# Pretend Benign: A Stealthy Adversarial Attack by Exploiting Vulnerabilities in Cooperative Perception

## 1. Results on Real Dataset V2V4Real

We added experiments on V2V4Real [11], a real-world cooperative perception dataset. The Table 1 below shows that PB outperforms other adversarial attack methods in terms of attack effectiveness (AP) and also achieves higher success rates (ASR) under defense methods. This demonstrates that PB retains strong attack performance and stealthiness in real-world scenarios.

| Method | No Defense AP | | ROBOSAC AP | | | FLD AP | | |
|---|---|---|---|---|---|---|---|---|
| | $IoU_{0.3}$ | $IoU_{0.5}$ | $IoU_{0.3}$ | $IoU_{0.5}$ | ASR | $IoU_{0.3}$ | $IoU_{0.5}$ | ASR |
| No Attack | 0.68 | 0.55 | 0.68 | 0.55 | - | 0.68 | 0.55 | - |
| PGD | 0.54 | 0.09 | 0.60 | 0.49 | 0.0 | 0.60 | 0.49 | 0.0 |
| C&W | 0.55 | 0.12 | 0.60 | 0.49 | 0.0 | 0.60 | 0.49 | 0.0 |
| James Attack | 0.50 | 0.07 | 0.60 | 0.49 | 0.0 | 0.60 | 0.49 | 0.0 |
| PB(Ours) | 0.24 | 0.15 | **0.24** | **0.15** | **0.98** | **0.31** | **0.22** | **0.83** |

Table 1. Results on V2V4Real.

## 2. Pretend Benign as a White-box Attack

The proposed Pretend Benign (PB) follows the settings in [2, 3, 6, 13], making it a white-box attack [5]. However, this does not prevent PB from being effectively deployed in real-world scenarios. In cooperative perception, if the perception models of different agents are heterogeneous and lack an alignment module, the entire cooperative perception system becomes ineffective [8]. Conversely, if an alignment module is trained, the scenario essentially becomes a white-box setting.

Furthermore, to ensure robust perception performance and facilitate practical deployment, all agents within the same cooperative system should adopt identical perception models. Given these considerations, white-box adversarial attacks are already applicable in most cooperative perception environments, making black-box attacks [4] less critical in such settings.

## 3. Exploring the Black-box Transferability of Pretend Benign

To address the remaining cases where black-box conditions exist [4, 6], we investigate the black-box transferability of PB across different cooperative perception systems. Specifically, we conduct experiments where adversarial attack signals are generated under varying perception models and fusion methods and then used to attack a heterogeneous victim model. The results are presented in Table 2.

From Table 2, we observe that due to the heterogeneity between the attacker's and the victim's perception models, PB's attack effectiveness without defense is generally weaker than in the white-box setting. However, when the victim model is SECOND [12] with Average fusion, PB still achieves strong attack performance, demonstrating a certain level of black-box transferability.

More importantly, PB maintains its stealth even under black-box conditions, successfully bypassing state-of-the-art defense methods such as ROBOSAC [3] and our improved FLD. These results confirm that PB exhibits notable black-box transferability while preserving its high stealthiness.

## 4. Potential defense strategies

Possible defense strategies include: (1) adopting more secure communication protocols to prevent attacks at the source; (2) using trusted agents to assist the ego agent in validating shared information in uncertain regions (AR and IR).

## 5. Additional Visualization Results

To further illustrate the effectiveness of PB, we present additional visual comparisons between PB and other attack methods on OPV2V [10] and V2XSet [9], as shown in Fig. 1 and Fig. 2.

## References


[1] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[2] 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 *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 18421–18427. IEEE, 2024. 1

| Attacker Model | Victim Model | No Attack AP $IoU_{0.7}$ | No Defense AP $IoU_{0.7}$ | ROBOSAC [3] AP $IoU_{0.7}$ | FLD AP $IoU_{0.7}$ |
|---|---|---|---|---|---|
| PointPillars [1]+Attention [7] | PointPillars [1]+Average | 0.81 | 0.76 | 0.72 | 0.76 |
| SECOND [12]+Average | PointPillars [1]+Average | 0.81 | 0.77 | 0.73 | 0.77 |
| PointPillars [1]+Average | PointPillars [1]+Attention [7] | 0.83 | 0.69 | 0.61 | 0.68 |
| SECOND [12]+Average | PointPillars [1]+Attention [7] | 0.83 | 0.72 | 0.61 | 0.72 |
| PointPillars [1]+Average | SECOND [12]+Average | 0.89 | 0.61 | 0.61 | 0.61 |
| PointPillars [1]+Attention [7] | SECOND [12]+Average | 0.89 | 0.50 | 0.50 | 0.50 |

Table 2. PB's Black-box Transferability Experiment Results on OPV2V.

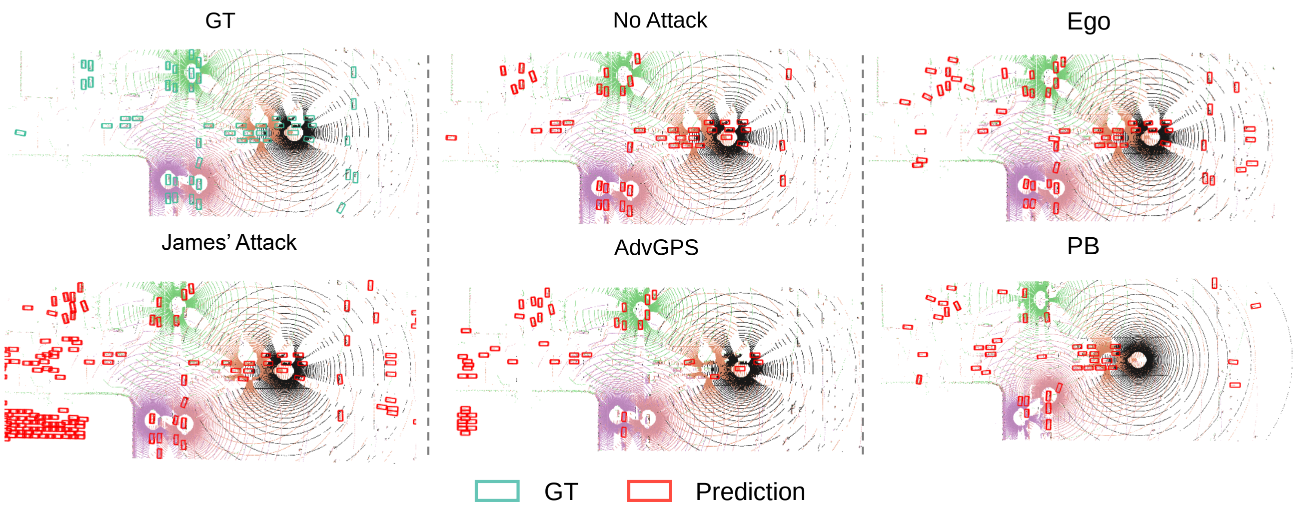


Figure 1. Visualization comparison of attack results between PB and other attack methods on OPV2V.

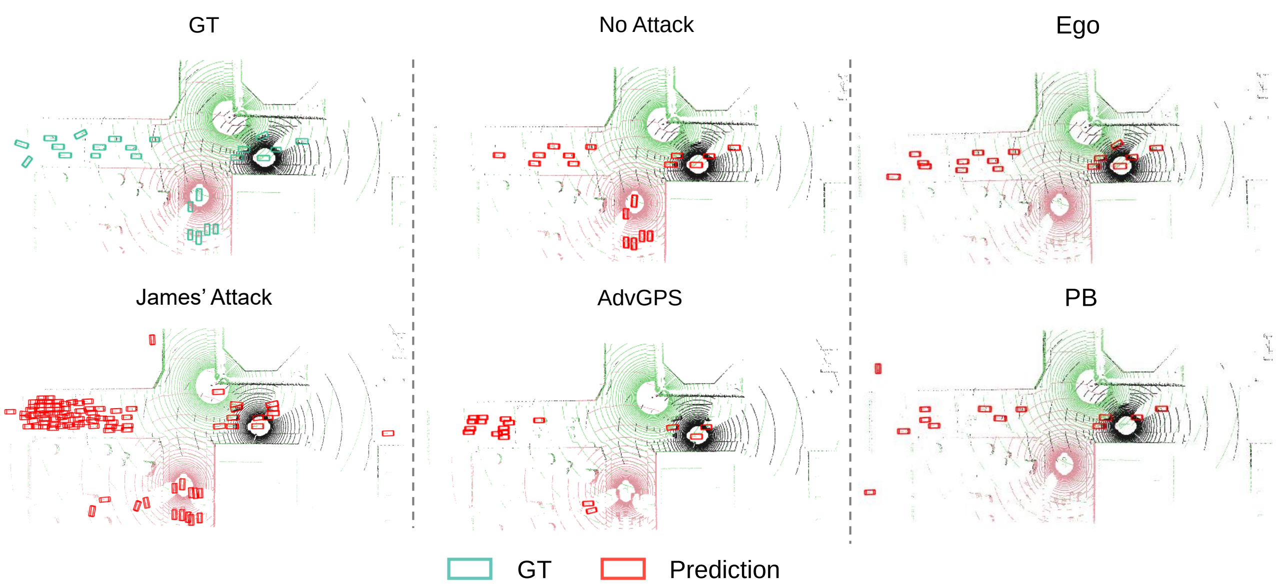


Figure 2. Visualization comparison of attack results between PB and other attack methods on V2XSet.

[3] 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. 1, 2

[4] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow,

Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1

[5] V Porkodi, Murugan Sivaram, Amin Salih Mohammed, and V Manikandan. Survey on white-box attacks and solutions. *Asian Journal of Computer Science and Technology*, 7(3): 28–32, 2018. 1

[6] James Tu, Tsunhsuan Wang, Jingkang Wang, Sivabalan Manivasagam, Mengye Ren, and Raquel Urtasun. Adversarial attacks on multi-agent communication. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7768–7777, 2021. 1

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 2

[8] Hao Xiang, Runsheng Xu, and Jiaqi Ma. Hm-vit: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 284–295, 2023. 1

[9] Runsheng Xu, Hao Xiang, Zhengzhong Tu, Xin Xia, Ming-Hsuan Yang, and Jiaqi Ma. V2x-vit: Vehicle-to-everything cooperative perception with vision transformer. In *European conference on computer vision*, pages 107–124. Springer, 2022. 1

[10] 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 *2022 International Conference on Robotics and Automation (ICRA)*, pages 2583–2589. IEEE, 2022. 1

[11] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13712–13722, 2023. 1

[12] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2

[13] Yangheng Zhao, Zhen Xiang, Sheng Yin, Xianghe Pang, Yanfeng Wang, and Siheng Chen. Made: Malicious agent detection for robust multi-agent collaborative perception. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 13817–13823. IEEE, 2024. 1