Cross-Modal Transformers for Robust Multi-Modal BEV Detection

Caixin Kang[1] Xinning Zhou[2] Chengyang Ying[2] Wentao Shang[3]
Xingxing Wei[1,*] Yinpeng Dong[2,*]

[1]Beihang University [2]Tsinghua University [3] Hefei University of Technology
{caixinkang,xxwei}@buaa.edu.cn;
{zxn21,ycy21,dongyinpeng}@mails.tsinghua.edu.cn;

Abstract

In this paper, we elaborate on the practical application and demonstration of Cross-Modal Transformer (CMT) for Track 5 – Robust Multi-Modal BEV Detection, in the 2024 RoboDrive Challenge. Track 5 mainly motivates the development of robust multi-modal 3D object detection models under sensor failures for safe perception of autonomous driving. Without explicit view transformations, CMT takes images and point cloud tokens as input and outputs accurate 3D bounding boxes directly. The simple structural design of the model achieved excellent performance in this track, with an improvement of 23.13% on NDS and 68.36% on mAP, respectively, compared to the baseline of Track5.

1. Introduction

Autonomous vehicles are usually deployed with LiDAR sensors, camera sensors, radar sensors, *etc*. [1]. The data collected by LiDAR sensors and camera sensors would be the most used by autonomous driving perception algorithms today [2–4]. The 2D image data captured by camera sensors contains rich texture information, and the 3D point cloud data collected by LiDAR sensors contains geometrical information that expresses the surrounding objects, and the two sensors mutually complement each other, thus becoming the data source for many multi-modal 3D object detection models [5–15].

Based on this fact, the **2024 RoboDrive Challenge** [16] considers in depth the problem of robustness of multi-modal 3D object detection models in the presence of sensor failures, provoking more researchers to develop suitable detection frameworks to handle this natural and realistic situation. The Out-of-Distribution (OOD) data under sensor failure is specifically designed in Track 5, including: (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. More and more methods choose to fuse multi-modal features under the BEV space based on the advantages of BEV unified representation. Typically, BEVFusion [17], and UniBEV [18] all fused image features and point cloud features in BEV space and achieved excellent perceptual performance [19].

Although BEV-based multi-modal 3D object detection models have achieved promising perceptual performance, there has been a relative lack of in-depth research in the face of reality under sensor failures. For example, what are the consequences of losing a particular camera view? Intuitively, such a situation would lead to a degradation of perceptual performance, which in turn causes safety accidents, which is unacceptable for autonomous driving. Therefore, this is the subject of Track 5 of the 2024 RoboDrive Challenge [16], *i.e.*, the design of more robust multi-modal 3D object detection models under sensor failures. Some work already exists to start focusing on the robustness of multi-modal 3D object detection models under sensor failures. RobustBEV [20] presented Y-mode and λ-mode camera sensor failure scenarios to evaluate the robustness of 3D object detection models and found that there are 3D object detection models with huge performance degradation. Robo3D [21] proposed more scenarios of sensor failures and evaluated the robustness of a large number of 3D object detection models, all of which came to the consistent conclusion that 3D object detection models struggle to cope with sensor failures.

Only comparatively little work has been done to develop robust 3D object detection algorithms to address sensor failures. M-BEV [22] simulated camera sensor failures by masking the camera feature and then reconstructing that camera feature, but lacked research on LiDAR sensor failures. Al-

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

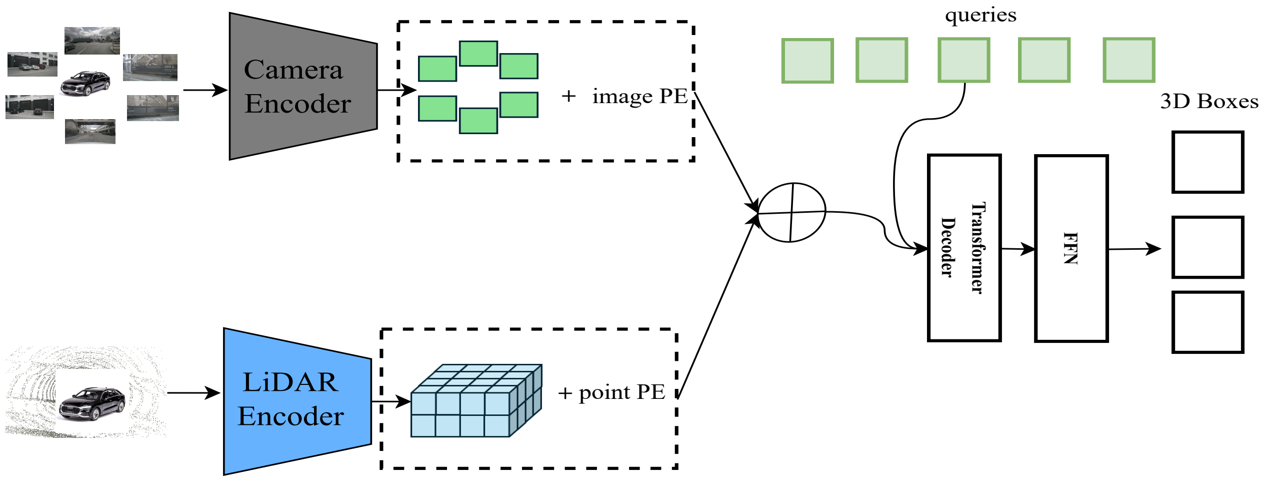

Figure 1. The framework of the CMT.

though CMT [23] was not developed to handle sensor failures, it was found that CMT still has better robustness when missing camera sensors or LiDAR sensors. We believe that the robustness is mainly due to its designed mask training strategy, i.e., randomly masking the image or point cloud during the training process.

Based on the above findings, we applied CMT to the data on sensor failures presented in Track5 and achieved excellent results. Specifically, CMT improved NDS from 39.13 to 48.18 and mAP from 21.59 to 36.35 compared to the baseline in Track5.

2. Approach

The overall architecture of CMT is shown in Figure 1. The ring-view camera images and LiDAR point cloud data are extracted with multi-modal tokens through two individual backbone networks. 3D coordinates are encoded into the multi-modal tokens via coordinate encoding module. Queries from the position-guided query generator are used to interact with the multi-modal tokens in the Transformer decoder, which then predicts the object class as well as the 3D bounding box.

Coordinates Encoding Module. Along the lines of what was done in PETR [24], the CMT generates coordinate encoding for the image. Since 3D point cloud data comes with spatial information, it is easier for coordinate encoding relative to images, and the CMT can directly sample along the Z-axis to further generate positional embeddings.

Position-guided Query Generator. Inspired by Anchor-DETR [25] and PETR [24], CMT initializes n reference points, i.e., anchor points. These anchor points were then transformed into the 3D world space by a linear transformation. Finally, these 3D anchor points were projected onto different modalities and the corresponding point sets were encoded by the coordinate encoding module. Thus, the positional embedding of the object query in CMT was obtained by summing up the point set embeddings of the different modalities.

The decoders in CMT used the original Transformer decoder in DETR [25] with the L decoder layer. For each decoder layer, the position-guided query interacts with the multi-modal token and updates its representation. Two feed-forward networks (FFNs) are used to predict the 3D bounding boxes and classes using the updated queries. Then bipartite matching was used for prediction, focal loss was used for classification, and L1 loss was used for 3D bounding box regression.

3. Experiments

This section details the dataset used, as well as detailed validation results.

3.1. Experimental Setups

This work follows the protocol in the 2024 RoboDrive Challenge [16] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [2] and tested on the held-out competition evaluation sets.

The nuScenes dataset [2] is a large-scale autonomous driving dataset with 3d object annotations. It has a full sensor suite (1 LiDAR, 5 RADAR, 6 cameras, IMU, GPS) with 1,000 scenes of 20 seconds each, with 1,400,000 camera images and 390,000 LiDAR sweeps. The data was collected from two different cities: Boston and Singapore, with detailed map information, 1.4M 3D bounding boxes, and manually annotated visibility, activity, and pose attributes for 23 object classes.

The evaluation data was created following RoboDepth

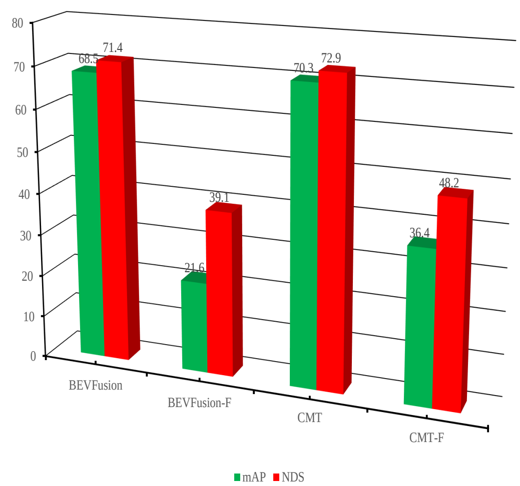

Figure 2. Comparison results on the vanilla nuScenes validation set and the nuScenes validation set proposed by Robodrive Track5.

[26–28], RoboBEV [19, 29, 30], and Robo3D [21, 31]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. To better simulate sensor failures, Track 5 reprocessed the data to include: (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. 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). The track follows these metrics to evaluate the perceptual capabilities of the model, with larger NDS and mAP indicating better perceptual performance.

3.3. Implementation Details

The framework is implemented using the PyTorch framework [32] and is based on the MMDetection3D codebase [33]. The implementation of CMT simply follows the setup in the paper and the model weights used in this competition are pre-trained model weights. We modified the inputs to the model appropriately to satisfy Track5's data.

3.4. Comparative Study

We compared the performance of CMT to the baseline in Track5, where "-F" denotes the performance of the model under the nuScenes validation set for sensor failures provided in Track5. The detailed results are shown in Figure 2, where firstly it can be observed that CMT outperforms BEVFusion both in terms of clean performance and perceived performance under sensor failures. Secondly, it can be seen that the performance degradation of CMT under sensor failure is much lower than BEVFusion, e.g., CMT degraded by 33.88% on NDS while BEVFusion degraded by 45.24%.

4. Conclusion

Based on the finding that CMT maintained excellent performance with missing camera data or point cloud data, we applied it to Track 5-Robust Multi-Modal BEV Detection in **The RoboDrive Challenge**. Surprisingly, CMT improved NDS from 39.13 to 48.18 and mAP from 21.59 to 36.35 compared to the baseline in Track5.

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] 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.

[4] 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.

[5] Zhe Liu, Tengteng Huang, Bingling Li, Xiwu Chen, Xi Wang, and Xiang Bai. Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[6] 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.

[7] 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.

[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] H Hu, F Wang, J Su, Y Wang, L Hu, W Fang, J Xu, and Z Zhang. Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection. *arXiv preprint arXiv:2303.17895*, 2, 2023.

[10] Zehui Chen, Zhenyu Li, Shiquan Zhang, Liangji Fang, Qinghong Jiang, Feng Zhao, Bolei Zhou, and Hang Zhao. Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection. *arXiv preprint arXiv:2201.06493*, 2022.

[11] 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.

[12] Yingwei Li, Adams Wei Yu, Tianjian Meng, Ben Caine, Jiquan Ngiam, Daiyi Peng, Junyang Shen, Yifeng Lu, Denny Zhou, Quoc V Le, et al. Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17182–17191, 2022.

[13] 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.

[14] Xiaopei Wu, Liang Peng, Honghui Yang, Liang Xie, Chenxi Huang, Chengqi Deng, Haifeng Liu, and Deng Cai. Sparse fuse dense: Towards high quality 3d detection with depth completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5418–5427, 2022.

[15] 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.

[16] 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.

[17] 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, 2023.

[18] Shiming Wang, Holger Caesar, Liangliang Nan, and Julian FP Kooij. Unibev: Multi-modal 3d object detection with uniform bev encoders for robustness against missing sensor modalities. *arXiv preprint arXiv:2309.14516*, 2023.

[19] 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.

[20] 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.

[21] 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.

[22] Siran Chen, Yue Ma, Yu Qiao, and Yali Wang. M-bev: Masked bev perception for robust autonomous driving. In *AAAI Conference on Artificial Intelligence*, volume 38, pages 1183–1191, 2024.

[23] 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.

[24] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *European Conference on Computer Vision*, pages 531–548. Springer, 2022.

[25] Yingming Wang, Xiangyu Zhang, Tong Yang, and Jian Sun. Anchor detr: Query design for transformer-based detector. In *AAAI Conference on Artificial Intelligence*, volume 36, pages 2567–2575, 2022.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[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.