ViewFormer: Spatiotemporal Modeling for Robust Occupancy Prediction

Jinke Li Xiao He Xiaoqiang Cheng

Uisee Foundation Research & Development
{jinke.li, xiao.he, xiaoqiang.cheng}@uisee.com

Abstract

This technical report outlines the solution that achieved top-ranking performance in the Robust Occupancy Prediction of the 2024 RoboDrive Challenge. Our approach is based on ViewFormer, a robust vision-centric spatiotemporal modeling method employing view-guided transformers. Expanding upon ViewFormer, we further investigate offline extensions that promote video temporal methods with future data, model scaling up, and effective post-processing strategies. These improvements and optimizations rank 1st place in the challenge with remarkable mIoU scores of 24.07% and 22.31% on the phase-1 and phase-2 corruption test sets, respectively.

1. Introduction

In recent years, the field of vision-centric autonomous driving has emerged as a research focus, driven by its remarkable cost-effectiveness and the potential to revolutionize transportation systems worldwide [1–3]. Central to this endeavor is accurately interpreting the real 3D world from 2D images captured by onboard cameras, which has stimulated the exploration of advanced techniques to extract informative spatial features from visual data [4–7].

One pivotal advancement in this area is the proposal of 3D occupancy representation, which unifies the notion of foreground and background and discretizes the entire 3D space into voxel-wise cells, with each cell annotated with a semantic label. This approach offers a comprehensive and structured framework for interpreting complex scenes.

In the context of the challenge, our solution is built on top of our proposed ViewFormer [8], a transformer-based framework designed to predict 3D occupancy with multi-camera images as input, featuring view attention for spatial interaction and multi-frame streaming temporal attention for temporal interaction. Regarding spatial interaction, our view attention facilitates multi-view feature aggregation distinguished from the projection-first deformable attention used by BEVFormer [9] as analyzed in [8], allowing for constructing more semantically informative and robust features. For temporal interaction, the streaming temporal attention utilizes online video data through a memory mechanism [10, 11] in both training and inference to reduce training time and maintain consistency. Moreover, as an offline extension, we introduce a reverse video playback mechanism that allows our online video temporal interaction method to benefit from future data frames.

2. Our Solution

Our solution is based on ViewFormer [8], featuring view attention and multi-frame streaming temporal attention. We give a short introduction to the method in Fig. 1, please refer to the official ViewFormer [8] paper for detailed methodology as well as experimental analyses.

2.1. Spatial Modeling of ViewFormer

Built upon existing bird's-eye-view (BEV) perception approach [9, 12, 13], our ViewFormer extracts multi-view image features F_t via an image backbone and subsequently refines queries through spatial interaction. Instead of adopting BEV queries, we directly utilize voxel queries V_t' to capture finer-grained 3D occupancy features. As the core of the spatial interaction, our view attention is introduced to adequately transform multi-view image features into 3D space, which, in fact, presents significant difference from the project-first deformable attention commonly employed in prior works [9, 14–16]. Limited by sensor distribution, the project-first method can only extract features from a single image for most queries. In contrast, our learning-first view attention aggregates multi-view features in a more reasonable data-driven manner. This module provides a solid foundation for us to construct robust 3D features.

Technical Report of the 2024 RoboDrive Challenge.
Track 3: Robust Occupancy Prediction.

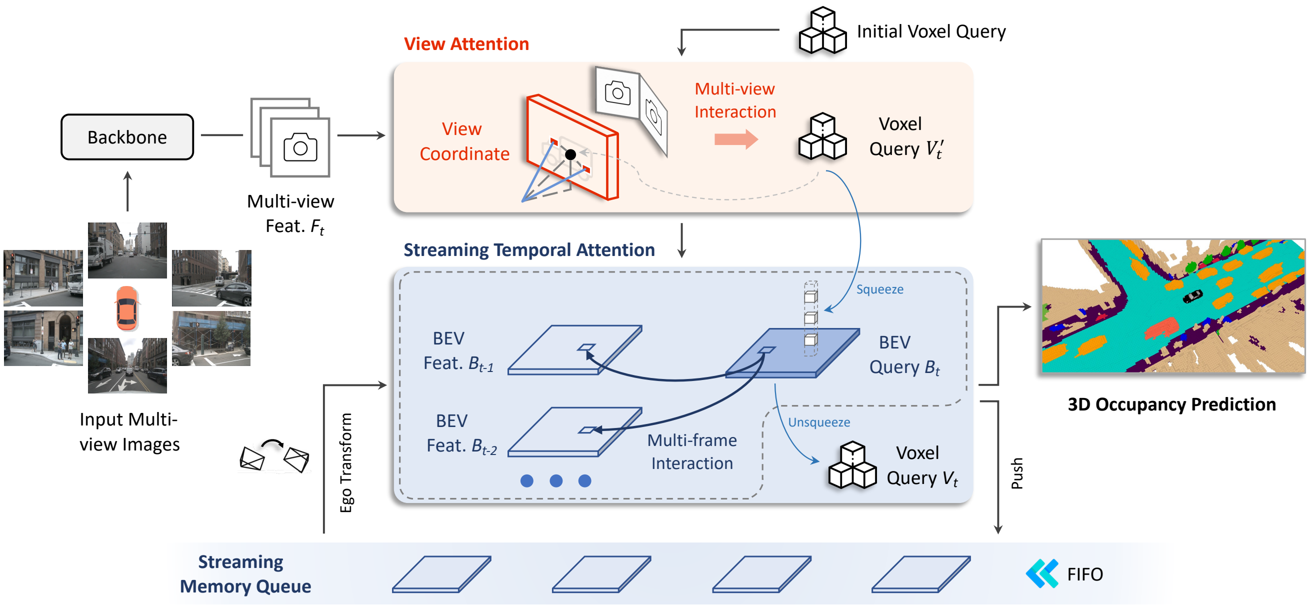

Figure 1. In our **ViewFormer** framework, the multi-view features F_t are first extracted from multiple images via a backbone. Then we introduce view attention, allowing us to aggregate multi-view features for voxels V_t' more adequately. In our streaming temporal attention, we squeeze voxel queries V_t' into BEV queries B_t. Each BEV cell of B_t interacts with historical multi-frame BEV features stored in the streaming memory queue. The voxels V_t obtained from unsqueezing the updated BEV features are subsequently fed into 3D occupancy prediction. We push the updated BEV queries into the memory queue for subsequent temporal interaction in the video stream pipeline.

2.2. Temporal Modeling of ViewFormer

Drawing inspiration from streaming video methods [10, 11], we establish a streaming memory queue to dynamically capture historical features spanning N frames in both the training and inference phases, following the first-in, first-out (FIFO) principle [10] for data entry and exit. Considering the increased storage and computational overhead, we process temporal modeling at the 2D BEV level. Specifically, the voxel queries V_t' are squeezed into BEV-level queries B_t along the z-axis, with each BEV cell of B_t interacting with the historical multi-frame BEV features stored in the memory queue. Here, we employ ego transformation to handle ego-motion. Subsequently, the voxels V_t, obtained by unsqueezing the updated BEV queries, are used for 3D occupancy prediction. Meanwhile, we push the updated BEV queries into the memory queue for subsequent temporal interaction in the video stream pipeline.

2.3. Reverse Video Playback

Since the challenge permits leveraging future frame data, we also propose a novel reverse video playback mechanism (RVP) to further enhance offline accuracy. Unlike BEVFormerv2, which repeats inference for each frame within a fixed sliding window, our method leverages future frames in a more lightweight way without modifying our online framework. By analyzing the discrepancy of perception accuracy between the front and rear area of the vehicle in the online video temporal interaction, we implement the reverse video playback mechanism by simply playing the video backward.

As illustrated in Fig. 2, let's take a stationary tree as an example. Suppose the car moves forward and the tree is in the rear area of the car, the corresponding voxel query can consistently retrieves historical features of the tree from the memory queue as in Fig. 2(a). On the contrary, as in Fig. 2(b), in case that the tree appears in front of the car, which means it is newly observed, as a result, the voxel query actually gains no benefit from temporal interaction. We evaluate the perception accuracy of the front and rear area of the ego-vehicle respectively, the finding indicates that the accuracy of the rear area can be approximately 3% more than that of front area, and obviously, reversing the video yields opposite result. Hence, we combine the inference outputs of normal video playback and reverse video playback to produce the final predictions. Additionally, to enhance the accuracy of the reverse video playback mechanism, we conduct fine-tuning for the trained model using reversed data for 8 epochs.

2.4. Test-Time Augmentation

We apply horizontal flipping to input images for spatial test-time augmentation, and notably, in our experiments on the

Table 1. **3D Occupancy Prediction on nuScenes Validation Set.** Our ViewFormer demonstrates significant performance improvements over previous SOTAs in terms of both the mIoU and IoU$_{geo}$.

Method	mIoU	IoU$_{geo}$	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	driv. surf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [17]	7.31	23.96	4.03	0.35	8.00	8.04	2.90	0.28	1.16	0.67	4.01	4.35	27.72	5.20	15.13	11.29	9.03	14.86
BEVFormer [9]	16.75	30.50	14.22	6.58	23.46	28.28	8.66	10.77	6.64	4.05	11.20	17.78	37.28	18.00	22.88	22.17	13.80	22.21
TPVFormer [15]	17.10	30.86	15.96	5.31	23.86	27.32	9.79	8.74	7.09	5.20	10.97	19.22	38.87	21.25	24.26	23.15	11.73	20.81
SurroundOcc [18]	20.30	31.49	20.59	11.68	28.06	30.86	10.70	15.14	14.09	12.06	14.38	22.26	37.29	23.70	24.49	22.77	14.89	21.86
Version A	20.85	34.86	20.35	9.17	27.49	31.20	12.19	14.64	9.55	6.94	10.34	22.42	45.62	27.17	30.84	27.72	13.50	24.46
Version B	22.44	36.24	22.63	12.34	29.69	32.06	12.70	16.74	11.24	7.97	10.89	24.28	46.40	30.17	31.62	28.98	15.40	25.99
Version C	22.35	35.77	21.64	11.57	29.94	32.87	12.90	18.21	11.41	7.48	11.09	24.17	46.43	29.69	31.63	28.60	15.48	24.48
Version D	24.71	38.87	24.03	12.13	30.51	33.44	16.91	19.80	13.40	8.63	13.68	26.41	48.10	32.75	34.00	31.43	20.57	29.50
Version E	25.66	40.68	25.65	12.13	30.51	33.44	17.95	19.80	13.40	9.75	13.68	26.41	49.38	35.20	36.09	33.67	22.21	31.36
Version F	26.67	39.75	26.70	18.05	32.21	35.42	16.08	23.77	17.16	14.66	15.66	28.03	47.01	31.32	34.08	31.56	22.18	32.77
Version G	30.68	45.57	30.38	19.81	34.03	37.43	22.88	27.74	19.62	17.55	19.23	30.96	52.09	36.45	38.36	36.40	29.16	38.74

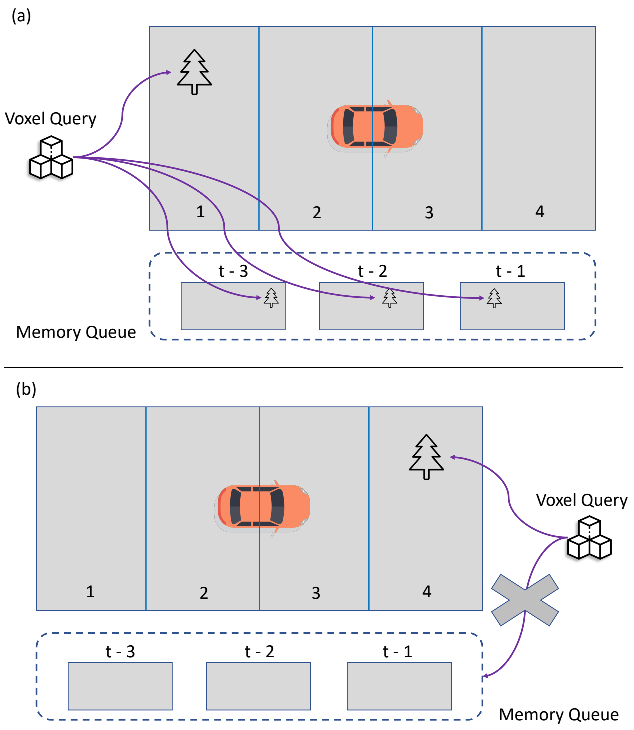

Figure 2. Online video temporal interaction.

RoboDrive dataset [19], we find that 3D spatial flips yield no improved accuracy. In addition, we introduce a temporal test-time augmentation strategy by transforming the inference results of the entire scene into a unified global coordinate system to facilitate result fusion, where the weighting factor assigned to each occupancy is determined by its distance to the ego-vehicle in the respective frame. We then map the occupancies of static semantic categories back to each individual frame and replace the corresponding static occupancy results.

3. Experiments

3.1. Datasets and Metrics

This work follows the protocol in the 2024 RoboDrive Challenge [19] 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 occupancy dataset is built based on the nuScenes dataset [20, 21]. The occupancy annotations are generated by [18]. The evaluation data was created following RoboDepth [22–24], RoboBEV [25–27], and Robo3D [28, 29]. 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 Mterics

In terms of evaluation metrics, we utilize the mean Intersection-over-Union (mIoU) across categories and the single-class IoU$_{geo}$ to evaluate the occupancy state.

3.3. Training Strategies

The framework is implemented using the PyTorch framework [30] and is based on the MMDetection3D codebase [31]. We train all ablation and large-scale models with a batch size of 8 on 8 NVIDIA A100 GPUs and utilize AdamW optimizer with a learning rate of 2×10^{-4} and a weight decay of 0.05, where the learning rate of the backbone undergoes layer-wise decay. Our models are trained for 24 epochs for occupancy tasks. Additionally, we fine-tune the trained model on RVP data for 8 epochs as mentioned above.

3.4. Ablation

During our exploration period, we evaluate our various methods with a smaller model scale, which adopts an input image size of 256 × 704 and a voxel resolution of 50 × 50 × 8, leveraging the InternImage-Tiny [32] image backbone. The remarkable milestones are outlined in Tab. 1. Version A serves as our baseline approach. In Version B, we introduce depth supervision following BEVDepth [33]. Version C showcases the results of the fine-tuned model with RVP data. Version D combines the outcomes of Version B and Version C, as discussed in Section 2.3. In Version E, we employ test-time augmentations. It is worth noting that the existing SOTA SurroundOcc [18], as shown in Tab. 1, adopts a large model scale with ResNet101 backbone, 900 × 1600 image size, and 200 × 200 × 16 voxel resolution.

3.5. Scaling Up

After the method exploration, we advance to upscale our model. For Version F and Version G in Tab. 1, we employ the larger ConvNeXt V2-H [34] backbone, an input image size of 960 × 1760 and a voxel resolution of 100 × 100 × 8. The other setups of Version F remains consistent with Version B, while the other setups of Version G remains consistent with Version E. Version G achieves 20.79% mIoU on the phase-2 corruption test set of the 2024 RoboDrive Challenge [19]. In our experiments, increasing the voxel resolution to 200 × 200 slightly improves accuracy on the validation set but marginally reduces accuracy on the corruption test set.

3.6. Ensemble

A model ensemble approach is applied to the final submission, where we employ popular image backbones pretrained on ImageNet [35], including ConvNeXt V2 [34], InternImage [32], and BEiT-based ViT-adapter [36]. By fusing the outputs of all the models through ensemble techniques, we achieve our best result on the phase-2 corruption test set with a mIoU score of 22.31%.

4. Conclusion

In the 2024 RoboDrive Challenge, our solution is based on our proposed ViewFormer, a robust vision-centric spatiotemporal modeling method with view-guided transformers. We also introduce offline techniques to further enhance the final results, including the reverse video playback mechanism to leverage future data, model scaling up, and effective post-processing strategies. Through all the enhancements and optimizations, we rank the 1st place in the challenge with exceptional mIoU scores of 24.07% and 22.31% on the phase-1 and phase-2 corruption test sets respectively.

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 *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] 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.

[4] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *Advances in Neural Information Processing Systems*, 36, 2024.

[5] Ruihang Miao, Weizhou Liu, Mingrui Chen, Zheng Gong, Weixin Xu, Chen Hu, and Shuchang Zhou. Occdepth: A depth-aware method for 3d semantic scene completion. *arXiv preprint arXiv:2302.13540*, 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] 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.

[8] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers. *arXiv preprint arXiv:2405.04299*, 2024.

[9] 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 *ECCV*, pages 1–18, 2022.

[10] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, pages 11618–11628, 2023.

[11] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *ICLR*, 2023.

[12] Yue Wang, Vitor Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. DETR3D: 3d object detection from multi-view images via 3d-to-2d queries. In *CoRL*, pages 180–191, 2021.

[13] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. PETR: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548, 2022.

[14] Chonghao Sima, Wenwen Tong, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, and Hongyang Li. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023.

[15] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023.

[16] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M. Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, pages 6919–6928, 2023.

[17] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022.

[18] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, October 2023.

[19] 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.

[20] 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 11618–11628, 2020.

[21] Whye Kit Fong, Rohit Mohan, Juana Valeria Hurtado, Lubing Zhou, Holger Caesar, Oscar Beijbom, and Abhinav Valada. Panoptic nuscenes: A large-scale benchmark for lidar panoptic segmentation and tracking. pages 3795–3802, 2022.

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

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

[32] Wenhai Wang, Jifeng Dai, Zhe Chen, Zhenhang Huang, Zhiqi Li, Xizhou Zhu, Xiaowei Hu, Tong Lu, Lewei Lu, Hongsheng Li, Xiaogang Wang, and Yu Qiao. Internimage: Exploring large-scale vision foundation models with deformable convolutions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14408–14419, 2023.

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

[34] Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. Convnext v2: Co-designing and scaling convnets with masked autoencoders. *arXiv preprint arXiv:2301.00808*, 2023.

[35] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[36] Zhe Chen, Yuchen Duan, Wenhai Wang, Junjun He, Tong Lu, Jifeng Dai, and Yu Qiao. Vision transformer adapter for dense predictions. *arXiv preprint arXiv:2205.08534*, 2022.