Improving Out-of-Distribution Robustness of Occupancy Prediction Networks with Advanced Loss Functions

Nanfei Ye, Lun Luo, Xun Wu, Yubo Tian, Zhe Cao, Yunfan Li, Yiwei Zuo, Wenjie Liu, Yi Ren

Haomo.AI

{yenanfei, luolun, wuxun, tianyubo, caozhe, zuoyiwei, liuwenjie}@haomo.ai

yunfan.li@stonybrook.edu, yrennnn@gmail.com

Abstract

We tested the state-of-the-art occupancy prediction method SurroundOcc in the RoboDrive Challenge 2024. We tried replacing its backbone and using more loss functions to improve its robustness to out-of-distribution data. We achieved 8.9% mIoU with ResNet101 as the backbone and a loss function consisting of cross-entropy loss, segmentation scale loss, geo scale loss, and Lovasz softmax loss. Our method ranked the 3rd in the competition.

1. Introduction

Occupancy [1] is a newly developed perception task for autonomous driving cars. It assigns an occupied probability to each voxel in the 3D space to build a drivable area in the 3D space [2–5]. occupancy unifies detection tasks such as object detection and road segmentation into one model, showing great potential to help achieve fully autonomous driving [6–9].

Modern occupancy prediction methods [1, 10–12] concentrate on improving occupancy prediction accuracy with multi-view image inputs. However, their evaluation benchmarks usually lack out-distribution evaluation protocols, which is essential for algorithms to adapt to real-world conditions. Specifically, the commonly used datasets such as unScenes [2] and Argoverse [13] have the problem that the test and training sets have large overlaps. Under these evaluation protocols, current methods that achieve high intersection-of-union (IoU) may not perform well in real-world scenes. In addition, these datasets lack enough simulation to weather condition changes, sensor failures, and image distortions. Thus, the methods trained on these datasets have little robustness to hardware failure and sensor noise, which could result in perception failure and traffic accidents.

The 2024 RoboDrive Challenge [14] targets probing the Out-of-Distribution (OoD) robustness of state-of-the-art autonomous driving perception models, centered around two mainstream topics: common corruptions and sensor failures. The challenge provides eighteen real-world corruption types in total, ranging from three perspectives:

- Weather and lighting conditions, such as bright, low-light, foggy, and snowy conditions.
- Movement and acquisition failures, such as potential blurs caused by vehicle motions.
- Data processing issues, such as noises and quantizations happen due to hardware malfunctions.

It provides several types of sensor failures including:

- Loss of certain camera frames during the driving system sensing process.
- Loss of one or more camera views during the driving system sensing process.
- Loss of the roof-top LiDAR view during the driving system sensing process.

The 2024 RoboDrive Challenge [14] tries to fill the gap in performance between academic studies and industrial applications and seeks to push the frontiers of robust autonomous driving perception [15, 16].

In this competition, we conducted various experiments to explore the Out-of-Distribution (OoD) robustness of the state-of-the-art occupancy prediction method. We tried different combinations of backbones and loss functions. We achieved the highest mIoU score of 8.94% with the backbone of ResNet101, the multiple loss functions with cross-entropy loss, segmentation scale loss, geo scale loss, and Lovasz softmax loss [17]. Our method ranked 3rd in the competition.

2. Approach

The pipeline of SurroundOcc [1] is illustrated in Fig. 1. It first uses a backbone network to extract N cameras' and M levels' multi-scale features $X = \{\{X_i^j\}_{i=1}^{N}\}_{j=1}^{M}$. For each

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

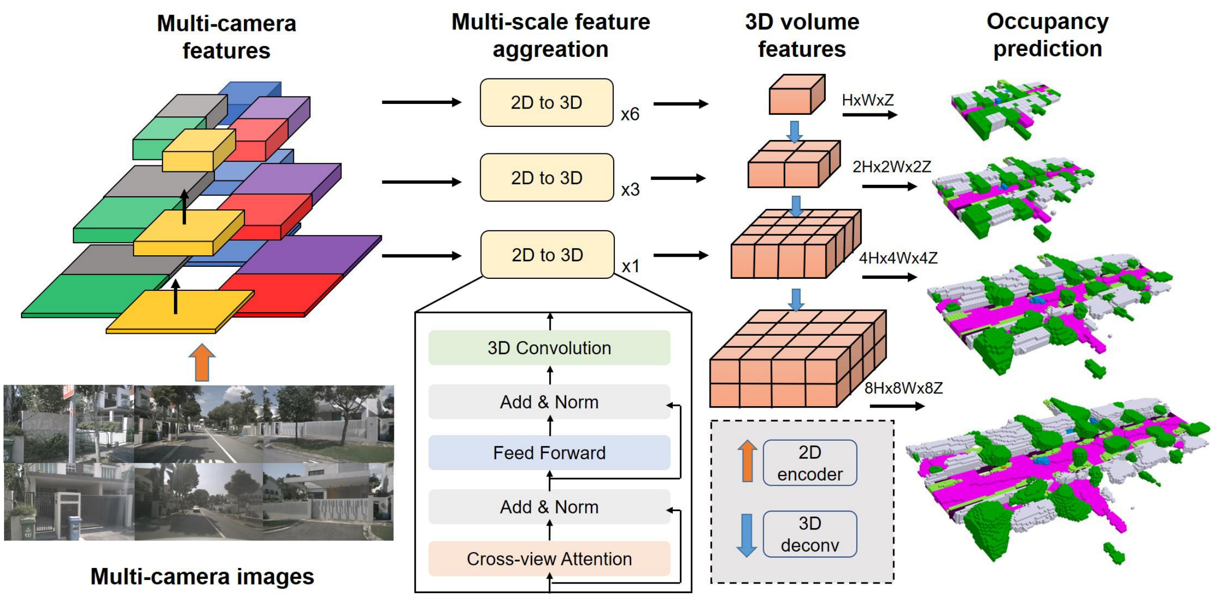

Figure 1. The pipeline of SurroundOcc. First, it uses a backbone to extract multi-scale features of multi-camera images. Then it adopts 2D-3D spatial attention to fuse multi-camera information and construct 3D volume features in a multi-scale fashion. Finally, the 3D deconvolution layer is used to upsample 3D volumes and occupancy prediction is supervised in each level.

level, it uses a transformer to fuse multi-camera features with spatial cross-attention. The output of the 2D-3D spatial attention layer is a 3D volume feature instead of the BEV feature. Then the 3D convolution network is utilized to upsample and combine multi-scale volume features. The occupancy prediction in each level is supervised by the generated dense occupancy ground truth with a decayed loss weight.

3. Experiments

3.1. Datasets

This work follows the protocol in the 2024 RoboDrive Challenge [14] when preparing the training and test data. Specifically, the model is trained on the official *train* split of the nuScenes dataset [2] and tested on the held-out competition evaluation sets. The evaluation data was created following RoboDepth [18–20], RoboBEV [21–23], and Robo3D [15, 24]. 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

We tested three backbones in this competition including ResNet34 [25], ResNet101 [25], and VoVNet-99 [26]. VoVNet-99 is a densely connected module. Compared to ResNet34, it aggregates features in early layers by concatenation to better preserve their characteristics in the output. In addition, it uses effective Squeeze-Excitation to perform the channel attention for the feature maps.

The occupancy prediction task can be considered as a 3D extension of 2D image pixel segmentation. Since the mIoU (mean Intersection-over-Union) metric is adopted for this competition, the Lovasz Softmax loss [17] designed for directly optimizing the mIoU in the multi-class image segmentation task could be an effective alternative. In addition to the cross-entropy loss, semantic classification loss, and geometry classification loss, there are 4 loss items used in our method. In this competition, we tried two strategies to balance these losses, i.e. loss normalization [27] and uncertainty loss [28]. The loss normalization scales each loss to ease the optimization, while the uncertainty loss considers each loss weight as a learnable parameter thus enabling adaptive learning for loss weights.

3.3. Implementation Details

The framework was implemented using the PyTorch framework [29] and was based on the MMDetection3D codebase [30]. We used 8 NVIDIA A100 GPUs for training, each with a batch size of 1. We optimized the method end-to-end with the AdamW optimizer for 24 epochs. We employed a cosine annealing learning adjustment strategy with a period of 500 iterations, setting the maximum and minimum learning rates to 2e-4 and 2e-7, respectively.

Table 1. mIoU of SurroundOcc under different configurations.

Backbone	Loss	Training Strategy	mIoU
VoVNet-99	w/o Lovasz Softmax loss	-	7.36
VoVNet-99	with Lovasz Softmax loss	-	7.45
VoVNet-99	with Lovasz Softmax loss	loss norm / uncertainty loss	7.80
baseline	-	-	8.66
ResNet101	with Lovasz Softmax loss	loss norm / uncertainty loss	8.94

3.4. Comparative Study

Table 1 shows the results under different backbones, loss combinations, and training strategies. As can be seen, Lovasz Softmax loss [17] introduces a minor performance improvement with VoVNet-99 as the backbone. The loss normalization and uncertainty loss further improve the mIoU to 7.80%. However, all these strategies do not reach the performance level of the official baseline. We finally achieved a mIoU of 8.94% by fine-tuning the official checkpoint with the Lovasz Softmax loss [17] and both loss balance strategies.

4. Conclusion

This work explored the out-of-distribution robustness of the state-of-the-art occupancy prediction method SorroundOcc. By setting different backbones and loss functions, we achieved an mIoU of 8.9% and ranked 3rd in the RoboDrive Challenge 2024. More performance improvement is likely achieved by deeply analyzing the statistics of PV and BEV feature changes under various distortions. Masked autoencoders are also beneficial in improving the robustness.

References

[1] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 21672–21683, 2023.

[2] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11621–11631, 2020.

[3] 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 (CVPR)*, pages 7020–7030, 2023.

[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 (CVPR)*, pages 21705–21715, 2023.

[5] 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 (NeurIPS)*, volume 36, 2024.

[6] 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 (NeurIPS)*, 36, 2024.

[7] 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.

[8] 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 (ICRA)*, pages 9338–9345, 2023.

[9] 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 (ICCV)*, pages 228–240, 2023.

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

[11] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9399–9409, 2023.

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

[13] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8740–8749, 2019.

[14] 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.

[15] 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 (ICCV)*, pages 19994–20006, 2023.

[16] 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.

[17] Maxim Berman, Amal Rannen Triki, and Matthew B. Blaschko. The lovasz-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4413–4421, 2018.

[18] 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 (NeurIPS)*, 36, 2024.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[23] 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.

[24] 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.

[25] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[26] 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. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 752–760, 2019.

[27] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17804–17813, 2023.

[28] Roberto Cipolla, Yarin Gal, and Alex Kendall. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7482–7491, 2018.

[29] 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 (NeurIPS)*, 2019.

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