TSMA-BEV: Towards Robust Multi-Camera 3D Object Detection through Temporal Sequence Mix Augmentation

Xu Cao[1,2] Hao Lu[1,2] Ying-Cong Chen[1,2]

[1]The Hong Kong University of Science & Technology (Guangzhou)
[2]The Hong Kong University of Science & Technology
xcao635@connect.hkust-gz.edu.cn, hlu585@connect.ust.hk, yingcongchen@ust.hk

Abstract

The advent of bird's-eye view (BEV) representation has witnessed significant advancements in camera-only 3D object detection. However, existing approaches usually struggle when applied to various corruptions that deviate from the original training domain. To address these vulnerabilities, we propose a novel framework, ***TSMA-BEV****, which combines a new image augmentation module* ***AugFFT*** *based on fast fourier transformation (FFT) with a mix-sequence augmentation strategy* ***SeqMixAug*** *to enhance the robustness and adaptability of 3D object detection algorithms. The proposed* ***AugFFT****, involves stochastic frequency cutoffs and amplitude scaling to generate augmented images, while* ***SeqMixAug*** *extends this augmentation to temporal sequences, maintaining consistency across frames. This approach ensures improved performance stability in the face of multiple corruptions. As demonstrated in our experiments, the effectiveness and superiority of* ***TSMA-BEV*** *in handling real-world corruptions are verified.*

1. Introduction

Multi-camera 3D object detection involves the identification and localization of objects within a 3D space, utilizing data derived from multiple cameras [1]. In recent years, significant advancements have been witnessed in the BEV representation-based camera-only 3D detection methods, showcasing remarkable breakthroughs across various challenging benchmarks [2–5]. In comparison to LiDAR-based methods [6–8], camera-only approaches are increasingly favored due to their cost-efficiency, computational effectiveness, and the provision of detailed semantic information. The BEV representation is pivotal in these advancements, which provides a unified learned representation of multi-view images, enables the interpretable fusion of information across different sensors and temporal instances, and is well-aligned with downstream applications. Consequently, BEV-based 3D perception approaches have attracted attention from both academia and the industry.

Despite the promising performance of off-the-shelf approaches on benchmarks such as nuScenes [9] and Waymo [10], the stability of these algorithms trained on single-domain data becomes a concern when confronted with out-of-domain or unseen scenarios, which poses challenges for meeting the high demands of autonomous driving in real-world scenarios [11, 12]. Robustness under common corruptions such as sensor failure, different noises, and severe weather conditions which usually occur in real driving scenarios, are vital for real-world applications, including autonomous driving [13], surveillance, and robotics [14].

To improve the robustness of the 3D perception algorithm and alleviate the performance degradation on out-of-domain data, various approaches have been proposed. Domain generalization (DG) addresses the specific camera attributes and environmental variables over-fitting problem by decoupling and eliminating the domain-specific factors, thus improving the general performance across various scenarios [15–17] . Unsupervised domain adaption (UDA) eliminates domain shift by generating pseudo-labels or aligning latent feature representations, which achieved considerable results in Sim2Real tasks [18–20].

This paper presents a novel approach to 3D object detection in the context of the 2024 RoboDrive Challenge [23] (in conjunction with **ICRA 2024**). The competition requires the development of an algorithm leveraging only nuScenes [9] training data, yet effectively dealing with multiple out-of-domain corruptions including severe weather conditions, sensor failure, and different noises, etc. To improve the robustness of the 3D object detection model under various out-of-domain corruptions, we propose **TSMA-BEV**, a

Technical Report of the 2024 RoboDrive Challenge.
Track 1: Robust BEV Detection.

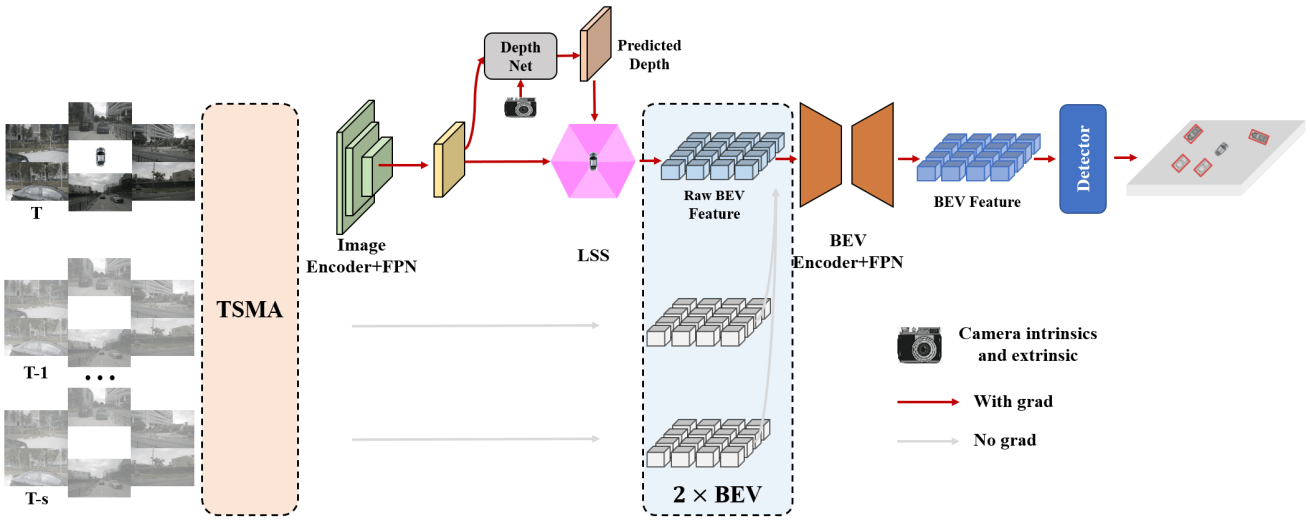

Figure 1. Overall framework of our **TSMA-BEV**. Multi-view images are randomly augmented according to sequence splits and fed into an image encoder to extract 2D image features. A view transformer [21] transforms 2D features to 3D space guided by estimated depth. To attain temporal information, the volume features from current and previous frames are aligned and concatenated. Consequently, an BEV encoder and CenterHead [22] predict object categories and locations.

novel framework consisting of an image augmentation module based on FFT and a sequence consistent augmentation strategy.

- **AugFFT:** For the augmentation module, we combine stochastic frequency cut-off and amplitude scaling operations in the frequency domain based on FFT to generate random augmented images.
- **SeqMixAug:** Different from common image augmentation in 3D object detection tasks, the sequence augmentation strategy we propose is not applied to a single frame, but to a temporal sequence. The augmentation configuration between adjacent frames is consistent, which avoids the performance degradation caused by random augmentation settings between adjacent frames, especially for the model leveraging long-term temporal information.

In Sec. 2, we introduce the FFT related applications in computer vision and deep learning. In Sec. 3, we introduce our technical approach in detail. In Sec. 4, we demonstrate and verify the effectiveness and satisfactory performance of the proposed framework under challenging out-of-domain scenarios.

2. Related Works

Frequency Domain Analysis As a cornerstone in signal processing, FFT pivotal across a broad spectrum of computer vision applications, including image analysis and feature extraction [24, 25]. In addition, frequency domain analysis plays a vital role in adversarial attacks and data augmentation. The AdvDrop attack introduces a novel approach by generating adversarial examples through the elimination of image details in the frequency domain, proving challenging for existing defensive frameworks to mitigate [26]. Amplitude-phase recombination proposes to augment the original input sample by recombining the phase spectrum of the sample and the amplitude spectrum of the distracting image so as to force the Convolutional Neural Networks (CNN) to pay more attention to the structured information of the phase component and remain robust to perturbation in amplitude, preventing the CNN from local optimum [27]. Inspired by the frequency adversarial attack and amplitude-phase recombination augmentation, we propose a novel augmentation module and adapt it to multi-camera 3D object detection frameworks to further enhance the model robustness.

3. Approach

In this section, we will present our solution in detail with the following aspects covered. Section 3.1 will elaborate on the design of the model structure, pretraining, and scaling. Section 3.2 will introduce the augmentation protocol and chain employed in our study.

3.1. Model Structure

Giving a multi-camera temporal sequence images $X_t = \{I_1, I_2, I_3, \cdots, I_N\}_t$, N is camera number and t is timestamp, the proposed **TSMA-BEV** framework aims to enhance the classification and bounding box regression robustness when facing out-of-domain scenarios. Inspired by this benchmark [28], temporal fusion is crucial to the robustness of camera-only 3D detection algorithms. Thus, we adopt both low-resolution long-term and high-resolution

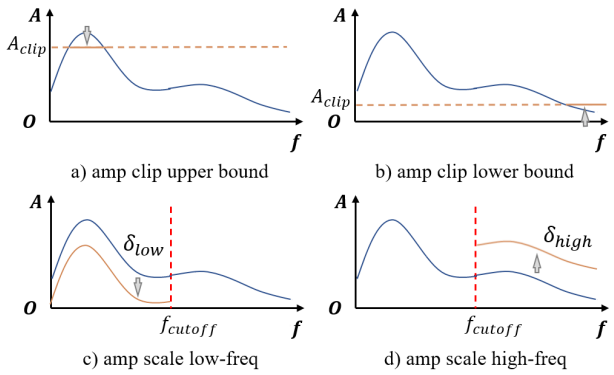

Figure 2. Frequency domain augmentation operations. The orange line is the transformed frequency distribution.

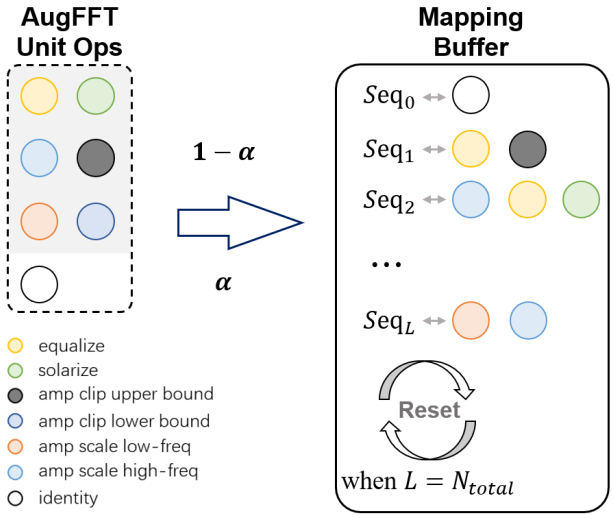

Figure 3. Overall pipeline of temporal sequence mix augmentation (TSMA). The mixing ratio α represents the proportion of sequences without AugFFT. α is the mix ratio.

short-term stereo depth potential from SOLOFusion [21] to make full use of temporal information.

Following the common pretrain protocols in recent BEV-based works, we leverage VoVNet-99 as an image encoder. We use publicly available pretrained weights* on DD3D datasets. Additionally, with the same detection range, a larger BEV grid size can represent more detailed scene information, which is friendly to accurate multi-scale object detection. Commonly used BEV feature size is $B_t \in \mathbb{R}^{B\times C\times H\times W}$ in which H and W are both 128, in this work we adopt $B_t \in \mathbb{R}^{B\times C\times 2H\times 2W}$, t is the timestamp.

3.2. Augmentation Module

Effective data augmentation is pivotal for enhancing the generalizability of camera-only 3D object detection models. Various strategies have been employed to boost model robustness during training, including the innovative use of adversarial losses [29]. While these methods tend to increase training latency and GPU memory demands. Consequently, there is a growing need for efficient augmentation techniques that can integrate into existing training workflows, and achieve a better balance between model performance and resource consumption.

3.2.1 Augmentation Protocol

In this study, we deliberately omit any image corruption operations included in the simulation methods that are leveraged to construct the evaluation dataset provided by the competition. In particular, we remove all unit operations from the original 18 corruption types in [28], only remain `equalize` and `solarize`. Besides, to avoid any potential overlap with the evaluation set, we do not use any image noising or image blurring operations.

*https://github.com/exiawsh/storage/releases/download/v1.0/dd3d_det_final.pth

In addition to these two operations, we propose a group of image augmentation operations in the frequency domain, as illustrated in Fig. 2. The operations a) and b) conduct amplitude clip based on randomly sampled scale σ. For upper bound clip, $A_{clip} = \sigma_{up} A_{max}, \sigma_{up} \in [0.7, 1]$. While as for lower bound clip, $A_{clip} = 1\times 10^{-4}(\sigma_{low}+1)A_{max}, \sigma_{low} \in [-0.3, 0.3]$.

The operations c) and d) conduct specific frequency band amplitude scaling based on random sampled cut-off frequency $f_{cutoff} \in [0.2 f_{bd}, 0.6 f_{bd}]$ and scale factor δ, in which f_{bd} is the maximize bandwidth. For low-frequency amplitude scaling, $A(f) = \delta_{low} A_{ori}(f), f \in (0, f_{cutoff}]$, in which $\delta_{low} \in [0.1, 0.5]$. While for high frequency amplitude scaling, $A(f) = \delta_{high} A_{ori}(f), f \in (f_{cutoff}, +\infty]$, in which $\delta_{high} \in [1.6, 2.0]$.

Finally, as mentioned in Fig. 3, the augmented operations we used consist of the following 6 operations `equalize`, `solarize`, `amp clip upper bound`, `amp clip lower bound`, `amp scale low-freq`, `amp scale high-freq`. The randomly augmented input images are shown in Fig. 4.

3.2.2 Augmentation Chain

NuScenes training data consists of 700 temporal sequences, which are usually split into $N_{total} = 1400$ mini-sequences [21]. To better facilitate the temporal information and avoid adjacent frame augmentation inconsistency, we propose to apply the augmentation protocol to each mini-sequence and update the mapping buffer after all mini-sequences are trained. The buffer reset will introduce more variety and avoid over-fitting on simple augmentation pat-

Table 1. Comparison of top-perform solutions on **Phase2** evaluation data. Our solution is marked with **red**. The baseline model is marked with green. The **best** scores of each corruption type are highlighted in **bold**.

Team Name	NDS↑	◦ brightness	◦ low-light	◦ fog	◦ frost	◦ snow	◦ contrast	◦ defocus blur	◦ glass blur	◦ motion blur	◦ zoom blur	◦ elastic transform	◦ quantization	◦ gaussian noise	◦ impulse noise	◦ shot noise	◦ ISO noise	◦ pixelate	◦ JPEG compression
DeepVision	**52.1**	39.5	**65.3**	36.7	**49.8**	**62.3**	**51.8**	53.5	**49.1**	41.8	**37.1**	**45.3**	**67.5**	**71.2**	**52.9**	**59.6**	**56.7**	**56.9**	40.8
Ponyville	50.2	**43.1**	62.7	**37.5**	46.6	60.9	49.2	**58.4**	46.5	**44.7**	18.8	44.8	66.7	70.6	42.4	56.0	50.8	56.8	**47.5**
CyberBEV	49.0	42.1	61.4	37.0	46.3	60.5	47.0	57.1	44.9	43.5	17.1	43.8	65.9	69.1	40.9	56.1	49.4	55.3	45.1
Safedrive-promax	48.1	39.2	59.8	37.4	39.8	62.5	47.6	59.0	43.9	41.4	14.3	45.4	63.3	68.5	42.3	53.3	49.5	56.2	42.2
drivingClass	47.8	39.1	60.0	28.1	48.7	56.3	44.1	52.5	46.9	38.9	34.9	44.3	62.9	63.8	50.9	51.6	52.0	56.3	33.4
BUPTMM	43.5	37.7	55.1	30.8	41.1	51.2	45.2	45.1	41.1	38.6	29.6	40.5	56.1	58.7	41.0	48.5	42.4	46.8	34.0
BEVFormer	22.8	28.5	34.9	21.4	10.5	28.0	15.7	28.7	21.1	19.2	6.4	35.0	26.9	20.6	12.3	24.9	25.4	30.3	21.3

Figure 4. Visualization of random generated augmented samples based on TSMA-BEV.

terns. For each mini-sequence, the augmentation chain consists of a maximum of 3 randomly sampled operations. Different from AugMix [30], we superimpose the randomly sampled operations on the input multi-camera images. Furthermore, as illustrated in Fig. 3, we only apply the above-mentioned augmentation protocol to a certain proportion $(1-\alpha)$ of the sequences and only apply common data augmentation (such as flip, rotate, resize, and crop, etc.) to the other parts. As discussed in Section. 4.4, the mixed sequence augmentation protocol can avoid over-fitting on either original data or augmented data.

4. Experiments

4.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [23] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [9] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [31–33], RoboBEV [14, 34, 35], and Robo3D [36, 37]. 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.

4.2. Implementation Details

The **TSMA-BEV** framework is implemented using the PyTorch framework [38] and is based on the MMDetection3D codebase [39]. All experiments are conducted on 8 NVIDIA A100 (80G) GPUs. The basic learning rate is 2e-4 when training with 8 GPUs and the total batch size is 64, actual learning rate is adapted to the actual batch size since we use auto-scale learning policy, that is, $lr_{act} = (lr_{basic}/64) \times (N_{GPU} \times B_{mini})$. The optimizer we use is AdamW [40]. As for our training pipeline, we leverage a 3 stage training strategy:

- **Stage 1**: We first train the model for 2 epochs with only short-term fusion in full precision and small batch size to make the model converge faster and ensure training stability;
- **Stage 2:** Then we further train the model for 4 epochs without long-term history fusion in FP16 precision with larger batch size;
- **Stage 3:** Finally, we train the model for extra 14 epochs with both short-term and long-term history fusion in FP16 precision. The proposed **TSMA** strategy is also applied.

Note that we leverage grad clip and loss scaling with default scale factor 128 when using FP16 precision training.

Table 2. Ablation results of the proposed AugFFT and SeqMixAug on **Phase2** evaluation data. 2× means using large BEV grid size as mentioned in Section. 3.1.

Method	Resolution	NDS↑	mAP↑	mATE↓	mASE ↓	mAOE↓	mAVE↓	mAAE↓
Base(2×)	1408 × 512	0.5027	0.4546	**0.5412**	0.4054	0.4808	**0.4667**	0.3524
Base(2×)+AugFFT	1408 × 512	0.5121	**0.4736**	0.5444	0.3905	0.4804	0.4865	0.3454
Base(2×)+AugFFT+SeqMixAug	1408 × 512	**0.5211**	0.4715	0.5493	**0.3840**	**0.4066**	0.4700	**0.3370**

Table 3. Ablation studies of mixing ratio α on **Phase1** evaluation data of the competition. Models are trained under an image resolution of 320 × 800.

Mixing Ratio α	0.5	0.2	0
NDS↑	0.5676	**0.5785**	0.5611

4.3. Comparative Study

As shown in Table. 1, we compared our solution with other approaches on the Phase2 evaluation data of this competition. Our solution outperforms the other methods by a large margin across many corruption types.

4.4. Ablation Study

As shown in Tab. 3, we also conduct ablation studies on the mix ratio α and find the optimal choice $\alpha = 0.8$, which is used in our best model. To further validate the effectiveness of the proposed data augmentation module, we assess the impact of the simple AugFFT module, the combination of AugFFT and SeqMixAug module, as shown in Tab. 2. The proposed data augmentation and sequence consistent augmentation strategy play an important role in improving the model performance under out-of-domain corruptions. Due to the long-term history fusion and sequence consistency augmentation design, our model performs better in mean Average Orientation Error (AOE), mean Average Velocity Error (AVE), and mean average attribute error (AAE).

5. Conclusion

This study introduces **TSMA-BEV**, a novel framework for 3D object detection tailored to the RoboDrive competition setting that challenges algorithms with out-of-domain scenarios such as severe weather and sensor failures. The proposed approach utilizes the nuScenes training dataset, enhances robustness through two novel augmentation strategies: AugFFT and SeqMixAug. AugFFT applies stochastic frequency cut-offs and amplitude scaling in the frequency domain to produce varied augmented images. While SeqMixAug extends AugFFT to temporal sequences, ensuring consistent augmentation settings across frames to facilitate the ability of the model to utilize temporal information. Experimental results confirm the efficacy of **TSMA-BEV** in handling out-of-domain conditions, marking a promising advancement in the robustness of 3D object detection algorithms.

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] 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.

[3] 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.

[4] 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.

[5] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3621–3631, 2023.

[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] 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.

[8] 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.

[9] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.

[10] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020.

[11] 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.

[12] 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.

[13] Fangzhou Hong, Lingdong Kong, Hui Zhou, Xinge Zhu, Hongsheng Li, and Ziwei Liu. Unified 3d and 4d panoptic segmentation via dynamic shifting networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3480–3495, 2024.

[14] 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.

[15] Shuo Wang, Xinhai Zhao, Hai-Ming Xu, Zehui Chen, Dameng Yu, Jiahao Chang, Zhen Yang, and Feng Zhao. Towards domain generalization for multi-view 3d object detection in bird-eye-view. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13333–13342, 2023.

[16] Hao Lu, Yunpeng Zhang, Qing Lian, Dalong Du, and Yingcong Chen. Towards generalizable multi-camera 3d object detection via perspective debiasing. *arXiv preprint arXiv:2310.11346*, 2023.

[17] Hao Lu, Jiaqi Tang, Xinli Xu, Xu Cao, Yunpeng Zhang, Guoqing Wang, Dalong Du, Hao Chen, and Yingcong Chen. Scaling multi-camera 3d object detection through weak-to-strong eliciting. *arXiv preprint arXiv:2404.06700*, 2024.

[18] Zhenyu Li, Zehui Chen, Ang Li, Liangji Fang, Qinhong Jiang, Xianming Liu, and Junjun Jiang. Unsupervised domain adaptation for monocular 3d object detection via self-training. In *European conference on computer vision*, pages 245–262. Springer, 2022.

[19] 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.

[20] Runsheng Xu, Xin Xia, Jinlong Li, Hanzhao Li, Shuo Zhang, Zhengzhong Tu, Zonglin Meng, Hao Xiang, Xiaoyu Dong, Rui Song, et al. V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13712–13722, 2023.

[21] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *The Eleventh International Conference on Learning Representations*, 2022.

[22] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021.

[23] 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.

[24] Libao Zhang and Kaina Yang. Region-of-interest extraction based on frequency domain analysis and salient region detection for remote sensing image. *IEEE Geoscience and Remote Sensing Letters*, 11(5):916–920, 2013.

[25] Chris Kreucher and Sridhar Lakshmanan. Lana: a lane extraction algorithm that uses frequency domain features. *IEEE Transactions on Robotics and automation*, 15(2):343–350, 1999.

[26] Ranjie Duan, Yuefeng Chen, Dantong Niu, Yun Yang, A Kai Qin, and Yuan He. Advdrop: Adversarial attack to dnns by dropping information. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7506–7515, 2021.

[27] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 458–467, 2021.

[28] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and analyzing bird's eye view perception robustness to corruptions.

[29] 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.

[30] Dan Hendrycks, Norman Mu, Ekin D Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. Augmix: A simple data processing method to improve robustness and uncertainty. *arXiv preprint arXiv:1912.02781*, 2019.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

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

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