Benchmarking and Analyzing Bird's Eye View Perception Robustness to Corruptions

Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu

Abstract—Recent advancements in bird's eye view (BEV) representations have shown remarkable promise for in-vehicle 3D perception. However, while these methods have achieved impressive results on standard benchmarks, their robustness in varied conditions, crucial for ensuring safe operations, remains insufficiently assessed. In this study, we present **RoboBEV**, an extensive benchmark suite designed to evaluate the resilience of BEV algorithms. This suite incorporates eight diverse image corruptions, namely Brightness, Darkness, Fog, Snow, Motion Blur, Color Quantization, Camera Crash, and Frame Lost, with each corruption examined over three severity levels. Significantly, our benchmark also considers the impact of complete sensor failures in multimodal perception models. Through RoboBEV, we rigorously assess 33 state-of-the-art BEV-based models spanning tasks like detection, map segmentation, depth estimation, and semantic occupancy prediction. Our analyses reveal a noticeable correlation between a model's performance on in-distribution datasets and its resilience to out-of-distribution challenges. Interestingly, while the absolute performance metrics showed consistency, relative performances varied significantly among different models. Our experimental results also underline the efficacy of strategies like pre-training and depth-free BEV transformations in enhancing robustness against out-of-distribution data. Furthermore, we observe that leveraging extensive temporal information significantly bolsters model robustness. The insights gleaned from this study pave the way for the development of future BEV models that seamlessly combine accuracy with real-world robustness. The benchmark toolkit and model checkpoints are publicly accessible at: https://github.com/Daniel-xsy/RoboBEV.

Index Terms—3D Object Detection, Bird's Eye View Segmentation, Semantic Occupancy Prediction, Out-of-Distribution Robustness.

1 Introduction

Deep neural network-based 3D perception methods have registered transformative breakthroughs, excelling in a range of demanding benchmarks [2], [3], [4], [5], [6], [7], [21], [22], [23], [32]. Among these, camera-centric methods [2], [3], [4], [5], [6], [7] have surged in popularity over their LiDAR-driven counterparts [21], [22], [23], [32], primarily due to advantages such as reduced deployment costs, augmented computational efficiency, and the ability to provide dense semantic insights. Central to many of these advancements is the bird's eye view (BEV) representation, which offers a trio of significant benefits:

- It facilitates unified learning from multi-view images.
- It encourages a physically interpretable methodology for fusing information across diverse sensors and temporal instances [33].
- Its output domain aligns seamlessly with several downstream applications like prediction and planning, fortifying the performance metrics of BEV-centric perception frameworks.

However, this blossoming landscape of BEV perception methodologies is not without its challenges. Despite their evident prowess, the resilience of these algorithms in the face of out-of-context or unforeseen scenarios remains under-examined. This oversight is particularly concerning given that many of these algorithms are envisioned to function in safety-critical realms such as autonomous driving. Traditionally, the robustness of algorithms can be bifurcated into adversarial robustness [8], [10], [11], [12], [19], [37], [38] – which delves into worst-case scenarios – and robustness under distribution shift [40], [47], [49], [49], [50] that examines average-case performance, often mirroring real-world conditions.

While adversarial robustness of 3D perception models has been studied by some [9], [13], [14], [52], this work seeks to explore a less-traveled avenue: robustness of BEV-centric 3D perception systems when subjected to natural, often unpredictable, corruptions.

In this work, to address the existing knowledge gap, we present a comprehensive benchmark dubbed *RoboBEV*. This benchmark evaluates the robustness of BEV perceptions against natural corruptions including exterior environments, interior sensors, and temporal factors. Specifically, the exterior environments include various light and weather conditions, which are simulated by incorporating *Brightness*, *Dark*, *Fog*, and *Snow* weathers. Additionally, the inputs may be corrupted by interior factors caused by sensors, such as *Motion Blur* and *Color Quant*. We further propose two novel corruptions in temporal space tailored for BEV-based temporal fusion strategies, namely *Camera Crash* and *Frame Lost*. Moreover, we consider complete sensor failure for camera-LiDAR fusion models [56], [58], [59] that are trained on multimodal input. The study involves a comprehensive investigation of diverse out-of-distribution corruption settings that are highly relevant to real-world autonomous driving applications.

Leveraging the proposed *RoboBEV* benchmark, we conduct an exhaustive analysis of 33 BEV perception models

- *S. Xie is with the Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA, US.*
- *L. Kong is with the School of Computing, Department of Computer Science, National University of Singapore.*
- *W. Zhang, J. Ren, L. Pan, and Z. Liu are with S-Lab, Nanyang Technological University, Singapore.*
- *K. Chen is with Shanghai AI Laboratory, China.*
- *Z. Liu serves as the corresponding author. E-mail: ziwei.liu@ntu.edu.sg.*

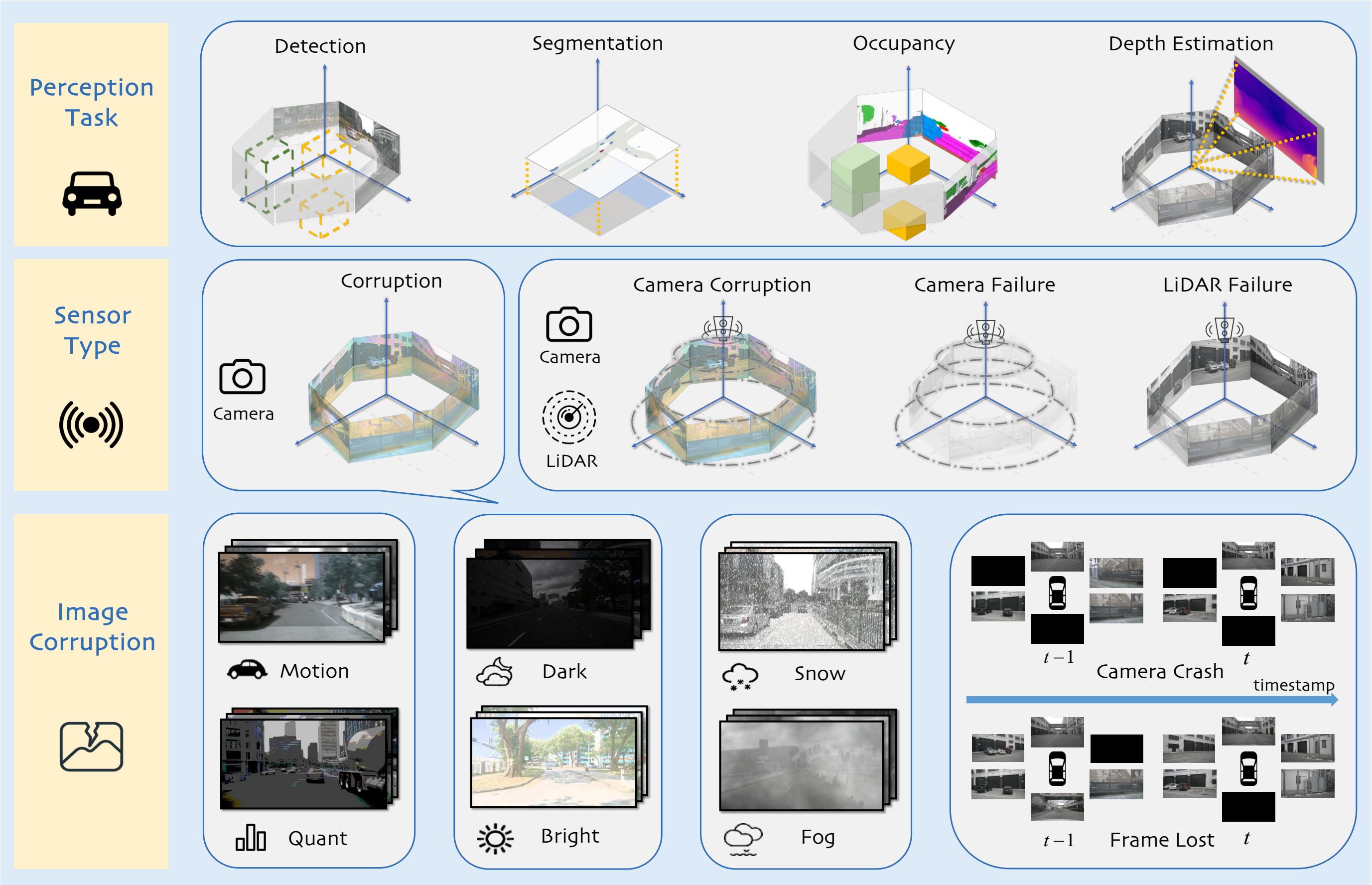


Fig. 1. ***RoboBEV* benchmark designs**. The benchmark comprehensively encompasses **four** distinct BEV perception tasks (detection, segmentation, occupancy prediction, and depth estimation), **four** diverse sensor type configurations in between LiDAR, cameras, and joint setups (camera corruption, camera failure, and LiDAR failure), and an array of **eight** natural image corruptions (Brightness, Darkness, Fog, Snow, Motion Blur, Color Quantization, Camera Crash, and Frame Lost), each categorized into **three** distinct severity levels.

under 8 corruptions across 3 severity levels. The key contributions of this work are summarized as follows:

1) We introduce *RoboBEV*, a comprehensive benchmark for evaluating BEV perception robustness under natural corruptions. Furthermore, we provide an open-source codebase, accessible through our repository. Additionally, we make the generated dataset publicly available, allowing the research community to replicate and extend our findings.
2) We conduct extensive experiments to assess the performance of 30 camera-based and 3 camera-LiDAR fusion-based BEV perception algorithms. These algorithms are evaluated across 8 distinct corruptions, each applied at 3 different severity levels, for a total of 4 perception tasks.
3) Our study offers valuable insights through in-depth analyses of the factors that contribute to superior robustness under corruption scenarios, which shed light on future model design. We mainly observe the following results: (a) the absolute performances show a strong correlation with the performances under corruption. However, the relative robustness does not necessarily increase as standard performance improves; (b) pre-training together with depth-free BEV transformation has great potential to enhance robustness; (c) utilizing long and rich temporal information largely helps with robustness.

2 Related Works

2.1 Camera-based Bird's Eye View Perception

BEV perception methodologies can be stratified into two primary branches predicated on the explicitness of their depth estimation. A segment of the literature, influenced by LSS [17], such as BEVDet [3], employs an auxiliary depth estimation branch to facilitate the transformation from perspective view to bird's eye view (PV2BEV). BEVDepth [16] refines this paradigm, enhancing depth estimation accuracy using explicit depth data from point clouds. Meanwhile, BEVerse [41] introduces a multi-task learning framework that achieves benchmark-setting outcomes. In contrast, an alternative research trajectory avoids explicit depth estimation. Drawing inspiration from DETR [15], models like DETR3D [4] and ORA3D [42] encapsulate 3D objects as queries, leveraging Transformers' cross-attention mechanisms. Following this, PETR [5] boosts performance by formulating 3D position-aware representations. Simultaneously, BEVFormer [2] and PolarFormer [43] venture into temporal cross-attention and polar coordinate-based 3D target predictions, respectively. Taking a leaf out of Sparse RCNN's [46] book, SRCN3D [44] and Sparse4D [45] pioneer sparse proposals for feature amalgamation. Meanwhile, SOLOFusion [55] pursues deeper historical data integration for temporal modeling. In addition to detection, BEV perception tasks also include map segmentation [18], [64], multiview depth estimation [60], and semantic occupancy prediction [61], [62], [65], [67], [68]. While these methodologies flaunt impressive outcomes on pristine datasets, their resilience against natural corruptions remains an enigma.

2.2 LiDAR-based 3D Perception

LiDAR, with its precision in capturing spatial relationships using laser beams, has paved the way for breakthroughs in 3D perception, central to applications like autonomous driving. Two primary tasks have gained prominence: 3D object detection and LiDAR semantic segmentation, both of which have inherent connections to BEV perception. In the realm of *3D object detection*, the focus has been on optimally representing LiDAR point cloud data [24]. Point-based approaches, such as those presented in [78], [79], [80], [81], shine in preserving the innate geometry of point clouds, capturing local structures and patterns. Meanwhile, voxel-based strategies, like [82], [83], [84], convert the irregular point clouds into structured grids, relying on sparse convolution techniques [82] to handle non-empty voxels efficiently. Pillar-based techniques, highlighted by works like [22], [85], offer a trade-off between detection accuracy and computational speed by fine-tuning the vertical resolution. Additionally, hybrid approaches, such as [86], [87], merge the strengths of both point and voxel representations to derive more enriched features. On the other hand, *semantic segmentation* techniques often pivot on the representation choice. Raw point methods, like [26], [88], emphasize the direct usage of irregular point clouds, while range view approaches, showcased in [27], [89], [90], [91], [97], convert these point clouds into 2D grids. This conversion aligns closely with BEV perception, transforming 3D data into a top-down perspective, essential for many applications. Further refining this idea are bird's eye view techniques, exemplified by [92], which offer a direct 2D top-down representation. Voxel-centric methods, such as [93], maintain the 3D spatial structure, often outperforming other singular modalities. Modern research, like [25], [28], [29], [30], [94], [95], [96], pushes the boundaries by exploring the fusion of multiple views, seeking to harness the complementary strengths of different representations. In essence, while LiDAR-based 3D perception methodologies, especially those linked with BEV perception, have exhibited significant promise, their resilience in real-world conditions warrants deeper exploration and validation.

2.3 Robustness under Adversarial Attacks

Modern neural networks, while showcasing staggering capabilities, remain vulnerable to adversarial onslaughts, where meticulously engineered perturbations in inputs can precipitate erroneous outputs [8], [11], [12]. The menace of adversarial examples has been a research epicenter across various vision domains: classification [8], [11], detection [10], [20], and segmentation [9], [10]. These adversarial stimuli can emerge in both digital domains [8], [11] and real-world environments [9], [51]. Alarming findings reveal that adversarial examples can cripple 3D perception systems, flagging potential safety concerns during practical deployments [9], [13], [14]. While Xie *et al.* [52] delve into the adversarial robustness of camera-centric detectors, our focus pivots towards more pervasive natural corruptions.

2.4 Robustness under Natural Corruptions

Assessing model tenacity against corruptions has burgeoned as a pivotal research domain. Several benchmarks, such as ImageNet-C [40], ObjectNet [50], ImageNetV2 [49], and more, evaluate the robustness of 2D image classifiers against an array of corruptions. For instance, ImageNet-C taints pristine ImageNet samples with simulated anomalies like compression artifacts and motion blur. On the other hand, ObjectNet offers a test set abundant in rotation, background, and viewpoint variances. Hendrycks *et al.* [48] underscore the correlation between synthetic corruption robustness and enhancements in real-world scenarios. Recently, Some works [66], [69], [70] endeavor to improve the robustness of 3D perception models. Kong *et al.* [75], [76] establish a robustness benchmark for monocular depth estimation under corruptions. Ren *et al.* [77] design atomic corruptions on indoor object-centric point clouds and CAD models to understand classifiers' robustness. Yet, a void persists concerning benchmarks for 3D BEV perception models, which play critical roles in safety-sensitive applications. While a concurrent study by Zhu *et al.* [63] explores a similar landscape, their narrative is predominantly adversarial-centric. In contrast, our benchmarks, spanning models, tasks, scenarios, and validation studies, offer a broader and more comprehensive lens into this domain.

3 Bird's Eye View Perception Preliminaries

3.1 Pre-training

Over the past few years, pre-training has staked its claim as an invaluable strategy, amplifying the efficiency of computer vision models in diverse tasks. Within the sphere of camera-driven 3D perception, initializing the ResNet backbone using FCOS3D [6] weights has become standard practice. To stabilize the training process, FCOS3D adjusts a depth weight from 0.2 to 1 during fine-tuning [6]. Another prevailing approach involves training the VoVNet-V2 [53] backbone on the DDAD15M [54] dataset, targeting depth estimation, before fine-tuning it using the nuScenes training set for detection. Semantically, these pre-training techniques fall into two categories: semantic and depth pre-training.

3.2 Temporal Fusion

The dynamic landscape of autonomous driving demands precise velocity estimates of moving entities, a challenge when relying on singular frame inputs. This accentuates the importance of temporal cues in fortifying vision systems' perception capabilities. Prior research has pioneered various methodologies to harness these temporal cues. For instance, BEVFormer [2] integrates history data and leverages temporal cross-attention to distill BEV features from multi-timestamp images. Meanwhile, BEVDet4D [34] appends features from antecedent frames to weave in temporal nuances, and SOLOFusion [55] aims for more inclusive temporal

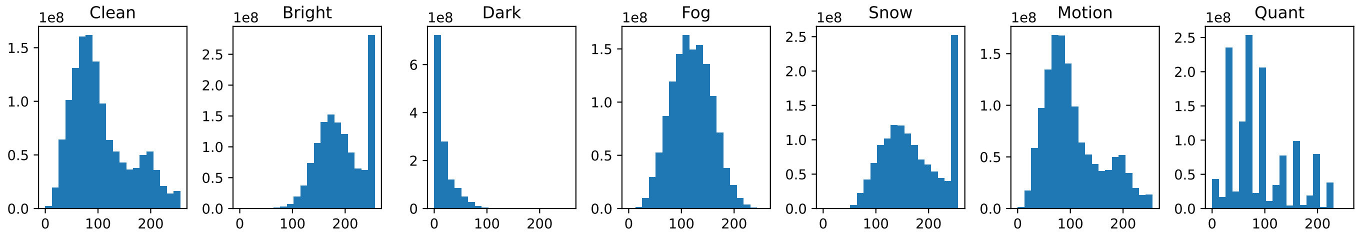


Fig. 2. Histograms of pixel distributions for different corruption types. While certain corruptions exhibit minimal shifts in pixel distribution (*e.g.*, Motion), it is noteworthy that these alterations predominantly have adverse effects on the overall performance of perception systems.

TABLE 1
Severity level setups in corruption simulations: Detailed parameters used for generating multi-level corruptions for each corruption type.

Corruption	Parameter	Easy	Moderate	Hard
Bright	adjustment in HSV space	0.2	0.4	0.5
Dark	scale factor	0.5	0.4	0.3
Fog	(thickness, smoothness)	(2.0, 2.0)	(2.5, 1.5)	(3.0, 1.4)
Snow	(mean, std, scale, threshold, blur radius, blur std, blending ratio)	(0.1, 0.3, 3.0, 0.5, 10.0, 4.0, 0.8)	(0.2, 0.3, 2, 0.5, 12, 4, 0.7)	(0.55, 0.3, 4, 0.9, 12, 8, 0.7)
Motion	(radius, sigma)	(15, 5)	(15, 12)	(20, 15)
Quant	bit number	5	4	3
Crash	number of dropped camera	2	4	5
Frame	probability of frame dropping	2/6	4/6	5/6

modeling by merging extensive historical data. However, the resilience of these sophisticated temporal models under corrupted conditions remains a territory largely uncharted.

3.3 Camera-LiDAR Fusion

The BEV paradigm streamlines the fusion of features mined from a variety of input modalities. While some algorithms focus on crafting BEV representations solely from images, a notable fraction of the literature, including works like [56], [57], [59], [71], [72], [73], advocates for a unified BEV space. This harmonizes features extracted from both images and point clouds. We delve deep into the performance of such multi-modal fusion algorithms, especially under circumstances where images face corruption, yet the LiDAR mechanism remains pristine. Furthermore, we address a common scenario where the model is trained using multi-modal input but deployed on vehicles equipped with only one of the sensors. To assess robustness, we evaluate the model's performance under conditions of complete sensor failure, where either the camera or LiDAR is missing.

3.4 BEV View Transformation

The body of work in BEV transformation is bifurcated based on depth estimation techniques. One faction, as exemplified by [3], [16], [41], [55], embeds a distinct depth-estimation branch within their systems. Given the inherent challenges in predicting 3D bounding boxes from singular images, these models first forecast a per-pixel depth map. This map then serves as a compass, guiding image features to their rightful 3D coordinates. The subsequent BEV transformation process often follows a bottom-up trajectory, as depicted in [4]. On the other side of the spectrum are models leveraging pre-ordained object queries [2], [4] or lean proposals [44], [45] to collate 2D features in a top-down manner. While both these paradigms have demonstrated their prowess on pristine datasets, we expand the horizon by examining their efficacy on data that deviates from the norm.

4 Benchmark Design

4.1 Dataset Generation

Emerging as our cornerstone is the *nuScenes-C* benchmark dataset, curated by introducing corruptions to the validation set of the renowned nuScenes dataset [1]. Given nuScenes' widespread application in modern BEV models, it stands as a fitting choice. Encompassing a vast expanse of eight distinct corruptions, our dataset mirrors challenges posed by external environmental elements, sensor-induced distortions, and our innovative temporal corruptions.

Mirroring the structure set by [40], we tier each corruption type across three intensities: easy, moderate, and hard. Striking a judicious balance, these severity levels ensure that while challenges are present, they do not entirely obliterate performance, thereby maintaining the relevance and integrity of our findings. Moreover, we infuse variability within each severity tier, bolstering the diversity of the dataset. Comprehensively, our benchmark consists of a staggering 866736 images, each with a resolution of 1600 $\times$ 900 pixels.

We also factor in scenarios simulating complete sensor blackouts in our camera-LiDAR fusion algorithms. While simulating the camera's absence, every pixel in the multi-view camera input is nullified. To emulate the lack of LiDAR readings, only the data points within a $[-45, 45]$ degree frontal field of view are retained, jettisoning the rest. Such a design choice is rooted in our observations that multi-modal trained models crumble when LiDAR readings are entirely absent.

TABLE 2
BEV model calibration. Pretrain: model initialized from pretrained FCOS3D [6] checkpoint; Temporal: model utilizes temporal information; Depth: model with explicit depth estimation branch used in the pipeline; CBGS: model uses the class-balanced group-sampling training strategy [36]. Bold: Best in the category. Underline: Second best in the category.

Model	Pretrain	Temporal	Depth	CBGS	Backbone	BEV Encoder	Image Size	NDS ↑	mCE (%) ↓	mRR (%) ↑
DETR3D [4]	✓				ResNet	Attention	1600 × 900	0.4224	100.00	70.77
$\text{DETR3D}_{\text{CBGS}}$ [4]	✓			✓	ResNet	Attention	1600 × 900	0.4341	99.21	70.02
BEVFormer (small) [2]	✓	✓			ResNet	Attention	1280 × 720	0.4787	101.23	59.07
BEVFormer-S (small) [2]	✓				ResNet	Attention	1280 × 720	0.2622	114.43	**76.87**
BEVFormer (base) [2]	✓	✓			ResNet	Attention	1600 × 900	**0.5174**	97.97	60.40
BEVFormer-S (base) [2]	✓				ResNet	Attention	1600 × 900	0.4129	101.87	69.33
PETR (r50) [5]					ResNet	Attention	1408 × 512	0.3665	111.01	61.26
PETR (vov) [5]	✓				VoVNet-V2	Attention	1600 × 640	0.4550	100.69	65.03
ORA3D [42]	✓				ResNet	Attention	1600 × 900	0.4436	99.17	68.63
PolarFormer (r101) [43]	✓				ResNet	Attention	1600 × 900	0.4602	**96.06**	70.88
PolarFormer (vov) [43]	✓				VoVNet-V2	Attention	1600 × 900	0.4558	98.75	67.51
SRCN3D (r101) [44]	✓				ResNet	CNN + Attn.	1600 × 900	0.4286	**99.67**	**70.23**
SRCN3D (vov) [44]	✓				VoVNet-V2	CNN + Attn.	1600 × 900	0.4205	102.04	67.95
Sparse4D (r101) [45]	✓	✓			ResNet	CNN + Attn.	1600 × 640	**0.5438**	100.01	55.04
BEVDet (r50) [3]			✓	✓	ResNet	CNN	704 × 256	0.3770	115.12	51.83
BEVDet (r101) [3]			✓	✓	ResNet	CNN	704 × 256	0.3877	113.68	53.12
BEVDet (r101) [3]	✓		✓	✓	ResNet	CNN	704 × 256	0.3780	112.80	56.35
BEVDet (tiny) [3]			✓	✓	SwinTrans	CNN	704 × 256	0.4037	116.48	46.26
BEVDepth (r50) [16]			✓	✓	ResNet	CNN	704 × 256	0.4058	110.02	56.82
BEVerse (swin-t) [41]		✓	✓	✓	SwinTrans	CNN	704 × 256	0.4665	110.67	48.60
BEVerse-S (swin-t) [41]			✓	✓	SwinTrans	CNN	704 × 256	0.1603	137.25	28.24
BEVerse (swin-s) [41]		✓	✓	✓	SwinTrans	CNN	1408 × 512	0.4951	117.82	49.57
BEVerse-S (swin-s) [41]			✓	✓	SwinTrans	CNN	1408 × 512	0.2682	132.13	29.54
SOLOFusion (short) [55]		✓	✓		ResNet	CNN	704 × 256	0.3907	108.68	61.45
SOLOFusion (long) [55]		✓	✓		ResNet	CNN	704 × 256	0.4850	97.99	64.42
SOLOFusion (fusion) [55]		✓	✓	✓	ResNet	CNN	704 × 256	**0.5381**	**92.86**	**64.53**

4.2 Natural Corruptions

A visual guide to our corruption taxonomy is presented in Figure 1. Broadly, we focus on three corruption categories. First, those induced by external environmental dynamics, such as varying illumination or meteorological extremes, are simulated via *Brightness*, *Dark*, *Fog*, and *Snow*. Considering the bulk of training data is captured under relatively benign conditions, testing models under these extremes is crucial.

Secondly, sensor-driven distortions can corrupt collected imagery. High-speed motion may induce blur, or memory conservation tactics might compel image quantization. To mirror these real-world challenges, we have integrated *Motion Blur* and *Color Quantization*.

Lastly, we tread into uncharted territories by simulating camera malfunctions, where entire image sets or random frames are omitted due to hardware issues. This is captured by our novel *Camera Crash* and *Frame Lost* corruptions. The intricacies of these processes are visually broken down in Figure 1. We visualize the pixel histogram analysis on our synthesized images, as shown in Figure 2. A notable observation was that the *Motion Blur* corruption, while inducing minimal pixel distribution shifts, still caused a significant performance dip. Additional experimental findings and results are discussed in detail in Section 5.

4.3 Robustness Metrics

We follow the official nuScenes metric [1] to calculate robustness metrics on the *nuScenes-C* dataset. We report nuScenes Detection Score (NDS) and mean Average Precision (mAP), along with mean Average Translation Error (mATE), mean Average Scale Error (mASE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE) and mean Average Attribute Error (mAAE).

To better compare the robustness among different BEV detectors, we introduce two new metrics inspired by [40] based on NDS. The first metric is the mean corruption error (mCE), which is applied to measure the *relative robustness* of candidate models compared to the baseline model:

$$\text{CE}_i = \frac{\sum_{l=1}^{3}(1-\text{NDS})_{i,l}}{\sum_{l=1}^{3}(1-\text{NDS}_{i,l}^{\text{baseline}})}~, ~~\text{mCE} = \frac{1}{N}\sum_{i=1}^{N}\text{CE}_i~, \tag{1}$$

where i denotes the corruption type and l is the severity level; N denotes the number of corruption types in our benchmark. To compare the *performance discrepancy* between *nuScenes-C* and the standard nuScenes dataset, we define a simple mean resilience rate (mRR) metric, which is calculated across three severity levels as follows:

$$\text{RR}_i = \frac{\sum_{l=1}^{3}\text{NDS}_{i,l}}{3\times\text{NDS}_{\text{clean}}}~, ~~\text{mRR} = \frac{1}{N}\sum_{i=1}^{N}\text{RR}_i~. \tag{2}$$

In our benchmark, we report both metrics for each candidate model and base our analyses on these.

5 Benchmark Experiments

5.1 Experimental Setup

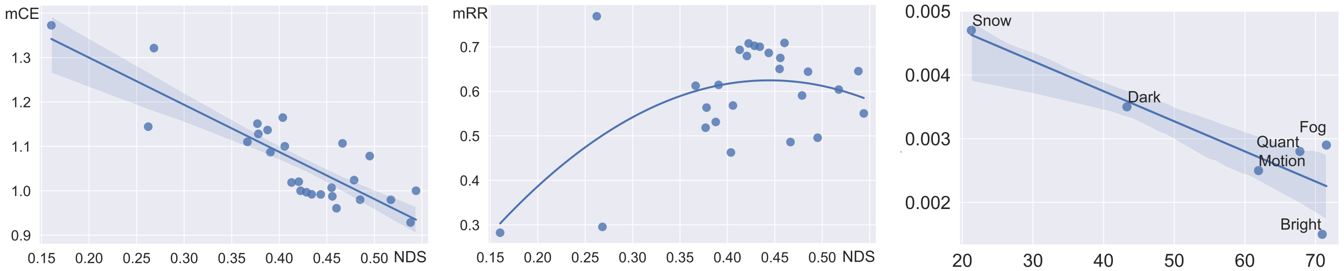
In our study, we use the official model configurations and public checkpoints provided by open-sourced codebases, whenever applicable; we also train additional model variants with minimal modifications to conduct experiments under controlled settings. To facilitate access to all model checkpoints and configurations, we have compiled a "model zoo", which can be accessed through our repository[1].

[1] Model zoo is publicly accessible at: https://github.com/Daniel-xsy/RoboBEV/blob/master/zoo/README.md.

TABLE 3

The Corruption Error (CE) of each BEV detector in our *RoboBEV* benchmark. Bold: Best in the category. Blue : Best in the row if improve upon baseline. Yellow : Worst in the row if decline upon baseline. †: distinguish pre-training version BEVDet.

Model	NDS ↑	mCE (%) ↓	Camera	Frame	Quant	Motion	Bright	Dark	Fog	Snow
DETR3D [4]	0.4224	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00
DETR3D$_{\text{CBGS}}$ [4]	0.4341	99.21	98.15	98.90	99.15	101.62	97.47	**100.28**	98.23	99.85
BEVFormer (small) [2]	0.4787	102.40	101.23	101.96	98.56	101.24	104.35	105.17	105.40	101.29
BEVFormer (base) [2]	**0.5174**	97.97	**95.87**	**94.42**	**95.13**	99.54	96.97	103.76	97.42	100.69
PETR (r50) [5]	0.3665	111.01	107.55	105.92	110.33	104.93	119.36	116.84	117.02	106.13
PETR (vov) [5]	0.4550	100.69	99.09	97.46	103.06	102.33	102.40	106.67	103.43	**91.11**
ORA3D [42]	0.4436	99.17	97.26	98.03	97.32	100.19	98.78	102.40	99.23	100.19
PolarFormer (r101) [43]	0.4602	**96.06**	96.16	97.24	**95.13**	**92.37**	**94.96**	103.22	**94.25**	95.17
PolarFormer (vov) [43]	0.4558	98.75	96.13	97.20	101.48	104.32	95.37	104.78	97.55	93.14
SRCN3D (r101) [44]	0.4286	**99.67**	**98.77**	**98.96**	**97.93**	**100.71**	**98.80**	**102.72**	**99.54**	99.91
SRCN3D (vov) [44]	0.4205	102.04	99.78	100.34	105.13	107.06	101.93	107.10	102.27	**92.75**
Sparse4D (r101) [45]	**0.5438**	100.01	99.80	99.91	98.05	102.00	100.30	103.83	100.46	95.72
BEVDet (r50) [3]	0.3770	115.12	105.22	109.19	111.27	108.18	123.96	123.34	123.83	115.93
BEVDet (tiny) [3]	0.4037	116.48	103.50	106.61	113.18	107.26	130.19	131.83	124.01	115.25
BEVDet (r101) [3]	0.3877	113.68	103.32	107.29	109.25	105.40	124.14	123.12	123.28	113.64
BEVDet (r101†) [3]	0.3780	112.80	105.84	108.68	101.99	100.97	123.39	119.31	130.21	112.04
BEVDepth (r50) [16]	0.4058	110.02	103.09	106.26	106.24	102.02	118.72	114.26	116.57	112.98
BEVerse (swin-t) [41]	0.4665	110.67	95.49	94.15	108.46	100.19	122.44	130.40	118.58	115.69
BEVerse (swin-s) [41]	0.4951	107.82	92.93	101.61	105.42	100.40	110.14	123.12	117.46	111.48
SOLOFusion (short) [55]	0.3907	108.68	104.45	105.53	105.47	100.79	117.27	110.44	115.01	110.47
SOLOFusion (long) [55]	0.4850	97.99	95.80	101.54	93.83	89.11	100.00	**99.61**	98.70	**105.35**
SOLOFusion (fusion) [55]	**0.5381**	**92.86**	**86.74**	**88.37**	**87.09**	**86.63**	**94.55**	102.22	**90.67**	106.64



(a) mCE *vs.* NDS

(b) mRR *vs.* NDS

(c) Depth estimation error *vs.* RR.

Fig. 3. (a): The mCE metric shows a linear relationship with "clean" performance while (b): the mRR metric confronts the risk of decreasing. (c): We observe strong correlations where large depth estimation errors under *Snow* and *Dark* tend to cause drastic performance drops.

We re-implemented several models, including BEVDet (r101) [3], PolarFormer (vov) [43], and SRCN3D (vov) [44], tailored to our investigative requirements. For BEVDet (r101), in the pursuit of fairness, we chose to preserve the input resolution consistent with BEVDet (r50). This decision, while producing results slightly lower than the official documentation [3], was a deliberate effort to emphasize our study's robustness metrics over purely optimizing for performance on the nuScenes dataset [1]. For the original versions of both PolarFormer (vov) and SRCN3D (vov), the models were initialized using checkpoints from DD3D [98], which had previously trained on the nuScenes *trainval* set. However, this method inadvertently caused information leakage, considering the *nuScenes-C* dataset originates from the nuScenes validation set. To mitigate this and ensure fair comparisons, we re-implemented the two models, initiating them via the FCOS3D [6] model, without further alterations. Specifically, the VoVNet-V2 iterations [53] of the FCOS3D models were first trained for depth estimation on the DDAD15M dataset [54] and then underwent fine-tuning on the nuScenes training set.

Furthermore, for a comprehensive overview, metrics for each corruption type were deduced by averaging results across all three severity levels. In our study, DETR3D [4] was designated as the baseline for the mCE metric. Our research methodology and the corresponding code were constructed atop the MMDetection3D codebase [31].

5.2 Camera Only Benchmarking Results

We undertook an exhaustive benchmark analysis of 30 contemporary BEV models on the *nuScenes-C* dataset. The primary outcomes of our investigations are encapsulated in Table 2. Our analysis revealed that all models manifest a decline in performance across the corrupted dataset.

5.2.1 3D Object Detection

A notable trend emerges when examining the absolute performances on both the nuScenes-C and its "clean" counterpart. Specifically, BEV detectors exhibiting proficiency on the standard dataset also tend to showcase commendable performance when faced with out-of-distribution datasets, a trend visually represented in Figure 3a. Nevertheless, delving deeper into these outcomes brings forth a layered narrative. Detectors, despite parallel performance on the "clean" dataset, display varied robustness when confronted with diverse corruption types. To illustrate, while BEVerse

TABLE 4
Benchmark results of perception tasks including Map Segmentation (MS), Depth Estimation (DE), and Semantic Occupancy Prediction (SOP).

Tasks	Model	Metric	Clean	Camera	Frame	Quant	Motion	Bright	Dark	Fog	Snow
MS	CVT [18]	IoU ↑	0.348	0.200	0.170	0.294	0.281	0.275	0.200	0.247	0.177
DE	SurroundDepth [60]	Abs Rel ↓	0.280	0.485	0.497	0.334	0.338	0.339	0.354	0.320	0.423
SOP	TPVFormer [62]	mIoU ↑	0.521	0.274	0.229	0.381	0.386	0.490	0.373	0.466	0.193
SOP	SurroundOcc [61]	SC IoU ↑	0.314	0.199	0.181	0.258	0.225	0.307	0.248	0.296	0.183

TABLE 5
NDS results of fusion model under different input modalities. Since *Fog* and *Snow* can also affect the LiDAR sensors, we do not consider these two corruptions of the fusion model.

Model	Camera	LiDAR	Clean	Camera	Frame	Quant	Motion	Bright	Dark	Fog	Snow
BEVFusion [56]	✓		0.4121	0.2777	0.2255	0.2763	0.2788	0.2902	0.1076	0.3041	0.1461
BEVFusion [56]		✓	0.6928	–	–	–	–	–	–	–	–
BEVFusion [56]	✓	✓	**0.7138**	**0.6963**	**0.6931**	**0.7044**	**0.6977**	**0.7018**	**0.6787**	–	–
TransFusion [58]	✓	✓	0.6887	0.6843	0.6447	0.6819	0.6749	0.6843	0.6663	–	–
AutoAlignV2 [59]	✓	✓	0.6139	0.5849	0.5832	0.6006	0.5901	0.6076	0.5770	–	–

TABLE 6
Benchmark results for complete sensor failure. The models are trained using multi-modal input while tested using single sensor input.

Model	Train	Camera	LiDAR	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFusion [56]	C	✓		0.4122	0.3556	0.6677	0.2727	0.5612	0.8954	0.2593
BEVFusion [56]	L		✓	0.6927	0.6468	0.2912	0.2530	0.3142	0.2627	0.1858
BEVFusion [56]	C+L	✓	✓	0.7138	0.6852	0.2874	0.2539	0.3044	0.2554	0.1874
BEVFusion [56]	C+L	✓		0.3340 (↓ 0.3798)	0.0789 (↓ 0.6063)	0.5044	0.3073	0.4999	0.5098	0.2338
BEVFusion [56]	C+L		✓	0.6802 (↓ 0.0605)	0.6247(↓ 0.0605)	0.2948	0.2590	0.3137	0.2697	0.1844
TransFusion [58]	C+L	✓	✓	0.6887	0.6453	0.2995	0.2552	0.3209	0.2765	0.1877
TransFusion [58]	C+L	✓		0.3470 (↓ 0.3417)	0.0343 (↓ 0.6110)	0.4087	0.3091	0.4446	0.3104	0.2282
TransFusion [58]	C+L		✓	0.6464 (↓ 0.0423)	0.5764 (↓ 0.0689)	0.3171	0.2761	0.3227	0.3124	0.1897
AutoAlignV2 [59]	C+L	✓	✓	0.6139	0.5649	0.3300	0.2699	0.4226	0.4644	0.1983
AutoAlignV2 [59]	C+L		✓	0.5651 (↓ 0.0448)	0.4794 (↓ 0.0855)	0.3463	0.2734	0.4361	0.4894	0.2007

(swin-s) [41] manifests heightened resilience during a *Camera Crash*, PETR (vov) [5] excels under *Snow* conditions. Yet, both falter significantly under *Dark* settings.

Our investigations further highlight a potential vulnerability in resilience rates across various corruptions. Even though the mCE metric displays a linear correlation between the nuScenes and *nuScenes-C* datasets, the mRR metric elucidates notable disparities among models with comparable baseline performance. This suggests potential overfitting of some models to the nuScenes dataset, thereby compromising their adaptability to the *nuScenes-C* dataset. For instance, despite Sparse4D [45] outpacing DETR3D [4] on the clean" dataset, it falls short in terms of mRR metrics across all corruption categories. Moreover, DETR3D's superior performance under *Dark* conditions contrasts starkly with BEVerse (swin-t), which, despite a better clean" performance, registers a relative performance of merely 12% under similar settings. Hence, it is evident that a multi-faceted assessment of cutting-edge models is imperative for a holistic evaluation of their capabilities.

To gain deeper insights into model robustness, we dissected BEV algorithms based on components like training strategies (*e.g.*, pre-training and CBGS [36] resampling strategy), model architectures (*e.g.*, backbone), and learning techniques (*e.g.*, temporal cue learning). The consequent results are detailed in Table 2.

5.2.2 *Other Perception Tasks*

Our inquiry also extended to associated tasks, including BEV-centric map segmentation, depth estimation, and occupancy prediction, with outcomes presented in Table 4. Adhering to setting 1 from [18], we reported the Intersection over Union (IoU) for vehicle map-view segmentation results. For depth estimation, we employed the Absolute Relative Difference (Abs Rel) score, and for semantic occupancy prediction, we used the mean Intersection over Union (mIoU). For comprehensive metric definitions, readers can consult the original publications [18], [60], [61], [62]. These results, spanning diverse perception tasks, offer an enriched perspective on BEV model capabilities and constraints.

It is worth noting that the performance of numerous BEV-centric perception models takes a hit under specific corruptions like *Dark* and *Snow*. This exposes a prevalent susceptibility across BEV models to such corruptions, compromising their reliability in real-world scenarios.

5.3 Camera-LiDAR Fusion

5.3.1 *Camera Sensor Corruption*

We studied scenarios where cameras are impaired while LiDAR operates optimally, a frequent occurrence in real-world conditions. For instance, LiDAR point cloud capture remains largely unhampered by lighting variations, whereas

TABLE 7
The Resilience Rate (RR) of each BEV detector in our RoboBEV benchmark. Bold: best within the category. Blue : Best across category. †: distinguish pre-training version BEVDet.

Model	NDS	mRR (%) ↑	Camera	Frame	Quant	Motion	Bright	Dark	Fog	Snow
DETR3D [4]	0.4224	70.77	67.68	61.65	75.21	63.00	94.74	**65.96**	**92.61**	45.29
DETR3D$_{\text{CBGS}}$ [4]	0.4341	70.02	**68.90**	61.85	74.52	58.56	**95.69**	63.72	**92.61**	44.34
BEVFormer (small) [2]	0.4787	59.07	57.89	51.37	68.41	53.69	78.15	50.41	74.85	37.79
BEVFormer (base) [2]	**0.5174**	60.40	60.96	58.31	67.82	52.09	80.87	48.61	78.64	35.89
PETR (r50) [5]	0.3665	61.26	63.30	59.10	67.45	62.73	77.52	42.86	78.47	38.66
PETR (vov) [5]	0.4550	65.03	64.26	61.36	65.23	54.73	84.79	50.66	81.38	**57.85**
ORA3D [42]	0.4436	68.63	68.87	**61.99**	75.74	59.67	91.86	58.90	89.25	42.79
PolarFormer (r101) [43]	0.4602	**70.88**	68.08	61.02	**76.25**	**69.99**	93.52	55.50	**92.61**	50.07
PolarFormer (vov) [43]	0.4558	67.51	68.78	61.67	67.49	51.43	93.90	53.55	89.10	54.15
SRCN3D (r101) [44]	0.4286	**70.23**	**68.76**	**62.55**	**77.41**	**60.87**	**95.05**	**60.43**	**91.93**	44.80
SRCN3D (vov) [44]	0.4205	67.95	68.37	61.33	67.23	50.96	92.41	54.08	89.75	**59.43**
Sparse4D(r101) [45]	**0.5438**	55.04	52.83	48.01	60.87	46.23	73.26	46.16	71.42	41.54
BEVDet (r50) [3]	0.3770	51.83	65.94	51.03	63.87	54.67	68.04	29.23	65.28	16.58
BEVDet (tiny) [3]	0.4037	46.26	64.63	52.39	56.43	52.71	54.27	12.14	60.69	16.84
BEVDet (r101) [3]	0.3877	53.12	67.63	53.26	65.67	58.42	65.88	28.84	64.35	20.89
BEVDet (r101†) [3]	0.3780	56.35	64.60	51.90	**80.45**	68.52	68.76	36.85	54.84	24.84
BEVDepth (r50) [16]	0.4058	56.82	65.01	52.76	67.79	61.93	70.95	43.30	71.54	21.27
BEVerse (swin-t) [41]	0.4665	48.60	68.19	**65.10**	55.73	56.74	56.93	12.71	59.61	13.80
BEVerse (swin-s) [41]	0.4951	49.57	67.95	50.19	56.70	53.16	68.55	22.58	57.54	19.89
SOLOFusion (short) [55]	0.3907	61.45	65.04	56.18	71.77	66.62	75.92	52.03	76.73	27.28
SOLOFusion (long) [55]	0.4850	64.42	65.13	51.34	74.19	**71.34**	**82.52**	**58.02**	82.29	**30.52**
SOLOFusion (fusion) [55]	**0.5381**	**64.53**	**70.73**	64.37	75.41	67.68	80.45	48.80	**83.26**	25.57

camera captures can degrade under limited light. Intentionally, we excluded conditions like *Snow* and *Fog*, as they could introduce noise to both camera and LiDAR readings. Results of these studies are depicted in Table 5.

Interestingly, multi-modal fusion models maintain high performance even when the camera data is compromised. When provided with pristine LiDAR and degraded camera inputs, BEVFusion [56] consistently outperforms its LiDAR-only counterpart, with a notably higher NDS score of 0.6928, across most types of camera corruptions, except *Dark*. This affirms the efficacy of using LiDAR data even when the camera data is suboptimal.

However, there are circumstances where corrupted camera inputs adversely affect the model's performance. For example, under conditions such as *Camera Crash* and *Motion Blur*, the benefits of incorporating camera features into the model are marginal. Moreover, in the presence of *Dark* corruption, corrupted camera features not only fail to provide useful information but also diminish the efficacy of LiDAR features, leading to a performance drop from an NDS score of 0.6928 to 0.6787. As a result, enhancing the robustness of multi-modal fusion models against input corruption emerges as a crucial avenue for future research.

5.3.2 *Complete Sensor Failure*

Multi-modal fusion models are typically trained using data from both camera and LiDAR sensors. However, the deployed model must function adequately even if one of these sensors fails. We evaluate the performance of our multi-modal model using input from only a single modality, with results presented in Table 6. When simulating camera failure, all pixel values are set to zero. For LiDAR sensor failure, we discovered that no model could perform adequately when all point data are absent (*i.e.*, the NDS falls to zero). Hence, we retain only the points within a $[-45, 45]$ degree range in front of the vehicle and discard all others.

Interestingly, our findings indicate that multi-modal models are disproportionately reliant on LiDAR input. In scenarios where LiDAR data is missing, the mAP metrics for BEVFusion [56] and Transfusion [58] drop by 89% and 95%, respectively. In contrast, the absence of image data leads to a much milder decline in performance. This phenomenon underscores that, during the training phase, point cloud features may disproportionately influence the model, thereby asserting dominance over image-based features in perception tasks.

Such a dependence on LiDAR data introduces a significant vulnerability to multi-modal perception models, particularly because LiDAR sensors are prone to data corruption under adverse weather conditions such as rain, snow, and fog. These observations necessitate further research focused on enhancing the robustness of multi-modal perception systems, especially when one sensory modality is entirely absent.

5.4 Validity Assessment

Since the corruption images are synthesized digitally, it is important to study how close they are compared to real-world corruption. To study the validity of synthesized images, we conducted two experiments, including the pixel distribution study and corruption-augmented training.

5.4.1 *Pixel Distribution*

Assuming that a corruption simulation is realistic enough to reflect real-world situations, the distribution of a corrupted ”clean” set should be similar to that of the real-world corruption set. We validate this using ACDC [99], nuScenes [1], Cityscapes [101], and Foggy-Cityscapes [100], since these datasets contain real-world corruption data and clean data collected by the same sensor types from the same physical locations. We simulate corruptions using ”clean” images and compare the distribution patterns with their corresponding real-world corrupted data. We do this to ensure that there is no extra distribution shift from aspects like sensor difference (*e.g.* FOVs and resolutions) and location discrepancy (*e.g.* environmental and semantic changes).

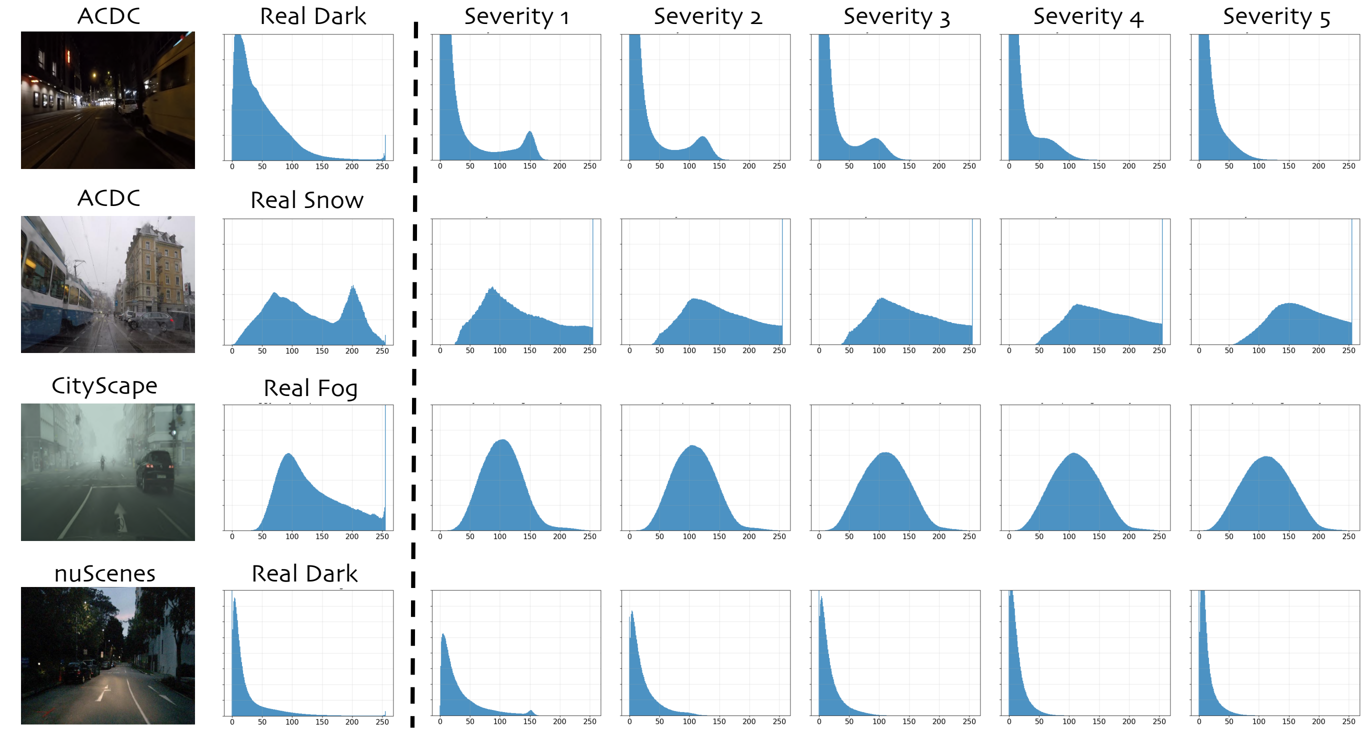


Fig. 4. The left two columns illustrate the appearances and pixel distribution patterns observed in genuine real-world corrupted data. The right five columns represent corresponding synthetic corruptions under different severity levels, which are of high fidelity compared to real-world data.

TABLE 8
Validity study for using corruption augmentation to improve model cross-domain robustness.

Model	Training Set	Testing Set	Corrupt Aug	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
DETR [4]	nuScenes train	nuScenes val		0.4224	0.3468	0.7647	0.2678	0.3917	0.8754	0.2108
DETR [4]	nuScenes train	nuScenes val	✓	$\mathbf{0.4242}$ ($\uparrow 0.0018$)	0.3511	0.7655	0.2736	0.4130	0.8487	0.2119
FCOS3D [6]	day train	day val		0.3867	0.3045	0.7651	0.2576	0.5001	1.2102	0.1321
FCOS3D [6]	day train	day val	✓	$\mathbf{0.3883}$ ($\uparrow 0.0196$)	0.3073	0.7630	0.2581	0.5043	1.1782	0.1286
FCOS3D [6]	day train	night val		0.0854	0.0162	1.0434	0.6431	0.8241	1.8505	0.7597
FCOS3D [6]	day train	night val	✓	$\mathbf{0.1245}$ ($\uparrow 0.0391$)	0.0265	1.0419	0.4658	0.8145	2.2727	0.6067
FCOS3D [6]	dry train	dry val		0.3846	0.2970	0.7744	0.2541	0.4721	1.3199	0.1380
FCOS3D [6]	dry train	dry val	✓	$\mathbf{0.3854}$	0.2992	0.7654	0.2582	0.4824	1.3334	0.1361
FCOS3D [6]	dry train	rain val		0.3203	0.2151	0.8994	0.2856	0.5253	1.7129	0.1619
FCOS3D [6]	dry train	rain val	✓	$\mathbf{0.3302}$ ($\uparrow 0.0099$)	0.2266	0.8595	0.2719	0.5559	1.5697	0.1439

As illustrated in Figure 4, the pixel distributions of our synthesized images exhibit a high degree of resemblance to those of real-world data, thereby affirming the dataset's validity from a pixel statistical perspective.

5.4.2 *Corruption-augmented Training*

Assuming that a corruption simulation is realistic enough to reflect real-world situations, a corruption-augmented model should achieve better generalizability than the "clean" model when tested on real-world corruption datasets. Also, the corruption-augmented model should also show better performance on the clean dataset. We validate this using nuScenes, nuScenes-Night, and nuScenes-Rain. We adopt FCOS3D as the baseline and train the model with corruption augmentation. For nuScenes-Night and nuScenes-Rain, we train the model on Day-train and Dry-train split and evaluate on Day-val, Night-val, Dry-val, and Rain-val split. The results can be seen in Table 8. We observe that using synthesized images as the data augmentation strategy successfully improves the cross-domain robustness. Specifically, in day-to-night domain transitions, we observe a significant performance drop from 0.3867 to 0.0854 in the baseline model due to the large domain gap. However, when trained with corruption augmentation, the model's cross-domain performance improves by 45.8%, thereby validating the validity of our synthesized images.

6 Analysis and Discussion

6.1 Depth Estimation

- *Depth-free BEV transformations show better robustness.* Our analysis reveals that depth-based approaches suffer from severe performance degradation when exposed to corrupted images as shown in Figure 6c and 6d. Moreover, we undertake a comparative study to evaluate the intermediate depth estimation results of BEVDepth [16] under corruptions. To this end, we compute the mean square error (MSE) between "clean" inputs and corrupted inputs. Our findings

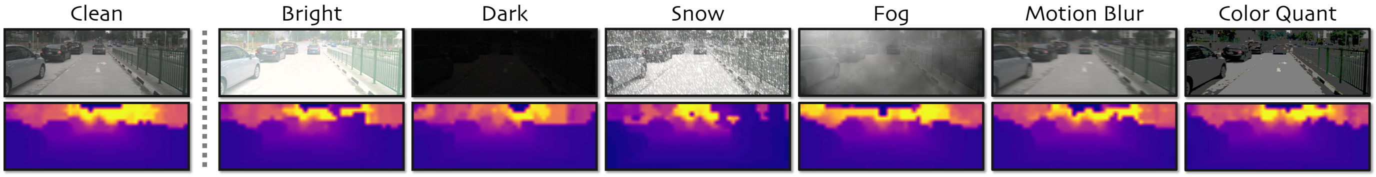
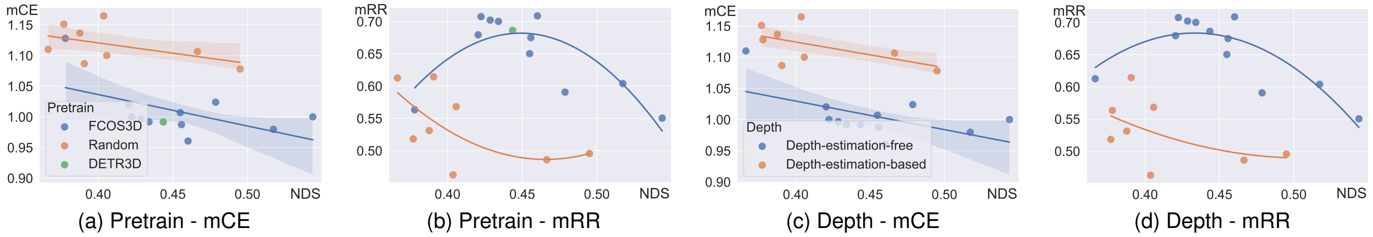


Fig. 5. Depth estimation results of BEVDepth [16] under different corruption types. The results exhibit a different sensitivity for each scenario.



(a) Pretrain - mCE (b) Pretrain - mRR (c) Depth - mCE (d) Depth - mRR

Fig. 6. Pre-training strategies together with depth-free bird's eye view transformation provide the models with better robustness. We do not consider SOLOFusion [55] long-term fusion here since it utilizes 16 frames, which is much larger than other methods.

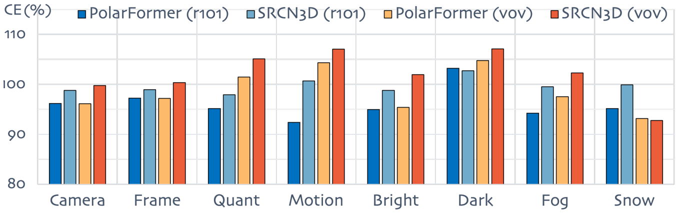


Fig. 7. ResNet *vs.* VoVNet-V2. Since the two versions have similar "clean" performances, we compare the absolute corruption error (the lower the better).

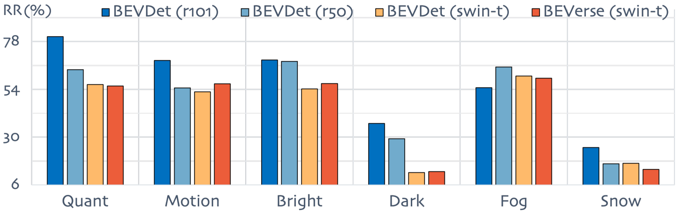


Fig. 8. ResNet *vs.* SwinTransformer. Since the models have different "clean" performances, we compare the relative resilience rate (the higher the better).

indicate an explicit correlation between vulnerability and depth estimation error, as presented in Figure 3c. Specifically, *Snow* and *Dark* corruptions significantly affect accurate depth estimation, leading to the largest performance drop. These results provide further support for our conclusion that the performance of depth-based approaches can suffer significantly if the depth estimation is not accurate enough. The depth estimation results under corruptions can be seen in Figure 5.

6.2 Pre-Training

- *Pre-training improves robustness across a wide range of semantic corruptions while does not help with temporal corruptions.* The effectiveness of these strategies for improving model robustness is illustrated in Figure 6a and Figure 6b, where models that utilize pre-training largely outperform those not. For

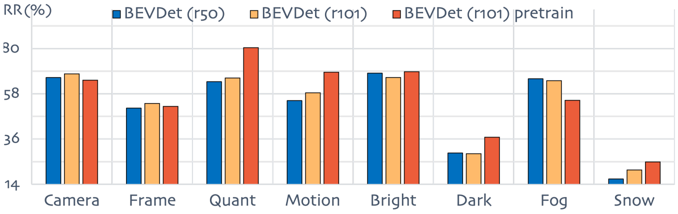
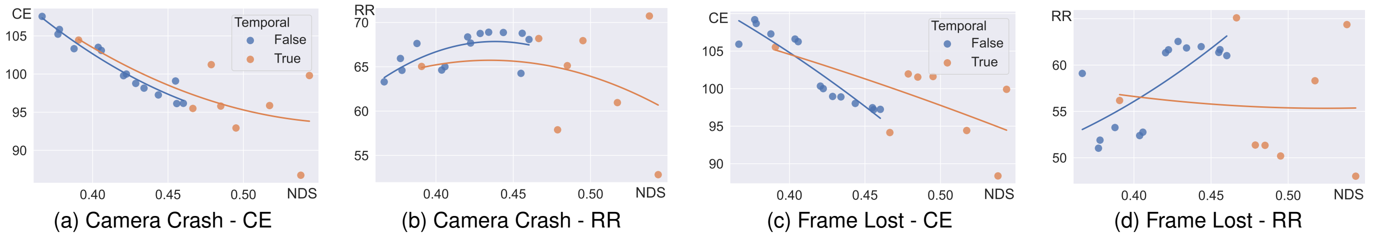


Fig. 9. Resilience rate comparisons of BEVDet [3] with and without pre-training. The higher the better.

controlled comparison, we re-implement the BEVDet (r101) model using the FCOS3D checkpoint as initialization. Our results, presented in Figure 9, show that pre-training can significantly improve mRR across a wide range of corruptions (except *Fog*) even if it has lower "clean" NDS (0.3780 *vs.* 0.3877). Specifically, under *Color Quant*, *Motion Blur*, and *Dark* corruptions, the mRR metric improves by 22.5%, 17.2%, and 27.8%, respectively. It is worth noting that pre-training mainly improves most semantic corruptions and does not improve temporal corruptions. Even though, the pre-trained BEVDet still largely lags behind those depth-free counterparts. Therefore, we can conclude that pre-training together with the depth-free bird's eye view transformation provides models with strong robustness.

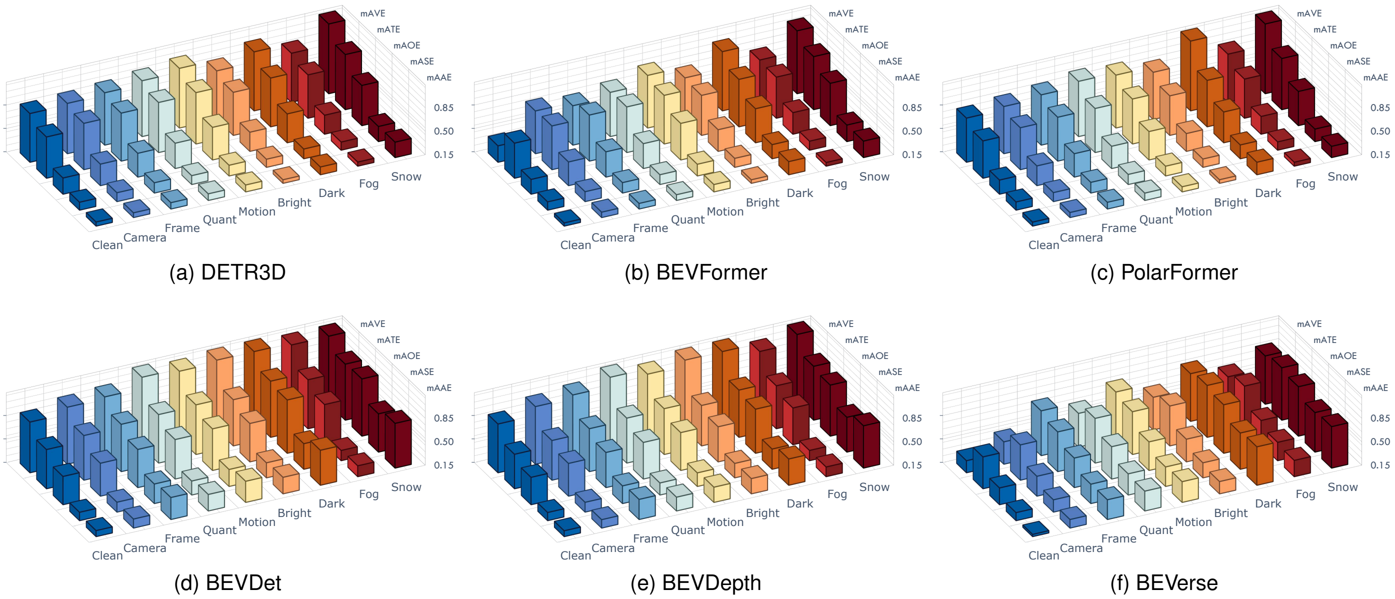
6.3 Temporal Fusion

- *Temporal fusion has the potential to yield better absolute performance under corruptions. Fusing longer temporal information largely helps with robustness.* We are particularly interested in examining how models utilizing temporal information perform under temporal corruptions. We find SOLOFusion [55] which fuses wider and richer temporal information performs extremely well compared to its short-only and long-only versions. In terms of *Camera Crash*, the short-only and long-only versions have close resilience rate performance (65.04 *vs.* 65.13). However, the fusion version improves to 70.73, which is the highest among all the candidate models. Similarly, the fusion version improves the resilience rate



(a) Camera Crash - CE (b) Camera Crash - RR (c) Frame Lost - CE (d) Frame Lost - RR

Fig. 10. Not all the models with temporal fusion exhibit better robustness under *Camera Crash* and *Frame Lost*. However, they have the potential since the lowest mCE metric models are always those that utilize temporal information.



(a) DETR3D (b) BEVFormer (c) PolarFormer

(d) BEVDet (e) BEVDepth (f) BEVerse

Fig. 11. Benchmark results of task-specific metrics reported on *nuScenes-C* other than NDS, under different corruption types in our benchmark.

by almost 10% compared to the other two versions under *Frame Lost* corruption. Moreover, the RR metric of its long-only version outperforms its short-only counterpart on a wide range of corruption types, which indicates the great potential of utilizing longer temporal information.

Surprisingly, we further find that not all models with temporal fusion exhibit better robustness under *Camera Crash* and *Frame Lost*. The robustness is highly correlated to how to fuse history frames and how many frames are used, which emphasizes the importance of evaluating temporal fusion strategies from wider perspectives. The results can be seen in Figure 10. Nonetheless, temporal fusion remains a potential method to enhance temporal robustness since the models with the lowest Corruption Error (or the highest Resilience Rate) are consistently those that utilize temporal information.

6.4 Backbone

- The Swin Transformer is more vulnerable towards the lighting changings; VoVNet-V2 is more robust against Snow while ResNet shows better robustness across a wide range of corruptions. Although ResNet and VoVNet [53] exhibit close standard performance, ResNet-based detectors exhibit consistently superior robustness across a wide range of corruptions, as illustrated in Figure 7.

Conversely, the VoVNet backbone consistently exhibits better robustness under *Snow* corruptions. Moreover, Swin Transformer [35] based BEVDet demonstrates significant vulnerability towards changes in lighting conditions (*e.g.*, *Bright* and *Dark*). A clear comparison can be found in Figure 8.

6.5 Corruption

The relationship between pixel distribution shifts and model performance degradation is not straightforward. We calculate the pixel distribution over 300 images sampled from the nuScenes dataset and visualize the pixel histograms shown in Figure 2. Interestingly, the *Motion Blur* causes the least pixel distribution shifts while causing a relatively large performance drop. On the other hand, *Bright* shifts the pixel distribution to higher values, and *Fog* makes fine-grained features more indistinct by shifting the pixel value more agminated. However, these two corruptions only lead to the smallest performance gap, which reveals that model robustness is not simply correlated with pixel distribution.

6.6 Detailed Metrics

- Velocity prediction errors amplify under corruptions, and attribution and scale errors differ across models. While our study

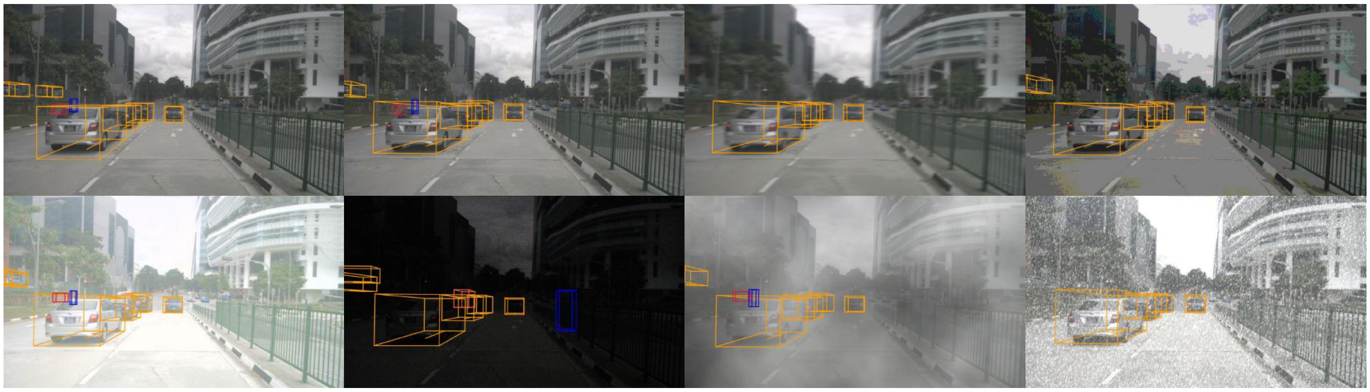


Fig. 12. Visualization detection results of BEVFormer [2]. From left to right: top: GT, *Clean*, *Motion*, *Quant*; bottom: *Bright*, *Dark*, *Fog*, *Snow*.

predominantly reports the nuScenes Detection Score (NDS) metrics, additional insights into model robustness are illustrated in Figure 11. We find that models incorporating temporal information, such as BEVFormer [2] and BEVerse [41], exhibit substantially lower mean Absolute Velocity Error (mAVE) compared to those that do not. Nonetheless, even models with temporal fusion are not immune to the adverse effects of image corruption; specifically, velocity prediction errors markedly escalate even under mild illumination alterations. Figure 11b and 11f illustrates that *Motion Blur* corruption detrimentally influences the velocity predictions for both BEVFormer and BEVerse, revealing a significant vulnerability in these models that incorporate temporal data.

Moreover, a closer examination of attribution and scale errors reveals considerable heterogeneity across models. Depth-free models demonstrate a consistent performance in these metrics, while depth-based models display pronounced variability. This observation underscores the heightened susceptibility of depth-based methods to image corruptions and emphasizes the need for further research to enhance their robustness.

7 Potential Limitation

Despite the eight distinct corruptions we introduce, they still cannot cover all the out-of-distribution contexts in real-world applications due to their unpredictable complexity. Additionally, we mainly analyze coarse-grained designs between models (*e.g.*, depth estimation) since it is considerably non-trivial to identify the trade-off between fine-grained network architecture designs.

8 Conclusion

In this study, we present the *RoboBEV* benchmark, crafted by incorporating a comprehensive set of eight different natural corruptions to form the *nuScenes-C* dataset. This benchmark serves as a rigorous testing ground for evaluating the out-of-distribution robustness of Bird's Eye View (BEV) perception models. Additionally, we extend our analysis to account for sensor failures in multi-modal perception frameworks, offering a more holistic view of model robustness. Through extensive experimentation, we scrutinize various factors influencing the robustness of BEV perception algorithms. Our findings elucidate critical vulnerabilities and strengths across different models and under diverse conditions. By shedding light on these aspects, we aim to furnish the research community with invaluable insights that can guide the development of more robust, future-ready BEV perception models.

References

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 621–11 631.

[2] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European Conference on Computer Vision*, 2022, pp. 1–18.

[3] J. Huang, G. Huang, Z. Zhu, and D. Du, "Bevdet: High-performance multi-camera 3d object detection in bird-eye-view," *arXiv preprint arXiv:2112.11790*, 2021.

[4] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, and J. Solomon, "Detr3d: 3d object detection from multi-view images via 3d-to-2d queries," in *Conference on Robot Learning*. PMLR, 2022, pp. 180–191.

[5] Y. Liu, T. Wang, X. Zhang, and J. Sun, "Petr: Position embedding transformation for multi-view 3d object detection," *arXiv preprint arXiv:2203.05625*, 2022.

[6] T. Wang, X. Zhu, J. Pang, and D. Lin, "Fcos3d: Fully convolutional one-stage monocular 3d object detection," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 913–922.

[7] T. Wang, X. Zhu, J. Pang, and D. Lin, "Probabilistic and geometric depth: Detecting objects in perspective," in *Conference on Robot Learning*. PMLR, 2022, pp. 1475–1485.

[8] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[9] G. Rossolini, F. Nesti, G. D'Amico, S. Nair, A. Biondi, and G. Buttazzo, "On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving," *arXiv preprint arXiv:2201.01850*, 2022.

[10] C. Xie, J. Wang, Z. Zhang, Y. Zhou, L. Xie, and A. Yuille, "Adversarial examples for semantic segmentation and object detection," in *IEEE/CVF International Conference on Computer Vision*, 2017, pp. 1369–1378.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[12] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1765–1773.

[13] J. Tu, M. Ren, S. Manivasagam, M. Liang, B. Yang, R. Du, F. Cheng, and R. Urtasun, "Physically realizable adversarial examples for lidar object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 716–13 725.

[14] Y. Cao, N. Wang, C. Xiao, D. Yang, J. Fang, R. Yang, Q. A. Chen, M. Liu, and B. Li, "Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks," in *IEEE Symposium on Security and Privacy*, 2021, pp. 176–194.

[15] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*, 2020, pp. 213–229.

[16] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, and Z. Li, "Bevdepth: Acquisition of reliable depth for multi-view 3d object detection," *arXiv preprint arXiv:2206.10092*, 2022.

[17] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *European Conference on Computer Vision*, 2020, pp. 194–210.

[18] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 760–13 769.

[19] T. B. Brown, D. Mané, A. Roy, M. Abadi, and J. Gilmer, "Adversarial patch," *arXiv preprint arXiv:1712.09665*, 2017.

[20] X. Liu, H. Yang, Z. Liu, L. Song, H. Li, and Y. Chen, "Dpatch: An adversarial patch attack on object detectors," *arXiv preprint arXiv:1806.02299*, 2018.

[21] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4604–4612.

[22] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.

[23] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[24] L. Kong, J. Ren, L. Pan, and Z. Liu, "Lasermix for semi-supervised lidar semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 705–21 715.

[25] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, and W. Wang, "Clip2scene: Towards label-efficient 3d scene understanding by clip," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 7020–7030.

[26] G. Puy, A. Boulch, and R. Marlet, "Using a waffle iron for automotive point cloud semantic segmentation," *arXiv preprint arXiv:2301.10100*, 2023.

[27] A. Ando, S. Gidaris, A. Bursuc, G. Puy, A. Boulch, and R. Marlet, "Rangevit: Towards vision transformers for 3d semantic segmentation in autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5240–5250.

[28] Y. Liu, R. Chen, X. Li, L. Kong, Y. Yang, Z. Xia, Y. Bai, X. Zhu, Y. Ma, Y. Li, Y. Qiao, and Y. Hou, "Uniseg: A unified multi-modal lidar segmentation network and the openpcseg codebase," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 662–21 673.

[29] R. Chen, Y. Liu, L. Kong, N. Chen, X. Zhu, Y. Ma, T. Liu, and W. Wang, "Towards label-free scene understanding by vision foundation models," in *Advances in Neural Information Processing Systems*, 2023.

[30] Y. Liu, L. Kong, J. Cen, R. Chen, W. Zhang, L. Pan, K. Chen, and Z. Liu, "Segment any point cloud sequences by distilling vision foundation models," in *Advances in Neural Information Processing Systems*, 2023.

[31] M. Contributors, "MMdetection3d: Openmmlab next-generation platform for general 3d object detection," *https://github.com/open-mmlab/mmdetection3d*, 2020.

[32] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.

[33] Y. Ma, T. Wang, X. Bai, H. Yang, Y. Hou, Y. Wang, Y. Qiao, R. Yang, D. Manocha, and X. Zhu, "Vision-centric bev perception: A survey," *arXiv preprint arXiv:2208.02797*, 2022.

[34] J. Huang and G. Huang, "Bevdet4d: Exploit temporal cues in multi-camera 3d object detection," *arXiv preprint arXiv:2203.17054*, 2022.

[35] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[36] B. Zhu, Z. Jiang, X. Zhou, Z. Li, and G. Yu, "Class-balanced grouping and sampling for point cloud 3d object detection," *arXiv preprint arXiv:1908.09492*, 2019.

[37] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[38] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *IEEE Symposium on Security and Privacy*, 2017, pp. 39–57.

[39] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[40] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," *arXiv preprint arXiv:1903.12261*, 2019.

[41] Y. Zhang, Z. Zhu, W. Zheng, J. Huang, G. Huang, J. Zhou, and J. Lu, "Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving," *arXiv preprint arXiv:2205.09743*, 2022.

[42] W. Roh, G. Chang, S. Moon, G. Nam, C. Kim, Y. Kim, S. Kim, and J. Kim, "Ora3d: Overlap region aware multi-view 3d object detection," *arXiv preprint arXiv:2207.00865*, 2022.

[43] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, and Y.-G. Jiang, "Polarformer: Multi-camera 3d object detection with polar transformers," *arXiv preprint arXiv:2206.15398*, 2022.

[44] Y. Shi, J. Shen, Y. Sun, Y. Wang, J. Li, S. Sun, K. Jiang, and D. Yang, "Srcn3d: Sparse r-cnn 3d surround-view camera object detection and tracking for autonomous driving," *arXiv preprint arXiv:2206.14451*, 2022.

[45] X. Lin, T. Lin, Z. Pei, L. Huang, and Z. Su, "Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion," *arXiv preprint arXiv:2211.10581*, 2022.

[46] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, M. Tomizuka, L. Li, Z. Yuan, C. Wang *et al.*, "Sparse r-cnn: End-to-end object detection with learnable proposals," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14 454–14 463.

[47] D. Hendrycks, K. Zhao, S. Basart, J. Steinhardt, and D. Song, "Natural adversarial examples," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 262–15 271.

[48] D. Hendrycks, S. Basart, N. Mu, S. Kadavath, F. Wang, E. Dorundo, R. Desai, T. Zhu, S. Parajuli, M. Guo *et al.*, "The many faces of robustness: A critical analysis of out-of-distribution generalization," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8340–8349.

[49] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do imagenet classifiers generalize to imagenet?" in *International Conference on Machine Learning*. PMLR, 2019, pp. 5389–5400.

[50] A. Barbu, D. Mayo, J. Alverio, W. Luo, C. Wang, D. Gutfreund, J. Tenenbaum, and B. Katz, "Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models," in *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Artificial Intelligence Safety and Security*, 2018, pp. 99–112.

[52] S. Xie, Z. Li, Z. Wang, and C. Xie, "On the adversarial robustness of camera-based 3d object detection," *arXiv preprint arXiv:2301.10766*, 2023.

[53] Y. Lee and J. Park, "Centermask: Real-time anchor-free instance segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 906–13 915.

[54] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2485–2494.

[55] J. Park, C. Xu, S. Yang, K. Keutzer, K. Kitani, M. Tomizuka, and W. Zhan, "Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection," *International Conference on Learning Representations*, 2023.

[56] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *IEEE International Conference on Robotics and Automation*, 2023.

[57] T. Liang, H. Xie, K. Yu, Z. Xia, Z. Lin, Y. Wang, T. Tang, B. Wang, and Z. Tang, "Bevfusion: A simple and robust lidar-camera fusion framework," *arXiv preprint arXiv:2205.13790*, 2022.

[58] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[59] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, and F. Zhao, "Autoalignv2: Deformable feature aggregation for dynamic multi-modal 3d object detection," *arXiv preprint arXiv:2207.10316*, 2022.

[60] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, Y. Rao, G. Huang, J. Lu, and J. Zhou, "Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation," in *Conference on Robot Learning*. PMLR, 2023, pp. 539–549.

[61] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[62] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9223–9232.

[63] Z. Zhu, Y. Zhang, H. Chen, Y. Dong, S. Zhao, W. Ding, J. Zhong, and S. Zheng, "Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 600–21 610.

[64] F. Bartoccioni, É. Zablocki, A. Bursuc, P. Pérez, M. Cord, and K. Alahari, "Lara: Latents and rays for multi-camera bird's-eye-view semantic segmentation," in *Conference on Robot Learning*. PMLR, 2023, pp. 1663–1672.

[65] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8406–8415.

[66] L. Kong, Y. Liu, X. Li, R. Chen, W. Zhang, J. Ren, L. Pan, K. Chen, and Z. Liu, "Robo3d: Towards robust and reliable 3d perception against corruptions," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 994–20 006.

[67] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," *arXiv preprint arXiv:2303.03991*, 2023.

[68] C. Min, X. Xu, D. Zhao, L. Xiao, Y. Nie, and B. Dai, "Occ-bev: Multi-camera unified pre-training via 3d scene reconstruction," *arXiv preprint arXiv:2305.18829*, 2023.

[69] C. Ge, J. Chen, E. Xie, Z. Wang, L. Hong, H. Lu, Z. Li, and P. Luo, "Metabev: Solving sensor failures for 3d detection and map segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8721–8731.

[70] Y. Man, L.-Y. Gui, and Y.-X. Wang, "Dualcross: Cross-modality cross-domain adaptation for monocular bev perception," *arXiv preprint arXiv:2305.03724*, 2023.

[71] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, "Autoalign: pixel-instance feature aggregation for multi-modal 3d object detection," *arXiv preprint arXiv:2201.06493*, 2022.

[72] C. Han, J. Sun, Z. Ge, J. Yang, R. Dong, H. Zhou, W. Mao, Y. Peng, and X. Zhang, "Exploring recurrent long-term temporal fusion for multi-view 3d perception," *arXiv preprint arXiv:2303.05970*, 2023.

[73] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "Futr3d: A unified sensor fusion framework for 3d detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 172–181.

[74] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, and J. Jia, "Unifying voxel-based representation with transformer for 3d object detection," in *Advances in Neural Information Processing Systems*, vol. 35, pp. 18 442–18 455, 2022.

[75] L. Kong, S. Xie, H. Hu, L. X. Ng, B. R. Cottereau, and W. T. Ooi, "Robodepth: Robust out-of-distribution depth estimation under corruptions," in *Advances in Neural Information Processing Systems*, 2023.

[76] L. Kong, Y. Niu, S. Xie, H. Hu, L. X. Ng, B. R. Cottereau, D. Zhao, L. Zhang, H. Wang, W. T. Ooi, R. Zhu, Z. Song, L. Liu, T. Zhang, J. Yu, M. Jing, P. Li, X. Qi, C. Jin, Y. Cheng, J. Hou, J. Zhang, Z. Kan, Q. Lin, L. Peng, M. Li, D. Xu, C. Yang, Y. Yao, G .Wu, J. Kuai, X. Liu, J. Jiang, J. Huang, B. Li, J. Chen, S. Zhang, S. Ao, Z. Li, R. Chen, H. Luo, F. Zhao, and J. Yu, "The robodepth challenge: Methods and advancements towards robust depth estimation," *arXiv preprint arXiv:2307.15061*, 2023.

[77] J. Ren, L. Pan, and Z. Liu, "Benchmarking and analyzing point cloud classification under corruptions," in *International Conference on Machine Learning*, 2022, pp. 18 559–18 575,.

[78] S. Shi, X. Wang, and H. Li, "Pointrcnn: 3d object proposal generation and detection from point cloud," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 770–779.

[79] W. Shi and R. Raj, "Point-gnn: Graph neural network for 3d object detection in a point cloud," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 1711–1719.

[80] Z. Yang, Y. Sun, S. Liu, X. Shen, and J. Jia, "Std: Sparse-to-dense 3d object detector for point cloud," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 1951–1960.

[81] Z. Yang, Y. Sun, S. Liu, and J. Jia, "3dssd: Point-based 3d single stage object detector," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 040–11 048.

[82] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, 2018, pp. 3337, vol. 18.

[83] Y. Zhou and O. Tuze, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4490–4499.

[84] T. Yin, X. Zhou, and P. Krahenbuhl, "Center-based 3d object detection and tracking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 784–11 793.

[85] G. Shi, R. Li, and C. Ma, "PillarNet: High-Performance Pillar-based 3D Object Detection," *arXiv preprint arXiv:2205.07403*, 2022.

[86] S. Shi, L. Jiang, J. Deng, Z. Wang, C. Guo, J. Shi, X. Wang, and H. Li "PV-RCNN++: Point-voxel feature set abstraction with local vector representation for 3D object detection," *International Journal of Computer Vision*, 2022.

[87] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10 529–10 583.

[88] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6411–6420.

[89] B. Wu, A. Wan, X. Yue, and K. Keutzer, "Squeezeseg: Convolutional neural nets with recurrent crf for real-time road-object segmentation from 3d lidar point cloud," in *IEEE International Conference on Robotics and Automation*, 2018, pp. 1887–1893.

[90] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "Rangenet++: Fast and accurate lidar semantic segmentation," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2019, pp. 4213–4220.

[91] L. Kong, Y. Liu, R. Chen, Y. Ma, X. Zhu, Y. Li, Y. Hou, Y. Qiao and Z. Liu, "Rethinking Range View Representation for LiDAR Segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2023, pp. 228–240.

[92] Y. Zhang, Z. Zhou, P. David, X. Yue, Z. Xi, B. Gong, and H. Foroosh, "Polarnet: An improved grid representation for online lidar point clouds semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9601-9610.

[93] C. Choy, J. Gwak, and S. Savarese, "4d spatio-temporal convnets: Minkowski convolutional neural networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3075–3084.

[94] V. E. Liong, T. N. T. Nguyen, S. Widjaja, D. Sharma, and Z. J. Chong, "Amvnet: Assertion-based multi-view fusion network for LiDAR semantic segmentation," *arXiv preprint arXiv:2012.04934*, 2020.

[95] J. Xu, R. Zhang, J. Dou, Y. Zhu, J. Sun, and S. Pu, "Rpvnet: A deep and efficient range-point-voxel fusion network for lidar point cloud segmentation," in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 024–16 033.

[96] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, and S. Han, "Searching efficient 3d architectures with sparse point-voxel convolution," in *European Conference on Computer Vision*, 2020, pp. 685–702.

[97] L. Kong, N. Quader, and V. E. Liong, "ConDA: Unsupervised domain adaptation for LiDAR segmentation via regularized domain concatenation," in *IEEE International Conference on Robotics and Automation*, 2023, pp. 9338–9345.

[98] D. Park, R. Ambrus, V. Guizilini, J. Li, and A. Gaidon, "Is pseudo-lidar needed for monocular 3d object detection?" in *IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3142–3152.

[99] C. Sakaridis, D. Dai, and L. Van Gool, "Acdc: The adverse conditions dataset with correspondences for semantic driving scene understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 765–10 775.

[100] C. Sakaridis, D. Dai, and L. Van Gool, "Semantic foggy scene understanding with synthetic data," *International Journal of Computer Vision*, vol. 126, pp. 973–992, 2018.

[101] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.