Depth Estimation from Indoor Panoramas with Neural Scene Representation

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

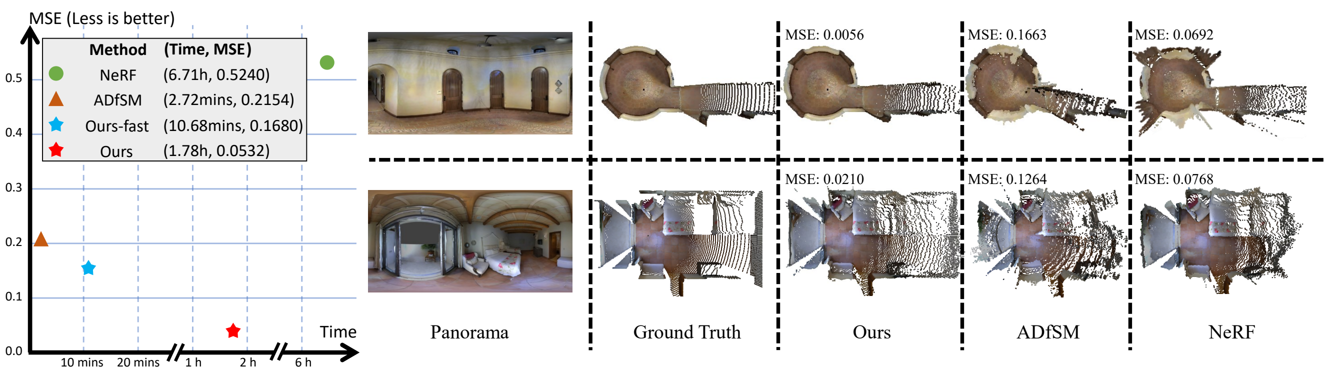

Figure 1. Our work aims to implicitly represent a scene with a neural network through fewer panoramas and thus obtain the relative depth measurements. The left diagram illustrates the efficiency and accuracy of our method against NeRF [18] and a traditional geometric-based Structure from Motion approach [12]. All methods adopt 3 panoramas with camera positions as inputs and are evaluated over 10 scenes from Matterport3D [4]. The right shows aerial views of point clouds that are transferred from depth estimations for better visualization. Compared with state-of-the-art methods, measurements from ours have clearer edges and more accurate geometry.

Abstract

Depth estimation from indoor panoramas is challenging due to the equirectangular distortions of panoramas and inaccurate matching. In this paper, we propose a practical framework to improve the accuracy and efficiency of depth estimation from multi-view indoor panoramic images with the Neural Radiance Field technology. Specifically, we develop two networks to implicitly learn the Signed Distance Function for depth measurements and the radiance field from panoramas. We also introduce a novel spherical position embedding scheme to achieve high accuracy. For better convergence, we propose an initialization method for the network weights based on the Manhattan World Assumption. Furthermore, we devise a geometric consistency loss, leveraging the surface normal, to further refine the depth estimation. The experimental results demonstrate that our proposed method outperforms state-of-the-art works by a large margin in both quantitative and qualitative evaluations. Our source code is available at https://github.com/WJ-Chang-42/IndoorPanoDepth.

1. Introduction

Panoramic imaging has emerged as an attractive imaging technique in many fields, such as computer vision and robotics. Different from traditional imaging devices, panoramic cameras capture a holistic scene and present it as a 2D image with equirectangular projection. Indoor panoramas, captured in the interior scenes by panoramic cameras, have been widely used in interior design and decoration. Recovering depth information aligned with RGB panoramic images benefits a line of down-streaming applications, such as augmented reality and indoor mapping.

Recent works on depth estimation from panoramas employ Convolutional Neural Network (CNN) structures with prior knowledge learned from depth labels and achieve excellent performance. Most of these works adopt a single panoramic image to predict the relative depth map [7, 23, 29, 31, 37, 39]. These methods require lots of RGB and depth pairs while training and encounter the problem of domain adaptation in practice. There are a few works attempting to employ multiview panoramic images in the depth estimation task [32, 38]. They recover depth information by finding the correspondence of different views. However, strict vertical or horizontal position relations are required for input images in these methods.

Panoramas show great distortions when presented as 2D images. Prior works adopt various technologies to overcome this problem, such as processing panoramas [7,26,27,

†Corresponding Author

31] with perspective projection and developing special convolution kernels [8, 30, 37]. Recently, the Neural Radiance Field (NeRF) [18] based on volume rendering has attracted great attention, which aims to synthesize novel views and recover the geometry of a complex scene. It considers image pixels as the rendering results of camera rays casting to the scene and learns geometric information from the correspondence among each ray, which eliminates affects from distortions while processing panoramic images. However, when applied to panoramas, the state-of-the-art scene representation methods still require a number of input images and take a long time to converge. It is a compelling research problem to explore how to leverage the omnidirectional information in panoramas to achieve satisfying depth estimation results with fewer images and faster convergence.

To exploit the holistic spatial information in panoramas, we propose a framework to achieve holistic depth estimation with a few panoramic images. Our framework consists of two main networks with a novel positional embedding scheme for learning a better representation from panoramas. The geometry network estimates the Signed Distance Function (SDF) to represent the 3D information of the scene, and the color network reconstructs the color texture. With the assistance of the rendering equation, the expected color of a pixel in an image is rendered with radiance values of the sampled 3D coordinates along camera rays. Both networks are optimized by minimizing the difference between the rendered colors. Inspired by [2], we propose a method to initialize the parameters of the geometry network based on the assumption that floors and ceilings are always vertical to the gravity direction in indoor panoramic images, which provides guidance to properly optimize the geometry network. Experimental results show that the proposed initialization scheme facilitates the network converge faster and achieves better results. In addition, considering that the geometric information from the depth is supposed to be consistent with the geometry from the surface normal, we devise the geometric consistency loss, which further refines the depth measurements. Moreover, we construct a synthetic dataset that provides RGB-D image pairs from various positions. We evaluate our method on our synthetic dataset and two real-world datasets. The experimental results demonstrate that our method achieves superior performance among state-of-the-art approaches. Even with fewer image views and a short training period, our method works well and outputs promising depth measurements. Our contributions are summarized as follows:

- We propose an unsupervised method for depth estimation from multi-view indoor panoramic images by utilizing a neural network with a specially designed positional embedding scheme to implicitly learn the SDF of the scene represented by panoramas.
- Inspired by the Manhattan World Assumption, we propose an initialization method for the network weights for better convergence.
- We devise a loss item based on geometric consistency that the geometric information from depth is supposed to be consistent with the surface norm.
- We release a synthetic panoramic RGB-D dataset rendered from photorealistic indoor scenes. Experimental results on our synthetic dataset and two realistic datasets demonstrate that our proposed method achieves superior performance in both quantitative and qualitative ways.

2. Related work

2.1. Indoor Panoramic Depth Estimation

Image-based depth estimation is a fundamental problem in 3D vision [3, 5, 11, 14, 22, 25]. For depth estimation from panoramas, the severe spatial distortion brought by equirectangular projection is challenging. To deal with this deformation, some methods dispose panorama images with a standard projection. Cheng *et al.* [7] adopted an additional cubemap projection branch which converts a panorama image to six faces of a cube with perspective projection. Wang *et al.* [31] proposed a bi-projection fusion component to leverage both projections, which was inspired by both peripheral and foveal vision of the human eye. HoHoNet [29] and SliceNet [23] extracted horizontal 1D feature maps from gravity-aligned equirectangular projections and recovered dense 2D predictions.

Others redesign the convolution kernels for panoramic depth estimation. Specifically, SphereNet [8] and DistConv [30] calculated the sampling positions for the convolution kernels with inverse gnomonic projection, and Fernandez-Labrador *et al.* [10] defined the convolution over the field of view on the spherical surface with longitudinal and latitudinal angles. Eder *et al.* [9] proposed a more general method to process images of any structured representation by introducing the corresponding mapping function. Zhuang *et al.* [37] proposed a combined dilated convolution to process panorama images. A few works fulfil the depth estimation task from multiview panorama images. Zioulis *et al.* [38] proposed a self-supervised method to train a monocular depth estimation network with two vertical panoramic images. Wang *et al.* [32] developed a stereo depth estimation spherical method which predicts disparity using the setting of top-bottom camera pairs. Both methods demand a strict spatial relation in the vertical or horizontal direction between different views.

2.2. Neural Scene Representation

Neural scene representation, which encodes a 3D scene with a neural network shows superior performance on 3D reconstruction and free-view rendering tasks [2, 6, 16–19, 21, 33, 35]. In particular, NeRF [18] has opened a series of research combining neural implicit functions together with volume rendering to achieve photo-realistic rendering results. NeRV [28] took a set of images in a scene illuminated by unconstrained known lighting as input and produced a 3D representation that can be rendered from novel viewpoints under arbitrary lighting conditions as output. Ost *et al.* [20] proposed a learned scene graph representation, which encodes object transformation and radiance, to efficiently render novel arrangements and views of the scene. Wei *et al.* [34] utilized both conventional SfM reconstruction and learning-based priors over NeRF [18] for multi-view depth estimation. Yariv *et al.* [36] modeled the SDF values as a function of geometry in contrast to previous works utilizing volume density. Wang *et al.* [33] proposed a new volume rendering method to train a neural SDF representation.

3. Background

We first introduce the volume rendering mechanism that obtains the expected color of a camera ray from a scene represented by SDF and directional emitted radiance, following NeRF [18], NeuS [33] and techniques in [15]. A camera ray $\mathbf{r}(t)$ passing through the scene with near and far bounds t_n and t_f is denoted by

$$\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}, \quad \mathbf{r}(t), \mathbf{o}, \mathbf{d} \in \mathbb{R}^3, \quad t \in [t_n, t_f], \tag{1}$$

where $\mathbf{o}$ is the camera origin and $\mathbf{d}$ is a unit direction vector. With a discrete set of 3D coordinates $\{\mathbf{r}(t_i)|i = 1, \dots, N, t_i < t_{i+1}\}$ sampled along the defined camera ray, the expected color is rendered as

$$C(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}_i, \qquad T_i = \prod_{j=1}^{i-1}(1 - \alpha_j), \tag{2}$$

where T_i is the discrete accumulated transmittance, $\mathbf{c}_i$ is the radiance value at the sampled 3D coordinates $\mathbf{r}(t_i)$ with direction $\mathbf{d}$. α_i is the discrete opacity value, which is defined with SDF values as

$$\alpha_i = max\left(\frac{\Phi_s(\sigma_i) - \Phi_s(\sigma_{i+1})}{\Phi_s(\sigma_i)}, 0\right), \tag{3}$$

where σ_i denotes the SDF value at the sampled location $\mathbf{r}(t_i)$, $\Phi_s(x)$ is the sigmoid function equated as $\Phi_s(x) = (1 + e^{-sx})^{-1}$ and s in $\Phi_s(x)$ is a learnable parameter. The rendered color $C(\mathbf{r})$ is then employed for the loss function.

4. Method

4.1. Proposed Network

In this section, we present the pipeline of the proposed method. First, pixels in panoramas are back-projected to a unit sphere and then the relative camera rays are calculated with camera positions. Afterwards, a set of 3D coordinates are sampled from each obtained camera ray. Our constructed network takes the embedded coordinates as input and outputs the corresponding SDF and radiance values. Finally, the expected color of each camera ray is rendered with the SDF and radiance values of the sampled coordinates for optimization. Details of each step are introduced separately.

Back-projection. For a point $[m, n]$ in Image Coordinate, the relative polar angle θ and azimuthal angle ϕ are calculated as

$$\theta = 2\pi\frac{m - c_x}{W} \qquad \phi = \pi\frac{n - c_y}{H}, \tag{4}$$

where W and H represent the width and height of the panoramic image respectively. c_x and c_y are coordinates of the principal points. The direction of the camera ray passing through $[m, n]$ is calculated as

$$\mathbf{d} = [\cos(\phi)\sin(\theta), \sin(\phi), \cos(\phi)\cos(\theta)]^T. \tag{5}$$

Sampling. For each obtained camera ray, we sample $N_c + N_f$ coordinates with a coarse-to-fine sampling strategy. In practice, we first sample N_c positions with a stratified sampling approach, which partitions the near and far bounds $[t_n, t_f]$ into N_c evenly spaced bins and then randomly draws one sample t_i from each bin. The mathematical expression of t_i which denotes the distance between the sampled 3D coordinates and the camera origin, is formulated as

$$t_i \backsim U[t_n + \frac{i-1}{N}(t_f - t_n), t_n + \frac{i}{N}(t_f - t_n)]. \tag{6}$$

The SDF values of the N_c samples predicted by networks are utilized to calculate a probability density function (PDF) as

$$PDF(n) = \frac{w_n}{\sum_{i=1}^{N_c} w_i}, \quad w_i = T_i \alpha_i. \tag{7}$$

Then, additional N_f samples are obtained with the calculated PDF and inverse transform sampling. Finally, the expected color of a camera ray is rendered with $N_c + N_f$ sampled 3D coordinates.

Positional Embedding. Rahaman *et al.* [24] showed that using high frequency functions to map the network's inputs to a higher dimensional space enables better fitting of data containing high frequency variation. As a result, NeRF [18] adopts a positional embedding scheme for the sampled 3D coordinates, which is commonly followed by Neural Rendering works and represented as

$$\begin{aligned}\gamma(x) = \big(&\sin\left(2^0 x\right), \cos\left(2^0 x\right), \cdots, \\ &\sin\left(2^{L-1} x\right), \cos\left(2^{L-1} x\right)\big).\end{aligned} \tag{8}$$

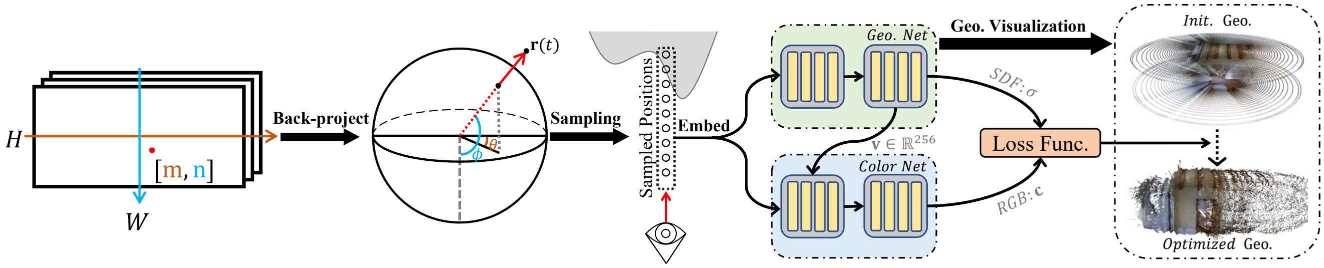

Figure 2. Illustration of the proposed method. Given the camera positions, we back project the pixels to a unit sphere to get camera rays. Then, a discrete set of sampled points along the ray is fed to our networks after positional embedding. Especially, geometry network adopts the same embedding way with NeRF and color network utilizes the proposed Sphere Embedding. Colors of each ray for optimizing are estimated from predicted SDF σ and RGB $\mathbf{c}$ with the volume rendering strategy. Init. Geo. is the visualization of the geometry field which is initialized with our proposed scheme before training. Optimized Geo. shows the scene denoted by geometry network after optimized.

Here γ is a mapping from $\mathbb{R}$ to a higher dimensional space $\mathbb{R}^{2L}$. However, two values x_1 and x_2 would be embedded as the same high dimensional vectors under the following condition

$$|x_1 - x_2| = 2\pi n, n = 0, 1, 2, \cdots . \tag{9}$$

To avoid this ambiguity, all coordinates in the sampled position $\mathbf{r}(t_i)$ are supposed to be in $[-\pi, \pi]$. NeRF [18] resolves this problem by scaling the whole scene to Normal Device Coordinate, encompassing a cube where the x, y, and z components range from -1 to 1. To address these limitations, we develop a new representation scheme, called Sphere Embedding. First, $\mathbf{r}(t_i) = [x, y, z]^T$ is converted in Sphere Coordinate, denoted as $[\theta, \phi, \rho]^T$. Then, it is encoded in a fusion representation, which is formulated as

$$\begin{aligned}\mu([\theta, \phi, \rho]^T) = [&\cos(\theta), \sin(\theta), \sin(\phi), \\ &\cos(\theta)\cos(\phi), \sin(\theta)\cos(\phi), 1/\rho]^T.\end{aligned} \tag{10}$$

The input of the MLP network is $\gamma(\mu(\mathbf{r}(t_i)))$. For this expression, the value space for ρ without ambiguity is $[\frac{1}{2\pi}, +\infty)$, which is more appropriate for indoor scenes.

Network Architecture. We construct two Multilayer-Perceptrons (MLPs) to learn the geometry field $\mathbb{F}_g$ and color field $\mathbb{F}_c$ separately. The geometry network adopts the embedded position $\gamma(\mathbf{r}(t_i))$, outputting SDF value σ and geometric feature vector $\mathbf{v} \in \mathbb{R}^{256}$. Here, σ determines the distance between the 3D location $\mathbf{r}(t_i)$ and its closest surface.

Then, we employ a color network to estimate color $\mathbf{c} \in \mathbb{R}^3 = [r, g, b]^T$ at point $\mathbf{r}(t_i)$. The color network takes the geometric feature vector $\mathbf{v}$, the embedded position $\gamma(\mu(\mathbf{r}(t_i)))$ and direction vector $\gamma(\mathbf{d})$ as inputs.

Both the geometry field $\mathbb{F}_g$ and the color field $\mathbb{F}_c$ utilize MLPs with 8 fully-connected layers and each layer consists of a linear transformation with the ReLU activation. Specifically, a skip connection operation is applied to the input of the 5^{th} layer in MLPs, which concatenates the embedded 3D information ($\gamma(\mathbf{r}(t_i))$ for the geometry network and $\gamma(\mathbf{d})$ for the color network) with feature vectors from the 4^{th} layer.

Optimizing. The constructed networks predict the SDF value σ_i and radiance value $\mathbf{c}_i$ of the samples $\{\mathbf{r}(t_i)|i = 1, \ldots, N_c + N_f, t_i < t_{i+1}\}$, and then the expected color of the camera ray $\mathbf{r}$ is calculated with Eq. 2. The main loss function minimizes the difference between the rendered and true pixel colors, which is formulated as

$$\mathcal{L}_C = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|, \tag{11}$$

where $\mathcal{R}$ is a batch of camera rays, $\hat{C}$ is the reference color and C is the rendered RGB colors.

4.2. Initialization

Atzmon and Lipman [2] proposed an initialization method that makes MLPs approximate an SDF representing a sphere in the 3D space before training. Indoor scenes are always different from spheres, but generally obey the Manhattan World Assumption, where floors and ceilings are always vertical to the gravity direction. Hence, it is a more reasonable choice to make the geometry field optimized from a shape that approximates the floors and ceilings. Inspired by [2] and the Manhattan World Assumption, we develop an initialization scheme for indoor panoramas and details are introduced in the following.

A single fully-connected layer in MLPs is denoted as

$$f_i(\mathbf{y}) = ReLU(\mathbf{W}_i\mathbf{y} + \mathbf{b}_i), \tag{12}$$

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

$$f([x, y, z]^T; \theta) = \mathbf{w}^T f_l \circ \cdots \circ f_1([x, y, z]^T) + b, \tag{13}$$

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 θ =

Algorithm 1 Initialization Scheme

Input: Network Parameters, $(\mathbf{W}_l, \mathbf{b}_l, \ldots, \mathbf{W}_1, \mathbf{b}_1, \mathbf{w}, b)$. Hyper-parameters, c.

Note: $\mathbf{w} \in \mathbb{R}^{d_l^{out}}$, $b \in \mathbb{R}$, $\mathbf{W}_i \in \mathbb{R}^{d_i^{out} \times d_i^{in}}$, $\mathbf{b}_i \in \mathbb{R}^{d_i^{out}}$, $\mathbf{W}_1 = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{d_1^{in}}]$, $\mathbf{a} \in \mathbb{R}^{d_1^{out}}$.

```
Let i = 1.
while i ≤ l do
    if i = 1 then
        Set b_1 = 0, W_1 = 0.
        Set a_2 in W_1 i.i.d. normal N ∽ (0, √2/√(d_1^out))
    else
        Set b_i = 0, W_i i.i.d. normal N ∽ (0, √2/√(d_i^out)).
    end if
    i = i + 1.
end while
Set w = √π/√(d_l^out), b = −c.
```

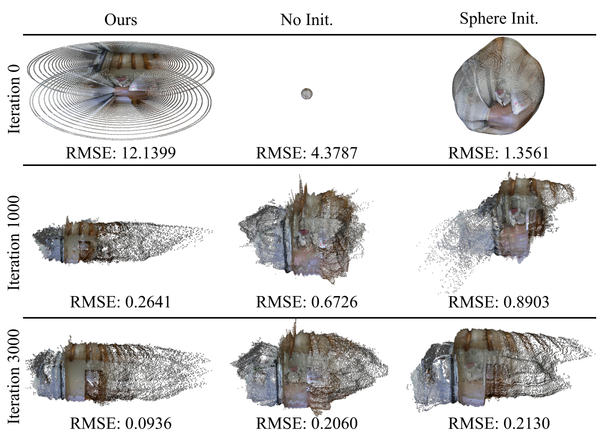

Figure 3. Illustration of the initialization scheme to the geometry field. We visualize the reconstruction results as point clouds in different training iterations. Our method shows a better reconstruction performance and faster convergence speed.

$(\mathbf{W}_l, \mathbf{b}_l, \ldots, \mathbf{W}_1, \mathbf{b}_1, \mathbf{w}, b)$ represents the parameters of the MLP. The initialization of parameters of the MLP is illustrated in Algorithm 1.

With the proposed scheme, the MLPs defined by Eq. 13 denotes an SDF, $f([x, y, z]^T, \theta) \approx |y| - c$. As Fig. 3 shows, after initialization, the geometry network is optimized from a shape approximating floors and ceilings in indoor scenes and converges more quickly. The proof of the initialization is provided in the supplementary material.

4.3. Geometric Consistency

The loss function in Eq. 11 only optimizes the proposed network with the rendered colors. As a result, we propose a new loss item to constrain the consistency of different properties learned by the geometric network. In addition to depth information, the surface normal is also one of the significant properties utilized to characterize the geometric of a scene. With our constructed geometric network, surface normal could be extracted from the points on the surface and the gradient of the network respectively.

Given a batch of camera rays $\mathcal{R} = \{\mathbf{r}_k(t) = \mathbf{o}_k + t\mathbf{d}_k | 1 \leq k \leq K\}$, the expected depth $D(\mathbf{r})$, which represents the distance from the camera origin to a point on the surface, is rendered with

$$D(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i t_i, \quad \mathbf{r} \in \mathcal{R}, \tag{14}$$

where α and T_i are introduced in Sec. 3 and t_i denotes the distance between the sampled 3D coordinates and the origin of ray $\mathbf{r}$. The 3D coordinates on surface along the ray are calculated by

$$\mathbf{a}_k = \mathbf{o}_k + D(\mathbf{r}_k)\mathbf{d}_k, \quad \mathbf{a}_k \in \mathbb{R}^3, \quad 1 \leq k \leq K, \tag{15}$$

The corresponding matrix is constructed as

$$\mathbf{A} = [\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_K]^T \in \mathbb{R}^{K \times 3}. \tag{16}$$

Then, the surface normal is calculated with

$$\boldsymbol{n}_d = \frac{\left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{1}}{\left\| \left(\mathbf{A}^\top \mathbf{A}\right)^{-1} \mathbf{A}^\top \mathbf{1} \right\|_2}, \tag{17}$$

where $\mathbf{1} \in \mathbb{R}^K$ is a vector with all 1 elements.

Meanwhile, the gradient of the geometric network could also be applied to calculate the surface normal. For a set of 3D points $\{\mathbf{r}(t_i) | i = 1, \ldots, N, t_i < t_{i+1}\}$ sampled from a camera ray $\mathbf{r}$, the normal vector at each point $\mathbf{r}(t_i) = [x_i, y_i, z_i]$ is defined by the partial derivative of the predicted SDF value α_i as

$$\boldsymbol{n}_i = -\left(\frac{\partial \alpha_i}{\partial x_i}, \frac{\partial \alpha_i}{\partial y_i}, \frac{\partial \alpha_i}{\partial z_i}\right) \tag{18}$$

Then, the normal vector of the surface hit by camera ray $\mathbf{r}$ is computed by

$$N(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{n}_i, \tag{19}$$

here α and T_i are introduced in Sec. 3. Finally, with the normal vectors $\{N(\mathbf{r}_k) | 1 \leq k \leq K\}$ calculated from Eq. 19 and $\mathbf{n}_d$ from Eq. 17, the consistency loss function is formulated as

$$\mathcal{L}_{GC} = 1 - cos\left(\mathbf{n}_d, \frac{1}{K}\sum_{k=1}^{K} N(\mathbf{r}_k)\right), \tag{20}$$

which minimizes the difference of the estimated surface normal from these two different ways and further refines the depth estimations in our evaluations.

5. Experiments

5.1. Datasets

IPMP. To evaluate our approach and other Neural Radiance Fields methods, we construct a synthetic dataset with multiview panorama images named the **I**ndoor **P**hotorealistic **M**ultiview **P**anoramas dataset. We select 5 different indoor scenes and render 12 photorealistic images in Blender at different viewpoints for each scene. All images are rendered with 512×1024 pixels from a panoramic camera. Different from other datasets that only offer multiview images at fixed positions with only 3 views, our dataset provides more views at various positions and does not follow a strict spatial relation in the vertical or horizontal direction between different views. Compared with other synthetic panorama RGB-D datasets, IPMP is rendered from high polygon models with the path-tracing algorithm, outputting high quality RGB images and precision depth maps.

Matterport3D. Matterport3D [4] is a large-scale RGB-D dataset containing panoramic views from RGB-D images of 90 building-scale scenes. In our experiments, we adopt the dataset processed by [38] and select 10 different scenes for evaluation.

Stanford2D3D. Stanford2D3D [1] is collected in 6 large-scale indoor areas that originate from 3 different buildings of educational and official use, providing equirectangular RGB images, as well as their corresponding depths. We select 5 different scenes for evaluation from the remake[1] by [38].

5.2. Implementation details

We implement our network with the PyTorch framework and train 100 epochs on a single NVIDIA RTX3090 GPU with approximately 6 GB video memory in 6.71 hours. Specifically, in the first 80 epochs, we randomly sample 512 camera rays for each optimizing iteration and train the networks only with the main color loss $\mathcal{L} = \mathcal{L}_C$. For the last 20 epochs, the camera rays are obtained from 32 image patches with 4×4 pixels, which means the K in Eq. 20 is set to 16. The loss function in this stage is $\mathcal{L} = \mathcal{L}_C + 0.01\mathcal{L}_{GC}$. We use the Adam optimizer [13] with a learning rate that begins with 5×10^{-4} and decays exponentially to 5×10^{-5} during optimization. The hyper-parameter c in the proposed initialization scheme is set to 1.5.

5.3. Evaluation

Baselines. To evaluate our approach, we conduct comparisons against scene representation methods which directly learn geometric information from multiview RGB images, a supervised stereo method and a geometric-based structure from motion (SfM) method. For scene representation methods, we compare with NeRF [18], NeuS [33] and VolSDF [36]. Particularly, NeRF adopts volume density to represent the 3D information, and NeuS [33] and VolSDF [36] utilize SDF to denote the surface of a scene. Moreover, we choose 360SD-Net [32], a supervised stereo method which achieves advanced depth estimation results with panoramic images, and ADfSM [12], a traditional geometric-based SfM method. In addition to our origin model, we also provide a fast model, named Ours-fast, which is trained for only 10 epochs without the learning rate decay.

Dataset	Method	MRE ↓	MSE ↓	δ_1 ↑
IPMP	NeRF	0.1890	0.5240	0.6712
	NeuS	1.0786	13.7711	0.4386
	VolSDF	0.5821	3.1328	0.0970
	Ours	**0.0641**	**0.0955**	**0.8975**
	Ours-fast	0.1546	0.2629	0.7483
M3D	NeRF	0.1006	0.5442	0.8551
	NeuS	0.8232	20.3833	0.5163
	VolSDF	0.3018	1.9464	0.5298
	Ours	**0.0258**	**0.0532**	**0.9902**
	Ours-fast	0.0801	0.1680	0.9331
S2D3D	NeRF	0.1209	0.4262	0.7960
	NeuS	0.4846	8.0183	0.6290
	VolSDF	0.5114	2.5315	0.2442
	Ours	**0.0352**	**0.0732**	**0.9790**
	Ours-fast	0.0550	0.1322	0.9423

Table 1. Evaluation results on IPMP, Matterport3D and Stanford2D3D. All results are trained with 3 views. For Matterport3D, metrics are averaged over 10 scenes. For Stanford2D3D and IPMP, metrics are evaluated on 5 scenes.

Following [31, 39], we use the following metrics to evaluate the performance: mean absolute error (MAE), mean relative error (MRE), mean square error of linear measures (MSE[2]) and relative accuracy δ_1 (the fraction of pixels where the relative error is within a threshold of 1.25). All errors are calculated in meters.

Quantitative Results. We compare with the advanced scene representation methods on 2 realistic datasets and the proposed dataset. All methods are trained with 3 views for a scene. As shown in Table 1, NeuS [33] and VolSDF [36] fail to reconstruct correctly and ours achieves the best results.

In addition, comparisons with prior advanced depth estimation works are constructed to demonstrate the outstanding performance of the proposed framework. As Table 3 shows, our approach presents competitive performance

[1]The processed dataset suffers from depth leakage problems, and we provide more discussion in the supplementary.

[2]Since many works adopt RMSE metric for evaluations, we also report the results with the standard RMSE in the supplementary.

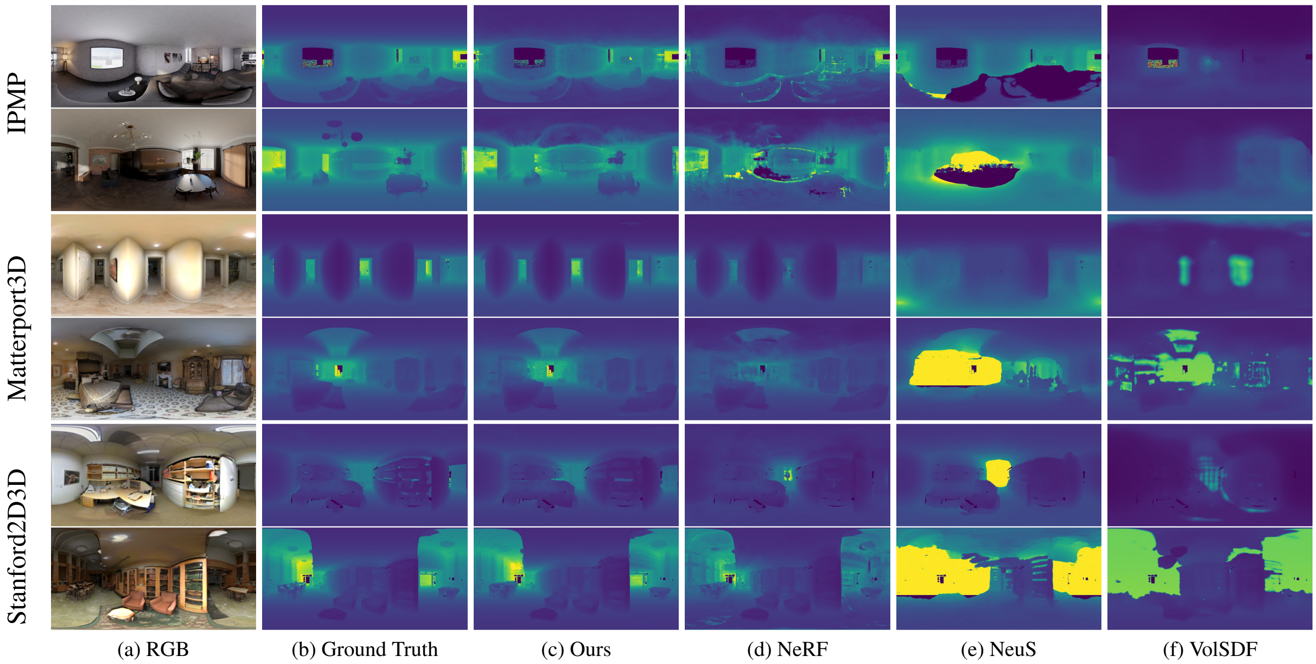

Figure 4. Qualitative comparisons on ours, NeRF [18], NeuS [33] and VolSDF [36]. (a) shows one view from the input images of a scene and (b) is the related ground truth depth map. All depth maps are visualized in a same depth range.

Method	Setting	MAE ↓	MRE ↓	MSE ↓	δ_1 ↑	Setting	MAE ↓	MRE ↓	MSE ↓	δ_1 ↑
NeRF	3-View	0.4270	0.1890	0.5240	0.6712	6-View	0.1943	0.0938	0.1783	0.9086
NeuS		2.0442	1.0786	13.7711	0.4386		0.2936	0.1407	1.2934	0.8870
VolSDF		1.3927	0.5821	3.1328	0.0970		0.5339	0.2403	1.0874	0.5545
Ours		**0.1266**	**0.0641**	**0.0955**	**0.8975**		**0.0953**	**0.0500**	**0.0874**	**0.9494**
NeRF	9-View	0.1411	0.0724	0.1112	0.9394	12-View	0.1231	0.0618	0.0939	0.9478
NeuS		0.1489	0.0727	0.1659	0.9273		0.1335	0.0635	0.1778	0.9423
VolSDF		0.3271	0.1356	0.6523	0.7688		0.2133	0.0856	0.3475	0.8504
Ours		**0.0711**	**0.0392**	**0.0560**	**0.9633**		**0.0747**	**0.0401**	**0.0650**	**0.9656**

Table 2. Evaluation results on our proposed dataset. We quantify the performance of each method using different numbers of images. The results are averaged on five scenes with image views used for training. Our method quantitatively outperforms all prior Neural Radiance Fields work in all settings.

against the supervised method and outperforms the traditional SfM method by a large margin. It should be noted that 360SD-Net [32] needs a top-bottom image pair that follows a strict vertical relation and is trained on a fusion dataset with RGB-D labels provided by [38]. Our approach does not require any prior knowledge from the ground truth and is evaluated on two settings. One is trained with 2 views (the same image pair as 360SD-Net) and the other is trained with 3 views. ADfSM is implemented with 3 views and the provided camera positions. 360SD-Net [32] is trained with RGB-D pairs from the total training set and evaluated on scenes selected from the evaluation set. Our method directly works on the panoramas from the evaluation set without any 3D supervision.

Qualitative Results. Fig. 4 shows qualitative results on our synthetic dataset IPMP and two real-world datasets Matterport3D and Stanford2D3D. All depth maps are shown in the same depth range for fair comparisons. The results on the synthetic dataset show that our method can better manage complex lighting conditions and obtain a more reasonable 3D representation (1^{st} and 2^{nd} rows). The results from the real-world dataset show that our method has a better performance in areas with severe distortion (4^{th} and 6^{th} rows), while a more accurate depth estimation result can be ob-

Method	MRE ↓	MSE ↓	δ_1 ↑
ADfSM	0.0493	0.1985	0.9598
360SD-Net	0.0322	0.0740	0.9760
Ours(2-view)	0.0713	0.1944	0.9427
Ours(3-view)	**0.0277**	**0.0525**	**0.9901**

Table 3. Comparisons with 360SD-Net [32] and ADfSM [12]. All metrics are evaluated on the depth estimations of the top image from the 360SD-Net's inputs and averaged over 10 scenes on Matterport3D. Our method shows comparable performance against the supervised method with one more view.

Embedding	Init.	MRE ↓	MSE ↓	δ_1 ↑
$[x, y, z]$	Ours	0.3606	2.1447	0.3348
$[x/s, y/s, z/s]$	Ours	0.4228	2.6902	0.2497
$[\theta, \phi, 1/\rho]$	Ours	0.1073	0.4034	0.8682
$S.E.$	Ours	**0.0258**	**0.0532**	**0.9902**
$S.E.$	None	0.1615	0.1767	0.8897
$S.E.$	Sphere	0.0430	0.1032	0.9717

Table 4. An ablation study of different positional embedding schemes for the inputs of the color network and the initialization scheme for the geometry network. Metrics are averaged over the 10 scenes from Matterport3D. Each scene is trained with 3 views. $[x, y, z]$ means positions in Cartesian Coordinate. $[x/s, y/s, z/s]$ means positions are scaled to the Normal Device Coordinate. $[\theta, \phi, 1/\rho]$ denotes positions that are converted to Sphere Coordinate. The three representations above are then embedded with Eq. 8 to acquire the inputs of the color network. $S.E.$ denotes the proposed Sphere Embedding.

tained in low-textured areas, such as walls, floors, and ceilings (3^{rd} row). Our method can overcome the problem of unsmooth depth estimation caused by repeatedly textured decorative items in interior scenes (4^{th} row).

5.4. Ablation Study

In this section, we construct a series of experiments to discuss the number of views in training and investigate the effectiveness of the proposed positional embedding scheme, the initialization method and the new loss item for geometric consistency.

Number of Views. The effects of using different numbers of image views on the depth estimation results are explored on our synthesized dataset. Table 2 shows the quantitative results. All methods are well converged with 6, 9 and 12 views. However, when fewer images are available, NeuS [33] and VolSDF [36] fail to generate satisfactory results. Our method achieves the best results among all settings and outperforms other methods by a large margin.

Positional Embedding. We perform ablation experiments to validate the proposed Sphere Embedding techniques with

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

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.

a series of experiments on different embedding schemes. Quantitative results presented in the $1^{st} - 4^{th}$ rows of Table 4 demonstrate that the proposed Sphere Embedding greatly improves the performance of the constructed networks when compared with other methods.

Initialization. To validate the effectiveness of the proposed initialization scheme, we deploy different initializing strategies for the geometry network, including no initialization, Sphere initialization [2], and our proposed initialization. The results are shown in the $4^{th} - 6^{th}$ rows of Table 4. Our proposed initialization method leads to the best depth estimation results. Table 5 shows the experiments on the robustness of the proposed initialization method when the normal of the ground in each scene deviates from the gravity direction by a few degrees. Even if the assumption that floors and ceilings are vertical to the gravity direction does not strictly hold, the proposed initialization scheme still has a superior performance.

Geometric Consistency Loss. To evaluate the effectiveness of the proposed Geometric Consistency Loss, we train the proposed network with only the color loss $\mathcal{L}_C$ for 100 epochs on 10 scenes from Matterport3D. The averaged evaluation metrics are MRE: 0.0288, MSE: 0.0592 and δ_1: 0.9884. Referring to the 4^{th} row in Table 4, the developed loss reduces the MSE by 10.14%.

6. Conclusion

In this paper, we propose a framework to estimate depth from multiview indoor panoramas with neural scene representation. The developed positional embedding scheme, initialization and geometric consistency loss improve our networks to obtain accurate measurements efficiently. Experimental results demonstrate that our method outperforms state-of-the-art NeRF-based methods in both qualitative and quantitative ways.

7. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under Grants 61901435, 62131003 and 62021001.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 6

[2] 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. 2, 3, 4, 8

[3] Shariq Farooq Bhat, Ibraheem Alhashim, and Peter Wonka. Adabins: Depth estimation using adaptive bins. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4009–4018, 2021. 2

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision*, 2017. 1, 6

[5] Xihao Chen, Zhiwei Xiong, Zhen Cheng, Jiayong Peng, Yueyi Zhang, and Zheng-Jun Zha. Degradation-agnostic correspondence from resolution-asymmetric stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12962–12971, June 2022. 2

[6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 3

[7] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360 videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429, 2018. 1, 2

[8] Benjamin Coors, Alexandru Paul Condurache, and Andreas Geiger. Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In *Proceedings of the European Conference on Computer Vision*, pages 518–533, 2018. 2

[9] Marc Eder, True Price, Thanh Vu, Akash Bapat, and Jan-Michael Frahm. Mapped convolutions. *arXiv preprint arXiv:1906.11096*, 2019. 2

[10] Clara Fernandez-Labrador, Jose M Facil, Alejandro Perez-Yus, Cédric Demonceaux, Javier Civera, and Jose J Guerrero. Corners for layout: End-to-end layout recovery from 360 images. *IEEE Robotics and Automation Letters*, 5(2):1255–1262, 2020. 2

[11] Xueyan Huang, Yueyi Zhang, and Zhiwei Xiong. High-speed structured light based 3d scanning using an event camera. *Opt. Express*, 29(22):35864–35876, Oct 2021. 2

[12] 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. 1, 6, 8

[13] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning and Representation*, 2015. 6

[14] Yue Li, Yueyi Zhang, and Zhiwei Xiong. Revisiting flipping strategy for learning-based stereo depth estimation. In *2021 International Conference on Visual Communications and Image Processing*, pages 1–4. IEEE, 2021. 2

[15] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3

[16] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3

[17] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 3

[18] 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. 1, 2, 3, 4, 6, 7

[19] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2020. 3

[20] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, June 2021. 3

[21] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[22] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1610–1621, 2022. 2

[23] Giovanni Pintore, Marco Agus, Eva Almansa, Jens Schneider, and Enrico Gobbetti. SliceNet: deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11536–11545, June 2021. 1, 2

[24] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 3

[25] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2

[26] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer

for indoor 360 depth estimation. *arXiv e-prints*, pages arXiv–2203, 2022. 1

[27] Zhijie Shen, Chunyu Lin, Lang Nie, Kang Liao, and Yao Zhao. Distortion-tolerant monocular depth estimation on omnidirectional images using dual-cubemap. In *2021 IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2021. 1

[28] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 3

[29] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Hohonet: 360 indoor holistic understanding with latent horizontal features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 1, 2

[30] Keisuke Tateno, Nassir Navab, and Federico Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *Proceedings of the European Conference on Computer Vision*, pages 707–722, 2018. 2

[31] 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. 1, 2, 6

[32] Ning-Hsu Wang, Bolivar Solarte, Yi-Hsuan Tsai, Wei-Chen Chiu, and Min Sun. 360sd-net: 360 stereo depth estimation with learnable cost volume. In *2020 IEEE International Conference on Robotics and Automation*, pages 582–588. IEEE, 2020. 1, 2, 6, 7, 8

[33] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 3, 6, 7, 8

[34] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[35] Ruikang Xu, Mingde Yao, Chang Chen, Lizhi Wang, and Zhiwei Xiong. Continuous spectral reconstruction from rgb images via implicit neural representation. In *Proceedings of the European Conference on Computer Vision Workshops: Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part V*, pages 78–94. Springer, 2023. 3

[36] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34, 2021. 3, 6, 7, 8

[37] Chuanqing Zhuang, Zhengda Lu, Yiqun Wang, Jun Xiao, and Ying Wang. Acdnet: Adaptively combined dilated convolution for monocular panorama depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 3653–3661, 2022. 1, 2

[38] 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. 1, 2, 6, 7

[39] 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. 1, 6