Identifying and Patching Vulnerabilities of Camera-LiDAR Based Autonomous Driving Systems

Final Report

by

Cihang Xie
University of California, Santa Cruz

Alvaro Cardenas, University of California, Santa Cruz
Murat Kantarcioglu, University of Texas at Dallas

May 2025

NATIONAL CENTER FOR TRANSPORTATION CYBERSECURITY AND RESILIENCY (TraCR)

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by the National Center for Transportation Cybersecurity and Resiliency (TraCR) under Grant No. 69A3552344812 and 69A3552348317 which is headquartered at Clemson University, South Carolina, USA, from the U.S. Department of Transportation's University Transportation Centers Program. The U.S. Government assumes no liability for the contents or use thereof.

CONTACTS
For more information:

Cihang Xie
1156 High St
Santa Cruz, CA 95064
Phone: 424-320-1038
Email: cixie@ucsc.edu

TraCR
Clemson University
One Research Dr
Greenville, SC 29607
tracr@clemson.edu

ACKNOWLEDGMENT

This work is based upon the work supported by the National Center for Transportation Cybersecurity and Resiliency (TraCR), a U.S. Department of Transportation National University Transportation Center headquartered at Clemson University, Clemson, South Carolina, USA. Any opinions, findings, conclusions, and recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of TraCR. The U.S. Government assumes no liability for the contents or use thereof.

Technical Report Documentation Page

1. Report No. 7

2. Government Accession No. N/A

3. Recipient's Catalog No. N/A

4. Title and Subtitle
Identifying and Patching Vulnerabilities of Camera-LiDAR Based Autonomous Driving Systems

5. Report Date: May 2025

6. Performing Organization Code: N/A

7. Author(s)
Cihang Xie, Ph.D.; https://orcid.org/0000-0003-1243-8045
Murat Kantarcioglu, Ph.D.; https://orcid.org/0000-0001-9795-9063
Alvaro Cardenas, Ph.D.; https://orcid.org/0000-0002-5142-9750

8. Performing Organization Report No. 7

9. Performing Organization Name and Address
National Center for Transportation Cybersecurity and Resiliency (TraCR),
Clemson University,
414 A One Research Dr, Greenville, SC 29607
University of California Santa Cruz
1156 High St, Santa Cruz, CA 95064

10. Work Unit No. N/A

11. Contract or Grant No.
69A3552344812 and 69A3552348317

12. Sponsoring Agency Name and Address
U.S. Department of Transportation,
Office of the Assistant Secretary for Research and Technology,
1200 New Jersey Avenue, SE, Washington, DC 20590

13. Type of Report and Period Covered
Final Report, 01/01/2024 - 12/31/2024

14. Sponsoring Agency Code OST-R

15. Supplementary Notes
Conducted under the U.S. DOT Office of the Assistant Secretary for Research and Technology's (OST-R) University Transportation Centers (UTC) program.

16. Abstract

Autonomous driving systems rely on advanced perception models to interpret their surroundings and make real-time driving decisions. Among these, Bird's Eye View (BEV) perception has emerged as a critical component, offering a unified 3D representation from multi-camera and sensor inputs. While BEV-based models have gained traction in industry-leading platforms, their security vulnerabilities remain largely underexplored in adversarial machine learning research. This study provides a multi-dimensional security analysis of BEV perception models, focusing on adversarial threats in both vision-only and multi-sensor fusion architectures. We examine the susceptibility of state-of-the-art models - including BEVDet, BEVDet4D, DAL, and BEVFormer—to adversarial attacks targeting their detection and decision-making capabilities. Unlike traditional adversarial research that primarily misleads perception models at the classification level, this study investigates real-world attack scenarios where adversaries can manipulate perception to cause practical disruptions, such as inducing traffic congestion or triggering unsafe vehicle behaviors. Our findings reveal significant security risks in BEV-based perception, with both vision-only and sensor-fusion models vulnerable to adversarial perturbations. Attack transferability across architectures further highlights the urgency of developing robust defense mechanisms to ensure the reliability of self-driving technology. This work underscores the need for adversarially resilient perception models to safeguard the future of autonomous driving.

17. Keywords
BEV, adversarial robustness

18. Distribution Statement
No restrictions.

19. Security Classif. (of this report)
Unclassified

20. Security Classif. (of this page)
Unclassified

21. No. of Pages
25

22. Price
N/A

TABLE OF CONTENTS

List of Tables

List of Figures

EXECUTIVE SUMMARY

Autonomous driving systems rely heavily on advanced perception models to interpret their surroundings and make real-time decisions. Among these, Bird’s Eye View (BEV) perception has become a crucial component, offering a top-down 3D representation generated from multi-camera and sensor inputs. This unified spatial understanding enables effective object detection, tracking, and path planning. Despite BEV's growing prominence, its security vulnerabilities remain underexplored, particularly in the context of adversarial machine learning. Existing research has largely focused on adversarial attacks targeting image classification or segmentation models, neglecting the more complex and safety-critical BEV-based systems. In this study, we present a comprehensive security analysis of BEV perception models, examining both vision-only and multi-sensor fusion architectures.

We evaluate the adversarial robustness of state-of-the-art BEV models, including BEVDet, BEVDet4D, Depth-Aware Learning (DAL), and BEVFormer. Unlike traditional attacks that aim to misclassify objects, we design real-world attack scenarios intended to disrupt driving behavior. Our findings reveal that BEV-based systems are highly susceptible to adversarial perturbations, with attacks often transferring across architectures and sensor configurations. Even multi-sensor fusion models show limited resilience, suggesting current fusion strategies are insufficient to counteract adversarial threats. This work highlights critical gaps in the adversarial robustness of BEV perception and emphasizes the urgent need for defense mechanisms tailored to these models. By exposing these vulnerabilities, we aim to catalyze research into more robust and secure perception systems, which are essential for the safe and reliable deployment of autonomous vehicles.

CHAPTER 1

Introduction

Autonomous driving technology has rapidly advanced in recent years, leveraging sophisticated perception models to interpret and navigate real-world environments. These systems rely on a variety of sensors, including cameras, LiDAR, and radar, to construct a comprehensive understanding of their surroundings. Among these, Bird-Eye-View (BEV) perception has emerged as a powerful approach, enabling self-driving vehicles to generate a unified spatial representation from multiple sensor inputs. Despite its growing adoption in industry-leading platforms such as Tesla Autopilot, BEV-based perception remains an underexplored area in adversarial machine learning research.

This report investigates the security vulnerabilities of advanced autonomous driving perception systems, focusing on the **susceptibility of BEV-based detection models to adversarial attacks,** and **its negative influence on the planning module for Autonomous driving**. Unlike traditional computer vision attacks that mislead object classification, adversarial threats against autonomous vehicles pose tangible safety risks, such as forcing vehicles into hazardous decisions or disrupting traffic flow. Specifically, we examine how adversarial perturbations in both vision-only and sensor-fusion models can degrade system performance, potentially leading to critical failures in real-world scenarios.

This report is structured as follows:

- Chapter 2 provides a comprehensive review of related adversarial research in autonomous perception.
- Chapter 3 outlines the experimental setup and co-simulation methodology.
- Chapter 4 and Chapter 5 present adversarial attack strategies against vision-only and vision-LiDAR fusion models, respectively.
- Chapter 6 extends the analysis to black-box attack transferability.
- Chapter 7 explores the effects of perception attacks on the planning module, leading to potentially dangerous driving decisions.
- Chapter 8 concludes with key findings and recommendations for enhancing the robustness of autonomous perception systems against adversarial threats.

Through this study, our findings underscore the need for resilient defense mechanisms to safeguard self-driving technology against emerging cyber-physical threats.

CHAPTER 2

Related Work

2.1 Bird-Eye-View Object Detection

Bird-Eye-View (BEV) detection has become a critical component of autonomous vehicle perception, transforming multi-camera and sensor inputs into a unified top-down spatial representation. BEVDet[1] pioneered vision-only BEV detection by using a two-stage encoding process to extract and transform multi-view image features into BEV space. BEVDet4D[2] and SOLOFusion[16] extend these by incorporating temporal cues, improving motion prediction and tracking, while BEVDepth extends these via depth information for enhanced object localization. Inspired by 3D object detection models[7,8,9,10,11,12] utilizing LiDAR for better object recognition, BEV models with LiDAR signals[3,13,14] enhance BEV perception further through multi-modal sensor fusion, integrating LiDAR signals for improved depth estimation and object localization. BEVFormer[4], a transformer-based model, introduces a historical BEV memory, leveraging attention mechanisms for long-term tracking and improved scene understanding.

While BEV detection enhances 3D perception, it also introduces security concerns. Vision-only models like BEVDet and BEVDet4D are vulnerable to adversarial perturbations that can manipulate object detection. Sensor-fusion models such as DAL[3] and BEVFormer add potential attack surfaces, including LiDAR spoofing and feature manipulation. These threats pose significant risks to autonomous driving safety. This report systematically analyzes adversarial vulnerabilities in BEV-based perception, evaluating attack strategies on BEVDet, BEVDet4D, DAL, and BEVFormer in simulated environments, with a focus on real-world security implications and defensive strategies.

2.2 Adversarial Attack for Vision and LiDAR

Adversarial attacks[5,6,17,18,19,20,21,22,23] pose significant challenges to the security of machine learning systems, particularly in the context of autonomous driving. Brown et al. [5] introduced the concept of the adversarial patch, a universal, robust, and targeted perturbation that, when added to any scene, can mislead image classifiers into predicting a specific target class. These patches are physically realizable and effective under various transformations, highlighting vulnerabilities in visual perception systems. Specifically in the field of autonomous driving, studies have explored the possibilities of attacking the perception module for autonomous driving via camera[24] or LiDAR signals[25,26]. Building upon these findings, MSF-ADV [6] examined the security of multi-sensor fusion (MSF) perception systems in autonomous vehicles. They developed a physically realizable adversarial 3D-printed object designed to be invisible to both camera and LiDAR sensors simultaneously. This attack challenges the assumption that MSF systems are inherently robust against single-sensor attacks by demonstrating that coordinated attacks can compromise all fusion sources, leading to critical perception failures.

These studies underscore the pressing need to develop robust defense mechanisms to safeguard autonomous systems against such adversarial threats.

CHAPTER 3

Co-simulation Development

3.1 Simulation Data Collection

To highlight how an attack on the perception module would actually affect the decision and driving behavior of an autonomous driving agent, we need to run the agent in a simulation world. To this end, we opt for CARLA, an open-source simulator for autonomous driving research, which has been developed from the ground up to support the development, training, and validation of autonomous driving systems. The simulation platform supports flexible specification of sensor suites and environmental conditions.

We set up CARLA to generate and collect simulation data in a nuScenes-like style, which is suitable for BEV detection models. Specifically, we load seven different routes and weather combinations, spawn over a hundred vehicles and pedestrians, and set them in autopilot mode to run the simulation. After that, we deploy six camera sensors, one LiDAR sensor, and six RADAR sensors to one specific vehicle and save the sensor capture to disk at a fixed frequency.
The sensor suite setup is shown in Fig. 1.

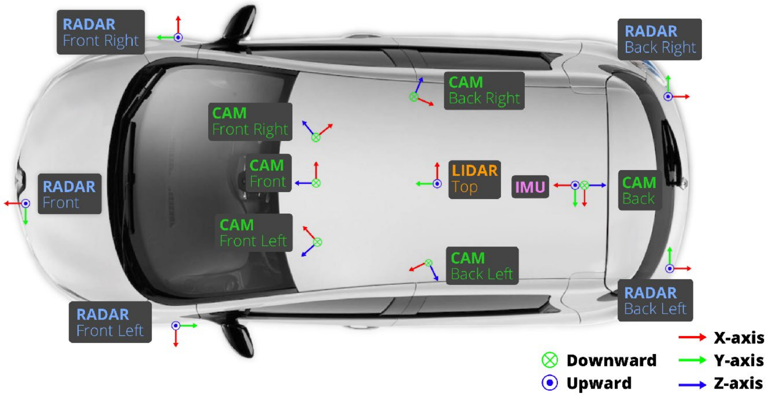

Figure 1: Sensor suite setup in CARLA simulator.

3.2 BEV Detection on Simulation Data

Perceiving 3D environments is essential for autonomous driving as it is crucial for subsequent onboard modules from prediction to planning. Thus, we focus on attacking the perception module and plan to extend to studying the subsequent driving behavior of an autonomous driving agent under attack afterward. The BEV-based detection framework is drawing extensive attention to offer a holistic feature representation space from multi-camera images, and owns the following inherent merits including (1) joint feature learning from multi-view images, (2) unified detection space without post fusion, (3) amenability for temporal fusion, and (4) convenient output representation for the downstream prediction and planning.

We opt for BEVDet as the vision-only baseline model by default due to its simplicity and efficiency, as well as its well-structured codebase. As shown in the following figure, it adopts a double-encoder structure for image-view and bird-eye-view representation learning and a 2D-3D view transformer to connect representations from these two different views.

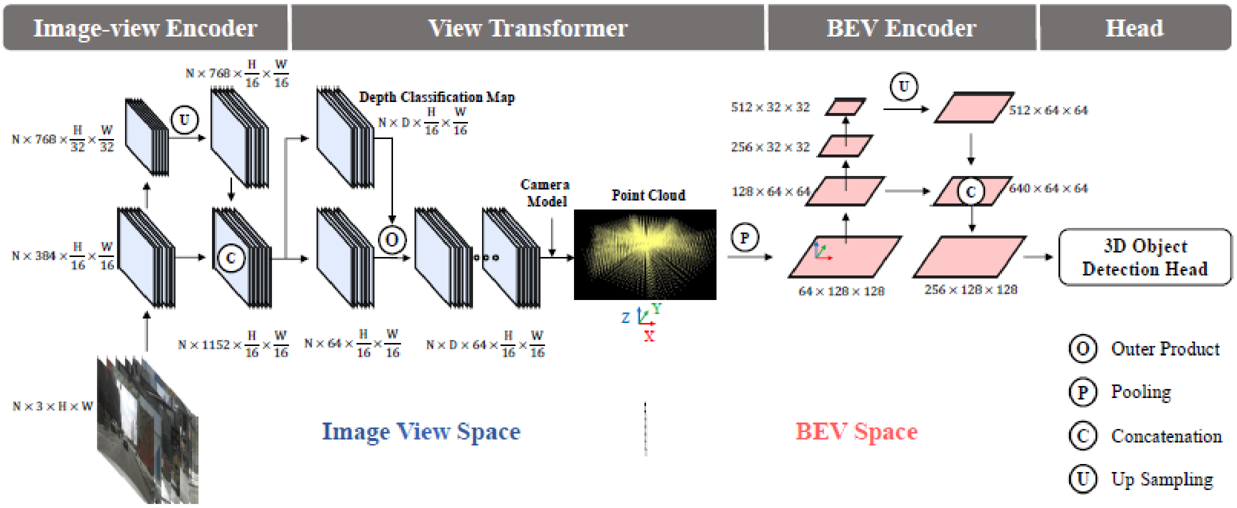

Figure 2: The structure overview of BEVDet.

CHAPTER 4

Adversarial Attack on Vision-only Model

4.1 Attacking BEVDet with Single-Frame Input

4.1.1 PatchAttack on BEVDet

As the first step of developing an adversarial attack framework, we conduct a comprehensive adversarial vulnerability analysis of vision-based BEV detection models in the digital world. Specifically, we adapt a range of existing white-box attacks, including PGD-Attack, FGSM-Attack, C&W-Attack, and AutoPGD, to the BEV setting. Let $\mathbf{I} \in R^{C\times H\times W}$ be an input image, comprising N targets given by $T = \{t_1, t_2, t_3, ..., t_N\}$. By feeding the image **I** into 3D object detectors, we can have n perception results, capturing class, 3D bounding boxes, and other attributes, represented as $f(\mathbf{I}) = \{y_1, y_2, y_3, ..., y_n\}$. Here, each y_i symbolizes a discrete detection attribute such as localization, class, velocity, etc. We then compare these predictions with the ground truth bounding boxes T, establishing a match when the 2D center distances on the ground plane are under a predefined threshold. We hereby consider both pixel-based attacks, where bounded perturbations are added to the whole image, and patch-based attacks, where unbounded perturbations are added into a pre-defined region of the image. Note for patch-based attack, considering a target within a 3D bounding box, it can be characterized by its eight vertices and a central point, collectively denoted as $\{c_o, c_1, ..., c_8\}$ with $c_i \in R^3$. Leveraging the camera parameters, we project these 3D points to 2D points on the image plane, yielding the transformed set $\{c'_o, c'_1, ..., c'_8\}$. We define the size of the adversarial patch to be proportional to the size of the rectangle formed by these 2D points, and strategically position the adversarial patch to be centered at the point c'_o.

Regarding the attack configuration, we consider untargeted attacks for each target and maximize the following objective:

$$L_{untargeted} = -\frac{1}{N}\sum_{i=1}^{N} \sum_{j=1}^{C} f_{cls}^{j}(I + r, t_i)\, log\, log\, p_{ij}$$

where C denotes the number of classes, and f_{cls}^{j} denotes the confidence score on $j - th$ class. The adversarial perturbation **r** is optimized iteratively using PGD-Adv, as $r_{i+1} = Proj_{\epsilon}\left(r_i + \alpha sgn\left(\nabla_{I+r_i}, L\right)\right)$. To facilitate an equitable comparison, the confidence scores undergo normalization within the range $[0,1]$ by using the sigmoid function, which mitigates sensitivity to unbounded logit ranges.

To attack the localization and other attributes, we adopt the straightforward L_1loss as the objective function,

$$L_{localization} = \frac{1}{N}\sum_{i=1}^{N} \left\|f_{loc}(I+r,t_i) - loc_i\right\|_1 + \left\|f_{orie}(I+r,t_i) - orie_i\right\|_1 + \left\|f_{vel}(I+r,t_i) - vel_i\right\|_1.$$

Using these objective functions together could complete our setup to adversarially attack the BEV-based object detectors.

4.1.2 Evaluation Results

We evaluate BEVDet-R50 on our collected CARLA simulation data and show the results in Table 1.

Table 1: mAP of BEVDet-R50 w/o and w/ PatchAttack.

	Car	*Pedestrian*
w/o Attack	*0.22*	*0.14*
w/ Attack	*0.00*	*0.00*

Additionally, we visualize the predicted bounding box before and after attack in Fig. 3, Fig. 4.

We can observe that most vehicles and pedestrians are successfully detected by our perception module, except the ones that are very far away without PatchAttack in Fig. 3, meanwhile in Fig. 4 we can observe that most vehicles and pedestrians now cannot be detected by the model, indicating the effectiveness of applying a small adversarially crafted patch on target object.

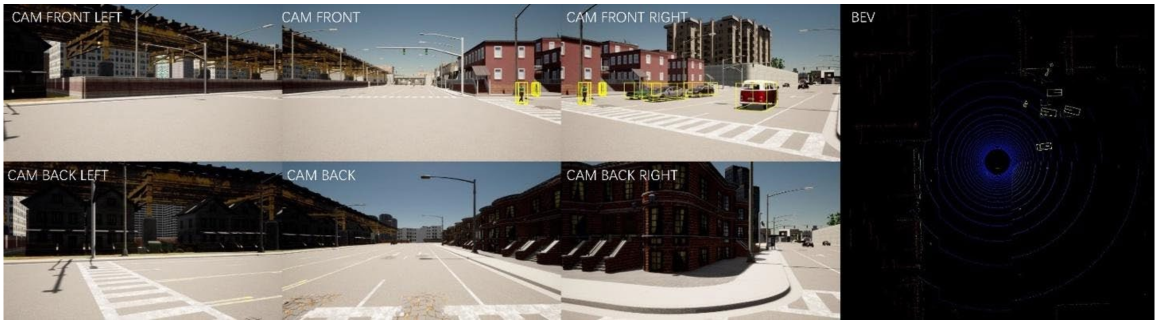

Figure 3: Detection Results w/o PatchAttack.

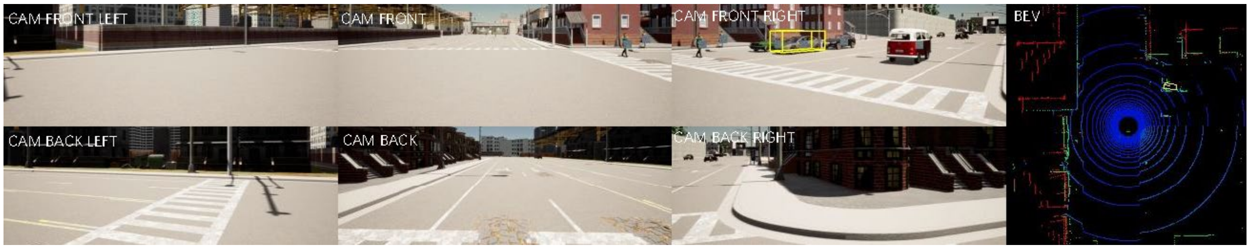

Figure 4: Detection Results w/ PatchAttack.

4.2 Attacking BEVDet with Temporal Input

Temporal cues are proven to be beneficial for accurate localization and velocity estimation in BEV detectors. However, it also gives a malicious adversary more channels to attack as information from multiple timestamps is gathered and processed together. To extend beyond simple single-frame images input in Chapter 4.1, and additionally consider temporal information in adversarial attacks, we experiment with the temporally extended version of BEVDet, BEVDet4D. It retains the intermediate BEV feature of the previous frame and concatenates it with the ones generated by the current frame before using the features for predictions.

4.2.1 PatchAttack on BEVDet4D

Given the nature of BEVDet4D, we consider three scenarios: (a) *Benign case*: The model processes clean input across multiple frames. (b) *Single adversarial attack*: The model processes multiple frames with only the last frame being adversarially impacted. (c) *Temporal-Continuous adversarial attack*: The adversarial input exists persistently across multiple timestamps. This results in all sequential inputs used for temporal information modeling being adversarial examples, causing an accumulation of errors within the model through retained temporal data. Note that in the last scenario, we assume a single fixed attack patch attached to an object (*e.g.,* a car or a person) results in multi-frame varying attack patches captured by the sensors (*e.g.,* camera), which is closer to the real-world setup.

Benign case is used to evaluate the baseline performance when there is no adversary attack.

Single adversarial attack is performed by simply applying the attack we developed in Chapter 4.1 onto the last frame in each frame sequence. This approach shows some effectiveness but not as much compared to its performance when the model only takes in a single frame in Chapter 4.1.

Temporal-continuous adversarial attack performs the attack on all the frames in the frame sequence instead of only the last one. To make the generated adversarial samples consistent with the movement of objects, the adversarial samples are generated in the 3D space and then translated into 2D space instead of directly generating the samples in the 2D space. As shown in Table 1, We can observe that the results are significantly improved as the attack achieved a 100% success rate.

Table 2: mAP of BEVDet4D-R50 w/ single adversarial attack and temporal-continuous adversarial attack.

Model	Attack	mAP	
		Car	Pedestrian
BEVDet	*Single*	*0.15*	*0.08*
	Temporal-Continuous	*0.00*	*0.00*

CHAPTER 5

Adversarial Attack on Vision-LiDAR Model

Compared with camera-captured images, 3D LiDAR data is another important data modality that is commonly used in the autonomous driving industry. Compared to cameras, LiDAR measures provide more accurate 3D geometric cues, such as depth and shapes, but are inherently more sparse and less semantic-oriented. Therefore, the complementary nature of different data modalities has motivated the design of multi-modal sensor-fusion detection models. In principle, the Multi-Sensor Fusion (MSF) model design can be more robust against malicious attacks under the assumption that not all sources are attacked at the same time.

5.1 BEV Models with Vision and LiDAR Input

Compared with camera-captured images, 3D LiDAR data is another important data modality that is commonly used in the autonomous driving industry. Compared to cameras, LiDAR measures provide more accurate 3D geometric cues, such as depth and shapes, but are inherently more sparse and less semantic-oriented. Therefore, the complementary nature of different data modalities has motivated the design of multi-modal sensor-fusion detection models. In principle, the Multi-Sensor Fusion (MSF) model design can be more robust against malicious attacks under the assumption that not all sources are attacked at the same time.

To study how additional LiDAR modality affects the detection performance when encountering attacks, we use the Detecting-As-Labeling (DAL) model, an extended version of BEVDet with the ability to deal with both LiDAR and temporal information, in our experiments. In each frame, an object is placed on top of a vehicle to interfere with the LiDAR measurements, altering the captured point cloud data to further interfere with the DAL's detection results.

5.1.1 Attacking DAL via Adversarial Vision and LiDAR Signal

To generate the adversarial LiDAR signal, we use MSF-ADV to alter the shape of an attack object represented by a 3D polygon mesh for each vehicle and pedestrian in the scene to generate an adversarial polygon mesh. That object is to be attached to the surface of a vehicle or pedestrian to alter the LiDAR signal. Similar to the patch-based attack, we propagate the gradient from the optimization objective to a benign 3D object. The gradient is then used to alter the shape of the said benign 3D object to make it adversarial. The pipeline overview is illustrated in Fig. 5.

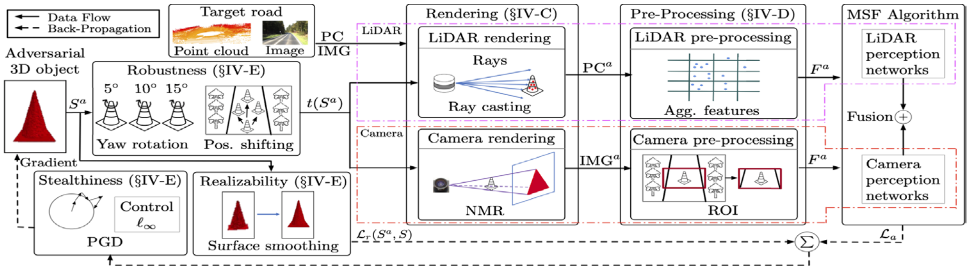

Figure 5: The pipeline overview of MSF-ADV.

The objective of optimization is composed of three loss functions. The first one is simply the confidence score of the vehicle or pedestrian to be undetected.

$$\mathcal{L}_a = y$$

where y is the model’s predicted confidence score for the object to be made invisible.

The second one is a laplacian smoothing loss defined as

$$\mathcal{L}_r = \sum_i \left\| \mathbf{v}_i - \frac{1}{|N(i)|} \sum_{j \in N(i)} \mathbf{v}_j \right\|^2$$

where M is the total number of vertices in the polygon mesh, v_i is the 3D coordinates of a vertex in the polygon mesh, v_j is the 3D coordinates of a neighboring vertex adjacent to v_i, N(i) is the total number of vertices adjacent to the v_i. The purpose of this loss is to smooth out the surface of the adversarial object, therefore increasing the realizability to 3D print the object.

The last one is the stealthiness loss to constrain the difference between the adversarial polygon mesh and the original polygon mesh so that it may look stealthier and natural. This loss is defined as the mean maximum absolute difference (l_∞) between the vertices in the adversarial polygon mesh and the ones in the original mesh.

$$\mathcal{L}_s = \frac{1}{M} \sum_{i=1}^{M} \|\mathbf{v}_i - \mathbf{v}'_i\|_\infty$$

where v'_i is the 3D coordinates of a vertex in the original polygon mesh corresponding to v_i.
A visualized example of the generated adversarial object to attack LiDAR is illustrated in Fig. 6.

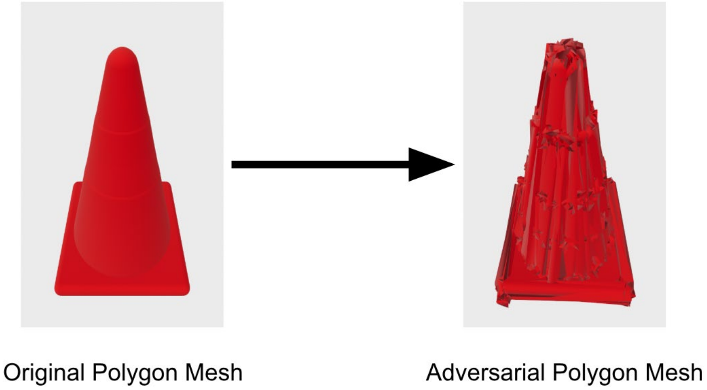

Figure 6: One example of the adversarial polygon mesh.

5.1.2 Evaluation of Adversarial Attack via Both Visual Sequence and LiDAR

Since BEVDet doesn’t take LiDAR signals as inputs, we use DAL to evaluate the performance of the LiDAR attack. In each frame, an object is placed on top of a vehicle to interfere with the LiDAR measurements, altering the captured point cloud data to further interfere with the DAL’s detection results. The patch-based attack designed for attacks along visual temporal sequences from Chapter 4.2 is also applied simultaneously, thus making sure the model receives adversarial signals via both modalities (*i.e.*, vision and LiDAR). Fig. 7 is a visual illustration of our attack: a patch is attached to the back of the vehicle to interfere with the camera, and an adversarial object is placed on the top of the vehicle to disturb the LiDAR sensor.

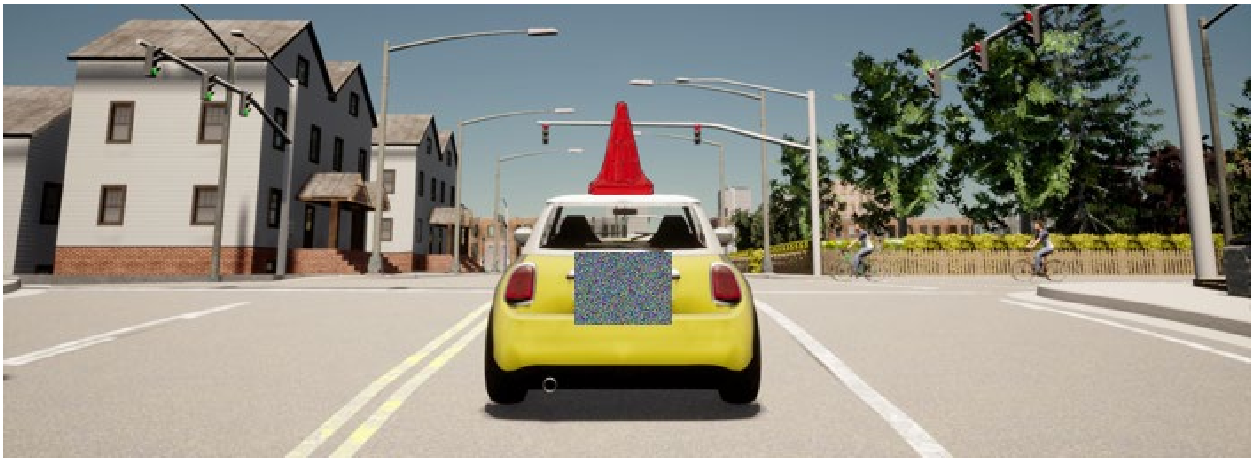

Figure 7: The front camera view w/ vision-lidar adversarial attack.

Although only applying the attack on the visual temporal sequences alone can greatly lower the detection accuracy of a vision-only model such as BEVDet4D, it is shown in Tab. 3 that the additional adversarial attack on the LiDAR signals still improves the attack performance.

Table 3: mAP of DAL w/ adversarial attack via Visual Sequence and LiDAR.

Model	Attacked Modality	mAP	
		Car	Pedestrian
BEVDet	*Visual Sequence*	*0.20*	*0.13*
	Visual Sequence + LiDAR	*0.00*	*0.00*

CHAPTER 6

Transitioning to Black-Box System Simulations

6.1 Experimental Setup

So far, we have only been working on BEVDet and its derivative models. However, an intriguing property of adversarial examples that makes them threaten in the real world is their transferability. Transferable attacks assume a realistic scenario where the adversarial examples generated on a (local) surrogate model can be directly transferred to the (*unknown*) target model. Such attacks require no interaction with the target model, nor any prior knowledge of the target model, and thus are more dangerous to safety-critical applications like autonomous driving.

Therefore, to study the security implications of transferable attacks, we choose BEVFormer as the target model and evaluate its performance on the adversarially impacted images generated based on BEVDet4D. BEVFormer is another popular BEV-based 3D detection model, with two major differences when compared with BEVDet and its various variant models (*e.g.,* BEVDet4D and DAL). First, the former is a Transformer-based model, while the latter ones are CNN-based models. Second, the temporal version of the former maintains and updates a historic BEV feature memory, while the temporal version of the latter simply uses the BEV feature from a previous timestamp. If the attack transfer success rate remains high even with these distinct differences, it means an even more severe security threat in the current perception modules.

6.2 Experimental Results

The experimental results in Tab. 4 demonstrate that attacking BEVFormer with the adversarial example generated based on BEVDet leads to a surprisingly high success rate.

Table 4: mAP of BEVFormer w/ and w/o adversarial attack.

Model	Attack	mAP	
		Car	Pedestrian
BEVFormer	*w/o Attack*	*0.24*	*0.13*
	w/ Attack	*0.07*	*0.03*

We additionally provide the visualization comparison below. In Fig. 8, it is shown that the BEVFormer has no trouble detecting the vehicle forward w/o attack. However, in Fig. 9, observe that most vehicles and pedestrians now cannot be detected by the model, indicating the effectiveness of our attack method.

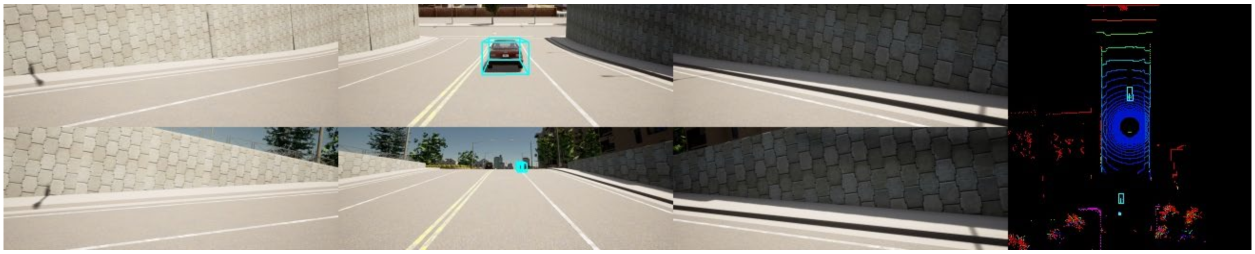

Figure 8: BEVFormer detection results w/o attack.

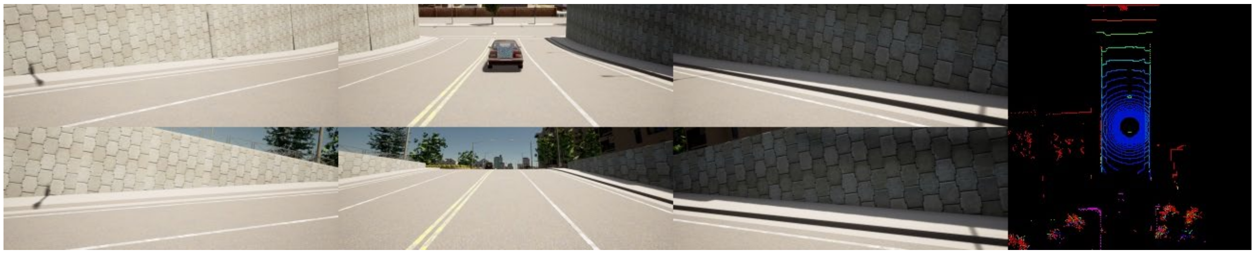

Figure 9: BEVFormer detection results w/ attack.

CHAPTER 7

Attacking the Planning Module

To investigate how adversarial attacks on perception models affect autonomous driving decisions, we aim to build an agent in CARLA with a perception module (e.g., BEVDet) and a planning module for driving actions like acceleration and braking. This involves integrating components like adversarial input generation and multi-camera 3D detection models into CARLA's codebase, enabling them to work together to produce real-time control signals for a CARLA agent (e.g., a car).

7.1 Planning Module Implementation

We first describe how we implement the planning module. Specifically, to highlight the effect of wrong detection results under attack and minimize the effect of other modules, we design a simple and straightforward planning module: Accelerate to 16 m/s when there is no obstacle detected within 20 meters ahead and decelerate to match the speed of the obstacle ahead if otherwise. Additionally, we create a simple scene in which two cars are created on a single, straightforward lane. One car is the main vehicle controlled by the autonomous agent, while another car is running at 8 m/s ahead. A 3D detection model will act as the perception module to detect any obstacles ahead for decision-making. If the perception module is successfully attacked and fails to recognize the parked car ahead, the car behind will crash into the car ahead, indicating severe safety repercussions.

7.2 Attacking Framework

In our implementation, to attach the adversarial patch to the back of the vehicle, the back facet of the front vehicle's bounding box is used to determine the 4 corners of the adversarial patch. The facet is resized to half of the original size, and its 4 corners' coordinates are translated to the coordinates in the camera view. The adversarial patch is then warped to fit the quadrilateral defined with these 4 corners in the camera view. The adversarial 3D object is rendered outside Carla with **open3d**. Box blur is then applied to the output image from open3d to suppress the noise in the background before the blank area is removed via cropping. The resulting image is then patched onto the top of the front vehicle in the camera view from Carla with a similar procedure for attaching the adversarial patch to the front vehicle. The lidar signal of the adversarial 3D object is also generated outside Carla and then merged with lidar signals from Carla before being fed to the detector.

7.3 Results Visualization

A simple visualization of the third-person view of an autonomous vehicle in CARLA is shown below. In Fig. 10, no attack is employed, and the front vehicle is being detected by the rear one. In Fig. 11, the attack is employed on the front vehicle, making it invisible to the rear one. In Fig. 12, as no obstacle is detected, the rear vehicle accelerates and collides with the front one.

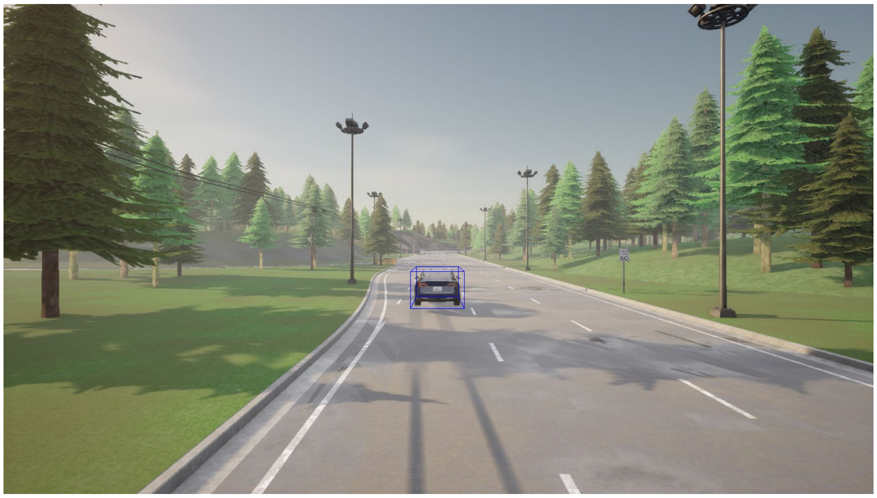

Figure 10: This is a No Attack Scenario, where the car can be successfully detected.

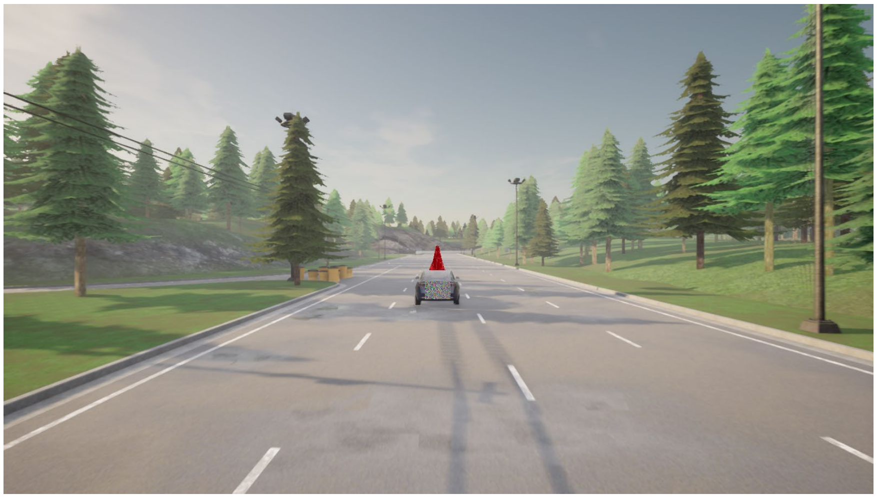

Figure 11: This is an Attack Scenario, where the car appears “invisible”.

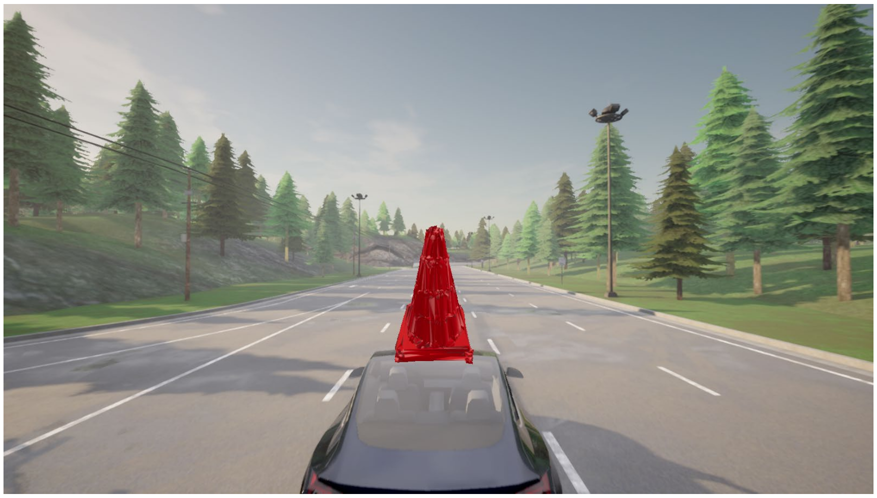

Figure 12: This is a successful Collision Scenario, which leads to a traffic accident.

CHAPTER 8

Discussion

Defense Mechanisms and Mitigation Strategies

Defending against multi-modal adversarial attacks is particularly challenging due to the complexity of cross-modal fusion and the high dimensionality of sensor inputs. Nonetheless, several directions can be explored. One such approach is adversarial training, which involves incorporating adversarial examples, especially those involving joint perturbations across modalities, into the training pipeline to improve model robustness. However, this method must generalize across diverse attack types and scale appropriately with increasing sensor complexity. Another strategy involves sensor cross-validation, where architectures are designed to perform internal consistency checks across modalities. For example, verifying LiDAR point cloud projections against vision-based outputs can help detect anomalies, with discrepancies serving as potential indicators of tampering or unreliable perception. Ensemble methods and redundancy-based defenses can also offer resilience by using multiple independently trained models or late fusion techniques. These systems can leverage consensus mechanisms, so that even if an attack succeeds against one model, others may still yield accurate results.

Certified robustness and formal verification can guarantee bounded robustness properties for high-dimensional and multi-modal models. In addition, input preprocessing and filtering may provide practical, lightweight defenses. For LiDAR data, this could include statistical filtering of anomalous point clouds, while for vision data, transformations like JPEG compression, Gaussian smoothing, or patch masking have shown some success in mitigating adversarial noise. No single defense is likely to be sufficient. Instead, a layered security approach—one that integrates detection mechanisms, redundancy, and robust training—is necessary to protect future AV perception systems.

Deployment Challenges in Real-World AV Systems

While the adversarial vulnerability analysis presented in our study is rigorous and highlights critical risks, translating these insights into real-world autonomous vehicle (AV) deployments presents numerous challenges. One major issue is computational overhead. Generating and defending against adversarial examples in real-time, especially in safety-critical environments, imposes substantial computational costs. As current AV stacks are already resource-intensive, any additional security layer must be highly efficient and optimized for deployment.

Another challenge stems from the fact that many commercial AV systems operate as black boxes, using proprietary models or closed software stacks. This makes it difficult to design or test robust defenses without full access to system internals, limiting the practical deployment of white-box-informed techniques. Additionally, domain adaptation and generalization remain significant hurdles. While our evaluations are grounded in datasets such as nuScenes, real-world environments are far more diverse in terms of geography, weather, lighting, and road conditions. Consequently, adversarial vulnerabilities may vary significantly, and defenses trained in one domain may not generalize well without further adaptation.

CHAPTER 9

Conclusions

In this paper, we conducted a systematic evaluation of adversarial vulnerabilities in Bird's Eye View (BEV) perception models used for autonomous driving. Our analysis focused on multiple BEV-based detection frameworks, including BEVDet, BEVDet4D, DAL, and BEVFormer, assessing their robustness against adversarial attacks on both vision-only and multi-sensor fusion systems. We evaluated these models across various adversarial scenarios, including patch-based attacks, temporal adversarial strategies, and LiDAR spoofing, with a particular focus on their impact on real-world driving safety.

Our findings indicate that vision-only models like BEVDet and BEVDet4D are highly susceptible to adversarial perturbations, leading to significant perception failures. While multi-modal sensor fusion models like DAL and BEVFormer improve perception accuracy, they remain vulnerable to coordinated vision-LiDAR attacks, highlighting the limitations of current multi-sensor security strategies. Furthermore, adversarial transferability across models underscores the broader risk to BEV-based perception systems, even when adversaries lack direct access to the target architecture.

These results emphasize the urgent need for robust adversarial defenses tailored for BEV-based perception, incorporating adaptive security mechanisms to mitigate real-world attack risks. Future work should extend adversarial evaluations to more complex driving environments and real-world scenarios to further strengthen the security of autonomous perception systems.

REFERENCES

1. Huang, J., Huang, G., Zhu, Z., Ye, Y. and Du, D., 2021. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. arXiv preprint arXiv:2112.11790.
2. Huang, J. and Huang, G., 2022. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv:2203.17054.
3. Huang, J., Ye, Y., Liang, Z., Shan, Y. and Du, D., 2024, September. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. In European Conference on Computer Vision (pp. 439-455). Cham: Springer Nature Switzerland.
4. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Yu, Q. and Dai, J., 2024. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence.
5. Brown, T.B., Mané, D., Roy, A., Abadi, M. and Gilmer, J., 2017. Adversarial patch. arXiv preprint arXiv:1712.09665.
6. Cao, Y., Wang, N., Xiao, C., Yang, D., Fang, J., Yang, R., Chen, Q.A., Liu, M. and Li, B., 2021, May. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In 2021 IEEE symposium on security and privacy (SP) (pp. 176-194). IEEE.1
7. Chen, X., Ma, H., Wan, J., Li, B. and Xia, T., 2017. Multi-view 3d object detection network for autonomous driving. In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (pp. 1907-1915).
8. Sindagi, V.A., Zhou, Y. and Tuzel, O., 2019, May. Mvx-net: Multimodal voxelnet for 3d object detection. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 7276-7282). IEEE.
9. Vora, S., Lang, A.H., Helou, B. and Beijbom, O., 2020. Pointpainting: Sequential fusion for 3d object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 4604-4612).
10. Wang, C., Ma, C., Zhu, M. and Yang, X. 2021 PointAugmenting: cross-modal augmentation for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June (pp. 11794–11803).
11. Wu, H., Wen, C., Shi, S., Li, X. and Wang, C., 2023. Virtual sparse convolution for multimodal 3d object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 21653-21662).
12. Wu, X., Peng, L., Yang, H., Xie, L., Huang, C., Deng, C., Liu, H. and Cai, D., 2022. Sparse fuse dense: Towards high quality 3d detection with depth completion. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5418-5427).
13. Liang, T., Xie, H., Yu, K., Xia, Z., Lin, Z., Wang, Y., Tang, T., Wang, B. and Tang, Z., 2022. Bevfusion: A simple and robust lidar-camera fusion framework. Advances in Neural Information Processing Systems, 35, pp.10421-10434.
14. Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D.L. and Han, S., 2023, May. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In 2023 IEEE international conference on robotics and automation (ICRA) (pp. 2774-2781). IEEE.
15. Li, Y., Ge, Z., Yu, G., Yang, J., Wang, Z., Shi, Y., Sun, J. and Li, Z., 2023, June. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 2, pp. 1477-1485).

16. Park, J., Xu, C., Yang, S., Keutzer, K., Kitani, K., Tomizuka, M. and Zhan, W., 2022. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. arXiv preprint arXiv:2210.02443.
17. Goodfellow, I.J., Shlens, J. and Szegedy, C., 2014. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
18. Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B. and Swami, A., 2016, March. The limitations of deep learning in adversarial settings. In 2016 IEEE European symposium on security and privacy (EuroS&P) (pp. 372-387). IEEE.
19. Xiao, C., Li, B., Zhu, J.Y., He, W., Liu, M. and Song, D., 2018. Generating adversarial examples with adversarial networks. arXiv preprint arXiv:1801.02610.
20. Xiao, C., Zhu, J.Y., Li, B., He, W., Liu, M. and Song, D., 2018. Spatially transformed adversarial examples. arXiv preprint arXiv:1801.02612.
21. Xiao, C., Deng, R., Li, B., Yu, F., Liu, M. and Song, D., 2018. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 217-234).
22. Pan, X., Xiao, C., He, W., Yang, S., Peng, J., Sun, M., Yi, J., Yang, Z., Liu, M., Li, B. and Song, D., 2019. Characterizing attacks on deep reinforcement learning. arXiv preprint arXiv:1907.09470.
23. Qiu, H., Xiao, C., Yang, L., Yan, X., Lee, H. and Li, B., 2020. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16 (pp. 19-37). Springer International Publishing.
24. Zhang, Y., Hassan Foroosh, P.D. and Gong, B. (2019) 'CAMOU: learning a vehicle camouflage for physical adversarial attack on object detections in the wild', *International Conference on Learning Representations (ICLR 2019)*.
25. Cao, Y., Xiao, C., Cyr, B., Zhou, Y., Park, W., Rampazzi, S., Chen, Q.A., Fu, K. and Mao, Z.M., 2019, November. Adversarial sensor attack on lidar-based perception in autonomous driving. In Proceedings of the 2019 ACM SIGSAC conference on computer and communications security (pp. 2267-2281).
26. Sun, J., Cao, Y., Chen, Q.A. and Mao, Z.M., 2020. Towards robust {LiDAR-based} perception in autonomous driving: General black-box adversarial sensor attack and countermeasures. In 29th USENIX Security Symposium (USENIX Security 20) (pp. 877-894).