

GaussianOcc: Fully Self-supervised and Efficient 3D Occupancy Estimation with Gaussian Splatting

Wanshui Gan^{1,2,*} Fang Liu^{1,*} Hongbin Xu³ Ningkai Mo⁴ Naoto Yokoya^{1,2,†}

¹The University of Tokyo, ²RIKEN, ³South China University of Technology

⁴Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences

* Equal contribution, † Corresponding author

{wanshuigan, fangliu2896, hongbinxu1013, nk.mo19941001}@gmail.com

yokoya@k.u-tokyo.ac.jp

Abstract

We introduce GaussianOcc, a systematic method that investigates the two usages of Gaussian splatting for fully self-supervised and efficient 3D occupancy estimation in surround views. First, traditional methods for self-supervised 3D occupancy estimation still require ground truth 6D poses from sensors during training. To address this limitation, we propose Gaussian Splatting for Projection (GSP) module to provide accurate scale information for fully self-supervised training from adjacent view projection. Additionally, existing methods rely on volume rendering for final 3D voxel representation learning using 2D signals (depth maps, semantic maps), which is both time-consuming and less effective. We propose Gaussian Splatting from Voxel space (GSV) to leverage the fast rendering properties of Gaussian splatting. As a result, the proposed GaussianOcc method enables fully self-supervised (no ground truth pose) 3D occupancy estimation in competitive performance with low computational cost (2.7 times faster in training and 5 times faster in rendering). The relevant code will be available in <https://github.com/GANWANSHUI/GaussianOcc.git>.

1. Introduction

In recent years, both vision-centric and grid-centric perception approaches have gained significant attention in the autonomous driving industry and academia [35, 39]. Among these, surround view 3D occupancy estimation [32, 45, 48, 51] has emerged as a core perception task and a promising alternative to bird’s-eye view (BEV) methods [9, 29, 30]. To facilitate 3D occupancy estimation, several benchmarks have been developed for supervised train-

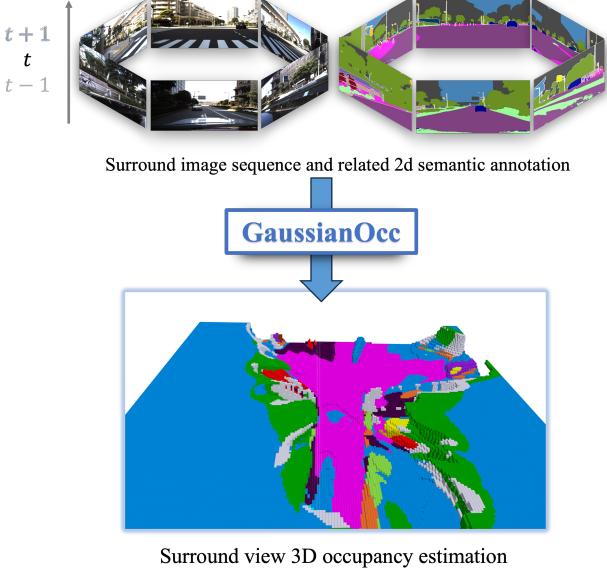


Figure 1. **Problem setting of GaussianOcc.** Given the surround image sequence and its corresponding 2D semantic annotation, GaussianOcc is able to perform 3D occupancy estimation without the need for the ground truth occupancy label and ground truth 6D pose for training.

ing [41–43, 45], though these require substantial effort in 3D annotation. To reduce the burden of annotation, self-supervised [5, 12, 18, 24, 54] and weakly-supervised [37] learning approaches based on volume rendering have been proposed [11, 36, 46]. Volume rendering allows 3D representation learning using 2D supervision signals, such as 2D semantic maps and depth maps, thereby eliminating the

need for extensive 3D annotation during training.

The existing methods [24, 54] achieve self-supervised learning through volume rendering, where the 2D semantic map supervision is derived from open-vocabulary semantic segmentation [55], and the depth map supervision is obtained from self-supervised depth estimation [13]. However, these approaches face two significant limitations. First, volume rendering is performed at real-world scale, which requires the availability of ground truth 6D pose to calculate the multi-view photometric loss across sequential images. Second, the inefficiency in volume rendering poses a challenge, the same as in novel view synthesis tasks [1, 19, 27], due to the dense sampling operation required. These limitations impede the development of a more general and efficient paradigm for self-supervised 3D occupancy estimation.

To address the aforementioned limitations, we explore a fully self-supervised and efficient approach to 3D occupancy estimation using Gaussian splatting [1, 27]. Specifically, we introduce the use of Gaussian splatting to perform cross-view splatting, where the rendered image constructs a cross-view loss that provides scale information during joint training with the 6D pose network. This eliminates the need for a ground truth 6D pose during training. To improve rendering efficiency, we move away from the dense sampling required in traditional volume rendering. Instead, we propose performing Gaussian splatting directly from the 3D voxel space. In this approach, each vertex in the voxel grid is treated as a 3D Gaussian, and we directly optimize the attributes of these Gaussians—such as semantic features and opacity—within the voxel space. Through this novel approach, our proposed method, GaussianOcc, makes a progress toward fully self-supervised and efficient 3D occupancy estimation, as outlined in Figure 1.

The main contributions of this work are summarized as follows.

- We introduce the first fully self-supervised method for efficient surrounding-view 3D occupancy estimation, featuring the exploration of the Gaussian splatting.
- We propose Gaussian splatting for cross-view projection, which could provide the scale information to get rid of the need of a ground truth 6D pose during the training.
- We propose Gaussian splatting from voxel space achieving competitive performance with 2.7 times faster training and 5 times faster rendering compared with the volume rendering.

2. Related work

2.1. Surround view depth estimation

The surround view setting offers an ego-centric 360-degree perception solution [3, 14, 35, 39]. [17] introduces the surround view benchmark in the supervised setting, which

learns the depth scale directly from the ground truth depth map. FSM [16] is a pioneering work in scale-aware surround view depth estimation relying on stereo constraint [53]. However, subsequent studies [28, 44] have found that reproducing the performance of [16] is challenging. Surrounddepth [44] improves the scale supervision signal using sparse point clouds from Structure-from-Motion (SfM). Building on the spatial and temporal constraints in FSM, [28] introduces a volume feature fusion module to enhance performance. For better performance, [38] proposes the temporal offline refinement strategy based on the multiple cameras and monocular depth refinement. These works [16, 28, 38, 44] all use traditional projection and index interpolation for cross-view synthesis, computing the loss between the synthesized view and target images. In contrast, our approach employs Gaussian splatting with a dedicated design, achieving better performance compared to the solely spatial constraint.

2.2. Surround view 3D occupancy estimation

Surround view 3D occupancy estimation has gained significant attention in recent years, with several benchmarks based on the Nuscene dataset [41, 43, 45]. In addition to the advanced architectures being proposed [31, 32, 51, 56], another research trend involves utilizing volume rendering for 3D supervision with 2D supervision [12, 24, 37, 54]. SimpleOcc [12] pioneered the use of volume rendering for 3D occupancy estimation, exploring both supervised and self-supervised learning. RenderOcc [37] extends semantic information for rendering. OccNeRF [54] and SelfOcc [24] share a similar approach by using 2D open-vocabulary semantic models to generate semantic maps for supervision. However, since volume rendering processes are conducted at real-world scale, these self-supervised methods [12, 24, 54] require ground truth 6D poses from sensors to provide the real-world scale for training. In contrast, we are exploring a solution that utilizes the overlap region in adjacent cameras to learn the real-world scale, eliminating the need for ground truth 6D poses.

2.3. 3D Gaussian splatting

3D Gaussian splatting has become a popular method for modeling 3D and 4D scenes using well-posed images [8, 26, 27, 47], which has the property of fast rendering compared to the volume rendering in neural radiance field [11, 19, 36]. In the context of driving scenes, a line of research has focused on scene-specific reconstruction [22, 49, 58]. Our work, however, investigates the function of Gaussian splatting in a generalized setting. In this setting, existing methods generally construct 3D Gaussians from the unprojection of learned 2D Gaussian attributes [7, 33, 40, 57]. We also employ this unprojection approach, but uniquely, our approach constructs cross-view information from adjacent

views to learn scale information through Gaussian splatting projection.

Additionally, two recent works, GaussianFormer [25] and GaussianBeV [6], are closely related to ours in their focus on 3D occupancy estimation and BEV prediction. However, our exploration diverges by focusing on two new properties that Gaussian splatting can contribute to occupancy estimation: scale-aware training and faster rendering.

3. Method

3.1. Preliminaries

In this paper, we investigate the use of 3D Gaussian splatting [27] for fully self-supervised 3D occupancy estimation. 3D Gaussian Splatting (3D-GS) is a technique for modeling static 3D scenes using point primitives. Each point is represented as a scaled Gaussian distribution characterized by a 3D covariance matrix Σ and a mean μ . The Gaussian distribution for a point \mathcal{X} is given by:

$$G(\mathcal{X}) = e^{-\frac{1}{2}(\mathcal{X}-\mu)^T \Sigma^{-1} (\mathcal{X}-\mu)} \quad (1)$$

To enable efficient optimization via gradient descent, the covariance matrix Σ is decomposed into a scaling matrix \mathbf{S} and a rotation matrix \mathbf{R} as follows:

$$\Sigma = \mathbf{R} \mathbf{S} \mathbf{S}^T \mathbf{R}^T \quad (2)$$

The projection of Gaussians from 3D space onto a 2D image plane involves a view transformation \mathbf{W} and the Jacobian of the affine approximation of the projective transformation \mathbf{J} . The 2D covariance matrix Σ' is computed as:

$$\Sigma' = \mathbf{J} \mathbf{W} \Sigma \mathbf{W}^T \mathbf{J}^T \quad (3)$$

Subsequently, an alpha-blend rendering technique, similar to that used in NeRF [36], is applied. This is formulated as:

$$\mathbf{C}_{\text{color}} = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \quad (4)$$

Here, c_i represents the color of each point, and the density α_i is determined by the product of a 2D Gaussian with covariance Σ' and a learned per-point opacity. The color is defined using spherical harmonics (SH) coefficients as described in [10, 27].

In summary, the basic 3D Gaussian Splatting approach characterizes each Gaussian point with the following attributes: (1) a 3D position $\mathcal{X} \in \mathbb{R}^3$, (2) a color defined by SH coefficients $c \in \mathbb{R}^k$ (where k denotes the dimensionality of the SH basis), (3) a rotation represented by a quaternion $r \in \mathbb{R}^4$, (4) a scaling factor $s \in \mathbb{R}_+^3$, and (5) an opacity $\alpha \in [0, 1]$.

3.2. Overview of the proposed GaussianOcc

The overview of the GaussianOcc is illustrated in Figure 2. The key contributions are the introduction of cross-view Gaussian splatting for scale-aware training and voxel grid Gaussian splatting for faster rendering. Unlike the original Gaussian splatting used for scene-specific 3D reconstruction, we explore a setting where Gaussian attributes are well-aligned in both 2D and 3D grids. This approach models the 3D scene in the 2D image plane as a depth map and in 3D grid space as a voxel (occupancy) format.

3.3. Scale-aware training by Gaussian Splatting

Scale from spatial camera rig: Similar to the previous work [16, 44], the scale information is from the surround camera rig. Specifically, the real-world scale could be obtained by leveraging camera extrinsic matrices which is to use spatial photometric loss in the overlap region between two adjacent views, *i.e.*, warping I_t^i to I_t^j :

$$p_t^{i \rightarrow j} = K^j(T^j)^{-1} T^i D_t^i(K^i)^{-1} p_t^i \quad (5)$$

where K^i, T^i are the intrinsic and extrinsic matrices of i th camera, D_t^i is the predict depth map of i th camera, p_t is the corresponding pixel during the warping. The warping operation is achieved by bilinear interpolation with the corresponding $p_t^{i \rightarrow j}$. However, as pointed out by [44], the mapping $p_t^{i \rightarrow j}$ is easy to be out of the image bounds with the weak supervision signal in such a small overlap region. Our practice also verifies this limitation. To provide a stronger supervision signal, [44] proposes to facilitate the Structure-from-Motion (SfM) to extract sparse depth information for the scale-aware training, which is time-consuming and not straightforward. Differently, we propose to use the Gaussian splatting for scale-aware projection to achieve the cross-view stereo constraint.

Gaussian Splatting for Projection (GSP): As illustrated in Figure 2, we use a depth network to predict the Gaussian attributes in 2D grid space, including the depth map, scale map, and rotation map. For each adjacent view, we first calculate the mask in the overlap region, then mask out one side of these overlap regions. Due to the presence of the other side's overlap region, the unprojected 3D scene remains complete, which is critical for providing scale training, as indicated in the experiment section. We then perform splatting rendering on the adjacent views to obtain the rendered image. If the depth map is accurately learned, the rendered image should resemble the original images, providing the necessary scale information for the entire pipeline.

The process of acquiring the overlap mask is shown in Figure 3. The acquisition of the overlap mask is based on volume rendering. We densely sample points along the ray in one view, and a pixel is considered part of the overlap region if more than one sampled 3D point falls within the

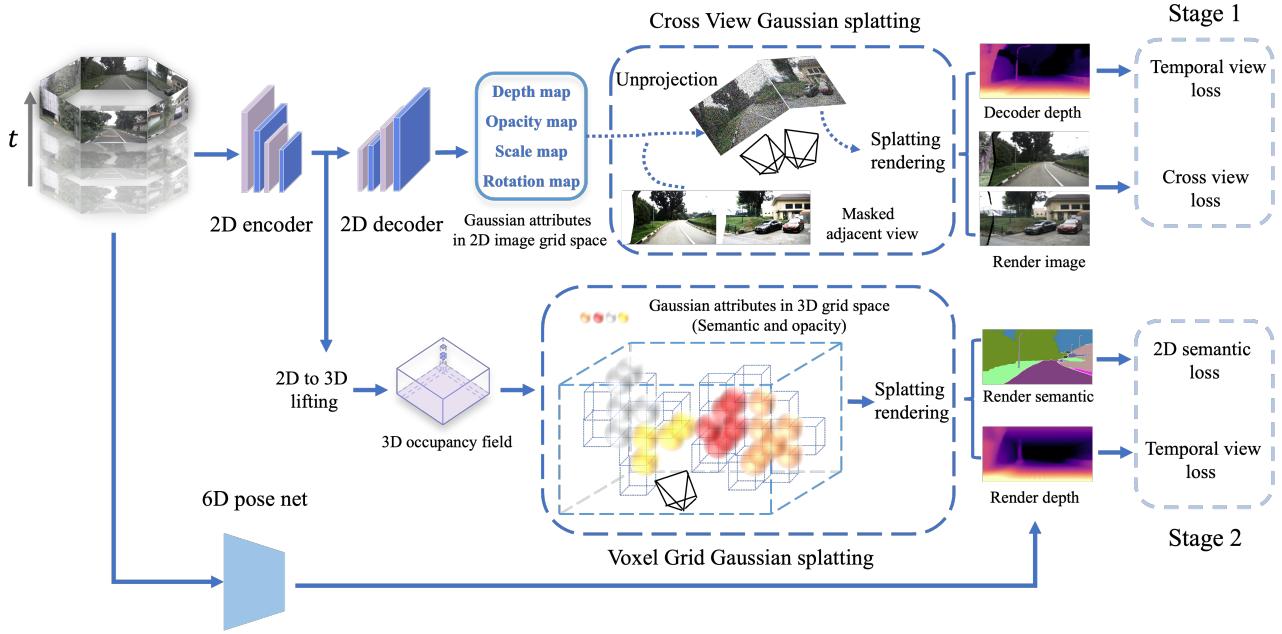


Figure 2. **The overview of the proposed GaussianOcc.** With a sequence of surround images, we employ a U-Net architecture to predict Gaussian attributes in the 2D image grid space for cross-view Gaussian splatting. This approach provides scale information in the joint training with the 6D pose network (Stage 1). For 3D occupancy estimation, we lift the 2D features to a 3D voxel space and propose voxel grid Gaussian splatting for fast rendering (Stage 2). We omit the line to from the 6D pose net to the loss in stage 1 for clarity.

adjacent view. Note that in the DDAD dataset [14], we exclude regions with self-occlusion (such as parts of the vehicle body). Finally, we apply an erosion operation from Open CV library [2] to the mask for purification.

3.4. Fast rendering by Gaussian Splatting

Inefficient performance in volume rendering: For 2D supervision (semantic and depth maps), previous methods [12, 24, 37, 54] employed volume rendering based on dense sampling. While the final 3D voxel representation for modeling the 3D scene is much quicker than the original implicit representation (MLPs) [36], it remains time-consuming, particularly when incorporating semantic map rendering. For example, in OccNeRF [54], the number of sampled points at a resolution of 180×320 is 108,735,066. However, the target optimized points correspond to the vertices in the 3D voxel grid, which total $320 \times 320 \times 24 = 2,457,600$. This redundancy in densely sampled points aids optimization with volume rendering but is highly inefficient.

Gaussian Splatting from Voxel (GSV): As analyzed above, the target optimized points are the vertices in the 3D voxel grid, prompting us to consider directly optimizing these vertices. Therefore, we investigate the use of Gaussian splatting to further accelerate the rendering process. For each vertex in the 3D voxel grid, we obtain the necessary attributes for Gaussian splatting, such as 3D position, semantic information, and opacity. Consequently, we pro-

pose treating each vertex as a 3D Gaussian. Since these vertices are arranged in the 3D voxel space, we can model the 3D scene by optimizing the related Gaussian attributes.

3.5. Loss function

We formulate the the total loss function in:

$$\mathcal{L}_{total} = \mathcal{L}_{temporal} + \mathcal{L}_{cross} + \mathcal{L}_{semantic} + \mathcal{L}_{tv}, \quad (6)$$

$$(\mathcal{L}_{temporal}, \mathcal{L}_{cross}) = \beta \frac{1 - SSIM(I_t, \hat{I}_t)}{2} + (1-\beta) \|I_t - \hat{I}_t\|, \quad (7)$$

where the terms I_t and \hat{I}_t refer to the target image and the corresponding synthesized image, respectively. Note that \hat{I}_t in the temporal-view photometric loss, $\mathcal{L}_{temporal}$, is generated by projecting pixels from the source image using the coordinate index. In contrast, \hat{I}_t in the cross-view photometric loss, \mathcal{L}_{cross} , is derived from our proposed cross-view Gaussian splatting method. $\mathcal{L}_{semantic}$ is the 2D semantic loss. \mathcal{L}_{tv} is the regularization loss for the whole voxel space. The balancing parameter β is set to 0.85 to balance the two loss terms.

4. Experiment

4.1. Tasks, Datasets, and Metric

nuScenes: We perform 3D occupancy estimation and depth estimation using the nuScenes dataset [3]. For 3D oc-

Method	GT Occ.	GT Pose	mIoU mIoU*																		
			barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suff.	sidewalk	terrain	manmade	vegetation				
MonoScene [4]	✓	✗	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 [21]	✓	✗	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 [30]	✓	✗	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 [56]	✓	✗	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 [37]	✓	✗	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 [23]	✓	✗	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 [41]	✓	✗	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 [12]	✗	✓	7.99	7.05	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		
SelfOcc [24]	✗	✓	10.54	9.30	0.15	0.66	5.46	12.54	0.00	0.80	2.10	0.00	0.00	8.25	55.49	26.30	26.54	14.22	5.60		
OccNeRF [54]	✗	✓	10.81	9.54	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		
GaussianOcc	✗	✗	11.26	9.94	1.79	5.82	14.58	13.55	1.30	2.82	7.95	9.76	0.56	9.61	44.59	20.10	17.58	8.61	10.29		

Table 1. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset in mIoU metric.** Since ‘other’ and ‘other flat’ classes are the invalid prompts for open-vocabulary models, we also calculate ‘mIoU’ as the result ignoring the classes that do not consider these two classes during evaluation, while ‘mIoU*’ is the original result. GT Occ. means using the ground truth occupancy label for the supervision. GT Pose is the ground truth pose from the sensor for self-supervised geometry learning.

Method	GT Occ.	GT Pose				RayIoU	RayIoU _{1m, 2m, 4m}				mIoU*	FPS
			Backbone	Input Size	Epoch		26.1	32.9	38.0	39.2		
BEVFormer (4f) [30]	✓	✗	R101	1600×900	24	32.4	26.1	32.9	38.0	39.2	3.0	-
RenderOcc [37]	✓	✗	Swin-B	1408×512	12	19.5	13.4	19.6	25.5	24.4	-	-
SimpleOcc [12]	✓	✗	R101	672×336	12	22.5	17.0	22.7	27.9	31.8	9.7	-
BEVDet-Occ (2f) [20]	✓	✗	R50	704×256	90	29.6	23.6	30.0	35.1	36.1	2.6	-
BEVDet-Occ-Long (8f)	✓	✗	R50	704×384	90	32.6	26.6	33.1	38.2	39.3	0.8	-
FB-Occ (16f) [31]	✓	✗	R50	704×256	90	33.5	26.7	34.1	39.7	39.1	10.3	-
SparseOcc (8f)	✓	✗	R50	704×256	24	34.0	28.0	34.7	39.4	30.1	17.3	-
SparseOcc (16f)	✓	✗	R50	704×256	48	36.1	30.2	36.8	41.2	30.9	12.5	-
OccNeRF [54]	✗	✓	R101	640×384	12	10.49	6.93	10.28	14.26	9.54	10.8	-
GaussianOcc	✗	✗	R101	640×384	12	11.85	8.69	11.90	14.95	9.94	10.8	-

Table 2. **3D Occupancy prediction performance on the Occ3D-nuScenes dataset in RayIoU metric.** GT Occ. means using the ground truth occupancy label for the supervision. GT Pose is the ground truth pose from the sensor for self-supervised geometry learning. “8f” and “16f” mean fusing temporal information from 8 or 16 frames. mIoU* is the mean Intersection over Union for all categories. FPS means frame per second for each method, which is measured on a Tesla A100 GPU.

cupancy estimation, we utilize annotations from Occ3D-nuScenes [41], which includes 600 scenes for training and 150 scenes for evaluation. Each sample has a perception range of [-40m, -40m, -1m, 40m, 40m, 5.4m] with a voxel size of 0.4m. For fair comparison, we use the 2D semantic map provided by OccNeRF [54]. We measure 3D occupancy estimation performance using the mean Intersection over Union (mIoU) metric. For depth estimation, we set the perception range in [-80m, -80m, -1m, 80m, 80m, 6m], while we clamp the ground truth to a range of 0.1m to 80m for evaluation, consistent with OccNeRF and SurroundDepth [44]. We evaluate depth maps using error met-

rics (Abs Rel, Sq Rel, RMSE, RMSE log) and threshold accuracy metrics (δ).

DDAD: We conduct 3D occupancy estimation and depth estimation using the DDAD dataset [14]. Due to the lack of ground truth labels, we only present qualitative results for the fully self-supervised 3D occupancy estimation setting. Following the procedure used with the nuScenes dataset, we obtain the 2D semantic labels as per the OccNeRF pipeline. For depth estimation, we also set the perception range in [-80m, -80m, -1m, 80m, 80m, 6m], but clamp the ground truth to a range of 0.1m to 200m for evaluation, consistent with SurroundDepth [44].

Method	<i>GT pose</i>	<i>Occ.</i>	Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
nuScenes [3]									
FSM [16]	×	×	0.297	-	-	-	-	-	-
FSM* [16]	×	×	0.319	7.534	7.860	0.362	0.716	0.874	0.931
SurroundDepth [44]	×	×	0.280	4.401	7.467	0.364	0.661	0.844	0.917
SA-FSM [50]	×	×	0.272	4.706	7.391	0.355	0.689	0.868	0.929
VFF [28]	×	×	0.289	5.718	7.551	0.348	0.709	0.876	0.932
R3D3 [38]	×	×	0.253	4.759	7.150	-	0.729	-	-
GaussianOcc \ddagger	×	×	0.258	5.733	7.222	0.343	0.753	0.888	0.934
SimpleOcc [12]	✓	✓	0.224	3.383	7.165	0.333	0.753	0.877	0.930
OccNeRF [54]	✓	✓	0.202	2.883	6.697	0.319	0.768	0.882	0.931
SelfOcc [24]	✓	✓	0.215	2.743	6.706	0.316	0.753	0.875	0.932
GaussianOcc	×	✓	0.211	3.115	7.131	0.326	0.762	0.878	0.931
DDAD [14]									
FSM* [16]	×	×	0.228	4.409	13.433	0.342	0.687	0.870	0.932
VFF [28]	×	×	0.218	3.660	13.327	0.339	0.674	0.862	0.932
SurroundDepth [44]	×	×	0.208	3.371	12.977	0.330	0.693	0.871	0.934
SA-FSM [50]	×	×	0.187	3.093	12.578	0.311	0.731	0.891	0.945
R3D3 [38]	×	×	0.162	3.019	11.408	-	0.811	-	-
GaussianOcc \ddagger	×	×	0.212	3.556	12.564	0.320	0.701	0.888	0.944
GaussianOcc	×	✓	0.228	3.854	14.326	0.357	0.660	0.853	0.922

Table 3. Comparisons for self-supervised multi-camera depth estimation on the nuScenes [3] and DDAD datasets [14]. The results are averaged over all views without median scaling at test time. ‘FSM*’ is the reproduced result in [28]. GaussianOcc \ddagger represents the depth estimation result from Stage 1. *Occ.* represents the ability of the method to predict the 3D occupancy.

Scale aware training in [44]		Scale aware training by ours			Abs Rel	Sq Rel	RMSE	RMSE log	$\delta < 1.25$	$\delta < 1.25^2$	$\delta < 1.25^3$
<i>Spatial loss</i>	<i>SFM loss</i>	<i>GS loss</i>	<i>Mask</i>	<i>Erode</i>	<i>Refine</i>						
✓*	✓					0.970	13.702	17.931	3.654	0.000	0.000
						0.921	30.393	14.444	0.808	0.057	0.187
						0.280	4.401	7.467	0.364	0.661	0.844
						0.672	25.405	11.999	0.568	0.419	0.808
	✓	✓	✓	✓	✓	0.798	11.571	15.251	1.472	0.006	0.015
						0.293	7.127	7.536	0.376	0.743	0.876
						0.281	6.986	7.347	0.354	0.766	0.885
						0.258	5.733	7.222	0.343	0.753	0.888

Table 4. Ablation study for scale-aware depth estimation on the Nuscene dataset [15]. The scores are averaged over all views **without** median-scaling at test time. ✓* means the result from the original paper and ✓ means the result from our implementation in New-CRFs network [52]. *Spatial loss* denotes the spatial context loss used in [44] and *SFM loss* denotes the sparse depth pretrain from the off-line Structure-from-Motion process. *GS loss* means using the spatial context constraint by our proposed Gaussian splatting for projection. *Mask* represents using the mask-out strategy before the unprojection. *Erode* means use the erode process to the binary overlap mask and *Refine* is the refinement of the depth estimation network by fixing the 6D pose net.

4.2. Implementation details

Network details: For U-Net architecture, we adopt the New-CRFs [52] as the depth estimation network, which is based on the Swin Transformer [34]. The 6D pose network is the same as that used in SurroundDepth [44]. For the 2D to 3D lifting, we follow the approach used in SimpleOcc [12]. In depth estimation benchmark, we use the network proposed by SimpleOcc, where the final output size is $256 \times 256 \times 16$. In our Gaussian splatting setting, we further upsample the final output to $512 \times 512 \times 32$ for improved performance, which only requires minimal additional com-

putational cost. For occupancy estimation, we use the same network as OccNeRF [54] to ensure a fair comparison.

Training details: We propose a two-stage training process for fully self-supervised 3D occupancy estimation. In the first stage, we jointly train the depth estimation network and the 6D pose network. In the second stage, we train the 3D occupancy network without optimizing the 6D pose network. During the first stage, we train the models for 8 epochs on the nuScenes dataset and 12 epochs on the DDAD dataset. In the second stage, we train the models for 12 epochs on both the nuScenes and DDAD datasets.

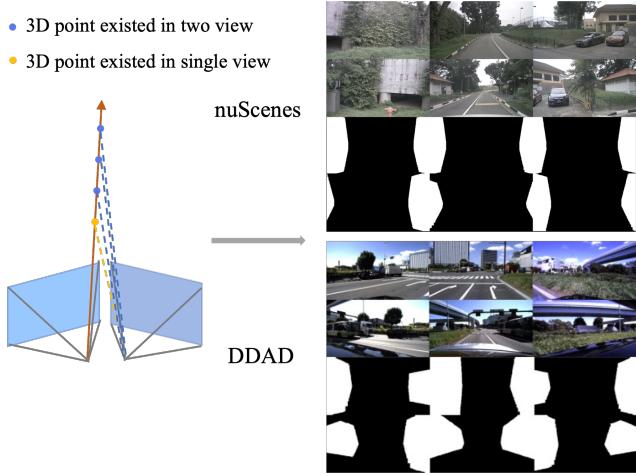


Figure 3. **Overlap mask in nuScenes [3] and DDAD [14].**

The optimizer and learning rate adjustment strategy are the same as those used in SimpleOcc [12] and OccNeRF [54].

4.3. Main results

3D occupancy estimation in mIoU metric: As shown in Table 1, the proposed GaussianOcc achieves the best performance compared with other self-supervised methods, without the need for ground truth occupancy labels and ground truth 6D poses for training. It’s worth noting that SelfOcc [24] differs slightly from methods like SimpleOcc, OccNeRF, and our approach. Unlike these methods, SelfOcc does not predict semantic information directly in the 3D occupancy space. Instead, it attaches 2D semantic information to the 3D occupied voxels, where the 2D semantic data is sourced from a third-party open-vocabulary model [55].

3D occupancy estimation in RayIoU metric: We also compare our method using a newly proposed metric, RayIoU, introduced by the recent work [32]. The RayIoU is a ray-based evaluation metric to solve the inconsistency penalty along the depth axis raised in traditional voxel-level mIoU criteria. As shown in Table 2, our approach also outperforms OccNeRF [54]. It’s important to note that the FPS is calculated excluding rendering time. Since GaussianOcc and OccNeRF utilize the same network architecture, they share the same inference time when the rendering process is not taken into account.

3D occupancy estimation in DDAD: Thanks to our fully self-supervised learning approach, we are able to perform 3D occupancy estimation on a similar dataset, DDAD [14]. To the best of our knowledge, our work is the first to introduce 3D occupancy estimation on this dataset, although we cannot provide quantitative results due to the lack of 3D annotations. We present the visualization results in Figure 4 and more results could be found in the supplement material.

Depth estimation: We present a comparison of depth estimation results in Table 3 for both the nuScenes and

DDAD datasets. In the first stage of monocular depth estimation, GaussianOcc \ddagger achieves top performance on the nuScenes dataset and delivers competitive results on the DDAD dataset. It is important to note that methods such as SurroundDepth [44] and SA-FSM [50] rely on third-party sparse point cloud extraction. Additionally, R3D3 [38] is a temporal offline refinement method that requires multi-frame optimization, as discussed in [24]. In the second stage, which involves depth estimation from rendering, our method also achieves competitive results compared to those trained with ground truth poses. Besides, we observe that the rendered depth in stage two outperforms the depth results from stage 1 on the nuScenes dataset, whereas the opposite is true for the DDAD dataset. This discrepancy might be attributed to differences in perception range—80 meters in nuScenes versus 200 meters in DDAD.

4.4. Ablation Study

Scale aware training: To demonstrate the effectiveness of the proposed scale-aware training using Gaussian Splatting, we conducted experiments comparing our method with existing approaches that use SFM loss for training [44]. As indicated in Table 4, we implemented scale-aware training from [44] on our network [52], but the results were suboptimal, and our proposed method outperformed the results reported in the original paper. In our method, we observed that a naive implementation of Gaussian splatting without a masking strategy is ineffective because it lacks cross-view rendering, which is essential for applying cross-view constraints. To enhance performance, we introduced an erosion operation on the binary mask to purify it, excluding regions that may fall outside the overlap area. This step ensures better alignment during training. Finally, we refined the depth estimation by fixing the 6D pose network and disabling the cross-view loss. This refinement helps reduce artifacts at the edges of the overlap region. We present the cross-view splatting render result in Figure 5, where the rendering result in overlap regions is reasonable in the correct depth estimation from cross view.

6D pose and training strategy: Building on the success of scale-aware training, we compared the depth estimation performance of our proposed two-stage training method with that of models trained using ground truth (GT) pose. As shown in Table 5, our two-stage approach delivers competitive results when compared to training with GT pose, indicating that the predicted pose is of high quality. Additionally, the GT pose provided by the nuScenes dataset is limited by sensor inaccuracies, particularly a lack of vertical movement. In contrast, the learned 6D pose in our method does not suffer from this limitation. Furthermore, in our one-stage experiment, we found that this approach was not effective. The cross-view loss in the 3D voxel space led to local optimization, which failed to generalize predictions to

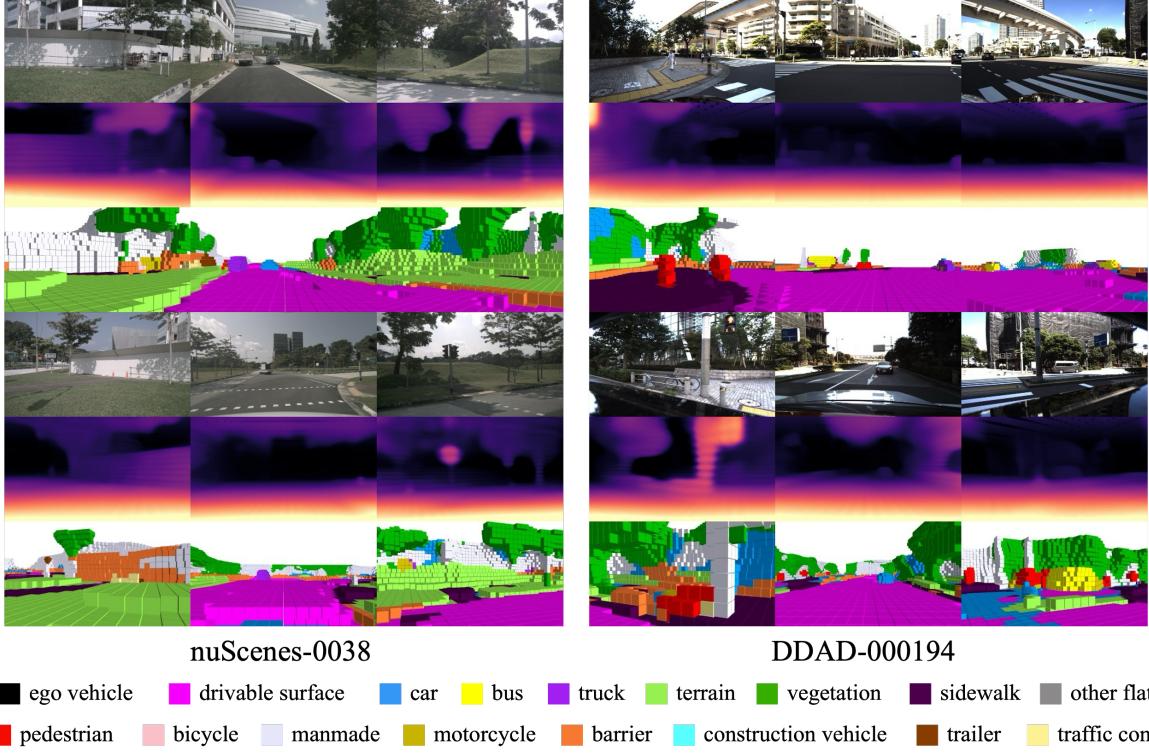


Figure 4. The visualization of the render depth map and 3D occupancy prediction on nuScenes and DDAD dataset.

non-overlapping regions as indicated in Figure 6. However, this issue did not arise in the depth maps generated by the 2D decoder in stage one.

Volume rendering and Splatting rendering: In Table 6, we examine the performance of volume rendering (VR) and splatting rendering (SR) in the voxel space. For the scale of each 3D Gaussian, we set a uniform scale s for all Gaussians, considering the well-arranged positions of vertices within the voxel grid. Since the scale s is consistent across all three dimensions, we set the rotation \mathbf{R} as the identity matrix. We first explored the optimal scale and found that a scale of 0.1 produced the best results. Notably, the performance of VR and SR was very close when using this scale. The render depth maps are shown in Figure 6. The depth map by the volume rendering is smoother thanks to dense sampling and the depth maps with splatting rendering have the graininess effect, especially at the small scale factor.

Pose and render types ablation study in occupancy task: In Table 7, we conduct an ablation study to evaluate the impact of using ground truth (GT) versus learned poses in the occupancy task, with training conducted using different rendering methods. The results show that splatting rendering achieves superior performance in both occupancy metrics (mIoU) and depth metrics. One observation is that depth estimation results from 3D semantic learning with volume rendering are significantly worse than those in Table 3 and the proposed splatting rendering method maintains consis-

Pose type	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$
GT pose	0.214	3.362	7.127	0.771
Two stages	0.211	3.115	7.131	0.762
One stage	0.946	17.008	16.397	0.103

Table 5. Comparison of the ground truth pose (GT pose) and our learned pose (Two stages) on the depth estimation task [15]. We also conducted the experiment that directly uses the cross-view loss to render a depth map (One stage).

tent performance, which indicates the splatting rendering in voxel grid may contribute better 3D geometry.

Rendering efficiency analysis: Finally, we analyze rendering efficiency in Table 8. The results indicate that volume rendering consumes 5 times more rendering time compared to splatting rendering, and as the resolution increases, both rendering time and GPU consumption rise significantly. In contrast, splatting rendering shows no significant increase in computational cost with higher resolutions, highlighting its efficiency and scalability.

5. Conclusion

In this paper, we introduce GaussianOcc, a fully self-supervised and efficient approach for 3D occupancy estimation. Through the carefully designed cross-view splatting rendering, we can accurately learn the real scale in depth

Render type	Abs Rel	Sq Rel	RMSE	$\delta < 1.25$
VR	0.215	3.508	7.113	0.775
SR ($s = 0.05$)	0.223	3.694	7.246	0.761
SR ($s = 0.1$)	<u>0.217</u>	<u>3.504</u>	<u>7.152</u>	<u>0.770</u>
SR ($s = 0.15$)	0.217	3.406	7.204	0.763

Table 6. Comparison of the render result between the volume rendering (VR) [54] and splatting rendering (SR, Ours) on depth estimation task [15]. The coefficient s means the scale of the 3D Gaussian. We use GT pose for this set of the ablation study.

Pose type	mIoU	Abs Rel	Sq Rel	$\delta < 1.25$
GT pose (VR)	10.81	0.456	12.682	0.704
GT pose (SR)	11.30	0.225	4.339	0.787
Learned pose (VR)	11.19	0.506	15.577	0.684
Learned pose (SR)	11.43	0.219	3.625	0.778

Table 7. Comparison of ground truth pose (GT Pose) and our learned pose (Two stages) on 3D occupancy estimation task [15] in volume rendering (VR) and splatting rendering (SR).

Render type	Render resolution and time (s)			Training time (h)
	180 × 320	240 × 520	360 × 640	
VR	≈ 0.85	≈ 1.57	N/A	≈ 2.68
SR	≈ 0.17	≈ 0.17	≈ 0.17	≈ 1

Table 8. Comparison of rendering efficiency between volume rendering (VR) [54] and splatting rendering (SR, Ours) on 3D occupancy estimation task [15]. The render time is calculated from surround 6 images. "N/A" indicates out-of-memory errors running in NVIDIA A 100 (40 GB). Training time is averaged per epoch.

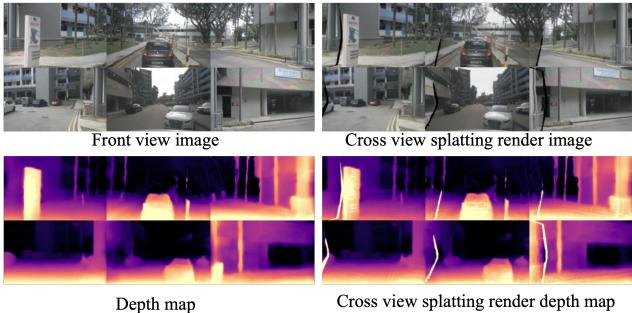


Figure 5. The visualization for the depth map, splatting render image and depth map from cross view in stage 1.

and the 6D pose net, enabling effective self-supervised 3D occupancy learning. Additionally, the proposed Gaussian splatting in voxel grids outperforms volume rendering in 3D occupancy estimation while reducing computational cost. We validate our method on the nuScenes and DDAD datasets, demonstrating its strong generalization ability to surround view scenes.

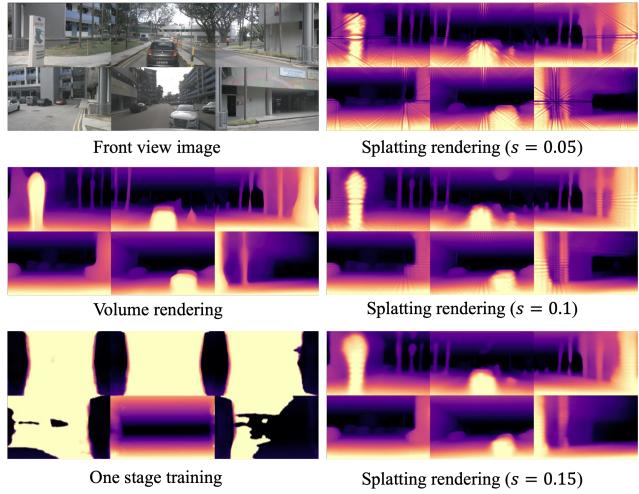


Figure 6. The comparison for the depth map in the different setting, corresponding to the training strategy in Table 5 and rendering type in Table 6.

Appendix

A. More implementation details

The detailed parameter setting in depth estimation network [52]. During the joint depth and 6D pose training in stage 1, we predict the 3D Gaussian parameters alongside the 2D depth map. Since the Gaussian parameters are well-arranged in the 2D image plane prior to unprojection, we maintain equal scaling across all three dimensions of each 3D Gaussian and constrain the maximum scale to 0.02. Given that the scale s is uniform across all dimensions, we set the rotation matrix \mathbf{R} to the identity matrix. Additionally, we assign an opacity value of 1 to each 3D Gaussian, ensuring that every 2D depth value corresponds to a valid point in 3D space.

The detailed parameter setting in voxel grid splatting rendering. We chose a fixed scale for each grid vertex to ensure a well-arranged structure that accurately models the 3D space. If we were to use a learnable scale, it could lead to a situation where the scale is small but the opacity is large, which might not be captured in the rendered depth map and semantic map but could still impact the 3D occupancy result. Therefore, using a fixed scale is sufficient for optimization, as demonstrated in the results presented in the main paper.

The detailed parameter setting in training. We follow the training setting as OccNeRF [54], the resolution of input images and rendered depth maps are set as 384 × 640 and 180 × 320 respectively. All experiments are conducted on 8 NVIDIA A100 (40 GB).

Detailed definition of depth map metric. Following the depth estimation task [44], we report the depth map

evaluation with the following metrics,

$$\begin{aligned}
 \text{Abs Rel:} & \frac{1}{|M|} \sum_{d \in M} |\hat{d} - d^*| / d^*, \\
 \text{Sq Rel:} & \frac{1}{|M|} \sum_{d \in M} \|\hat{d} - d^*\|^2 / d^*, \\
 \text{RMSE:} & \sqrt{\frac{1}{|M|} \sum_{d \in M} \|\hat{d} - d^*\|^2}, \\
 \text{RMSE log:} & \sqrt{\frac{1}{|M|} \sum_{d \in M} \|\log \hat{d} - \log d^*\|^2}, \\
 \delta < t : \% \text{ of } d \text{ s.t. } \max\left(\frac{\hat{d}}{d^*}, \frac{d^*}{\hat{d}}\right) = \delta < t,
 \end{aligned} \tag{8}$$

where M is the valid pixel, \hat{d} is the ground truth depth and d^* is the predicted depth.

B. More experiment results

More visualization. We provide more visualization for nuScenes dataset in Figure A and DDAD dataset in Figure B. Please check the attached videos for sequence visualization.

C. Limitation and future work

The proposed method achieves reasonable predictions in most scenes; however, we observe that some cases still present challenges, as shown in Figure C. Specifically, in the DDAD dataset, incorrect predictions occur in the back camera in certain situations as marked with the red circle, where the drivable surface is mistakenly projected into the car due to extensive self-occlusion. Notably, this issue is absent in the nuScenes dataset, which has less self-occlusion. We believe that this problem could be mitigated with better 2D semantic maps for supervision, which warrants further investigation. Additionally, we plan to explore the use of splatting rendering for 3D occupancy flow in future work.

References

- [1] Yanqi Bao, Tianyu Ding, Jing Huo, Yaoli Liu, Yuxin Li, Wenbin Li, Yang Gao, and Jiebo Luo. 3d gaussian splatting: Survey, technologies, challenges, and opportunities. *arXiv preprint arXiv:2407.17418*, 2024. 2
- [2] Gary Bradski. The opencv library. *Dr. Dobb's Journal: Software Tools for the Professional Programmer*, 25(11):120–123, 2000. 4
- [3] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Lioung, 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, 4, 6, 7
- [4] Anh-Quan Cao and Raoul de Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 5
- [5] Anh-Quan Cao and Raoul de Charette. Scenerf: Self-supervised monocular 3d scene reconstruction with radiance fields. In *ICCV*, pages 9387–9398, 2023. 1
- [6] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. *arXiv preprint arXiv:2407.14108*, 2024. 3
- [7] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024. 2
- [8] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d-rotor gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024. 2
- [9] Shaoheng Fang, Zi Wang, Yiqi Zhong, Junhao Ge, and Si-heng Chen. Tbp-former: Learning temporal bird's-eye-view pyramid for joint perception and prediction in vision-centric autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1368–1378, 2023. 1
- [10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinrong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5501–5510, 2022. 3
- [11] Wanshui Gan, Hongbin Xu, Yi Huang, Shifeng Chen, and Naoto Yokoya. V4d: Voxel for 4d novel view synthesis. *IEEE Transactions on Visualization and Computer Graphics*, 2023. 1, 2
- [12] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A comprehensive framework for 3d occupancy estimation in autonomous driving. *IEEE Transactions on Intelligent Vehicles*, 2024. 1, 2, 4, 5, 6, 7
- [13] 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
- [14] Vitor Guizilini, Rares Ambrus, Sudeep Pillai, Allan Raventos, and Adrien Gaidon. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2485–2494, 2020. 2, 4, 5, 6, 7
- [15] 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, 8, 9
- [16] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *RAL*, 7(2):5397–5404, 2022. 2, 3, 6
- [17] Xianda Guo, Wenjie Yuan, Yunpeng Zhang, Tian Yang, Chenming Zhang, Zheng Zhu, and Long Chen. A sim-

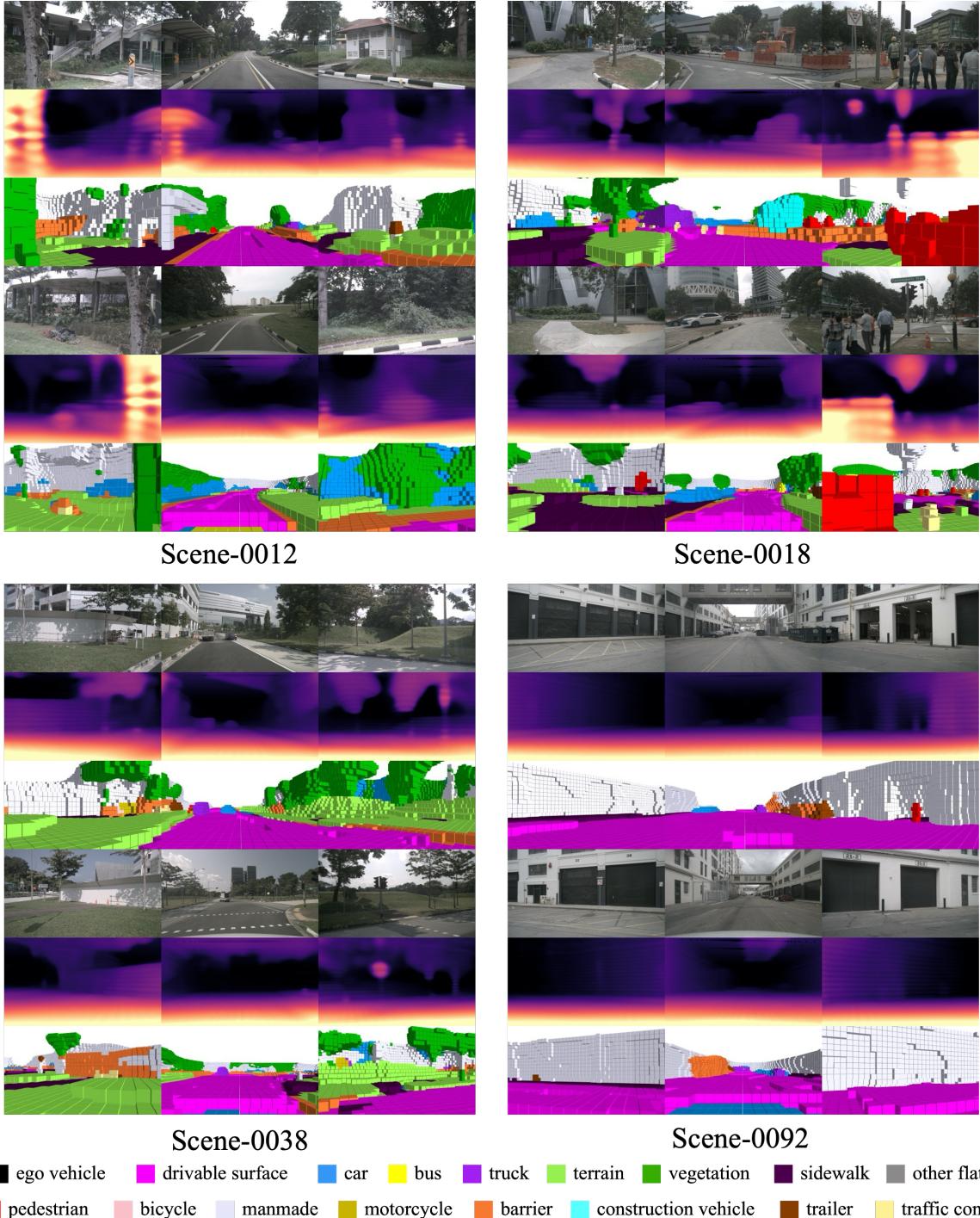


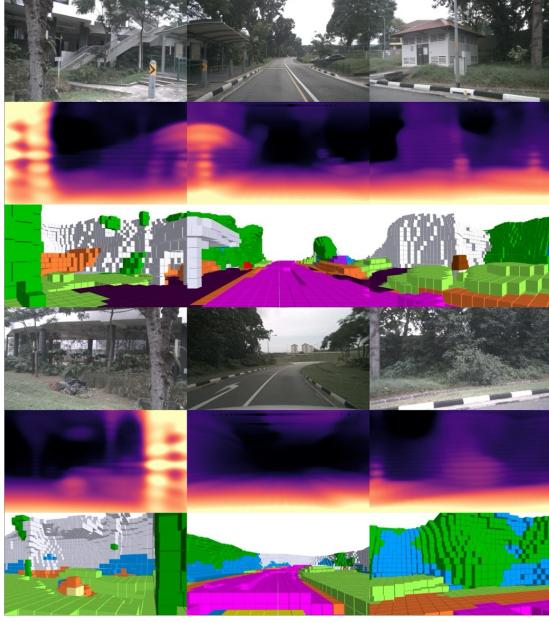
Figure A. The visualization of the render depth map and 3D occupancy prediction on nuScenes dataset.

ple baseline for supervised surround-view depth estimation. *arXiv preprint arXiv:2303.07759*, 2023. [2](#)

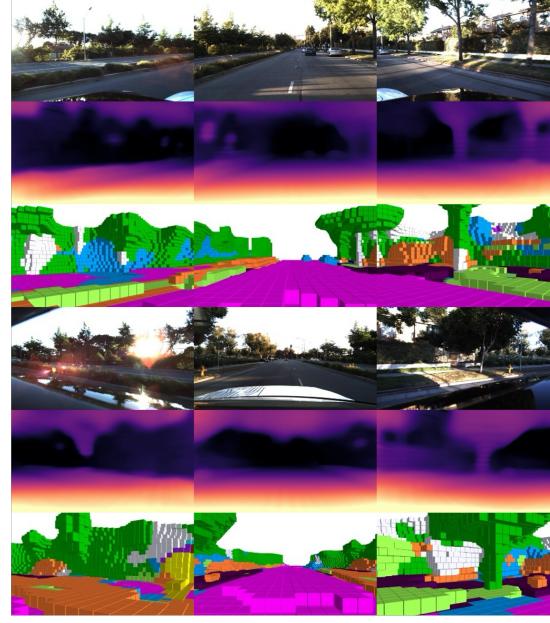
- [18] Adrian Hayler, Felix Wimbauer, Dominik Muhle, Christian Rupprecht, and Daniel Cremers. S4c: Self-supervised semantic scene completion with neural fields. In *2024 International Conference on 3D Vision (3DV)*, pages 409–420.

IEEE, 2024. [1](#)

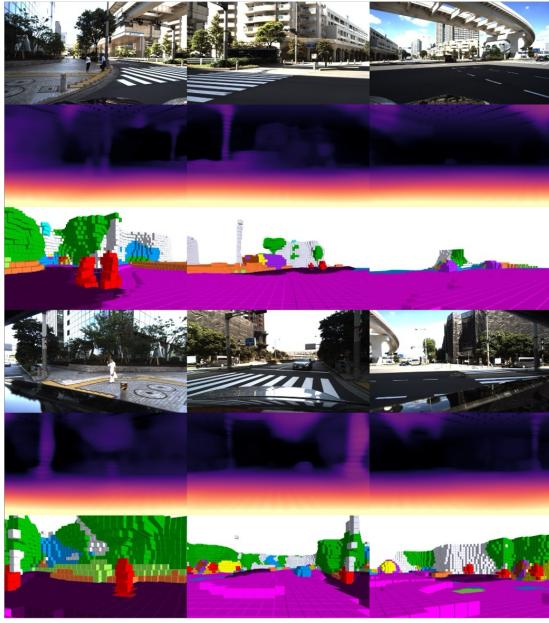
- [19] Lei He, Leheng Li, Wenchao Sun, Zeyu Han, Yichen Liu, Sifa Zheng, Jianqiang Wang, and Keqiang Li. Neural radiance field in autonomous driving: A survey. *arXiv preprint arXiv:2404.13816*, 2024. [2](#)
- [20] Junjie Huang and Guan Huang. Bevdet4d: Exploit tempo-



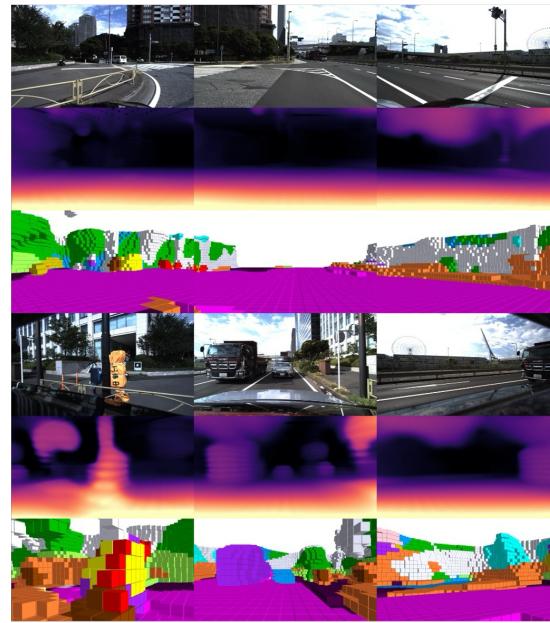
DDAD-000159



DDAD-000180



DDAD-000194



DDAD-000195

■ ego vehicle	■ drivable surface	■ car	■ bus	■ truck	■ terrain	■ vegetation	■ sidewalk	■ other flat
■ pedestrian	■ bicycle	■ manmade	■ motorcycle	■ barrier	■ construction vehicle	■ trailer	■ traffic cone	

Figure B. The visualization of the render depth map and 3D occupancy prediction on DDAD dataset

ral cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 5

[21] 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. 5

[22] Nan Huang, Xiaobao Wei, Wenzhao Zheng, Pengju An,

Ming Lu, Wei Zhan, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. S3 gaussian: Self-supervised street gaussians for autonomous driving. *arXiv preprint arXiv:2405.20323*, 2024. 2

[23] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d se-

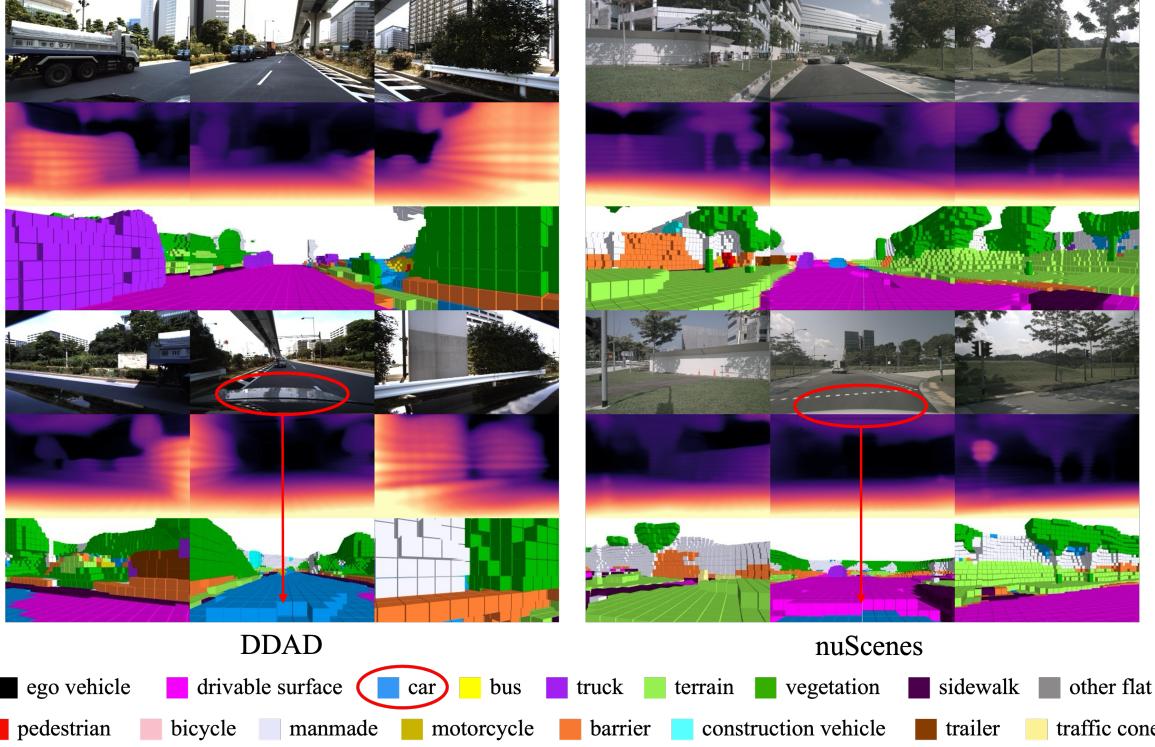


Figure C. Some wrong predictions due to the large self-occlusion on DDAD dataset.

mantic occupancy prediction. In *CVPR*, pages 9223–9232, 2023. 5

- [24] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024. 1, 2, 4, 5, 6, 7
- [25] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024. 3
- [26] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv preprint arXiv:2311.12897*, 2023. 2
- [27] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3
- [28] 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. 2, 6
- [29] 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
- [30] Zhiqi Li, Wenhui Wang, Hongyang Li, Enze Xie, Chong-hao 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, 5

- [31] 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. 2, 5
- [32] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d panoptic occupancy prediction. *arXiv preprint arXiv:2312.17118*, 2023. 1, 2, 7
- [33] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Fast generalizable gaussian splatting reconstruction from multi-view stereo. *arXiv preprint arXiv:2405.12218*, 2024. 2
- [34] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 6
- [35] Yuexin Ma, Tai Wang, Xuyang Bai, Huitong Yang, Yuenan Hou, Yaming Wang, Yu Qiao, Ruigang Yang, Dinesh Manocha, and Xinge Zhu. Vision-centric bev perception: A survey. *arXiv preprint arXiv:2208.02797*, 2022. 1, 2
- [36] 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. 1, 2, 3, 4

- [37] 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. 1, 2, 4, 5
- [38] 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. 2, 6, 7
- [39] Yining Shi, Kun Jiang, Jiusi Li, Junze Wen, Zelin Qian, Mengmeng Yang, Ke Wang, and Diange Yang. Grid-centric traffic scenario perception for autonomous driving: A comprehensive review. *arXiv preprint arXiv:2303.01212*, 2023. 1, 2
- [40] Stanislaw Szymanowicz, Christian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10208–10217, 2024. 2
- [41] 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
- [42] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023.
- [43] 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
- [44] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surround-depth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *CoRL*, pages 539–549. PMLR, 2023. 2, 3, 5, 6, 7, 9
- [45] 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
- [46] 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
- [47] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20310–20320, 2024. 2
- [48] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *arXiv preprint arXiv:2405.05173*, 2024. 1
- [49] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 2
- [50] Yuchen Yang, Xinyi Wang, Dong Li, Lu Tian, Ashish Sirasao, and Xun Yang. Towards scale-aware full surround monodepth with transformers. *arXiv preprint arXiv:2407.10406*, 2024. 6, 7
- [51] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zong-dai Liu, Jiangyong Yu, Dawei Yang, Hui Li, and Yan Chen. Flashocc: Fast and memory-efficient occupancy prediction via channel-to-height plugin. *arXiv preprint arXiv:2311.12058*, 2023. 1, 2
- [52] Weihao Yuan, Xiaodong Gu, Zuozhuo Dai, Siyu Zhu, and Ping Tan. Newcrfs: Neural window fully-connected crfs for monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 6, 7, 9
- [53] Zhenlong Yuan, Jiakai Cao, Zhaoxin Li, Hao Jiang, and Zhaoqi Wang. Sd-mvs: Segmentation-driven deformation multi-view stereo with spherical refinement and em optimization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6871–6880, 2024. 2
- [54] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 1, 2, 4, 5, 6, 7, 9
- [55] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianfeng Gao, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. *arXiv preprint arXiv:2303.08131*, 2023. 2, 7
- [56] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, 2023. 2, 5
- [57] Shunyuan Zheng, Boyao Zhou, Ruizhi Shao, Boning Liu, Shengping Zhang, Liqiang Nie, and Yebin Liu. Gps-gaussian: Generalizable pixel-wise 3d gaussian splatting for real-time human novel view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19680–19690, 2024. 2
- [58] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21634–21643, 2024. 2