao Man, *Independent Researcher, U.S.;* Ryan Gerdes, *Virginia Tech;* Z. Berkay Celik, *Purdue University;* Ming Li, *University Of Arizona*

https://www.usenix.org/conference/usenixsecurity25/presentation/wang-chenyi

This paper is included in the Proceedings of the
34th USENIX Security Symposium.

August 13–15, 2025 • Seattle, WA, USA

978-1-939133-52-6

From Threat to Trust: Exploiting Attention Mechanisms for Attacks and Defenses in Cooperative Perception

Chenyi Wang[1] Raymond Muller[2] Ruoyu Song[2] Jean-Philippe Monteuuis[3] Jonathan Petit[3]
Yanmao Man[5] Ryan Gerdes[4] Z. Berkay Celik[2] Ming Li[1]

[1]University of Arizona [2]Purdue University [3]Qualcomm [4]Virginia Tech [5]Independent Researcher, U.S.

{chenyiw, yman, lim}@arizona.edu, {mullerr, song464, zcelik}@purdue.edu
{jmonteuu, petit}@qti.qualcomm.com, rgerdes@vt.edu

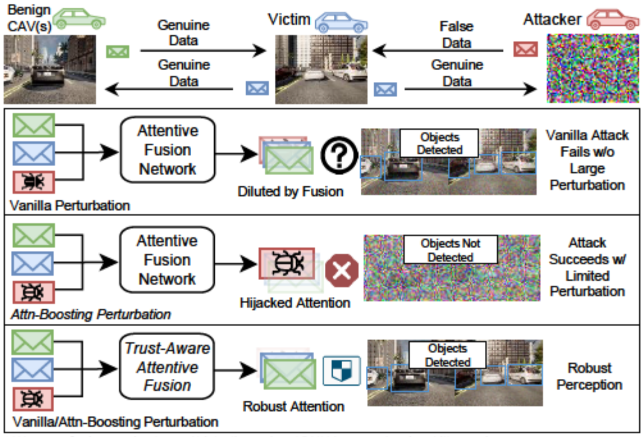
Abstract

Cooperative perception (CP) extends detection range and situational awareness in connected and autonomous vehicles by aggregating information from multiple agents. However, attackers can inject fabricated data into shared messages to achieve adversarial attacks. While prior defenses detect object spoofing, object *removal* attacks remain a serious threat. Nevertheless, prior attacks require unnaturally large perturbations and rely on unrealistic assumptions such as complete knowledge of participant agents, which limits their attack success. In this paper, we present SOMBRA, a stealthy and practical object removal attack exploiting the attentive fusion mechanism in modern CP algorithms. SOMBRA achieves 99% success in both targeted and mass object removal scenarios (a 90%+ improvement over prior art) with less than 1% perturbation strength and no knowledge of benign agents other than the victim. To address the unique vulnerabilities of attentive fusion within CP, we propose LUCIA, a novel trustworthiness-aware attention mechanism that proactively mitigates adversarial features. LUCIA achieves 94.93% success against targeted attacks, reduces mass removal rates by over 90%, restores detection to baseline levels, and lowers defense overhead by 300× compared to prior art. Our contributions set a new state-of-the-art for adversarial attacks and defenses in CP.

1 Introduction

Connected and autonomous vehicles (CAVs) are transforming transportation by making it safer and more efficient [47], and are predicted to capture 50-90% of the market by 2040 [7,34]. A key technology behind this transformation is cooperative perception (CP), where CAVs exchange information to better understand their surroundings [50]. By working collaboratively, vehicles can detect objects beyond their own sensor range, reduce blind spots, and perceive occluded obstacles, enabling smarter decision-making on the road.

In CP, vehicles exchange and aggregate perceptual information. Depending on the stage at which data is shared and fused, current CP systems can be broadly divided into three categories: early fusion, intermediate fusion, and late fusion. Among these, intermediate fusion is the dominant choice for state-of-the-art CP algorithms, as it strikes a balance between detection accuracy and communication efficiency [68].



* We use RGB images in place of high-dimensional DNN features for visual illustration.

Figure 1: Attention mechanisms in cooperative perception's fusion networks can be exploited for attacks and defenses.

To effectively merge messages from multiple vehicles, an integral part of state-of-the-art (SOTA) intermediate fusion CP algorithms is the attentive fusion mechanism [19,68]: a DNN that learns to focus on the most informative parts of the input features. Through attention weighting, the fusion module discerns the contribution of CAV's feature maps to the overall perception results, and thus, selectively amplifies or suppresses different inputs to maximize the detection accuracy.

However, the increased complexity of CP systems expands the attack surface of vehicular perception. Recently, attacks against CP have been demonstrated in which an adversary-controlled CAV in the network can inject adversarial messages to mislead the perception of a remote agent [58,72] (e.g., remove objects otherwise perceivable by the victim) and cause serious consequences such as collisions [4].

Nevertheless, existing attacks on CP systems use conventional single-agent adversarial loss functions, while neglecting the unique vulnerabilities of CP systems that lie in their fusion process. Hence, such attacks require unrealistically large perturbations that can be detected by simple sanity checks. Furthermore, prior works rely on an unrealistic assumption. They assume the adversary is connected to all CAVs working with the victim, allowing the attacker to estimate the CP outcomes from the victim's perspective using identical inputs for the CP algorithms. Such an assumption ignores the real-world scenario of hidden terminal problems in wireless communication. In such scenarios, the attacker may not detect the presence of other benign agents connected to the victim due to limited transmission range, interference [63], or preference to remain far and covert. Their attack success is therefore greatly reduced in a more practical setting.

In this work, we propose SOMBRA, the first object removal attack that exploits the unique vulnerability in the attentive fusion mechanism of CP, to enhance both the attack success rates and its stealthiness (i.e., requiring less perturbation), especially under limited attacker knowledge of other benign CAVs. As shown in Figure 1, by manipulating the victim's attention distribution, the adversary can misdirect the victim's focus heavily onto the malicious feature. Our approach departs from conventional, single-agent adversarial losses by designing an attention-focused loss function that leverages the interplay between shared feature vectors, attention weights, and the final fused perception.

On the other hand, existing defenses in CP systems are *reactive*. They either validate with additional (yet unverified) information from the same set of input sources, or compare across multi-round inference results using partial information. Therefore, prior defenses have several limitations in terms of their effectiveness and practicality, including: (1) high false-positive rates, indistinguishability from benign errors and ineffectiveness against adaptive attacks [72], (2) hyperparameter sensitivity and unrealistic prior knowledge assumption about the attacker [29], and (3) high computation overhead due to multi-iteration design and information loss (e.g., randomly discarding benign messages) [29].

To bridge this gap, we propose LUCIA, the first *proactive* defense against CP threats by harnessing and improving the attention mechanisms in SOTA CP systems. Our CP defense adjusts each agent's attention to the fused feature map based on trustworthiness scores derived from a lightweight feature consistency check. By modulating the focus to more consistent and trusted input, LUCIA can prevent a single compromised source from significantly swaying the fused outcome with stealthy adversarial manipulations, while maintaining real-time feasibility and high information utilization rate.

Our contributions can be summarized as follows.

- We design a novel object removal attack SOMBRA against CP systems that manipulates attention weights to amplify the contribution of malicious features in the fused CP result, achieving over 99% success rates in targeted and mass object removal scenarios, outperforming existing attacks [29, 58, 72] by more than 90%. Unlike prior methods, SOMBRA achieves high success rates even when the attacker is connected only to the victim, without receiving information from other benign agents.
- We propose a lightweight and proactive trustworthiness-aware defense LUCIA that computes feature-level consistency scores to dynamically adjust attention weights. By neutralizing adversarial manipulations while preserving utilization of cooperative messages from trustworthy agents, LUCIA achieves up to 94.69% success rates under targeted attacks, outperforming prior art [29] by up to 91%, and restores perception performance to near baseline level for mass object removal and general spoofing/removal attacks. LUCIA maintains over 90% success rates against adaptive attacks.
- We evaluate SOMBRA and LUCIA on four SOTA CP algorithms [19, 28, 38, 68] using the benchmark dataset OPV2V [68]. Compared with prior art [29, 58, 72] in CP attacks and defenses, SOMBRA demonstrates superior attack efficacy even with 1% of the perturbation required by prior methods, while LUCIA achieves higher robust perception accuracy and reduces computation overhead by over 300×.

Our code is open-source at `https://github.com/WiSeR-Lab/SOMBRA_LUCIA/`.

2 Background and Related Work

2.1 Cooperative Perception Systems

Connected and Autonomous Vehicles (CAVs) are autonomous vehicles (AV) equipped with advanced wireless communication systems that enable them to share information with other networked vehicles (V2V), infrastructure (V2I), and devices (V2X) [47]. In addition to the capability of navigating autonomously based on local sensors, such as cameras and LiDAR, their connectivity enables real-time data exchange. This enhances situational awareness beyond the sensing range of a single vehicle, enabling collaborative driving.

Cooperative Perception (CP) enhances the perception capabilities of CAVs by enabling them to share and fuse sensor data with nearby CAVs. Depending on the stage at which perceptual information is exchanged among CAVs and fused for inference, current CP algorithms broadly fall into one of the three categories: (1) early fusion, such as Cooper [9], where raw sensor data (e.g., LiDAR point clouds) are exchanged and combined, by data concatenation, before DNN-based feature extraction; (2) intermediate fusion, such as F-Cooper [8], AttFusion [68], and V2VAM [28] where spatially-aligned intermediate feature values (e.g., bird's eye view feature) extracted from the sensor data are being shared and aggregated using DNN-based fusion backbone into a single feature for inference (as shown in Figure 2); and (3) late fusion, exemplified by [51], where detection results made by individual vehicles based on local data are aggregated into a single set of collective detection results.

However, due to AV system's stringent requirements for real-time applicability (e.g., 100 ms end-to-end latency [39]), early fusion faces challenges from the limited communication bandwidth offered for V2X systems (e.g., 20 MHz [2]) and prohibitively large raw sensor data volumes (e.g., 4 MB for each LiDAR frame [8]). Meanwhile, late fusion suffers from suboptimal detection performance due to its reliance on local information [68]. Therefore, recent works have been focusing on intermediate fusion as it strikes a balance between communication overhead and detection performance [8, 68].

2.2 Related Work

Single-Vehicle Perception Attacks and Defenses. Perception systems in AVs are vulnerable to adversarial attacks. In single-vehicle perception, attacks have been extensively studied, including LiDAR spoofing, where adversaries inject fake points into point clouds [5, 23], and physically realizable attacks, such as using a monitor or projector to display adversarial patterns [44, 46, 64]. Other methods, such as adversarial trajectories [55, 61], manipulate the vehicle's perception of object dynamics by maneuvering along an adversarially crafted path. On the other hand, single-vehicle defenses typically validate detected objects by focusing on their physical plausibility or movement patterns. For instance, a misclassification attack of a person as a car could be detected by observing object dynamics and/or attributive features [41, 45, 71], whereas spoofed LiDAR points and objects can be identified using occlusion constraints (e.g., LiDAR shadows) [16, 56]. However, such methods rely on cross-validation with the raw sensor data (which is prohibitively large for real-time transmission in V2X applications) or physics-based constraints of a single sensor source, which are ineffective against CP attacks where remote adversaries inject adversarial perturbations into the shared messages. We review such existing CP attacks below.

Cooperative Perception Attacks. CP introduces a new threat model in which adversaries exploit the data-sharing mechanism. In contrast to physical sensor attacks [23, 64]—which demand deep knowledge of sensor hardware and are typically tailored to a specific brand or manufacturer—false data injection attacks in CP circumvent the physical constraints of sensor attacks, and are more scalable and easier to launch. Consequently, a remote adversary can target surrounding CAVs that rely on shared features. This distributed nature and reliance on multi-agent data exchange pose unique risks that remain largely under-explored compared to single-vehicle scenarios. Depending on the attacker's capabilities, existing CP attacks can be broadly categorized into insider and outsider attacks.

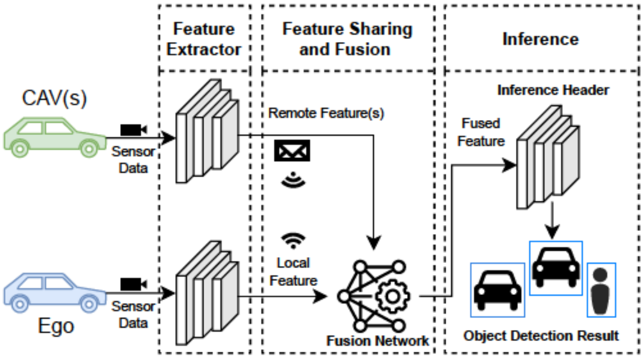


Figure 2: Illustration of the intermediate fusion CP pipeline.

In an insider attack, the adversary is an active participant in the CP system, equipped with valid credentials or physical access to a compromised vehicle [43]. Tu et al. [58] presented the first insider adversarial attack on CP systems, where compromised agents inject perturbations into their shared intermediate features by adapting the loss function proposed in single-agent attacks [57]. The attack reduces the victim's perception accuracy by introducing missing detections (false negatives) and spoofed objects (false positives). Zhang et al. [72] extended this concept by applying the same loss function at specific regions, allowing the attacker to remove particular objects or insert spoofed ones perceivable by the victim. However, their method requires precise prior knowledge of the targets (e.g., bounding box coordinates) and demands extensive fabrication on both raw sensor data and shared features with additional LiDAR point clusters and unbounded perturbation, respectively. Therefore, the attack can be defeated by simple sanity checks (e.g., whether values fall out of range) or sensor integrity defenses [18].

Outsider attacks do not originate from a compromised CAV, but instead exploit vulnerabilities in data transmission or sensor spoofing channels. For example, wireless jamming can cause the loss of sensing data exchange and CP performance degradation [33]. However, they cannot achieve subtle manipulations of the detection results, such as removing the detection of a single object. Also, even in the extreme case where all messages shared with the victim are blocked, the CP algorithm still operates with the genuine input from the victim's local sensor. In this case, the perception range and performance are reduced, yet the results remain truthful. Another line of work [27] examined how GPS spoofing can reduce detection performance in CP systems. Their proposed attack, AdvGPS, primarily causes localization offsets that lead to mismatches in the shared features and, consequently, higher false-positive and false-negative rates. However, the evaluated CP algorithms presume near-accurate positioning prior to fusion, and the resulting degradation is only marginally higher than random noise.

Key Observations. Existing CP attacks overlooked two crit-

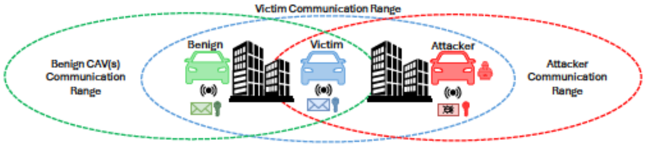

Victim Communication Range
Benign CAV(s) Communication Range
Benign
Victim
Attacker
Attacker Communication Range

Figure 3: Illustration of the hidden terminal scenario.

ical aspects of modern CP frameworks. First, they failed to exploit the *unique designs* in CP algorithms, which lie in the fusion module. This leads to reduced stealthiness or attack efficacy. Second, except for wireless jamming [33], they assumed the adversary is connected to and can receive messages from all agents that communicate with the victim, hence having the same inputs as the victim to facilitate gradient-based optimization. This assumption neglects real-world *hidden terminal* scenarios, as shown in Figure 3. In such scenarios, the attacker may not detect the presence of other benign agents connected to the victim due to limited transmission range, interference [63], or preference to remain far and covert.

Insights from Dot-Product Attention. Recent research in other domains demonstrated the vulnerability of scaled dot-product attention to adversarial manipulation. For instance, Lovisotto et al. [36] showed that conventional gradient-based attacks primarily focus on value tokens while neglecting the potential to manipulate attention weights. They proposed an Attention-Fool strategy that modifies pre-softmax dot-product similarities and misdirects the attention of all queries to a single adversarial key in the model, thereby significantly reducing accuracy in image classification and detection tasks. Similarly, Sharma et al. [53] demonstrated how attention maps in Visual Question Answering can be exploited to yield targeted adversarial samples with minimal noise, causing models to provide incorrect answers by redirecting focus onto malicious regions. These findings illustrate how attention can become a critical vulnerability if attackers manage to bias attention distributions towards particular areas of a single input source.

Based on these insights, we propose the SOMBRA attack, which targets the core component of CP algorithms, i.e., the attentive fusion mechanism, to gain two key advantages. *First*, by guiding the victim's attention toward maliciously crafted features, our method achieves potent object-removal results with *significantly less perturbation* (e.g., $< 1\%$), thereby improving the attack stealthiness. *Second*, it remains highly effective even under limited attacker knowledge, mitigating the need for full observability of all collaborating agents.

3 System and Threat Model

3.1 System Model

We consider a CP system (Figure 2) where each CAV captures perceptual data using its sensors and extracts intermediate feature representations using DNNs, which are then shared with nearby CAVs through wireless communication. Each CAV combines these features using intermediate fusion algorithms such as AttFusion [68] and Where2comm [19] that are based on the attention architecture [60] or its variants, upon which object detection is performed. The results are then delivered to downstream modules such as prediction and planning.

3.2 Threat Model

Attacker Capability and Knowledge. We consider an insider attack scenario, where the attacker poses as a normal participating CAV with CP model access and capability of perceptual message exchange. Following existing insider attack models against CP [58,72], we assume that the attacker has access to a compromised CAV and can manipulate the data being signed and transmitted to other CAVs [15]. Such insider attacks represent a recognized threat vector for modern connected vehicles [49].

We primarily consider a white-box attacker who has knowledge of the CP models and weights, in line with previous work [58, 72]. With such information, the attacker can perform gradient-based optimization to generate adversarial perturbations. This aligns with standard practice in adversarial machine learning security analysis, establishing a worst-case scenario assessment [48]. The white-box assumption, while strong, is relevant in the automotive context for several reasons. Firstly, CP models can be shared/aligned between manufacturers to ensure interoperability, making them widely known/accessible [37]. Secondly, models can potentially be reverse-engineered from captured vehicle hardware, software updates, or diagnostic interfaces [10,42]. Thirdly, model details can be leaked through supply chain compromises [49]. An insider attacker having already compromised a CAV inherently increases the likelihood of gaining knowledge to the deployed model [42]. While the white-box assumption aids SOMBRA's optimal design and analysis, our evaluations (Appendix A) show high efficacy (e.g., >90% success) with transfer attacks, expanding SOMBRA's threat potential.

Attacker Goal. The primary objective of the adversary is object removal in the victim's perception output. We consider two variants: (1) Targeted Object Removal Attack (TOR): the attacker aims to suppress the detection of one or more specific objects in the victim's final detection (e.g., a particular car or pedestrian on the victim's planned path). (2) Mass Object Removal Attack (MOR): the attacker attempts to suppress as many otherwise perceivable objects as possible from the victim's inference output, which leaves the victim mostly 'blind' to surrounding obstacles, posing an immediate safety risk.

4 SOMBRA: Attacking Attentive Fusion

We present our methodology for launching SOMBRA against CP systems. We first present the problem statement and discuss the challenge of performing gradient-based attacks in CP under a realistic threat model. We then review the attentive fusion mechanism—the core design of SOTA intermediate-fusion pipelines—and show how an adversary can exploit it to amplify malicious feature contributions. Finally, we detail our object removal losses for both targeted and mass object removal attacks, which are combined with the attention-boosting loss into single loss functions.

4.1 Problem Statement

Let $\mathcal{V} = \{V_1, \ldots, V_n\}$ be the set of CAVs participating in CP, and let $V_a \in \mathcal{V}$ denote the attacker-controlled vehicle, while $V_v \in \mathcal{V}$ is the chosen victim. At each time step t, each CAV V_i generates and shares its local feature map $X_i^t \in \mathbb{R}^{C\times W\times H}$. Here, C, W, H represent the number of channels, width, and height of the feature map, respectively. These features are spatially aligned to a common coordinate, by modern CP design [68]. Under the benign scenario, the victim V_v fuses the received feature maps and the local feature via an attentive fusion network $\mathcal{F}$:

$$\tilde{X}_v^t = \mathcal{F}\left(\{X_j^t\}_{V_j \in \mathcal{V}_v}, X_v^t\right) \in \mathbb{R}^{C\times W\times H}, \tag{1}$$

where $\mathcal{V}_v \subseteq \mathcal{V} \setminus \{V_v\}$ is the set of neighbors transmitting features to V_v. The fused feature $\tilde{X}_v^t$ is then passed to the inference header I to obtain the detection results $\mathcal{D}_v^t$.

To achieve the attack, the attacker generates and injects an adversarial perturbation δ^t into its own feature X_a^t for the current time step, based on the victim's shared feature X_v^{t-1} from $t-1$ due to transmission delay (❶). The perturbation is computed onboard the adversary's compromised CAV leveraging its GPU accelerators. [1] (❷). The perturbed feature $X_a^t + \delta^t$ is then sent to the victim (❸), where the victim fuses it with its own feature and other features it receives (❹).

$$\tilde{X}_v^t = \mathcal{F}\left(\{X_j^t\}_{j\in\mathcal{V}_v, j\neq a} \cup \{X_a^t + \delta^t\}, X_v^t\right). \tag{2}$$

which is used for producing the final detection results $\mathcal{D}_v^t$.

Object Removal Objective. For Targeted Object Removal Attack (TOR), the attacker specifies a subset of objects $\mathcal{O}_{\text{target}}$ that appear in the victim's ground-truth scene. The attacker can estimate such information by applying a lightweight object detector on X_v^{t-1} [72]. The goal is to suppress the targeted objects from the victim's final detection $\mathcal{D}_v^t = I(\tilde{X}_v^t) \not\supseteq \mathcal{O}_{\text{target}}$. For Mass Object Removal (MOR), the attacker aims for $\min|\mathcal{D}_v^t| = \min\left|I(\tilde{X}_v^t)\right|$, which encourages the victim to overlook as many ground-truth objects present in the scene as possible.

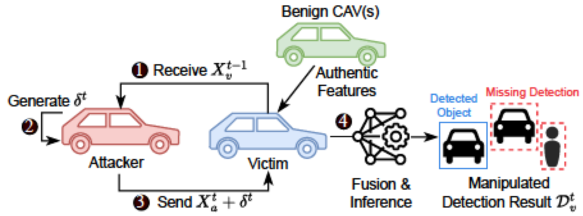


Figure 4: Stages of SOMBRA.

Challenges. To perform gradient-based attacks and manipulate the victim's CP output $\mathcal{D}_v^t = I(\tilde{X}_v^t)$, the attacker needs to have all the input features to the victim's CP algorithm including X_v^t and $\{X_j^t\}_{j\in\mathcal{V}_v}$. Previous works have shown that X_v^t can be effectively approximated using X_v^{t-1} available to the attacker by spatial warping [38, 58, 72] or first-order feature flow [70]. However, existing attacks assumed that the attacker has access to all the other input features $\{X_j^t\}_{j\in\mathcal{V}_v}$ the victim receives, which is unrealistic under a more practical setting due to potential hidden terminal scenarios [63]. Consequently, these attacks suffer from greatly reduced attack success and unrealistically large perturbations required under a limited attacker model. As we detail below, we address this challenge by exploiting the unique weakness in CP's attentive fusion mechanism, which allows us to boost the impact of the malicious feature, significantly increasing attack success while with minimal perturbation required.

4.2 Attentive Fusion Mechanism

Most SOTA intermediate fusion algorithms adopt the attention mechanism [60] or its variant as the fusion backbone [19, 28, 38, 68] that dynamically weighs each vehicle's shared feature map $X_i \in \mathbb{R}^{C\times W\times H}$. For convenience, we denote $X_i(x) \in \mathbb{R}^C$ as the vector at location $x \in \Omega$ in the feature map X_i, where $|\Omega| = WH$. Conceptually, a typical attentive fusion network $\mathcal{F}$ computes an attention score $\alpha_j(x) \in [0,1]$ for each vehicle $V_j \in \{V_1,\ldots,V_N\}$ at each spatial location x based on the corresponding contribution to the final detection. Specifically, at each location, the set of feature vectors from different vehicles are rearranged into a matrix $\mathbf{X}(x) = [X_1(x), \ldots, X_N(x)]^T \in \mathbb{R}^{N\times C}$. The attention score matrix at this location is then obtained by

$$\mathbf{A}(x) = \operatorname{softmax}\left(\frac{\mathbf{X}(x)\mathbf{X}(x)^T}{\sqrt{C}}\right) \in \mathbb{R}^{N\times N}, \tag{3}$$

where $(i, j)-$th entry represents the attention score vehicle V_i assigns to V_j. Let $(\alpha_1(x), \ldots, \alpha_N(x))$ be the i-th row of $\mathbf{A}(x)$, then vehicle V_i's fused feature vector at location x becomes

$$\tilde{X}_i(x) = \sum_{j\in\mathcal{V}_i\cup\{V_i\}} \alpha_j(x) X_j(x), \text{ where } \sum_j \alpha_j(x) = 1. \tag{4}$$

The fused feature map $\tilde{X}_i \in \mathbb{R}^{C\times W\times H}$ is then obtained by assembling the fused vectors. The whole fusion process is efficiently computed using batch computing and parallelization at each CAV. By calculating $\alpha_j(x)$, the fusion network

[1] Alternatively, an attacker can employ off-board computation resources in conjunction with the compromised vehicle's V2X transceiver and credentials.

dynamically quantifies the contextual 'relevance' of different vehicles' messages for each spatial location, which helps improve situational awareness under benign conditions [68].

Exploiting the Attention Mechanism. An attacker controlling V_a can craft a small perturbation δ such that their feature $X'_a(x) = X_a(x) + \delta$ shapes the victim's attention distribution in two ways - (1) *Encourage high weights*: by subtly crafting $X_a + \delta$, the attacker can increase the attention scores $\alpha_a(x)$ so that the malicious feature map disproportionately influences the fused representation. Intuitively, the attention module interprets $X_a + \delta$ as highly 'informative' or 'reliable', thus assigning higher weights to the adversarial message. (2) *Dilute benign features*: as $\sum_i \alpha_i(x) = 1$, amplifying $\alpha_a(x)$ reduces the weights of benign vehicles ($\alpha_j(x) \forall j \neq a$), ensuring that small adversarial perturbations become the dominant signal in the fusion. A straightforward way to encourage this outcome is to optimize for an attention-boosting loss $\mathcal{L}_{\text{attn}}$:

$$\mathcal{L}_{\text{attn}} = \frac{1}{|\Omega|} \sum_{x \in \Omega} \alpha_a(x). \tag{5}$$

Maximizing $\mathcal{L}_{\text{attn}}$ encourages the attacker's feature to appear more salient at each spatial location as perceived by the victim, effectively overriding or diluting the contributions from benign agents and the victim's own local features.

Attacker Influence Analysis. Under the assumption that the pre-softmax scores between the victim and any benign agent (or the unperturbed attacker), $S_{vj}(x) = \frac{\mathbf{X}_v(x)^T \mathbf{X}_j(x)}{\sqrt{C}}$, are bounded[2], the victim's attention share to the perturbed attacker feature is upper-bounded by:

$$\alpha_a(x) \leq \frac{e^{\Delta}}{e^{\Delta} + N - 1}. \tag{6}$$

Here, N is the total number of agents participating in the fusion, and Δ represents the increase in the pre-softmax score achieved by the adversarial perturbation δ. Specifically, $\Delta \leq \frac{\varepsilon_{max} \|X_v(x)\|_1}{\sqrt{C}}$, where ε_{max} is the L_∞ norm of the perturbation[3]. The bound shows that the attacker's influence grows *exponentially* with Δ but decreases linearly as the number of other agents. Sombra exploits this key vulnerability by efficiently hijacking the victim's attention to maximize the adversarial impact. We provide more details in Appendix B.

4.3 Object Removal Loss

Our primary goal is to remove objects from the victim's final detection $\mathcal{D}_v$, which can lead to severe consequences such as collisions [4]. We consider two variants: (1) Targeted Object Removal (TOR) and (2) Mass Object Removal (MOR). In both cases, we employ gradient-based adversarial optimization, where the attacker updates according to the gradient of a loss function $\mathcal{L}_{\text{removal}} \in \{\mathcal{L}_{\text{TOR}}, \mathcal{L}_{\text{MOR}}\}$.

Targeted Object Removal. In a targeted attack, the adversary selects areas corresponding to a set of target object(s) $\mathcal{O}_{\text{target}}$ in the scene and attempts to suppress any detection from being made at these areas. Let $\mathcal{R}(\mathcal{O}_{\text{target}})$ represent the locations in the feature map that corresponds to the target areas. To push the model to classify these locations as background (i.e., no object), we employ a focal loss–like term $\mathcal{L}_{\text{TOR}}$ that encourages high probabilities of the background class are assigned to the target areas:

$$\mathcal{L}_{\text{TOR}} = - \sum_{x \in \mathcal{R}(\mathcal{O}_{\text{target}})} (1 - p_x)^{\gamma} \log(p_x), \tag{7}$$

where p_x is the predicted probability that location x has no object, and γ is the focusing parameter of the focal loss [32].

Mass Object Removal. For mass removal, the adversary aims to eliminate as many objects perceivable by the victim as possible, effectively encouraging an empty detection output produced by the victim $\mathcal{D}_v = \emptyset$. Similar to the targeted approach, we extend the focal loss concept to *all* spatial locations in the feature map:

$$\mathcal{L}_{\text{MOR}} = - \sum_{x \in \Omega} (1 - p_x)^{\gamma} \log(p_x). \tag{8}$$

In other words, we encourage the prediction of every location as if it belongs to the background class, so the victim's perception network becomes confident in 'nothing there'. This strategy eliminates the need for object-specific knowledge.

Advantages of Focal Loss. Previous object removal attacks [57, 58, 72] rely on precise prior knowledge of the victim's perceivable objects, such as exact bounding box coordinates. However, without accurate prior information, these attacks become unstable (side-effect of creating false objects near the removal target) or unsuccessful. For instance, if the initial victim's prediction does not contain the object of interest, the gradient becomes zero, making optimization ineffective.

To overcome this challenge, we adopt the focal loss that does not require prior knowledge of object bounding boxes. Also, note that the focal loss is designed to address class imbalance by down-weighting easy examples and focusing on hard negatives, making it suitable for our purpose of reducing objectness scores for target areas. In both TOR and MOR, the modulating factor $(1 - p_x)^{\gamma}$ reduces the loss contribution from easy negatives (where p_x is close to 1), allowing the optimization to focus on harder examples (where p_x is lower). It effectively emphasizes locations where the network is uncertain about the background class. Meanwhile, the logarithmic component $\log(p_x)$ encourages p_x to approach 1 (i.e., maximizing the probability of the background class). By minimizing $\mathcal{L}_{\text{removal}}$, the victim's confidence on obstacle

[2] The scores are bounded since both the (1) input LiDAR point count and intensity and (2) feature extraction model weights are bounded.

[3] The equality holds when the optimization is performed using PGD.

detected at targeted/all areas are reduced, encouraging no corresponding detections being made as the output.

Combined Loss Function. To fully exploit the attention mechanism while enforcing object removal, we combine $\mathcal{L}_{\text{attn}}$ and $\mathcal{L}_{\text{removal}}$ into a single loss function:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} + \lambda_{\text{removal}}\mathcal{L}_{\text{removal}}, \tag{9}$$

where λ_{attn} and λ_{removal} hyperparameters that balance the two terms and $\mathcal{L}_{\text{removal}} \in \{\mathcal{L}_{\text{TOR}}, \mathcal{L}_{\text{MOR}}\}$. By explicitly *increasing* the attention assigned to $X_a + \delta$, the adversary ensures that even small perturbations can drastically impact the fused representation $\tilde{X}_v$, achieving potent object removal without requiring knowledge of other agents' data or overly large δ that might be easily detected.

5 LUCIA: Harnessing Attention for Trust

Although the vanilla attention mechanisms in SOTA CP algorithms present a unique vulnerability, they can be harnessed in turn for enhanced adversarial robustness. In this section, we present our defense LUCIA, based on a novel trustworthiness-aware attentive fusion, which can be embedded directly into the existing CP fusion pipelines without additional training. LUCIA proactively adjusts each agent's contribution based on a lightweight, on-the-fly consistency check of their intermediate features, achieving significantly reduced overheads and improved effectiveness compared to prior art [29].

Before detailing our proposed methodology, we introduce existing CP defenses and highlight their key limitations.

5.1 Existing Defenses & Limitations

To prevent direct V2X message alterations, the C-V2X [1] and IEEE 1609.2 [21] standards adopt cryptographic mechanisms for message authentication, yet they do not prevent false data injection before authentication [3]. Also, existing V2X-MBDs [59] can only counteract threats for basic safety applications, and are not easily applicable to advanced CP (requiring sensor data sharing and fusion). To date, defenses against false data injection attacks in CP systems remain under-explored, where existing works can be broadly categorized into consensus-based and consistency-based approaches.

Consensus-Based Defenses. A representative example is ROBOSAC [29], which adapts the RANSAC (RANdom SAmple Consensus) approach to multi-agent perception. ROBOSAC randomly samples small subsets of agents and fuses their shared data to generate a detection result; it then compares this outcome against the ego agent's local perception to see if they fall within a predefined similarity threshold. Once such a detection result is found (e.g., more than 70% of objects can be matched with the local detection), it is accepted as the 'robust' detection output. Another work CP-Guard [17] follows a similar RANSAC paradigm and specifically applies to the semantic segmentation task.

Despite its conceptual appeal, RANSAC-based defenses suffer from three main drawbacks. *First*, although measuring the difference between the inference results from multiple input sources with the trusted local result can help identify obvious discrepancies caused by untargeted attacks (e.g., FGSM [14], C&W [6]), they struggle to identify more subtle attacks such as the removal of single object detections. *Second*, these algorithms need to run multiple rounds of sampling and inference in each perception cycle. Notably, the sampling budget in terms of the number of iterations required is exponential with respect to the number of collaborators demanded. Therefore, it incurs high computational overhead that is incompatible with strict real-time requirements for autonomous driving (e.g., 100 ms end-to-end [31]). *Third*, it requires multiple *a priori* hyperparameters such as the attacker ratio, and suffers from hyperparameter sensitivity. Such assumptions are unrealistic given that vehicles may join and leave the network dynamically, and that adversaries can strategically time their attacks. Also, the dependence on input parameters makes them brittle such that small changes in hyperparameters can dramatically shift the information utilization rate. Not only does it randomly discard information from other agents, to ensure real-time feasibility, the algorithm often has to discard *all* benign agents' information only suspecting the presence of a single attacker. The incurred tradeoff between information utilization and computation efficiency is challenging to balance. We present more detailed analysis in Appendix C.

Consistency-Based Defenses. Collaborative Anomaly Detection (CAD) [72] introduces the only consistency-based defense for (LiDAR-based) CP to date. Rather than sampling random subsets of agents, CAD validates the consistency between the final detection results and an occupancy map aggregated from each agent's local counterpart derived by a separate point segmentation model on raw LiDAR point cloud. Specifically, it generates the fused occupancy map by filtering out conflicted regions across shared occupancy data, then cross-checks this fused occupancy map against the ego vehicle's final detection results to alert suspected anomalies.

Although this technique can alert blatant discrepancies, it suffers from several limitations. *First*, the algorithm has a strong dependency on accurate validation data. However, a single attacker can forge both the adversarial messages and the corresponding validation data that deliberately conflicts with the occupancy maps from other agents. By its design, CAD is forced to discard conflicted occupancy maps in the final comparison with the detection, effectively leaving potential attacked areas unchecked. *Second,* the additionally required modules (e.g., LiDAR point segmentation, ground fitting) are non-trivial to maintain in real-time pipelines and incur accumulating errors from sensor noise to inference errors, resulting in significant challenges in distinguishing benign errors from adversarial attacks. As a result, CAD can only detect around 40% object removal attacks and incur up to 60% false alarms [72], while its AUC for object removal attack detection is about 0.5, indicating its ability to detect such attacks is comparable to random guessing (see Figure 14 in [72]). *Third,* CAD does not recover or produce robust perception results in addition to anomaly detection, limiting its applicability in the real world.

Key Observations. Both consensus-and consistency-based defenses are *reactive* in nature: they require obtaining the final perception results before attempting to discern abnormal output. In the time and safety-critical application of autonomous driving, such approaches are difficult to deploy. By contrast, as we show next, our defense is designed under fundamentally different principles, which is *proactive* and lightweight, and integrated at the feature level-an earlier stage in the CP pipeline, instead of reactively detecting anomalies from non-robust inference results. The core insight is that existing attention mechanisms in CP compare raw feature vectors across agents, without validating inter-agent consistency, leaving the system vulnerable to modifications from a single malicious source. By dynamically assessing the trustworthiness of each agent and informing the attentive fusion module accordingly, we can harness it for robust CP algorithm designs. This not only drastically reduces the additional communication and computation overhead, but also inherently counters the attacker's influence on the fused feature and the inference results made thereon.

5.2 Our Defense Methodology

To address the issues in existing CP defenses, we propose our defense LUCIA based on a novel trustworthiness-aware attention mechanism that complements existing CP algorithms without incurring high overhead or depending on additional validation data. Figure 5 shows the stages of LUCIA.

Trust Score Computation via Feature Consistency. Each agent's shared intermediate feature map is compressed (e.g., via average pooling) and normalized to reduce dimensionality that mitigates feature misalignment due to benign sensor error (e.g., localization error) and reduces additional computation overhead (❶). We then compute pairwise L_1 distances among these compressed representations to measure how similarly each agent 'sees' the environment relative to others. Agents whose features deviate substantially accumulate a higher total L_1 distance. We apply a softmax to transform these distances into scores and invert the result, producing a final trust score $T_i \in [0,1]$ per agent (❷).

$$T_i = 1 - \text{softmax}\left(\left\{\sum_{j\neq i} \|X_i - X_j\|_1 \mid j = 1,\dots,N\right\}\right)[i] \quad (10)$$

This process operates entirely on data that the CP pipeline already exchanges (i.e., the intermediate features), avoiding additional communication overhead.

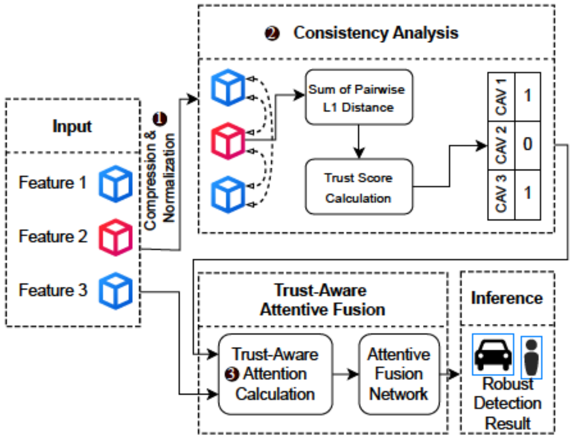


Figure 5: Stages of LUCIA.

Incorporating Trust Scores into Attention. We then integrate T_i directly into the attention module. We multiply both the logits and the final normalized attention weights by T_i (❸). Recall that the original attention score matrix is calculated as $\mathbf{A}(x) = \text{softmax}\left(\frac{\mathbf{X}(x)\mathbf{X}(x)^T}{\sqrt{C}}\right) \in \mathbb{R}^{N\times N}$. Let $\mathbf{T} = \text{diag}(T_1,\dots,T_n)$, then the trustworthiness-aware attention score matrix $\hat{\mathbf{A}}_T(x)$ is calculated by

$$\hat{\mathbf{A}}(x) = \text{softmax}\left(\mathbf{X}(x)\mathbf{X}(x)^T\mathbf{T}/\sqrt{C}\right)\mathbf{T} \quad (11)$$

$$\hat{\mathbf{A}}_T(x) = \hat{\mathbf{A}}(x) \oslash \left(\hat{\mathbf{A}}(x)\mathbf{J}\right) \quad (12)$$

where $\oslash$ represents elementwise division and $\mathbf{J}$ is the matrix of all ones. Intuitively, even if an adversarial feature initially has a high raw attention score, a low trust score T_i ensures that it contributes significantly less to the fused feature (e.g., a trust score of 0 implies no attention paid to this feature). By weighting attention according to observed cross-agent consistency, our defense proactively discounts suspicious feature maps in a single pass, negating the need for iterative or external validation. Also, the fused feature map remains a weighted sum of features from multiple agents, where the weights are normalized as the vanilla model does, yet informed by the trustworthiness of the participating agents. We detail more discussions on leveraging feature-level consistency for defense in Appendix D.

5.3 Advantages of Our Defense

Proactive and Low Overhead. Instead of conducting multiple sampling and partial inferences (like ROBOSAC [29] or CP-Guard [17]), LUCIA computes trust scores once per perception cycle and incorporate them into the existing attention layer. The average-pooling and L_1 distance calculations add minimal computational cost, preserving real-time performance without multiplicative increase in inference time.

Seamless Integration. LUCIA modifies only the fusion step, making it straightforward to deploy in existing intermediate-fusion frameworks without additional training. There is no reliance on specialized LiDAR segmentation or occupancy map generation, simplifying implementation and maintenance, and avoiding additional error accumulation.

Improved Information Utilization. LUCIA does not discard entire feature maps based on rigid pre-defined thresholds or random sampling. RANSAC-based methods' conservative random sampling frequently discards a significant portion of the legitimate data contributed by other agents (e.g., not collaborating with certain vehicles at all if it suspects just one attacker in a group). On the other hand, consistency-based methods suffer from the indistinguishability of benign errors from adversarial attacks, and the inability to produce robust perception results as output. As contrast, our method smoothly modulates attention informed by the trust score, accommodating natural variations in sensor data without triggering a complete rejection. This balance helps maintain the benefits of CP while enhancing the robustness of the system.

6 Evaluation

6.1 Experimental Setup

Dataset and Algorithms. We test SOMBRA and LUCIA using the widely adopted CP dataset OPV2V [68], which is a large-scale benchmark for V2V perception, collected across 70 diverse scenes from 8 towns in the digital twin simulator CARLA [11]. It includes 11,464 frames of LiDAR point cloud data from 2-5 CAV agents and 232,913 annotated 3D vehicle bounding boxes. We regard the CAVs with the lowest and second-lowest indices as the victim and attacker vehicles, respectively. Also, to align with our realistic threat model, we limit the attacker's knowledge about other CAVs to only the victim without access to other benign CAVs messages that are available to the victim. We evaluate SOMBRA and LUCIA on representative intermediate fusion-based CP algorithms that achieve SOTA 3D object detection accuracy (mAP@0.5>0.9) [65] and real-time computation efficiency (< 100 ms on our testbed): AttFusion [68], Where2comm (W2C) [19], CoAlign [38], and V2VAM [28], using publicly available pre-trained weights [65].

Comparison Baselines. We compare SOMBRA with the targeted object removal attack loss used in prior art [57, 58, 72]. Their method requires the attacker to have precise prior knowledge of the target objects' bounding boxes to generate adversarial perturbations, for which we supply the ground truth. Note that to achieve the MOR attack, their attacker is required to have complete knowledge of all objects in the scene, which is a significantly stronger assumption than ours.

LUCIA is compared to ROBOSAC [29] because it is the only defense that aims at outputting robust object detection results. Especially, we assess each defense's detection accuracy under benign and attacked scenarios and computation overhead.

Target Object Selection. For TOR, to evaluate the attack's effectiveness and generalizability, we randomly select an object within each frame to serve as the *target* for removal, while considering whether the target is within victim's Line-of-Sight (LoS): (1) *Objects within LoS*: non-occluded objects that fall within the victim vehicle's local LiDAR sensor range. In these cases, the victim has partial local information about the target. (2) *Objects beyond LoS*: objects that lie entirely outside the victim's sensor range or are occluded by other obstacles. The victim relies solely on the shared features from other CAVs to detect these objects. The randomized selection approach enables a comprehensive evaluation of our defense, as the attacker must handle a wide variety of target sizes, distances, and occlusion conditions.

6.1.1 Implementation

Implementation Details. For our attack SOMBRA, the perturbation is optimized via PGD [40] with 10 iterations and a learning rate ε of 0.1. We set $\lambda_{\text{attn}} = \lambda_{\text{removal}} = 1$, and $\gamma = 2$. Additionally, we show the attack results for different optimization hyperparameters demonstrating the advantage of SOMBRA in reducing the required perturbation strength than the baseline. Note that stealthiness is inversely proportional to perturbation strength measured by ε, where smaller perturbations produce greater stealthiness.

We integrate our defense LUCIA into the fusion backbones of the evaluated CP algorithms. Specifically, we replace the standard scaled dot-product attention with our trust score–modulated variant. For trust score computation, we apply a 32× average pooling across spatial dimensions, followed by normalization and the cross-agent consistency scoring. To test the generalization of our defense to other attacks, we evaluate it against Basic Iterative Method (BIM) [25], a white-box adversarial attack that simultaneously removes and spoofs objects at random. BIM is configured with a learning rate of 0.1 and 10 iterations. For comparison with ROBOSAC [29], we followed the official implementation and supply the algorithm with ground-truth attacker ratio (single attacker) and the same hyperparameters as reported by the authors. The sampling budget is set to 10.

Testing Hardware. To demonstrate system deployment and obtain performance measurements on real vehicles, we ran the experiments with a DataSpeed vehicle testbed running a NUVO-8208GC computer [22], equipped with Intel Xeon E-22278GE CPU, NVIDIA RTX 2080-Super GPU and Ubuntu 20.04, as shown in Figure 11. On our testbed, the PGD perturbation optimization takes 51.21 ± 15.14 ms per frame with PyTorch 2.4.1 and CUDA 11.8, fitting within one typical LiDAR cycle [72]. The OPV2V dataset is loaded through developer I/O to emulate sensor input and message exchange.

6.1.2 Evaluation Metrics

Attack Success Rate (ASR). The ASR quantifies the proportion of successful attacks over the total number of frames N in the dataset. It measures the attack's effectiveness in achieving specific goals. The ASR is calculated as:

$$\mathrm{ASR}(\mathcal{C}) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\mathcal{D}_i \models \mathcal{C}), \tag{13}$$

where $\mathbb{I}(\cdot)$ is the indicator function that evaluates to 1 if the detections $\mathcal{D}_i$ satisfy the attack objective related condition $\mathcal{C}$, and 0 otherwise. This metric is agnostic to specific downstream modules, focusing solely on the attack effectiveness against the perception system. For TOR, an attack succeeds if the target is not detected (i.e., no detection has a non-zero Intersection-over-Union (IoU) with the target), and the number of false positives FP_i (i.e., detections that have zero IoU with any ground truth object) is within a threshold τ_{TOR}:

$$\mathcal{D}_i \models \mathcal{C}_{\mathrm{TOR}} \iff (\mathrm{IoU}_{\max} = 0) \wedge (\mathrm{FP}_i \leq \tau_{\mathrm{TOR}}). \tag{14}$$

Here, $\mathrm{IoU}_{\max}$ is the maximum IoU between any detected bounding box and the ground truth bounding box of the target object in frame i, and τ_{TOR} is a small threshold to limit false positives that might interfere with the attack's objective (e.g., in the way of a potential collision trajectory between the target and the victim). Ideally, the rate of false positives should not exceed that when no attack is performed. Since $\mathrm{FPR} \approx 2/\mathrm{frame}$ on OPV2V for SOTA CP algorithms under the benign case, we consider $\tau_{\mathrm{TOR}} = 2$.

For MOR, an attack succeeds if the total number of detections $|\mathcal{D}_i|$ in frame i is within a threshold τ_{mor}:

$$\mathcal{D}_i \models \mathcal{C}_{\mathrm{MOR}} \iff |\mathcal{D}_i| \leq \tau_{\mathrm{mor}} \tag{15}$$

One can set τ_{blinding} to 0 to evaluate complete blinding or 1 allowing a minimal number of detections to be made. For convenience, we denote the ASR setting τ_{MOR} to 0 and 1 as $\mathrm{ASR}^{0}_{\mathrm{MOR}}$ and $\mathrm{ASR}^{1}_{\mathrm{MOR}}$, respectively.

We also report the Object Removal Rate (ORR), which measures the effectiveness of the attack in reducing the number of detected objects relative to the ground truth:

$$\mathrm{ORR} = \frac{1}{N}\sum_{i=1}^{N} \max\left(0, 1 - \frac{|\mathcal{D}_i|}{|\mathcal{G}_i|}\right), \tag{16}$$

where $|\mathcal{G}_i|$ is the total number of ground truth objects in frame i. The ORR ranges between 0 to 1, representing the average proportion of ground truth objects missed.

Defense Metrics. For defense against TOR, we define *Defense Success Rate (DSR)* as the fraction of frames for which the target object remains correctly detected (with $\mathrm{IoU} \geq 0$) despite the attacker's manipulations. For defense against MOR, where the attacker aims to *blind* the victim by removing all or most objects, we measure the defense's capability to preserve detection via ORR, where *smaller* ORR indicates the defense prevents large-scale object removal.

Table 1: Results on targeted object removal. Notably, SOMBRA achieves consistent and high ASRs, and maintains higher mAP. This is because it selectively removes the target without the side effect of introducing large numbers of false positives like the baseline method, thereby maintaining stealthiness.

		Within Victim LoS				Beyond Victim LoS			
	No Attack	Baseline		SOMBRA		Baseline		SOMBRA	
Model	mAP @0.5	ASR↑	mAP @0.5	ASR↑	mAP @0.5	ASR↑	mAP @0.5	ASR↑	mAP @0.5
AttFusion	0.91	15.26%	0.28	**99.61%**	0.90	45.39%	0.49	**98.79%**	0.90
CoAlign	0.91	1.07%	0.05	**99.12%**	0.89	2.93%	0.09	**98.39%**	0.88
W2C	0.91	2.58%	0.11	**78.21%**	0.20	7.99%	0.15	**77.76%**	0.20
V2VAM	0.93	**99.80%**	0.01	97.37%	0.81	93.71%	0.13	**97.22%**	0.82

Table 2: Results on mass object removal. For AttFusion and CoAlign, SOMBRA achieves near complete removal of all objects perceivable by the victim, exceeding the baseline method by over 90% in ASRs and ORRs.

Attack	No Attack	Baseline			SOMBRA		
Model	ORR	$\mathrm{ASR}^{0}_{\mathrm{MOR}}$ ↑	$\mathrm{ASR}^{1}_{\mathrm{MOR}}$ ↑	ORR↑	$\mathrm{ASR}^{0}_{\mathrm{MOR}}$ ↑	$\mathrm{ASR}^{1}_{\mathrm{MOR}}$ ↑	ORR↑
AttFusion	3.47%	0.05%	0.15%	0.15%	**99.95%**	**100.00%**	**100.00%**
CoAlign	2.28%	0.00%	0.00%	2.13%	**99.76%**	**100.00%**	**99.99%**
W2C	3.68%	0.00%	0.15%	21.90%	**55.24%**	**76.67%**	**94.26%**
V2VAM	3.12%	99.95%	100.00%	100.00%	**100.00%**	**100.00%**	**100.00%**

Mean Average Precision (mAP): We report mAP, a metric used to measure the overall accuracy of object detectors, at 0.5 and 0.7 IoU, which computes the average precision of producing detections that match ground truths.

6.2 Evaluation of Attacks

Targeted Object Removal (TOR). Table 1 shows the attack results on TOR under the two visibility scenarios, where SOMBRA maintains high ASRs in both within-LoS and beyond-LoS scenarios when no defense is present. For instance, AttFusion yields 98–99% ASR in both cases, while CoAlign and V2VAM also remain above 97%. In contrast, the baseline method [72] shows lower ASRs overall (e.g., 15.26% vs. 99.61% for AttFusion within-LoS). Notably, it also exhibits a side effect of reduced mAP due to unintended injection of additional false positives (extra bounding boxes) in regions unrelated to the target. This broader disruption increases the detectability of the attack itself. Our approach, by comparison, removes the chosen target object without notable harm to the detection of other objects, making the attack stealthier.

Mass Object Removal (MOR). Table 2 reports both the attack success rates $\mathrm{ASR}^{0}_{\mathrm{MOR}}$ (strict zero detections) and $\mathrm{ASR}^{1}_{\mathrm{MOR}}$ (at most one detection), alongside ORR—the proportion of ground-truth objects that vanish from the victim's detections. For AttFusion, SOMBRA reaches 100% success in both $\mathrm{ASR}^{1}_{\mathrm{MOR}}$ and ORR, versus negligible baseline values.

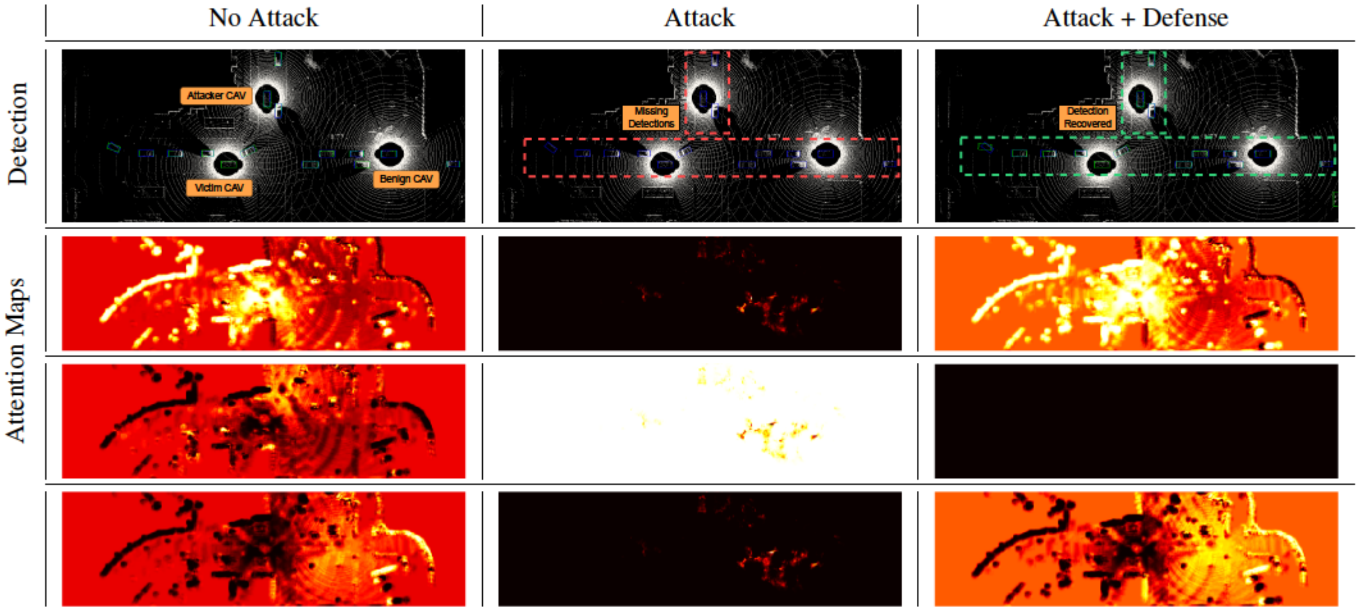
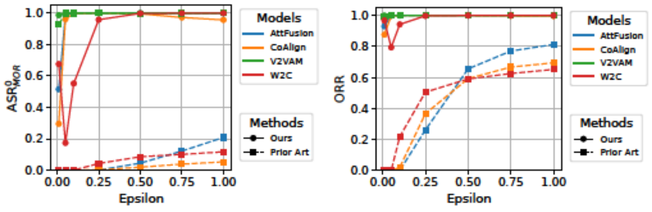


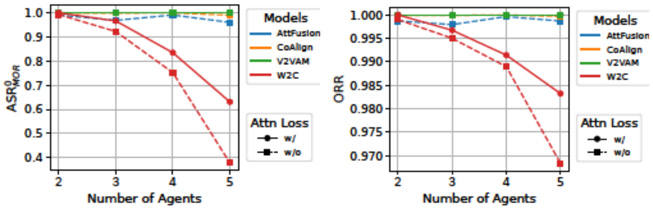
Figure 6: Example of detection results and victim's attention maps to others, under our attack and defense. (Blue bounding boxes: ground truth objects, green bounding boxes: detected objects. Attention maps top: victim, mid: attacker, bottom: benign agent.)



(a) ASR vs epsilon.

(b) ORR vs epsilon.

Figure 7: ASR and ORR increase with higher perturbation strengths, while SOMBRA achieve higher ASR and ORR compared to prior art with (over 100×) smaller perturbation.



(a) ASR vs number of agents.

(b) ORR vs number of agents.

Figure 8: Ablation on the attention-boosting loss ($\epsilon = 0.25$). Including the attention-boosting loss increases both the ASRs and ORRs, especially when the number of agents increases.

Even W2C, which exhibits a somewhat robust baseline, experiences a jump from effectively 0% success (baseline) to 55–77% for $\mathrm{ASR}^0_{\mathrm{MOR}}$ and $\mathrm{ASR}^1_{\mathrm{MOR}}$. These sharp increases confirm that attentive fusion can be exploited to remove *all* objects in the scene via subtle yet powerful perturbations.

Figure 7 extends these findings by showing how *less* perturbation (lower ϵ) is required for our method to achieve high attack efficacy, whereas the baseline needs (over 100×) more perturbation strength to approach similar removal rates. Both $\mathrm{ASR}^0_{\mathrm{MOR}}$ and ORR climb to near 100% significantly faster than the baseline, highlighting our approach's efficiency.

Ablation Study. The incorporation of the attention-boosting loss plays a crucial role in enhancing the effectiveness of SOMBRA while adhering to realistic perturbation constraints. To assess the impact of the attention-boosting loss component, we conduct an ablation study comparing the performance of our attacks with and without the attention manipulation. Figure 8 examines the impact of explicitly boosting the attacker's attention term by comparing our MOR results *with* (solid lines) and *without* (dashed lines) the attention-boosting loss. Across all four models, enabling the attention-boosting term yields consistently higher $\mathrm{ASR}^0_{\mathrm{MOR}}$ and ORR as the network size increases from 2 to 5 agents. Notably, W2C exhibits the sharpest decline when attention boosting is disabled—its $\mathrm{ASR}^0_{\mathrm{MOR}}$ drops from nearly 1.0 to below 0.8 at five agents. These results underscore that shaping the victim's attention distribution is crucial for maintaining high removal rates, even in scenarios with more cooperating agents.

6.3 Case Study: High-Density Traffic Scenario

To assess the scalability and robustness of SOMBRA in extreme conditions, we evaluate its performance in a high-density 'traffic jam' scenario involving 50 collaborating CAVs at a busy in-

Table 3: SOMBRA maintains high ASR and ORR in high-density traffic jam scenario with 50 CAVs, resisting dilution of attention paid to attacker due to increased number of agents.

Attack	No Attack	Baseline			SOMBRA		
Model	ORR	$\text{ASR}^0_{\text{MOR}}$ ↑	$\text{ASR}^1_{\text{MOR}}$ ↑	ORR↑	$\text{ASR}^0_{\text{MOR}}$ ↑	$\text{ASR}^1_{\text{MOR}}$ ↑	ORR↑
AttFusion	39.89%	0.00%	0.00%	56.89%	**47.82%**	**86.96%**	**99.28%**
CoAlign	10.95%	0.00%	0.00%	59.42%	**100.00%**	**100.00%**	**100.00%**
W2C	24.59%	0.00%	0.00%	28.26%	0.00%	0.00%	**71.01%**
V2VAM	1.79%	0.00%	33.33%	97.46%	**100.00%**	**100.00%**	**100.00%**

tersection, with the same attack hyperparameters. This setup is designed to test the attack's effectiveness as the number of participating agents N grows significantly, simulating a challenging scenario for the attacker due to naturally diluted attention. Details on this customized dataset are in Appendix E.

Impact of High Agent Density. As the number of agents increases in dense traffic with limited Field-of-View overlap, the contribution of each individual benign agent to the fused feature map tends to diminish. Table 3 shows the baseline ORR under no attack, indicating the percentage of ground truth objects missed even without adversarial interference. For models like AttFusion (39.89% ORR) and W2C (24.59% ORR), this inherent information loss is substantial. The attentive fusion mechanism naturally assigns lower weights to each agent as N grows, diluting the signals from individual benign vehicles, which can degrade overall perception performance.

Attack Performance. Despite the challenging scenario, Table 3 demonstrates that SOMBRA remains highly effective, outperforming the baseline attack across all tested CP models. For instance, on AttFusion, SOMBRA achieves a 99.28% ORR and an 86.96% $\text{ASR}^1_{\text{MOR}}$, compared to the baseline's 56.89% ORR and 0% ASR. Similarly, for CoAlign and V2VAM, SOMBRA achieves near-complete object removal (100% ASR/ORR). Even for W2C, where SOMBRA does not achieve complete removal of all objects, it still dramatically increases the ORR from 28.26% (baseline attack) to 71.01%. This indicates that by directly manipulating attention weights, SOMBRA can effectively counteract the signal dilution effect that hampers baseline attacks in dense environments.

Intriguingly, V2VAM exhibits the best benign performance (lowest ORR at 1.79%) yet is highly susceptible to SOMBRA (100% ASR/ORR). This vulnerability stems from its unique design that incorporates an agent-wise max-out operation added to the fused feature, which, while potentially preserving strong benign signals, can be readily hijacked by a dominant adversarial signal from SOMBRA. This highlights a fundamental trade-off in CP fusion design between robustness against signal dilution and vulnerability to attention manipulation.

Adversarial Signal Propagation Analysis. In CP, higher attention scores imply higher contribution of the adversarial signal to the victim's final fused feature. To understand how SOMBRA maintains its influence as N grows, Figure 9 visualizes the average attention score the victim assigns to the attacker. The results show that incorporating the attention-boosting loss enables the attacker to sustain significantly higher attention scores compared to omitting this loss term. Even with 50 agents, SOMBRA with $\mathcal{L}_{attn}$ ensures the attacker hijacks substantial attention (e.g., >0.25 for $\varepsilon = 0.01$, >0.9 for $\varepsilon = 0.1$), preventing the adversarial signal from being diluted. This sustained attention directly translates to the high attack success, reinforcing the central role of attention manipulation in SOMBRA's effectiveness, especially in large-scale CP systems.

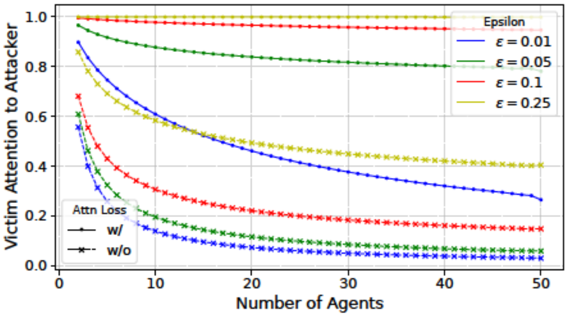


Figure 9: Victim attention paid to attacker in the traffic jam scenario, where the incorporation of attention-boosting loss ensures that the malicious signal dominates the attention share and propagates effectively through the attentive fusion network. Attacks without attention-boosting loss get diluted attention closer to the expected average of $1/N$.

6.4 Evaluation of Defenses

Defense Against TOR. Table 4 demonstrates the performance of LUCIA and ROBOSAC across different target visibility scenarios. LUCIA achieves consistently higher DSR compared to ROBOSAC. For instance, in the beyond-LoS scenario with CoAlign, DSR increases from 1.51% (ROBOSAC) to 59.97% with our defense. Similarly, AttFusion improves from 1.46% to 52.75% for beyond-LoS. This indicates that our defense effectively excludes adversarial features while preserving cross-agent consistency. For beyond-LoS scenarios, where targets can only be detected through CP, remain particularly challenging for ROBOSAC. LUCIA, however, achieves significantly higher DSR and maintains strong mAP. For example, W2C improves from 41.19% DSR (ROBOSAC) to 71.92% with our defense while also achieving higher mAP. ROBOSAC struggles for TOR due to its design that accepts a detection result if it falls within a predefined similarity threshold with local perception. It incurs a rigid tradeoff between trust and distrust that either makes it vulnerable to subtle manipulations by one adversarial source or denial of collaboration even under benign errors.

Defense Against MOR. Table 5 shows the defense effectiveness under the MOR attack. LUCIA significantly reduces ORR

Table 4: Defense results on targeted object removal. In both visibility cases, Lucia achieves notably higher DSRs compared to ROBOSAC, indicating it better restores the detection of the target object under attacks. When the perception of a single object is adversarially manipulated, ROBOSAC's fail to identify the anomaly as it is based on a predefined result similarity threshold.

Target Visibility	Within Victim LoS						Beyond Victim LoS					
Defense	ROBOSAC			Lucia			ROBOSAC			Lucia		
Model	DSR↑	mAP@0.5	mAP@0.7	DSR↑	mAP@0.5	mAP@0.7	DSR↑	mAP@0.5	mAP@0.7	DSR↑	mAP@0.5	mAP@0.7
AttFusion	3.66%	0.89	0.78	**93.95%**	0.84	0.72	1.46%	0.89	0.78	**52.75%**	0.84	0.72
CoAlign	3.17%	0.88	0.80	**94.93%**	0.85	0.76	1.51%	0.88	0.79	**59.97%**	0.85	0.76
Where2comm	81.67%	0.71	0.46	**94.69%**	0.85	0.66	41.19%	0.70	0.45	**71.92%**	0.85	0.65
V2VAM	7.11%	0.81	0.66	**85.86%**	0.82	0.76	2.44%	0.82	0.66	**37.68%**	0.82	0.76

Table 5: Defense evaluation results for mass object removal. Lucia achieves lower ORR while restoring higher overall detection performance than ROBOSAC.

Method	No Defense			ROBOSAC			Lucia		
Model	ORR	mAP @0.5	mAP @0.7	ORR↓	mAP @0.5	mAP @0.7	ORR↓	mAP @0.5	mAP @0.7
AttFusion	100%	0	0	10.52%	0.71	0.55	**7.52%**	0.84	0.72
CoAlign	99.99%	0	0	8.03%	0.72	0.59	**5.73%**	0.85	0.76
W2C	94.26%	0	0	5.65%	0.75	0.51	**4.52%**	0.85	0.65
V2VAM	100.00%	0	0	**12.97%**	0.77	0.63	19.09%	0.82	0.76

Table 6: Defense evaluation results for BIM attack. Lucia restore the overall detection performance and achieves consistently higher mAP than ROBOSAC.

Method	No Defense		ROBOSAC		Lucia	
Model	mAP @0.5	mAP @0.7	mAP @0.5	mAP @0.7	mAP @0.5	mAP @0.7
AttFusion	0	0	0.71	0.55	**0.84**	**0.72**
CoAlign	0	0	0.72	0.59	**0.85**	**0.76**
Where2comm	0	0	0.75	0.51	**0.85**	**0.65**
V2VAM	0	0	0.77	0.63	**0.82**	**0.76**

compared to ROBOSAC. For AttFusion, ORR drops from 100% (no defense) and 10.52% (ROBOSAC) to 7.52% with Lucia. Similarly, CoAlign achieves 5.73% ORR, compared to 8.03% for ROBOSAC. This demonstrates our defense's ability to preserve the majority of objects in detection outputs, even under severe adversarial conditions. Our defense consistently restores mAP values to levels approaching the benign baseline. For instance, W2C achieves 0.85 mAP@0.5 with our defense compared to 0.75 for ROBOSAC, showing superior robustness against object removal.

Generalization to Other Attacks. Table 6 shows the results of Lucia in untargeted attacks using BIM, where the attacker performs object spoofing and removal simultaneously. For all models, mAPs with our defense are consistently higher than those with ROBOSAC, similar to the results from MOR. This indicates that our defense generalizes and defends against other types of attack including object spoofing attacks.

Defense Robustness. Note that we tested our defense against attacks where the perturbation is optimized with 10 iterations and various different learning rates ϵ ranging from 0.01 to 1. Results remain consistent across these settings including where $\epsilon = 0.01$, demonstrating the robustness of our method. Upon investigation, our defense consistently identified the attacker with a trustworthiness score of 0 and assigned a trustworthiness score of 1 to benign agents, achieving 100% TPR and 0% FPR across all evaluations. This is reflected in Figure 6, where the victim's attention paid to the attacker, when our defense is implemented, is 0 across all areas due to a zero trust being assigned to the attacker. Unlike ROBOSAC, which suffers from iterative sampling and often reduces to single-vehicle perception, our defense maximizes the information available by excluding only the attacker and neutralizing the adversarial feature efficiently. This ensures higher detection accuracy while maintaining computational efficiency.

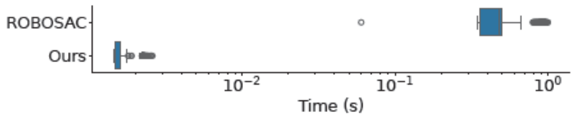


Figure 10: Per frame computation overhead of defenses on our autonomous vehicle testbed. Our defense incurs 300× less computation overhead than prior art.

Computation Overhead. Figure 10 highlights the computational efficiency of Lucia compared to ROBOSAC, on our real-world autonomous vehicle testbed. Lucia maintains tightly clustered inference times around 0.001 seconds, introducing only negligible computational overhead. ROBOSAC incurs significantly higher additional overhead to the system, ranging from 0.3 to 1 seconds, which is orders of magnitude slower than Lucia, due to its iterative sampling mechanism. This variability and high overhead make it unsuitable for time-critical applications, while our defense consistently operates well within the 100 ms latency budget required for autonomous driving [31].

6.5 Defense against Adaptive Attacker

In addition to standard attacker models, we evaluate the strongest adaptive attacker under our threat model. We consider the attacker to be aware of the defense and aims to minimize the L_1 distance between the perturbed message

Table 7: Defense evaluation results for adaptive attacks. Lucia remains resilient against adaptive attacks, that it maintains high DSR against target removal and low ORR for mass removal, and restores the overall perception performance.

Attack	Target Object Removal			Mass Object Removal		
Model	DSR↑	mAP @0.5	mAP @0.7	ORR↓	mAP @0.5	mAP @0.7
AttFusion	86.74%	0.84	0.72	7.23%	0.84	0.72
CoAlign	87.47%	0.86	0.76	8.84%	0.82	0.73
W2C	90.35%	0.85	0.65	4.53%	0.85	0.65
V2VAM	55.73%	0.86	0.83	52.25%	0.42	0.39

$X_a^t + \delta^t$ with the victim's feature X_v^t to increase its consistency hence trustworthiness score. Specifically, we assume that the attacker gains access to the victim's exact feature map X_v^t in real-time and substitutes its own feature $X_a^t = X_v^t$. To craft adversarial perturbations, the attacker employs PGD optimization with a small learning rate $\varepsilon = 0.01$ for 10 iterations based on the exact same feature used by the victim.

Table 7 presents the defense results under this adaptive attack. Lucia maintains robust performance. For instance, AttFusion still achieves an 86.74% DSR in TOR and a low 7.23% ORR for MOR. The average trust score assigned to the adaptive attacker remains close to zero (e.g., < 0.001 in AttFusion, < 0.08 in CoAlign), indicating that while the attacker's feature mimics the victim's state, the inconsistency caused by the minimum perturbation makes it fail to blend in sufficiently to avoid down-weighting by the defense. Interestingly, V2VAM appears to be more subject to attacks (either adaptive or non-adaptive) and harder to defend. We provide discussions on the robustness of V2VAM in Appendix F.

7 Discussions and Future Work

Comparison with RANSAC. Lucia shares conceptual parallels with RANSAC-based methods, as both approaches implicitly determine the 'level of collaboration' with other agents. Specifically, the subset sampling operation in RANSAC-based methods is equivalent to a random subset of agents being assigned trust scores of one and zero for excluded agents. Such a binary inclusion rule inherently discards the contributions of randomly excluded agents entirely. In contrast, our framework is more general and uses soft trust scores to scale each agent's contribution dynamically. This approach accommodates more informed and subtle ways to determine these scores beyond random binary assignment. Yet, these two paradigms have the potential for a synergetic integration. We will explore possible interplay and multi-layered trust defenses as future work.

Limitations. Our evaluations focus on scenarios with a single compromised vehicle. While this aligns with the scope of many prior studies [27, 57, 72], real-world deployments may face coordinated attacks involving multiple adversaries. While existing CP defenses [29, 72] struggle against single attackers, as we detail in Appendix C, extending our defense to handle collusion among multiple attackers is an important direction for future work.

The evaluation is conducted using our real-world vehicle testbed, however, the input is based on the benchmark dataset OPV2V [68] collected with the digital twin simulator CARLA [11], which captures diverse yet simulated conditions. Real-world inputs introduce additional challenges such as dynamic traffic densities, environmental noise, and intermittent communication failures. Nevertheless, existing real-world CP benchmark datasets either has exclusive access [69] or only involves two CAVs [67]. Therefore, further validation in field deployments or emerging real-world datasets is necessary to evaluate performance under such complexities, especially in the setting of heterogeneous CP systems.

Potential Software/Hardware Security Controls. While Sombra assumes that the attacker bypasses message authentication using compromised credentials, certain platform-level security measures might hinder such attacks. Technologies such as trusted execution environments (TEEs) [18] could be used to protect the integrity and confidentiality of the CP models. If the model parameters and critical computations are isolated within a TEE, obtaining the white-box knowledge required by the attacker becomes harder. Additionally, secure boot mechanisms [52] and runtime integrity monitoring could help detect unauthorized modifications [24]. However, designing and deploying such defenses effectively against a determined insider attacker who has already compromised the host system remains an active area of research [49], as TEEs themselves can have vulnerabilities [72], and system-level compromises might bypass monitoring. Developing a robust security platform specifically tailored for CP systems is a crucial direction for future work.

8 Conclusion

We explore how attention mechanisms–the core component of cooperative perception (CP) systems–can be exploited by attackers, but also how it can be harnessed for practical defenses. We present Sombra, a highly effective and stealthy object removal attack on CP, and Lucia, a lightweight defense mechanism that proactively mitigates adversarial features. Our evaluations on four state-of-the-art CP algorithms demonstrated that Sombra surpasses existing methods by over 90% in attack efficacy, consistently achieving above 99% success in both targeted and mass object-removal scenarios—while requiring 1/100 the perturbation strength of prior work. Meanwhile, Lucia achieved up to 94.93% success in blocking targeted removal attacks, reducing mass removal rates by more than 90%, and cutting defense overhead by more than 300×.

Ethics Considerations

CP systems are integral to the future of autonomous vehicles, and their success directly impacts road safety. While we develop effective attacks, our understanding of the unique design of CP systems enables us to develop and strengthen defensive strategies, ensuring that CP systems are safer for real-world deployment, which is the goal of this work. All evaluations are conducted using benchmark dataset under controlled in-vehicle environment, ensuring no harm to public systems. By demonstrating both vulnerabilities and effective defenses, we contribute to improving safety in the critical application of automated driving. The results presented in this work underscore the deployability of our defense in resource-constrained environments and its relevance to industry and regulators.

Open Science

In compliance with the open science policy, our research exclusively uses the publicly available benchmark dataset OPV2V [68] and our customized case study data in the same format, with the corresponding pre-trained models released by the authors [65]. Moreover, to enable full transparency and facilitate future research, we open-source our implementation (including scripts for data preprocessing, attack generation, and defense integration) along with instructions to replicate our results. This ensures that the broader community can independently validate, build on, and extend our work. Our code and the collected case study dataset [62] are permanently available at: https://doi.org/10.5281/zenodo.15523768.

Acknowledgments

We thank our shepherd and the anonymous reviewers for their valuable suggestions. This work has been partially supported by the FY25 Eighteenth Mile TRIF Funding from the University of Arizona, and by the National Science Foundation (NSF) through grants CNS-2144645 and IIS-2229876.

References

[1] 3rd Generation Partnership Project (3GPP). Security aspects for LTE-based V2X services. Technical Specification TS 33.185, 3GPP, 2018. https://www.3gpp.org/DynaReport/33-series.htm.

[2] Hamza Ijaz Abbasi, Ralph Gholmieh, Tien Viet Nguyen, Shailesh Patil, and Jim Misener. LTE-V2X (C-V2X) performance in congested highway scenarios. In *IEEE ICC*, 2022.

[3] Mohammad Raashid Ansari, Jean-Philippe Monteuuis, Jonathan Petit, and Cong Chen. V2X misbehavior and collective perception service: Considerations for standardization. In *IEEE CSCN*, 2021.

[4] Yulong Cao, S Hrushikesh Bhupathiraju, Pirouz Naghavi, Takeshi Sugawara, Z Morley Mao, and Sara Rampazzi. You can't see me: Physical removal attacks on LiDAR-based autonomous vehicles driving frameworks. In *USENIX Security Symposium*, 2023.

[5] Yulong Cao, Chaowei Xiao, Benjamin Cyr, Yimeng Zhou, Won Park, Sara Rampazzi, Qi Alfred Chen, Kevin Fu, and Z Morley Mao. Adversarial sensor attack on LiDAR-based perception in autonomous driving. In *ACM CCS*, 2019.

[6] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE S&P*, 2017.

[7] Ioannis Chatziioannou, Stefanos Tsigdinos, Panagiotis G. Tzouras, Alexandros Nikitas, and Efthimios Bakogiannis. Connected and autonomous vehicles and infrastructure needs: Exploring road network changes and policy interventions. In *Deception in Autonomous Transport Systems: Threats, Impacts and Mitigation Policies*. Springer, 2024.

[8] Qi Chen, Xu Ma, Sihai Tang, Jingda Guo, Qing Yang, and Song 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.

[9] Qi Chen, Sihai Tang, Qing Yang, and Song Fu. Cooper: Cooperative Perception for Connected Autonomous Vehicles Based on 3D Point Clouds . In *IEEE ICDCS*, 2019.

[10] Roderick Currie. Hacking the can bus: basic manipulation of a modern automobile through can bus reverse engineering. *SANS Institute*, 2017.

[11] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, 2017.

[12] European Telecommunications Standards Institute (ETSI). ETSI EN 303 613 V1.1.1: Intelligent Transport Systems (ITS); Congestion Control Mechanisms for C-V2X PC5 Interface; Access Layer Part, 2019. [Online].

[13] Federal Communications Commission (FCC). Dedicated Short Range Communications (DSRC) Service, 2025. [Online].

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

[15] Maanak Gupta, James Benson, Farhan Patwa, and Ravi Sandhu. Dynamic groups and attribute-based access control for next-generation smart cars. In *ACM Conf. on Data and App. Sec. and Pri.*, 2019.

[16] Zhongyuan Hau, Soteris Demetriou, and Emil C Lupu. Using 3D shadows to detect object hiding attacks on autonomous vehicle perception. In *IEEE Security and Privacy Workshops (SPW)*, 2022.

[17] Senkang Hu, Yihang Tao, Guowen Xu, Yiqin Deng, Xianhao Chen, Yuguang Fang, and Sam Kwong. CP-Guard: Malicious agent detection and defense in collaborative bird's eye view perception. *arXiv:2412.12000*, 2024.

[18] Shengtuo Hu, Qi Alfred Chen, Jiwon Joung, Can Carlak, Yiheng Feng, Z Morley Mao, and Henry X Liu. Cvshield: Guarding sensor data in connected vehicle with trusted execution environment. In *ACM Workshop on Automotive and Aerial Vehicle Security*, 2020.

[19] Yue Hu, Shaoheng Fang, Zixing Lei, Yiqi Zhong, and Siheng Chen. Where2comm: Communication-efficient collaborative perception via spatial confidence maps. *NeurIPS*, 2022.

[20] Zilong Huang, Xinggang Wang, Lichao Huang, Chang Huang, Yunchao Wei, and Wenyu Liu. Ccnet: Criss-cross attention for semantic segmentation. In *CVPR*, 2019.

[21] IEEE. Standard for wireless access in vehicular environments–security services for applications and management messages. *Std 1609.2*, 2016.

[22] DataSpeed Inc. Dataspeed home. https://www.dataspeedinc.com/, 2024. [Online].

[23] Zizhi Jin, Xiaoyu Ji, Yushi Cheng, Bo Yang, Chen Yan, and Wenyuan Xu. Pla-LiDAR: Physical laser attacks against lidar-based 3D object detection in autonomous vehicle. In *IEEE S&P*, 2023.

[24] Philip Koopman and Michael Wagner. Challenges in autonomous vehicle testing and validation. *SAE International Journal of Transportation Safety*, 2016.

[25] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*. Chapman and Hall/CRC, 2018.

[26] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, 2019.

[27] Jinlong Li, Baolu Li, Xinyu Liu, Jianwu Fang, Felix Juefei-Xu, Qing Guo, and Hongkai Yu. Advgps: Adversarial gps for multi-agent perception attack. In *ICRA*, 2024.

[28] Jinlong Li, Runsheng Xu, Xinyu Liu, Jin Ma, Zicheng Chi, Jiaqi Ma, and Hongkai Yu. Learning for vehicle-to-vehicle cooperative perception under lossy communication. *IEEE Trans. on Intelligent Vehicles*, 2023.

[29] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *ICCV*, 2023.

[30] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022.

[31] Shih-Chieh Lin, Yunqi Zhang, Chang-Hong Hsu, Matt Skach, Md E. Haque, Lingjia Tang, and Jason Mars. The architectural implications of autonomous driving: Constraints and acceleration. In *ACM International Conference on Architectural Support for Programming Languages and Operating Systems*, 2018.

[32] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017.

[33] Zhiping Lin, Liang Xiao, Hongyi Chen, Zefang Lv, Yunjun Zhu, Yanyong Zhang, and Yong-Jin Liu. Edge-assisted collaborative perception against jamming and interference in vehicular networks. *IEEE Transactions on Wireless Communications*, 24, 2025.

[34] Todd Litman. Autonomous vehicle implementation predictions: Implications for transport planning. *Victoria Transport Policy Institute*, 2023.

[35] Xingbin Liu, Huafeng Kuang, Hong Liu, Xianming Lin, Yongjian Wu, and Rongrong Ji. Latent feature relation consistency for adversarial robustness. *arXiv preprint arXiv:2303.16697*, 2023.

[36] Giulio Lovisotto, Nicole Finnie, Mauricio Munoz, Chaithanya Kumar Mummadi, and Jan Hendrik Metzen. Give me your attention: Dot-product attention considered harmful for adversarial patch robustness. In *CVPR*, 2022.

[37] Yifan Lu, Yue Hu, Yiqi Zhong, Dequan Wang, Siheng Chen, and Yanfeng Wang. An extensible framework for open heterogeneous collaborative perception. In *ICLR*, 2024.

[38] Yifan Lu, Quanhao Li, Baoan Liu, Mehrdad Dianati, Chen Feng, Siheng Chen, and Yanfeng Wang. Robust collaborative 3D object detection in presence of pose errors. In *ICRA*, 2023.

[39] Yujia Luo. Time constraints and fault tolerance in autonomous driving systems. *Tech. rep, Tech. Rep*, 2019.

[40] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.

[41] Yanmao Man, Raymond Muller, Ming Li, Z. Berkay Celik, and Ryan Gerdes. That person moves like a car: Misclassification attack detection for autonomous systems using spatiotemporal consistency. In *USENIX Security Symposium*, 2023.

[42] Carsten Maple, Matthew Bradbury, Anh Tuan Le, and Kevin Ghirardello. A connected and autonomous vehicle reference architecture for attack surface analysis. *Applied Sciences*, 9(23), 2019.

[43] Jean-Philippe Monteuuis, Jonathan Petit, Jun Zhang, Houda Labiod, Stefano Mafrica, and Alain Servel. Attacker model for connected and automated vehicles. In *ACM Computer Science in Car Symposium*, 2018.

[44] Raymond Muller, Yanmao Man, Z Berkay Celik, Ming Li, and Ryan Gerdes. Physical hijacking attacks against object trackers. In *ACM CCS*, 2022.

[45] Raymond Muller, Yanmao Man, Ming Li, Ryan Gerdes, Jonathan Petit, and Z Berkay Celik. {VOGUES}: Validation of object guise using estimated components. In *USENIX Security Symposium*, 2024.

[46] Raymond Muller, Ruoyu Song, Chenyi Wang, Yuxia Zhan, Jean-Phillipe Monteuuis, Yanmao Man, Ming Li, Ryan Gerdes, Jonathan Petit, and Z. Berkay Celik. Investigating physical latency attacks against camera-based perception. In *IEEE S&P*, 2025.

[47] U.S. Department of Transportation. Connected and Automated Vehicles - Transportation Planning Capacity Building Program — planning.dot.gov. https://www.planning.dot.gov/planning/topic_CVAV.aspx.

[48] Alina Oprea and Apostol Vassilev. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical report, National Institute of Standards and Technology, 2023.

[49] Simon Parkinson, Paul Ward, Kyle Wilson, and Jonathan Miller. Cyber threats facing autonomous and connected vehicles: Future challenges. *IEEE Transactions on Intelligent Transportation Systems*, 18, 2017.

[50] Hang Qiu, Fawad Ahmad, Ramesh Govindan, Marco Gruteser, Fan Bai, and Gorkem Kar. Augmented vehicular reality: Enabling extended vision for future vehicles. In *ACM International Workshop on Mobile Computing Systems and Applications*, 2017.

[51] Andreas Rauch, Felix Klanner, Ralph Rasshofer, and Klaus Dietmayer. Car2x-based perception in a high-level fusion architecture for cooperative perception systems. In *IEEE Intelligent Vehicles Symposium*, 2012.

[52] Steffen Sanwald, Liron Kaneti, Marc Stöttinger, and Martin Böhner. Secure boot revisited: challenges for secure implementations in the automotive domain. *SAE International Jour. of Trans. Cybersec. and Priv.*, 2020.

[53] Vasu Sharma, Ankita Kalra, Sumedha Chaudhary Vaibhav, Labhesh Patel, and Louis-Phillippe Morency. Attend and attack: Attention guided adversarial attacks on visual question answering models. In *Conf. Neural Inf. Process. Syst. Workshop Secur. Mach. Learn*, 2018.

[54] Eduardo Soares, Plamen Angelov, and Neeraj Suri. Similarity-based deep neural network to detect imperceptible adversarial attacks. In *IEEE Symposium Series on Computational Intelligence*, 2022.

[55] Ruoyu Song, Muslum Ozgur Ozmen, Hyungsub Kim, Raymond Muller, Z Berkay Celik, and Antonio Bianchi. Discovering adversarial driving maneuvers against autonomous vehicles. In *USENIX Security Symposium*, 2023.

[56] Jiachen Sun, Yulong Cao, Qi Alfred Chen, and Z Morley Mao. Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In *USENIX Security Symposium*, 2020.

[57] James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *CVPR*, 2020.

[58] James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *ICCV*, 2021.

[59] Rens Wouter van der Heijden, Stefan Dietzel, Tim Leinmüller, and Frank Kargl. Survey on misbehavior detection in cooperative intelligent transportation systems. *IEEE Communications Surveys & Tutorials*, 2019.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Chenyi Wang, Yanmao Man, Raymond Muller, Ming Li, Z. Berkay Celik, Ryan Gerdes, and Jonathan Petit. Physical id-transfer attacks against multi-object tracking via adversarial trajectory. In *Annual Computer Security Applications Conference (ACSAC)*. IEEE Computer Society, 2024.

[62] Chenyi Wang, Raymond Muller, Ruoyu Song, Jean-Philippe, Z. Berkay Celik, Yanmao Man, Jonathan Petit, Ryan Gerdes, and Ming Li. Usenix security 25' cycle2-592-attention-exploit- artifact-evaluation, May 2025.

[63] Philip Wendland and Guenter Schaefer. Feedback-based hidden-terminal mitigation for distributed scheduling in cellular V2X. In *IFIP Networking Conference*, 2020.

[64] Qi Xia and Qian Chen. Moiré injection attack(mia): Compromising autonomous vehicle safety via exploiting camera's color filter array (cfa) to inject hidden traffic sign. In *Annual Computer Security Applications Conference (ACSAC)*. IEEE Computer Society, 2024.

[65] Runsheng Xu. Opencood. https://github.com/DerrickXuNu/OpenCOOD, 2023.

[66] Runsheng Xu, Yi Guo, Xu Han, Xin Xia, Hao Xiang, and Jiaqi Ma. Opencda: An open cooperative driving automation framework integrated with co-simulation. In *2021 IEEE International Intelligent Transportation Systems Conference*, 2021.

[67] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, Hongkai Yu, Bolei Zhou, and Jiaqi Ma. V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *CVPR*, 2023.

[68] Runsheng Xu, Hao Xiang, Xin Xia, Xu Han, Jinlong Li, and Jiaqi Ma. OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication. In *ICRA*, 2022.

[69] Haibao Yu, Yizhen Luo, Mao Shu, Yiyi Huo, Zebang Yang, Yifeng Shi, Zhenglong Guo, Hanyu Li, Xing Hu, Jirui Yuan, et al. DAIR-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection. In *CVPR*, 2022.

[70] Haibao Yu, Yingjuan Tang, Enze Xie, Jilei Mao, Ping Luo, and Zaiqing Nie. Flow-based feature fusion for vehicle-infrastructure cooperative 3d object detection. *NeurIPS*, 2024.

[71] Zhiyuan Yu, Ao Li, Ruoyao Wen, Yijia Chen, and Ning Zhang. Physense: Defending physically realizable attacks for autonomous systems via consistency reasoning. In *ACM CCS*, 2024.

[72] Qingzhao Zhang, Shuwei Jin, Jiachen Sun, Xumiao Zhang, Ruiyang Zhu, Qi Alfred Chen, and Z Morley Mao. On data fabrication in collaborative vehicular perception: Attacks and countermeasures. In *USENIX Security Symposium*, 2024.

Appendix A Transfer Attack

To further assess the practicality and robustness of SOMBRA, particularly in scenarios where the attacker may not possess exact knowledge of the victim's deployed model (a black-box setting), we evaluate its transferability. In a transfer attack, adversarial perturbations are generated using a known surrogate model (source) but are then applied against a different, potentially unknown model used by the victim (destination).

Experimental Setup. For each source model, we generate adversarial perturbations using SOMBRA under white-box assumptions with the same optimization hyperparameters in Section 6. These generated perturbations are then added to the attacker's feature map and transmitted to a victim vehicle assumed to be running one of the four models.

Table 8: Transfer attack results, where SOMBRA demonstrates high transferability across state-of-the-art CP models

Destination Model	AttFusion		CoAlign		Where2comm		V2VAM	
Source Model	ASR^0_{MOR} ↑	ORR↑	ASR^0_{MOR} ↑	ORR↑	ASR^0_{MOR} ↑	ORR↑	ASR^0_{MOR} ↑	ORR↑
AttFusion	99.95%	100.00%	99.85%	99.99%	99.66%	99.98%	99.71%	99.97%
CoAlign	99.81%	99.99%	99.76%	99.99%	99.81%	99.98%	99.66%	99.98%
Where2comm	52.14%	93.72%	53.46%	94.15%	55.24%	94.26%	52.05%	93.95%
V2VAM	99.95%	100.00%	99.95%	100.00%	99.95%	100.00%	100.00%	100.00%

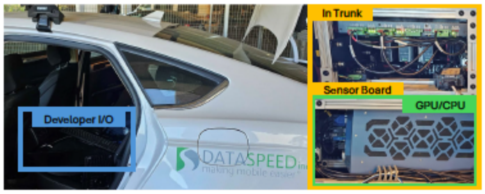


Figure 11: Our vehicle testbed used for evaluation.

Analysis. The results in Table 8 indicate that SOMBRA, by targeting the fundamental attention mechanism, generates highly transferable adversarial examples. Perturbations crafted using AttFusion, CoAlign, or V2VAM as source models achieve near-perfect MOR success (>99% ASR^0_{MOR} and ORR) when applied against any of the other three models, demonstrating robustness in black-box settings. The performance using Where2comm as a source model is lower, suggesting its learned features or attention patterns might be less generalizable, although it still achieves a high ORR (>93%) when transferred. Conversely, Where2comm as a destination is still highly vulnerable to attacks generated from the other three models. The minimal difference between diagonal (white-box) and off-diagonal (black-box) results for the top-performing models underscores the practical threat posed by SOMBRA, as precise knowledge of the victim's model is often not required for a successful attack. This transferability reinforces the findings discussed in Section 3 regarding the feasibility of the attack beyond strict white-box assumptions.

Appendix B Derivation of Attacker Influence Bound

This section provides a detailed derivation for the upper bound on the attention score $\alpha_a(x)$ assigned by a victim vehicle V_v to an attacker V_a, as presented in Eq. 6.

We start with the definition of the attention score computed via scaled dot-product attention for the victim vehicle V_v focusing on agent V_j at spatial location x:

$$\alpha_j(x) = \frac{\exp(S_{vj}(x))}{\sum_{k=1}^{N} \exp(S_{vk}(x))} \tag{17}$$

where N is the total number of participating agents, and $S_{vk}(x)$ is the pre-softmax score between the victim's feature vector

$X_v(x)$ and agent k's feature vector $X_k(x)$:

$$S_{vk}(x) = \frac{X_v(x)^T X_k(x)}{\sqrt{C}} \tag{18}$$

Here, C is the number of channels in the feature vectors.

Let V_a be the attacker, who modifies their feature vector to $X'_a(x) = X_a(x) + \delta$, where δ is the adversarial perturbation. The victim V_v receives $X'_a(x)$ from the attacker and benign features $X_j(x)$ from other agents $j \neq a$. The attention score the victim assigns to the attacker is:

$$\alpha_a(x) = \frac{\exp(S'_{va}(x))}{\exp(S'_{va}(x)) + \sum_{j\neq a} \exp(S_{vj}(x))} \tag{19}$$

where $S'_{va}(x) = \frac{X_v(x)^T (X_a(x)+\delta)}{\sqrt{C}} = S_{va}(x) + \frac{X_v(x)^T \delta}{\sqrt{C}}$.

We make the following assumptions: (1) The scaled dot-product similarity between the victim and any other agent j (including the victim itself, $j = v$, and the attacker before perturbation, $j = a$) is bounded: $S_{vj}(x) \leq \beta$ for all j, which holds due to bounded (a) LiDAR point count and intensity, and (b) feature extractor weights. (2) The attacker uses PGD optimization to craft the perturbation δ with the goal of maximizing the victim's attention towards the attacker, subject to a perturbation constraint L_∞: $\|\delta\|_\infty \leq \varepsilon_{max}$.

Under the L_∞ constraint $\|\delta\|_\infty \leq \varepsilon_{max}$, the term $X_v(x)^T\delta$ is maximized when $\delta = \varepsilon_{max} \cdot \text{sign}(X_v(x))$. In this case, the maximum value is:

$$\begin{aligned}\max_{\|\delta\|_\infty \leq \varepsilon_{max}} X_v(x)^T\delta &= X_v(x)^T(\varepsilon_{max} \cdot \text{sign}(X_v(x))) \\ &= \varepsilon_{max} \sum_{c=1}^{C} |X_{v,c}(x)| = \varepsilon_{max} \|X_v(x)\|_1 \end{aligned} \tag{20}$$

where $\|X_v(x)\|_1$ is the L_1 norm of the victim's feature vector at location x. Let $\Delta = \frac{\varepsilon_{max}\|X_v(x)\|_1}{\sqrt{C}}$ represent the increase in the scaled dot-product similarity due to the optimally crafted perturbation. Then, the similarity score for the attacker becomes:

$$S'_{va}(x) = S_{va}(x) + \frac{X_v(x)^T\delta}{\sqrt{C}} \leq \beta + \Delta \tag{21}$$

where we use the bound β for the original similarity $S_{va}(x)$.

Now, we substitute these into the softmax expression for $\alpha_a(x)$: $\alpha_a(x) \approx \frac{\exp(\beta+\Delta)}{\exp(\beta+\Delta)+\sum_{j\neq a}\exp(S_{vj}(x))}$. Using assumption 1, we can approximate the summation term:

$$\sum_{j\neq a} \exp(S_{vj}(x)) \leq \sum_{j\neq a} \exp(\beta) = (N-1)e^{\beta} \tag{22}$$

This assumes that all other $N-1$ agents (benign agents and the victim itself) have roughly the same baseline similarity β with the victim. Plugging this back, we get:

$$\alpha_a(x) \leq \frac{e^{\beta+\Delta}}{e^{\beta+\Delta} + (N-1)e^{\beta}} \tag{23}$$

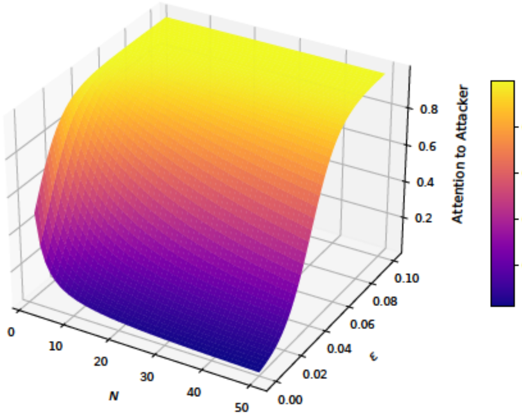


Figure 12: Theoretical upper bound of attention paid to attacker (α_a), with $\beta = 5, C = 384$.

Factoring out e^{β} from the numerator and denominator yields the final approximate bound:

$$\alpha_a(x) \leq \frac{e^{\Delta}}{e^{\Delta} + N - 1} \tag{24}$$

where $\Delta = \frac{\varepsilon_{max}\|X_v(x)\|_1}{\sqrt{C}}$.

Discussion. This derived bound provides insight into how the attacker's influence scales. The term e^{Δ} shows the exponential dependence on the perturbation's effectiveness (Δ), which itself depends on the allowed perturbation magnitude (ε_{max}), the victim's feature vector's magnitude ($\|X_v(x)\|_1$), and inversely on the square root of the feature dimension ($\sqrt{C}$). The $N-1$ term in the denominator reflects the dilution effect from other participants; as N increases, the attacker needs a larger Δ to maintain the same level of attention. SOMBRA's attention-boosting loss term ($\mathcal{L}_{attn}$) is explicitly designed to maximize the attacker's attention share, effectively maximizing Δ during optimization, thus making the attack potent even when N is large, as demonstrated empirically in Section 6.3 and Figure 9. This analysis underscores the criticality of the attention mechanism as an attack vector in CP systems.

Appendix C Hyperparameter Sensitivity of RANSAC-Based Defenses

As the representative RANSAC-based defense for CP, ROBOSAC's performance is heavily influenced by its hyperparameters, particularly the assumed attacker-to-benign ratio η, sampling budget N, and resulting number of randomly sampled collaborators (corresponding to different levels of information utilization rates). These parameters directly affect its ability to detect adversarial agents while maintaining effective collaboration among benign agents.

Attacker Ratio. ROBOSAC assumes a known attacker ratio, which dictates how many agents are treated as potentially compromised. If the ratio is underestimated, adversarial subsets may evade detection, while overestimation risks discarding benign agents. In real-world settings with dynamic networks, accurately estimating this ratio is highly challenging where the attacker-ratio estimation is subject to an adaptive attacker.

Sampling Budget. The sampling budget N, representing the allowed number of iterations per perception cycle, determines the maximum number of subsets to be evaluated to find a robust collaboration group. A higher sampling budgets increase robustness but imposes significant computational overhead, especially as the number of agents grows. As the sampling budget grows exponentially with respect to both the attacker ratio and number of collaborators demanded, under the stringent constraints of end-to-end AV delay and the restricted computation resources, ROBOSAC would refuses to sample anyone even a single attacker presence is suspected under a practical setting.

Number of Collaborators. ROBOSAC sacrifices information utilization by conservatively excluding agents during robust sampling. The maximum achievable utilization rate declines with increasing agent subsets, as only a fraction of agents are included in the final collaboration. This underutilization limits the benefits of cooperative perception, particularly in adversarial scenarios.

Similarity Threshold. The similarity threshold ε between two sets of inference results, in ROBOSAC's hyperparameter tuning impacts the acceptance of detection results. The Jaccard index was used to evaluate the consistency between bounding box sets by calculating the ratio of matched bounding boxes to the total number of bounding boxes across agents minus the matched ones. A higher similarity threshold ensures stricter validation but increases the likelihood of discarding benign agent contributions due to minor discrepancies caused by sensor noise or environmental factors.

In practical deployments, this rigid requirement can lead to reduced information utilization, as the system may frequently reject collaborative results even when they are accurate. Lowering the threshold to improve inclusion, on the other hand, risks allowing adversarial contributions to pass undetected (e.g., single object removal). Striking a balance between these competing goals is non-trivial and further emphasizes the limitations of reactive RANSAC-based approaches, which rely on such global metrics to determine trustworthy collaborations.

Appendix D Feature Consistency

Alignment via Ego-Vehicle Pose Information. Modern CP systems align feature maps by leveraging pose information from participating vehicles [68]. By applying affine transformations to warp local coordinate systems, features corresponding to the same physical objects are co-registered into a unified spatial representation in the latent semantic domain. This alignment ensures that differences in viewpoint are effectively neutralized, allowing consistency to be measured meaningfully.

Bird's Eye View (BEV) Feature Representations. BEV features are commonly used by CP systems and are well-suited for ensuring consistency across agents. BEV encoders aggregate information from camera images [30] or LiDAR point clouds [26] into a 2D top-down perspective. Such representations focus on horizontal spatial distributions, reducing the sensitivity to varying viewpoints. Therefore, features corresponding to overlapping areas remain largely consistent, even when agents observe them from different viewpoints.

Effect of Feature Misalignment. In CP, feature misalignment can happen due to benign localization errors or sensor spoofing attacks. Such misalignment can decrease the perception performance. To address the challenge brought by benign localization errors, SOTA CP algorithms leverage multiple CNN kernel sizes and effectively learn to fuse at different resolutions [38]. Intuitively, spacial misalignment is mitigated when the feature is compressed to a lower resolution, which is also integrated into the design of LUCIA. A substantial feature misalignment (e.g., an offset of 10 meters) is challenging for CP algorithms and will result in a significantly reduced perception performance (e.g., >50% reduction in accuracy [27]). However, such misalignment would incur high inconsistencies with other features and hence lower trust scores under LUCIA, which results in low contribution in the final fused feature. In other words, LUCIA effectively neutralized the contribution of the misaligned feature, either due to attack or harsh environment (e.g., dense urban topology), and restores the perception performance of the system.

Effect of Field-of-View Overlap. CP benefits the perception range of CAVs by enabling them to see through occlusions. Intuitively, feature consistencies and hence trustworthiness scores in LUCIA increases as the overlap in Field-of-View (FoV) increases among agents. The similar holds for ROBOSAC [29] where larger FOV implies smaller distance between local inference results and that made on the fused feature of *randomly sampled subset*. Instead of a rigid similarity threshold applied to the final detection results like ROBOSAC (e.g., 70% of objects must be matched in order to be considered truthful), LUCIA dynamically compares across all agents' features simultaneously and assigns trustworthiness-informed attention score accordingly. Nevertheless, it encapsulates the similar philosophy to ROBOSAC that a message that sufficiently diffesr from the others' and the local view is either alarming or is of lesser use (e.g., low overlap in FoV due to far distance).

Feature Consistency as Defense. Recent studies have revealed a compelling observation regarding the behavior of deep neural networks when processing benign and adversarial examples. Specifically, these studies have discovered a

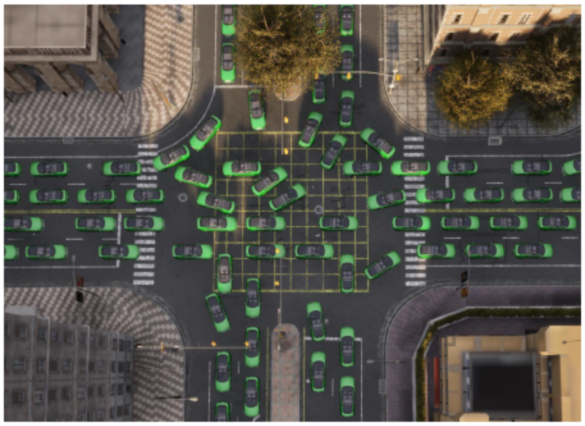


Figure 13: Top-down visualization of the custom simulated traffic jam data.

lack of consistency in feature representations between these two types of inputs [35]. It has been found that natural examples exhibit more compact similarity matrices compared to their adversarial counterparts. This suggests that benign samples from the same class tend to cluster more tightly in the latent feature space, whereas adversarial examples exhibit greater dispersion. This observation has been leveraged to develop defense mechanisms, such as Bit Plane Feature Consistency (BPFC) and Latent Feature Relation Consistency (LFRC) [35,54]. These methods aim to enforce consistency in feature representations by either promoting agreement between features extracted from different bit planes of an image or constraining the similarity between the latent feature relationships of natural and adversarial examples within a batch. By encouraging such consistency, these techniques guide the model to learn more robust features that are less susceptible to adversarial perturbations.

Appendix E Dataset Collection for the Traffic Jam Scenario

To evaluate SOMBRA under extreme conditions with a high density of collaborating vehicles, we generated a customized dataset simulating a traffic jam scenario (Figure 13). This dataset serves as the basis for the case study in Section 6.3.

Simulation Environment and Tools. We utilized the CARLA digital driving simulator [11] in conjunction with the OpenCDA framework [66] for coordinated multi-agent simulation and data recording. The data was collected following the same format and conventions as the public OPV2V benchmark dataset, including sensor configurations, coordinate systems, and CAV parameters, to ensure compatibility with existing CP models and evaluation pipelines.

Scenario Setup. The scenario was staged within CARLA's 'Town10HD' map, specifically focusing on a large, multi-lane intersection known for enabling complex traffic interactions. We simultaneously spawned and controlled 50 CAVs within the vicinity of this intersection, ensuring dense traffic conditions representative of a jam. Each CAV was equipped with LiDAR sensors and participated in CP.

Context and Motivation. It is important to note that scaling cooperative perception to 50 simultaneous agents in the real world remains a significant research challenge. Current V2X communication standards (e.g., C-V2X, DSRC) offer limited bandwidth [12, 13], which is often insufficient for reliably transmitting the large volumes of data required for raw or intermediate feature sharing among numerous vehicles, especially dense sensor data like LiDAR point clouds [8]. Consequently, much of the existing CP research focuses on scenarios involving a smaller number of agents, typically ranging from 2 to 5 CAVs. Furthermore, in extremely dense and slow-moving traffic, the potential benefits of CP (e.g., extended perception range) might diminish compared to less congested conditions.

Despite these practical limitations, we constructed this idealized 50-CAV scenario specifically to rigorously test the scalability of the SOMBRA methodology. Our goal was not to perfectly replicate current real-world communication constraints, but rather to demonstrate how effectively the attack leverages the attention mechanism even when the number of participating agents is large. By showing SOMBRA's potency in this worst-case setting, we highlight the critical vulnerability posed by attention manipulation, irrespective of the number of benign contributors.

Appendix F Robustness of V2VAM

We observed that V2VAM [28] appears to be harder to protect by LUCIA. Note that V2VAM employs Criss-Cross attention (CCNet) [20], a variant of the dot-product attention [60] employed in other CP algorithms, as its fusion backbone. In CCNet, features are exchanged more locally across spatial dimensions, which can amplify partial manipulations from the adversarial source. Consequently, when the attacker mimics the victim's exact feature and only introduces slight deviations, V2VAM's localized attention updates can be more easily hijacked. This narrower attention scope also limits the effectiveness of global, consistency-based defenses, making it harder for trust scores to isolate and downweight the attacker's carefully aligned perturbations. Additionally, V2VAM has a unique fusion design of an agent-wise max-out operation as an addition to the fused feature. Although the additional component helps preserve strong benign signals, it can be readily exploited by an attacker by ensuring at least half of the final attention comes from (dominant) malicious signal. This design sacrifices its robustness and makes it more vulnerable to adversarial attacks even under non-adaptive attacks and the baseline attacks.