The Visual Computer
https://doi.org/10.1007/s00371-023-02863-5

ORIGINAL ARTICLE

RDNeRF: relative depth guided NeRF for dense free view synthesis

Jiaxiong Qiu[1] · Yifan Zhu[1] · Peng-Tao Jiang[2] · Ming-Ming Cheng[1] · Bo Ren[1]

Accepted: 27 March 2023

Abstract
In this paper, we focus on dense view synthesis with free movements in indoor scenes for better user interactions than sparse views. Neural radiance field (NeRF) handles sparsely and spherically captured scenes well, while it struggles in scenes with dense free views. We extend NeRF to handle these views of indoor scenes. We present a learning-based approach named relative depth guided NeRF (RDNeRF), which jointly renders RGB images and recovers scene geometry in dense free views. To recover the geometry of each view without the ground-truth depth, we propose to directly learn the relative depth by implicit functions and transform it as a geometric volume bound for geometry-aware sampling and integration of NeRF. With correct scene geometry, we further model the implicit internal relevance of inputs to enhance the representation ability of NeRF in dense free views. We conduct extensive experiments in indoor scenes for dense free view synthesis. RDNeRF outperforms current state-of-the-art methods and achieves 24.95 PSNR score and 0.77 SSIM score. Besides, it recovers more accurate geometry than basic models.

Keywords Dense free view synthesis · Neural radiance fields · Relative depth · Internal relevance

1 Introduction

Rendering novel views of a scene is a fundamental and challenging task in computer graphics and vision. Image-based rendering methods [1, 4, 14, 29, 30, 48, 52] have been proposed to fix scene geometry into synthesizing images. However, they have limited control over the quality of their results in novel views.

NeRF [27] tackles this issue and achieves impressive results of novel view synthesis, which combines the standard volume rendering [20] and neural implicit representations. It packs the whole scene into a fixed-bound volume, which is available to fit the correct geometry of spherical views because the rendered depth can be estimated by integrating learned volumetric density. Despite the success in rendering spherical views, we observe that NeRF achieves poor rendering results in dense free views synthesis. As shown in Fig. 1, the rendered depth of NeRF illustrates the wrong geometry in **dense free views** of indoor scenes, which results in a noisy rendered image. Unlike sparse spherical views in which the camera moves at a fixed distance around the rendered objects, dense free views [5] are captured by a camera moves with arbitrary motion. In this work, we aim to improve NeRF for rendering realistic dense free views.

In the following, we have summarized the reasons why NeRF often fails in dense free views synthesis. (i) The geometric relationship between objects and the camera is continuously changing in dense free views. For example, when the camera moves to a chair, the distance between the camera and the chair decreases. This geometric relationship is hard to be modeled by neural volume rendering with a fixed volume bound. (ii) The same object can be observed in each spherical view of NeRF's scenes [27], while they only exist in a few views captured by the freely moving camera (e.g., the chair shown in Fig. 1). Thus, dense free views suffer

Jiaxiong Qiu and Yifan Zhu have contributed equally to this work.

✉ Bo Ren
rb@nankai.edu.cn

Jiaxiong Qiu
qiujiaxiong727@gmail.com

Yifan Zhu
zhuyifan@mail.nankai.edu.cn

Peng-Tao Jiang
pt.jiang@mail.nankai.edu.cn

Ming-Ming Cheng
cmm@nankai.edu.cn

[1] VCIP, College of Computer Science, Nankai University, Tianjin, China

[2] Zhejiang University, Hangzhou, China

Published online: 05 May 2023

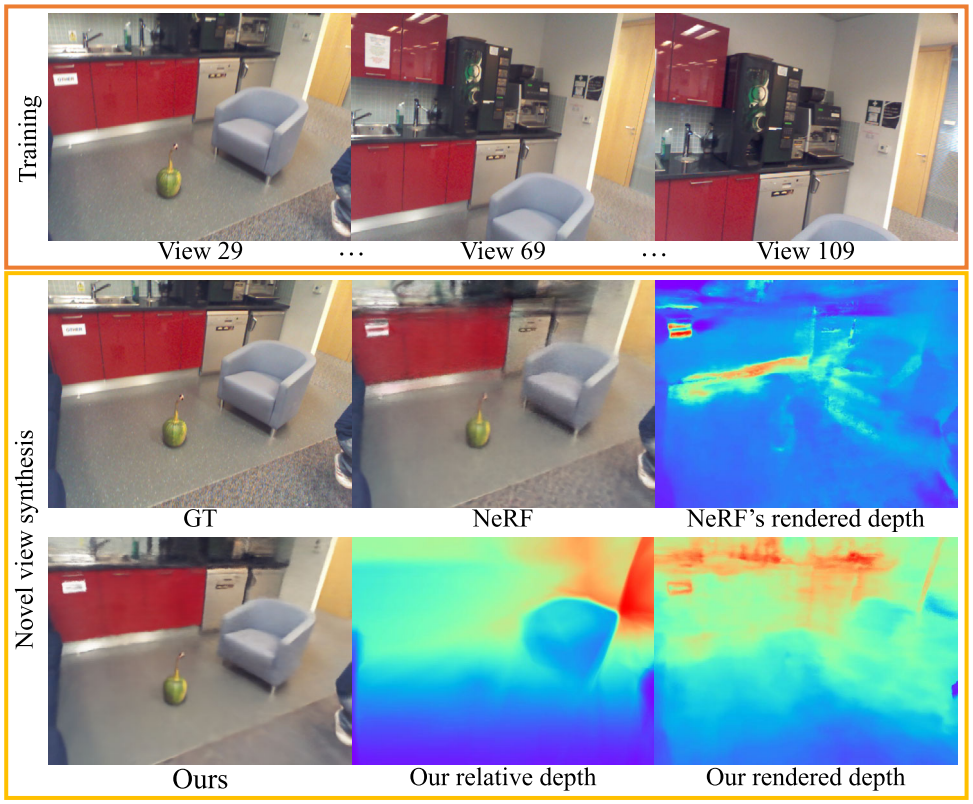

Fig. 1 In dense free views (more than 500 views) of a scene, the content of a view may appear at an arbitrary distance. This ambiguity makes NeRF [27] fail to extract the correct geometry of this scene, then synthesize noisy views. We tackle this ambiguity and synthesize a more realistic novel view with more accurate geometry

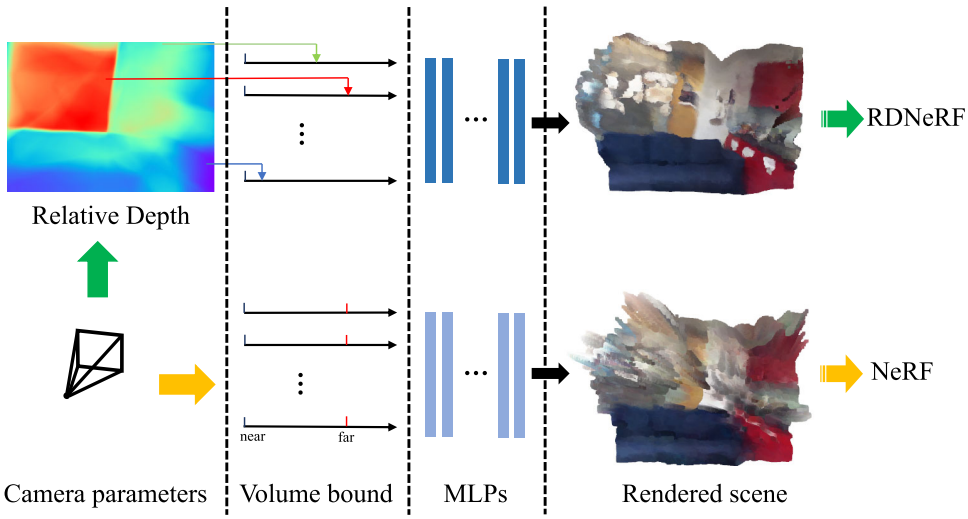

Fig. 2 Overview about RDNeRF and NeRF [27]. In NeRF, the volume bound of each camera ray is usually set as a constant value in a view. In RDNeRF, we learn the relative depth from a generalizable model [46] to recover the correct scene geometry and guide the sampling and integration of NeRF. We also enhance the representation ability of MLPs. The point cloud demonstrates that our model achieves more realistic and geometry-aware rendering

from more ambiguity than spherical views. NeRF consists of vanilla MLPs with a limited ability of scene representation, which makes NeRF hard to handle the tremendous ambiguity of dense free views. As shown in Fig. 2, NeRF reconstructs an incorrect 3D view because of the incorrectly rendered depth and low-quality rendered image.

To remedy these issues, we propose a framework called RDNeRF, an extension of NeRF that renders dense free views and recovers the scene geometry simultaneously. The overall pipeline of our framework is visualized in Fig. 3. Specifically, we tackle this challenging task in two aspects: (i) recovering scene geometry by learning the relative depth. The depth of the real-world scenes needs expensive devices to be acquired. And the obtained depth often loses some important parts of objects. We propose to utilize a generalizable model [46] to generate the relative depth, recovering the holistic scene geometry. Due to the lack of actual scale in the relative depth, we take it as the relative distance between objects and the camera to build the spatial context of camera rays. (ii) Modeling the internal relevance of the sampled points

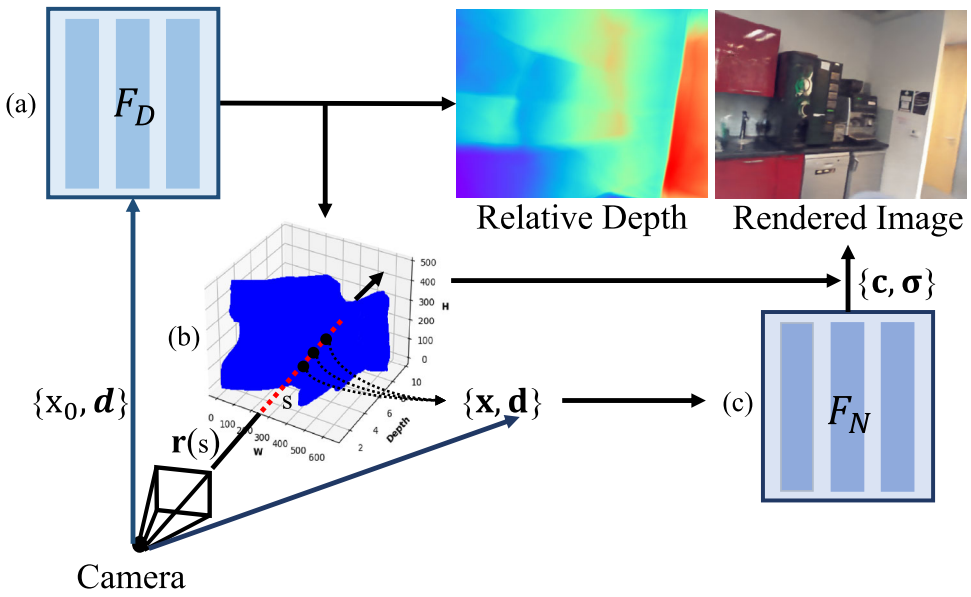

Fig. 3 The architecture of RDNeRF. Our framework consists of two parts: relative depth prediction and novel view synthesis. **a** Predicting the relative depth for a geometric volume bound. **b** Sampling spatial points **x** along the camera ray (**r**). **c** The backbone of RDNeRF with the self-attention mechanism. **a**, **c** are parameterized by MLPs. The sampling procedure and integration of RDNeRF are guided by the geometric volume bound

along camera rays for the view ambiguity. We observe that there are more sampled points in the volume for dense free views than sparse views for rendering. We consider the internal relevance between the sampled points along a camera ray in a view cannot be modeled well by NeRF. We enhance the internal relevance among the sampled points by using the self-attention mechanism.

In summary, the main contributions of our paper are:

- We design a NeRF-based framework for dense free view synthesis, named RDNeRF, which can render high-quality images with accurate geometry in real-world indoor scenes.
- We model the spatial context between each camera ray by learning the relative depth and transforming it as a geometric volume bound.
- We build the internal relevance of the sampled points along camera rays by the self-attention mechanism, enhancing the representation ability of neural radiance fields in dense free views.

2 Related work

2.1 Novel view synthesis

Novel view synthesis is a long-standing problem in computer graphics and vision. Learning-based methods have led to significant progress toward novel view synthesis [7, 11, 19, 22, 24, 25, 27, 31, 35, 36, 39, 45, 47, 49]. These methods used volume rendering to generate the rendered image and depth. NeRF++ [49] extended NeRF to unbounded scenes. It partitioned the scene space into two volumes, an inner unit sphere and an outer volume represented by an inverted sphere covering the complement of the inner volume. This strategy made NeRF++ handle unbounded scenes well. Similar to unbounded scenes, indoor scenes with dense free views also suffer from internal ambiguity. Neural sparse voxel field (NSVF) [24] learned the underlying voxel structures with a differentiable ray-marching operation from posed RGB images. However, it needed the ground-truth depth to initialize the volume bound, and its training process cost many computational resources.

Recently, COLMAP-based methods [10, 37, 44] adopt the sparse metric depth estimated from COLMAP [38]. The metric depth represents the actual distance between objects and the camera. It can supervise the rendered depth to recover the scene geometry. NerfingMVS (NMVS) [44] and dense depth priors (DDP) [37] trained an extra scene-specific depth network to get dense metric depth of a scene. Depth supervised NeRF (DS-NeRF) [10] used the sparse metric depth from COLMAP to supervise the rendered depth. These methods are limited by sparse views of a scene and the reconstruction quality of COLMAP. Dense views have better user interactions when compared with sparse views, while they encode more ambiguity for accurate rendering. When we applied these methods to dense free views, they failed when COLMAP estimated the inaccurate geometry. Depth Oracle NeRF (DONeRF) [28] degraded sharply without ground-truth depth as the supervision. In contrast to these approaches, our method removes the dependence on the COLMAP and ground-truth depth, by building the spatial context and modeling the internal relevance to handle dense free views of real-world scenes.

2.2 Learning-based monocular depth estimation

The depth estimation task is proposed to recover higher-dimensional depth information from low-dimensional image information. With big achievements made in machine learning and deep learning, the supervised learning methods had achieved impressive performance for depth estimation [6, 13, 15, 21, 23, 41]. These methods needed ground-truth depth maps for training, which was commonly tricky to acquire. Furthermore, these approaches worked well in specific scenes but did not generalize well to other scenes due to the limitation of training data. Unsupervised learning also made significant progress for depth estimation [2, 3, 17, 18, 32, 51]. These approaches did not need the ground truth depth map. During training, these methods applied differentiable warping and minimized photometric reprojection error. However, unsupervised learning suffers from limited generalization ability. For generalizable monocular depth estimation, Ranft et al. [33] and Yin et al. [46] proposed to mix several datasets to enhance generalization and used deep neural networks to predict depth. Our model takes camera parameters instead of RGB images as the input to estimate the relative depth of a view, with the supervision of results from the generalizable model [46].

3 Methodology

3.1 Revisiting neural radiance field (NeRF)

NeRF takes 5-dimensional (5D) vectors as the input, including 3D pixel coordinates and 2D view directions. 5D vectors are generated from the camera intrinsic parameters and poses, which are then formed as camera rays. Formally, NeRF models the view-dependent representations in a fixed-bound volume bound by implicit functions F_N: $(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, which maps a point $\mathbf{x}$ with the view direction $\mathbf{d}$ to RGB values $\mathbf{c}$ and a density value σ.

The implicit representations of $\mathbf{c}$ and σ are ray-traced to render each pixel p of an image. A camera ray can be formally defined as: $\mathbf{r}(s) = x_0 + s\mathbf{d}$, which denotes that the camera ray initialized with camera center x_0, sampled by the depth s and controlled by the unit camera direction vector $\mathbf{d}$. The volumetric radiance field along a camera ray can be rendered into an RGB image by:

$$\hat{I}(p) = \int_{d_n}^{d_f} T(s)\sigma(\mathbf{r}(s))c(\mathbf{r}(s), \mathbf{d})\mathrm{d}s, \tag{1}$$

where

$$T(s) = \exp\left(-\int_{d_n}^{s} \sigma(\mathbf{r}(h))\mathrm{d}h\right) \tag{2}$$

and a rendered depth map can also be generated as:

$$D^*(p) = \int_{d_n}^{d_f} T(s)\sigma(\mathbf{r}(s))s\mathrm{d}s \tag{3}$$

The integral is computed between pre-defined near and far depth d_n and d_f, which are usually two constants in NeRF. In practice, NeRF samples 3D points along each camera ray and performs this approximate integral with numerical quadrature [26].

NeRF consists of two MLPs, including a coarse sampling network and a refining network. The coarse network sparsely samples the volume with a fixed-bound grid and then learns the rough boundary of objects in a scene. The refining network utilizes the coarse density to produce a dense sampling pattern along the same camera ray according to the location of high-density gradients. In our work, we use relative depth to constrain the geometric volume bound (Sect. 3.2) and enhance the scene representations of MLPs by modeling the internal relevance of sampled points (Sect. 3.3).

3.2 Relative depth guided sampling and integration

In NeRF [27], the quality of scene geometry can be represented in the rendered depth, and the high-quality scene geometry usually exhibits well-rendered images. As shown in Fig. 1, NeRF estimates incorrect scene geometry and renders a noisy RGB image for dense free view synthesis. To remedy this issue, we improve NeRF in rendering scenes with complex camera motion by recovering correct scene geometry first and then use this prior to guide NeRF rendering images with more accurate geometry. Figure 3 shows the overview of our framework.

Most metric depth maps captured by 3D sensors or estimated by COLMAP are sparse [37, 44]. In contrast, the relative depth generated from the generalizable method [46] can supply dense and stable depth without expensive sensors. We adopt the relative depth instead of the sparse metric depth to recover the holistic scene depth. We utilize neural networks to learn the relative depth in training views and generate it in novel views. NeRF proves that neural networks can learn implicit functions for rendering images from camera parameters. Essentially, NeRF models a one-to-one correspondence from camera parameters to the rendered image at each viewpoint of a scene by implicit functions. We build another one-to-one correspondence from camera parameters to the relative depth by implicit functions at each free viewpoint. To obtain the view-dependent relative depth d_r, we design a sub-network with implicit functions F_D, which can be denoted as: $F_D : (x_0, \mathbf{d}) \rightarrow d_r$. It is composed of four fully connected layers followed by ReLU and takes the camera center and the view direction of each pixel as the input.

To utilize the learned relative depth guide the rendering procedure, we build a connection between the relative depth estimation and image synthesis. Because the relative depth models the geometric relationship between objects and the camera in a view, we consider transforming it as the relative distance to be a geometric view bound. When acquiring the learned relative depth, we normalize it as d_{rn} and use a linear transformation: $T_d : \{d_f = a * d_{rn} + b\}$, which maps the relative depth to the far depth d_f of Eqn. 1. This operation produces a geometric volume bound which encodes the spatial context between each camera ray of a view. With the help of the geometric volume bound, we can achieve the geometric-aware sampling and integrate more accurate rendered depth and rendered RGB images by volume rendering.

Related methods [37, 44] are limited by the quality of the metric depth from COLMAP. In contrast, we use the relative depth (relative distance) instead of the metric depth (actual distance) to guide both sampling and integration.

3.3 Modeling internal relevance of sampled points

In forward-facing or spherically captured scenes, the same objects in these scenes almost exist at each sparse viewpoint. It makes these scenes easy to be parameterized by learning to map sampled points to color values. In dense free views, the number of totally sampled points in the global volume increases dramatically which augments the ambiguity of mapping. It is extremely hard for NeRF to render accurate images due to the ambiguity [11].

Transformers [12] have been successfully applied to several vision tasks. The core part of the transformer is the multi-head self-attention mechanism. It can reduce the reliance on external information and capture the internal relevance of learned features to improve the representation ability of neural networks [40]. The geometric volume bound builds the spatial context of camera rays in each view. For the totally sampled points, we notice that the internal relevance of them along camera rays has not been well modeled in vanilla MLPs of NeRF. This relevance between each sampled point can supply more information for neural networks to handle the ambiguity from the large number of sampled points. Considering the advantages of the self-attention mechanism, we propose to utilize it for modeling the internal relevance of sampled points to boost the representation ability of vanilla MLPs.

As illustrated in Fig. 4, we model this relevance by integrating a spatial self-attention module into the backbone of NeRF. Firstly, we feed the embedded location into an initial layer to acquire features. Then, we use the multi-head attention module and short-cut connection following the initial layer to model the internal relevance. The rest components of our backbone are the same as NeRF.

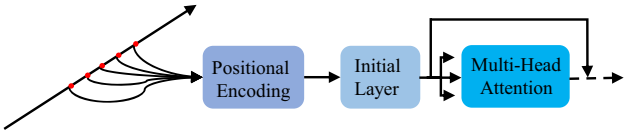

Fig. 4 Modeling the internal relevance of sampled points. We apply the self-attention mechanism to model the internal relevance of the sampled spatial points. Unlike other self-attention-based modules [34, 42], the input of our module only contains the sampled points without the source image features

Related self-attention modules [34, 42] rely on the limited source view features to build the context between each camera ray, while dense free views contain a large number of source views which makes our task ill-suited for these modules. In contrast to them, the input of our module only consists of the spatial points in a view since the spatial context between each camera ray has been well modeled by the geometric volume bound. Figure 5 shows the effectiveness of the modeled internal relevance; it faithfully helps NeRF render a more accurate view.

3.4 Loss function

Similar to NeRF, we optimize coarse and fine models simultaneously. For the rendering function F_N, the loss function L_N is defined as the mean squared error between the rendered and ground-truth color of pixels. For the relative depth estimation function F_D, the loss function L_D is the mean squared error between the predicted relative depth $\hat{D}_{\text{rel}}$ and the result D_{rel} of the generalizable model [46]. The loss functions of our method are formulated as follows:

$$L_N = \frac{1}{N_R} \sum_{r \in R} \left[||\hat{I}_c(r) - I_c(r)||_2^2 + ||\hat{I}_f(r) - I_f(r)||_2^2 \right] \tag{4}$$

$$L_D = \frac{1}{N_R} \sum_{r \in R} ||\hat{D}_{\text{rel}}(r) - D_{\text{rel}}(r)||_2^2 \tag{5}$$

$$L = L_N + \lambda L_D \tag{6}$$

where N_R is the total number of camera rays R. λ adjusts the weights between L_N and L_D. In practice, we set $\lambda = 0.1$.

4 Experiments

4.1 Datasets and preparation

We focus on dense views synthesis of real-world indoor scenes and conduct adequate experiments on 9 scenes from two datasets, including 7-Scenes [16] and ScanNet [8].
7-Scenes [16] is a collection of images with 640 × 480 resolution captured from the hand-held Kinect RGB-D camera,

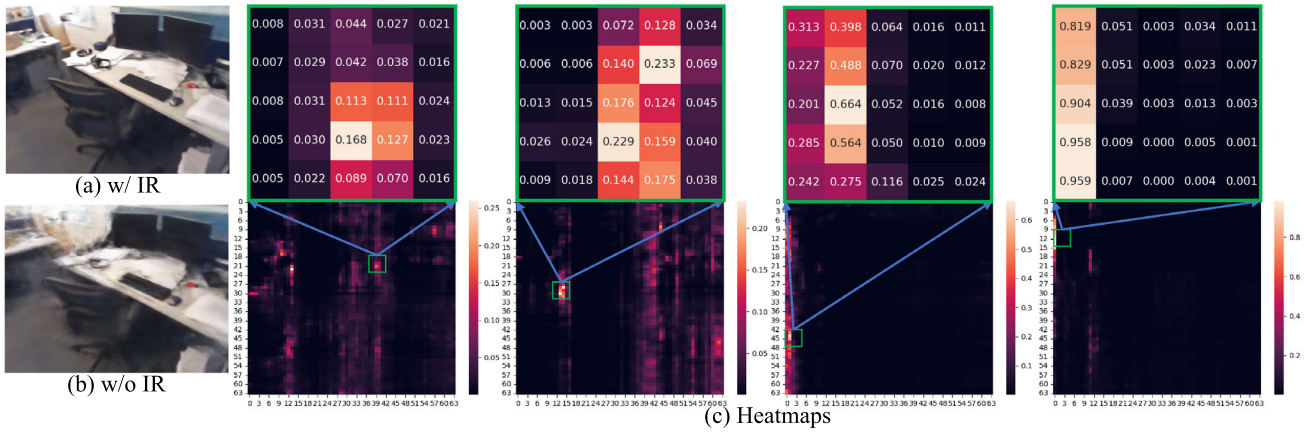

Fig. 5 Effect of modeling the internal relevance. 'IR': modeling internal relevance of sampled points along camera rays. **a** Effect of adding IR to NeRF. **b** The test view result of NeRF. **c** Heatmaps of different attention heads illustrate that the self-attention module faithfully models different relevance between a sampled point and other points along a camera ray. These internal relevances enhance the representation ability of MLPs and make RDNeRF achieve more accurate rendering

Table 1 Quantitative comparison of novel views between our method and other methods (NeRF [27], NeRF++ [49] and NMVS [44]) on 7-Scene [16]

Scene	PSNR↑				SSIM↑				LPIPS↓			
	NeRF	NeRF++	NMVS	Ours	NeRF	NeRF++	NMVS	Ours	NeRF	NeRF++	NMVS	Ours
Pumpkin	22.99	22.68	24.75	**25.27**	0.75	0.74	**0.81**	0.79	0.49	0.48	**0.41**	0.47
Office	22.59	21.23	19.76	**25.04**	0.76	0.74	0.65	**0.79**	**0.51**	0.53	0.53	0.54
Heads	22.96	22.57	21.10	**25.30**	0.77	0.76	0.76	**0.82**	0.48	0.50	**0.46**	0.47
Chess	22.22	21.96	22.36	**24.48**	0.70	0.69	0.77	**0.78**	0.49	0.49	**0.40**	0.43
Fire	22.31	22.18	21.36	**24.66**	0.62	0.62	**0.69**	0.68	0.54	0.53	**0.47**	0.52
Mean	22.61	22.12	21.87	**24.95**	0.72	0.71	0.74	**0.77**	0.50	0.51	**0.45**	0.49

Best results are bold. Our model outperforms other models with PSNR and SSIM on average

Table 2 Quantitative comparison of novel views between our method and recent approaches (DONeRF [28], DSNeRF [10] and NMVS [44]) on 7-Scene [16]

Method	PSNR↑	SSIM↑	LPIPS↓
DONeRF	21.49	0.72	0.533
DSNeRF	21.53	0.75	0.498
NMVS	21.87	0.74	**0.450**
Ours	**24.95**	**0.77**	0.490

Best results are bold. Our model outperforms other models with PSNR and SSIM on average

including ground-truth camera tracks. We select five scenes from 7-Scenes. Specifically, we pick one frame from every 20 frames for testing. The remaining frames are used for training. To reduce the frame similarity between the training and testing set, we drop the neighboring frames (within 10 frames) of every testing frame from the training set. The average number of training images of a scene is 510.

ScanNet [8] contains 1613 indoor scenes with ground-truth camera poses, depth maps, and RGB images. We follow the previous works [1, 9]) for choosing four scenes and splitting training, and testing datasets. The average number of training images of a scene is 2588.

4.2 Experiment setup

Evaluation metrics We report several quantitative metrics, including PSNR and SSIM [43] to measure the rendered image quality. Besides, we also introduce LPIPS [50], which can reflect the perception of humans more precisely for all evaluations. For the evaluation of the relative depth estimation, we follow the method [46], aligning the scale of the rendered and learned depth to the scale of ground-truth depth by a linear transformation. Then, we use the absolute mean relative error (AbsRel) and the percentage error of pixels ($\delta_1 = max(\frac{d_i}{d_i^*}, \frac{d_i^*}{d_i}) < 1.25$) as quantitative metrics.

Implementation details We re-implement all experiments of NeRF and RDNeRF based on the official code of NeRF++ [49]. The IRM consists of three parts: positional encoding, initial layer and multi-head attention. The positional encoding is following the traditional NeRF. The initial

Table 3 Quantitative comparison of novel views between our method and other methods (NeRF [27] and NeRF++ [49]) on ScanNet [8]

Scene	PSNR↑			SSIM↑			LPIPS↓		
	NeRF	NeRF++	Ours	NeRF	NeRF++	Ours	NeRF	NeRF++	Ours
Scan00	23.90	23.82	**24.49**	**0.72**	0.71	**0.72**	**0.48**	0.49	0.49
Scan10	25.32	25.13	**26.17**	0.81	0.81	**0.83**	**0.42**	0.44	0.43
Scan16	24.51	–	**24.91**	0.76	–	**0.77**	**0.42**	–	**0.42**
Scan24	21.39	–	**22.01**	**0.71**	–	**0.71**	**0.48**	–	0.50
Mean	23.78	–	**24.40**	0.75	–	**0.76**	**0.45**	–	0.46

Best results are bold. Our method outperforms the basic methods with PSNR and SSIM and achieves competitive results in LPIPS across most scenes. Due to NeRF++ reporting NaN loss during training on 'Scan16' and 'Scan24,' we fail to evaluate it in these two scenes

Table 4 Quantitative comparison of novel scene geometry between our method and other methods (NeRF [27] and NeRF++ [49]) on 7-Scene [16] and ScanNet [8]

Method	7-Scene		ScanNet (00,10)		ScanNet (16,24)	
	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑	AbsRel↓	δ_1 ↑
NeRF	17.39	73.18	26.67	79.50	38.26	78.80
NeRF++	18.36	71.54	29.20	73.53	–	–
Ours (Rendered)	13.55	83.20	24.47	86.03	35.45	85.01
Ours (Relative)	**9.85**	**90.92**	**23.85**	**86.43**	**34.46**	**86.00**

Best results are bold

layer is a linear layer. The multi-head attention is following the traditional Transformer [40]. All approaches use the same input views for model training in each scene. We train each method for 300K iterations with a batch size of 1024 and use the Adam optimizer. The initial learning rate is 0.0005, which decreased by a factor of 10 at 100k and 250k steps. All the experiments are conducted on a Tesla P100 GPU.

4.3 Comparison with related methods

Our proposed method is the composition of the dense free view synthesis and the relative depth estimation. We evaluate RDNeRF in indoor scenes for these two tasks against the basic model NeRF [27] and other related methods quantitatively and qualitatively on selected datasets. Because we focus on real-world indoor scenes, we use non-normalized device coordinates and set the far depth of NeRF as 10 m. 'a' and 'b' of the linear transformation T_d in our framework are set as 9 and 1, respectively.

NeRF++ [49] is an improved version of NeRF and adopts normalized device coordinates to handle real-world unbounded scenes. The scenes with dense free views can also be considered unbounded scenes because both of them suffer from internal ambiguity. We take it as the baseline to evaluate its performance in our task.

DDP [37] **and NerfingMVS** [44] **(NMVS)** utilize a depth completion model to generate dense metric depth from the sparse depth prior of COLMAP by the scene-specific training. Note that DDP needs the ground-truth depth to supervise the depth completion model. This setting is unfair to our method. NMVS adopts scene-specific finetuning without the ground-truth depth. Hence, we pick NerfingMVS for a fair comparison.

DS-NeRF [10] is proposed to tackle sparse views while we aim to deal with dense free views. We conduct a comparison with DS-NeRF [10] because it also drops the dependence on the ground-truth depth.

DO-NeRF [28] utilizes a modified NeRF to generate the scene depth firstly and then adopts it to render scenes. We also conduct a comparison with this method.

We evaluate all approaches in the tasks of dense free view synthesis at novel viewpoints. Given camera parameters at a novel viewpoint, our model renders an RGB image with its rendered depth and predicts a relative depth map from the same camera pose. We count the metrics of these outputs with the corresponding ground-truth data.

4.4 Experiments on 7-Scenes

Figure 6 shows the qualitative results of our method and other methods on 7-Scenes [16]. NeRF and NeRF++ fail to extract the correct scene geometry. The rendered depth of NMVS is over-smoothing and has unstable quality. Our model achieves more accurate geometry in both learned relative depth and rendered depth. The relative depth plays an important role in rendering RGB images. When scene geometry cannot be modeled correctly, more artifacts and errors occur in the rendered RGB image. NeRF and NeRF++ generate noisy and blurry images with ghosting artifacts, especially on the chessboard and bookcase of the top scene and the white walls of the

Table 5 Quantitative comparison of novel views between our method and different settings of our method on 7-Scene and ScanNet

Method	7-scene			ScanNet		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	22.59	0.76	0.51	25.58	0.83	0.43
+GVB	23.16	0.74	0.52	25.67	0.84	0.42
+IRM	24.46	0.79	0.56	26.07	0.84	0.44
Full	25.04	0.79	0.54	26.25	0.84	0.44

IRM, internal relevance module; GVB, geometric volume bound

bottom scene. On the display screen of the top scene, the local color shifts obviously appear in their rendered RGB images. When the correct scene geometry cannot be extracted, the quality of RGB images synthesized by NMVS decreases rapidly. With the correct spatial context from the learned relative depth, our method generates cleaner, sharper objects and smoother backgrounds than other approaches.

Quantitative results of novel view synthesis and geometry estimation are reported in Tables 1 and 4. Our method outperforms the previous methods in terms of PSNR for all scenes. In detail, RDNeRF achieves over 2.3dB better than NeRF in terms of PSNR and reaches up to 2% advancement than NMVS in terms of SSIM. In Table 4, we can see that the learned relative depth of RDNeRF surpasses the rendered depth from NeRF and NeRF++ on all metrics. Besides, our rendered depth is more accurate than the rendered depth of NeRF, which demonstrates that our model extracts more accurate scene geometry. Table 2 presents the results of our method and recent related approaches on novel view synthesis. RDNeRF achieves the best performance and outperforms state-of-the-art methods over 3dB in terms of PSNR.

4.5 Experiments on ScanNet

Results on ScanNet [8] for novel view synthesis are shown in Fig. 7. Due to the huge number of training views, DS-NeRF [10] and NMVS [44] fail to handle these scenes. NeRF and NeRF++ produce results with global and local color shifts. These effects easily make the result unreal at first glance. Our method preserves the boundaries of objects, while basic models suffer from ghosting artifacts in these regions, for example, the guitar and chairs in the scenes shown.

Table 3 reports the quantitative results on ScanNet [8] about synthesized novel views. Our method outperforms NeRF over 0.6dB on PSNR and achieves competitive performance on SSIM and LPIPS on average. NeRF++ fails to converge in the last two scenes, and our results also have better performance on all metrics in the other two scenes than it. For scene geometry estimation in Table 4, our learned depth and rendered depth both achieve better performance than the rendered depth of basic models.

Table 6 Quantitative comparison of relative depth on 'Pumpkin' of 7-Scene [16]

Method	AbsRel↓	δ_1 ↑
DP_2	8.21	95.69
+GVB (Ours)	6.31	97.97

GVB, geometric volume bound; DP_2, relative depth estimation by the original generalizable model [46] with the rendered RGB image of our model (**w/o** GVB) as the input

Moreover, we conduct a user study to compare our method against NeRF and NeRF++ on the visual video quality. We render four videos of two different scenes of ScanNet in novel views for each method. In the user study, we present these videos, including the results of three methods, and invite the volunteer to select the best sub-video at each time. And they are requested to consider which one is most similar to the ground-truth video. We invite 27 participants to do the user study. Each user picks four times with four videos in a survey. The total number of picking times is accumulated to 108. Our method gets 75 picks and achieves more picks than basic models in total. It indicates that our method can render more realistic scenes than NeRF.

4.6 Ablation study

Our full model consists of two parts, including the geometric volume bound (GVB) and the internal relevance module (IRM). To investigate the impact of each component of our full model, we conduct the ablation study in a scene of 7-Scene by enabling each component, respectively, to the basic model and showing how performance improves. With each component embedded, the performance is enhanced reasonably. Quantitative results are reported in Table 5. Qualitative results are shown in Figs. 8, 9 and 10.

Effect of the internal relevance module (IRM) To verify the necessity of modeling the internal relevance, we integrate our designed internal relevance module into the backbone of NeRF ('+IRM'). The result is reported in

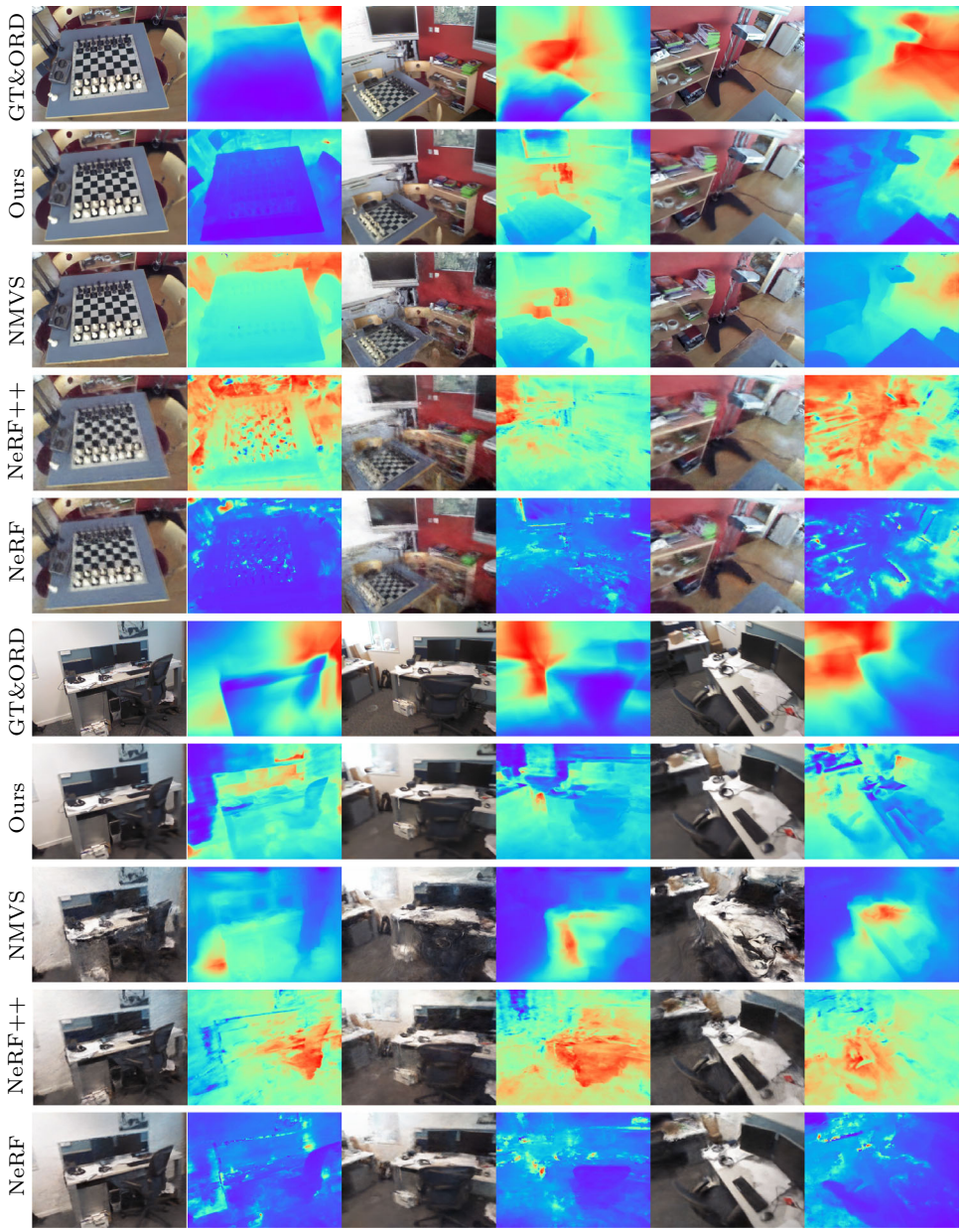

Fig. 6 Qualitative comparison of novel views and relative depth maps between our method and other methods (NeRF [27], NeRF++ [49] and NMVS [44]) on 7-Scene [16]. 'ORD': Our relative depth, which is predicted by F_D. The colorized depth maps shown are all from the rendered depth except our relative depth

Table 5. With the help of effective internal relevance, the performance increases significantly. Specifically, PSNR elevates over 1.8dB when compared to the basic model. However, without the spatial context, the performance of LIPIS degrades.

Effect of the geometric volume bound (GVB) To build the spatial context between each camera ray, we connect the learned relative depth and the volume bound to generate the geometric volume bound (GVB). On the one hand, the geometric volume bound improves the accuracy of extracted scene geometry, as reported in Table 6. Figure 8 illustrates objects that have sharper boundaries in the learned relative depth with GVB ('Full') than objects without GVB. On the other hand, the qualitative results in Fig. 9 show more realistic and geometric renderings (including RGB images and rendered depth maps) when adding GVB. This demonstrates that the geometric volume bound is effective in recovering the correct scene geometry for accurate rendering. As Table 5 presents, our full model ('IRM + GVB') achieves the best performance on PSNR and SSIM with comparable LPIPS.

4.7 Analysis

Parameters and effects Compared to NeRF, RDNeRF's parameters only have a 0.1 M gain, while RDNeRF renders more accurate RGB images than NeRF.

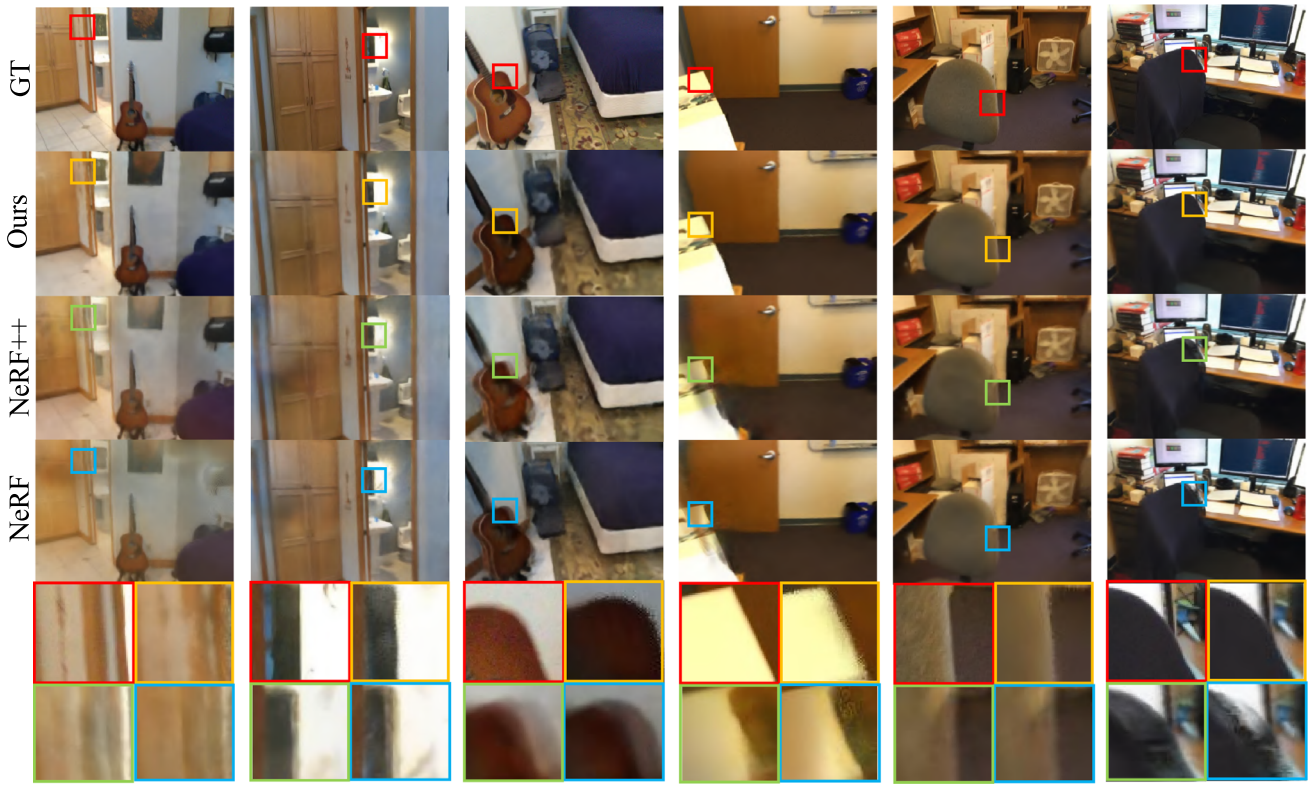

Fig. 7 Qualitative comparison of novel views between our method and other methods (NeRF [27] and NeRF++ [49]) on ScanNet [8]. Our method performs good results in occlusion areas, whereas other methods generate more obvious color shifts and blurry effects

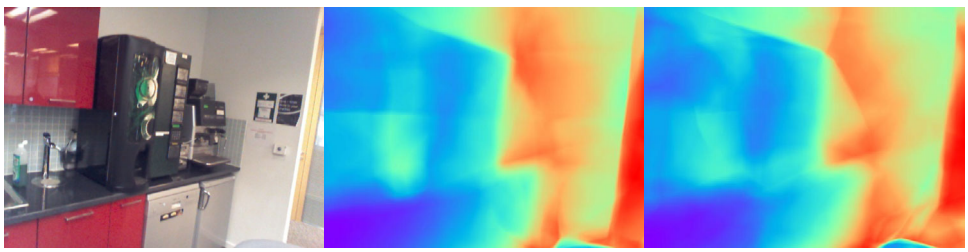

Fig. 8 Qualitative comparison of learned relative depth maps in novel views between our full model and different settings of it. (a) RGB image. (b) Depth map from our full model. (c) Depth map without GVB. We feed (a) into the generalized model [46] and obtain the relative depth as the supervision of our model for scene geometry estimation

Limitations Although we preserve the object boundary in novel views well, our method is mainly limited by the basic model parameters for handling the scenes with large number of training views on ScanNet. The failure cases are shown in Fig. 11. These two views are very different from training views then suffer from artifacts.

5 Conclusion

In this paper, we propose RDNeRF, an end-to-end framework built upon NeRF for dense free view synthesis in real-world indoor scenes. RDNeRF improves the performance of NeRF to render more photo-realistic novel views and recover more accurate scene geometry. Specifically, we adopt the implicit functions to learn the relative depth and transform it to be the volume bound of rendering. With the spatial context built by the geometric volume bound, we further construct the internal relevance of camera rays by a spatial self-attention module. We conduct extensive experiments to evaluate our method and compare it with relative approaches. Our method outperforms other methods in the novel view synthesis and scene geometry extraction qualitatively and quantitatively.

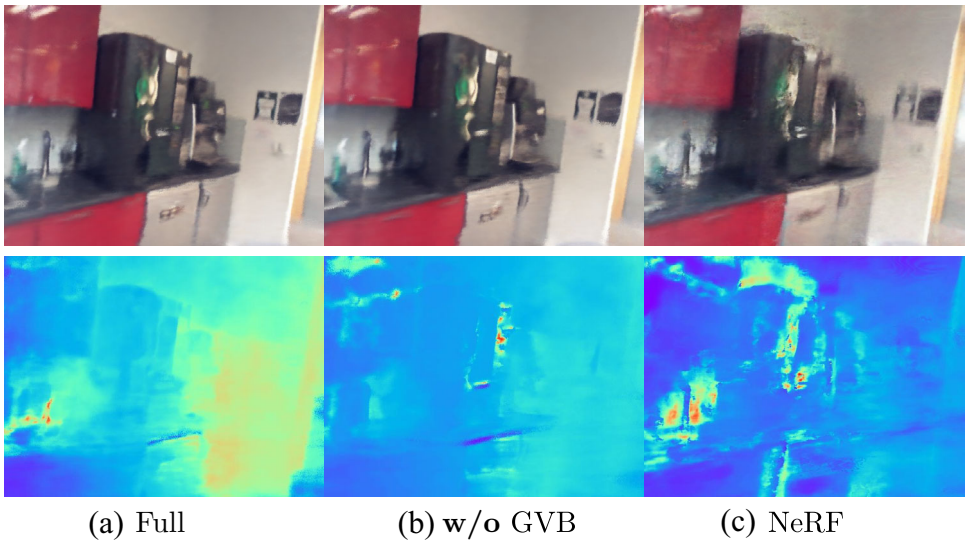

Fig. 9 Qualitative comparison of novel view synthesis (top) and rendered depth maps (bottom) between our full model and different settings of it. (a) Our full model. (b) Without GVB. (c) Original NeRF

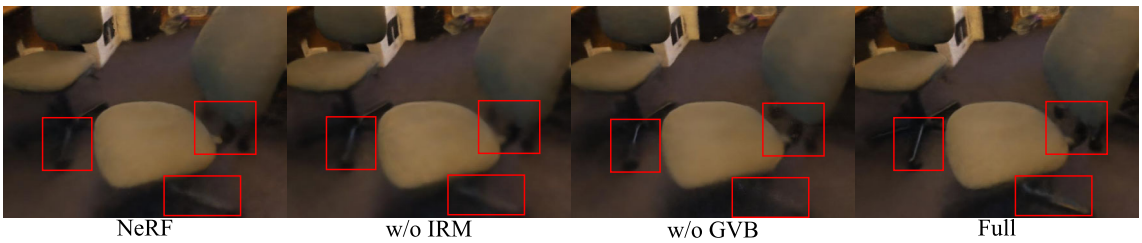

Fig. 10 Qualitative comparison of novel view synthesis between our full model and different settings of it on ScanNet

Fig. 11 Failure cases on ScanNet

Author Contributions JQ and YZ were involved in conceiving, designing the analysis and writing; P-TJ and BR contributed to writing—review and editing; M-MC assisted in the supervision.

Funding This work is supported by the National Key Research and Development Program of China Grant (No. 2018AAA0100400), NSFC (No.61922046) and NSFC (No. 62132012).

Data availability We used two common datasets in this work: 7-Scenes [16] https://www.microsoft.com/en-us/research/project/rgb-d-dataset-7-scenes and ScanNet [8] http://www.scan-net.org/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Aliev, KA., Ulyanov, D., Lempitsky, V.: Neural point-based graphics 2(3):4 (2019). arXiv preprint arXiv:1906.08240
2. Andraghetti, L., Myriokefalitakis, P., Dovesi, P.L., et al.: Enhancing self-supervised monocular depth estimation with traditional visual odometry. In: 2019 International Conference on 3D Vision (3DV), pp. 424–433. IEEE (2019)
3. Battiato, S., Curti, S., La Cascia, M., et al.: Depth map generation by image classification. In: Three-Dimensional Image Capture and Applications VI, International Society for Optics and Photonics, pp. 95–104 (2004)
4. Chan, S., Shum, H.Y., Ng, K.T.: Image-based rendering and synthesis. IEEE Signal Process. Mag. **24**(6), 22–33 (2007)
5. Chen, D., Sang, X., Wang, P., et al.: Dense-view synthesis for three-dimensional light-field display based on unsupervised learning. Opt. Express **27**(17), 24,624–24,641 (2019)
6. Chen, W., Fu, Z., Yang, D., et al.: Single-image depth perception in the wild. arXiv preprint arXiv:1604.03901 (2016)

7. Chen, Z., Wang, C., Guo, Y.C, et al.: Structnerf: Neural radiance fields for indoor scenes with structural hints. arXiv preprint arXiv:2209.05277 (2022)
8. Dai, A., Chang, A.X., Savva, M., et al.: Scannet: Richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5828–5839 (2017)
9. Dai, P., Zhang, Y., Li, Z., et al.: Neural point cloud rendering via multi-plane projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 7830–7839 (2020)
10. Deng, K., Liu, A., Zhu, J.Y., et al.: Depth-supervised nerf: Fewer views and faster training for free. arXiv preprint arXiv:2107.02791 (2021)
11. DeVries, T., Bautista, M.A., Srivastava, N., et al .: Unconstrained scene generation with locally conditioned radiance fields. arXiv preprint arXiv:2104.00670 (2021)
12. Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al .: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
13. Eigen, D., Fergus, R .: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2650–2658 (2015)
14. Flynn, J., Broxton, M., Debevec, P., et al.: Deepview: View synthesis with learned gradient descent. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2367–2376 (2019)
15. Fu, H., Gong, M., Wang, C., et al .: Deep ordinal regression network for monocular depth estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2002–2011 (2018)
16. Glocker, B., Izadi, S., Shotton, J., et al.: Real-time RGB-D camera relocalization. In: 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR), pp. 173–179. IEEE (2013)
17. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)
18. Gordon, A., Li, H., Jonschkowski, R., et al.: Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 8977–8986 (2019)
19. Hedman, P., Philip, J., Price, T., et al.: Deep blending for free-viewpoint image-based rendering. ACM Trans. Graph. (TOG) **37**(6), 1–15 (2018)
20. Kajiya, J.T., Von Herzen, B.P.: Ray tracing volume densities. ACM SIGGRAPH Comput. Graph. **18**(3), 165–174 (1984)
21. Laina, I., Rupprecht, C., Belagiannis, V., et al.: Deeper depth prediction with fully convolutional residual networks. In: 2016 Fourth International Conference on 3D Vision (3DV), pp. 239–248. IEEE (2016)
22. Lindell, D.B., Martel, J.N., Wetzstein, G.: Autoint: Automatic integration for fast neural volume rendering. arXiv preprint arXiv:2012.01714 (2020)
23. Liu, F., Shen, C., Lin, G., et al.: Learning depth from single monocular images using deep convolutional neural fields. IEEE Trans. Pattern Anal. Mach. Intell. **38**(10), 2024–2039 (2015)
24. Liu, L., Gu, J., Lin, K.Z., et al.: Neural sparse voxel fields. arXiv preprint arXiv:2007.11571 (2020)
25. Martin-Brualla, R., Radwan, N., Sajjadi, M.S., et al .: Nerf in the wild: Neural radiance fields for unconstrained photo collections. arXiv preprint arXiv:2008.02268 (2020)
26. Max, N.: Optical models for direct volume rendering. IEEE Trans. Visual Comput. Graph. **1**(2), 99–108 (1995)
27. Mildenhall, B., Srinivasan, P.P., Tancik, M., et al.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision, pp. 405–421. Springer (2020)
28. Neff, T., Stadlbauer, P., Parger, M., et al.: Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. In: Computer Graphics Forum, Wiley Online Library, pp. 45–59 (2021)
29. Nguyen, H.T., Do, M.N.: Error analysis for image-based rendering with depth information. IEEE Trans. Image Process. **18**(4), 703–716 (2009)
30. Penner, E., Zhang, L.: Soft 3d reconstruction for view synthesis. ACM Trans. Graph. (TOG) **36**(6), 1–11 (2017)
31. Pumarola, A., Corona, E., Pons-Moll, G., et al.: D-nerf: Neural radiance fields for dynamic scenes. arXiv preprint arXiv:2011.13961 (2020)
32. Qi, X., Liao, R., Liu, Z., et al.: Geonet: Geometric neural network for joint depth and surface normal estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 283–291 (2018)
33. Ranftl, R., Lasinger, K., Hafner, D., et al.: Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2020)
34. Reizenstein, J., Shapovalov, R., Henzler, P., et al.: Common objects in 3D: Large-scale learning and evaluation of real-life 3D category reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10,901–10,911 (2021)
35. Riegler, G., Koltun, V.: Free view synthesis. In: European Conference on Computer Vision (2020)
36. Riegler, G., Koltun, V.: Stable view synthesis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2021)
37. Roessle, B., Barron, J.T., Mildenhall, B., et al.: Dense depth priors for neural radiance fields from sparse input views. arXiv preprint arXiv:2112.03288 (2021)
38. Schonberger, J.L., Frahm, J.M .: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
39. Srinivasan, P.P., Tucker, R., Barron , J.T., et al.: Pushing the boundaries of view extrapolation with multiplane images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 175–184 (2019)
40. Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
41. Wang, C., Lucey, S., Perazzi, F., et al.: Web stereo video supervision for depth prediction from dynamic scenes. In: 2019 International Conference on 3D Vision (3DV), pp. 348–357. IEEE (2019)
42. Wang, P., Chen, X., Chen, T., et al.: Is attention all nerf needs? arXiv preprint arXiv:2207.13298 (2022)
43. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. **13**(4), 600–612 (2004)
44. Wei, Y., Liu, S., Rao, Y., et al.: Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5610–5619 (2021)
45. Wu, X., Xu, J., Zhu, Z., et al.: Scalable neural indoor scene rendering. ACM Trans. Graph. (TOG) **41**(4), 1–16 (2022)
46. Yin, W., Zhang, J., Wang, O., et al.: Learning to recover 3d scene shape from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 204–213 (2021)
47. Yu, A., Ye, V., Tancik, M., et al.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4578–4587 (2021)

48. Zhang, C., Chen, T.: A survey on image-based rendering-representation, sampling and compression. Signal Process. Image Commun. **19**(1), 1–28 (2004)
49. Zhang, K., Riegler, G., Snavely, N., et al.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
50. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
51. Zhou, T., Brown, M., Snavely, N., et al.: Unsupervised learning of depth and ego-motion from video. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1851–1858 (2017)
52. Zhou, T., Tucker, R., Flynn, J., et al.: Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817 (2018)

Jiaxiong Qiu is a Ph.D. student from the College of Computer Science at Nankai University. He received his master degree supervised at University of Electronic Science and Technology of China in 2020. He obtained his bachelor degree at Dalian Maritime University in 2017. His research interests include computer vision, computer graphics, robotics, and deep learning.

Yifan Zhu is a master student from the College of Computer Science at Nankai University under the supervision of Bo Ren. Before that, he received the bachelor's degree from Hebei University of Technology in 2020. His research interests include deep learning and computer vision.

Peng-Tao Jiang is a Post-doc Researcher at Zhejiang University, working with Prof. Chunhua Shen. Prior to that, He received my Ph.D. at Nankai University, advised by Prof. Ming-Ming Cheng. Moreover, He has spent several months as an intern at SenseTime and Tencent YouTu. His research interests include explainable AI (XAI), weakly supervised semantic segmentation, regularization tool for image classification.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. Then, he did 2-year research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program. He is on the editorial boards of IEEE TIP.

Bo Ren received the PhD degree from Tsinghua University in 2015. He is currently an associate professor in the College of Computer Science, Nankai University, Tianjin. His research interests include physically based simulation, 3D scene reconstruction and analysis.