Language Driven Occupancy Prediction

Zhu Yu[1,†] Bowen Pang[1,2,†] Lizhe Liu[2,✉] Runmin Zhang[1] Qiang Li[2]
Si-Yuan Cao[3,1] Maochun Luo[2] Mingxia Chen[2] Sheng Yang[2] Hui-Liang Shen[1,✉]
[1]Zhejiang University [2]Unmanned Vehicle Dept., CaiNiao Inc., Alibaba Group
[3]Ningbo Global Innovation Center, Zhejiang University

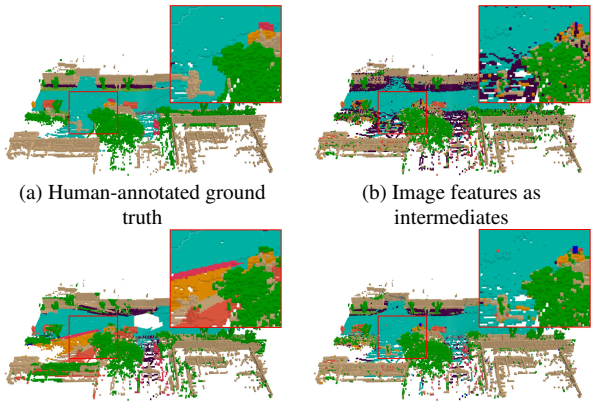
Project Page: https://github.com/pkqbajng/LOcc

Abstract

*We introduce **LOcc**, an effective and generalizable framework for open-vocabulary occupancy (OVO) prediction. Previous approaches typically supervise the networks through coarse voxel-to-text correspondences via image features as intermediates or noisy and sparse correspondences from voxel-based model-view projections. To alleviate the inaccurate supervision, we propose a **semantic transitive labeling** pipeline to generate dense and fine-grained 3D language occupancy ground truth. Our pipeline presents a feasible way to dig into the valuable semantic information of images, transferring text labels from images to LiDAR point clouds and ultimately to voxels, to establish precise voxel-to-text correspondences. By replacing the original prediction head of supervised occupancy models with a geometry head for binary occupancy states and a language head for language features, LOcc effectively uses the generated language ground truth to guide the learning of 3D language volume. Through extensive experiments, we demonstrate that our transitive semantic labeling pipeline can produce more accurate pseudo-labeled ground truth, diminishing labor-intensive human annotations. Additionally, we validate LOcc across various architectures, where all models consistently outperform state-of-the-art zero-shot occupancy prediction approaches on the Occ3D-nuScenes dataset.*

1. Introduction

Vision-based occupancy prediction aims to estimate both the complete scene geometry and semantics using only image inputs, which serves as a critical foundation for various 3D perception tasks, such as autonomous driving [10], embodied agent [15, 36], mapping and planning [34, 42]. Existing supervised occupancy prediction approaches [26, 27, 33, 34, 37, 39] are generally constrained to predicting a fixed set of semantic categories, as the dense occupancy ground truth is generated by merging multi-frame LiDAR point clouds with semantic annotations. In this process, human annotators must label each point across thousands of LiDAR frames with a semantic class, which remains highly labor-intensive and costly. Instead, open-vocabulary occupancy (OVO) facilitates predicting the occupancy with arbitrary sets of vocabularies using only unlabeled image-LiDAR data for training [35, 56].



(a) Human-annotated ground truth (b) Image features as intermediates (c) Voxel-based model-view projection (d) Semantic transitive labeling (ours)

Figure 1. Comparison of pseudo-Labeled 3D language occupancy ground truth from different pipelines.

Due to the lack of large-scale occupancy datasets with language annotations, distilling knowledge from vision-language models [4, 18, 31, 35] into occupancy prediction networks presents a viable solution. Some approaches [18, 35] use image features from the pretrained vision-language models [18, 31] as intermediates [4, 35], subsequently bridging the gap between 3D voxel and language features. However, the feature values of objects with the same semantics may vary, as image features inherently encode both semantic and appearance information, leading to inconsistencies in semantic representations across different inputs. Consequently, the pseudo-labeled ground truth derived from these features is inherently noisy in terms of

†: Equal contribution. ✉: Corresponding author.

semantics, failing to provide precise guidance for voxel-level understanding. Moreover, the single-scan sparse LiDAR [35] can only provide sparse geometry supervision, further resulting in sparse and coarse voxel-to-text correspondences. To address this, VEON [56] enhances geometry supervision by employing dense binary occupancy maps, which are generated through offline post-processing. However, it treats voxels as point clouds and assigns semantic labels to them by model-view projection, which overlooks the occlusion problem and only uses single-frame images, leading to coarse voxel-to-text correspondences, too.

To address the aforementioned issues, we revisit the core aspect of OVO: generating dense and fine-grained 3D language occupancy ground truth. In this work, we propose a **semantic transitive labeling** pipeline to achieve this goal, without laborious procedures. It can be roughly divided into two steps: label transferring and scene reconstruction. For the first step, we use large vision-language models (LVLM) [1, 2, 24] and open-vocabulary segmentation models (OV-Seg) [8, 43, 44] to assign text labels to each pixel, obtaining consistent semantic representations. Then, through the projection of LiDAR points, the text labels are transferred to the LiDAR points, bridging the gap between 3D data and texts. In the reconstruction process, we model the occlusion relationships of the point clouds during point projection, preventing the wrong label assignment of voxel-based model-view projection. With the pseudo-labeled LiDAR data, we reconstruct a temporally dense scene by merging multi-frame LiDAR points to obtain a dense 3D spatial representation. Furthermore, to reduce the impact of segmentation noise from individual frames, we apply majority-voting voxelization to assign the most frequent label to each voxel. The resulting ground truth enhances 3D language supervision for open-vocabulary occupancy networks. A comparison of different pseudo labeled ground truth from different pipelines is illustrated in Fig. 1.

Based on the semantic transitive labeling pipeline, we devise **LOcc**, an effective and generalizable framework that is compatible with most existing supervised occupancy models [11, 12, 23] for OVO. To efficiently align high-dimensional CLIP [31] embeddings, we devise a language autoencoder that maps them to a low-dimensional latent space, reducing computational cost. We validate our framework on several mainstream approaches, referred to as LOcc-BEVFormer, LOcc-BEVDet, and LOcc-BEVDet4D. All three models consistently surpass all the previous state-of-the-art zero-shot occupancy prediction approaches on the Occ3D-nuScenes dataset. Notably, even the simpler LOcc-BEVDet, with an input resolution of 256×704, achieves a mIoU of 20.29, outperforming previous methods that rely on temporal image inputs, higher-resolution inputs, or larger backbone networks. In summary, our contributions are summarized as follows:

- We propose LOcc, an effective and generalizable framework that is compatible with most existing supervised methods for OVO. To reduce the computational cost for the language feature alignment, a text-based autoencoder is introduced to map the high-dimensional CLIP embeddings to a low-dimensional latent space.
- We propose a semantic transitive labeling pipeline for generating dense and fine-grained 3D language occupancy ground truth. This pipeline ensures that text semantic labels can be effectively transferred from images to LiDAR point clouds and ultimately to voxels. We experimentally verify that our pipeline can produce more accurate pseudo-labeled ground truth, diminishing time-consuming human annotations.
- We validate our LOcc framework with several mainstream models, which all demonstrate superior performance over previous state-of-the-art methods, proving its generablizability and effectiveness.

2. Related Work

2.1. Vision-based Occupancy Prediction

Occupancy prediction jointly estimates the occupancy state and semantic label for every voxel in the scene using images. Voxels are classified as free, occupied, or unobserved, with occupied voxels assigned corresponding semantic labels. Early works [3, 6, 22, 52, 55, 57] typically use monocular or stereo images as inputs, also referred to as semantic scene completion (SSC). With advancements in 3D perception [11–13, 19, 21, 23, 29, 46, 47, 50, 51], the occupancy of surrounding scenes using multi-camera images [13, 26, 40, 41, 49] has garnered significant attention due to its strong representation capabilities in 3D space. SurroundOcc [39] introduces a new pipeline for constructing dense occupancy ground truth. Occ3D [33] establishes benchmarks for surround occupancy prediction through a three-stage data generation pipeline: voxel densification, occlusion reasoning, and image-guided voxel refinement. OpenOccupancy [37] provides a benchmark with higher-resolution ground truth, while OccNet [34] offers flow annotations.

2.2. Open-Vocabulary Understanding

The recent advances in vision-language models [31] have shown a remarkable zero-shot performance by using only paired images and captions for model training, demonstrating the strong connection between images and natural language. Then, numerous works explore the potential of language driven downstream applications, such as object detection [7, 53], tracking [20], and segmentation [8, 18, 43, 44]. Taking the image features of these 2D foundation models as intermediaries, OpenScene [28] distills 2D vision-language knowledge into a 3D point cloud network and

achieves zero-shot 3D segmentation. LERF [16] is the first to embed CLIP features into NeRF for controllable rendering. LangSplat [30] uses SAM [17] to tackle the point ambiguity issue for 3D language field modeling.

2.3. Open-Vocabulary Occupancy Prediction

Benefiting from the development of vision-language foundation models, OVO enables the prediction of occupancy associated with arbitrary vocabulary sets without requiring manual annotations of extensive 3D voxel data. POP-3D [35] employs only unlabeled images and sparse LiDAR point clouds for training. It distills vision-language knowledge from the pretrained LSeg [18] into the occupancy networks by computing cosine similarity between 3D features and their corresponding 2D features, obtained by projecting the point cloud onto the image plane. VEON [56] introduces a side adaptor network to integrate CLIP [31] into the occupancy prediction model. However, the extracted image features may exhibit inconsistencies across different input images, lacking accurate guidance for voxel-level understanding and leading to coarse voxel-to-text correspondences. Furthermore, current works typically treat the voxels as point clouds and employ voxel-based model-view projection to obtain supervision, leading to a lots of misprojections.

3. Method

Our proposed LOcc framework mainly consists of two aspects: generating dense and fine-grained pseudo-labeled 3D language occupancy ground truth, diminishing laborious human annotations (Sec. 3.1), and using the generated ground truth to train the OVO networks (Sec. 3.2). To reduce computational cost, an autoencoder is introduced to map the CLIP [31] embeddings into a lower dimensional latent space (Sec. 3.3).

3.1. Pseudo-Labeled Ground Truth Generation

We propose a semantic transitive labeling pipeline to generate dense and fine-grained 3D language ground truth. The overview framework is illustrated in Fig. 2. Specifically, we first use vision-language foundation models, including Large Vision-Language Model (LVLM) and Open-Vocabulary Segmentation Model (OV-Seg), to associate each pixel with a text-based pseudo label. We then project the unlabeled LiDAR point clouds onto the image plane to propagate these pseudo labels to each point. The resultant pseudo-labeled LiDAR point clouds can further be used for scene reconstruction. Each voxel subsequently derives its pseudo label from the pseudo-labeled LiDAR points it contains, ultimately generating dense and fine-grained pseudo-labeled language occupancy ground truth.

Vocabulary Extraction with LVLM. Benefiting from the advent of LVLMs [1, 24], we can decode visual information in images using textual descriptions. Here, we employ the LVLM as an initial perception model to list the objects within each image. Fig. 3 illustrates the process of requiring the LVLM to list all classes within an image using Qwen-VL [2]. The conversation process follows the chain-of-thought [38] approach. Specifically, rather than directly requesting the LVLM to generate class nouns, we first ask it to provide a description of the scene, followed by a dialogue in which we prompt the LVLM to list the names of all identified classes. To enhance the LVLM's understanding of outdoor environments, we provide an example in advance to assist this process when analyzing a large number of cases. The results extracted from single-frame surround images are then consolidated to represent the overall classes for that frame.

Pixel-to-Text Association with OV-Seg. Having obtained the vocabularies within a frame of surround images, we can establish pixel-to-text association with open-vocabulary semantic segmentation models [8, 43, 44]. This process begins by mapping the image and vocabularies into a shared feature space via the feature-aligned vision encoder Φ_I and text encoder Φ_T, respectively. Each pixel is then assigned a text label by computing cosine similarity with the set of vocabulary embeddings, with the highest-scoring text designated as the label for that pixel. Specifically, for the k-th frame surround-view images $\mathcal{I}_k = \left\{\mathbf{I}_{i,k} \in \mathbb{R}^{H\times W\times 3}, i \in \{1,\dots,N_c\}\right\}$ and vocabularies $\mathcal{T}_k = \{\mathbf{T}_{j,k}, j \in \{1,...,N_{t,k}\}\}$, their corresponding segmentation maps $\mathcal{S}_k = \left\{\mathbf{S}_{i,k} \in \mathbb{R}^{H\times W}, i \in \{1,...,N_c\}\right\}$ can be obtained as

$$\begin{aligned}\mathbf{F}_{i,k} = \Phi_I\left(\mathbf{I}_{i,k}\right) \in \mathbb{R}^{H\times W\times C}, \mathbf{E}_{j,k} = \Phi_T\left(\mathbf{T}_{j,k}\right) \in \mathbb{R}^{C},\\ \mathbf{S}_{i,k}(x,y) = \underset{\mathbf{T}_{j,k}}{\arg\max} \cos(\mathbf{F}_{i,k}(x,y), \mathbf{E}_{j,k}),\end{aligned} \tag{1}$$

where $\mathbf{F}_i$ represents the feature maps of images and $\mathbf{E}_j$ represents the embeddings of texts. Here, $N_{t,k}$ indicates the length of the vocabulary set for the k-th frame. Typically, Φ_I generates feature maps at a lower resolution and the final high-resolution segmentation maps are produced using an upsampling decoder [8]. For simplicity, we let $\mathbf{F}_i$ have the same resolution with the original image.

In our investigation, we found that the LVLM occasionally fails, overlooking many important classes and listing only a limited vocabulary set within a single frame. To address this issue, we treat multiple frames as a unified sequence, integrating the vocabularies from each frame into a cohesive set $\mathcal{T} = \{\mathbf{T}_j, j \in \{1,...,N_t\}\}$, where $N_t = \sum_k^K N_{t,k}$. This integrated set is then used to perform open-vocabulary segmentation as described in Eq. 1. The key motivation behind this approach is the overlap between consecutive frames, where recurring objects display a degree of similarity. Consequently, frames with incomplete text classes can be supplemented by the outputs from adjacent

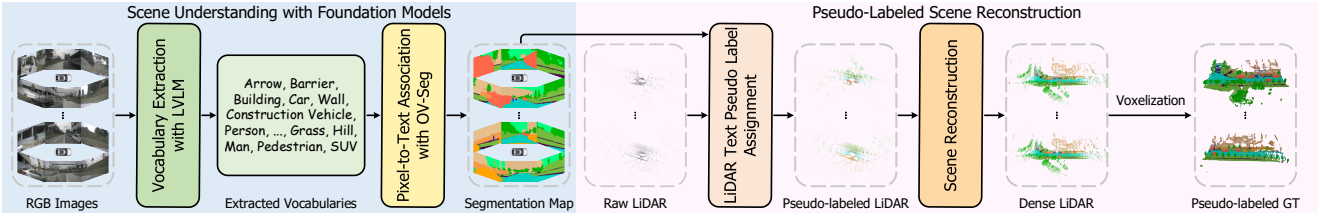


Figure 2. Framework of our semantic transitive labeling pipeline for generating dense and fine-grained pseudo-labeled 3D language occupancy ground truth. Given surround images of the entire scene, we first map them to a vocabulary set using a large vision-language model (LVLM) and associate pixels with these terms through an open-vocabulary segmentation model (OV-Seg). Based on the segmentation results, we assign pseudo labels to the points by projecting them onto the corresponding image plane and merge all the point clouds to form a complete scene. Finally, we voxelize the scene reconstruction results to establish dense and fine-grained voxel-to-text correspondences, generating the pseudo-labeled ground truth.

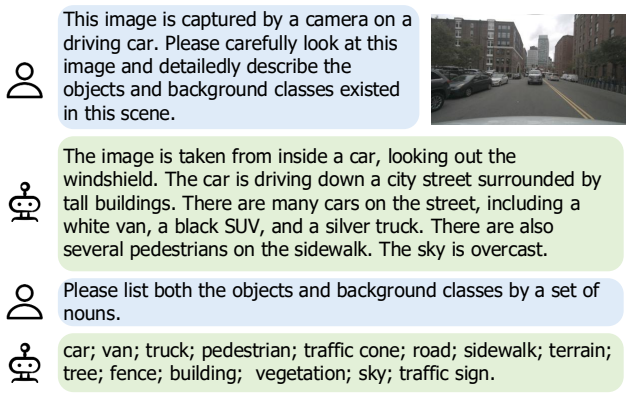


Figure 3. An example of the conversation process with Qwen-VL [2] to extract vocabularies from an individual image.

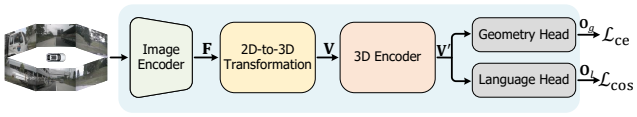


Figure 4. Overview of the open-vocabulary occupancy (OVO) model architecture.

frames within the same sequence.

LiDAR Text Pseudo Label Assignment. Using the generated segmentation maps, we can assign text pseudo labels to the unlabeled LiDAR $\mathcal{P} = \{\mathbf{P}_k, k \in \{1, ..., K\}\}$. Specifically, for a point $\mathbf{P}_k(p)$ in the k-th frame point cloud, its corresponding image coordinates on the N_c surround images are defined as

$$(x_i, y_i, z_i) = \mathbf{K}_i \left(\mathbf{R}_i \mathbf{P}_k(p) + \mathbf{t}_i\right), i \in \{1, ..., N_c\}, \quad (2)$$

where $\mathbf{K}_i, \mathbf{R}_i, \mathbf{t}_i$ are camera intrinsic and extrinsic parameters. Among these coordinates, the one that lies within the image boundaries and has the smallest depth value z_i serves as the pixel corresponding to point $\mathbf{P}_k(p)$. This pixel is then used to sample the segmentation map to obtain the pseudo label $\hat{\mathbf{S}}(p)$ for $\mathbf{P}_k(p)$. This process can be summarized as

$$n = \arg\min_m z_m, 0 < x_m < W, 0 < y_m < H,$$
$$\hat{\mathbf{S}}(p) = \phi\left(\mathbf{S}_{n,k}, (x_n, y_n)\right), \quad (3)$$

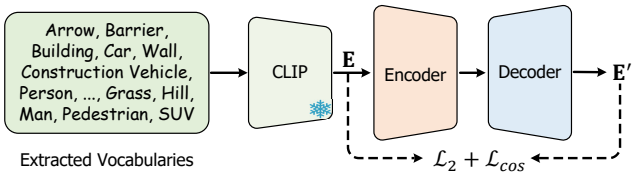


Figure 5. Framework of the autoencoder to map the language features into a low-dimensional latent space.

where ϕ indicates the nearest neighbor sampling of $\mathbf{S}_{n,k}$ at the coordinate (x_n, y_n).

Scene Reconstruction. The pseudo-labeled LiDAR data is then used to perform scene reconstruction to establish correspondences between the texts and the voxels. To combine the multi-frame point clouds, we first transform their coordinates into the world coordinate system using the calibration matrices and ego poses. In alignment with VEON [56], we distinguish moving and static objects by leveraging the geometric information of 3D bounding boxes. For the k-th frame, we transform the unified point cloud to this frame and apply majority-voting voxelization on it to obtain the pseudo-labeled 3D language occupancy ground truth. Each voxel takes the most frequent text within it as its text label, and voxels without point clouds are set to empty. Compare to the voxel-based model-view projection, our method obtains text label by projecting LiDAR point clouds onto image planes, which are easily calibrated with the corresponding images during the data collection process. Besides, the majority-voting is more robust and less susceptible to the influence of single-point projection errors, generating ground truth with higher metrics compared to the nearest-point voxelization, as shown in the ablation experiments.

3.2. Open Vocabulary Occupancy Prediction

Based on the processes described above, we successfully establish dense and fine-grained voxel-to-text correspondences. By minimally modifying most existing supervised occupancy methods [12, 12, 23], we can leverage the obtained pseudo-labeled 3D language occupancy ground truth

to train a robust OVO model, without the need for additional complex network designs. The overall network architecture is shown in Fig. 4. Similar to most existing occupancy network structures, it involves an image encoder, a 2D-to-3D view transformation module, a 3D encoder, except the original prediction head is replaced by a geometry head to predict the binary occupancy state and a language head to estimate the language features.

Image Encoder. The image encoder includes an image backbone network for extracting multi-scale features, complemented by an FPN [45] network to enhance features across different levels. The enhanced features $\mathbf{F} \in \mathbb{R}^{H \times W \times C}$ are then used to construct 3D features.

2D-to-3D Transformation. The 2D-to-3D view transformation aims to project multiview image features into the 3D voxel space. This can be achieved by either forward projection [29] or backward projection [23]. In this work, we verify the effectiveness of our proposed method on both projection approaches [11, 12, 23]. The constructed 3D volume $\mathbf{V}$ has a shape of $X \times Y \times Z$ for voxel representation or $X \times Y$ when pooled into BEV representation

3D Encoder. After constructing the 3D volume $\mathbf{V}$, the 3D encoder further refines its representation by performing 3D convolution [9] or stacking self-attention layers [23] to obtain the fine-grained volume $\mathbf{V}'$. This refined volume $\mathbf{V}'$ is then fed to a geometry head to estimate binary occupancy states $\mathbf{O}_g$ and a language head to generate 3D language volume $\mathbf{O}_l$.

Training and Evaluation. During training, we jointly supervise the geometry and language heads to optimize the neural network. Given the binary occupancy ground truth $\hat{\mathbf{O}}_g$, we adopt cross entropy loss to supervise $\mathbf{O}_g$,

$$\mathcal{L}_{ce} = \text{CrossEntropy}(\hat{\mathbf{O}}_g, \mathbf{O}_g). \tag{4}$$

Given the pseudo-labeled 3D language ground truth, each voxel of it corresponds to a text. We perform alignment between $\mathbf{O}_l$ and the CLIP embeddings of these texts $\hat{\mathbf{O}}_l$ using cosine similarity,

$$\mathcal{L}_{cos} = \sum_{p \in \mathcal{C}(\hat{\mathbf{O}}_g)} 1 - \cos(\hat{\mathbf{O}}_l(p), \mathbf{O}_l(p)), \tag{5}$$

where $\mathcal{C}(\hat{\mathbf{O}}_g)$ denotes the set of occupied voxels in $\hat{\mathbf{O}}_g$. The final loss function $\mathcal{L}$ is a combination of $\mathcal{L}_{\text{ce}}$ and $\mathcal{L}_{\text{cos}}$,

$$\mathcal{L} = \mathcal{L}_{\text{ce}} + \mathcal{L}_{\text{cos}}. \tag{6}$$

During evaluation, $\mathbf{O}_g$ first defines the set of occupied voxels. Then for each occupied voxel, its corresponding language feature is used to compute cosine similarity with the predefined text features, where the text with the highest score represents its semantics.

3.3. Language Autoencoder

Generally, the dimension of CLIP features is 512. To reduce the memory requirements, we introduce a language autoencoder to map the language features into a low-dimensional latent space. The extracted vocabularies of the entire dataset contains approximately one thousand unique classes, allowing the autoencoder to be trained efficiently. Fig. 5 shows the framework of the autoencoder.

Specifically, given the vocabulary set $\mathcal{T}_a = \{\mathbf{T}_i, i \in 1, ..., N_a\}$ of the entire dataset, the encoder Φ_E maps the D-dimensional CLIP features $\mathbf{E} \in \mathbb{R}^D$ to $\mathbf{E}' \in \mathbb{R}^{D'}$, where $D' < D$. Then we learn a decoder Φ_D to reconstruct the original embedding from the compressed $\mathbf{E}'$. The autoencoder is trained with a combination of L2 loss and cosine loss,

$$\mathcal{L}_a = ||\mathbf{E} - \hat{\mathbf{E}}||_2 + \cos(\mathbf{E}, \hat{\mathbf{E}}), \tag{7}$$

where $\hat{\mathbf{E}}$ denotes the reconstructed embedding. After training the autoencoder, the language volume of the OVO model is aligned with the compressed embedding.

4. Experiments

4.1. Datasets and Metrics

Datasets. We evaluate LOcc on the nuScenes dataset [5, 33], a large-scale autonomous driving dataset collected across various cities and weather conditions. Each frame includes a LiDAR scan and six RGB images from different viewpoints. It contains 700 training scenes and 150 validation scenes. The spatial volume ranges from $-40m$ to $40m$ for the X and Y axes, and $-1m$ to $5.4m$ for the Z axis. Voxelization of this volume produces a set of 3D grids with a resolution of $200 \times 200 \times 16$, where each voxel measures $0.4m \times 0.4m \times 0.4m$. The ground truth includes 17 unique labels (16 semantic classes and 1 free class).

Metrics. Following previous approaches [33], we list the intersection over union (IoU) and mean IoU (mIoU) metrics for occupied voxel grids and voxel-wise semantic predictions, respectively, providing a comprehensive analysis for geometry and semantic aspects of the scene.

4.2. Evaluation Benchmarks

The evaluation benchmarks focus on two settings, i.e., zero-shot and open-vocabulary. In the zero-shot setting, none of the ground truth classes are seen during training. In contrast, the open-vocabulary setting assumes that only a subset of classes is available during training, and the model is expected to recognize both seen and unseen classes during evaluation

a

Table 1. Zero-shot occupancy prediction results on the Occ3D-nuScenes dataset. We report the mean IoU (mIoU) for semantics across different categories, along with per-class semantic IoUs. The symbol * indicates reproduced results from [49]. The methods marked with † indicate that they specifically design the set of vocabularies for training. The methods marked with ‡ are implemented the official open-source codebase. The top three results among the methods supervised by pseudo-labeled ground truth are in **red**, green, and blue, respectively.

Model	Image Backbone	Image Size	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Supervised by human-labeled ground truth.																				
MonoScene [6]	ResNet-101	900 × 1600	6.06	1.75	7.23	4.26	4.93	9.38	5.67	3.98	3.01	5.90	4.45	7.17	14.91	6.32	7.92	7.43	1.01	7.65
OccFormer [55]	ResNet-101	900 × 1600	21.93	5.94	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	30.96	34.66	22.73	6.76	6.97
TPVFormer [13]	ResNet-101	900 × 1600	27.83	7.22	38.90	13.67	40.78	45.90	17.23	19.99	18.85	14.30	26.69	34.17	55.65	35.47	37.55	30.70	19.40	16.78
CTF-Occ [33]	ResNet-101	900 × 1600	28.5	8.09	39.33	20.56	38.29	42.24	16.93	24.52	22.72	21.05	22.98	31.11	53.33	33.84	37.98	33.23	20.79	18.0
BEVFormer wo TSA* [23]	ResNet-101	900 × 1600	38.05	9.11	45.68	22.61	46.19	52.97	20.27	26.5	26.8	26.21	32.29	37.58	80.5	40.6	49.93	52.48	41.59	35.51
BEVFormer* [23]	ResNet-101	900 × 1600	39.04	9.57	47.13	22.52	47.61	54.14	20.39	26.44	28.12	27.46	34.53	39.69	81.44	41.14	50.79	54.00	43.08	35.60
CVT-Occ* [49]	ResNet-101	900 × 1600	40.34	9.45	49.46	23.57	49.18	55.63	23.10	27.85	28.88	29.07	34.97	40.98	81.44	40.92	51.37	54.25	45.94	39.71
BEVDet [12]	ResNet-50	256 × 704	33.62	8.18	38.46	13.75	41.10	45.17	18.99	18.39	19.52	19.36	29.48	31.77	79.42	37.55	49.08	51.70	36.94	32.63
BEVDet4D [11]	ResNet-50	256 × 704	39.11	11.03	46.68	21.59	44.13	50.70	25.30	24.12	25.89	25.21	34.13	38.94	81.33	39.08	51.96	56.13	47.41	41.15
Supervised by images.																				
SelfOcc (BEV) [14]	ResNet-50	384 × 800	6.76	0.00	0.00	0.00	0.00	9.82	0.00	0.00	0.00	0.00	0.00	6.97	47.03	0.00	18.75	16.58	11.93	3.81
SelfOcc (TPV) [14]	ResNet-50	384 × 800	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 [54]	ResNet-101	928 × 1600	10.81	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
Supervised by images and unlabeled LiDAR.																				
VEON-B† [56]	ViT-B	256 × 704	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
VEON-L† [56]	ViT-L	512 × 1408	15.14	**0.90**	10.40	**6.20**	17.7	12.7	8.50	7.60	6.50	5.50	**8.20**	11.80	54.50	0.40	25.50	30.20	25.40	25.40
POP-3D-BEVFormer [35]	ResNet-101	256 × 704	11.24	0.00	7.32	0.95	11.23	9.18	2.59	2.44	2.21	0.78	1.7	8.26	56.02	0.00	19.28	21.47	21.54	26.18
POP-3D-BEVFormer [35]	ResNet-101	512 × 1408	12.68	0.00	9.43	2.43	15.21	10.39	3.54	3.78	3.98	2.91	1.96	10.01	57.28	0.00	21.01	23.07	22.67	27.87
POP-3D-BEVDet [35]	ResNet-50	256 × 704	12.30	0.00	8.72	1.94	13.11	11.48	3.12	3.29	3.47	1.78	1.21	9.43	56.81	0.00	20.17	23.29	23.33	28.12
POP-3D-BEVDet [35]	ResNet-50	512 × 1408	12.85	0.00	7.43	2.21	15.13	12.25	3.40	3.47	4.21	2.22	2.42	10.67	58.23	0.00	21.46	24.33	22.16	28.88
POP-3D-BEVDet4D [35]	ResNet-50	256 × 704	14.57	0.00	9.94	3.09	15.83	13.51	4.38	3.57	4.64	3.84	2.21	12.20	59.71	0.00	26.60	28.16	27.88	32.20
POP-3D-BEVDet4D [35]	ResNet-50	512 × 1408	15.64	0.00	10.06	3.63	16.74	14.32	5.38	4.01	5.14	3.52	2.46	15.20	61.13	0.00	28.82	30.16	31.92	33.31
LOcc-BEVFormer (ours)	ResNet-101	256 × 704	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
LOcc-BEVFormer (ours)	ResNet-101	512 × 1408	19.85	0.00	14.65	1.01	27.99	35.34	5.61	4.16	11.70	3.05	2.72	23.08	72.00	0.00	39.35	34.18	32.03	30.66
LOcc-BEVDet (ours)	ResNet-50	256 × 704	20.29	0.00	14.03	2.81	31.75	34.55	7.52	3.94	8.99	3.64	3.04	25.58	71.43	0.01	38.74	35.62	32.61	30.68
LOcc-BEVDet (ours)	ResNet-50	512 × 1408	20.84	0.00	13.85	3.28	31.85	37.09	8.22	4.84	10.78	3.24	2.47	26.95	72.35	0.00	39.58	35.29	33.35	31.23
LOcc-BEVDet4D (ours)	ResNet-50	256 × 704	22.95	0.00	15.01	5.76	29.86	38.90	**11.61**	4.61	13.46	6.99	2.64	**30.69**	73.61	0.00	40.38	**39.16**	38.80	38.68
LOcc-BEVDet4D (ours)	ResNet-50	512 × 1408	**23.84**	0.00	**16.92**	5.89	**32.94**	**40.08**	10.18	**9.85**	**17.11**	**7.21**	3.10	30.48	**74.13**	0.00	**41.25**	36.97	**39.57**	**39.66**

Table 2. Ablation study on the framework for generating dense and fine-grained pseudo-labeled 3D language occupancy ground truth. For label transferring process, we analyze three settings: (a) using image features from vision-language foundation models as intermediates, (b) using single-frame vocabularies, and (c) integrating vocabularies across multiple consecutive frames. For scene reconstruction process, we replace the majority-voting voxelization with (d) nearest-point voxelization and (e) voxel-based model-view projection. The best results for each model are in **bold**.

Setting	Label Transferring Process			Scene Reconstruction Process			mIoU			
	Image Feat.	Single-frame vocab.	Consecutive-frame vocab.	Point-based Projection: Voting voxelization	Point-based Projection: Nearest voxelization	Voxel-based Projection	LOcc-BEVFormer	LOcc-BEVDet	LOcc-BEVDet4D	Pseudo-labeled GT
(a)	✓			✓			15.04	17.56	19.37	22.12
(b)		✓		✓			17.87	19.61	22.15	22.66
(c)			✓	✓			**18.62**	**20.29**	**22.95**	**25.53**
(d)			✓		✓		18.32	19.98	22.68	23.80
(e)			✓			✓	13.77	14.49	15.41	19.55

4.3. Implementation Details

Network Structures. To demonstrate the generalizability of the proposed LOcc, we evaluate it on BEVFormer [23], BEVDet [12], and BEVDet4D [11], including both forward- and backward-projection methods. For BEVDet [12] and BEVDet4D [11], we employ ResNet-50 [9] as the image backbone. The resolution of the voxel features after the 2D-to-3D transformation is $200 \times 200 \times 16$, with a feature dimension of 64. For BEVFormer [23], we adopt ResNet101-DCN [9] as the backbone, with the default size for BEV queries set to 200×200, with a feature dimension of 256. We remove the temporal self-attention layer and use 6 cross-attention layers. For each query, it corresponds to 4 target points with different heights in 3D space, with height anchors uniformly predefined from $-1m$ to $5.4m$. The number of sampling points around each reference point is set to 8. For all models, the image size is resized to either 256×704 or 512×1408, and the language features are compressed to a dimension of 128.

Training Setup. We implement the networks using Py-

Table 3. Detailed metrics of the pseudo-labeled ground truth from different generation settings on the Occ3D-nuScenes dataset. The settings match those outlined in Table 2. We report the mean IoU (mIoU) for semantics across different categories. The best results of different settings are in **bold**.

Setting	mIoU	others	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
(a)	22.12	0.00	**10.20**	7.18	29.59	35.22	**15.70**	12.80	11.19	1.86	**15.28**	21.56	58.52	0.00	25.80	22.72	55.53	52.92
(b)	22.66	0.00	8.07	4.51	23.33	37.48	9.62	11.42	14.04	10.50	4.61	19.86	60.30	0.24	30.94	25.97	59.17	65.22
(c)	**25.53**	0.00	9.40	**8.47**	**30.16**	**38.00**	11.43	**20.20**	**11.39**	**11.54**	9.55	**25.73**	**61.86**	0.40	**36.41**	**33.00**	**59.42**	**67.10**
(d)	23.80	0.00	8.90	6.75	27.86	34.21	10.19	15.75	10.05	8.68	8.83	22.67	60.91	**0.50**	34.91	31.82	56.91	65.70
(e)	19.55	0.00	7.19	4.28	24.62	21.15	8.30	8.43	6.71	4.63	8.65	16.01	55.41	**0.50**	26.32	31.01	50.95	58.22

Torch. The models are trained for 25 epochs on 8 NVIDIA 3090 GPUs, with a batch size of 1 on each GPU. We employ the AdamW [25] optimizer with $\beta_1 = 0.9, \beta_2 = 0.99$. The maximum learning rate is set to 3×10^{-4}, and the cosine annealing learning rate strategy is adopted for the learning rate decay, where the cosine warm up strategy is applied for the first 5% iterations.

4.4. Zero-shot Occupancy Prediction

We evaluate the effectiveness of the proposed LOcc framework on three models: BEVFormer [23], BEVDet [12] and BEVDet4D [11], referred to as LOcc-BEVFormer, LOcc-BEVDet, and LOcc-BEVDet4D, respectively. The comparison methods includes those that rely solely on images [14, 54] and those that utilize unlabeled LiDAR data to assist in geometry prediction [35, 56]. To ensure a fair comparison with POP-3D, which uses single-frame LiDAR for geometry supervision, we replace the single-frame LiDAR data with dense binary occupancy ground truth [56]. For OV-Seg models, we replace LSeg [18] with more powerful SAN [44].

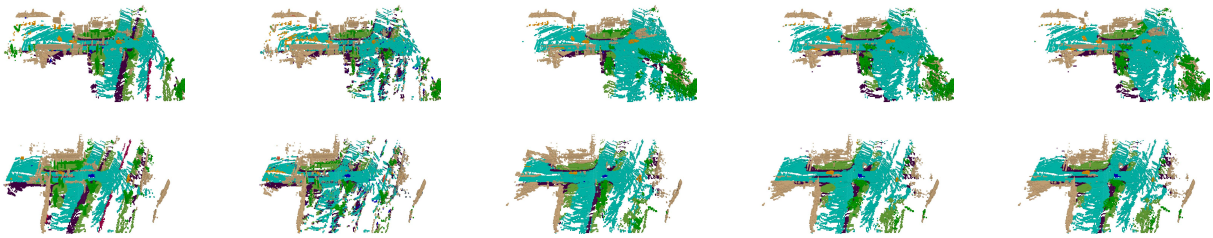
Table 1 displays detailed results of zero-shot occupancy prediction. With the input image of size 256×704, LOcc-BEVFormer, LOcc-BEVDet, and LOcc-BEVDet4D achieve an mIoU of 18.62, 20.29, and 22.95, respectively. Notably, LOcc-BEVDet, a more simpler model, surpasses all previous approaches, even though many of these approaches use larger image backbones or higher-resolution inputs. Furthermore, when the image resolution is increased to 512×1408, the models attain an better mIoU of 19.85, 20.84, and 23.84.

4.5. Ablation Study

We conduct ablation study to evaluate the effectiveness of the proposed LOcc framework.

Label Transferring Process. Table 2 provides an analysis of the label transferring process. First, we replace the segmentation maps with image features extracted from SAN [44], requiring the learned 3D features to align with these image features. This configuration indirectly establishes correspondences with textual labels. Setting (a) demonstrates that it leads to a performance drop in the generated pseudo-labeled ground truth, which in turn degrades the performance of all models. Next, we apply our proposed semantic transitive labeling pipeline using only single-frame vocabularies. As shown in setting (b), the mIoU of the pseudo-labeled ground truth increases to 22.66, and the metrics for all models improve to 17.87, 19.61, and 22.15, respectively. Finally, by treating multiple consecutive frames as a unified sequence and consolidating the vocabulary from each frame into a cohesive set for pseudo-label assignment, as indicated in setting (c), we observe additional performance gains, further validating the effectiveness of our label transferring process.

Scene Reconstruction Process. The voxel-based model-view projection treats all voxels as point clouds, making it challenging to determine whether a voxel lies on the surface after projection. This coarse approach overlooks the occlusion problem, leading to noisy voxel-to-text correspondences. Setting (e) demonstrates that the pseudo-labeled ground truth from voxel-based model-view projection can only achieve an mIoU of 19.55. Using this unsatisfied ground truth to train the OVO models results in a huge performance drop to 13.77, 14.49, and 15.41. Instead, each frame of LiDAR can be accurately calibrated with corresponding images during the data collection process, significantly reducing misprojections. Thus, performing scene reconstruction with nearest-point voxelization yields notable improvements, as shown in setting (d). Furthermore, majority-voting voxelization can mitigate the issue of wrong segmentation results, achieving better results with the mIoU of pseudo-labeled and models improving to



(a) Ground Truth (b) Pseudo Ground Truth (c) LOcc-BEVFormer (d) LOcc-BEVDet (e) LOcc-BEVDet4D

Figure 6. Quantitative visualization results on the Occ3D-nuScenes [33] dataset.

Table 4. Open-vocabulary occupancy prediction results on the Occ3D-nuScenes dataset. We report the mean IoU (mIoU) for semantics across different categories, along with per-class semantic IoUs. The best results among the methods are in **bold**.

Model	Image Backbone	Image Size	mIoU	bicycle	motorcycle	traffic cone	sidewalk	others	barrier	bus	car	const. veh.	pedestrian	trailer	truck	drive. suf.	other flat	terrain	manmade	vegetation
LOcc-BEVFormer (ours)	ResNet-101	256 × 704	23.26	18.71	22.21	24.33	48.14	0.00	14.48	28.16	33.88	4.32	9.19	2.37	22.45	72.09	0.00	34.32	31.22	29.58
LOcc-BEVDet (ours)	ResNet-50	256 × 704	23.50	12.21	17.33	17.24	47.03	0.00	13.43	32.67	35.54	8.68	10.01	2.54	27.33	72.16	0.00	36.31	34.47	32.57
LOcc-BEVDet4D (ours)	ResNet-50	256 × 704	26.91	20.34	23.77	23.93	51.47	0.00	15.63	29.54	39.89	10.99	14.01	6.21	31.72	73.23	0.00	40.11	39.00	37.76

Table 5. Ablation study on the OV-Seg models. The best results among different methods are in **bold**.

Method	mIoU		
	LOcc-BEVDet	LOcc-BEVDet4D	LOcc-BEVFormer
ODISE [43]	20.17	22.52	**18.69**
CAT-Seg [8]	**20.68**	**22.76**	18.57
SAN [44]	20.29	22.68	18.62

25.53, 18.62, 20.29, and 22.68, as illustrated in setting (c). These results prove that majority-voting is more robust and less susceptible to the influence of single-point projection errors. The overall ablation results prove that the quality of the pseudo-labeled ground truth highly influences the performance of OVO models. In Table 3, we present detailed metrics for the pseudo-labeled ground truth generated under various settings. Consistent with prior analyses, our pipeline yields better results on most of categories, further confirming the effectiveness of our proposed pipeline.

OV-Seg Models. Table 5 lists the results of the OVO networks trained on the pseudo-labeled ground truth generated from different OV-Seg models: SAN [44], ODISE [43] and CAT-Seg [8]. In this work, we use SAN [44] for fair comparison with previous approaches [56].

4.6. Open-vocabulary Occupancy Prediction

Table 4 lists the detailed metrics of open vocabulary occupancy prediction. We use *bicycle*, *motorcycle*, *traffic cone*, and *sidewalk* as base classes, while treating the remaining categories as novel classes. With more seen classes, the overall performance for the models rise. Besides, the IoUs on unseen classes are always competitive, proving the open-vocabulary capability of our methods.

4.7. Qualitative Results

Fig. 6 presents visualizations of the pseudo-labeled ground truth, along with results from LOcc-BEVFormer, LOcc-BEVDet, and LOcc-BEVDet4D trained on it. Each model demonstrates clear predictions, despite the noise present in the training dataset. This demonstrates the ability of the models to learn useful information from semi-labeled data. These findings are also observed in other vision tasks, such as image segmentation [17], video segmentation [32] ,and depth estimation [48],

5. Conclusions

In this paper, we have presented LOcc, an effective and generalizable framework for open-vocabulary occupancy prediction. We propose a semantic transitive labeling pipeline for generating dense and fine-grained 3D language occupancy ground truth. We experimentally demonstrates that text semantic labels can be effectively transferred from images to LiDAR point clouds and ultimately to voxels, diminishing labor-intensive human annotations. By replacing the original prediction head with a geometry head and a language head, LOcc is compatible with most existing supervised networks. Based on the BEVFormer, BEVDet, and BEVDet4D, our propose method can achieve better results than all the previous state-of-the-art zero-shot occupancy prediction methods.

Acknowledgments

We thank the reviewers for the valuable discussions. This research was supported by the National Key Research and Development Program of China under grant 2023YFB3209800, in part by the National Natural Science Foundation of China under grant 62301484, in part by the Ningbo Natural Science Foundation of China under grant 2024J454, in part by the Aeronautical Science Foundation of China under grant 2024M071076001 and in part by the Zhejiang Provincial Natural Science Foundation of China under Grant No. LD24F030001.

References

[1] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, and et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 2, 3

[2] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 2, 3, 4

[3] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jürgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9297–9307, 2019. 2

[4] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Self-supervised open vocabulary occupancy estimation via volume rendering. *arXiv preprint arXiv:2407.17310*, 2024. 1

[5] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020. 5

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

[7] Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. Yolo-world: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16901–16911, 2024. 2

[8] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024. 2, 3, 8

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5, 6

[10] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17853–17862, 2023. 1

[11] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 2, 5, 6, 7

[12] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 4, 5, 6, 7

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

[14] Yuanhui Huang, Wenzhao Zheng, Borui Zhang, Jie Zhou, and Jiwen Lu. Selfocc: Self-supervised vision-based 3d occupancy prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19946–19956, 2024. 6, 7

[15] Bu Jin, Yupeng Zheng, Pengfei Li, Weize Li, Yuhang Zheng, Sujie Hu, Xinyu Liu, Jinwei Zhu, Zhijie Yan, Haiyang Sun, et al. Tod3cap: Towards 3d dense captioning in outdoor scenes. In *Proceedings of the European Conference on Computer Vision*, pages 367–384, 2025. 1

[16] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 3

[17] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 3, 8

[18] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2023. 1, 2, 3, 7

[19] Hongyang Li, Hao Zhang, Zhaoyang Zeng, Shilong Liu, Feng Li, Tianhe Ren, and Lei Zhang. Dfa3d: 3d deformable attention for 2d-to-3d feature lifting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6684–6693, 2023. 2

[20] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. Ovtrack: Open-vocabulary multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5567–5577, 2023. 2

[21] Wuyang Li, Zhu Yu, and Alexandre Alahi. Voxdet: Rethinking 3d semantic occupancy prediction as dense object detection. *arXiv preprint arXiv:2506.04623*, 2025. 2

[22] Yiming Li, Zhiding Yu, Christopher B. Choy, Chaowei Xiao, José M. Álvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for

camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9087–9098, 2023. 2

[23] 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 *Proceedings of the European Conference on Computer Vision*, pages 1–18, 2022. 2, 4, 5, 6, 7

[24] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[25] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[26] Qihang Ma, Xin Tan, Yanyun Qu, Lizhuang Ma, Zhizhong Zhang, and Yuan Xie. Cotr: Compact occupancy transformer for vision-based 3d occupancy prediction. *arXiv preprint arXiv:2312.01919*, 2023. 1, 2

[27] 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 *IEEE International Conference on Robotics and Automation*, pages 12404–12411, 2024. 1

[28] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 815–824, 2023. 2

[29] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Proceedings of the European Conference on Computer Vision*, pages 194–210, 2020. 2, 5

[30] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 3

[31] 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 *Proceedings of the International Conference on Machine Learning*, pages 8748–8763, 2021. 1, 2, 3

[32] 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. 8

[33] 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 *Advances in Neural Information Processing Systems*, pages 64318–64330, 2023. 1, 2, 5, 6, 8

[34] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8406–8415, 2023. 1, 2

[35] Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. In *Advances in Neural Information Processing Systems*, 2024. 1, 2, 3, 6, 7

[36] Tai Wang, Xiaohan Mao, Chenming Zhu, Runsen Xu, Ruiyuan Lyu, Peisen Li, Xiao Chen, Wenwei Zhang, Kai Chen, Tianfan Xue, et al. Embodiedscan: A holistic multimodal 3d perception suite towards embodied ai. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19757–19767, 2024. 1

[37] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17850–17859, 2023. 1, 2

[38] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022. 3

[39] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21729–21740, 2023. 1, 2

[40] Yuan Wu, Zhiqiang Yan, Zhengxue Wang, Xiang Li, Le Hui, and Jian Yang. Deep height decoupling for precise vision-based 3d occupancy prediction. *arXiv preprint arXiv:2409.07972*, 2024. 2

[41] Yuan Wu, Zhiqiang Yan, Yigong Zhang, Xiang Li, and Jian Yang. See through the dark: Learning illumination-affined representations for nighttime occupancy prediction. *arXiv preprint arXiv:2505.20641*, 2025. 2

[42] Huaiyuan Xu, Junliang Chen, Shiyu Meng, Yi Wang, and Lap-Pui Chau. A survey on occupancy perception for autonomous driving: The information fusion perspective. *arXiv preprint arXiv:2405.05173*, 2024. 1

[43] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2955–2966, 2023. 2, 3, 8

[44] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2945–2954, 2023. 2, 3, 7, 8

[45] Yan Yan, Yuxing Mao, and Bo Li. Second: Sparsely embedded convolutional detection. *Sensors*, 18(10):3337, 2018. 5

[46] Zhiqiang Yan, Yuankai Lin, Kun Wang, Yupeng Zheng, Yufei Wang, Zhenyu Zhang, Jun Li, and Jian Yang. Tri-perspective view decomposition for geometry-aware depth

completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4874–4884, 2024. 2

[47] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 2

[48] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10371–10381, 2024. 8

[49] Zhangchen Ye, Tao Jiang, Chenfeng Xu, Yiming Li, and Hang Zhao. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. *arXiv preprint arXiv:2409.13430*, 2024. 2, 6

[50] Zhu Yu, Zehua Sheng, Zili Zhou, Lun Luo, Si-Yuan Cao, Hong Gu, Huaqi Zhang, and Hui-Liang Shen. Aggregating feature point cloud for depth completion. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8732–8743, 2023. 2

[51] 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. 2

[52] Zhu Yu, Runmin Zhang, Jiacheng Ying, Junchen Yu, Xiaohai Hu, Lun Luo, Si-Yuan Cao, and Hui-Liang Shen. Context and geometry aware voxel transformer for semantic scene completion. In *Advances in Neural Information Processing Systems*, 2024. 2

[53] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14393–14402, 2021. 2

[54] Chubin Zhang, Juncheng Yan, Yi Wei, Jiaxin Li, Li Liu, Yansong Tang, Yueqi Duan, and Jiwen Lu. Occnerf: Self-supervised multi-camera occupancy prediction with neural radiance fields. *arXiv preprint arXiv:2312.09243*, 2023. 6, 7

[55] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9433–9443, 2023. 2, 6

[56] Jilai Zheng, Pin Tang, Zhongdao Wang, Guoqing Wang, Xiangxuan Ren, Bailan Feng, and Chao Ma. Veon: Vocabulary-enhanced occupancy prediction. *arXiv preprint arXiv:2407.12294*, 2024. 1, 2, 3, 4, 6, 7, 8

[57] Yupeng Zheng, Xiang Li, Pengfei Li, Yuhang Zheng, Bu Jin, Chengliang Zhong, Xiaoxiao Long, Hao Zhao, and Qichao Zhang. Monoocc: Digging into monocular semantic occupancy prediction. *arXiv preprint arXiv:2403.08766*, 2024. 2