Cam4DOcc: Benchmark for Camera-Only 4D Occupancy Forecasting in Autonomous Driving Applications: Supplementary Material

Junyi Ma[1,3*] Xieyuanli Chen[2*] Jiawei Huang[3] Jingyi Xu[4] Zhen Luo[5]
Jintao Xu[3] Weihao Gu[3] Rui Ai[3] Hesheng Wang[1†]
[1]IRMV Lab, Department of Automation, Shanghai Jiao Tong University
[2]College of Intelligence Science and Technology, National University of Defense Technology
[3]HAOMO.AI [4]NAV, Shanghai Jiao Tong University [5]IVRC, Beijing Institute of Technology

A. Dataset Setup Details

We provide more details about our new dataset format for our Cam4DOcc benchmark by presenting statistics on the instance duration $[t_{in}, t_{out}]$ after splitting the original nuScenes and Lyft-Level5 datasets to separate sequences mentioned in Sec. 3.2. As shown in Fig. 1, most general movable objects (GMO) appear in at least two historical observations and all future observations ($[-2, 4]$ and $[-1, 4]$) in our benchmark. The long instance duration leads to an effective training strategy for the occupancy forecasting model. Besides, over 30% instances in the two datasets first appear in the current frame ($t = 0$), which makes the model learn to forecast the object motion only according to their current location and surrounding conditions.

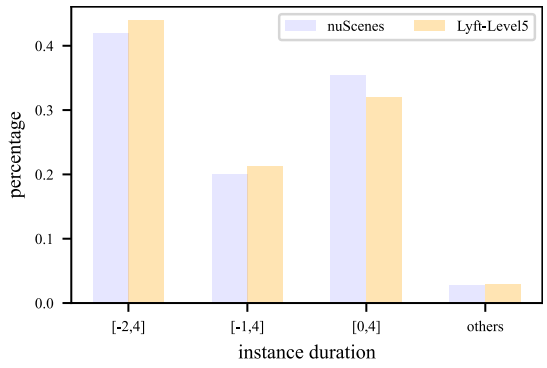

Figure 1. Instance duration on nuScenes and Lyft-Level5.

In addition, we further provide a detailed illustration of inflated GMO and fine-grained GMO defined in our Cam4DOcc introduced in Sec. 3.3, as shown in Fig. 2. Compared to the fine-grained labels, the inflated bounding-box-wise annotation overall provides more comprehensive training signals for the occupancy forecasting model. In addition, the motion of GMO with a structured format from the instance bounding box is easier to capture (validated in Sec. 5.2). From the second row of Fig. 2 we can also see that sometimes fine-grained voxel annotation cannot accurately represent the sophisticated shape of GMO while the bounding-box-wise annotation can totally encompass the holistic GMO instance grids. The third row of Fig. 2 also presents that fine-grained annotation may miss some occluded objects compared to the original instance bounding box labels, affecting the rationality of the training and evaluation on these scenarios. Therefore, Cam4DOcc suggests using inflated GMO annotations to train current-stage camera-based models for more reliable 4D occupancy forecasting and safer navigation in autonomous driving. We also hope that the preset tasks with fine-grained GMO labels in Cam4DOcc can be the foundation for developing more advanced camera-only 4D occupancy forecasting approaches in future research.

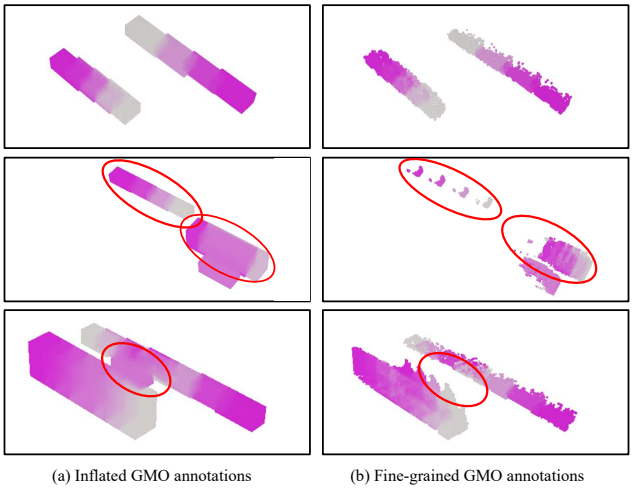

Figure 2. Comparison of GMO categories defined in Cam4DOcc.

*Equal contribution
†Corresponding author: wanghesheng@sjtu.edu.cn

B. OCFNet Model Details

Our proposed OCFNet receives 6 images with the size of 900×1600 captured by surround-view cameras mounted on the vehicle. We use ResNet50 [4] pretrained on ImageNet [3] with FPN [8] as the image encoder in OCFNet. LSS-based 2D-3D Lifting module [9] transforms and fuses image features from multiple camera images to unified voxel features. We use the vanilla 3D-ResNet18 as the Voxel Encoder and use 3D-FPN as the Voxel Decoder in both the occupancy forecasting head and flow prediction head of the Future State Prediction Module. The prediction module containing stacked residual convolutional blocks orderly encodes historical 3D features, expands channel dimensions according to the future time horizon N_f, and produces future 3D features, as shown in Fig. 3. Referring to the setups of PowerBEV [7], the numbers of the three types of residual convolutional blocks in the prediction module are set to 2, 1, and 2, with the kernel size of (3, 3, 1).

To extend our occupancy forecasting model to 3D instance prediction, our OCFNet predicts occupancy and 3D flow over $t \in [0, N_f]$, corresponding to 5 continuous estimations specifically in our work. Local maxima are first extracted from the estimated occupancy probabilities at $t = 0$ following [7], determining the instances' centers. Then, the instances in the following future frames are associated consecutively with the predicted flow.

To train our OCFNet using the loss defined in Eq. (4), we set $\lambda_1 = \lambda_3 = 0.5$ and $\lambda_2 = 0.05$ to balance the optimization for occupancy forecasting, depth reconstruction, and 3D backward centripetal flow prediction. The total parameter number of our OCFNet is 370 M, the GFLOPs are 6434, and the training-time GPU memory is 57 GB. We believe that our model can serve as a foundational codebase to facilitate future 4D occupancy forecasting works.

C. Study on Future Time Horizons

We further conduct a study on forecasting performance drops with different future time horizons. Since the occupancy grids of static objects do not change in the future time steps unless ground-truth annotations jitter, here we solely focus on the ability to forecast the future occupancy state of movable objects. In this experiment, we post the performance of OpenOccupancy-C, PowerBEV-3D, and our OCFNet for the first-level task and the second-level task since the baseline SPC fails to forecast the inflated GMO mentioned in Sec. 5.2. As shown in Tab. 1, our OCFNet† remains the best performance for different time horizons in both tasks. In addition, all the baseline approaches show better performance on Lyft-Level5 than nuScenes as the time period for evaluating on Lyft-Level5 is relatively shorter. The closer the timestamp is to the current moment, the easier it is for all the baselines to forecast the occupancy status.

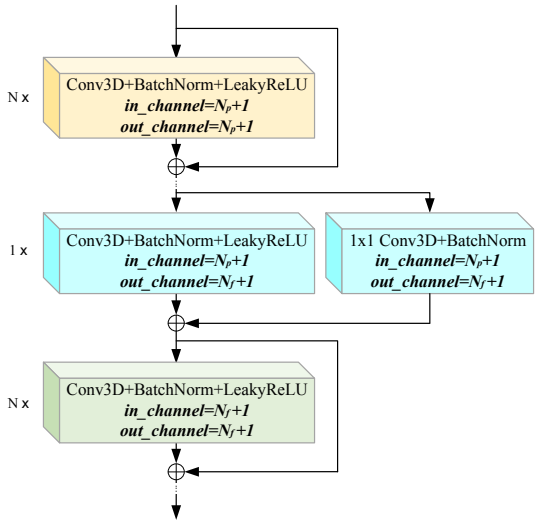

Figure 3. The prediction module in our OCFNet.

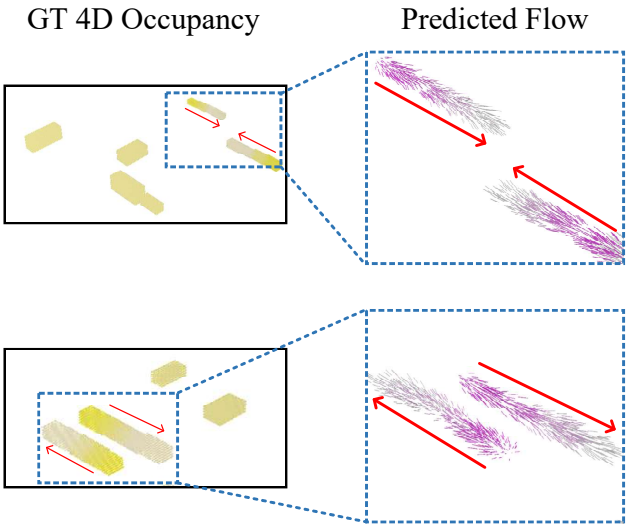

Figure 4. Visualization of predicted 3D backward flow ($t \in [1, N_f]$). The output flow vectors and ground-truth occupancy from timestamps 1 to N_f are assigned colors from dark to light respectively. The motion trend of each selected moving object is represented by red arrows.

D. 3D Flow Prediction

Our proposed novel end-to-end occupancy forecasting network OCFNet is trained to reasonably estimate future occupancy state and 3D motion flow simultaneously. We notice that the multi-task learning scheme can help to improve forecasting performance, as shown in Sec. 5.3. Here, we illustrate the predicted 3D backward centripetal flow in Fig. 4. As can be seen, the predicted flow vectors of the

Table 1. Comparison of performance on forecasting GMO in different future time horizons

approach	nuScenes				Lyft-Level5				nuScenes-Occupancy			
	0.5 s	1.0 s	1.5 s	2.0 s	0.2 s	0.4 s	0.6 s	0.8 s	0.5 s	1.0 s	1.5 s	2.0 s
OpenOccupancy-C [10]	12.07	11.80	11.63	11.45	13.87	13.77	13.65	13.53	9.17	8.64	8.29	8.02
PowerBEV-3D [7]	22.48	22.07	21.65	21.25	25.70	25.25	24.82	24.47	5.74	5.56	5.41	5.25
OCFNet (ours)	25.95	24.92	24.33	23.89	31.51	30.87	30.17	29.56	9.17	8.72	8.53	8.35
OCFNet† (ours)	**29.36**	**28.30**	**27.44**	**26.82**	**35.58**	**34.96**	**34.28**	**33.56**	**10.64**	**10.20**	**9.89**	**9.68**

Table 2. Comparison of performance on 3D instance prediction

approaches	nuScenes	Lyft-Level5
PowerBEV-3D [7]	20.02	27.39
OCFNet	14.26	24.82
OCFNet†	18.57	28.23
OCFNet*	**21.36**	**28.81**

moving object approximately point from the voxel grids of the new coming frame to the ones of the past frame belonging to the same instance. Therefore, the predicted flow can further guide occupancy forecasting by explicitly capturing the motion of GMO in each time interval. Thanks to the flow vectors predicted by Cam4DOcc, we can further associate consistent instances between adjacent future frames, leading to 3D instance prediction beyond occupancy state forecasting.

E. 3D Instance Prediction

Most existing instance prediction methods [1, 2, 5, 7] can only forecast the future position of objects of interest on BEV representation, while our work extends this task to more complex 3D space. To achieve instance prediction, models are required to output reasonable 3D flow after the training process. The ground-truth 3D backward centripetal flow in our Cam4DOcc is directly calculated from the annotations of the original datasets, given the positions of instance bounding boxes as well as the corresponding IDs. During the forecasting process, we first extract the centers of instances at $t = 0$ and then associate pixel-wise instance ID over time $t \in [1, N_f]$ using the predicted 3D backward centripetal flow. Following the previous work [7], we calculate the centers C of instances by non-maximum suppression (NMS) at $t = 0$. The predicted backward flow starts from the occupancy grid o_1 at $t = 1$ to o_0 at $t = 0$, and the instance ID of the center in C closest to o_0 is propagated to o_1. Since there is no real observation from $t = 1$ on, we then directly use the predicted flow from o_2 at $t = 2$ to o_1 at $t = 1$ to propagate the instance ID for the forecasted frames. The same association is then implemented over the following time steps.

To report the instance prediction quality, we extend the metric video panoptic quality (VPQ) [6] from the previous 2D instance prediction [5, 7] to our 3D instance prediction, which is calculated by

$$\text{VPQ}_f(\hat{\mathbf{O}}_f^{inst}, \mathbf{O}_f^{inst}) = \frac{1}{N_f}\sum_{t=0}^{N_f}\frac{\sum_{(p_t,q_t)\in TP_t}\text{IoU}(p_t,q_t)}{|TP_t| + \frac{1}{2}|FP_t| + \frac{1}{2}|FN_t|}, \tag{1}$$

where TP_t, FP_t, and FN_t represent true positives, false positives, and false negatives at timestamp t. Note that in our work the predicted instance is regarded as one true positive once its IoU is greater than 0.2 (adaptively chosen according to the level of IoU) and the corresponding instance ID is correctly tracked. The experimental results are shown in Tab. 2. Note that the instance prediction results of PowerBEV-3D are also from the duplication of forecasted 2D flow along the height dimension (same as its 3D extension of forecasted occupancy introduced in Sec. 3.4). As can be seen, our proposed OCFNet† shows better 3D instance prediction ability than PowerBEV-3D on Lyft-Level5 while PowerBEV-3D outperforms our approach on nuScenes. In addition, OCFNet† improves the prediction of OCFNet by 30.2% and 13.7% on nuScenes and Lyft-Level5 respectively. The 2D-3D instance-based prediction baseline presents good instance prediction ability on nuScenes because 2D backward centripetal flow is easier to forecast than the 3D counterpart. On the contrary, our proposed method produces better forecasting results on Lyft-Level5, dominated by far better GMO occupancy forecasting quality of OCFNet† than that of PowerBEV-3D on this dataset. Therefore, in the 3D instance prediction task, we further propose a new baseline namely OCFNet*, which combines the advantages of our original OCFNet† and PowerBEV-3D. The principle is that the 3D flow of the intersection GMO occupancy forecasted by the two methods follows PowerBEV-3D's results, while the other GMO occupancy grids forecasted by OCFNet† have the motion flow generated by OCFNet† itself. Based on this setup, whether an occupancy grid is occupied totally depends on OCFNet†, and its flow depends on the choice between OCFNet† and PowerBEV-3D. From Tab. 2, we can see that OCFNet* has the best 3D instance prediction performance, which en-

hances PowerBEV-3D by 6.7% on nuScenes and improves OCFNet† by 2.1% on Lyft-Level5.

F. Study on Movable Objects of Multiple Categories

We additionally report the results of each class of movable objects in Tab. 3, providing a more detailed and nuanced analysis of OCFNet forecasting performance across various categories. To extend the use of our proposed benchmark, we further provide forecasting performance evaluation on the OCFNet with LiDAR inputs (OCFNet-L), compared with our original camera-only OCFNet (OCFNet-C). OCFNet-L replaces the image encoder and the lifting module in the vanilla OCFNet architecture with voxel feature encoding as well as sparse convolution middle layers [11]. Here we only use point clouds from sample observations as OCFNet-L inputs rather than aggregated multiple sweeps on the nuScenes dataset. As can be seen, OCFNet has a better ability to forecast moving objects with larger sizes. OCFNet-C outperforms OCFNet-L in the bicycle, motorcycle, and pedestrian classes, which reveals that OCFNet-L tends to be affected by the sparsity of LiDAR observations since relatively small objects are hit by fewer rays, thus harder to capture future motion. We will report more results of multi-modal baselines in our open-source repository.

G. Visualization of future GMO occupancy forecasted by OCFNet on Lyft-Level5

In this section, we present our proposed OCFNet forecasting inflated general movable objects of the Lyft-Level5 dataset. Fig. 5 and Fig. 6 show the results in small-scale and large-scale scenes respectively. The prediction results and ground truth from timestamps 1 to N_f are assigned colors from dark to light. As to the small-scale scenes, the valid GMO over the future time horizon occupy relatively fewer volumes and both OCFNet and OCFNet† can capture their motion accurately. When it comes to the large-scale scenes with more complicated traffic conditions, OCFNet† significantly outperforms OCFNet which only uses $\frac{1}{6}$ sequences for training. Therefore, when the driving scenario of the ego vehicle has few movable obstacles, such as in rural areas, OCFNet trained with limited data is enough to forecast the future occupancy of surrounding traffic participators. This can significantly improve the deployment efficiency of forecasting modules in autonomous driving systems by decreasing memory consumption and training period.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. In *ECCV*, pages 444–460, 2022. 3

[2] Sergio Casas, Abbas Sadat, and Raquel Urtasun. Mp3: A unified model to map, perceive, predict and plan. In *CVPR*, pages 14403–14412, 2021. 3

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 2

[5] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 3

[6] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, pages 9859–9868, 2020. 3

[7] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: A powerful yet lightweight framework for instance prediction in bird's-eye view. In *IJCAI*, pages 1080–1088, 2023. 2, 3

[8] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 2

[9] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210, 2020. 2

[10] 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 *ICCV*, pages 17850–17859, 2023. 3

[11] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10), 2018. 4

Table 3. IoU_f results of multiple movable categories in the first-level task on nuScenes

approach	bicycle	bus	car	construction	motorcycle	trailer	truck	pedestrian	mean
OCFNet-C	10.75	28.87	24.62	12.70	9.62	22.12	23.52	10.26	17.81
OCFNet-L	5.80	25.31	25.83	13.02	7.55	24.29	23.94	9.61	16.92

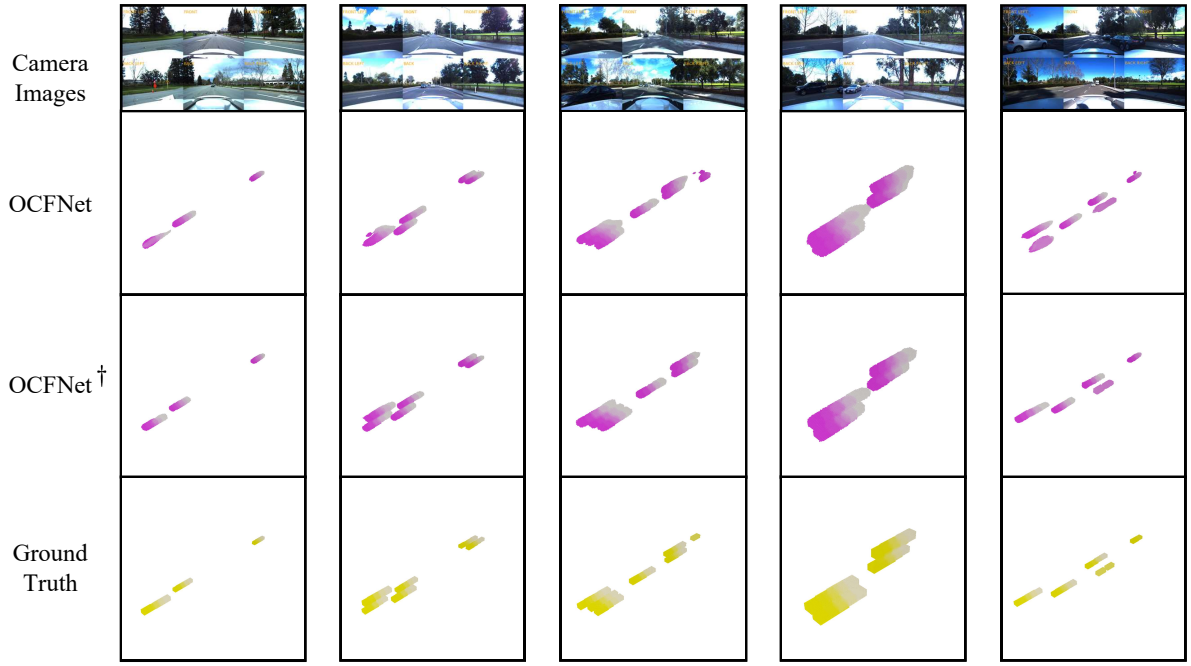

Figure 5. Visualization of forecasting inflated GMO by our proposed OCFNet in small-scale scenes of Lyft-Level5.

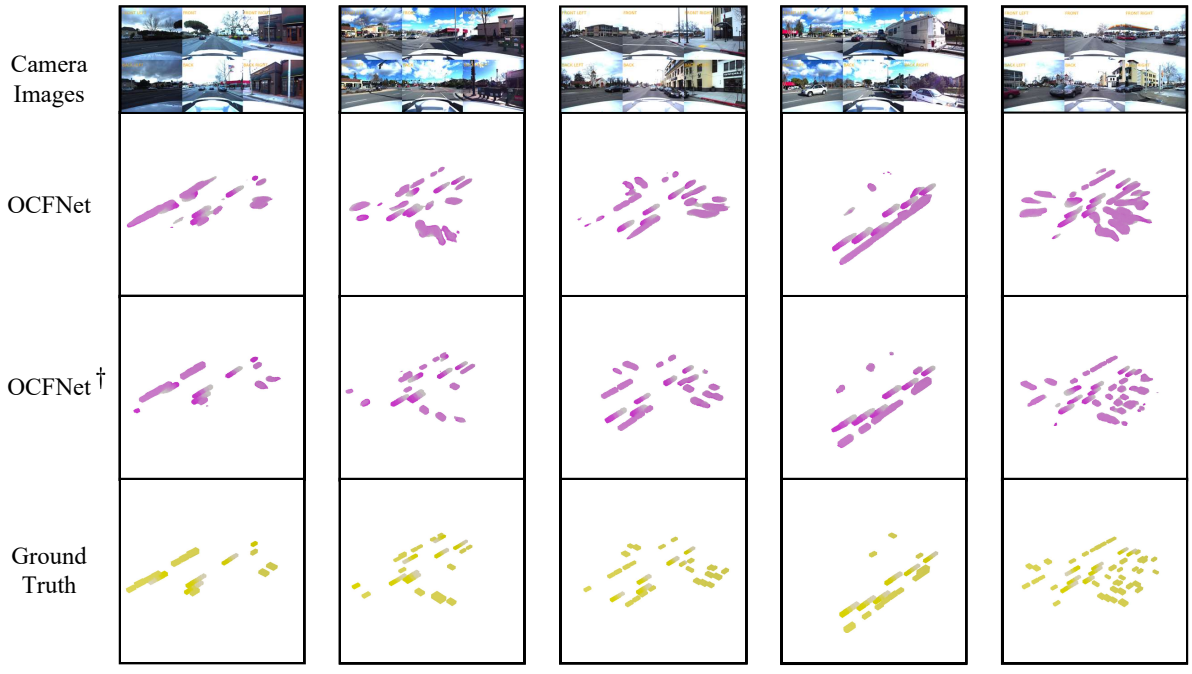

Figure 6. Visualization of forecasting inflated GMO by our proposed OCFNet in large-scale scenes of Lyft-Level5.