MAESTRO: Task-Relevant Optimization via Adaptive Feature Enhancement and Suppression for Multi-task 3D Perception

Supplementary Material

In this supplementary material, we provide additional details that could not be included in the main paper due to space constraints. Section 1 presents *Experimental Details*, including training configurations, implementation specifics, and dataset descriptions. Sections 2 and 3 provide *Extensive Quantitative Results* and *Additional Ablation Studies*, offering further empirical analysis and detailed component evaluation. Section 4 presents *Extensive Qualitative Results*, illustrating additional visual comparisons to further demonstrate the effectiveness of MAESTRO across diverse scenarios.

1. Experimental Details

1.1. Datasets

We conducted our experiments on the nuScenes dataset [2] for 3D object detection and BEV map segmentation, and on the Occ3D-nuScenes dataset [24] for 3D occupancy prediction. To evaluate generalization performance, we further evaluated the proposed method on the Waymo Open Dataset (WOD) [23] for 3D object detection and the Occ3D-Waymo dataset [24] for 3D occupancy prediction.

nuScenes Datasets. The nuScenes dataset consists of 700 training, 150 validation, and 150 test scenes, each spanning 20 seconds. Annotations are provided at 2 Hz, resulting in 28,130 training samples and 6,019 validation samples. Each sample includes six camera images covering a full 360° field of view. The Occ3D-nuScenes dataset extends nuScenes dataset with voxel-level semantic occupancy ground truth in the ego coordinate frame, covering a spatial range of $[-40m, -40m, -1m]$ to $[40m, 40m, 5.4m]$ with a resolution of $[0.4m, 0.4m, 0.4m]$. The dataset includes annotations for 18 classes: 17 semantic categories and an additional "free" class representing unoccupied space. Visibility masks are also provided to facilitate evaluation within regions observable by the sensors.

Waymo Open Datasets. The WOD consists of 798 training, 202 validation, and 150 test segments, each lasting 20 seconds and recorded at 10 Hz, resulting in 159,600 training and 40,400 validation samples. Each sample includes five synchronized cameras and five LiDARs. For semantic occupancy prediction, we use the Occ3D-Waymo dataset [24], which provides dense voxel-level ground truth annotations across 16 semantic categories, including a "free" class representing unoccupied space. The spatial range of Occ3D-Waymo matches that of Occ3D-nuScenes.

1.2. Evaluation Metrics

3D Object Detection on nuScenes dataset. For 3D object detection on the nuScenes dataset, we adopt the official nuScenes evaluation protocol [2], which includes mean Average Precision (mAP) and the nuScenes Detection Score (NDS). The mAP is calculated based on center distance thresholds of 0.5m, 1.0m, 2.0m, 4.0m, averaged across 10 object categories. The NDS evaluates 3D object detection performance by combining mAP with five error metrics: Average Translation Error (ATE), Average Scale Error (ASE), Average Orientation Error (AOE), Average Velocity Error (AVE), and Average Attribute Error (AAE).

BEV Map Segmentation on nuScenes dataset. For BEV map segmentation, we adopt the evaluation setup from BEVFusion [17], computing binary segmentation metrics separately for six background categories. The performance is measured using the mean Intersection-over-Union (mIoU), selecting the maximum IoU across multiple thresholds to ensure a fair evaluation of BEV elements with varying spatial scales.

3D Object Detection on Waymo Open Dataset. 3D object detection performance on WOD is measured using the official evaluation metrics [23], reporting mean Average Precision (mAP) and its heading-aware counterpart, mAPH. Each metric is calculated over three distance intervals (0-30m, 30-50m, 50m-∞), and their average provides the final evaluation score.

3D Occupancy Prediction on nuScenes and Waymo Open Dataset. For semantic occupancy prediction on Occ3D-nuScenes and Occ3D-Waymo, we employ mean Intersection-over-Union (mIoU) as the primary evaluation metric. The mIoU quantifies the overlap between predicted and ground-truth voxel-wise labels and is computed as:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^{C} \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c + \text{FN}_c}, \tag{1}$$

where C denotes the number of semantic categories (excluding the "free" class), and TP_c, FP_c, and FN_c represent the true positives, false positives, and false negatives for class c, respectively.

1.3. Implementation Details

Backbone and Input Representation. MAESTRO employs ResNet-50 [5] with Feature Pyramid Network (FPN) [15] as the image backbone, resizing input images to a resolution of 256×704. The perception range spans

Method	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVDet [7]	29.8	37.9	0.73	0.28	0.59	0.86	0.25
DETR3D [25]	30.3	37.4	0.86	0.28	0.44	0.97	0.24
BEVDepth * [13]	33.7	41.4	0.65	0.27	0.57	0.84	0.22
BEVFormer v2 [28]	35.1	41.4	0.73	0.27	0.51	0.90	0.20
Dual-BEV † [12]	35.2	42.5	0.64	0.27	0.52	0.84	0.22
Ours-MTL	36.4	43.2	0.60	0.29	0.56	0.80	0.25

Table 1. Performance comparison with existing methods for object detection benchmark on the nuScenes validation set. "*" denotes the performance reported by Dual-BEV [12]. † indicates the method using CBGS for training.

Method	Drivable ↑	Ped. Cross. ↑	Walkway ↑	Stop Line ↑	Carpark ↑	Divider ↑	Mean ↑
CVT [32]	74.8	25.8	45.2	16.2	36.7	27.6	37.7
LSS [20]	76.0	33.3	45.5	20.1	39.8	31.4	41.0
BEVFusion [17]	78.0	42.8	49.7	31.3	43.1	37.8	47.1
DifFUSER [11]	77.8	44.2	50.6	33.3	45.6	38.5	48.3
Ours-MTL	80.3	45.9	55.4	36.1	48.3	41.8	51.3

Table 2. Performance comparison with previous methods for BEV map segmentation on the nuScenes validation set. "Drivable" and "Ped. Cross." denote drivable surface and pedestrian crosswalk. All methods were trained without employing class-balanced grouping and sampling (CBGS), and adopted ResNet-50 [5] as the image backbone for fair comparison.

Method	Modality	image backbone	mIoU (Map)
BEVFusion [17]	C	Swin-T	56.6
MapPrior [34]	C	Swin-T	56.7
Ours-MTL	C	Swin-T	**57.5**

Table 3. Performance comparison with existing camera-based map segmentation methods trained with Swin-T [16] image backbone and CBGS [33] on the nuScenes validation set.

$[-51.2m, 51.2m]$ for nuScenes and $[-75.2m, 75.2m]$ for the WOD along the X and Y axes, and $[-1m, 5.4m]$ for nuScenes and $[-5m, 7.8m]$ for WOD along the Z axis, all with a voxel resolution of 0.4m. For occupancy prediction, features within the spatial region of $\pm 40m$ along the X and Y axes are cropped to align with the annotated regions of the Occ3D-nuScenes and Occ3D-Waymo datasets [24], enabling dense voxel-based predictions.

Task-Specific Heads. We employed task-specific heads optimized for their respective tasks, each processing task-specific features to generate independent predictions. The 3D object detection head adopts the anchor-free architecture of CenterPoint [30], operating on task-specific BEV features to predict precise 3D bounding boxes. The BEV map segmentation head utilizes the segmentation head from BEVFusion [17], directly processing task-specific features for BEV map segmentation. The 3D occupancy prediction head adopts the transformer-based architecture of OccFormer [31], utilizing scene prototypes and task-specific features to produce high-resolution occupancy predictions.

Training Setup. MAESTRO was trained for 24 epochs without employing class-balanced grouping and sampling (CBGS) [33]. We employed the AdamW optimizer with an initial learning rate of 1×10^{-4} and a weight decay of 0.01. The batch size was set to 8, and random flipping was applied as a data augmentation strategy.

Hardware and Software Environment. Model training was conducted on a system running Ubuntu 18.04, equipped with two Intel Xeon CPUs and four RTX 3090 GPUs.

2. Extensive Quantitative Results

2.1. Quantitative results on the nuScenes dataset

3D Object Detection Performance. Table 1 presents a detailed performance comparison of MAESTRO against state-of-the-art 3D object detection methods on the nuScenes validation set. MAESTRO achieves 36.4% mAP and 43.2% NDS, outperforming existing methods [7, 13, 25, 28]. Notably, our method surpasses the previous state-of-the-art approach, Dual-BEV [12], by 1.2% mAP and 0.7% NDS.

BEV Map Segmentation Performance. Table 2 reports the class-wise performance for BEV map segmentation on the nuScenes validation set. MAESTRO achieves an mIoU of 51.3%, outperforming the previous state-of-the-art method, Diffuser [11], by 3.0%. Our method consistently achieves higher segmentation accuracy across all categories, particularly for pedestrian crossings and dividers, where precise localization is crucial for autonomous driving. These results demonstrate the benefit of leveraging task-aware representations for fine-grained BEV map segmentation.

In Table 3, we compare our framework with previous camera-only BEV map segmentation methods on the nuScenes validation set. Our proposed method achieves

Method	Backbone	Resolution	others	barrier	bicycle	bus	car	construction	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation	mIoU
MonoScene [3]	R101	600×928	1.8	7.2	4.3	4.9	9.4	5.7	4.0	3.0	5.9	4.5	7.2	14.9	6.3	7.9	7.4	1.0	7.7	6.1
OccFormer [31]	R101	928×1600	5.9	30.3	12.3	34.4	39.2	14.4	16.5	17.2	9.3	13.9	26.4	51.0	31.0	34.7	22.7	6.8	7.0	21.9
TPVFormer [8]	R101	600×928	7.2	38.9	13.7	40.8	45.9	17.2	20.0	18.9	14.3	26.7	34.2	55.7	35.5	37.6	30.7	19.4	16.8	27.8
CTF-Occ [24]	R101	640×960	8.1	39.3	20.6	38.3	42.2	16.9	24.5	22.7	21.1	23.0	31.1	53.3	33.8	38.0	33.2	20.8	18.0	28.5
SurroundOcc [26]	R101	800×1333	9.5	38.5	22.1	39.8	47.0	20.5	22.5	23.8	23.0	27.3	34.3	78.3	37.0	46.3	49.7	35.9	32.1	34.6
BEVDet [7]	R50	256×704	4.4	30.3	0.2	32.4	34.5	13.0	10.3	10.4	6.3	8.9	23.7	52.3	24.6	26.1	22.3	15.0	15.1	19.4
Vampire [27]	R50	256×704	7.5	32.6	16.2	36.7	41.4	16.6	20.6	16.6	15.1	21.0	28.5	68.0	33.7	41.6	40.8	24.5	20.3	28.3
FB-Occ [14]	R50	256×704	**12.2**	44.8	**25.7**	42.6	48.0	**23.2**	25.2	25.8	**26.7**	**31.3**	34.9	78.8	41.4	49.1	52.2	39.1	34.6	37.4
Ours	R50	256×704	10.7	**46.5**	24.5	**44.5**	**52.0**	19.8	**26.2**	**26.7**	26.6	30.6	**36.9**	**81.8**	**44.8**	**52.4**	**55.2**	**41.7**	**34.7**	**38.6**

Table 4. Comparison of class-wise performance with previous single-frame methods on the Occ3D-nuScenes validation set. The "R50" and "R101" respectively correspond to ResNet-50 and ResNet-101 [5].

Method	Frames	CBGS	mAP	NDS
BEVStereo	2	O	37.2	50.0
DualBEV	2	O	38.0	50.4
SOGDet	2	O	38.8	50.6
Ours-MTL	2	O	**41.8**	**52.2**

Table 5. Performance comparison with existing multi-frame methods on the nuScenes validation set.

Method	mAP	NDS	mIoU (Map)	mIoU (Occ)
MoE	34.8	42.1	46.7	36.8
Ours	**36.4**	**43.2**	**51.3**	**38.6**

Table 6. Performance comparison with Mixture of Experts (MoE) on the nuScenes validation set.

57.5% mIoU, surpassing the officially reported BEVFusion [17] (56.6% mIoU) and MapPrior [34] (56.7% mIoU) on the same backbone. These results demonstrate that MAESTRO maintains high performance on the different backbone settings.

3D Occupancy Prediction Performance. Table 4 compares class-wise 3D occupancy prediction performance on the nuScenes validation set. MAESTRO achieves 38.6% mIoU, surpassing previous methods [3, 8, 24, 26, 31], including those with higher-resolution inputs or more complex backbones. In particular, MAESTRO outperforms the previous best method, FB-Occ [14], by 1.2% mIoU, demonstrating the effectiveness of the proposed TSFG and SPA modules in enhancing task-specific feature representations through cross-task semantic aggregation.

Comparisons of performance on the multi-frame setting. We extended our method to incorporate two-frame input and compared its performance with other two-frame methods. As shown in the table 5, our approach achieves superior performance in the multi-frame setting.

Comparison with task-specific feature method. It is challenging to directly compare our method with the task-specific feature modeling method, such as InvPT++ [29], which relies on 2D image features rather than BEV or voxel representations. Similarly, MetaBEV [4] utilizes both LiDAR and camera inputs, whereas our method uses only camera images, making a direct comparison unfair. Therefore, we incorporated the Mixture-of-Experts (MoE) module proposed in MetaBEV [4] into our baseline to reduce task interference. As shown in the table 6, our method outperforms the baseline enhanced with the MoE module.

2.2. Quantitative results on the WOD.

Comparisons of Waymo Open Datasets. Table 7 compares our proposed method with existing camera-only single-task methods on the Waymo Open Dataset [23] validation split. Notably, our method outperforms previous object detection and occupancy prediction methods, achieving 6.51% mAP and 6.44% mAPH at Level 1 and 5.54% mAP and 5.49% mAPH at Level 2 for 3D object detection, and 24.19 % mIoU for semantic occupancy prediction.

3. Additional Ablation Study

We conducted additional ablation studies on 1/4 of the nuScenes training set for 24 epochs to further evaluate the effectiveness of the proposed components.

3.1. Effect of pooling methods for detection prototype generation in SPA

Table 8 shows the impact of different pooling methods for generating detection prototypes in SPA. MAESTRO achieves the best performance when employing RoIAlign [6], indicating its efficacy in extracting task-oriented features.

Method	Difficulty	mAP/mAPH ($\text{IoU}_{3D} \geq 0.7$)	mIoU (Occ)
3D Object Detection			
M3D-RPN [1]	Level 1 Level 2	0.35/0.34 0.33/0.33	-
CaDNN [22]	Level 1 Level 2	5.03/4.99 4.49/4.45	-
DEVIANT [10]	Level 1 Level 2	2.69/2.67 2.52/2.50	-
NeuROCS [19]	Level 1 Level 2	2.44/2.43 2.29/2.28	-
MonoUNI [9]	Level 1 Level 2	3.20/3.16 3.04/3.00	-
MonoDGP [21]	Level 1 Level 2	4.28/4.23 4.00/3.96	-
Semantic Occupancy Prediction			
BEVDet [7]	-	-	9.88
TPVFormer [8]	-	-	16.76
CTF-Occ [24]	-	-	18.73
Ours-MTL	Level 1 Level 2	**6.51/6.44** **5.54/5.49**	**24.19**

Table 7. Comparison of performance on Waymo Open Dataset validation split.

Method	***RoI Align***	*RoI Pooling*
mIoU (Occ)	**36.9**	36.5

Table 8. Ablation study of the pooling method for Detection prototypes.

Method	***Summation***	*Concatenation*
mIoU (Occ)	**36.9**	36.6

Table 9. Ablation study of the aggregation method for SPA.

3.2. Effect of different prototype aggregation methods in SPA

Table 9 presents the effectiveness of different aggregation methods employed in SPA. MAESTRO achieves superior performance with *Summation*, attaining an mIoU of 36.9% compared to 36.6% with *Concatenation*. These results indicate that direct summation effectively preserves semantic information beneficial for occupancy prediction.

In the Table 10, we compare our proposed strategy with a similarity-based approach that aligns task-oriented prototypes with prototype groups using cosine similarity followed by summation. The similarity-based method yields inferior performance, demonstrating the effectiveness of our explicit fusion strategy.

Method	mAP	NDS	mIoU (Map)	mIoU (Occ)
Similarity-based	32.5	**34.3**	44.1	35.7
Ours	**32.6**	**34.3**	**44.2**	**36.9**

Table 10. Ablation study for different prototype aggregation methods in SPA.

4. Extensive Qualitative Analysis

4.1. Additional Qualitative Results

To further illustrate the effectiveness of MAESTRO, we present additional qualitative comparisons against existing methods across multiple 3D perception tasks.

Figure 1 provides a qualitative comparison of BEV map segmentation results. MAESTRO accurately delineates various road elements, including drivable surfaces, dividers, pedestrian crossings, and parking areas, demonstrating superior spatial consistency and structural alignment compared to existing methods. Notably, MAESTRO produces finer-grained segmentation boundaries, reducing false positive predictions in critical areas.

Figure 2 compares occupancy prediction results with state-of-the-art approaches. Our method generates more precise occupancy predictions, particularly in occluded or distant regions, as highlighted by the red-dashed regions. MAESTRO effectively reconstructs fine-grained scene structures while suppressing erroneous predictions, demonstrating superior scene understanding and spatial reasoning capabilities.

Figures 3 and 4 evaluate MAESTRO's robustness under diverse environmental conditions, including sunny, cloudy, rainy, and nighttime scenarios. Despite significant variations in illumination, scene texture, and visibility, our method consistently produces stable and semantically coherent occupancy predictions. Specifically, MAESTRO accurately preserves the structure of drivable surfaces and surrounding obstacles, even in adverse conditions such as heavy rain or extreme darkness.

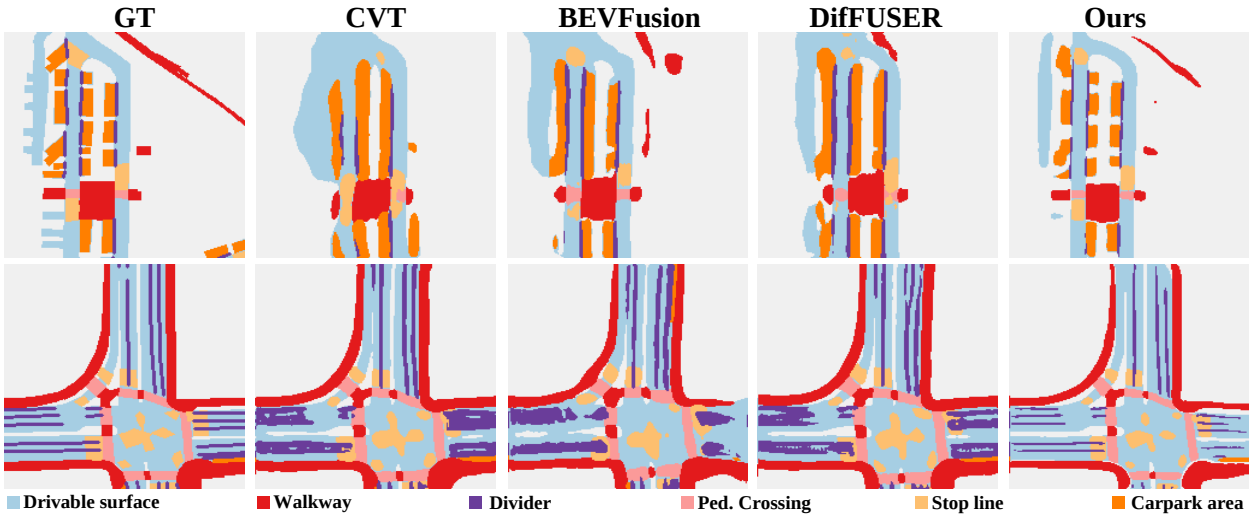

Figure 1. **Comparison with existing methods using qualitative visualization on the nuScenes validation set.**

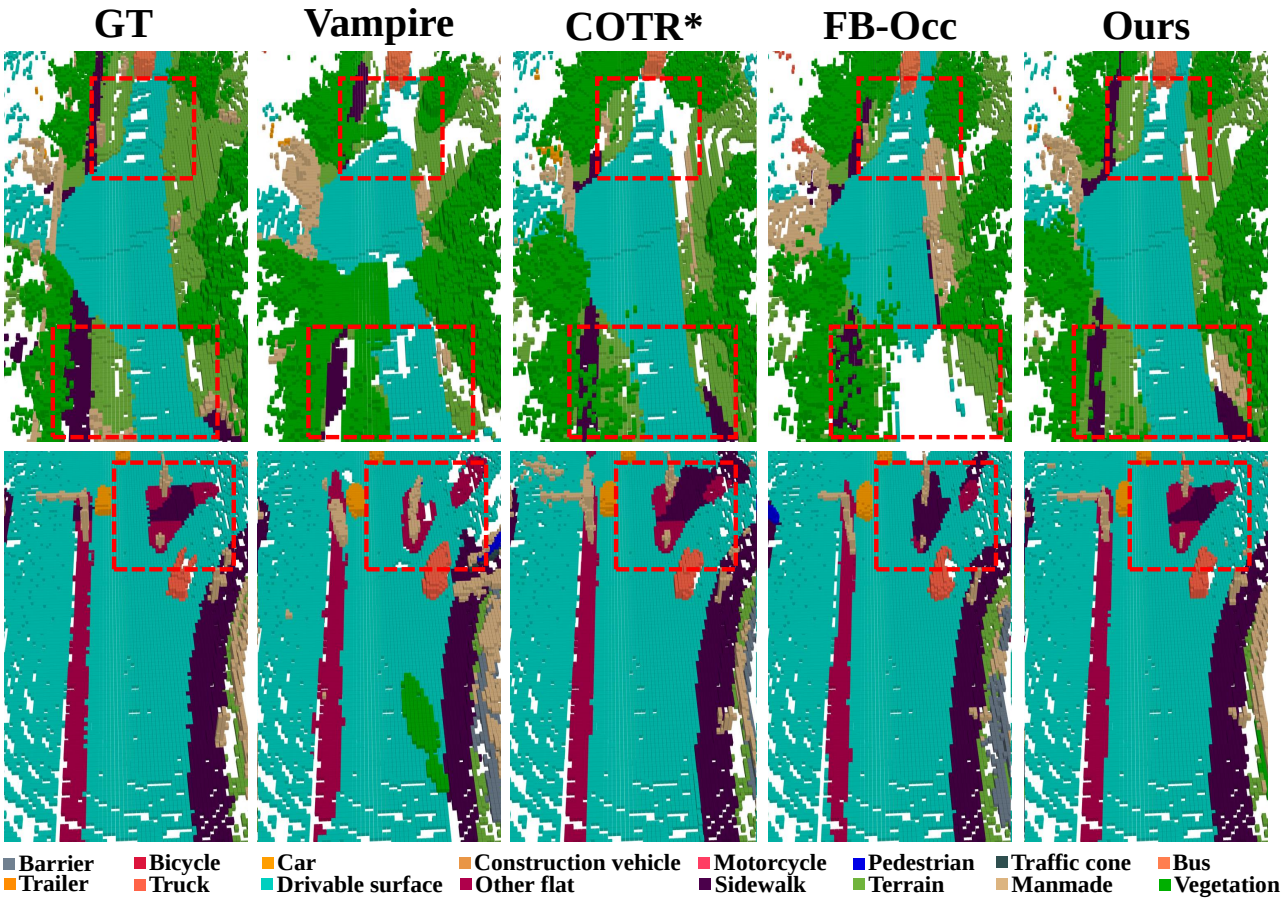

Figure 2. **Comparison with existing methods using qualitative visualization on the Occ3D-nuScenes validation set.** COTR [18] is reproduced by leveraging official code on the single-frame setting.

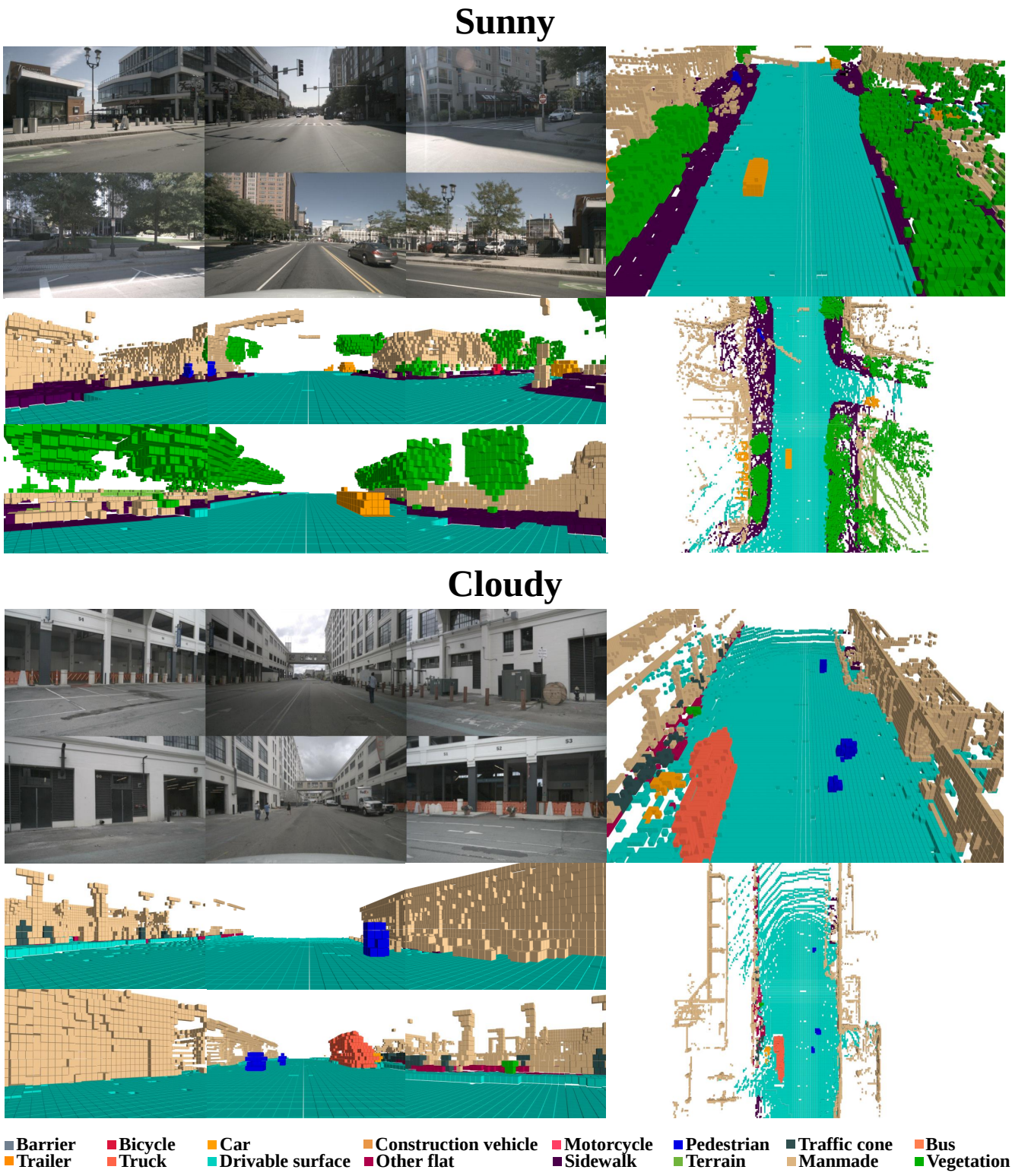

Figure 3. **Qualitative results under sunny and cloudy conditions on Occ3D-nuScenes validation set.**

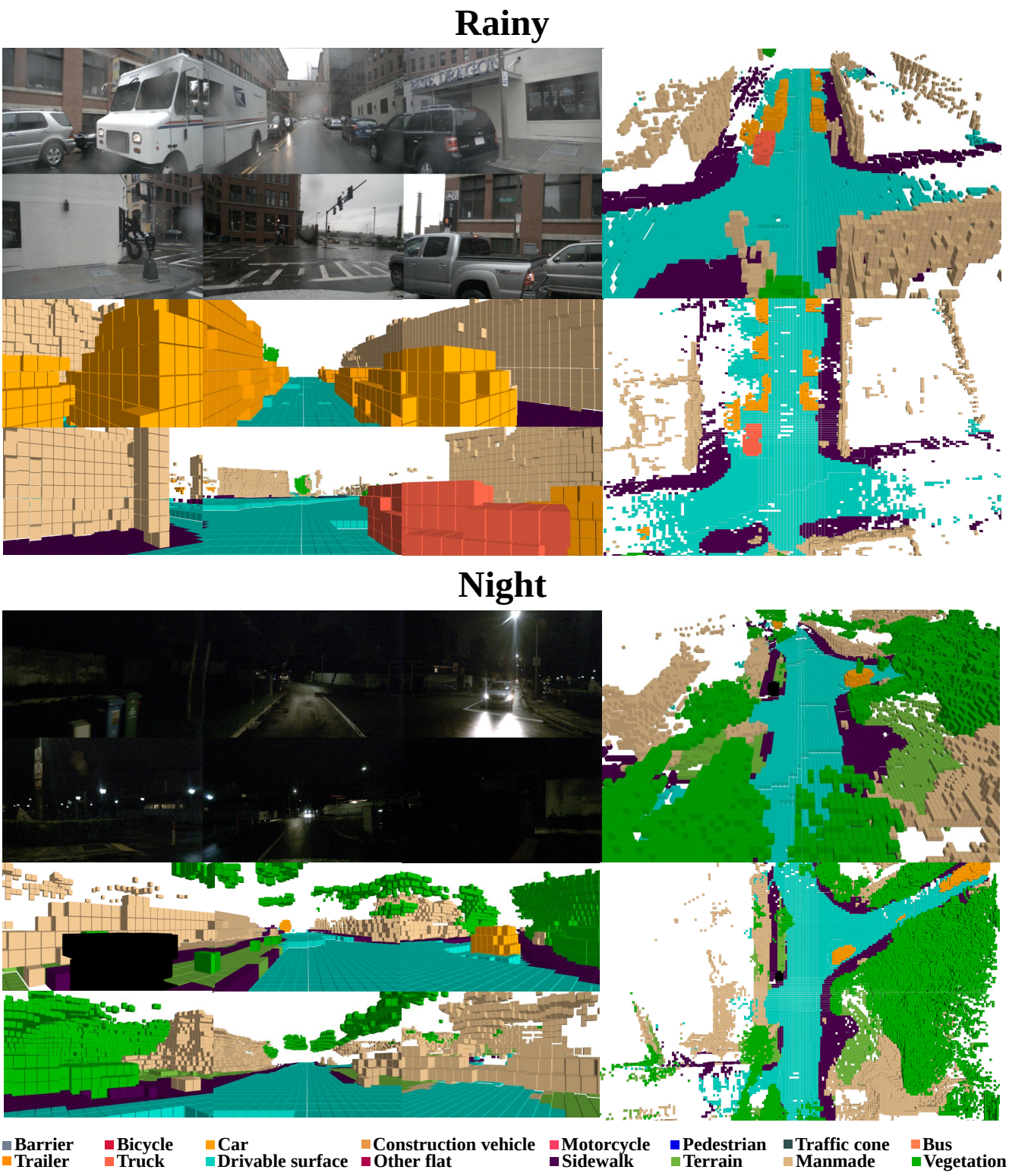

Figure 4. **Qualitative results under rainy and night conditions on Occ3D-nuScenes validation set.**

References

[1] Garrick Brazil and Xiaoming Liu. M3d-rpn: Monocular 3d region proposal network for object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 4

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1

[3] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 3

[4] Chongjian Ge, Junsong Chen, Enze Xie, Zhongdao Wang, Lanqing Hong, Huchuan Lu, Zhenguo Li, and Ping Luo. Metabev: Solving sensor failures for 3d detection and map segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8721–8731, 2023. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 2, 3

[6] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 3

[7] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 3, 4

[8] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9223–9232, 2023. 3, 4

[9] Jia Jinrang, Zhenjia Li, and Yifeng Shi. Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues. *Advances in Neural Information Processing Systems*, 36:11703–11715, 2023. 4

[10] Abhinav Kumar, Garrick Brazil, Enrique Corona, Armin Parchami, and Xiaoming Liu. Deviant: Depth equivariant network for monocular 3d object detection. In *European Conference on Computer Vision*, pages 664–683. Springer, 2022. 4

[11] Duy-Tho Le, Hengcan Shi, Jianfei Cai, and Hamid Rezatofighi. Diffusion model for robust multi-sensor fusion in 3d object detection and bev segmentation. In *European Conference on Computer Vision*, pages 232–249. Springer, 2024. 2

[12] Peidong Li, Wancheng Shen, Qihao Huang, and Dixiao Cui. Dualbev: Unifying dual view transformation with probabilistic correspondences. In *European Conference on Computer Vision*, pages 286–302. Springer, 2024. 2

[13] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1477–1485, 2023. 2

[14] 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. *arXiv preprint arXiv:2307.01492*, 2023. 3

[15] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 1

[16] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 2

[17] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 1, 2, 3

[18] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19936–19945, 2024. 5

[19] Zhixiang Min, Bingbing Zhuang, Samuel Schulter, Buyu Liu, Enrique Dunn, and Manmohan Chandraker. Neurocs: Neural nocs supervision for monocular 3d object localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21404–21414, 2023. 4

[20] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[21] Fanqi Pu, Yifan Wang, Jiru Deng, and Wenming Yang. Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6520–6530, 2025. 4

[22] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8555–8564, 2021. 4

[23] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 1, 3

[24] 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*, 36:64318–64330, 2023. 1, 2, 3, 4

[25] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *Conference on Robot Learning*, pages 180–191. PMLR, 2022. 2

[26] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 3

[27] Junkai Xu, Liang Peng, Haoran Cheng, Linxuan Xia, Qi Zhou, Dan Deng, Wei Qian, Wenxiao Wang, and Deng Cai. Regulating intermediate 3d features for vision-centric autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6306–6314, 2024. 3

[28] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 2

[29] Hanrong Ye and Dan Xu. Invpt++: Inverted pyramid multi-task transformer for visual scene understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 3

[30] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 2

[31] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 2, 3

[32] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022. 2

[33] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. *arXiv preprint arXiv:1908.09492*, 2019. 2

[34] Xiyue Zhu, Vlas Zyrianov, Zhijian Liu, and Shenlong Wang. Mapprior: bird's-eye view map layout estimation with generative models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8228–8239, 2023. 2, 3