RIOcc: Efficient Cross-Modal Fusion Transformer with Collaborative Feature Refinement for 3D Semantic Occupancy Prediction

Supplementary Material

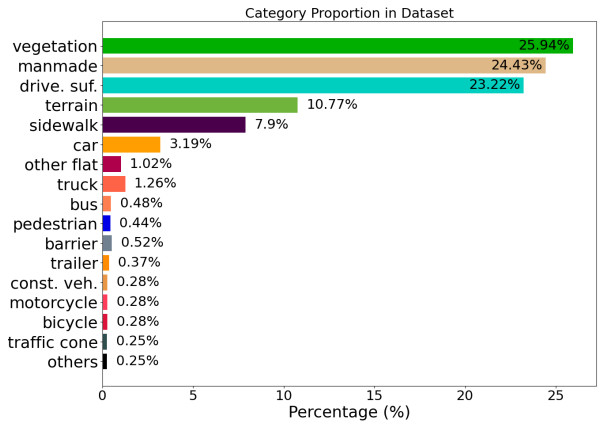

Figure 8. Class distribution of the datasets used in our paper.

6. Rationale

The supplementary document is organized as follows:

- Detailed calculation method of evaluation metrics;
- Class distribution figures for the Occ3D-nuScenes and nuScenes-Occupancy datasets.
- The generalization study of the proposed modules;
- More ablation about different backbones, computational complexity, GFLOPs, inference speed and performance.
- Experiments on module-level performance contributions.
- More visualization of the model's prediction results.

6.1. Metric

For 3D semantic occupancy prediction, the intersection over union (IoU) is used as the evaluation metric for occupied voxels in the scene completion (SC) task, disregarding their semantic class. For the semantic scene completion (SSC) task, the mean intersection over union (mIoU) across all semantic classes is used. The IoU is defined as:

$$\text{IoU} = \frac{TP}{TP + FP + FN} \tag{12}$$

The mIoU is computed as:

$$\text{mIoU} = \frac{1}{C} \sum_{i=1}^{C} \frac{TP_i}{TP_i + FP_i + FN_i} \tag{13}$$

where TP, FP, and FN represent the counts of true positives, false positives, and false negatives, respectively. C is the number of classes.

Method	DBP	DDA	mAP ↑	NDS ↑
BEVFusion (MIT)			68.52	71.38
	✓		69.00↑0.48	71.69↑0.31
		✓	69.54↑1.02	72.19↑0.81
	✓	✓	69.88↑1.37	72.32↑0.94

Table 8. **Generalization of DBP and DDA on the nuScenes val datasets.**

Methods	Modality	DBP	DDA	mIoU	Datasets
BEVDet-Occ	C		-	31.6	nuScenes -Occ3D
		✓	-	**33.0**	
FlashOcc(M1)	C		-	32.4	
		✓	-	**33.9**	
M-Baseline	C+L			15.1	nuScenes -Occupancy
		✓		16.3	
			✓	17.9	
		✓	✓	**20.4**	

Table 9. **Generalization of DBP and DDA across competitive 3D occupancy models.** All experiments conducted without the temporal module.

6.2. Dataset Class Distribution

As shown in Figure 8, we present the class distribution of the datasets. It is worth noting that the class distributions of the Occ3D-nuScenes and nuScenes-Occupancy datasets are highly similar. Therefore, we use a single chart to represent their category distributions approximately.

6.3. The Generalization Study of Our Modules

Dual-branch pooling (DBP) and Deformable Dual-Attention (DDA) in BEVFusion [27]. To validate the generalization ability of the proposed modules, we apply DBP and DDA to the 3D object detection task (BEVFusion [27]). As shown in Table 8, with DBP alone, BEVFusion achieves a notable improvement of 0.48% mAP and 0.31% NDS. Meanwhile, DDA contributes a gain of 1.02% mAP and 0.94% NDS. These results demonstrate that the proposed modules provide relatively significant improvements across different frameworks. This can be attributed to the modules' ability to extract refined structural information and semantic features while effectively bridging the gap between different modalities.

DBP and DDA in 3D semantic occupancy task. We also conduct extensive ablation experiments on several competitive 3D occupancy models. As shown in Table 9, for vision-

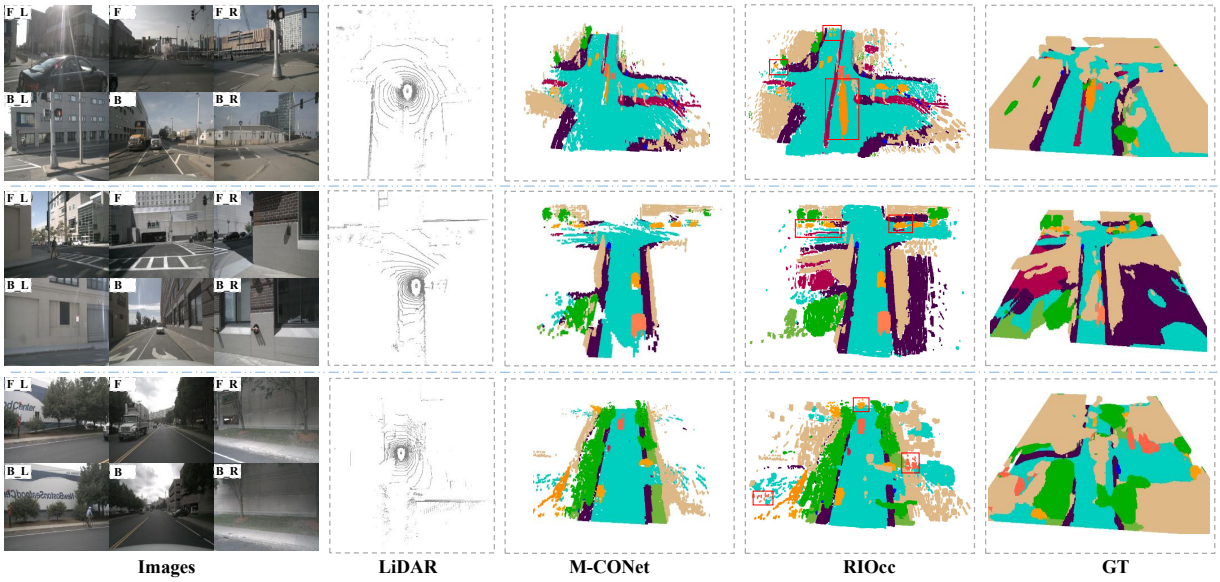

Figure 9. **Visual comparison between RIOcc and M-CONet.** Semantic occupancy visualizations on Occ3D-nuScenes are presented. The leftmost column displays the input surrounding images and LiDAR sweeps. This is followed by the results from M-CONet, our RIOcc, and the annotations from Occ3D-nuScenes. The red box highlights the effectiveness in dealing with distant and occluded targets.

Methods	Backbone	Params	GFLOPs	FPS	mIOU
Co-Occ	R101	205M	2028	2.4	21.9
M-CONet	R50	137M	3089	2.8	20.1
RIOcc	**R101**	133M	1712	2.4	26.7
RIOcc	**R50**	102M	1448	2.7	25.9
RIOcc	**R18**	87M	1308	3.5	23.5

Table 10. Comparison of different metrics.

#	DBP	Wavelet	Semantic	DDA	mIOU
1					48.01
2	✓				49.15
3	✓	✓			50.02
4	✓	✓	✓		51.40
5	✓	✓	✓	✓	54.21

Table 11. Analysis on module-level contributions.

based occupancy models, we apply DBP solely to the BEV features extracted by the respective models. The results show that applying DBP led to an mIoU improvement of 1.4 and 1.5 for BEVDet-Occ and FlashOcc, respectively. Furthermore, for multi-modal occupancy models, we adopt the M-Baseline of OpenOccupancy [47] as a foundation and introduce targeted modifications for evaluation. Specifically, all features extracted by the M-Baseline are transformed into the BEV space, subsequently processed through the Channel-to-Height module, and finally passed to an Occupancy Head for the final occupancy prediction. Importantly, DBP is integrated into the LiDAR branch. The experimental results demonstrate a substantial improvement of 5.3 mIoU, which can be primarily credited to the refined feature extraction and enhanced interaction mechanisms facilitated by our proposed modules.

6.4. More Ablation

To more intuitively demonstrate the effectiveness of the proposed model, we conducted additional ablation experiments on different backbones, computational complexity, GFLOPs, and inference speed. As shown in Table 10, A fair comparison is conducted using the same backbone (R50) with M-CONet. RIOcc achieves 25% reduction in parameters while maintaining comparable inference speed to M-CONet, with **5.8 mIOU** performance improvement.

6.5. Module-level Performance Contributions

In the Table 11, we conducted additional experiments to demonstrate the impact of each of the proposed modules on the overall performance of the model. In fact, we analyzed the impact of different components on the model's performance in the original paper. Now, we have additionally aggregated all the results. Note that our baseline is not Openoccupancy, but a multi-scale model in the BEV space (excluding the modules we propose).

6.6. More Visualization

Due to space limitations, we conduct additional visual comparison experiments in Figure 9 to further validate the effectiveness of our method. The results show that RIOcc performs better at capturing distant targets and achieves a more comprehensive spatial representation.