Depth Estimation from Indoor Panoramas with Neural Scene Representation Supplementary Material

Wenjie Chang, Yueyi Zhang†, Zhiwei Xiong
University of Science and Technology of China
changwj@mail.ustc.edu.cn, {zhyuey, zwxiong}@ustc.edu.cn

1. Proof to Initialization Scheme

Inspired by [1], we propose an initialization scheme which is illustrated as follows.

Background. A single fully-connected layer in MLPs is denoted as

$$f_i(\mathbf{y}) = v(\mathbf{W}_i\mathbf{y} + \mathbf{b}_i), \tag{1}$$

where $\mathbf{W}_i \in \mathbb{R}^{d_i^{out}\times d_i^{in}}$, $\mathbf{b}_i \in \mathbb{R}^{d_i^{out}}$, $\mathbf{y} \in \mathbb{R}^{d_i^{in}}$ and v denotes the ReLU activation function. Then, the MLP used in our networks is formulated as

$$f([x,y,z]^T;\theta) = \mathbf{w}^T f_l \circ f_{l-1} \circ \cdots \circ f_1([x,y,z]^T) + b, \tag{2}$$

where $\mathbf{w} \in \mathbb{R}^{d_l^{out}}$, $b \in \mathbb{R}$, $[x,y,z]$ denotes the input 3D location in Cartesian Coordinate and $\theta = (\mathbf{W}_l, \mathbf{b}_l, \ldots, \mathbf{W}_1, \mathbf{b}_1, \mathbf{w}, b)$ represents the parameters of the MLP.

Initialization. For the first layer in the defined MLP, $f_1(\mathbf{y}) = v(\mathbf{W}_1\mathbf{y} + \mathbf{b_1})$, we set $\mathbf{b_1} = 0$. As for $\mathbf{W}_1$, we represent it as column vectors,

$$\mathbf{W}_1 = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{d_1^{in}}], \quad \mathbf{a} \in \mathbb{R}^{d_1^{out}}, \tag{3}$$

where $\mathbf{a}_i$ is the i^{th} column vector in the matrix $\mathbf{W}_1$. Then we normalize $\mathbf{a}_2$ with a normal distribution $\mathcal{N} \backsim (0, \sqrt{2}/\sqrt{d_1^{out}})$ and set $\mathbf{a_n} = \mathbf{0}$ ($n \neq 2$). For other layers, we set all entries in $\mathbf{W}_i (1 < i \leq l)$ i.d.d. normal distribution $\mathcal{N} \backsim (0, \sqrt{2}/\sqrt{d_i^{out}})$ and $\mathbf{b}_i = 0$. Finally, set $\mathbf{w} = \sqrt{\pi}/\sqrt{d_l^{out}}$ and $b = -c$, where c is a hyper-parameter and is set to 1.5 in our experiments.

Proposition. With this initialization, the defined MLP is approximated to a signed distance function, $f([x,y,z]^T, \theta) \approx |y| - c$.

Proof. For the first layer of the defined MLP, $f_1(\mathbf{x}) = v(\mathbf{W}_1\mathbf{x} + \mathbf{b_1})$, if setting $k_1 = d_1^{out}$, we have

$$\|f_1(\mathbf{x})\|^2 = \sum_{j=1}^{k_1} v^2(\mathbf{u}_j \cdot \mathbf{x}) = \frac{1}{k_1}\sum_{j=1}^{k_1} v^2(\sqrt{k_1}\mathbf{u}_j \cdot \mathbf{x}), \tag{4}$$

where $\mathbf{u}_j$ denotes the j^{th} row of $\mathbf{W}_1$ and $\mathbf{m}_j = \sqrt{k}\mathbf{u}_j$. After applying the initialization scheme to parameters of the defined MLP in Eq. 2 and $\mathbf{m} = [0, m_2, 0, \ldots, 0]$, m_2 is i.d.d. from a normal distribution $\mathcal{N} \backsim (0, \sqrt{2})$. $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_{d^{in}}]$ is the input of each layer. Eq. 4 is calculated with the law of large numbers as

$$\frac{1}{k_1}\sum_{j=1}^{k_1} v^2(\sqrt{k_1}\mathbf{u}_j \cdot \mathbf{x}) \tag{5}$$

$$= E(v(\mathbf{m_j} \cdot \mathbf{x})^2) = \int_{\mathbb{R}^{k_1}} v^2(\mathbf{m}\cdot\mathbf{x})\mu(\mathbf{m})d\mathbf{m} \tag{6}$$

$$= \int_{\mathbb{R}} v^2(m_2 x_2)\mu(m_2)dm_2 \tag{7}$$

$$= \int_{\mathbb{R}} v^2\left(m_2\frac{x_2}{|x_2|}|x_2|\right)\mu(m_2)dm_2 \tag{8}$$

$$= |x_2|^2 \int_{\mathbb{R}} v^2\left(m_2\frac{x_2}{|x_2|}\right)\frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-m_2^2}{2\sigma^2}}dm_2 \tag{9}$$

$$= |x_2|^2 \int_{\mathbb{R}+} m_2^2 \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-m_2^2}{2\sigma^2}}dm_2 \tag{10}$$

$$= \frac{|x_2|^2}{2}\int_{\mathbb{R}} m_2^2 \frac{1}{\sqrt{2\pi\sigma^2}}e^{\frac{-m_2^2}{2\sigma^2}}dm_2 \tag{11}$$

$$= |x_2|^2, \tag{12}$$

where μ denotes the probability density function. As a result, The following equation holds between the output and input $\mathbf{x} = [x_1, x_2, x_3, \ldots, x_{d^{in}}]$ of the first layer

$$\|f_1(\mathbf{x})\|^2 \approx |x_2|^2. \tag{13}$$

When $1 < i < l$, we have $f_i(\mathbf{x}) = v(\mathbf{W}_i\mathbf{x} + \mathbf{b_i})$. Then $\|f_i(\mathbf{x})\|^2$ converges to

$$\int_{\mathbb{R}^k_i} v^2(\mathbf{m}\cdot\mathbf{x})\mu(\mathbf{m})d\mathbf{m} \tag{14}$$

$$= \|x\|^2 \int_{\mathbb{R}^{k_i}} v^2\left(\mathbf{m}\cdot\frac{\mathbf{x}}{\|\mathbf{x}\|}\right)\mu(\mathbf{m})d\mathbf{m} \tag{15}$$

where $\mathbf{m}$ is the row vector of the matrix $\sqrt{k_i}\mathbf{W}_i$ and is i.d.d with a normal distribution $\mathcal{N} \backsim (0, \sqrt{2})$. Let $\mathbf{m} = \mathbf{R}\mathbf{m}'$, where $\mathbf{R} \in \mathbb{R}^{k\times k}$ is an orthogonal matrix and $\mathbf{R}^T \frac{\mathbf{x}}{\|\mathbf{x}\|} = [1, 0, 0, \ldots, 0]$. Hence, $\mu(\mathbf{R}\mathbf{m}') = \mu(\mathbf{m}')$. Then, Eq. 14 is calculated as

$$\|\mathbf{x}\|^2 \int_{\mathbb{R}^{k_i}} v^2 \left(\mathbf{m} \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \mu(\mathbf{m}) d\mathbf{m} \tag{16}$$

$$= \|\mathbf{x}\|^2 \int_{\mathbb{R}^{k_i}} v^2 \left(\mathbf{m}'^T \mathbf{R}^T \frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \mu(\mathbf{m}') d\mathbf{m}' \tag{17}$$

$$= \|\mathbf{x}\|^2 \int_{\mathbb{R}^{k_i}} v^2 \,(m'_1) \mu(\mathbf{m}') d\mathbf{m}' \tag{18}$$

$$= \|\mathbf{x}\|^2 \int_{\mathbb{R}+} {m'_1}^2 \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1 \tag{19}$$

$$= \frac{\|\mathbf{x}\|^2}{2} \int_{\mathbb{R}} {m'_1}^2 \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1 \tag{20}$$

$$= \|\mathbf{x}\|^2 \tag{21}$$

Thus, for the input $\mathbf{x}$ of each layer, the output is formulated as

$$\|f_i(\mathbf{x})\|^2 \approx \|\mathbf{x}\|^2, 1 < i < l \tag{22}$$

Further more, for the last layer of the defined MLP, $f(\mathbf{x}) = \mathbf{w}^T v(\mathbf{W}_l \mathbf{x} + \mathbf{b}_l) + b$. After initializing, we get $f(\mathbf{x}) = \frac{\sqrt{\pi}}{k_l} \sum_{j=1}^{k_l} v(\mathbf{m}_j \cdot \mathbf{x}) - c$. By the law of large numbers, $\frac{\sqrt{\pi}}{k_l} \sum_{j=1}^{k_l} v(\mathbf{m}_j \cdot \mathbf{x})$ converges to

$$\sqrt{\pi}\|\mathbf{x}\| \int_{\mathbb{R}^k} v \left(\mathbf{m} \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \mu(\mathbf{m}) d\mathbf{m} \tag{23}$$

$$= \sqrt{\pi}\|\mathbf{x}\| \int_{\mathbb{R}^k} v \left(\mathbf{m}'^T \mathbf{R}^T \frac{\mathbf{x}}{\|\mathbf{x}\|}\right) \mu(\mathbf{m}') d\mathbf{m}' \tag{24}$$

$$= \sqrt{\pi}\|\mathbf{x}\| \int_{\mathbb{R}^k} v\,(m'_1) \mu(\mathbf{m}') d\mathbf{m}' \tag{25}$$

$$= \sqrt{\pi}\|\mathbf{x}\| \int_{\mathbb{R}^k} v\,(m'_1) \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1 \tag{26}$$

$$= \sqrt{\pi}\|\mathbf{x}\| \int_{\mathbb{R}+} m'_1 \tfrac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1 \tag{27}$$

$$= \frac{\sqrt{\pi}\|\mathbf{x}\|}{2} \int_{\mathbb{R}+} m'_1 \tfrac{\sqrt{2}}{\sqrt{\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1 \tag{28}$$

$$= \|\mathbf{x}\| \tag{29}$$

where $\mathbf{m} = \mathbf{R}\mathbf{m}'$ and $\mathbf{R} \in \mathbb{R}^{k\times k}$ is an orthogonal matrix which $\mathbf{R}^T \frac{\mathbf{x}}{\|\mathbf{x}\|} = [1, 0, 0, \ldots, 0]$. $\int_{\mathbb{R}+} m'_1 \frac{\sqrt{2}}{\sqrt{\pi\sigma^2}} e^{\frac{-{m'_1}^2}{2\sigma^2}} dm'_1$ denotes the mean of the half-normal distribution and is calculated as $\frac{2}{\sqrt{\pi}}$ when $\sigma = \sqrt{2}$. As a result, for the input $\mathbf{x}$ of the last layer,

$$f(\mathbf{x}) = \mathbf{w}^T v(\mathbf{W}\mathbf{x} + \mathbf{b}) + b \approx \|\mathbf{x}\| - c \tag{30}$$

From Eq. 13, Eq. 22 and Eq. 30, the MLP defined by Eq. 2 is initialized to denote the Signed Distance Function: $f([x, y, z]; \theta) \approx |y| - c$, which represents two planes at the distance c from the origin of the coordinate and thus approximates to the floors and ceilings in indoor scenes.

2. Evaluation Metrics

We evaluate our method with the same error metrics used in prior depth estimation works [5, 7]. The mathematical expressions of the evaluation metrics are presented in the following:

$$\text{Mean Absolute Error: } MAE = \frac{1}{n}\sum_p^n \frac{|y_p - \hat{y}_p|}{\hat{y}_p}$$

$$\text{Mean Square Error: } MSE = \frac{1}{n}\sum_p^n (y_p - \hat{y}_p)^2$$

$$\text{Mean Relative Error: } MRE = \frac{1}{n}\sum_p^n |y_p - \hat{y}_p|$$

$$\text{Threshold: } \delta_1 = \text{ \% of } y_p \text{ s.t.} \max\left(\frac{y_p}{\hat{y}_p}, \frac{\hat{y}_p}{y_p}\right) = \delta < 1.25$$

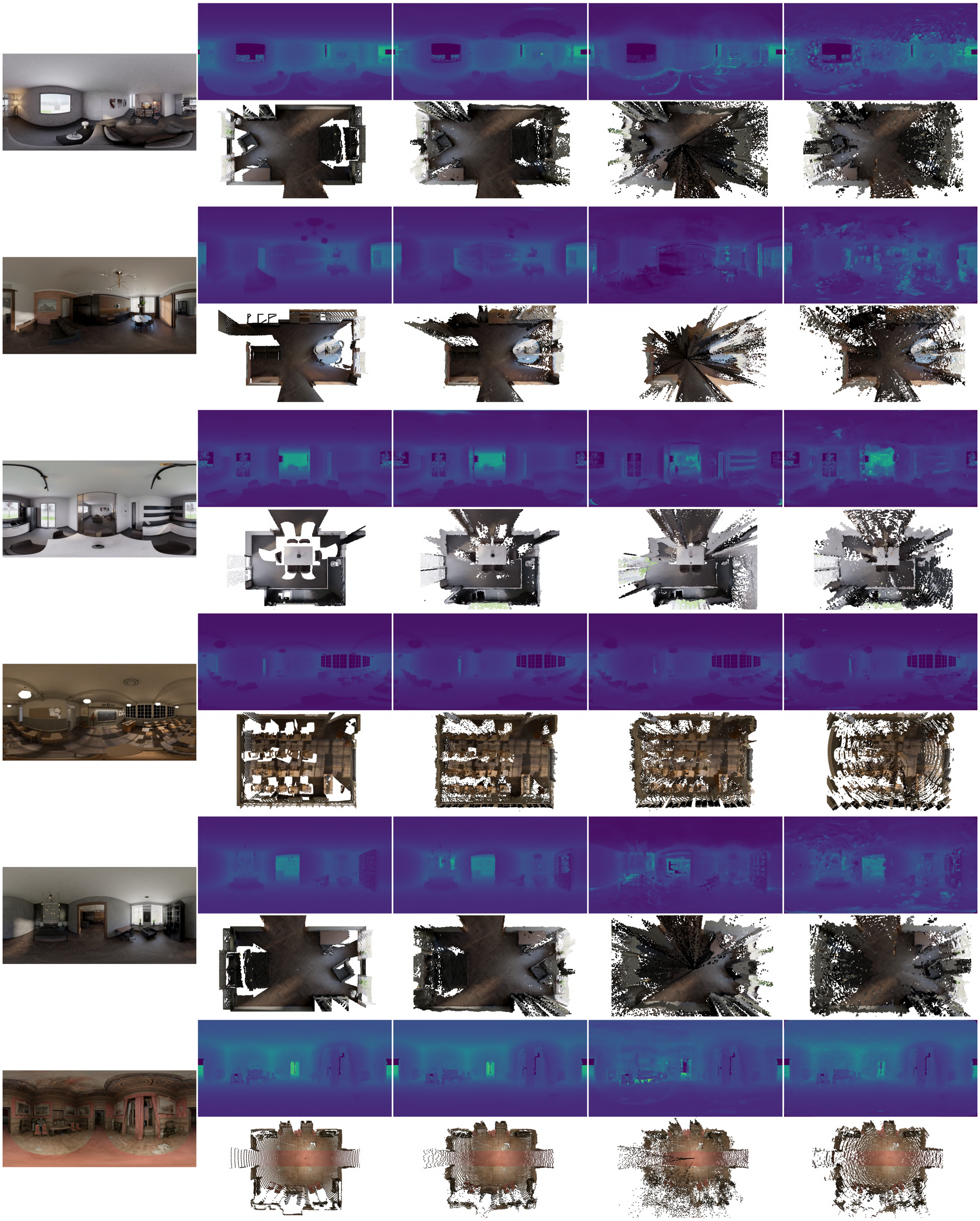
where n denotes the number of valid pixels of a panorama image, $\hat{y}_p$ denotes the depth value in ground truth and y_p is the depth value predicted from networks. Since many works on depth estimations adopt RMSE as a important metric for evaluation, we also present the quantitative results in standard RMSE for reference.

Dataset	Method	MRE ↓	RMSE ↓	δ_1 ↑
IPMP	NeRF	0.1890	0.6820	0.6712
	NeuS	1.0786	3.3349	0.4386
	VolSDF	0.5821	1.7501	0.0970
	Ours	**0.0641**	**0.3024**	**0.8975**
M3D	NeRF	0.1006	0.6892	0.8551
	NeuS	0.8232	3.9928	0.5163
	VolSDF	0.3018	1.3385	0.5298
	Ours	**0.0258**	**0.2164**	**0.9902**
S2D3D	NeRF	0.1209	0.6396	0.7960
	NeuS	0.4846	2.0591	0.6290
	VolSDF	0.5114	1.5369	0.2442
	Ours	**0.0352**	**0.2637**	**0.9790**

Table 1. Quantitative results evaluated with RMSE metric on IPMP, Matterport3D and Stanford2D3D datasets.

3. Qualitative Results

In this section, we show more qualitative comparisons in Figure 1.



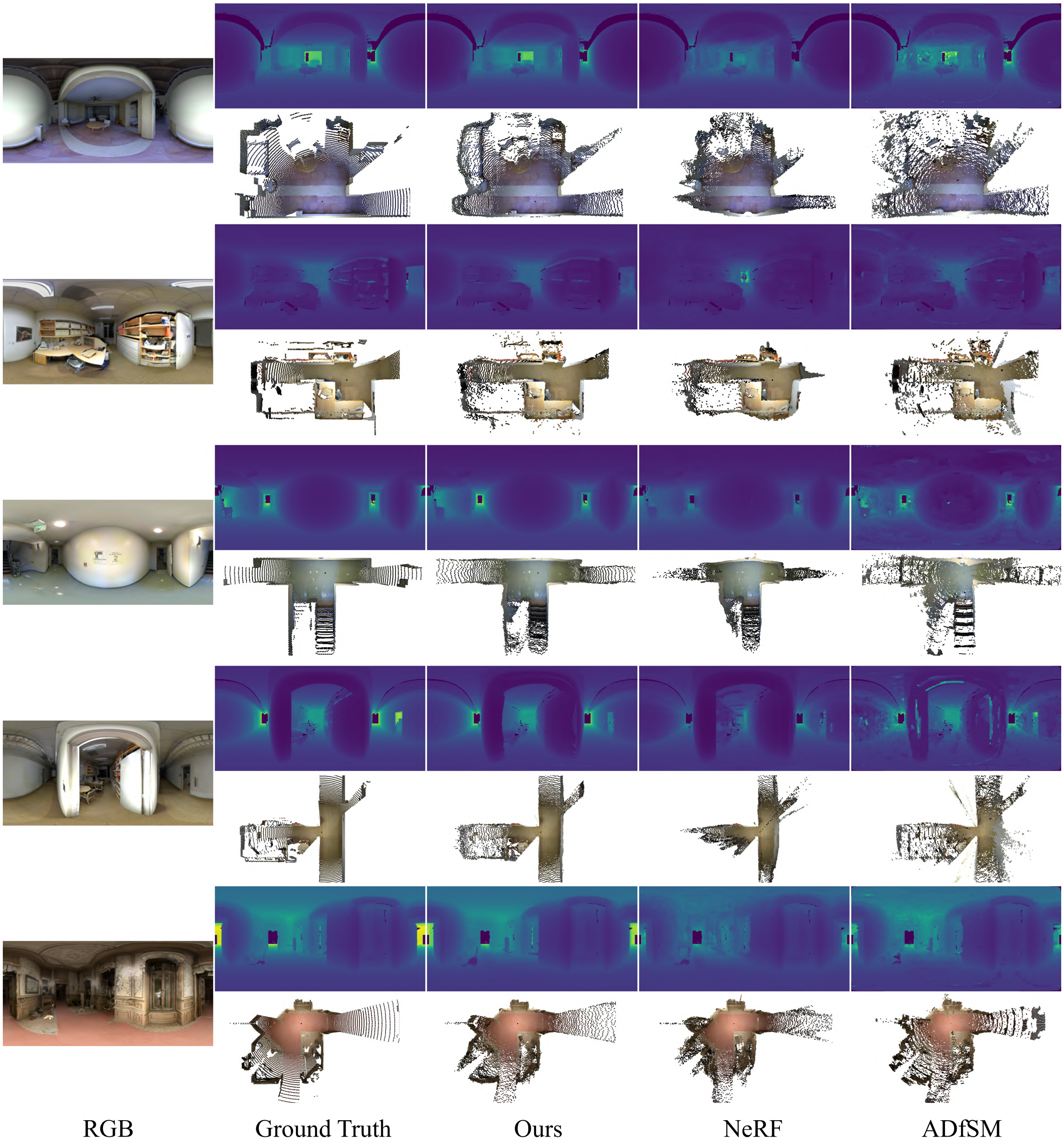


Figure 1. Qualitative results of our method, NeRF [4] and ADfSM [3]. Compared with current advanced approaches, our methods show accurate layout of the reconstructed scenes. All depth maps are visualized in the range of 0 to 10 meters. Point clouds in aerial views are transformed from the estimated depth maps and RGB images for better visualization.

4. The hyper-parameter c in the initialization

In this section, we discuss the hyper-parameter c, which is used in our proposed initialization approach to configure the distance between the approximated floors and ceilings. As Table 2 shows, the selection of c will affect the performance of the proposed initialization scheme to some extent. In our experiments, we set c to 1.5.

c	MAE ↓	MRE ↓	MSE ↓	δ_1 ↑
0	0.1252	0.0417	0.0997	0.9694
0.5	0.0967	0.0387	0.0787	0.9684
1.5	**0.0731**	0.0258	**0.0532**	**0.9902**
2.5	0.0738	**0.0257**	0.0574	0.9854
3.5	0.0793	0.0297	0.0606	0.9803

Table 2. An ablation study of c for quantitative evaluations. All results are trained with 3 views and metrics are averaged over the 10 scenes from Matterport3D. Setting c to 1.5 achieves better results.

5. Parameters

In this section, we discuss about the number of parameters in different methods. As Table 3 shows, our method, which outperforms other methods by a large margin, has similar parameter number with other methods.

Method	Parameters	Method	Parameters
NeRF	1.192M	VolSDF	0.803M
NeuS	1.407M	Ours	1.220M

Table 3. Comparisons of the number of parameters among different models.

6. Cylinder Initialization

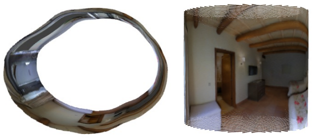


Figure 2. Illustration of the Cylinder Initialization.

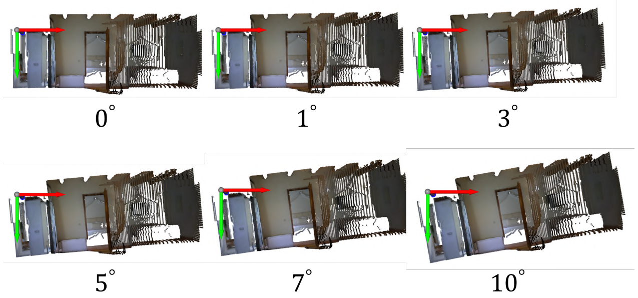
Based on the Manhattan World Assumption, we also develope an initialization method to approximate the walls of indoor scenes, which are parallel to the direction of gravity and perform like a Cylinder in the space as shown in Figure 2. Quantitative comparisons are shown in Table 4. The proposed initialization that approximates the floors and ceilings achieves the best performance.

Init.	MRE ↓	MSE ↓	δ_1 ↑
Sphere	0.0430	0.1032	0.9717
Cylinder	0.0556	0.1028	0.9590
Floors&Ceilings	**0.0258**	**0.0532**	**0.9902**

Table 4. Comparisons among different initialization methods. All results are trained with 3 views and metrics are averaged over the 10 scenes from Matterport3D.

7. Robustness of the Initialization Method

We evaluate the robustness of our proposed initialization scheme on several circumstances where the assumption that floors and ceilings are vertical to the gravity direction does not strictly hold. As shown in Figure 3, the vertical direction of the scene is deviated from the gravity direction by 1°, 3°, 5°, 7° and 10°. It is obvious that slight deviations do not cause dramatic performance degradation.



0° 1° 3°

5° 7° 10°

Figure 3. Illustration of the misalignment that the direction of gravity offset from the vertical direction to the ground by a few degrees. The Green Arrow denotes the direction of gravity.

Degree	MAE↓	MRE ↓	MSE ↓	δ_1 ↑
0°	**0.0731**	**0.0258**	**0.0532**	0.9902
1°	0.0785	0.0269	0.0610	0.9892
3°	0.0840	0.0310	0.0616	0.9826
5°	0.0843	0.0294	0.0629	0.9886
7°	0.0815	0.0279	0.0655	**0.9906**
10°	0.0849	0.0301	0.0684	0.9892

Table 5. Experimental results on scenes that are not strictly following the Manhattan World Assumption. Metrics are averaged over 10 artificially disturbed scenes from Matterport3D and each scene is trained with 3 views.

8. Comparisons with Supervised Methods

We evaluated the performance of the supervised monocular method. SliceNet is pre-trained on 3D60 [6], OmniFusion is pre-trained on Stanford2D3D. Ours is trained with

3 views. All methods are evaluated on our proposed IPMP dataset. As shown in the Table 6, the supervised methods suffer from domain adaptation issues and significant performance degradation in practical applications.

Method	Time	MAE ↓	MRE ↓	MSE ↓	δ_1 ↑
OmniFusion	CVPR 2022	0.6667	0.3101	0.7925	0.2993
SliceNet	CVPR 2021	0.4974	0.2100	0.7462	0.5776
Ours		**0.1266**	**0.0641**	**0.0955**	**0.8975**

Table 6. Comparisons with supervised methods. Metrics are averaged over 5 scenes.

9. Evaluations on the scene captured by a Commercial Panoramic Camera

We apply our method to a real-world dataset, the Coffee Area dataset, which is captured by a Ricoh-Theta-S spherical camera and first released in SOMSI [2]. Figure 4 visualizes the depth estimation and point cloud reconstruction results of three different methods for a single scene. Our method outperforms the supervised methods, which encounter domain adaptation issues in real-captured scenarios. We demonstrate the effectiveness and robustness of our method in terms of accuracy and completeness.

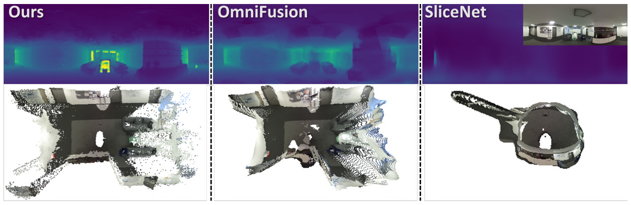


Figure 4. Qualitative results of different methods on the Coffee Area dataset.

10. Discussion of the processed dataset

The Matterport3D and Stanford2D3D datasets rerendered by [6] suffer the depth leaking issue that the brighter regions have smaller depth values (the left image of Fig. 5). As a result, the network may directly learn the depth information from the variety of the brightness, especially for a supervised method. To verify that the proposed method does not benefit from the leaked depth information, we brightened up the dark regions of panoramas from the processed Matterport3D dataset (the right image of Fig. 5). Table 7 shows the quantitative results of the original dataset and the brightness-adjusted dataset. Our method still generates satisfying results with the adjusted scenes, which demonstrate the proposed method has good robustness to brightness changes. The brightness adjusted dataset is released with the code.

Input	MAE ↓	MRE ↓	MSE ↓	δ_1 ↑
Original	**0.0731**	**0.0258**	0.0532	0.9902
Brightness Adjust	0.0776	0.0261	**0.0474**	**0.9916**

Table 7. Evaluations on the original dataset and Brightness-adjusted dataset. Metrics are averaged over 10 scenes.

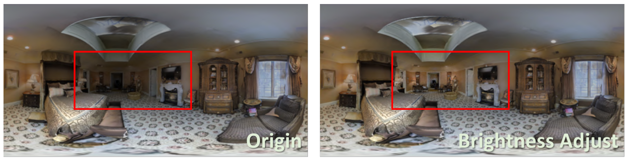


Figure 5. Illustration of the depth leaking issue and the brightness-adjusted dataset.

References

[1] Matan Atzmon and Yaron Lipman. Sal: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 1

[2] Tewodros Habtegebrial, Christiano Gava, Marcel Rogge, Didier Stricker, and Varun Jampani. Somsi: Spherical novel view synthesis with soft occlusion multi-sphere images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15725–15734, 2022. 6

[3] Sunghoon Im, Hyowon Ha, François Rameau, Hae-Gon Jeon, Gyeongmin Choe, and In So Kweon. All-around depth from small motion with a spherical panoramic camera. In *Proceedings of the European Conference on Computer Vision*, pages 156–172. Springer, 2016. 4

[4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision*, pages 405–421. Springer, 2020. 4

[5] Fu-En Wang, Yu-Hsuan Yeh, Min Sun, Wei-Chen Chiu, and Yi-Hsuan Tsai. Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 2

[6] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360 depth estimation. In *International Conference on 3D Vision*, pages 690–699. IEEE, 2019. 5, 6

[7] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision*, pages 448–465, 2018. 2