ASF: Robust 3D Object Detection by Solving Sensor Failures

Hai Chen
Tsinghua University
Beijing China
chber_ahu@hotmail.com

Xiao Yang
Tsinghua University
Beijing China
yangxiao19@mails.tsinghua.edu.cn

Lizhong Wang
Tsinghua University
Beijing China
wanglizhong99@outlook.com

Abstract

This paper describes the methodology and results of Track 5 - Robust Multi-Modal BEV Detection in the 2024 RoboDrive Challenge. This track focuses on 3D scene perception robustness under the camera and LiDAR sensor failures, where sensor failures are included: (1) loss of certain camera frames during the driving system sensing process; (2) loss of one or more camera views during the driving system sensing process; (3) loss of the roof-top LiDAR view during the driving system sensing process. To improve the robustness of the 3D object detection model under these conditions, we propose a novel method called Against Sensor Failure, abbreviated as ***ASF****. ASF utilizes self-supervised methods to reconstruct image features when facing camera sensor failures. In addition, we also propose the* ***I****mage feature* ***E****nhancement* ***L****iDAR feature (****IEL****) module, designed to alleviate the negative impact of LiDAR sensor failure. Our results demonstrate obvious improvements over the baseline, with the ASF method elevating the nuScenes Detection Score (NDS) from 39.13 to 49.68 and the mean Average Precision (mAP) from 21.59 to 39.46.*

1. Introduction

In the rapidly evolving domain of autonomous driving, the accuracy and resilience of perception systems are paramount [1–6]. Recent advancements, particularly in bird's eye view (BEV) representations and LiDAR sensing technologies, have significantly improved in-vehicle 3D scene perception. Yet, the robustness of 3D scene perception methods under varied and challenging conditions — integral to ensuring safe operations — has been insufficiently assessed [7, 8]. Therefore, the **2024 RoboDrive Challenge** [9] breaks this limitation and promotes the development of more robust autonomous driving perception algorithms.

The competition was centered around two prevailing themes: common corruptions and sensor failures [7]. We focus on the second theme, sensor failures. Sensor failures are an inevitable problem for autonomous vehicles in real-world scenarios, and they are also one of the situations that can easily arise to endanger the safety of autonomous vehicles. Thanks to the powerful expressive ability of BEV, more and more BEV-based 3D object detection models emerge and show strong perceptual capabilities. Typically, BEVDet [10] transforms the ring-viewed 2D image features into BEV features by View Transformer, and then implements further feature extraction on the BEV features to achieve effective performance under the camera-only 3D object detection task. After that, more BEV-based 3D object detection algorithms appeared, such as BEVDepth [11] that adds depth supervision, BEVFormer [12] that introduces temporal information, and so on Since the features obtained from LiDAR data can be easily converted to BEV representation, the BEV-based multi-modal 3D object detection model was proposed [13–16]. One of the most representative works is BEVFusion [14], which fused image features and point cloud features in BEV space and achieved excellent performance on several tasks [17].

The rapid development of BEV-based 3D object detection models has also brought further thinking, that is, how safe is it? Zhu et al. [18] conducted a detailed study and analysis of BEV-based 3D object detection models under natural corruptions and adversarial attacks. Recently, RoboBEV [19] comprehensively evaluated the robustness of BEV-based 3D object detection models under natural corruptions. Robo3D [20] has been further extended to analyze the robustness of the 3D perception algorithms for more hazardous conditions, including severe weather conditions, data blurring due to external disturbances, and internal sensor failure. One of the more practical situations that autonomous vehicles will face is sensor failure, as practical situations such as deterioration, bumps, etc. may cause a particular sensor failure, such as the camera sensor not being able to capture its surroundings. On the other hand, mutual compensation between sensors can somewhat mitigate the effects of a particular sensor failure,

Technical Report of the 2024 RoboDrive Challenge.
Track 5: Robust Multi-Modal BEV Detection.

so it is important to investigate the boundaries of multimodal 3D object detection models under sensor failures.

There are some recent studies that are beginning to focus on this topic. For example, M-BEV [21] tackled camera sensor failures by reconstructing image features of a particular failed camera sensor. However, the method has not considered other gains or effects from the introduction of LiDAR sensors. Considering the low performance of LiDAR sensors and camera sensors due to damage or failures, MetaBEV [22] addressed sensor failures through the meta-BEV query and BEV-evolving decoder of the setup. However, research for sensor failures is still in its infancy, especially studying the robustness of multi-modal 3D object detection models in this setting [23]. Although MetaBEV began its research, it did not demonstrate high performance and still falls short of expectations.

Therefore, attracted by **The RoboDrive Challenge**, we focused on Track 5-Robust Multi-Modal BEV Detection. Not limited to us, the competition has attracted attention from major universities and companies. We proposed a new multi-modal 3D object detection model against sensor failures in this competition, named ASF. The method significantly improves the performance of the model under sensor failures, greatly exceeding the baseline in the competition, with 26.96% improvement in NDS and 82.77% improvement in mAP.

2. Approach

In this section, we describe the proposed ASF in detail. Our method is based on CMT [24] improved by adding the self-supervised pre-training process for the image pipeline and proposing an **I**mage feature **E**nhancement **L**iDAR feature (**IEL**) module, which significantly improves the robustness of the model under sensor failures, and which we rename as ASF. The detailed pipeline is shown in Figure 1.

Self-supervised pre-training. In this process, the main solution is proposed for the camera sensor failure. Following CMT, we use VoVNet [25] as the image backbone to extract the 2D image features. In the nuScenes [2] dataset, the 6 ring-view cameras capture the surrounding environment, and there may be unavoidable failures of the 6 cameras, i.e., one or more of them fail. Inspired by M-BEV [21], we designed the self-supervised pre-training process for image pipelines to face this situation. Specifically, we randomly mask a certain 2D image feature, name the masked-off feature as V_{mask}, and then initialize the V_{mask} using the spatial feature cues around it. Then, a similar method in [26] was utilized to generate the 2D positional embedding P_{2D} for it, and finally, the feature is represented as:

$$U_{mask} = V_{mask} + P_{2D} \tag{1}$$

Finally, we stacked multiple layers of transformers to regenerate the U_{mask}, and multiple layers of transformers form the self-supervised pretrained decoder module:

$$U_{mask} = decoder(U_{mask}) \tag{2}$$

Where the decoder consists of a self-attention mechanism and cross-attention mechanism, the cross-attention mechanism reconstructs the U_{mask} from the surrounding spatial cue features, and the self-attention mechanism further helps in the U_{mask} reconstruction process. Finally, we minimize the L2 loss between the original feature F_{mask} and the reconstructed feature U_{mask}:

$$\mathcal{L}_{pre} = \|F_{mask} - U_{mask}\|_2 \tag{3}$$

IEL. Typically, LiDAR sensors deteriorate 3D perception due to practicalities such as incomplete echoes or non-detection of dark-colored instances (e.g., black cars) and crosstalk between multiple sensors. To address the situation, we first reduced the point cloud in the nuScenes [2] from 32 to 16 lines, significantly increasing the sparsity of the point cloud as a way to simply simulate LiDAR sensor failures. However, simulating this case would result in more than 0 elements in the voxelized 3D point cloud features, i.e., many meaningful features are discarded. Therefore, we consider the use of image features to enhance the point cloud features. Specifically, we augment the point cloud features by projecting the point cloud onto the image coordinate system and taking out the features of the pixels corresponding to the point cloud located on the image. However, there are a large number of background points (as shown in Figure 2), which not only consumes a huge graphic memory space but also wastes training time. Therefore, we project the reference point in the CMT onto the image and take out only the features of the N pixel points around the reference point to enhance the LiDAR features. This operation not only allows the taken-out features to gather on the object but also ensures the disturbance of redundant noise information. Finally, we stack multiple layers of cross-attention mechanisms to enhance LiDAR features.

The rest of the structure of the ASF remains consistent with the CMT, so the training of the ASF is divided into two phases, first self-supervised pre-training and then end-to-end training. In the inference phase, the ASF removes the masking strategy from the self-supervised pre-training.

3. Experiments

In this section, we present the detailed experimental setup and the results of the competition.

3.1. Experimental Setups

This work follows the protocol in the 2024 RoboDrive Challenge [9] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [2] (which contains 700 scenes) and tested

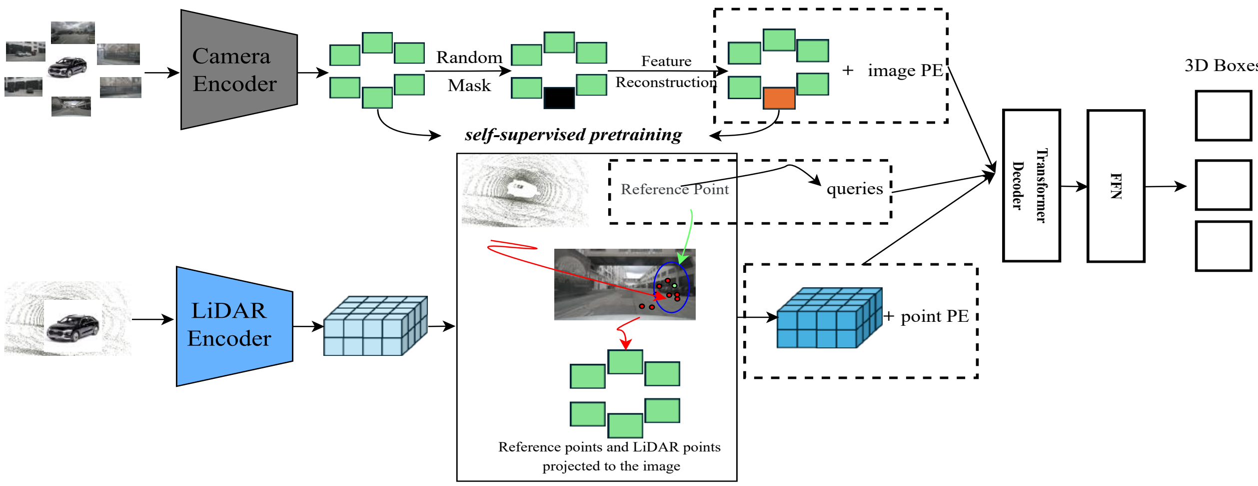


Figure 1. The framework of ASF

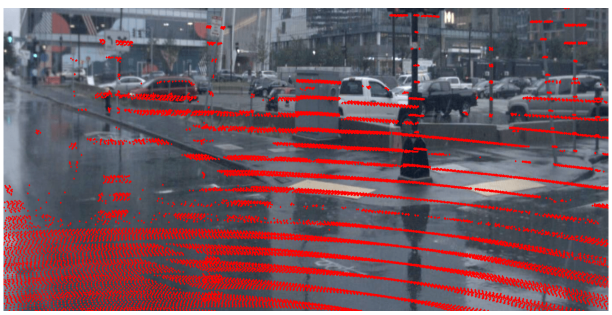


Figure 2. Visualization of point cloud projects to images

on the held-out competition evaluation sets (which contains 150 scenes). The evaluation data was created following RoboDepth [27–29], RoboBEV [7, 19, 30], and Robo3D [20, 31]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. For more details, please refer to the corresponding GitHub repositories.

3.2. Evaluation Metrics

The mean Average Precision (mAP) and nuScenes Detection Score (NDS) are the default evaluation metrics for the nuScenes dataset [2], where half of the NDS is based on mAP and the other half is based on the quality of detection of mean Average Translation Error (mATE), mean Average Orientation Error (mAOE), mean Average Velocity Error (mAVE), mean Average Attribute Error (mAAE) and mean Average Scale Error (mASE). Following the nuScenes evaluation and the Track 5 evaluation, we still use these metrics to measure the effectiveness of the ASF.

Table 1. Comparison results with nuScenes validation set.

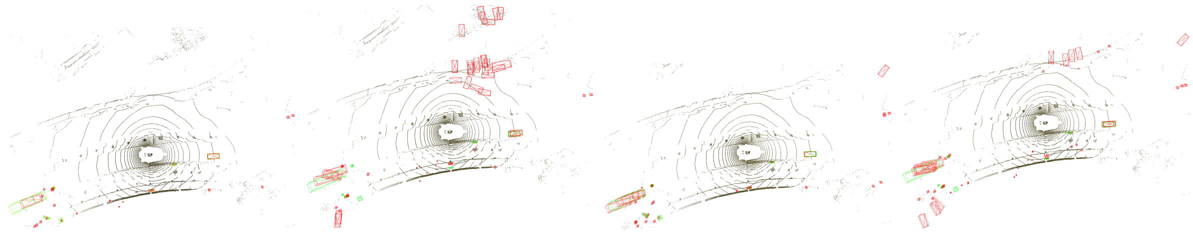
Models	Modality	mAP	NDS
BEVFusion [14]	C+L	68.5	71.4
FocalFormer3D [36]	C+L	70.5	73.1
CMT [24]	C+L	70.3	72.9
ASF	C+L	67.8	71.3

3.3. Implementation Details

The framework is implemented using the PyTorch framework [32] and is based on the MMDetection3D codebase [33]. We use the pre-trained weights provided by CMT as the initial weights of the ASF. Freezing the image encoder during self-supervised pre-training. To reconstruct the image features, we stacked 6 layers of decoer and each layer of decoer consisted of cross-attention and self-attention. In this process, we set the learning rate to 0.0001, the batch size to 20, and the number of iterations to 48. At the end of this phase of pre-training, the relevant parameters are frozen and no parameter updates are performed in subsequent end-to-end training. In end-to-end training, we train the ASF for 20 epochs at the learning rate of 0.0001 with a batch size of 10. Note that we used CBGS [34] to load the data and the AdamW [35] optimizer for optimization. The GT sample augmentation was used for the first 15 epochs and turned off for the last 5 epochs. In addition, we set N to 10 in the IEL module.

3.4. Comparative Study

We validated our proposed method on the vanilla nuScenes validation set and the nuScenes validation set proposed by Track 5, respectively. BEVFusion [14] serves as the baseline in Track 5, and the rest of the methods are state-of-the-art



(a) (b) (c) (d)

Figure 3. The 3D perception results with a confidence score greater than 0.1. Figure 2(a) and Figure 2(b) show the results of CMT under the vanilla nuScenes validation set and the nuScenes validation set provided by Track 5, respectively. Figure 2(c) and Figure 2(d) show the results of ASF under the vanilla nuScenes validation set and the nuScenes validation set provided by Track 5, respectively. Note that the green box represents the ground truth 3D bounding box, and the red box represents the predicted 3D bounding box.

Table 2. Comparison results under the nuScenes validation set provided by Track 5

Model	Modality	mAP	NDS
BEVFusion [14]	C+L	21.6	39.1
FocalFormer3D [36]	C+L	27.1	43.2
CMT [24]	C+L	36.4	48.2
ASF	C+L	**39.5**	**49.7**

multi-modal 3D object detection models.

It can be found from Table 1 that ASF achieves the worst clean performance, which is mainly due to the fact that ASF reduces the 32-line LiDAR to 16-line LiDAR at the point cloud input, thus leading to some performance degradation. However, comparable results to these methods have been achieved, with the best clean performance of FocalFormer3D dropping from 73.1 to 71.3 on the NDS. Surprisingly, our proposed ASF shows the most robust performance under the nuScenes validation set under sensor failure. On the other hand, FocalFormer3D no longer shows surprising results, but instead exposes serious potential threats. Although our approach loses weak clean performance, it greatly improves the robustness under sensor failure, which is acceptable and satisfactory.

In addition, we visualize the perceptual results of CMT and ASF under the vanilla nuScenes validation set and the nuScenes validation set provided by Track 5. It can be seen from Figure 3 that a large number of false positive instances occur under sensor failure compared to normal conditions, and are accompanied by a certain amount of detection bias and missed detections. ASF has fewer instances of false positives and relatively low detection bias compared to CMT. It is also confirmed in Table 2 that ASF is more robust under sensor failure.

4. Conclusion

In Track 5-Robust Multi-Modal BEV Detection in the **2024 RoboDrive Challenge**, we present a 3D object detection model against sensor faults, called ASF. ASF significantly improves perceptual robustness under sensor failures. ASF proposes two main core components. One is a self-supervised pre-training process proposed for camera sensor failures, which mitigates the performance degradation associated with the situation by reconstructing the features of the failed camera. The other is the IEL module that enhances LiDAR features with image features to face LiDAR sensor failures. Compared to the baseline in Track 5, ASF significantly improves the robustness of the 3D object detection model under sensor failure, and ASF is located in the first place on the demonstrated leaderboard.

References

[1] 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.

[2] 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*, pages 11621–11631, 2020.

[3] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7020–7030, 2023.

[4] 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*, pages 2446–2454, 2020.

[5] Youquan Liu, Lingdong Kong, Jun Cen, Runnan Chen, Wenwei Zhang, Liang Pan, Kai Chen, and Ziwei Liu. Segment any point cloud sequences by distilling vision foundation models. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

[6] Lingdong Kong, Youquan Liu, Runnan Chen, Yuexin Ma, Xinge Zhu, Yikang Li, Yuenan Hou, Yu Qiao, and Ziwei Liu. Rethinking range view representation for lidar segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 228–240, 2023.

[7] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *arXiv preprint arXiv:2405.17426*, 2024.

[8] Lingdong Kong, Jiawei Ren, Liang Pan, and Ziwei Liu. Lasermix for semi-supervised lidar semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21705–21715, 2023.

[9] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Yaru Niu, Wei Tsang Ooi, Benoit R. Cottereau, Lai Xing Ng, Yuexin Ma, Wenwei Zhang, Liang Pan, Kai Chen, Ziwei Liu, Weichao Qiu, Wei Zhang, Xu Cao, Hao Lu, Ying-Cong Chen, Caixin Kang, Xinning Zhou, Chengyang Ying, Wentao Shang, Xingxing Wei, Yinpeng Dong, Bo Yang, Shengyin Jiang, Zeliang Ma, Dengyi Ji, Haiwen Li, Xingliang Huang, Yu Tian, Genghua Kou, Fan Jia, Yingfei Liu, Tiancai Wang, Ying Li, Xiaoshuai Hao, Yifan Yang, Hui Zhang, Mengchuan Wei, Yi Zhou, Haimei Zhao, Jing Zhang, Jinke Li, Xiao He, Xiaoqiang Cheng, Bingyang Zhang, Lirong Zhao, Dianlei Ding, Fangsheng Liu, Yixiang Yan, Hongming Wang, Nanfei Ye, Lun Luo, Yubo Tian, Yiwei Zuo, Zhe Cao, Yi Ren, Yunfan Li, Wenjie Liu, Xun Wu, Yifan Mao, Ming Li, Jian Liu, Jiayang Liu, Zihan Qin, Cunxi Chu, Jialei Xu, Wenbo Zhao, Junjun Jiang, Xianming Liu, Ziyan Wang, Chiwei Li, Shilong Li, Chendong Yuan, Songyue Yang, Wentao Liu, Peng Chen, Bin Zhou, Yubo Wang, Chi Zhang, Jianhang Sun, Hai Chen, Xiao Yang, Lizhong Wang, Dongyi Fu, Yongchun Lin, Huitong Yang, Haoang Li, Yadan Luo, Xianjing Cheng, and Yong Xu. The robodrive challenge: Drive anytime anywhere in any condition. *arXiv preprint arXiv:2405.08816*, 2024.

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

[11] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023.

[12] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022.

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

[14] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *IEEE International Conference on Robotics and Automation*, pages 2774–2781. IEEE, 2023.

[15] Zequn Qin, Jingyu Chen, Chao Chen, Xiaozhi Chen, and Xi Li. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view. In *IEEE/CVF International Conference on Computer Vision*, pages 8690–8699, 2023.

[16] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21960–21969, 2023.

[17] Lingdong Kong, Niamul Quader, and Venice Erin Liong. Conda: Unsupervised domain adaptation for lidar segmentation via regularized domain concatenation. In *IEEE International Conference on Robotics and Automation*, pages 9338–9345, 2023.

[18] Zijian Zhu, Yichi Zhang, Hai Chen, Yinpeng Dong, Shu Zhao, Wenbo Ding, Jiachen Zhong, and Shibao 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*, pages 21600–21610, 2023.

[19] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Robobev: Towards robust bird's eye view perception under corruptions. *arXiv preprint arXiv:2304.06719*, 2023.

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

[21] Siran Chen, Yue Ma, Yu Qiao, and Yali Wang. M-bev: Masked bev perception for robust autonomous driving. In *AAAI Conference on Artificial Intelligence*, pages 1183–1191, 2024.

[22] Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *IEEE/CVF International Conference on Computer Vision*, pages 8721–8731, 2023.

[23] Lingdong Kong, Xiang Xu, Jiawei Ren, Wenwei Zhang, Liang Pan, Kai Chen, Wei Tsang Ooi, and Ziwei Liu. Multi-modal data-efficient 3d scene understanding for autonomous driving. *arXiv preprint arXiv:2405.05258*, 2024.

[24] Junjie Yan, Yingfei Liu, Jianjian Sun, Fan Jia, Shuailin Li, Tiancai Wang, and Xiangyu Zhang. Cross modal transformer: Towards fast and robust 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 18268–18278, 2023.

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

[26] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022.

[27] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, and Wei Tsang Ooi. Robodepth: Robust out-of-distribution depth estimation under corruptions. *Advances in Neural Information Processing Systems*, 36, 2024.

[28] Lingdong Kong, Shaoyuan Xie, Hanjiang Hu, Benoit Cottereau, Lai Xing Ng, and Wei Tsang Ooi. The robodepth benchmark for robust out-of-distribution depth estimation under corruptions. https://github.com/ldkong1205/RoboDepth, 2023.

[29] Lingdong Kong, Yaru Niu, Shaoyuan Xie, Hanjiang Hu, Lai Xing Ng, Benoit Cottereau, Ding Zhao, Liangjun Zhang, Hesheng Wang, Wei Tsang Ooi, Ruijie Zhu, Ziyang Song, Li Liu, Tianzhu Zhang, Jun Yu, Mohan Jing, Pengwei Li, Xiaohua Qi, Cheng Jin, Yingfeng Chen, Jie Hou, Jie Zhang, Zhen Kan, Qiang Lin, Liang Peng, Minglei Li, Di Xu, Changpeng Yang, Yuanqi Yao, Gang Wu, Jian Kuai, Xianming Liu, Junjun Jiang, Jiamian Huang, Baojun Li, Jiale Chen, Shuang Zhang, Sun Ao, Zhenyu Li, Runze Chen, Haiyong Luo, Fang Zhao, and Jingze Yu. The robodepth challenge: Methods and advancements towards robust depth estimation. *arXiv preprint arXiv:2307.15061*, 2023.

[30] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robobev benchmark for robust bird's eye view detection under common corruption and domain shift. https://github.com/Daniel-xsy/RoboBEV, 2023.

[31] Lingdong Kong, Youquan Liu, Xin Li, Runnan Chen, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. The robo3d benchmark for robust and reliable 3d perception. https://github.com/ldkong1205/Robo3D, 2023.

[32] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

[33] MMDetection3D Contributors. MMDetection3D: OpenMMLab next-generation platform for general 3D object detection. https://github.com/open-mmlab/mmdetection3d, 2020.

[34] 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.

[35] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[36] Yilun Chen, Zhiding Yu, Yukang Chen, Shiyi Lan, Anima Anandkumar, Jiaya Jia, and Jose M Alvarez. Focalformer3d: focusing on hard instance for 3d object detection. In *IEEE/CVF International Conference on Computer Vision*, pages 8394–8405, 2023.