Understanding the Robustness of 3D Object Detection with Bird's-Eye-View Representations in Autonomous Driving

Zijian Zhu[1*] Yichi Zhang[2,5*] Hai Chen[3] Yinpeng Dong[2✉]
Shu Zhao[3] Wenbo Ding[4] Jiachen Zhong[4] Shibao Zheng[1✉]
[1] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[2] Dept. of Comp. Sci. and Tech., Institute for AI, THBI Lab, BNRist Center, Tsinghua University
[3] Key Laboratory of Intelligent Computing and Signal Processing, Ministry of Education,
School of Computer Science and Technology, Anhui University,
Information Materials and Intelligent Sensing Laboratory of Anhui Province
[4] SAIC Motor AI Lab [5] Zhongguancun Laboratory
{zzj403,sbzh}@sjtu.edu.cn, zyc22@mails.tsinghua.edu.cn, dongyinpeng@tsinghua.edu.cn
{chber_ahu,zhaoshuzs2002}@hotmail.com, {dingwenbo,zhongjiachen}@saicmotor.com

Abstract

3D object detection is an essential perception task in autonomous driving to understand the environments. The Bird's-Eye-View (BEV) representations have significantly improved the performance of 3D detectors with camera inputs on popular benchmarks. However, there still lacks a systematic understanding of the robustness of these vision-dependent BEV models, which is closely related to the safety of autonomous driving systems. In this paper, we evaluate the natural and adversarial robustness of various representative models under extensive settings, to fully understand their behaviors influenced by explicit BEV features compared with those without BEV. In addition to the classic settings, we propose a 3D consistent patch attack by applying adversarial patches in the 3D space to guarantee the spatiotemporal consistency, which is more realistic for the scenario of autonomous driving. With substantial experiments, we draw several findings: 1) BEV models tend to be more stable than previous methods under different natural conditions and common corruptions due to the expressive spatial representations; 2) BEV models are more vulnerable to adversarial noises, mainly caused by the redundant BEV features; 3) Camera-LiDAR fusion models have superior performance under different settings with multi-modal inputs, but BEV fusion model is still vulnerable to adversarial noises of both point cloud and image. These findings alert the safety issue in the applications of BEV detectors and could facilitate the development of more robust models.

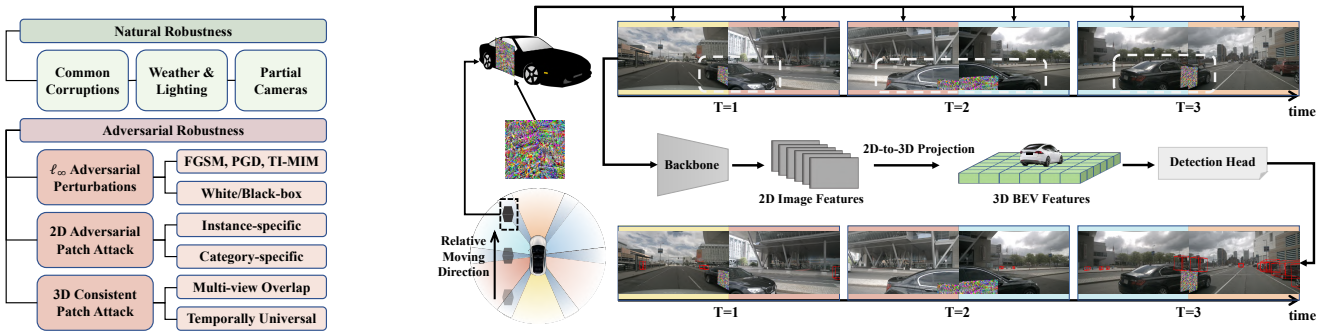
1. Introduction

Autonomous driving systems have great demand for reliable 3D object detection models [21], which aim to predict 3D bounding boxes and categories of road objects, in order to understand the surroundings. To extract holistic representations in the 3D space, the Bird's-Eye-View (BEV) is commonly adopted as a unified representation [30], since it contains both locations and semantic features of objects without being affected by occlusion, and shows promise for various 3D perception tasks in autonomous driving, such as map restoration [41, 44]. Although being broadly used for LiDAR point clouds [28, 65], the BEV representation has recently achieved great success for 3D object detection with multiple cameras, arousing tremendous attention from both industry and academia due to low cost of camera sensors and better exploitation of semantic information in images. These *vision-dependent BEV models*[1] typically project 2D image features to explicit BEV feature maps in the 3D space and make predictions based on BEV features [25, 26, 31, 32, 35]. As representative models, BEVDet [26], BEVDepth [31] and BEVFusion [35] distribute the 2D features into 3D space according to the estimated depth map, while BEVFormer [32] adopts cross attention to query BEV features from 2D images. With expressive spatial semantics of BEV, these models achieve the state-of-the-art results on popular benchmarks (*e.g.*, nuScenes [8]).

Despite the excellent performance, these models are still far from practical deployment due to the robustness issues. Previous works have shown that deep learning models are

[*] Equal Contribution. [✉] Corresponding authors. This work was done when Zijian Zhu and Hai Chen were visiting Tsinghua University.

[1] In this paper, we use the term vision-dependent BEV models to indicate both camera-only and LiDAR-camera fusion BEV models.



(a) Settings for Robustness Evaluation

(b) Pipeline of 3D Consistent Patch Attack

Figure 1. **Overview.** (a) We measure the natural and adversarial robustness of vision-dependent BEV models in 3D object detection under various settings to thoroughly understand the influence of explicit BEV representations on robustness. (b) By pasting an adversarial patch on a car in the 3D space and projecting it to the 2D images, the generated patch is aligned spatially (across adjacent cameras) and temporally (across continuous frames). The patch cloaks the car in all 3 frames from BEVDepth [31], bringing high safety risks to autonomous driving.

vulnerable to adversarial examples [20, 51], common corruptions [22], natural transformations [14, 15], *etc*. The robustness issues rooted in the data-driven deep learning based 3D object detectors can raise severe concerns about the safety and reliability of autonomous driving, making it imperative to evaluate and understand model robustness before being deployed. As vision-dependent BEV models achieve superior performance and become increasingly prevalent in the field, it is of particular importance to compare their robustness to other models that do not rely on BEV representations, given the inherent trade-off between accuracy and robustness [47, 61].

In this paper, we take the first step to systematically analyze and understand the robustness of representative vision-dependent BEV models, by performing thorough experimental evaluations ranging from natural robustness to adversarial robustness as illustrated in Fig. 1(a). We draw several important findings as below:

- We first evaluate the natural robustness under common corruptions, various weather and lighting conditions, and partially missing cameras. We find that camera-based BEV models are generally more robust to natural corruptions of images as a result of the rich spatial information carried by BEV representations.
- We then evaluate the adversarial robustness under the global ℓ_p adversarial perturbations, instance-level and category-level adversarial patches. We observe that BEV models are more vulnerable to adversarial noises, owing to the redundant spatial features represented by BEV based on an in-depth analysis.
- Based on the results, we find that camera-LiDAR fusion models have superior performance under all settings due to the aid of multi-modal inputs. Besides, BEVFusion [35] is less robust when both point cloud and image perturbations are imposed.

In addition to digital adversarial patches, we propose a novel attack method called **3D consistent patch attack**. As shown in Fig. 1(b), adversarial patches are attached to objects for sptiotemporal consistency in the 3D space. We provide two case studies of 3D consistent patch attack. First, we paste patches on objects falling into the overlap regions of multiple cameras, which are observed in different shapes from different viewpoints. Second, we generate temporally universal patches for objects across a continuous sequence of frames in a certain scene, which is a step further from case one. Both spatial alignment and temporal consistency are considered, which distinguishes 3D object detection for autonomous driving cars from the traditional 2D object detection task. The conclusions are consistent with those of adversarial robustness above and can inspire more works to guarantee safe autonomous driving.

2. Related Work

2.1. Vision-Dependent 3D Object Detection

3D object detection [28, 44, 54, 55, 65] is important for autonomous driving cars to understand the surrounding objects and therefore safely navigate through the scenes. Camera-based methods [36, 45, 54, 55] have been widely studied recently because of its low cost for deployment. Camera inputs are also fused with point clouds in camera-LiDAR fusion models [4, 10, 53, 59], to capture both rich semantic information in images and spatial information in point clouds. Learning from map restoration [41, 44] that BEV can serve as an effective representation of the environments for 3D perception tasks, methods with BEV representations are proposed for vision-dependent detectors [25, 26, 31, 32, 35]. They project 2D image features into explicit BEV feature maps in 3D space in order to perform more accurate 3D object detection with expressive spatial information. Specifically, the transformation from 2D to

	Modality	With Transformer	BEV Repr.	#Params	mAP / NDS
FCOS3D [54]	camera-only	No	No	55M	29.8 / 37.7
BEVDet [26]	camera-only	No	Yes	48M	29.2 / 37.2
BEVDepth [31]	camera-only	No	Yes	53M	33.2 / 40.4
DETR3D [55]	camera-only	Yes	No	54M	34.7 / 42.2
BEVFormer [32]	camera-only	Yes	Yes	60M	37.0 / 47.9
TransFusion [4]	camera-LiDAR	Yes	No	37M	67.2 / 70.9
BEVFusion [35]	camera-LiDAR	Yes	Yes	41M	68.5 / 71.4

Table 1. **3D object detection models.** Basic information of different models evaluated in this paper, including input modality, Transformer block, BEV representation, number of parameters and clean detection performance on nuScenes validation set.

3D varies across different methods, including projecting 2D features based on depth estimation [26, 31, 35] and querying 2D features from 3D space with cross attention [32]. Since the perception in auto-driving is performed on a sequence of temporally continuous frames, some methods [32] take the history BEV features of previous frames into account when detecting in the current frame. The holistic spatial representation by BEV boosts the performance of vision-dependent methods to lead the popular benchmarks [8, 18, 50].

2.2. Robustness Evaluation of Object Detection

Robustness has been a crucial issue for deep neural networks (DNN) [11, 20, 22, 51], especially concerning their applications like auto-driving. It has been shown that DNN-based object detectors can be vulnerable to multiple threats including adversarial examples [1, 9, 57, 58, 63], common corruptions [38, 39, 49], *etc*., in either 2D or 3D domains. Specifically for 3D object detection, there has been some works evaluating the robustness of LiDAR-based or fusion models [2, 39, 43, 49, 60]. However, the robustness of camera-based 3D object detectors [62], especially those with BEV representations, has not been fully exploited. This is where our work stands.

Efforts have been devoted to designing effective adversarial patches [24, 33, 52, 56, 64] for attacking object detectors. Many works take use of Expectation of Transformation (EOT) [3, 29] and physically-printable constraints [52, 56] to extend 2D-applied patches to successful physical attack, which is a popular topic for the adversarial patch. These methods are further adopted in attacking 3D object detectors [43, 62], but is not sufficient for scenarios like auto-driving, where the detection is performed along the time and multi-view cameras are used, which contains plentiful spatial and temporal information. Based on this, we propose 3D consistent patch attack in this paper to apply adversarial patches that are aligned in 3D space.

3. Preliminary

Before evaluating the robustness of vision-dependent BEV detectors and non-BEV detectors in the following sections, we first introduce some notations and experimental setups for better readability.

Formally, for 3D object detection, we consider the dataset $\mathcal{D} = \cup_{i=1}^{N} S_i$ containing N scenes, where each scene $S_i = \{(x_t^{(i)}, \hat{y}_t^{(i)}) | t \in \mathcal{T}\}$ is a sequence of observations $x_t^{(i)}$ and ground-truth labels $\hat{y}_t^{(i)}$ at T continuous timestamps $\mathcal{T} = \{0, \cdots, T-1\}$. $x \in [0, 255]^{N_c \times 3 \times H \times W}$ represents N_c views of RGB images for camera-based models and x can also include point clouds data for fusion models. For a 3D object detector, we denote it as $y = f_\theta(x)$ in general, which is trained by an objective function $\mathcal{L}$. In the following context, we will use these notations by default unless otherwise stated.

For experimental evaluation, we use the validation set of nuScenes [8] to study 7 modern 3D object detectors as shown in Tab. 1. There are 4 BEV models and 3 non-BEV counterparts. Among them, FCOS3D [54], BEVDet [26] and BEVDepth [31] are CNN-based, while DETR3D [55], BEVFormer [32], TransFusion [4] and BEVFusion [35] have Transformer blocks in their architectures. The last two are fusion models and have better performance compared to the aforementioned camera-only models. Models for comparison are chosen to have similar numbers of parameters, since larger models tend to have better robustness [6]. We further rule out additional impacts from training techniques like data augmentation. The detailed strategies for the chosen models are listed in Appendix E and there is an insignificant difference within the three comparing subgroups. Though BEVDet and BEVDetpth take in images with lower resolution, but our experiments in Appendix G show that the conclusions are consistent and the influence is nonessential. For evaluation metrics, we adopt mean Average Precision (mAP) and nuScenes Detection Score (NDS) which is a composite indicator of a set of True Positive metrics along with mAP.

4. Natural Robustness

In this section, we first evaluate the natural robustness of vision-dependent BEV models under common corruptions, different weather and lighting conditions, and partial cameras. The natural robustness of a model is a critical issue in real world applications [22, 23], which indicates its reliabil-

	Noise			Blur				Digital				
	Gaussian	Shot	Impulse	Defocus	Glass	Motion	Zoom	Brightness	Contrast	Elastic	Pixel	JPEG
FCOS3D	2.7/11.5	3.0/11.9	3.0/11.9	7.0/18.9	7.0/19.7	7.4/18.3	0.3/1.8	15.1/26.4	11.4/24.8	18.9/30.3	11.3/24.0	11.2/24.4
BEVDet	3.3/10.6	4.1/12.4	2.6/8.5	12.8/24.2	18.0/28.1	14.5/25.7	1.2/4.6	21.1/29.3	10.6/22.1	28.9/37.1	28.2/36.3	17.3/27.5
BEVDepth	4.7/11.4	6.1/15.4	4.6/11.2	**20.1**/30.8	**22.8**/32.2	19.8/30.7	1.8/7.9	23.6/32.8	15.8/25.1	32.6/40.1	31.6/39.3	23.8/33.6
DETR3D	15.7/27.8	17.0/28.5	16.2/28.0	20.0/30.3	16.6/28.2	18.8/29.8	2.4/11.8	33.5/41.6	29.1/38.3	32.5/40.4	28.6/37.7	25.8/35.4
BEVFormer	**16.5/32.6**	**18.3/33.9**	**16.8/32.7**	19.8/**35.6**	21.0/**35.7**	**24.2/38.9**	**2.7/14.2**	**35.0/46.7**	**29.2/42.7**	**36.0/47.0**	**33.7/45.9**	**30.2/43.5**
TransFusion	**65.9/70.2**	**66.0/70.3**	**65.8/70.2**	**66.5/70.5**	66.5/**70.5**	66.4/**70.5**	**65.5/70.0**	**67.1/70.8**	**66.3/70.4**	67.2/70.9	67.1/70.8	**66.9/70.7**
BEVFusion	63.4/68.7	63.4/68.7	63.4/68.7	66.1/70.2	**66.6**/70.4	**66.6**/70.4	61.9/67.7	66.8/70.4	66.0/70.0	**68.4/71.4**	**68.3/71.3**	66.7/70.6

Table 2. **Common Corruptions.** mAP/NDS of vision-dependent models under 12 common corruptions at level 3 as in ImageNet-C [22].

	Day	Night	Sunny	Rainy
FCOS3D	30.1/37.9	13.9/22.6	29.5/37.6	30.1/37.6
BEVDet	30.2/38.1	12.0/21.7	29.4/36.5	30.1/43.2
BEVDepth	33.7/41.0	13.2/22.9	33.1/39.4	33.3/44.7
DETR3D	35.0/42.7	16.0/23.0	34.3/41.0	36.1/47.5
BEVFormer	**37.2/48.1**	**20.1/27.3**	**36.6/47.0**	**38.3/50.8**
TransFusion	67.3/71.0	39.8/44.7	67.0/70.5	67.5/71.9
BEVFusion	**68.5/71.5**	**43.9/46.7**	**68.4/71.3**	**69.5/72.2**

Table 3. **Weather & Lighting.** mAP/NDS of vision-dependent models under different weather and lighting conditions.

ity, stability and consistency.

4.1. Common Visual Corruptions

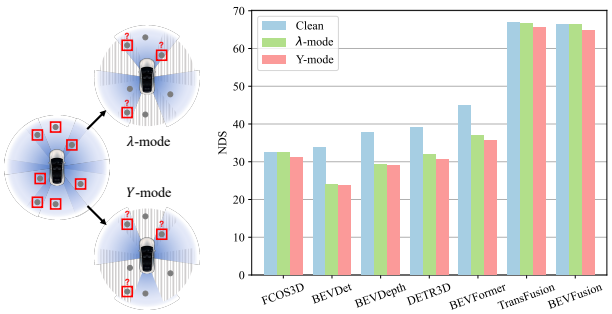
Motivation and Setting. Various forms of vision corruptions like noise, blur and digital distortions could happen in the scenario of autonomous driving due to high speed or malfunctioning sensors [22], presenting potential safety risks. Four main categories of common corruptions are proposed in [22] to evaluate the robustness of image classifiers. Similarly, we analyze the performance of vision-dependent BEV detectors under 12 different corruptions, including noise, blur and digital distortions at level 3.

Results. The model performance in terms of mAP and NDS under various corruptions is shown in Tab. 2. We make the observations below. First, the corruptions imposed on the images barely influence the detection of TransFusion and BEVFusion. The complementary information contained in point clouds makes the fusion models robust under image corruptions. Second, methods with Transformer blocks are more robust to natural corruptions compared to FCOS3D, BEVDet and BEVDetpth, which are CNN-based. This finding is consistent with the conclusions in previous studies that claim the better robustness of Transformers in image classification [6, 40]. Third, besides the negligible difference between fusion models, there are significant advantages of camera-only BEV models compared to non-BEV methods, especially for corruptions of blur and digital distortions. This indicates that BEV models have better natural robustness to image corruptions in general. The holistic modeling of spatial features by BEV representations fusing the multi-view information contributes to this superiority.

4.2. Weather and Lighting Conditions

Motivation and Setting. It is common to encounter different weather and lighting conditions for autonomous driving, which can cause the loss of image quality as well as a shift of data distribution [48], and further lead to worse performance. Following the setting in [35], we analyze the performance of models under different weather and lighting conditions by dividing the dataset into 4 subsets, including Day, Night, Sunny and Rainy, according to the annotations in nuScenes [8].

Results. The results of different models under various weather and lighting conditions are shown in Tab. 3. Generally, though poor lighting condition (Night) brings great difficulties to the detection, BEV models have better performance on 4 subsets, which is consistent to clean performance on the complete dataset.



(a) Two modes (b) Results of NDS

Figure 2. **Partial Cameras.** We consider two modes of partially missing cameras, λ-mode and Y-mode, where 3 nonadjacent cameras are masked out and only objects falling into the multi-view overlap regions are focused, as depicted in (a). The performance with partial cameras in terms of NDS respectively across different models is shown in (b).

4.3. Partial Cameras

Motivation and Setting. The multi-view detection with BEV features enables the global perception of the surrounding environment, but may suffer from incomplete data when some cameras break down. To study model robustness in this case, we consider objects falling into the overlap regions of adjacent cameras [55], *i.e.*, they are captured by multiple cameras, thus the detection of objects is feasible in the absence of one view due to the global representation of BEV. As shown in Fig. 2(a), we mask out 3 non-adjacent cameras out of the 6 multi-view cameras in nuScenes, resulting in 2 settings—λ-mode and Y-mode.

Results. The corresponding NDS is plotted in Fig. 2(b).

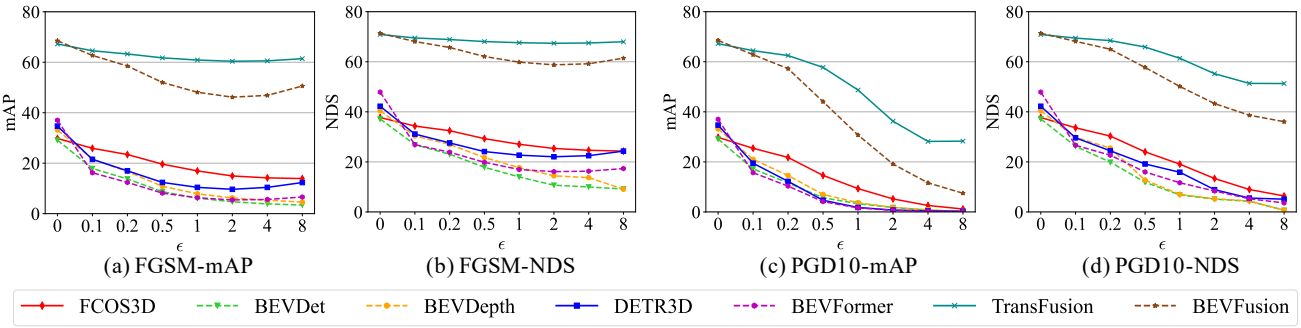


Figure 3. ℓ_∞ **Image Perturbations.** Results of ℓ_∞ adversarial perturbations generated by FGSM [20] and PGD10 [37] with different ϵ ranging from 0 to 8.

With partial cameras, the detection performance decreases to different extents for all models. Fusion models with partial camera inputs and complete LiDAR inputs are slightly affected. FCOS3D, a monocular detector, exhibits a negligible performance drop since it conducts detection on individual images. BEVFormer outperforms DETR3D in terms of two metrics, which indicates that holistic BEV representations from vision input contribute to the spatial modeling of multi-view overlap regions. Meanwhile, the metrics in λ-mode are universally higher than that in Y-mode. This suggests that there is a bias in multi-view detection among these cameras and the detection in the front is better.

5. Adversarial Robustness

Adversarial examples [20, 51] are inputs with carefully crafted noises by an adversary to mislead the models. They help to measure the worst-case performance especially under malicious attacks. In this section, we conduct a series of classic untargeted adversarial attacks to assess the adversarial robustness of BEV methods, by maximizing the training objectives of the detectors with adversarial noises. The formulations of adversarial attacks in this section are introduced in Appendix A as a background guide.

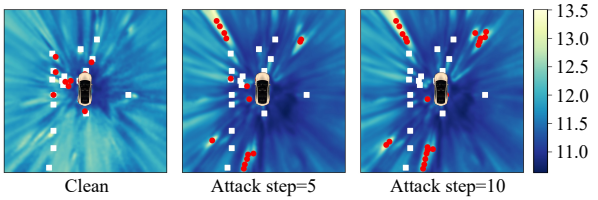
5.1. ℓ_∞ Adversarial Perturbations

Motivation. The pioneering methods [13, 20, 51] first propose to add small, human-imperceptible noises to legitimate examples to induce wrong model outputs. This attack scheme allows us to primarily understand the behaviors of models when encountering small malicious perturbations in the digital space. The scale of the perturbations can be measured by the ℓ_p distance between the adversarial example and the clean input and the ℓ_∞ norm is used here.

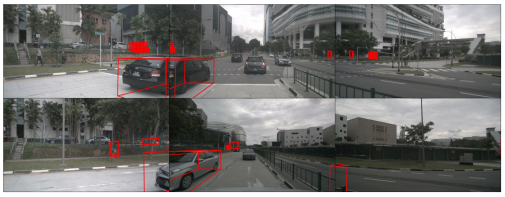
Settings. We adopt two typical gradient-based attack methods—Fast Gradient Sign Method (FGSM) [20] and Projected Gradient Descent (PGD) [37], to generate ℓ_∞ adversarial perturbations for each model. The budget of perturbation ϵ varies from 0 (clean) to 8 under the ℓ_∞ norm. The number of iterations for PGD is set to 10 as PGD10.

Results. We show the curves of mAP and NDS of different models under attacks along with ϵ in Fig. 3. Overall, all models are affected by the adversarial noises to varying degrees and the impact grows with more iterations and larger ϵ. First, the changes in the performance of fusion models are less than camera-only models. For instance, the NDS of BEVFusion at $\epsilon = 8$ under FGSM drops 10.0 compared to the clean performance, while for BEVFormer, it drops 30.6. This implies that fusion models are more robust to adversarial perturbations imposed on images due to the multi-modality inputs including point clouds, which is similar to the conclusion in Sec. 4. Second, though BEV models have better performance on clean images, there is a consistent phenomenon that the impacts on the performance of BEV models are more severe and their resulted mAP and NDS are lower than their non-BEV counterparts. A significant result is that BEVFormer achieves the highest mAP of 37.0 on clean images and declines to one of the lowest mAP of 6.3 with PGD at $\epsilon = 1$ among camera-only methods. This leads to the conclusion that although BEV models have better performance in 3D object detection, they have inferior adversarial robustness. Besides, we notice that the performance of some models rises a little after the ϵ comes 4 and 8 in FGSM attack. We attribute the unexpected rise to the highly-nonlinear loss landscape of 3D detectors. With a large step size, generated perturbations could fail to cross the decision boundary and therefore result in inferior performance.

Analysis. To further pinpoint the underlying reason why BEV models are vulnerable to ℓ_∞ perturbations, we conduct some qualitative analysis by comparing the behaviors between DETR3D [55] and BEVFormer [32], which share the same backbone of ResNet101 and use Transformer blocks to perform detection. We statistically compute the changes on image features extracted by ResNet101 in terms of normalized mean square error (NMSE), when the models are under attack by PGD10 with $\epsilon = 1$. The NMSE of DETR3D on 200 samples is 4.7488 (σ=0.2863) while that of BEVFormer is 4.8042 (σ=0.2675), which indicates that for two different models, the influence on feature extraction from PGD10 is similar. Thus, the difference in adversarial robustness should come from stages after feature extrac-



(a) Visualized BEV feature map with predictions (red dots) and ground truth (white squares) of BEVFormer at different attack steps.



(b) Visualized predictions (red bounding boxes) of BEVFormer under PGD10 attack.

Figure 4. **Qualitative Analysis of BEV model.** Visualization of BEV feature map and predictions under PGD10 attack.

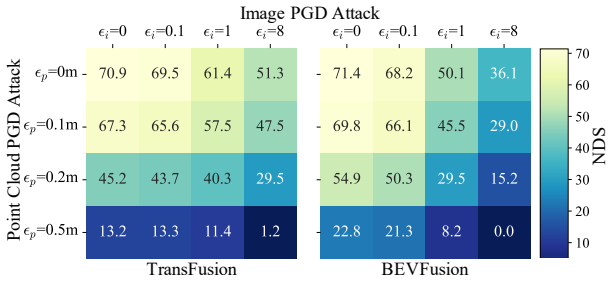


Figure 5. ℓ_∞ **Perturbations for Fusion Models.** NDS of Camera-LiDAR fusion models under different adversarial perturbations added to images and point clouds.

Patch Size	1%	2%	5%	10%
FCOS3D	**18.3/27.2**	**14.6**/23.2	**8.4**/16.2	**3.9**/11.8
BEVDet	10.1/18.8	6.6/13.9	3.0/7.9	1.1/2.9
BEVDepth	13.3/21.7	9.2/18.6	4.3/9.1	1.7/7.3
DETR3D	16.8/ 26.9	12.1/**23.3**	6.0/**16.8**	2.4/**12.5**
BEVFormer	12.7/24.5	9.4/20.2	4.9/15.9	2.1/11.0
TransFusion	**62.5/68.3**	**60.5/67.2**	**55.6/64.7**	**46.5/60.1**
BEVFusion	57.0/64.9	50.9/61.5	39.1/55.0	29.1/49.5

Table 4. **Instance-Specific Adversarial Patches.** mAP/NDS of vision-dependent models with instance-specific adversarial patches of different size ratios.

tion, *i.e.*, BEV modeling for BEVFormer. We analyze the BEV feature maps under attack qualitatively to study why BEV models are vulnerable. One example of visualization is shown in Fig. 4. We find that adversarial perturbations lead to greater activations in areas without objects and further generate numerous false positives. Since BEV representations model the whole 3D space, there are redundant features for the adversarial attack to corrupt, which leads to worse robustness of BEV models.

Point Cloud Perturbations. Although being less affected by image perturbations, fusion models are also exposed to adversarial noises to both images and point clouds. We still adopt PGD10 with a 4×4 design of ℓ_∞ budgets for image and point cloud perturbations, as shown in Fig. 5. We find that there is a trade-off between robustness to image perturbations and point cloud perturbations. BEVFusion is less robust than TransFusion when only image perturbations are imposed, but more robust with only point cloud perturbations. Also, when two kinds of perturbations are both applied, BEVFusion tends to be more vulnerable. The reason is that BEVFusion relies more on image inputs in its pipeline, *i.e.*, it aggregates image and point cloud features to holistic BEV representations to generate object proposals, while TransFusion only uses point cloud features in its early proposal stage. Therefore, larger image perturbations will further decrease the adversarial robustness to point cloud perturbations for BEVFusion. This can also account for the slightly inferior performance of BEVFusion compared to TransFusion under previous image corruptions.

Transfer Attack. Adversarial examples are shown to have cross-model transferability [13, 34, 42], which enables practical black-box attacks. We further study the transferability of adversarial examples between different 3D detection models. We find that the transferability is better across models with similar architectures, while the effect of BEV features is insignificant. Details are listed in Appendix B.

5.2. Instance-specific Adversarial Patches

Motivation. Adversarial examples with ℓ_p perturbations are infeasible for physical attack due to pixel-wise modifications. Patch attack with localized and visible perturbations is therefore studied [7, 33, 46, 52, 57, 63]. The adversarial patch is often a square pattern masked over the victim object for attack in object detection. We first consider the instance-specific adversarial patches, *i.e.*, for each frame, we generate adversarial patches for every single object under each view respectively.

Settings. We set the locations of 2D patches at the 2D positions projected from the centers of 3D bounding boxes, and their sizes are proportional to the sizes of the projected 2D bounding boxes at different ratios, ranging from 1% to 10%. Patch colors for each frame are optimized in 20 steps by Adam [27] optimizer with learning rate of 0.1.

Results. The results of instance-specific patch attack are shown in Tab. 4. With the size ratio increasing, the performance of all models keeps declining. In general, the results show the similar trends with ℓ_∞ adversarial perturbations. Fusion models still demonstrate better robustness to the local adversarial noises. Also, the performance of all BEV methods is worse than non-BEV methods and the gaps are significant especially for camera-only models without Transformer block (BEVDet, BEVDepth *vs.* FCOS3D) and fusion models (BEVFusion *vs.* TransFusion). Since the frame-by-frame attack setting is only distinguished from

Patch Size	1%	2%	5%	10%
FCOS3D	24.9/28.8	22.0/26.7	14.5/22.2	**14.5**/22.2
BEVDet	16.1/27.8	9.3/21.9	3.6/15.4	2.0/8.2
BEVDepth	21.9/31.9	16.3/26.4	8.3/20.4	2.4/8.5
DETR3D	31.3/39.4	26.1/36.8	15.5/29.5	7.3/22.8
BEVFormer	**33.2/45.0**	**30.9/43.4**	**19.8/35.9**	8.1/**24.0**
TransFusion	66.6/**70.6**	66.3/**70.4**	**65.5/70.0**	**65.0/69.7**
BEVFusion	**67.2/70.6**	**66.4**/70.2	64.3/69.2	61.4/67.7

Table 5. **Category-Specific Adversarial Patches.** mAP/NDS of vision-dependent models with category-specific adversarial patches of different size ratios.

ℓ_∞ perturbations with localized noises, it is reasonable that BEV models have the similar conclusion of worse adversarial robustness and the underlying reasons are similar.

5.3. Category-specific Adversarial Patches

Motivation. Furthermore, universal attack is often considered in the context of adversarial patch [7, 12, 33, 66], because it's more feasible to perform physical attack when the objects and models are not certain. We consider category-specific adversarial patches, *i.e.*, for each category of objects in the dataset, we generate a adversarial patch for it.

Settings. The patch location is the same as it in Sec. 5.2. We pre-define a 100×100 patch for each category and resize them to the proper size when applying them to the object. The patch colors are optimized across the dataset for 3 epochs by the Adam optimizer with learning rate of 0.01.

Results. With similar settings but universal attack, we draw a similar finding that fusion models have better robustness and BEV models have worse robustness from the results in Tab. 5. However, an interesting phenomenon occurs that BEVFormer has better performance than DETR3D when applied with category-level adversarial patches. A possible explanation is that BEVFormer predicts based on history BEV features, which neutralize the computation of the expectation of gradients over the dataset and make it harder to conduct the universal attack.

6. 3D Consistent Patch Attack

Adversarial attacks in Sec. 5 are mostly proposed in 2D vision tasks of image classification and object detection. Considering that 3D object detection in autonomous driving involves multi-view cameras and continuous frames, we propose a **3D consistent patch attack**, which applies adversarial patches on objects in the real 3D space. This attack considers the 3D consistency and alignment, which is more appropriate for autonomous driving.

6.1. Method

Rather than patch attack in digital space as in Sec. 5, where the process of applying a patch is a pixel-wise multiplication between masking matrix and 2D images, 3D consistent patch attack requires an applying function $\mathcal{A}$ to project the adversarial patches in the 3D space to 2D perspective views of cameras.

We then introduce the process of $\mathcal{A}$ formally. For an object in a single frame, we paste the corresponding adversarial patch, whose size is $H_p \times W_p$ in the real world, to an appointed 3D position of the object. Given the information of ground-truth 3D bounding box, the 3D positions of the corners of the pasted patch can be computed, which are located in the LiDAR coordinate system and represented by (p_1, p_2, p_3, p_4), where $p_i \in \mathbb{R}^3$. For each camera view, there is a projection matrix $\mathcal{M}_{3d-2d} \in \mathbb{R}^{4\times4}$ according to the camera intrinsic parameters and a 3D point $p = (x_p, y_p, z_p)^\top$ in LiDAR coordinates can be projected to point $p' \in \mathbb{R}^2$ on 2D image plane following

$$\begin{bmatrix} x_c \\ y_c \\ z_c \\ 1 \end{bmatrix} = \mathcal{M}_{3d-2d} \begin{bmatrix} x_p \\ y_p \\ z_p \\ 1 \end{bmatrix}, \; p' = (x_c/z_c, y_c/z_c)^\top. \quad (1)$$

Then, we digitally apply the patch to the quadrangle defined by the projected corners (p'_1, p'_2, p'_3, p'_4) with perspective transformation. A vector of projection coefficients $[a, b, c, d, e, f, g, h]^\top \in \mathbb{R}^8$ is used and the pixel (h_t, w_t) inside the target quadrangle are projected back to the location (h_s, w_s) on the source $H_p \times W_p$ patch following

$$\begin{bmatrix} h_s \\ w_s \end{bmatrix} = \begin{bmatrix} a \cdot h_t + b \cdot w_t + c/g \cdot h_t + h \cdot w_t + 1 \\ d \cdot h_t + e \cdot w_t + f/g \cdot h_t + h \cdot w_t + 1 \end{bmatrix}, \quad (2)$$

where the coefficients can be solved with the corresponding corner coordinates. Then, the pixel color at (h_t, w_t) on the 2D image can be interpolated across the neighbor pixels around (h_s, w_s) on the patch. The process is differentiable and the original adversarial patches can be optimized in a similar way as 2D patches in Sec. 5.2 and Sec. 5.3.

In the following, we consider two cases that can apply the proposed 3D consistent patches and line with the distinguished properties of 3D object detection in auto-driving.

6.2. Multi-view Patch Attack

Motivation and Formulation. Multi-view detection is a unique property of autonomous driving. An adversarial patch in 3D space can be captured from different angles and thus deforms to different shapes on 2D images. In this case, 3D patches are attached to the objects in multi-view overlap regions. The patch has different ratios of size similar to those in Sec. 5.2, and its plane is vertical to the direction pointing the car. The multi-view patch attack is conducted frame by frame and formalized as

$$\max_{\mathbf{p}=\{p_1,\cdots,p_{N_{ol}}\}} \mathcal{L}(f_\theta(\mathcal{A}(x, \mathbf{p})), \hat{y}), \quad (3)$$

where $\mathbf{p}$ is the set of 3D adversarial patches for the N_{ol} objects in the overlap regions in one specific frame, which are applied to the model inputs with $\mathcal{A}$.

Case	Multi-view Patch			Temp. Univ. Patch		
Patch Size	0%	5%	10%	0%	5%	10%
FCOS3D	32.5	**22.2**	**17.5**	37.7	20.6	15.3
BEVDet	33.8	11.1	6.2	37.2	12.7	5.9
BEVDepth	37.7	11.8	8.8	40.4	17.7	8.7
DETR3D	39.1	15.2	10.3	42.2	28.3	23.2
BEVFormer	45.0	14.8	9.6	47.9	**35.0**	**29.0**
TransFusion	**66.8**	**63.7**	**62.8**	70.9	**68.0**	**66.4**
BEVFusion	66.4	54.2	50.8	**71.4**	63.9	60.7

Table 6. **3D Consistent Patch Attack.** NDS of vision-dependent models with 3D consistent patches in the cases of Multi-view Patch Attack and Temporally Universal Patch Attack respectively. 0% denotes clean images.

Results. The performance is shown in Tab. 6, which is only evaluated on objects in overlap regions. Fusion models are still superior to camera-only methods either in clean detection or with 3D consistent patches. BEV models demonstrate better or closed clean detection of overlap regions, but the performance is worse than non-BEV models under attack. Though the form of adversarial noises has changed to spatially aligned 3D consistent patches, the attack is still white-box and performed frame by frame. Therefore, the conclusions of adversarial robustness for both fusion models and BEV models in Sec. 5.1 and Sec. 5.2 still hold.

6.3. Temporally Universal Patch Attack

Motivation and Formulation. The frame-to-frame temporal continuity is another distinguished property in the auto-driving system. Most attack settings considered in previous sections are for individual frame and the temporal consistency is not taken into account. We propose a setting where adversarial patches are applied to the surrounding surfaces of every object appeared in a sequence of frames and these patches are spatial-temporally consistent. For a sequence of observations in the i-th scene S_i, the attack is formalized as

$$\max_{\mathbf{p}^{(i)}=\{p_1^{(i)},\cdots,p_{N_i}^{(i)}\}} \mathbb{E}_{t\sim\mathcal{T}}[\mathcal{L}(f_\theta(\mathcal{A}(x_t^{(i)},\mathbf{p}^{(i)})),\hat{y}_t^{(i)})], \quad (4)$$

where $\mathbf{p}^{(i)}$ is the set of 3D adversarial patches for the N_i objects appeared in scene S_i and remains consistent across the timeline to perform the temporally universal attack.

Results. We show the results in Tab. 6. Despite the impressive performance of fusion models and better adversarial robustness of FCOS3D and TransFusion, we find that BEVFormer outperforms DETR3D by at least 2.2 in mAP and 5.8 in NDS, demonstrating better robustness in this case. This is similar to the category-specific patch attack. Both settings require universal noises across instances in a category or frames in a sequence, and therefore the better robustness of BEVFormer should also come from the temporal dependency of history BEV features.

7. Discussion and Conclusion

In this paper, we delve into the robustness of vision-dependent BEV models in 3D object detection [26, 31, 32, 35]. We have studied the natural and adversarial robustness of 7 popular models under different settings, and propose a novel method of 3D consistent patch attack specifically designed for the scenario of autonomous driving. Various phenomena are observed and many conclusions are drawn from them, some of which confirm the previous observations while others are inspiring for the design of BEV models in 3D object detection:

- The holistic representations of BEV with rich spatial and semantic features improve the detection accuracy as well as the natural robustness of 3D detectors to common corruptions of images and other forms of disturbance like partially missing cameras.
- Under intentional adversarial attacks, BEV models tend to be more vulnerable to the adversarial noises on images (*e.g.*, ℓ_∞ perturbations and adversarial patches), which results from the extra activation on the BEV feature maps as analyzed on BEVFormer.
- Fusion models generally have better performance and robustness than camera-only methods due to the multi-modality, and BEVFusion is more vulnerable to adversarial attacks when perturbations on both images and point clouds are applied because of its greater dependence on vision inputs.
- Consistent with [5, 6] in image classification, Transformer improves the performance and natural robustness of 3D detectors, while they are not more adversarially robust than pure CNN models.
- As shown by results of BEVFormer in category-specific patch attack and temporally universal patch attack, temporal association in feature space can lead to better robustness to universal adversarial noises.

As BEV representations begin to be deployed in real-world autonomous driving systems, we explore their robustness and ring the alarm of the safety issue in their usage. To further enhance the robustness of BEV models while preserving their accuracy in future works, temporal information could be adopted as in BEVFormer [29] and the redundant spatial features in the explicit BEV representations can be reinforced with dense supervisory signals.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2020AAA0104304), NSFC Projects (62076147, 62071292, 61771303, 62276149, 61876001,U19B2034, U1811461, U19A2081), STCSM Project (No. 18DZ2270700), and the General Project of the Natural Science Foundation of Anhui Province (No. 1808085MF175), Tsinghua-Alibaba Joint Research Program, Tsinghua-OPPO Joint Research Center for Future Terminal Technology. Y. Dong was also supported by the China National Postdoctoral Program for Innovative Talents and Shuimu Tsinghua Scholar Program.

References

[1] Mazen Abdelfattah, Kaiwen Yuan, Z Jane Wang, and Rabab Ward. Towards universal physical attacks on cascaded camera-lidar 3d object detection models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 3592–3596. IEEE, 2021. 3

[2] Fatima Albreiki, Sultan Abughazal, Jean Lahoud, Rao Anwer, Hisham Cholakkal, and Fahad Khan. On the robustness of 3d object detectors. *arXiv preprint arXiv:2207.10205*, 2022. 3

[3] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 3

[4] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1090–1099, 2022. 2, 3, 12, 14

[5] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 34:26831–26843, 2021. 8

[6] Srinadh Bhojanapalli, Ayan Chakrabarti, Daniel Glasner, Daliang Li, Thomas Unterthiner, and Andreas Veit. Understanding robustness of transformers for image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10231–10241, 2021. 3, 4, 8

[7] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 6, 7

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 1, 3, 4

[9] Chang Chen and Teng Huang. Camdar-adv: generating adversarial patches on 3d object. *International Journal of Intelligent Systems*, 36(3):1441–1453, 2021. 3

[10] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. *arXiv preprint arXiv:2203.10642*, 2022. 2

[11] Yuefeng Chen, Xiaofeng Mao, Yuan He, Hui Xue, Chao Li, Yinpeng Dong, Qi-An Fu, Xiao Yang, Wenzhao Xiang, Tianyu Pang, et al. Unrestricted adversarial attacks on imagenet competition. *arXiv preprint arXiv:2110.09903*, 2021. 3

[12] Edward Chou, Florian Tramer, and Giancarlo Pellegrino. Sentinet: Detecting localized universal attacks against deep learning systems. In *2020 IEEE Security and Privacy Workshops (SPW)*, pages 48–54. IEEE, 2020. 7

[13] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 5, 6

[14] Yinpeng Dong, Shouwei Ruan, Hang Su, Caixin Kang, Xingxing Wei, and Jun Zhu. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. In *Neural Information Processing Systems (NeurIPS)*, 2022. 2

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. Exploring the landscape of spatial robustness. In *International conference on machine learning*, pages 1802–1811. PMLR, 2019. 2

[16] C. Michaelis et al. Benchmarking robustness in object detection: Autonomous driving when winter is coming. 2019. 13

[17] Y. Dong et al. Benchmarking adversarial robustness on image classification. In *CVPR*, 2020. 13

[18] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013. 3

[19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3354–3361, 2012. 14

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 3, 5

[21] Sorin Grigorescu, Bogdan Trasnea, Tiberiu Cocias, and Gigel Macesanu. A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, 37(3):362–386, 2020. 1

[22] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2019. 2, 3, 4

[23] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021. 3

[24] Hao Huang, Yongtao Wang, Zhaoyu Chen, Zhi Tang, Wenqiang Zhang, and Kai-Kuang Ma. Rpattack: Refined patch attack on general object detectors. In *2021 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2021. 3

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

[26] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 2, 3, 8, 14

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

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

[29] Mark Lee and Zico Kolter. On physical adversarial patches for object detection. *arXiv preprint arXiv:1906.11897*, 2019. 3

[30] Hongyang Li, Chonghao Sima, Jifeng Dai, Wenhai Wang, Lewei Lu, Huijie Wang, Enze Xie, Zhiqi Li, Hanming Deng, Hao Tian, et al. Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe. *arXiv preprint arXiv:2209.05324*, 2022. 1

[31] 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. *arXiv preprint arXiv:2206.10092*, 2022. 1, 2, 3, 8, 14

[32] 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 multi-camera images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 1, 2, 3, 5, 8, 14

[33] Xin Liu, Huanrui Yang, Ziwei Liu, Linghao Song, Hai Li, and Yiran Chen. Dpatch: An adversarial patch attack on object detectors. *arXiv preprint arXiv:1806.02299*, 2018. 3, 6, 7

[34] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016. 6

[35] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 1, 2, 3, 4, 8, 12, 14

[36] Zechen Liu, Zizhang Wu, and Roland Tóth. Smoke: Single-stage monocular 3d object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 996–997, 2020. 2

[37] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 5

[38] Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking robustness in object detection: Autonomous driving when winter is coming. *arXiv preprint arXiv:1907.07484*, 2019. 3

[39] Muhammad Jehanzeb Mirza, Cornelius Buerkle, Julio Jarquin, Michael Opitz, Fabian Oboril, Kay-Ulrich Scholl, and Horst Bischof. Robustness of object detectors in degrading weather conditions. In *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*, pages 2719–2724. IEEE, 2021. 3

[40] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021. 4

[41] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 1, 2

[42] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016. 6

[43] Won Park, Nan Liu, Qi Alfred Chen, and Z Morley Mao. Sensor adversarial traits: Analyzing robustness of 3d object detection sensor fusion models. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 484–488. IEEE, 2021. 3

[44] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *European Conference on Computer Vision*, pages 194–210. Springer, 2020. 1, 2

[45] Danila Rukhovich, Anna Vorontsova, and Anton Konushin. Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2397–2406, 2022. 2

[46] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th USENIX workshop on offensive technologies (WOOT 18)*, 2018. 6

[47] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018. 2

[48] Adarsh Subbaswamy, Roy Adams, and Suchi Saria. Evaluating model robustness and stability to dataset shift. In *International Conference on Artificial Intelligence and Statistics*, pages 2611–2619. PMLR, 2021. 4

[49] Jiachen Sun, Qingzhao Zhang, Bhavya Kailkhura, Zhiding Yu, Chaowei Xiao, and Z Morley Mao. Benchmarking robustness of 3d point cloud recognition against common corruptions. *arXiv preprint arXiv:2201.12296*, 2022. 3

[50] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2446–2454, 2020. 3

[51] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 2, 3, 5

[52] Simen Thys, Wiebe Van Ranst, and Toon Goedemé. Fooling automated surveillance cameras: adversarial patches to attack person detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 0–0, 2019. 3, 6

[53] 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

[54] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 913–922, 2021. 2, 3, 14

[55] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191, 2022. 2, 3, 4, 5, 14

[56] Yajie Wang, Haoran Lv, Xiaohui Kuang, Gang Zhao, Yu-an Tan, Quanxin Zhang, and Jingjing Hu. Towards a physical-world adversarial patch for blinding object detection models. *Information Sciences*, 556:459–471, 2021. 3

[57] Shudeng Wu, Tao Dai, and Shu-Tao Xia. Dpattack: Diffused patch attacks against universal object detection. *arXiv preprint arXiv:2010.11679*, 2020. 3, 6

[58] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1369–1378, 2017. 3

[59] Tianwei Yin, Xingyi Zhou, and Philipp Krähenbühl. Multi-modal virtual point 3d detection. *Advances in Neural Information Processing Systems*, 34:16494–16507, 2021. 2

[60] Kaicheng Yu, Tang Tao, Hongwei Xie, Zhiwei Lin, Zhongwei Wu, Zhongyu Xia, Tingting Liang, Haiyang Sun, Jiong Deng, Dayang Hao, et al. Benchmarking the robustness of lidar-camera fusion for 3d object detection. *arXiv preprint arXiv:2205.14951*, 2022. 3

[61] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 2

[62] Jindi Zhang, Yang Lou, Jianping Wang, Kui Wu, Kejie Lu, and Xiaohua Jia. Evaluating adversarial attacks on driving safety in vision-based autonomous vehicles. *IEEE Internet of Things Journal*, 9(5):3443–3456, 2021. 3

[63] Yichi Zhang, Zijian Zhu, Xiao Yang, and Jun Zhu. Adversarial semantic contour for object detection. *arXiv preprint arXiv:2109.15009*, 2021. 3, 6

[64] Yusheng Zhao, Huanqian Yan, and Xingxing Wei. Object hider: Adversarial patch attack against object detectors. *arXiv preprint arXiv:2010.14974*, 2020. 3

[65] 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. 1, 2

[66] Alon Zolfi, Moshe Kravchik, Yuval Elovici, and Asaf Shabtai. The translucent patch: A physical and universal attack on object detectors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15232–15241, 2021. 7