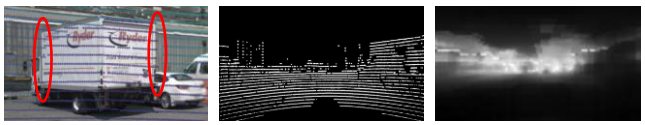
Co-Occ: Coupling Explicit feature Fusion with Volume Rendering Regularization for Multi-Modal 3D Semantic Occupancy Prediction –Supplementary Material–

Jingyi Pan[1], Zipeng Wang[1], Lin Wang[1,2,*]

I. Details

A. The Voxel Interaction of LiDAR and Camera

Fig. 1 (a) illustrates the challenge of aligning images and point clouds caused by inaccurate extrinsic parameters. Direct geometric alignment is difficult to achieve. To address the accumulation of errors resulting from misalignment, we propose GSFusion. This method searches for nearby features to ensure both geometric and semantic alignment, enabling each LiDAR voxel feature to interact with K neighboring lifted pixel features in the fusion process. This expands the perception field, allowing for a more comprehensive and robust fusion of image and point features. Furthermore, Fig. 1 (b) highlights the impact of the sparsity of LiDAR point clouds on voxel interaction with the camera. To address this, the rendering process ensures dense representations for LiDAR features, camera features, or LiDAR-camera features, as depicted in Fig. 1 (c). This ensures sufficient voxel interaction and improves overall performance.



(a) (b) (c)

Fig. 1. (a) illustrates the inaccurate calibrations between the projected points and the corresponding images. (b) displays the sparse distribution of projected LiDAR points. (c) showcases the rendered depth obtained from the fusion of LiDAR-camera features.

B. Other Implementation Details

In our implicit volume rendering regularization, we employ two Multi-Layer Perceptron (MLP) networks as our density head and color head. These MLP networks are utilized to generate the density grid and color grid from the sampled frustum features. The color head consists of three layers of MLP, while the density head consists of either one layer of MLP or a Linear layer. This configuration strikes a balance between training memory requirements and rendering performance.

*Corresponding author

[1]J. PAN and Z. Wang are with the AI Thrust, The Hong Kong University of Science and Technology (Guangzhou), Guangdong 511458, China. {jpan305, zwang253}@connect.hkust-gz.edu.cn

[1,2]L. Wang is with AI/CMA Thrust, HKUST(GZ) and Dept. of CSE, HKUST, Hong Kong SAR, China, Email: linwang@ust.hk

C. Evaluation Metrics

IoU metrics. The Intersection over Union (IoU) is a metric used to determine whether a voxel is occupied or empty [1], [2]. It treats all occupied voxels as the occupied class and all others as the empty class. The IoU is calculated as follows:

$$\mathrm{IoU} = \frac{\mathrm{TP_o}}{\mathrm{TP_o} + \mathrm{FP_o} + \mathrm{FN_o}}, \tag{1}$$

where TP_o, FP_o, FN_o represent the number of true positives, false positives, and false negatives of the occupied class, respectively.

mIoU metrics. The mean Intersection over Union (mIoU) is a metric that calculates the average IoU for each semantic class. It is defined as follows:

$$\mathrm{mIoU} = \frac{1}{\mathrm{N_c}} \sum_{\mathrm{c}=1}^{\mathrm{N_c}} \frac{\mathrm{TP_c}}{\mathrm{TP_c} + \mathrm{FP_c} + \mathrm{FN_c}}, \tag{2}$$

where TP_c, FP_c, FN_c represent the number of true positives, false positives, and false negatives for class c, respectively, and N_c is the total number of classes.

II. More Results

More Quantitative Results on NuScenes dataset. We supply the 3D semantic occupancy prediction results on different voxel resolutions. As the openoccupancy benchmark [2], the voxel resolution is $0.2m$ in a volume of $512 \times 512 \times 40$ for occupancy predictions. As presented in Tab. I, we test other method on nuScenes-occupancy validation set [1], [2]. our Co-Occ method obtain **1.8%** mIoU improvement among the camera-LiDAR fusion-based M-CONet. Besides, the scores of most semantic classes have a large margin improvement. These experiments validate the effectiveness of our method that not only obatin better performance on one voxel resolution.

More Visualization Results on NuScenes Dataset. Due to space constraints, we have included additional visual results in Table I. These results demonstrate that our method has more precise details compared to the camera-only method [9], while also achieving greater consistency than other LiDAR-camera fusion-based methods [2]. Furthermore, the effectiveness of our methods in challenging conditions is validated through the *video demo*.

More Visualization Results on SemanticKITTI Dataset. Similarly, we present additional qualitative results using the SemanticKITTI validation dataset in Fig. 3. Our method's semantic predictions clearly outperform not only in dynamic

TABLE I

3D SEMANTIC OCCUPANCY PREDICTION RESULTS ON NUSCENES-OCCUPANCY VALIDATION SET. WE REPORT THE GEOMETRIC METRIC IOU, SEMANTIC METRIC MIOU, AND THE IOU FOR EACH SEMANTIC CLASS. THE C, D, AND L DENOTES CAMERA, DEPTH, AND LIDAR, RESPECTIVELY. **BOLD** REPRESENTS THE BEST SCORE.

Method	Modality	IoU	mIoU	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [3]	C	18.4	6.9	7.1	3.9	9.3	7.2	5.6	3.0	5.9	4.4	4.9	4.2	14.9	6.3	7.9	7.4	10.0	7.6
TPVFormer [4]	C	15.3	7.8	9.3	4.1	11.3	10.1	5.2	4.3	5.9	5.3	6.8	6.5	13.6	9.0	8.3	8.0	9.2	8.2
3DSketch [5]	C&D	25.6	10.7	12.0	5.1	10.7	12.4	6.5	4.0	5.0	6.3	8.0	7.2	21.8	14.8	13.0	11.8	12.0	21.2
AICNet [6]	C&D	23.8	10.6	11.5	4.0	11.8	12.3	5.1	3.8	6.2	6.0	8.2	7.5	24.1	13.0	12.8	11.5	11.6	20.2
LMSCNet [7]	L	27.3	11.5	12.4	4.2	12.8	12.1	6.2	4.7	6.2	6.3	8.8	7.2	24.2	12.3	16.6	14.1	13.9	22.2
JS3C-Net [8]	L	30.2	12.5	14.2	3.4	13.6	12.0	7.2	4.3	7.3	6.8	9.2	9.1	27.9	15.3	14.9	16.2	14.0	24.9
M-CONet [2]	C&L	29.5	20.1	23.3	13.3	21.2	24.3	**15.3**	15.9	18.0	13.3	15.3	20.7	33.2	21.0	22.5	21.5	19.6	23.2
Co-Occ (Ours)	C&L	**30.6**	**21.9**	**26.5**	**16.8**	**22.3**	**27.0**	10.1	**20.9**	**20.7**	**14.5**	**16.4**	**21.6**	**36.9**	**23.5**	**25.5**	**23.7**	**20.5**	**23.5**

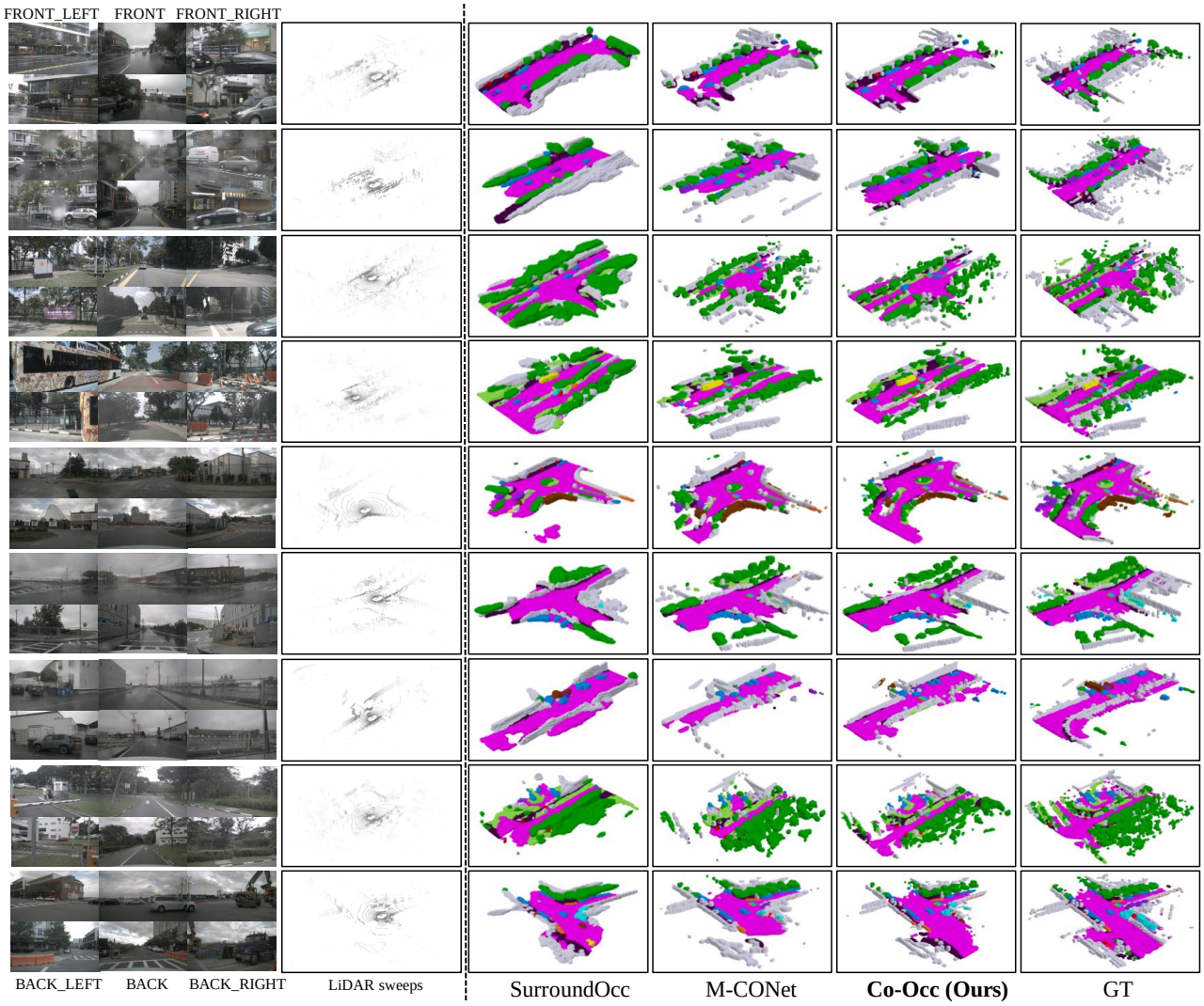


Fig. 2. The additional qualitative comparisons results on nuScenes validation set [10]. The leftmost column shows the input surrounding images and LiDAR sweeps, the following three columns visualize the 3D semantic occupancy prediction from SurroundOcc [9] (SurroundOcc predicts results using only cameras), M-CONet [2], our Co-Occ, and the annotation from [9]. **Better viewed when zoomed in.**

objects like cars but also in capturing the complementary aspects of road and vegetation. This showcases the effec-

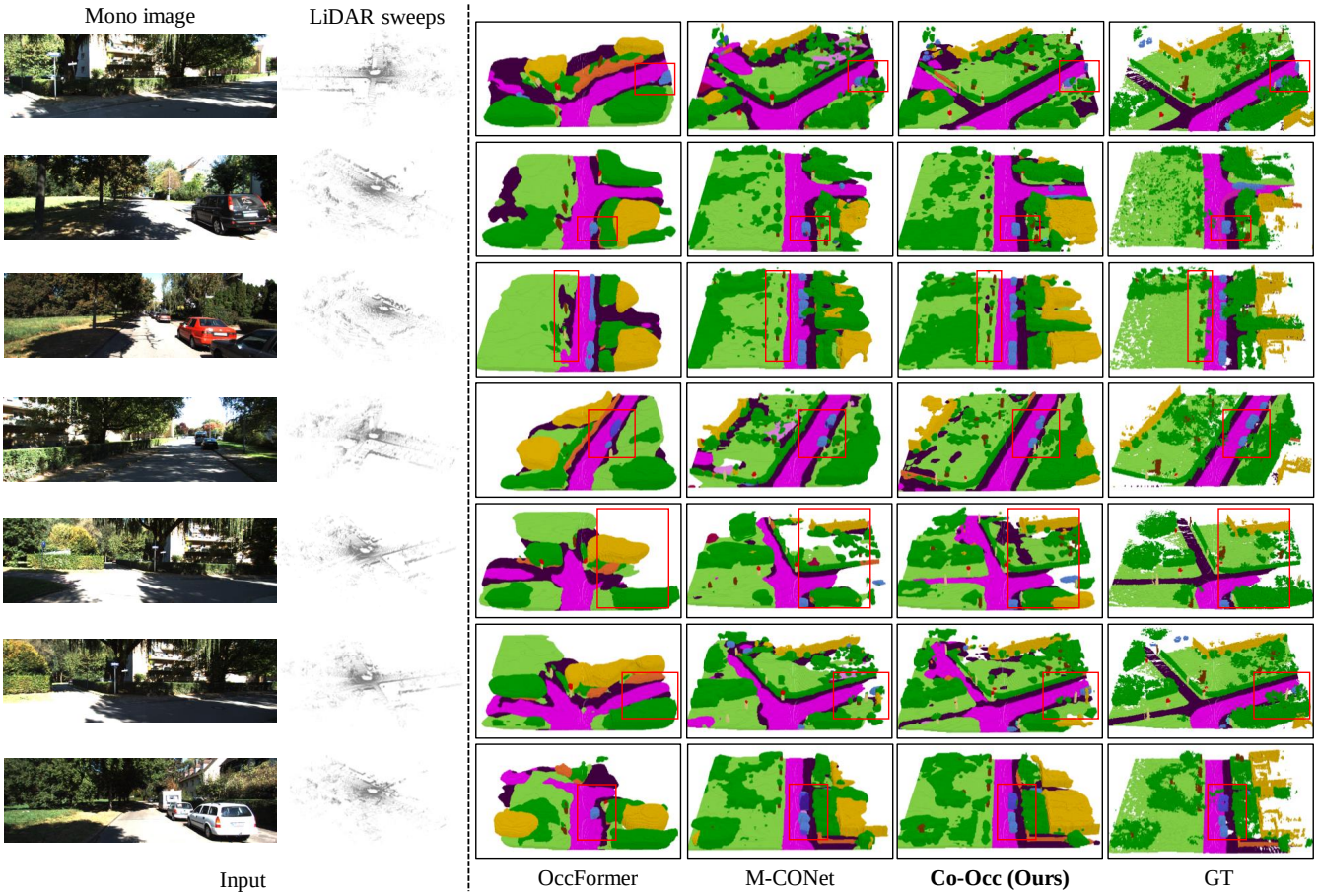


Fig. 3. The additional qualitative comparisons results on SemanticKITTI validation set. The input monocular image and LiDAR sweeps are shown on the left and the 3D semantic occupancy results from OccFormer [11] (OccFormer predicts results using only mono image), M-CONet [2], our Co-Occ, and the annotations are then visualized sequentially. **Better viewed when zoomed in.**

tiveness and versatility of our approach.

Video Demo. To provide a comprehensive understanding of our method and showcase the dynamic performance of our results, we have prepared a video demo. This demo visually presents our continuous video visualization results, allowing you to observe our method in action and gain a deeper understanding of its process.

REFERENCES

[1] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[2] X. Wang, Z. Zhu, W. Xu, Y. Zhang, Y. Wei, X. Chi, Y. Ye, D. Du, J. Lu, and X. Wang, "Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception," *arXiv preprint arXiv:2303.03991*, 2023.

[3] A.-Q. Cao and R. de Charette, "Monoscene: Monocular 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3991–4001.

[4] Y. Huang, W. Zheng, Y. Zhang, J. Zhou, and J. Lu, "Tri-perspective view for vision-based 3d semantic occupancy prediction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 9223–9232.

[5] X. Chen, K.-Y. Lin, C. Qian, G. Zeng, and H. Li, "3d sketch-aware semantic scene completion via semi-supervised structure prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 4193–4202.

[6] J. Li, K. Han, P. Wang, Y. Liu, and X. Yuan, "Anisotropic convolutional networks for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3351–3359.

[7] L. Roldao, R. de Charette, and A. Verroust-Blondet, "Lmscnet: Lightweight multiscale 3d semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 111–119.

[8] X. Yan, J. Gao, J. Li, R. Zhang, Z. Li, R. Huang, and S. Cui, "Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3101–3109.

[9] Y. Wei, L. Zhao, W. Zheng, Z. Zhu, J. Zhou, and J. Lu, "Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 21 729–21 740.

[10] J. Behley, M. Garbade, A. Milioto, J. Quenzel, S. Behnke, C. Stachniss, and J. Gall, "Semantickitti: A dataset for semantic scene understanding of lidar sequences," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9297–9307.

[11] Y. Zhang, Z. Zhu, and D. Du, "Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction," *arXiv preprint arXiv:2304.05316*, 2023.