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.

AGO: Adaptive Grounding for Open World 3D Occupancy Prediction

Peizheng Li[1,2], Shuxiao Ding[1,4], You Zhou[1], Qingwen Zhang[5], Onat Inak[1,6],
Larissa Triess[1], Niklas Hanselmann[1,2,3], Marius Cordts[1], Andreas Zell[2]
[1]Mercedes-Benz AG, Sindelfingen, [2]University of Tübingen, [3]Tübingen AI Center,
[4]University of Bonn, [5]RPL, KTH Royal Institute of Technology, [6]TU Berlin
{firstname.lastname}@mercedes-benz.com, qingwen@kth.se, andreas.zell@uni-tuebingen.de

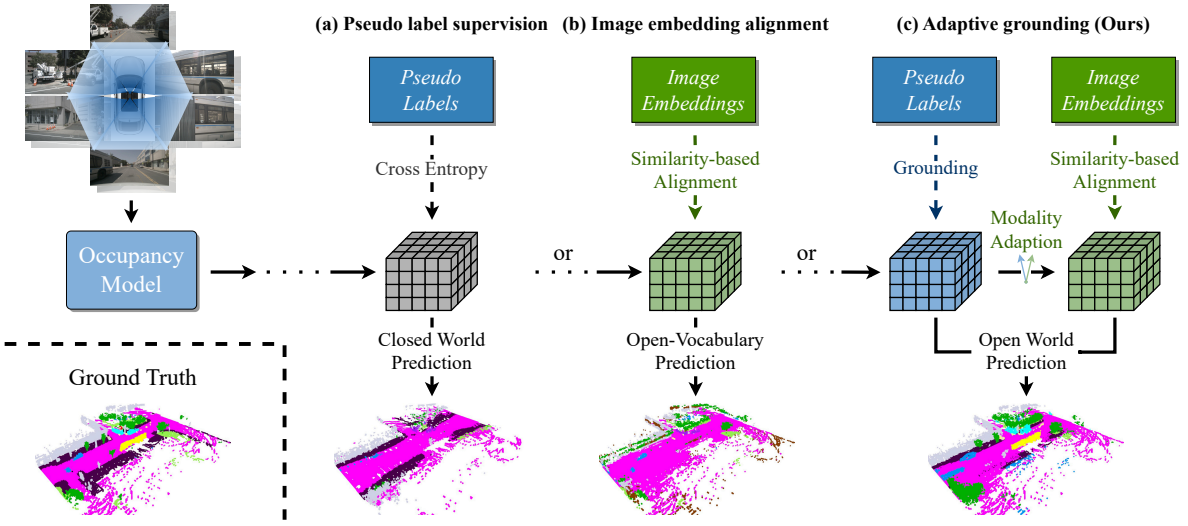


Figure 1. **Open-world 3D semantic occupancy prediction.** (a) Supervision based on pseudo labels with fixed classes cannot predict novel categories, such as the "construction vehicle" and "vegetation". (b) Similarity-based alignment suffers from significant mismatches due to issues like modality discrepancy, leading to confusions between *e.g.* "sidewalk" and "driveable surface". (c) Our proposed Adaptive Grounding flexibly accommodates both known and unknown objects, achieving more precise open world occupancy prediction.

Abstract

Open-world 3D semantic occupancy prediction aims to generate a voxelized 3D representation from sensor inputs while recognizing both known and unknown objects. Transferring open-vocabulary knowledge from vision-language models (VLMs) offers a promising direction but remains challenging. However, methods based on VLM-derived 2D pseudo-labels with traditional supervision are limited by a predefined label space and lack general prediction capabilities. Direct alignment with pretrained image embeddings, on the other hand, often fails to achieve reliable performance because of inconsistent image and text representations in VLMs. To address these challenges, we propose AGO, a novel 3D occupancy prediction framework with adaptive grounding to handle diverse open-world scenarios. AGO first encodes surrounding images and class prompts into 3D and text embeddings, respectively, leveraging similarity-based grounding training with 3D pseudo-labels. Additionally, a modality adapter maps 3D embeddings into a space aligned with VLM-derived image embeddings, reducing modality gaps. Experiments on Occ3D-nuScenes show that AGO improves unknown object prediction in zero-shot and few-shot transfer while achieving state-of-the-art closed-world self-supervised performance, surpassing prior methods by 4.09 mIoU. Code is available at: https://github.com/EdwardLeeLPZ/AGO.

1. Introduction

Robust 3D scene representations are essential to autonomous driving systems, as they inherently encode rich geometric and semantic details. One effective way for achieving such representations is 3D semantic occupancy prediction, which constructs voxelized structures with semantic labels from sensor inputs. Although supervised ap-

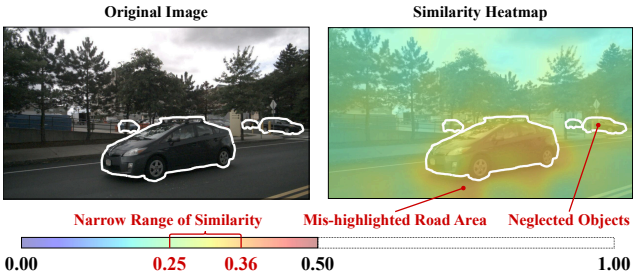


Figure 2. **Modality gaps between image and text embeddings in pre-trained VLMs.** The similarity heatmap generated using the image and prompt "car" only covers a small range of about 0.1 (with a peak of only 0.36), while the low resolution of the embeddings also results in spatial misalignment and semantic ambiguity.

proaches [4, 18, 27, 40, 43, 47, 55, 56] deliver strong performance, their reliance on extensive 3D annotations within closed label spaces restricts both training scalability and generalization to unlabeled unknown objects.

Recent advances incorporate open-vocabulary knowledge through vision-language models (VLMs) [52] to address these limitations. As shown in Fig 1.a and 1.b, several works generate 2D semantic pseudo-labels for 3D self-supervised training [15, 19, 51, 57], while others directly align 3D features with VLM-derived 2D embeddings via knowledge distillation [3, 42, 45]. However, most VLMs are pre-trained on 2D vision-text data and lack explicit 3D geometric reasoning, which poses challenges in transferring their knowledge effectively. Additionally, limited closed label spaces and inherent modality gaps (Figure 2) can impair both semantic fidelity and geometric completeness.

To address these issues, we propose **A**daptive **G**rounding for 3D **O**ccupancy Prediction (**AGO**) illustrated in Figure 1.c. Our method constructs voxel representations for both closed-world and open-world objects and applies targeted alignment strategies to bridge the gap between the text and 2D image embeddings of the VLM. During training, VLM-generated pseudo-labels enable grounding between 3D voxel and text embeddings, enhanced by negative sampling of additional noise text prompts. A modality adapter further maps 3D embeddings into a space aligned with VLM-derived image embeddings, facilitating knowledge transfer beyond predefined labels. During inference, an open-world identifier selects the most reliable embeddings based on information entropy for the final occupancy prediction.

Experimental results show that AGO outperforms the state-of-the-art self-supervised methods on the Occ3D dataset [43]. To further assess its open-world effectiveness, we introduce a benchmark consisting of 3 stages: pre-training on a limited label space, zero-shot evaluation on expanded classes, and few-shot fine-tuning, all without manual annotations. Our proposed AGO maintains strong performance on known objects while achieving a clear advantage in predicting unknown objects, outperforming existing closed-world self-supervised methods.

The contributions of this paper are as follows:

- We introduce AGO, a novel open-world occupancy prediction framework that adaptively distills knowledge from pretrained VLMs into 3D perception for autonomous driving.
- We introduce grounding training with random text noise as negative samples to enhance the discriminative power of the 3D representation.
- We design a modality adapter to mitigate the differences between text and 2D image embeddings and an open world identifier to adaptively select features for known and unknown objects during inference in the open world.
- Our method achieves state-of-the-art performance on the closed-world self-supervised Occ3D-nuScenes benchmark and significantly outperforms existing methods in open-world occupancy prediction.

2. Related Work

3D scene understanding To accurately capture 3D environmental information, a common approach is to encode LiDAR point clouds into a global representation for various downstream tasks [9, 11, 12, 23, 39, 48, 50, 53, 54, 60]. Recently, vision-centric approaches have gained increasing attention, extracting image features from perspective views and lifting them into a BEV [17, 21, 26, 28, 35, 58] or 3D voxel representation [4, 6, 18, 27, 29, 34, 44, 46, 47, 49, 55, 56]. Among them, MonoScene [6] first attempts to extract monocular 2D image features into 3D voxels using a line-of-sight projection, paving the way for future works. Inspired by BEV-based models [21, 28, 58], TPVFormer [18] introduced a tri-perspective view representation, complementing BEV with two additional perpendicular planes for richer spatial encoding. Subsequent methods refine prediction accuracy by improving feature representation. OccFormer [55] enhances long-range and fine-grained feature interactions through global-local path decoupling, while VoxFormer [27] improves spatial consistency with a coarse-to-fine strategy. However, these approaches heavily rely on large-scale 3D annotations. The high cost of manual labeling restricts their ability to adapt and scale to various, constantly evolving open-world driving situations.

Multi-modal representation learning Multi-modal representation learning has expanded the scope of visual tasks. Building on early works [8, 24, 33, 41], CLIP [36] introduced a dual-stream architecture with contrastive learning to align textual and visual modalities. Trained on 400 million image-text pairs, CLIP demonstrated strong zero-shot capabilities and inspired further research into large-scale cross-modal pre-training. As visual tasks became more complex, GLIP [25] unified object detection and phrase grounding by aligning visual regions with text de-

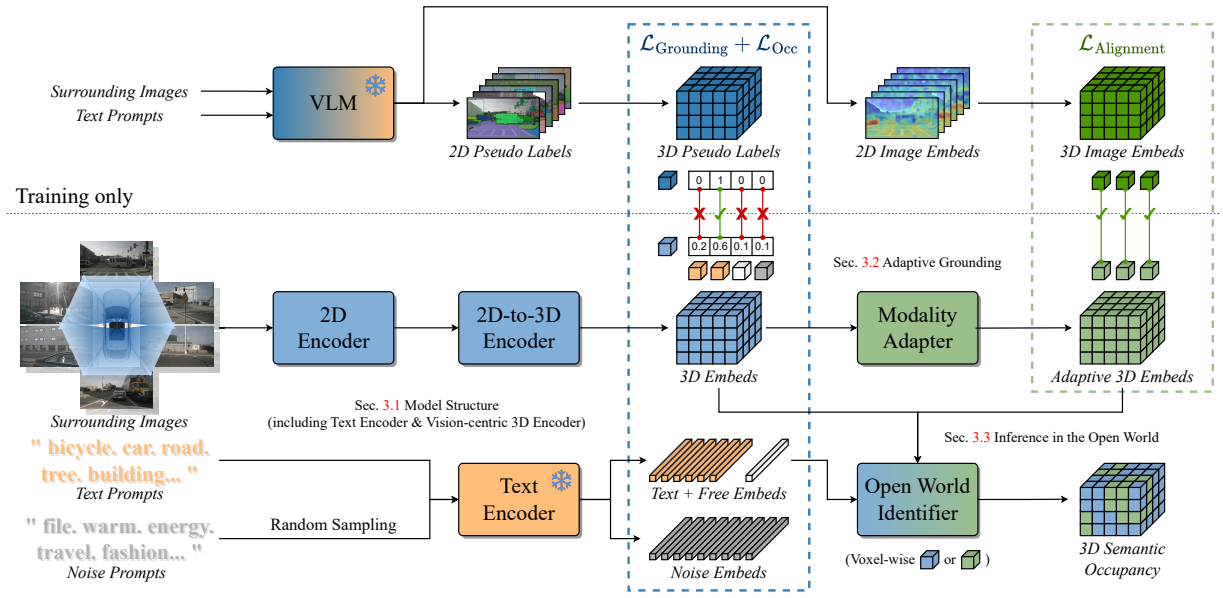


Figure 3. **AGO architecture.** The upper part illustrates the generation process of 3D pseudo-labels and image embeddings based on pre-trained VLMs during training. The lower part depicts the main architecture of our AGO framework, which comprises a frozen pre-trained text encoder, a vision-centric 3D encoder, a modality adapter, and an open-world identifier. The detailed illustration in the middle showcases our training paradigm, which consists of a noise-augmented grounding training and an adaptive image embedding alignment.

scriptors. Grounding DINO [30] extended this approach with a transformer-based deformable attention mechanism. MaskCLIP+ [59] leveraged pre-trained CLIP features for pixel-level tasks, achieving zero-shot semantic segmentation through pseudo-labeling. More recently, Grounded SAM [38] combined SAM [22] with Grounding DINO to enable fine-grained segmentation using arbitrary text inputs, showing strong potential for open-world applications. Despite their advances, these models are not designed for 3D reasoning and require further adaptation for autonomous driving.

Open-vocabulary occupancy prediction Recently, many works have explored integrating multi-modal representations into 3D scene understanding to extend knowledge boundaries with their open-vocabulary capabilities. One common paradigm [1, 7, 15, 19, 31, 51, 57] is using 2D pseudo-labels generated by pre-training multimodal models as supervision. OccNeRF [51] constructs unbounded parameterized 3D occupancy fields and then renders these fields into depth maps, supervised through multi-frame photometric consistency. SelfOcc [19] follows the same paradigm and proposes a MVS-embedded strategy to improve depth optimization. But the above methods are limited by the closed label space and do not have the ability of open-world prediction. VEON [57] further improves scene understanding by integrating pre-trained depth estimation [37] and vision-language model (VLM) [36] through delicate structure design and two-stage training. Alternatively, some methods [3, 20, 45] directly align 3D or reconstructed 2D features with multi-modal pre-trained embeddings. POP-3D [45] performs pixel-level alignment of 3D features and pre-trained CLIP [36] image embeddings based on the LiDAR point cloud. GaussTR [20], on the other hand, constructs a sparse 3D Gaussian representation and splats it in the image plane for 2D alignment, reducing the dependence on LiDAR. However, these methods achieved suboptimal performance due to the modality gaps within pre-trained models, low resolution, and incomplete scenes. In summary, developing a 3D occupancy prediction framework that balances strong performance with open-world generalization remains an urgent challenge.

3. Method

3D semantic occupancy prediction aims to convert sensor inputs into a voxelized semantic representation of the environment. The vision-centric model takes N_{cam} surrounding camera images $\mathbf{I} = \{\mathbf{I}_n\}_{n=1}^{N_{\text{cam}}}$ from a single timestamp as inputs and outputs semantic voxels $\mathbf{O} \in \mathbb{C}^{H \times W \times D}$, where H, W and D represent the spatial resolution of the 3D voxel space, and $\mathbb{C}$ is the label space containing N_{cls} classes.

As illustrated in Figure 3, our proposed AGO consists of four modules. The *text encoder* converts label and noise prompts into text embeddings. The *vision-centric 3D encoder* extracts 2D features from the input images and lifts them into 3D voxel embeddings. The *modality adapter* projects these embeddings into a semantic space aligned with pretrained VLM image embeddings. Finally, the *open world identifier* selects the most appropriate features for occupancy prediction in the open-world scenarios.

3.1. Model Structure

Text encoder To extend a fixed, closed label space into a flexible, open semantic space, we take the labels as textual inputs and employ a text encoder to extract their semantic information. Specifically, unlike the full sentences used in traditional 2D visual grounding, we decompose the given category labels into predefined finer-grained subcategory phrases serving text prompts. These prompts are then tokenized and further encoded into text embeddings by a pre-trained language model, *e.g*. BERT [10] in MaskCLIP+ [59]. Besides, we randomly add noise prompts by selecting phrases from a general dictionary. The resulting noise embeddings serve as negative samples during grounding training, thereby enhancing the discrimination capability of the model for closed-world objects. The process described above can be summarized as:

$$\mathbf{F}_T = \mathcal{E}_T(\mathbf{T}), \quad \mathbf{F}_{\text{noise}} = \mathcal{E}_T(\mathbf{T}_{\text{noise}}), \tag{1}$$

where $\mathbf{F}_T \in \mathbb{R}^{N_{\text{text}} \times C}$ and $\mathbf{F}_{\text{noise}} \in \mathbb{R}^{N_{\text{noise}} \times C}$. N_{text} and N_{noise} are the number of tokens of text prompts and noise prompts, which are roughly the same. C is the dimensionality of text embeddings. We freeze the text encoder throughout the training process, thus preserving its open-vocabulary performance. Compared to traditional supervised learning, a clear advantage is that the label space does not have to remain consistent between training and inference.

Vision-centric 3D encoder Our framework follows the architecture of TPVFormer [18] to construct an efficient 3D representation. First, the surrounding images are fed into an image backbone $\mathcal{E}_I$ to obtain 2D perspective features $\mathbf{F}_I^{\text{2D}} = \mathcal{E}_I(\mathbf{I})$. Next, a 3D encoder lifts these $\mathbf{F}_I^{\text{2D}}$ to mutually orthogonal tri-perspective views $\mathbf{F}_I^{TPV} = [\mathbf{F}_I^{HW}, \mathbf{F}_I^{DH}, \mathbf{F}_I^{WD}] = \mathcal{E}_{\text{3D}}(\mathbf{F}_I^{\text{2D}})$. Finally, they are expanded and superimposed to form the full 3D embeddings $\mathbf{F}_I^{\text{3D}} \in \mathbb{R}^{H \times W \times D \times C}$, where C denotes their feature dimensionality, identical to above text embeddings. This process can be formulated as:

$$\mathbf{F}_I^{\text{3D}} = \mathcal{E}_{\text{3D}}(\mathcal{E}_I(\mathbf{I})). \tag{2}$$

3.2. Adaptive Grounding

To achieve precise closed-world occupancy prediction while also enabling open-world transfer, we propose adaptive grounding that decouples training for known and unknown categories through grounding training and adaptive alignment, respectively.

Closed-world grounding training For known categories, we construct 3D voxel pseudo-labels from 2D semantic masks generated by a pre-trained VLM [38]. Specifically, we define the label space according to common autonomous driving scenarios [3, 45, 57] and use the label phrases as text prompts to generate corresponding 2D semantic masks of the surrounding images. These masks are then transformed into 3D occupancy pseudo-labels $\mathbf{O}_{\text{pl}} \in \mathbb{C}^{H \times W \times D}$ via the geometric correspondence between LiDAR point clouds and camera pixels. By incoporating the post-processing of multi-frame aggregation, point cloud ray casting, and semantic voting, this process produces a more comprehensive 3D semantic scene than what single-frame 2D supervision can achieve, thereby enhancing closed-world self-supervised performance.

Traditional classifiers trained on pseudo-labels can only distinguish known categories and require retraining to recognize unlabeled objects. Instead, we adopt grounding training by directly comparing the 3D voxel features $\mathbf{F}_I^{\text{3D}}$ with text embeddings $\mathbf{F}_T$ and noise embeddings $\mathbf{F}_{\text{noise}}$. In order to represent free (unoccupied) voxels, we also introduce a learnable embedding $\mathbf{F}_{\text{free}}$ with the same dimension C as $\mathbf{F}_T$. These embeddings are concatenated into $\mathbf{F}_{T+\text{noise}+\text{free}} = [\mathbf{F}_T; \mathbf{F}_{\text{noise}}; \mathbf{F}_{\text{free}}]$. We then compute the voxel-wise similarity score by taking the dot product:

$$\mathbf{S}^{\text{3D}} = \mathbf{F}_I^{\text{3D}} \cdot \mathbf{F}_{T+\text{noise}+\text{free}}^{\top}. \tag{3}$$

The resulting $\mathbf{S}^{\text{3D}} \in \mathbb{R}^{H \times W \times D \times (N_{\text{token}} + N_{\text{noise}} + 1)}$ acts as generalized logits, allowing us to apply standard occupancy prediction losses used in traditional supervised training.

We follow TPVFormer [18], using a combination of the cross-entropy loss and the Lovász-softmax loss [2] for grounding training:

$$\begin{aligned} \mathcal{L}_{\text{Grounding}} &= \mathcal{L}_{\text{CE}}(\mathbf{S}_{\text{sem}}^{\text{3D}}, \mathbf{O}_{\text{pl}}) + \mathcal{L}_{\text{Lovász}}(\mathbf{S}_{\text{sem}}^{\text{3D}}, \mathbf{O}_{\text{pl}}), \text{ with} \\ \mathbf{S}_{\text{sem},k}^{\text{3D}} &= \frac{1}{|\mathcal{T}_k|} \sum_{i \in \mathcal{T}_k} \mathbf{S}_i^{\text{3D}}, \end{aligned} \tag{4}$$

where $\mathcal{T}_k$ denotes the set of token indices for category k and $|\mathcal{T}_k|$ is the number of tokens in the set. We average the scores $\{\mathbf{S}_i^{\text{3D}}\}$ from tokens belonging to the same category to form the semantic logits for both text and noise inputs.

Furthermore, we incorporate an occupancy loss to balance the occupied and free space:

$$\begin{aligned} \mathcal{L}_{\text{Occ}} &= \mathcal{L}_{\text{CE}}(\mathbf{S}_{\text{binary}}^{\text{3D}}, \mathbf{O}_{\text{binary}}), \text{ with} \\ \mathbf{S}_{\text{binary}}^{\text{3D}} &= [\max(\mathbf{S}_{\text{sem}}^{\text{3D}} \setminus \mathbf{S}_{\text{sem, noise+free}}^{\text{3D}}); \ \mathbf{S}_{\text{sem, free}}^{\text{3D}}], \\ \mathbf{O}_{\text{binary}} &= \begin{cases} 0 & \text{if } \mathbf{O}_{\text{pl}} = \text{free} \\ 1 & \text{otherwise} \end{cases} \end{aligned} \tag{5}$$

Open-world adaptive alignment While grounding is an effective closed-world training paradigm, it cannot recognize unknown objects beyond fixed pseudo-labels. In contrast, VLMs benefit from large-scale pretraining and produce image embeddings with richer semantics that implicitly represent unlabeled objects in autonomous driving scenes. By using similarity-based alignment, these embeddings can be distilled into 3D representations, expanding their semantic

scope. However, as shown in Section 1 and Figure 2, alignment alone yields limited prediction performance. A more promising approach is to combine similarity-based alignment with the grounding paradigm described above.

One straightforward implementation is to apply both loss functions simultaneously to the same 3D embedding. However, since the two alignment objectives target different modalities—text and vision—naively adding their losses can lead to modality conflicts during training, ultimately degrading performance. This is also supported by the experimental comparison in Table 3.

To address these issues, we propose an adaptive alignment mechanism that assists grounding training to solve the unknown object prediction. We first map the 3D embeddings $\mathbf{F}_I^{\text{3D}}$ into a new semantic space as $\tilde{\mathbf{F}}_I^{\text{3D}} \in \mathbb{R}^{H\times W\times D\times C}$. They are then aligned with the image embeddings of a pre-trained VLM, *e.g.* MaskCLIP+ [59], distilling the semantic-rich information contained in the image embeddings into 3D. Specifically, the 2D image embeddings $\mathbf{E}_I^{\text{2D}} \in \mathbb{R}^{H_I\times W_I\times C}$ are projected into the 3D voxels $\mathbf{E}_I^{\text{3D}} \in \mathbb{R}^{H\times W\times D\times C}$ following the same geometric projection used for pseudo-label generation. Here, we employ a lightweight MLP as the modality adapter ($\mathcal{MA}$):

$$\tilde{\mathbf{F}}_I^{\text{3D}} = \mathcal{MA}(\mathbf{F}_I^{\text{3D}}). \tag{6}$$

For the alignment loss, we adopt cosine similarity. It computes the loss on non-empty voxels visible in the camera views, thereby avoiding mismatches due to occlusion.

$$\mathcal{L}_{\text{Alignment}} = \frac{1}{|\mathcal{V}_{\text{visible}}|} \sum_{v\in\mathcal{V}_{\text{visible}}} \left(1 - \cos(\tilde{\mathbf{F}}_{I,v}^{\text{3D}}, \mathbf{E}_{I,v}^{\text{3D}})\right), \tag{7}$$

where $\mathcal{V}_{\text{visible}}$ denotes the set of visible non-empty voxels. Finally, the overall loss function is formulated as:

$$\mathcal{L}_{\text{Total}} = \mathcal{L}_{\text{Grounding}} + \mathcal{L}_{\text{Occ}} + \mathcal{L}_{\text{Alignment}}. \tag{8}$$

3.3. Inference in the Open World

We propose an open world identifier for balanced decision-making between grounding and adaptive alignment in occupancy prediction. For each voxel $v \in \mathcal{V}$, it compares the similarity obtained from both 3D embedding and adaptive 3D embedding to determine which one to use for the final prediction. Specifically, we compute the dot products ($\mathbf{S}_{\text{sem},v}^{\text{3D}}$ and $\tilde{\mathbf{S}}_{\text{sem},v}^{\text{3D}}$) between query text embeddings and the 3D embeddings as well as adaptive 3D embeddings according to the formula Eqs. (3, 4). They are then passed through the softmax function and become the confidence scores ($\mathbf{P}_v$ and $\tilde{\mathbf{P}}_v$) for the voxel belonging to a particular category.

Based on statistical analysis in Figure 6 and experimental verification in Table 4, we use minimum information entropy as the criterion: By comparing the entropy of $\mathbf{P}_v$ and $\tilde{\mathbf{P}}_v$, we select the probability distribution with lower entropy:

$$\begin{aligned} \mathbf{P}_{\text{final},v} &= \mathbf{c}_v\mathbf{P}_v + (1-\mathbf{c}_v)\tilde{\mathbf{P}}_v, \text{ with} \\ \mathbf{c}_v &= \mathbb{I}(\mathcal{H}(\mathbf{P}_v) \le \mathcal{H}(\tilde{\mathbf{P}}_v)) \end{aligned} \tag{9}$$

where $\mathbb{I}(\cdot)$ is the indicator function, $\mathbf{c}_v$ represents the criterion for voxel v and $\mathcal{H}(\cdot)$ denotes the entropy.

4. Experiments

4.1. Experimental Settings

Dataset and metrics Our experiments are conducted on the Occ3D-nuScenes benchmark [43], which is derived from the nuScenes dataset [5] that consists of 700/150/150 scenes for train/val/test splits, totaling 40,000 frames. This benchmark uses nuScenes sensor data and each frame includes six-camera images and a 32-beam LiDAR point cloud. Occ3D-nuScenes provides voxelized 3D semantic occupancy labels with a spatial range of [-40m, 40m]×[-40m, 40m]×[-1m, 5.4m] and a resolution of 0.4m×0.4m×0.4m. We evaluate performance using per-category IoU and the overall mIoU.

Task settings We evaluate baselines and our method under two settings. The ***self-supervised task*** follows the Occ3D-nuScenes benchmark [43] but only uses pseudo-labels during training. In the ***open-world task***, we simulate real-world conditions by gradually expanding the label space. As shown in Table 2, the model is first pretrained with pseudo-labels with only five major categories (the first five columns in green and blue), with the last two being supercategories composed of multiple nuScenes classes. It is then evaluated in zero-shot settings on the full Occ3D label space. By this time, the supercategories (blue) are split into their original classes (orange) and the remaining classes are added as fully unknown categories (red). Afterwards, the model is fine-tuned based on a small number of samples to validate few-shot generalization ability of unknown classes. Further details are provided in the supplementary material.

Implementation details We use TPVFormer [18] with ResNet-101 [16] image backbone as the vision-centric 3D encoder. It produces a voxel grid with the shape 200×200×16. Moreover, we adopt BERT [10] with MaskCLIP+ [59] pre-trained weights as the text encoder. For higher resolution, we utilize the MaskCLIP+ [59] image features upsampled by FeatUp [13] as 2D image embeddings. For the modal adapters in AGO, we use a 2-layer MLP with a softplus activation. Given its superior mask quality, we employ Grounded SAM [38] to generate 2D pseudo-labels. During training, we utilize the AdamW [32] optimizer with a learning rate of 10^{-3} as well as the cosine learning rate scheduler, which starts with a linear warm-up from 10^{-5} for 500 steps and finishes at 10^{-6}. For self-supervised training and open-world pre-training tasks, we train our model for 24

Method	Image Backbone	oth.	bar.	bic.	bus	car	c. v.	mot.	ped.	t. c.	tra.	tru.	d. s.	o. f.	sid.	ter.	man.	veg.	mIoU*	mIoU
SimpleOcc [14]	ResNet-101	0.00	0.67	1.18	3.21	7.63	1.02	0.26	1.80	0.26	1.07	2.81	40.44	0.00	18.30	17.01	13.42	10.84	7.99	7.05
POP-3D† [45]	ResNet-101	0.06	0.02	0.46	1.83	4.87	0.00	0.00	1.29	0.00	0.65	2.62	**55.90**	1.60	9.99	25.17	15.75	21.11	9.42	8.31
SelfOcc [19]	ResNet-50	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	10.54	9.30
OccNeRF [51]	ResNet-101	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	10.81	9.53
GaussianOcc [15]	Swin	0.00	1.79	5.82	**14.58**	13.55	1.30	2.82	**7.95**	**9.76**	0.56	9.61	44.59	0.00	20.10	17.58	8.61	10.29	11.26	9.94
GaussTR [20]	VFMs	0.00	2.09	5.22	14.07	**20.34**	**5.70**	7.08	5.12	3.93	0.92	**13.36**	39.44	0.00	15.68	22.89	21.17	21.87	13.26	11.70
LangOcc [3]	ResNet-50	0.00	3.10	**9.00**	6.30	14.20	0.40	**10.80**	6.20	**9.00**	3.80	10.70	43.70	**2.23**	9.50	26.40	19.60	**26.40**	13.27	11.84
VEON [57]	ViT-L	**0.90**	**10.40**	6.20	**17.70**	12.70	**8.50**	7.60	6.50	5.50	**8.20**	11.80	54.50	0.40	25.50	**30.20**	**25.40**	25.40	**17.07**	**15.14**
AGO (ours)	ResNet-101	**1.53**	**6.75**	**6.43**	14.00	**22.82**	5.57	**16.66**	**13.20**	6.80	**10.53**	**15.89**	**71.48**	**4.48**	**34.48**	**41.37**	**29.33**	**25.66**	**21.39**	**19.23**

Table 1. **3D occupancy prediction performance under the self-supervised setting on the Occ3D-nuScenes [43] dataset.** The full names of all abbreviated categories can be found by color in Figure 4. "VFMs" stands for the ensemble of multiple vision foundation models. † indicate values obtained from our retraining. We also calculate the result as "mIoU*" ignoring the "other" and "other flat" categories, while "mIoU" is the original result. Results are highlighted in **bold & underlined** for the best performance and **bold** for the second-best performance.

Training Stages	Method	ped.	d. s.	sid.	veh.	cyc.	k. mIoU	car	bus	c. v.	tra.	tru.	bic.	mot.	bar.	t. c.	ter.	man.	veg.	u. mIoU	mIoU
Pretraining	POP-3D† [45]	0.00	58.77	13.80	8.27	1.10	16.39	-	-	-	-	-	-	-	0.00	0.00	3.95	**0.13**	0.60	0.94	8.66
	SelfOcc† [19]	0.98	60.29	14.68	7.11	0.00	16.61	-	-	-	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	8.31
	GaussTR† [20]	6.11	60.06	18.02	6.77	2.25	18.64	-	-	-	-	-	-	-	0.00	0.00	4.95	0.07	8.05	2.61	10.63
	AGO (ours)	**7.82**	**63.09**	**25.53**	**9.19**	**5.03**	**22.13**	-	-	-	-	-	-	-	0.00	0.00	**7.04**	0.03	**10.88**	**3.59**	**12.86**
Zero-shot Evaluation	POP-3D† [45]	0.00	58.77	13.80	-	-	24.19	6.72	0.00	0.00	0.59	4.34	1.17	**1.20**	0.00	0.00	3.95	0.13	0.60	1.56	6.08
	SelfOcc† [19]	0.98	60.29	14.68	-	-	25.23	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	5.06
	GaussTR† [20]	6.11	60.06	18.02	-	-	28.06	5.07	**1.65**	0.00	0.04	1.84	2.58	0.27	0.00	0.00	4.95	0.07	8.05	2.04	7.25
	AGO (ours)	**7.82**	**63.09**	**25.53**	-	-	**32.15**	**7.67**	0.00	0.00	**1.33**	**6.50**	**4.50**	0.00	0.00	0.00	**7.04**	0.03	**10.88**	**3.16**	**8.96**
Few-shot Finetuning	POP-3D† [45]	0.00	44.90	12.79	-	-	19.23	5.59	0.03	0.00	0.29	2.05	1.26	1.03	0.00	0.00	5.72	0.21	6.75	1.91	5.37
	SelfOcc† [19]	7.85	65.65	25.29	-	-	32.93	1.41	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	3.63	6.04	10.96	1.84	8.06
	GaussTR† [20]	7.84	66.36	25.55	-	-	33.25	10.85	1.58	0.00	0.00	1.32	1.42	0.00	0.00	0.00	12.74	9.12	8.16	3.77	9.66
	AGO (ours)	**13.00**	**71.54**	**29.91**	-	-	**38.15**	**18.73**	**5.49**	0.00	**0.41**	**2.16**	**3.72**	**2.22**	**0.43**	0.00	**29.63**	**21.43**	**17.73**	**8.50**	**14.43**

Table 2. **3D occupancy prediction performance under the open-world setting on the Occ3D-nuScenes [43] dataset.** † indicate values obtained from our retraining. The background color represents whether the category is known or unknown during the **pre-training stage**: green and blue indicate known IoU, orange and red indicate unknown IoU. The orange categories are the refined version of the blue categories. Results are highlighted in **bold** for the best performance.

epochs following the Occ3D [43] setting, while for open-world fine-tuning tasks, we train our model for 12 epochs. All experiments are conducted on 8 × NVIDIA A100 GPUs with a batch size of 1 per GPU. Further details can be found in the supplementary material.

4.2. Results and Analysis

Closed-world self-supervised benchmark Table 1 shows the prediction performance of AGO in the closed-world self-supervised task. Our method demonstrates notable performance improvement, surpassing the previous best approach [57] by 4.09 mIoU. Notably, AGO not only obtains significant gains in many static classes (*e.g.* +15.58 in "driveable surface", +8.18 in "sidewalk", +11.17 in "terrain"), but also achieves the state of the art in dynamic categories (*e.g.* +2.48 in "car", +5.86 in "motorcycle", +5.25 in "pedestrian", +2.53 in "truck"). Unlike methods that rely on volume rendering or Gaussian splatting for 3D-to-2D projections, AGO directly optimizes 3D voxel features without requiring specific voxel settings [51], additional learnable modules [3], or complex consistency losses [15, 19, 20]. Some approaches also directly integrate large pretrained foundational models [20, 57], leading to high computational costs (*e.g.*, VEON [57] with 678.1M parameters). In contrast, AGO achieves the state-of-the-art performance without relying on delicate training setting or large-scale models (requiring only 62.5M parameters). Additionally, we present the qualitative results. As shown in Figure 4, our method can produce more accurate predictions for dynamic objects and long-tailed classes. These results highlight the robustness and effectiveness of AGO in learning accurate 3D semantic occupancy representations in a closed-world setting.

Open-world evaluation and analysis We evaluate the generalization capability of AGO in the open-world task, as shown in Table 2. To compare the open-world prediction performance of different training paradigms, we selected POP-3D [45], SelfOcc [19] and GaussTR [20] as baselines. POP-3D and GaussTR represent image embedding alignment methods, while SelfOcc relies on predefined closed-set pseudo-labels for supervision. During pretraining (first four rows in Table 2), AGO maintains strong performance on known classes (22.13 mIoU) while also achieving non-trivial prediction for unknown categories such as 7.04 IoU for "terrain" and 10.88 IoU for "vegetation". SelfOcc, restricted by its fixed label space, completely fails to predict

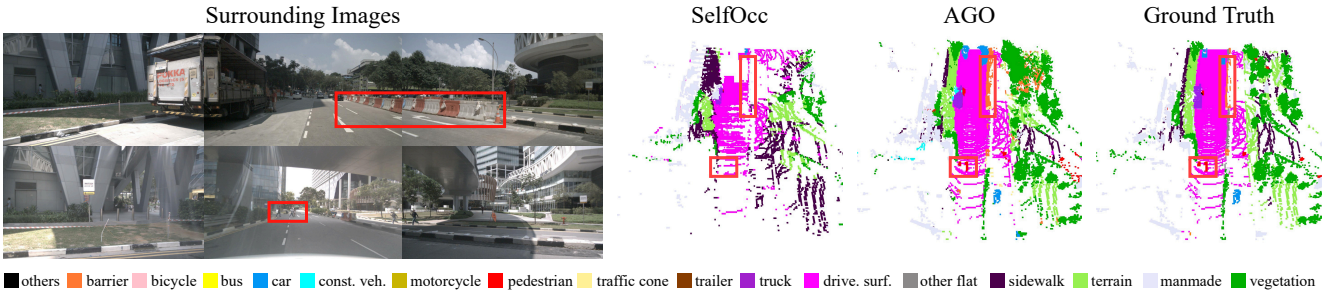


Figure 4. **Visualization of self-supervised 3D semantic occupancy prediction on the Occ3D-nuScenes occupancy benchmark.** Our method demonstrates more detailed predictions for dynamic ("pedestrian") and long-tailed ("barrier") objects.

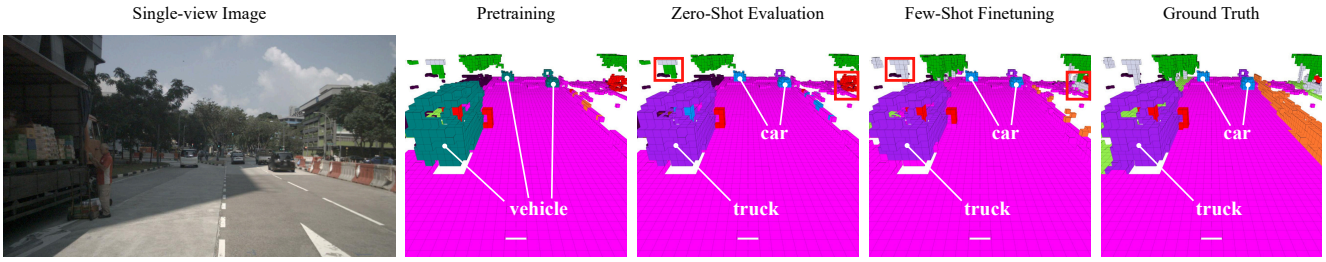


Figure 5. **Visualization of Open-world Zero-&Few-Shot Transfer.** Our method can not only adapt to category changes from coarse to fine but also easily accommodate newly introduced, previously unknown categories with only a small amount of few-shot fine-tuning.

unknown classes, yielding 0 mIoU across all of them. POP-3D, despite leveraging a pre-trained VLM, struggles due to the lack of geometric and semantic cues in its pseudo-labels, reaching only 0.94 mIoU on average for unknown categories. Thanks to the multiple integrated foundation models, GaussTR obtains better results but still performs lower than the 12.86 mIoU achieved by AGO.

When evaluated in a zero-shot setting with splitting supercategories into original classes, AGO outperforms POP-3D with IoU improvements of +0.95 ("car"), +2.16 ("truck"), and +3.33 ("bicycle"). Similarly, GaussTR also struggles in these categories. Notably, the IoU scores of "truck" (6.50) and "bicycle" (4.50) even surpass those of several methods in the closed-world benchmark Table 1. Although the closed-world SelfOcc may capture semantic similarity, they are not trained on the new labels of supercategories and thus cannot handle the coarse-to-fine category transition.

Few-shot transfer aims to expand the knowledge of the model with minimal additional labeling. We fine-tune all methods using a small amount of supercategories data. POP-3D, relying only on alignment, shows little improvement and overfits image embeddings, leading to a 4.95 mIoU drop in known classes (24.19 mIoU to 19.23 mIoU). SelfOcc also struggles in the open-world task, achieving only a +1.84 mIoU gain for unknown categories. While GaussTR benefits from fine-tuning, its alignment-based training limits improvement to just 2.41 mIoU. Our approach, in contrast, achieves consistent improvements in both known and unknown categories, with gains of +6.00 and +5.34 in mIoU, respectively, demonstrating its superior few-shot transfer capability in the open-world task.

Training Paradigm	mIoU Self.	O.W. Pre.	O.W. Z.S.	O.W. F.S.
Align	10.28	15.4 / 0.8 / 8.1	23.5 / 1.2 / 5.6	24.4 / 4.0 / 8.1
Gro.	19.08	20.6 / 0.0 / 10.3	**33.0** / 0.3 / 6.8	**38.4** / 3.0 / 10.1
Gro. + Align	18.89	18.3 / 2.2 / 10.2	29.3 / 1.4 / 7.0	37.3 / 5.7 / 12.0
AGO	**19.23**	**22.1 / 3.6 / 12.9**	32.2 / **3.2 / 9.0**	38.2 / **8.5 / 14.4**

Table 3. **Ablation study of training paradigm.** "Self." is short for Self-supervised and "O.W." for Open World. "Pre.", "Z.S." and "F.S." represent three open-world stages: pretraining, zero-shot evaluation and few-shot finetuning. Each of their corresponding items is composed of three parts: **known mIoU / unknown mIoU / mIoU**. The "Gro." here stands for Grounding.

As shown in Figure 5, AGO generalizes effectively in zero-shot supercategory expansion (e.g., vehicle → car & truck). Few-shot fine-tuning further refines subcategory completeness while improving recognition of novel classes like "manmade" and "vegetation". Overall, AGO balances open-world adaptability with strong closed-world performance.

4.3. Ablation Study

Training paradigm Table 3 compares different training paradigms in both closed- and open-world settings. Similarity-based alignment with image embeddings from pre-trained VLMs is currently the common practice for self-supervised occupancy prediction. To analyze its effective-

Open World Identifier	Discrimination Criteria	Open World Pretraining mIoU	Open World Zero-shot mIoU
✗	None	22.2 / 1.1 / 11.7	**32.8** / 1.6 / 7.8
✓	Max Confidence	**22.4** / 3.1 / 12.8	32.6 / 2.8 / 8.7
✓	Min Entropy	22.1 / **3.6** / **12.9**	32.2 / **3.2** / **9.0**

Table 4. **Ablation study of open-world inference strategy.**

Method	Noise Prompts	$\mathcal{L}_{\text{Occ}}$	Self-supervised IoU	Self-supervised mIoU
AGO			53.10	18.01
AGO		✓	55.18 (+2.08)	18.97 (+0.96)
AGO	✓	✓	**55.45** (+0.27)	**19.23** (+0.26)

Table 5. **Ablation study of architecture components.**

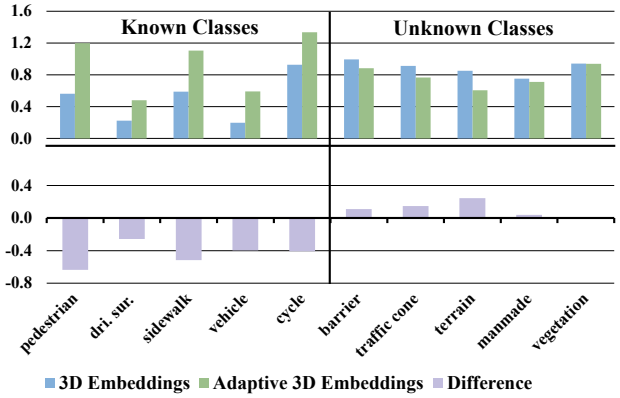


Figure 6. **Average information entropy for each class.**

ness, we replace $\mathcal{L}_{\text{Grounding}}$ in AGO with a cosine similarity loss and remove the modality adapter (denoted as "Align"). Additionally, we evaluate a pure grounding approach without the modality adapter ("Gro.") and a hybrid approach that applies both grounding and alignment to the same 3D embeddings ("Gro.+Align"). As shown in Table 3, alignment alone is not competitive in either setting, as the modality gap, low resolution, and incomplete scene representation of image embeddings limit prediction capability Grounding alone improves known category prediction but almost fails for unknown objects, yielding unknown mIoU scores of only 0.0 and 0.3 in pretraining and zero-shot evaluation. When both grounding and alignment are applied to the same embedding, the 3D representation is forced to adapt to text and image embeddings simultaneously. This conflict results in a 0.19 mIoU drop in the self-supervised setting and over a 1 mIoU decrease in known categories for the open-world task. These findings indicate that simply combining losses is insufficient to address modality conflicts. In contrast, our proposed adaptive grounding alleviates these gaps, leading to more robust prediction performance in both closed- and open-world scenarios.

Open-world inference strategy Table 4 compares the impact of the open world identifier and different criteria. Since fine-tuning provides pseudo-labels for all categories, we focus on pre-training and zero-shot evaluation. Even without an identifier, the model demonstrates some open-world prediction capability through adaptive grounding. We also observe differences in prediction distributions between the original and adaptive 3D embeddings. As shown in Figure 6, adaptive 3D embeddings generally exhibit lower information entropy for unknown categories. Similar difference also exists in the maximum confidence distribution. Based on this, we design an open world identifier with two criteria to further leverage the general perception capability of adaptive 3D embeddings (see the supplementary material for further details.). As shown in Table 4, both criteria improve unknown category prediction while minimally affecting known categories. The minimum entropy criterion provides a slightly higher gain (3.6 vs. 3.1 unknown mIoU in pretraining) compared to the confidence-based approach.

Architecture components Table 5 highlights the impact of $\mathcal{L}_{\text{Occ}}$ and the use of additional random noise prompts. Since these components affect overall occupancy prediction performance regardless of task settings, we report comparisons based on self-supervised training. Although post-processing techniques such as multi-frame aggregation and ray casting are applied, the pseudo-labels remain spatially incomplete. The occupancy loss, $\mathcal{L}_{\text{Occ}}$, helps balance the free and occupied voxels during training, leading to improvements in both IoU and mIoU. The incorporation of noise prompts as negative samples in grounding training optimizes the semantic space, thereby enhancing the discriminative capability of the 3D embeddings. Random sampling strategy from a general dictionary further reduces potential conflicts between noise prompts and label prompts, resulting in a 0.26 mIoU improvement.

5. Conclusion

In this paper, we proposed AGO, an effective 3D semantic occupancy prediction framework for open-world scenarios in autonomous driving. By integrating grounding training with noise prompts, our AGO refines 3D features to achieve enhanced discriminative power. The adaptive projection effectively transfers knowledge from pretrained VLMs, avoiding the impact of the modality gaps, while an open-world identifier based on information entropy ensures robust voxel selection for both known and unknown objects. Experiments on the Occ3D-nuScenes dataset demonstrate that AGO outperforms previous methods in both closed-world self-supervised tasks and open-world settings. While promising, there are still research directions to be explored in the future, such as better integration of temporal sequence information and the usage of semantic-rich text prompts for grounding training. Nevertheless, AGO offers a promising solution for 3D scene understanding in real-world autonomous driving applications.

Acknowledgement

This work is a result of the joint research project STADT:up (Förderkennzeichen 19A22006O). The project is supported by the German Federal Ministry for Economic Affairs and Climate Action (BMWK), based on a decision of the German Bundestag. The author is solely responsible for the content of this publication.

References

[1] S Abualhanud, E Erahan, and M Mehltretter. Self-supervised 3d semantic occupancy prediction from multi-view 2d surround images. *PFG–Journal of Photogrammetry, Remote Sensing and Geoinformation Science*, 92(5):483–498, 2024. 3

[2] Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4413–4421, 2018. 4

[3] Simon Boeder, Fabian Gigengack, and Benjamin Risse. Langocc: Self-supervised open vocabulary occupancy estimation via volume rendering. *arXiv preprint arXiv:2407.17310*, 2024. 2, 3, 4, 6, 1

[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

[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, 3

[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

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

[8] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2

[9] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3075–3084, 2019. 2

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 4, 5

[11] Juncong Fei, Wenbo Chen, Philipp Heidenreich, Sascha Wirges, and Christoph Stiller. Semanticvoxels: Sequential fusion for 3d pedestrian detection using lidar point cloud and semantic segmentation. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, pages 185–190, 2020. 2

[12] Juncong Fei, Kunyu Peng, Philipp Heidenreich, Frank Bieder, and Christoph Stiller. Pillarsegnet: Pillar-based semantic grid map estimation using sparse lidar data. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 838–844, 2021. 2

[13] Stephanie Fu, Mark Hamilton, Laura Brandt, Axel Feldman, Zhoutong Zhang, and William T Freeman. Featup: A model-agnostic framework for features at any resolution. *arXiv preprint arXiv:2403.10516*, 2024. 5

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

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

[16] 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

[17] 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

[18] 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, 4, 5

[19] 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. 2, 3, 6, 4, 5

[20] Haoyi Jiang, Liu Liu, Tianheng Cheng, Xinjie Wang, Tianwei Lin, Zhizhong Su, Wenyu Liu, and Xinggang Wang. Gausstr: Foundation model-aligned gaussian transformer for self-supervised 3d spatial understanding. *arXiv preprint arXiv:2412.13193*, 2024. 3, 6, 5

[21] Yanqin Jiang, Li Zhang, Zhenwei Miao, Xiatian Zhu, Jin Gao, Weiming Hu, and Yu-Gang Jiang. Polarformer: Multi-camera 3d object detection with polar transformer. In *Proceedings of the AAAI conference on Artificial Intelligence*, pages 1042–1050, 2023. 2

[22] 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, 1

[23] Alex H Lang, Sourabh Vora, Holger Caesar, Lubing Zhou, Jiong Yang, and Oscar Beijbom. Pointpillars: Fast encoders

for object detection from point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12697–12705, 2019. 2

[24] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2

[25] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10965–10975, 2022. 2

[26] Peizheng Li, Shuxiao Ding, Xieyuanli Chen, Niklas Hanselmann, Marius Cordts, and Juergen Gall. Powerbev: a powerful yet lightweight framework for instance prediction in bird's-eye view. *arXiv preprint arXiv:2306.10761*, 2023. 2

[27] Yiming Li, Zhiding Yu, Christopher Choy, Chaowei Xiao, Jose M Alvarez, Sanja Fidler, Chen Feng, and Anima Anandkumar. Voxformer: Sparse voxel transformer for camera-based 3d semantic scene completion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9087–9098, 2023. 2

[28] 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 *European conference on computer vision*, pages 1–18. Springer, 2022. 2

[29] Zhiwei Lin, Hongbo Jin, Yongtao Wang, Yufei Wei, and Nan Dong. Teocc: Radar-camera multi-modal occupancy prediction via temporal enhancement. In *ECAI 2024*, pages 129–136. IOS Press, 2024. 2

[30] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 3

[31] Yili Liu, Linzhan Mou, Xuan Yu, Chenrui Han, Sitong Mao, Rong Xiong, and Yue Wang. Let occ flow: Self-supervised 3d occupancy flow prediction. *arXiv preprint arXiv:2407.07587*, 2024. 3

[32] I Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[33] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2

[34] Junyi Ma, Xieyuanli Chen, Jiawei Huang, Jingyi Xu, Zhen Luo, Jintao Xu, Weihao Gu, Rui Ai, and Hesheng Wang. Cam4docc: Benchmark for camera-only 4d occupancy forecasting in autonomous driving applications. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21486–21495, 2024. 2

[35] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 194–210. Springer, 2020. 2

[36] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 3

[37] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 3

[38] 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. 3, 4, 5, 1

[39] Luis Roldao, Raoul de Charette, and Anne Verroust-Blondet. Lmscnet: Lightweight multiscale 3d semantic completion. In *2020 International Conference on 3D Vision (3DV)*, pages 111–119. IEEE, 2020. 2

[40] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *European Conference on Computer Vision*, pages 72–87. Springer, 2024. 2

[41] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

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

[43] 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. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 5, 6, 1, 3, 4

[44] 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. 2

[45] Antonin Vobecky, Oriane Siméoni, David Hurych, Spyridon Gidaris, Andrei Bursuc, Patrick Pérez, and Josef Sivic. Pop-3d: Open-vocabulary 3d occupancy prediction from images. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3, 4, 6, 1, 5

[46] Yuqi Wang, Yuntao Chen, Xingyu Liao, Lue Fan, and Zhaoxiang Zhang. Panoocc: Unified occupancy representation for camera-based 3d panoptic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17158–17168, 2024. 2

[47] 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. 2

[48] Zhaoyang Xia, Youquan Liu, Xin Li, Xinge Zhu, Yuexin Ma, Yikang Li, Yuenan Hou, and Yu Qiao. Scpnet: Semantic scene completion on point cloud. In *Proceedings of*

the IEEE/CVF conference on computer vision and pattern recognition, pages 17642–17651, 2023. 2

[49] Ziyang Yan, Wenzhen Dong, Yihua Shao, Yuhang Lu, Liu Haiyang, Jingwen Liu, Haozhe Wang, Zhe Wang, Yan Wang, Fabio Remondino, et al. Renderworld: World model with self-supervised 3d label. *arXiv preprint arXiv:2409.11356*, 2024. 2

[50] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 2

[51] 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

[52] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2

[53] Qingwen Zhang, Yi Yang, Heng Fang, Ruoyu Geng, and Patric Jensfelt. Deflow: Decoder of scene flow network in autonomous driving. *arXiv preprint arXiv:2401.16122*, 2024. 2

[54] Qingwen Zhang, Yi Yang, Peizheng Li, Olov Andersson, and Patric Jensfelt. Seflow: A self-supervised scene flow method in autonomous driving. In *European Conference on Computer Vision*, pages 353–369. Springer, 2025. 2

[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

[56] Xiao Zhao, Bo Chen, Mingyang Sun, Dingkang Yang, Youxing Wang, Xukun Zhang, Mingcheng Li, Dongliang Kou, Xiaoyi Wei, and Lihua Zhang. Hybridocc: Nerf enhanced transformer-based multi-camera 3d occupancy prediction. *IEEE Robotics and Automation Letters*, 2024. 2

[57] 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. 2, 3, 4, 6, 1

[58] Brady Zhou and Philipp Krähenbühl. Cross-view transformers for real-time map-view semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13760–13769, 2022. 2

[59] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 3, 4, 5, 1

[60] Yin Zhou and Oncel Tuzel. Voxelnet: End-to-end learning for point cloud based 3d object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4490–4499, 2018. 2