ALOcc: Adaptive Lifting-Based 3D Semantic Occupancy and Cost Volume-Based Flow Predictions

Supplementary Material

A. Metrics

The Occ3D dataset [24] contains 18 categories, comprising 17 semantic categories: *others*, *barrier*, *bicycle*, *bus*, *car*, *construction vehicle*, *motorcycle*, *pedestrian*, *traffic cone*, *trailer*, *truck*, *driveable surface*, *other flat*, *sidewalk*, *terrain*, *manmade*, *vegetation*, and an additional category representing non-occupied space, termed *empty*. The OpenOcc dataset [25] contains all categories from Occ3D except for the *others* category. Our approach is evaluated across multiple metrics, including $\text{mIoU}_\text{D}^\text{m}$, mIoU^m, mIoU, *RayIoU*, *mAVE*, mAVE_TP, and *Occ Score* [13, 25].

$mIoU^m$ and mIoU: mIoU^m and mIoU represent the mean Intersection over Union (IoU) over all semantic categories (the superscript m indicating whether the camera visibility mask was used during training):

$$\text{mIoU} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\text{IoU}(c),$$

where $\mathcal{C}$ denotes the set of all semantic categories, and $\text{IoU}(c)$ represents the IoU for category c.

$mIoU_\text{D}^m$: The $\text{mIoU}_\text{D}^\text{m}$ metric measures the mean IoU for the eight dynamic object categories (*i.e.*, $\mathcal{C}_\text{D} = \{$ *bicycle*, *bus*, *car*, *construction vehicle*, *motorcycle*, *pedestrian*, *trailer*, *truck* $\}$):

$$\text{mIoU}_\text{D}^\text{m} = \frac{1}{|\mathcal{C}_\text{D}|}\sum_{c\in\mathcal{C}_\text{D}}\text{IoU}(c).$$

RayIoU: Ray-based mIoU [13] calculates the mIoU using query rays instead of voxels, simulating LiDAR by projecting rays into the predicted 3D occupancy volume. A query ray is considered a true positive (TP) if both the predicted and ground-truth class labels match, and the L1 error between the predicted and ground-truth depth is within a certain threshold (*e.g.*, 2m):

$$\text{RayIoU} = \frac{1}{|\mathcal{C}|}\sum_{c=1}^{|\mathcal{C}|}\frac{\text{TP}_c}{\text{TP}_c+\text{FP}_c+\text{FN}_c},$$

where TP_c represents the true positives based on both semantic accuracy and depth error threshold, and FP_c and FN_c are the false positives and false negatives for class c, respectively. The final RayIoU is computed as the average of RayIoU values at thresholds of 1m, 2m, and 4m.

$mAVE_{TP}$: The absolute velocity error (AVE) is defined for dynamic object categories $\mathcal{C}_\text{D}$. The mAVE_TP is computed for the true positives of RayIoU when the depth error threshold is 2m, and represents the average velocity error for those voxels:

$$\text{mAVE}_\text{TP} = \frac{1}{|\mathcal{C}_\text{D}|}\sum_{c\in\mathcal{C}_\text{D}}\frac{1}{|\mathcal{V}_\text{TP}^\text{c}|}\sum_{v\in\mathcal{V}_\text{TP}^\text{c}}\text{AVE}(v),$$

where $\mathcal{V}_\text{TP}^\text{c}$ denotes the set of true positive voxels of class c.

mAVE: The mAVE is computed as the average velocity error across all voxels *w.r.t.* dynamic object categories:

$$\text{mAVE} = \frac{1}{|\mathcal{C}_\text{D}|}\sum_{c\in\mathcal{C}_\text{D}}\frac{1}{|\mathcal{V}^c|}\sum_{v\in\mathcal{V}^c}\text{AVE}(v).$$

Occ Score: The occupancy score is a comprehensive metric for joint evaluation of semantic occupancy and motion flow, defined as a weighted sum of RayIoU and mAVE_TP. The Occ Score is given by:

$$\text{OccScore} = \text{RayIoU}\times 0.9 + \max(1-\text{mAVE}_\text{TP}, 0.0)\times 0.1.$$

B. Additional Experimental Results

B.1. Per-Class IoU

Tab. A.1 presents the per-class 3D occupancy prediction results *w.r.t. training with camera visible mask* on Occ3D. ALOcc achieves the best performance in most classes. Notably, ALOcc demonstrates significant improvements in rare but traffic-critical categories such as *pedestrian* and *truck*, highlighting the practical applicability of our method in real-world scenarios.

B.2. Results on Surroundocc Benchmark

We additionally evaluate our method on the SurroundOcc benchmark [28] to facilitate a broader comparison with recent work. Although also built on the nuScenes dataset, SurroundOcc differs from Occ3D by omitting the “others” semantic class and providing ground truth annotations of slightly lower quality. Despite this, its adoption in many recent studies [7, 8, 28] makes it a relevant benchmark. A key characteristic of its evaluation protocol is that, unlike Occ3D, metrics are computed over all voxels. To ensure a fair comparison with prior work that seldom incorporates temporal fusion, we report our results without this component, marking them with an asterisk (*). As shown in Tab. A.2, ALOcc achieves SOTA performance, outperforming all other competitors across all reported metrics.

Method	Input	Backbone	Image Size	mIoU_D^m	mIoU^m	others ■	barrier ■	bicycle ■	bus ■	car ■	cons. veh. ■	motor. ■	pedes. ■	tfc. cone ■	trailer ■	truck ■	drv. surf. ■	other flat ■	sidewalk ■	terrain ■	manmade ■	vegetation ■
BEVDetOcc-SF [5]	C	ResNet-50	256 × 704	34.4	41.9	12.1	50.0	22.1	43.9	53.9	29.1	23.8	25.8	28.5	34.9	41.8	84.3	44.4	57.5	61.0	53.1	46.7
UniOCC [20]	C	ResNet-50	256 × 704	-	39.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
FB-Occ [12]	C	ResNet-50	256 × 704	34.2	39.8	13.8	44.5	27.1	46.2	49.7	24.6	27.4	28.5	28.2	33.7	36.5	81.7	44.1	52.6	56.9	42.6	38.1
SurroundSDF [15]	C	ResNet-50	256 × 704	36.2	42.4	13.9	49.7	27.8	44.6	53.0	30.0	29.0	28.3	31.1	35.8	41.2	83.6	44.6	55.3	58.9	49.6	43.8
FlashOCC [31]	C	ResNet-50	256 × 704	24.7	32.0	6.2	39.6	11.3	36.3	44.0	16.3	14.7	16.9	15.8	28.6	30.9	78.2	37.5	47.4	51.4	36.8	31.4
COTR [18]	C	ResNet-50	256 × 704	38.6	44.5	13.3	52.1	32.0	46.0	55.6	32.6	32.8	30.4	34.1	37.7	41.8	84.5	46.2	57.6	60.7	52.0	46.3
ViewFormer [9]	C	ResNet-50	256 × 704	35.0	41.9	12.9	50.1	28.0	44.6	52.9	22.4	29.6	28.0	29.3	35.2	39.4	84.7	**49.4**	57.4	59.7	47.4	40.6
OPUS [26]	C	ResNet-50	256 × 704	33.3	36.2	11.9	43.5	25.5	41.0	47.2	23.9	25.9	21.3	29.1	30.1	35.3	73.1	41.1	47.0	45.7	37.4	35.3
ALOcc-2D-mini	C	ResNet-50	256 × 704	35.4	41.4	14.2	48.6	28.7	44.8	52.8	24.7	29.2	29.0	32.0	34.6	39.6	82.4	46.9	54.8	57.7	44.7	39.3
ALOcc-2D	C	ResNet-50	256 × 704	38.7	44.8	**15.4**	52.2	**32.2**	46.2	55.4	28.2	**34.1**	**32.4**	**36.4**	38.0	42.8	84.2	48.8	57.4	60.0	52.9	45.6
ALOcc-3D	C	ResNet-50	256 × 704	**39.3**	**45.5**	15.3	**52.5**	30.8	**47.2**	**55.9**	**32.7**	33.3	**32.4**	36.2	**38.9**	**43.7**	**84.9**	48.5	**58.8**	**61.9**	**53.5**	**47.3**
ALOcc-2D-mini	C	Intern-T	256 × 704	37.9	43.7	14.8	50.1	31.3	48.1	55.7	23.8	32.8	31.6	33.5	36.3	43.7	84.0	49.2	57.1	59.8	48.2	42.4
ALOcc-2D	C	Intern-T	256 × 704	40.7	46.6	16.3	53.3	35.0	48.3	57.6	28.7	35.3	34.6	38.1	40.0	46.5	85.2	50.5	59.1	61.8	54.3	47.1
ALOcc-3D	C	Intern-T	256 × 704	41.5	47.5	17.0	54.6	34.5	50.6	58.2	28.6	36.5	34.8	39.6	41.1	47.6	85.7	51.5	60.0	63.5	55.0	48.3
HyDRa [29]	C+R	ResNet-50	256 × 704	40.6	44.4	15.1	51.1	32.7	52.3	56.3	29.4	35.9	35.1	33.7	39.1	44.1	80.4	45.1	52.0	55.3	52.1	44.4
EFFOcc [23]	C+L	ResNet-50	256 × 704	50.1	52.8	12.1	**59.7**	33.4	**61.8**	**65.0**	35.5	**46.0**	**57.1**	41.0	47.9	**54.6**	82.8	44.0	56.4	60.2	**71.1**	**69.6**
SDGOcc [3]	C+L	ResNet-50	256 × 704	47.7	51.7	13.2	57.8	24.3	60.3	64.3	36.2	39.4	52.4	35.8	**50.9**	53.7	84.6	47.5	58.0	61.6	70.7	67.7
ALOcc-2D-mini	C+D	ResNet-50	256 × 704	46.2	50.0	15.7	54.6	36.6	55.7	60.6	34.8	41.0	44.9	39.3	44.5	51.1	83.6	48.5	57.3	60.2	62.7	58.2
ALOcc-2D	C+D	ResNet-50	256 × 704	50.3	53.5	16.5	57.8	**41.6**	57.9	63.8	**37.6**	45.0	52.1	45.8	49.6	54.4	85.3	50.5	59.7	62.3	67.1	62.0
ALOcc-3D	C+D	ResNet-50	256 × 704	**50.6**	**54.5**	**17.0**	59.0	40.9	58.3	64.4	37.2	45.9	52.7	**46.8**	50.5	54.5	**86.3**	**51.5**	**61.7**	**64.8**	69.1	65.1
ALOcc-2D-mini	C+D	Intern-T	256 × 704	48.9	52.1	17.4	56.5	39.2	60.4	62.8	34.7	45.2	45.9	41.1	48.6	54.5	85.3	50.3	59.5	62.3	63.2	58.8
ALOcc-2D	C+D	Intern-T	256 × 704	52.0	54.9	17.4	59.0	41.9	60.8	65.1	38.3	48.9	53.1	46.7	51.3	56.6	86.3	52.8	61.5	63.9	67.7	62.4
ALOcc-3D	C+D	Intern-T	256 × 704	52.4	55.6	18.3	60.1	42.9	61.6	65.5	38.4	48.5	53.5	46.8	51.6	57.5	86.6	52.2	62.1	65.0	69.1	65.1
BEVFormer [11]	C	ResNet-101	900 × 1600	37.2	39.2	5.0	44.9	26.2	**59.7**	55.1	27.9	29.1	34.3	29.6	29.1	50.5	44.4	22.4	21.5	19.5	39.3	31.1
VoxFormer [10]	C	ResNet-101	900 × 1600	-	40.7	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SurroundOcc [28]	C	ResNet-50	900 × 1600	31.2	37.2	9.0	46.3	17.1	46.5	52.0	20.1	21.5	23.5	18.7	31.5	37.6	81.9	41.6	50.8	53.9	42.9	37.2
FastOcc [4]	C	ResNet-101	640 × 1600	34.5	39.2	2.1	43.5	28.0	44.8	52.2	23.0	29.1	29.7	27.0	30.8	38.4	82.0	41.9	51.9	53.7	41.0	35.5
PanoOcc [27]	C	ResNet-101	640 × 1600	37.3	42.1	11.7	50.5	29.6	49.4	55.5	23.3	33.3	30.6	31.0	34.4	42.6	83.3	44.2	54.4	56.0	45.9	40.4
OSP [22]	C	ResNet-101	900 × 1600	37.0	41.2	11.0	49.0	27.7	50.2	56.0	23.0	31.0	30.9	30.3	35.6	41.2	82.1	42.6	51.9	55.1	44.8	38.2
BEVDetOcc [5]	C	Swin-Base	512 × 1408	36.9	42.0	12.2	49.6	25.1	52.0	54.5	27.9	28.0	28.9	27.2	36.4	42.2	82.3	43.3	54.6	57.9	48.6	43.6
COTR [18]	C	Swin-Base	512 × 1408	41.3	46.2	14.9	53.3	35.2	50.8	57.3	**35.4**	34.1	33.5	37.1	39.0	45.0	84.5	48.7	57.6	61.1	51.6	46.7
ALOcc-2D	C	Swin-Base	512 × 1408	44.5	49.3	16.3	56.9	39.2	55.9	61.8	30.4	38.9	38.8	40.3	42.0	49.3	85.8	52.2	60.6	63.6	56.3	49.0
ALOcc-3D	C	Swin-Base	512 × 1408	**46.1**	**50.6**	**17.0**	**58.3**	**39.7**	56.6	**63.2**	33.2	**41.3**	**40.3**	**40.8**	**43.7**	**51.0**	**87.0**	**52.7**	**62.0**	**65.1**	**57.7**	**50.9**
OccFusion [19]	C+L	ResNet-101	900 × 1600	45.3	46.8	11.6	47.8	32.1	57.3	57.5	31.8	40.1	47.3	33.7	45.8	50.3	78.8	37.2	44.4	53.4	63.2	63.2
BEVFusion [17]	C+L	Swin-Base	512 × 1408	48.7	54.0	16.2	61.9	39.3	58.2	62.5	38.1	41.6	46.7	47.7	50.6	52.7	85.7	49.4	60.7	64.3	71.7	70.2
EFFOcc [23]	C+L	Swin-Base	512 × 1408	50.7	54.1	15.7	61.0	36.2	62.2	66.4	38.7	43.9	52.1	42.4	50.3	56.1	84.9	48.0	58.6	62.0	71.3	69.5
FusionOcc [33]	C+L	Swin-Base	512 × 1408	53.1	56.6	17.1	62.6	43.1	63.8	66.2	37.9	49.7	53.7	49.8	53.1	57.5	86.2	49.8	61.6	65.1	73.5	**71.9**
ALOcc-2D	C+D	Swin-Base	512 × 1408	56.8	58.7	17.9	63.0	48.0	**66.1**	70.0	41.1	55.4	60.2	51.0	53.4	59.9	87.6	55.2	63.6	65.8	72.3	67.5
ALOcc-3D	C+D	Swin-Base	512 × 1408	**57.8**	**60.0**	**18.7**	**64.6**	**50.5**	65.5	**70.9**	**42.1**	**56.2**	**61.5**	**52.6**	**54.4**	**61.5**	**88.3**	**55.3**	**64.8**	**67.9**	**74.2**	70.2

Table A.1. **3D semantic occupancy prediction results *w.r.t. training with camera visible mask* on Occ3D, showing per-class IoU, $\text{mIoU}_\text{D}{}^\text{m}$ and mIoU^m.** Input modalities include *Camera (C)*, *Radar (R)*, *LIDAR (L)*, and *Depth (D)*, where Depth represents a sparser signal mapped from LIDAR points. The best results among similar conditions (*i.e.*, comparable image size, backbone, and input modalities) are **bolded**. ALOcc outperforms all competing methods in both $\text{mIoU}_\text{D}{}^\text{m}$ and mIoU^m, and also achieves the highest IoU scores for the majority of classes. Notably, even without a dedicated point cloud backbone, ALOcc achieves SOTA performance in multimodal scenarios.

B.3. Results on Occ3D-Waymo

To validate the generalization capability of our method, we conduct experiments on the large-scale Occ3D-Waymo dataset [24]. This dataset is significantly more extensive than nuScenes, containing 5.6× more training frames. This dataset presents a significant challenge that few methods have reported on at full scale. For this evaluation, we use a ResNet-50 backbone and follow CVT-Occ [30] in setting the input image resolution to 640 × 960, while all other hyperparameters remain consistent with our nuScenes experiments. As presented in Tab. A.3, our method surpasses all competing approaches, underscoring its robust generalization to a different and more challenging scenario.

B.4. Visualization

As shown in Fig. A.1, we provide qualitative visualizations to compare our 3D semantic occupancy predictions against other methods. The results highlight several key advantages of our approach. In the first row, our model demonstrates superior sensitivity by successfully identifying pedestrians, which competing methods overlook. The second row showcases its ability to reconstruct large, complex structures, accurately capturing a dense cluster of trees that are largely omitted by the others. Finally, the third row underscores our method's robustness across different object scales: it precisely predicts both a small traffic cone, which BEVDetOcc-SF fails to render correctly, and a large con-

Method	Backbone	Image Size	$\mathrm{mIoU_D}$	mIoU	IoU	barrier ■	bicycle ■	bus ■	car ■	cons. veh. ■	motor. ■	pedes. ■	tfc. cone ■	trailer ■	truck ■	drv. surf. ■	other flat ■	sidewalk ■	terrain ■	manmade ■	vegetation ■
BEVFormer [11]	ResNet-101	900 × 1600	14.2	16.8	30.5	14.2	6.6	23.5	28.3	8.7	10.8	6.6	4.1	11.2	17.8	37.3	18.0	22.9	22.2	13.8	22.2
TPVFormer [6]	ResNet-101	900 × 1600	14.0	17.1	30.9	16.0	5.3	23.9	27.3	9.8	8.7	7.1	5.2	11.0	19.2	38.9	21.3	24.3	23.2	11.7	20.8
SurroundOcc [28]	ResNet-101	900 × 1600	18.4	20.3	31.5	20.6	11.7	28.1	30.9	10.7	15.1	14.1	12.1	14.4	22.3	37.3	23.7	24.5	22.8	14.9	21.9
GaussianFormer [7]	ResNet-101	900 × 1600	19.1	17.3	29.8	19.5	11.3	26.1	29.8	10.5	13.8	12.6	8.7	12.7	21.6	39.6	23.3	24.5	23.0	9.6	19.1
GaussianFormer2 [8]	ResNet-101	900 × 1600	18.8	20.8	31.7	21.4	13.4	28.5	30.8	10.9	15.8	13.6	10.5	14.0	22.9	40.6	24.4	26.1	24.3	13.8	22.0
BEVDetOcc [5]	ResNet-50	900 × 1600	14.1	17.5	29.2	18.1	2.1	25.5	29.5	11.6	9.5	7.0	4.4	7.3	20.1	40.4	21.3	26.3	23.8	11.5	21.6
ALOcc-2D-mini*	ResNet-50	900 × 1600	19.5	21.5	31.5	21.8	15.7	27.3	30.7	12.7	17.4	15.7	14.0	13.9	22.4	40.0	24.7	26.3	24.4	14.4	22.3
ALOcc-2D*	ResNet-50	900 × 1600	21.5	23.7	34.5	23.5	17.2	28.0	33.0	17.0	19.2	17.1	16.1	15.0	25.1	41.9	26.8	28.3	26.5	18.5	26.5
ALOcc*	ResNet-50	900 × 1600	21.7	24.0	34.7	23.8	17.4	28.0	32.9	17.0	20.2	17.2	**16.9**	15.4	25.5	41.8	26.7	28.3	27.0	18.6	27.0
ALOcc-2D-mini*	Intern-T	900 × 1600	21.0	23.3	33.6	23.4	16.5	29.1	33.0	12.3	20.0	17.4	15.6	15.1	24.5	41.3	26.3	28.2	26.6	17.2	25.4
ALOcc-2D*	Intern-T	900 × 1600	22.6	24.9	35.8	**25.0**	18.7	29.7	34.1	16.1	**22.0**	17.8	16.6	15.8	**26.3**	42.5	27.9	29.4	28.0	20.1	28.4
ALOcc*	Intern-T	900 × 1600	**22.8**	**25.1**	**36.1**	24.8	**19.0**	**30.2**	**34.3**	**17.5**	21.4	**18.2**	**16.9**	**15.9**	**26.3**	**42.9**	**28.0**	**29.6**	**28.2**	**20.2**	**28.8**

Table A.2. **3D semantic occupancy prediction results on SurorundOcc, reporting mIoU, $\mathbf{mIoU_D}$, and per-class IoU.** The metrics are computed over all voxels, a notable difference from the Occ3D benchmark. For a fair comparison against existing methods, we report results without voxel-level temporal fusion (denoted by *). Our approach sets a new SOTA on this benchmark, surpassing all competing methods on every metric.

Method	mIoU	Go	Vehicle	Pedestrian	Sign	Bicyclist	Traffic Light	Pole	Cons. Cone	Bicycle	Building	Vegetation	Tree Trunk	Road	Walkable
BEVFormer-w/o TSA	23.87	**7.50**	34.54	21.07	9.69	**20.96**	11.48	11.48	14.06	14.51	23.14	21.82	8.57	78.45	56.89
BEVFormer [11]	24.58	7.18	36.06	21.00	9.76	20.23	12.61	14.52	14.70	16.06	23.98	22.50	9.39	79.11	57.04
SOLOFusion [21]	24.73	4.97	32.45	18.28	10.33	17.14	8.07	17.83	16.23	19.3	31.49	28.98	16.93	70.95	53.28
BEVFormer-WrapConcat	25.07	6.2	36.17	20.95	9.56	20.58	12.82	16.24	14.31	16.78	25.14	23.56	12.81	79.04	56.83
CVT-Occ [30]	27.37	7.44	**41.0**	23.93	11.92	20.81	12.07	18.03	16.88	**21.37**	29.4	27.42	14.67	**79.12**	59.09
ALOcc-3D	**30.03**	6.51	39.61	**24.14**	**20.84**	20.56	**20.56**	**24.28**	**17.95**	12.22	**35.67**	**37.25**	**22.45**	78.42	**59.91**

Table A.3. **3D semantic occupancy prediction results on Occ3D-Waymo.** All methods use image input. We use an input image size of 640 × 960 and a backbone of ResNet-50 for comparison with other methods, with settings for other approaches detailed in CVT-Occ [30].

struction vehicle missed entirely by FB-Occ. Collectively, these visualizations demonstrate that ALOcc produces more complete and semantically accurate scene representations.

B.5. Experiments on Model Architecture

Regarding Adaptive Lifting. As shown in Tab. A.4, we conducted experiments on the benchmark of *training without mask* to investigate the impact of the adaptive lifting module. The results indicate that converting depth weights to occlusion weights significantly improves the prediction of dynamic objects, while depth denoising further enhances overall performance. Furthermore, we set the value of m to 3 for managing inter-object occlusion, based on the experimental results in Tab. A.5.

As illustrated in Tab. A.6, we analyze the effect of different epochs for depth denoising on the performance. The parameter E in Eq. (5) was set as the number of epochs multiplied by the number of iterations per epoch. Our default model uses an epoch number of 6.

Exp.	Condition	mIoU^m_D	mIoU^m
0	ALOcc-2D-40 (*w/o* SP)	36.0	42.1
1	Exp. 0 *w/o* DD	35.8	41.8
2	Exp. 1 *w/o* D2IO	35.7	41.6
3	Exp. 1 *w/o* D2O	34.9	41.2

Table A.4. **Ablation study on adaptive lifting. SP** denotes the Semantic Prototype-based occupancy head. **DD** represents depth denoising. **D2O** indicates converting depth weights to occlusion weights. **D2IO** refers to converting depth weights to inter-object occlusion.

Effect of Point Sampling Density in Supervision. We conduct an ablation study in Tab. A.7 to analyze the impact of the number of sampled points, K, used during occupancy supervision. Using the 12544 points sampled by default in Mask2Former [2] as a baseline (1×), we experiment with sampling factors of 2× and 4×. Based on the results, we adopt a sampling density of 2× (25088 points) as the default setting for our models, as it provides a favorable balance of performance and efficiency.

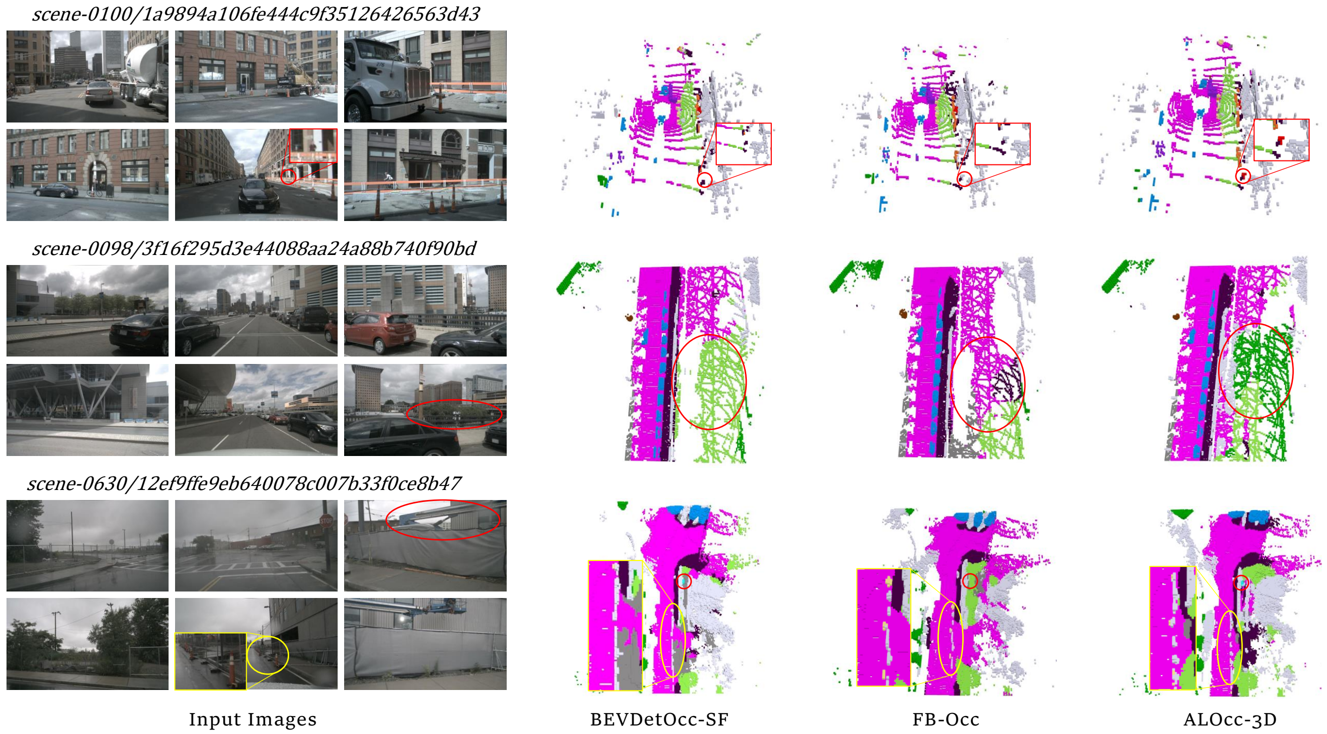
scene-0100/1a9894a106fe444c9f35126426563d43
scene-0098/3f16f295d3e44088aa24a88b740f90bd
scene-0630/12ef9ffe9eb640078c007b33f0ce8b47
Input Images
BEVDetOcc-SF
FB-Occ
ALOcc-3D

Figure A.1. **Qualitative results on Occ3D.** From left to right, the columns represent the input images, visualization results of BEVDetOcc-SF, FB-Occ, and our ALOcc-3D. Key differences are highlighted in red or orange. The corresponding categories for different colors can be found in Tab. A.1.

Num.	mIoU_D^m	mIoU^m
0	35.7	41.6
1	35.9	41.9
3	36.0	42.1
5	35.7	41.8

Table A.5. **Effects of different m values (regarding inter-object occlusion).**

Num.	mIoU_D^m	mIoU^m
2	38.1	44.2
4	38.3	44.4
6	38.5	44.5
8	38.5	44.5

Table A.6. **Effects of depth denoising *w.r.t.* the *number of epochs* of conducting depth denoising.**

Num	mIoU_D^m	mIoU^m
1×	37.8	43.9
2×	38.5	44.5
4×	38.0	44.2

Table A.7. **Effects of point sampling number K.** The notations 1×, 2×, and 4× represent a multiplication factor of 12544.

C. Additional Experimental Details

C.1. Model Details

We primarily provide three models: ALOcc-3D, ALOcc-2D, and ALOcc-2D-mini. The main difference between ALOcc-3D and ALOcc-2D lies in the processing before the volume encoder. ALOcc-2D compresses the height of the volume features into the channel dimension before passing it to the volume encoder, utilizing 2D convolution for feature encoding [31]. Before the prediction head, a simple convolution layer is used to recover the height dimension from the channel dimension, thus avoiding the high computational cost of 3D convolution. For the ResNet-50 and the Intern-T baselines, during feature encoding, ALOcc-3D has feature dimensions of $200 \times 200 \times 16 \times 32$, while ALOcc-2D has feature dimensions of $200 \times 200 \times 80$. ALOcc-2D-mini further simplifies the depth prediction module by switching from stereo depth estimation to single-view depth estimation and using a smaller channel dimension of 40. For the Swin-Base baseline, ALOcc-3D uses a channel dimension of 64, while ALOcc-2D uses a channel dimension of 160.

C.2. Training Details

Our models are initialized using publicly available checkpoints: we adopt the BEVDet [5] checkpoint for ResNet-50 backbones and the GeoMIM [14] checkpoint for Swin-Base backbones. For additional experiments on the Intern-T backbone, we pre-train it ourselves with BEVDet. The models in Tab. 1 and Tab. 2 differ in whether a camera-visible mask was used during training. Models in Tab. 1 use this mask, constraining their objective to mapping observed image content into 3D space. In contrast, models in Tab. 2 are trained without this constraint, tasking them with the more challenging goal of inferring both visible and occluded content. Additionally, our models in Tab. 3 use a ray-visible mask [1] during training, which was generated with ray queries.

D. Data Source

To facilitate data traceability, we additionally document the data sources for the compared methods presented in Tab. 1, Tab. 2, and Tab. 3. The results of BEVFormer and RenderOcc (Tab. 2) were cited from SparseOcc [13], while the results of SurroundOcc were cited from FastOcc [4]. The results of VoxFormer were cited from COTR [18]. The results of OccFormer, TPVFormer, and CTF-Occ were cited from Occ3D [24]. The results of BEVFusion were cited from FusionOcc [33]. The results of OccNet, OccNerf, and RenderOcc (Tab. 3) were cited from LetOccFlow [16]. The results of UniOCC, SurroundSDF, COTR, HyDRa, FastOcc, PanoOcc, EFFOcc, SDGOcc, OccFusion, FusionOcc, SparseOcc, LetOccFlow, CascadeFlow, and F-Occ were cited from their original papers. The results of FlashOcc, Panoptic-FlashOCC, ViewFormer, OPUS, OSP, and BEVDetOcc were evaluated using the official checkpoints [5, 9, 22, 26, 31, 32]. We reproduced the results for FBOcc using the official code. BEVDetOcc-SF was implemented by ourselves. We extended it using SoloFusion [21] to create a long history (16-frame) fusion version. All FPS were measured by ourselves using a single RTX 4090 GPU.

E. Definition of the Occluded Length

We give a mathematical formulation of the occluded length.

Definition 1 (Occluded Length.) *Let $(u, v, d) \in \mathbb{R}^3$ represent the coordinates of the surface point in the camera frustum space, with (u, v) denoting the coordinate of the corresponding pixel and d representing depth. The Occluded length l at this point is defined as the maximal extension within the object's range from the camera's perspective. It satisfies the following conditions:*

1. $\forall \lambda \in (0, 1], \mathrm{CLASS}(u, v, d) = \mathrm{CLASS}(u, v, d + \lambda l)$,
2. *There exists a unique* $l \in \mathbb{R}_+$ *such that:* $\lim_{\epsilon \to 0^+} \mathrm{CLASS}(u, v, d+l) \neq \mathrm{CLASS}(u, v, d+l+\epsilon)\}$,

where $\mathrm{CLASS} : \mathbb{R}^3 \to \mathcal{C}$ *maps frustum space coordinates to the class space* $\mathcal{C}$.

References

[1] Dubing Chen, Wencheng Han, Jin Fang, and Jianbing Shen. Adaocc: Adaptive forward view transformation and flow modeling for 3d occupancy and flow prediction. *arXiv preprint arXiv:2407.01436*, 2024. 5

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3

[3] ZaiPeng Duan, ChenXu Dang, Xuzhong Hu, Pei An, Junfeng Ding, Jie Zhan, YunBiao Xu, and Jie Ma. Sdgocc: Semantic and depth-guided bird's-eye view transformation for 3d multimodal occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6751–6760, 2025. 2

[4] Jiawei Hou, Xiaoyan Li, Wenhao Guan, Gang Zhang, Di Feng, Yuheng Du, Xiangyang Xue, and Jian Pu. Fastocc: Accelerating 3d occupancy prediction by fusing the 2d bird's-eye view and perspective view. In *IEEE International Conference on Robotics and Automation*, 2024. 2, 5

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

[6] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Tri-perspective view for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2302.07817*, 2023. 3

[7] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. In *European Conference on Computer Vision*, pages 376–393. Springer, 2024. 1, 3

[8] Yuanhui Huang, Amonnut Thammatadatrakoon, Wenzhao Zheng, Yunpeng Zhang, Dalong Du, and Jiwen Lu. Gaussianformer-2: Probabilistic gaussian superposition for efficient 3d occupancy prediction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27477–27486, 2025. 1, 3

[9] Jinke Li, Xiao He, Chonghua Zhou, Xiaoqiang Cheng, Yang Wen, and Dan Zhang. Viewformer: Exploring spatiotemporal modeling for multi-view 3d occupancy perception via view-guided transformers. In *Computer Vision–ECCV 2024: 18th European Conference*, 2024. 2, 5

[10] 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

[11] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Qiao Yu, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera

images via spatiotemporal transformers. *arXiv preprint arXiv:2203.17270*, 2022. 2, 3

[12] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. *arXiv preprint arXiv:2307.01492*, 2023. 2

[13] Haisong Liu, Haiguang Wang, Yang Chen, Zetong Yang, Jia Zeng, Li Chen, and Limin Wang. Fully sparse 3d occupancy prediction. In *Computer Vision–ECCV 2024: 18th European Conference*, 2024. 1, 5

[14] Jihao Liu, Tai Wang, Boxiao Liu, Qihang Zhang, Yu Liu, and Hongsheng Li. Geomim: Towards better 3d knowledge transfer via masked image modeling for multi-view 3d understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 17839–17849, 2023. 5

[15] Lizhe Liu, Bohua Wang, Hongwei Xie, Daqi Liu, Li Liu, Zhiqiang Tian, Kuiyuan Yang, and Bing Wang. Surroundsdf: Implicit 3d scene understanding based on signed distance field. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[16] 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. 5

[17] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. *arXiv preprint arXiv:2205.13542*, 2022. 2

[18] 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. 2, 5

[19] Zhenxing Ming, Julie Stephany Berrio, Mao Shan, and Stewart Worrall. Occfusion: Multi-sensor fusion framework for 3d semantic occupancy prediction. *IEEE Transactions on Intelligent Vehicles*, 2024. 2

[20] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 2

[21] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. *arXiv preprint arXiv:2210.02443*, 2022. 3, 5

[22] Yiang Shi, Tianheng Cheng, Qian Zhang, Wenyu Liu, and Xinggang Wang. Occupancy as set of points. In *Computer Vision–ECCV 2024: 18th European Conference*, 2024. 2, 5

[23] Yining Shi, Kun Jiang, Jinyu Miao, Ke Wang, Kangan Qian, Yunlong Wang, Jiusi Li, Tuopu Wen, Mengmeng Yang, Yiliang Xu, et al. Effocc: Learning efficient occupancy networks from minimal labels for autonomous driving. *arXiv preprint arXiv:2406.07042*, 2024. 2

[24] 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. 1, 2, 5

[25] 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

[26] Jiabao Wang, Zhaojiang Liu, Qiang Meng, Liujiang Yan, Ke Wang, Jie Yang, Wei Liu, Qibin Hou, and Mingming Cheng. Opus: occupancy prediction using a sparse set. In *Advances in Neural Information Processing Systems*, 2024. 2, 5

[27] 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*, 2024. 2

[28] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. *arXiv preprint arXiv:2303.09551*, 2023. 1, 2, 3

[29] Philipp Wolters, Johannes Gilg, Torben Teepe, Fabian Herzog, Anouar Laouichi, Martin Hofmann, and Gerhard Rigoll. Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception. *arXiv preprint arXiv:2403.07746*, 2024. 2

[30] Zhangchen Ye, Tao Jiang, Chenfeng Xu, Yiming Li, and Hang Zhao. Cvt-occ: Cost volume temporal fusion for 3d occupancy prediction. In *European Conference on Computer Vision*, pages 381–397. Springer, 2024. 2, 3

[31] 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, 4, 5

[32] Zichen Yu, Changyong Shu, Qianpu Sun, Junjie Linghu, Xiaobao Wei, Jiangyong Yu, Zongdai Liu, Dawei Yang, Hui Li, and Yan Chen. Panoptic-flashocc: An efficient baseline to marry semantic occupancy with panoptic via instance center. *arXiv preprint arXiv:2406.10527*, 2024. 5

[33] Shuo Zhang, Yupeng Zhai, Jilin Mei, and Yu Hu. Fusionocc: Multi-modal fusion for 3d occupancy prediction. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 787–796, 2024. 2, 5