SelfOcc: Self-Supervised Vision-Based 3D Occupancy Prediction

Yuanhui Huang*   Wenzhao Zheng*   Borui Zhang   Jie Zhou   Jiwen Lu[†]

Beijing National Research Center for Information Science and Technology, China

Department of Automation, Tsinghua University, China

{huangyh22,zhang-br21}@mails.tsinghua.edu.cn; wenzhao.zheng@outlook.com;
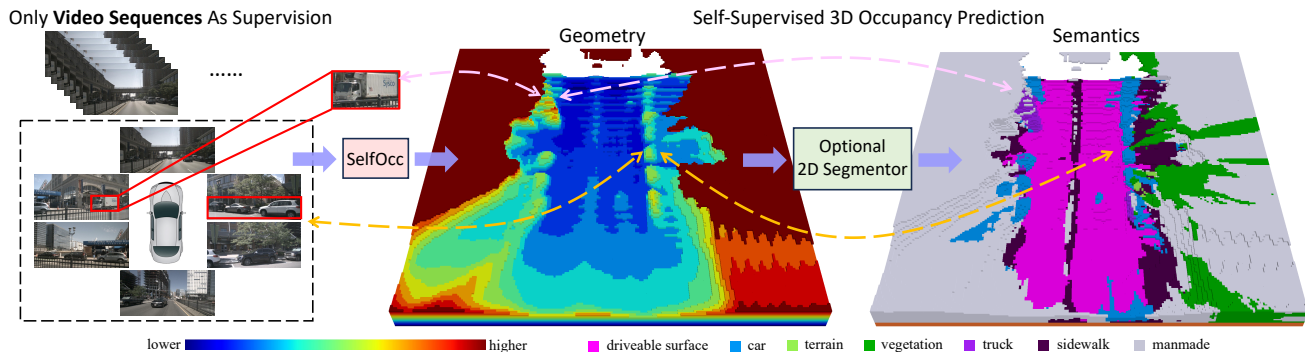
{jzhou,lujiwen}@tsinghua.edu.cn

Figure 1. Trained with only video sequences as supervision, our model can predict meaningful geometry for the scene given surround-camera RGB images, which can be further extended to semantic occupancy prediction if 2D segmentation maps are available e.g. from an off-the-shelf segmentor. This task is challenging because it completely depends on video sequences to reconstruct scenes without any 3D supervision. We observe that our model can produce dense and consistent occupancy prediction and even infer the back side of cars.

Abstract

3D occupancy prediction is an important task for the robustness of vision-centric autonomous driving, which aims to predict whether each point is occupied in the surrounding 3D space. Existing methods usually require 3D occupancy labels to produce meaningful results. However, it is very laborious to annotate the occupancy status of each voxel. In this paper, we propose SelfOcc to explore a self-supervised way to learn 3D occupancy using only video sequences. We first transform the images into the 3D space (e.g., bird's eye view) to obtain 3D representation of the scene. We directly impose constraints on the 3D representations by treating them as signed distance fields. We can then render 2D images of previous and future frames as self-supervision signals to learn the 3D representations. We propose an MVS-embedded strategy to directly optimize the SDF-induced weights with multiple depth proposals. Our SelfOcc outperforms the previous best method SceneRF by 58.7% using a single frame as input on SemanticKITTI and is the first self-supervised work that produces reasonable 3D occupancy for surround cameras on nuScenes. SelfOcc produces high-quality depth and achieves state-of-the-art results on novel depth synthesis, monocular depth estimation, and surround-view depth estimation on the SemanticKITTI, KITTI-2015, and nuScenes, respectively. Code: https://github.com/huang-yh/SelfOcc.

1. Introduction

The vision-centric paradigm has demonstrated its potential for efficient autonomous driving [29, 31, 32, 40], where the key procedure is to obtain a perceptual 3D representation, e.g. the bird's eye view (BEV) [50] or tri-perspective view (TPV) [31] of the surrounding scene given only 2D images [40, 54]. Most existing methods achieve this using 3D annotations as supervision and achieve promising results on various tasks including semantic map construction [28, 50, 54], 3D object detection [30, 39, 40, 81], and 3D semantic occupancy prediction [6, 31, 63, 66].

Despite the strong performance of supervised 3D representation that facilitates accurate subsequent prediction and planning [29, 32, 64], it is very laborious to annotate 3D labels and thus difficult to scale to large-scale training data. As one workaround, data auto-labeling creates data cycles and train large networks to assist human annotation, yet it is still time-consuming and resource-hungry [11, 55, 74]. It is also noticed to be the bottleneck towards building and updating large models for autonomous driving [9]. Therefore, it is desired to learn meaningful 3D representations only from recorded video sequences without 3D labels. While existing works have explored learning 3D information (e.g., depth) from images [7, 23, 71], they are still based on 2D representations in each monocular camera space. It remains challenging to obtain reasonable and comprehensive 3D representations useful for further autonomous driving prediction and planning [29] in a self-supervised manner.

In this paper, we explore the problem of learning self-

---

*Equal contribution. †Corresponding author.

supervised 3D representation (BEV and TPV) and adopt 3D occupancy prediction as the main task to evaluate the quality of the learned 3D representation, which aims to classify each point in the 3D space into occupied or empty. We first lift 2D image features into 3D space with deformable attention layers [31, 40] to enable feature interactions in the 3D space and avoid ambiguities from multiple cameras. Inspired by recent advances in neural implicit surface reconstruction [52, 65], we then transform the 3D representation into a signed distance function (SDF) field to allow for meaningful regularization and straightforward determination of occupancy boundaries. When exploiting the temporal consistency inherent in video sequences to optimize the SDF field, we propose an MVS-embedded strategy that directly optimizes the SDF-induced weight values with multiple depth proposals. Combined with the temporal supervision scheme and loss formulation tailored for 3D occupancy prediction, our method can reconstruct meaningful 3D occupancy in a self-supervised manner.

We conduct extensive experiments on various datasets to demonstrate the effectiveness of the proposed SelfOcc. We apply SelfOcc to two types of 3D representations BEV and TPV. For surround-view 3D occupancy prediction, SelfOcc is the first self-supervised method that is able to produce reasonable occupancy results using only video supervision (with an IoU of 45.01 and mIoU of 9.30 on Occ3D [63]). For monocular 3D occupancy prediction, SelfOcc outperforms the previous best method SceneRF [7] by 58.7% with an IoU of 21.97 over 13.84 on SemanticKITTI [3]. To further evaluate the usefulness of the learned 3D representations, we render depth maps for each camera from the self-supervised BEV and TPV representation. Our SelfOcc outperforms the mainstream method SceneRF [7], MonoDepth2 [23], and SurroundDepth [69] on novel depth synthesis, monocular depth estimation, and surround-view depth estimation on the SemanticKITTI [3], KITTI-2015 [21], and nuScenes [4], respectively.

2. Related Work

3D Occupancy Prediction: While detection, tracking, prediction and planning have long been the focal tasks and composed a standard pipeline for autonomous driving, 3D occupancy prediction [31] or semantic scene completion [62] has recently gained attention as a more fundamental task encompassing the most fine-grained prediction of both occupancy and semantics of the environment. Pioneering works rely on 3D inputs such as depth [13, 20, 34–36, 38, 44], occupancy grids [20, 59, 72, 73], point cloud [57, 83], or truncated signed distance function (TSDF) [5, 10, 12, 13, 16, 17, 35, 62, 67, 78, 80]. Cao et al. [6] and Huang et al. [31] first attempt to predict semantic occupancy from image input only. Recent advancements related to 3D occupancy prediction include modal-

ity fusion [66], multi-task learning [64], and end-to-end autonomous driving [29]. Despite the promising performance, all these methods require 3D ground truth from laborious annotation for supervision. Most related to our work is BTS [71] and SceneRF [7], which learn 3D occupancy in a self-supervised manner in monocular scenarios. However, it is nontrivial to adapt them to surround views, while our work can inherently handle monocular and surround cases.

Neural Radiance Fields: One of the most predominant paradigms for self-supervised 3D reconstruction is neural radiance fields [47], which learn a multi-layer perceptron (MLP) per scene to map spatial locations to radiance and density values. To synthesize a novel view, volume rendering [46] is often leveraged which queries the MLP at multiple locations along rays and integrates radiance over density-induced weights. Training of NeRFs [1, 2, 8, 18, 47, 49] depends on a large number of images of the same scene with ground truth poses. To facilitate sparse view reconstruction and generalization for object-level data, several methods [42, 48, 75] propose to condition NeRFs on image features which establish the foundation for online NeRFs. Recently, this boundary has been further extended to scene-level online reconstruction by [7, 37, 71]. However, these works still use image features as conditions for NeRFs, ignoring 3D feature encoding crucial for 3D perception, while our method learns 3D representations explicitly.

Self-supervised Depth Prediction: Training of NeRFs is time-consuming and prone to overfitting due to geometry-appearance coupling, which can be alleviated with additional depth supervision from separately trained depth prediction modules [15, 58, 68] or LiDAR data [56]. To integrate explicit depth optimization in self-supervised generalizable NeRFs, BTS [71] and SceneRF [7] leverage the reprojection loss from self-supervised depth estimation literature [22, 23, 76, 85], which minimizes the discrepancy between the target pixel and its warped counterpart. However, the extra degrees of freedom from volume rendering and the local receptive field from bilinear interpolation may still hinder depth optimization, while we propose a novel MVS-embedded approach to learn depth in NeRFs.

3. Proposed Approach

In this section, we present a self-supervised vision-based 3D occupancy prediction method for autonomous driving.

3.1. From Image to Occupancy

To understand the complicated interactions among traffic participants and their surrounding environment, it is fundamental for autonomous driving systems to reconstruct the accurate 3D structure of the scene. 3D occupancy prediction is one prevalent proxy for scene reconstruction given its fine granularity and comprehensiveness, which aims at producing a voxelized prediction $\mathbf{O} \in \mathcal{C}^{H \times W \times D}$
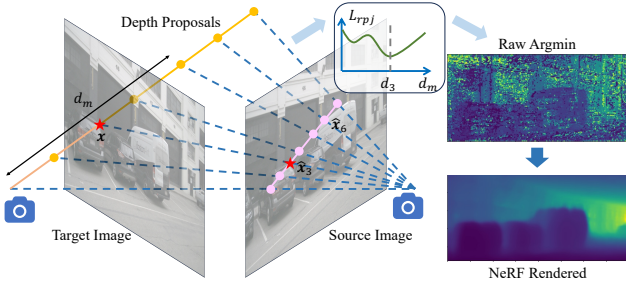
Figure 2. Our MVS-embedded strategy effectively enlarges the receptive field of the depth optimization process across the whole epipolar line, which provides a good depth prior (Raw Argmin).

encoding per-voxel occupancy (and semantic) information. Here H, W, D and $\mathcal{C}$ denote the resolution of the occupancy grid and the set of predefined classes. First proposed by MonoScene [6], vision-based methods have become a promising approach to 3D occupancy prediction because of their low cost and effectiveness compared with 3D-based methods. Vision-based methods generally learn a mapping $\mathcal{M} = \mathcal{D} \circ \mathcal{F} \circ \mathcal{E}$ from RGB images $\mathbf{I} = \{\mathbf{I}_n | n = 1, ..., N\}$ captured by N cameras to 3D occupancy $\mathbf{O}$:

$$\mathbf{O} = \mathcal{M}(\mathbf{I}) = \mathcal{D}(\mathcal{F}(\mathcal{E}(\mathbf{I}_1), ..., \mathcal{E}(\mathbf{I}_N))), \quad (1)$$

where a 2D backbone $\mathcal{E}$ first encodes N input images into multi-scale image features $\mathbf{F} = \mathcal{E}(\mathbf{I})$, and a 3D encoder $\mathcal{F}$ further lifts 2D features into 3D representation $\mathbf{R} = \mathcal{F}(\mathbf{F})$ through attention mechanism [31, 40], lift-splat-shoot (LLS) [39] or identical transformation for monocular scenario [6], followed by a decoder network $\mathcal{D}$ to generate the final occupancy prediction result $\mathbf{O} = \mathcal{D}(\mathbf{R})$.

Although recent supervised vision-based approaches have achieved encouraging performance on 3D occupancy prediction given only visual inputs at inference, they still rely on 3D (semantic) supervision during training, e.g. Li-DAR point cloud or dense occupancy annotation. On the contrary, self-supervised counterparts [7, 71] take only temporal correspondences inherent in video sequences as supervision to predict meaningful occupancy, and therefore can effortlessly exploit the vast amount of unlabeled driving images. However, existing self-supervised methods [7, 71] consider only one monocular camera and decode forward occupancy $\mathbf{O}_{front}$ directly from 2D image features by instantiating the 3D encoder $\mathcal{F}$ with identical transformation:

$$\mathbf{O}_{front} = \mathcal{D}(\mathcal{F}(\mathcal{E}(\mathbf{I}_{mono}))) = \mathcal{D}(\mathcal{E}(\mathbf{I}_{mono})). \quad (2)$$

Despite being straightforward, ignoring $\mathcal{F}$ may cause multi-camera inconsistency in the surround view scenario and inadequate geometry reasoning in the 3D space.

To this end, we propose to transform the images into the BEV or TPV space to obtain a 3D representation of the scene in order to enable 3D feature interactions even in the monocular setting and also avoid ambiguities from multiple cameras. Similar to the supervised literature [31, 40], we leverage the deformable cross-attention (CA) to adaptively aggregate information from the image features $\mathbf{F}$, in which a set of learnable 3D tokens $\mathbf{Q}$ serve as queries and the corresponding local image features serve as keys and values. Each of these learnable tokens represents a pillar area as in BEV or TPV representation. Moreover, we determine the correspondences between 3D tokens and image features with the projection matrices $\mathbf{T} = \{\mathbf{T}_n | n = 1, ..., N\}$ from the ego car to pixel coordinates. We interleave deformable self-attention (SA), deformable cross-attention (CA) and feed forward network (FFN) to build the 3D encoder $\mathcal{F}$.

$$\mathcal{F}_l = \text{FFN}_l(\text{CA}_l(\text{SA}_l(\mathbf{Q}_l); \mathbf{F}, \mathbf{T})), \quad (3)$$

where the subscription l represents the l-th block.

3.2. From Occupancy to Image

Because 3D ground truth is unavailable, we project occupancy $\mathbf{O}$ from Eq. (1) back into 2D views and leverage multi-view consistency inherent in video sequences to optimize a self-supervised vision-based occupancy prediction system. Similar to previous work [7, 71], we employ differentiable volume rendering [47] to synthesize color and depth views, which enables the seamless integration of gradient information into the rendering pipeline and effective exploitation of supervision from multiple viewpoints.

We first transform the 3D representation $\mathbf{R}$ into an SDF field $\mathbf{S} \in \mathbb{R}^{H \times W \times D}$, which indicates the distance of every voxel center to the nearest surface of an object, by implementing the decoder network $\mathcal{D}$ with an MLP. We use different MLPs for different heights for BEV representation, while we construct a 3D feature volume from TPV representation with broadcasting and summation [31] before feeding the feature volume through a single MLP. For a continuous 3D coordinate $\mathbf{p}$, we employ bicubic interpolation (BI) to predict the SDF value $s_{\mathbf{p}}$ and determine its occupancy status $o_{\mathbf{p}}$ according to the sign of $s_{\mathbf{p}}$:

$$o_{\mathbf{p}} = \text{sgn}(s_{\mathbf{p}}), \quad s_{\mathbf{p}} = \text{BI}(\mathbf{S}, \mathbf{p}) = \text{BI}(\mathcal{D}(\mathbf{R}), \mathbf{p}). \quad (4)$$

The preference for SDF field over density field as in [7, 71] is based on two considerations: 1) An SDF field has more explicit physical meaning compared with a density field and inherently holds assumptions about the gradient magnitudes [24], thus allowing for easier regularization and optimization. 2) The signed nature of SDF facilitates straightforward determination of whether a point lies inside or outside a surface, thus enabling accurate differentiation of intricate geometries. Similar to SDF values, we can also decode color $\mathbf{c}$ from $\mathbf{R}$ by using a separate MLP $\mathcal{D}_c$ as the decoder:

$$\mathbf{c}_{\mathbf{p}} = \text{BI}(\mathbf{C}, \mathbf{p}) = \text{BI}(\mathcal{D}_c(\mathbf{R}), \mathbf{p}). \quad (5)$$

Take the volume rendering process of one single ray as an example. We first determine the origin and direction of
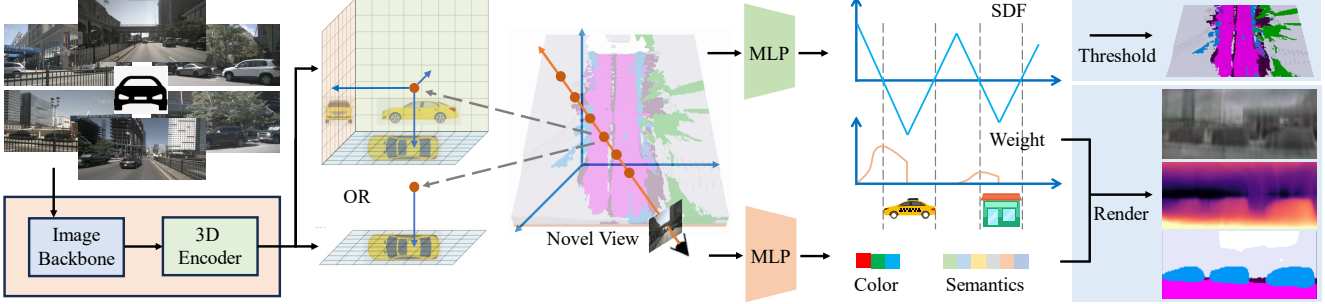
Figure 3. Framework of the proposed SelfOcc for self-supervised vision-based 3D occupancy prediction. We employ an image backbone and a 3D encoder to generate 3D representations as in BEVFormer [40] or TPVFormer [31]. To render a novel view, we apply a lightweight MLP on the 3D features to predict the SDF values, color and semantic vectors. We then perform volume rendering to synthesize color, depth and semantic views. We use simple 0-thresholding to predict the occupancy volume.

the ray and uniformly sample M points $\mathbf{P} = \{\mathbf{p}_m | m = 1, ..., M\}$ between the origin and the intersection of the ray with the border of the 3D representation. The SDF values of these points $\mathbf{s} = \{s_m | m = 1, ..., M\}$ are then calculated with Eq. (4). According to NeuS [65], the discrete opacity value α_m, which is the probability of the ray terminating between $\mathbf{p}_m$ and $\mathbf{p}_{m+1}$, can be derived with

$$\alpha_m = \max\left(\frac{\Phi_a(s_m) - \Phi_a(s_{m+1})}{\Phi_a(s_m)}, 0\right), \quad (6)$$

where $\Phi_a(x) = (1 + e^{-ax})^{-1}$ is the sigmoid function with a being a learnable parameter. Further, the probability of the ray passing through $\mathbf{p}_m$, i.e. the discrete accumulated transmittance, can be defined by $T_m = \prod_{i=1}^{m-1}(1 - \alpha_i)$. Finally, we can calculate the probability of the ray ending at $\mathbf{p}_m$ after emission as $w_m = T_m \alpha_m$, and the rendering process is an integration of per-point attribute over the probability distribution $\mathbf{w} = \{w_m | m = 1, ..., M\}$. Eq. (7) calculates the rendered color $\mathbf{c}_r$ and depth d_r of the ray:

$$\mathbf{c}_r = \sum_{m=1}^{M} w_m \mathbf{c}_m, \quad d_r = \sum_{m=1}^{M} w_m d_m, \quad (7)$$

where $\mathbf{c}_m, d_m$ denote the color and depth of the m-th point. Figure 3 shows the overall framework of our method.

3.3. Occupancy-Oriented Supervision

In this section, we elaborate on the supervision formulation that facilitates self-supervised vision-based 3D occupancy prediction for autonomous driving.

MVS-embedded Depth Learning. Depth supervision enhances the convergence of neural radiance fields [68], enabling faster and more effective learning of accurate geometry, especially with sparse views. Self-supervised depth estimation in NeRFs usually relies on the photometric reprojection loss L_{rpj} which aims to maximize the similarity between the target image $\mathbf{I}_t$ and sampled pixels on the

source image $\mathbf{I}_s$ by refining the depth prediction $\mathbf{D}(\mathbf{I}_t; \boldsymbol{\theta})$:

$$\begin{aligned} L_{rpj}(\mathbf{x}, \mathbf{I}_t, \mathbf{I}_s; \boldsymbol{\theta}) &= \|\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_s(\hat{\mathbf{x}})\| \\ &= \|\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_s(\pi(\mathbf{x}, \mathbf{D}(\mathbf{I}_t; \boldsymbol{\theta})|_{\mathbf{x}}, \mathbf{\Pi}))\|, \end{aligned} \quad (8)$$

where $\boldsymbol{\theta}$, $\mathbf{x}$, $\hat{\mathbf{x}}$, $\mathbf{\Pi}$ denote the learnable parameters, a random pixel on the target image, the warped pixel on the source image and the projection matrix from the target to the source image coordinate, respectively. $\mathbf{I}(\mathbf{x})$ and $\mathbf{D}|_{\mathbf{x}}$ denote bilinear interpolation of corresponding 2D tensors, and $\pi(\mathbf{x}, d_{\mathbf{x}}, \mathbf{\Pi})$ warps a pixel $\mathbf{x}$ to another image according to the depth $d_{\mathbf{x}}$ and projection matrix $\mathbf{\Pi}$. The inherent local receptive field of the bilinear interpolation $\mathbf{I}_s(\hat{\mathbf{x}})$ has an adverse effect on the optimization property of L_{rpj} since the differences between $\mathbf{I}_t(\mathbf{x})$ and the colors of the four adjacent grid pixels of $\hat{\mathbf{x}}$ are the only reference information for the optimization of $\boldsymbol{\theta}$, which could be easily misled by poor initialization or low-texture regions.

Therefore, we propose an MVS-embedded strategy to extend the receptive field of the depth optimization for a ray across the whole epipolar line. Specifically, we first take the uniformly spaced depths $\mathbf{d} = \{d_m | m = 1, ..., M\}$ associated with $\mathbf{P} = \{\mathbf{p}_m | m = 1, ..., M\}$ as depth proposals. Then we define L_{mvs} as an average of the dissimilarity at each proposal d_m with weight w_m from volume rendering:

$$\begin{aligned} L_{mvs}(\mathbf{x}, \mathbf{I}_t, \mathbf{I}_s; \boldsymbol{\theta}) \\ = \sum_{m=1}^{M} w_m(\boldsymbol{\theta}) \|\mathbf{I}_t(\mathbf{x}) - \mathbf{I}_x(\pi(\mathbf{x}, d_m, \mathbf{\Pi}))\|. \end{aligned} \quad (9)$$

As Figure 2 shows, L_{mvs} enlarges the receptive field of depth optimization by associating $\boldsymbol{\theta}$ with multiple depth proposals along the ray, thus stabilizing and facilitating self-supervised depth optimization in NeRFs. Similar to Monodepth2 [23], we formulate the final depth loss by taking the minimum of two L_{mvs}s using the previous image $\mathbf{I}_{t-1}$ or the next image $\mathbf{I}_{t+1}$ as the source image and leverage the auto-masking strategy to filter out low texture regions:

$$L_{dep}(\mathbf{x}, \mathbf{I}_t) = \text{automask}(\min(L_{mvs}^{(t-1)}, L_{mvs}^{(t+1)})), \quad (10)$$

Table 1. **3D occupancy prediction performance on the Occ3D-nuScenes [63] dataset.** Cons. veh. and drive. surf. represent construction vehicle and driveable surface, respectively. Our method learns meaningful geometry and reasonable semantics compared with 3D-supervised methods, and achieves even higher IoU than LiDAR-supervised TPVFormer [31].

Method	Supervision	IoU	mIoU	others	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
MonoScene [6]	3D	-	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
OccFormer [82]	3D	-	21.93	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97
BEVFormer [40]	3D	-	26.88	5.85	37.83	17.87	40.44	42.43	7.36	23.88	21.81	20.98	22.38	30.70	55.35	28.36	36.0	28.06	20.04	17.69
CTF-Occ [63]	3D	-	_28.53_	8.09	_39.33_	_20.56_	38.29	42.24	16.93	_24.52_	_22.72_	_21.05_	22.98	31.11	53.33	33.84	37.98	33.23	_20.79_	18.0
TPVFormer [31]	3D	-	27.83	7.22	38.90	13.67	_40.78_	_45.90_	_17.23_	19.99	18.85	14.30	_26.69_	_34.17_	55.65	_35.47_	37.55	30.70	19.40	16.78
FB-OCC [41]	3D	-	**42.06**	**14.30**	**49.71**	**30.00**	**46.62**	**51.54**	**29.30**	**29.13**	**29.35**	**30.48**	**34.97**	**39.36**	**83.07**	**47.16**	**55.62**	**59.88**	**44.89**	**39.58**
RenderOcc [53]	L	-	23.93	5.69	27.56	14.36	19.91	20.56	11.96	12.42	12.14	14.34	20.81	18.94	_68.85_	33.35	_42.01_	_43.94_	17.36	_22.61_
TPVFormer [31]	L	17.20	13.57	0.00	14.80	9.36	21.27	16.81	14.45	13.76	11.23	5.32	16.05	19.73	10.75	9.43	9.50	11.16	16.51	17.04
SelfOcc (BEV)	C	_44.33_	6.76	0.00	0.00	0.00	0.00	9.82	0.00	0.00	0.00	0.00	0.00	6.97	47.03	0.00	18.75	16.58	11.93	3.81
SelfOcc (TPV)	C	**45.01**	9.30	0.00	0.15	0.66	5.46	12.54	0.00	0.80	2.10	0.00	0.00	8.25	55.49	0.00	26.30	26.54	14.22	5.60

where we omit the learnable parameters $\boldsymbol{\theta}$.

Supervision from temporal frames. Unlike depth estimation which predicts only the visible surfaces, 3D occupancy prediction reconstructs the complete geometry of the scene which requires self-supervision signals from multiple viewpoints as in NeRFs. Considering the limited overlapping field of view for cameras, we take advantage of the temporal frames as supervision. Given a video sequence $\mathbf{V} = \{\mathbf{V}_1, ..., \mathbf{V}_T\}$ with T frames and the current timestamp t, we use $\mathbf{V}_t$ as the input to our model $\mathcal{M}$ to predict the SDF and radiance fields $\{\mathbf{S}_t, \mathbf{C}_t\}$. Then we decide whether to use a temporal frame as supervision with probability p. If positive, we randomly sample one frame t' for supervision satisfying that the distance between the ego cars at timestamps t and t' falls inside $[l_1, l_2]$, which guarantees diversity and adheres to the range of the 3D representation. Otherwise, we use the current frame t for supervision.

SDF Field Regularization. Neural radiance fields with sparse views suffer from overfitting because of a large solution set [51], which is exacerbated in autonomous driving with outward-facing cameras. Therefore, it is crucial to regularize the SDF field in order to restrict the solution set. To learn a smooth SDF field, we minimize the magnitude of the second-order derivatives of the SDF field with a Hessian loss [79] $L_H(\mathbf{p}) = \|\mathbf{H}(\mathbf{p})\|_1$, where $\|\cdot\|_1$ is the element-wise matrix 1-norm, and $\mathbf{H}(\mathbf{p})$ denotes the Hessian matrix of the SDF field with respect to the 3D coordinates at location $\mathbf{p}$. To impose sparsity of occupancy in the invisible area such as those blocked by a wall, we propose a simple yet effective regularization $L_s(\mathbf{p}) = \max(-s_{\mathbf{p}}, 0)$, which encourages negative SDF values to become larger. In addition, we include the Eikonal term [24] $L_E(\mathbf{p}) = |\|\nabla s(\mathbf{p})\| - 1|$ for compliance with physical meaning of SDF. For optional semantic prediction, we use 2D segmentation maps from off-the-shelf segmentors as supervision similar to colors.

Til now, we define the training loss L as follows:

$$L = L_{dep} + \lambda_c L_{rgb} + \lambda_e L_E + \lambda_H L_H + \lambda_s L_s \quad (11)$$

Table 2. **3D occupancy prediction performance on the SemanticKITTI [3] dataset.** MonoScene* is supervised by depth predictions of Monodepth2 [23] trained with ground-truth poses.

Method	Supervision 3D	Supervision Depth	Supervision Image	IoU	Prec.	Rec.
MonoScene [6]	✓			**37.14**	**49.90**	59.24
LMSCNet^rgb [59]		✓		12.08	13.00	_63.16_
3DSketch^rgb [10]		✓		12.01	12.95	62.31
AICNet^rgb [36]		✓		11.28	11.84	**70.89**
MonoScene [6]		✓		13.53	16.98	40.06
MonoScene* [6]			✓	11.18	13.15	40.22
SceneRF [7]			✓	13.84	17.28	40.96
SelfOcc (BEV)			✓	20.95	_37.27_	32.37
SelfOcc (TPV)			✓	_21.97_	34.83	37.31

where the four λs are hyperparameters and L_{rgb} measures the dissimilarity between the rendered color $\mathbf{c}_r$ of a ray and the ground truth color on the target image. For semantic prediction, we optionally add a semantic loss L_{sem} similar to L_{rbg} and an off-the-shelf 2D segmentor.

4. Experiments

4.1. Task Descriptions

The 3D occupancy prediction task aims to predict the dense occupancy (and semantic) states for a voxel grid, which is a fundamental task in autonomous driving due to its fine granularity and least abstraction of the real world. We conduct this task on the Occ3D-nuScenes [63] dataset, and use IoU and mIoU as geometric and semantic metrics, respectively. In addition, we evaluate our method on the SemanticKITTI [3] dataset for the geometry-only 3D occupancy prediction task and report IoU, Precision and Recall.

The novel depth synthesis task synthesizes depth maps from novel viewpoints within a certain distance given the input frame [7]. This task is more reliable as a proxy for 3D reconstruction compared with novel view synthesis because

Table 3. **Novel depth synthesis on SemanticKITTI [3] and nuScenes [4].** We use the results reported in SceneRF [7] and all methods are trained using only images with ground truth poses. We adapt SceneRF [7] to the nuScenes dataset based on the official implementation.

Method	Dataset	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ1↑	δ2↑	δ3↑
MonoDepth2 [23]		0.5259	7.113	14.43	1.0292	10.44	26.32	41.43
SynSin [70]		0.3611	3.483	8.824	0.4290	52.61	74.56	86.50
PixelNeRF [75]		0.2364	2.080	6.449	0.3354	65.81	85.43	92.90
MINE [37]	SemanticKITTI	0.2248	1.787	6.343	0.3283	65.87	85.52	93.30
VisionNerf [42]		0.2054	1.490	5.841	0.3073	69.11	88.28	94.37
SceneRF [7]		0.1681	1.291	5.781	0.2851	75.07	89.09	94.50
SelfOcc (TPV)		**0.1562**	**1.153**	**5.349**	**0.2534**	**78.92**	**91.12**	**95.92**
SceneRF [7]	nuScenes	0.7631	15.34	11.50	0.6240	40.26	64.86	77.02
SelfOcc (TPV)		**0.4003**	**6.723**	**8.460**	**0.4165**	**64.50**	**80.40**	**88.17**

Table 4. **Self-supervised depth estimation on the nuScenes [4] and KITTI-2015 [21] dataset.** M: trained with monocular sequence. MS: trained with both monocular sequence and stereo images. SV: trained with surround-view images. TO: temporal offline refinement. We train SurroundDepth* with ground truth poses based on the official implementation.

Method	Dataset	Type	Scale Method	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ1↑	δ2↑	δ3↑
Monodepth2 [23]		SV	Median	0.287	3.349	7.184	0.345	0.641	0.845	0.925
FSM [26]		SV	Median	0.299	-	-	-	-	-	-
SurroundDepth [69]	nuScenes	SV	SfM Pretrain	0.280	4.401	7.467	0.364	0.661	0.844	0.917
SurroundDepth* [69]		SV	Pose GT	0.342	7.917	8.019	0.360	0.716	0.858	0.918
SimpleOcc [19]		SV	Pose GT	0.224	3.383	7.165	0.333	0.753	0.877	0.930
SelfOcc (TPV)		SV	Pose GT	**0.215**	**2.743**	**6.706**	**0.316**	0.753	0.875	**0.932**
R3D3 [60]	nuScenes	TO	Extrinsics	0.253	4.759	7.150	-	0.729	-	-
Johnston [33]		M	Median	0.106	0.861	4.699	0.185	0.889	0.962	0.982
FeatDepth(R50) [61]		M	Median	0.104	0.729	4.481	0.179	**0.893**	0.965	0.984
PackNet-SfM [25]	KITTI-2015	M	Median	0.107	0.802	4.538	0.186	0.889	0.962	0.981
R-MSFM6 [86]		M	Median	0.108	0.748	4.470	0.185	0.889	0.963	0.982
DevNet [84]		M	Median	**0.100**	**0.699**	**4.412**	**0.174**	**0.893**	**0.966**	**0.985**
SelfOcc (TPV)		M	Pose GT	0.103	0.792	4.673	0.187	0.877	0.959	0.982
MonoDepth2(R50) [23]		MS	Extrinsics	0.106	0.818	4.75	0.196	0.874	0.957	0.979
FeatDepth(R50) [61]		MS	Extrinsics	0.099	0.697	4.427	0.184	0.889	0.963	0.982
R-MSFM6 [86]	KITTI-2015	MS	Extrinsics	0.108	0.753	4.469	0.185	0.888	0.963	0.982
DevNet [84]		MS	Extrinsics	**0.095**	**0.671**	**4.365**	**0.174**	**0.895**	**0.970**	**0.988**
BTS [71]		MS	Pose GT	0.102	0.751	4.407	0.188	0.882	0.961	0.982
SelfOcc (TPV)		MS	Pose GT	0.099	0.711	4.586	0.186	0.880	0.960	0.982

the rendered color is an integration of predicted radiance at sampled locations, which introduces unnecessary degrees of freedom. We conduct experiments on nuScenes [4] and SemanticKITTI [3] and use Abs Rel, Sq Rel, RMSE, RMSE log and threshold accuracies ($\delta1$, $\delta2$, $\delta3$) as metrics.

The depth estimation task is a traditional task for 3D perception, which predicts the pixel-aligned depth map for a given frame. We conduct surround-view depth estimation on nuScenes [4] and monocular depth estiamtion on KITTI-2015 [21]. The same metrics are used as in novel depth synthesis. We detail the datasets in the supplementary.

4.2. Implementation Details

We use the identical network architecture for all three tasks, while the temporal supervision and the loss formulation may differ according to their respective objectives. We adopt ResNet50 [27] pretrained on ImageNet [14] as the 2D backbone and an feature pyramid network (FPN) [43] to generate multi-scale image features. The 3D encoder $\mathcal{F}$ is the same as in BEVFormer [40] or TPVFormer [31] depending on the 3D representation we use. We further employ a two-layer MLP as the decoder network $\mathcal{D}$ to generate the SDF field, color and semantic logits (optional).

We activate temporal supervision for 3D occupancy prediction and novel view synthesis which both involve scene reconstruction from multiple viewpoints, while we use only the current frame as supervision for depth estimation since it predicts the geometry only for the input view. As for loss formulation, we use L_{dep}, L_E and an edge loss L_{edg} [22] for depth estimation since RGB supervision from the input view is meaningless, while L_H and L_s focus on improving the overall geometry, thus helpless for depth es-

Table 5. **Architecture ablation on the nuScenes dataset.** L_{mvs}: the proposed L_{mvs} (✓) or L_{rpj} for depth optimization. SDF: SDF field (✓) or density field for volume rendering.

L_{mvs}	SDF	Occ.		Novel Depth		Depth	
		IoU	mIoU	Abs Rel↓	RMSE↓	Abs Rel↓	RMSE↓
	✓	18.90	2.81	0.8101	12.408	0.655	13.242
✓		36.68	6.10	**0.4007**	**8.149**	0.318	8.524
✓	✓	**38.43**	**6.94**	0.4376	8.734	**0.312**	**7.544**

Table 6. **Temporal supervision ablation on nuScenes [4].** We do not use semantic supervision for occupancy prediction here.

Supervision Ratio			Occ.			Novel Depth		Depth	
Curr.	Prev.	Next	IoU	Prec.	Rec.	Abs Rel	RMSE	Abs Rel	RMSE
1	0	0	39.67	66.96	49.33	0.471	8.852	**0.312**	**7.544**
2	1	1	40.50	67.51	50.31	0.446	8.827	0.365	8.046
1	1	1	40.47	63.98	52.41	**0.438**	8.734	0.363	7.975
1	2	2	40.43	63.35	**52.78**	0.442	8.781	0.366	7.998
0	1	1	**42.48**	**70.08**	51.89	0.462	**8.732**	0.392	8.075

Table 7. **Loss and regularization ablation** on the SemanticKITTI [3] (occupancy prediction and novel depth synthesis) and KITTI-2015 [21] (depth estimation) datasets.

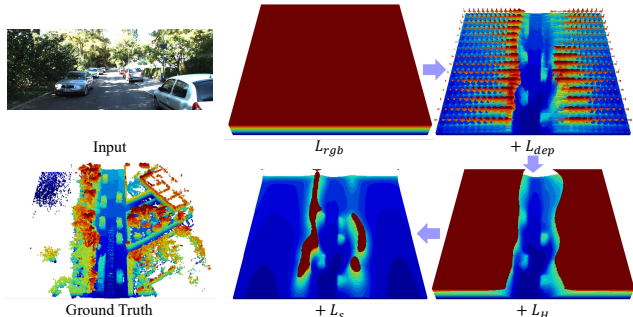| L_{dep}	L_{rgb}	L_H	L_s	Occ.			Novel Depth		Depth	
				IoU	Prec.	Rec.	Abs Rel	RMSE	Abs Rel	RMSE
	✓			9.81	9.85	**96.18**	0.3852	11.67	-	-
✓				23.51	**41.01**	35.51	0.1579	**5.348**	0.102	**4.676**
✓	✓			**24.07**	37.78	39.88	**0.1573**	5.352	**0.100**	4.773
✓	✓	✓		12.27	13.02	68.14	0.1722	5.748	0.107	5.091
✓	✓	✓	✓	22.16	33.54	39.50	0.1694	5.601	0.113	5.078

Figure 4. Visualization of the effect of different loss functions. The occupancy quality gradually improves with L_{dep}, L_H and L_s.

timation. For novel depth synthesis, we leverage L_{dep}, L_{rgb} and L_E. And for 3D occupancy prediction, we further activate L_H and L_s on the basis of novel depth synthesis to impose smoothness and sparsity constraints on the predicted occupancy grid. For semantic prediction, we use the off-the-shelf open-vocabulary segmentation network OpenSeeD [77] to generate pseudo segmentation labels for nuScenes [4]. We use the AdamW [45] optimizer with an initial learning rate of 1e-4 which decays to zero following the cosine schedule. We train our models for 12 epochs on the nuScenes [4] dataset and 24 epochs on the SemanticKITTI [3] and KITTI-2015 [21] datasets.

4.3. Results and Analysis

3D Occupancy Prediction. We report the results for the 3D occupancy prediction task in Table 1 and 2. On the Occ3D-nuScenes [63] dataset, our methods achieve comparable IoU and reasonable mIoU without any form of 3D supervision compared with supervised approaches. The IoU metric (45.01%) indicates that SelfOcc learns meaningful geometry given only video sequences. We notice that SelfOcc is better at predicting background classes than foreground classes, which might be related to the semantic misalignment of the open-vocabulary 2D segmentation network we use. Nonetheless, our method still outperforms MonoScene [6] supervised by ground truth in mIoU and LiDAR-supervised TPVFormer [31] in IoU.

On the SemanticKITTI [3] dataset, we use the results reported in SceneRF [7], where baselines are grouped into 3D, depth and image supervised categories, and the superscript ^{rgb} indicates the adaptation from 3D input to image input. From Table 2, SelfOcc sets the new state-of-the-art for vision-based depth-supervised and self-supervised 3D occupancy prediction tasks, which outperforms SceneRF [7] by 58.7% in IoU, with much higher Precision and slightly lower Recall. Visualizations for 3D occupancy predictions are shown in Figure 5.

Novel Depth Synthesis. We report the results for novel depth synthesis in Table 3, which averages depth metrics over all frames no further than 10 meters away from the input frame. Following SceneRF [7], we compare against methods based on monocular depth estimation [23], generalizable NeRF [37, 42, 75] and 3D-aware GAN [70]. Our method outperforms previous state-of-the-art SceneRF [7] on all metrics on the SemanticKITTI [3] dataset. In addition, we modify SceneRF for the nuScenes [4] dataset by randomly selecting one camera every iteration during training, and directly report the depth metric for the input frame considering the complexity of volume rendering conditioned on image features from surround cameras. SelfOcc still achieves better results than SceneRF, which confirms its superiority in 3D reconstruction from surround views. We provide novel depth visualizations in the supplementary.

Depth Estimation. We report the results for depth estimation in Table 4. Our method achieves state-of-the-art performance on nuScenes for self-supervised surround-view depth estimation. We also train SurroundDepth [69] using the official implementation with ground-truth poses to eliminate their influence, which indicates that ground-truth poses do not necessarily boost performance compared with median-scaling applied for scale-ambiguous methods. Additionally, SelfOcc performs on par with the state-of-the-art for self-supervised depth estimation on KITTI-2015 [21]. Visualizations are provided in the supplementary.
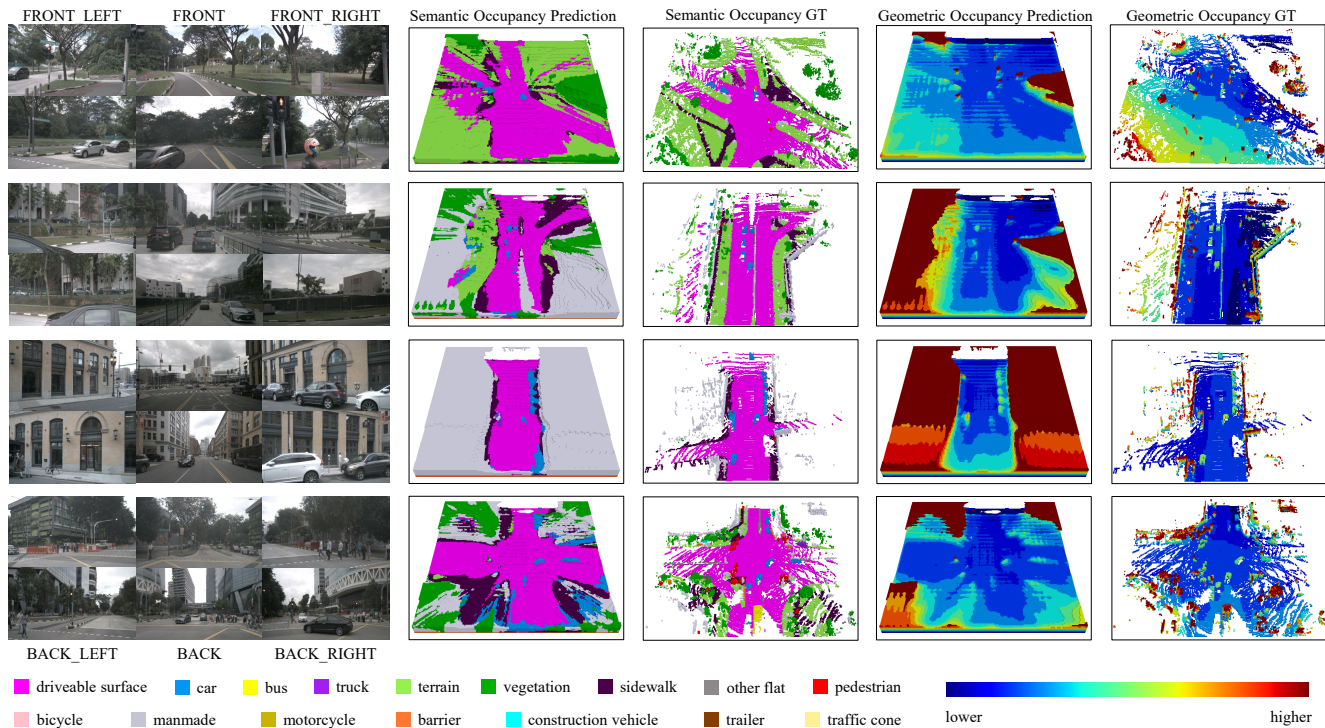
FRONT_LEFT   FRONT   FRONT_RIGHT   Semantic Occupancy Prediction   Semantic Occupancy GT   Geometric Occupancy Prediction   Geometric Occupancy GT

BACK_LEFT   BACK   BACK_RIGHT

■ driveable surface ■ car ■ bus ■ truck ■ terrain ■ vegetation ■ sidewalk ■ other flat ■ pedestrian
■ bicycle ■ manmade ■ motorcycle ■ barrier ■ construction vehicle ■ trailer ■ traffic cone   lower — higher

Figure 5. Visualization results for 3D occupancy prediction on the Occ3D-nuScenes [63] dataset. Our method achieves comparable visualization quality with ground truth for both semantic and geometric occupancy prediction tasks.

4.4. Ablation Study

We conduct comprehensive ablation studies on architectural choices, temporal supervision, and loss formulations to validate the effectiveness of every component of our method. All experiments are trained for half of the original epochs.

Architecture. We investigate the effect of L_{mvs} and the SDF field in Table 5. MVS-embedded depth optimization brings consistent improvement for all three tasks. While the SDF field outperforms the density field in occupancy prediction and depth estimation, it achieves worse results in novel depth synthesis. We think it is because the SDF field holds an inherent smoothness prior and is thus more difficult to optimize for novel depth synthesis.

Temporal Supervision. We report the results of different temporal supervision settings in Table 6. It is shown that increasing temporal supervision benefits occupancy prediction, but deteriorates depth estimation performance. In addition, the best score for novel depth synthesis is achieved with a moderate level of temporal supervision (1:1:1).

Loss Formulations. We ablate the effect of loss functions in Table 7. Note that we replace L_{rbg} with L_{edg} for depth estimation task. Models trained with depth loss L_{dep} and color loss L_{rgb} achieve best performance for depth-related tasks. As shown in Figure 4, we use the Hessian loss L_H and sparsity loss L_s to regularize the SDF field for better visualization quality, although models trained without them can also achieve high IoU during evaluation. In fact,

the sparsity loss makes the prediction more similar to the ground truth, but it may not improve the IoU under the evaluation protocol of using camera masks as in Occ3D [63].

4.5. 3D Occupancy Visualization

We provide the visualization results for 3D occupancy prediction on nuScenes [4] in Figure 5. Despite being trained with video sequences only, our model produces meaningful geometric and reasonable semantic predictions.

5. Conclusion

In this paper, we have presented a self-supervised vision-based 3D occupancy prediction method (SelfOcc) for autonomous driving, which predicts meaningful geometry and semantics from video sequences only. To facilitate feature extraction in the 3D space, we have proposed to condition NeRFs on 3D representations. Furthermore, we have designed an MVS-embedded strategy to boost depth optimization for NeRFs. SelfOcc has set new state-of-the-art for self-supervised vision-based occupancy prediction on SemanticKITTI and has predicted reasonable 3D occupancy for surrounding cameras on Occ3D for the first time.

Acknowledgement

References

[1] 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

[2] 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

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *ICCV*, pages 9297–9307, 2019. 2, 5, 6, 7

[4] 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*, 2020. 2, 6, 7, 8

[5] Yingjie Cai, Xuesong Chen, Chao Zhang, Kwan-Yee Lin, Xiaogang Wang, and Hongsheng Li. Semantic scene completion via integrating instances and scene in-the-loop. In *CVPR*, pages 324–333, 2021. 2

[6] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3991–4001, 2022. 1, 2, 3, 5, 7

[7] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9387–9398, 2023. 1, 2, 3, 5, 6, 7

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

[9] Li Chen, Penghao Wu, Kashyap Chitta, Bernhard Jaeger, Andreas Geiger, and Hongyang Li. End-to-end autonomous driving: Challenges and frontiers. *arXiv preprint arXiv:2306.16927*, 2023. 1

[10] Xiaokang Chen, Kwan-Yee Lin, Chen Qian, Gang Zeng, and Hongsheng Li. 3d sketch-aware semantic scene completion via semi-supervised structure prior. In *CVPR*, 2020. 2, 5

[11] Xuesong Chen, Shaoshuai Shi, Benjin Zhu, Ka Chun Cheung, Hang Xu, and Hongsheng Li. Mppnet: Multi-frame feature intertwining with proxy points for 3d temporal object detection. In *ECCV*, pages 680–697. Springer, 2022. 1

[12] Yueh-Tung Chen, Martin Garbade, and Juergen Gall. 3d semantic scene completion from a single depth image using adversarial training. In *ICIP*, pages 1835–1839. IEEE, 2019. 2

[13] Ran Cheng, Christopher Agia, Yuan Ren, Xinhai Li, and Liu Bingbing. S3cnet: A sparse semantic scene completion network for lidar point clouds. In *CoRL*, pages 2148–2161. PMLR, 2021. 2

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. Ieee, 2009. 6

[15] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, 2022. 2

[16] Aloisio Dourado, Hansung Kim, Teofilo E de Campos, and Adrian Hilton. Semantic scene completion from a single 360-degree image and depth map. In *VISIGRAPP (5: VISAPP)*, pages 36–46, 2020. 2

[17] Aloisio Dourado, Teofilo E De Campos, Hansung Kim, and Adrian Hilton. Edgenet: Semantic scene completion from a single rgb-d image. In *ICPR*, pages 503–510. IEEE, 2021. 2

[18] 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

[19] 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. 6

[20] Martin Garbade, Yueh-Tung Chen, Johann Sawatzky, and Juergen Gall. Two stream 3d semantic scene completion. In *CVPR Workshops*, pages 0–0, 2019. 2

[21] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, pages 3354–3361. IEEE, 2012. 2, 6, 7

[22] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, pages 270–279, 2017. 2, 6

[23] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J Brostow. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3828–3838, 2019. 1, 2, 4, 5, 6, 7

[24] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 3, 5

[25] 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. 6

[26] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *arXiv preprint arXiv:2104.00152*, 2021. 6

[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016. 6

[28] 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*, 2021. 1

[29] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Goal-oriented autonomous driving. *arXiv preprint arXiv:2212.10156*, 2022. 1, 2

[30] 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

[31] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9223–9232, 2023. 1, 2, 3, 4, 5, 6, 7

[32] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. *arXiv preprint arXiv:2303.12077*, 2023. 1

[33] Adrian Johnston and Gustavo Carneiro. Self-supervised monocular trained depth estimation using self-attention and discrete disparity volume. In *CVPR*, 2020. 6

[34] Jie Li, Yu Liu, Dong Gong, Qinfeng Shi, Xia Yuan, Chunxia Zhao, and Ian Reid. Rgbd based dimensional decomposition residual network for 3d semantic scene completion. In *CVPR*, pages 7693–7702, 2019. 2

[35] Jie Li, Yu Liu, Xia Yuan, Chunxia Zhao, Roland Siegwart, Ian Reid, and Cesar Cadena. Depth based semantic scene completion with position importance aware loss. *IEEE Robotics and Automation Letters*, 5(1):219–226, 2019. 2

[36] Jie Li, Kai Han, Peng Wang, Yu Liu, and Xia Yuan. Anisotropic convolutional networks for 3d semantic scene completion. In *CVPR*, 2020. 2, 5

[37] 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 *ICCV*, 2021. 2, 6, 7

[38] Siqi Li, Changqing Zou, Yipeng Li, Xibin Zhao, and Yue Gao. Attention-based multi-modal fusion network for semantic scene completion. In *AAAI*, pages 11402–11409, 2020. 2

[39] 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, 3

[40] 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. *arXiv preprint arXiv:2203.17270*, 2022. 1, 2, 3, 4, 5, 6

[41] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 5

[42] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yichang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *WACV*, 2023. 2, 6, 7

[43] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[44] Shice Liu, Yu Hu, Yiming Zeng, Qiankun Tang, Beibei Jin, Yinhe Han, and Xiaowei Li. See and think: Disentangling semantic scene completion. *NIPS*, 31, 2018. 2

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

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

[47] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3

[48] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulo, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *CVPR*, pages 3971–3980, 2022. 2

[49] 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

[50] Mong H Ng, Kaahan Radia, Jianfei Chen, Dequan Wang, Ionel Gog, and Joseph E Gonzalez. Bev-seg: Bird's eye view semantic segmentation using geometry and semantic point cloud. *arXiv preprint arXiv:2006.11436*, 2020. 1

[51] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 5

[52] 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

[53] 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. 5

[54] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 1

[55] Charles R Qi, Yin Zhou, Mahyar Najibi, Pei Sun, Khoa Vo, Boyang Deng, and Dragomir Anguelov. Offboard 3d object detection from point cloud sequences. In *CVPR*, pages 6134–6144, 2021. 1

[56] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 2

[57] Christoph B Rist, David Emmerichs, Markus Enzweiler, and Dariu M Gavrila. Semantic scene completion using local deep implicit functions on lidar data. *TPAMI*, 44(10):7205–7218, 2021. 2

[58] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 2022. 2

[59] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *ThreeDV*, 2020. 2, 5

[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. 6

[61] Chang Shu, Kun Yu, Zhixiang Duan, and Kuiyuan Yang. Feature-metric loss for self-supervised learning of depth and egomotion. In *ECCV*, 2020. 6

[62] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *CVPR*, pages 1746–1754, 2017. 2

[63] 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, 5, 7, 8

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

[65] 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. *arXiv preprint arXiv:2106.10689*, 2021. 2, 4

[66] 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. *arXiv preprint arXiv:2303.03991*, 2023. 1, 2

[67] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Forknet: Multi-branch volumetric semantic completion from a single depth image. In *ICCV*, pages 8608–8617, 2019. 2

[68] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 2, 4

[69] 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. 2, 6, 7

[70] Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. SynSin: End-to-end view synthesis from a single image. In *CVPR*, 2020. 6, 7

[71] Felix Wimbauer, Nan Yang, Christian Rupprecht, and Daniel Cremers. Behind the scenes: Density fields for single view reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9076–9086, 2023. 1, 2, 3, 6

[72] Shun-Cheng Wu, Keisuke Tateno, Nassir Navab, and Federico Tombari. Scfusion: Real-time incremental scene reconstruction with semantic completion. In *3DV*, pages 801–810. IEEE, 2020. 2

[73] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In *AAAI*, pages 3101–3109, 2021. 2

[74] Bin Yang, Min Bai, Ming Liang, Wenyuan Zeng, and Raquel Urtasun. Auto4d: Learning to label 4d objects from sequential point clouds. *arXiv preprint arXiv:2101.06586*, 2021. 1

[75] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 2, 6, 7

[76] Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *CVPR*, 2018. 2

[77] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, pages 1020–1031, 2023. 7

[78] Jiahui Zhang, Hao Zhao, Anbang Yao, Yurong Chen, Li Zhang, and Hongen Liao. Efficient semantic scene completion network with spatial group convolution. In *ECCV*, pages 733–749, 2018. 2

[79] Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Critical regularizations for neural surface reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6270–6279, 2022. 5

[80] Pingping Zhang, Wei Liu, Yinjie Lei, Huchuan Lu, and Xiaoyun Yang. Cascaded context pyramid for full-resolution 3d semantic scene completion. In *ICCV*, pages 7801–7810, 2019. 2

[81] 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

[82] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2304.05316*, 2023. 5

[83] Min Zhong and Gang Zeng. Semantic point completion network for 3d semantic scene completion. In *ECAI 2020*, pages 2824–2831. IOS Press, 2020. 2

[84] Kaichen Zhou, Lanqing Hong, Changhao Chen, Hang Xu, Chaoqiang Ye, Qingyong Hu, and Zhenguo Li. Devnet: Self-supervised monocular depth learning via density volume construction. In *ECCV*, pages 125–142. Springer, 2022. 6

[85] 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

[86] Zhongkai Zhou, Xinnan Fan, Pengfei Shi, and Yuanxue Xin. R-MSFM: Recurrent multi-scale feature modulation for monocular depth estimating. In *ICCV*, 2021. 6