GS-Occ3D: Scaling Vision-only Occupancy Reconstruction with Gaussian Splatting

Supplementary Material

1. Prelimnary

1.1. 3DGS

3DGS [5] represents a 3D scene by a set of Gaussian Primitives, each defined as:

$$\mathbf{G}(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right), \quad (1)$$

where $x \in \mathbb{R}^{3\times1}$ is Gaussian's 3D position in the scene, $\boldsymbol{\mu} \in \mathbb{R}^{3\times1}$ is the mean vector, and $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$ is the covariance matrix. To ensure positive semi-definiteness, $\boldsymbol{\Sigma}$ is parameterized as $\boldsymbol{\Sigma} = \mathbf{RSS}^T\mathbf{R}^T$, where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix and $\mathbf{S} \in \mathbb{R}^{3\times3}$ is a scaling matrix.

To render an image, 3D Gaussians are projected onto a 2D plane, sorted by depth, and alpha-blended to compute pixel colors:

$$\mathbf{C} = \sum_{i=1}^{n} T_i\alpha_i\mathbf{c}i, \quad T_i = \prod j = 1^{i-1}\left(1-\alpha_j\right), \quad (2)$$

where n is the number of contributing 2D Gaussians, T_i is the transmission factor, and $\mathbf{c}_i$ represents the spherical harmonics-based color of the i-th Gaussian.

1.2. Scaffold-GS

Scaffold-GS [6] represents 3D scenes using anchor points, decoding Gaussian properties (R, S, c, α) from the anchor feature $\mathcal{F} \in \mathbb{R}^d$ via multiple MLPs. The relative positions of Gaussians with respect to the parent anchor center are embedded in $\mathcal{F}$, allowing for implicit encoding of Gaussian components while utilizing the MLP's fitting capability. Gaussians are instantiated dynamically each iteration and removed after updates, optimizing memory efficiency.

2. Loss Function

The overall loss is formulated as a weighted sum of five components: RGB loss, geometry loss, object loss, ground loss, and sky loss. It can be written as:

$$L = L_{rgb} + \lambda_{geo}L_{geo} + \lambda_{obj}L_{obj} + \lambda_{road}L_{road} + \lambda_{sky}L_{sky}, \quad (3)$$

where λ_{geo}, λ_{obj}, λ_{road}, and λ_{sky} denote the corresponding weights.

As in 3DGS [5], we apply a combination of L1 loss and D-SSIM loss to supervise the RGB reconstruction, forming the RGB loss L_{rgb}. For the geometry loss L_{geo}, we introduce a surfel regularization term to flatten Gaussians into surfels, along with geometry constraints from 2DGS and GOF, which encourage the surfels to better conform to the underlying scene geometry. To model dynamic objects, we apply an entropy-based object loss L_{obj} following [9], encouraging a clearer decoupling between foreground and background within the object opacity map. For the road surface, we introduce an smoothness loss L_{road} [2] to ensure the road remains flat and even. For the sky region, we apply a binary cross-entropy loss L_{sky} [9] to the rendered opacity.

The detailed geometry loss L_{geo} is composed of three terms: the surfel regularization loss, depth distortion loss, and depth-normal consistency loss, formulated as:

$$L_{geo} = \lambda_s L_s + \lambda_d L_d + \lambda_n L_n, \quad (4)$$

where λ_s, λ_d, and λ_n are the corresponding weights for each term.

The surfel regularization term L_s is defined as:

$$L_s = \lambda_f L_f + \lambda_r L_r, \quad (5)$$

where L_f is the flatten loss from Neusg [1], and L_r is a ratio loss. L_f is defined as:

$$L_f = \| \min(s_1, s_2, s_3) \|_1, \quad (6)$$

where s_1, s_2, and s_3 denote the scale factors of the Gaussian, and minimizing the smallest scale encourages surfel flattening.

To mitigate the needle-like artifacts, we apply a ratio loss ensures the Gaussian approximates a circular shape:

$$L_r = (s_1/s_2 + s_2/s_1 - 2), \quad (7)$$

where s_1 is the longest scaling factor, and s_2 is the second longest.

The depth distortion loss L_d and depth-normal consistency loss L_n are defined as:

$$L_d = \sum_{i,j} \omega_i\omega_j \left|d_i - d_j\right|, \quad (8)$$

$$L_n = \sum_i \omega_i \left(1 - \mathbf{n}_i^{\mathrm{T}}\mathbf{N}\right). \quad (9)$$

The object loss L_{obj} is an entropy loss applied to the opacity map of the decomposed dynamic objects $\mathbf{O}_{\text{obj}}$. This term encourages a clearer separation between foreground and background:

$$\mathcal{L}_{\text{reg}} = -\sum(\mathbf{O}_{\text{obj}} \log \mathbf{O}_{\text{obj}} + (1 - \mathbf{O}_{\text{obj}}) \log(1 - \mathbf{O}_{\text{obj}})). \tag{10}$$

For the road surface, we introduce an elevation smoothness term to ensure the surface remains flat and even.

$$L_{road} = \frac{1}{K} \sum_{i=1}^{N} \sum_{j \in N(i)} \|z_i - z_j\|_2^2, \tag{11}$$

where z represents the z-coordinate of each road surfel, and $N(i)$ denotes its K nearest neighbors.

Finally, to constrain the sky region, we apply a binary cross-entropy loss between the rendered opacity $\mathbf{O}_g$ and the predicted sky mask $\mathbf{M}_{\text{sky}}$:

$$L_{sky} = -\sum((1 - \mathbf{M}_{\text{sky}}) \log \mathbf{O}_g + \mathbf{M}_{\text{sky}} \log(1 - \mathbf{O}_g)). \tag{12}$$

During training we set the weight of losses as follows: $\lambda_{obj} = 0.1$, $\lambda_{road} = 0.003$, $\lambda_{sky} = 0.05$, $\lambda_f = 100.0$, $\lambda_r = 1.0$, $\lambda_d = 100.0$, and $\lambda_n = 0.01$.

3. Detailed Experimental Setup

3.1. Geometry Reconstruction Experimental Setup

Implementation Details. To ensure fair comparison, we align the training and densification schedules across all methods. Training runs for 40,000 iterations, with densification concluding at 30,000 iterations. For our method, the voxel size of the intermediate anchor grid is set to 0.02. All experiments are conducted on a single NVIDIA A100 80GB GPU.

Metrics. We evaluate reconstruction quality across geometry, rendering, and efficiency. For geometry, we follow StreetSurf [3] to use Chamfer Distance (CD) between the reconstructed and original LiDAR point clouds $\widehat{G}, G$:

$$\text{CD}(\widehat{G}, G) = \frac{1}{|\widehat{G}|} \sum_{\mathbf{x} \in \widehat{G}} \min_{\mathbf{y} \in G} \|\mathbf{x} - \mathbf{y}\|_2^2 + \frac{1}{|G|} \sum_{\mathbf{y} \in G} \min_{\mathbf{x} \in \widehat{G}} \|\mathbf{y} - \mathbf{x}\|_2^2 \tag{13}$$

For rendering quality, we report peak signal-to-noise ratio (PSNR). For efficiency, we record storage requirements, GPU memory usage, and training time.

3.2. 3D Occupancy Prediction Experimental Setup

Datasets. We conduct experiments on the Occ3D-Waymo and Occ3D-nuScenes dataset. The spatial range is set to $[-40\text{ m},\ 40\text{ m}]$ for both x and y axes, and $[-1\text{ m},\ 5.4\text{ m}]$ for the z axis. The voxel grid size is $(0.4\text{ m},\ 0.4\text{ m},\ 0.4\text{ m})$, resulting in a resolution of $(200 \times 200 \times 16)$ for (H, W, Z). Occ3D-Waymo is one of the most diverse and comprehensively labeled open-source 3D occupancy datasets.

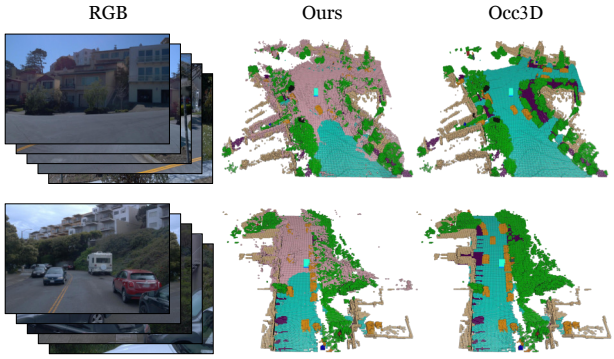

Figure 1. **More Qualitative Results of Our Curated Labels.** Within the camera's field of view, our geometry is comparable to Occ3D's, providing reliable, prior-free supervision for occupancy model training. The metrics are slightly degraded because the lack of a rear view prevents reconstruction behind the initial frame. Pink indicates the binary labels, while other colors represent the Occ3D semantic labels.

Method	Variant	Geometry CD ↓	Rendering PSNR ↑
2DGS [4]	w/ Ground Gaussians	**1.23**	**25.60**
	w/o Ground Gaussians	1.32	23.71
GVKF [7]	w/ Ground Gaussians	**0.87**	**26.22**
	w/o Ground Gaussians	0.91	25.89

Table 1. **Ablation for Ground Gaussians on the Waymo Static-32 Split.** We show the effectiveness of our Ground Gaussians.

Implementation Details. We train SOTA occupancy model CVT-Occ [10] for 8 epochs using AdamW with a cosine annealing learning rate schedule starting from 4×10^{-4}. Training is conducted on 8 NVIDIA A100 GPUs with a batch size of 1 per GPU. Input images are resized to 960 × 640 pixels.

4. Additional Visualization Results

Fig. 1 visualizes our curated labels on the Waymo. Our approach achieves geometry on par with Occ3D within the camera's field of view, ensuring accurate and consistent supervision for occupancy model training without requiring geometric priors. However, the inherent limitation of the camera's rear visibility prevents the effective reconstruction of labels behind the initial frame of the scene. As a result, regions outside the camera's field of view receive insufficient supervision, leading to a slight decline in evaluation metrics.

5. Additional Ablation Results

We also evaluate the effectiveness of our ground gaussians on 2DGS [4] and GVKF [7] using the Waymo Static-32 split [8], averaged across all scenes. As shown in Tab. 1, incorpo-

rating ground gaussians significantly enhances both visual and geometry reconstruction in both methods, demonstrating its effectiveness.

6. Detailed Experiment Results

Tab. 2 and Tab. 3 show detailed experimental results for the different camera setup on the Waymo Static-32 Split. We find that: (1) Unlike other methods that can degrade with more views, our approach benefits from a 5-camera input. (2) Directly reconstructing the point cloud is a more suitable and scalable approach.

References

[1] Hanlin Chen, Chen Li, and Gim Hee Lee. Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance. *arXiv preprint arXiv:2312.00846*, 2023. 1

[2] Zhiheng Feng, Wenhua Wu, Tianchen Deng, and Hesheng Wang. Rogs: Large scale road surface reconstruction with meshgrid gaussian. *arXiv preprint arXiv:2405.14342*, 2024. 1

[3] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views. *arXiv preprint arXiv:2306.04988*, 2023. 2

[4] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. *arXiv preprint arXiv:2403.17888*, 2024. 2

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[6] Tao Lu, Mulin Yu, Linning Xu, Yuanbo Xiangli, Limin Wang, Dahua Lin, and Bo Dai. Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20654–20664, 2024. 1

[7] Gaochao Song, Chong Cheng, and Hao Wang. Gvkf: Gaussian voxel kernel functions for highly efficient surface reconstruction in open scenes. *Advances in Neural Information Processing Systems*, 37:104792–104815, 2025. 2

[8] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2446–2454, 2020. 2

[9] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 1

[10] 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

Sequence	CD-Point Cloud ↓				CD-Mesh ↓			PSNR ↑			
	Ours	GVKF	2DGS	PGSR	GVKF	2DGS	PGSR	Ours	GVKF	2DGS	PGSR
003	**0.51**	0.84	0.73	5.14	**1.53**	1.77	5.41	**27.19**	25.78	22.89	19.74
019	**0.48**	0.84	1.03	4.89	**1.02**	1.79	5.74	**27.07**	24.91	23.40	16.81
036	**0.45**	0.87	1.36	3.00	**1.45**	2.01	3.32	27.19	**27.22**	22.93	22.15
069	**0.64**	0.97	1.78	4.84	**1.26**	2.91	5.17	27.16	**27.62**	25.18	18.04
081	**0.49**	0.90	1.08	3.74	**2.16**	3.76	5.72	**28.17**	25.43	22.56	19.45
126	**0.39**	0.56	0.83	1.87	**1.01**	1.86	2.14	**29.23**	28.25	28.30	24.99
139	**0.62**	0.80	1.14	2.89	**1.25**	2.01	3.77	25.23	**25.64**	23.11	19.27
140	**0.54**	0.72	1.37	2.73	**1.08**	1.83	3.49	**25.63**	25.10	23.11	18.23
146	**0.70**	**0.70**	1.37	2.76	**0.97**	2.41	4.70	**26.16**	26.00	25.04	16.96
148	**0.43**	0.70	1.25	4.73	**1.10**	1.87	6.92	**27.55**	26.02	24.20	19.99
157	**0.45**	0.73	0.55	1.57	**1.05**	1.19	1.82	**28.26**	26.19	23.81	19.78
181	**0.47**	0.62	0.90	3.98	**0.81**	1.26	5.21	**26.79**	25.77	23.61	19.92
200	**0.81**	1.05	1.27	3.00	**1.18**	2.18	2.72	26.62	**29.58**	27.53	18.07
204	**0.53**	0.86	1.25	4.48	**1.27**	1.87	6.12	**27.44**	26.73	23.09	20.81
226	**0.42**	0.52	0.64	4.61	**1.20**	1.63	5.23	**25.84**	22.03	19.43	18.56
232	**0.40**	0.87	1.01	4.40	**1.06**	1.25	5.96	**26.93**	25.45	23.06	20.90
237	**0.95**	1.41	1.97	4.68	**1.42**	2.79	5.69	25.14	**27.20**	26.12	18.82
241	**0.47**	0.88	1.38	4.37	**1.26**	2.47	4.44	**28.48**	26.79	23.50	19.16
245	**0.66**	0.92	1.00	3.06	**1.04**	2.03	4.12	**26.87**	25.50	23.34	16.38
246	**0.60**	0.98	1.45	3.87	**1.46**	2.44	4.32	**27.34**	25.95	22.39	21.09
271	**0.59**	0.87	0.80	2.88	**1.25**	1.43	3.54	**26.41**	24.60	23.36	16.95
297	**0.62**	0.87	3.47	3.05	**1.48**	3.57	3.30	**27.53**	23.54	21.51	17.52
302	0.67	**0.64**	1.48	3.47	**0.94**	2.14	3.83	**25.15**	23.69	22.35	18.20
312	0.61	**0.59**	1.20	3.53	**0.95**	2.80	4.05	**24.57**	23.98	21.89	15.30
314	**0.51**	0.90	0.76	3.47	**1.05**	1.47	3.66	**27.56**	24.70	23.01	16.28
362	0.75	**0.69**	1.40	3.32	**1.08**	1.92	4.77	**26.54**	26.20	25.07	18.37
482	**0.50**	1.08	1.17	3.86	**2.22**	2.37	4.39	**28.30**	24.97	22.04	20.06
495	**0.59**	0.76	1.50	3.49	**0.86**	2.24	4.68	29.46	**29.69**	28.46	25.09
524	**0.50**	0.64	1.18	4.36	**0.88**	2.54	4.35	24.85	**25.03**	21.88	19.73
527	**0.49**	0.81	0.96	1.89	**1.16**	2.11	3.28	**28.11**	26.38	23.93	18.35
753	**0.46**	0.74	1.19	3.37	**1.01**	2.28	4.15	26.65	**26.80**	24.07	19.36
780	**0.53**	1.04	1.66	4.96	**1.51**	2.37	5.02	24.99	**25.12**	21.27	19.35
Average	**0.56**	0.82	1.25	3.63	**1.22**	2.14	4.41	**26.89**	25.87	23.42	19.18

Table 2. **Detailed Experimental Results for the 5-Camera Setup on the Waymo Static-32 Split.**

Sequence	CD-Point Cloud ↓				CD-Mesh ↓			PSNR ↑			
	Ours	GVKF	2DGS	PGSR	GVKF	2DGS	PGSR	Ours	GVKF	2DGS	PGSR
003	0.68	**0.67**	0.74	3.11	**1.14**	1.51	4.19	**28.63**	26.22	25.19	22.22
019	**0.55**	0.92	1.32	3.14	**0.79**	2.01	3.04	**27.91**	25.04	23.63	24.18
036	**0.49**	0.68	0.93	2.69	**1.18**	1.32	2.39	**28.78**	27.77	27.44	24.88
069	**0.70**	0.85	1.21	4.20	**0.94**	2.04	3.53	27.65	**28.23**	27.72	25.49
081	**0.62**	0.94	1.12	3.16	**2.06**	3.55	4.67	**27.55**	25.35	25.57	21.17
126	**0.47**	0.75	0.95	1.69	**0.97**	1.51	1.77	**30.48**	28.62	30.14	27.60
139	**0.78**	1.06	1.35	3.10	**1.31**	1.76	2.87	25.56	**25.78**	24.98	19.92
140	**0.59**	0.69	1.32	2.36	**0.99**	1.66	2.05	25.70	25.68	**25.72**	23.73
146	**0.62**	0.97	1.45	3.12	**1.11**	2.05	2.9	25.37	**26.44**	26.42	20.15
148	**0.47**	0.58	1.07	3.26	**1.06**	1.23	3.53	**28.80**	26.47	25.69	23.92
157	**0.55**	0.67	**0.55**	1.44	**1.05**	1.41	1.60	**27.44**	26.63	25.80	21.98
181	**0.51**	0.82	0.87	3.84	**0.66**	1.04	3.99	25.07	**25.71**	24.74	23.39
200	**0.82**	1.38	1.67	2.43	**1.52**	2.64	2.23	27.00	**30.01**	29.34	22.71
204	**0.60**	0.75	1.13	2.72	**1.12**	1.53	2.84	**28.12**	26.98	26.73	25.53
226	**0.44**	0.49	0.62	3.17	**0.78**	1.20	3.48	**25.34**	22.22	21.38	19.41
232	**0.51**	0.62	0.84	4.03	**0.83**	1.43	4.41	**27.54**	25.86	24.89	22.19
237	**1.09**	1.88	2.19	2.69	**0.91**	2.38	2.44	25.11	**27.31**	26.77	25.36
241	**0.56**	0.88	1.16	3.53	**0.66**	2.59	3.19	**28.08**	27.38	24.99	21.44
245	**0.81**	1.29	1.39	2.04	**1.05**	2.22	2.01	**26.40**	25.95	25.20	19.72
246	**1.01**	1.04	1.47	3.65	**0.78**	2.46	3.79	**27.22**	26.45	24.77	24.03
271	**0.71**	0.95	0.87	1.91	**1.21**	1.25	2.00	**27.13**	24.78	25.22	20.50
297	**0.95**	1.00	3.48	4.11	**1.34**	3.63	4.20	**25.94**	24.15	24.30	20.21
302	**0.64**	0.68	1.49	2.71	**0.73**	1.66	2.70	**24.22**	23.92	23.58	20.55
312	0.72	**0.61**	1.01	2.33	**0.83**	1.87	2.49	**24.84**	24.14	24.00	22.03
314	0.89	0.91	**0.60**	2.33	**0.71**	1.35	2.70	**28.14**	24.95	25.10	19.24
362	**0.77**	0.84	1.75	2.56	**1.11**	2.06	2.72	26.36	**26.73**	26.33	24.12
482	**0.74**	1.00	0.91	3.33	**1.24**	1.34	3.91	**27.37**	25.16	23.45	22.28
495	**0.57**	0.65	1.27	3.00	**0.67**	1.73	3.12	28.34	**29.85**	29.6	26.41
524	**0.55**	0.66	1.05	2.86	**0.65**	2.07	3.08	24.45	**25.58**	23.82	23.36
527	**0.70**	0.94	1.08	1.62	**1.17**	1.66	1.93	**29.04**	27.00	26.51	21.37
753	**0.51**	0.70	1.02	2.27	**0.97**	1.39	2.47	**27.35**	27.06	27.16	24.01
780	**0.60**	0.90	1.39	4.27	**1.23**	1.77	4.60	**25.91**	25.50	22.93	20.46
Average	**0.66**	0.87	1.23	2.90	**1.02**	1.85	3.03	**26.96**	26.22	25.60	22.61

Table 3. **Detailed Experimental Results for the 3-Camera Setup on the Waymo Static-32 Split.**