MVE: Multi-View Enhancer for Robust Bird's Eye View Object 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 the 2024 RoboDrive Challenge, specifically Track 1: Robust BEV Detection, the "Multi-View Enhancer (MVE)" method leverages innovative approaches to significantly improve the robustness of 3D object detection from multiple camera perspectives. Building on the foundation of the RayDN architecture, MVE integrates a modified backbone using EVA ViT-Large, pre-trained on ImageNet to ensure deep and robust feature extraction. This method is further enhanced by a strategic combination of Augmix and DeepAug data augmentation techniques, meticulously tailored to avoid overlapping corruptions with those encountered in the challenge test sets. By adopting depth-aware hard negative sampling, MVE not only refines the detection capabilities but also ensures the model's adaptability to varied and unforeseen environmental conditions. The training process is systematically structured to evolve from clean, unaltered datasets to increasingly complex scenarios, ensuring that each step contributes to building a more resilient detection system. This method has shown promising results in preliminary tests, highlighting its potential as a robust solution for BEV detection challenges in autonomous driving applications.

1. Introduction

The advent of autonomous driving technologies has catalyzed an unprecedented focus on the development of robust and reliable detection systems capable of accurately interpreting and navigating complex environments [1, 2]. Among the various challenges, Bird's Eye View (BEV) detection remains pivotal, offering a comprehensive perspective that is critical for the safe operation of autonomous vehicles. The 2024 RoboDrive Challenge [3], particularly Track 1: Robust BEV Detection, presents an opportunity to address this challenge by leveraging advanced computer vision techniques. Our solution, titled "Multi-View Enhancer (MVE)", aims to significantly enhance the accuracy and robustness of BEV detection across multiple camera perspectives.

BEV detection systems are essential for understanding the vehicle's surrounding environment from a top-down view, integrating data from multiple sensors to create a consolidated and actionable understanding of road conditions, obstacles, and navigational cues [4]. However, the dynamic nature of driving environments, coupled with the inherent limitations of current detection technologies, poses significant challenges. These include the variability of environmental conditions, the presence of occlusions, and the need for high precision in object detection and depth estimation from 2D images.

Our approach is designed to overcome these challenges by integrating a novel pipeline for camera-only 3D object detection, a sophisticated feature extraction backbone, and innovative data augmentation techniques. The Multi-View Enhancer (MVE) employs a combination of state-of-the-art technologies and methodologies to ensure high performance and adaptability in real-world driving scenarios, setting a new standard for BEV detection systems in autonomous vehicles.

2. Related work

2.1. 3D Object Detection

The field of 3D object detection has evolved significantly with advancements in deep learning and computer vision. Early efforts predominantly utilized geometric properties and stereoscopic vision to estimate depth and object positioning [5]. With the advent of deep convolutional neural net-

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

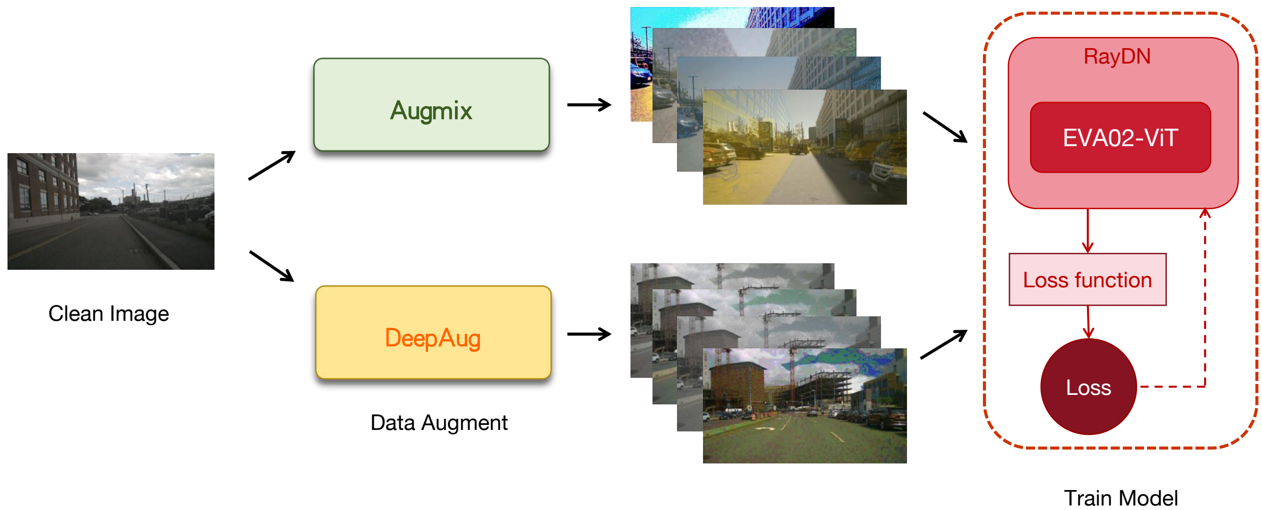

Figure 1. Pipeline of MVE.

works, researchers have shifted towards more sophisticated methods that leverage large volumes of data for training more accurate models. Notable developments include the use of point clouds generated by LiDAR sensors [6–8], as seen in models like PointNet and PointRCNN, which have set benchmarks for accuracy in 3D object detection [9, 10]. More recently, methods that infer 3D information from 2D images have gained prominence due to their cost-effectiveness and ease of integration with existing camera-based systems, such as MonoDIS and Pseudo-LiDAR approaches [11, 12]. More advanced methods have been developed for multi-view camera 3D object detection in BEV space, such as BEVFormer [13], BEVDepth [14], and Sparse4D [15]. These methods have even achieved performance comparable to LiDAR-based detection [16], marking significant progress [17].

2.2. Robustness of Visual Systems

Recent research has extensively explored adversarial robustness, focusing on how models can withstand malicious inputs designed to induce errors. Studies by Madry et al. and Goodfellow et al. have laid foundational work in understanding and defending against adversarial examples, highlighting techniques like adversarial training as effective countermeasures and son on [18–20]. On the other hand, natural robustness pertains to a model's ability to perform reliably across a range of environmental conditions and sensor noises, a vital attribute for systems deployed in variable real-world settings. Efforts to enhance natural robustness often involve data augmentation techniques and robust training frameworks that mimic real-world disturbances. Research in this area has been propelled by benchmarks like ImageNet-C, which tests models against common visual corruptions and has spurred the development of more resilient architectures [21–23].

3. Approach

Our approach for the RoboDrive Challenge, titled "Multi-View Enhancer (MVE)", harnesses advanced computational techniques and innovative methodologies to significantly bolster the robustness and accuracy of Bird's Eye View (BEV) detection across multiple camera perspectives. By integrating sophisticated systems for data processing, augmentation, and adaptive training, MVE is designed to address the multifaceted challenges associated with 3D object detection in dynamic driving environments. The pipeline of MVE is illustrated in Fig. 1.

3.1. Pipeline

MVE approach follows a novel pipeline RayDN [24] for camera-only 3D object detection, the enhancement specifically developed for multi-view 3D object detection. This method strategically mitigates the common issues of redundant and incorrect detections, which are prevalent due to the inherent difficulties in depth estimation from 2D images. By implementing depth-aware hard negative sampling directly along camera rays, Ray Denoising creates hard negative examples that are visually indistinguishable from true positives. These challenging examples force the model to refine its ability to discern depth-related features, significantly improving its capability to distinguish between true and false positives. Ray Denoising functions as a plug-and-play module, easily integrating with any DETR-style multi-view 3D detector. It offers a substantial boost in detection accuracy, demonstrating an improvement in mean Average Precision (mAP) over existing state-of-the-art methods like StreamPETR on the nuScenes dataset [1], without increasing training computational overhead or affecting inference speeds

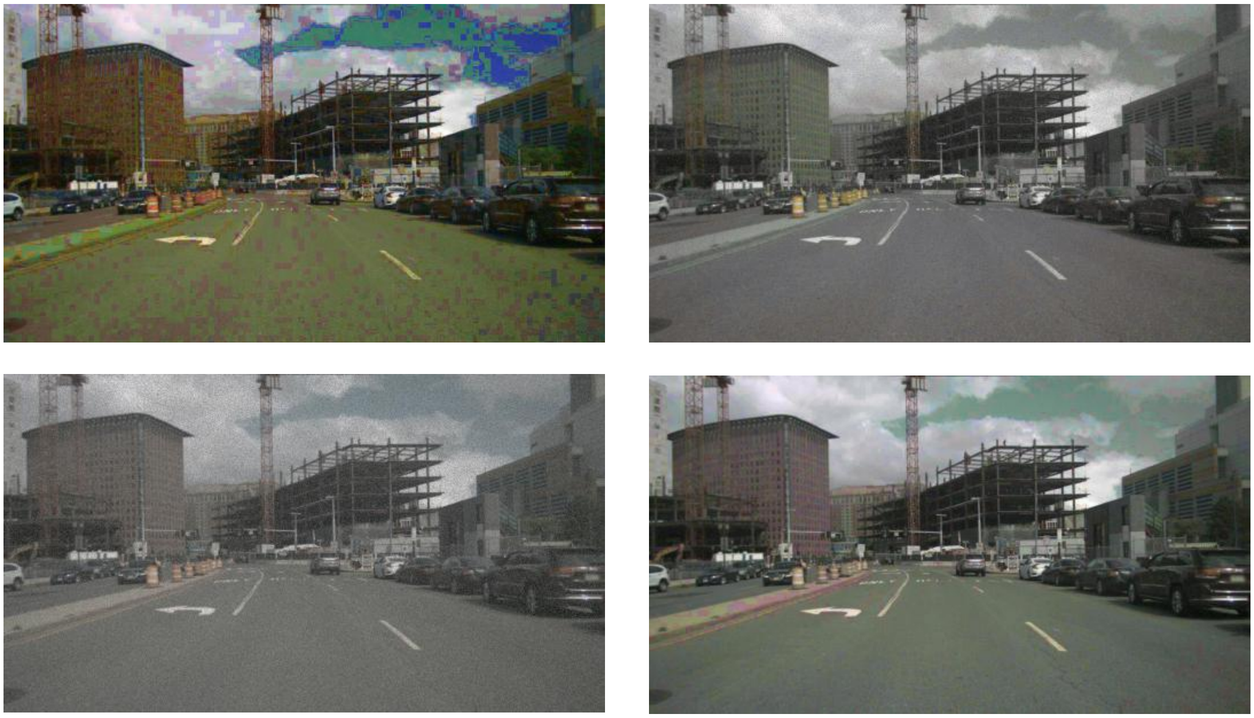

Figure 2. Visualization of Augmix-Enhanced Data.

Figure 3. Visualization of DeepAug-Enhanced Data.

3.2. Backbone

For superior feature extraction, MVE employs the EVA ViT-Large, a next-generation Transformer-based model that has been pre-trained on the extensive ImageNet dataset. The EVA-02 variant of this backbone utilizes an updated plain Transformer architecture and has been extensively trained to reconstruct robust, language-aligned vision features via masked image modeling. This allows the EVA ViT-Large to excel in extracting high-quality features that are crucial for precise object detection, even under variable environmental conditions. With its exceptional capability to maintain high performance using significantly fewer parameters, the EVA-02 backbone ensures that our model is not only effective but also efficient, making it ideal for real-time applications in autonomous driving.

3.3. Data Augmentation

To ensure that MVE performs reliably across varied and unforeseen operational conditions, our approach incorporates two advanced data augmentation strategies: Augmix and DeepAug. Visualization of enhanced data can be seen in Fig. 2 and Fig. 3.

Augmix is designed to enhance model robustness by applying a combination of simple image processing techniques such as pixel shuffle, random hue, and random saturation in a manner that preserves the semantic content of the images while introducing realistic, unseen variations. This method significantly improves the model's uncertainty estimates and resilience against data corruptions not present during training, effectively bridging the gap between clean data and real-world performance.

The Augmix method uses a combination of image processing operations and mixes the resulting images using a convex combination, maintaining the semantic integrity of the images while introducing diverse variations. Each augmentation chain consists of a sequence of operations applied to the image. Let x be the original image, and $\mathcal{O}_1, \mathcal{O}_2, \ldots, \mathcal{O}_n$ be the image processing operations (like pixel shuffle, random hue, and random saturation). An augmentation chain for a single image can be expressed as:

$$x' = \mathcal{O}_n(\ldots(\mathcal{O}_2(\mathcal{O}_1(x)))\ldots)$$

The outputs of different augmentation chains are mixed together using element-wise convex combinations. If $x'_1, x'_2, \ldots, x'_k$ are the outputs from k different augmentation chains, and $\mathbf{w}$ is the vector of mixing weights sampled from a Dirichlet distribution $\mathrm{Dir}(\alpha, \ldots, \alpha)$, then the mixed image $\tilde{x}$ can be represented as:

$$\tilde{x} = w_1 \cdot x'_1 + w_2 \cdot x'_2 + \cdots + w_k \cdot x'_k$$

where $w_1, w_2, \ldots, w_k$ are the components of $\mathbf{w}$.

Finally, the mixed image $\tilde{x}$ is combined with the original image x using a second random convex combination sampled from a Beta distribution $\mathrm{Beta}(\alpha, \alpha)$. Let β be the mixing coefficient from the Beta distribution, and the final image y is given by:

$$y = \beta \cdot x + (1 - \beta) \cdot \tilde{x}$$

The full Augmix process, combining several sources of randomness—choice of operations, severity, lengths of augmentation chains, and mixing weights—helps ensure robustness and generalization, preparing the model to handle unseen variations and corruptions effectively.

DeepAug, on the other hand, represents a more radical departure from traditional data augmentation techniques. Instead of applying transformations directly to the raw images, DeepAug manipulates the internal representations within deep neural networks. By passing clean images through image-to-image networks like CAE and EDSR and introducing random perturbations at various layers, DeepAug generates images that maintain semantic integrity but differ significantly in appearance from their original versions. These perturbations include operations such as zeroing, negating, and convolving, which introduce a rich tapestry of visual variations, thereby training the model to recognize and adapt to a broader range of visual data.

This dual approach of Augmix and DeepAug not only prepares the model to handle diverse environmental changes but also ensures that it can adapt to potential shifts in input data distributions encountered in real deployment scenarios.

3.4. Training Strategy

The training regimen of MVE is meticulously planned to maximize the model's exposure to a wide range of scenarios, starting with training on clean, unaltered data from the nuScenes dataset. This foundational phase establishes baseline accuracy and robustness. Subsequent phases introduce complexity incrementally, first integrating data enhanced with Augmix, followed by data simultaneously enhanced by both Augmix and DeepAug. This staged training strategy not only helps in layering the robustness attributes of the model but also ensures that the system develops the ability to generalize well across different types of environmental and operational conditions, ultimately leading to a more resilient and dependable detection system.

By deploying these strategic implementations, MVE sets a new benchmark for robustness and accuracy in multi-view BEV detection, providing a comprehensive, adaptable solution for the evolving demands of autonomous driving technology.

4. Experiments

4.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [3] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [1] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [25–27], RoboBEV [4, 28, 29], and Robo3D [22, 30]. The corruption types are mainly from three sources, namely camera corruptions, camera failures, and LiDAR failures. For more details, please refer to the corresponding

Table 1. NDS of corruption categories on Robodrive Track1 Phase2 test dataset. (Part 1)

Corruptions	Bright	Dark	Fog	Frost	Snow	Contrast	Defocus Blur	Glass Blur	Motion Blur	Zoom Blur
RayDN	0.354	0.528	0.334	0.256	0.616	0.336	0.493	0.451	0.380	0.119
MVE (Augmix)	0.421	0.627	0.336	0.439	0.648	0.480	0.587	0.434	0.413	0.156
MVE (Augmix+DeepAug)	0.431	0.627	0.375	0.466	0.609	0.492	0.584	0.465	0.447	0.188

Table 2. NDS of corruption categories on Robodrive Track1 Phase2 test dataset. (Part 2)

Corruptions	Elastic Transform	Color Quant	Gaussian Noise	Impluse Noise	Shot Noise	ISO Noise	Pixelate	JPEG	**Average**
RayDN	0.470	0.487	0.588	0.363	0.483	0.482	0.559	0.429	0.429
MVE (Augmix)	0.434	0.661	0.691	0.468	0.532	0.511	0.566	0.382	0.488
MVE (Augmix+DeepAug)	0.448	0.667	0.706	0.424	0.560	0.508	0.568	0.475	0.502

Table 3. Clean performance on nuScenes dataset validation split.

Models	NDS	mAP	mATE	mASE	mAOE	mAVE	mAAE
BEVFormer (Baseline)	0.517	0.415	0.672	0.274	0.369	0.397	0.198
RayDN	0.624	0.541	0.518	0.252	0.274	0.230	0.195
MVE (Augmix)	0.623	0.541	0.509	0.253	0.268	0.248	0.194
MVE (Augmix+DeepAug)	0.619	0.536	0.506	0.256	0.294	0.248	0.187

GitHub repositories.

4.2. Experimental Setups

Our proposed approach was implemented using the PyTorch framework [31] and was based on the MMDetection3D codebase [32]. The Multi-View Enhancer (MVE) model was trained using eight NVIDIA GeForce RTX 4090 GPUs. During the training process, only images from the training split of the nuScenes dataset were utilized. The training regimen was structured, to begin with 24 epochs on the clean nuScenes train split, followed by 16 epochs on Augmix-enhanced data, and concluded with 16 epochs using a combination of both Augmix and DeepAug enhanced data.

4.3. Implementation Details

Augmix. The initial implementation of Augmix overlapped significantly with the corruptions used in the 2024 RoboDrive competition. Due to competition rules that prohibit the use of identical corruptions during training, we selected pixel shuffle, random hue, and random saturation as augmentation methods that differ from the competition's corruptions to simulate data degradation and enhance the generalization capabilities of the detection model.

DeepAug. The DeepAug enhancement includes augmented data processed by CAE and EDSR models. This approach is implemented during the image loading phase, either on the fly or through pre-generated augmented data to optimize computational efficiency. In practice, augmented data is pre-generated, and during model training, images are loaded based on a random probability rd (threshold t experimentally set to 0.6). If rd exceeds t, data processed by EDSR is used; otherwise, CAE-processed data is employed. Additionally, to maintain consistency in detection outcomes, operations such as horizontal and vertical flips, which are typically part of DeepAug, were excluded.

4.4. Comparative Study

RoboDrive Track 1: Robust BEV Detection competition involves 18 types of corruptions designed to evaluate the algorithm's recovery capabilities against various environmental and sensor-based damages. Tab. 1 and Tab. 2 display the results of our baseline RayDN and the MVE method on the 18 corruption types of the NDS metrics. It was observed that post-Augmix processing, there was a performance improvement on most corruption types, with gains of 0.182 and 0.173 NDS on Frost and Color Quant respectively, achieving an average NDS of 0.488. However, slight decreases were noted on corruptions such as Glass Blur, Elastic Transform, and JPEG. Following the addition of DeepAug enhancements, overall robustness further improved, with the average NDS reaching 0.502. This indicates that both Augmix and DeepAug enhancements contribute to improved NDS across the dataset.

4.5. Results on the nuScenes Dataset

Furthermore, the performance of our methods on the nuScenes validation split clean data is evaluated, as shown in Table Tab. 3. Compared to BEVFormer, RayDN showed an NDS improvement of 0.1068. Selecting RayDN as the pipeline laid a solid foundation for our approach. The MVE method, after augmentation with Augmix data, retained almost complete performance on clean data. After an additional 16 epochs of training on DeepAug data, the NDS on clean data slightly decreased to 0.619. However, at this point, the MVE method achieved the best robustness NDS values, illustrating a trade-off between clean data NDS performance and robust data NDS performance. This also highlights that MVE's data augmentation techniques do not overly impact performance on clean data, preserving the method's detection capabilities on uncorrupted datasets.

5. Conclusion

In this study, we introduced the Multi-View Enhancer (MVE), an advanced approach designed to improve the robustness and accuracy of BEV detection in autonomous vehicles. By integrating the Ray Denoising technique with the EVA ViT-Large backbone and innovative data augmentation methods like Augmix and DeepAug, MVE significantly enhances the detection capabilities under various environmental conditions. Our results demonstrated marked improvements in handling diverse types of data corruptions in the RoboDrive Challenge, maintaining high performance on clean data from the nuScenes dataset. This work lays a solid foundation for further research into reliable and efficient BEV detection systems for autonomous driving, aiming to balance high performance with robustness in real-world scenarios.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018.

[6] 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.

[7] 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.

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

[9] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[10] Shaoshuai Shi, Xiaogang Wang, and Hongsheng Li. Pointrcnn: 3d object proposal generation and detection from point cloud. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 770–779, 2019.

[11] Andrea Simonelli, Samuel Rota Bulo, Lorenzo Porzi, Manuel López-Antequera, and Peter Kontschieder. Disentangling monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1991–1999, 2019.

[12] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019.

[13] 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.

[14] 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 *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 1477–1485, 2023.

[15] 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.

[16] 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.

[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] 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.

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

[20] Caixin Kang, Yinpeng Dong, Zhengyi Wang, Shouwei Ruan, Hang Su, and Xingxing Wei. Diffender: Diffusion-based adversarial defense against patch attacks in the physical world. *arXiv preprint arXiv:2306.09124*, 2023.

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

[22] 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.

[23] Yinpeng Dong, Caixin Kang, Jinlai Zhang, Zijian Zhu, Yikai Wang, Xiao Yang, Hang Su, Xingxing Wei, and Jun Zhu. Benchmarking robustness of 3d object detection to common corruptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1022–1032, 2023.

[24] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection, 2024.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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