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

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

Abstract

Understanding how the surrounding environment changes is crucial for performing downstream tasks safely and reliably in autonomous driving applications. Recent occupancy estimation techniques using only camera images as input can provide dense occupancy representations of large-scale scenes based on the current observation. However, they are mostly limited to representing the current 3D space and do not consider the future state of surrounding objects along the time axis. To extend camera-only occupancy estimation into spatiotemporal prediction, we propose Cam4DOcc, a new benchmark for camera-only 4D occupancy forecasting, evaluating the surrounding scene changes in a near future. We build our benchmark based on multiple publicly available datasets, including nuScenes, nuScenes-Occupancy, and Lyft-Level5, which provides sequential occupancy states of general movable and static objects, as well as their 3D backward centripetal flow. To establish this benchmark for future research with comprehensive comparisons, we introduce four baseline types from diverse camera-based perception and prediction implementations, including a static-world occupancy model, voxelization of point cloud prediction, 2D-3D instance-based prediction, and our proposed novel end-to-end 4D occupancy forecasting network. Furthermore, the standardized evaluation protocol for preset multiple tasks is also provided to compare the performance of all the proposed baselines on present and future occupancy estimation with respect to objects of interest in autonomous driving scenarios. The dataset and our implementation of all four baselines in the proposed Cam4DOcc benchmark are released as open source at https://github.com/haomo-ai/Cam4DOcc.

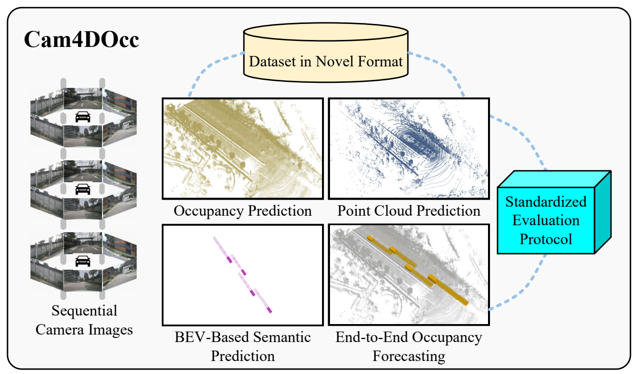


Figure 1. Cam4DOcc focuses on providing a novel dataset format, creating baselines modified from off-the-shelf camera-based perception and prediction approaches, and proposing a standardized evaluation protocol for the 4D occupancy forecasting task.

1. Introduction

Accurately perceiving the status of objects in surrounding environments using cameras is important for autonomous vehicles or robots to make reasonable downstream planning and action decisions. Traditional camera-based perception methods for object detection [30, 37, 44, 51], semantic segmentation [5, 24, 31, 49], and panoptic segmentation [9, 16, 27, 43] focus on predefined specific object categories, making them less effective at recognizing uncommon objects. To tackle this limitation, a shift towards camera-based occupancy estimation [17, 29, 41, 45, 47] has emerged by estimating the spatial occupancy states over classifying specific objects. It reduces the complexity of multi-class classification tasks and emphasizes general occupied state estimation, enhancing the reliability and adaptability of autonomous mobile systems.

Despite the increasing attention to camera-based occupancy estimation, existing methods only estimate the current and past occupancy status. However, advanced collision avoidance and trajectory optimization methods employed by autonomous vehicles [10, 11, 39] require the

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

ability to forecast future environmental conditions to ensure the safety and reliability of driving. Some semantic/instance prediction algorithms [14, 15, 25, 34, 48] have been proposed to forecast the motion of objects of interest, but they are mostly limited to 2D bird's eye view (BEV) format and can only recognize specific objects, mainly in the vehicle category. As to existing occupancy forecasting algorithms [19, 20, 42] without considering semantics, they need LiDAR point clouds as necessary prior information to perceive the surrounding spatial structure, while LiDAR-based solutions are more resource-intensive and expensive than the camera counterparts. It is natural to anticipate the next significant challenge in autonomous driving will be camera-only 4D occupancy forecasting. This task aims to not only extend temporal occupancy prediction with camera images as input but also broaden semantic/instance prediction beyond BEV format and predefined categories. To this end, we propose *Cam4DOcc* as shown in Fig. 1, the first camera-only 4D occupancy forecasting benchmark comprising the new format of dataset, various types of baselines, and standardized evaluation protocol, to facilitate the advancements in this emerging domain. In this benchmark, we construct a dataset by extracting continuous occupancy changes along the time axis from the original nuScenes [2], nuScenes-Occupancy [45], and Lyft-Level5 [18]. This dataset includes sequential semantic and instance annotations and 3D backward centripetal flow indicating the motion of occupancy grids. Furthermore, to achieve camera-based 4D occupancy forecasting, we introduce four baseline methods, including a static-world occupancy model, voxelization of point cloud prediction, 2D-3D instance-based prediction, and an end-to-end 4D occupancy forecasting network. Finally, we evaluate the performance of these baseline methods for both present and future occupancy estimation using a proposed standardized protocol.

The main contributions of this paper are fourfold: (1) We propose Cam4DOcc, the first benchmark to facilitate future work on camera-based 4D occupancy forecasting. (2) We propose a new dataset format for the forecasting task in autonomous driving scenarios by leveraging existing datasets in the field. (3) We provide four novel baselines for camera-based 4D occupancy forecasting. Three of them are the extension of off-the-shelf approaches. Additionally, we introduce a novel end-to-end 4D occupancy forecasting network that demonstrates strong performance and can serve as a valuable reference for future research. (4) We introduce a novel standardized evaluation protocol and conduct comprehensive experiments with detailed analysis based on this protocol with our Cam4DOcc.

2. Related Work

Occupancy prediction. Occupancy prediction/estimation is a trendy technique to comprehensively estimate the occupancy state of the surrounding environments. It represents the space with geometric details, significantly enhancing the expressiveness of complex scenes. MonoScene proposed by Cao *et al*. [3] first addresses 3D scene semantic completion from camera images, but only considers the front-view voxels. In contrast, Huang *et al*. [17] replace the Features Line of Sight Projection of MonoScene with TPVFormer to enhance the performance of surround-view occupancy prediction based on cross attention mechanism. UniOcc by Pan *et al*. [36] combines voxel-based neural radiance field (NeRF) with occupancy prediction to implement geometric and semantic rendering. Wang *et al*. [45] propose a large-scale benchmark named OpenOccupancy which establishes the nuScenes-Occupancy dataset with high-resolution occupancy ground-truth, and further provides several baselines using different modalities. Tong *et al*. [41] also propose an occupancy prediction benchmark OpenOcc and exploit the occupancy estimated by their OccNet on various tasks, including semantic scene completion, 3D object detection, BEV segmentation, and motion planning. More recently, Occ3D [40] utilizes occlusion reasoning and image-guided refinement to further improve the annotation quality. Similar to OpenOcc, SurroundOcc by Wei *et al*. [47] also produces dense occupancy labels and uses spatial attention to reproject 2D camera features back to the 3D volumes.

Occupancy forecasting. Occupancy forecasting is utilized to foresee how the surrounding occupancy changes in the near future beyond the present moment. Existing occupancy forecasting approaches [19, 20, 42] mainly use LiDAR point clouds as input to capture the change of surrounding structures. For example, Khurana *et al*. [20] propose a differentiable raycasting method to forecast 2D occupancy states by pose-aligned LiDAR sweeps. More recently, they propose rendering future pseudo LiDAR points with estimated occupancy [20]. Other Point cloud prediction methods [12, 32, 33, 35] directly forecast the future laser points, which can be voxelized to future occupancy estimation. However, they still need sequential LiDAR point clouds and lose semantic consistency during prediction. In contrast to the above-mentioned LiDAR-based occupancy forecasting, directly predicting future 3D occupancy with multiple semantic categories using only camera images in large-scale scenes remains challenging. Therefore, some camera-only semantic/instance prediction methods turn to forecast the motion of objects of interest, e.g., general vehicle classes on 2D BEV occupancy representation [1, 15, 25, 50]. For example, FIERY by Hu *et al*. [15] directly extracts BEV features from multi-view 2D camera images and then combines a temporal convolution model and a recurrent network to estimate future instance distributions. After that, StretchBEV [1] and BEVerse [50] are proposed for further enhancement on longer time horizons. Towards the over-supervision with redundant outputs, Power-

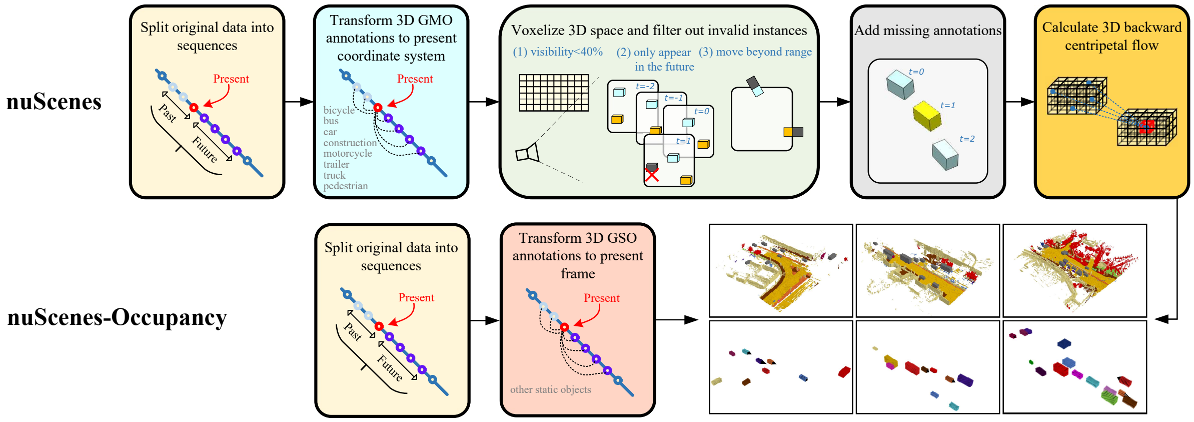


Figure 2. Overall pipeline of constructing dataset in our Cam4DOcc based on the original nuScenes and nuScenes-Occupancy. The dataset is reorganized into a novel format that considers both general movable and static categories for the unified 4D occupancy forecasting task.

BEV [25] is recently proposed to improve the forecasting performance on accuracy and efficiency.

The abovementioned methods cannot directly achieve the camera-only 4D occupancy forecasting task. In this work, we propose a novel benchmark on this topic where several baselines are created by converting the implementation of the existing state-of-the-art occupancy prediction, point cloud prediction, and BEV-based semantic/instance prediction algorithms. In addition, we develop a novel camera-based 4D occupancy forecasting network that can simultaneously forecast the future occupancy state of the general movable and static objects end-to-end. Standardized dataset format and evaluation protocol are also proposed to train and test all the baselines, which can further support future work in this literature.

3. Cam4DOcc Benchmark

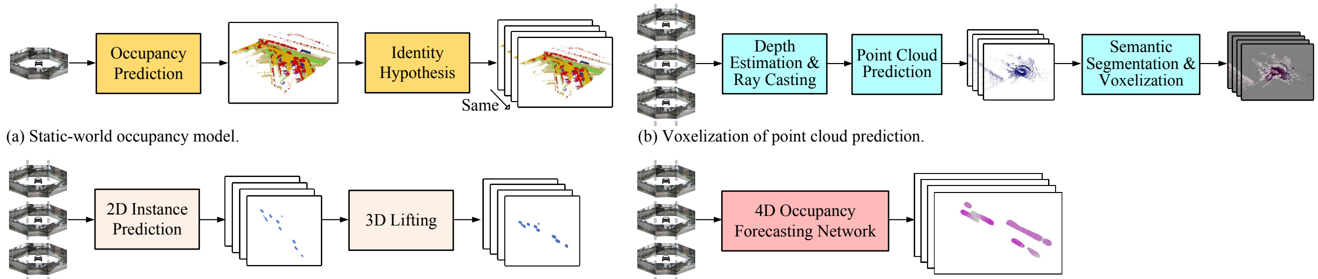
3.1. Task Definition

Given N_p past and the current consecutive camera images $\mathcal{I} = \{I_t\}_{t=-N_p}^{0}$ as input, 4D occupancy forecasting aims to output the current occupancy $\mathbf{O}_c \in \mathbb{R}^{1\times H\times W\times L}$ and the future occupancy $\mathbf{O}_f \in \mathbb{R}^{N_f\times H\times W\times L}$ in a short time interval N_f, where H, W, L represent the height, width, and length of the specific range defined in the present coordinate system ($t = 0$). Each voxel of $\mathbf{O}_f$ has N_f sequential states $\mathcal{S} = \{S_t\}_{t=1}^{N_f}$ to represent whether it is free or occupied in each future timestamp.

Cam4DOcc considers two categories regarding their motion characteristics, *general movable objects (GMO)*, and *general static objects (GSO)*, as the semantic labels of occupied voxel grids. GMO usually have higher dynamic motion characteristics compared to GSO, thus requiring more attention during traffic activities for safety reasons. Accurately estimating the behavior of GMO and predicting their potential motion changes significantly affect the decision making and motion planning of the ego vehicle. Compared to the previous semantic scene completion task [6, 17, 23, 38, 41, 45] considering multiple semantic categories, we focus more on investigating the ongoing change of voxel states for movable objects because we believe that motion characteristics of traffic participants deserve increased attention in the context of autonomous driving applications. Compared to the existing semantic/instance prediction task [1, 4, 15, 25, 50], we not only emphasize the prediction of neighboring foreground objects but also focus on the occupancy estimation for the background of surrounding environments towards the requirement of more reliable navigation for autonomous vehicles.

3.2. Dataset in New Format

Our Cam4DOcc benchmark introduces a new dataset format based on original nuScenes [2], nuScenes-Occupancy [45], and Lyft-Level5. As Fig. 2 illustrates, we first split the original nuScenes dataset into sequences with the time length of $N = N_p + N_f + 1$. Then sequential semantic and instance annotations of movable objects are extracted for each sequence and collected into the GMO class, including *bicycle*, *bus*, *car*, *construction*, *motorcycle*, *trailer*, *truck*, and *pedestrian*. They are all transformed to the present coordinate system ($t = 0$). After that, we voxelize the present 3D space and attach semantic/instance labels to the grids of movable objects using bounding boxes annotation. Notably, the invalid instance is discarded in this process once: (1) its visibility is under 40% over the 6 camera images if it is a newly appeared object in N_p historical frames, (2) it first appears in N_f incoming frames, or (3) it moves beyond the range (H, W, L) predefined at $t = 0$. The visibility is quantified by the visible proportion of all pixels of the instance showing in camera images [2]. The sequential annotations are exploited to fill in missing intermediate instances based on constant velocity assumption [8, 25]. The same operations are also applied to the Lyft-Level5 dataset. The distribution of instance duration $[t_{in}, t_{out}]$ after the process-



(a) Static-world occupancy model. (b) Voxelization of point cloud prediction. (c) 2D-3D instance-based prediction. (d) End-to-end occupancy forecasting network.

Figure 3. Four types of baselines are proposed in the Cam4DOcc benchmark from the extension of occupancy prediction, point cloud prediction, and 2D instance prediction, as well as our end-to-end 4D occupancy forecasting network.

ing mentioned above is presented in supplementary Sec. A. Lastly, we generate 3D backward centripetal flow using the instance association in the annotations. Li *et al*. [25] introduced 2D backward centripetal flow to improve the efficiency of 2D instance prediction. Inspired by that, we calculate 3D backward centripetal flow pointing from the voxel at time t to its corresponding 3D instance center at $t-1$. We exploit this 3D flow to improve the accuracy of camera-based 4D occupancy forecasting. More details about the generation and utilization of 3D flow are further clarified in Sec. E in supplementary materials.

We aim not only to forecast future positions of GMO but also to estimate the occupancy state of GSO and free space necessary for safe navigation. Thus, we further concatenate the sequential instance annotations from the original nuScenes with the sequential occupancy annotations transformed to the present frame from nuScenes-Occupancy. This combination balances safety and precision for downstream navigation in autonomous driving applications. GMO labels are borrowed from the bounding box annotations of the original nuScenes, which can be regarded as performing a dilation operation on the movable obstacles. GSO and free labels are provided by nuScenes-Occupancy to concentrate on more fine-grained geometric structures of surrounding large-scale environments.

3.3. Evaluation Protocol

To fully access the camera-only 4D occupancy forecasting performance, we establish various evaluation tasks and metrics with varying levels of complexity in our Cam4DOcc.

Multiple tasks. We introduce four-level occupancy forecasting tasks in the standardized evaluation protocol: (1) *Forecasting inflated GMO*: the categories of all the occupancy grids are divided into GMO and others, where the voxel grids within the instance bounding boxes from nuScenes and Lyft-Level5 are annotated as GMO. (2) *Forecasting fine-grained GMO*: the categories are also divided into GMO and others but the annotation of GMO are directly from voxel-wise labels of nuScenes-Occupancy removing invalid grids introduced in Sec. 3.2. (3) *Forecasting inflated GMO, fine-grained GSO, and free space*: the categories are divided into GMO from bounding box annotations, GSO following fine-grained annotations, and free space. (4) *Forecasting fine-grained GMO, fine-grained GSO, and free space*: the categories are divided into GMO and GSO both following fine-grained annotations, and free space. Since the Lyft-Level5 dataset lacks occupancy labels, we only conduct the evaluation for the first task on it.

Metrics. For all four tasks, we use intersection over union (IoU) as the performance metric. We separately evaluate the current moment ($t=0$) occupancy estimation and the future time ($t \in [1, N_f]$) forecasting by

$$\mathrm{IoU}_c(\hat{\mathbf{O}}_c, \mathbf{O}_c) = \frac{\sum_{H,W,L} \hat{S}_c \cdot S_c}{\sum_{H,W,L} \hat{S}_c + S_c - \hat{S}_c \cdot S_c}, \tag{1}$$

$$\mathrm{IoU}_f(\hat{\mathbf{O}}_f, \mathbf{O}_f) = \frac{1}{N_f} \sum_{t=1}^{N_f} \frac{\sum_{H,W,L} \hat{S}_t \cdot S_t}{\sum_{H,W,L} \hat{S}_t + S_t - \hat{S}_t \cdot S_t}, \tag{2}$$

where $\hat{S}_t$ and S_t represent the estimated and ground-truth voxel state at timestamp t respectively.

We also provide a singular quantitative indicator to evaluate forecasting performance within the whole time horizon using one value calculated by

$$\tilde{\mathrm{IoU}}_f(\hat{\mathbf{O}}_f, \mathbf{O}_f) = \frac{1}{N_f} \sum_{t=1}^{N_f} \frac{1}{t} \sum_{k=1}^{t} \frac{\sum_{H,W,L} \hat{S}_k \cdot S_k}{\sum_{H,W,L} \hat{S}_k + S_k - \hat{S}_k \cdot S_k}. \tag{3}$$

IoU of timestamps closer to the current moment contributes more to the final $\tilde{\mathrm{IoU}}_f$. This aligns with the principle that occupancy predictions at near timestamps are more crucial for subsequent motion planning and decision making.

3.4. Baselines

We propose four methods as baselines in Cam4DOcc to assist future comparison for the camera-only 4D occupancy forecasting task as shown in Fig. 3.

Static-world occupancy model. The existing camera-based occupancy prediction approaches [17, 22, 29, 41, 45, 47] can only estimate the present occupancy grids based on the current observation. Therefore, one of the most

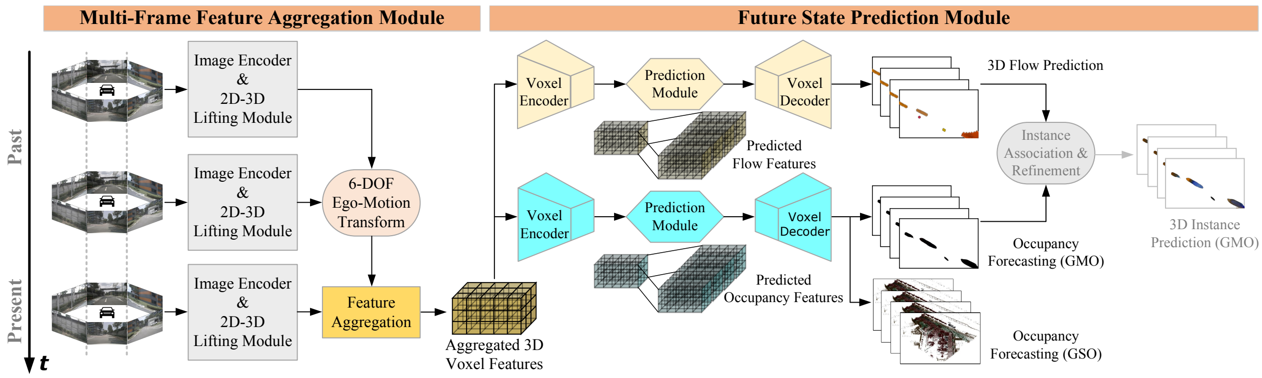


Figure 4. System overview of our proposed OCFNet.

straightforward baselines is to assume the environment remains static over a short time interval. Thus, we can use the present estimated occupancy grids as predictions for all future time steps based on the static-world hypothesis, as illustrated in Fig. 3 a.

Voxelization of point cloud prediction. Another type of baseline can be the occupancy grid voxelization based on the point clouds forecasting results from existing point clouds prediction methods [12, 32, 33, 35]. Here, we use surround-view depth estimation to generate depth maps across multiple cameras, followed by ray casting to generate 3D point clouds, which is applied with point cloud prediction to obtain predicted future pseudo points. Based on that, we then apply point-based semantic segmentation [7, 26, 52] to obtain movable and static labels for each voxel, resulting in the final occupancy predictions (see Fig. 3 b).

2D-3D instance-based prediction. Many off-the-shelf 2D BEV-based instance prediction methods [14, 15, 25, 34, 48] can forecast semantics for a near future with surround-view camera images. The third type of baseline is to obtain forecasted GMO in 3D space by replicating the BEV occupancy grids along the z-axis to the height of the vehicle, as shown in Fig. 3 c. It can be seen that this baseline assumes that the driving surface is flat and all moving objects have the same height. We do not evaluate this baseline on forecasting GSO since boosting 2D results by replication is unsuitable for simulating large-scale backgrounds with much more complex structures compared to GMO.

End-to-end occupancy forecasting network. None of the above baselines can directly predict the future occupancy state of 3D space. They all need additional post-processing based on certain hypotheses to extend and transform the existing results into 4D occupancy forecasting, inevitably introducing inherent artifacts. To fill this gap, we propose a novel approach shown in Fig. 3 d to achieve camera-only 4D occupancy forecasting in an end-to-end manner, introduced in detail in the next section.

4. End-to-End 4D Occupancy Forecasting

To our best knowledge, no existing camera-only 4D occupancy forecasting baseline is capable of simultaneously predicting future occupancy and extracting 3D general objects in an end-to-end fashion. In this paper, we introduce a novel end-to-end spatio-temporal network dubbed OCFNet, depicted in Fig. 4. OCFNet receives sequential past surround-view camera images to predict the present and future occupancy states. It utilizes the multi-frame feature aggregation module to extract warped 3D voxel features and the future state prediction module to forecast future occupancy as well as 3D backward centripetal flow.

4.1. Multi-Frame Feature Aggregation Module

The multi-frame feature aggregation module takes a sequence of past surround-view camera images as input and employs an image encoder backbone to extract 2D features. These 2D features are subsequently lifted and integrated into 3D voxel features by the 2D-3D lifting module. All the resulting 3D feature volumes are transformed to the current coordinate system through the application of 6-DOF ego-car poses, yielding the aggregated feature $F_p \in \mathbb{R}^{(N_p+1)c \times h \times w \times l}$. Here, we collapse the time and feature dimensions into one dimension to implement the following 3D spatiotemporal convolution. Subsequently, we concatenate it with the 6-DOF relative ego-car poses between adjacent frames, leading to the motion-aware feature $F_{pm} \in \mathbb{R}^{(N_p+1)(c+6) \times h \times w \times l}$.

4.2. Future State Prediction Module

With the motion-aware feature aggregated from sequential features as input, the future state prediction module uses two heads to forecast future occupancy as well as the motion of the grids simultaneously. Firstly, a voxel encoder downsamples F_{pm} to multi-scale features $F^i_{pm} \in \mathbb{R}^{(N_p+1)c_i \times \frac{h}{2^i} \times \frac{w}{2^i} \times \frac{l}{2^i}}$, where $i = 0, 1, 2, 3$. Then, the prediction module expands the channel dimension of each F^i_{pm} to $(N_f+1)c_i$ using stacked 3D residual convolutional blocks (see Sec. B in supplementary materials), resulting in $F_{pf}^{i} \in \mathbb{R}^{(N_f+1)c_i \times \frac{h}{2^i} \times \frac{w}{2^i} \times \frac{l}{2^i}}$. They are further concatenated with the feature upsampled by a voxel decoder, after which a softmax function is exploited in the occupancy forecasting head to produce the coarse occupancy feature $F_f^{occ} \in \mathbb{R}^{(N_f+1) \times cls \times h \times w \times l}$. In the flow prediction head, an additional 1×1 convolutional layer instead of the softmax function is utilized to produce the coarse flow feature $F_f^{flow} \in \mathbb{R}^{(N_f+1) \times 3 \times h \times w \times l}$. Lastly, we utilize trilinear interpolation on F_f^{occ} and F_f^{flow}, and an additional argmax function on the occupancy state dimension to generate the final occupancy estimation $\hat{\mathbf{O}}_t \in \mathbb{R}^{(N_f+1) \times H \times W \times L}$ and flow-based motion prediction $\hat{\mathbf{M}}_t \in \mathbb{R}^{(N_f+1) \times 3 \times H \times W \times L}$. Here, we need to estimate the present and forecast the future occupancy with semantics of general objects simultaneously according to the evaluation protocol described in Sec. 3.3. In addition, OCFNet not only forecasts occupancy but also predicts 3D backward centripetal flow as grid motion within the space, which can be utilized to achieve instance prediction (see Sec. E in supplementary materials).

4.3. Loss function

We use the cross-entropy loss as the occupancy forecasting loss L_{occ} and use smooth l_1 distance as the flow prediction loss L_{flow}. The explicit depth loss L_{depth} [28] is also used as the previous work [45] suggests, but here it is only calculated for supervising the present occupancy ($t = 0$) to improve training efficiency and decrease memory consumption. The overall loss for training OCFNet is given by

$$L_{all} = \frac{1}{N_f + 1}\Big(\sum_{t=0}^{N_f} \lambda_1 L_{occ}(\hat{\mathbf{O}}_t, \mathbf{O}_t) + \lambda_2 L_{flow}(\hat{\mathbf{M}}_t, \mathbf{M}_t)\Big) + \lambda_3 L_{depth}(\hat{\mathbf{D}}_0, \mathbf{D}_0), \qquad (4)$$

where $\hat{\mathbf{D}}_0, \mathbf{D}_0$ are the depth image estimated by the 2D-3D Lifting module and ground-truth range image projected from LiDAR data, respectively. λ_1, λ_2, and λ_3 are the weights to balance the optimization for occupancy forecasting, flow prediction, and depth reconstruction.

5. Experiments on Cam4DOcc

Using the proposed Cam4DOcc benchmark, we evaluate the occupancy estimation and forecasting performance of the proposed baselines, including our OCFNet, for four tasks in autonomous driving scenarios.

5.1. Experimental Setups

Dataset details. Following [25, 45], we use 700 out of 850 scenes with ground-truth annotations in the nuScenes and nuScenes-Occupancy datasets, and 130 out of 180 scenes in the Lyft-Level5 for training the proposed baselines and our OCFNet. The remaining scenes are used for evaluation. The length N of each sequence in our benchmark is set to 7 ($N_p = 2$ and $N_f = 4$), which means we use three observations, including the present one, to forecast occupancy in four incoming time steps. Because nuScenes is annotated at 2 Hz while Lyft-Level5 is annotated at 5 Hz, we report the forecasting performance with different time intervals. The predefined range of each sequence is set as [-51.2 m, 51.2 m] for x-axis and y-axis, and [-5 m, 3 m] for z-axis. The voxel resolution is 0.2 m, leading to occupancy grids with the size of $512 \times 512 \times 40$ in the present coordinate system of each sequence. After the data reorganizing of our Cam4DOcc benchmark, the number of sequences for training and test are 23930 and 5119 in nuScenes and nuScenes-Occupancy, and 15720 and 5880 in Lyft-Level5.

Baseline setups. We choose the state-of-the-art camera-based approaches as the outset of each baseline proposed in Sec. 3.4. For the static-world occupancy model, we use the camera baseline of OpenOccupancy [45] (OpenOccupancy-C) to estimate the occupancy state of the present frame, which is then regarded as the prediction of all the future time steps. For the voxelization of point cloud prediction, we use SurroundDepth [46] to estimate continuous surrounding depth maps, which are then downsampled to generate pseudo point clouds by ray casting. Based on sequential pseudo point clouds input, we then use PCPNet [33] to forecast incoming 3D point clouds, followed by Cylinder3D [52] to extract point-level GMO and GSO labels, and further voxelize the results into occupancy grids (SPC). For the 2D-3D instance-based prediction, we choose PowerBEV [25] to forecast occupancy semantics on BEV and then lift the 2D results to 3D space (PowerBEV-3D). As to our proposed OCFNet, we directly implement 4D occupancy forecasting end-to-end. Notably, PowerBEV is trained by the 2D ground-truth semantics and 2D flow projected to the BEV plane. Besides, only PowerBEV and OCFNet are trained with flow annotations from Cam4DOcc simultaneously since they both have the flow head. To show that our proposed OCFNet can generate good forecasted results even seeing limited training data, we report the performance of OCFNet only trained on $\frac{1}{6}$ training sequences as well as the performance of the one trained on all training sequences (OCFNet†). OpenOccupancy-C, PowerBEV, and OCFNet are trained for 15 epochs using AdamW optimizer [21] with an initial learning rate 3e-4 and a weight decay of 0.01. SurroundDepth and Cylinder3D used in the point cloud prediction baseline are fine-tuned as their open sources suggest. PCPNet is firstly pretrained by range loss for 40 epochs using the same optimizer, but the initial learning rate is set to 1e-3. After that, it is further fine-tuned by Chamfer distance loss [13] for 10 epochs with a learning rate of 6e-4. All the networks mentioned above are trained with a batch size of 8 on 8 A100 GPUs. More details about the model parameters of our OCFNet are provided in supplementary Sec. B.

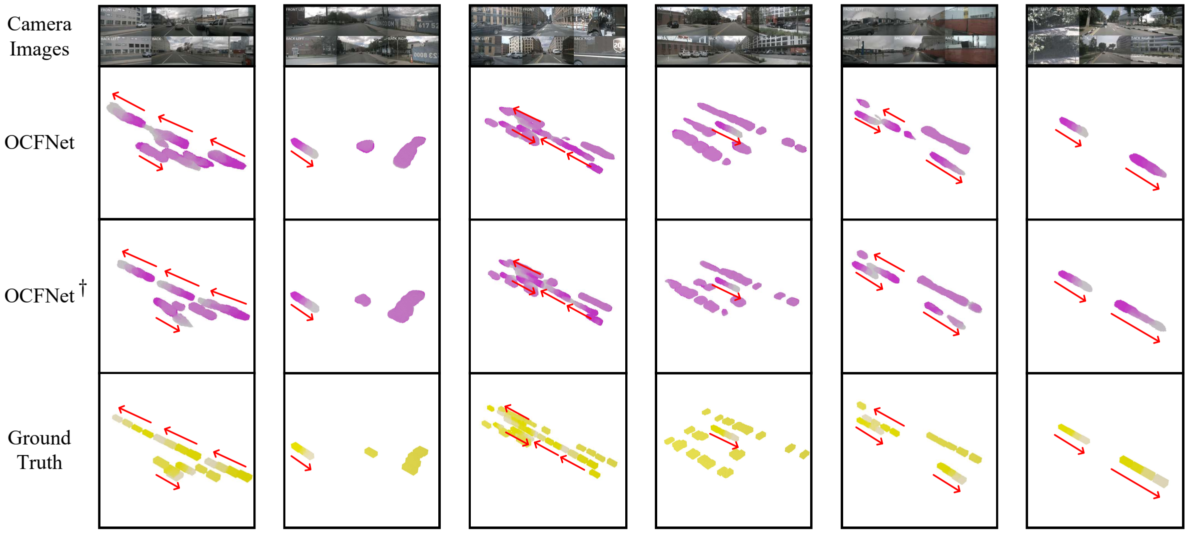


Figure 5. Visualization of forecasting inflated GMO by our proposed OCFNet. The prediction results and ground truth from timestamps 1 to N_f are assigned colors from dark to light. The motion trend of each moving object is represented by red arrows.

5.2. 4D Occupancy Forecasting Assessment

Evaluation on forecasting inflated GMO. Results of the first task, forecasting inflated GMO on nuScenes and Lyft-Level5, are presented in Tab. 1. Here, OpenOccupancy-C, PowerBEV, and OCFNet are trained only with inflated GMO labels, while PCPNet is trained by holistic point clouds. As shown, OCFNet and OCFNet† outperform all other baselines, surpassing the BEV-based method by 12.4% and 13.3% in IoU_f and $\tilde{\text{IoU}}_f$ on nuScenes. On Lyft-Level5, our OCFNet and OCFNet† consistently outperforms PowerBEV-3D by 20.8% and 21.8% in IoU_f and $\tilde{\text{IoU}}_f$. In addition, Fig. 5 shows the results of nuScenes GMO occupancy forecasted by our OCFNet and CFNet†, which indicates that OCFNet trained only with limited data can still capture the motion of GMO occupancy grids reasonably. The visualization on Lyft-Level5 is shown in supplementary Sec. G. The baseline SPC cannot work well for the present frame and even tends to fail while forecasting future occupancy state. This is because movable objects are labeled as the inflated dense voxel grids in this task, while the voxelization of PCPNet outputs is from sparse point-level prediction. In addition, the shape of the predicted objects loses consistency significantly in future time steps. The performance of OpenOccupancy-C is much better than that of the point cloud prediction baseline but still has a weak ability to estimate present occupancy and forecast future occupancy compared to PowerEBV-3D and OCFNet.

Evaluation on forecasting fine-grained GMO. We further report the occupancy estimation and forecasting performance on fine-grained general movable objects with nuScenes-Occupancy (the second-level task). In Tab. 2, we exhibit how the IoU of the forecasted objects changes once the GMO annotations have fine-grained voxel format

Table 1. Comparison of performance on forecasting inflated GMO

approach	nuScenes			Lyft-Level5		
	IoU_c	IoU_f (2 s)	$\tilde{\text{IoU}}_f$	IoU_c	IoU_f (0.8 s)	$\tilde{\text{IoU}}_f$
OpenOccupancy-C [45]	12.17	11.45	11.74	14.01	13.53	13.71
SPC [33, 46, 52]	1.27	failed	failed	1.42	failed	failed
PowerBEV-3D [25]	23.08	21.25	21.86	26.19	24.47	25.06
OCFNet (ours)	27.86	23.89	24.77	32.12	29.56	30.53
OCFNet† (ours)	**31.30**	**26.82**	**27.98**	**36.41**	**33.56**	**34.60**

SPC: SurroundDepth [46] + PCPNet [33] + Cylinder3D [52]

rather than the inflated one in the first-level task for training as well as evaluation. It can be seen that the IoU of GMO forecasted by all the methods except the point cloud prediction baseline decreases significantly because it is rather difficult to predict sophisticated moving 3D structures using past continuous camera images. In contrast, SPC presents slightly better performance compared to the results in Tab. 1 since the ground-truth labels are also fine-grained and sparser than the counterparts in the first-level task. However, due to the loss of shape consistency, it still has the worst performance among the baselines. Besides, we can also see in Tab. 2 that OCFNet and OCFNet† still have the best performance. This experiment reveals the reason why Cam4DOcc suggests the inflated labels for GMO annotation in the occupancy forecasting task: Forecasting sophisticated future 3D structures of movable objects only using camera images is very difficult while forecasting inflated GMO potentially promotes more reliable and safer navigation in autonomous driving applications.

Evaluation on forecasting inflated GMO, fine-grained GSO, and free space. Next, we compare the performance of different methods on forecasting inflated general movable objects, fine-grained general static objects, and free space (the third-level task). Here, we do not report the GSO results from the 2D-3D instance-based prediction since the

Table 2. Comparison on forecasting fine-grained GMO

approach	nuScenes-Occupancy		
	IoU$_c$	IoU$_f$ (2 s)	$\widetilde{\text{IoU}}_f$
OpenOccupancy-C [45]	10.82	8.02	8.53
SPC [33, 46, 52]	5.85	1.08	1.12
PowerBEV-3D [25]	5.91	5.25	5.49
OCFNet (ours)	10.15	8.35	8.69
OCFNet† (ours)	**11.45**	**9.68**	**10.10**

Table 3. Comparison of performance on forecasting inflated GMO, fine-grained GSO, and free space simultaneously

approach	IoU$_c$			IoU$_f$ (2 s)			$\widetilde{\text{IoU}}_f$
	GMO	GSO	mean	GMO	GSO	mean	GMO
OpenOccupancy-C [45]	13.53	16.86	15.20	12.67	17.09	14.88	12.97
SPC [33, 46, 52]	1.27	3.29	2.28	failed	1.40	–	failed
PowerBEV-3D [25]	23.08	–	–	21.25	–	–	21.86
OCFNet (ours)	26.41	16.95	21.68	22.21	17.14	19.68	23.06
OCFNet† (ours)	**29.84**	**17.72**	**23.78**	**25.53**	**17.81**	**21.67**	**26.53**

Table 4. Comparison of performance on forecasting fine-grained GMO, fine-grained GSO, and free space simultaneously

approach	IoU$_c$			IoU$_f$ (2 s)			$\widetilde{\text{IoU}}_f$
	GMO	GSO	mean	GMO	GSO	mean	GMO
OpenOccupancy-C [45]	9.62	17.21	13.42	7.41	17.30	12.36	7.86
SPC [33, 46, 52]	5.85	3.29	4.57	1.08	1.40	1.24	1.12
PowerBEV-3D [25]	5.91	–	–	5.25	–	–	5.49
OCFNet (ours)	9.54	17.30	13.42	8.23	17.32	12.78	8.46
OCFNet† (ours)	**11.02**	**17.79**	**14.41**	**9.20**	**17.83**	**13.52**	**9.66**

fine-grained 3D structure of static foreground and background objects cannot be approximately estimated by lifting 2D voxel grids to 3D space. The experimental results are shown in Tab. 3. SPC remains the worst in this experiment where the IoU of inflated GMO is consistent with the results of Tab. 1. OCFNet and OCFNet† outperform OpenOccupancy-C significantly in terms of estimating GMO occupancy in both present moment and future time steps. It also can be seen that by aggregating features of multiple past frames, OCFNet† enhances the performance of GSO occupancy estimation of single-frame-based OpenOccupancy-C by 5.1% and 4.2% on IoU$_c$ and IoU$_f$ respectively. For OpenOccupancy-C and our OCFNet, the IoU values of future GSO slightly increase due to the jitter of ground truth annotations from nuScenes-Occupancy.

Evaluation on forecasting fine-grained GMO, fine-grained GSO, and free space. In the fourth-level task, only OpenOccupancy-C and our OCFNet need to be retrained. As seen in Tab. 4, OCFNet† remains the best performance against all the other approaches on forecasting fine-grained objects of interest. Compared to the results in Tab. 2, the GMO forecasting performance of OpenOccupancy-C and our OCFNet drops slightly due to additional artifacts introduced by the fine-grained GSO class.

Table 5. Ablation study on flow prediction head

approach	IoU$_c$	IoU$_f$				$\widetilde{\text{IoU}}_f$
		0.5 s	1.0 s	1.5 s	2.0 s	
OCFNet w/o flow	26.84	25.01	24.04	23.38	22.99	23.86
OCFNet	**27.86**	**25.95**	**24.92**	**24.33**	**23.89**	**24.77**
improvement ↑	3.8%	3.8%	3.7%	4.1%	3.9%	3.8%

5.3. Ablation Study on Multi-Task Learning

We conduct an ablation study on the flow prediction head to present the enhancement from the multi-task learning scheme. As Tab. 5 shows, the complete OCFNet enhances the one without the flow prediction head by around 4% in both present and future occupancy estimation. The reason could be that 3D flow guides learning GMO motion in each time interval, as shown in Sec. D in supplementary materials, and thus helps the model determine the change of occupancy estimation in the next timestamp. With this analysis, using 3D backward centripetal flow in our Cam4DOcc is suggested for future end-to-end 4D Occupancy forecasting models to achieve better forecasting performance.

6. Conclusion

This paper proposes a novel benchmark, Cam4DOcc, for camera-only 4D occupancy forecasting in autonomous driving applications. We first establish the devised dataset in a new format based on several publicly available datasets. Then the standardized evaluation protocol as well as four types of baselines are further proposed to provide basic reference in our Cam4DOcc benchmark. Moreover, we propose the first camera-based 4D occupancy forecasting network OCFNet to estimate future occupancy states in an end-to-end manner. Multiple experiments with four different tasks are conducted based on our Cam4DOcc benchmark to thoroughly evaluate the proposed baselines as well as our OCFNet. The experimental results show that OCFNet outperforms all the baselines and can still produce reasonable future occupancy even seeing limited training data.

Insights: By comparing four different types of baselines, we demonstrated that end-to-end spatiotemporal network could be the most promising research direction for camera-only occupancy forecasting. Besides, using inflated GMO annotation and additional 3D backward centripetal flow is also verified to be beneficial for 4D occupancy forecasting.

Limitation: While our OCFNet has achieved notable results, camera-only 4D occupancy forecasting remains challenging, especially for predicting over longer time intervals with many moving objects. Our Cam4DOcc benchmark and comprehensive analysis aim to enhance understanding of the strengths and limitations of current occupancy perception models. We envision this benchmark as a valuable evaluation tool, and our OCFNet can serve as a foundational codebase for future research on 4D occupancy forecasting.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. In *ECCV*, pages 444–460, 2022. 2, 3

[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 *CVPR*, pages 11621–11631, 2020. 2, 3

[3] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 2

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

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 40(4):834–848, 2017. 1

[6] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, pages 4193–4202, 2020. 3

[7] Xieyuanli Chen, Shijie Li, Benedikt Mersch, Louis Wiesmann, Jürgen Gall, Jens Behley, and Cyrill Stachniss. Moving object segmentation in 3d lidar data: A learning-based approach exploiting sequential data. *RA-L*, 6(4):6529–6536, 2021. 5

[8] Xieyuanli Chen, Benedikt Mersch, Lucas Nunes, Rodrigo Marcuzzi, Ignacio Vizzo, Jens Behley, and Cyrill Stachniss. Automatic Labeling to Generate Training Data for Online LiDAR-Based Moving Object Segmentation. *RA-L*, 7(3): 6107–6114, 2022. 3

[9] Bowen Cheng, Maxwell D Collins, Yukun Zhu, Ting Liu, Thomas S Huang, Hartwig Adam, and Liang-Chieh Chen. Panoptic-deeplab: A simple, strong, and fast baseline for bottom-up panoptic segmentation. In *CVPR*, pages 12475–12485, 2020. 1

[10] Wenchao Ding, Lu Zhang, Jing Chen, and Shaojie Shen. Safe trajectory generation for complex urban environments using spatio-temporal semantic corridor. *RA-L*, 4(3):2997–3004, 2019. 1

[11] Wenchao Ding, Lu Zhang, Jing Chen, and Shaojie Shen. Epsilon: An efficient planning system for automated vehicles in highly interactive environments. *TRO*, 38(2):1118–1138, 2021. 1

[12] Hehe Fan and Yi Yang. Pointrnn: Point recurrent neural network for moving point cloud processing. *arXiv preprint arXiv:1910.08287*, 2019. 2, 5

[13] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, pages 605–613, 2017. 6

[14] Noureldin Hendy, Cooper Sloan, Feng Tian, Pengfei Duan, Nick Charchut, Yuesong Xie, Chuang Wang, and James Philbin. Fishing net: Future inference of semantic heatmaps in grids. In *CVPRW*, 2020. 2, 5

[15] 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. 2, 3, 5

[16] Jie Hu, Linyan Huang, Tianhe Ren, Shengchuan Zhang, Rongrong Ji, and Liujuan Cao. You only segment once: Towards real-time panoptic segmentation. In *CVPR*, pages 17819–17829, 2023. 1

[17] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 1, 2, 3, 4

[18] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet. Lyft level 5 perception dataset 2020, 2019. 2

[19] Tarasha Khurana, Peiyun Hu, Achal Dave, Jason Ziglar, David Held, and Deva Ramanan. Differentiable raycasting for self-supervised occupancy forecasting. In *ECCV*, pages 353–369, 2022. 2

[20] Tarasha Khurana, Peiyun Hu, David Held, and Deva Ramanan. Point cloud forecasting as a proxy for 4d occupancy forecasting. In *CVPR*, pages 1116–1124, 2023. 2

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 6

[22] Hongyu Li, Zhengang Li, Neşet Ünver Akmandor, Huaizu Jiang, Yanzhi Wang, and Taşkın Padır. Stereovoxelnet: Real-time obstacle detection based on occupancy voxels from a stereo camera using deep neural networks. In *ICRA*, pages 4826–4833, 2023. 4

[23] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, pages 3351–3359, 2020. 3

[24] Liulei Li, Tianfei Zhou, Wenguan Wang, Jianwu Li, and Yi Yang. Deep hierarchical semantic segmentation. In *CVPR*, pages 1246–1257, 2022. 1

[25] 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, 4, 5, 6, 7, 8

[26] Shijie Li, Xieyuanli Chen, Yun Liu, Dengxin Dai, Cyrill Stachniss, and Juergen Gall. Multi-scale interaction for real-time lidar data segmentation on an embedded platform. *RA-L*, 7(2):738–745, 2022. 5

[27] Wentong Li, Yuqian Yuan, Song Wang, Jianke Zhu, Jianshu Li, Jian Liu, and Lei Zhang. Point2mask: Point-supervised panoptic segmentation via optimal transport. In *ICCV*, pages 572–581, 2023. 1

[28] 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 *AAAI*, pages 1477–1485, 2023. 6

[29] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anand-

kumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, pages 9087–9098, 2023. 1, 4

[30] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*, pages 531–548, 2022. 1

[31] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, pages 3431–3440, 2015. 1

[32] Fan Lu, Guang Chen, Zhijun Li, Lijun Zhang, Yinlong Liu, Sanqing Qu, and Alois Knoll. Monet: Motion-based point cloud prediction network. *TITS*, 23(8):13794–13804, 2021. 2, 5

[33] Zhen Luo, Junyi Ma, Zijie Zhou, and Guangming Xiong. Pcpnet: An efficient and semantic-enhanced transformer network for point cloud prediction. *RA-L*, 2023. 2, 5, 6, 7, 8

[34] Reza Mahjourian, Jinkyu Kim, Yuning Chai, Mingxing Tan, Ben Sapp, and Dragomir Anguelov. Occupancy flow fields for motion forecasting in autonomous driving. *RA-L*, 7(2): 5639–5646, 2022. 2, 5

[35] Benedikt Mersch, Xieyuanli Chen, Jens Behley, and Cyrill Stachniss. Self-supervised point cloud prediction using 3d spatio-temporal convolutional networks. In *CoRL*, pages 1444–1454, 2022. 2, 5

[36] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 2

[37] Cody Reading, Ali Harakeh, Julia Chae, and Steven L Waslander. Categorical depth distribution network for monocular 3d object detection. In *CVPR*, pages 8555–8564, 2021. 1

[38] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *3DV*, pages 111–119, 2020. 3

[39] Haoran Song, Wenchao Ding, Yuxuan Chen, Shaojie Shen, Michael Yu Wang, and Qifeng Chen. Pip: Planning-informed trajectory prediction for autonomous driving. In *ECCV*, pages 598–614, 2020. 1

[40] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 2

[41] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *ICCV*, pages 8406–8415, 2023. 1, 2, 3, 4

[42] Maneekwan Toyungyernsub, Esen Yel, Jiachen Li, and Mykel J Kochenderfer. Dynamics-aware spatiotemporal occupancy prediction in urban environments. In *IROS*, pages 10836–10841, 2022. 2

[43] Niclas Vödisch, Kürsat Petek, Wolfram Burgard, and Abhinav Valada. Codeps: Online continual learning for depth estimation and panoptic segmentation. *RSS*, 2023. 1

[44] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao. Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In *CVPR*, pages 7464–7475, 2023. 1

[45] 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. 1, 2, 3, 4, 6, 7, 8

[46] 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*, pages 539–549, 2023. 6, 7, 8

[47] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *ICCV*, pages 21729–21740, 2023. 1, 2, 4

[48] Pengxiang Wu, Siheng Chen, and Dimitris N Metaxas. Motionnet: Joint perception and motion prediction for autonomous driving based on bird's eye view maps. In *CVPR*, pages 11385–11395, 2020. 2, 5

[49] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *NeurIPS*, 34:12077–12090, 2021. 1

[50] Yunpeng Zhang, Zheng Zhu, Wenzhao Zheng, Junjie Huang, Guan Huang, Jie Zhou, and Jiwen Lu. Beverse: Unified perception and prediction in birds-eye-view for vision-centric autonomous driving. *arXiv preprint arXiv:2205.09743*, 2022. 2, 3

[51] Yunsong Zhou, Quan Liu, Hongzi Zhu, Yunzhe Li, Shan Chang, and Minyi Guo. Mogde: Boosting mobile monocular 3d object detection with ground depth estimation. *NeurIPS*, 35:2033–2045, 2022. 1

[52] Xinge Zhu, Hui Zhou, Tai Wang, Fangzhou Hong, Yuexin Ma, Wei Li, Hongsheng Li, and Dahua Lin. Cylindrical and asymmetrical 3d convolution networks for lidar segmentation. In *CVPR*, pages 9939–9948, 2021. 5, 6, 7, 8