This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

AutoOcc: Automatic Open-Ended Semantic Occupancy Annotation via Vision-Language Guided Gaussian Splatting

Xiaoyu Zhou[1] Jingqi Wang[1] Yongtao Wang[1]* Yufei Wei[2]
Nan Dong[2] Ming-Hsuan Yang[3]
[1]Wangxuan Institute of Computer Technology, Peking University
[2]Chongqing Changan Automobile Co., Ltd [3]University of California, Merced

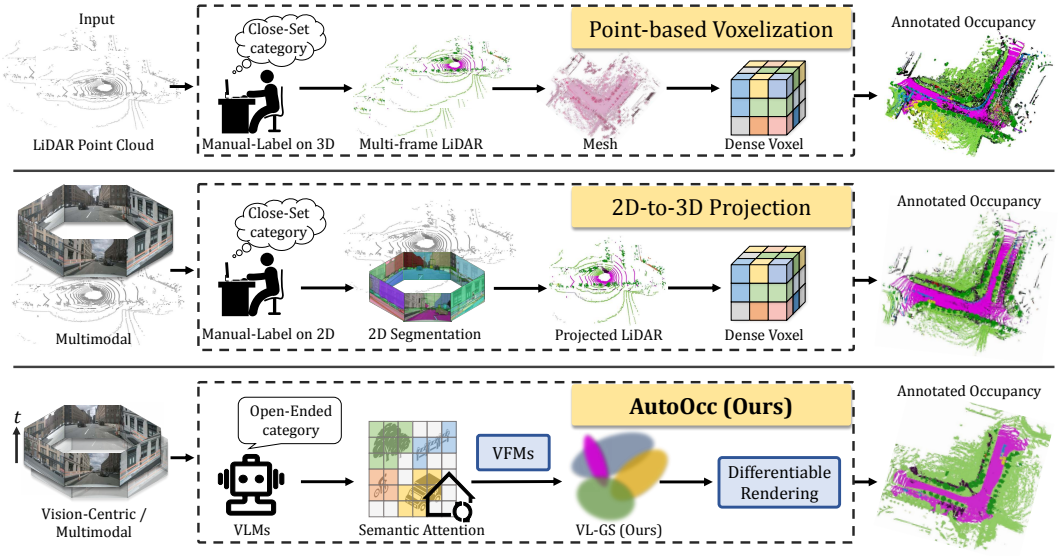


Figure 1. **AutoOcc** is a fully automatic, vision-centric pipeline for open-ended semantic 3D occupancy annotation. Our method achieves more efficient and effective semantic occupancy auto-labeling by integrating vision-language guidance with differentiable reconstruction. AutoOcc supports open-ended semantic annotation and effectively handles dynamic objects, without relying on any human annotations.

Abstract

*Obtaining high-quality 3D semantic occupancy from raw sensor data remains an essential yet challenging task, often requiring extensive manual labeling. In this work, we propose **AutoOcc**, a vision-centric **auto**mated pipeline for open-ended semantic **occ**upancy annotation that integrates differentiable Gaussian splatting guided by vision-language models. We formulate the open-ended semantic 3D occupancy reconstruction task to automatically generate scene occupancy by combining attention maps from vision-language models and foundation vision models. We devise semantic-aware Gaussians as intermediate geometric descriptors and propose a cumulative Gaussian-to-voxel splatting algorithm that enables effective and efficient occupancy annotation. Our framework outperforms existing automated occupancy annotation methods without human labels. AutoOcc also enables open-ended semantic occupancy auto-labeling, achieving robust performance in both static and dynamically complex scenarios.*

1. Introduction

3D semantic occupancy has attracted a considerable amount of attention in autonomous driving [47, 48, 50] and embodied intelligence [7, 38, 39], demonstrating great potential to facilitate understanding of 3D scenes and perception of irregular objects. Despite its promising applications, automatic generation of precise and complete semantic occupancy annotations from raw sensor data remains a fundamental challenge, particularly in the pursuit of cost-effective solutions for real-world deployment.

Vision-centric automated 3D semantic occupancy annotation has long been undervalued, while existing occupancy annotation pipelines heavily rely on LiDAR point

*Corresponding author.

Table 1. **Comparisons between AutoOcc and existing semantic occupancy annotation pipelines.** The definitions of closed-set, open-set, and open-ended are introduced in Section 2. Our method achieves high-quality occupancy annotation without additional manual labeling or post-processing while maintaining superior speed and generalization. C represents camera, and L denotes LiDAR.

Method	Categories	Modality	Manual-label	Post-processing	Speed	Zero-shot	Dynamic
Point-based Voxelization [47, 49, 51]	Close-set	L	3D GT	Human	Slow	✗	✓
2D-to-3D Projection [32, 60]	Close/Open-set	C&L	2D GT	Auto&Human	Slow	✗	✗
Ours (AutoOcc)	Open-ended	C or C&L	N/A	N/A	Fast	✓	✓

clouds (Table 1), requiring human pre-annotations and labor-intensive post-processing (over 4k+ human hours for nuScenes [47]). Current automated or semi-automated annotation pipelines primarily follow three paths. (1) Automated-assisted manual annotation, which is labor-intensive and costly. (2) Point cloud voxelization guided by manual annotation priors relies heavily on manual priors and multi-stage post-processing, making it time-consuming. (3) 2D-to-3D projection-based methods, which simply merge 2D segmentation results into 3D point clouds or meshes, struggle to ensure precise 3D consistency. These annotation methods heavily rely on LiDAR point clouds while overlooking semantic and geometric cues from multi-view images. Given that LiDAR point clouds are inherently sparse and incomplete, they are insufficient for comprehensive scene modeling. These approaches also employ voxel-based scene representations that require excessive parameters and incur redundant computational costs. Recent self-supervised occupancy models [4, 13, 14, 17, 59] have eliminated the need for extensive labeled training data by leveraging 2D features from image inputs and semantic information from visual foundation models (VFMs), such as SAM [20] and OpenSeed [61]. Nevertheless, these methods struggle to ensure complete, consistent scene occupancy, and exhibit limited generalization across diverse scenes.

Additionally, these pipelines are all confined to closed-set or open-set occupancy classes that require predefined categories. However, real-world scenes often involve open-ended occupancy—objects outside any predefined category, making it unwise to label all undefined semantics as "others." For example, self-driving vehicles may encounter collapsed poles or plastic sheets on road surfaces that require distinct occupancy annotations for safe driving strategies.

To address these limitations, we present AutoOcc, a fully automated framework for open-ended semantic occupancy annotation that requires neither manual labeling nor predefined categories. To achieve open-ended semantic occupancy labeling, we employ semantic attention maps generated by vision-language models (VLMs) to describe the scene, constructing a continuously evolving semantic query list. The generated attention maps are used simultaneously to prompt segmentation in SAM and guide instance-level depth estimation from UniDepth, thereby eliminating the need for manual annotations. We further introduce a self-estimated flow module to identify and manage dynamic objects in temporal rendering. We further propose Gaussian Splatting with open-ended semantic awareness (VL-GS) as an intermediate representation, offering a more comprehensive modeling, improved spatiotemporal consistency, and finer geometry with fewer primitives. Compared to densified point clouds and voxels, VL-GS achieves higher representation efficiency, greater accuracy, and reduced memory consumption. The semantic occupancy annotation is then automatically generated end-to-end through cumulative Gaussian-to-voxel splatting. Extensive experiments demonstrate that AutoOcc outperforms existing automated occupancy annotation methods. Our method further exhibits excellent open-ended and zero-shot generalization capabilities, as evidenced by cross-dataset experiments. Our main contributions include:

- We present AutoOcc, a vision-centric automatic annotation pipeline that supports open-ended semantic 3D occupancy label generation, based on vision-language guided differentiable reconstruction.
- We devise VL-GS, an efficient and comprehensive scene representation for semantic occupancy annotation. By integrating vision-language attention with visual foundation models, VL-GS effectively handles dynamic objects over time while enhancing both spatiotemporal consistency and 3D geometric detail in semantic occupancy.
- AutoOcc gains notable improvements over the existing automatic occupancy annotation pipelines, even without relying on manual priors. Our method also demonstrates strong generalization and open-ended understanding capabilities across unseen categories and diverse scenes.

2. Related Work

Semantic Occupancy Annotation. Semantic occupancy annotation aims to label semantic 0-1 occupancy from sensor data. However, current automated and semi-automated methods [47, 49, 51] heavily rely on LiDAR point clouds and human pre-annotated 2D or 3D ground truth. Most of these methods also require time-consuming post-processing and expensive manual purification. In contrast, we design a vision-centric fully automated occupancy annotation pipeline that eliminates the reliance on LiDAR. Our method also integrates VLMs and VFMs, supporting open-ended semantic category annotation.

3D Occupancy Estimation. Semantic 3D occupancy estimation [46, 49, 62] aims to estimate the occupancy states

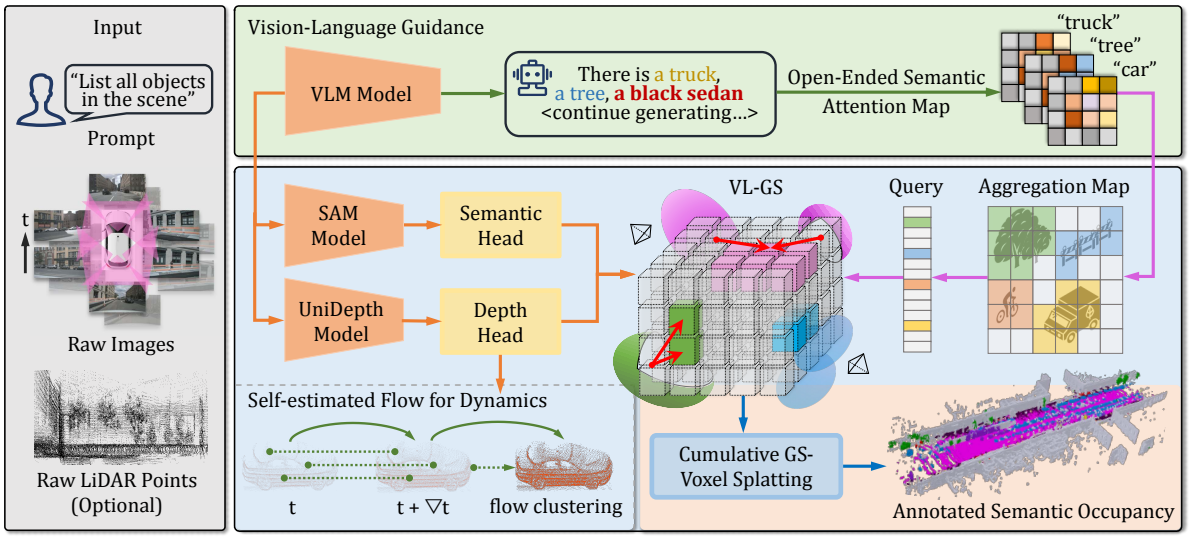


Figure 2. **Overall pipeline of our method.** AutoOcc is a vision-centric automated pipeline for semantic occupancy annotation. Our method starts with multi-view image inputs (optionally with LiDAR), extracts semantic attention maps from VLMs, and refines a dynamic semantic query list. We then propose Vision-Language Guided Gaussian Splatting (VL-GS), incorporating semantic-aware scalable Gaussians and self-estimated flow for dynamic objects. The final occupancy annotation is generated through a forward-pass Cumulative GS-Voxel Splatting. AutoOcc demonstrates strong generalization and open-ended annotation capabilities without relying on manual priors or LiDAR.

and semantics of complex scenes, which is crucial for 3D perception and planning. Existing learning-based occupancy models [16, 26, 27, 34, 43, 56, 62] are heavily reliant on the extensive labeled training data generated by annotation pipelines. Recent advances in self-supervised methods [3, 4, 13, 17, 59] for estimating 3D occupancy have diminished reliance on costly annotations and can be regarded as online occupancy labeling techniques. However, these approaches introduce ambiguity and illusions, resulting in misaligned geometry and temporal inconsistencies due to their limited awareness of intricate spatial structures and dynamic objects. They are also hampered by limited cross-dataset and scene-aware generalization capabilities.

To address these limitations, we propose a reconstruction-based occupancy annotation framework that requires no manual 2D or 3D annotations, achieving high-precision open-ended understanding, zero-shot learning, and cross-dataset generalization.

Scene Representation and Reconstruction. Efficient scene representation is the core of occupancy annotation. Dense voxel-based methods [5, 8, 23, 24] assign each voxel a feature vector, inevitably suffering from high computational cost due to redundant grids. As a compressed representation, BEV [6, 15, 25, 33] encodes 3D information on the ground plane, but struggles to capture diverse 3D geometry using flattened vectors. By implicitly modeling 3D space, [14, 17, 35, 59] create a NeRF-style 3D volume to estimate scene occupancy. However, the continuous implicit neural fields struggle with modeling complex dynamic scenes, and dense sampling leads to redundant, memory-intensive operations. Most recently, 3D Gaussian splatting (3DGS) [19, 37, 57] has demonstrated its powerful capability in reconstruction, even for driving scenes [12, 45, 67]. By treating each vertex as a Gaussian, [13] adopts a self-supervised approach for occupancy estimation but results in a dramatic increase in computational cost.

In contrast to prior art, we propose VL-GS, specifically designed to reconstruct semantic instances and dynamic objects, leveraging semantic attention clues from vision-language models. As a more efficient representation, VL-GS achieves high precision and versatile occupancy annotation with reduced cost.

Open-World Understanding. Existing open-world understanding methods [52] are confined to 2D images and can be broadly classified into two types: open-set [42] and open-ended [28]. Open-set methods [10, 22, 31] focus on text-image embedding matching using a predefined vocabulary bank. In contrast, open-ended methods [29, 29] continuously update observed object categories via language models. The key difference lies in the reliance on predefined categories, which allows open-ended approaches to produce more precise and comprehensive semantic representations, ultimately enhancing semantic occupancy annotation in open-world scenarios.

VLMs and VFMs. Vision language models (VLM) [21, 53] and visual foundation models (VFMs) [40, 42] have shown promising results and generalization ability in various visual tasks. However, their application to 3D occupancy annotation has received limited attention. Unlike direct training with 3D annotations, existing foundational vi-

sion models [18, 20, 30, 41] are primarily trained on 2D images, which may challenge the consistency of 3D occupancy across different cameras and frames. In this work, we explore the potential of applying VLMs [20, 30, 41] to occupancy annotation.

3. Method

As shown in Figure 2, we provide an overview of our proposed auto-annotation pipeline. Given a multi-view image sequence as input, we employ a fixed text prompt to enumerate all possible objects within the scene. Concurrently, our method supports LiDAR input, serving as a robust geometric prior constraint.

3.1. Vision-Language Guidance

Human annotations are both costly and labor-intensive. In contrast, world prior knowledge acquired from Vision-Language Models (VLMs) offers a cost-effective and efficient alternative, supporting open-ended semantic category perception. Current VLMs and VFMs are limited to specific 2D single-image tasks, such as captioning and segmentation. These methods often struggle with multimodal interactions and multi-view consistency, potentially leading to mismatches and 3D semantic ambiguities. Moreover, they lack a comprehensive understanding of the entire 3D space. To overcome these limitations, we propose a guidance framework centered around semantic attention maps and resolve ambiguities through scene reconstruction, thereby preserving 3D semantic and geometric coherence.

Semantic Attention Map. We employ semantic attention maps to integrate and guide the acquisition of desired prior knowledge from vision-language models at the semantic level. Given a multi-view image sequence, we prompt the VLM [9] to consistently generate all possible object categories within each image. Specifically, we use the attention map generation method [1, 29] to compute and aggregate the attentions from transformer decoder, with N output tokens $S = s_1, \cdots, s_N$ and the attention tensor $A \in \mathbb{R}^{H \times L \times N \times N}$, with H attention heads, L layers:

$$Attn(s_n^l) = \sum_{l=0}^{L} \left(\frac{1}{|H'|} \sum_{h \in H'} A_{h,l,k,j} \right), \tag{1}$$

where $s_n^l \in S$ is the output of n-th semantic from the transformer layer $l \in L$, $A_{h,l,k,j}$ is the attention tensor between query j and key k in the head h among subset of heads H'. We then rasterize the attention maps corresponding to these semantic categories into 2D feature maps, with each category represented by an aggregated attention map M. Notably, we establish a dynamically updated query list that incorporates the semantic information generated by VLMs. We implement a semantic integration strategy that merges similar sub-vocabularies with excessive gradients into unified semantic categories, thereby enhancing efficiency and mitigating visual ambiguity. For instance, we consolidate "tree" and "shrub" under the general term "vegetation".

Attention-guided Visual Prior. Semantic attention maps unveil category-related visual cues, which we subsequently leverage to guide the generation of semantic-aligned masks and depth information. Concretely, we input semantic attention maps as prompt cues into the off-the-shelf segmentation models [58, 65], which then generate multiple masks within the region of interest. These masks are merged into instance-level candidate masks to fully delineate the targeted semantic regions. The mask with the highest similarity score to the embeddings of the semantic attention query is then selected.

In parallel, we employ semantic attention maps to guide depth estimation [36, 55] at the semantic level, decoupling background and foreground objects while excluding sky regions to avoid interference from infinite distances. We then aggregate depth information from multi-view images using semantic attention cues, where pixels within each area of interest yield a set of pseudo 3D point clouds that represent an individual instance.

3.2. VL-GS

Although vision-language guidance provides valuable priors, 3D occupancy annotation still encounters three major challenges: 1) Semantic conflicts across multiple views make naïve 2D-to-3D projection prone to misalignment and ambiguity; 2) Errors in geometry and depth estimation can lead to distortions and misplacements in 3D space; 3) Dynamic objects disrupt both spatial and temporal consistency in semantics and geometry.

To overcome these challenges, we propose Vision-Language Guided Gaussian Splatting (VL-GS), which efficiently reconstructs the entire scene while maintaining both semantic and geometric 3D consistency by combining attention-based priors and differentiable rendering. The core of VL-GS is the semantic-aware scalable GS, guided by semantic attention maps from vision-language models. During reconstruction, VL-GS smooths out 2D semantic ambiguities at the instance level and optimizes the geometric details of objects. We also introduce a self-estimated flow module to capture and reconstruct dynamic objects using temporally-aware dynamic Gaussians. 3D Semantic occupancy is then directly annotated through cumulative GS-Voxel splatting, which is both efficient and precise.

Semantic-aware Scalable Gaussian. Obviously, different semantic objects occupy varying "weights" within a scene, which is intuitively reflected in their semantic occupancy across scales. Meanwhile, the ability to model at multiple granularities is expected to represent the diverse geometric complexities of instances. Based on this, we

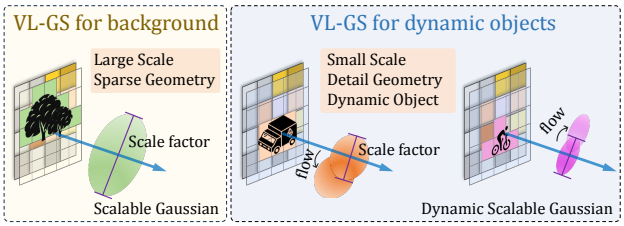


Figure 3. **Vision-Language Guided Gaussian Splatting (VL-GS)** efficiently reconstructs semantic instances using a scalable strategy guided by semantic attention maps from VLMs. Additionally, VL-GS models dynamic objects through dynamic Gaussians driven by self-estimated flow.

propose designing a semantic-aware scalable Gaussian that adaptively scales and reconstructs different semantic objects. Unlike dense voxels or point clouds, our method allows for representing regions of interest with sparse Gaussians, aided by scalability and semantic attention maps.

Given semantic attention cues from VLMs, we assign semantic attributes and corresponding scaling factors to each Gaussian. The blended semantic category of Gaussians can be obtained via α-blending:

$$\Gamma_i = \sum_{i=1}^{N} \text{softmax}(\gamma_i)\alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{2}$$

where Γ_i is the rendered semantic for each pixel, weighted by the Gaussians' semantic attributes γ and opacity α. The scaling factor needs to be linearly related to the space occupied by each Gaussian, which cannot be simply calculated from the Gaussian centroid position $\{o_x, o_y, o_z\}$ due to the variations in anisotropic shape and spatial overlap. Thus, we estimate the occupied range of each Gaussian by considering the distance from the nearest tangent surface of the Gaussian ellipsoid to the voxel as:

$$d = o_z - \frac{\eta^{-1}\Sigma_{0,2}^{-1}(o_x - \kappa_x) + \eta^{-1}\Sigma_{1,2}^{-1}(o_y - \kappa_y)}{\Sigma_{2,2}^{-1}}, \tag{3}$$

where d is the occupied depth from the voxel to the Gaussian ellipsoid, η is the ray direction from the voxel center $k = (\kappa_x, \kappa_y, \kappa_z)$ to the Gaussian. Σ is the covariance matrix, with $\Sigma_{i,j}$ denoting the corresponding matrix elements. The Gaussian value $G(x)$ can be formulated as:

$$G(x) = e^{-\frac{1}{2}(\kappa-o)^\top \Sigma^{-1}(\kappa-o)}. \tag{4}$$

The scaling factor is then adaptively adjusted based on the gradients of Gaussian values and the occupied range of the Gaussians. Notably, Gaussians of the same semantic category share similar scaling factor ranges, as objects with the same semantics exhibit comparable scales and geometries. As shown in Figure 3, semantic-aware scalable gaussians enable the representation of large background areas (e.g., buildings) with sparse gaussians at a larger scale, while capturing finer geometries (e.g., cyclists) with denser gaussians at a smaller scale.

Self-estimated Flow for Dynamic Objects. Dynamic objects could cause trailing effects due to temporal variations, thereby reducing the accuracy of occupancy annotation. Independently handling dynamic objects facilitates the enhancement of temporal and spatial consistency in semantics. Thus, we introduce a self-estimated 3D flow module, which is used to capture and aggregate dynamic objects. We also assign dynamic attributes to dynamic Gaussians to better model the motion of objects.

Specifically, we model the translation of each Gaussian kernel p from time t to time $t + \Delta t$ as a flow vector f. Our goal is to minimize the point distances between the object's source points U_1 and target points U_2 to estimate the flow by applying Chamfer distance (CD) [11]. Since the same dynamic object is often represented by spatially adjacent Gaussians with the same semantics, we search for correspondences between paired points among the nearest Gaussian neighbors that share the same semantic:

$$CD(p, p') = \sum_{p \in U_1} \min_{p' \in U_2} ||p - p'||_2^2 + \sum_{p' \in U_2} \min_{p \in U_1} ||p' - p||_2^2, \tag{5}$$

where p and p' are the positions of Gaussian kernels with the same semantic at time t and $t + \Delta t$, respectively. We define a dynamic indicator function between paired Gaussians to determine whether an object is in motion:

$$\mathbb{1}(D) = \rho - \frac{1}{m}\sum_{i=1}^{m} \|p_{t+\Delta t}^i - p_t^j\|_2, \tag{6}$$

where D is the average distance between paired Gaussians with the same semantic, ρ is the dynamic threshold, and m denotes the number of Gaussian ellipsoids. The centroid position at the i-th frame is denoted by o_i. Subsequently, we aggregate all temporally paired Gaussians based on the semantic attention map and motion cues.

Geometric constraints from LiDAR. LiDAR points are widely used by existing occupancy annotation methods due to their precise geometric priors. Our pipeline also supports the use of LiDAR to obtain geometric constraints and continuously optimize the distribution of Gaussians.

Similar to [63, 64], a point $p_{i,t}$ in the LiDAR sweep L_t is projected onto the frame I_t, and its initial semantic label can be obtained by $K^{-1}[R^\top \phi_{x,y,t} + T]$, where (K, R, T) are the corresponding camera parameters and homogenous transformation matrix, and ϕ is the pixel-level semantic label. We aggregate the multi-frame of LiDAR points over time and compute the anchor centers $p_c = (x_c^i, y_c^i, z_c^i)$. We then implement a geometry-aware loss to enforce the alignment of Gaussian ellipsoid distributions with the geometric

Table 2. **Semantic occupancy annotation on Occ3D-nuScenes [46].** C represents camera, and L denotes LiDAR. "cons. veh." and "drive. surf." stand for construction vehicles and driveable surfaces, respectively. AutoOcc-V uses only images as input, while AutoOcc-M integrates both camera and LiDAR data. The intersection over union (IoU) and mean IoU of semantic classes (mIoU) are calculated over all voxels. For fair comparisions, we replicate SurroundOcc* [51] and OpenOcc* [49] by replacing the manually annotated results with the semantic point clouds projected from VLMs.

Method	Input	IoU ↑	mIoU ↑	barrier	bicycle	bus	car	cons. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	sidewalk	terrain	manmade	vegetation
GaussianOcc [13]	C	51.22	12.59	1.88	6.42	13.94	16.75	2.02	3.41	6.84	12.33	1.75	10.32	41.28	19.32	18.26	12.41	21.88
LangOcc [3]	C	46.55	12.04	2.73	7.21	5.78	13.92	0.51	10.80	6.42	8.67	3.24	11.02	42.10	12.44	27.17	14.13	14.55
VEON [66]	C	57.92	14.51	5.03	4.65	13.88	11.04	9.63	10.25	4.51	10.99	4.32	12.63	47.50	11.43	20.52	25.43	25.76
SurroundOcc* [51]	L	68.87	18.59	18.68	17.23	18.19	18.31	10.27	18.29	17.34	14.95	21.19	19.88	21.33	20.74	18.11	23.26	21.02
OpenOcc* [49]	C&L	70.59	17.76	23.73	8.06	26.10	22.95	11.72	11.59	10.36	9.72	5.60	19.13	39.51	22.15	20.87	13.19	21.81
VLM-LiDAR	C&L	73.28	16.32	13.34	10.37	17.04	20.65	7.26	15.20	14.61	5.88	19.40	21.47	15.13	13.32	15.74	28.17	27.24
OVIR-3D [32]	C&L	54.30	18.47	18.54	10.69	15.30	23.82	9.42	13.13	11.57	8.32	10.19	20.49	36.85	24.22	21.84	16.30	36.33
AutoOcc-V	C	83.01	20.92	12.70	10.45	7.81	20.42	5.79	17.58	18.50	24.25	4.23	12.88	55.54	24.23	27.14	35.62	36.61
AutoOcc-M	C&L	88.62	25.84	21.19	16.08	18.42	25.90	4.32	14.58	25.62	27.18	3.51	20.93	58.38	32.03	29.80	46.15	43.59

priors of their corresponding semantic regions:

$$L_{geo} = -\sum_{c=1}^{C}\sum_{i=1}^{M}\frac{1}{\|o_c(i) - p_c(i)\|_2^2}, \tag{7}$$

where C denotes the number of semantic categories, M is the number of Gaussian ellipsoid centers within the anchor range, and o_i is the coordinate of the i-th Gaussian center.

Cumulative GS-Voxel Splatting. Finally, we cumulatively splat VL-GS onto the voxel grid at an arbitrary voxel size, with each voxel's semantic label determined by weighting the occupied range and opacity from Gaussians:

$$F(o) = \sum_{i=1}^{N} d_i G(x_i)\alpha_i \mathrm{softmax}(\gamma_i), \tag{8}$$

where d_i is the occupied depth of the Gaussian-to-3D voxel, treated as the splatting weight coefficient. α_i is the opacity, and $\mathrm{softmax}(\gamma_i)$ computes the semantic probability.

4. Experiments

4.1. Implementation Details

We use two benchmarks for evaluation: Occ3D-nuScenes, which is used to compare the performance of our method with other occupancy annotation methods for specific categories, while SemanticKITTI is used to assess the zero-shot capability across datasets and unseen categories. We set the resolutions of images as 900 × 1600 for Occ3D-nuScenes and 370 × 1226 for SemanticKITTI. During optimization, we scale the image size to 225 × 400 and double it every 300 steps until reaching the original resolution. We use the AdamW optimizer for optimization with an initial learning rate of 0.005. The learning rate for the position parameters decays every 250 steps with a decay rate of 0.98. Similar to [59, 66], we evaluate without the "other" and "other flat" classes. All experiments are carried out on 8 A100 GPUs.

4.2. Performance Evaluation and Analysis

We evaluate our method against the state-of-the-art (SOTA) methods for automatic semantic occupancy annotation, including offline methods [32, 49, 51] and self-supervised online methods [3, 13, 66].

Compared with point-based voxelization pipelines. Point-based voxelization annotation pipelines directly use LiDAR with 3D annotations (semantic points and 3D bounding boxes) as input. SurroundOcc [51] performs mesh reconstruction and a nearest neighbors algorithm to densify semantic points. OpenOcc [49] proposes the AAP pipeline to densify the voxel, followed by human post-processing to purify artifacts. For fair comparisons, we replicate these methods by replacing the manually annotated results with the semantic point clouds projected from VLMs. As shown in Table 2, our vision-centric method outperforms these pipelines that utilize LiDAR point clouds.

Compared with 2D-to-3D projection methods. Projecting annotated or generated 2D labels back onto a 3D representation is a natural idea, which is further refined by several methods [32, 54]. However, these methods rely on pre-built 3D representations (e.g., point clouds or meshes) and employ multi-stage post-processing, including voting, filtering, and merging, to eliminate overlapping information. Undoubtedly, this strategy leads to the loss of crucial details and misalignment between semantics and representations. AutoOcc performs well against SAMPro3D [54] and OVIR-3D [32], both of which project the outputs of SAM [41] onto 3D point clouds. We also design a baseline that directly projects the results of VLM and SAM onto LiDAR point clouds (VLM-LiDAR) and voxelizes them into semantic occupancy. Table 2 shows that AutoOcc still

Table 3. **Zero-shot cross-dataset performance on SemanticKITTI [2].** Novel class refers to entirely new, unseen semantics in nuScenes, while base class includes those seen during training. All compared methods are trained on Occ3D-nuScenes and evaluated on SemanticKITTI. Metric mIoU-base denotes the mIoU computed solely on base classes from Occ3D-nuScenes.

	(a) Val: SemanticKITTI			(b) Novel Class							(c) Base Class											
Method	Input	IoU ↑	mIoU ↑	bicyclist	moto-cyc.	parking	fence	trunk	pole	traffic-sign	bicycle	car	other-veh.	motorcycle	pedestrian	truck	road	sidewalk	terrain	building	vegetation	mIoU-base ↑
GaussianOcc [13]	C	22.42	4.18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.33	7.10	2.81	3.06	2.91	3.42	15.80	10.43	3.78	2.55	22.11	6.84
OVO [44]	C	20.94	5.83	0.90	0.0	0.68	3.50	2.31	0.60	2.20	0.40	12.70	3.50	0.20	0.74	0.70	19.44	24.81	4.86	11.70	15.62	8.61
SurroundOcc [51]	L	27.83	6.39	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.52	23.19	4.81	6.71	4.37	3.16	24.32	11.98	9.95	5.79	19.14	10.45
VLM-LiDAR	C&L	28.12	5.32	0.0	0.0	0.0	0.0	0.0	0.0	0.0	2.04	19.17	3.31	2.13	2.64	5.89	19.02	16.58	6.31	3.59	14.98	8.69
AutoOcc-V	C	35.64	9.36	1.38	3.60	0.59	4.34	5.36	14.32	6.62	4.71	22.29	3.89	10.35	7.54	8.78	26.14	15.66	9.84	4.14	18.87	12.02
AutoOcc-M	C&L	41.23	12.76	1.27	5.23	0.33	5.71	5.97	15.17	8.72	7.83	24.60	4.92	9.30	11.18	8.39	44.74	24.43	5.85	17.01	29.12	17.03

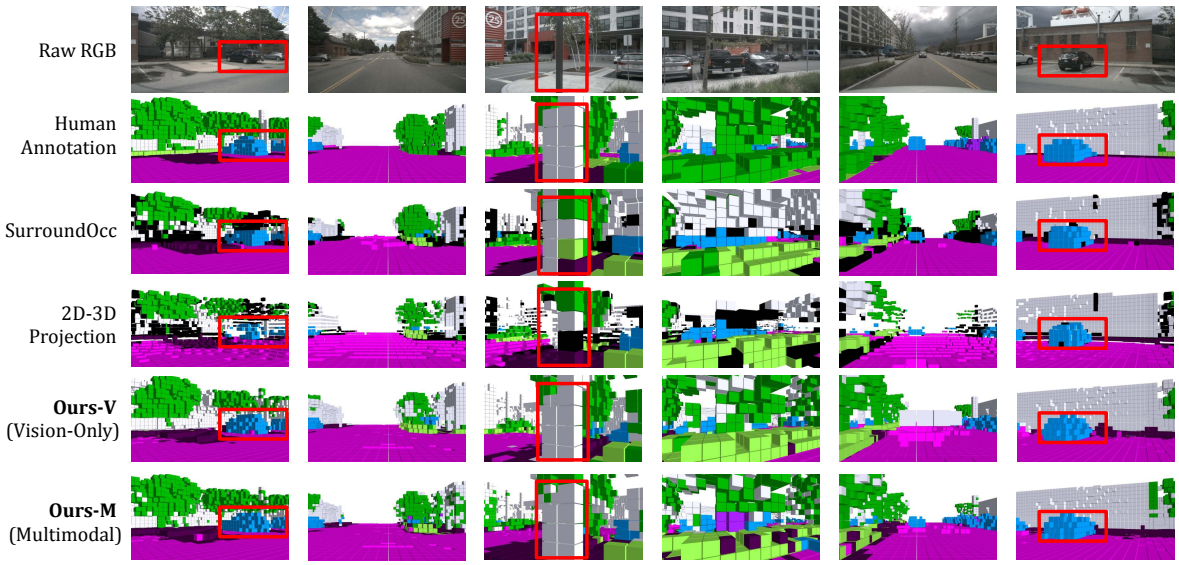


Figure 4. **Qualitative results of semantic occupancy annotation on Occ3D-nuScenes [46].** Our method enables high-quality annotation of semantic 3D occupancy, capturing fine-grained geometry, structurally challenging regions, and dynamic objects across complex scenes.

demonstrates better performance, based on the deep integration of VLM guidance and differentiable reconstruction.

Compared with self-supervised methods. Self-supervised methods enable occupancy estimation from image features without relying on manual annotations. For a fair comparison, we extend existing self-supervised approaches by incorporating image sequences as historical frames and performing multi-frame feature aggregation. We further perform temporal fusion of the above outputs in the global coordinate system. As shown in Table 2, using pure visual input, our method outperforms GaussianOcc [13], which utilizes vanilla GS as an intermediate representation. AutoOcc also performs well against LangOcc [3] and VEON [66], which are specifically designed for open-vocabulary occupancy estimation in surrounding-view scenes. While the aforementioned approaches do not require additional supervision, they struggle with efficiently modeling semantic geometry and neglect dynamic objects, leading to performance degradation.

Qualitative results. Figure 4 and 5 show that our method excels in semantic occupancy annotation, showcasing superior scene completeness, consistency, and dynamic object handling, even without the use of LiDAR. In extreme weather conditions (e.g., rain and nighttime), our method maintains robust performance, achieving annotation results comparable to or even surpassing manually labeled ground truth. For instance, in areas where ground truth is missing due to rain, AutoOcc successfully reconstructs both the geometry and semantics of the road surface.

4.3. Zero-shot and Generalization Ability

SemanticKITTI differs from Occ3D-nuScenes in terms of semantic categories, sensor parameters, camera distribution, and voxel size. We evaluate on SemanticKITTI to verify the zero-shot and cross-dataset generalization capability.

Table 4. **Comparisons of annotation efficiency.** Open-ended stands for the annotation capability for undefined classes. Label-free means training without any human-labeled annotations. † indicates the use of VLMs to obtain 2D semantics instead of human labeling.

Method	Anno. Time	Input Modality	Representation	Memory	Number	Open-Ended	Label-Free
Auto+Human [49]	4000+ human hours	L	Point Clouds	-	1.2 M	✗	✗
GaussianOcc [13] †	≈60 GPU hours	C	Vanilla GS	32 G	0.8 M	✗	✓
SurroundOcc [51] †	1000+ GPU hours	L	Mesh & Voxel	73 G	3.0 M	✗	✗
VLM-LiDAR †	≈50 GPU hours	C&L	Point Clouds	34 G	1.2 M	✗	✗
AutoOcc †	≈30 GPU hours	C or C&L	VL-GS	5.0 G	0.3 M	✓	✓

To evaluate the zero-shot and open-ended semantic annotation ability, we select novel classes from SemanticKITTI as the test set, which are not visible during the annotation process. Table 3 shows that all self-supervised methods [13, 44] suffer significant performance degradation, as they are tailored to specific camera parameters and occupancy distributions. For novel classes unseen during learning, these methods fail to label undefined semantic occupancy. Compared to offline annotation pipelines, including point-based voxelization and semantic projection, our method shows better robustness and enhanced capability for open-ended semantic annotation.

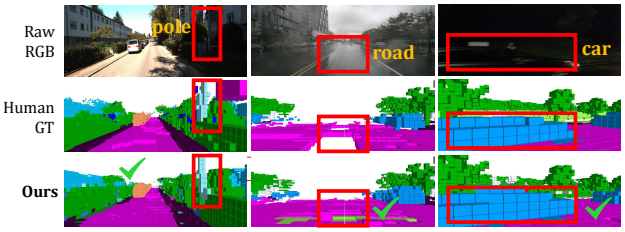


Figure 5. Qualitative comparison of our method with human annotations under complex lighting and extreme weather conditions.

4.4. Annotation Efficiency

Table 4 presents evaluations on representation characteristics and model efficiency. Notably, AutoOcc demonstrates an advantage in computational cost, delivering better performance with reduced memory requirements. In contrast, scene representations based on dense voxels and Point Clouds incur redundant computational costs. In addition, AutoOcc strikes a balance between efficiency and flexibility, enabling open-ended scene-aware occupancy reconstruction, supporting open-vocabulary semantic occupancy annotation, and requiring no human-labeled annotations.

4.5. Ablation Studies

We analyze the effect of the self-estimated flow module for dynamic objects by disabling the clustering of dynamic objects and optimizing them together with static foregrounds. Table 5 and Figure 6 show that the self-estimated flow module effectively mitigates the challenges of dynamic trailing and spatial occlusion in the annotation of occupancy. We further ablate the effect of semantic-aware scalable Gaussians and LiDAR geometric priors by replacing the scalable

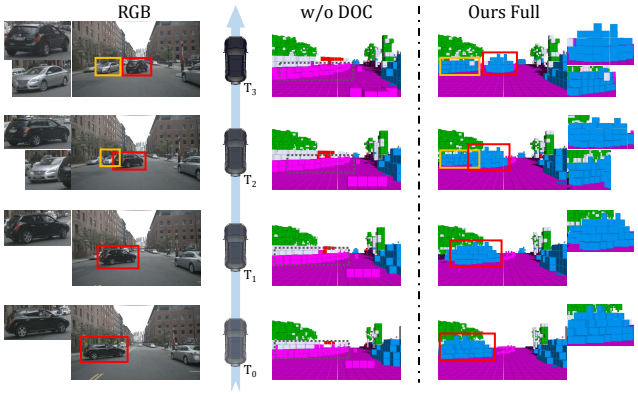


Figure 6. **Semantic occupancy of dynamics.** AutoOcc accurately annotates the semantic occupancy of dynamic objects, maintains spatiotemporal consistency, and infers occluded parts.

Table 5. **The effect of each module in our method.** SFM is short for the self-estimated flow module, and SSG denotes the employment of the semantic-aware scalable Gaussians.

Model	**IoU** ↑	**mIoU** ↑
w/o SFM	82.65	16.84
w/o SSG	80.27	17.67
w/o L_{geo}	81.49	20.36
AutoOcc-V	83.01	20.92
AutoOcc-M	88.62	25.84

Gaussians and removing the L_{geo} loss from our framework. The degraded results highlight the importance of geometric priors in constraining the shape and distribution of Gaussians, leading to more accurate occupancy reconstruction.

5. Conclusion

In this paper, we propose AutoOcc, an vision-centric automated pipeline for open-ended semantic 3D occupancy annotation that integrates differentiable Gaussian splatting guided by vision-language models. To facilitate scene understanding, we leverage VLMs and build an efficient and comprehensive scene representation for occupancy annotation. AutoOcc integrates vision-language attention with visual foundation models, effectively handles dynamic objects over time, and enhances both spatiotemporal consistency and geometric detail. Our framework achieves state-of-the-art performance on open-ended semantic occupancy annotation and performs favorably against other automated annotation pipeline, without using any human annotations.

Acknowledgment

This work was supported by National Key R&D Program of China (Grant No. 2022ZD0160305). This work was also a research achievement of Key Laboratory of Science, Technology, and Standard in Press Industry (Key Laboratory of Intelligent Press Media Technology). Ming-Hsuan Yang was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korean Government (MSIT) (No. RS-2024-00457882, National AI Research Lab Project).

References

[1] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928*, 2020. 4

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

[3] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Self-supervised open vocabulary occupancy estimation via volume rendering. *arXiv preprint arXiv:2407.17310*, 2024. 3, 6, 7

[4] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Occflownet: Towards self-supervised occupancy estimation via differentiable rendering and occupancy flow. *arXiv preprint arXiv:2402.12792*, 2024. 2, 3

[5] Anh-Quan Cao and Raoul De Charette. Monoscene: Monocular 3d semantic scene completion. In *CVPR*, pages 3991–4001, 2022. 3

[6] 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. 3

[7] Devendra Singh Chaplot, Murtaza Dalal, Saurabh Gupta, Jitendra Malik, and Russ R Salakhutdinov. Seal: Self-supervised embodied active learning using exploration and 3d consistency. *NIPS*, 34:13086–13098, 2021. 1

[8] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *CVPR*, pages 21674–21683, 2023. 3

[9] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*, 2024. 4

[10] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *CVPR*, 2024. 3

[11] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 5

[12] Tobias Fischer, Jonas Kulhanek, Samuel Rota Bulo, Lorenzo Porzi, Marc Pollefeys, and Peter Kontschieder. Dynamic 3d gaussian fields for urban areas. *arXiv preprint arXiv:2406.03175*, 2024. 3

[13] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv e-prints*, 2024. 2, 3, 6, 7, 8

[14] 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. 2, 3

[15] Adam W Harley, Zhaoyuan Fang, Jie Li, Rares Ambrus, and Katerina Fragkiadaki. Simple-bev: What really matters for multi-sensor bev perception? In *ICRA*, pages 2759–2765. IEEE, 2023. 3

[16] 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. 3

[17] 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, 3

[18] Lei Ke, Mingqiao Ye, Martin Danelljan, Yu-Wing Tai, Chi-Keung Tang, Fisher Yu, et al. Segment anything in high quality. *NIPS*, 36, 2024. 4

[19] 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. 3

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 2, 4

[21] Weicheng Kuo, Yin Cui, Xiuye Gu, AJ Piergiovanni, and Anelia Angelova. F-vlm: Open-vocabulary object detection upon frozen vision and language models. *arXiv preprint arXiv:2209.15639*, 2022. 3

[22] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded language-image pre-training. In *CVPR*, 2022. 3

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

[24] Yanwei Li, Xiaojuan Qi, Yukang Chen, Liwei Wang, Zeming Li, Jian Sun, and Jiaya Jia. Voxel field fusion for 3d object detection. In *CVPR*, pages 1120–1129, 2022. 3

[25] Yangguang Li, Bin Huang, Zeren Chen, Yufeng Cui, Feng Liang, Mingzhu Shen, Fenggang Liu, Enze Xie, Lu Sheng, Wanli Ouyang, et al. Fast-bev: A fast and strong bird's-eye view perception baseline. *TPAMI*, 2024. 3

[26] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, pages 1–18. Springer, 2022. 3

[27] 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. 3

[28] Chuang Lin, Yi Jiang, Lizhen Qu, Zehuan Yuan, and Jianfei Cai. Generative region-language pretraining for open-ended object detection. In *CVPR*, 2024. 3

[29] Zhiwei Lin, Yongtao Wang, and Zhi Tang. Training-free open-ended object detection and segmentation via attention as prompts. *arXiv preprint arXiv:2410.05963*, 2024. 3, 4

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 4

[31] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *ECCV*, 2024. 3

[32] 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 *Conference on Robot Learning*, pages 1610–1620. PMLR, 2023. 2, 6

[33] Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Bev-guided multi-modality fusion for driving perception. In *CVPR*, pages 21960–21969, 2023. 3

[34] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 3

[35] 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. 3

[36] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 4

[37] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *CVPR*, pages 20051–20060, 2024. 3

[38] Santhosh K Ramakrishnan, Ziad Al-Halah, and Kristen Grauman. Occupancy anticipation for efficient exploration and navigation. In *ECCV*, pages 400–418. Springer, 2020. 1

[39] Santhosh K Ramakrishnan, Dinesh Jayaraman, and Kristen Grauman. An exploration of embodied visual exploration. *IJCV*, 129(5):1616–1649, 2021. 1

[40] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3

[41] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 4, 6

[42] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *TPAMI*, pages 1757–1772, 2012. 3

[43] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. *arXiv preprint arXiv:2407.04049*, 2024. 3

[44] Zhiyu Tan, Zichao Dong, Cheng Zhang, Weikun Zhang, Hang Ji, and Hao Li. Ovo: Open-vocabulary occupancy. *arXiv preprint arXiv:2305.16133*, 2023. 7, 8

[45] Qijian Tian, Xin Tan, Yuan Xie, and Lizhuang Ma. Drivingforward: Feed-forward 3d gaussian splatting for driving scene reconstruction from flexible surround-view input. *arXiv preprint arXiv:2409.12753*, 2024. 3

[46] 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. *NIPS*, 36, 2024. 2, 6, 7

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

[48] Lizi Wang, Hongkai Ye, Qianhao Wang, Yuman Gao, Chao Xu, and Fei Gao. Learning-based 3d occupancy prediction for autonomous navigation in occluded environments. In *IROS*, pages 4509–4516. IEEE, 2021. 1

[49] 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. 2, 6, 8

[50] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *CVPR*, pages 17158–17168, 2024. 1

[51] 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. 2, 6, 7, 8

[52] Jianzong Wu, Xiangtai Li, Shilin Xu, Haobo Yuan, Henghui Ding, Yibo Yang, Xia Li, Jiangning Zhang, Yunhai Tong, Xudong Jiang, et al. Towards open vocabulary learning: A survey. *TPAMI*, 2024. 3

[53] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 3

[54] 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. 6

[55] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 4

[56] Zichen Yu, Changyong Shu, Jiajun Deng, Kangjie Lu, Zongdai 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. 3

[57] Zehao Yu, Anpei Chen, Binbin Huang, Torsten Sattler, and Andreas Geiger. Mip-splatting: Alias-free 3d gaussian splatting. In *CVPR*, pages 19447–19456, 2024. 3

[58] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023. 4

[59] 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 e-prints*, pages arXiv–2312, 2023. 2, 3, 6

[60] Dingyuan Zhang, Dingkang Liang, Hongcheng Yang, Zhikang Zou, Xiaoqing Ye, Zhe Liu, and Xiang Bai. Sam3d: Zero-shot 3d object detection via segment anything model. *arXiv preprint arXiv:2306.02245*, 2023. 2

[61] Hao Zhang, Feng Li, Xueyan Zou, Shilong Liu, Chunyuan Li, Jianwei Yang, and Lei Zhang. A simple framework for open-vocabulary segmentation and detection. In *ICCV*, pages 1020–1031, 2023. 2

[62] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *ICCV*, pages 9433–9443, 2023. 2, 3

[63] Lin Zhao, Hui Zhou, Xinge Zhu, Xiao Song, Hongsheng Li, and Wenbing Tao. Lif-seg: Lidar and camera image fusion for 3d lidar semantic segmentation. *IEEE Transactions on Multimedia*, 26:1158–1168, 2023. 5

[64] Xiangmo Zhao, Pengpeng Sun, Zhigang Xu, Haigen Min, and Hongkai Yu. Fusion of 3d lidar and camera data for object detection in autonomous vehicle applications. *IEEE Sensors Journal*, 20(9):4901–4913, 2020. 5

[65] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023. 4

[66] 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, 2025. 6, 7

[67] 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 *CVPR*, pages 21634–21643, 2024. 3