Probing Vulnerabilities of Vision-LiDAR Based Autonomous Driving Systems

Siwei Yang[1], Zeyu Wang[1], Diego Ortiz[1], Luis Burbano[1],
Murat Kantarcioglu[2], Alvaro A. Cardenas[1], Cihang Xie[1]
[1]University of California, Santa Cruz [2]Virginia Tech

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. However, in the meantime, the security vulnerabilities of BEV-based models are starting to be examined within the scope of adversarial machine learning research. This study provides a preliminary security analysis of BEV perception models, focusing on adversarial attacks employed in different modalities, including both visual signals from cameras and point-cloud signals from LiDAR. Specifically, we examine the vulnerabilities of state-of-the-art models–including BEVDet, BEVDet4D, DAL, and BEVFormer–to different forms of adversarial attacks. In addition to the white-box setup, we also check the transferability of these attacks to black-box models.

Our findings reveal that although multi-modal inputs significantly improve BEV models' detection performance, they also introduce new channels for adversarial attacks and hence increase vulnerability. As long as the adversarial attack is applied to all modalities that a model takes in, e.g.*, adversarial perturbation is added to both vision and LiDAR signals for a vision-LiDAR model, and the attack can always achieve an almost complete success rate, while utilizing incomplete modalities to attack results in sub-optimal outcomes, as this strategy allows detection models to still capture undisturbed information. Moreover, we show that the designed attack can transfer across totally different BEV architectures. For example, adversarial input trained with DAL, a CNN-based model, can still transfer to BEVFusion and significantly degrade its performance, despite BEVFusion using a transformer-based architecture. We hope that this work can help to raise the community's attention to the vulnerability of BEV-based autonomous driving systems.*

1. Introduction

Autonomous driving technology has advanced rapidly 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 [6–8, 10–12] has emerged as a powerful approach, enabling self-driving vehicles to generate a unified spatial representation from multiple sensor inputs. Given its growing adoption on industry-leading platforms such as Waymo and the possibilities that adversarial threats against autonomous vehicles can pose tangible safety risks in real life, adversarial attacks on BEV-based perception are starting to attract interest from the research community [3, 4, 9, 24].

This work investigates the security vulnerabilities of advanced autonomous driving perception systems, focusing on how adversarial attacks employed on different modalities affect multi-modal BEV-based detection models. Specifically, we develop adversarial attacks through different modalities for models with different input formats, *e.g.*, single-frame vision-only attack, multi-frame vision-only attack, and multi-frame vision-LiDAR attack. Our research shows several interesting findings. First, we find that vision-based models such as BEVDet [7] and BEVDet4D [6] are highly susceptible to adversarial attacks, leading to perceptual failures. Multi-sensor models such as DAL [8] and BEVFormer [11] improve accuracy but remain vulnerable to synchronized vision-LiDAR attacks, revealing flaws in current multi-sensor security. Additionally, we found that although extra input signals such as multi-frame camera images and LiDAR can enhance models' robustness against adversarial perturbation, they also introduce new channels for more adversarial inputs, hence more vulnerabilities. For white-box attacks, as long as the adversarial input is applied to all input signals, *e.g.*, the adversarial perturbation is added to both the LiDAR signals and all frames in the input image sequence for a multi-frame vision-LiDAR BEV model, the adversarial attack can achieve a complete suc-

cess rate.

More interestingly, we note that the adversarial attack that we developed shows strong transferability to black-box models when testing, even if the testing-time model and training-time model use completely different architectures, *e.g.*, CNN-based and Transformer-based architectures. This strong transferability among models further stresses the feasibility of employing such adversarial attacks in real-life scenarios. We also include a straightforward simulation showing how our generated adversarial image patch and polygon mesh, when attached to a vehicle, lead to the failure of the rear car's BEV perception module. This, in turn, disrupts the rear vehicle's planning module, ultimately resulting in a practical traffic accident.

We believe that the vulnerability findings in this paper also highlight the urgent need for robust adversarial defenses tailored for BEV-based perception systems. Future research should broaden adversarial assessments in complex and real-world driving scenarios to enhance the security of autonomous systems, emphasizing the importance of resilient defenses against cyber-physical threats.

2. Related Works

Bird-Eye-View Detection 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 [7] pioneered vision-only BEV detection by using a two-stage encoding process to extract and transform multi-view image features into BEV space. BEVDet4D [6] and SOLOFusion [13] extend these by incorporating temporal cues, improving motion prediction and tracking, while BEVDepth extends these via depth information for improved object localization.

Inspired by 3D object detection models [5, 15, 17–20] that utilize LiDAR for better object recognition, BEV models with LiDAR signals [8, 12] enhance BEV perception further through multi-modal sensor fusion, integrating LiDAR signals for improved depth estimation and object localization. BEVFormer [11], a transformer-based model, introduces a historical BEV memory, leveraging attention mechanisms for long-term tracking and improved scene understanding.

Although BEV detection improves 3D perception, it also introduces security concerns. Vision-only models such as BEVDet and BEVDet4D are vulnerable to adversarial perturbations that can manipulate object detection. Sensor-fusion models such as DAL [8] and BEVFusion [11] add potential attack surfaces, including LiDAR spoofing and feature manipulation. These threats pose significant risks to the safety of autonomous driving. 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.

Adversarial Attacks. Adversarial attacks [1, 3, 14, 14, 21–23] pose significant challenges to the security of machine learning systems, particularly in the context of autonomous driving. Brown *et al*. [1] introduced the concept of an 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 through camera [25] or LiDAR signals [2, 16]. Based on these findings, MSF-ADV [3] 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 show that both vision-only and sensor-fusion BEV models are susceptible to adversarial perturbation in input signals, making autonomous driving systems relying on these models vulnerable to malicious attackers.

3. Methodology

3.1. Adversarial Attack on Vision-Only Model

3.1.1. Vision-Only Model with Single-Frame Input

As the first step in developing an adversarial attack framework, we adapt an existing white-box attack method, *e.g.* PGD-Attack, to the BEV setting. Let $I \in \mathrm{R}^{C\times H\times W}$ be an input image comprising the N targets given by $T = \{t_1, t_2, \ldots, t_N\}$. By feeding the image I, into the 3D object detectors, we get n perception results, capturing class, 3D bounding boxes, and other attributes, represented as $f(I) = \{y_1, y_2, \ldots, y_n\}$. Here, y_i symbolizes a discrete detection attribute such as location, category, 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 below 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 that for the 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 $\{c_0, c_2, \ldots, c_8\}$

with each $c_i \in \mathrm{R}^3$. Using the camera parameters, we project these 3D points onto 2D points in the image plane, which yields the transformed set $\{c'_0, c'_2, \ldots, c'_8\}$. 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'_0.

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

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

where C denotes the number of classes and f_{cls}^{j} denotes the confidence score in the j-th class.

For a fair comparison, confidence scores undergo normalization within the range [0,1] using the sigmoid function, which mitigates sensitivity to unbounded logit ranges.

To attack the localization and other geometrical attributes such as orientation and velocity, we adopt the straightforward L1 loss as the objective function,

$$L_{\text{geo}} = \frac{1}{N}\sum_{i=1}^{N}(\|f_{\text{loc}}(I + r, t_i) - \text{loc}_i\|_1 \tag{2}$$

$$+ \|f_{\text{orie}}(I + r, t_i) - \text{orie}_i\|_1 \tag{3}$$

$$+ \|f_{\text{vel}}(I + r, t_i) - \text{vel}_i\|_1). \tag{4}$$

Using these objective functions together completes our adversarial attack to BEV-based object detectors:

$$L = L_{\text{cls}} + L_{\text{geo}} \tag{5}$$

The adversarial perturbation r is optimized iteratively using Projected Gradient Ascent, as

$$r_{i+1} = r_i + \alpha \times sgn(\nabla_{I+r_i} L). \tag{6}$$

3.1.2. Vision-Only Model with Multi-Frame Input

Temporal signals are beneficial for accurate location and velocity estimation in BEV detectors; however, they increase the attack surface as information from multiple timestamps is gathered and processed together.

Given the BEV detection model with multi-frame input, we consider two kinds of vision-only adversarial attacks:

- **Single-frame adversarial attack** is performed by simply applying the attack we developed in Sec. 3.1.1 onto the last frame in each frame sequence;
- **Multi-frame 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 to 2D space instead of directly generating the samples in the 2D space.

The BEV detection model's performance under each scenario is discussed in Sec. 4.3.2.

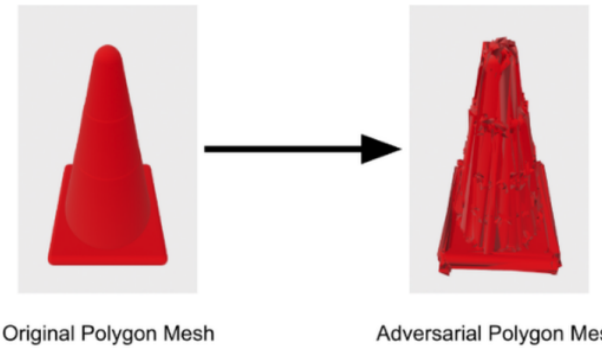

Figure 1. One example of our adversarial polygon mesh.

3.2. Adversarial Attack on Vision-LiDAR Model

Compared to 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 adversarily attack a BEV detection model via LiDAR signals, an object can be placed on top of a vehicle to interfere with the LiDAR measurements, altering the captured point-cloud data to influence the model detection results.

We use MSF-ADV [3] 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. Similarly to the patch-based attack, we propagate the gradient from the optimization objective to a benign 3D object, as shown in Fig. 1. The gradient is then used to alter the shape of that benign 3D object to make it adversarial.

The optimization objective consists of three loss functions. The first is simply the confidence score that a vehicle or pedestrian will be undetected.

$$L_{\text{a}} = y, \tag{7}$$

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

$$L_r = \sum_i \left\| v_i - \frac{1}{|N(i)|}\sum_{j \in N(i)} v_j \right\|^2 \tag{8}$$

Figure 2. The front camera view w/ Vision-LiDAR adversarial attack.

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 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 between the vertices in the adversarial polygon mesh and the ones in the original mesh.

$$L_s = \frac{1}{M} \sum_{i=1}^{M} ||v_i - v_i'||_\infty \tag{9}$$

where v_i' are the 3D coordinates of a vertex in the original polygon mesh corresponding to v_i.

3.2.1. Temporal-Continuous Vision-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 model's detection results. The patch-based attack designed for attacks along visual temporal sequences as mentioned in Sec. 3.2 is also applied simultaneously, making sure that the model receives adversarial signals via both modalities (ie, vision and LiDAR). Fig. 2 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.

4. Experiments

4.1. Simulation Settings

To highlight how the 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.

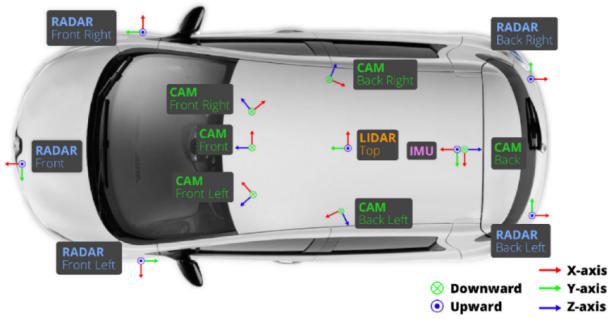

Figure 3. Sensor suite setup in CARLA simulator.

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

Table 1. mAP of BEVDet-R50 w/o and w/ vision-only single-frame attack.

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 a specific vehicle and save the sensor capture to disk at a fixed frequency. The setup of the sensor suite is shown in Fig. 3.

4.2. Evaluation Settings

When developing vision-only adversarial attacks, we use BEVDet [7] and BEVDet [6] to train our adversarial image patches. For vision-LiDAR models, we choose DAL [8] to train both our adversarial image patches and our adversarial polygon mesh.

The adversarial inputs we develop are tested on white-box models by default, *e.g.* the test-time model and the development-time model being the same. The transferability of adversarial inputs to a black-box model is discussed in Sec. 4.5.

4.3. Adversarial Attack on Vision-only Model

4.3.1. Attacking BEVDet with Single-Frame Input

We choose BEVDet-R50 [7] as the tested model with the single-frame vision-only attack discussed in Sec. 3.1.1 with our collected CARLA simulation data and show the results in Tab. 1. We can see that most detected objects in Fig. 4 are not recognized when a single-frame attack is used in Fig. 5.

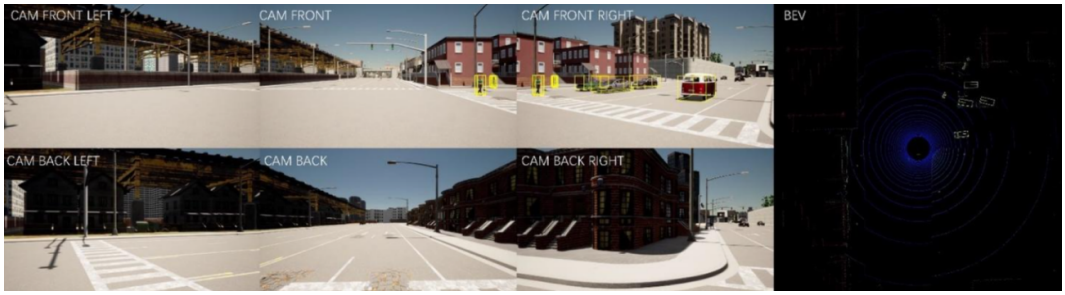

Figure 4. Detection results without attack.

Figure 5. Detection results with vision-only single-frame attack.

Attack Method	*Car*	*Pedestrian*
Single-frame	0.15	0.08
Multi-frame	0.00	0.00

Table 2. mAP of BEVDet4D-R50 w/ single-frame adversarial attack and multi-frame adversarial attack.

Attack Method	*Car*	*Pedestrian*
Multi-frame	0.20	0.13
Multi-frame and LiDAR	0.00	0.00

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

4.3.2. Attacking BEVDet4D with Multi-Frame Input

To extend beyond simple single-frame input for BEV detection, and consider additional temporal information in adversarial attacks, we experiment with the temporally extended version of BEVDet, BEVDet4D [6], which retains the intermediate BEV feature of the previous frame and concatenates it with those generated by the current frame before using the features for predictions.

As shown in Tab. 2, we can observe that the mAP detection performance decreased significantly as the attack achieved a 100% attack success rate.

Our results suggest that models relying solely on visual data, such as BEVDet and BEVDet4D, are quite vulnerable to adversarial disturbances, which can cause major perceptual breakdowns.

4.4. Vision-LiDAR Attack on Vision-LiDAR Model

Since BEVDet does not take LiDAR signals as inputs, we use DAL [8] to evaluate the performance of the LiDAR attack. As shown in Tab. 3, attacking with vision-only attack alone is not enough for a vision-LiDAR BEV model, while attacking from both modalities can achieve a complete success rate.

These experimental results indicate that while multi-modal sensor fusion models like DAL improve perception accuracy, they remain vulnerable to coordinated vision-LiDAR attacks, highlighting the limitations of current multi-sensor security strategies. We can also conclude that, for adversarially attacking a white-box BEV detection model, it is crucial to achieve a complete success rate in attacking via all the modalities and channels.

4.5. Black-Box Attack

So far, we have only been working on BEVDet and its derivative models. However, an intriguing property of adversarial examples that makes them threatening in the real world is their transferability. Transferable attacks assume a realistic scenario in which adversarial examples generated on a known surrogate model can be directly transferred to the unknown target model. Such attacks require no interaction with the target model or any prior knowledge of the target model and thus are more dangerous to safety-critical applications such as autonomous driving.

Therefore, to study the security implications of transferable attacks, we choose BEVFormer [11] as the target model and evaluate its performance on adversarial pertur-

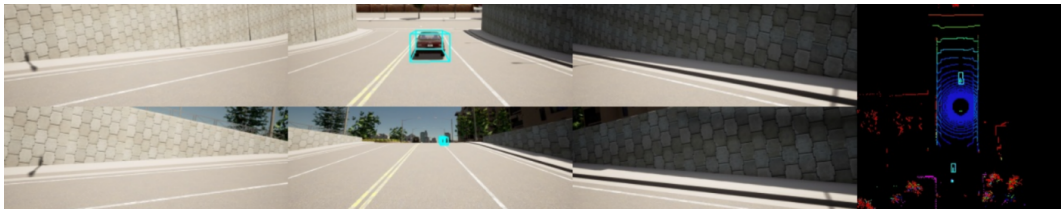

Figure 6. Detection results of BEVFormer without attack.

Figure 7. Detection results of BEVFormer with attack.

Attack	*Car*	*Pedestrian*
	0.24	0.13
✓	0.07	0.03

Table 4. mAP of BEVFormer with and without multi-frame vision-LiDAR adversarial attack.

bations generated based on BEVDet4D. BEVFormer is another popular BEV-based 3D detection model, with two major differences compared to BEVDet and its various variant models (*e.g*. BEVDet4D and DAL). First, the former is a Transformer-based model, while the latter is a CNN-based model. Second, the temporal version of the former maintains and updates a special memory feature which can be seen to be encoded with BEV features from the previous timestamps, while the latter directly uses the previous BEV features. 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.

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.

We also provide a visualization comparison in Fig. 6, showing that BEVFormer does not have difficulty detecting the vehicle forward without attack. However, in Fig. 7, it can be observed that most vehicles and pedestrians can no longer be detected by the model, indicating the effectiveness of our attack method.

4.6. Effects on Planning Module

To investigate how adversarial attacks on perception models affect autonomous driving decisions, we build an agent in CARLA with a perception module (*e.g*. BEVDet) and a planning module for driving actions such as 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).

Figure 8. This is a No Attack Scenario, where no attack is employed and the car can be successfully detected.

4.6.1. 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 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. In addition, we create a simple scene in which two cars are created in a single lane. One car is the main vehicle controlled by the autonomous agent, while another car is running ahead at 8 m/s. A 3D detection model will act as the perception module to detect any obstacles for decision-making. If the perception module is successfully attacked and does not recognize the parked car ahead, the

Figure 9. This is an Attack Scenario, where the adversarial patch and polygon mesh are attached to the car thus the car appears "invisible".

Figure 10. Undetected vehicle causes the planning module to decide to accelerate, resulting in a collision.

car behind will crash into the car ahead, indicating severe safety consequences.

4.6.2. Attacking Framework Implementation

In our implementation, to attach the adversarial patch to the back of the vehicle, the back facet of the front vehicle boundary box is used to determine the 4 corners of the adversarial patch. The facet is resized to half the original size and the coordinates of its 4 corners 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. The box blur is then applied to the open3d output image to suppress the noise in the background before the blank area is removed by cropping. The resulting image is then patched onto the 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 the LiDAR signals from CARLA before being fed to the detector.

4.6.3. Visualized Results

A simple visualization of the third-person view of an autonomous vehicle in CARLA is shown below. In Fig. 8, no attack is employed and the vehicle in front is detected by the ego vehicle. In Fig. 9, the attack is applied to the front vehicle, making it invisible to the rear vehicle. In Fig. 10, as no obstacle is detected, the rear vehicle accelerates and collides with the car in front of it.

5. Conclusion

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 in 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, multi-modality fusion introduce new vulnerabilities therefore 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 if adversaries cannot access the target model.

These results emphasize the urgent need to investigate the robustness of current widely adopted BEV perception systems against adversarial attacks, especially attacks carried out via multiple modalities. Future work should extend adversarial evaluations to more complex driving environments and real-world scenarios to further reveal vulnerabilities in current BEV detection systems.

Acknowledge

This work is 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 (USDOT Grant #69A3552344812). 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, and the U.S. Government assumes no liability for the contents or use thereof.

References

[1] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 2

[2] 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

[3] Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. 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)*, pages 176–194, 2021. 1, 2, 3

[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 *32nd USENIX Security Symposium (USENIX Security 23)*, pages 2993–3010, Anaheim, CA, 2023. USENIX Association. 1

[5] 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*, pages 1907–1915, 2017. 2

[6] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 1, 2, 4, 5

[7] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2, 4

[8] Junjie Huang, Yun Ye, Zhujin Liang, Yi Shan, and Dalong Du. Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection. In *European Conference on Computer Vision*, pages 439–455. Springer, 2024. 1, 2, 4, 5

[9] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robo3d: Towards robust and reliable 3d perception against corruptions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19994–20006, 2023. 1

[10] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI conference on artificial intelligence*, pages 1477–1485, 2023. 1

[11] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2, 5

[12] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *Advances in Neural Information Processing Systems*, 35:10421–10434, 2022. 1, 2

[13] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 2

[14] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. 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*, pages 19–37. Springer, 2020. 2

[15] Vishwanath A Sindagi, Yin Zhou, and Oncel Tuzel. Mvx-net: Multimodal voxelnet for 3d object detection. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 7276–7282. IEEE, 2019. 2

[16] 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 *29th USENIX Security Symposium (USENIX Security 20)*, pages 877–894, 2020. 2

[17] Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4604–4612, 2020. 2

[18] Chunwei Wang, Chao Ma, Ming Zhu, and Xiaokang Yang. Pointaugmenting: Cross-modal augmentation for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11794–11803, 2021.

[19] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21653–21662, 2023.

[20] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5418–5427, 2022. 2

[21] Chaowei Xiao, Ruizhi Deng, Bo Li, Fisher Yu, Mingyan Liu, and Dawn Song. Characterizing adversarial examples based on spatial consistency information for semantic segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 217–234, 2018. 2

[22] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018.

[23] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018. 2

[24] Shaoyuan Xie, Zichao Li, Zeyu Wang, and Cihang Xie. On the adversarial robustness of camera-based 3d object detection. *arXiv preprint arXiv:2301.10766*, 2023. 1

[25] Yang Zhang, Hassan Foroosh, Philip David, and Boqing Gong. CAMOU: Learning physical vehicle camouflages to adversarially attack detectors in the wild. In *International Conference on Learning Representations*, 2019. 2