Hi-Gaussian: Hierarchical Gaussians under Normalized Spherical Projection for Single-View 3D Reconstruction

Supplementary Material

1. More Theoretical Study

1.1. Normalized Spherical Projection

In Sec. 3.2 of the main text, we propose normalized spherical projection, which introduces less spatial compression than the un-normalized one. In other words, the curvature of the upper edges in the images obtained from normalized spherical projection is flatter, which is shown in Fig. 8 of the main text. Now we provide a theoretical analysis of the phenomenon described above.

The formula for un-normalized spherical projection is given as follows:

$$\begin{bmatrix} \theta \\ \phi \end{bmatrix} = \begin{bmatrix} \pi - \arctan\left(x^{-1}\right) \\ \arccos\left(-y/r\right) \end{bmatrix} \tag{1}$$

where $[x, y]^\top$ is coordinates in the camera coordinate system and $r = \sqrt{x^2 + y^2 + 1}$. We consider θ and ϕ as functions of x and y, denoting them as $\theta = h_1(x, y)$ and $\phi = h_2(x, y)$ respectively. From Eqn. 1, we can see that $h_1(x, y) = \pi - \arctan\left(x^{-1}\right)$ is solely dependent on x, while $h_2(x, y) = \arccos\left(-y/r\right) = \arccos\left(-y/\sqrt{x^2 + y^2 + 1}\right)$ is dependent on both x and y. Taking the partial derivative of $h_2(x, y)$ with respect to x, we obtain

$$\frac{\partial h_2(x,y)}{\partial x} = -\frac{xy}{(x^2 + y^2 + 1)\sqrt{x^2 + y^2 + 1}} \begin{cases} > 0 & \text{if} \quad xy < 0 \\ = 0 & \text{if} \quad xy = 0 \\ < 0 & \text{if} \quad xy > 0 \end{cases} \tag{2}$$

Thus, for a given constant $y_0 > 0$, $h_2(x, y_0)$ is strictly monotonically increasing for $x \in (-\infty, 0)$, and strictly monotonically decreasing for $x \in (0, +\infty)$. Similarly, $h_2(x, -y_0)$ is strictly monotonically decreasing for $x \in (-\infty, 0)$, and strictly monotonically increasing for $x \in (0, +\infty)$. Moreover, since $h_2(x, y) = h_2(-x, y)$, it is evident that $h_2(x, y)$ is an even function with respect to x. Therefore, for any $x \in \mathbb{R}$, for any $\hat{x} \in \{x' \mid |x'| < |x|\}$, we have $h_2(\hat{x}, y_0) > h_2(\pm x, y_0)$ and $h_2(\hat{x}, -y_0) < h_2(\pm x, -y_0)$.

Since normalization typically transforms x into a smaller $\hat{x}$, our normalized spherical projection maps Cartesian coordinates to a smaller range of spherical coordinates. This results in the upper and lower edges of the images obtained from normalized spherical projection appear relatively flatter than the un-normalized one.

2. More Experiments

Additional quantitative and qualitative results are presented to further elaborate on the superiority of our proposed method. We start by showcasing a more comprehensive comparison with other methods, then present a more detailed ablation study.

2.1. Quantitative Comparison with other Methods

Cross-dataset Novel View Synthesis. We present quantitative results on cross-dataset generalization. Models are trained on BundleFusion [3] and are tested on all 8 sequences in NeRF-LLFF [7] dataset. Table 1 shows that our method achieves the best performance on most sequences in NeRF-LLFF dataset.

Novel Depth Synthesis. To better showcase the depth estimation capability of our method, we conduct evaluation on novel depth synthesis following the experimental setup of SceneRF [2]. Our approach outperforms other methods across all metrics, which is shown in Table 2 and Table 3.

Scene Reconstruction. We also evaluate mesh reconstruction following SceneRF [2] for a fair comparison. As demonstrated in Table 4, our method achieves the best reconstruction performance. It is worth noting that if we proposed an improved approach for mesh reconstruction, we could get better results.

Ablation Study. To further demonstrate the effectiveness of our method, we conduct more ablation studies. We compare the quantitative results of novel view synthesis at different distances from the input view. The results in Fig. 1 show that our Normalized Spherical Projection module and Hierarchical Gaussian Sampling strategy consistently enhance the performance of our model, regardless of the distance between the novel view and the input view.

2.2. Qualitative Comparison with Other Methods

Cross-dataset Novel View Synthesis. To better demonstrate the generalization capability of our method, we conduct cross-dataset evaluations on novel view synthesis. Models are trained on BundleFusion [3] and are tested on NeRF-LLFF [7] dataset. The qualitative results in Fig. 2 indicate that our method renders the sharpest and clearest images in cross-dataset generalization.

Methods	Fern			Flower			Fortress			Horns		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SceneRF	13.64	0.130	0.702	13.31	**0.263**	0.584	15.46	0.290	0.636	14.55	0.222	0.650
Splatter Image	11.89	0.133	0.610	12.36	0.107	0.523	8.95	0.218	0.565	14.48	0.266	0.469
Ours	**16.93**	**0.439**	**0.359**	**14.01**	0.167	**0.468**	**17.59**	**0.302**	**0.318**	**15.84**	**0.398**	**0.379**
Methods	Leaves			Orchids			Room			Trex		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
SceneRF	**13.77**	**0.161**	0.548	11.31	0.128	0.716	12.74	0.286	0.743	10.65	0.093	0.757
Splatter Image	12.66	0.120	0.488	9.88	0.067	0.615	9.62	0.207	0.648	9.22	0.074	0.638
Ours	13.00	0.158	**0.443**	**11.76**	**0.145**	**0.507**	**16.26**	**0.454**	**0.328**	**16.91**	**0.453**	**0.274**

Table 1. Quantitative evaluations on cross-dataset generalization from BundleFusion to NeRF-LLFF dataset.

Methods	SemanticKITTI						
	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	δ_1 ↑	δ_2 ↑	δ_3 ↑
PixelNeRF [9]	0.2364	2.080	6.449	0.3354	65.81	85.43	92.90
MINE [5]	0.2248	1.787	6.343	0.3283	65.87	85.52	93.30
VisionNeRF [6]	0.2054	1.490	5.841	0.3073	69.11	88.28	94.37
SceneRF [2]	0.1681	1.291	5.781	0.2851	75.07	89.09	94.50
SplatterImage [8]	0.2519	2.127	7.282	0.4205	58.41	79.30	89.02
Ours	**0.1165**	**0.812**	**4.702**	**0.2397**	**80.99**	**90.02**	**94.67**

Table 2. Novel depth synthesis on SemanticKITTI datasets.

Mesh Visualization of Scene Reconstruction. To offer a more intuitive representation of scene reconstruction, we display 3D meshes on the validation set of SemanticKITTI [1, 4] and BundleFusion [3]. These meshes are produced from the scene TSDF, which is obtained through the conversion of rendered images and depths. Fig. 4 and Fig. 5 demonstrate that our method reconstructs the clearest and sharpest meshes.

Ablation study. Due to the limitation of the main text's length, we have only presented the qualitative results of ablation study on SemanticKITTI. Therefore, the results for BundleFusion are provided here, as shown in Fig. 3. Without Normalized Spherical Projection, our model fails to reconstruct outside the input FOV, leading to ghosting artifacts. Without Hierarchical Gaussian Sampling, there will be ripple artifacts and holes in the rendered novel views.

References

[1] Jens Behley, Martin Garbade, Andres Milioto, Jan Quenzel, Sven Behnke, Cyrill Stachniss, and Jurgen Gall. Semantickitti: A dataset for semantic scene understanding of lidar sequences. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9297–9307, 2019. 2

[2] 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. 1, 2, 3

[3] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 1, 2

[4] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 2

[5] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, 2021. 2, 3

[6] Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 806–815, 2023. 2, 3

[7] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (ToG)*, 38(4):1–14, 2019. 1

[8] Stanislaw Szymanowicz, Chrisitian Rupprecht, and Andrea Vedaldi. Splatter image: Ultra-fast single-view 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Com-*

Methods	BundleFusion						
	Abs Rel↓	Sq Rel↓	RMSE↓	RMSE log↓	$\delta_1 \uparrow$	$\delta_2 \uparrow$	$\delta_3 \uparrow$
PixelNeRF [9]	0.6029	2.312	1.750	0.5904	46.34	72.38	83.89
MINE [5]	0.1839	0.098	0.386	0.2386	65.53	91.78	98.21
VisionNeRF [6]	0.5958	2.468	1.783	0.5586	55.47	79.29	86.68
SceneRF [2]	0.1766	0.094	0.368	0.2100	72.71	94.89	99.23
SplatterImage [8]	0.2407	0.142	0.454	0.2710	57.06	89.00	97.99
Ours	**0.0792**	**0.041**	**0.225**	**0.1101**	**95.28**	**99.30**	**99.75**

Table 3. Novel depth synthesis on BundleFusion datasets.

Methods	SemanticKITTI			BundleFusion		
	IoU↑	Prec.↑	Rec.↑	IoU↑	Prec.↑	Rec.↑
SceneRF [2]	13.84	17.28	40.96	20.16	25.82	47.92
Splatter Image [8]	10.30	11.30	53.93	13.89	22.22	27.04
Ours	**15.56**	**17.39**	**59.72**	**40.42**	**48.91**	**69.96**

Table 4. Reconstruction evaluations on SemanticKITTI and BundleFusion datasets. We outperform all other methods across all metrics.

puter Vision and Pattern Recognition, pages 10208–10217, 2024. 2, 3

[9] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4578–4587, 2021. 2, 3

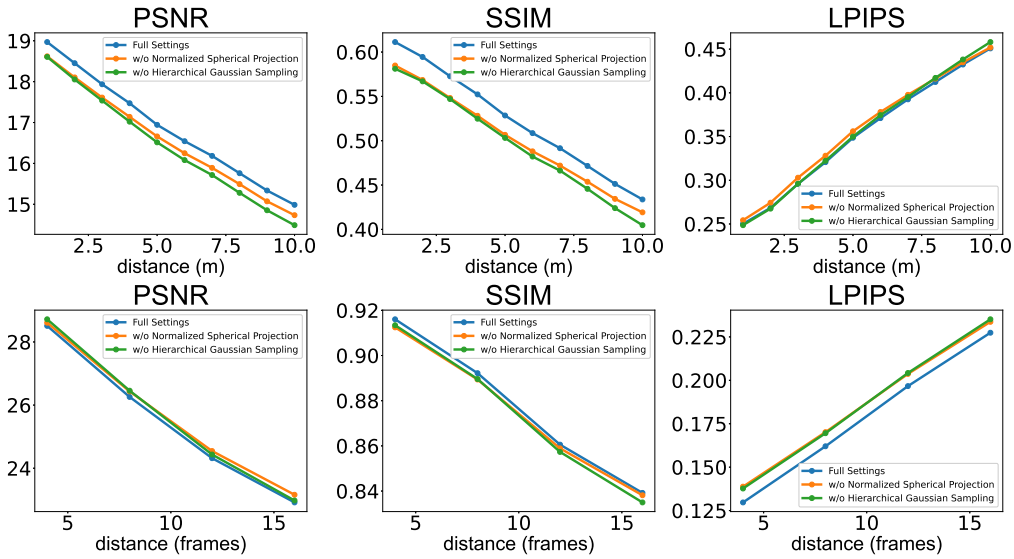

Figure 1. Performances at varying input view distances on SemanticKITTI (the first row) and BundleFusion (the second row) for ablation study. Our model with full settings usually achieves the best performance.

Figure 2. Qualitative evaluations on cross-dataset generalization from BundleFusion to NeRF-LLFF dataset.

Figure 3. Ablations on BundleFusion (val). Without Normalized Spherical Projection, our model fails to reconstruct outside the input FOV, leading to ghosting artifacts. Without Hierarchical Gaussian Sampling, there will be ripple artifacts and holes in the rendered novel views.

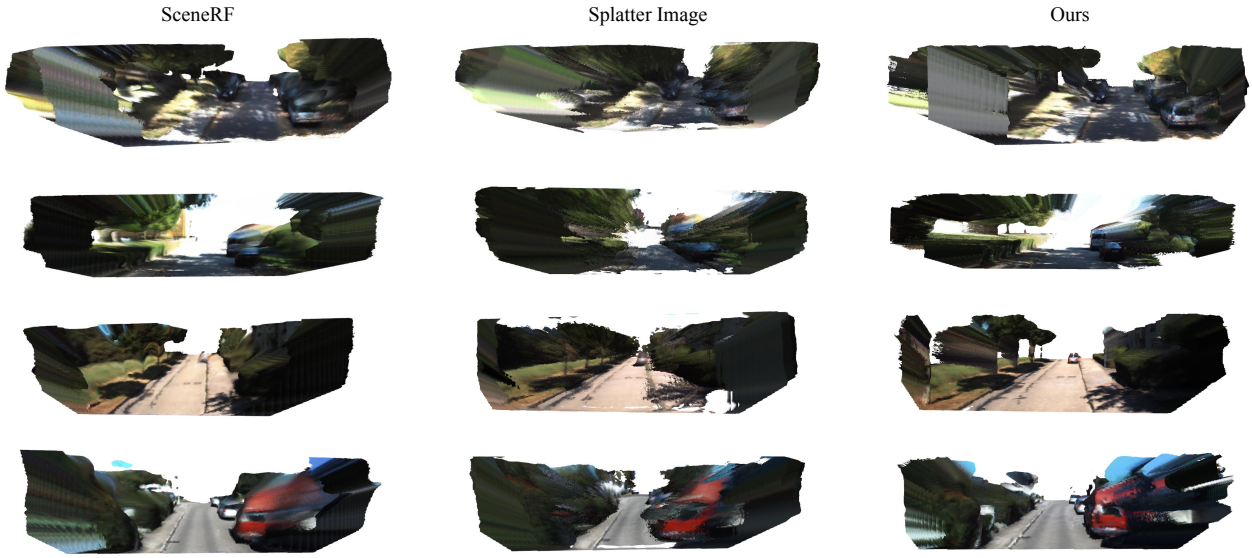

Figure 4. 3D meshes on SemanticKITTI (val.).

Figure 5. 3D meshes on BundleFusion (val.).