# Point Cloud Adversarial Examples by Optimizing Point Intensity Values

Kondreddy Rohith Sai Reddy
*University of Florida*

Satya Aakash Chowdary Obellaneni
*University of Florida*

Vivek Reddy Gangula
*University of Florida*

Kushi Vardhan Reddy Pasham
*University of Florida*

## 1 Introduction

Securing point cloud-based object detection systems is crucial, particularly as they play a central role in autonomous vehicles and other advanced applications requiring accurate environmental perception. Investigating and fortifying these systems against intensity-based adversarial attacks is essential to prevent potential failures with severe consequences. Point clouds represent sets of data points situated within a three-dimensional coordinate framework, commonly captured by 3D scanning technologies like LiDAR (Light Detection and Ranging). Point cloud-based object detection is a method that identifies objects within a scene by processing data points in three-dimensional space [7]. Point cloud-based object detection is utilized in autonomous vehicles for obstacle recognition, drones for terrain mapping, virtual reality for realistic object placement, and mapping for creating detailed 3D maps [11].

In the study of point cloud-based object detection vulnerabilities, various methods have been developed to exploit these systems. [11] introduced an adversarial attack that perturbs points in the point cloud to fool LiDAR-based 3D object detectors by generating sparse point perturbations. [9] crafted an attack that involves adding a small number of fake points into the scene, creating ghost objects that confuse both the object detection and tracking modules. [1] developed an algorithm that optimizes the location and number of points to be added or removed, effectively evading detection by altering the point cloud's distribution subtly. Lastly, [4] proposed an adversarial framework that not only adds but also removes points from a point cloud to generate adversarial examples capable of bypassing detectors.

Unlike traditional approaches that manipulate the spatial components of point clouds to compromise detection systems, our research explores novel intensity-based manipulations in LiDAR-generated point clouds. Here, point intensities, which represent the reflectivity of surfaces as measured by LiDAR sensors, are crucial for performing object detections. This research primarily investigates how modifications to the intensity values in 3D point clouds impact the performance of deep learning models. Additionally, it examines whether optimizing these intensity values can serve as an novel method for generating adversarial examples against 3D detection models.

Despite varying the intensity values within the 3D point clouds, our observations showed that the accuracy of the PointRCNN model remained unchanged, suggesting inherent ability to withstand against such modifications.

## 2 Background and Related Work

A. *Point Cloud-based object Detection:*

Point cloud-based object detection methods are pivotal for autonomous systems, offering a three-dimensional representation of the environment that is crucial for navigation and interaction within various spaces. These methods work by converting sensor data, typically from LiDAR, into a 'cloud' consisting of numerous points in a 3D coordinate system. This data is then processed to identify and classify objects based on their geometric and reflective properties.

Emerging approaches in this area have leveraged various aspects of point clouds for enhanced detection. For example, the use of intensity values, which reflect the return signal strength of the LiDAR sensor, has been explored in works like [3] for improving object detection accuracy. PointRCNN [8] advanced the field by directly generating 3D bounding boxes from raw point cloud data, enabling precise object localization. Other notable contributions include [2], where deep learning models were specifically tuned to utilize the spatial distribution of points within clouds for better object classification, and [6], which introduced a method for segmenting point clouds into meaningful clusters, thereby aiding in object recognition tasks.

Despite these technological advancements, the robustness of point cloud-based detection systems against adversarial attacks remains a pressing concern.

B. *Adversarial Attacks:*

Adversarial attacks are deliberate manipulations of model inputs, designed to cause incorrect model outputs. These are generally categorized into two types: evasion attacks, which modify inputs to cause misclassification at inference, and poisoning attacks, which corrupt training data to mislead the learning process. In point cloud-based systems, evasion attacks are particularly concerning as they can cause models to misinterpret or overlook critical objects like pedestrians or vehicles. Researchers have demonstrated techniques that subtly shift the positions of points or add noise to the data, effectively camouflaging or misrepresenting objects, thus evading detection or causing incorrect classifications [8], [10].

Building on this foundation, our research focuses on the less-explored aspect of point cloud intensity values and their potential exploitation in adversarial attack scenarios. Intensity values, which provide reflectivity characteristics essential for accurate object interpretation, have been sparingly considered in previous research on adversarial vulnerabilities. By redirecting focus to this attribute, we aim to unveil new insights into the security and resilience of point cloud-based object detection systems against adversarial manipulations.

## 3 Methodology

This section describes the structured approach for evaluating the susceptibility of 3D object detection systems to intensity-based adversarial attacks.

A.*Threat Model*

We try to identify the vulnerabilities in deep learning systems that analyze LiDAR data for object detection in autonomous vehicles. The specific vulnerability addressed is the manipulation of intensity values within 3D point clouds, essential for the accurate interpretation of the vehicle's surroundings. Such manipulations, subtly executed either through direct interference with LiDAR sensors or by altering environmental light reflections, are designed to mislead AI models. This could lead to misinterpretations of object types or locations, ultimately resulting in flawed navigational decisions or operational failures. We assume that attackers have the capability to digitally alter point cloud data and possess black-box access to the detection models, allowing them to observe the impact of their manipulations without detailed knowledge of the model's architecture.

B. *Methodology and Technical Approach*

1. *Data Preparation:*

We utilize the KITTI dataset, which includes LiDAR-generated 3D point clouds. Each data point consists of spatial coordinates (x, y, z) and an intensity value, which quantifies the reflectivity characteristic of the surfaces detected by sensors like LiDAR. This setup allows us to accurately simulate real-world conditions that AI models encounter.

2. *Initial Model Training:*

The KITTI dataset's preprocessed 3D point cloud information serves as input for PointRCNN, a specialized deep learning model tailored for 3D object detection. This model functions through two primary phases: initially segmenting the point cloud into foreground and background, then producing 3D object proposals. These proposals are then refined to accurately localize and detect objects based on their spatial and intensity attributes. Alongside these processes, the model assesses and assigns confidence levels to its predictions, providing a measure of reliability for each detected object.

3. *Object Selection using Open3D:*

We use Open3D to precisely identify and isolate target objects, such as cyclists or pedestrians, within the KITI dataset's 3D point clouds. The selection is based on spatial coordinates for specific intensity value alterations. These tailored modifications enable an in-depth analysis of how PointRCNN responds to localized adversarial changes.

4. *Intensity Manipulation and Evaluation:*

In our methodology, the Iterative Gradient Method is meticulously applied to manipulate the intensity values of the selected object 3D point clouds, aimed at crafting effective adversarial examples.Specifically, the process begins by setting the initial adversarial intensity to be the original intensity. In each subsequent iteration, the adversarial intensity is updated by adjusting in the direction that maximizes the loss. This direction is determined by the sign of the gradient of the loss function J with respect to the adversarial example from the previous iteration, scaled by a small step size ε.

$x_0^{\text{adv}} = x$ [5]

$x_t^{\text{adv}} = x_{t-1}^{\text{adv}} + \varepsilon \cdot \text{sign}(\nabla_{x_{t-1}^{\text{adv}}} J(x_{t-1}^{\text{adv}}, \theta, y))$ [5]

After each adjustment, the modified point cloud is fed back into the PointRCNN model to evaluate the effect of the changes. Specifically, we assess the confidence level of the model's predictions. If the desired level of disruption (decreased confidence in correct classifications) is not achieved, the process repeats—adjusting the intensities further and re-evaluating the model's response.

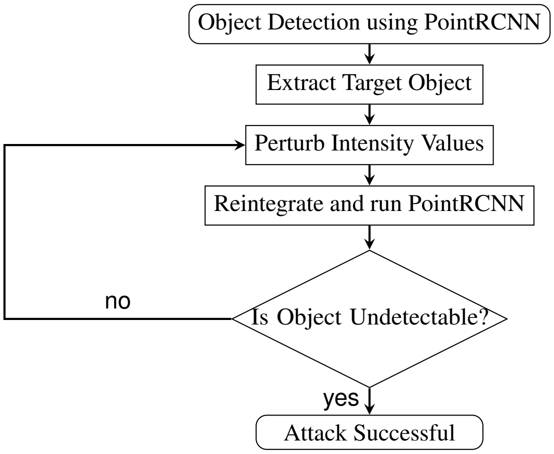


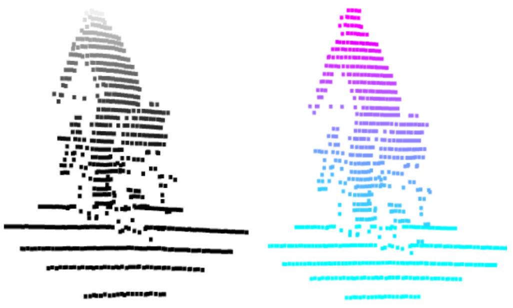
Figure 1: Flowchart illustrating the process of intensity-based adversarial attack on PointRCNN object detection.

5. *Evaluation Criteria:*

We evaluate the efficacy of adversarial attacks on point cloud-based object detection models using confidence scores which serve as a primary indicator of model performance. These encompass analyzing model performance across different distances to test range-based robustness and varying object classes, such as pedestrians, cyclists, and cars, to assess class-specific vulnerabilities. This multi-faceted approach allows us to comprehensively determine the model's accuracy and reliability in diverse scenarios that mimic real-world conditions.

## 4 Experimental Results

We observed distinct trends in object detection accuracy across different distance ranges and iterations, particularly focusing on the iterative gradient attack variations in intensities. Notably, we found that the accuracy of detecting cyclists, pedestrians, and cars tends to decrease as the distance increases. This trend is evident in the declining detection percentages for cyclists (from 70.7% to 52.31%), pedestrians (from 48.98% to 37.63%), and cars (from 84.34% to 67.03%) across distance ranges from 0-10 meters to 30+ meters. Conversely, when examining the relationship between iterations of the iterative gradient attack and confidence score, we noted that the model's confidence stabilizes around 70.70% starting from 73% after approximately 400 iterations. We also observed that there is no considerable change with perturbation of the intensities. This suggests a consistent level of certainty in the model's predictions over time, even amidst varying attack intensities. The accuracy was reduced by around 3% for cyclist class objects and a similar trend was observed for car and pedestrian object attacks.These findings highlight the model's ability to maintain stable confidence levels with iterative refinement, despite adversarial attacks.



(a) Cyclist Point Cloud Before Intensity Perturbation

(b) Cyclist Point Cloud After Intensity Perturbation

Figure 2: These are the visualizations of the intensities of attack vectors(cyclist point clouds) using the open3d module, this represents the attack vectors which are the point clouds of a cyclist object extracted from the whole point cloud.

| Distance Range (m) | Cyclist | Pedestrian | Car |
|---|---|---|---|
| 0-10 | 70.7% | 48.98% | 84.34% |
| 10-20 | 64.65% | 45.34% | 79.34% |
| 20-30 | 58.4% | 40.28% | 74.53% |
| 30+ | 52.31% | 37.63% | 67.03% |

Table 1: Comparisons of cyclist, pedestrian and car objects confidence scores with differing the distance range

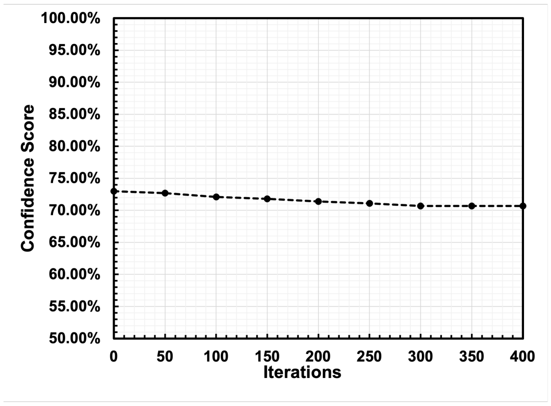


Figure 3: Illustrating the confidence scores of cyclist object detection across iterations

## 5 Analysis and Discussion

During our work with intensity-varying adversarial attacks in point cloud-based object detection systems, we found that perturbation of the intensity attacks were generally ineffective in reducing the model's confidence in object classification. Despite varying the intensity of the attacks, there was minimal impact on the model's confidence scores and classification outcomes for detected objects.

Specifically, we observed that objects located farther from the sensor and the cyclist objects were more susceptible to misclassification or undetection, highlighting the need to tailor defense mechanisms to the unique characteristics of different object categories. However, interestingly, there was no discernible change in the classification of certain objects, and the confidence scores associated with detected objects remained largely unchanged. These findings highlight the resilience of deep neural networks processing 3D data against intensity-based adversarial attacks. They also underscore the need for future work of more sophisticated attack strategies that incorporate both spatial and intensity perturbations to effectively evaluate and fortify the defenses of point cloud-based object detection systems.

## References

[1] 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 *Proceedings of the 2019 ACM SIGSAC conference on computer and communications security*, pages 2267–2281, 2019.

[2] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[3] Loic Landrieu and Martin Simonovsky. Large-scale point cloud semantic segmentation with superpoint graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[4] Xinke Li, Zhirui Chen, Yue Zhao, Zekun Tong, Yabang Zhao, Andrew Lim, and Joey Tianyi Zhou. Pointba: Towards backdoor attacks in 3d point cloud. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 16492–16501, 2021.

[5] Daniel Liu, Ronald Yu, and Hao Su. Extending adversarial attacks and defenses to deep 3d point cloud classifiers. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 2279–2283, 2019.

[6] Charles R. Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J. Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[7] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[8] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Chong Xiang, Charles R. Qi, and Bo Li. Generating 3d adversarial point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[10] Hengshuang Zhao, Li Jiang, Chi-Wing Fu, and Jiaya Jia. Pointweb: Enhancing local neighborhood features for point cloud processing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[11] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.