All Vehicles Can Lie: Efficient Adversarial Defense in Fully Untrusted-Vehicle Collaborative Perception via Pseudo-Random Bayesian Inference

Supplementary Material

1. Hypothesis Verification Experiments

To validate our assumption regarding the temporal stability of LiDAR-based perception, we conduct a dedicated hypothesis verification experiment using the V2X-Sim dataset [5] within the collaborative perception framework. Our goal is to examine whether perception outputs across consecutive frames maintain consistently high similarity under normal conditions, while experiencing a significant drop in similarity under adversarial perturbations.

Experimental Setup. We consider three representative types of driving scenes in V2X-Sim dataset: *urban*, *suburban*, and *intersection*, as shown in Fig. 2, where we provide representative LiDAR point cloud visualizations for each scene type. For each category, we randomly select 5–10 scenes, each containing multiple consecutive frames of multi-vehicle LiDAR data. Within each scene, we compute the Jaccard similarity between consecutive-frame detection outputs, using V2VNet [8] as the perception backbone. The evaluation is carried out for both:

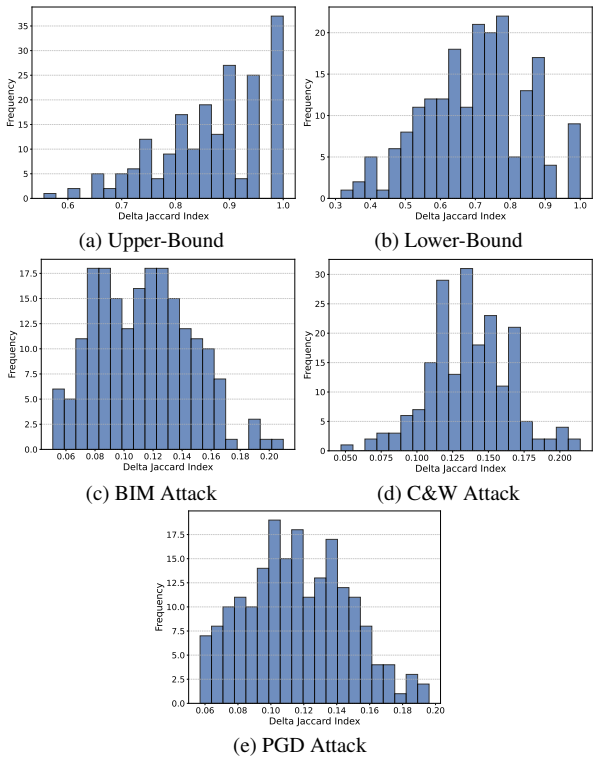
- **Normal settings:**
  1. *Upper-Bound:* All vehicles are collaborative and benign.
  2. *Lower-Bound:* Only the ego vehicle's own perception is used (*single-vehicle baseline*).
- **Adversarial settings:** All scenes re-evaluated under three widely used white-box attacks—BIM [3], C&W [1], and PGD [7]. Perturbations are injected into the feature maps of randomly selected malicious vehicles, following standard attack parameters.

Observations and Results. For Fig. 1, we visualize distribution statistics from *two randomly picked scenes (scene 4 & 78)* out of the entire test scenarios. This sub-selection is purely for clarity and readability of the plots; the trends are representative of the complete experimental results across all selected scenes. Fig. 1 clearly illustrates an evident separation between normal and adversarial regimes:

- Under **normal conditions** (both Upper- and Lower-Bound), inter-frame similarity consistently clusters around $\mathbf{0.8}$, indicating strong temporal stability and spatial continuity of perception outputs.
- Under **adversarial attacks** (BIM, C&W, PGD), similarity values drop sharply, *often well below 0.3, and peaking only around 0.2*, reflecting the severe disruption of temporal consistency caused by injected perturbations.

These results hold across scene types and independent random samplings, confirming that the temporal stability signal is strong and scenario-agnostic.

Conclusion of Hypothesis Verification. The persistent and large margin between benign and adversarial similarity distributions validates our assumption: temporal similarity between successive LiDAR perception outputs is consistently high under benign conditions and substantially degraded under attack. This robust gap can serve as a reliable self-referential signal, forming a sound foundation for our proposed frame-wise self-supervision mechanism, without the need to assume a trusted ego vehicle.



(a) Upper-Bound (b) Lower-Bound

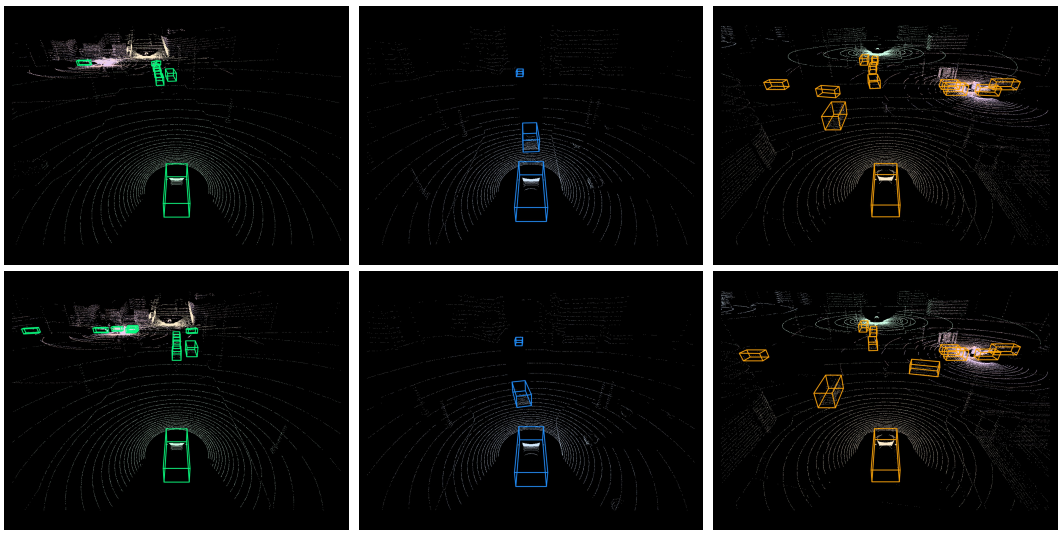
(c) BIM Attack (d) C&W Attack

(e) PGD Attack

Figure 1. Hypothesis verification experiments. We report the Jaccard similarity of detected object sets between consecutive frames for: *(a)-(b)* normal conditions (Upper-Bound: all vehicles collaborate; Lower-Bound: ego-only perception), and *(c)-(e)* adversarial conditions (BIM, C&W, and PGD). Plots are based on statistics from two randomly selected representative scenes for readability; trends are consistent across all tested scenes and scenarios (urban, suburban, intersection).

2. The Pseudocode of PRBI Algorithm

In this section, we provide a detailed pseudocode of the proposed Pseudo-Random Bayesian Inference (PRBI) algo-



Urban Scene Suburban Scene Intersection Scene

Figure 2. Example LiDAR point clouds illustrating three representative driving scenes in V2X-Sim dataset: urban, suburban, and intersection.

rithm. The PRBI framework is designed to detect malicious vehicles in collaborative perception with minimal verification attempts, thereby safeguarding the victim vehicle's perception performance under adversarial conditions.

The full algorithm is outlined in Algorithm 1, which formalizes the PRBI defense mechanism for adversarial collaborative perception.

3. Theoretical Proof

3.1. Proof of Theorem 1

In **Soft Sampling**, we adopt a pseudo-random grouping strategy that assigns the top $\lfloor m \rfloor$ most suspicious vehicles into one group, and the remaining vehicles into the second group. A key question naturally arises — ***Can*** m ***converge to the true number of attackers*** k ***?*** This subsection presents a proof showing why pseudo-random grouping ensures that m will indeed converge to the true attacker count k.

Assume the estimated number of attackers at frame i is m. We aim to prove that the estimate at the next frame, m', will move toward k. The proof is divided into two parts:
(1) When $m < k$, then m' will increase: According to the definition of soft sampling, at frame $i+1$, the grouping consists of "m suspicious vehicles + $(n-m)$ remaining vehicles". Since $m < k$, there must still be at least one attacker in the remaining group. As a result, the normal detection count β for both groups will not increase and remains unchanged. The new estimate m' is:

$$m' = \log_2\left(\frac{n\cdot(N+1)}{\sum\beta}\right) > \log_2\left(\frac{n\cdot N}{\sum\beta}\right) = m. \quad (1)$$

Thus, part (1) is proven based on Eq. (1).
(2) When $m \geq k$, then m' will decrease: Again, the frame $i+1$ grouping consists of "m suspicious vehicles + $(n-m)$ remaining vehicles". However, this time, the remaining group contains no attackers. The normal detection count for the first group remains unchanged, while the second group contributes $n-m$ additional normal detections. The updated estimate becomes:

$$m' = \log_2\left(\frac{n\cdot(N+1)}{n-m+\sum\beta}\right). \quad (2)$$

To prove part (2), we must prove the next inequality:

$$m' = \log_2\left(\frac{n\cdot(N+1)}{n-m+\sum\beta}\right) < \log_2\left(\frac{n\cdot(N)}{\sum\beta}\right) = m. \quad (3)$$

Removing the logarithm and rewriting the right-hand side using 2^m, the Eq. (3) becomes:

$$\frac{n\cdot(N+1)}{n-m+\sum\beta} < 2^m. \quad (4)$$

Divide by $\sum\beta$ to get:

$$\frac{2^m + n/\sum\beta}{1+(n-m)/\sum\beta} < 2^m. \quad (5)$$

Simplifying further and rearranging terms, the goal is to prove:

$$m + \frac{m}{2^m-1} < n. \quad (6)$$

Define $f(m) = m + \frac{m}{2^m-1}$ for $m \in [1, n-1]$. Taking the derivative of $f(m)$ and we can easily know that $f(m)$ is

Algorithm 1 Pseudo-Random Bayesian Inference

Input: Number of CAVs n; Detection count arrays $\mathbf{c}_{\text{normal}}$, $\mathbf{c}_{\text{abnormal}}$; Similarity threshold ϵ; Hypothesis testing window size w.

Output: Malicious vehicles: $Attackers = \{id_1, \ldots, id_k\}$

$\mathbf{c}_{\text{normal}} \leftarrow \mathbf{0}$, $\mathbf{c}_{\text{abnormal}} \leftarrow \mathbf{0}$, $N \leftarrow 0$
$m \leftarrow n-1$, $P_{\text{benign}} \leftarrow \mathbf{0}$
$W \leftarrow \text{Queue}()$, $convergence \leftarrow$ **False**
while system running, $i \leftarrow i+1$ **do**
  $D_1^i \leftarrow \text{Perception}(F_1^i, \ldots, F_n^i)$
  if $i = 0$ **then**
    $D_{\text{ref}} \leftarrow D_1^i$ ▷ Initialize reference frame
    D_1^i as *PerceptionOutput*
    continue
  end if
  if $\text{Jaccard}(D_1^i, D_{\text{ref}}) < \epsilon$ **and** $convergence =$ **False then**
    ▷ *Soft Sampling*
    $Group_1 \leftarrow \text{argsort}(P_{\text{benign}})_{0:\lfloor m \rfloor}$
    $Group_2 \leftarrow \text{argsort}(P_{\text{benign}})_{\lfloor m \rfloor:n}$
    ▷ *Consistency Validation*
    $D_{\text{Group}_{1,2}} \leftarrow \text{Perception}(Group_{1,2})$
    Compare $\text{Jaccard}(D_{\text{Group}_{1,2}}, D_{\text{ref}})$ with ϵ
    Update $\mathbf{c}_{\text{normal}}$, N, and $\mathbf{c}_{\text{abnormal}}$
    ▷ *Attacker Evaluation*
    $m \leftarrow \min\left(\log_2\left(\frac{n \cdot N}{\sum(\mathbf{c}_{\text{normal}})}\right), n-1\right)$
    for $id = 1$ **to** n **do**
      $P_{\text{benign}}[id] \leftarrow \text{Bayesian}(\mathbf{c}_{\text{nor}}, \mathbf{c}_{\text{abnor}}, N)$
    end for
    $Attackers \leftarrow \text{argsort}(P_{\text{benign}})_{0:\lfloor m \rceil}$
    ▷ *Hypothesis Test*
    W.push(m) **with** size limit w
    if T-test(W) indicates convergence **then**
      $convergence \leftarrow$ **True**
      Output $Attackers$
    else
      $Collaborators \leftarrow \{i \mid P_{\text{benign}}[i] \neq 0\}$
      if $Collaborators = \emptyset$ **and** first abnormal frame **then**
        $Collaborators \leftarrow$ first valid subset from recursive split of $Group_{1,2}$
        Update $\mathbf{c}_{\text{normal}}$, N, and $\mathbf{c}_{\text{abnormal}}$
      end if
      $D_1^i \leftarrow \text{Perception}(Collaborators)$
      $D_{\text{ref}} \leftarrow D_1^i$
      D_1^i as *PerceptionOutput*
    end if
  else
    Collaborative perception excluding $Attackers$
  end if
end while

increasing for $m \in [1, n-1]$. Thus, we have:

$$f(m) \leq f(n-1) = n + \frac{2n - 2^n}{2^n - 2}. \tag{7}$$

When $n = 2$, m can only be 1. In this special case, equality holds, and m' always stays at 1. When $n > 2$, since $2n - 2^n < 0$, we have:

$$f(m) \leq n + \frac{2n - 2^n}{2^n - 2} < n. \tag{8}$$

Based on Eq. (8), the Eq. (6) holds and part (2) is proven.

3.2. Proof of Theorem 2

In Sec. 3.1, we have proven that under the pseudo-random grouping strategy, the estimated number of attackers m will always converge toward the true number of attackers k. In practical applications, it is essential to address another important issue — ***How should*** m ***be rounded ?*** Since m tends to converge to k, directly applying standard rounding (i.e., rounding to the nearest integer) when evaluating suspicious malicious vehicles may suffice. However, during the **pseudo-random grouping** process, m may not yet have fully converged, and different rounding strategies can influence the convergence outcome. We now provide a detailed derivation of these effects under the same settings and leveraging the proven proof **(1), (2)** in Sec. 3.1:

(a) Rounding to the nearest integer $\lfloor m \rceil$: When $m \in (-\infty, k-0.5)$, we have $\lfloor m \rceil < k$. According to proof **(1)**, the next frame's m' will increase. When $m \in [k-0.5, +\infty)$, $\lfloor m \rceil \geq k$, and m' will decrease based on proof **(2)**. Therefore, m ultimately converges to approximately $k - 0.5$. The subsequent derivations are carried out in the same way.

(b) Ceiling $\lceil m \rceil$: When $m \in (-\infty, k-1)$, we have $\lceil m \rceil < k$, and m' will increase. When $m \in [k-1, +\infty)$, $\lceil m \rceil \geq k$, and m' will decrease. Thus, m ultimately converges to approximately $k - 1$. Note that when the true number of attackers $k = 1$, m will not tend toward 0, because the argument of a logarithm in Eq. (1) can never be one. In this case, apart from the malicious vehicle, each of the remaining $(n-1)$ benign vehicles undergoes one successful normal detection per frame. Thus, the final m is:

$$m \cong \log_2\left(\frac{n \cdot N}{(n-1) \cdot N}\right) = \log_2\left(\frac{n}{n-1}\right). \tag{9}$$

(c) Floor $\lfloor m \rfloor$: When $m \in (-\infty, k)$, we have $\lfloor m \rfloor < k$, and m' will increase. When $m \in [k, +\infty)$, $\lfloor m \rfloor \geq k$, and m' will decrease. Hence, m ultimately converges to k.

In conclusion, **only by applying floor rounding ($\lfloor m \rfloor$) in the pseudo-random grouping strategy can we ensure that m converges to k.**

4. Experimental Results

4.1. Experimental Details

Adversarial Optimization. Perturbations are added to all malicious vehicles' features and optimized jointly. For adversarial attacks, the step size is set to 0.1, the perturbation stealth metric to 0.3, and the number of iterations to 15.

Baselines and Implementation. We randomly select 5 scenarios from the V2X-Sim test set, each consisting of 100 frames. Collaborative perception scenarios involve 6 collaborators (1 RSU and 5 vehicles, with the first vehicle being the ego/victim). Detection model training follows the default setup in DiscoNet [4]. For baseline defenses, ROBOSAC [6] strictly follows the official configuration and always selects the maximum possible number of collaborators for joint perception to ensure fairness and stability. PASAC [2] also adopts its official hyperparameter settings for the consistency loss module. All baseline implementations are reproduced under the same perception backbone, communication setup, and attack conditions for a fair comparison.

Defense Hyperparameters. The Jaccard similarity threshold ϵ is 0.35, the verification window size w is 10, the confidence level α for the T-test is 0.01, and the probability weights γ and λ are set to 0.35 and 0.65, respectively. The setting of the similarity threshold is based on the following considerations: (1) a relatively large threshold may occasionally misclassify benign vehicles as malicious in a single frame, but such errors can be corrected in subsequent frames; (2) setting the threshold too small risks misclassifying malicious vehicles as benign, yet this is tolerable in our framework. Specifically, a false negative occurs when malicious vehicles exhibit high similarity, implying that the attack has not significantly impacted perception and thus does not cause irreversible consequences. Conversely, effective attacks induce a sustained drop in similarity, and our probabilistic updating mechanism ensures that such malicious vehicles will eventually be identified. Combined with the experimental results in Sec. 1 (where adversarial cases rarely exceed 0.3), we adopt 0.35 as a balanced threshold.

4.2. Varying Test Parameters

In the scenario where the total number of vehicles is set to 5, the results of the hypothesis testing parameter (α and w) experiments for all attacker ratios are presented in Tab. 1. We document the total number of frames required to achieve convergence under each setting ***(Average Converge Frames)***, the identification rate of malicious vehicles upon convergence ***(Identification Rate)***, and the misclassification rate of benign vehicles upon convergence ***(Misclassification Rate)***. The results show that regardless of the proportion of attackers, the convergence speed is always inversely proportional to the confidence level α and directly proportional to the window size w. For instance, when the α is set to 0.2, the average maximum number of frames required for convergence is 10.01 frames; while the α is

Table 1. Comprehensive experimental results of different test parameters on defense effect. *ID Rate*: malicious vehicle identification rate. *MC Rate*: benign vehicle misclassification rate.

Settings	Count		Avg. Count		Avg.	Avg.	ID	MC
	Min	Max	Min	Max		Frames	Rate	Rate
Confidence Level (20% Attackers)								
Conf. 0.20	2	2	2.00	2.00	2.00	2.53	100%	0%
Conf. 0.15	2	2	2.00	2.00	2.00	2.27	100%	0%
Conf. 0.10	2	2	2.00	2.00	2.00	2.29	100%	0%
Conf. 0.05	2	2	2.00	2.00	2.00	2.19	100%	0%
Conf. 0.01	2	2	2.00	2.00	2.00	2.25	100%	0%
Confidence Level (40% Attackers)								
Conf. 0.20	2	4	2.00	3.02	2.38	3.13	100%	9%
Conf. 0.15	2	4	2.00	3.00	2.37	3.04	100%	8%
Conf. 0.10	2	4	2.00	2.96	2.35	2.98	100%	8%
Conf. 0.05	2	4	2.00	2.91	2.34	2.77	100%	4%
Conf. 0.01	2	4	2.00	2.88	2.35	2.77	100%	6%
Confidence Level (60% Attackers)								
Conf. 0.20	2	6	2.00	4.22	2.61	3.99	100%	0%
Conf. 0.15	2	6	2.00	4.12	2.62	3.51	100%	0%
Conf. 0.10	2	6	2.00	3.94	2.53	3.44	100%	0%
Conf. 0.05	2	6	2.00	4.26	2.67	3.40	100%	0%
Conf. 0.01	2	6	2.00	4.06	2.61	3.36	100%	0%
Confidence Level (80% Attackers)								
Conf. 0.20	2	8	2.00	6.06	2.50	10.01	100%	0%
Conf. 0.15	2	8	2.00	5.86	2.51	9.31	100%	0%
Conf. 0.10	2	8	2.00	6.24	2.80	6.58	100%	0%
Conf. 0.05	2	8	2.00	6.04	2.77	6.51	100%	0%
Conf. 0.01	2	8	2.00	5.70	2.86	4.27	100%	0%
Window Size (20% Attackers)								
Size 10	2	2	2.00	2.00	2.00	2.25	100%	0%
Size 8	2	2	2.00	2.00	2.00	2.37	100%	0%
Size 6	2	2	2.00	2.00	2.00	2.35	100%	0%
Size 4	2	2	2.00	2.00	2.00	2.33	100%	0%
Window Size (40% Attackers)								
Size 10	2	4	2.00	2.88	2.35	2.77	100%	6%
Size 8	2	4	2.00	3.12	2.46	2.80	100%	7%
Size 6	2	4	2.00	3.00	2.39	2.84	100%	4%
Size 4	2	4	2.00	3.02	2.40	2.94	100%	11%
Window Size (60% Attackers)								
Size 10	2	6	2.00	4.06	2.61	3.36	100%	0%
Size 8	2	6	2.00	4.08	2.62	3.33	100%	0%
Size 6	2	6	2.00	4.32	2.69	3.34	100%	0%
Size 4	2	6	2.00	4.14	2.63	3.41	100%	0%
Window Size (80% Attackers)								
Size 10	2	8	2.00	5.70	2.86	4.27	100%	0%
Size 8	2	8	2.00	6.16	2.91	4.60	100%	0%
Size 6	2	8	2.00	6.10	2.89	4.57	100%	0%
Size 4	2	8	2.00	6.34	2.88	4.89	100%	0%

Table 2. Defense performance under intermittent attacks.

	V2VNet		DiscoNet	
Methods	AP@0.5 / 0.7	ID / MC	AP@0.5 / 0.7	ID / MC
1 frame	68.91 / 65.57	100% / 0%	69.37 / 65.89	100% / 0%
3 frames	70.97 / 65.91	100% / 0%	69.12 / 66.27	100% / 0%
5 frames	73.52 / 67.04	100% / 0%	70.43 / 66.91	100% / 0%

reduced to 0.01, the average minimum number of frames needed for convergence is only 2.25 frames. Additionally, the experimental results show that under all test conditions, the identification rate of malicious vehicles reaches 100%, fully verifying the high robustness of the proposed method.

4.3. Robustness to Intermittent Attacks

We further evaluate PRBI under intermittent attacks, where adversarial perturbations are injected periodically every 1, 3, or 5 frames. The total number of vehicles is set to $n = 5$, and all other experimental settings remain consistent with the main text.

PRBI is inherently robust to intermittent attacks due to its temporal Bayesian updating mechanism. Even when attacks occur sporadically, the posterior benign probability of malicious vehicles decreases cumulatively over time, leading to their progressive exclusion from the collaborator subset. As shown in Tab. 2, PRBI consistently achieves 100% attacker identification with 0% benign misclassification under all attack frequencies. Detection performance remains stable across both V2VNet and DiscoNet backbones, demonstrating strong robustness against temporally sparse adversarial behavior.

4.4. Sensitivity Analysis of Similarity Threshold ϵ

We evaluate the sensitivity of PRBI to the similarity threshold ϵ under standard adversarial settings. The total number of vehicles is set to $n = 5$, and other experimental configurations remain consistent with the main text.

As shown in Tab. 3, PRBI maintains 100% attacker identification and 0% benign misclassification across a wide range of thresholds $\epsilon \in [0.30, 0.50]$. This indicates that the correctness of PRBI does not critically depend on precise threshold tuning. From a performance perspective, smaller ϵ values lead to slightly faster convergence but may introduce marginal reductions in detection precision, whereas larger ϵ values enforce stricter consistency validation, resulting in improved stability at the cost of modestly increased convergence latency. This behavior reflects a controllable strictness–efficiency trade-off.

Overall, the robustness of PRBI primarily stems from its probabilistic temporal accumulation mechanism rather than the specific value of ϵ. Empirically, $\epsilon = 0.35$ achieves the best balance between detection accuracy and convergence speed.

Table 3. Sensitivity of PRBI to similarity threshold ϵ under standard adversarial settings.

ϵ	AP@0.5 / 0.7	Avg. Convergence Frames
0.30	69.17 / 64.14	2.41
0.35	69.44 / 66.00	2.74
0.40	68.61 / 64.14	2.90
0.45	69.40 / 65.80	2.93
0.50	67.47 / 64.96	2.88
ID Rate: 100%, MC Rate: 0% across all settings		

References

[1] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of 2017 IEEE Symposium on Security and Privacy*, pages 39–57. IEEE, 2017. 1

[2] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 23203–23211, 2025. 4

[3] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pages 99–112. Chapman and Hall/CRC, 2018. 1

[4] Yiming Li, Shunli Ren, Pengxiang Wu, Siheng Chen, Chen Feng, and Wenjun Zhang. Learning distilled collaboration graph for multi-agent perception. *Advances in Neural Information Processing Systems*, 34:29541–29552, 2021. 4

[5] Yiming Li, Dekun Ma, Ziyan An, Zixun Wang, Yiqi Zhong, Siheng Chen, and Chen Feng. V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving. *IEEE Robotics and Automation Letters*, 7(4):10914–10921, 2022. 1

[6] Yiming Li, Qi Fang, Jiamu Bai, Siheng Chen, Felix Juefei-Xu, and Chen Feng. Among us: Adversarially robust collaborative perception by consensus. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 186–195, 2023. 4

[7] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1

[8] Tsun-Hsuan Wang, Sivabalan Manivasagam, Ming Liang, Bin Yang, Wenyuan Zeng, and Raquel Urtasun. V2vnet: Vehicle-to-vehicle communication for joint perception and prediction. In *Proceedings of Computer Vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part II 16*, pages 605–621. Springer, 2020. 1