OccNeRF: Self-Supervised Multi-Camera Occupancy Prediction with Neural Radiance Fields

Chubin Zhang[1]*, Juncheng Yan[1]*†, Yi Wei[1,2]*,
Jiaxin Li[3], Li Liu[4], Yansong Tang[1]‡, Yueqi Duan[1], Jiwen Lu[1,2]
[1]Tsinghua University [2]Beijing National Research Center for Information Science and Technology
[3]Gaussian Robotics [4]Xiaomi Car
{zhangcb19, yanjc20, y-wei19}@mails.tsinghua.edu.cn;
{lijx1992, liuli.ll9412}@gmail.com;
{tang.yansong@sz., duanyueqi@, lujiwen@}tsinghua.edu.cn

Abstract

As a fundamental task of vision-based perception, 3D occupancy prediction reconstructs 3D structures of surrounding environments. It provides detailed information for autonomous driving planning and navigation. However, most existing methods heavily rely on the LiDAR point clouds to generate occupancy ground truth, which is not available in the vision-based system. In this paper, we propose an OccNeRF method for self-supervised multi-camera occupancy prediction. Different from bounded 3D occupancy labels, we need to consider unbounded scenes with raw image supervision. To solve the issue, we parameterize the reconstructed occupancy fields and reorganize the sampling strategy. The neural rendering is adopted to convert occupancy fields to multi-camera depth maps, supervised by multi-frame photometric consistency. Moreover, for semantic occupancy prediction, we design several strategies to polish the prompts and filter the outputs of a pretrained open-vocabulary 2D segmentation model. Extensive experiments for both self-supervised depth estimation and semantic occupancy prediction tasks on nuScenes dataset demonstrate the effectiveness of our method. Code is available at https://github.com/LinShan-Bin/OccNeRF.

1. Introduction

Recent years have witnessed the great process of autonomous driving [37, 39, 45, 78]. As a crucial component, 3D perception helps the model to understand the real 3D world. Although LiDAR provides a direct means to capture the geometric data, its adoption is hindered by the expense of sensors and the sparsity of scanned points. In contrast, as a cheap while effective solution, the vision-centric methods [1, 27, 40, 45, 85] have received more and more attention. Among various 3D scene understanding tasks, multi-camera 3D object detection [28, 37, 39, 44] plays an important role in autonomous systems. However, it struggles to detect objects from infinite classes and suffer from long-tail problem.

Complementary to 3D object detection, 3D occupancy prediction [8, 67, 73, 86] reconstructs the geometric structure of surrounding scenes, which can facilitate downstream tasks. As mentioned in [78], 3D occupancy is a good 3D representation for multi-camera scene reconstruction since

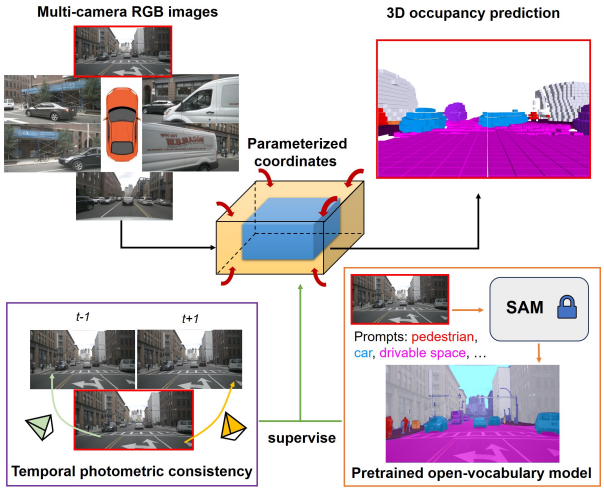


Figure 1. The overview of OccNeRF. To represent unbounded scenes, we propose a parameterized coordinate to contract infinite space to the bounded occupancy fields. Without using any annotated labels, we leverage temporal photometric constraints and pretrained open-vocabulary segmentation models to provide geometric and semantic supervision.

*Equal contribution.
†Work done during Juncheng Yan's internship at Xiaomi Car.
‡Corresponding author.

it has the potential to reconstruct occluded parts and guarantees multi-camera consistency. Recently, some methods have been proposed to lift image features to the 3D space and further predict 3D occupancy. However, most of these methods need 3D supervision. Although some previous works [67, 78] autolabel the occupancy ground truth by accumulating multi-frame LiDAR points, we cannot use LiDAR sensors in vision-centric systems. In other words, we need special vehicles that equip LiDARs to collect data, which is expensive and wastes a large amount of unlabeled multi-camera images. Therefore, it is a valuable direction to explore self-supervised occupancy prediction with only image data.

To address this, we propose an OccNeRF method, which targets at self-supervised multi-camera occupancy prediction. We first utilize a 2D backbone to extract multi-camera 2D features. To save memory, we directly interpolate 2D features to obtain 3D volume features instead of using heavy cross-view attention. In previous works, the volume features are supervised by the bounded occupancy labels (*e.g*. 50m range) and they only need to predict the occupancy with finite resolution (*e.g*. $200 \times 200 \times 16$). Differently, for self-supervised training, we should consider unbounded scenes since the RGB images perceive an infinite range. To this end, we parameterize the occupancy fields to represent unbounded environments. Specifically, we split the whole 3D space into the inside and outside regions. The inside one maintains the original coordinate while the outside one adopts a contracted coordinate. Moreover, A specific sampling strategy is designed to transfer parameterized occupancy fields to multi-camera depth maps with neural rendering.

A straightforward way to supervise predicted occupancy is to calculate loss between rendered images and training images, which is the same as the loss function used in NeRF [49]. Unfortunately, our experimental results show that it does not work well. Instead, we leverage the temporal photometric loss as the supervision signals, which is commonly used in self-supervised depth estimation methods [21, 22, 46, 82, 89]. To better leverage temporal cues, we perform multi-frame photometric constraints. For semantic occupancy, we propose three strategies to map the class names to the prompts, which are fed to a pre-trained open-vocabulary segmentation model [33, 43] to get 2D semantic labels. Then an additional semantic head is employed to render semantic images and supervised by these labels. To verify the effectiveness of our method, we conduct experiments on both self-supervised multi-camera depth estimation and semantic occupancy prediction tasks. Experimental results show that our OccNeRF outperforms other depth estimation methods by a large margin and achieves comparable performance with some fully-supervised occupancy methods on nuScenes [7] dataset.

2. Related Work

2.1. 3D Occupancy Prediction

Due to the significance to the vision-centric autonomous driving systems, more and more researchers begin to focus on 3D occupancy prediction tasks [8, 9, 25, 29, 67, 68, 73, 78, 86, 91]. In the industry community, 3D occupancy is treated as an alternative to LiDAR perception. As one of the pioneering works, MonoScene [8] extracts the voxel features generated by sight projection to reconstruct scenes from a single image. TPVFormer [29] further extends it to multi-camera fashion with tri-perspective view representation. Beyond TPVFormer, SurroundOcc [78] designs a pipeline to generate dense occupancy labels instead of using sparse LiDAR points as the ground truth. In addition, a 2D-3D UNet network with cross-view attention layers is proposed to predict dense occupancy. RenderOcc [54] uses the 2D depth and semantic labels to train the model, reducing the dependence on expensive 3D occupancy annotations. Occ3D [67] builds a 3D occupancy prediction benchmark on nuScenes and Waymo datasets and proposes the CTF-Occ network. Compared with these methods, our method does not need any annotated 3D or 2D labels. Recently, as a preprint work, SimpleOccupancy [17] presents a simple while effective framework for occupancy estimation. Although SimpleOccupancy investigates self-supervised learning, it does not consider infinite range and semantic prediction.

2.2. Neural Radiance Fields

As one of the most popular topics in 3D area, neural radiance fields (NeRF) [6, 10, 14, 31, 36, 41, 47, 50, 52, 64, 66, 74, 80] have made great achievement in recent years. NeRF [49] learns the geometry of the scene by optimizing a continuous volumetric scene function with a set of multi-view images. To obtain the novel views, volume rendering is performed to convert the radiance fields to RGB images. As a follow-up, mip-NeRF [2] represents the scene at a continuously-valued and replaces rays as anti-aliased conical frustums. Beyond mip-NeRF, Zip-NeRF [4] integrates mip-NeRF with a grid-based model for faster training and better quality. There are several extensions of original NeRF, including dynamic scenes [18, 38, 55, 56, 70], model accelerating [53, 72, 76, 81, 83], 3D reconstruction [11, 15, 19, 51, 58, 65], etc. As one of these extensions, some works aim to describe unbounded scenes [3, 84]. NeRF++ [84] split the 3D space as an inner unit sphere and an outer volume and proposes inverted sphere parameterization to represent outside regions. Further, mip-NeRF 360 [3] embeds this idea into mip-NeRF and applies the smooth parameterization to volumes. Inspired by these methods, we also design a parameterization scheme to model the unbounded scene for occupancy prediction.

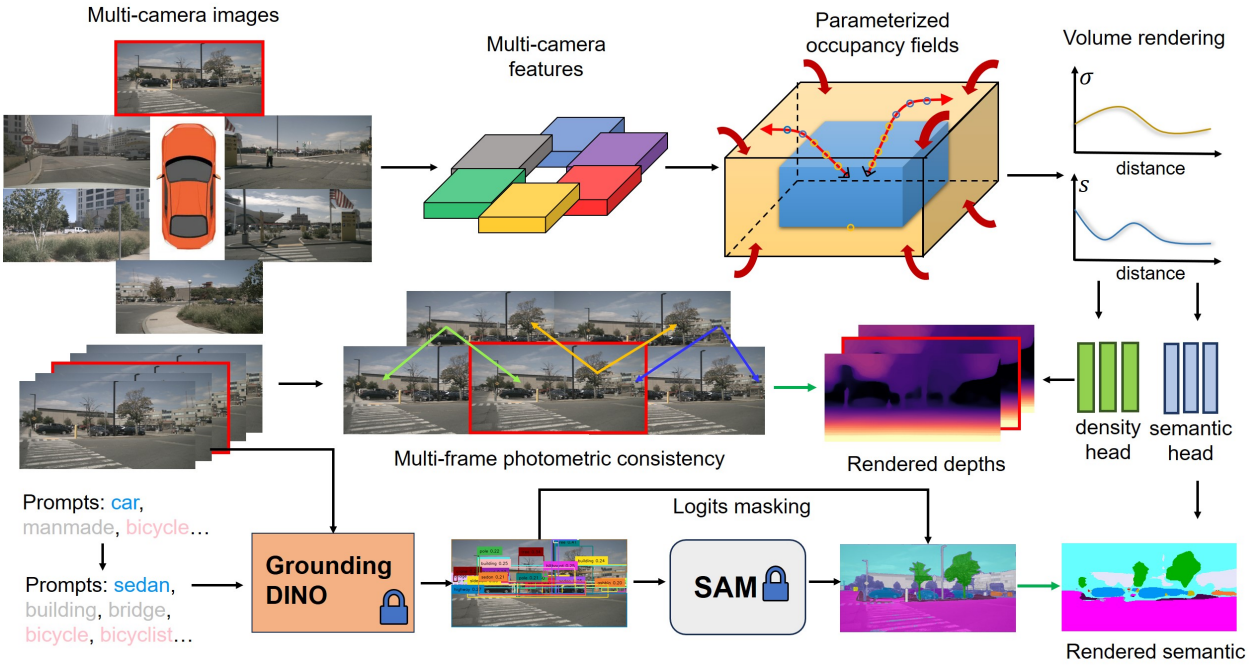


Figure 2. The pipeline of OccNeRF. We first use a 2D backbone to extract multi-camera features, which are lifted to 3D space to get volume features with interpolation. The parameterized occupancy fields are reconstructed to describe unbounded scenes. To obtain the rendered depth and semantic maps, we perform volume rendering with our reorganized sampling strategy. The multi-frame depths are supervised by photometric loss. For semantic prediction, we adopted pretrained Grounded-SAM with prompts cleaning. The green arrow indicates supervision signals.

2.3. Self-supervised Depth Estimation

While early works [16, 35, 42, 59, 87] require dense depth annotations, recent depth estimation methods [5, 12, 22, 34, 46, 57, 61, 62, 69, 71, 75, 82, 88, 90] are designed in a self-supervised manner. Most of these methods predict depth maps and ego-motions simultaneously, adopting the photometric constraints [20, 89] between successive frames as the supervision signal. As a classical work in this field, Monodepth2 [21] proposes some techniques to improve the quality of depth predictions, including the minimum re-projection loss, full-resolution multi-scale sampling, and auto-masking loss. Since modern self-driving vehicles are usually equipped with multiple cameras to capture the full view of the surrounding environment, researchers begin to focus on multi-camera self-supervised depth estimation [23, 32, 60, 63, 77, 79]. FSM [23] is the first work to extend monocular depth estimation to full surrounding views by leveraging spatio-temporal contexts and pose consistency constraints. To predict real-world scale, SurroundDepth [77] uses structure-from-motion to generate scale-aware pseudo depths to pretrain the models. Further, it proposes the cross-view transformer and joint pose estimation to incorporate the multi-camera information. Recently, R3D3 [60] combines the feature correlation with bundle adjustment operators for robust depth and pose estimation. Different from these methods, our approach directly extracts features in 3D space, achieving multi-camera consistency and better reconstruction quality.

3. Approach

3.1. Overview

Figure 2 shows the pipeline of our approach. With the multi-camera images $\{I^i\}_{i=1}^N$ as inputs, we first utilize a 2D backbone to extract N cameras' features $\{X^i\}_{i=1}^N$. Then the 2D features are interpolated to the 3D space to obtain the volume features with known intrinsic $\{K^i\}_{i=1}^N$ and extrinsic $\{T^i\}_{i=1}^N$. As discussed in Section 3.2, to represent the unbounded scenes, we propose a coordinate parameterization to contract the infinite range to a limited occupancy field. The volume rendering is performed to convert occupancy fields to multi-frame depth maps, which are supervised by photometric loss. Section 3.3 introduces this part in detail. Finally, Section 3.4 shows how we use a pretrained open-vocabulary segmentation model to get 2D semantic labels.

3.2. Parameterized Occupancy Fields

Different from previous works [78, 86], we need to consider unbounded scenes in the self-supervised setting. On the one hand, we should preserve high resolution for the inside re-

gion (*e.g.* [-40m, -40m, -1m, 40m, 40m, 5.4m]), since this part covers most regions of interest. On the other hand, the outside region is necessary but less informative and should be represented within a contracted space to reduce memory consumption. Inspired by [3], we propose a transformation function with adjustable regions of interest and contraction threshold to parameterize the coordinates $r = (x, y, z)$ of each voxel grid:

$$f(r) = \begin{cases} \alpha \cdot r' & |r'| \leq 1 \\ \frac{r'}{|r'|} \cdot \left(1 - \frac{(1-\alpha)^2}{\alpha|r'|-2\alpha+1}\right) & |r'| > 1 \end{cases}, \quad (1)$$

where $r' = r/r_b$ is the normalized coordinate of the input r and $f(r) \in (-1, 1)$ indicates the normalized parameterized coordinate. r_b is the bound of the inside region, which is different for x, y, z direction. $\alpha \in [0, 1]$ represents the proportion of the region of interest in the parameterized space. Higher α indicates we use more space to describe the inside region. Note that the two functions in Equation 1 have the same value and gradient at $r = r_b$. Please refer to the supplementary material for the derivation details.

To obtain 3D voxel features from 2D views, we first generate the corresponding points $\mathcal{P}_{pc} = [\boldsymbol{x}_{pc}, \boldsymbol{y}_{pc}, \boldsymbol{z}_{pc}]^T$ for each voxel in the parameterized coordinate system and map them back to the ego coordinate system:

$$\mathcal{P} = [f_x^{-1}(\boldsymbol{x}_{pc}), f_y^{-1}(\boldsymbol{y}_{pc}), f_z^{-1}(\boldsymbol{x}_{pc})]^T. \quad (2)$$

Then we project these points to the 2D image feature planes and use bilinear interpolation to get the 2D features:

$$\mathcal{F}^i = X^i \left\langle \text{proj}(\mathcal{P}, T^i, K^i) \right\rangle. \quad (3)$$

where proj is the function projecting 3D points $\mathcal{P}$ to the 2D image plane defined by the camera extrinsic T^i and intrinsic K^i, $\langle\rangle$ is the bilinear interpolation operator, $\mathcal{F}^i$ is the interpolation result. To simplify the aggregation process and reduce computation costs, we directly average the multi-camera 2D features to get volume features, which is the same as the method used in [17, 24]. Finally, a 3D convolution network is employed to extract features and predict final occupancy outputs.

3.3. Multi-frame Depth Estimation

To project the occupancy fields to multi-camera depth maps, we adopt volume rendering [48], which is widely used in NeRF-based methods [2, 49, 84]. To render the depth value of a given pixel, we cast a ray from the camera center $\mathbf{o}$ along the direction $\mathbf{d}$ pointing to the pixel. The ray is represented by $\mathbf{v}(t) = \mathbf{o} + t\mathbf{d}, t \in [t_n, t_f]$. Then, we sample L points $\{t_k\}_{k=1}^L$ along the ray in 3D space to get the density $\sigma(t_k)$. For the selected L quadrature points, the depth of the corresponding pixel is computed by:

$$D(\mathbf{v}) = \sum_{k=1}^{L} T(t_k)(1 - \exp(-\sigma(t_k)\delta_k))t_k, \quad (4)$$

where $T(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_k)\delta_k\right)$, and $\delta_k = t_{k+1} - t_k$ are intervals between sampled points.

A vital problem here is how to sample $\{t_k\}_{k=1}^L$ in our proposed coordinate system. Uniform sampling in the depth space or disparity space will result in an unbalanced series of points in either the outside or inside region of our parameterized grid, which is to the detriment of the optimization process. With the assumption that o is around the coordinate system's origin, we directly sample $L(\mathbf{r})$ points from $U[0, 1]$ in parameterized coordinate and use the inverse function of Equation 1 to calculate the $\{t_k\}_{k=1}^{L(\mathbf{v})}$ in the ego coordinates. The specific $L(\mathbf{v})$ and $r_b(\mathbf{v})$ for a ray are calculated by:

$$\begin{aligned} r_b(\mathbf{v}) &= \frac{\sqrt{(\mathbf{d} \cdot \mathbf{i}l_x)^2 + (\mathbf{d} \cdot \mathbf{j}l_y)^2 + (\mathbf{d} \cdot \mathbf{k}l_z)^2}}{2\|\mathbf{d}\|}, \\ L(\mathbf{v}) &= \frac{2r_b(\mathbf{v})}{\alpha d_v} \end{aligned} \quad (5)$$

where $\mathbf{i}, \mathbf{j}, \mathbf{k}$ are the unit vectors in the x, y, z directions, l_x, l_y, l_z are the lengths of the inside region, and d_v is the voxel size. To better adapt to the occupancy representation, we directly predict the rendering weight instead of the density.

A conventional supervision method is to calculate the difference between rendered RGB images and raw RGB images, which is employed in NeRF [49]. However, our experimental results show that it does not work well. The possible reason is that the large-scale scene and few view supervision is difficult for NeRF to converge. To better make use of temporal information, we employ the photometric loss proposed in [21, 89]. Specifically, we project adjacent frames to the current frames according to the rendered depths and given relative poses. Then we calculate the reconstruction error between projected images and raw images:

$$\mathcal{L}_{pe}^i = \frac{\beta}{2}(1 - \text{SSIM}(I^i, \hat{I}^i)) + (1 - \beta)\|I^i, \hat{I}^i\|_1, \quad (6)$$

where $\hat{I}^i$ is the projected image and $\beta = 0.85$. Moreover, we adopt the techniques introduced in [21], *i.e.* per-pixel minimum reprojection loss and auto-masking stationary pixels. For each camera view, we render a short sequence instead of a single frame and perform multi-frame photometric loss.

3.4. Open-vocabulary Semantic Supervision

2D semantic labels of multi-camera images provide pixel-level semantic supervision for semantic 3D occupancy prediction, which helps the network capture geometry consistency and spatial relationships among voxels. To obtain 2D labels, previous works [54] project 3D LiDAR points

Method	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
FSM [23]	0.297	-	-	-	-	-	-
FSM* [23]	0.319	7.534	7.860	0.362	0.716	0.874	0.931
SurroundDepth [77]	0.280	4.401	7.467	0.364	0.661	0.844	0.917
Kim *et al*. [32]	0.289	5.718	7.551	0.348	0.709	0.876	0.932
R3D3 [60]	0.253	4.759	7.150	-	0.729	-	-
SimpleOcc [17]	0.224	3.383	7.165	0.333	0.753	0.877	0.930
OccNeRF	**0.202**	**2.883**	**6.697**	**0.319**	**0.768**	**0.882**	**0.931**

Table 1. Comparisons for self-supervised multi-camera depth estimation on the nuScenes dataset [7]. The results are averaged over all views without median-scaling at test time. 'FSM*' is the reproduced result in [32] .

(a) RGB image (b) LiDAR projection labels

(c) Our Grounding DINO bboxs (d) Our Grounded-SAM labels

Figure 3. Detection bounding boxes generated by our Grounding DINO and semantic labels predicted by SAM in our method exhibit precision, which is comparable with that of LiDAR points projection labels.

with segmentation labels to the image space to avoid the expensive cost of annotating dense 3D occupancy. However, we aim to predict semantic occupancy in a fully vision-centric system and use 2D data only. To this end, we leverage a pretrained open-vocabulary model Grounded-SAM [30, 33, 43] to generate 2D semantic segmentation labels. Without any 2D or 3D ground truth data, the pretrained open-vocabulary model enables us to obtain 2D labels which closely match the semantics of the given category names. This method can easily extend to any dataset, making our approach efficient and generalizable.

Specifically, when dealing with $\boldsymbol{c}$ categories, we employ three strategies to determine the prompts provided to the Grounding DINO. These strategies consist of synonymous substitution, where we replace words with their synonyms (e.g., changing 'car' to 'sedan' to enable the model to distinguish it from 'truck' and 'bus'); splitting single words into multiple entities (e.g., 'manmade' is divided into 'building', 'billboard', and 'bridge' etc. to enhance differentiation); and incorporating additional information (e.g., introducing 'bicyclist' to facilitate the detection of a person on a bike). Subsequently, we obtain detection bounding boxes along with their corresponding logits and phrases, which are fed to SAM [33] to generate M precise segmentation binary masks. After multiplying the Grounding DINO logits with binary masks, every pixel has $\{l_i\}_{i=1}^{M}$ logits. We get the per-pixel label $\mathcal{S}^{pix}$ using:

$$\mathcal{S}^{pix} = \psi(\arg\max_i l_i), \tag{7}$$

where $\psi(\cdot)$ is a function that maps the index of l_i to the category label according to the phrases. If a pixel does not belong to any categories and gets M zero logits, we will give it an 'uncertain' label. The generated detection bounding boxes and semantic labels are shown in Figure 3.

To leverage the 2D semantic supervision, we initially utilize a semantic head with $\boldsymbol{c}$ output channels to map volume features extracted to semantic outputs, denoted as $S(x)$. Similar to the method outlined in Section 3.3, we engage in volume rendering once more using the subsequent equation:

$$\hat{\mathcal{S}}^{pix}(\mathbf{r}) = \sum_{k=1}^{L_s} T(t_k)(1 - \exp(-\sigma(t_k)\delta_k))S(t_k), \tag{8}$$

where $\hat{\mathcal{S}}^{pix}$ represents the per-pixel semantic rendering output. To save the memory and improve efficiency, we do not render the pixels that are assigned with 'uncertain' labels. Moreover, we only render the central frame instead of multiple frames and reduce the sample ratio to $L_s = L/4$. Our overall loss function is expressed as:

$$\mathcal{L}_{total} = \sum_i \mathcal{L}^i_{pe} + \lambda \mathcal{L}^i_{sem}(\hat{\mathcal{S}}^{pix}, \mathcal{S}^{pix}) \tag{9}$$

where $\mathcal{L}_{sem}$ is the cross-entropy loss function and λ is the semantic loss weight.

4. Experiments

4.1. Experimental Setup

Dataset: We conduct experiments on the popular large-scale autonomous driving dataset nuScenes [7], which contains 600 scenes for training, 150 scenes for validation, and 150 for testing. The dataset has about 40000 frames and 17 classes in total. For self-supervised depth estimation, we project LiDAR point clouds to each view to get depth

Method	GT	mIoU	mIoU*	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	sidewalk	terrain	manmade	vegetation
MonoScene [8]	✓	6.33	6.06	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	7.92	7.43	1.01	7.65
BEVDet [28]	✓	20.03	19.38	30.31	0.23	32.26	34.47	12.97	10.34	10.36	6.26	8.93	23.65	52.27	26.06	22.31	15.04	15.10
BEVFormer [39]	✓	24.64	23.67	38.79	9.98	34.41	41.09	13.24	16.50	18.15	17.83	18.66	27.70	48.95	29.08	25.38	15.41	14.46
OccFormer [86]	✓	22.39	21.93	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	34.66	22.73	6.76	6.97
RenderOcc [54]	✓	24.53	23.93	27.56	14.36	19.91	20.56	11.96	12.42	12.14	14.34	20.81	18.94	68.85	42.01	43.94	17.36	22.61
TPVFormer [29]	✓	28.69	27.83	38.90	13.67	40.78	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	37.55	30.70	19.40	16.78
CTF-Occ [67]	✓	29.54	28.53	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	37.98	33.23	20.79	18.00
SimpleOcc [17]	×	7.99	-	0.67	1.18	3.21	7.63	1.02	0.26	1.80	0.26	1.07	2.81	40.44	18.30	17.01	13.42	10.84
OccNeRF	×	10.81	-	0.83	0.82	5.13	12.49	3.50	0.23	3.10	1.84	0.52	3.90	52.62	20.81	24.75	18.45	13.19

Table 2. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset.** Since 'other' and 'other flat' classes are the invalid prompts for open-vocabulary models, we do not consider these two classes during evaluation. 'mIoU*' is the original result and 'mIoU' is the result ignoring the classes.

ground truth for evaluation. Following SurroundDepth [77], we clip the depth prediction and ground truth from 0.1m to 80m. To evaluate the semantic occupancy prediction, we use Occ3D-nuScenes [67] benchmark. The range of each sample is [-40m, -40m, -1m, 40m, 40m, 5.4m] and the voxel size is 0.4m. Among 17 classes, we do not consider 'other' and 'other flat' classes for evaluation since open-vocabulary models cannot recognize the semantic-ambiguous text. Following [67, 77], we evaluate models on validation sets.

Implementation Details: We adopt ResNet-101 [26] with ImageNet [13] pretrained weights as the 2D backbone to extract multi-camera features. The resolution of input images and rendered depth maps are set as 336x672 and 180x320 respectively. The predicted occupancy field has the shape 300x300x24. The central 200x200x16 voxels represent inside regions: -40m to 40m for X and Y axis, and -1m to 5.4m for the Z axis, which is the same as the scope defined in Occ3D-nuScenes. We render 3 frames depth maps, which are supervised by the photometric loss with a sequence of 5 frames raw images (1 keyframe with 4 neighbored non-key frames). The α is set as 0.667. To predict semantic occupancy, the Grounded-SAM [33, 43] is employed as our pretrained open-vocabulary model. The text and box thresholds are set as 0.2 and we use the loss weight $\lambda = 0.05$. All experiments are conducted on 8 A100.

Evaluation Metric: For depth estimation, we use the commonly used depth evaluation metrics [21, 77, 89]: Abs Rel, Sq Rel, RMSE, RMSE log and $\delta < t$. The Abs Rel is the main metric and see supplementary for the details of these metrics. During evaluation, we do not perform median scaling since our method can predict real-world scale. For semantic occupancy prediction, we use the mean intersection over union (mIoU) of all classes for evaluation. Following the evaluation tool in Occ3D-nuScenes, the evaluation is only performed on the 'observed' voxels in camera views.

4.2. Self-supervised Depth Estimation

Table 1 shows the self-supervised multi-camera depth estimation results on nuScenes dataset. We do not use pretrained segmentation model in this experiment. The results are averaged over 6 cameras and 'FSM*' is the reproduced FSM [23] result reported in [32]. We can see that our method outperforms other state-of-the-art methods by a large margin, demonstrating the effectiveness of OccNeRF. Compared with depth estimation methods, our method directly predicts occupancy in 3D space, naturally guaranteeing multi-camera consistency. Moreover, we do not need to lift 2D depth maps to 3D point clouds with post-processing.

4.3. Semantic Occupancy Prediction

We conduct semantic occupancy prediction experiments on the Occ3D-nuScenes dataset. Since the pretrained open-vocabulary model [33, 43] cannot recognize ambiguous prompts such as 'other' and 'other flat', we remove these two classes during evaluation. For another self-supervised method SimpleOcc [17], we use the same 2D semantic labels from the pretrained model for fair comparison. As shown in Table 2, our method outperforms SimpleOcc by a large margin and even gets comparable performance with some full-supervised methods. For some classes, such as 'drivable space' and 'manmade', our method surprisingly surpasses all supervised methods. However, we note that for some small object categories (*e.g.* bicycle and pedestrian), the gap between our method and state-of-the-art supervised methods is large. The possible reason is that the current open-vocabulary model often misses small objects and fails

Depth	Multi	Abs Rel	RMSE	$\delta < 1.25$
		0.627	15.901	0.051
	✓	0.489	9.352	0.362
✓		0.216	6.752	0.764
✓	✓	**0.202**	**6.697**	**0.768**

Table 3. The ablation study of supervision method. 'Depth' means whether we use the temporal photometric constraints to train the model. If not, we directly utilize the supervision method in NeRF [49]. 'Multi' indicates whether we employ multi-frame rendering and supervision.

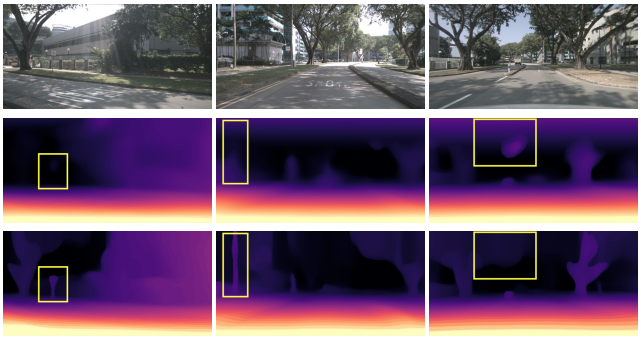


Figure 4. Qualitative comparison of different coordinates. The second line indicates the results without using coordinate parameterization. With the ability to represent unbounded environments, our method can get better results in far scenes, such as the sky.

to provide strong supervision.

4.4. Ablation Study

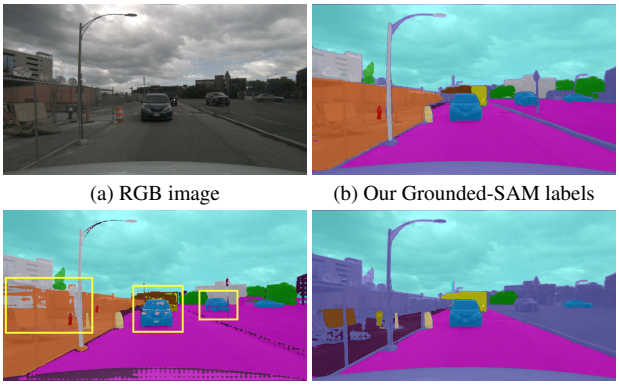
Supervision Method: A straightforward supervision signal is the difference between the rendered and true pixel colors, which is the same as the loss function used in NeRF [49]. However, as shown in Table 3, this supervision method yields terrible performance. We attribute this to the challenge NeRF faces in learning the scene structure with only six views. On the contrary, temporal photometric loss can better leverage geometric cues in adjacent frames, which is the golden metric in self-supervised depth estimation methods. Moreover, multi-frame training provides stronger supervision, further boosting the model's performance.

Coordinate Parameterization: Table 4 shows the ablation study of coordinate parameterization. Different from occupancy labels, the photometric loss assumes that the images perceive an infinite range. The aim of the contracted coordinate is to represent the unbounded scene in a bounded occupancy. From the table, we can see that the contracted coordinate greatly improves the model's performance. In addition, since the parameterized coordinate is not the Euclidean 3D space, the proposed sampling strategy works better than normal uniform sampling in the original ego coordinate.

Semantic Label Generation: In this subsection, we conduct ablation studies of semantic label generation on the

CC	Resample	Abs Rel	Sq Rel	$\delta < 1.25$
		0.216	8.465	0.694
✓		0.208	7.339	0.743
✓	✓	**0.202**	**6.697**	**0.768**

Table 4. The ablation study of coordinate parameterization. 'CC' means that whether we adopt contracted coordinates. 'Resample' indicates whether we leverage the proposed sampling strategy.



(a) RGB image (b) Our Grounded-SAM labels

(c) Labels with SAM logits (d) Labels w/o prompting strategies

Figure 5. Comparison of different semantic label generation methods. Compared with generating semantic labels with SAM logits or feeding raw category names, our semantic labels are preciser and have better continuity.

Occ3d-nuScenes dataset. First, we change grounding DINO [43] logits as SAM logits [33] to get semantic labels. As shown in Table 5 and Figure 5, we find that the SAM logits are more noisy and discontinuous. Then, we also feed raw category names to the open-vocabulary model without proposed prompting strategies. However, this method leads to worse results since the original class names cannot provide fine-grained semantic guidance and bring ambiguity.

Method	mIoU
SAM logits	7.50
category names	8.23
Ours	10.81

Table 5. The ablation study of semantic label generation. 'SAM logits' means that we directly use the logits from SAM [33]. 'Catagory names' means that we feed raw catagory names to the pretrained open-vocabulary segmentation model and do not adopt any prompting strategy.

4.5. Visualization

To further demonstrate the superiority of our method, we provide some qualitative results in Figure 6 and 7. From Figure 6, we can see that our method can generate high-quality depth maps and occupancy with fine-grained details.

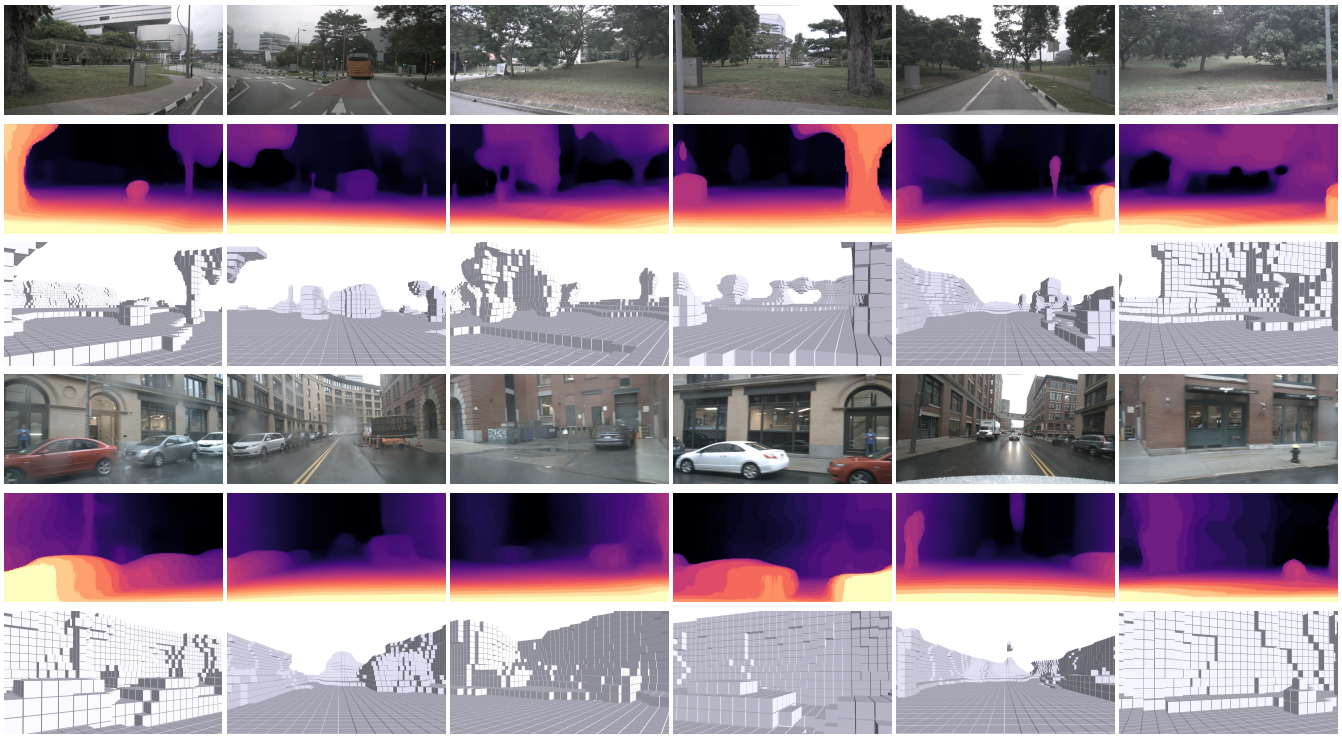


Figure 6. Qualitative results on nuScenes dataset [7] . Our method can predict visually appealing depth maps with texture details and fine-grained occupancy. **Better viewed when zoomed in.**

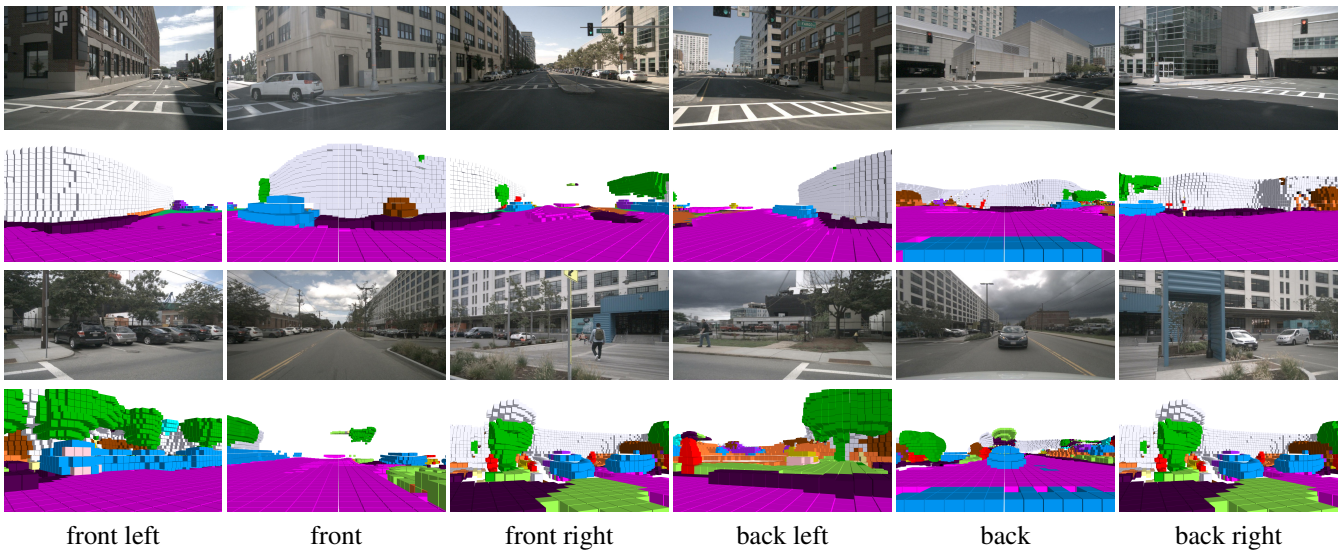


Figure 7. Qualitative results of semantic occupancy on nuScenes dataset [7]. Our method can predict visually appealing semantic occupancy with well geometry correspondence. **Better viewed when zoomed in.**

See supplementary material for more qualitative comparisons with other methods. For semantic occupancy prediction, as shown in Figure 7, our OccNeRF can reconstruct dense results of the surrounding scenes, especially for the large-area categories, such as 'drivable space' and 'man-made'.

5. Limitations and Future Works

During inference, we investigate single-frame occupancy prediction and do not consider multi-frame information as inputs. Therefore, our method is unable to predict occupancy flow. For future work, we will try to use a pretrained optical flow model to supervise the occupancy flow and

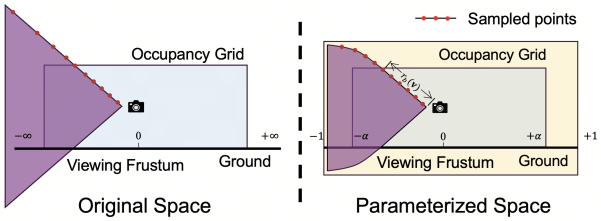


Figure 8. Comparison between original space and parameterized space. The original space utilizes the conventional Euclidean space, emphasizing linear mapping. The parameterized space is divided into two parts: an inner space with linear mapping to preserve high-resolution detail and an outer space where point distribution is scaled inversely with distance, facilitating the representation of an infinite range within a finite spatial domain.

adopt multi-frame multi-camera images as inputs. Further, another limitation is that the performance of our method is bounded by the outputs of open-vocabulary segmentation models, which often neglect small objects.

6. Conclusion

In this paper, we propose OccNeRF for self-supervised multi-camera 3D occupancy prediction. To tackle the unbounded scenes, we propose the parameterized occupancy fields to contract the infinite space to a bounded voxel. To leverage temporal photometric loss, volume rendering is performed on parameterized coordinates to obtain multi-frame multi-camera depth maps. For the semantic occupancy prediction, we utilize an open-vocabulary model to get 2D semantic pseudo labels with the proposed prompt cleaning strategies. The experimental results on nuScenes dataset demonstrate the effectiveness of our method.

Appendix

A. Parameterized Occupancy Fields Derivation

The objective of utilizing parameterized coordinates is to encapsulate an infinite range within a confined spatial domain. This concept is illustrated in Figure 8, where the spatial domain is bifurcated into two distinct regions: the inner space and the outer space. The inner space retains a linear mapping to ensure the preservation of high-resolution details. Conversely, in the outer space, point distribution is executed in proportion to disparity, which inversely relates to distance. Consequently, the transformation function is articulated as follows:

$$f(r) = \begin{cases} \alpha \cdot r' & |r'| \leq 1 \\ \frac{r'}{|r'|} \cdot \left(1 - \frac{a}{|r'|+b}\right) & |r'| > 1 \end{cases}, \tag{10}$$

Original labels	Ours
car	sedan
bicycle	bicycle bicyclist
vegetation	tree
motorcycle	motorcycle motorcyclist
drivable surface	highway
traffic cone	cone
construction vehicle	crane
manmade	building compound bridge pole billboard light ashbin

Table 6. Implementation details of prompt strategy.

where $r' = r/r_b$ denotes the normalized coordinate based on the input r. The parameters a and b are introduced to maintain the continuity of the first derivative. The determination of these parameters is achieved through the resolution of the ensuing equations:

$$\begin{cases} \lim\limits_{r \to r_b^+} f(r) = \lim\limits_{r \to r_b^-} f(r) \\ \lim\limits_{r \to r_b^+} f'(r) = \lim\limits_{r \to r_b^-} f'(r) \end{cases}, \tag{11}$$

The derived solutions are presented as:

$$\begin{cases} a = \frac{(1-\alpha)^2}{\alpha} \\ b = \frac{1-2\alpha}{\alpha} \end{cases}. \tag{12}$$

B. Semantic Label Generation Details

In Section 3.4 of main paper, we present a concise overview of the generation of semantic labels using our open-vocabulary model, Grounded-SAM [30, 33, 43]. We employ three prompt strategies to manually refine the category names fed into Grounding DINO [43]. To illustrate these strategies, we focus on the semantic labels for occupancy in the Occ3D-nuScenes [67] benchmark and provide a detailed explanation.

Specifically, for the synonymous substitution strategy, we substitute 'car' with 'sedan' to enhance model discrimination, replace 'vegetation' with 'tree' to improve detection rates, and change 'driveable surface' to 'highway' to aid the model in distinguishing it from 'sidewalk'. In the case of splitting the word strategy, we change 'manmade' to 'building', 'compound', 'bridge', 'pole', 'billboard', 'light', and 'ashbin', among others. Moreover, we employ the incorporating additional information strategy, introducing prompts such as 'bicyclist', 'motorcyclist', and 'barricade'. Finally, we modify 'traffic cone' to 'cone' and 'construction vehicle' to 'crane' due to the bad performance of Grounding DINO [43] when processing original phrases. Readers can refer to Table 6 for our prompts replacement. For hyperparameter, we set both 'BOX_TRESHOLD' and 'TEXT_TRESHOLD' of Grounding DINO [43] to 0.20. We find that the open-vocabulary model has difficulty to

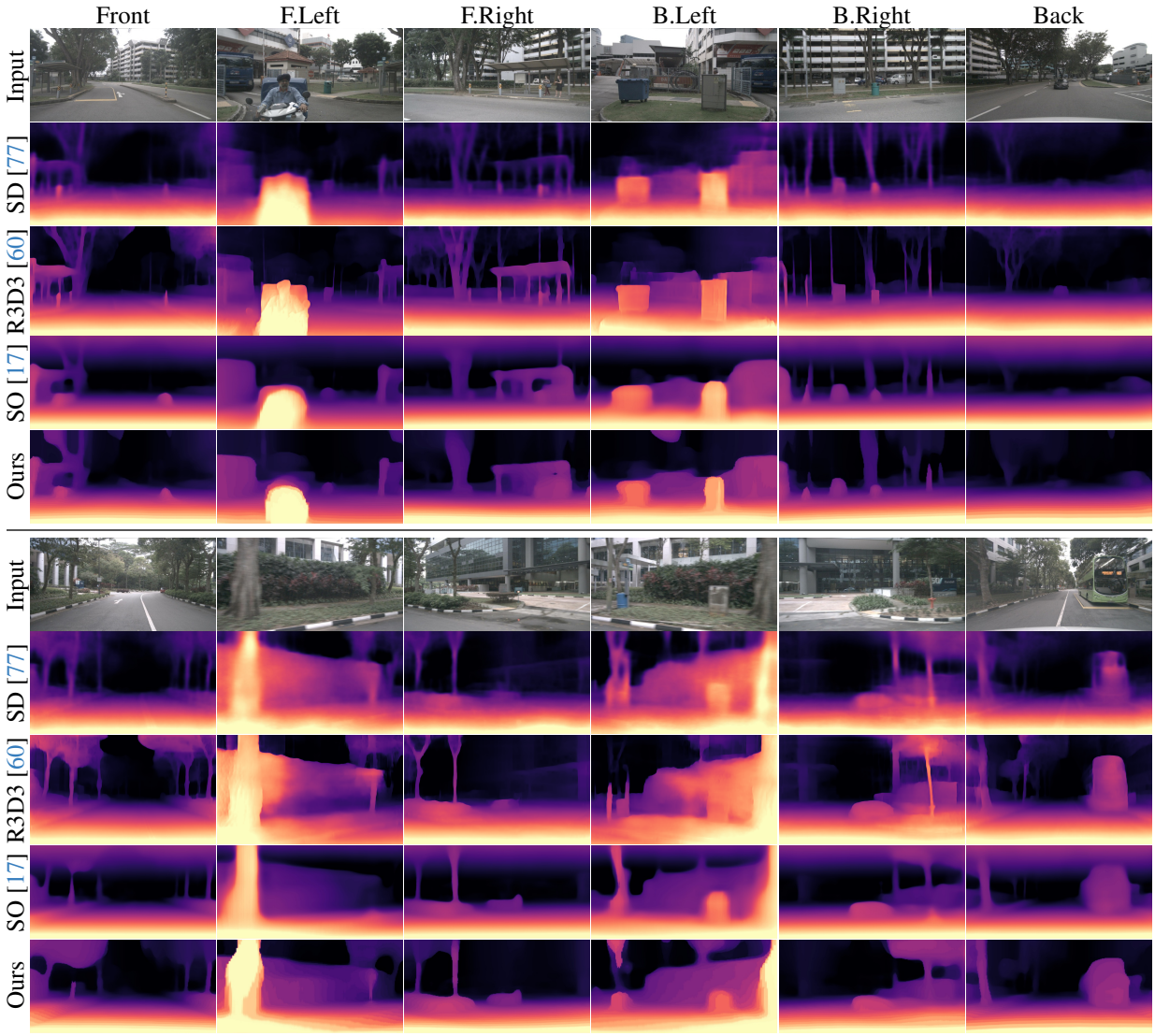


Figure 9. Qualitative comparison on the nuScenes dataset. **Better viewed when zoomed in.**

deal with long prompts. Thus, we further organize our refined prompts into several groups instead of feeding them together, allowing our open-vocabulary model to generate detection frames for each group sequentially.

C. Evaluation Metrics

The detailed evaluation metrics of self-supervised depth estimation can be described as follows:

- Abs Rel: $\frac{1}{|T|}\sum_{d\in T}|d-d^*|/d^*$,
- Sq Rel: $\frac{1}{|T|}\sum_{d\in T}|d-d^*|^2/d^*$,
- RMSE: $\sqrt{\frac{1}{|T|}\sum_{d\in T}|d-d^*|^2}$,
- RMSE log: $\sqrt{\frac{1}{|T|}\sum_{d\in T}|\log d-\log d^*|^2}$,
- $\delta < t$: % of d s.t. $\max(\frac{d}{d^*},\frac{d^*}{d})=\delta<t$,

where d and d^* indicate predicted and ground truth depths respectively, and T indicates all pixels on the depth image D. In our experiments, all the predicted depth maps are scale-aware and we do not perform any scale alignment.

D. More Experimental Results

Per-camera evaluation: We give the per-camera comparisons of our method with previous works on the nuScenes [7] dataset in Table 7. Our method outperforms other methods across all cameras, with a particularly high improvement in side views.

Qualitative Comparisons: Figure 9 shows qualitative comparisons on nuScenes [7] validation set. We visualize several state-of-the-art depth estimation and occupancy prediction methods' results with their official codes. Compared with the these methods, our occupancy-based method has

Method	Abs Rel ↓					
	Front	*F.Left*	*F.Right*	*B.Left*	*B.Right*	*Back*
FSM [23]	0.186	0.287	0.375	0.296	0.418	0.221
SurroundDepth [77]	0.179	0.260	0.340	0.282	0.403	0.212
R3D3 [60]	0.174	0.230	0.302	0.249	0.360	0.201
Ours	**0.132**	**0.190**	**0.227**	**0.204**	**0.289**	**0.169**

Table 7. Per-camera comparisons for scale-aware multi-camera depth estimation on the nuScenes dataset. Tests are conducted within 80 meters.

fewer artifacts and better overall accuracy.

References

[1] Adil Kaan Akan and Fatma Güney. Stretchbev: Stretching future instance prediction spatially and temporally. *arXiv preprint arXiv:2203.13641*, 2022. 1

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 2, 4

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, pages 5470–5479, 2022. 2, 4

[4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. In *ICCV*, 2023. 2

[5] Jia-Wang Bian, Zhichao Li, Naiyan Wang, Huangying Zhan, Chunhua Shen, Ming-Ming Cheng, and Ian Reid. Unsupervised Scale-consistent Depth and Ego-motion Learning from Monocular Video. In *NeurIPS*, pages 35–45, 2019. 3

[6] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, pages 12684–12694, 2021. 2

[7] 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, 5, 8, 10

[8] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 1, 2, 6

[9] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *ICCV*, pages 9387–9398, 2023. 2

[10] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021. 2

[11] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, pages 333–350. Springer, 2022. 2

[12] Yuhua Chen, Cordelia Schmid, and Cristian Sminchisescu. Self-supervised learning with geometric constraints in monocular video: Connecting flow, depth, and camera. In *ICCV*, pages 7063–7072, 2019. 3

[13] 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, 2009. 6

[14] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2022. 2

[15] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 2

[16] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *CVPR*, pages 2002–2011, 2018. 3

[17] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *arXiv preprint arXiv:2303.10076*, 2023. 2, 4, 5, 6, 10

[18] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pages 5712–5721, 2021. 2

[19] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *ICCV*, pages 14346–14355, 2021. 2

[20] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 3

[21] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *ICCV*, pages 3828–3838, 2019. 2, 3, 4, 6

[22] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *CVPR*, pages 2485–2494, 2020. 2, 3

[23] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *RAL*, 7(2):5397–5404, 2022. 3, 5, 6, 11

[24] Adam W. Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-BEV: What really matters for multi-sensor bev perception? In *IEEE International Conference on Robotics and Automation (ICRA)*, 2023. 4

[25] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. *arXiv preprint arXiv:2310.07522*, 2023. 2

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 6

[27] 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. 1

[28] Junjie Huang, Guan Huang, Zheng Zhu, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection

in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 1, 6

[29] 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. 2, 6

[30] IDEA-Research. Grounded segment anything. https://github.com/IDEA-Research/Grounded-Segment-Anything. 5, 9

[31] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *TOG*, 42(4):1–14, 2023. 2

[32] Jung-Hee Kim, Junhwa Hur, Tien Phuoc Nguyen, and Seong-Gyun Jeong. Self-supervised surround-view depth estimation with volumetric feature fusion. *NeurIPS*, 35:4032–4045, 2022. 3, 5, 6

[33] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, 2023. 2, 5, 6, 7, 9

[34] Maria Klodt and Andrea Vedaldi. Supervising the new with the old: learning SFM from SFM. In *ECCV*, pages 698–713, 2018. 3

[35] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019. 3

[36] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2

[37] 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. *arXiv preprint arXiv:2206.10092*, 2022. 1

[38] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, pages 6498–6508, 2021. 2

[39] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 1, 6

[40] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *arXiv preprint arXiv:2205.13790*, 2022. 1

[41] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, pages 5741–5751, 2021. 2

[42] Fayao Liu, Chunhua Shen, Guosheng Lin, and Ian Reid. Learning depth from single monocular images using deep convolutional neural fields. *TPAMI*, 38(10):2024–2039, 2015. 3

[43] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 5, 6, 7, 9

[44] Yingfei Liu, Tiancai Wang, Xiangyu Zhang, and Jian Sun. Petr: Position embedding transformation for multi-view 3d object detection. *arXiv preprint arXiv:2203.05625*, 2022. 1

[45] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 1

[46] Reza Mahjourian, Martin Wicke, and Anelia Angelova. Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints. In *CVPR*, pages 5667–5675, 2018. 2, 3

[47] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021. 2

[48] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 4

[49] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, pages 405–421. Springer, 2020. 2, 4, 7

[50] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *CVPR*, pages 16190–16199, 2022. 2

[51] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *TOG*, 41(4):1–15, 2022. 2

[52] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 2

[53] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, pages 5589–5599, 2021. 2

[54] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. *arXiv preprint arXiv:2309.09502*, 2023. 2, 4, 6

[55] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5865–5874, 2021. 2

[56] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 2

[57] Anurag Ranjan, Varun Jampani, Lukas Balles, Kihwan Kim, Deqing Sun, Jonas Wulff, and Michael J Black. Competitive Collaboration: Joint Unsupervised Learning of Depth, Camera Motion, Optical Flow and Motion Segmentation. In *CVPR*, pages 12240–12249, 2019. 3

[58] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *ICCV*, pages 14335–14345, 2021. 2

[59] Anirban Roy and Sinisa Todorovic. Monocular depth estimation using neural regression forest. In *CVPR*, pages 5506–5514, 2016. 3

[60] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *ICCV*, pages 3216–3226, 2023. 3, 5, 10, 11

[61] Tianwei Shen, Zixin Luo, Lei Zhou, Hanyu Deng, Runze Zhang, Tian Fang, and Long Quan. Beyond Photometric Loss for Self-Supervised Ego-Motion Estimation. *arXiv preprint arXiv:1902.09103*, 2019. 3

[62] Tianwei Shen, Lei Zhou, Zixin Luo, Yao Yao, Shiwei Li, Jiahui Zhang, Tian Fang, and Long Quan. Self-Supervised Learning of Depth and Motion Under Photometric Inconsistency. In *ICCVW*, pages 0–0, 2019. 3

[63] Yunxiao Shi, Hong Cai, Amin Ansari, and Fatih Porikli. Ega-depth: Efficient guided attention for self-supervised multi-camera depth estimation. In *CVPRW*, pages 119–129, 2023. 3

[64] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, pages 7495–7504, 2021. 2

[65] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5459–5469, 2022. 2

[66] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, pages 8248–8258, 2022. 2

[67] 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. 1, 2, 6, 9

[68] 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. 2

[69] Fabio Tosi, Filippo Aleotti, Matteo Poggi, and Stefano Mattoccia. Learning monocular depth estimation infusing traditional stereo knowledge. In *CVPR*, pages 9799–9809, 2019. 3

[70] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, pages 12959–12970, 2021. 2

[71] Chaoyang Wang, José Miguel Buenaposada, Rui Zhu, and Simon Lucey. Learning depth from monocular videos using direct methods. In *CVPR*, pages 2022–2030, 2018. 3

[72] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 34:27171–27183, 2021. 2

[73] 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*, 2023. 1, 2

[74] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[75] Jamie Watson, Michael Firman, Gabriel J Brostow, and Daniyar Turmukhambetov. Self-Supervised Monocular Depth Hints. In *ICCV*, pages 2162–2171, 2019. 3

[76] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, pages 5610–5619, 2021. 2

[77] 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. PMLR, 2023. 3, 5, 6, 10, 11

[78] 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, 3

[79] Jialei Xu, Xianming Liu, Yuanchao Bai, Junjun Jiang, Kaixuan Wang, Xiaozhi Chen, and Xiangyang Ji. Multi-camera collaborative depth prediction via consistent structure estimation. In *ACMMM*, pages 2730–2738, 2022. 3

[80] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *CVPR*, pages 5438–5448, 2022. 2

[81] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 34: 4805–4815, 2021. 2

[82] Zhichao Yin and Jianping Shi. Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In *CVPR*, pages 1983–1992, 2018. 2, 3

[83] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 35:25018–25032, 2022. 2

[84] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2, 4

[85] 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. 1

[86] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, 2023. 1, 2, 3, 6

[87] Zhenyu Zhang, Chunyan Xu, Jian Yang, Junbin Gao, and Zhen Cui. Progressive hard-mining network for monocular depth estimation. *TIP*, 27(8):3691–3702, 2018. 3

[88] Junsheng Zhou, Yuwang Wang, Kaihuai Qin, and Wenjun Zeng. Moving Indoor: Unsupervised Video Depth Learning in Challenging Environments. In *ICCV*, pages 8618–8627, 2019. 3

[89] Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *CVPR*, pages 1851–1858, 2017. 2, 3, 4, 6

[90] Yuliang Zou, Zelun Luo, and Jia-Bin Huang. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *ECCV*, pages 36–53, 2018. 3

[91] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Pointocc: Cylindrical tri-perspective view for point-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2308.16896*, 2023. 2