AGO: Adaptive Grounding for Open World 3D Occupancy Prediction

Supplementary Material

In this supplementary material, we first describe the implementation details of our method in Sec. A. Following that, additional experimental results and ablation study are presented in Sec. B. Finally, Sec. C provides more visualization comparisons and qualitative analysis.

A. Additional Implementation Details

A.1. Label space design

Choosing appropriate label prompts is crucial for effectively establishing semantic associations between categories in the language priors. However, some of the category names in the Occ3D-nuScenes dataset [43] are too vague or broad to be directly encoded into semantically rich text embeddings. For example, in the definition of Occ3D-nuScenes [43], "others" and "other flat" represent structures and horizontal ground-level planes that cannot be classified into any other category, respectively. They are just general terms for many categories that have not been specifically annotated and thus do not have a clear semantic meaning. Therefore, we follow existing works [3, 45, 57] and divide them into the subcategories as shown in Table A. The above design is employed in all supervised and self-supervised experiments mentioned in this paper. For pretraining as well as zero-shot and few-shot transfer in the open-world setting, we adopt the original labels as corresponding prompts after removing the semantically ambiguous "others" and "other flat". The reason for this is to avoid the influence of different label space designs on open-world performance.

A.2. Pseudo-label generation

Thanks to the integrated SAM [22], Grounded SAM [38] can generate semantic masks with more detail and more accurate boundaries than MaskCLIP+ [59]. Therefore, we utilize pre-trained Grounded SAM to generate pseudo-labels. We feed each surrounding image and the label prompts defined in Appendix A.1 into this model to generate 2D pseudo semantic masks corresponding to the image. In this process, the box threshold is set to 0.2 and the text threshold is set to 0.15. These 2D masks are projected into the 3D voxels based on the calibration matrices of the LiDAR and surrounding cameras.

Considering the sparsity of the LiDAR point cloud, we aggregate multiple frames to densify it. Specifically, we first select a certain number N_{sweep} of camera sweeps and pseudo-synchronize each of them with the temporally closest LiDAR sweep. It is worth noting that the above selection is not continuous, but has a sampling interval of N_{interval}. This is to allow the pseudo labels to cover as much space as possible while using the same number of sweeps. Thereafter, the generated voxelized 3D pseudo-labels of each sweep are warped to the reference sweep and superimposed to obtain dense pseudo-labels. In our implementation, we set $N_{\text{sweep}} = 30$ and $N_{\text{interval}} = 2$ corresponding to a 15 s time range and take the key frame as the reference sweep. Despite this, there are still many false negatives in the above pseudo-labels, *i.e.*, many occupied voxels are marked as free, which is caused by occlusion. Therefore, we employ ray casting for free voxel assignment, that is, only unoccupied voxels between the LiDAR and the reflection point on each ray are set as free, while the remaining occluded voxels are ignored. In addition, several LiDAR points may be located in the same voxel. For this reason, we apply a semantic voting mechanism for each voxel, which selects the category with the most corresponding points as the category of the voxel.

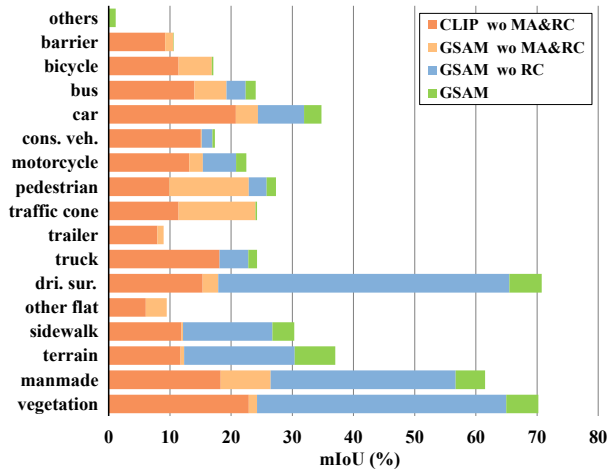

Figure A. **Pseudo-label evaluation on the Occ3D-nuScenes [43] dataset.** "MA" denotes multi-frame aggregation and "RC" refers to ray casting. "GSAM" is an acronym for Grounded SAM.

Table B compares the evaluation results of the pseudo-labels with different model bases and post-processing. As can be seen, the mIoU of the pseudo-labels generated based on Grounded SAM is 3.24 higher than that of MaskCLIP+. Multi-frame aggregation brings an improvement of 9.89 mIoU, while ray casting further increases the mIoU by 2.08. It is worth noting that we did not conduct extensive prompt engineering to enhance the quality of pseudo-labels. However, our method even outperforms the pseudo labels in many categories, such as "driveable surface", "sidewalk" and "terrain". This further demonstrates the effectiveness of our proposed AGO.

Classes	Subclasses
others	animal, skateboard, segway, scooter, stroller, wheelchair, trash bag, dolley, wheel barrow, trash bin, shopping cart, bicycle rack, ambulance, police vehicle, cyclist
barrier	barrier
bicycle	bicycle
bus	bendy bus, rigid bus
car	car, van, suv
construction vehicle	construction vehicle
motorcycle	motorcycle
pedestrian	adult pedestrian, child pedestrian, worker, police officer
traffic cone	traffic cone
trailer	trailer
truck	truck
driveable surface	road
other flat	traffic island, traffic delimiter, rail track, lake, river
sidewalk	sidewalk
terrain	lawn
manmade	building, sign, pole, traffic light
vegetation	tree, bush

Table A. **Subclass description in label space design.** These subclasses apply to self-supervised and open-world training.

Method	Model Base	others	barrier	bicycle	bus	car	construction vehicle	motorcycle	pedestrian	traffic cone	trailer	truck	driveable surface	other flat	sidewalk	terrain	manmade	vegetation	mIoU
CLIP wo MA&RC	MaskCLIP+	0.05	9.23	11.41	13.99	20.79	15.05	13.15	9.90	11.43	7.97	18.11	15.36	6.08	11.88	11.74	18.32	22.90	12.79
GSAM wo MA&RC	Grounded SAM	0.04	10.58	16.86	19.26	24.37	15.19	15.38	22.91	23.98	8.96	12.37	17.90	9.48	12.14	12.34	26.48	24.24	16.03
GSAM wo RC	Grounded SAM	0.03	10.44	15.09	22.38	31.94	16.97	20.81	25.79	23.55	7.44	17.09	65.49	4.84	26.76	30.39	56.69	64.95	25.92
GSAM	Grounded SAM	1.12	10.50	15.33	24.02	34.77	17.34	22.50	27.34	23.82	6.78	18.51	70.77	4.14	30.31	37.03	61.50	70.21	28.00
Self-supervised AGO	-	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	19.23

Table B. **Pseudo-label evaluation on the Occ3D-nuScenes [43] dataset.** "MA" denotes multi-frame aggregation and "RC" refers to ray casting. "GSAM" is an acronym for Grounded SAM.

A.3. Framework details

As described in Section 3.1, AGO features a dual-stream architecture with a text encoder and a vision-centric 3D encoder. Since the traditional TPVFormer [18] is designed for the LiDAR point cloud segmentation task, it involves both point-level and voxel-level supervision. But for 3D semantic occupancy prediction, point-level supervision is not necessary. Therefore, we remove this part from our implementation.

Since we split the original categories into subcategories according to Appendix A.1, during supervised grounded training, one class in the ground truth may correspond to the text embeddings and similarity scores of multiple subclasses. To solve this problem, for each voxel, we take the maximum score across all subclasses belonging to the same class as its similarity score. In other words, as long as one subcategory exhibits an extremely high similarity, the original category containing that subcategory should be considered the occupancy prediction for the corresponding voxel.

In AGO, we use a dictionary obtained from Natural Language Toolkit (NLTK) library of Python as the source of noise prompts. For each step, we randomly select N_{noise} prompts from it and encode them into corresponding noise embeddings using the same text encoder. In our implementation, we set $N_{\text{noise}} = 100$.

In addition, the main purpose of the open world identifier is to flexibly select suitable features based on the prediction distribution of the original 3D embeddings and adaptive 3D embeddings. Considering that in the closed-world setting, all categories are known during grounding training, the original 3D embedding has stronger discriminative ability. Therefore, during closed-world prediction, the open world identifier directly selects the original 3D embedding for the final prediction.

Method	Image Backbone	Training Epochs	Image Resolution	Self-supervised IoU	Self-supervised mIoU
AGO	ResNet-101	24	900×1600	55.45	19.32
AGO	ResNet-50	24	900×1600	50.76	15.23
AGO	ResNet-50	12	900×1600	50.06	14.84
AGO	ResNet-50	12	450×800	50.24	14.78

Table C. **Ablation study of image backbones, traninig epochs and resolutions.**

$\mathcal{L}_{\text{Alignment}}$	Open World		
	Pretraining mIoU	Zero-shot mIoU	Few-shot mIoU
Cosine	22.1 / 3.6 / 12.9	32.2 / 3.2 / 9.0	38.2 / 8.5 / 14.4
MSE (L2)	21.1 / 0.3 / 11.1	31.7 / 0.4 / 6.7	36.9 / 7.1 / 12.4
MAE (L1)	20.8 / 0.1 / 10.5	30.1 / 0.3 / 6.4	37.1 / 6.2 / 12.0

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

Number of MLP Layers	Open World		
	Pretraining mIoU	Zero-shot mIoU	Few-shot mIoU
1	18.4 / 2.6 / 10.5	29.2 / 1.7 / 6.8	38.0 / 6.5 / 13.2
2	22.1 / 3.6 / 12.9	32.2 / 3.2 / 9.0	38.2 / 8.5 / 14.4
3	14.3 / 3.7 / 10.4	22.0 / 3.4 / 8.3	37.3 / 8.2 / 13.9
4	13.6 / 3.8 / 10.3	19.5 / 3.4 / 8.1	37.1 / 7.9 / 13.7

Table E. **Ablation study of MLP layer number.**

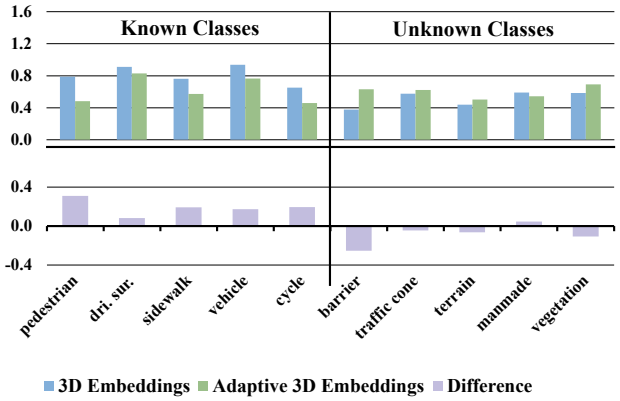

Figure B. **Average maximum confidence for each class.**

A.4. Closed-world task details

For the vast majority of closed-world methods, we directly adopt the performance values reported in their papers. However, POP-3D [45], as a pioneering alignment-based zero-shot method, has not yet been compared on the Occ3D [43] benchmark. Its evaluation is based on the original nuScenes dataset [5] and considers only the voxels traversed by LiDAR rays in a single frame, with a resolution of 100×100×8 (see Sec. 4.1 in POP-3D [45]). For fairness, we retrain it using the 200×200×16 resolution consistent with Occ3D [43] setting. The drop in reported mIoU largely stems from the stricter evaluation protocol (see Sec. 3.3 & 6.1 in Occ3D [43]).

A.5. Open-world task details

In the open-world task, the entire progress is divided into three stages: pre-training, zero-shot evaluation and few-shot finetuning. In the pretraining stage, only the pseudo-labels of 5 major categories are known, namely $\mathbb{C}_{pre.} = \{$ "pedestrian", "driveable surface", "sidewalk", "vehicle", "cycle" $\}$. Among them, "vehicle" and "cycle" are two supercategories, which are formed by the original category sets $\{$ "car", "bus", "construction vehicle", "trailer", "truck" $\}$ and $\{$ "bicycle", "motorcycle" $\}$, respectively, to simulate the common coarse-to-fine labeling process in real-world applications. The remaining Occ3D classes as well as corresponding voxels are ignored during pre-training but included in the subsequent zero-shot evaluation. During the few-shot fine-tuning stage, only a small number of samples are provided. They have the complete original Occ3D label space and each category appears at least in k samples (k-shot setting). We set $k = 100$ and repeat every few-shot experiment 5 times to calculate the average, thus reducing randomness. This is to validate model's few-shot generalization ability of unknown classes. Noting that the original classes "others" and "other flat" are semantically ambiguous, they are not included in all stages.

Due to different categories in open-world pretraining stage, all methods need to be retrained (based on their original code). The specific retraining settings are as follows:

- POP-3D [45]: It is trained solely with the original alignment loss, without pseudo-labels. The only difference across stages lies in the different class text prompts used for inference and evaluation.
- SelfOcc [19]: We define its output label space as the full class set ($\mathbb{C}_k \cup \mathbb{C}_{uk}$). In the open-world pretraining stage, only outputs of $\mathbb{C}_k$ are supervised by pseudo-labels, while all categories are considered during zero-shot evaluation, resulting in 0 unknown mIoU. In the open-world fine-tuning stage, all class outputs are supervised.
- GaussTR [20]: The alignment loss is the same at all open-world stages. In the open-world pretraining stage, only pseudo-labels of $\mathbb{C}_k$ are used for the extra loss to refine the semantic boundaries, while in the open-world fine-tuning stage, pseudo-labels of all classes are used.

B. Additional Experiments and Analyses

B.1. Additional ablation study

Considering the closed-world self-supervised methods that we compared to used different image backbones, traninig epochs and resolutions, we show the corresponding ablation

study results in Tab. C. As can be seen, the image backbone has the greatest impact on performance. However, even using only ResNet-50, our AGO still outperforms all existing methods with 15.23 mIoU on the self-supervised benchmark. In contrast, the training epochs and the resolution of the input images have relatively small influence. But even under the most challenging setting, our method is still at the same level as the current state-of-the-art model (with only a 0.36 mIoU gap). This further indicates that the effectiveness of our method does not come from large numbers of parameters, long training durations, or high-resolution input images, but from the framework design itself.

Table D illustrates the impact of different alignment loss functions on the open-world performance of the model. It can be observed that replacing the cosine similarity loss with either mean squared error (MSE) or mean absolute error (MAE) loss leads to a degradation in prediction performance across all stages, regardless of whether the objects are from known or unknown categories. Notably, in both the pretraining and zero-shot phases, the model almost entirely loses its ability to recognize unknown instances. This finding underscores that, in contrast to cosine similarity loss, MSE and MAE losses are not suitable for cross-modal alignment tasks, thereby impairing the perception capability of open-world scenes.

As shown in the Table E, we further compare the impact of the number of MLP layers in the modality adapter on the open-world prediction capability. Notably, when the adapter consists of only a single layer MLP, it does not include any non-linear activation functions. In this case, the semantic space before and after adaptation remains highly similar, leading to performance comparable to Gro.+Align in Table 3. As the number of non-linear projection layers increases, a clear trend is observed: while the mIoU for unknown categories has a slight improvement, the mIoU for known categories degrades significantly. Considering the overall predictive performance, we use a two-layer MLP as the modality adapter.

B.2. Confidence analysis of 3D embeddings

In addition to the information entropy of the predicted probability distribution, we also analyze the maximum confidence score of each category during the pretraining phase, *i.e*. the maximum value of the probability distribution.. As shown in the Figure B, the adaptive 3D embedding exhibits generally higher maximum confidence for unknown categories, while demonstrating more confident predictions for known categories. This observation aligns with our previous entropy-based analysis in Section 4.3, where predictions with lower entropy tend to correspond to higher confidence scores. Therefore, maximum confidence can also serve as a criterion in the open world identifier.

Method	Param.	Seen/Unseen mIoU 0/17	9/8	13/4
VEON [57]	678.1M	15.14	15.16	19.94
AGO	62.5M	**19.23**	**22.42**	**25.90**

Table F. **VEON's open-world benchmark.**

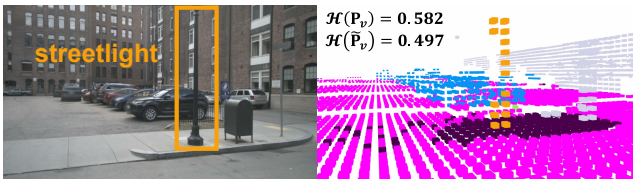

Figure C. **Visualization of open-vocabular retrieval.**

B.3. Comparison under VEON's open-world setting

VEON [57] defines another open-world benchmark with partial semantic labels. Specifically, it divides the complete label set into X seen categories with GT annotations and Y unseen categories without annotations for training, and then performs inference and evaluation on the complete label set. In Table F, we also compare AGO with it under the same settings. It can be observed that, regardless of the X/Y setting, our method consistently outperforms VEON [57] by a significant margin, while utilizing less than 10% of its parameters.

B.4. Comparison on Occ3D-Waymo dataset

In Table G, we provide the closed-world comparison based on the Occ3D-Waymo [43] dataset. In addition, Table H presents the prediction performance comparison under the open-world setting, with the "GO" category excluded due to its semantic ambiguity.

B.5. Detailed open-world results

In Table I, we provide the detailed prediction results of the open-world ablation experiments in Table 3 and 4.

C. Additional Visualization

Figure C shows the visualization of AGO's open vocabulary retrieval results, where $\mathcal{H}(P_v)$ and $\mathcal{H}(\tilde{P}_v)$ represent the average entropy of the corresponding voxel predictions before and after the modality adapter. We also show more visual comparisons of self-supervised 3D semantic occupancy prediction in Figure D. It can be seen that compared to SelfOcc [19], our AGO provides more complete and fine-grained predictions. Especially for dynamic categories with small scales, such as cars and pedestrians, the results of our method are closer to the ground truth. However, due to the natural flaws of volume rendering-based methods in dynamic objects, SelfOcc [19] has massive false positives and false negatives in the predictions of these categories. In addition, limited by the computational complexity, 3D

Method	Image Backbone	GO	vehicle	bicyclist	pedestrian	sign	traffic light	pole	cons. cone	bicycle	motorcycle	building	vegetation	tree trunk	road	sidewalk	mIoU
POP-3D† [45]	ResNet-101	0.31	20.31	5.46	**0.83**	0.00	7.11	10.02	7.29	9.36	0.65	12.62	8.90	1.60	65.51	18.89	11.26
SelfOcc† [19]	ResNet-50	1.06	22.90	6.38	0.52	0.00	9.03	14.88	6.68	11.25	0.23	15.10	8.81	2.96	69.77	**22.32**	12.79
GaussTR† [20]	VFMs	**2.10**	**23.13**	**7.15**	0.15	0.00	**10.71**	**15.59**	**8.21**	**12.86**	**0.89**	**19.52**	**12.25**	**3.69**	**70.95**	**22.41**	**13.97**
AGO (ours)	ResNet-101	**2.49**	**26.97**	**6.94**	**1.01**	0.00	**25.02**	**17.95**	**11.47**	**14.15**	**1.87**	**22.06**	**14.56**	**3.77**	**71.72**	20.26	**16.02**

Table G. **3D occupancy prediction performance under the self-supervised setting on the Occ3D-Waymo [43] dataset.** "cons. cone" stands for construction cone. † indicate values obtained from our retraining. Results are highlighted in **bold & underlined** for the best performance and **bold** for the second-best performance.

Training Stages	Method	pedestrian	road	sidewalk	vehicle	cycle	k. mIoU	bicycle	motorcycle	bicyclist	sign	traffic light	pole	cons. cone	building	vegetation	tree trunk	u. mIoU	mIoU
Pretraining	POP-3D† [45]	0.00	49.84	6.89	10.43	0.00	13.43	-	-	0.00	0.00	0.00	0.00	0.00	1.26	0.57	0.44	0.28	5.34
	SelfOcc† [19]	**0.27**	67.45	16.07	28.05	0.00	22.37	-	-	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	8.60
	GaussTR† [20]	0.22	**69.97**	**22.85**	12.05	0.00	21.02	-	-	0.00	**0.04**	0.00	0.00	0.00	4.20	0.99	0.29	0.69	8.51
	AGO (ours)	0.09	67.59	19.50	**28.29**	**6.92**	**24.48**	-	-	**0.03**	0.00	**0.01**	0.00	**0.05**	**5.16**	**1.92**	**0.61**	**0.97**	**10.01**
Zero-shot Evaluation	POP-3D† [45]	0.00	49.84	6.89	10.43	-	16.79	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.26	0.57	0.44	0.23	4.96
	SelfOcc† [19]	**0.27**	67.45	16.07	28.05	-	27.96	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	7.99
	GaussTR† [20]	0.22	**69.97**	**22.85**	12.05	-	26.27	0.00	0.00	0.00	**0.04**	0.00	0.00	0.00	4.20	0.99	0.29	0.55	7.90
	AGO (ours)	0.09	67.59	19.50	**28.29**	-	**28.87**	**5.56**	**0.14**	**0.03**	0.00	**0.01**	0.00	**0.05**	**5.16**	**1.92**	**0.61**	**1.35**	**9.21**
Few-shot Finetuning	POP-3D† [45]	0.00	37.96	6.26	9.12	-	13.34	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.33	0.49	0.66	0.25	3.99
	SelfOcc† [19]	0.55	68.21	18.20	29.60	-	29.14	1.09	0.00	0.08	0.00	0.91	0.00	0.00	3.11	1.02	0.03	0.62	8.77
	GaussTR† [20]	0.61	69.71	23.10	12.05	-	26.37	5.15	0.02	2.12	0.00	2.06	1.03	1.27	6.84	3.15	1.12	2.28	9.16
	AGO (ours)	**0.82**	**70.15**	**23.50**	**30.03**	-	**31.13**	**9.26**	**0.56**	**2.13**	0.00	**10.86**	**5.28**	**4.15**	**11.23**	**8.81**	**2.57**	**5.49**	**12.81**

Table H. **3D occupancy prediction performance under the open-world setting on the Occ3D-Waymo [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.

features can only be sampled along the rays at a relatively low sampling rate in volume rendering, which results in a large number of periodic blank strip textures in the final prediction. This phenomenon is highly evident in the "driveable surface" predictions shown in the third and fifth rows of the figure. Even with constraints such as Hessian loss $\mathcal{L}_H$, regularization $\mathcal{L}_S$, and Eikonal term $\mathcal{L}_E$ [19], this issue cannot be fundamentally solved. These qualitative comparisons further indicate that existing self-supervised models are insufficient for 3D scene understanding. In contrast, our proposed AGO framework based on grounded training has greater potential in this regard.

Training Stages	Method	pedestrian	driveable surface	sidewalk	vehicle	cycle	known mIoU	car	bus	construction vehicle	trailer	truck	bicycle	motorcycle	barrier	traffic cone	terrain	manmade	vegetation	unknown mIoU	mIoU
Pretraining	Align	0.00	58.10	12.29	6.72	0.00	15.42	-	-	-	-	-	-	-	0.00	0.00	3.81	0.00	0.16	0.79	8.11
	Gro.	1.84	68.68	28.52	3.15	0.99	20.64	-	-	-	-	-	-	-	0.00	0.00	0.00	0.00	0.00	0.00	10.32
	Gro. + Align	4.31	57.73	25.96	2.36	0.90	18.25	-	-	-	-	-	-	-	0.00	0.00	2.42	0.02	8.49	2.19	10.22
	AGO w/o OWI	7.25	65.27	25.93	7.82	4.92	22.24	-	-	-	-	-	-	-	0.00	0.00	0.78	0.00	4.63	1.08	11.66
	AGO w/ Max Confi.	7.36	64.76	25.68	8.92	5.32	22.41	-	-	-	-	-	-	-	0.00	0.00	6.90	0.01	8.73	3.13	12.77
Zero-shot Evaluation	Align	0.00	58.10	12.29	-	-	23.46	6.81	0.00	0.00	0.57	2.91	0.00	0.00	0.00	0.00	3.81	0.00	0.16	1.19	5.64
	Gro.	1.84	68.68	28.52	-	-	33.01	2.21	0.00	0.00	0.02	0.87	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.26	6.81
	Gro. + Align	4.31	57.73	25.96	-	-	29.33	2.78	0.00	0.00	0.15	1.29	1.36	0.00	0.00	0.00	2.42	0.02	8.49	1.38	6.97
	AGO w/o OWI	7.25	65.27	25.93	-	-	32.82	5.26	0.00	0.00	0.82	4.36	3.08	0.00	0.00	0.00	0.78	0.00	4.63	1.58	7.83
	AGO w/ Max Confi.	7.36	64.76	25.68	-	-	32.60	7.63	0.00	0.00	0.71	6.08	3.13	0.00	0.00	0.00	6.90	0.01	8.73	2.77	8.73
Few-shot Finetuning	Align	0.00	58.95	14.18	-	-	24.38	5.64	0.00	0.00	0.50	3.97	0.00	0.00	0.00	0.00	11.70	11.34	15.28	4.04	8.10
	Gro.	13.34	70.95	30.90	-	-	38.40	15.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.02	8.14	12.25	3.03	10.10
	Gro. + Align	12.01	71.57	28.44	-	-	37.34	13.52	4.11	0.00	0.00	1.36	1.24	0.00	0.00	0.00	26.83	8.86	12.68	5.72	12.04
	AGO w/o OWI	12.58	72.02	30.12	-	-	38.24	18.69	5.20	0.00	0.21	2.56	3.59	2.18	0.13	0.00	29.39	21.54	17.80	8.44	14.40
	AGO w/ Max Confi.	12.97	71.53	29.63	-	-	38.04	18.58	5.03	0.00	0.23	2.74	3.56	2.04	0.30	0.00	29.33	21.32	17.81	8.41	14.34

Table I. **Detailed 3D occupancy prediction results under the open-world setting on the Occ3D-nuScenes [43] dataset.** 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.

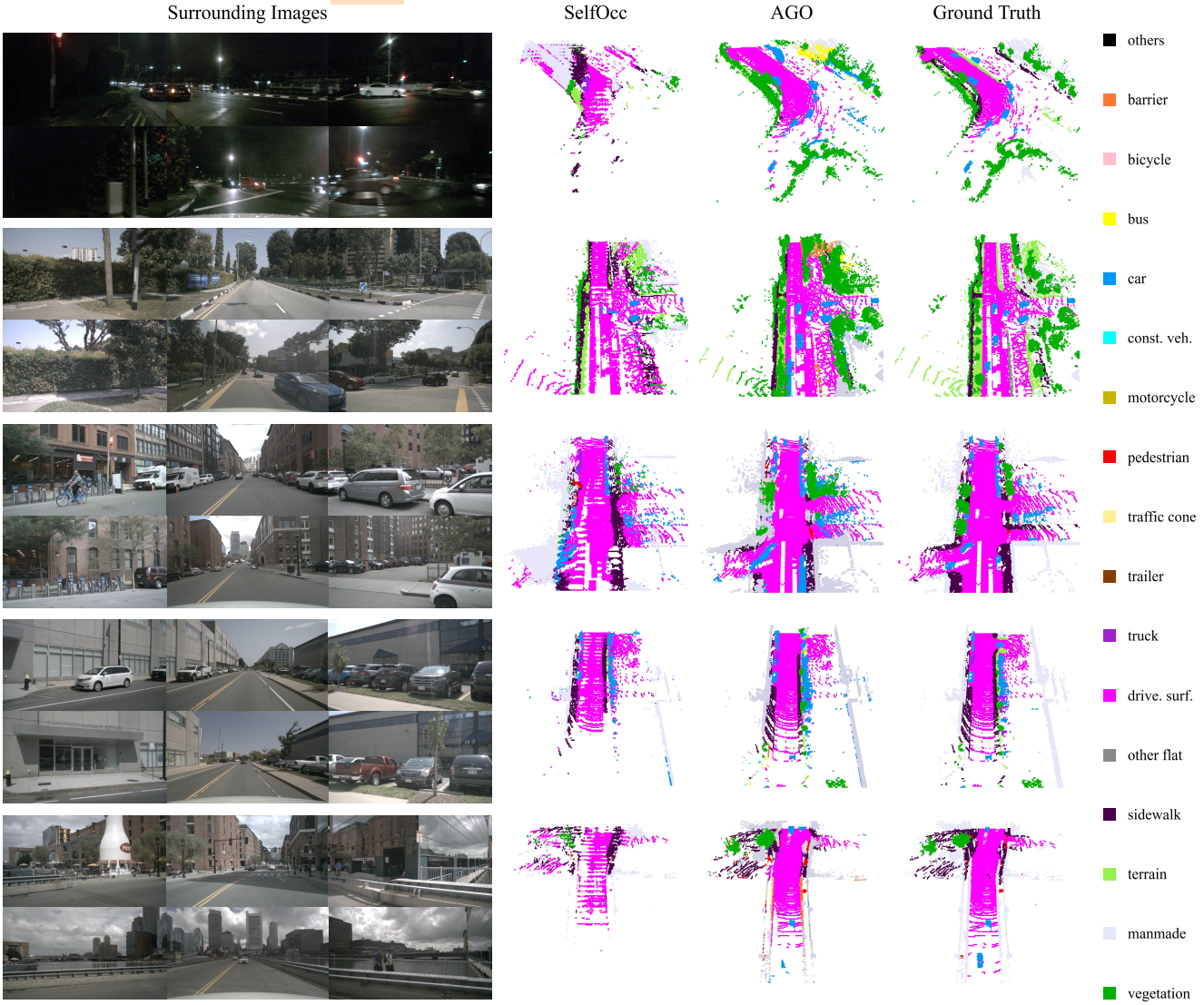

Figure D. **Additional visualization of self-supervised 3D semantic occupancy prediction on the Occ3D-nuScenes occupancy benchmark.**