Robust Occupancy Prediction based on Enhanced SurroundOcc

Bingyang Zhang[1], Lirong Zhao[2], Dianlei Ding[1], Fangsheng Liu[1], Yixiang Yan[1], Hongming Wang[1]

[1]Beijing APEC Blue Technology Co., Ltd
[2]Beihang University
zhangbingyang1@foxmail.com

Abstract

This technical report summarizes our team - APEC Blue's solution to the RoboDrive Challenge (Track 3) in conjunction with ICRA 2024. The enhanced SurroundOcc framework is introduced to refine occupancy prediction in complex driving environments. The report delves into three key aspects to enhance algorithm performance: (1) Fine-tuning the SurroundOcc network, (2) Optimizing model network structure, and (3) Ensembling multiple algorithmic models. Our approach yielded a notable mIoU score of 10.32% on the test dataset, securing 2nd place in the challenge track. Code and models have been released at: https://github.com/amazingzby/robodriveChallenge.

1. Introduction

In the rapidly evolving domain of autonomous driving, the accuracy and resilience of perception systems are paramount [1–4]. Recent advancements, particularly in bird's eye view (BEV) representations and LiDAR sensing technologies, have significantly improved in-vehicle 3D scene perception [5–9]. Yet, the robustness of 3D scene perception methods under varied and challenging conditions — integral to ensuring safe operations — has been insufficiently assessed []. To fill in the existing gap, the 2024 RoboDrive Challenge [10, 11], seeking to push the frontiers of robust autonomous driving perception, is introduced.

Our team (APEC Blue) was selected as one of the top-performing teams in Track 3: Robust Occupancy Prediction of the 2024 RoboDrive Challenge [10]. During the challenge, we investigated approaches for improving algorithm performance across three aspects based on SurroundOcc [12]: (1) Fine-tuning baseline models, (2) Optimizing model network structures, and (3) Ensembling multiple algorithmic models.

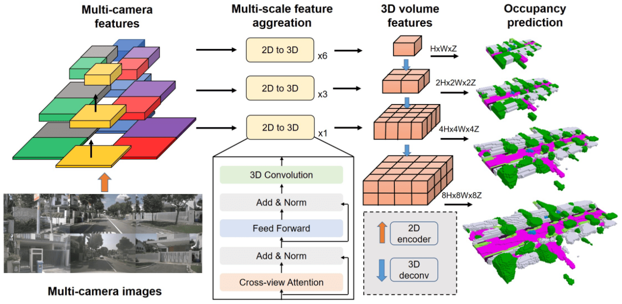

Figure 1. Overview of the SurroundOcc framework [12].

2. Approach

In this section, we will provide an introduction to our method. We conducted a comparison between the occupancy networks SurroundOcc [12] (baseline method) and FB-Occ [13] (state-of-the-art method as of CVPR 2023). It is observed that FB-Occ exhibits a much lower mean Intersection over Union (mIoU) on the validation set. Consequently, our focus lies in the analysis and optimization of the occupancy prediction task based on SurroundOcc. The pipeline of SurroundOcc is illustrated in Fig. 1.

In our model analysis, we focused on examining the influence of backbone networks and 2D-3D feature transformations on the mIoU metric. Specifically, we compared the performance of ResNet-101 [14] and VoVNet-99 [15] as backbone networks, and found that VoVNet-99 surpassed ResNet-101 in mIoU on the validation set. However, on the test set, VoVNet-99 exhibited notably inferior performance compared to ResNet-101. Nevertheless, the integration of VoVNet-99 into a multi-model ensemble led to enhancements in the overall algorithm performance. Regarding 2D-3D feature transformation, we primarily experimented with different voxel sizes to assess their impact on mIoU performance. Our experiments revealed that increasing the voxel size from [100,100,8] to [200,200,16] resulted in a slight improvement in algorithm performance on both the validation

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

Table 1. Occupancy prediction performance of different settings on the test dataset.

Method	brigh	dark	fog	frost	snow	contr	defoc	glass	motio	zoom	elast	quant	gauss	impul	shot	iso	pixel	jpeg	mIoU(%)
A	12.07	10.87	7.83	8.79	7.49	4.45	11.83	12.74	4.33	7.38	13.43	7.62	8.07	8.04	9.88	10.48	14.21	9.71	9.40
B	12.61	11.51	8.44	8.96	7.48	4.44	12.20	12.85	3.91	7.23	13.95	7.27	8.20	8.73	10.13	11.15	14.54	9.68	9.63
C	9.92	7.56	6.77	7.18	5.20	4.46	9.52	12.70	3.91	5.82	15.16	5.99	5.41	7.11	7.69	8.45	12.24	8.91	8.00
D	11.88	10.15	6.96	8.36	7.19	4.92	11.35	12.32	3.88	6.77	13.59	6.40	7.04	7.63	9.46	9.84	13.45	9.31	8.92
E	11.17	10.20	5.96	8.39	6.20	4.91	10.88	12.82	4.65	7.50	13.49	7.29	7.66	8.13	9.40	9.97	14.04	9.73	9.02
F	13.00	11.97	8.58	9.35	7.81	4.85	12.64	13.46	4.26	7.67	14.41	7.93	8.63	9.05	10.54	11.43	14.97	10.31	10.05
G	13.14	11.85	8.52	9.47	7.31	5.06	12.68	14.03	4.21	7.43	15.66	7.75	8.56	9.10	10.55	11.45	15.12	10.58	10.14
H	13.28	12.16	8.67	9.71	7.80	5.23	13.01	14.01	4.32	7.68	15.28	7.96	8.80	9.25	10.80	11.72	15.27	10.76	10.32

and test sets.

In terms of algorithm optimization, our primary focus was on exploring ways to enhance algorithm performance through model fine-tuning and loss optimization. In the initial stages of the competition, we initiated the process by fine-tuning the model using loss solely from the final layer. Our rationale is that channeling the model's attention towards the loss associated with the final layer, as opposed to considering losses from all feature layers, would result in superior algorithm performance through fine-tuning. Subsequently, we delved deeper into optimizing various loss functions to boost the algorithm's performance. These included Focal Loss, weighted softmax loss, Lovasz-softmax loss, L_{scal}^{geo}, and L_{scal}^{sem}. By incorporating a combination of these diverse loss functions, we were able to further elevate the algorithm's performance.

Finally, we explored the enhancement of algorithm performance through model ensemble. Initially, integrating multiple models contributed to the improvement of the mIoU metric. Subsequently, increasing both the quantity and diversity of model ensembles further elevated the algorithm's performance.

3. Experiments

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [10] 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 [16–18], RoboBEV [8, 19, 20], and Robo3D [21, 22]. 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. Experimental Setups

The framework is implemented using the PyTorch framework [23] and is based on the MMDetection3D codebase [24]. We train our models using a batch size of 2 on 2 NVIDIA V100 GPUs. The resolutions of all models are set to 1600x900. The summary of results is presented in Tab. 1.

3.3. Ablation Study

Model Fine-tuning: Model A and B in Tab. 1 are the results of the model fine-tuning section. Model A(9.40) utilizes a smaller learning rate and trains solely on the loss from the final layer, building upon the baseline model through fine-tuning. Model B(9.63), based on Model A, further increases the number of iterations and incorporates additional loss methods. In the single-model section, both Model A and Model B achieve better performance metrics compared to other models. We attribute this not only to improvements in fine-tuning and loss optimization but also to the increased number of training iterations. Due to limitations in training resources, subsequent model iterations are limited to 15 - 20 epochs, far fewer than the training iterations in the fine-tuning section. We believe that with sufficient training resources, subsequent models would achieve better results.

Network Structure: Models C to E are the results of the model structure optimization section. In this part, models were trained from scratch, and results from different network structures were compared. Model C(8.0) utilizes the VoVNet-99 backbone network and achieves the best results among all models on the validation set. However, it performs the worst on the test set, indicating poor generalization performance. Model D(8.92) employs ResNet-101, consistent with the baseline model, and adopts the training strategy of the best model in the fine-tuning section, resulting in slightly better results than the baseline(8.66). Model E(9.02) further improves the mIoU metric by increasing the voxel size from [100,100,8] to [200,200,16] in the structure of Model D.

Ensemble: Models F to H are the results of the model ensemble section. The model ensemble strategy is as follows: (1) For each voxel, the predicted class with the highest frequency is chosen as the final prediction. (2) If a voxel has multiple equally predicted classes, the model with a higher mIoU value is given higher priority. Comparing Model F (10.05, ensemble of A, B, E) with Model G (10.14, ensemble of A, C, E) demonstrates that improving model diversity (backbone diversity and voxel size diversity) can enhance ensemble performance. The comparison between Models F & G, and H (10.32, ensemble of all 5 models) indicates that increasing the number of ensemble models can further enhance performance.

4. Conclusion

In this report, we summarized our winning solution for the RoboDrive Challenge (Track 3) in conjunction with ICRA 2024. In the future work, we think that sufficient training resources and the incorporation of temporal information have the potential to enhance algorithm performance.

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] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] https://codalab.lisn.upsaclay.fr/competitions/17063.

[12] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving, 2023. arXiv preprint arXiv:2303.09551.

[13] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. FB-OCC: 3D occupancy prediction based on forward-backward view transformation, 2023. arXiv:2307.01492.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2016.

[15] Youngwan Lee, Joong won Hwang, Sangrok Lee, Yuseok Bae, and Jongyoul Park. An energy and gpu-computation efficient backbone network for real-time object detection, 2019. IEEE/CVF Conference on Computer Vision and Pattern Recognition.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

[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] 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.

[23] 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.

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