Supplementary Material for "SurroundOcc: Multi-Camera 3D Occupancy Prediction for Autonomous Driving"

Yi Wei[1,2]*, Linqing Zhao[3]*, Wenzhao Zheng[1,2], Zheng Zhu[4], Jie Zhou[1,2], Jiwen Lu[1,2]†

[1]Beijing National Research Center for Information Science and Technology, China
[2]Department of Automation, Tsinghua University, China
[3]School of Electrical and Information Engineering, Tianjin University, China
[4]PhiGent Robotics

{y-wei19,zhengwz18}@mails.tsinghua.edu.cn; linqingzhao@tju.edu.cn;
zhengzhu@ieee.org; {jzhou, lujiwen}@tsinghua.edu.cn

A. Baseline Method Details

We compare with several baseline methods on nuScenes dataset, which can be roughly classified as four categories:

Depth estimation: SurroundDepth [13], AdaBins [2], NeWCRFs [14]. Since SurroundDepth method is multi-camera self-supervised method, we use depth groundtruth to supervise the network along with self-supervised photometric loss. AdaBins [2] and NeWCRFs [14] are the state-of-the-art depth estimation methods both in outdoor and indoor scenes. To implement these two methods, we use their official released code with the dataloader in SurroundDepth. The depth results are fused by the TSDF fusion algorithm [5, 11] with the voxel size $0.5m$, which is same to our method.

3D scene reconstruction: Atlas [10] and Transformerfusion [3]. These two methods are state-of-the-art indoor scene reconstruction methods. We use our dense occupancy groundtruth to supervise them instead of tsdf ground truth. To fairly compare, we also adopt ResNet101-DCN [7, 6] with the initial weight from FCOS3D [12] as the backbone to extract image features.

Occupancy reconstruction: MonoScene [4] and TPVFormer [8]. To extend MonoScene to multi-camera setting, we project occupancy labels to each camera's coordinate and the shape of each camera's prediction is $(128, 104, 16)$ with 0.5m voxel size. We fuse multi-camera results in LiDAR coordinate with camera extrinsics. The final result has the same shape and voxel size with ours. For TPVFormer, the resolution is set as 200x200x16 and the feature dimension is 64.

BEV perception: BEVFormer [9]. We use the full-resolution 200x200 BEV features. To lift BEV features to the 3D space, we split 256 dimensions BEV features as 16 grids and the feature of each grid has 16 dimensions. Then we adopt a 3D encoder-decoder network [10] as a segmentation head to predict occupancy. Following the setting in TPVFormer, we employ both cross entropy loss and lovasz-softmax [1] as the supervision signals.

Acc	$\text{mean}_{p\in P}(\min_{p^*\in P^*}\|\|p-p^*\|\|)$
Comp	$\text{mean}_{p^*\in P^*}(\min_{p\in P}\|\|p-p^*\|\|)$
Prec	$\text{mean}_{p\in P}(\min_{p^*\in P^*}\|\|p-p^*\|\|<0.5)$
Recal	$\text{mean}_{p^*\in P^*}(\min_{p\in P}\|\|p-p^*\|\|<0.5)$
CD	Acc + Comp
F-score	$(2\times\text{Prec}\times\text{Recal})/(\text{Prec}+\text{Recal})$

Table 1. Evaluation metrics for 3D scene reconstruction. p and p^* are the predicted and ground truth point clouds.

B. More Visualizations

Figure 1 shows the qualitative comparison with other methods. We can see that our predictions are more accurate and denser. We also provide some video demos in the material. Specifically, 'demo-nuscenes' shows the results on nuScenes validation set and 'demo-gt' visualizes our generated groundtruth. 'demo-comparison' illustrates the comparison with other methods and 'demo-wild' shows the occupancy predictions on Beijing street (trained on nuScenes training set).

References

[1] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *CVPR*, pages 4413–4421, 2018. 1

*Equal contribution.
†Corresponding author.

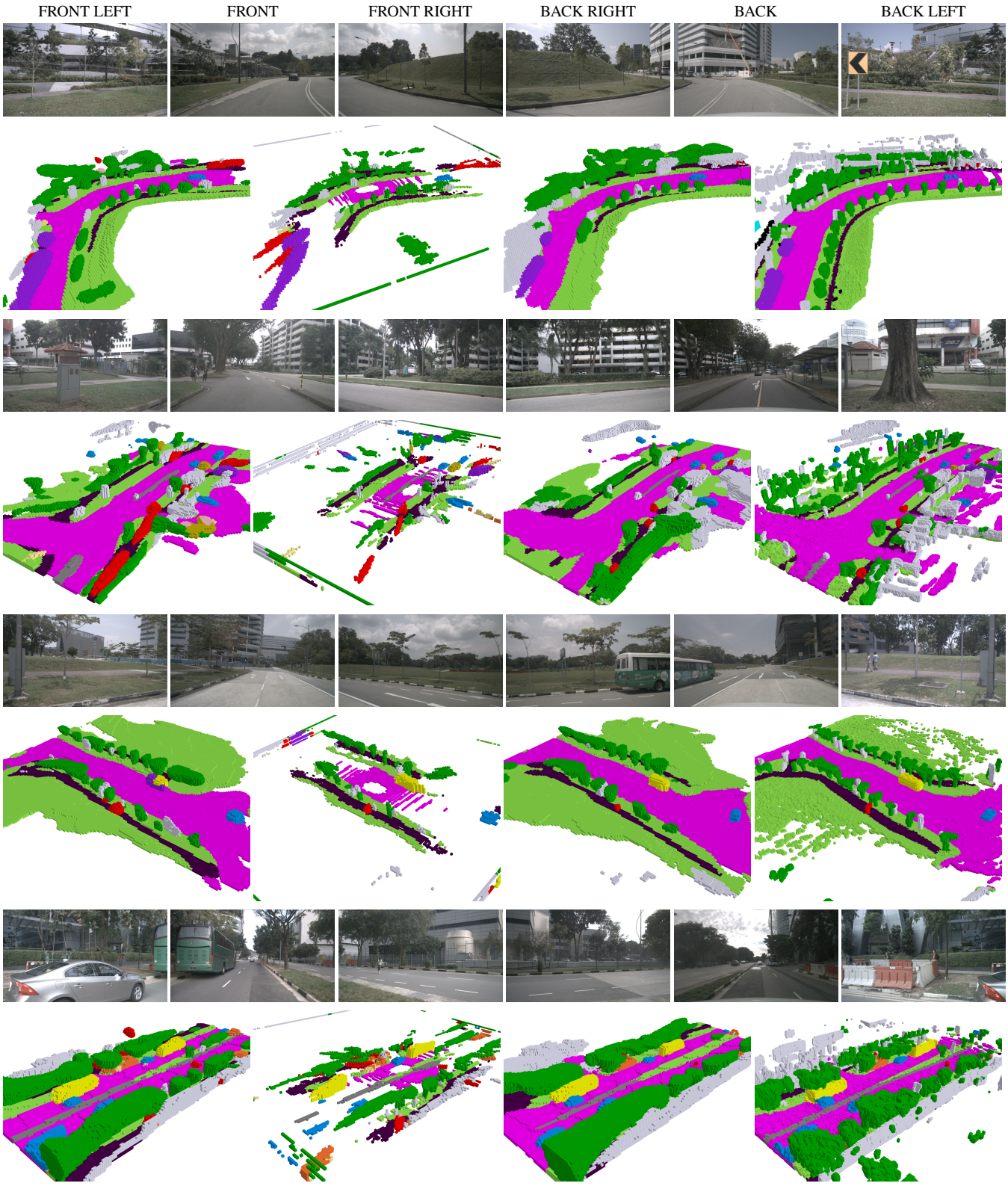

Figure 1. Qualitaative comparison on nuScenes validation set. Our mrthod can predict more accurate and denser occupancy. **Better viewed when zoomed in.**

[2] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *CVPR*, pages 4009–4018, 2021. 1

[3] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 1

[4] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 1

[5] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *SIGGRAPH*, pages 303–312, 1996. 1

[6] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *ICCV*, pages 764–773, 2017. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[8] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2302.07817*, 2023. 1, 2

[9] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 2

[10] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *ECCV*, pages 414–431, 2020. 1

[11] Richard A Newcombe, Shahram Izadi, Otmar Hilliges, David Molyneaux, David Kim, Andrew J Davison, Pushmeet Kohi, Jamie Shotton, Steve Hodges, and Andrew Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *ISMAR*, pages 127–136, 2011. 1

[12] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, 2021. 1

[13] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *CoRL*, 2022. 1

[14] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. New crfs: Neural window fully-connected crfs for monocular depth estimation. In *CVPR*, 2022. 1