ShelfGaussian: Shelf-Supervised Open-Vocabulary Gaussian-based 3D Scene Understanding

Lingjun Zhao*† Yandong Luo* James Hays Lu Gan
Georgia Institute of Technology, GA, USA
{lzhao360, yluo471, hays, lgan}@gatech.edu

Abstract

*We introduce **ShelfGaussian**, an open-vocabulary multi-modal Gaussian-based 3D scene understanding framework supervised by off-the-shelf vision foundation models (VFMs). Gaussian-based methods have demonstrated superior performance and computational efficiency across a wide range of scene understanding tasks. However, existing methods either model objects as closed-set semantic Gaussians supervised by annotated 3D labels, neglecting their rendering ability, or learn open-set Gaussian representations via purely 2D self-supervision, leading to degraded geometry and limited to camera-only settings. To fully exploit the potential of Gaussians, we propose a Multi-Modal Gaussian Transformer that enables Gaussians to query features from diverse sensor modalities, and a Shelf-Supervised Learning Paradigm that efficiently optimizes Gaussians with VFM features jointly at 2D image and 3D scene levels. We evaluate ShelfGaussian on various perception and planning tasks. Experiments on Occ3D-nuScenes demonstrate its state-of-the-art zero-shot semantic occupancy prediction performance. ShelfGaussian is further evaluated on an unmanned ground vehicle (UGV) to assess its in-the-wild performance across diverse urban scenarios. Project website: https://lunarlab-gatech.github.io/ShelfGaussian/.*

1. Introduction

3D scene understanding plays a crucial role in autonomous driving and robotic navigation. Among various tasks, 3D semantic occupancy prediction is particularly challenging, as it requires a fine-grained, joint semantic and geometric understanding of the surroundings [64, 65, 70]. Traditional semantic occupancy frameworks [26, 40, 72, 86] employ a dense voxel-based scene representation that is computationally and memory-intensive and unsuitable for onboard real-time deployment. Recently, 3D Gaussian Splatting (3DGS) [34] is naturally extended into an efficient and expressive scene representation for occupancy prediction, leveraging its object-centric modeling and high-fidelity rendering capabilities. Some work [11, 19, 75] retains voxels as the scene representation and use Gaussian rendering to improve feature learning. However, they inherit the fundamental limitations of voxel-based representations. In contrast, GaussianFormer family [28, 29, 59, 87, 93] directly adopts Gaussians as an alternative to voxels but still trains Gaussians to predict predefined semantics in a fully-supervised manner, which is limited to closed-set scenarios.

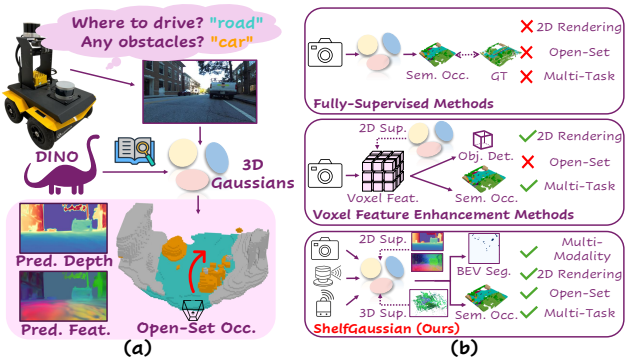

Figure 1. We propose **ShelfGaussian** for Gaussian-based 3D scene understanding under *open-vocabulary*, *multi-modal* and *multi-task* scenario. **(a)** Our model is able to assist a robot in predicting open-set occupancy from any sensor modalities with the help of VFMs. **(b)** Compared to existing Gaussian-based methods, ours provides a generalizable solution for 3D scene understanding.

This closed-vocabulary constraint is a bottleneck for real-world deployment. Autonomous vehicles inevitably encounter semantics beyond a predefined set and lead to unpredictable behaviors. The reliance on fully-supervised training also impedes model's ability to scale with data. These challenges have spurred a significant push towards self-supervised, open-vocabulary 3D scene understanding. VFMs such as CLIP [57] and DINO [8, 52, 58], demonstrate superior performance in open-set 2D visual perception, making them key enablers for extending this success to 3D. However, effectively leveraging 2D VFMs for 3D scene understanding remains challenging due to fundamen-

*Equal contribution.
†Corresponding author.

tal data and representation discrepancies between domains.

Recent attempts, such as GaussTR [32] and GaussianFlowOcc [4], learn feed-forward open-vocabulary Gaussians from camera-only information by using outputs from VFMs for 2D self-supervision. While these methods established a VFM-enabled self-supervised pipeline for 3D occupancy prediction, a significant performance gap remains compared to fully-supervised ones. These observations motivate us to explore stronger supervision signals and complementary input modalities beyond 2D images to fully unlock the expressiveness of Gaussian-based scene representations for open-vocabulary 3D scene understanding.

However, effectively learning open-vocabulary Gaussian features directly from unannotated 3D data is non-trivial. One bottleneck is the lack of an efficient, general Gaussian-to-voxel splatting module to support high-dimensional feature gradient backpropagation from voxel-based supervision. Another limitation is the absence of a systematic mechanism to bridge Gaussians with data from different 3D input modalities. Driven by these opportunities and challenges, we dig deeper into the application of feed-forward Gaussians for 3D scene understanding under *multi-modal*, *open-vocabulary*, and *multi-task* settings, and propose a **ShelfGaussian** framework. Our contributions are:

- We propose a **multi-modal Gaussian transformer** architecture that queries features from different sensor modalities to predict feed-forward 3D Gaussians, supporting camera-only, LiDAR-only, LiDAR-camera, camera-radar and LiDAR-camera-radar sensor input.
- We introduce a novel **shelf-supervised learning paradigm** to optimize Gaussians using off-the-shelf VFMs at both 2D image and 3D scene levels. A highly efficient, CUDA-accelerated Gaussian-to-Voxel (G2V) splatting module is designed to enable high-dimensional VFM feature distillation and speed up training.
- ShelfGaussian achieves **state-of-the-art zero-shot performance** on semantic occupancy prediction on nuScenes dataset, superior performance on Bird's-Eye-View (BEV) segmentation, and reduced collision rate in trajectory planning when integrated into our Gaussian-Planner.
- We further test ShelfGaussian on an unmanned ground vehicle (UGV) across diverse urban scenes, demonstrating its **superior in-the-wild performance**.

2. Related Work

2.1. Open-Vocabulary 3D Scene Understanding

Open-vocabulary 3D scene understanding typically includes 3D semantic segmentation [13, 15, 16, 38, 53, 55, 78], 3D instance segmentation [20, 30, 36, 48, 50, 51, 53, 60, 61, 71, 73, 74, 76, 79, 81], 3D object classification [25, 30, 85, 92], and 3D object detection [30, 35, 49, 92]. For outdoor driving scenarios, open-vocabulary occupancy prediction is particularly prominent due to its fine-grained voxel-level output. OVO [63], Veon [88] and POP-3D [67] are pioneers in distilling CLIP [57] features into 3D voxels to enable open-set querying abilities. Building on the success of 2D self-supervision in semantic occupancy prediction [18, 27, 54], recent works such as OccNeRF [84], DistillNeRF [69], and LangOcc [5] leverage volume rendering to align voxel features with foundation models, thereby enabling open-set occupancy prediction. CAL [62] and LOcc [83] propose a CLIP-based 3D pseudo-labeling engine and leverage it to supervise occupancy networks.

2.2. 3D Gaussians for Occupancy Prediction

3D Gaussian Splatting [34] demonstrates superior reconstruction and rendering quality, and has been extended to semantic occupancy prediction. GaussianFormer [28, 29] models scenes as dense semantic Gaussians supervised by closed-set voxel-wise labels. This representation is further extended by GaussianWorld [93] to temporal fusion and by GaussianFormer3D [87] to multi-modality scenarios, respectively. GaussianOcc [19], GaussRender [11], and GaussianPretrain [75] all employ a voxel-based representation and enhance voxel features through Gaussian rendering. GaussTR [32] learns sparse 3D Gaussians as the scene representation by aligning Gaussian renderings with 2D vision foundation features. GaussianFlowOcc [4] further incorporates a temporal module for Gaussians to model scene dynamics, achieving improved performance. To the best of our knowledge, our proposed ShelfGaussian is the first framework that unifies *open-vocabulary*, *multi-modality* and *Gaussian-based* scene completion.

3. Method

This section detailed the proposed ShelfGaussian framework, with an overview illustrated in Fig. 2.

3.1. Preliminary: Scene as Sparse 3D Gaussians

Inspired by generalizable 3D reconstruction methods [12, 82] which predict Gaussian parameters in a feed-forward manner, we utilize sparse feed-forward Gaussians to represent the scenes, akin to the Gaussians employed in previous work [4, 32]. Each Gaussian $\mathbf{G}$ is parameterized by its 3D mean $\mathbf{m}_{ego} \in \mathbb{R}^3$ in ego space, a 3D covariance matrix $\mathbf{\Sigma}$ composed of a rotation quaternion $\mathbf{r} \in \mathbb{R}^4$ and a scaling factor $\mathbf{s} \in \mathbb{R}^3$, an opacity $\alpha \in [0, 1]$, and a high-dimensional feature vector $\mathbf{f} \in \mathbb{R}^{C_f}$. For initialization, we first assign a 2D pixel location $\mathbf{m}_c^{2D} = (u, v)$ and retrieve its corresponding depth value d from a monocular estimated depth map $\mathbf{D}$, which can be optionally refined by LiDAR projected depth. These are combined into a 3D mean $\mathbf{m}_c^{3D} = (u, v, d)$ in the image frame, and then transformed to the ego frame using

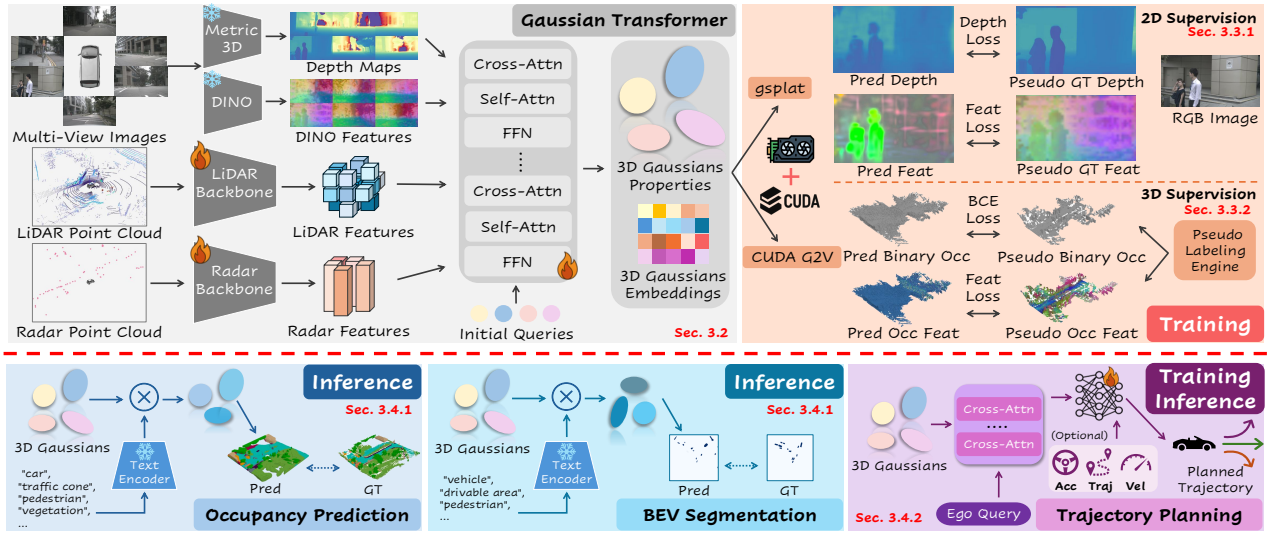

Figure 2. **Overview of ShelfGaussian.** ShelfGaussian employs off-the-shelf VFMs to extract depth and DINO feature maps from multi-view images, and trains LiDAR and radar backbones to extract related features. These are then fed into our multi-modal Gaussian transformer to predict sparse sets of 3D Gaussians to represent the scene. During training, Gaussians are rendered into camera views for VFM-based 2D supervision, while being converted into voxels via our CUDA-accelerated G2V module for 3D supervision. The shelf-supervised Gaussians support zero-shot semantic occupancy prediction, BEV segmentation, and are further evaluated for trajectory planning.

the camera intrinsics K and extrinsics E_{e2c}:

$$\mathbf{m}_{ego} = E_{e2c}^{-1} K^{-1} \mathbf{m}_c^{3D}. \tag{1}$$

The rotation quaternion $\mathbf{r}$ is initialized to be orthogonal to the camera view, while the scaling factor $\mathbf{s}$ is adjusted to be proportional to d based on perspective principles. The opacity and feature vector are left uninitialized. Thus, 3D Gaussian primitive can be parameterized and initialized as:

$$\mathbf{G} = (\mathbf{m}_{ego}, \mathbf{s}, \mathbf{r}, \alpha, \mathbf{f}). \tag{2}$$

3.2. Multi-Modal Gaussian Transformer

Multi-modal transformer [66] has achieved great success in learning unified BEV or voxel representations from multiple sensors (e.g., LiDAR, camera) for 3D object detection [2, 14, 39, 43, 47, 68, 80]. Motivated by this success, we propose a novel and generalizable multi-modal transformer for Gaussian-based 3D scene understanding. The learning objectives of our framework consist of a set of learnable Gaussian queries $\mathbf{Q} \in \mathbb{R}^{N_g \times C_g}$ and their 3D mean positions $\mathbf{M}_c^{3D} \in \mathbb{R}^{N_g \times 3}$ in the image frame, where N_g and C_g denote the number of Gaussians and the embedding dimension, respectively. We first extract feature maps from camera, LiDAR and radar data separately. For camera data, multi-scale feature maps $\mathbf{F}_c$ are extracted via a pre-trained DINO [52, 58] backbone followed by a feature pyramid network (FPN) [44]. For LiDAR data, multi-scale BEV feature maps $\mathbf{F}_l$ are extracted through a 3D backbone SECOND [77] and FPN [44]. For radar data, its BEV feature map $\mathbf{F}_r$ is generated from a 3D backbone PointPillars [37]. We then transform $\mathbf{M}_c^{3D}$ into the LiDAR frame using the camera intrinsics K and extrinsics E_{l2c} to obtain 3D positions of Gaussian queries on LiDAR and radar feature maps:

$$\mathbf{M}_l^{3D} = E_{l2c}^{-1} K^{-1} \mathbf{M}_c^{3D}. \tag{3}$$

Afterwards, through a sequence of transformer decoder layers, we gradually aggregate features from different sensor modalities via deformable attention [91]:

$$\mathbf{Q}_i = \text{DeAttn}(\mathbf{F}_i, \mathbf{Q}, \mathbf{M}_i^{3D}),\ i \in \{c, l, r\}. \tag{4}$$

The sampled features are then concatenated and fused via a multi-layer perceptron (MLP) to update the queries:

$$\mathbf{Q} = \mathbf{Q} + \text{MLP}(\mathbf{Q}_c \oplus \mathbf{Q}_l \oplus \mathbf{Q}_r). \tag{5}$$

After cross-attention to multi-modal features, we further update Gaussian queries through self-attention layers [66] to interact Gaussian features across the scene:

$$\mathbf{Q} = \text{SelfAttn}(\mathbf{Q}, \text{PE}(\mathbf{M}_c^{3D})), \tag{6}$$

where $\text{PE}(\cdot)$ represents positional encoding. Finally, a Gaussian-wise nonlinear mapping is applied to further enhance Gaussian queries via a feed-forward network (FFN):

$$\mathbf{Q} = \text{FFN}(\mathbf{Q}). \tag{7}$$

The query positions $\mathbf{M}_c^{3D}$ are subsequently updated via a MLP by $\mathbf{M}_c^{3D} = \mathbf{M}_c^{3D} + \text{MLP}(\mathbf{Q})$ within the decoder. To

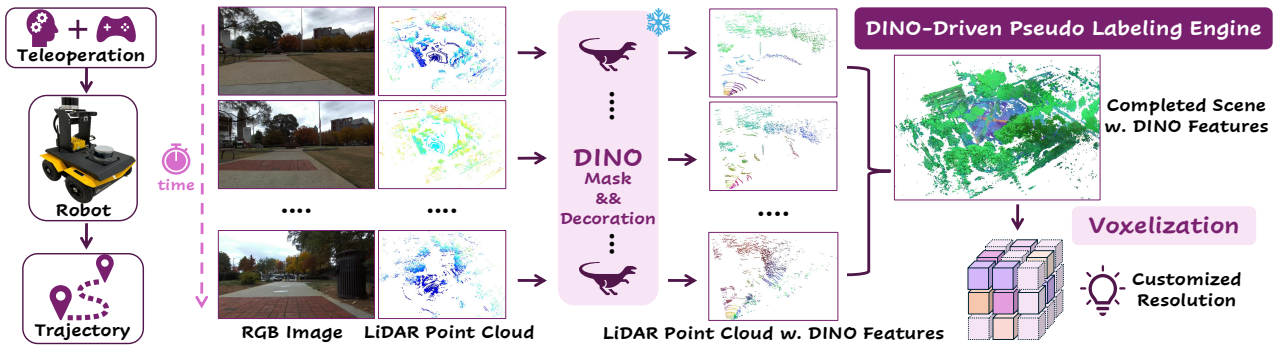

Figure 3. **Overview of DINO-Driven Pseudo Labeling Engine.** We teleoperate our UGV through urban scenarios to collect paired image and point cloud sequences along with trajectories from onboard camera and LiDAR. LiDAR points are then projected to image and decorated with pixel-wise DINO features. These points are aggregated and voxelized at a customized resolution to be 3D pseudo labels.

obtain other Gaussian properties, a set of MLPs are employed to regress these attributes based on learned queries:

$$\{\Delta s_i, \alpha_i, \mathbf{f}_i\} = \text{MLP}(\mathbf{Q}),\ i \in \{1, \dots, N_g\}. \tag{8}$$

The final predicted Gaussians can be passed to different heads to enable downstream perception and planning tasks.

3.3. Shelf-Supervised Learning for Gaussians

Instead of only aligning 2D Gaussian renderings with VFM features, we propose to jointly align 2D image-level and 3D scene-level Gaussian outputs with VFM features to narrow the performance gap to fully supervised methods.

3.3.1. 2D Image-Level Alignment with VFMs

Leveraging the high-fidelity rendering capability of 3D Gaussians, we first splat the Gaussians onto source views and then align the renders with estimated depth maps and VFM features [32]. To improve rendering efficiency, principal component analysis (PCA) [1] is used to compress both the high-dimensional Gaussian features $\mathbf{f}$ into $\bar{\mathbf{f}}$, and the target feature maps $\mathbf{F}^{tgt}$ into $\bar{\mathbf{F}}^{tgt}$. The rendered depth map $\bar{\mathbf{D}}$ and feature map $\bar{\mathbf{F}}$ are then computed by alpha blending:

$$\bar{\mathbf{F}} = \sum_{i=1}^{N} \bar{\mathbf{f}}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{9}$$

$$\bar{\mathbf{D}} = \sum_{i=1}^{N} d_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{10}$$

where d_i denotes the distance of each Gaussian to the camera along the viewing direction. Afterwards, a cosine similarity-based feature loss $\mathcal{L}^{2D}_{feat}$, a L1 depth loss $\mathcal{L}^{2D}_{depth}$, and a Scale-Invariant Logarithmic (SILog) depth loss [17] $\mathcal{L}^{2D}_{SILog}$ are combined for 2D image-level supervision.

3.3.2. 3D Scene-Level Alignment with VFMs

Current Gaussian-based approaches are supervised by either 2D feature maps or 3D closed-set labels, which present a trade-off and cannot achieve high efficiency and superior performance simultaneously. To alleviate this, we propose a *DINO-Driven Pseudo Labeling Engine* to automatically generate high-fidelity 3D occupancy pseudo labels, and a *CUDA-Accelerated Gaussian-to-Voxel Splatting* module to enable highly efficient Gaussian-to-voxel splatting, capable of supporting both forward and backward propagation of 3D Gaussians with high-dimensional features.

DINO-Driven Pseudo Labeling Engine. As shown in Fig. 3, the pseudo labeling engine begins by collecting paired RGB image and LiDAR point cloud sequences using onboard sensors (e.g.: LiDAR, camera) on a driving vehicle or mobile robot. DINO [52, 58] is then employed to *decorate* the raw LiDAR point clouds via point-to-image projection, i.e., points are projected onto 2D image plane to retrieve their corresponding VFM features. Notably, we apply a visibility mask to filter out points outside of the camera's field of view. Afterwards, LiDAR points of key frames are then accumulated for scene reconstruction. Lastly, the completed scene with DINO features is voxelized into pseudo occupancy labels at customized resolution. Our proposed pipeline can be inserted into any occupancy framework for automatic and high-quality 3D VFM feature labeling.

CUDA-Accelerated Gaussian-to-Voxel Splatting. An efficient Gaussian-to-Voxel splatting module can significantly reduce overall training time. However, PyTorch-based implementations in GaussTR [32] remain computationally expensive, and the CUDA-based implementation in GaussianFormer [28] is primarily designed for probability computation with a small set of semantic categories and only supports operations for Gaussians centered within a fixed voxel-space range. As a result, Gaussians that are located outside this range while still overlapping with target voxels are ignored, leading to inaccurate G2V conversion. In contrast, our approach is able to accurately and efficiently account for contributions from all Gaussians, regardless of their spatial locations. To this, we implement an efficient computation strategy based on a dual Compressed Sparse Row (CSR) [6] structure for mapping between Gaussians

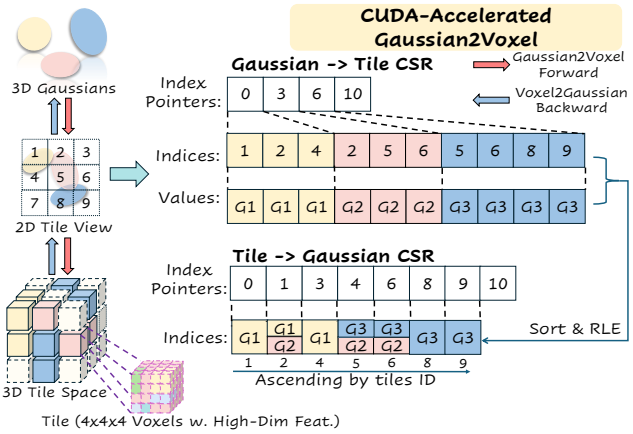

Figure 4. **Dual-CSR Structure for CUDA-Accelerated Gaussian2Voxel**. **Gaussian→Tile CSR:** index pointers store tile offsets per Gaussian, indices record tile IDs, and values store Gaussian IDs. **Tile→Gaussian CSR:** index pointers store Gaussian offsets per tile, and indices record Gaussian IDs obtained by sorting and run-length encoding (RLE) tile-Gaussian pairs.

and tiles, where each tile is defined as a block of $4 \times 4 \times 4$ voxels. As illustrated in Fig. 4, for the forward, based on Tile-to-Gaussian CSR, we assign one tile to each CUDA block for parallel voxel accumulation, which optimizes L1 cache usage and reduces redundant access. For the backward, based on Gaussian-to-Tile CSR, we assign one Gaussian per block to propagate gradients without atomic operations, thereby improving computational efficiency.

3D Shelf-Supervision. For 3D supervision, based on learned Gaussian properties and embeddings, ShelfGaussian first predicts occupancy densities $\boldsymbol{\rho} \in \mathbb{R}^{N_v \times 1}$ and corresponding VFM features $\mathbf{O} \in \mathbb{R}^{N_v \times C_v}$, where N_v and C_v denote the number of voxels and feature dimension. An MLP is then adopted to regress binary occupancy logits via $\mathbf{o} = \text{MLP}(\boldsymbol{\rho})$, and a predicted binary occupancy mask $\hat{\mathbf{o}}^{pred} \in \mathbb{R}^{N_v \times 1}$ is obtained by $\hat{\mathbf{o}}^{pred} = \mathbf{o} > \tau$, where τ is an empirical threshold. By multiplying $\hat{\mathbf{o}}^{pred}$ with pseudo binary occupancy label $\hat{\mathbf{o}}^{pseudo}$ and camera visibility mask $\hat{\mathbf{o}}^{vis}$, we derive the final feature supervision mask $\hat{\mathbf{o}}$. The final predicted occupancy features are formed by $\mathbf{O}^{pred} = \hat{\mathbf{o}} \times \mathbf{O}$. A binary cross entropy loss $\mathcal{L}^{3D}_{bce}$ between $\mathbf{o}$ and $\hat{\mathbf{o}}^{p}$, and a cosine similarity-based feature loss $\mathcal{L}^{3D}_{feat}$ between $\mathbf{O}^{pred}$ and pseudo feature label $\mathbf{O}^{pseudo}$ are applied for 3D supervision. Finally, by combining 2D and 3D losses, we train the model with the overall loss as follows:

$$\mathcal{L} = \mathcal{L}^{2D}_{feat} + \mathcal{L}^{2D}_{depth} + \mathcal{L}^{2D}_{SILog} + \mathcal{L}^{3D}_{bce} + \mathcal{L}^{3D}_{feat}. \quad (11)$$

3.4. Multi-Task Evaluation

3.4.1. Zero-Shot 3D Scene Understanding

We evaluate the learned VFM-aligned Gaussian representations on zero-shot semantic occupancy prediction and BEV map segmentation. During inference, we first leverage pretrained models from Talk2DINO [3] or DINO.txt [33] to map the CLIP [57] text embeddings of open-set queries into the DINO feature space, denoted as $\mathbf{f}_{txt} \in \mathbb{R}^{N_c \times C_f}$, where N_c is the number of semantic categories. Then we measure the cosine similarity between Gaussian features $\mathbf{f}$ and text embeddings $\mathbf{f}_{txt}$ to calculate the semantic logits for the Gaussian by $\mathbf{s} = \text{Softmax}(\mathbf{f} \cdot \mathbf{f}^{\text{T}}_{txt})$. Finally, we voxelize the Gaussians into a target resolution for zero-shot semantic occupancy prediction and BEV segmentation via Gaussian-to-voxel and Gaussian-to-BEV splatting separately.

3.4.2. Gaussian-based Trajectory Planning

Most existing methods adopt BEV [23, 24, 31, 42] or voxel [45, 65, 89] representations for end-to-end trajectory planning. Recent frameworks such as GaussianFusion [46] and GaussianAD [90] adopt a Gaussian representation similar to that in GaussianFormer [28], trained with auxiliary perception loss based on closed-set human-annotated labels. In contrast, we propose a simple yet effective *Gaussian-Planner* framework to directly apply our shelf-supervised open-vocabulary Gaussians to open-loop trajectory planning. To illustrate, we first initialize an ego query $\mathbf{Q}_{ego} \in \mathbb{R}^{N_{ego} \times C_{ego}}$ as a learnable embedding, where N_{ego} and C_{ego} denote the number of ego queries and query dimension, respectively. We then perform cross attention [66] between ego query and learned Gaussian embeddings $\mathbf{Q}$ to update the ego query. Finally, the trajectory is predicted by decoding the updated ego query with a MLP. The ego status (e.g., velocity, acceleration) can be concatenated to the ego query optionally for further enhancing vehicle self-awareness and motion feasibility. The prediction of future trajectory $\mathbf{X} \in \mathbb{R}^{N_{step} \times 2}$ can be formulated as follows:

$$\mathbf{X} = \text{MLP}(\text{CrossAttn}(\mathbf{Q}_{ego}, \mathbf{Q})). \quad (12)$$

An L_1 loss between predicted and ground-truth trajectories is adopted for supervision.

4. Experiment

4.1. Datasets and Metrics

Datasets. **NuScenes** [7] dataset provides 1000 sequences of driving scenes collected with 6 surrounding cameras, 1 LiDAR, 5 radars and 1 IMU. Occ3D [64] offers semantic occupancy annotation across 18 classes for nuScenes dataset within the range of $[-40\text{m}, 40\text{m}] \times [-40\text{m}, 40\text{m}] \times [-1\text{m}, 5.4\text{m}]$ with a voxel resolution of 0.4m. To evaluate the in-the-wild performance, we built a **Custom Dataset** by collecting paired monocular RGB images and LiDAR point clouds across diverse urban scenes using a Clearpath Jackal robot. We use a ZED 2i stereo camera to collect images and a combination of Velodyne VLP16 and Hesai XT16 LiDARs for point clouds. Our custom dataset consists of 2638

Method	Mod.	IoU	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
Supervised by 2D Image																				
SimpleOcc [18]	C	-	7.05	0.00	0.67	1.18	3.21	7.63	1.02	0.26	1.80	0.26	1.07	2.81	40.44	0.00	18.30	17.01	13.42	10.84
SelfOcc [27]	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
OccNeRF [84]	C	22.81	9.53	0.00	0.83	0.82	5.13	12.49	3.50	0.23	3.10	1.84	0.52	3.90	52.62	0.00	20.81	24.75	18.45	13.19
DistillNeRF [69]	C	29.11	8.93	0.03	1.35	2.08	10.21	10.09	2.56	1.98	5.54	4.62	1.43	7.90	43.02	0.00	16.86	15.02	14.06	15.06
GaussianOcc [19]	C	-	9.94	0.00	1.79	5.82	14.58	13.55	1.30	2.82	7.95	9.76	0.56	9.61	44.59	0.00	20.10	17.58	8.61	10.29
LangOcc [5]	C	51.76	11.84	0.00	3.10	9.00	6.30	14.20	0.40	10.80	6.20	9.00	3.80	10.70	43.70	**2.23**	9.50	26.40	19.60	26.40
GaussTR [32]	C	44.54	12.27	0.00	6.50	8.54	21.77	24.27	6.26	15.48	7.94	1.86	6.10	17.16	36.98	0.00	17.21	7.16	21.18	9.99
Supervised by 3D Pseudo Label or Temporal Information																				
VEON-B [88]	C	-	12.38	0.50	4.80	2.70	14.70	10.90	11.00	3.80	4.70	4.00	5.30	9.60	46.50	0.70	21.10	22.10	24.80	23.70
GaussianFlowOcc [4]	C	46.91	17.08	0.00	6.75	9.68	18.98	17.15	4.19	11.78	9.27	**10.30**	1.83	12.33	61.03	0.00	31.17	**34.78**	14.66	12.40
LOcc [83]	C	-	18.62	0.00	**14.21**	0.90	27.21	**32.61**	3.09	2.06	8.67	1.74	1.96	21.45	**71.22**	0.00	**38.18**	33.32	30.68	29.32
ShelfGaussian	C	63.25	19.07	0.36	2.96	10.84	30.20	30.20	11.34	18.93	14.49	6.54	9.47	17.77	60.02	0.14	29.66	21.62	29.63	30.08
	L	66.10	19.34	0.23	3.70	4.17	31.25	29.87	15.12	8.65	**24.34**	2.87	**14.62**	22.06	53.34	0.06	23.97	18.63	**36.02**	**39.96**
	C+R	62.84	19.42	0.51	2.77	11.72	31.15	30.46	12.40	19.56	15.35	6.60	11.19	19.28	58.89	0.15	28.17	21.36	29.58	31.01
	L+C	69.24	21.52	**0.60**	3.08	11.80	**36.36**	31.40	**14.71**	21.04	18.57	6.62	14.44	22.85	58.43	0.15	28.45	22.88	35.34	39.11
	L+C+R	**69.45**	**21.78**	0.58	3.30	**12.16**	34.20	31.62	14.59	**21.84**	20.88	7.06	14.24	**22.89**	59.91	0.14	29.14	22.77	35.35	39.64

Table 1. **Zero-shot semantic occupancy prediction performance of ShelfGaussian on Occ3D-nuScenes [64] dataset.** C, L and R denote camera, LiDAR and radar for simplicity. Mod. denotes the modality input during *inference*. The best is in **bolded** and the second best is underlined. Ours achieves the state-of-the-art performance with least number of scene queries, and works for any modality as input.

Method	Mod.	GT	Sup.	IoU		
				Vehicle	Pedestrian	Drivable Area
LSS [56]	C	✓	FS	-	15.0	72.9
FIERY [21]	C	✓	FS	35.8	17.2	-
BEVFormer [41]	C	✓	FS	35.8	-	80.1
PointBeV [10]	C	✓	FS	38.7	18.5	72.5
GaussianBEV [9]	C	✓	FS	**41.6**	**21.2**	**82.6**
ShelfGaussian	C	✗	ZS	**21.1**	**6.2**	**38.4**

Table 2. **Performance of ShelfGaussian on zero-shot BEV map segmentation on nuScenes [7] dataset.** GT: ground-truth. Sup.: supervision. FS: fully-supervised. ZS: zero-shot. Results of other methods are reported in [9].

Method	L2 (m) ↓				CR (%) ↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
ST-P3 [23]	1.59	2.64	3.73	2.65	0.69	3.62	8.39	4.23
UniAD* [24]	0.20	0.42	0.75	0.46	0.02	0.25	0.84	0.37
VAD-Base* [31]	0.17	0.34	0.60	0.37	0.04	0.27	**0.67**	0.33
OccWorld [89]	0.52	1.27	2.41	1.40	0.12	0.40	2.08	0.87
OccNet [65]	1.29	2.13	2.99	2.14	0.21	0.59	1.37	0.72
GaussianAD(5f) [90]	0.40	0.64	0.88	0.64	0.09	0.38	0.81	0.42
BEV-Planner(4f)* [42]	**0.16**	**0.32**	**0.57**	**0.35**	0.04	0.29	0.94	0.42
Gaussian-Planner*	**0.16**	**0.32**	0.58	**0.35**	**0.00**	**0.18**	0.78	**0.32**

Table 3. **Performance comparison of open-loop trajectory planning methods on nuScenes [7] dataset.** (xf) and * denote the number of past frames and ego trajectory in usage, respectively. L2 error and CR are computed the same way as ST-P3 [23].

training and 294 testing frames, collected in *garden*, *street*, *park* and *grassland* scenarios. Pseudo occupancy labels are generated at 2Hz in a range of $[0\text{m}, 20\text{m}] \times [-10\text{m}, 10\text{m}] \times [-1\text{m}, 4\text{m}]$ with a voxel resolution of 0.2m.

Metrics. Intersection-over-Union (IoU) and mean IoU (mIoU) across all classes are used for evaluation of our model on zero-shot semantic occupancy prediction and BEV map segmentation. L2 error and collision rate (CR) are reported to assess the trajectory planning performance. L2 error is defined as the mean Euclidean distance between predicted and ground-truth trajectories while CR is the percentage of the planned trajectory intersecting an obstacle.

4.2. Implementation Details

We evaluate our method with DINOv2 [52] and DINOv3 [58] ViT-B models, and image resolutions are set as 378×672 and 432×768, respectively. The DINO feature dimension is compressed to 128 via PCA [1]. Metric3Dv2 [22] is used for depth supervision. SECOND [77] and PointPillars [37] serve as the 3D backbone for LiDAR and radar, respectively. Talk2DINO [3] and DINO.txt [33] are utilized for vision-language alignment. The number of Gaussians per view is set as 1000 in our main experiments. The Gaussian embedding dimension is set to 256. We train our perception model for 24 epochs on 8 Nvidia L40S GPUs with a learning rate of 2×10^{-4} and a batch size of 8. The planning model is further trained for 12 epochs on 8 Nvidia V100 GPUs with a learning rate of 1×10^{-4} and a batch size of 32. Please see supplementary material for more details.

4.3. Quantitative Results

Zero-Shot Semantic Occupancy Prediction. We demonstrate zero-shot semantic occupancy prediction performance of ShelfGaussian in Tab. 1. Ours outperforms all existing self-supervised and weakly supervised methods, especially on vehicles (e.g., *bus*, *car*, *trailer*), dynamic agents (e.g., *bicycle*, *motorcycle*, *pedestrian*), and unstructured surfaces (e.g., *manmade*, *vegetation*), due to Gaussians' superior object-centric modeling property. Notably, ours surpasses the SOTA method LOcc [83] in most metrics with only **15%** of its number of queries (i.e., 6000 vs. 40000). As more sensor modalities are fused, our model exhibits steady performance gains, with ShelfGaussian-LCR yielding **+6.2** IoU and **+2.71** mIoU over CO, demonstrating the effectiveness of our multi-modal Gaussian transformer.

Zero-Shot BEV Map Segmentation. Existing BEV segmentation models rely heavily on ground-truth labels and

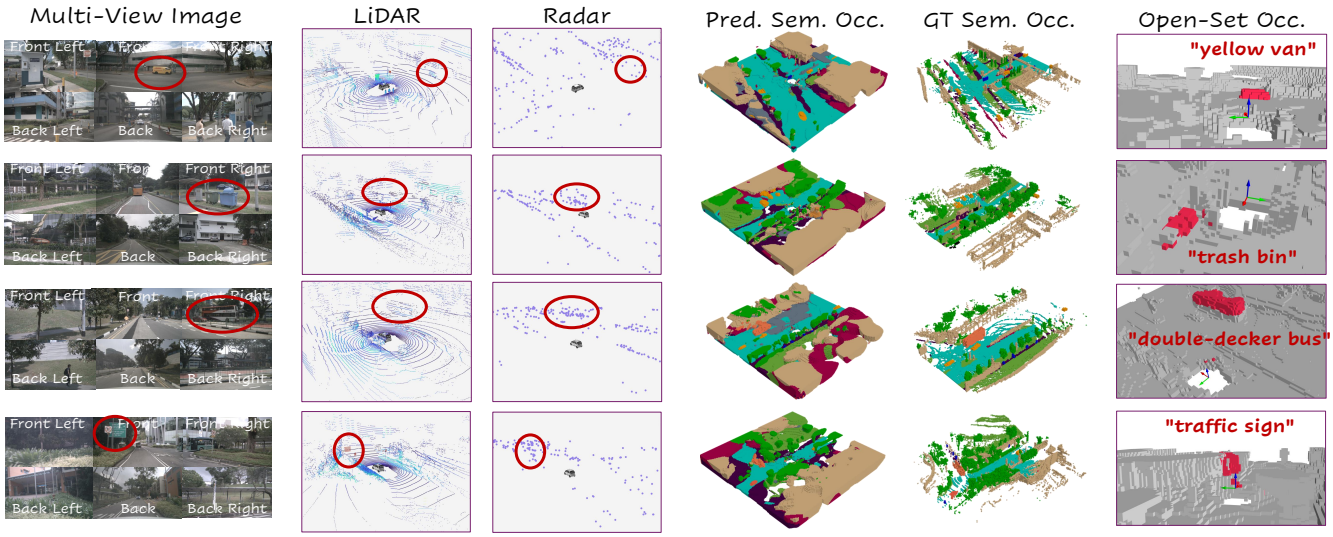

Figure 5. **Qualitative results of ShelfGaussian on nuScenes dataset.** The figure demonstrates the predicted semantic occupancy queried by semantic classes in Tab. 1, ground-truth labels from Occ3D [64] and occupancy of open-set queries from ShelfGaussian-LCR model. Best viewed on screen and color bar is given in Tab. 1.

Mod.	DINOv3	DINOv2	IoU	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	other flat	sidewalk	terrain	manmade	vegetation
C	✓		61.03	13.27	1.21	3.28	7.33	23.41	35.01	17.78	8.54	3.03	8.82	14.06	12.58	55.82	1.05	0.03	13.97	7.29	12.36
		✓	63.25	19.07	0.36	2.96	10.84	30.20	30.20	11.34	18.93	14.49	6.54	9.47	17.77	60.02	0.14	29.66	21.62	29.63	30.08
L+C	✓		69.24	15.41	1.05	4.15	8.63	26.31	39.76	25.12	10.84	3.65	11.61	22.22	14.97	54.17	1.22	0.08	14.52	8.72	14.99
		✓	69.24	21.52	0.60	3.08	11.80	36.36	31.40	14.71	21.04	18.57	6.62	14.44	22.85	58.43	0.15	28.45	22.88	35.34	39.11

Table 4. **Ablation on alignment with different VFMs.** DINOv3 [58]: ViT-B/16, 432 × 768 image input. DINOv2 [52]: ViT-B/14, 378 × 672 image input. Better results with DINOv2 and DINOv3 are highlighted in orange and cyan colors, respectively.

Mod.	2D Loss	3D BCE Loss	3D Feat. Loss	IoU	mIoU
C	✓			47.26	12.39
	✓	✓		59.92	14.49
		✓	✓	58.64	18.41
	✓	✓	✓	63.25	19.07

Table 5. **Ablation on shelf-supervision loss.**

Method	18k Gauss. w. 1024D FW(s)	BW(s)	Mem.(GB)	9k Gauss. w. 768D FW(s)	BW(s)	Mem.(GB)
GaussTR [32]	14.1	596.7	22.3	12.4	446.4	14.8
GaussianFormer [28]	5.4	13.4	12.9	1.9	4.2	7.7
ShelfGaussian	**0.5**	**2.7**	**4.9**	**0.2**	**1.0**	**3.7**

Table 6. **Comparison of G2V splatting module on one Nvidia L40S GPU.** FW / BW: Forward / Backward time. Mem.: memory.

require separate training for each semantic class (e.g., vehicle, pedestrian). In contrast, ours eliminates the need for human annotation and requires only a single training process. As shown in Tab. 2, ShelfGaussian achieves competitive performance compared with existing SOTA methods.

Gaussian-based Trajectory Planning. As shown in Tab. 3, Gaussian-Planner performs on par with existing top-performing methods while reducing the average CR to **0.32%**, benefiting from the intrinsically object-centric characteristics of Gaussian representations. These results validate the effectiveness of our shelf-supervised Gaussians in vehicle trajectory planning and demonstrate their potential to mitigate collision risk in real-world autonomous driving.

4.4. Ablation Study

Ablation on Alignment with Different VFMs. We perform experiments on aligning Gaussians with DINOv2 [52] and DINOv3 [58], and using Talk2DINO [3] to query semantics. Results in Tab. 4 lead to an empirical finding that our model with DINOv3 excels DINOv2 on some vehicles (e.g., *car*, *trailer*) and small long-tail classes (e.g., *traffic cone*, *others*) but degrades catastrophically on dynamic road agents (e.g., *motorcycle*, *pedestrian*) and static unstructured surfaces (e.g., *sidewalk*, *manmade*). This indicates that although DINOv3 provides stronger visual representations, it still faces challenges in urban driving scenarios.

Ablation on Shelf-Supervision Losses. We conduct an ablation study on different supervision losses. As revealed in Tab. 5, 3D scene-level supervision alone yields more accurate geometric and semantic understanding than 2D image-level supervision, with performance gains primarily stemming from the 3D BCE loss and the 3D feature loss, respectively. The results also show that combining 2D and 3D supervision achieves the best performance by leveraging complementary information from both domains.

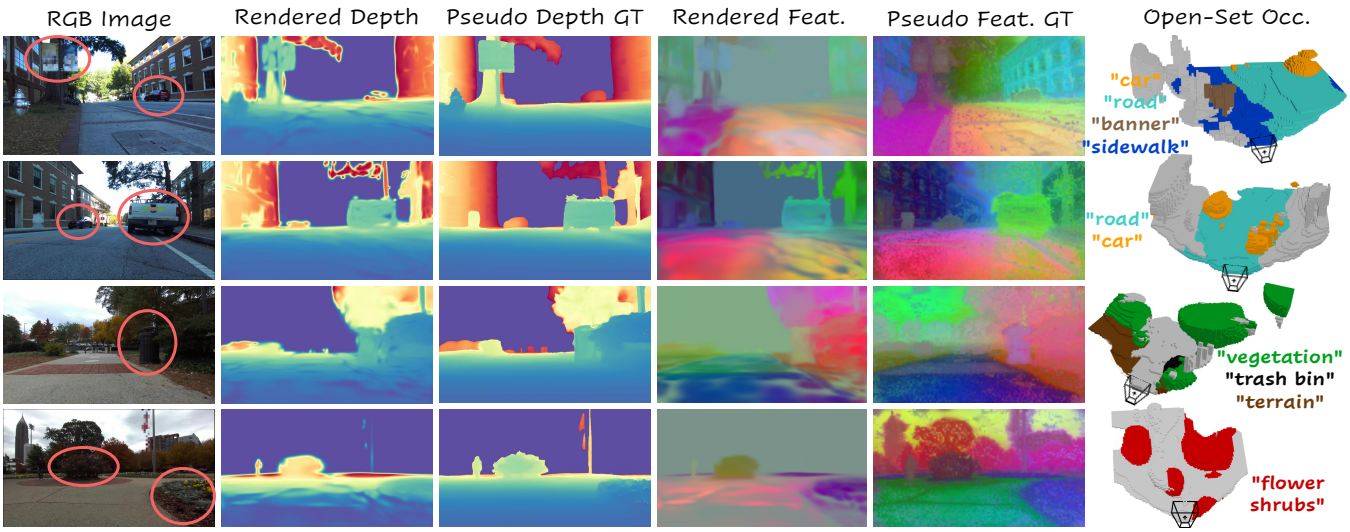

Figure 6. **Qualitative results of ShelfGaussian on custom dataset collected by a UGV.** The figure shows the rendered depth map, DINO feature map, and occupancy of novel categories from ShelfGaussian-CO model. Best viewed on screen and in color.

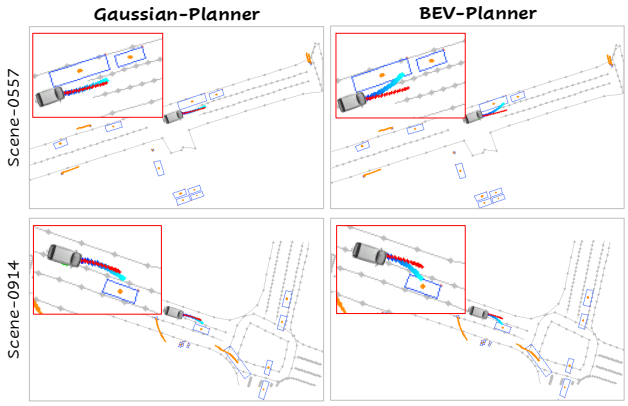

Figure 7. **Qualitative comparison of BEV-Planner [42] and Gaussian-Planner on nuScenes [7] dataset.** Red and cyan lines denote the ground-truth and predicted trajectories separately.

Benchmark of G2V Splatting Module. We benchmark our G2V spalting module against other open-source methods [28, 32] in two settings: 18k Gaussians with 1024-dim features and 9k Gaussians with 768-dim features, with a target voxel number of 640k. As shown in Tab. 6, our module achieves **10x** faster forward and **5x** faster backward passes than GaussianFormer [28], while consuming only around **40%** of the memory. Our G2V module enables Gaussians to learn high-dimensional open-vocabulary features from 3D data with high efficiency, laying the foundation for generalizable Gaussian-based 3D scene understanding.

4.5. Qualitative Results

Semantic Occupancy Prediction. Qualitative results of ShelfGaussian on nuScenes and custom datasets are given in Fig. 5 and Fig. 6, respectively. In on-road driving scenarios, as shown in Fig. 5, ShelfGaussian predicts fine-grained occupancy with accurate semantics and detailed geometry, exhibiting high similarity to the annotated closed-vocabulary ground-truth while being open-vocabulary. Given a natural language query (e.g., *double-decker bus*), our model can localize the target object and complete its shape precisely. In other urban scenarios, ShelfGaussian also demonstrates superior open-vocabulary occupancy prediction capability. For example, as shown in Fig. 6, given the text query *"flower shrubs"*, our model successfully recovers all flower shrub areas, highlighted in red, despite their sparse features and low texture.

Trajectory Planning. Fig. 7 illustrates a failure case of BEV-Planner [42] where our Gaussian-Planner successfully prevents the collision. This indicates the crucial role of object-centric Gaussian representations in reducing collision risk for safe, robust vehicle trajectory planning.

5. Conclusion

In this work, we present ShelfGaussian, an open-vocabulary, multi-modal, Gaussian-based 3D scene understanding framework supervised by off-the-shelf VFMs. Unlike existing methods that align 3D Gaussians with VFMs only at the 2D image level, ours unlocks the capability of foundation model alignment at the 3D scene level, achieved by a DINO-driven pseudo labeling engine and a CUDA-accelerated G2V splatting module. This 3D alignment capability significantly improves the model's zero-shot performance, narrowing the performance gap to fully-supervised methods while alleviating scalability and generalization limitations. Furthermore, our model is general, supporting a combination of multi-sensor inputs, and versatile, achieving superior performance on three downstream tasks. Notably, ShelfGaussian achieves SOTA zero-shot performance on semantic occupancy prediction on nuScenes dataset and superior in-the-wild performance on a UGV platform.

Acknowledgments

Lingjun Zhao is supported by the Institute for Robotics and Intelligent Machines (IRIM) Ph.D. Fellowship at Georgia Institute of Technology. This work used Delta GPU and DeltaAI at NCSA through allocation CIS251067 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

References

[1] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010. 4, 6

[2] Xuyang Bai, Zeyu Hu, Xinge Zhu, Qingqiu Huang, Yilun Chen, Hongbo Fu, and Chiew-Lan Tai. Transfusion: Robust lidar-camera fusion for 3d object detection with transformers. In *CVPR*, pages 1090–1099, 2022. 3

[3] Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to dino: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *ICCV*, pages 22025–22035, 2025. 5, 6, 7

[4] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Gaussianflowocc: Sparse and weakly supervised occupancy estimation using gaussian splatting and temporal flow. *arXiv preprint arXiv:2502.17288*, 2025. 2, 6

[5] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Open vocabulary occupancy estimation via volume rendering. In *3DV*, pages 200–210. IEEE, 2025. 2, 6

[6] Aydin Buluç, Jeremy T Fineman, Matteo Frigo, John R Gilbert, and Charles E Leiserson. Parallel sparse matrix-vector and matrix-transpose-vector multiplication using compressed sparse blocks. In *Proceedings of the twenty-first annual symposium on Parallelism in algorithms and architectures*, pages 233–244, 2009. 4

[7] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020. 5, 6, 8

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1

[9] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. In *WACV*, pages 2250–2259. IEEE, 2025. 6

[10] Loick Chambon, Eloi Zablocki, Mickaël Chen, Florent Bartoccioni, Patrick Pérez, and Matthieu Cord. Pointbev: A sparse approach for bev predictions. In *CVPR*, pages 15195–15204, 2024. 6

[11] Loick Chambon, Eloi Zablocki, Alexandre Boulch, Mickael Chen, and Matthieu Cord. Gaussrender: Learning 3d occupancy with gaussian rendering. In *ICCV*, pages 27010–27020, 2025. 1, 2

[12] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *CVPR*, pages 19457–19467, 2024. 2

[13] Runnan Chen, Youquan Liu, Lingdong Kong, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, Yu Qiao, and Wenping Wang. Clip2scene: Towards label-efficient 3d scene understanding by clip. In *CVPR*, pages 7020–7030, 2023. 2

[14] Xuanyao Chen, Tianyuan Zhang, Yue Wang, Yilun Wang, and Hang Zhao. Futr3d: A unified sensor fusion framework for 3d detection. In *CVPR*, pages 172–181, 2023. 3

[15] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Pla: Language-driven open-vocabulary 3d scene understanding. In *CVPR*, pages 7010–7019, 2023. 2

[16] Runyu Ding, Jihan Yang, Chuhui Xue, Wenqing Zhang, Song Bai, and Xiaojuan Qi. Lowis3d: Language-driven open-world instance-level 3d scene understanding. *IEEE TPAMI*, 46(12):8517–8533, 2024. 2

[17] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *NeurIPS*, 27, 2014. 4

[18] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *CoRR*, 2023. 2, 6

[19] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. In *ICCV*, pages 28980–28990, 2025. 1, 2, 6

[20] Qingdong He, Jinlong Peng, Zhengkai Jiang, Kai Wu, Xiaozhong Ji, Jiangning Zhang, Yabiao Wang, Chengjie Wang, Mingang Chen, and Yunsheng Wu. Unimov3d: Uni-modality open-vocabulary 3d scene understanding with fine-grained feature representation. *arXiv preprint arXiv:2401.11395*, 2024. 2

[21] Anthony Hu, Zak Murez, Nikhil Mohan, Sofía Dudas, Jeffrey Hawke, Vijay Badrinarayanan, Roberto Cipolla, and Alex Kendall. Fiery: Future instance prediction in bird's-eye view from surround monocular cameras. In *ICCV*, pages 15273–15282, 2021. 6

[22] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. *IEEE TPAMI*, 2024. 6

[23] Shengchao Hu, Li Chen, Penghao Wu, Hongyang Li, Junchi Yan, and Dacheng Tao. St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning. In *ECCV*, pages 533–549. Springer, 2022. 5, 6

[24] Yihan Hu, Jiazhi Yang, Li Chen, Keyu Li, Chonghao Sima, Xizhou Zhu, Siqi Chai, Senyao Du, Tianwei Lin, Wenhai Wang, et al. Planning-oriented autonomous driving. In *CVPR*, pages 17853–17862, 2023. 5, 6

[25] Tianyu Huang, Bowen Dong, Yunhan Yang, Xiaoshui Huang, Rynson WH Lau, Wanli Ouyang, and Wangmeng

Zuo. Clip2point: Transfer clip to point cloud classification with image-depth pre-training. In *ICCV*, pages 22157–22167, 2023. 2

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

[27] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *CVPR*, pages 19946–19956, 2024. 2, 6

[28] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *ECCV*, pages 376–393. Springer, 2024. 1, 2, 4, 5, 7, 8

[29] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *CVPR*, pages 27477–27486, 2025. 1, 2

[30] Zhening Huang, Xiaoyang Wu, Xi Chen, Hengshuang Zhao, Lei Zhu, and Joan Lasenby. Openins3d: Snap and lookup for 3d open-vocabulary instance segmentation. In *ECCV*, pages 169–185. Springer, 2024. 2

[31] 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. In *ICCV*, pages 8340–8350, 2023. 5, 6

[32] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. In *CVPR*, pages 11960–11970, 2025. 2, 4, 6, 7, 8

[33] Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, et al. Dinov2 meets text: A unified framework for image-and pixel-level vision-language alignment. In *CVPR*, pages 24905–24916, 2025. 5, 6

[34] 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. 1, 2

[35] Mehar Khurana, Neehar Peri, James Hays, and Deva Ramanan. Shelf-supervised cross-modal pre-training for 3d object detection. In *CoRL*, 2024. 2

[36] Sebastian Koch, Narunas Vaskevicius, Mirco Colosi, Pedro Hermosilla, and Timo Ropinski. Open3dsg: Open-vocabulary 3d scene graphs from point clouds with queryable objects and open-set relationships. In *CVPR*, pages 14183–14193, 2024. 2

[37] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders for object detection from point clouds. In *CVPR*, pages 12697–12705, 2019. 3, 6

[38] Ruihuang Li, Zhengqiang Zhang, Chenhang He, Zhiyuan Ma, Vishal M Patel, and Lei Zhang. Dense multimodal alignment for open-vocabulary 3d scene understanding. In *ECCV*, pages 416–434. Springer, 2024. 2

[39] Yanwei Li, Yilun Chen, Xiaojuan Qi, Zeming Li, Jian Sun, and Jiaya Jia. Unifying voxel-based representation with transformer for 3d object detection. *NeurIPS*, 35:18442–18455, 2022. 3

[40] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *CVPR*, pages 9087–9098, 2023. 1

[41] 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 lidar-camera via spatiotemporal transformers. *IEEE TPAMI*, 2024. 6

[42] Zhiqi Li, Zhiding Yu, Shiyi Lan, Jiahan Li, Jan Kautz, Tong Lu, and Jose M Alvarez. Is ego status all you need for open-loop end-to-end autonomous driving? In *CVPR*, pages 14864–14873, 2024. 5, 6, 8

[43] Tingting Liang, Hongwei Xie, Kaicheng Yu, Zhongyu Xia, Zhiwei Lin, Yongtao Wang, Tao Tang, Bing Wang, and Zhi Tang. Bevfusion: A simple and robust lidar-camera fusion framework. *NeurIPS*, 35:10421–10434, 2022. 3

[44] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, pages 2117–2125, 2017. 3

[45] Ruixun Liu, Lingyu Kong, Derun Li, and Hang Zhao. Occvla: Vision-language-action model with implicit 3d occupancy supervision. *arXiv preprint arXiv:2509.05578*, 2025. 5

[46] Shuai Liu, Quanmin Liang, Zefeng Li, Boyang Li, and Kai Huang. Gaussianfusion: Gaussian-based multi-sensor fusion for end-to-end autonomous driving. *arXiv preprint arXiv:2506.00034*, 2025. 5

[47] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *ICRA*, 2023. 3

[48] Shiyang Lu, Haonan Chang, Eric Pu Jing, Abdeslam Boularias, and Kostas Bekris. Ovir-3d: Open-vocabulary 3d instance retrieval without training on 3d data. In *CoRL*, pages 1610–1620. PMLR, 2023. 2

[49] Yuheng Lu, Chenfeng Xu, Xiaobao Wei, Xiaodong Xie, Masayoshi Tomizuka, Kurt Keutzer, and Shanghang Zhang. Open-vocabulary point-cloud object detection without 3d annotation. In *CVPR*, pages 1190–1199, 2023. 2

[50] Rafay Mohiuddin, Sai Manoj Prakhya, Fiona Collins, Ziyuan Liu, and André Borrmann. Opensu3d: Open world 3d scene understanding using foundation models. In *ICRA*, pages 13560–13566. IEEE, 2025. 2

[51] Phuc Nguyen, Tuan Duc Ngo, Evangelos Kalogerakis, Chuang Gan, Anh Tran, Cuong Pham, and Khoi Nguyen. Open3dis: Open-vocabulary 3d instance segmentation with 2d mask guidance. In *CVPR*, pages 4018–4028, 2024. 2

[52] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 1, 3, 4, 6, 7

[53] Aljoša Ošep, Tim Meinhardt, Francesco Ferroni, Neehar Peri, Deva Ramanan, and Laura Leal-Taixé. Better call sal: Towards learning to segment anything in lidar. In *ECCV*, pages 71–90. Springer, 2024. 2

[54] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, pages 12404–12411. IEEE, 2024. 2

[55] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *CVPR*, pages 815–824, 2023. 2

[56] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, pages 194–210. Springer, 2020. 6

[57] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 1, 2, 5

[58] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 1, 3, 4, 6, 7

[59] Ke Song, Yunhe Wu, Chunchit Siu, and Huiyuan Xiong. Graphgsocc: Semantic and geometric graph transformer for 3d gaussian splating-based occupancy prediction. *arXiv preprint arXiv:2506.14825*, 2025. 1

[60] Ayça Takmaz, Elisabetta Fedele, Robert W Sumner, Marc Pollefeys, Federico Tombari, and Francis Engelmann. Openmask3d: Open-vocabulary 3d instance segmentation. *arXiv preprint arXiv:2306.13631*, 2023. 2

[61] Ayca Takmaz, Alexandros Delitzas, Robert W Sumner, Francis Engelmann, Johanna Wald, and Federico Tombari. Search3d: Hierarchical open-vocabulary 3d segmentation. *IEEE Robotics and Automation Letters*, 2025. 2

[62] Ayça Takmaz, Cristiano Saltori, Neehar Peri, Tim Meinhardt, Riccardo de Lutio, Laura Leal-Taixé, and Aljoša Ošep. Towards learning to complete anything in lidar. In *ICML*, 2025. 2

[63] Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023. 2

[64] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. In *NeurIPS*, pages 64318–64330, 2023. 1, 5, 6, 7

[65] 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, 5, 6

[66] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 30, 2017. 3, 5

[67] Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *NeurIPS*, 36:50545–50557, 2023. 2

[68] Haiyang Wang, Hao Tang, Shaoshuai Shi, Aoxue Li, Zhenguo Li, Bernt Schiele, and Liwei Wang. Unitr: A unified and efficient multi-modal transformer for bird's-eye-view representation. In *ICCV*, pages 6792–6802, 2023. 3

[69] Letian Wang, Seung Wook Kim, Jiawei Yang, Cunjun Yu, Boris Ivanovic, Steven Waslander, Yue Wang, Sanja Fidler, Marco Pavone, and Peter Karkus. Distillnerf: Perceiving 3d scenes from single-glance images by distilling neural fields and foundation model features. *NeurIPS*, 37:62334–62361, 2024. 2, 6

[70] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In *ICCV*, pages 17850–17859, 2023. 1

[71] Zhigang Wang, Yifei Su, Chenhui Li, Dong Wang, Yan Huang, Xuelong Li, and Bin Zhao. Open-vocabulary octree-graph for 3d scene understanding. In *ICCV*, pages 7037–7047, 2025. 2

[72] 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

[73] Jianyun Xu, Song Wang, Ziqian Ni, Chunyong Hu, Sheng Yang, Jianke Zhu, and Qiang Li. Sam4d: Segment anything in camera and lidar streams. In *ICCV*, 2025. 2

[74] Mutian Xu, Xingyilang Yin, Lingteng Qiu, Yang Liu, Xin Tong, and Xiaoguang Han. Sampro3d: Locating sam prompts in 3d for zero-shot scene segmentation. *arXiv preprint arXiv:2311.17707*, 2023. 2

[75] Shaoqing Xu, Fang Li, Shengyin Jiang, Ziying Song, Li Liu, and Zhi-xin Yang. Gaussianpretrain: A simple unified 3d gaussian representation for visual pre-training in autonomous driving. *arXiv preprint arXiv:2411.12452*, 2024. 1, 2

[76] Mi Yan, Jiazhao Zhang, Yan Zhu, and He Wang. Maskclustering: View consensus based mask graph clustering for open-vocabulary 3d instance segmentation. In *CVPR*, pages 28274–28284, 2024. 2

[77] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 3, 6

[78] Jihan Yang, Runyu Ding, Weipeng Deng, Zhe Wang, and Xiaojuan Qi. Regionplc: Regional point-language contrastive learning for open-world 3d scene understanding. In *CVPR*, pages 19823–19832, 2024. 2

[79] Yunhan Yang, Xiaoyang Wu, Tong He, Hengshuang Zhao, and Xihui Liu. Sam3d: Segment anything in 3d scenes. *arXiv preprint arXiv:2306.03908*, 2023. 2

[80] Zeyu Yang, Jiaqi Chen, Zhenwei Miao, Wei Li, Xiatian Zhu, and Li Zhang. Deepinteraction: 3d object detection via modality interaction. *NeurIPS*, 35:1992–2005, 2022. 3

[81] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *CVPR*, pages 3292–3302, 2024. 2

[82] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 2

[83] Zhu Yu, Bowen Pang, Lizhe Liu, Runmin Zhang, Qiang Li, Si-Yuan Cao, Maochun Luo, Mingxia Chen, Sheng Yang, and Hui-Liang Shen. Language driven occupancy prediction. In *ICCV*, pages 7548–7558, 2025. 2, 6

[84] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Advancing 3d occupancy prediction in lidar-free environments. *IEEE TIP*, 2025. 2, 6

[85] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *CVPR*, pages 8552–8562, 2022. 2

[86] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023. 1

[87] Lingjun Zhao, Sizhe Wei, James Hays, and Lu Gan. Gaussianformer3d: Multi-modal gaussian-based semantic occupancy prediction with 3d deformable attention. *arXiv preprint arXiv:2505.10685*, 2025. 1, 2

[88] Jilai Zheng, Pin Tang, Zhongdao Wang, Guoqing Wang, Xiangxuan Ren, Bailan Feng, and Chao Ma. Veon: Vocabulary-enhanced occupancy prediction. In *ECCV*, pages 92–108. Springer, 2024. 2, 6

[89] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. In *ECCV*, pages 55–72. Springer, 2024. 5, 6

[90] Wenzhao Zheng, Junjie Wu, Yao Zheng, Sicheng Zuo, Zixun Xie, Longchao Yang, Yong Pan, Zhihui Hao, Peng Jia, Xianpeng Lang, et al. Gaussianad: Gaussian-centric end-to-end autonomous driving. *arXiv preprint arXiv:2412.10371*, 2024. 5, 6

[91] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 3

[92] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *ICCV*, pages 2639–2650, 2023. 2

[93] Sicheng Zuo, Wenzhao Zheng, Yuanhui Huang, Jie Zhou, and Jiwen Lu. Gaussianworld: Gaussian world model for streaming 3d occupancy prediction. In *CVPR*, pages 6772–6781, 2025. 1, 2