# BENCHMARKING BIRD'S EYE VIEW DETECTION ROBUSTNESS TO REAL-WORLD CORRUPTIONS

**Shaoyuan Xie**[1], **Lingdong Kong**[2,3], **Wenwei Zhang**[2,4], **Jiawei Ren**[4], **Liang Pan**[4], **Kai Chen**[2], **Ziwei Liu**[4,✉]

[1]Huazhong University of Science and Technology [2]Shanghai AI Laboratory
[3]National University of Singapore [4]S-Lab, Nanyang Technological University

shaoyuanxie@hust.edu.cn {konglingdong,zhangwenwei,chenkai}@pjlab.org.cn
{jiawei011,liang.pan,ziwei.liu}@ntu.edu.sg

## ABSTRACT

The recent advent of camera-based bird's eye view (BEV) detection algorithms exhibits great potential for in-vehicle 3D object detection. Despite the progressively achieved results on the standard benchmark, the robustness of BEV detectors has not been thoroughly examined, which is critical for safe operations. To fill in this gap, we introduce **nuScenes-C**, a test suite that encompasses eight distinct corruptions with a high likelihood to occur in real-world applications, including *Bright*, *Dark*, *Fog*, *Snow*, *Motion Blur*, *Color Quant*, *Camera Crash*, and *Frame Lost*. Based on **nuScenes-C**, we extensively evaluate a wide range of BEV detection models to understand their resilience and reliability. Our findings indicate a strong correlation between the absolute performance on in-distribution and out-of-distribution datasets. Nonetheless, there is considerable variation in relative performance across different approaches. Our experiments further demonstrate that pre-training and depth-free BEV transformation have the potential to enhance out-of-distribution robustness. The benchmark is openly accessible at `https://github.com/Daniel-xsy/RoboBEV`.

## 1 INTRODUCTION

Deep neural network-based 3D object detectors, such as those presented in Li et al. (2022b); Wang et al. (2022b); Huang et al. (2021); Liu et al. (2022a); Wang et al. (2022a; 2021); Lang et al. (2019); Vora et al. (2020); Zhou & Tuzel (2018); Yan et al. (2018), have demonstrated promising performance on various challenging real-world benchmarks, including the KITTI Geiger et al. (2012), nuScenes Caesar et al. (2020), and Waymo Open Sun et al. (2020). These popular approaches utilize either point clouds Lang et al. (2019); Vora et al. (2020); Zhou & Tuzel (2018); Yan et al. (2018) or images Wang et al. (2021; 2022a;b); Li et al. (2022b;a); Huang et al. (2021); Liu et al. (2022a) as inputs for detection tasks. In comparison to LiDAR-based methods, camera-based approaches have caught significant attention due to their low deployment cost, high computational efficiency, and dense semantic information. Additionally, camera-based detection exhibits inherent advantages in detecting long-range objects and identifying vision-based traffic signs.

Despite remarkable progress achieved by recent camera-based 3D object detection methods, their ability to withstand natural corruption remains inadequately understood. The reliability of camera-based 3D object detectors under corruption is crucial because these methods are usually applied in highly safety-critical systems (*e.g.*, autonomous driving systems) in the real world, which introduces a range of external factors, including fluctuations in lighting and diverse meteorological conditions. Moreover, images captured by the camera may become blurred or even fail to register, due to changes in velocity and sensor malfunction, respectively. Therefore, we are motivated to investigate the robustness of these approaches in the face of diverse natural corruptions.

To bridge this knowledge gap, we undertake the initial comprehensive benchmark for assessing the out-of-distribution robustness of camera-based BEV detectors. To facilitate this, we develop a novel dataset, namely **nuScenes-C**, which comprises eight distinct natural corruptions. Each corruption type is divided into three different levels – easy, moderate, and hard since corruptions can manifest

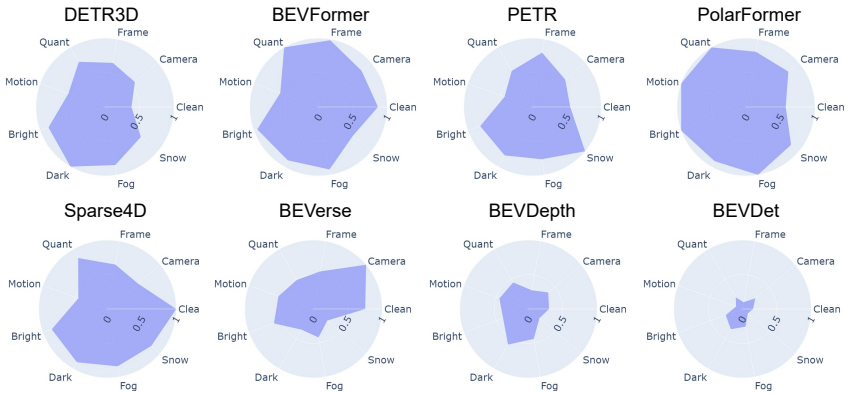



Figure 1: The radar charts of existing BEV detectors' nuScenes Detection Score (NDS) Caesar et al. (2020) under eight corruption types. We observe diverse behaviors of different models even with competitive "clean" performance. The NDS is normalized across all the benchmarking BEV detection models to lie between 0.1 and 1.

themselves at varying intensities Hendrycks & Dietterich (2019). Our study entails a comprehensive investigation of diverse corruption settings, encompassing the failure of sensors, which often occurs in real-time systems where individual cameras malfunction or a series of frames are lost during processing. We categorize this type of corruption as *Camera Crash* and *Frame Lost*, respectively. The second category of corruption results from vehicle motion and image processing. Specifically, images captured by cameras while moving at high speed may be *Motion Blur*, or the images may be *Color Quant* when memory is limited. Lastly, real-world scenarios contain diverse light and weather conditions. To analyze model performance under these conditions, we incorporate *Brightness*, *Low Light*, *Fog*, and *Snow* weather to cover the above situations.

Leveraging the proposed nuScenes-C dataset, we present an exhaustive analysis of camera-based 3D object detection models. In particular, our study benchmarks **10** existing models with **22** variants versions and provides noteworthy insights into their out-of-distribution robustness. As shown in Figure 1, the benchmark results indicate that models present competitive performance on clean datasets vary largely on corrupted data and different corrupted types. We find that methods using pre-training and depth-free BEV transformation exhibit better out-of-distribution robustness than those not. Additionally, temporal fusion can provide better robustness only when they are carefully designed to avoid error accumulation. The above findings provide valuable insights for designing future models to strive for better out-of-distribution robustness.

The key contributions of this work are summarized as follows.

- To the best of our knowledge, we make the first attempt to benchmark the robustness of camera-based BEV 3D object detectors under natural corruptions.
- We conduct exhaustive experiments to benchmark 22 existing camera-only BEV 3D object detectors on the proposed nuScenes-C dataset.
- Our study provides an in-depth analysis of the experimental results, highlighting the factors that contribute to the superior robustness of certain camera-only BEV 3D object detectors on the out-of-distribution dataset.

## 2 Related Work

**Camera-based Bird's Eye View Detection**. Compared to the monocular image input Wang et al. (2021); Park et al. (2021), conducting 3D detection by consuming multi-view images and building bird's eye view (BEV) representations Li et al. (2022b); Huang et al. (2021); Li et al. (2022a); Wang et al. (2022b); Liu et al. (2022a); Zhang et al. (2022); Lin et al. (2022); Roh et al. (2022); Shi et al. (2022) enjoys several advantages. First, it allows for joint learning from multi-view images. Second, the bird's eye perspective provides a physics-interpretable way for information fusion from different


Figure 2: Corruption examples from nuScenes-C dataset. **Left:** Corruption taxonomy. **Right:** Temporal corruptions. *Camera Crash* drop fixed set of images along timestamps; *Frame Lost* randomly drop frames along timestamps.

sensors and different timestamps Ma et al. (2022). Third, BEV output space can be readily applied in many downstream tasks, like prediction and planning, which leads to substantial improvements in performance for BEV-based detection frameworks. Existing works can be basically divided into two categories based on whether they perform depth estimation explicitly. Inspired by LSS Philion & Fidler (2020), BEVDet Huang et al. (2021) utilizes an extra depth estimation branch to perform the transform from perspective view to bird's eye view (PE2BEV). BEVDepth Li et al. (2022a) makes more accurate depth estimation via explicit depth supervision from point clouds and further improves the performance. Based on the above pipeline, BEVerse Zhang et al. (2022) proposes a unified framework for multi-task learning and achieving state-of-the-art performance. Another line of work performs PV2BEV without explicit depth estimation. Following DETR Carion et al. (2020), DETR3D Wang et al. (2022b) represents 3D objects as queries and performs cross-attention via Transformer decoder. PETR Liu et al. (2022a;b) further improves performance by proposing 3D position-aware representations. BEVFormer Li et al. (2022b) introduces temporal cross-attention to extract BEV features from multi-timestamps images. PolarFormer Jiang et al. (2022) explores predicting 3D targets in polar coordinates. Inspired by the success of Sparse RCNN Sun et al. (2021), SRCN3D Shi et al. (2022) introduce sparse proposals for feature aggregation. Despite these works showing competitive results on the standard dataset, it's unclear how they behave in front of natural corruptions.

**Robustness under Adversarial Attacks**. Contemporary neural networks are prone to adversarial attacks, where a deliberately created perturbation added to the input can lead the network to produce incorrect predictions Szegedy et al. (2013); Goodfellow et al. (2014); Moosavi-Dezfooli et al. (2017). Adversarial examples have been extensively researched in various vision tasks, including classification Szegedy et al. (2013); Goodfellow et al. (2014); Moosavi-Dezfooli et al. (2017); Madry et al. (2017); Brown et al. (2017); Liu et al. (2016), detection Xie et al. (2017); Liu et al. (2018); Tu et al. (2020); Cao et al. (2021), and segmentation Xie et al. (2017); Rossolini et al. (2022). These adversarial examples can be generated both in digital Szegedy et al. (2013); Goodfellow et al. (2014); Moosavi-Dezfooli et al. (2017); Madry et al. (2017); Brown et al. (2017); Liu et al. (2016); Xie et al. (2017) and physical environments Rossolini et al. (2022); Kurakin et al. (2018); Cao et al. (2021). The recent study demonstrates that the 3D perception system tends to crash when exposed to adversarial examples, which could pose potential safety risks in the deployment stages Rossolini et al. (2022); Cao et al. (2021); Tu et al. (2020). While Xie *et al* Xie et al. (2023) conducts a comprehensive study of the adversarial robustness of camera-based detectors, we focus on natural corruptions, which are more likely to occur in the real world.

**Robustness under Natural Corruptions**. The assessment of model robustness has been an active research area in computer vision, where the evaluation is mainly focused on worst-case robustness against adversarial attacks. On the other hand, natural corruption robustness assesses the average-case performance of models against common corrupted images that are prevalent in real-world applications. Various benchmarks have been proposed to evaluate the robustness of 2D image classification models, including ImageNet-C Hendrycks & Dietterich (2019), ObjectNet Barbu et al.

(2019), ImageNetV2 Recht et al. (2019), ImageNet-A Hendrycks et al. (2021b), and ImageNet-R Recht et al. (2019). ImageNet-C corrupts the ImageNet's test set with simulated corruptions like compression loss and motion blur, while ObjectNet collects a test set with rich variations in rotation, background, and viewpoint. ImageNetV2 re-collects a test set following ImageNet's protocol and evaluates the performance gap due to the natural distribution shift. Moreover, ImageNet-A and ImageNet-R benchmark the classifier's robustness to natural adversarial examples and abstract visual renditions, respectively. In this context, Hendrycks et al. (2021a) highlights the correlation between the robustness of simulated corruptions and the improvement of real-world corruptions. However, there is still a lack of comparable benchmarks to evaluate the out-of-distribution robustness of vision-based 3D object detection models, which are commonly used in safety-critical applications. To this end, we make the first attempt to understand the robustness of these models under natural corruptions.

## 3 ROBOBEV BENCHMARK

**Benchmark Design**. Our benchmark is composed of eight distinct corruption types, which are applied to the validation set of nuScenes Caesar et al. (2020), which contains 6019 images with 360-degree views. We adhere to the experimental protocol established in Hendrycks & Dietterich (2019) and establish three different levels of corruption intensity (*i.e.*, easy, moderate, and hard) for each type of corruption. These severity levels are carefully set to avoid drastic performance drops that could hinder drawing meaningful conclusions. Additionally, we incorporate variation within each severity level of corruption to augment the diversity of the dataset.

**Natural Corruptions**. The initial category of corruption that we examine is sensor failure, which can arise in real-world settings due to physical damage to cameras. In our experiments, we simulate this scenario by introducing the *Camera Crash* corruption. More specifically, we randomly select a fixed set of cameras according to the corruption intensity and discard the corresponding images consistently. Additionally, we propose *Frame Lost*, where we randomly drop various cameras independently with the same probability at every timestamp, emulating situations where some frames are lost in consecutive time sequences. This procedure is illustrated in Figure 2. The third corruption category, *Motion Blur*, occurs when images are blurred due to camera motion at high speeds. Furthermore, we account for images that are quantized when deployed on devices with strict memory constraints by considering the *Color Quantization* corruption. The last four types of corruption pertain to lighting and weather conditions. We address these situations by including *Brightness*, *Low Light*, *Fog*, and *Snow*. The corrupted images are displayed in Figure 2.

**Robustness Metrics**. We follow the nuScenes Caesar et al. (2020) official metric to report performance 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), mean Average Attribute Error(mAAE).

To better compare the robustness among different BEV detectors, we introduce two new metrics inspired by Hendrycks & Dietterich (2019) 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:

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

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:

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

In our benchmark, we report both metrics for each candidate BEV algorithm and draw key analyses upon them.

Table 1: Comparisons of configurations and performance among camera-based BEV detectors. **Pre-train** denotes the model is initialized from the pre-trained FCOS3D Wang et al. (2021) checkpoint. **Temporal** indicates the use of temporal information. **Depth** denotes models with explicit depth estimation branch. **CBGS** highlight models use class-balanced group-sampling Zhu et al. (2019).

| Model | Pre-train | Temporal | Depth | CBGS | Backbone | BEV Encoder | NDS ↑ | mCE (%) ↓ | mRR (%) ↑ |
|---|---|---|---|---|---|---|---|---|---|
| DETR3D | ✓ | | | | ResNet | Attention | 0.4224 | 100.00 | 70.77 |
| DETR3D (cbgs) | ✓ | | | ✓ | ResNet | Attention | 0.4341 | 99.21 | 70.02 |
| BEVFormer (small) | ✓ | ✓ | | | ResNet | Attention | 0.4787 | 101.23 | 59.07 |
| BEVFormer-S (small) | ✓ | | | | ResNet | Attention | 0.2622 | 114.43 | 76.87 |
| BEVFormer (base) | ✓ | ✓ | | | ResNet | Attention | 0.5174 | 97.97 | 60.40 |
| BEVFormer-S (base) | ✓ | | | | ResNet | Attention | 0.2622 | 101.87 | 69.33 |
| PETR (r50) | | | | | ResNet | Attention | 0.3665 | 111.01 | 61.26 |
| PETR (vov) | ✓ | | | | VoVNet-V2 | Attention | 0.4550 | 100.69 | 65.03 |
| ORA3D | ✓ | | | | ResNet | Attention | 0.4436 | 99.17 | 68.63 |
| PolarFormer (r101) | ✓ | | | | ResNet | Attention | 0.4602 | 96.06 | 70.88 |
| PolarFormer (vov) | ✓ | | | | VoVNet-V2 | Attention | 0.4558 | 98.75 | 67.51 |
| SRCN3D (r101) | ✓ | | | | ResNet | CNN + Attn. | 0.4286 | 99.67 | 70.23 |
| SRCN3D (vov) | ✓ | | | | VoVNet-V2 | CNN + Attn. | 0.4205 | 102.04 | 67.95 |
| Sparse4D (r101) | ✓ | ✓ | | | ResNet | CNN + Attn. | 0.5438 | 100.01 | 55.04 |
| BEVDet (r50) | | | ✓ | ✓ | ResNet | CNN | 0.3770 | 115.12 | 51.83 |
| BEVDet (r100) | | | ✓ | ✓ | ResNet | CNN | 0.3877 | 113.68 | 53.12 |
| BEVDet (tiny) | | | ✓ | ✓ | Swin | CNN | 0.4037 | 116.48 | 46.26 |
| BEVDepth (r50) | | | ✓ | ✓ | ResNet | CNN | 0.4058 | 110.02 | 56.82 |
| BEVerse (tiny) | | ✓ | ✓ | ✓ | Swin | CNN | 0.4665 | 110.67 | 48.60 |
| BEVerse-S (tiny) | | | ✓ | ✓ | Swin | CNN | 0.1603 | 137.25 | 28.24 |
| BEVerse (small) | | ✓ | ✓ | ✓ | Swin | CNN | 0.4951 | 117.82 | 49.57 |
| BEVerse-S (small) | | | ✓ | ✓ | Swin | CNN | 0.2682 | 132.13 | 29.54 |

# 4 EXPERIMENT

In our study, we conduct a comprehensive benchmarking analysis of 22 models on the nuScenes-C dataset. We use DETR3D Wang et al. (2022b) as our baseline for the mCE metric, given that many of the state-of-the-art approaches are based on it. The main results are in Table 2 and Table 3.

Broadly speaking, we observe a solid linear relationship between the absolute performance on the nuScenes-C dataset and the performance on the clean nuScenes Caesar et al. (2020) dataset. Specifically, models with leading performance on the clean dataset are also more likely to perform better on the out-of-distribution dataset, as demonstrated in Figure 3(a). However, a closer examination reveals that the situation becomes more complex. We find that models with similar performance on the clean dataset exhibit diverse robustness under different types of corruption. For example, BEVerse (small)Zhang et al. (2022) significantly improves the robustness of *Camera Crash* over the baseline, while PETR (vov)Liu et al. (2022a) performs well in *Snow* weather. However, both models struggle to make accurate predictions under *Dark* conditions.

Although the mCE metric reveals a clear linear relationship between the absolute performance on the clean dataset and that on the out-of-distribution nuScenes-C dataset, the mRR metric shows a large variation among models that exhibit similar performance on the clean dataset, as depicted in Figure 3(b). This indicates that models may overfit the clean dataset and fail to generalize well to the nuScenes-C dataset. Moreover, we observe that models with superior performance on the clean dataset do not necessarily perform better in terms of the mRR metric compared to the mCE metric. For example, Sparse4D Lin et al. (2022) outperforms DETR3D Wang et al. (2022b) on nuScenes with a significant margin (0.5504 *vs.* 0.4224), while all the mRR metrics of Sparse4D are lower than that of DETR3D. Additionally, DETR3D exhibits the most robust performance towards *Dark* conditions, while BEVerse (tiny), which has better clean performance (0.4665 *vs.* 0.4224), only achieves a relative performance of approximately 12% under dark lighting conditions. These results suggest that the performance on the clean nuScenes Caesar et al. (2020) dataset cannot provide a comprehensive understanding of the model's performance, and therefore it is crucial to benchmark the state-of-the-art models from multiple perspectives.

To further investigate the behavior of model robustness, we break down the models into various components, namely, training strategies (*e.g.*, FCOS3D Wang et al. (2021) pretraining and CBGS Zhu et al. (2019) resampling strategy for addressing unbalanced classes), model architectures (*e.g.*, backbone and detection head), and approach pipelines (*e.g.*, temporal cue learning and depth estimation). The results of this analysis are in Table 1. Specifically, we focus on FCOS3D Wang et al. (2021)

Table 2: Results for the Corruption Error (CE) of each model in our *RoboBEV* benchmark. **Bold**: Best in the column. Underline: Second best in the column. Blue : Best in the row if improve upon baseline. Yellow : Worst in the row if decline upon baseline.

| Model | NDS ↑ | mCE ↓ | Camera | Frame | Quant | Motion | Bright | Dark | Fog | Snow |
|---|---|---|---|---|---|---|---|---|---|---|
| DETR3D | 0.4224 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 | **100.00** | 100.00 | 100.00 |
| DETR3D (cbgs) | 0.4341 | 99.21 | 98.15 | 98.90 | 99.15 | 101.62 | 97.47 | 100.28 | 98.23 | 99.85 |
| BEVFormer (small) | 0.4787 | 102.40 | 101.23 | 101.96 | 98.56 | 101.24 | 104.35 | 105.17 | 105.40 | 101.29 |
| BEVFormer (base) | 0.5174 | 97.97 | 95.87 | 94.42 | **95.13** | 99.54 | 96.97 | 103.76 | 97.42 | 100.69 |
| PETR (r50) | 0.3665 | 111.01 | 107.55 | 105.92 | 110.33 | 104.93 | 119.36 | 116.84 | 117.02 | 106.13 |
| PETR (vov) | 0.4550 | 100.69 | 99.09 | 97.46 | 103.06 | 102.33 | 102.40 | 106.67 | 103.43 | **91.11** |
| ORA3D | 0.4436 | 99.17 | 97.26 | 98.03 | 97.32 | 100.19 | 98.78 | 102.40 | 99.23 | 100.19 |
| PolarFormer (r101) | 0.4602 | **96.06** | 96.16 | 97.24 | **95.13** | **92.37** | **94.96** | 103.22 | **94.25** | 95.17 |
| PolarFormer (vov) | 0.4558 | 98.75 | 96.13 | 97.2 | 101.48 | 104.32 | 95.37 | 104.78 | 97.55 | 93.14 |
| SRCN3D (r101) | 0.4286 | 99.67 | 98.77 | 98.96 | 97.93 | 100.71 | 98.80 | 102.72 | 99.54 | 99.91 |
| SRCN3D (vov) | 0.4205 | 102.04 | 99.78 | 100.34 | 105.13 | 107.06 | 101.93 | 107.1 | 102.27 | 92.75 |
| Sparse4D$_{(R101)}$ | **0.5438** | 100.01 | 99.80 | 99.91 | 98.05 | 102.00 | 100.30 | 103.83 | 100.46 | 95.72 |
| BEVDet (r50) | 0.3770 | 115.12 | 105.22 | 109.19 | 111.27 | 108.18 | 123.96 | 123.34 | 123.83 | 115.93 |
| BEVDet (r101) | 0.3877 | 113.68 | 103.32 | 107.29 | 109.25 | 105.40 | 124.14 | 123.12 | 123.28 | 113.64 |
| BEVDet (tiny) | 0.4037 | 116.48 | 103.50 | 106.61 | 113.18 | 107.26 | 130.19 | 131.83 | 124.01 | 115.25 |
| BEVDepth (r50) | 0.4058 | 110.02 | 103.09 | 106.26 | 106.24 | 102.02 | 118.72 | 114.26 | 116.57 | 112.98 |
| BEVerse (tiny) | 0.4665 | 110.67 | 95.49 | **94.15** | 108.46 | 100.19 | 122.44 | 130.40 | 118.58 | 115.69 |
| BEVerse (small) | 0.4951 | 107.82 | **92.93** | 101.61 | 105.42 | 100.40 | 110.14 | 123.12 | 117.46 | 111.48 |

Table 3: Results for the Resilience Rate (RR) of each model in our *RoboBEV* benchmark. **Bold**: best in column. Underline: second best in column.

| Model | mRR ↑ | Camera | Frame | Quant | Motion | Bright | Dark | Fog | Snow |
|---|---|---|---|---|---|---|---|---|---|
| DETR3D | 70.77 | 67.68 | 61.65 | 75.21 | 63.00 | 94.74 | **65.96** | **92.61** | 45.29 |
| DETR3D (cbgs) | 70.02 | **68.90** | 61.85 | 74.52 | 58.56 | **95.69** | 63.72 | **92.61** | 44.34 |
| BEVFormer (small) | 59.07 | 57.89 | 51.37 | 68.41 | 53.69 | 78.15 | 50.41 | 74.85 | 37.79 |
| BEVFormer (base) | 60.40 | 60.96 | 58.31 | 67.82 | 52.09 | 80.87 | 48.61 | 78.64 | 35.89 |
| PETR (r50) | 61.26 | 63.30 | 59.10 | 67.45 | 62.73 | 77.52 | 42.86 | 78.47 | 38.66 |
| PETR (vov) | 65.03 | 64.26 | 61.36 | 65.23 | 54.73 | 84.79 | 50.66 | 81.38 | **57.85** |
| ORA3D | 68.63 | 68.87 | 61.99 | 75.74 | 59.67 | 91.86 | 58.90 | 89.25 | 42.79 |
| PolarFormer (r101) | **70.88** | 68.08 | 61.02 | 76.25 | **69.99** | 93.52 | 55.50 | **92.61** | 50.07 |
| PolarFormer (vov) | 67.51 | 68.78 | 61.67 | 67.49 | 51.43 | 93.9 | 53.55 | 89.10 | 54.15 |
| SRCN3D (r101) | 70.23 | 68.76 | 62.55 | **77.41** | 60.87 | 95.05 | 60.43 | 91.93 | 44.80 |
| SRCN3D (vov) | 67.95 | 68.37 | 61.33 | 67.23 | 50.96 | 92.41 | 54.08 | 89.75 | 59.43 |
| Sparse4D$_{(R101)}$ | 55.04 | 52.83 | 48.01 | 60.87 | 46.23 | 73.26 | 46.16 | 71.42 | 41.54 |
| BEVDet (r50) | 51.83 | 65.94 | 51.03 | 63.87 | 54.67 | 68.04 | 29.23 | 65.28 | 16.58 |
| BEVDet (r101) | 53.12 | 67.63 | 53.26 | 65.67 | 58.42 | 65.88 | 28.84 | 64.35 | 20.89 |
| BEVDet (tiny) | 46.26 | 64.63 | 52.39 | 56.43 | 52.71 | 54.27 | 12.14 | 60.69 | 16.84 |
| BEVDepth (r50) | 56.82 | 65.01 | 52.76 | 67.79 | 61.93 | 70.95 | 43.30 | 71.54 | 21.27 |
| BEVerse (tiny) | 48.60 | 68.19 | **65.10** | 55.73 | 56.74 | 56.93 | 12.71 | 59.61 | 13.80 |
| BEVerse (small) | 49.57 | 67.95 | 50.19 | 56.70 | 53.16 | 68.55 | 22.58 | 57.54 | 19.89 |

pretraining and CBGS Zhu et al. (2019) resampling strategy, as they are commonly utilized in these approaches and have demonstrated effectiveness in improving the performance of the clean dataset. We find using pre-training and depth-free BEV transformation shows superior performance than those not. Additionally, different ways to fuse temporal information also lead to opposite results.

## 5 Further Analysis

### 5.1 Depth Estimation

Prior works in this area can be classified into two categories based on their approach to depth estimation. The first category, including works such as BEVDet Huang et al. (2021), BEVDepth Li et al. (2022a), and BEVerse Zhang et al. (2022), incorporates an explicit depth-estimation branch in the pipeline. This is done because predicting 3D bounding boxes from monocular images is an ill-posed problem. These approaches first predict a per-pixel depth map, which is then used to map image


(a) mCE *vs.* NDS (b) mRR *vs.* NDS

Figure 3: The performance on nuScenes-C is improved as the performance on the clean nuScenes dataset. The relation of absolute performance is close to linear. However, when considering the relative performance, the mRR metric is more randomly distributed without a clear trend to increase.

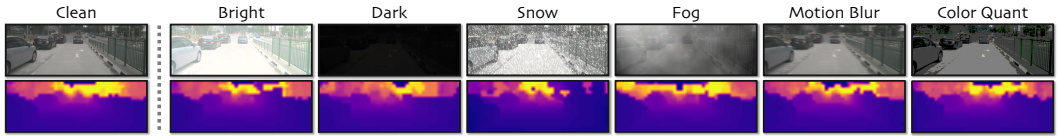
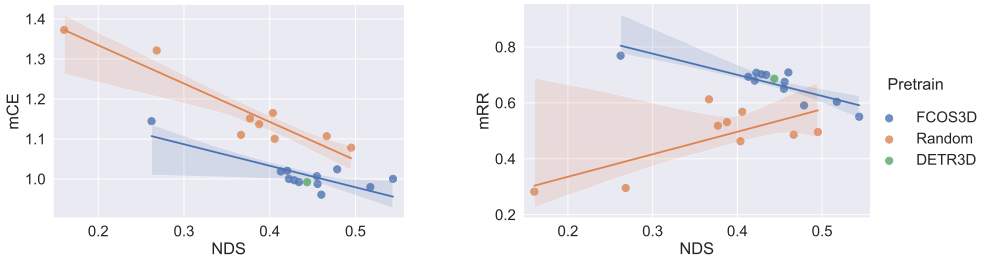



Figure 4: Depth estimation of BEVDepth Li et al. (2022a) under corruptions.

features to corresponding 3D locations. Subsequently, they predict 3D targets in the bird's-eye view (BEV) perspective by aggregating BEV features in a bottom-up fashion Wang et al. (2022b).

The second category of works uses pre-defined object queries Li et al. (2022b); Wang et al. (2022b) or sparse proposals Lin et al. (2022); Shi et al. (2022) to index 2D features in a top-down manner. These designs exhibit competitive performance on clean data, and we extend the analysis by examining their performance on out-of-distribution datasets.

Our evaluation results, as shown in Figure 1, demonstrate that the depth-based approaches, including BEVerse Zhang et al. (2022), BEVDet Huang et al. (2021), and BEVDepth Li et al. (2022a), struggle under most of corruption types. In addition, as shown in Table 2 and Table 3, all the depth-based approaches experience severe performance degradation when exposed to corrupted images, which may be attributed to inaccurate depth estimation, as illustrated in Figure 4. Due to the inaccurate depth intermediate, all the errors increase drastically compared to depth-free methods.

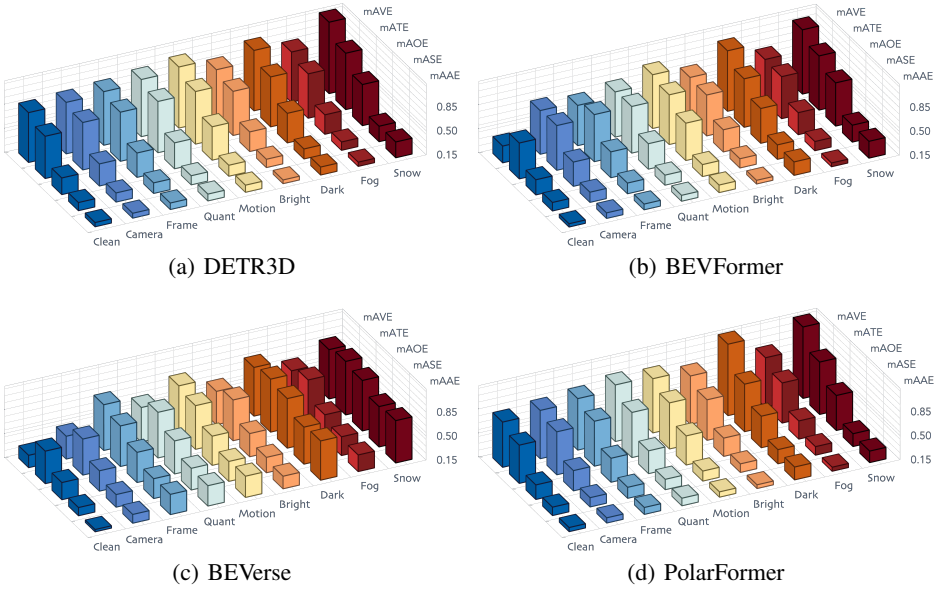



(a) mCE *vs.* NDS (b) mRR *vs.* NDS

Figure 5: Pretrain largely improves the robustness of models on the out-of-distribution dataset.

### 5.2 Pre-Training

In recent years, pretraining has emerged as a promising technique for enhancing the performance of computer vision models across a range of tasks. In the context of 3D detection, it is common




(a) DETR3D (b) BEVFormer

(c) BEVerse (d) PolarFormer

Figure 6: Other metric reported on nuScenes Caesar et al. (2020) other than NDS.

to use the FCOS3D Wang et al. (2021) weights as initialization. For the ResNet He et al. (2016) backbone. FCOS3D employs a depth weight of 0.2 for stable training, which is then switched to 1 for fine-tuning, as described in Wang et al. (2021). Alternatively, the VoVNet-V2 Lee & Park (2020) backbone is first trained on the DDAD15M Guizilini et al. (2020) dataset for depth estimation and then fine-tuned on the nuScenes Caesar et al. (2020) train set for the detection task. These two pretraining strategies can be categorized as semantic and depth pre-training, respectively. The results of these strategies are shown in Figure 5(a) and Figure 5(b), which demonstrates the benefits of pre-training for improving model robustness.

### 5.3 Temporal Modeling

In the context of autonomous driving applications, it remains challenging for models to accurately estimate the velocity of moving objects based on a single-frame input. Therefore, it is crucial for vision systems to leverage temporal cues to better perceive the surrounding environment. Previous works have proposed various approaches to learning effective temporal cues. In this study, we investigate whether temporal information can mitigate the influence of corrupted images. To this end, we compare the performance of BEVFormer Li et al. (2022b) and BEVerse Zhang et al. (2022), both of which have single-frame and multi-frame versions that allow us to analyze the impact of temporal information. Our experimental results, shown in Table 1, indicate that the use of multi-frame information largely improves the robustness for both the mCE and mRR metrics for BEVerse (*e.g.*, the BEVerse Small Zhang et al. (2022) improves from 132.12 to 117.82 on mCE and from 29.54 to 49.57 on mRR). Surprisingly, for BEVFormer Li et al. (2022b), the temporal cross-attention operation seems to improve the mCE metric but significantly degrades the mRR metric from 69.33 to 60.40. This can be attributed to that BEVFormer Li et al. (2022b) stores all the history information inside and updates it on-the-fly, the error caused by corruption can accumulate over time. On the other hand, BEVerse Zhang et al. (2022) only utilizes a fixed window of past timestamps.

## 6 Conclusion

In this work, we utilize a corpus of eight diverse natural corruptions to create the nuScenes-C dataset, which is used to assess the out-of-distribution robustness of leading camera-only BEV 3D object detection approaches. Exhaustive experiments are conducted to investigate the factors that impact the model's robustness. The findings of this study provide valuable insights for designing future models that can achieve better out-of-distribution robustness.

## REFERENCES

Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in Neural Information Processing Systems*, 32, 2019.

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

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11621–11631, 2020.

Yulong Cao, Ningfei Wang, Chaowei Xiao, Dawei Yang, Jin Fang, Ruigang Yang, Qi Alfred Chen, Mingyan Liu, and Bo Li. Invisible for both camera and lidar: Security of multi-sensor fusion based perception in autonomous driving under physical-world attacks. In *IEEE Symposium on Security and Privacy*, pp. 176–194, 2021.

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*, pp. 213–229, 2020.

Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3354–3361, 2012.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2485–2494, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

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 *IEEE/CVF International Conference on Computer Vision*, pp. 8340–8349, 2021a.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021b.

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

Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformers. *arXiv preprint arXiv:2206.15398*, 2022.

Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*, pp. 99–112. 2018.

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

Youngwan Lee and Jongyoul Park. Centermask: Real-time anchor-free instance segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13906–13915, 2020.

Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. *arXiv preprint arXiv:2206.10092*, 2022a.

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*, pp. 1–18, 2022b.

Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022.

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

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

Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022a.

Yingfei Liu, Junjie Yan, Fan Jia, Shuailin Li, Qi Gao, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petrv2: A unified framework for 3d perception from multi-camera images. *arXiv preprint arXiv:2206.01256*, 2022b.

Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1765–1773, 2017.

Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *IEEE/CVF International Conference on Computer Vision*, pp. 3142–3152, 2021.

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

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International Conference on Machine Learning*, pp. 5389–5400. PMLR, 2019.

Wonseok Roh, Gyusam Chang, Seokha Moon, Giljoo Nam, Chanyoung Kim, Younghyun Kim, Sangpil Kim, and Jinkyu Kim. Ora3d: Overlap region aware multi-view 3d object detection. *arXiv preprint arXiv:2207.00865*, 2022.

Giulio Rossolini, Federico Nesti, Gianluca D'Amico, Saasha Nair, Alessandro Biondi, and Giorgio Buttazzo. On the real-world adversarial robustness of real-time semantic segmentation models for autonomous driving. *arXiv preprint arXiv:2201.01850*, 2022.

Yining Shi, Jingyan Shen, Yifan Sun, Yunlong Wang, Jiaxin Li, Shiqi Sun, Kun Jiang, and Diange Yang. Srcn3d: Sparse r-cnn 3d surround-view camera object detection and tracking for autonomous driving. *arXiv preprint arXiv:2206.14451*, 2022.

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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2446–2454, 2020.

Peize Sun, Rufeng Zhang, Yi Jiang, Tao Kong, Chenfeng Xu, Wei Zhan, Masayoshi Tomizuka, Lei Li, Zehuan Yuan, Changhu Wang, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14454–14463, 2021.

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

James Tu, Mengye Ren, Sivabalan Manivasagam, Ming Liang, Bin Yang, Richard Du, Frank Cheng, and Raquel Urtasun. Physically realizable adversarial examples for lidar object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13716–13725, 2020.

Sourabh Vora, Alex H Lang, Bassam Helou, and Oscar Beijbom. Pointpainting: Sequential fusion for 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4604–4612, 2020.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 913–922, 2021.

Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Probabilistic and geometric depth: Detecting objects in perspective. In *Conference on Robot Learning*, pp. 1475–1485. PMLR, 2022a.

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*, pp. 180–191. PMLR, 2022b.

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Yuyin Zhou, Lingxi Xie, and Alan Yuille. Adversarial examples for semantic segmentation and object detection. In *IEEE/CVF International Conference on Computer Vision*, pp. 1369–1378, 2017.

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

Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018.

Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022.

Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4490–4499, 2018.

Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019.

# A APPENDIX

## A.1 FULL RESULT

In this section, we provide the full results of our benchmark. The results can be seen from Table 4 to Table 25.

| **Corruption** | **NDS ↑** | **mAP ↑** | **mATE ↓** | **mASE ↓** | **mAOE ↓** | **mAVE ↓** | **mAAE ↓** |
|---|---|---|---|---|---|---|---|
| Clean | 0.4224 | 0.3468 | 0.7647 | 0.2678 | 0.3917 | 0.8754 | 0.2108 |
| Cam Crash | 0.2859 | 0.1144 | 0.8400 | 0.2821 | 0.4707 | 0.8992 | 0.2202 |
| Frame Lost | 0.2604 | 0.0898 | 0.8647 | 0.3030 | 0.5041 | 0.9297 | 0.2439 |
| Color Quant | 0.3177 | 0.2165 | 0.8953 | 0.2816 | 0.5266 | 0.9813 | 0.2483 |
| Motion Blur | 0.2661 | 0.1479 | 0.9146 | 0.3085 | 0.6351 | 1.0385 | 0.2526 |
| Brightness | 0.4002 | 0.3149 | 0.7915 | 0.2703 | 0.4348 | 0.8733 | 0.2028 |
| Low Light | 0.2786 | 0.1559 | 0.8768 | 0.2947 | 0.5802 | 1.0290 | 0.2654 |
| Fog | 0.3912 | 0.3007 | 0.7961 | 0.2711 | 0.4326 | 0.8807 | 0.2110 |
| Snow | 0.1913 | 0.0776 | 0.9714 | 0.3752 | 0.7486 | 1.2478 | 0.3797 |

Table 4: DETR3D Wang et al. (2022b) results.

| **Corruption** | **NDS ↑** | **mAP ↑** | **mATE ↓** | **mASE ↓** | **mAOE ↓** | **mAVE ↓** | **mAAE ↓** |
|---|---|---|---|---|---|---|---|
| Clean | 0.4341 | 0.3494 | 0.7163 | 0.2682 | 0.3798 | 0.8421 | 0.1997 |
| Cam Crash | 0.2991 | 0.1174 | 0.7932 | 0.2853 | 0.4575 | 0.8471 | 0.2131 |
| Frame Lost | 0.2685 | 0.0923 | 0.8268 | 0.3135 | 0.5042 | 0.8867 | 0.2455 |
| Color Quant | 0.3235 | 0.2152 | 0.8571 | 0.2875 | 0.5350 | 0.9354 | 0.2400 |
| Motion Blur | 0.2542 | 0.1385 | 0.8909 | 0.3355 | 0.6707 | 1.0682 | 0.2928 |
| Brightness | 0.4154 | 0.3200 | 0.7357 | 0.2720 | 0.4086 | 0.8302 | 0.1990 |
| Low Light | 0.2766 | 0.1539 | 0.8419 | 0.3262 | 0.5682 | 1.09520 | 0.2847 |
| Fog | 0.4020 | 0.3012 | 0.7552 | 0.2710 | 0.4237 | 0.8302 | 0.2054 |
| Snow | 0.1925 | 0.0702 | 0.9246 | 0.3793 | 0.7648 | 1.2585 | 0.3577 |

Table 5: DETR3D Wang et al. (2022b) with cbgs results.

| **Corruption** | **NDS ↑** | **mAP ↑** | **mATE ↓** | **mASE ↓** | **mAOE ↓** | **mAVE ↓** | **mAAE ↓** |
|---|---|---|---|---|---|---|---|
| Clean | 0.4787 | 0.3700 | 0.7212 | 0.2792 | 0.4065 | 0.4364 | 0.2201 |
| Cam Crash | 0.2771 | 0.1130 | 0.8627 | 0.3099 | 0.5398 | 0.8376 | 0.2446 |
| Frame Lost | 0.2459 | 0.0933 | 0.8959 | 0.3411 | 0.5742 | 0.9154 | 0.2804 |
| Color Quant | 0.3275 | 0.2109 | 0.8476 | 0.2943 | 0.5234 | 0.8539 | 0.2601 |
| Motion Blur | 0.2570 | 0.1344 | 0.8995 | 0.3264 | 0.6774 | 0.9625 | 0.2605 |
| Brightness | 0.3741 | 0.2697 | 0.8064 | 0.2830 | 0.4796 | 0.8162 | 0.2226 |
| Low Light | 0.2413 | 0.1191 | 0.8838 | 0.3598 | 0.6470 | 1.0391 | 0.3323 |
| Fog | 0.3583 | 0.2486 | 0.8131 | 0.2862 | 0.5056 | 0.8301 | 0.2251 |
| Snow | 0.1809 | 0.0635 | 0.9630 | 0.3855 | 0.7741 | 1.1002 | 0.3863 |

Table 6: BEVFormer-Small Li et al. (2022b) results.

## A.2 ADDITIONAL VISUALIZATION RESULT

More visualization results of the nuScenes-C dataset can be seen in Figure7.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.2622 | 0.1324 | 0.9352 | 0.3024 | 0.5556 | 1.1106 | 0.2466 |
| Camera Crash | 0.2013 | 0.0425 | 0.9844 | 0.3306 | 0.6330 | 1.0969 | 0.2556 |
| Frame Lost | 0.1638 | 0.0292 | 1.0051 | 0.4294 | 0.6963 | 1.1418 | 0.3954 |
| Color Quant | 0.2313 | 0.1041 | 0.9625 | 0.3131 | 0.6435 | 1.1686 | 0.2882 |
| Motion Blur | 0.1916 | 0.0676 | 0.9741 | 0.3644 | 0.7525 | 1.3062 | 0.3307 |
| Brightness | 0.2520 | 0.1250 | 0.9484 | 0.3034 | 0.6046 | 1.1318 | 0.2486 |
| Low Light | 0.1868 | 0.0624 | 0.9414 | 0.3984 | 0.7185 | 1.3064 | 0.3859 |
| Fog | 0.2442 | 0.1181 | 0.9498 | 0.3055 | 0.6343 | 1.1806 | 0.2592 |
| Snow | 0.1414 | 0.0294 | 1.0231 | 0.4242 | 0.8644 | 1.3622 | 0.4444 |

Table 7: BEVFormer-SingleFrame-Small Li et al. (2022b) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.5174 | 0.4164 | 0.6726 | 0.2734 | 0.3704 | 0.3941 | 0.1974 |
| Cam Crash | 0.3154 | 0.1545 | 0.8015 | 0.2975 | 0.5031 | 0.7865 | 0.2301 |
| Frame Lost | 0.3017 | 0.1307 | 0.8359 | 0.3053 | 0.5262 | 0.7364 | 0.2328 |
| Color Quant | 0.3509 | 0.2393 | 0.8294 | 0.2953 | 0.5200 | 0.8079 | 0.2350 |
| Motion Blur | 0.2695 | 0.1531 | 0.8739 | 0.3236 | 0.6941 | 0.9334 | 0.2592 |
| Brightness | 0.4184 | 0.3312 | 0.7457 | 0.2832 | 0.4721 | 0.7686 | 0.2024 |
| Low Light | 0.2515 | 0.1394 | 0.8568 | 0.3601 | 0.6571 | 1.0322 | 0.3453 |
| Fog | 0.4069 | 0.3141 | 0.7627 | 0.2837 | 0.4711 | 0.7798 | 0.2046 |
| Snow | 0.1857 | 0.0739 | 0.9405 | 0.3966 | 0.7806 | 1.0880 | 0.3951 |

Table 8: BEVFormer-Base Li et al. (2022b) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4129 | 0.3461 | 0.7549 | 0.2832 | 0.4520 | 0.8917 | 0.2194 |
| Cam Crash | 0.2879 | 0.1240 | 0.8041 | 0.2966 | 0.5094 | 0.8986 | 0.2323 |
| Frame Lost | 0.2642 | 0.0969 | 0.8352 | 0.3093 | 0.5748 | 0.8861 | 0.2374 |
| Color Quant | 0.3207 | 0.2243 | 0.8488 | 0.2992 | 0.5422 | 1.0003 | 0.2522 |
| Motion Blur | 0.2518 | 0.1434 | 0.8845 | 0.3248 | 0.7179 | 1.1211 | 0.2860 |
| Brightness | 0.3819 | 0.3093 | 0.7761 | 0.2861 | 0.4999 | 0.9466 | 0.2201 |
| Low Light | 0.2381 | 0.1316 | 0.8640 | 0.3602 | 0.6903 | 1.2132 | 0.3622 |
| Fog | 0.3662 | 0.2907 | 0.7938 | 0.2870 | 0.5162 | 0.9702 | 0.2254 |
| Snow | 0.1793 | 0.0687 | 0.9472 | 0.3954 | 0.8004 | 1.2524 | 0.4078 |

Table 9: BEVFormer-SingleFrame-Base Li et al. (2022b) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.3665 | 0.3174 | 0.8397 | 0.2796 | 0.6158 | 0.9543 | 0.2326 |
| Cam Crash | 0.2320 | 0.1065 | 0.9383 | 0.2975 | 0.7220 | 1.0169 | 0.2585 |
| Frame Lost | 0.2166 | 0.0868 | 0.9513 | 0.3041 | 0.7597 | 1.0081 | 0.2629 |
| Color Quant | 0.2472 | 0.1734 | 0.9121 | 0.3616 | 0.7807 | 1.1634 | 0.3473 |
| Motion Blur | 0.2299 | 0.1378 | 0.9587 | 0.3164 | 0.8461 | 1.1190 | 0.2847 |
| Brightness | 0.2841 | 0.2101 | 0.9049 | 0.3080 | 0.7429 | 1.0838 | 0.2552 |
| Low Light | 0.1571 | 0.0685 | 0.9465 | 0.4222 | 0.9201 | 1.4371 | 0.4971 |
| Fog | 0.2876 | 0.2161 | 0.9078 | 0.2928 | 0.7492 | 1.1781 | 0.2549 |
| Snow | 0.1417 | 0.0582 | 1.0437 | 0.4411 | 1.0177 | 1.3481 | 0.4713 |

Table 10: PETR (r50) Liu et al. (2022a) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4550 | 0.4035 | 0.7362 | 0.2710 | 0.4316 | 0.8249 | 0.2039 |
| Cam Crash | 0.2924 | 0.1408 | 0.8167 | 0.2854 | 0.5492 | 0.9014 | 0.2267 |
| Frame Lost | 0.2792 | 0.1153 | 0.8311 | 0.2909 | 0.5662 | 0.8816 | 0.2144 |
| Color Quant | 0.2968 | 0.2089 | 0.8818 | 0.3455 | 0.5997 | 1.0875 | 0.3123 |
| Motion Blur | 0.2490 | 0.1395 | 0.9521 | 0.3153 | 0.7424 | 1.0353 | 0.2639 |
| Brightness | 0.3858 | 0.3199 | 0.7982 | 0.2779 | 0.5256 | 0.9342 | 0.2112 |
| Low Light | 0.2305 | 0.1221 | 0.8897 | 0.3645 | 0.6960 | 1.2311 | 0.3553 |
| Fog | 0.3703 | 0.2815 | 0.8337 | 0.2778 | 0.4982 | 0.8833 | 0.2111 |
| Snow | 0.2632 | 0.1653 | 0.8980 | 0.3138 | 0.7034 | 1.1314 | 0.2886 |

Table 11: PETR (VoVNet-V2) Liu et al. (2022a) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4602 | 0.3916 | 0.7060 | 0.2718 | 0.3610 | 0.8079 | 0.2093 |
| Cam Crash | 0.3133 | 0.1425 | 0.7746 | 0.2840 | 0.4440 | 0.8524 | 0.2250 |
| Frame Lost | 0.2808 | 0.1134 | 0.8034 | 0.3093 | 0.4981 | 0.8988 | 0.2498 |
| Color Quant | 0.3509 | 0.2538 | 0.8059 | 0.2999 | 0.4812 | 0.9724 | 0.2592 |
| Motion Blur | 0.3221 | 0.2117 | 0.8196 | 0.2946 | 0.5727 | 0.9379 | 0.2258 |
| Brightness | 0.4304 | 0.3574 | 0.7390 | 0.2738 | 0.4149 | 0.8522 | 0.2032 |
| Low Light | 0.2554 | 0.1393 | 0.8418 | 0.3557 | 0.6087 | 1.2004 | 0.3364 |
| Fog | 0.4262 | 0.3518 | 0.7338 | 0.2735 | 0.4143 | 0.8672 | 0.2082 |
| Snow | 0.2304 | 0.1058 | 0.9125 | 0.3363 | 0.6592 | 1.2284 | 0.3174 |

Table 12: PolarFormer (r101) Jiang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4558 | 0.4028 | 0.7097 | 0.2690 | 0.4019 | 0.8682 | 0.2072 |
| Cam Crash | 0.3135 | 0.1453 | 0.7626 | 0.2815 | 0.4519 | 0.8735 | 0.2216 |
| Frame Lost | 0.2811 | 0.1155 | 0.8019 | 0.3015 | 0.4956 | 0.9158 | 0.2512 |
| Color Quant | 0.3076 | 0.2000 | 0.8846 | 0.2962 | 0.5393 | 1.0044 | 0.2483 |
| Motion Blur | 0.2344 | 0.1256 | 0.9392 | 0.3616 | 0.6840 | 1.0992 | 0.3489 |
| Brightness | 0.4280 | 0.3619 | 0.7447 | 0.2696 | 0.4413 | 0.8667 | 0.2065 |
| Low Light | 0.2441 | 0.1361 | 0.8828 | 0.3647 | 0.6506 | 1.2090 | 0.3419 |
| Fog | 0.4061 | 0.3349 | 0.7651 | 0.2743 | 0.4487 | 0.9100 | 0.2156 |
| Snow | 0.2468 | 0.1384 | 0.9104 | 0.3375 | 0.6427 | 1.1737 | 0.3337 |

Table 13: PolarFormer (vov) Jiang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4436 | 0.3677 | 0.7319 | 0.2698 | 0.3890 | 0.8150 | 0.1975 |
| Cam Crash | 0.3055 | 0.1275 | 0.7952 | 0.2803 | 0.4549 | 0.8376 | 0.2145 |
| Frame Lost | 0.2750 | 0.0997 | 0.8362 | 0.3075 | 0.4963 | 0.8747 | 0.2340 |
| Color Quant | 0.3360 | 0.2382 | 0.8479 | 0.2848 | 0.5249 | 0.9516 | 0.2432 |
| Motion Blur | 0.2647 | 0.1527 | 0.8656 | 0.3497 | 0.6251 | 1.0433 | 0.3160 |
| Brightness | 0.4075 | 0.3252 | 0.7740 | 0.2741 | 0.4620 | 0.8372 | 0.2029 |
| Low Light | 0.2613 | 0.1509 | 0.8489 | 0.3445 | 0.6207 | 1.2113 | 0.3278 |
| Fog | 0.3959 | 0.3084 | 0.7822 | 0.2753 | 0.4515 | 0.8685 | 0.2048 |
| Snow | 0.1898 | 0.0757 | 0.9404 | 0.3857 | 0.7665 | 1.2890 | 0.3879 |

Table 14: ORA3D Roh et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.3770 | 0.2987 | 0.7336 | 0.2744 | 0.5713 | 0.9051 | 0.2394 |
| Cam Crash | 0.2486 | 0.0990 | 0.8147 | 0.2975 | 0.6402 | 0.9990 | 0.2842 |
| Frame Lost | 0.1924 | 0.0781 | 0.8545 | 0.4413 | 0.7179 | 1.0247 | 0.4780 |
| Color Quant | 0.2408 | 0.1542 | 0.8718 | 0.3579 | 0.7376 | 1.2194 | 0.3958 |
| Motion Blur | 0.2061 | 0.1156 | 0.8891 | 0.4020 | 0.7693 | 1.1521 | 0.4645 |
| Brightness | 0.2565 | 0.1787 | 0.8380 | 0.3736 | 0.7216 | 1.2912 | 0.3955 |
| Low Light | 0.1102 | 0.0470 | 0.9867 | 0.5308 | 0.9443 | 1.2841 | 0.6708 |
| Fog | 0.2461 | 0.1404 | 0.8801 | 0.3018 | 0.7483 | 1.1610 | 0.3112 |
| Snow | 0.0625 | 0.0254 | 0.9853 | 0.7204 | 1.0029 | 1.1642 | 0.8160 |

Table 15: BEVDet (r50) Huang et al. (2021) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.3877 | 0.3008 | 0.7035 | 0.2752 | 0.5384 | 0.8715 | 0.2379 |
| Cam Crash | 0.2622 | 0.1042 | 0.7821 | 0.3004 | 0.6028 | 0.9783 | 0.2715 |
| Frame Lost | 0.2065 | 0.0805 | 0.8248 | 0.4175 | 0.6754 | 1.0578 | 0.4474 |
| Color Quant | 0.2546 | 0.1566 | 0.8457 | 0.3361 | 0.6966 | 1.1529 | 0.3716 |
| Motion Blur | 0.2265 | 0.1278 | 0.8596 | 0.3785 | 0.7112 | 1.1344 | 0.4246 |
| Brightness | 0.2554 | 0.1738 | 0.8094 | 0.3770 | 0.7228 | 1.3752 | 0.4060 |
| Low Light | 0.1118 | 0.0426 | 0.9659 | 0.5550 | 0.8904 | 1.3003 | 0.6836 |
| Fog | 0.2495 | 0.1412 | 0.8460 | 0.3269 | 0.7007 | 1.1480 | 0.3376 |
| Snow | 0.0810 | 0.0296 | 0.9727 | 0.6758 | 0.9027 | 1.1803 | 0.7869 |

Table 16: BEVDet (r101) Huang et al. (2021) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4037 | 0.3080 | 0.6648 | 0.2729 | 0.5323 | 0.8278 | 0.2050 |
| Cam Crash | 0.2609 | 0.1053 | 0.7786 | 0.3246 | 0.5761 | 0.9821 | 0.2822 |
| Frame Lost | 0.2115 | 0.0826 | 0.8174 | 0.4207 | 0.6710 | 1.0138 | 0.4294 |
| Color Quant | 0.2278 | 0.1487 | 0.8236 | 0.4518 | 0.7461 | 1.1668 | 0.4742 |
| Motion Blur | 0.2128 | 0.1235 | 0.8455 | 0.4457 | 0.7074 | 1.1857 | 0.5080 |
| Brightness | 0.2191 | 0.1370 | 0.8300 | 0.4523 | 0.7277 | 1.2995 | 0.4833 |
| Low Light | 0.0490 | 0.0180 | 0.9883 | 0.7696 | 1.0083 | 1.1225 | 0.8607 |
| Fog | 0.2450 | 0.1396 | 0.8459 | 0.3656 | 0.6839 | 1.2694 | 0.3520 |
| Snow | 0.0680 | 0.0312 | 0.9730 | 0.7665 | 0.8973 | 1.2609 | 0.8393 |

Table 17: BEVDet (swin-tiny) Huang et al. (2021) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4058 | 0.3328 | 0.6633 | 0.2714 | 0.5581 | 0.8763 | 0.2369 |
| Cam Crash | 0.2638 | 0.1111 | 0.7407 | 0.2959 | 0.6373 | 1.0079 | 0.2749 |
| Frame Lost | 0.2141 | 0.0876 | 0.7890 | 0.4134 | 0.6728 | 1.0536 | 0.4498 |
| Color Quant | 0.2751 | 0.1865 | 0.8190 | 0.3292 | 0.6946 | 1.2008 | 0.3552 |
| Motion Blur | 0.2513 | 0.1508 | 0.8320 | 0.3516 | 0.7135 | 1.1084 | 0.3765 |
| Brightness | 0.2879 | 0.2090 | 0.7520 | 0.3646 | 0.6724 | 1.2089 | 0.3766 |
| Low Light | 0.1757 | 0.0820 | 0.8540 | 0.4509 | 0.8073 | 1.3149 | 0.5410 |
| Fog | 0.2903 | 0.1973 | 0.7900 | 0.3021 | 0.6973 | 1.0640 | 0.2940 |
| Snow | 0.0863 | 0.0350 | 0.9529 | 0.6682 | 0.9107 | 1.2750 | 0.7802 |

Table 18: BEVDepth (r50) Li et al. (2022a) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4665 | 0.3214 | 0.6807 | 0.2782 | 0.4657 | 0.3281 | 0.1893 |
| Cam Crash | 0.3181 | 0.1218 | 0.7447 | 0.3545 | 0.5479 | 0.4974 | 0.2833 |
| Frame Lost | 0.3037 | 0.1466 | 0.7892 | 0.3511 | 0.6217 | 0.6491 | 0.2844 |
| Color Quant | 0.2600 | 0.1497 | 0.8577 | 0.4758 | 0.6711 | 0.6931 | 0.4676 |
| Motion Blur | 0.2647 | 0.1456 | 0.8139 | 0.4269 | 0.6275 | 0.8103 | 0.4225 |
| Brightness | 0.2656 | 0.1512 | 0.8120 | 0.4548 | 0.6799 | 0.7029 | 0.4507 |
| Low Light | 0.0593 | 0.0235 | 0.9744 | 0.7926 | 0.9961 | 0.9437 | 0.8304 |
| Fog | 0.2781 | 0.1348 | 0.8467 | 0.3967 | 0.6135 | 0.6596 | 0.3764 |
| Snow | 0.0644 | 0.0251 | 0.9662 | 0.7966 | 0.8893 | 0.9829 | 0.8464 |

Table 19: BEVerse (tiny) Zhang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.1603 | 0.0826 | 0.8298 | 0.5296 | 0.8771 | 1.2639 | 0.5739 |
| Cam Crash | 0.0639 | 0.0165 | 0.9135 | 0.7574 | 0.9522 | 1.1890 | 0.8201 |
| Frame Lost | 0.0508 | 0.0141 | 0.9455 | 0.8181 | 0.9221 | 1.1765 | 0.8765 |
| Color Quant | 0.0642 | 0.0317 | 0.9478 | 0.7735 | 0.9723 | 1.2508 | 0.8397 |
| Motion Blur | 0.0540 | 0.0230 | 0.9556 | 0.8028 | 0.9339 | 1.2137 | 0.8826 |
| Brightness | 0.0683 | 0.0360 | 0.9369 | 0.7315 | 0.9878 | 1.3048 | 0.8531 |
| Low Light | 0.0100 | 0.0005 | 1.0097 | 0.9474 | 1.0048 | 1.1073 | 0.9561 |
| Fog | 0.0402 | 0.0179 | 0.9789 | 0.8230 | 1.0094 | 1.3083 | 0.8962 |
| Snow | 0.0107 | 0.0017 | 1.0021 | 0.9468 | 0.9968 | 1.1652 | 0.9612 |

Table 20: BEVerse-SingleFrame (tiny) Zhang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4951 | 0.3512 | 0.6243 | 0.2694 | 0.3999 | 0.3292 | 0.1827 |
| Cam Crash | 0.3364 | 0.1156 | 0.6753 | 0.3331 | 0.4460 | 0.4823 | 0.2772 |
| Frame Lost | 0.2485 | 0.0959 | 0.7413 | 0.4389 | 0.5898 | 0.8170 | 0.4445 |
| Color Quant | 0.2807 | 0.1630 | 0.8148 | 0.4651 | 0.6311 | 0.6511 | 0.4455 |
| Motion Blur | 0.2632 | 0.1455 | 0.7866 | 0.4399 | 0.5753 | 0.8424 | 0.4586 |
| Brightness | 0.3394 | 0.1935 | 0.7441 | 0.3736 | 0.4873 | 0.6357 | 0.3326 |
| Low Light | 0.1118 | 0.0373 | 0.9230 | 0.6900 | 0.8727 | 0.8600 | 0.7223 |
| Fog | 0.2849 | 0.1291 | 0.7858 | 0.4234 | 0.5105 | 0.6852 | 0.3921 |
| Snow | 0.0985 | 0.0357 | 0.9309 | 0.7389 | 0.8864 | 0.8695 | 0.7676 |

Table 21: BEVerse (small) Zhang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.2682 | 0.1513 | 0.6631 | 0.4228 | 0.5406 | 1.3996 | 0.4483 |
| Cam Crash | 0.1305 | 0.0340 | 0.8028 | 0.6164 | 0.7475 | 1.2273 | 0.6978 |
| Frame Lost | 0.0822 | 0.0274 | 0.8755 | 0.7651 | 0.8674 | 1.1223 | 0.8107 |
| Color Quant | 0.1002 | 0.0495 | 0.8923 | 0.7228 | 0.8517 | 1.1570 | 0.7850 |
| Motion Blur | 0.0716 | 0.0370 | 0.9117 | 0.7927 | 0.8818 | 1.1616 | 0.8833 |
| Brightness | 0.1336 | 0.0724 | 0.8340 | 0.6499 | 0.8086 | 1.2874 | 0.7333 |
| Low Light | 0.0132 | 0.0041 | 0.9862 | 0.9356 | 1.0175 | 0.9964 | 0.9707 |
| Fog | 0.0910 | 0.0406 | 0.8894 | 0.7200 | 0.8700 | 1.0564 | 0.8140 |
| Snow | 0.0116 | 0.0066 | 0.9785 | 0.9385 | 1.0000 | 1.0000 | 1.0000 |

Table 22: BEVerse-SingleFrame (small) Zhang et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4286 | 0.3373 | 0.7783 | 0.2873 | 0.3665 | 0.7806 | 0.1878 |
| Cam Crash | 0.2947 | 0.1172 | 0.8369 | 0.3017 | 0.4403 | 0.8506 | 0.2097 |
| Frame Lost | 0.2681 | 0.0924 | 0.8637 | 0.3303 | 0.4798 | 0.8725 | 0.2349 |
| Color Quant | 0.3318 | 0.2199 | 0.8696 | 0.3041 | 0.4747 | 0.8877 | 0.2458 |
| Motion Blur | 0.2609 | 0.1361 | 0.9026 | 0.3524 | 0.5788 | 0.9964 | 0.2927 |
| Brightness | 0.4074 | 0.3133 | 0.7936 | 0.2911 | 0.3974 | 0.8227 | 0.1877 |
| Low Light | 0.2590 | 0.1406 | 0.8586 | 0.3642 | 0.5773 | 1.1257 | 0.3353 |
| Fog | 0.3940 | 0.2932 | 0.7993 | 0.2919 | 0.3978 | 0.8428 | 0.1944 |
| Snow | 0.1920 | 0.0734 | 0.9372 | 0.3996 | 0.7302 | 1.2366 | 0.3803 |

Table 23: SRCN3D (r101) Shi et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.4205 | 0.3475 | 0.7855 | 0.2994 | 0.4099 | 0.8352 | 0.2030 |
| Cam Crash | 0.2875 | 0.1252 | 0.8435 | 0.3139 | 0.4879 | 0.8897 | 0.2165 |
| Frame Lost | 0.2579 | 0.0982 | 0.8710 | 0.3428 | 0.5324 | 0.9194 | 0.2458 |
| Color Quant | 0.2827 | 0.1755 | 0.9167 | 0.3443 | 0.5574 | 1.0077 | 0.2747 |
| Motion Blur | 0.2143 | 0.1102 | 0.9833 | 0.3966 | 0.7434 | 1.1151 | 0.3500 |
| Brightness | 0.3886 | 0.3086 | 0.8175 | 0.3018 | 0.4660 | 0.8720 | 0.2001 |
| Low Light | 0.2274 | 0.1142 | 0.9192 | 0.3866 | 0.6475 | 1.2095 | 0.3435 |
| Fog | 0.3774 | 0.2911 | 0.8227 | 0.3045 | 0.4646 | 0.8864 | 0.2034 |
| Snow | 0.2499 | 0.1418 | 0.9299 | 0.3575 | 0.6125 | 1.1351 | 0.3176 |

Table 24: SRCN3D (vov) Shi et al. (2022) results.

| Corruption | NDS ↑ | mAP ↑ | mATE ↓ | mASE ↓ | mAOE ↓ | mAVE ↓ | mAAE ↓ |
|---|---|---|---|---|---|---|---|
| Clean | 0.5438 | 0.4409 | 0.6282 | 0.2721 | 0.3853 | 0.2922 | 0.1888 |
| Cam Crash | 0.2873 | 0.1319 | 0.7852 | 0.2917 | 0.4989 | 0.9611 | 0.2510 |
| Frame Lost | 0.2611 | 0.1050 | 0.8175 | 0.3166 | 0.5404 | 1.0253 | 0.2726 |
| Color Quant | 0.3310 | 0.2345 | 0.8348 | 0.2956 | 0.5452 | 0.9712 | 0.2496 |
| Motion Blur | 0.2514 | 0.1438 | 0.8719 | 0.3553 | 0.6780 | 1.0817 | 0.3347 |
| Brightness | 0.3984 | 0.3296 | 0.7543 | 0.2835 | 0.4844 | 0.9232 | 0.2187 |
| Low Light | 0.2510 | 0.1386 | 0.8501 | 0.3543 | 0.6464 | 1.1621 | 0.3356 |
| Fog | 0.3884 | 0.3097 | 0.7552 | 0.2840 | 0.4933 | 0.9087 | 0.2229 |
| Snow | 0.2259 | 0.1275 | 0.8860 | 0.3875 | 0.7116 | 1.1418 | 0.3936 |

Table 25: Sparse4D Lin et al. (2022) results.


Figure 7: More visualization results of nuScenes-C. From left to right: easy, moderate, and hard.