Benchmarking Robustness of 3D Object Detection to Common Corruptions in Autonomous Driving

Yinpeng Dong[1,5], Caixin Kang[2], Jinlai Zhang[3], Zijian Zhu[4], Yikai Wang[1], Xiao Yang[1], Hang Su[1,6*], Xingxing Wei[2], Jun Zhu[1,5,6*]
[1] Dept. of Comp. Sci. and Tech., Institute for AI, Tsinghua-Bosch Joint ML Center, Tsinghua-China Mobile Communications Group Co., Ltd. Joint Institute, Tsinghua University
[2] Institute of Artificial Intelligence, Beihang University [3] Guangxi University
[4] Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University
[5] RealAI [6] Peng Cheng Laboratory, Pazhou Laboratory (Huangpu), Guangzhou, China
{dongyinpeng,suhangss,dcszj}@tsinghua.edu.cn, caixinkang@buaa.edu.cn, 1711302013@st.gxu.edu.cn

Abstract

3D object detection is an important task in autonomous driving to perceive the surroundings. Despite the excellent performance, the existing 3D detectors lack the robustness to real-world corruptions caused by adverse weathers, sensor noises, etc., provoking concerns about the safety and reliability of autonomous driving systems. To comprehensively and rigorously benchmark the corruption robustness of 3D detectors, in this paper we design 27 types of common corruptions for both LiDAR and camera inputs considering real-world driving scenarios. By synthesizing these corruptions on public datasets, we establish three corruption robustness benchmarks—KITTI-C, nuScenes-C, and Waymo-C. Then, we conduct large-scale experiments on 24 diverse 3D object detection models to evaluate their corruption robustness. Based on the evaluation results, we draw several important findings, including: 1) motion-level corruptions are the most threatening ones that lead to significant performance drop of all models; 2) LiDAR-camera fusion models demonstrate better robustness; 3) camera-only models are extremely vulnerable to image corruptions, showing the indispensability of LiDAR point clouds. We release the benchmarks and codes at `https://github.com/thu-ml/3D_Corruptions_AD` to be helpful for future studies.

1. Introduction

As a fundamental task in autonomous driving, 3D object detection aims to identify objects of interest (*e.g.*, vehicles, pedestrians, or cyclists) in the surrounding environment by predicting their categories and the corresponding 3D bounding boxes. LiDAR and camera are two important types of sensors for 3D object detection, where the former provides the depth information of road objects as sparse point clouds, while the latter captures abundant semantic information of the scene as color images. Based on the complementary nature of the two modalities, 3D object detection models can be categorized into LiDAR-only [29,47,48,60,69], camera-only [39,56–58], and LiDAR-camera fusion [11,28,34,53] models. Since autonomous driving is safety-critical, it is of paramount importance to assess the robustness of 3D object detectors under diverse circumstances before deployed.

Although the recent progress of 3D object detection has led to significant improvements in typical benchmarks (*e.g.*, KITTI [17], nuScenes [6], and Waymo [51]), the existing models based on data-driven deep learning approaches often generalize poorly to the corrupted data caused by, *e.g.*, adverse weathers [21, 22, 27], sensor noises [7, 25, 44], and uncommon objects [9, 31], posing a formidable obstacle to safe and reliable autonomous driving [1]. To perform robustness evaluation, recent works construct new datasets of road anomalies [9,23,31,40] or under extreme weather conditions [4, 15, 41]. Nevertheless, they are usually of small sizes due to the high data collection costs and the rareness of corner cases or adverse weathers. Other works synthesize common corruptions on clean datasets to benchmark robustness on image classification [25] and point cloud recognition [44, 50], but they only consider several simple corruptions, which could be insufficient and unrealistic for 3D object detection. Therefore, it remains challenging to comprehensively characterize different corruptions considering diverse driving scenarios and fairly evaluate corruption robustness of existing models within a unified framework.

In this paper, we systematically design **27** types of common corruptions in 3D object detection for both LiDAR and camera sensors to comprehensively and rigorously evaluate the corruption robustness of current 3D object detectors.

*Corresponding authors.

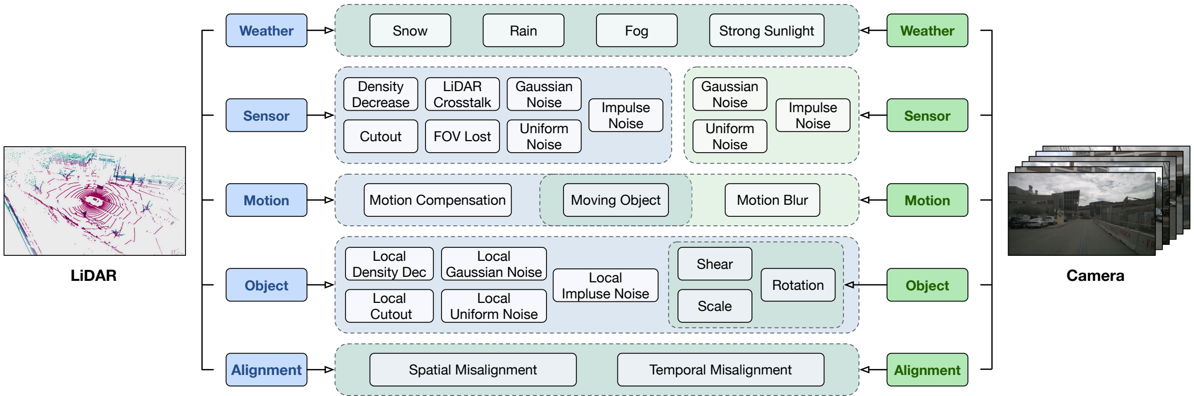

Figure 1. An overview of 27 corruptions for 3D object detection, which are categorized into weather, sensor, motion, object, and alignment levels. As shown, some corruptions are effective for one modality, while the others are applied to both (*e.g.*, *Snow*, *Moving Object*, *Shear*).

The corruptions are grouped into *weather*, *sensor*, *motion*, *object*, and *alignment* levels, covering the majority of real-world corruption cases, as demonstrated in Fig. 1. Most of them are specifically designed for autonomous driving (*e.g.*, motion-level ones), which have not been explored before. Following [25], every corruption has five severities, leading to a total number of **135** distinct corruptions. By applying them to typical autonomous driving datasets—KITTI [17], nuScenes [6], and Waymo [51], we establish three corruption robustness benchmarks—**KITTI-C**, **nuScenes-C**, and **Waymo-C**. We hope that they can serve as general datasets for comprehensively benchmarking corruption robustness of 3D object detectors and facilitating future research.

We conduct large-scale experiments to compare the corruption robustness of existing 3D object detection models. Specifically, we evaluate 11 models on KITTI-C, 10 models on nuScenes-C, and 3 models on Waymo-C. The models are of great variety with different input modalities, representation methods, and detection heads. Based on the evaluation results, we find that: 1) the corruption robustness of 3D object detectors is highly correlated with their clean accuracy; 2) motion-level corruptions impair the model performance most, while being rarely explored before; 3) LiDAR-camera fusion models are more resistant to corruptions, but there is a trade-off between robustness under image corruptions and point cloud corruptions of fusion models. More discussions are provided in Sec. 6. Moreover, we study data augmentation strategies [14, 64, 67] as potential solutions to improve corruption robustness, but find that they provide a little robustness gain, leaving robustness enhancement of 3D object detection an open problem for future research.

2. Related Work

2.1. 3D Object Detection

Based on the input modality, we categorize 3D object detection models into LiDAR-only, camera-only, and LiDAR-camera fusion models.

LiDAR-only models: LiDAR point clouds are sparse, irregular, and unordered by nature. To learn useful representations, *voxel-based* methods project point clouds to compact grids. Typically, VoxelNet [69] rasterizes point clouds into voxels, which are processed by PointNets [43] and 3D CNNs. To speed up, SECOND [60] introduces sparse 3D convolutions and PointPillars [29] elongates voxels into pillars. Other works exploit information of object parts [49] or shape [70] to improve the performance. On the other hand, *point-based* methods take raw point clouds as inputs and make predictions on each point. PointRCNN [48] proposes a two-stage framework that first generates 3D proposals and then refines the proposals in the canonical coordinates. 3DSSD [61] is a lightweight one-stage detector with a fusion sampling strategy. To have the best of both worlds, *point-voxel-based* methods are then explored. PV-RCNN [47] integrates 3D voxel CNN and PointNet-based set abstraction to efficiently create high-quality proposals.

Camera-only models: 3D object detection based on images is challenging due to the lack of depth information, but attracts extensive attention considering the advantage of low cost. The most straightforward approach is to take *monocular* detection methods [10, 36, 39, 56, 57] and apply post-processing across cameras. For example, Mono3D [10] generates 3D object proposals scored by semantic features. SMOKE [36] combines a single keypoint estimation with regressed 3D variables. To address the limitation of post-processing in monocular methods, *multi-view* methods fuse information from all cameras in the intermediate layers. DETR3D [58] adopts a transformer-based detector [8] that fetches the image features by projecting object queries onto images. BEVFormer [33] exploits spatial-temporal information from multi-view images based on BEV queries.

LiDAR-camera fusion models: To leverage the complementary information from LiDAR and camera inputs, fusion methods are also extensively studied. Following [35], we classify the newly developed methods into *point-level*, *proposal-level*, and *unified representation* fusion methods. Point-level methods augment LiDAR point clouds with semantic image features and then apply existing LiDAR-only models for 3D detection, including PointPainting [53], EP-Net [26], PointAugmenting [54], Focals Conv [13], *etc*. Proposal-level fusion methods [11, 42] generate 3D object proposals and integrate image features into these proposals. FUTR3D [12] and TransFusion [2] employ a query-based transformer decoder, which fuses image features with object queries. Moreover, BEVFusion [35] unifies the image feature and point cloud feature in a BEV representation space, which stands out as a new fusion strategy.

2.2. Robustness Benchmarks

It is well-known that deep learning models lack the robustness to adversarial examples [20, 52], common corruptions [25], and other kinds of distribution shifts [18, 19, 24]. In autonomous driving, many works collect new datasets to evaluate model robustness under different conditions. For example, the Seeing Through Fog (STF) [4], Canadian Adverse Driving Conditions (CADC) [41], and Ithaca365 [15] datasets are collected in adverse weathers; and others gather road anomalies of 2D images [9, 23, 31, 40]. Despite the efforts, these datasets only cover limited scenarios due to the high collection costs of rare data. Moreover, as mainly used for evaluation, these datasets have a big domain gap from the large-scale training datasets since they were collected in different cities with varying vehicles and sensors, making it hard for us to examine the effects of different factors (*e.g*., weather *vs*. city) on model robustness.

One promising direction is to synthesize real-world corruptions on clean datasets to benchmark model robustness. For example, ImageNet-C [25] is first introduced in image classification with 15 corruption types, ranging from noise, blur, weather to digital corruptions. The similar methodology is further applied to 2D object detection [38] and point cloud recognition [44, 50]. However, many of these studied corruptions are hypothetical and thus unrealistic in the scenario of autonomous driving. It is still challenging to build a comprehensive benchmark for robustness evaluation of 3D object detection considering diverse real-world driving cases. We notice that two concurrent works [32, 63] to ours also study robustness of 3D object detection in autonomous driving. However, they mainly consider specific kinds of 3D detection models (*i.e*., LiDAR-only models in [32] and fusion models in [63]) and include limited types of corruptions with less evaluations, as compared in Appendix A.2.

3. Corruptions in 3D Object Detection

Real-world corruptions arise from diverse scenarios in autonomous driving, based on which we systematically categorize the corruptions into *weather*, *sensor*, *motion*, *object*, and *alignment* levels. We identify common corruption types for each level considering real-world driving scenarios, resulting in **27** distinct corruptions in total, as shown in Fig. 1. Among them, some corruptions are applied to both modalities simultaneously, such as weather-level ones, while the others are designed for a single modality, such as sensor-level ones. We visualize a subset of corruptions in Fig. 2.

Weather-level corruptions: Weather change is usually encountered in autonomous driving, which can dramatically disrupt both LiDAR and camera inputs. For example, *fog* reduces the visibility of objects in images and causes scattered points due to attenuation and backscattering [4, 22, 65]. Consequently, 3D detectors trained on data collected in normal weather tend to perform poorly under adverse weathers [4]. To study the robustness under weather changes, we consider 4 weather-level corruptions: *Snow*, *Rain*, *Fog*, and *Strong Sunlight*, as they are more common [4, 15, 41]. For LiDAR, we adopt physically based methods [21, 22, 27] to simulate the effects of rain, snow, and fog on point clouds from normal weather. We simulate the effect of strong sunlight by applying strong Gaussian noises to points along the sun direction [7]. For camera, we apply image augmentations [25] to simulate visually realistic weathers.

Sensor-level corruptions: The sensors, when affected by numerous internal or external factors (*e.g*., sensor vibration [46], lighting conditions [25, 33] and reflective materials), can induce various kinds of corruptions to the captured data. Based on prior discussions on sensor noises [3, 7, 25, 44], we design 10 practical sensor-level corruptions—7 for point clouds and 3 for images. The point cloud corruptions are: *Density Decrease*, *Cutout*, *LiDAR Crosstalk*, *FOV Lost*, *Gaussian Noise*, *Uniform Noise*, and *Impulse Noise*. Density decrease simulates missing points commonly observed in typical datasets [17]. Cutout occurs when laser pulses have no echo in a local region (*e.g*., puddle) and is simulated by dropping points in a randomly selected area. LiDAR crosstalk [5] happens when multiple LiDARs operate at close range, which is simulated by applying strong Gaussian noises to a small subset of points. FOV lost simulates a limited field-of-view of LiDAR caused by occlusion. Moreover, due to the ranging inaccuracy of LiDAR, we consider 3 noise corruptions that apply Gaussian, uniform, and impulse noises to point coordinates, respectively. The 3 image corruptions include: *Gaussian Noise*, *Uniform Noise*, and *Impulse Noise* to simulate the visual noise patterns due to low-lighting conditions or defects of camera [25]. Although we design sensor-level corruptions for LiDAR and camera separately, they can occur for both sensors at the same time, affecting LiDAR-camera fusion models further.

Motion-level corruptions: An autonomous vehicle will encounter several types of corruptions during driving. In this paper, we introduce 3 motion-level corruptions: *Motion Compensation*, *Moving Object*, and *Motion Blur*, which are practical in the real world and studied for the first time. Vehicle ego-motion induces distortions to point clouds since the points in a frame are not obtained in the same coordi-

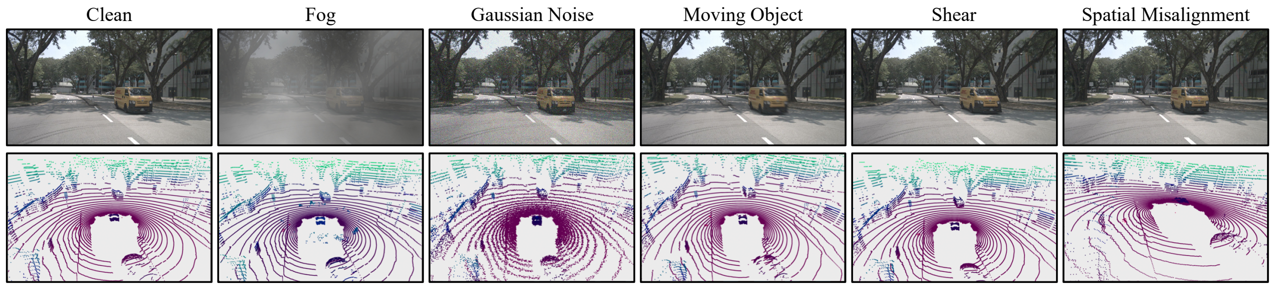

Figure 2. Visualization of typical corruption types of each level in our benchmark (best viewed when zoomed in). Full visualization results of all corruptions are shown in Appendix A.3.

nate system [68]. To obtain accurate point clouds, motion compensation is typically used with the aid of the localization information [6, 17]. However, this process can introduce noises, which we call motion compensation corruption, simulated by adding small Gaussian noises to the rotation and translation matrices of the vehicle's ego pose. The moving object corruption denotes the case that an object is moving rapidly in the scene. It can cause shifting points within the object's 3D bounding box [59] and blur the image patch of the object. The last corruption is motion blur on camera images, which is caused by driving too fast.

Object-level corruptions: Objects in the real world always come in a variety of shapes and materials [9,31], making it challenging to correctly recognize them. The viewing direction can also lead to wrong recognition of objects [16]. Based on this, we introduce 8 object-level corruptions: *Local Density Decrease*, *Local Cutout*, *Local Gaussian Noise*, *Local Uniform Noise*, *Local Impluse Noise*, *Shear*, *Scale*, and *Rotation*. The first five corruptions are only applied to LiDAR point clouds to simulate the distortions caused by different object materials or occlusion. As their names indicate, these corruptions only make changes to local sets of points within the objects' 3D bounding boxes. The last three corruptions simulate shape deformation of objects, and *Rotation* can also simulate different view directions of objects. They can affect both LiDAR and camera inputs. To make consistent distortions to two modalities, we apply the same transformation of shear, scale, or rotation to both points and image patches belonging to the objects in the scene.

Alignment-level corruptions: It was typically assumed that LiDAR and camera inputs are well aligned before feeding to the fusion models. However, this assumption can be invalid during long-time driving, *e.g.*, the collection of the ONCE dataset [37] needs re-calibration almost every day to avoid misalignment between different sensors. In practice, an autonomous vehicle can encounter *Spatial Misalignment* and *Temporal Misalignment* [63]. Spatial misalignment can be caused by sensor vibration due to bumps of the vehicle. We simulate it by adding random noises to the calibration matrices. Temporal misalignment happens when the data is stuck or delayed for a sensor. We keep the input of one modality the same as that at the previous timestamp to simulate temporal misalignment between the two modalities.

Discussion about the gap between synthetic and real-world corruptions. Real-world corruptions can come from multiple and diverse sources. For example, an autonomous vehicle can encounter adverse weather and uncommon objects at the same time, leading to much more complicated corruptions. Although it is impossible to enumerate all real-world corruptions, we systematically design 27 corruption types grouped into five levels, which can serve as a practical testbed to perform controllable robustness evaluation. Especially, for weather-level corruptions, we adopt the state-of-the-art methods for simulation, which are shown to approximate real data well [21,22]. Although there inevitably exists a gap, we validate that the model performance on synthetic weathers are consistent with that on real data under adverse weathers. More discussions are provided in Appendix A.4.

4. Corruption Robustness Benchmarks

To comprehensively evaluate the corruption robustness of 3D object detection models, we establish three corruption robustness benchmarks based on the most widely used datasets in autonomous driving—KITTI [17], nuScenes [6], and Waymo [51]. We apply the aforementioned corruptions to the validation sets of these datasets and obtain **KITTI-C**, **nuScenes-C**, and **Waymo-C**, respectively. Note that although several corruptions naturally appear in few samples of the datasets, we still apply the synthetic corruptions to all data to fairly compare model robustness under different corruptions and reduce the efforts of filtering data. Besides, we build a unified toolkit comprising of all corruptions, that can be used for other datasets as well. Below we introduce the dataset details, evaluation metrics, and evaluated models of the three benchmarks, respectively.

4.1. KITTI-C

The KITTI dataset [17] contains 3712 training, 3769 validation, and 7518 test samples. As we do not have access to the test set, KITTI-C is constructed upon the validation set. Among the corruptions, we do not include *FOV Lost*, *Motion Compensation* and *Temporal Misalignment* since: 1) 3D object detection models usually take front-view point clouds of 90° FOV as inputs since the KITTI dataset only

Model	Modality	Representation	Detection
SECOND [60]	LiDAR-only	voxel-based	one-stage
PointPillars [29]	LiDAR-only	voxel-based	one-stage
PointRCNN [48]	LiDAR-only	point-based	two-stage
Part-A^2 [49]	LiDAR-only	voxel-based	two-stage
PV-RCNN [47]	LiDAR-only	point-voxel-based	two-stage
3DSSD [61]	LiDAR-only	point-based	one-stage
SMOKE [36]	camera-only	monocular	one-stage
PGD [55]	camera-only	monocular	one-stage
ImVoxelNet [45]	camera-only	monocular	one-stage
EPNet [26]	fusion	point-level	two-stage
Focals Conv [13]	fusion	point-level	two-stage

(a) Evaluated models on KITTI-C.

Model	Modality	Representation	Detection
PointPillars [29]	LiDAR-only	voxel-based	one-stage
SSN [70]	LiDAR-only	voxel-based	one-stage
CenterPoint [62]	LiDAR-only	voxel-based	two-stage
FCOS3D [56]	camera-only	monocular	one-stage
PGD [55]	camera-only	monocular	one-stage
DETR3D [58]	camera-only	multi-view	query-based
BEVFormer [33]	camera-only	multi-view	query-based
FUTR3D [12]	fusion	proposal-level	query-based
TransFusion [2]	fusion	proposal-level	query-based
BEVFusion [35]	fusion	unified	query-based

(b) Evaluated models on nuScenes-C.

Table 1. The 3D object detection models adopted for corruption robustness evaluation on KITTI-C and nuScenes-C. We show the input modality, representation learning method (see Sec. 2.1), and detection head of each model.

provides box annotations in front of the vehicle; 2) the localization and timestamp information of each frame is not provided in the dataset. Therefore, there are 24 corruptions in KITTI-C with 5 severities for each following [25].

The standard evaluation is performed on *Car*, *Pedestrian* and *Cyclist* categories at *Easy*, *Moderate* and *Hard* levels of difficulty. The evaluation metric is the Average Precision (AP) with 40 recall positions at an IoU threshold 0.7 for cars and 0.5 for pedestrians/cyclists. We denote model performance on the original validation set as $\mathrm{AP}_{\mathrm{clean}}$. For each corruption type c at each severity s, we adopt the same metric to measure model performance as $\mathrm{AP}_{c,s}$. Then, the *corruption robustness* of a model is calculated by averaging over all corruption types and severities as

$$\mathrm{AP}_{\mathrm{cor}} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{1}{5}\sum_{s=1}^{5}\mathrm{AP}_{c,s}, \tag{1}$$

where $\mathcal{C}$ is the set of corruptions in evaluation. Note that for different kinds of 3D object detectors, the set of corruptions can be different (*e.g.*, we do not evaluate camera noises for LiDAR-only models), thus the results of $\mathrm{AP}_{\mathrm{cor}}$ are *not* directly comparable between different kinds of models and we perform a fine-grained analysis under each corruption. We also calculate *relative corruption error (RCE)* by measuring the percentage of performance drop as

$$\mathrm{RCE}_{c,s} = \frac{\mathrm{AP}_{\mathrm{clean}} - \mathrm{AP}_{c,s}}{\mathrm{AP}_{\mathrm{clean}}};\ \mathrm{RCE} = \frac{\mathrm{AP}_{\mathrm{clean}} - \mathrm{AP}_{\mathrm{cor}}}{\mathrm{AP}_{\mathrm{clean}}}. \tag{2}$$

We select 11 representative 3D object detection models trained on KITTI, including 6 LiDAR-only models: *SECOND* [60], *PointPillars* [29], *PointRCNN* [48], *Part-A^2* [49], *PV-RCNN* [47], and *3DSSD* [61]; 3 camera-only models: *SMOKE* [36], *PGD* [55], and *ImVoxelNet* [45]; and 2 LiDAR-camera fusion models: *EPNet* [26] and *Focals Conv* [13]. The details regarding their representations and detection heads are shown in Table 1(a).

4.2. nuScenes-C

The nuScenes dataset [6] contains 1000 sequences of approximately 20s duration with a LiDAR frequency of 20 FPS. The box annotations are provided for every 0.5s. Each frame has one point cloud and six images covering 360° horizontal FOV. In total, there are 40k frames which are split into 28k, 6k, 6k for training, validation, and testing. As the dataset provides full annotations and information of vehicle pose and timestamp, we can simulate all corruptions. Thus, we apply all 27 corruptions to the nuScenes validation set with 5 severities to obtain nuScenes-C.

For 3D object detection, the main evaluation metrics are mean Average Precision (mAP) and nuScenes detection score (NDS) computed on 10 object categories. The mAP is calculated using the 2D center distance on the ground plane instead of the 3D IoU. The NDS metric consolidates mAP and other aspects (*e.g.*, scale, orientation) into a unified score. Similar to KITTI-C, we denote model performance on the validation set as $\mathrm{mAP}_{\mathrm{clean}}$ and $\mathrm{NDS}_{\mathrm{clean}}$, and measure the corruption robustness $\mathrm{mAP}_{\mathrm{cor}}$ and $\mathrm{NDS}_{\mathrm{cor}}$ by averaging over all corruptions and severities. We also compute the relative corruption error RCE under both mAP and NDS metrics similar to Eq. (2).

On nuScenes-C, we select 10 3D detectors, including 3 LiDAR-only models: *PointPillars* [29], *SSN* [70], and *CenterPoint* [62]; 4 camera-only models: *FCOS3D* [56], *PGD* [55], *DETR3D* [58], and *BEVFormer* [33]; and 3 LiDAR-camera fusion models: *FUTR3D* [12], *TransFusion* [2], and *BEVFusion* [35]. The model details are shown in Table 1(b).

4.3. Waymo-C

The Waymo open dataset [51] consists of 798 scenes for training and 202 scenes for validation. Similar to nuScenes-C, Waymo-C is constructed by applying all 27 corruptions to the Waymo validation set with 5 severities. The official evaluation metrics are mAP and mAPH by taking the heading accuracy into consideration. We similarly calculate the corruption robustness and relative corruption error on Waymo-C. Due to the license agreement, there are no pre-train models publicly. Thus, we train the LiDAR-only *PointPillars* [29], camera-only *BEVFormer* [33], and LiDAR-camera fusion *TransFusion* [2] on a subset of training data [33] for robustness evaluation.

Corruption		LiDAR-only						Camera-only			LC Fusion	
		SECOND	PointPillars	PointRCNN	Part-A^2	PV-RCNN	3DSSD	SMOKE	PGD	ImVoxelNet	EPNet	Focals Conv
None ($\mathrm{AP}_{\mathrm{clean}}$)		81.59	78.41	80.57	82.45	**84.39**	80.03	7.09	8.10	**11.49**	82.72	**85.88**
Weather	Snow	**52.34**	36.47	50.36	42.70	**52.35**	27.12	2.47	0.63	0.22	34.58	34.77
	Rain	**52.55**	36.18	51.27	41.63	51.58	26.28	3.94	3.06	1.24	36.27	41.30
	Fog	74.10	64.28	72.14	71.61	**79.47**	45.89	5.63	0.87	1.34	44.35	44.55
	Sunlight	78.32	62.28	62.78	76.45	79.91	26.09	6.00	7.07	10.08	69.65	**80.97**
Sensor	Density	80.18	76.49	80.35	80.53	82.79	77.65	-	-	-	82.09	**84.95**
	Cutout	73.59	70.28	73.94	76.08	76.09	73.05	-	-	-	76.10	**78.06**
	Crosstalk	80.24	70.85	71.53	79.95	82.34	46.49	-	-	-	82.10	**85.82**
	Gaussian (L)	64.90	74.68	61.20	60.73	65.11	59.14	-	-	-	60.88	**82.14**
	Uniform (L)	79.18	77.31	76.39	77.77	81.16	74.91	-	-	-	79.24	**85.81**
	Impulse (L)	81.43	78.17	79.78	80.80	82.81	78.28	-	-	-	81.63	**85.01**
	Gaussian (C)	-	-	-	-	-	-	1.56	1.71	2.43	80.64	**80.97**
	Uniform (C)	-	-	-	-	-	-	2.67	3.29	4.85	81.61	**83.38**
	Impulse (C)	-	-	-	-	-	-	1.83	1.14	2.13	**81.18**	80.83
Motion	Moving Obj.	52.69	50.15	50.54	54.62	54.60	52.47	1.67	2.64	5.93	**55.78**	49.14
	Motion Blur	-	-	-	-	-	-	3.51	3.36	4.19	74.71	**81.08**
Object	Local Density	75.10	69.56	74.24	79.57	77.63	77.96	-	-	-	76.73	**80.84**
	Local Cutout	68.29	61.80	67.94	75.06	72.29	73.22	-	-	-	69.92	**76.64**
	Local Gaussian	72.31	76.58	69.82	77.44	70.44	75.11	-	-	-	75.76	**82.02**
	Local Uniform	80.17	78.04	77.67	80.77	82.09	78.64	-	-	-	81.71	**84.69**
	Local Impulse	81.56	78.43	80.26	82.25	84.03	79.53	-	-	-	82.21	**85.78**
	Shear	41.64	39.63	39.80	37.08	**47.72**	26.56	1.68	2.99	1.33	41.43	45.77
	Scale	73.11	70.29	71.50	75.90	**76.81**	75.02	0.13	0.15	0.33	69.05	69.48
	Rotation	76.84	72.70	75.57	77.50	**79.93**	76.98	1.11	2.14	2.57	74.62	77.76
Alignment	Spatial	-	-	-	-	-	-	-	-	-	35.14	43.01
Average ($\mathrm{AP}_{\mathrm{cor}}$)		70.45	65.48	67.74	69.92	**72.59**	60.55	2.68	2.42	**3.05**	67.81	**71.87**

Table 2. The benchmarking results of 11 3D object detectors on **KITTI-C**. We show the performance under each corruption and the overall corruption robustness $\mathrm{AP}_{\mathrm{cor}}$ averaged over all corruption types. The results are evaluated based on the car class at moderate difficulty.

5. Benchmarking Results

We present the evaluation results on KITTI-C in Sec. 5.1, nuScenes-C in Sec. 5.2, and leave the results on Waymo-C in Appendix D. We summarize the key findings in Sec. 6.

5.1. Results on KITTI-C

We show the corruption robustness of 11 3D object detection models on KITTI-C in Table 2, in which we only report the results on the car class at moderate difficulty, while leaving full results of other classes and difficulties in Appendix B. Overall, the corruption robustness is highly correlated with the clean accuracy, as the models (*e.g.*, PV-RCNN, Focals Conv) with higher $\mathrm{AP}_{\mathrm{clean}}$ also achieve higher $\mathrm{AP}_{\mathrm{cor}}$. It is not surprising due to the consistent performance degradation of different models. We further show the relative corruption error RCE of these models under each level of corruptions in Fig. 3. Based on the evaluation results, we provide the analyses below.

Comparison of corruption types. Based on Table 2 and Fig. 3, we can observe that weather-level and motion-level corruptions affect the performance of LiDAR-only and fusion models most, while all corruptions cause significant performance drop for camera-only models. For example, *Snow* and *Rain* lead to more than 35% RCE for all models, demonstrating the threats of adverse weathers on 3D object detectors. Besides, *Moving Object* and *Shear* are also challenging for all models, while *Spatial Misalignment* has a great impact on fusion models. On the other hand, most models exhibit negligible performance drop under sensor-level and object-level corruptions, mainly due to their ubiquity in the training dataset.

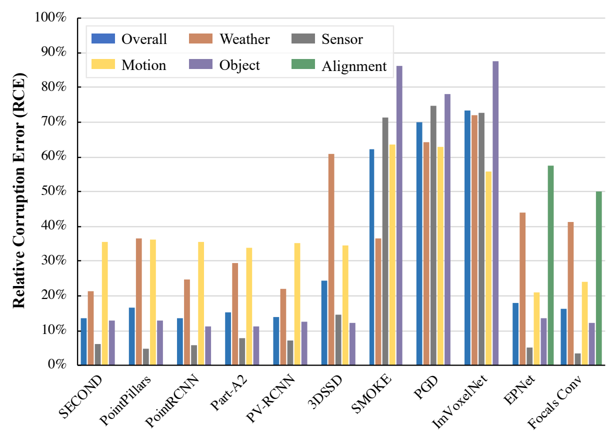

Figure 3. The relative corruption error RCE of 11 3D object detectors on **KITTI-C**. We show the overall results under all corruptions and the results under each level of corruptions.

Comparison of 3D object detectors. Due to the inferior performance of camera-only models, we mainly compare LiDAR-only and LiDAR-camera fusion models. We notice that for corruptions that affect both modalities (*e.g.*, *Snow*, *Moving Object*, *Shear*), LiDAR-only models lead to better performance. But for those that only corrupt point clouds (*e.g.*, sensor noises), fusion models are more competitive. This is due to that the accurate image data can endow fusion models with better robustness under point cloud noises, but when images are also corrupted, fusion models are affected by both inputs, resulting in inferior performance. To further validate this, we apply sensor noises to LiDAR and camera inputs at the same time. We show the performance of Fo-

Corruption		LiDAR-only			Camera-only				LC Fusion		
		PointPillars	SSN	CenterPoint	FCOS3D	PGD	DETR3D	BEVFormer	FUTR3D	TransFusion	BEVFusion
None ($\mathrm{mAP_{clean}}$)		27.69	46.65	**59.28**	23.86	23.19	34.71	**41.65**	64.17	66.38	**68.45**
Weather	Snow	27.57	46.38	55.90	2.01	2.30	5.08	5.73	52.73	**63.30**	62.84
	Rain	27.71	46.50	56.08	13.00	13.51	20.39	24.97	58.40	65.35	**66.13**
	Fog	24.49	41.64	43.78	13.53	12.83	27.89	32.76	53.19	53.67	**54.10**
	Sunlight	23.71	40.28	54.20	17.20	22.77	34.66	41.68	57.70	55.14	**64.42**
Sensor	Density	27.27	46.14	58.60	-	-	-	-	63.72	65.77	**67.79**
	Cutout	24.14	40.95	56.28	-	-	-	-	62.25	63.66	**66.18**
	Crosstalk	25.92	44.08	56.64	-	-	-	-	62.66	64.67	**67.32**
	FOV Lost	8.87	15.40	20.84	-	-	-	-	26.32	24.63	**27.17**
	Gaussian (L)	19.41	39.16	45.79	-	-	-	-	58.94	55.10	**60.64**
	Uniform (L)	25.60	45.00	56.12	-	-	-	-	63.21	64.72	**66.81**
	Impulse (L)	26.44	45.58	57.67	-	-	-	-	63.43	65.51	**67.54**
	Gaussian (C)	-	-	-	3.96	4.33	14.86	15.04	54.96	**64.52**	64.44
	Uniform (C)	-	-	-	8.12	8.48	21.49	23.00	57.61	65.26	**65.81**
	Impulse (C)	-	-	-	3.55	3.78	14.32	13.99	55.16	**64.37**	64.30
Motion	Compensation	3.85	10.39	11.02	-	-	-	-	**31.87**	9.01	27.57
	Moving Obj.	19.38	35.11	44.30	10.36	10.47	16.63	20.22	45.43	51.01	**51.63**
	Motion Blur	-	-	-	10.19	9.64	11.06	19.79	55.99	64.39	**64.74**
Object	Local Density	26.70	45.42	57.55	-	-	-	-	63.60	65.65	**67.42**
	Local Cutout	17.97	32.16	48.36	-	-	-	-	61.85	63.33	**63.41**
	Local Gaussian	25.93	43.71	51.13	-	-	-	-	62.94	63.76	**64.34**
	Local Uniform	27.69	46.87	57.87	-	-	-	-	64.09	66.20	**67.58**
	Local Impulse	27.67	46.88	58.49	-	-	-	-	64.02	66.29	**67.91**
	Shear	26.34	43.28	49.57	17.20	16.66	17.46	24.71	55.42	**62.32**	60.72
	Scale	27.29	45.98	51.13	6.75	6.57	12.02	17.64	56.79	64.13	**64.57**
	Rotation	27.80	46.93	54.68	17.21	16.84	27.28	33.97	59.64	63.36	**65.13**
Alignment	Spatial	-	-	-	-	-	-	-	63.77	66.22	**68.39**
	Temporal	-	-	-	-	-	-	-	**51.43**	43.65	49.02
Average ($\mathrm{mAP_{cor}}$)		23.42	40.37	**49.81**	10.26	10.68	18.60	**22.79**	56.99	58.73	**61.03**

Table 3. The benchmarking results of 10 3D object detectors on **nuScenes-C**. We show the performance under each corruption and the overall corruption robustness $\mathrm{mAP_{cor}}$ averaged over all corruption types.

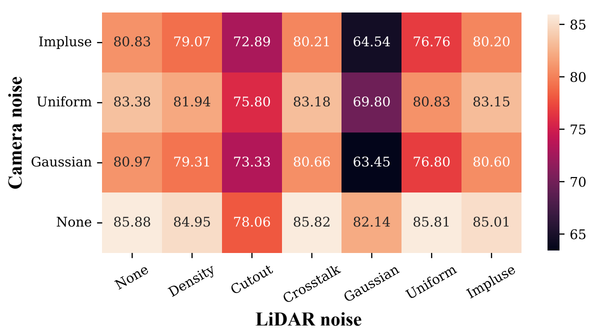

Figure 4. The performance of Focals Conv [13] under the concurrence of LiDAR and camera noises.

cals Conv [13] under the concurrence of LiDAR and camera noises in Fig. 4. It can be seen that the accuracy of Focals Conv further drops in the presence of both LiDAR and camera noises, leading to worse performance than LiDAR-only models that cannot be affected by camera noises. The results demonstrate that although fusion models are more robust to noises of one modality, they are potentially exposed to corruptions from multiple sensors.

Comparison of LiDAR-only models. Among the six LiDAR-only detectors, we find that SECOND [60], PointRCNN [48], and PV-RCNN [47] possess better relative corruption robustness than the others, whose RCE is 13.65%, 13.61%, and 13.99%. The worst model is 3DSSD, exhibiting a 24.34% performance drop. In general, there does not exist a clear margin of robustness between voxel-based and point-based detectors, or between one-stage and two-stage detectors, different from previous findings [32]. However, we notice that the worst two models PointPillars [29] and 3DSSD [61] are developed for improving the efficiency of 3D object detection, which may indicate a trade-off between corruption robustness and efficiency.

5.2. Results on nuScenes-C

We report the corruption robustness of 10 3D detectors on nuScenes-C in Table 3 under the mAP metric, and leave the results under the NDS metric in Appendix C. The model performance is consistent for both metrics. We further show the relative corruption error RCE under each level of corruptions in Fig. 5. Similar to the results on KITTI-C, models that have higher clean accuracy generally achieve better corruption robustness. But differently, the nuScenes dataset provides multi-view images, thus the camera-only models achieve competitive clean accuracy with LiDAR-only models, enabling us to compare their performance. We provide more detailed analyses below.

Comparison of corruption types. From Fig. 5, we can observe that motion-level corruptions are significantly more detrimental to LiDAR-only and LiDAR-camera fusion models. They give rise to more than 50% performance drop for LiDAR-only models and about 30% drop for fusion models. Similar to KITTI-C, all corruptions remarkably degrade the performance of camera-only models. A notable difference from KITTI-C is that most models are resistant to weather-level corruptions. We think that the adverse weathers (*e.g.*, rain) contained in the nuScenes dataset

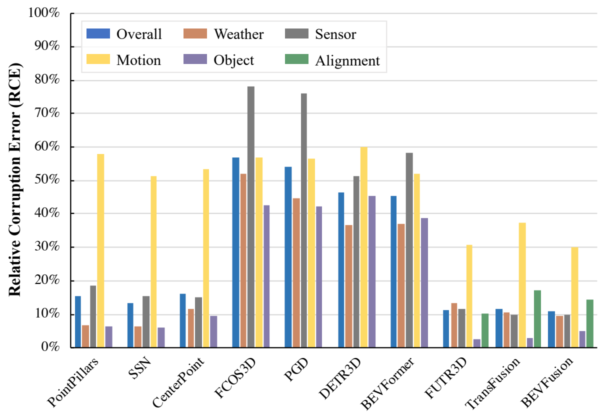

Figure 5. The relative corruption error RCE of 10 3D object detectors on **nuScenes-C**. We show the overall results under all corruptions and the results under each level of corruptions.

enable the detectors to predict robustly under weather-level corruptions. Among all corruptions, *FOV Lost* and *Motion Compensation* impair the models most, mainly due to the large distortions of the LiDAR point clouds.

Comparison of 3D object detectors. For different categories of 3D object detectors, camera-only models are more prone to common corruptions, whose performance drops more than 40% under RCE. On the contrary, LiDAR-only and fusion models exhibit less than 20% performance drop. The reason is that LiDAR point clouds are inherently noisy due to the ranging inaccuracy [7] and self-occlusion, such that the models trained on point clouds are relatively robust to corruptions. The results may suggest the indispensability of LiDAR point clouds for reliable 3D object detection.

Comparison of camera-only models. Though camera-only detectors are greatly affected by corruptions, we find that multi-view methods outperform monocular methods in terms of both clean and corruption accuracy. From Fig. 5, the overall performance drop of FCOS3D and PGD is 57% and 54%, while that of DETR3D and BEVFormer is 46% and 45%, respectively. Since monocular methods directly predict 3D objects from single images without considering 3D scene structure, they are more prone to noises [58] and exhibit inferior performance. Besides, BEVFormer performs better than DETR3D, especially under object-level corruptions (*e.g.*, *Shear*, *Rotation*), since it can capture both semantic and location information of objects in the BEV space with being less affected by varying object shapes [30].

Comparison of LiDAR-camera fusion models. Based on the above analysis, fusion models demonstrate superior corruption robustness on nuScene-C. By carefully examining their performance, we find that there exists a trade-off between robustness under image corruptions and point cloud corruptions. Specifically, FUTR3D suffers from the largest performance drop (12.9% on average) under *Gaussian*, *Uniform* and *Impluse* noises of images, compared with 2.5% of TransFusion and 5.3% of BEVFusion. However, under *Motion Compensation* that significantly distorts point clouds, FUTR3D obtains the highest mAP of 31.87% while TransFusion only has 9.01% mAP. The reason behind this trade-off is that fusion models have varying reliance on images or point clouds, resulting in the inconsistent robustness under the corresponding corruptions of different sensors.

6. Discussion and Conclusion

In this paper, we systematically design 27 types of common corruptions in 3D object detection to benchmark corruption robustness of existing 3D object detectors. We establish three corruption robustness benchmarks—KITTI-C, nuScenes-C, and Waymo-C by synthesizing the corruptions on public datasets. By conducting large-scale experiments on 24 diverse 3D object detection models under corruptions, we draw some important findings, as summarized below:

1) In general, the corruption robustness of 3D object detection models is largely correlated with their clean performance, similar to the observation in [25].
2) Among all corruption types, motion-level ones degrade the model performance most, which pose a significant threat to autonomous driving. Weather-level corruptions are also influential to models trained on normal weather.
3) Among all 3D detectors, LiDAR-camera fusion models have better corruption robustness, especially under those that apply distortions to only one modality. However, they are also exposed to corruptions from both sensors, leading to degraded performance in this case. Besides, there is a trade-off between robustness under image corruptions and point cloud corruptions of fusion models.
4) Camera-only models are more easily affected by common corruptions, demonstrating the indispensability of LiDAR point clouds for reliable 3D detection or the necessity of developing more robust camera-only models.
5) In Appendix E, we further try several data augmentation strategies, including those applied to point clouds [14, 67] and images [64, 66]. The experiments validate that they can hardly improve corruption robustness, leaving robustness enhancement of 3D object detection an open problem for future research.

We hope our comprehensive benchmarks, in-depth analyses, and insightful findings can be helpful for understanding the corruption robustness of 3D object detection models and improving their robustness in future.

Acknowledgement

This work was supported by the National Key Research and Development Program of China (Nos. 2020AAA0106302, 2017YFA0700904, 2020AAA0104304), NSFC Projects (Nos. 62276149, 62061136001, 62076145, 62076147, U19B2034, U1811461, U19A2081, 61972224), BNRist (BNR2022RC01006), Tsinghua Institute for Guo Qiang, and the High Performance Computing Center, Tsinghua University. Y. Dong was also supported by the China National Postdoctoral Program for Innovative Talents and Shuimu Tsinghua Scholar Program. J. Zhu was also supported by the XPlorer Prize.

References

[1] Safe driving cars. *Nature Machine Intelligence*, 4:95–96, 2022. 1

[2] 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 (CVPR)*, pages 1090–1099, 2022. 3, 5

[3] Matthew Berger, Andrea Tagliasacchi, Lee Seversky, Pierre Alliez, Joshua Levine, Andrei Sharf, and Claudio Silva. State of the art in surface reconstruction from point clouds. *Eurographics 2014-State of the Art Reports*, 1(1):161–185, 2014. 3

[4] Mario Bijelic, Tobias Gruber, Fahim Mannan, Florian Kraus, Werner Ritter, Klaus Dietmayer, and Felix Heide. Seeing through fog without seeing fog: Deep multimodal sensor fusion in unseen adverse weather. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11682–11692, 2020. 1, 3

[5] Lara Brinon-Arranz, Tiana Rakotovao, Thierry Creuzet, Cem Karaoguz, and Oussama El-Hamzaoui. A methodology for analyzing the impact of crosstalk on lidar measurements. In *2021 IEEE Sensors*, pages 1–4, 2021. 3

[6] 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 (CVPR)*, pages 11621–11631, 2020. 1, 2, 4, 5

[7] Alexander Carballo, Jacob Lambert, Abraham Monrroy, David Wong, Patiphon Narksri, Yuki Kitsukawa, Eijiro Takeuchi, Shinpei Kato, and Kazuya Takeda. Libre: The multiple 3d lidar dataset. In *IEEE Intelligent Vehicles Symposium (IV)*, pages 1094–1101, 2020. 1, 3, 8

[8] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European Conference on Computer Vision (ECCV)*, pages 213–229, 2020. 2

[9] Robin Chan, Krzysztof Lis, Svenja Uhlemeyer, Hermann Blum, Sina Honari, Roland Siegwart, Pascal Fua, Mathieu Salzmann, and Matthias Rottmann. Segmentmeifyoucan: A benchmark for anomaly segmentation. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 1, 3, 4

[10] Xiaozhi Chen, Kaustav Kundu, Ziyu Zhang, Huimin Ma, Sanja Fidler, and Raquel Urtasun. Monocular 3d object detection for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2147–2156, 2016. 2

[11] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1907–1915, 2017. 1, 3

[12] 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. 3, 5

[13] Yukang Chen, Yanwei Li, Xiangyu Zhang, Jian Sun, and Jiaya Jia. Focal sparse convolutional networks for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5428–5437, 2022. 3, 5, 7

[14] Jaeseok Choi, Yeji Song, and Nojun Kwak. Part-aware data augmentation for 3d object detection in point cloud. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3391–3397, 2021. 2, 8

[15] Carlos A Diaz-Ruiz, Youya Xia, Yurong You, Jose Nino, Junan Chen, Josephine Monica, Xiangyu Chen, Katie Luo, Yan Wang, Marc Emond, et al. Ithaca365: Dataset and driving perception under repeated and challenging weather conditions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21383–21392, 2022. 1, 3

[16] 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. 4

[17] 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 (CVPR)*, pages 3354–3361, 2012. 1, 2, 3, 4

[18] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations (ICLR)*, 2019. 3

[19] Robert Geirhos, Carlos RM Temme, Jonas Rauber, Heiko H Schütt, Matthias Bethge, and Felix A Wichmann. Generalisation in humans and deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 7549–7561, 2018. 3

[20] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 3

[21] Martin Hahner, Christos Sakaridis, Mario Bijelic, Felix Heide, Fisher Yu, Dengxin Dai, and Luc Van Gool. Lidar snowfall simulation for robust 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16364–16374, 2022. 1, 3, 4

[22] Martin Hahner, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Fog simulation on real lidar point clouds for 3d object detection in adverse weather. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, pages 15283–15292, 2021. 1, 3, 4

[23] Dan Hendrycks, Steven Basart, Mantas Mazeika, Andy Zou, Joseph Kwon, Mohammadreza Mostajabi, Jacob Steinhardt, and Dawn Song. Scaling out-of-distribution detection for real-world settings. In *International Conference on Machine Learning (ICML)*, pages 8759–8773, 2022. 1, 3

[24] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8340–8349, 2021. 3

[25] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations (ICLR)*, 2019. 1, 2, 3, 5, 8

[26] Tengteng Huang, Zhe Liu, Xiwu Chen, and Xiang Bai. Epnet: Enhancing point features with image semantics for 3d object detection. In *European Conference on Computer Vision (ECCV)*, pages 35–52, 2020. 3, 5

[27] Velat Kilic, Deepti Hegde, Vishwanath Sindagi, A Brinton Cooper, Mark A Foster, and Vishal M Patel. Lidar light scattering augmentation (lisa): Physics-based simulation of adverse weather conditions for 3d object detection. *arXiv preprint arXiv:2107.07004*, 2021. 1, 3

[28] Jason Ku, Melissa Mozifian, Jungwook Lee, Ali Harakeh, and Steven L Waslander. Joint 3d proposal generation and object detection from view aggregation. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1–8, 2018. 1

[29] 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 (CVPR)*, pages 12697–12705, 2019. 1, 2, 5, 7

[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. 8

[31] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. *arXiv preprint arXiv:2203.07724*, 2022. 1, 3, 4

[32] Shuangzhi Li, Zhijie Wang, Felix Juefei-Xu, Qing Guo, Xingyu Li, and Lei Ma. Common corruption robustness of point cloud detectors: Benchmark and enhancement. *arXiv preprint arXiv:2210.05896*, 2022. 3, 7

[33] 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. In *European Conference on Computer Vision (ECCV)*, 2022. 2, 3, 5

[34] Ming Liang, Bin Yang, Shenlong Wang, and Raquel Urtasun. Deep continuous fusion for multi-sensor 3d object detection. In *European Conference on Computer Vision (ECCV)*, pages 641–656, 2018. 1

[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. 2, 3, 5

[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, 5

[37] Jiageng Mao, Minzhe Niu, Chenhan Jiang, Jingheng Chen, Xiaodan Liang, Yamin Li, Chaoqiang Ye, Wei Zhang, Zhenguo Li, Jie Yu, et al. One million scenes for autonomous driving: Once dataset. In *Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. 4

[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] Arsalan Mousavian, Dragomir Anguelov, John Flynn, and Jana Kosecka. 3d bounding box estimation using deep learning and geometry. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7074–7082, 2017. 1, 2

[40] Peter Pinggera, Sebastian Ramos, Stefan Gehrig, Uwe Franke, Carsten Rother, and Rudolf Mester. Lost and found: detecting small road hazards for self-driving vehicles. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1099–1106, 2016. 1, 3

[41] Matthew Pitropov, Danson Evan Garcia, Jason Rebello, Michael Smart, Carlos Wang, Krzysztof Czarnecki, and Steven Waslander. Canadian adverse driving conditions dataset. *The International Journal of Robotics Research*, 40(4-5):681–690, 2021. 1, 3

[42] Charles R Qi, Wei Liu, Chenxia Wu, Hao Su, and Leonidas J Guibas. Frustum pointnets for 3d object detection from rgb-d data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 918–927, 2018. 3

[43] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017. 2

[44] Jiawei Ren, Liang Pan, and Ziwei Liu. Benchmarking and analyzing point cloud classification under corruptions. In *International Conference on Machine Learning (ICML)*, pages 18559–18575, 2022. 1, 3

[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. 5

[46] Birgit Schlager, Thomas Goelles, Marco Behmer, Stefan Muckenhuber, Johann Payer, and Daniel Watzenig. Automotive lidar and vibration: Resonance, inertial measurement unit, and effects on the point cloud. *IEEE Open Journal of Intelligent Transportation Systems*, 2022. 3

[47] Shaoshuai Shi, Chaoxu Guo, Li Jiang, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. Pv-rcnn: Point-voxel feature set abstraction for 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision*

and Pattern Recognition (CVPR)*, pages 10529–10538, 2020. 1, 2, 5, 7

[48] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–779, 2019. 1, 2, 5, 7

[49] Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(8):2647–2664, 2020. 2, 5

[50] 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. 1, 3

[51] 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 (CVPR)*, pages 2446–2454, 2020. 1, 2, 4, 5

[52] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. 3

[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 (CVPR)*, pages 4604–4612, 2020. 1, 3

[54] 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. 3

[55] Tai Wang, ZHU Xinge, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning (CoRL)*, pages 1475–1485, 2022. 5

[56] 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 (ICCV)*, pages 913–922, 2021. 1, 2, 5

[57] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8445–8453, 2019. 1, 2

[58] 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 (CoRL)*, pages 180–191, 2022. 1, 2, 5, 8

[59] Weisong Weisong Wen, Guohao Zhang, and Li-Ta Hsu. Gnss nlos exclusion based on dynamic object detection using lidar point cloud. *IEEE Transactions on Intelligent Transportation Systems*, 22(2):853–862, 2019. 4

[60] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 1, 2, 5, 7

[61] Zetong Yang, Yanan Sun, Shu Liu, and Jiaya Jia. 3dssd: Point-based 3d single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11040–11048, 2020. 2, 5, 7

[62] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11784–11793, 2021. 5

[63] 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, 4

[64] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6023–6032, 2019. 2, 8

[65] Shizhe Zang, Ming Ding, David Smith, Paul Tyler, Thierry Rakotoarivelo, and Mohamed Ali Kaafar. The impact of adverse weather conditions on autonomous vehicles: How rain, snow, fog, and hail affect the performance of a self-driving car. *IEEE Vehicular Technology Magazine*, 14(2):103–111, 2019. 3

[66] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2018. 8

[67] Jinlai Zhang, Lyujie Chen, Bo Ouyang, Binbin Liu, Jihong Zhu, Yujin Chen, Yanmei Meng, and Danfeng Wu. Pointcutmix: Regularization strategy for point cloud classification. *Neurocomputing*, 505:58–67, 2022. 2, 8

[68] Ji Zhang and Sanjiv Singh. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*, pages 1–9, 2014. 4

[69] 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 (CVPR)*, pages 4490–4499, 2018. 1, 2

[70] Xinge Zhu, Yuexin Ma, Tai Wang, Yan Xu, Jianping Shi, and Dahua Lin. Ssn: Shape signature networks for multi-class object detection from point clouds. In *European Conference on Computer Vision (ECCV)*, pages 581–597, 2020. 2, 5