2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

GenerOcc: Self-supervised Framework of Real-time 3D Occupancy Prediction for Monocular Generic Cameras

Xianghui Pan[1], Jiayuan Du[1], Shuai Su[1], Wenhao Zong[2], Xiao Wang[2], Chengju Liu[1] and Qijun Chen[1], *Senior Member, IEEE*

***Abstract*— In the context of 3D scene perception tasks, the significance of 3D occupancy prediction has been progressively growing, aiming to forecast the occupancy state of voxels in a discrete 3D space. However, existing methods typically exhibit several limitations, such as restricted adaptability to non-pinhole cameras due to fixed camera parameters, heavy reliance on 3D annotations because of the inability to project 3D output back to the camera plane, and inferior real-time inference performance resulting from the conversion process from 2D to 3D features. To address these constrains, we introduce GenerOcc, a self-supervised framework of real-time 3D occupancy prediction for monocular generic cameras. We have collected the fisheye Dominant dataset to confirm the compatibility of our ray-based camera model with non-pinhole cameras. By transforming the occupancy prediction task into a depth estimation task in a self-supervised manner, we eliminate dependency on 3D annotations. Furthermore, we propose a parametric voxel probability distribution module that leverages 2D features to quickly predict 3D occupancy without 3D representations of the scene. Additionally, our GenerOcc has been extensively evaluated on public pinhole Occ3D-nuScenes dataset and our proprietary fisheye Dominant dataset, both yielding impressive performance.**

I. Introduction

3D occupancy prediction has emerged as a novel vision-centric perception technology in the field of autonomous driving involving the task of predicting occupancy or semantic labels for voxels within a discrete 3D space [1]. Unlike 3D object detection based on bounding box [2], 3D occupancy prediction providing autonomous vehicles a more comprehensive understanding of their surroundings by detail description of geometric shape and background.

Currently, vehicles are often equipped with fisheye cameras in automotive industry to provide a wider field of view for enhanced driver assistance. However, there is a significant scarcity of fisheye datasets [3], and available methods compatible with fisheye cameras are also rare in 3D scene perception tasks. Furthermore, the majority of the models employed for 3D occupancy prediction rely on annotations such as point clouds [4], [5], 3D occupancy labels [6], depth images [7], [8] or even truncated signed distance function (TSDF) [9], [10], which are both costly and time-consuming. Moreover, many methods utilize 3D conversion module to transform 2D features from images into 3D represents of scene [11], [12], [13], leading to substantial computational demands and extra difficulties in actual deployment.

Addressing aforementioned issues, we propose GenerOcc, a self-supervised framework of real-time 3D occupancy prediction for monocular generic cameras. We have collected the Dominant dataset to validate the compatibility of our ray-based camera model with various camera types. The Dominant dataset contains images from distorted fisheye camera, corresponding ego-pose information and point clouds from 128-beam LiDAR, which has shorter interval and higher resolution than public Occ3d-nuScene dataset. Recognizing the significance of depth information in mapping features from the image plane to the 3D scene, we transform the task of occupancy prediction into a depth estimation problem, which has been extensively studied [14], [15], [16]. Given that voxelization process from depth to voxels is non-differentiable, we employ two routes to get rid of 3D annotations: self-supervised depth estimation for generating the occupancy label, and direct occupancy prediction. During the latter, we adopt a parametric voxel probability distribution to maintain the differentiability of module and accelerate inference by avoiding 3D convolutions or other time-consuming methods.

To evaluate the predictive capabilities of our method, we conducted experiments on two datasets: the widely used pinhole Occ3D-nuScenes [4], [17] dataset and our proprietary fisheye Dominant dataset. Results are presented in Sec. IV-C, which highlights the adaptability of our method to generic cameras and effectiveness on occupancy prediction without any 3D annotations.

Our main contributions are summarized as follows:

- We have collected the fisheye Dominant dataset during vehicle drives, containing fisheye images, ego-pose and point clouds, which will be released at `https://github.com/RONGSHENG0404/dominant3062`.
- We build a ray-based camera model to broaden the compatibility of our approach with any types of cameras. We are the first to predict occupancy for generic cameras.
- We transform occupancy prediction into depth estimation, and utilize a self-supervised manner to cast off dependency on 3D annotations.
- We design a parametric voxel probability distribution module to ensure the differentiability and accelerate inference for 3D occupancy prediction.

This paper is supported by the National Natural Science Foundation of China under Grants (62073245, 62173248, 62233013). Shanghai Municipal Science and Technology Major Project (2021SHZDZX0100) and the Fundamental Research Funds for the Central Universities. (Corresponding author: Chengju Liu, Qijun Chen)

[1]Department of Control Science and Engineering, Tongji University, Shanghai, China (e-mail:panxianghui, dujiayuan, sushuai, liuchengju, qjchen@tongji.edu.cn).

[2]DominantTech (e-mail:daniel.zong, xiao.wang@dominant-tech.com).

II. Related Work

Generic Camera Models. Fisheye cameras have become a widely applied choice for automobile manufacturers due to higher resolution in near-field. However, for a long time, many approaches have grappled with the limitations of image clipping and stretching deformations when converting fisheye images into pinhole images for training, resulting from lack of fisheye dataset for 3D perception task. Kumar et al. have made significant advancements in this area. They introduced super-resolution networks and variable convolutions for feature extraction from fisheye images [15], and incorporated robust loss functions and self-attention mechanisms [18] on depth estimation. Nevertheless, these computational improvements have led to exponentially increased consumption. Vasiljevic et al. [16] proposed the NRS model, aiming to accommodate any camera model by employing ray surfaces with learnable parameters. However, the fitting effect of learnable ray surfaces is not satisfactory. Inspired by those pioneer works, we build our ray-based camera model for occupancy prediction, and collect our proprietary fisheye Dominant dataset to validate our method.

Annotations of 3D Occupancy Prediction. In the field of scene perception, 3D occupancy prediction has recently drawn more attention in the context of 3D object detection. The task is often heavily dependent on the availability of accurate 3D information such as point clouds, depth, or 3D occupancy labels. Chen et al. [10] even utilize conrresponding TSDF with image as input to build a multi-modal scene perception architecture. While most of methods rely on the 3D annotations, Cao et al. [11] have made significant contributions to image-only semantic occupancy prediction research on monocular camera, and Pan et al. [19] propose a novel paradigm to train model with 2D labels by projecting 3D represents to camera plane on multi-view. Under the influence of NeRF [20] and SceneRF [21], Huang et al. [22] further put forward self-supervised occupancy prediction using neural radiance fields. All these methods are designed for multi-task, and employ the depth information extracted from images implicitly, leading to a huge expense on training. We focus solely on the 3D geometric occupancy prediction task, and explicitly estimate depth in self-supervised manner, which avoids both the high cost of training models and reliance on annotations.

Represents for 3D Occupancy Prediction. It is widely believed that 3D represents of scene is the key point for achieving complete 3D scene understanding. Thus, various 3D feature extracting modules emerge in an endless stream. Cao et al. [11] employ 3D UNet [23] and 3D context relation prior module to enhance spatio-semantic awareness. Li et al. [12] utilize spatial cross-attention mechanism to interact 2D features with different cameras. Based on that, Huang et al. [13] subjoin hybrid-attention module to communicate with three perpendicular planes. All of these approaches have achieved good results by extracting 3D features from 2D images, at the price of time-consuming and challenging to deploy in practical applications. In contrast, we design a parametric distribution module to forecast the occupied probability of voxels, using only 2D features extracted from images, with a good real-time performance.

III. Methods

A. Problem Statement

In this work, we attempt to predict the 3D occupancy of surrounding scene with only one image $I \in \mathbb{R}^{H\times W\times 3}$ and ray surface $S \in \mathbb{R}^{H\times W\times 3}$ for different camera, which is a vector set of incident rays determined by camera parameters. Formally, the 3D occupancy prediction can be represented as

$$O = \mathbb{G}(I,S),\ O \in \mathbb{R}^{X\times Y\times Z} \tag{1}$$

where $\mathbb{G}$ is a mapping function from image and ray surface to occupancy. Occupancy O is composed by $X \times Y \times Z$ voxels, representing the occupied probability of each voxel.

B. Overall Framework

Our framework of self-supervised 3D occupancy prediction, as shown in Fig. 1, comprise of two fundamental components: depth estimation and voxel prediction. The self-supervised depth estimation provides supervision for occupancy prediction. The parametric voxel probability distribution module forecast the voxel probability distribution along the same incident ray, enabling the fusion of these probabilities to derive occupancy prediction. Notably, both components can be readily applied to datasets captured by any type of camera while maintaining excellent real-time performance. In cases where the dataset includes point clouds from LiDAR or depth information from RGBD, we can leverage voxelization techniques to construct occupancy labels, which eliminates the need for the depth estimation component in such scenarios.

C. Ray-based Camera Model

Unlike a simple transformation matrix for the pinhole camera, the generic camera unprojection model requires individual description of each incident ray. Taking the classic fisheye camera model as an example, we need to calculate ray surface $S(v,u) = [x_m, y_m, z_m]^T$ for each pixel by unprojection equations in Eq. 2.

$$\varphi = \arctan2\left(\frac{v-c_y}{f_y}, \frac{u-c_x}{f_x}\right) \tag{2a}$$

$$\rho = \theta + k_1\theta^3 + k_2\theta^5 + k_3\theta^7 + k_4\theta^9 \tag{2b}$$

$$\rho^2 = \left(\frac{u-c_x}{f_x}\right)^2 + \left(\frac{v-c_y}{f_y}\right)^2 \tag{2c}$$

$$\begin{bmatrix} x_m \\ y_m \end{bmatrix} = \frac{\tan\theta}{\sqrt{1+\tan^2\varphi}} \begin{bmatrix} 1 \\ \tan\varphi \end{bmatrix} \tag{2d}$$

where c_i, f_i and k_i are intrinsic of fisheye camera. For pixels in field of view, we set $z_m = 1$, otherwise $z_m = 0$. In this manner, we establish a correspondence between the pixel coordinates in the image and the direction vector of the incident rays, serving as a foundation for extending to any type of camera.

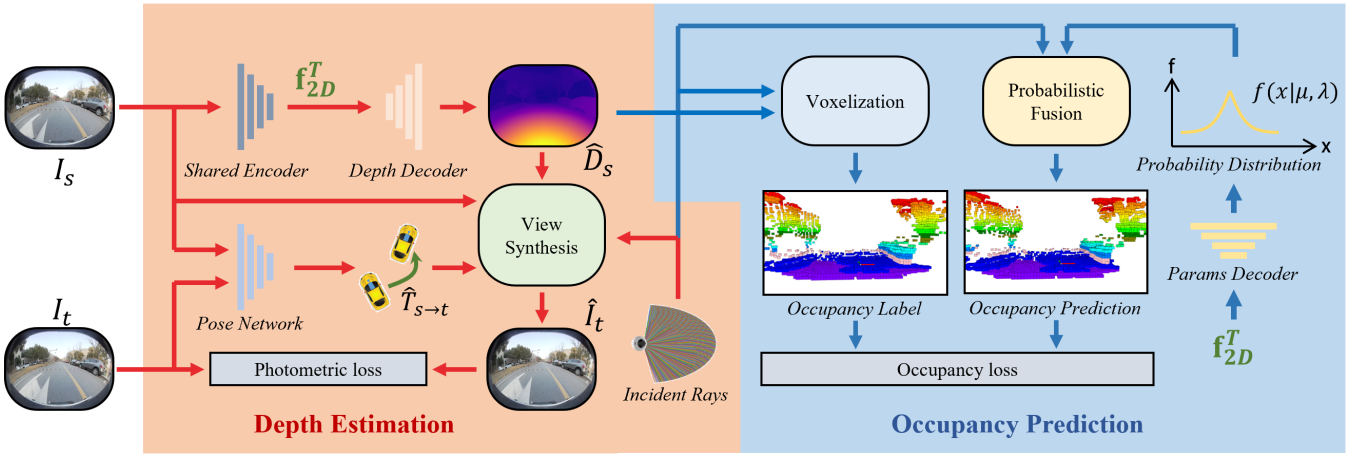


Fig. 1. **Overall framework of GenerOcc.** Our approach utilizes a self-supervised depth estimation module to generate occupancy supervision and a parametric voxel prediction model to forecast 3D occupancy. As both of these components capture the spatial characteristics of objects, we share a common feature extractor. Furthermore, our depth estimation and voxel prediction methods are explicitly designed based on the camera projection model, enabling their applicability to datasets of different cameras.

Generally, mapping matrix from an image to occupancy is a huge sparse matrix, whose dimension is $H \times W \times X \times Y \times Z$. To avoid such a huge expense, we calculate the voxels that each incident ray passes through in advance, and then predict the occupied probability distribution in these voxels, so that we reduce the dimension of the mapping matrix to less than $H \times W \times (X+Y+Z) \times 3$. The process can be described as:

$$V_0 = \mathbb{M}(S),\ V_0 \in \mathbb{R}^{H\times W\times L\times 3},\ L < X+Y+Z \tag{3}$$

where $\mathbb{M}$ is mapping function from ray surface S to voxel predistribution V_0, V_0 contains the coordinates of the voxels on the same ray at the pixel level, and L means the maximum number of voxels in a ray. Based on that, we can take all the voxels on the same ray into consideration when calculate voxel distribution, instead of part voxels where the discrete sampling points are as LSS [24] and SelfOcc [22] did.

D. Self-supervised Depth Estimation

Following SfM (Structure from Motion) method [25], we utilize depth D_s, pose transition matrix $T_{s\to t}$ and the camera intrinsic matrix to calculate the pixel position of source image I_s in the target image I_t coordinate. For a generic camera, however, maximizing the cosine similarity between the points in I_t coordinate and the ray surface is an alternative operation [16], where softmax is utilized instead of argmax to ensure differentiability in this process. While Vasiljevic [16] incorporated additional network to aid the pinhole camera in reconstructing the ray surface of other cameras, our empirical testing revealed suboptimal results with this approach. Consequently, we opted to utilize the original ray surface S of the camera in our method as shown in Eq. 4.

$$r_t(v,u) = \hat{R}_{s\to t} * S(v,u) * \hat{D}_s(v,u) + \hat{p}_{s\to t} \tag{4a}$$
$$Q = \mathrm{softmax}(\cos(r_t(v,u), S(v_i,u_j))) \tag{4b}$$
$$\hat{I}_t(v,u) = I_s(\sum_{i,j} Q(i,j) * v_i, \sum_{i,j} Q(i,j) * u_j) \tag{4c}$$

where $\hat{R}_{s\to t}$ and $\hat{p}_{s\to t}$ indicate the rotation matrix and displacement vector of predicted pose transition matrix $\hat{T}_{s\to t}$ respectively, $v_i \in [v-\delta, v+\delta]$ and $u_j \in [u-\delta, u+\delta]$ represent the coordinates of adjacent pixels, δ means pixel patch and Q represents coefficient matrix by applying the softmax operation to cosine similarity matrix.

Under the influence of photometric and pose loss in Sec. III-F, we get the pseudo-LiDAR points $r_t(v,u)$ of scene. After allocating points to voxels, we get the occupancy label for occupancy prediction.

E. Parametric Voxel Distribution Module

For a given image, we predict parameters of voxel distribution for each pixel using a shared encoder network with depth esimation, followed by a decoder network. We adopt Laplacian distribution to calculate occupied probability for each voxel in an incident ray as described in Eq. 5.

$$f(x|\mu,\lambda) = \frac{1}{2\lambda}\exp(-\frac{|x-\mu|}{\lambda}) \tag{5a}$$
$$p(x=l) = \sum_{i=1}^{n} \frac{1}{n} f(l+\frac{i}{n}) \tag{5b}$$

where $\mu,\lambda \in \mathbb{R}$ represent location and scale parameter of Laplacian distribution respectively, $f(x|\mu,\lambda)$ is occupied probability density function, n is the number of samples between two adjacent voxels, and $l \in \{0,1,2,...,L-1\}$ is the ordinal number of voxel. Then, we get probability matrix $P \in \mathbb{R}^{H\times W\times L}$ for whole rays with their voxels. As shown in Fig. 2, for voxels passed by more than one ray, we utilize parallel fusion to aggregate probability from multiple rays in the same voxel (x,y,z), described as Eq. 6.

$$M(x,y,z) = \{(v,u,n) \mid V_0(v,u,n) = [x,y,z]^T\} \tag{6a}$$
$$\hat{O}(x,y,z) = 1 - \prod_{m\in M}(1-P(m)) \tag{6b}$$

where v_i, u_i and n_i represents indices of the ray passing through the voxel and ordinal number of the voxel on this ray,

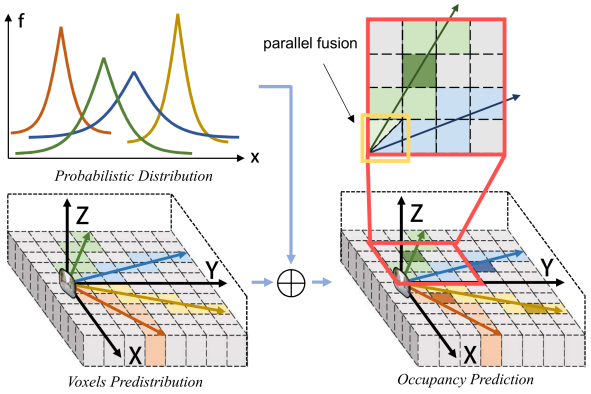


Fig. 2. **Probabilistic distribution and fusion.** Based on the probability distribution along each ray, we assign probabilities to the voxels traversed by the ray. Higher probabilities are associated with darker colors. In cases where multiple rays pass through a voxel, the probabilities are fused in parallel, such as the voxel within the highlighted gold box in this figure.

$M(x,y,z)$ means assemble of (v,u,n), and $\hat{O}(x,y,z)$ contains probability prediction of this voxel.

F. Training Strategy

Following the work of Vasiljevic et al [16], we train the depth and pose network for generic images in a self-supervised manner to generate supervision for 3D occupancy. As shown in Fig 1, we optimize the depth network and pose network by minimizing a per-pixel photometric reprojection loss, which is described as Eq. 7.

$$L_d(I_t,\hat{I}_t) = \alpha\frac{1-\mathbf{SSIM}(I_t,\hat{I}_t)}{2} + (1-\alpha)||I_t-\hat{I}_t|| \quad (7)$$

where I_t and $\hat{I}_t$ represent the target image and its synthesized image, respectively. However, an issue of scale ambiguity arises in monocular depth estimation through photometric loss, wherein the photometric loss remains constant when both depth and pose are magnified by a certain factor. To address this, additional pose supervision is essential, restoring the results of both depth and pose estimation to the real scale. Given that pose transition matrix $T_{s\to t}$ can be decomposed into a displacement vector $\mathbf{p}$ and an angle vector $\theta = [\theta_1,\theta_2,\theta_3]$ in Euler manner, the supervision for pose transition is structured as Eq. 8.

$$L_p(T,\hat{T}) = \frac{||\mathbf{p}-\hat{\mathbf{p}}||}{||\mathbf{p}||} + \frac{1}{3}\sum_{i=1}^{3}\left[1-\cos\frac{1}{2}\left(\theta_i-\hat{\theta}_i\right)\right] \quad (8)$$

We employ cross-entropy loss to oversee 3D occupancy predictions $\hat{O}\in\mathbb{R}^{X\times Y\times Z}$ derived from voxel distribution, utilizing occupancy labels $O\in\mathbb{R}^{X\times Y\times Z}$ voxelized by depth estimation as Eq. 9. It is important to highlight that the process of occupancy generated through depth is non-differentiable, thus serving solely as a supervisory label rather than being directly optimized as a model prediction output.

$$L_{occ}(\hat{O},O) = -\sum_{x,y,z\in O}\hat{O}(x,y,z)log(O(x,y,z)) \quad (9)$$

We use the training loss L, to consider all loss function mentioned above:

$$L = L_d + \lambda_p L_p + \lambda_{occ} L_{occ} \quad (10)$$

where λ_p and λ_{occ} are hyperparameters, and we set $\lambda_p = \lambda_{occ} = 0.05$.

IV. Experiments

To demonstrate the capabilities of our GenerOcc, we performed a comprehensive set of experiments on both the publicly available Occ3D-nuScenes dataset and our proprietary Dominant dataset, representing fisheye and pinhole camera configurations respectively. Furthermore, we are conducting extensive ablation experiments to gain deeper insights into the behavior and performance of our model in Sec. IV-D.

A. Dataset and Metrics

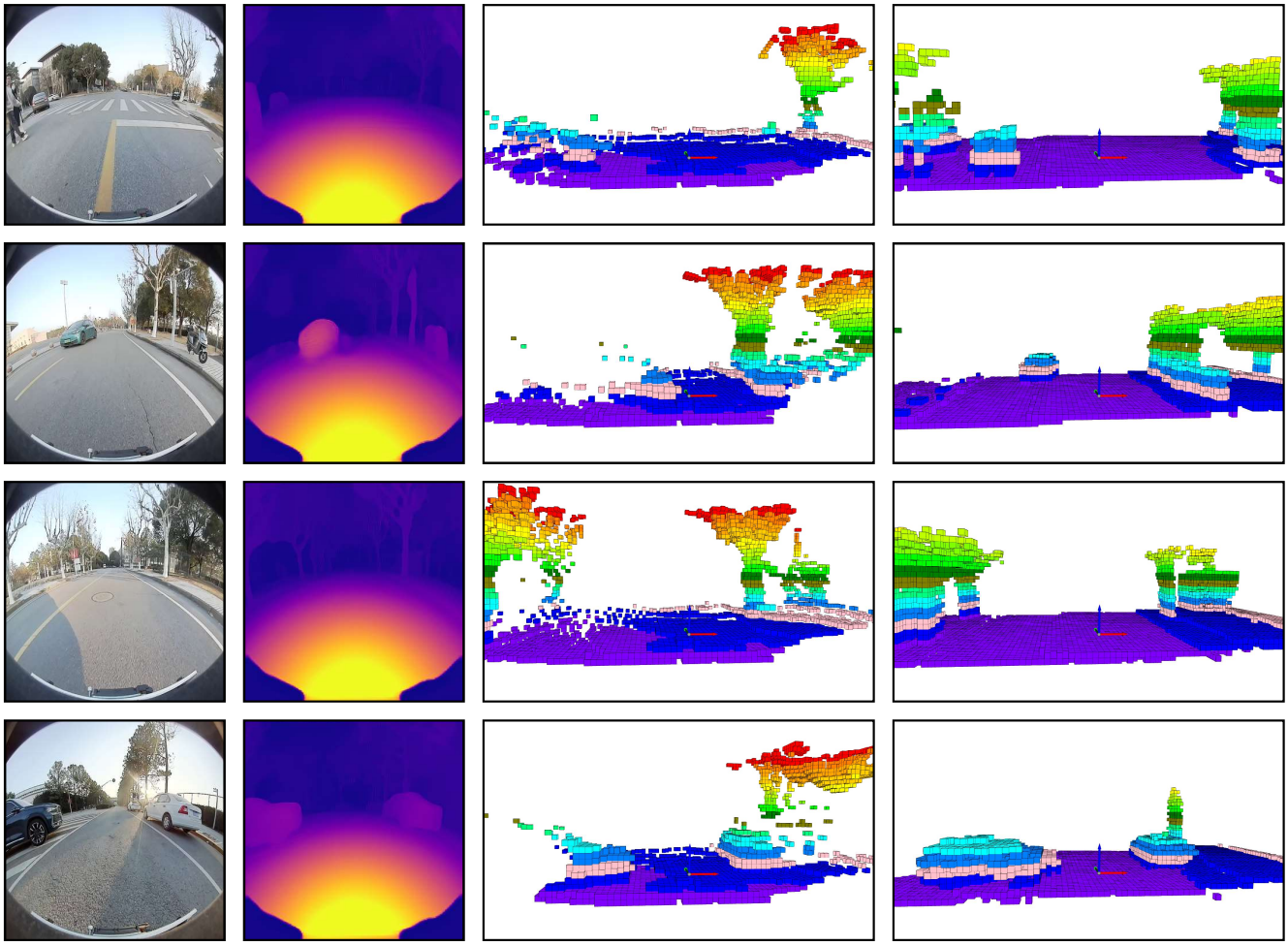
We evaluate our method on Occ3D-nuScenes and Dominant for the same front-view setting. Occ3D-nuScenes comprises over 800 outdoor scenes, with each scene containing approximately 40 images. It provides 3D occupancy ground truth covering a range of [-40m, -40m, -1m, 40m, 40m, 5.4m] at a resolution of 0.4m, generated from a 32-beam LiDAR for each image, with a capture interval of 0.5s. The Dominant dataset consists of over 2000 outdoor images, accompanied by corresponding point clouds captured by a 128-beam LiDAR, with a interval of 0.1s. Notably, on both datasets, 3D occupancy ground truth are solely used for result evaluation and are not involved in the training process. We employed metrics such as IoU, Prec. (precision), and Rec. (recall) for evaluation.

B. Architecture and Implementation Details

We utilize ResNet18 [26] as the image encoder with a MLP as decoder, resize image to 384×384, and use Adam as the optimizer with a learning rate of 2e-4. All experiments are conducted on a NVIDIA A6000 GPU with 48GB of memory. We set the batch size to 4, and train approximately 6 hours. Based on the higher resolution from LiDAR and shorter interval from images, we voxelized the point clouds of Dominant into 3D occupancy with a range of [-10m, 0m, -1m, 10m, 10m, 2.2m] and a resolution of 0.2m. For the sake of monocular setting of our method, it will only take the objects into consideration that fall within the field of view of the front camera on both datasets.

C. Main Results

Results on Occ3D-nuScenes. The experimental results of GenerOcc and some baselines on the Occ3D dataset are outlined in Tab. I. Our GenerOcc utilized a 384 × 384 ray surface for each image, achieving 14.32% IoU with 13.5 FPS (Frames Per Second). Notably, our approach without any 3D supervision attains comparable results with the TPVFormer model trained under LiDAR supervision. At the same time, our inference speed greatly exceeds other methods. It is evident that our model exhibits capability in high-efficiency predicting occupancy within its field of view.



(a) Fisheye Image (b) Depth Estimation (c) GenerOcc Prediction (d) Ground Truth

Fig. 3. **Visualization results for Dominant dataset.** The range of occupancy prediction and ground truth (GT) is [-10m, 0m, -1m, 10m, 10m, 2.2m], with a resolution of 0.2m, where colors differentiate voxel heights. GT are derived from the voxelization of point clouds from LiDAR.

TABLE I

3D OCCUPANCY PREDICTION PERFORMANCE ON OCC3D-NUSCENES DATASET. 'SUP.' COLUMN ILLUSTRATES THE SUPERVISION MECHANISM, WHERE 'C' REPRESENTS CAMERA AND 'L' MEANS LIDAR.

Method	Input	Sup.	IoU↑	FPS↑
VoxelNet	L	L	37.59	1.9
SurroundOcc	C	L	31.50	1.8
TPVFormer	C	L	17.20	1.9
SelfOcc	C	C	**44.33**	1.1
GenerOcc (ours)	C	C	14.32	**13.5**

TABLE II

3D OCCUPANCY PREDICTION PERFORMANCE ON DOMINANT DATASET. GENEROCC* SUPERVISED BY LIDAR. BACKBONE OF GENEROCC IS RESNET18.

Method	Sup.	IoU↑	Prec.↑	Rec.↑	FPS↑
NRS-ResNet18	C	9.51	30.07	10.19	**14.5**
NRS-PackNet	C	16.73	40.36	22.58	8.6
GenerOcc (ours)	C	15.92	46.81	16.12	13.5
GenerOcc* (ours)	L	**22.93**	**52.82**	**32.68**	13.5

Results on Dominant. The training results on the Dominant dataset are presented in Tab. II. GenerOcc get a 15.92% IoU, 46.81% precision and 16.12% recall using solely the images. To benchmark our results, we trained NRS [16], a self-supervised depth estimation model for fisheye cameras, using the same dataset settings as GenerOcc for fair comparison. We employed ResNet18 and PackNet [27] as feature extraction backbone of NRS and voxelized the depth output for assessment. Notably, GenerOcc utilizing ResNet18, demonstrates a performance level that closely matches that of NRS-PackNet, while also standing out for its impressive speed enhancement. As illustrated in Fig. 4, GenerOcc exhibits superior performance in extracting geometric features of local details compared to NRS-PackNet. Moreover, when GenerOcc incorporates ground truth (GT) for supervision during training, it achieves a remarkable IoU of 22.93%. Some outcomes are shown in Fig. 3, where diffenrent colors differentiate voxel heights.

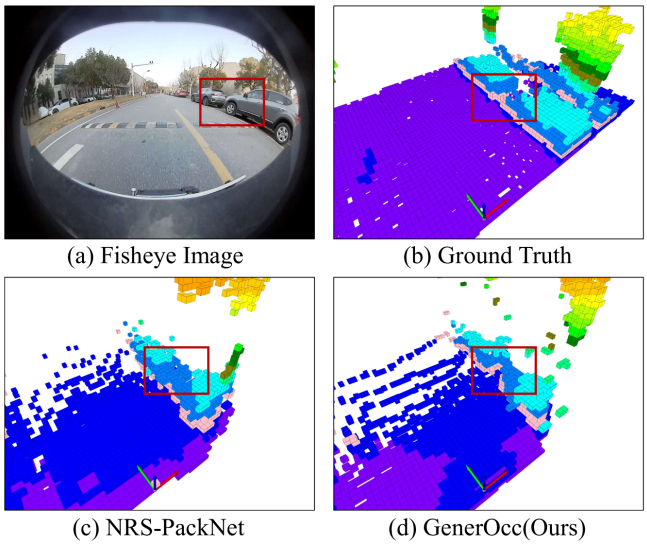


Fig. 4. **Details for Dominant dataset.** GenerOcc can detect the distance between two cars in the red box, while NRS-PackNet recognizes the two cars as a single entity. GenerOcc excels in distinguishing geometric intricacies.

TABLE III

ABLATION STUDY OF RAY-BASED CAMERA MODEL ON DOMINANT.

Method	IoU↑	Prec.↑	Rec.↑	FPS↑
Ray-based Camera Model	**15.92**	**46.81**	**16.12**	13.5
Fixed Intrinsic Matrix	8.37	29.42	10.75	**15.4**

TABLE IV

ABLATION STUDY OF OCCUPANCY LABELS ON DOMINANT. IOU* MEANS IOU OF OCCUPANCY LABEL AND GROUND TRUTH.

Backbone	Sup.	IoU*	IoU↑	Prec.↑	Rec.↑	FPS↑
ResNet18	C	14.65	15.92	46.81	**16.61**	**13.5**
ResNet50	C	**17.59**	17.39	45.72	16.37	12.7
Packnet	C	17.58	**18.42**	**49.07**	13.20	8.5
ResNet18	L	/	22.93	52.82	32.68	**15.4**
ResNet50	L	/	**23.15**	51.46	**32.78**	13.8
Packnet	L	/	23.12	**54.26**	31.43	9.3

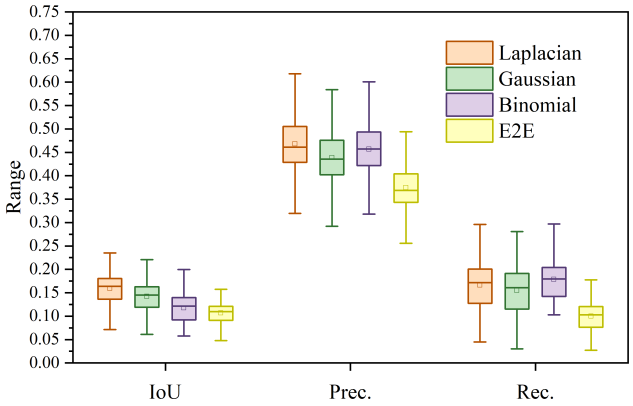


Fig. 5. **Ablation study of probability distribution module on Dominant dataset.** E2E represents directly predicting voxel occupied probability from 2D features instead of parameters for distribution.

D. Ablation Study and Analysis

Influence of Ray-based Camera Model. We employ a fixed camera intrinsic matrix discard the distortion coefficient for projection and voxel predistribution to simulate the elimination of the ray-based camera model. Results shown in Tab. III, indicate that though simplified projection process accelerate the inference speed, it lead to a significant decline in performance of prediction. This verifies the validity of our ray-based model.

Impact of Occupancy Labels. To assess the impact of occupancy labels on output of GenerOcc, we conduct experiments utilizing 3D occupancy labels generated by different backbone through self-supervised depth estimation. As a supplement, we also show the results training with 3D occupancy ground truth as label in Tab. IV, where we calculate IoU of occupancy label and ground truth. We observed that IoU of occupancy prediction is close to or higher than occupancy label in self-supervised setting, indicating that our voxel prediction can improve the depth estimation results to some extent. Moreover, transitioning to a deeper network is not guaranteed yielding a significant progress in evaluation metrics. For example, the precision of ResNet50 and the recall of PackNet did not surpass the performance achieved by ResNet18 in self-supervised manner, and IoU of PackNet was lower than ResNet18 and ResNet50 in supervised manner. This suggests that our method is robust across different backbone architectures, highlighting the lightweight potential of our method for practical deployment and applications.

Parametric Voxel Distribution Module. To explore the affect of the parametric voxel probability distribution module on our method, we increase the output dimension of MLP decoder to predict voxel probability along the ray end-to-end. Furthermore, we substituted the Laplacian distribution with Gaussian and Binomial distributions in our experiments to investigate influence of different probability distribution models on GenerOcc results, with the outcomes detailed in the Fig. 5. Surprisingly, E2E method has the worst performance on all metrics, which means our distribution module has a positive effect. Notably, the Laplacian distribution exhibited higher IoU and precision metrics, albeit with a slightly lower recall compared to the Binomial distributions. Our experiments suggest that utilizing the Laplacian distribution, with its concentrated probability at extreme values, is a worthwhile trade-off, even if it results in a scarcity of voxels within object interiors. This choice enhances the ability of GenerOcc to recognize object edges.

V. CONCLUSION

In this paper, we introduce a self-supervised framework GenerOcc of real-time 3D occupancy prediction, applicable to generic cameras. We collect the fisheye Dominant dataset to validate the effectiveness of our ray-based camera model on any type of camera. By occupancy labels generated from self-supervised depth estimation, our parametric voxel probability distribution module performs efficiently on both public Occ3d-nuScene and our proprietary Dominant dataset.

REFERENCES

[1] W. Tong, C. Sima, T. Wang, L. Chen, S. Wu, H. Deng, Y. Gu, L. Lu, P. Luo, D. Lin *et al.*, "Scene as occupancy," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 8406–8415.

[2] R. Qian, X. Lai, and X. Li, "3d object detection for autonomous driving: A survey," *Pattern Recognition*, vol. 130, p. 108796, 2022.

[3] V. R. Kumar, C. Eising, C. Witt, and S. Yogamani, "Surround-view fisheye camera perception for automated driving: Overview, survey & challenges," *IEEE Transactions on Intelligent Transportation Systems*, 2023.

[4] X. Tian, T. Jiang, L. Yun, Y. Mao, H. Yang, Y. Wang, Y. Wang, and H. Zhao, "Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] R. Cheng, C. Agia, Y. Ren, X. Li, and L. Bingbing, "S3cnet: A sparse semantic scene completion network for lidar point clouds," in *Conference on Robot Learning*. PMLR, 2021, pp. 2148–2161.

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

[7] J. Li, Y. Liu, D. Gong, Q. Shi, X. Yuan, C. Zhao, and I. Reid, "Rgbd based dimensional decomposition residual network for 3d semantic scene completion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 7693–7702.

[8] S.-C. Wu, K. Tateno, N. Navab, and F. Tombari, "Scfusion: Real-time incremental scene reconstruction with semantic completion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 801–810.

[9] J. Zhang, H. Zhao, A. Yao, Y. Chen, L. Zhang, and H. Liao, "Efficient semantic scene completion network with spatial group convolution," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 733–749.

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

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

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

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

[14] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, "Digging into self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3828–3838.

[15] V. R. Kumar, S. A. Hiremath, M. Bach, S. Milz, C. Witt, C. Pinard, S. Yogamani, and P. Mäder, "Fisheyedistancenet: Self-supervised scale-aware distance estimation using monocular fisheye camera for autonomous driving," in *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2020, pp. 574–581.

[16] I. Vasiljevic, V. Guizilini, R. Ambrus, S. Pillai, W. Burgard, G. Shakhnarovich, and A. Gaidon, "Neural ray surfaces for self-supervised learning of depth and ego-motion," in *2020 International Conference on 3D Vision (3DV)*. IEEE, 2020, pp. 1–11.

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

[18] V. R. Kumar, S. Yogamani, S. Milz, and P. Mäder, "Fisheyedistancenet++: Self-supervised fisheye distance estimation with self-attention, robust loss function and camera view generalization," *Electronic Imaging*, vol. 33, pp. 1–11, 2021.

[19] M. Pan, J. Liu, R. Zhang, P. Huang, X. Li, L. Liu, and S. Zhang, "Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision," *arXiv preprint arXiv:2309.09502*, 2023.

[20] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[21] A.-Q. Cao and R. de Charette, "Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 9387–9398.

[22] Y. Huang, W. Zheng, B. Zhang, J. Zhou, and J. Lu, "Selfocc: Self-supervised vision-based 3d occupancy prediction," *arXiv preprint arXiv:2311.12754*, 2023.

[23] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3d u-net: learning dense volumetric segmentation from sparse annotation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 424–432.

[24] J. Philion and S. 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*. Springer, 2020, pp. 194–210.

[25] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[27] V. Guizilini, R. Ambrus, S. Pillai, A. Raventos, and A. Gaidon, "3d packing for self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2485–2494.