SurfelOcc: Self-supervised Occupancy Prediction via 2D Surfel Splatting

Supplementary Material

1. Additional Experiments

1.1. Quantitative Results upon Occupancy Categories

As shown in Tab. 1, we provide detailed per-class mIoU results of SurfelOcc on the Occ3D-nuScenes [12] dataset. It can be observed that the performance of SurfelOcc is highly competitive in dynamic foreground object categories, including bus, car, pedestrian, trailer, truck, etc., without feature-level supervision that is utilized in GaussTR [8]. In addition, SurfelOcc can preserve accurate background scene modeling, remaining high performances on static categories. This further representing the excellent overall performance of SurfelOcc, demonstrating the effectiveness of our surfel-based scene representation.

1.2. Quantitative Results on Surround-View Depth Estimation

As a depth-aware occupancy prediction method, we also represent the ability of SurfelOcc to perform depth prediction on nuScenes [1] dataset and compare with other state-of-the-art weakly-supervised surround-view depth perception methods and depth-aware occupancy prediction methods. To evaluate depth prediction results, we utilize absolute relative error (Abs Rel), square relative error (Sq Rel), root mean square error (RMSE), and its log version (RMSE log) as followings:

$$AbsRel = \frac{1}{|M|}\sum_{d\in M}\frac{||d-d^*||}{d^*}, \tag{1}$$

$$SqRel = \frac{1}{|M|}\sum_{d\in M}\frac{||d-d^*||^2}{d^*}, \tag{2}$$

$$RMSE = \sqrt{\frac{1}{|M|}\sum_{d\in M}||d-d^*||^2}, \tag{3}$$

$$RMSE\ log = \sqrt{\frac{1}{|M|}\sum_{d\in M}||\log d-\log d^*||^2}, \tag{4}$$

where M is the valid pixel, and d, d^* denotes the ground truth depth and the predicted depth respectively. Also we deploy threshold accuracy metrics $\delta < t$ with threshold t, which can be represented as the percentage of d subjected to $max(\frac{d}{d^*}, \frac{d^*}{d}) = \delta < t$. Note that for Abs Rel, Sq Rel, RMSE, RMSE log, the lower value indicates the better performance. And for $\delta < t$ metrics, the higher value indicates the better performance.

In Tab. 2, we show the quantitative results for surround-view depth prediction. It can be noticed that SurfelOcc achieves the top performance across all metrics compared to state-of-the-art weakly-supervised depth-aware methods. Notably, compared to previous best-performed GaussianOcc [3], our method can improve by about 27.4% and 19.8% in terms of absolute relative error and root mean square error, respectively. Similar trends can be observed on other metrics, where our method surpasses other methods with large margins. Thus, it can be deduced that our surfel-based scene representation paradigm demonstrates superiority in perceiving spatial relationships and reconstructing scene representations, which contributes to achieving accurate occupancy prediction results.

2. Additional Implementation Details

2.1. 2D Surfel Rasterization

To conduct self-supervision, we perform differentiable rasterization operation[5], projecting 2D surfels onto the image plane for depth and semantic maps. As mentioned in the main paper, we formulate a 2D-to-2D transformation from image plane to local surfel plane, which can describe a splatting or intersecting process to obtain 2D semantic or depth maps from surfel primitives. Given 4×4 combined transformation matrix W from the world space to image plane, these processes are presented as:

$$\begin{aligned} p_{sem} &= \sum_j Splat(o_j, \mathbf{s}_j, \mathbf{H}_j|W), \\ p_{depth} &= \sum_j Intersect(o_j, \mathbf{H}_j|W), \end{aligned} \tag{5}$$

where p_{sem}, p_{depth} refer to the pixel value in the semantic map and depth map, respectively. The surfel attributes include opacity o, semantics $\mathbf{s}$, and 4×4 homogeneous transformation matrix $\mathbf{H}$, and j refers to the number of surfels intersected with the ray from the camera center to the pixel. Given an image coordinate $\mathbf{x} = (x, y)$ and the combined 4×4 transformation matrix W from world space to image plane, we can get the corresponding coordinates $(u(\mathbf{x}), v(\mathbf{x}))$ on surfel local plane by the following equations:

$$\begin{aligned} \mathbf{h}_u &= (W\mathbf{H})^T \cdot (-1, 0, 0, x)^T \\ \mathbf{h}_v &= (W\mathbf{H})^T \cdot (0, -1, 0, y)^T \\ u(\mathbf{x}) = &\frac{\mathbf{h}_u^2\mathbf{h}_v^4 - \mathbf{h}_u^4\mathbf{h}_v^2}{\mathbf{h}_u^1\mathbf{h}_v^2 - \mathbf{h}_u^2\mathbf{h}_v^1}, v(\mathbf{x}) = \frac{\mathbf{h}_u^4\mathbf{h}_v^1 - \mathbf{h}_u^1\mathbf{h}_v^4}{\mathbf{h}_u^1\mathbf{h}_v^2 - \mathbf{h}_u^2\mathbf{h}_v^1}, \end{aligned} \tag{6}$$

Methods	GT Free.	mIoU	mIoU*	barrier	bicycle	bus	car	const. veh.	motorcycle	pedestrian	traffic cone	trailer	truck	drive. surf.	sidewalk	terrain	manmade	vegetation
BEVDet [6]	×	20.03	19.38	30.31	0.23	32.26	34.47	12.97	10.34	10.36	6.26	8.93	23.65	52.27	26.06	22.31	15.04	15.10
BEVFormer [10]	×	24.64	23.67	38.79	9.98	34.41	41.09	13.24	16.50	18.15	17.83	18.66	27.70	48.95	29.08	25.38	15.41	14.46
OccFormer [16]	×	22.39	21.93	30.29	12.32	34.40	39.17	14.44	16.45	17.22	9.27	13.90	26.36	50.99	34.66	22.73	6.76	6.97
SimpleOcc[2]	✓	7.99	7.05	0.67	1.18	3.21	7.63	1.02	0.26	1.80	0.26	1.07	2.81	40.44	18.30	17.01	13.42	10.84
SelfOcc[7]	✓	10.54	9.30	0.15	0.66	5.46	12.54	0.00	0.80	2.10	0.00	0.00	8.25	55.49	26.30	26.54	14.22	5.60
OccNeRF[15]	✓	10.81	9.54	0.83	0.82	5.13	12.49	3.50	0.23	3.10	1.84	0.52	3.90	52.62	20.81	24.75	18.45	13.19
GaussianOcc[3]	✓	11.26	9.94	1.79	5.82	14.58	13.55	1.30	2.82	7.95	9.76	0.56	9.61	44.59	20.10	17.58	8.61	10.29
GaussTR$_{(\text{FeatUp})}$[8]	✓	13.26	11.70	2.09	5.22	14.07	20.43	5.70	7.08	5.12	3.93	0.92	13.36	39.44	15.68	22.89	21.17	21.87
GaussTR$_{(\text{Talk2DINO})}$[8]	✓	13.91	12.27	6.50	8.54	21.77	24.27	6.26	15.48	7.94	1.86	6.10	17.16	36.98	17.21	7.16	21.18	9.99
SurfelOcc	✓	**16.70**	**15.54**	6.16	6.65	24.69	30.54	11.46	9.27	8.45	12.08	1.83	18.25	48.14	10.38	15.70	20.55	19.24

Table 1. A comprehensive quantitative comparison of occupancy prediction methods on Occ3D-nuScenes [12] dataset. Note that we ignore "other" and "other flat" classes when calculating "mIoU" and present original results as "mIoU*".

Methods	Occ.	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	$\delta < 1.25$ ↑	$\delta < 1.25^2$ ↑	$\delta < 1.25^3$ ↑
FSM [4]	×	0.319	7.534	7.860	0.362	0.716	0.874	0.931
SurroundDepth [13]	×	0.280	4.401	7.467	0.364	0.661	0.844	0.917
SA-FSM [14]	×	0.272	4.706	7.391	0.355	0.689	0.868	0.929
VFF [9]	×	0.289	5.718	7.551	0.348	0.709	0.876	0.932
R3D3 [11]	×	0.253	4.759	7.150	-	0.729	-	-
M^{2}Depth [17]	×	0.259	4.599	6.898	0.332	0.734	0.871	0.928
SimpleOcc [2]	✓	0.224	3.383	7.165	0.333	0.753	0.877	0.930
OccNeRF [15]	✓	0.202	2.883	6.697	0.319	0.768	0.882	0.931
SelfOcc [7]	✓	0.215	2.743	6.706	0.316	0.753	0.875	0.932
GaussianOcc [3]	✓	0.197	1.846	6.733	0.312	0.746	0.873	0.931
SurfelOcc	✓	**0.144**	**1.194**	**5.402**	**0.268**	**0.821**	**0.921**	**0.955**

Table 2. Quantitative comparison of state-of-the-art weakly supervised depth prediction methods on nuScenes [1] dataset. Note that we apply **bold** format to the best value of each metric. The second column Occ. refers to whether the methods can be adopted to predict occupancy. It can observed that our method achieves the highest across all metrics.

where $\mathbf{h}_u^i$, $\mathbf{h}_v^i$ are the i-th parameter of $\mathbf{h}_u$, $\mathbf{h}_v$.

2.2. Splatting for Semantic Map

For semantic map rendering, we need to perform semantic surfel splatting, which can be described as a general 2D-to-2D mapping in homogeneous coordinates:

$$\begin{aligned} p_{sem} &= \sum_j Splat(o_j, \mathbf{s}_j, \mathbf{H}_j | W) \\ &= \sum_j \mathbf{s}_j \alpha_j \hat{\mathcal{G}}(\mathbf{u}(\mathbf{x})) \prod_{k=1}^{j-1} (1 - \alpha_k \hat{\mathcal{G}}(\mathbf{u}(\mathbf{x}))), \end{aligned} \tag{7}$$

where $\mathbf{u}(\mathbf{x})$ denotes the coordinates $\mathbf{u}$ in local tangential plane of the surfel from the coordinates $\mathbf{x}$ in image plane and α_j denotes the alpha values. $\hat{\mathcal{G}}(x) = max\{\mathcal{G}(\mathbf{u}(\mathbf{x})), \mathcal{G}(\frac{\mathbf{x}-\mathbf{p}}{\sigma}))\}$ denotes 2D Gaussian distribution $\mathcal{G}(\mathbf{x}) = exp(-\frac{\mathbf{x}\cdot\mathbf{x}^T}{2})$ with a low-pass filter in object space of surfel center $\mathbf{p}$ and $\sigma = \sqrt{2}/2$.

2.3. Intersecting for Depth Map

We can calculate the depth z at $\mathbf{x} = (x, y)$ by the following equation to acquire the depth map:

$$(xz, yz, z, z)^T = WP(u, v) = W\mathbf{H}(u, v, 1, 1)^T. \tag{8}$$

The final depth value p_{depth} can be represented as:

$$\begin{aligned} p_{depth} &= \sum_j Intersect(o_j, \mathbf{H}_j | W) = \frac{\sum_j \omega_j z_j}{\sum_j \omega_j + \epsilon}, \\ \omega_j &= \alpha_j \hat{\mathcal{G}}(\mathbf{u}(\mathbf{x})) \prod_{k=1}^{j-1} (1 - \alpha_k \hat{\mathcal{G}}(\mathbf{u}(\mathbf{x}))) \end{aligned} \tag{9}$$

where ϵ refers to a negligible positive constant to avoid zero division.

2.4. Occupancy Grids Acquisition

Referring to previous works [3, 15], we consider unbounded scene for occupancy grid acquisition to preserve high reso-

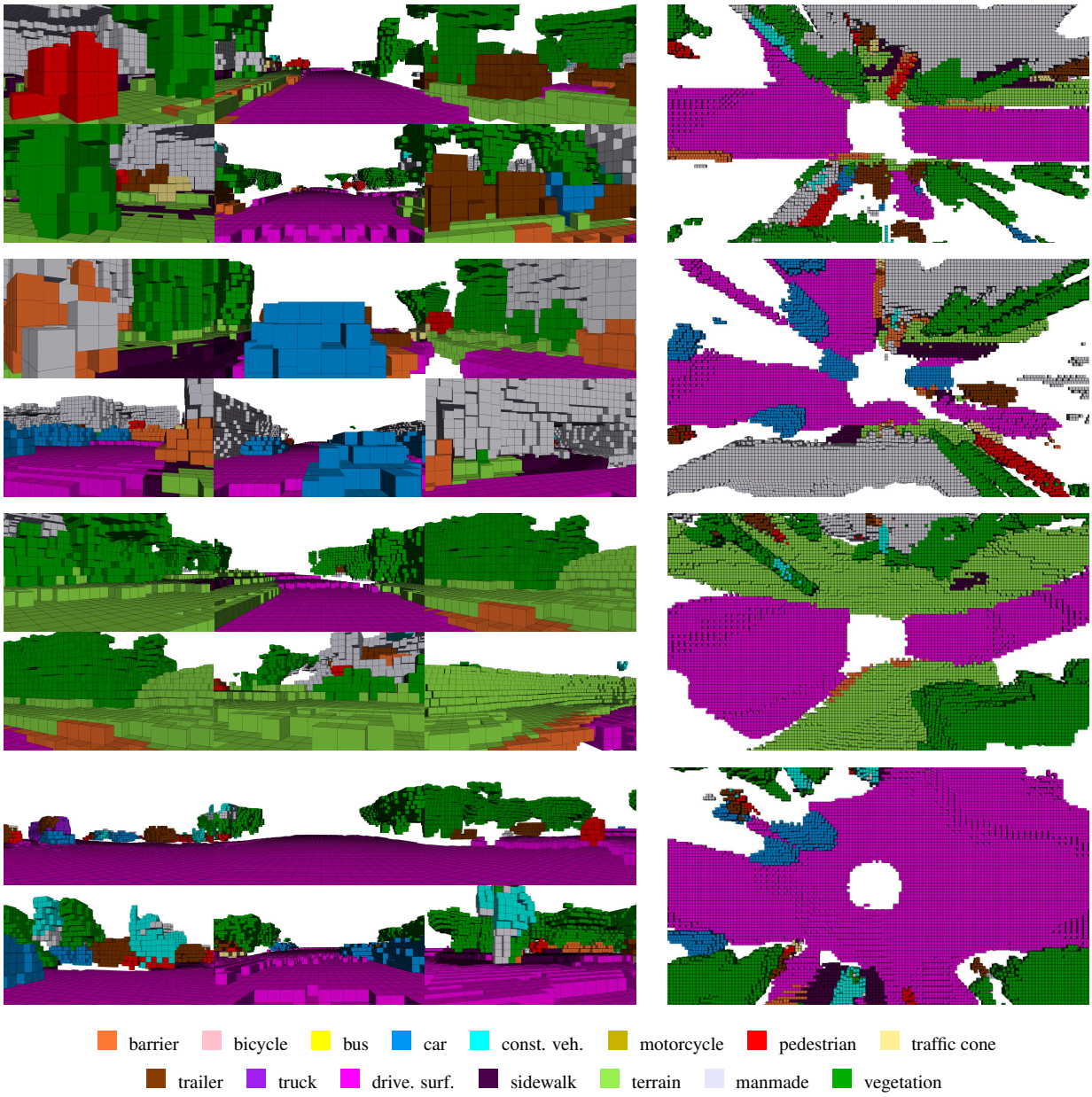

Figure 1. Additional visualization results of SurfelOcc on NuScenes [1] datasets. We provide each camera's view on the left and bird's-eye view on the right.

lution for the inside space [-40m, -40m, -1m, 40m, 40m, 5.4m] that covers the most of regions of interest (RoI) and contract outside space to reduce memory consumption. In specific, we apply a transformation function f with adjustable RoI and contraction threshold to parameterize the coordinates $r = (x, y, z)$ of each voxel grid, which can be formulated as:

$$f(r) = \begin{cases} \alpha \cdot r', & |r'| \leq 1 \\ \frac{r'}{|r'|} \cdot (1 - \frac{a}{|r'|+b}), & |r'| > 1 \end{cases}, \tag{10}$$

where $\alpha \in [0, 1]$ refers to the proportion of RoI in the parameterized space and $r' = r/r_b$ denotes the normalized coordinates with pre-defined inside region bound r_b. a, b are set to $\frac{(1-\alpha)^2}{\alpha}$, $\frac{1-2\alpha}{\alpha}$, respectively, to maintain the continuity of f and its first derivative f' at $r = r_b$. We first uniformly sample points $\mathbf{x}_P$ in the parameterized coordinate system and map them back to the ego coordinate system to acquire occupancy grid points $\mathbf{x}_{OG}$:

$$\mathbf{x}_{OG} = f^{-1}(\mathbf{x}_P). \tag{11}$$

Note that we set $\alpha = 2/3$ to ensure the central $200 \times 200 \times 16$ occupancy grids are evenly distributed in the inside space.

2.5. Surfel Normal Reorientation

We perform a normal reorientation step before the Surfel2Voxel module to ensure that all surfel normals consistently point towards the ego center. Specifically, for each surfel centered at $\mathbf{p}_k = (p_x^k, p_y^k, p_z^k)$, we construct a reference vector $\mathbf{v}_{ref} = (-p_x^k, -p_y^k, 0)$ pointing from the surfel center to its projection on ego-centric pillar $(0, 0, p_z^k)$. The alignment between the normal vector $\mathbf{n}$ and the reference vector is measured by their inner product $n_{proj} = \mathbf{n} \cdot \mathbf{v}_{ref}$. The reoriented normal $\hat{\mathbf{n}}$ is defined as:

$$\hat{\mathbf{n}} = \begin{cases} \mathbf{n}, & n_{proj} \geq 0 \\ -\mathbf{n}, & n_{proj} < 0 \end{cases} . \tag{12}$$

Correspondingly, for the rotation angles $\mathbf{r} = (\theta, \phi)$, the reoriented angles $\hat{\mathbf{r}} = (\hat{\theta}, \hat{\phi})$ are given by:

$$\hat{\mathbf{r}} = (\hat{\theta}, \hat{\phi}) = \begin{cases} (\theta, \phi), & n_{proj} \geq 0 \\ (\theta + \pi, \pi - \phi), & n_{proj} < 0 \end{cases} . \tag{13}$$

The reoriented parameters are then used to compute the modified transformation matrix $\mathbf{H}'$ for the Surfel2Voxel module.

3. Additional Visualization Results

We provide additional visualization results on the nuScenes dataset [1] from multiple views in Fig. 1. Our method not only captures accurate geometry and semantics in the surrounding views but also yields high-fidelity occupancy predictions in the bird's-eye-view (BEV) representation. Notably, SurfelOcc effectively mitigates the trailing artifact, particularly for moving objects. These demonstrates the robustness of our method in modeling dynamic traffic agents under 2D multi-view supervision.

References

[1] 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. *arXiv preprint arXiv:1903.11027*, 2019. 1, 2, 3, 4

[2] Wanshui Gan, Ningkai Mo, Hongbin Xu, and Naoto Yokoya. A simple attempt for 3d occupancy estimation in autonomous driving. *CoRR*, 2023. 2

[3] Wanshui Gan, Fang Liu, Hongbin Xu, Ningkai Mo, and Naoto Yokoya. Gaussianocc: Fully self-supervised and efficient 3d occupancy estimation with gaussian splatting. *arXiv preprint arXiv:2408.11447*, 2024. 1, 2

[4] Vitor Guizilini, Igor Vasiljevic, Rares Ambrus, Greg Shakhnarovich, and Adrien Gaidon. Full surround monodepth from multiple cameras. *IEEE Robotics and Automation Letters*, 7(2):5397–5404, 2022. 2

[5] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024. 1

[6] 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

[7] 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

[8] 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. 1, 2

[9] Jung-Hee Kim, Junhwa Hur, Tien Phuoc Nguyen, and Seong-Gyun Jeong. Self-supervised surround-view depth estimation with volumetric feature fusion. *Advances in Neural Information Processing Systems*, 35:4032–4045, 2022. 2

[10] 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 lidar-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(03):2020–2036, 2025. 2

[11] Aron Schmied, Tobias Fischer, Martin Danelljan, Marc Pollefeys, and Fisher Yu. R3d3: Dense 3d reconstruction of dynamic scenes from multiple cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3216–3226, 2023. 2

[12] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. *arXiv preprint arXiv:2304.14365*, 2023. 1, 2

[13] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Yongming Rao, Guan Huang, Jiwen Lu, and Jie Zhou. Surrounddepth: Entangling surrounding views for self-supervised multi-camera depth estimation. In *Conference on robot learning*, pages 539–549. PMLR, 2023. 2

[14] Yuchen Yang, Xinyi Wang, Dong Li, Lu Tian, Ashish Sirasao, and Xun Yang. Towards scale-aware full surround monodepth with transformers. *arXiv preprint arXiv:2407.10406*, 2024. 2

[15] 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

[16] Yunpeng Zhang, Zheng Zhu, and Dalong Du. Occformer: Dual-path transformer for vision-based 3d semantic occupancy prediction. In *Proceedings of the IEEE/CVF Inter-*

national Conference on Computer Vision, pages 9433–9443, 2023. 2

[17] Yingshuang Zou, Yikang Ding, Xi Qiu, Haoqian Wang, and Haotian Zhang. M 2 depth: Self-supervised two-frame m ulti-camera m etric depth estimation. In *European Conference on Computer Vision*, pages 269–285. Springer, 2024. 2