
DäRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation

Jiuhn Song* Seonghoon Park* Honggyu An*
 Seokju Cho Min-Seop Kwak Sungjin Cho Seungryong Kim†

Korea University

Abstract

Neural radiance field (NeRF) shows powerful performance in novel view synthesis and 3D geometry reconstruction, but it suffers from critical performance degradation when the number of known viewpoints is drastically reduced. Existing works attempt to overcome this problem by employing external priors, but their success is limited to certain types of scenes or datasets. Employing monocular depth estimation (MDE) networks, pretrained on large-scale RGB-D datasets, with powerful generalization capability would be a key to solving this problem: however, using MDE in conjunction with NeRF comes with a new set of challenges due to various ambiguity problems exhibited by monocular depths. In this light, we propose a novel framework, dubbed DäRF, that achieves robust NeRF reconstruction with a handful of real-world images by combining the strengths of NeRF and monocular depth estimation through online complementary training. Our framework imposes the MDE network’s powerful geometry prior to NeRF representation at both seen and unseen viewpoints to enhance its robustness and coherence. In addition, we overcome the ambiguity problems of monocular depths through patch-wise scale-shift fitting and geometry distillation, which adapts the MDE network to produce depths aligned accurately with NeRF geometry. Experiments show our framework achieves state-of-the-art results both quantitatively and qualitatively, demonstrating consistent and reliable performance in both indoor and outdoor real-world datasets. Project page is available at <https://ku-cvlab.github.io/DaRF/>.

1 Introduction

Neural radiance field (NeRF) [31] has gained significant attention for its powerful performance in reconstructing 3D scenes and synthesizing novel views. However, despite its impressive performance, NeRF often comes with a considerable limitation in that its performance highly relies on the presence of densely well-calibrated input images which are difficult to acquire. As the number of input images is reduced, NeRF’s novel view synthesis quality drops significantly, displaying failure cases such as erroneous overfitting to the input images [18, 33], artifacts clouding empty spaces [33], or degenerate geometry that yields incomprehensible jumble when rendered at unseen viewpoints [19]. These challenges derive from its under-constrained nature, causing it to have extreme difficulty mapping a pixel in input images to a correct 3D location. In addition, NeRF’s volume rendering allows the model to map a pixel to multiple 3D locations [14], exacerbating this problem.

Previous *few-shot* NeRF methods attempt to solve these issues by imposing geometric regularization [33, 19, 23] or exploiting external 3D priors [14, 41] such as depth information extracted from

*Equal contribution

†Corresponding author

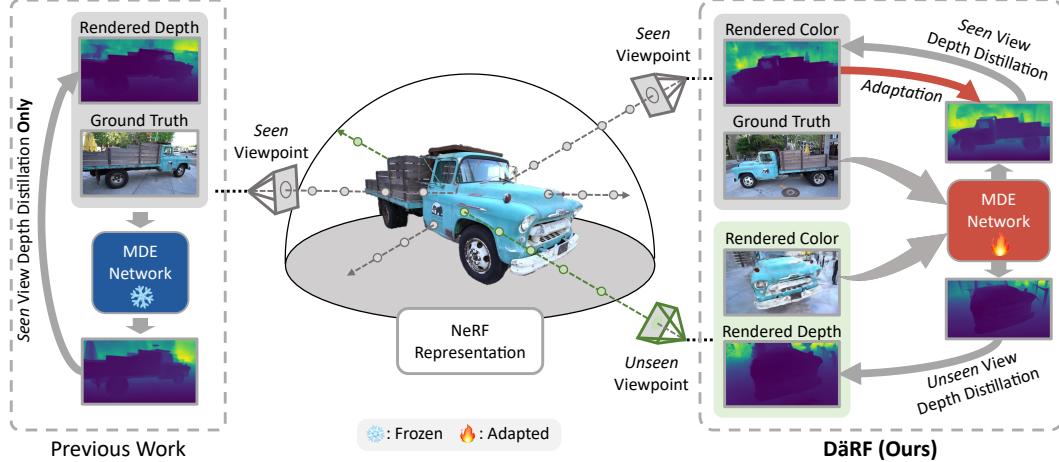


Figure 1: **Motivation.** Unlike existing work [50] that distills depths by pretrained MDE to NeRF at seen view only, our DäRF fully exploits the ability of MDE by jointly optimizing NeRF and MDE at a specific scene, and distilling the monocular depth prior to NeRF at both seen and unseen views.

input images by COLMAP [43]. However, these methods have weaknesses in that they use 3D priors extracted from a few input images only, which prevents such guidance from encompassing the entire scene. To effectively tackle all the issues mentioned above, pretrained monocular depth estimation (MDE) networks with strong generalization capability [39, 38, 5] could be used to inject an additional 3D prior into NeRF that facilitates robust geometric reconstruction. Specifically, geometry prediction by MDE can constrain NeRF into recovering smooth and coherent geometry, while their bias towards predicting smooth geometry helps to filter out fine-grained artifacts that clutter the scene. More importantly, NeRF’s capability to render any unseen viewpoints enables fully exploiting the capability of the MDE, as MDE could provide depth prior to the numerous renderings of unseen viewpoints as well as the original input viewpoints. This allows injecting additional 3D prior to effectively covering the entire scene instead of being constrained to a few input images.

However, applying MDE to few-shot NeRF is not trivial, as there are ambiguity problems that hinder the monocular depth from serving as a good 3D prior. Primarily, relative depths predicted by MDEs are not multiview-consistent [6, 12]. Moreover, MDEs perform poorly in estimating depth differences between multiple objects: this prevents global scale-shift fitting [62, 30] from being a viable solution, as alignment to one region of the scene inevitably leads to misalignment in many other regions. There also exists a convexity problem [30], in which the MDE has difficulty determining whether the surface is planar, convex, or concave, are also present. To overcome these challenges, we introduce a novel method to adapt MDE to NeRF’s absolute scaling and multiview consistency as NeRF is regularized by MDE’s powerful 3D priors, creating a complementary cycle.

In this paper, we propose DäRF, short for Monocular Depth Adaptation for boosting Radiance Fields from Sparse Input Views, which achieves robust optimization of few-shot NeRF through MDE’s geometric prior, as well as MDE adaptation for alignment with NeRF through complementary training (see Fig. 1). We exploit MDE for robust geometry reconstruction and artifact removal in both *unseen* and *seen* viewpoints. In addition, we leverage NeRF to adapt MDE toward multiview-consistent geometry prediction and introduce novel patch-wise scale-shift fitting to more accurately map local depths to NeRF geometry. Combined with a confidence modeling technique for verifying accurate depth information, our method achieves state-of-the-art performance in few-shot NeRF optimization. We evaluate and compare our approach on real-world indoor and outdoor scene datasets, establishing new state-of-the-art results for the benchmarks.

2 Related Work

Neural radiance field. Neural radiance field (NeRF) [31] represents photo-realistic 3D scenes with MLP. Owing to its remarkable performance, there has been a variety of follow-up studies [3, 59, 29]. These studies improve NeRF such as dynamic and deformable scenes [35, 49, 37, 2], real-time rendering [59, 40, 32], unbounded scene [4, 46, 54] and generative modeling [44, 34, 56, 8]. However,

these works still encounter challenges in synthesizing novel views with a limited number of images in a single scene, limiting their applicability in real-world scenarios.

Few-shot NeRF. Numerous *few-shot* NeRF works attempted to address few-shot 3D reconstruction problem through various techniques, such as pretraining external priors [60, 11], meta-learning [47], regularization [18, 33, 19, 57, 23] or off-the-shelf modules [18, 33]. Recent approaches [33, 19, 57, 23] emphasize the importance of geometric consistency and apply geometric regularization at unknown viewpoints. However, these regularization methods show limitations due to their heavy reliance on geometry information recovered by NeRF. Other works such as DS-NeRF [14], DDP-NeRF [41] and SCADE [50] exploit additional geometric information, such as COLMAP [43] 3D points or monocular depth estimation, for geometry supervision. However, these works have critical limitations of only being able to provide geometry information corresponding to existing input viewpoints. Unlike these works, our work demonstrates methods to provide geometric prior even at unknown viewpoints with MDE for more effective geometry reconstruction.

Monocular depth estimation. Monocular depth estimation (MDE) is a task that aims to predict a dense depth map given a single image. Early works on MDE used handcrafted methods such as MRF for depth estimation [42]. After the advent of deep learning, learning-based approaches [15, 17, 20, 24] were introduced to the field. In this direction, the models were trained on ground-truth depth maps acquired by RGB-D cameras or LiDAR sensors to predict absolute depth values [27, 26]. Other approaches trained the networks on large-scale diverse datasets [9, 25, 38, 39], which demonstrates better generalization power. These approaches struggle with depth ambiguity caused by ill-posed problem, so the following works LeRes [58] and ZoeDepth [5] opt to recover absolute depths using additional parameters.

Incorporating MDE into 3D representation. As both NeRF and monocular depth estimation are closely related, there have been some works that utilize MDE models to enhance NeRF’s performance. NeuralLift [55] and MonoSDF [61] leverage depths predicted by pretrained MDE for depth ordering and detailed surface reconstruction, respectively. Other works optimize scene-specific parameters, such as depth predictor utilizing depth recovered by COLMAP [52] or learnable scale-shift values for reconstruction in noisy pose setting [7]. As a concurrent work, SCADE [50] utilizes monocular depths for scene reconstruction by providing explicit geometry priors for sparse view inputs. However, these previous approaches were limited in that MDEs were used to provide prior to only the input viewpoints, which constrains their effectiveness when input views are reduced, e.g., in the few-shot setting. In addition, they only used pretrained MDE models without fine-tuning on a specific scene, which inherits the limitations of pretrained MDE, such as estimating relative and view-inconsistent depths [39, 38]. In contrast, our method exploits pretrained MDEs even at unknown viewpoints, along with finetuning for absolute depth prediction, for effective suppression of artifacts and divergent behaviors of few-shot NeRF.

3 Preliminaries

NeRF [31] represents a scene as a continuous function $\mathcal{F}_\theta(\cdot)$ represented by a neural network with parameters θ . During optimization, 3D points are sampled along rays represented by \mathbf{r} coming from a set of input images $\mathcal{S} = \{I_i\}$, whose ground truth camera poses are given, for evaluation by the neural network. For each sampled point, $\mathcal{F}_\theta(\cdot)$ takes as input its coordinate $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{R}^2$ with a positional encoding $\gamma(\cdot)$ that facilitates learning high-frequency details [48], and outputs a color $\mathbf{c} \in \mathbb{R}^3$ and a density $\sigma \in \mathbb{R}$ such that $\{\mathbf{c}, \sigma\} = \mathcal{F}_\theta(\gamma(\mathbf{x}), \gamma(\mathbf{d}))$. With a ray parameterized as $\mathbf{r}_p(t) = \mathbf{o} + t\mathbf{d}_p$, starting from camera center \mathbf{o} along the direction \mathbf{d}_p , color and depth value at the pixel \mathbf{p} are rendered as follows:

$$\bar{I}(\mathbf{p}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_p(t))\mathbf{c}(\mathbf{r}_p(t))dt, \quad \bar{D}(\mathbf{p}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_p(t))tdt, \quad (1)$$

where $\bar{I}(\mathbf{p})$ and $\bar{D}(\mathbf{p})$ are rendered color and depth values at the pixel \mathbf{p} along the ray $\mathbf{r}_p(t)$ from t_n to t_f , and $T(t)$ denotes an accumulated transmittance along the ray from t_n to t as follows:

$$T(t) = \exp \left(- \int_{t_n}^t \sigma(\mathbf{r}_p(s))ds \right). \quad (2)$$

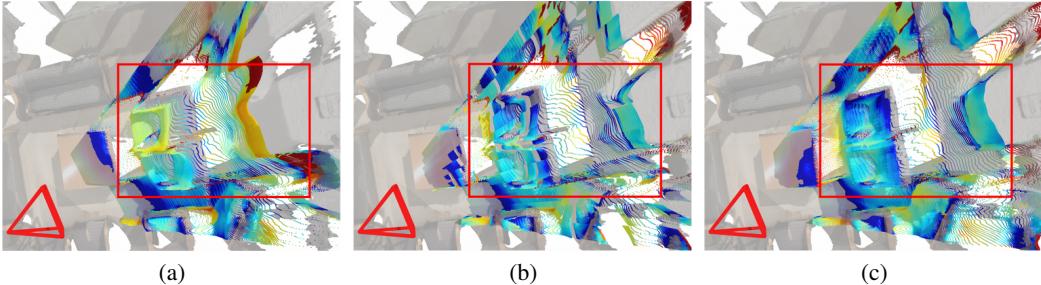


Figure 2: **Visualization of the effectiveness of patch-wise scale and shift adjustment:** (a) monocular depth with image-level adjustment, (b) monocular depth with patch-level adjustment, and (c) rendered depth by NeRF trained with patch-level adjustment. For visualization, depth maps are unprojected into 3D space. The proposed patch-wise adjustment helps to minimize the errors caused by inconsistency in depth differences among objects.

Based on this volume rendering, $\mathcal{F}_\theta(\cdot)$ is optimized by the reconstruction loss $\mathcal{L}_{\text{recon}}$ that compares rendered color $\bar{I}(\mathbf{p})$ to corresponding ground-truth $I(\mathbf{p})$, with \mathcal{R} as a set of pixels for training rays:

$$\mathcal{L}_{\text{recon}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{R}} \|I_i(\mathbf{p}) - \bar{I}_i(\mathbf{p})\|_2^2. \quad (3)$$

Our work explores the setting of few-shot optimization with NeRF [19, 23]. Whereas the number of input viewpoints $|\mathcal{S}|$ is normally higher than one hundred in the standard NeRF setting [31], the task of few-shot NeRF considers scenarios when $|\mathcal{S}|$ is drastically reduced to a few viewpoints (e.g., $|\mathcal{S}| < 20$). With such a small number of input viewpoints, NeRF shows high divergent behaviors such as geometry breakdown, overfitting to input viewpoints, and generation of artifacts that cloud the empty space between the camera and object, which causes its performance to drop sharply [18, 19, 33]. To overcome this problem, existing few-shot NeRF frameworks applied regularization techniques at unknown viewpoints to constrain NeRF with additional 3D priors [41, 14] and enhance the robustness of geometry, but they showed limited performance.

4 Methodology

4.1 Motivation and Overview

Our framework leverages the complementary benefits of few-shot NeRF and monocular depth estimation networks for the goal of robust 3D reconstruction. The benefits that pretrained MDE can provide to few-shot NeRF are clear and straightforward: because they predict dense geometry, they provide guidance for the NeRF to recover more smooth geometry. In cases where few-shot NeRF’s geometry undergoes divergent behaviors, MDE provides strong constraints to prevent the global geometry from breaking down.

However, there are difficult challenges that must be overcome if the depths estimated by MDE are to be used as 3D prior to NeRF. These challenges, which can be summarized as depth ambiguity problems [30], stem from the inherent ill-posed nature of the monocular depth estimation. Most importantly, MDE networks only predict relative depth information inferred from an image, meaning it is initially not aligned to NeRF’s absolute geometry [5]. Global scaling and shifting may seem to be the answer, but this approach leads us to another depth ambiguity problem, as predicted scales and spacings of each instance are inconsistent with one another, as demonstrated in (a) of Fig. 2. Additionally, MDE’s weakness in predicting the convexity of a surface, whether it is flat, convex, or concave - also poses a problem in using this depth for NeRF guidance.

In this light, we adapt a pre-trained monodepth network to a single NeRF scene so that its powerful 3D prior can be leveraged to its maximum capability in regularizing the few-shot NeRF. In the following, we first explain how to distill geometric prior from off-the-shelf MDE model [39] from both seen and unseen viewpoints (Sec. 4.2). We also provide a strategy for adapting the MDE model to handle ill-posed problems to a specific scene, while keeping its 3D prior knowledge (Sec. 4.3). Then, we demonstrate a method to handle inaccurate depths (Sec. 4.4). Fig. 1 shows an overview of our method, compared to previous works using MDE prior [50, 61].

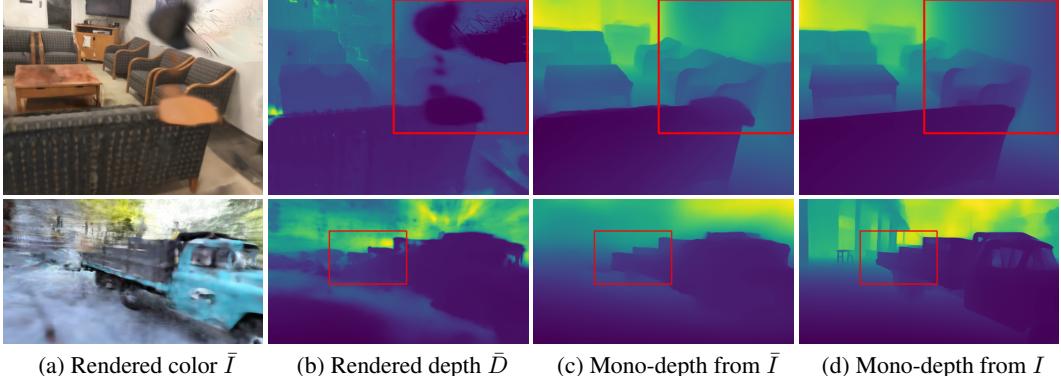


Figure 3: **Robustness of MDE model for multi-view scale ambiguity and artifacts:** (a-b) color and depth of NeRF rendered in the early stage of the training, (c-d) monocular depths estimated from rendered image \bar{I} and input image I . The results show that MDE model ignores the artifacts of rendered images by NeRF, enabling reliable supervision for seen and unseen viewpoint.

4.2 Distilling Monocular Depth Prior into Neural Radiance Field

To prevent the degradation of reconstruction quality in few-shot NeRF, we propose to distill monocular depth prior to the neural radiance field during optimization. By exploiting pre-trained MDE networks [38, 39], which have high generalization power, we enforce a dense geometric constraint on both *seen* and *unseen* viewpoints by using estimated monocular depth maps as pseudo ground truth depth for training few-shot NeRF. We describe the details of this process below.

Monocular depth regularization on seen views. We leverage a pre-trained MDE model, denoted as $\mathcal{G}_\phi(\cdot)$ with parameters ϕ , to predict pseudo depth map from given *seen* view image I_i as $D_i^* = \mathcal{G}_\phi(I_i)$. Since D_i^* is initially a relative depth map, it needs to be scaled and shifted into an absolute depth [62] and aligned with NeRF’s rendered depth \bar{D} in order for it to be used as pseudo-depth D^* . However, the scale and shift parameters inferred from the global statistic may undermine local statistic [62]. For example, as shown in Fig. 2 (a), global scale fitting tends to favor dominant objects in the image, leading to ill-fitted depths in less dominant sections of the scene due to inconsistencies in predicted depth differences between the objects. Naïvely employing such inaccurately estimated depths for distillation can adversely impact the overall geometry of the NeRF.

To alleviate this issue, we propose a patch-wise adjustment of scale and shift parameters, reducing the impact of erroneous depth differences, as illustrated in Fig. 2 (b). The depth consistency loss is defined as follows:

$$\mathcal{L}_{\text{seen}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} \| (w_i \text{sg}(D_i^*(\mathbf{p})) + q_i) - \bar{D}_i(\mathbf{p}) \|, \quad (4)$$

where w_i and q_i denote the scale and shift parameters obtained by least square [39] between D_i^* and \bar{D}_i , \mathcal{P} denotes a set of pixels within a patch, and $\text{sg}(\cdot)$ denotes stop-gradient operation [10]. Thus patch-based approach also helps to overcome the computational bottleneck of full image rendering.

Monocular depth regularization on unseen views. We further propose to give supervision even at *unseen* viewpoints. As NeRF has the ability to render any unseen viewpoint of the scene, we render color \bar{I}_l and depth \bar{D}_l from a sampled patch of l -th novel viewpoint. Sequentially, we extract a monocular depth map from the rendered image as $\bar{D}_l^* = \mathcal{G}_\phi(\bar{I}_l)$. Then, we enforce consistency between our rendered depth \bar{D}_l and the monocular depth \bar{D}_l^* of l -th novel viewpoint as follows:

$$\mathcal{L}_{\text{unseen}} = \sum_{I_l \in \mathcal{U}} \sum_{\mathbf{p} \in \mathcal{P}} \| (w_l \text{sg}(\bar{D}_l^*(\mathbf{p})) + q_l) - \bar{D}_l(\mathbf{p}) \|, \quad (5)$$

where \mathcal{U} denotes a set of unseen view images, w_l and q_l denotes the scale and shift parameters used to align \bar{D}_l^* towards \bar{D}_l , and \mathcal{P} denotes randomly sampled patch.

A valid concern regarding this approach is that monocular depth obtained from noisy NeRF rendering may be affected by fine-grained rendering artifacts that frequently appear in unseen viewpoints of few-shot NeRF, resulting in noisy and erroneous pseudo-depths. However, we demonstrate in

Fig. 3 that a strong geometric prior within the MDE model exhibits robustness against such artifacts, effectively filtering out the artifacts and thereby providing reliable supervision for the unseen views.

It should be noted that our strategy differs from previous methods [14, 41, 61, 50] that exploit monocular depth estimation [38] and external depth priors such as COLMAP [43]. These methods only impose depth priors upon the input viewpoints, and thus their priors only influence the scene partially due to self-occlusions and sparsity of known views. Our method, on the other hand, enables external depth priors to be applied to any arbitrary viewpoint and thus allows guidance signals to thoroughly reach every location of the scene, leading to more robust and coherent NeRF optimization.

4.3 Adaptation of MDE via Neural Radiance Field

Although the patch-wise distillation of monocular depth provides invariance to depth difference inconsistency in MDE, the ill-posed nature of monocular depth estimation often introduces additional ambiguities, such as the inability to distinguish whether the surface is concavity, convexity, or planar or difficulty in determining the orientation of flat surfaces [30]. We argue that these ambiguities arise due to the MDE lacking awareness of the scene-specific absolute depth priors and multiview consistency. To address this issue, we propose providing the scene priors optimized NeRF to MDE, whose knowledge of canonical space and absolute geometry helps eliminate the ambiguities present within MDE. Therefore, we propose to adapt the MDE to the absolute scene geometry, formally written as:

$$\mathcal{L}_{\text{MDE}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} \left\{ \|\text{sg}(\bar{D}_i(\mathbf{p})) - \bar{D}_i^*(\mathbf{p})\| + \|(w_i \text{sg}(\bar{D}_i(\mathbf{p})) + q_i) - \bar{D}_i^*(\mathbf{p})\| \right\}. \quad (6)$$

In addition to the patch-wise loss in Eq. 4, we add an l -1 loss without scale-shift adjustment to adapt the MDE with absolute depth prior. We also introduce a regularization term to preserve the local smoothness of MDE, given by:

$$\mathcal{L}_{\text{reg}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} \|(w_i \text{sg}(D_i^{*,\text{init}}(\mathbf{p})) + q_i) - D_i^*(\mathbf{p})\|, \quad (7)$$

where $D_i^{*,\text{init}}$ denotes monocular depth map of I_i extracted from MDE with initial pre-trained weight.

4.4 Confidence Modeling

Our framework must take into account the errors present in both few-shot NeRF and estimated monocular depths, which will propagate [45] and intensify during the distillation process if left unchecked. To prevent this, we adopt confidence modeling [23, 45] inspired by self-training approaches [45, 1], to verify the accuracy and reliability of each ray before the distillation process.

The homogeneous coordinates of a pixel \mathbf{p} in the seen viewpoint are transformed to \mathbf{p}' at the target viewpoint using the viewpoint difference $R_{i \rightarrow l}$ and the camera intrinsic parameter K , as follows:

$$\mathbf{p}' \sim KR_{i \rightarrow l}D_i(\mathbf{p})K^{-1}\mathbf{p}. \quad (8)$$

We generate the confidence map M_i by measuring the distance between rendered depth of the unseen viewpoint and MDE output of seen viewpoint such that

$$M_i(\mathbf{p}) = [\|(w_i D_i^*(\mathbf{p}) + q_i) - \bar{D}_i(\mathbf{p}')\| < \tau], \quad (9)$$

where τ denotes threshold parameter, $[\cdot]$ is Iverson bracket, and $D_l(\mathbf{p}')$ refers to depth value of the corresponding pixel at l -th unseen viewpoint for reprojected target pixel \mathbf{p} of i -th seen viewpoint. We fit D_i^* to absolute scale, where scale and shift parameters, w_i and q_i , are obtained by least square [39] between D_i^* and \bar{D}_i .

4.5 Overall Training

With the incorporation of confidence modeling, the loss functions for both the radiance field and MDE can redefined. $\mathcal{L}_{\text{seen}}$ and $\mathcal{L}_{\text{unseen}}$ can be redefined as:

$$\mathcal{L}_{\text{seen}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} M_i(\mathbf{p}) \|(w_i \text{sg}(D_i^*(\mathbf{p})) + q_i) - \bar{D}_i(\mathbf{p})\|, \quad (10)$$

$$\mathcal{L}_{\text{unseen}} = \sum_{I_l \in \mathcal{U}} \sum_{\mathbf{p} \in \mathcal{P}} M_l(\mathbf{p}) \|(w_l \text{sg}(\bar{D}_l^*(\mathbf{p})) + q_l) - \bar{D}_l(\mathbf{p})\|. \quad (11)$$

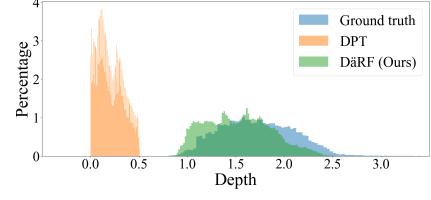
Table 1: **Quantitative comparison on ScanNet [13] and Tanks and Temples [22]**. The best results are highlighted in bold, while the second best results are underlined.

Methods	Depth prior	ScanNet [13]						Tanks and Temples [22]		
		9 - 10 views			18 - 20 views			PSNR ↑	SSIM ↑	LPIPS ↓
		PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NerfingMVS [52]	✓	N/A	N/A	N/A	16.29	0.626	0.502	N/A	N/A	N/A
<i>K</i> -planes [16]	✗	16.01	0.618	0.494	18.70	0.708	0.400	12.57	0.453	0.607
RegNeRF [33]	✗	<u>16.38</u>	<u>0.624</u>	<u>0.493</u>	18.93	0.676	0.450	<u>14.12</u>	<u>0.469</u>	0.580
DS-NeRF [14]	✓	N/A	N/A	N/A	20.85	0.713	0.344	N/A	N/A	N/A
DDP-NeRF [41]	✓	N/A	N/A	N/A	19.29	0.695	0.368	N/A	N/A	N/A
SCADE [50]	✓	-	-	-	21.54	0.732	0.292	-	-	-
DäRF (Ours)	✓	18.29	0.690	0.412	21.58	0.765	<u>0.325</u>	15.70	0.514	<u>0.583</u>

Table 2: **Evaluation of depth quality:** (a) quantitative evaluation of the adapted MDE, compared with other monocular depth estimation models and (b) visualization of depth distributions. The adapted MDE by our method shows a similar distribution to that of the ground truth.

Methods	AbsRel ↓	SqRel ↓	RMSE ↓	RMSE log ↓
LeRes [58]	0.391	0.472	0.999	0.661
MiDaS [39]	<u>0.152</u>	0.095	0.452	0.183
DPT [38]	0.191	0.135	0.563	0.220
DäRF (9 - 10 views)	0.154	<u>0.074</u>	<u>0.361</u>	0.171
DäRF (18 - 20 views)	0.151	0.071	0.356	0.168

(a) Quantitative comparison



(b) Depth distribution comparison

In addition, the loss for the adaptation of the MDE module can be redefined considering M :

$$\mathcal{L}_{\text{MDE}} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} M_i(\mathbf{p}) (\|\text{sg}(\bar{D}_i(\mathbf{p})) - \bar{D}_i^*(\mathbf{p})\| + \|(w_i \text{sg}(\bar{D}_i(\mathbf{p})) + q_i) - \bar{D}_i^*(\mathbf{p})\|). \quad (12)$$

With these losses, we train both NeRF and MDE simultaneously, enhancing both models by complementing each other. MDE provides a strong geometric prior to NeRF while having the inherent limitation of obliviousness to the scene-specific prior, whereas NeRF provides it with its absolute geometry.

5 Experiments

5.1 Experimental Settings

Implementation details. DäRF is implemented based on *K*-planes [36] as NeRF. We use DPT-hybrid [38] as MDE model. We use Adam [21] as an optimizer, with a learning rate of $1 \cdot 10^{-2}$ for NeRF and $1 \cdot 10^{-5}$ for the MDE, along with a cosine warmup learning rate scheduling. See supplementary material for more details. The code and pre-trained weights will be made publicly available.

Datasets. We evaluate our method in real-world scenes captured at both indoor and outdoor locations. Following previous works [41, 50], we use a subset of sparse-view ScanNet data [13] comprised with three indoor scenes, each consisting of 18 to 20 training images and 8 test images. We also conduct evaluations on more challenging setting with 9 to 10 train images. For outdoor reconstruction, we further test on 5 challenging scenes from the Tanks and Temples dataset [22]. The scenes are real-world outdoor dataset, with a wide variety of scene scales and lighting conditions. Note that these setups are extremely sparse compared to full image setups, where we use approximately 0.5 to 5 percent of the whole training inputs.

Baselines. We adopt the following six recently proposed methods as baselines: standard neural radiance field method: *K*-planes [16], few-shot NeRF method: RegNeRF [33], and depth prior based methods: NerfingMVS [52], DS-NeRF [14], DDP-NeRF [41], and SCADE [50]. For methods whose code has not open-sourced, we leave the result as blank.

Evaluation metrics. For quantitative comparison, we follow the NeRF [31] and report the PSNR, SSIM [51], LPIPS [63]. We report standard evaluation metrics for depth estimation [15], absolute

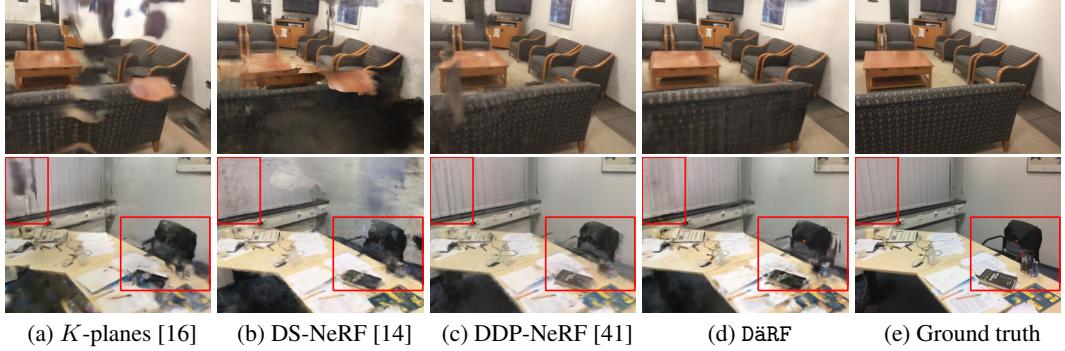


Figure 5: Qualitative results of on ScanNet [13] with 18 - 20 input views.

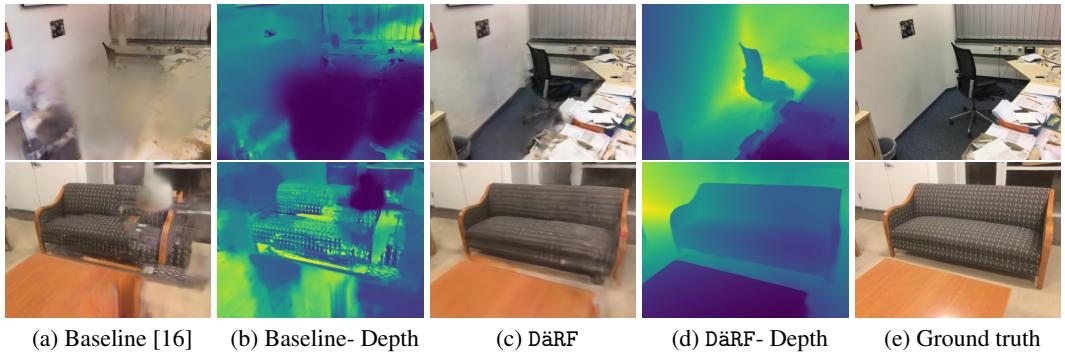


Figure 6: Qualitative results on ScanNet [13] with 9 - 10 input views.

relative error (Abs Rel), squared relative error (SqRel), root mean squared error (RMSE), root mean squared log error (RMSE log). To evaluate view consistency, we utilize a single scaling factor s for each scene, which is the median scaling [64] value averaged across all test views.

5.2 Comparisons

Indoor scene reconstruction. We conducted experiments in two settings: (1) a standard few-shot setup as described in literature [41, 50], and (2) an extreme few-shot setup with approximately 0.5 percent of the full images. As shown in Tab. 1, our approach outperforms the baseline methods in both settings in most of the metrics. Additionally, we provide quantitative results of the adapted MDE model in ScanNet dataset in Tab. 2, and qualitative results in Fig. 4. As shown in Fig. 5 for the setting of standard few-shot, DS-NeRF [14] and DDP-NeRF [41] still show floating artifacts in the novel view and show limitation in capturing details in the chair, smoothing into nearby object. Our method shows better qualitative results compared to other baselines, showing better geometry understanding and detailed view synthesis in the small objects near the chair. In the extreme few-shot setup, we conducted a visual comparison between our method and a baseline [16] in Fig. 6. This is a more complex setting than standard, but our method outperforms the baseline, showing better geometric understanding. More qualitative images are included in the supplementary material.

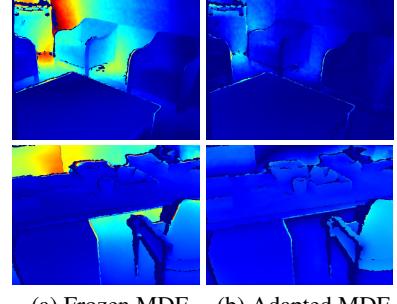


Figure 4: Error map visualization. MDE adaptation results in a reduction of errors.

Outdoor scene reconstruction. We conduct the qualitative and quantitative comparisons on the Tanks and Temples dataset in Tab. 1 and Fig. 7. Since COLMAP [43] with sparse images is not available, we provide comparisons with baselines without explicit depth prior. The quantitative results show that our approach outperforms the baseline methods on this complex outdoor dataset in all metrics. As shown in Fig. 7, our baseline shows limited performance, despite its feasible results of

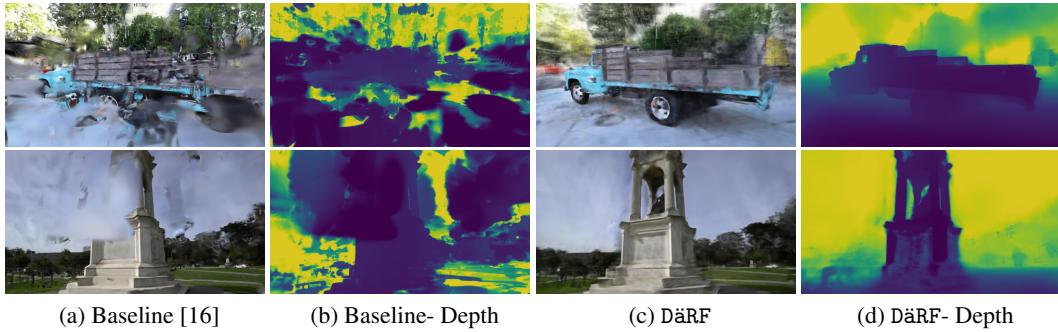


Figure 7: Qualitative results on Tanks and Temples [22].

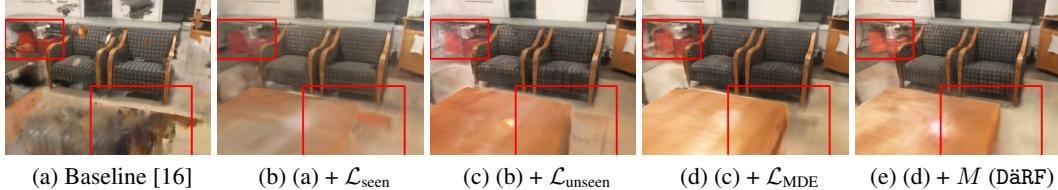


Figure 8: Visualization of ablation studies on ScanNet [13].

view synthesis in novel viewpoint, its depth results show that the network totally fails to understand 3D geometry. Our method shows rich 3D understanding, even in this real-world outdoor setting which is more complicated than other scenes. More qualitative images are included in the supplementary material.

5.3 Ablation Study

Ablation on core components. In Tab. 3 and Fig. 8, we evaluate the effect of each proposed component. The quantitative results show effectiveness of each component. For qualitative results, we found out that $\mathcal{L}_{\text{unseen}}$ suppresses the artifacts in novel viewpoint, compared to when only $\mathcal{L}_{\text{seen}}$ is given. With adaptation of the MDE network to this scene, red basket in the background shows more accurate results and artifacts near the table are removed. In our model, with confidence modeling, view synthesis results show to be more structurally confident in the overall scene.

Analysis of local fitting. In Tab. 4, we further investigate the effectiveness of local scale-shift fitting and global scale-fitting. For global scale-shift fitting, we give learnable scale and shift parameters for each input image and convert MDE’s output to the absolute value in a global manner. For a fair comparison, we compare our model only with MDE distillation on seen viewpoints. The results show that our method local scale-shift fitting is more effective on giving accurate depth supervision.

6 Conclusion

We propose DäRF, a novel method that addresses the limitations of NeRF in few-shot settings by fully leveraging the ability of monocular depth estimation networks. By integrating MDE’s geometric priors, DäRF achieves robust optimization of few-shot NeRF, improving geometry reconstruction and artifact removal in both unseen and seen viewpoints. We further introduce patch-wise scale-shift fitting for accurate mapping of local depths to 3D space, and adapt MDE to NeRF’s absolute scaling and multiview consistency, by distilling NeRF’s absolute geometry to monocular depth estimation. Through complementary training, DäRF establishes a strong synergy between MDE and NeRF, leading to a state-of-the-art performance in few-shot NeRF. Extensive evaluations on real-world scene datasets demonstrate the effectiveness of DäRF.

Table 3: Ablation study.

Components	PSNR↑	SSIM↑	LPIPS↓
(a) Baseline [16]	18.70	0.708	0.400
(b) (a) + $\mathcal{L}_{\text{seen}}$	19.71	0.730	0.380
(c) (b) + $\mathcal{L}_{\text{unseen}}$	21.21	0.758	0.333
(d) (c) + MDE Adapt. (\mathcal{L}_{MDE})	21.39	0.763	0.327
(e) (d) + Conf. Modeling (DäRF)	21.58	0.765	0.325

Table 4: Local fitting ablation.

Components	PSNR↑	SSIM↑	LPIPS↓
Baseline [16]	18.65	0.706	0.502
w/ global fitting	19.05	0.698	0.399
w/ local fitting (DäRF)	19.71	0.730	0.380

References

- [1] Eric Arazo, Diego Ortego, Paul Albert, Noel E O’Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2020.
- [2] Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O’Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in neural information processing systems*, 34, 2021.
- [3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.
- [4] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [5] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023.
- [6] Amlaan Bhoi. Monocular depth estimation: A survey. *arXiv preprint arXiv:1901.09402*, 2019.
- [7] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. *arXiv preprint arXiv:2212.07388*, 2022.
- [8] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [9] Weifeng Chen, Zhao Fu, Dawei Yang, and Jia Deng. Single-image depth perception in the wild, 2017.
- [10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15750–15758, 2021.
- [11] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7911–7920, 2021.
- [12] Jaehoon Choi, Dongki Jung, Yonghan Lee, Deokhwa Kim, Dinesh Manocha, and Donghwan Lee. Selftune: Metrically scaled monocular depth estimation through self-supervised learning. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6511–6518. IEEE, 2022.
- [13] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Niessner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.
- [14] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.
- [15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014.
- [16] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. *arXiv preprint arXiv:2301.10241*, 2023.
- [17] Huan Fu, Mingming Gong, Chaohui Wang, Kayhan Batmanghelich, and Dacheng Tao. Deep ordinal regression network for monocular depth estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2002–2011, 2018.
- [18] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5885–5894, 2021.

- [19] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [20] Seungryong Kim, Kihong Park, Kwanghoon Sohn, and Stephen Lin. Unified depth prediction and intrinsic image decomposition from a single image via joint convolutional neural fields. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII 14*, pages 143–159. Springer, 2016.
- [21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *iclr*. 2015. *arXiv preprint arXiv:1412.6980*, 9, 2015.
- [22] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017.
- [23] Minseop Kwak, Jiuhan Song, and Seungryong Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. *arXiv preprint arXiv:2301.10941*, 2023.
- [24] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *2016 Fourth international conference on 3D vision (3DV)*, pages 239–248. IEEE, 2016.
- [25] Jae-Han Lee and Chang-Su Kim. Monocular depth estimation using relative depth maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2019.
- [26] Jin Han Lee, Myung-Kyu Han, Dong Wook Ko, and Il Hong Suh. From big to small: Multi-scale local planar guidance for monocular depth estimation. *arXiv preprint arXiv:1907.10326*, 2019.
- [27] Bo Li, Chunhua Shen, Yuchao Dai, Anton Van Den Hengel, and Mingyi He. Depth and surface normal estimation from monocular images using regression on deep features and hierarchical crfs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1119–1127, 2015.
- [28] Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. Nerfacc: A general nerf acceleration toolbox. *arXiv preprint arXiv:2210.04847*, 2022.
- [29] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *NeurIPS*, 2020.
- [30] S Mahdi H Miangoleh, Sebastian Dille, Long Mai, Sylvain Paris, and Yagiz Aksoy. Boosting monocular depth estimation models to high-resolution via content-adaptive multi-resolution merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9685–9694, 2021.
- [31] 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 *ECCV*, 2020.
- [32] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.
- [33] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.
- [34] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.
- [35] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.
- [36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

- [37] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.
- [38] 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.
- [39] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.
- [40] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021.
- [41] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *arXiv preprint arXiv:2112.03288*, 2021.
- [42] Ashutosh Saxena, Min Sun, and Andrew Y Ng. Make3d: Learning 3d scene structure from a single still image. *IEEE transactions on pattern analysis and machine intelligence*, 31(5):824–840, 2008.
- [43] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [44] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020.
- [45] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. *Advances in neural information processing systems*, 33:596–608, 2020.
- [46] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, 2022.
- [47] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *CVPR*, 2021.
- [48] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.
- [49] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021.
- [50] Mikaela Angelina Uy, Ricardo Martin-Brualla, Leonidas Guibas, and Ke Li. Scade: Nerfs from space carving with ambiguity-aware depth estimates. *arXiv preprint arXiv:2303.13582*, 2023.
- [51] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.
- [52] 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*, pages 5610–5619, 2021.
- [53] Cho-Ying Wu, Jialiang Wang, Michael Hall, Ulrich Neumann, and Shuochen Su. Toward practical monocular indoor depth estimation. *arXiv preprint arXiv:2112.02306*, 2021.

- [54] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 106–122. Springer, 2022.
- [55] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360 $\{\backslash\deg\}$ views. *arXiv preprint arXiv:2211.16431*, 2022.
- [56] Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.
- [57] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. *arXiv preprint arXiv:2303.07418*, 2023.
- [58] Wei Yin, Jianming Zhang, Oliver Wang, Simon Niklaus, Long Mai, Simon Chen, and Chunhua Shen. Learning to recover 3d scene shape from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 204–213, 2021.
- [59] Alex Yu, Rui long Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021.
- [60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021.
- [61] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022.
- [62] Chi Zhang, Wei Yin, Billzb Wang, Gang Yu, Bin Fu, and Chunhua Shen. Hierarchical normalization for robust monocular depth estimation. *Advances in Neural Information Processing Systems*, 35:14128–14139, 2022.
- [63] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [64] Zhoutong Zhang, Forrester Cole, Richard Tucker, William T Freeman, and Tali Dekel. Consistent depth of moving objects in video. *ACM Transactions on Graphics (TOG)*, 40(4):1–12, 2021.

Appendix

A Implementation Details

A.1 Architecture

We implement DäRF with K -planes [16] as the base model. It represents a radiance field using tri-planes with three multi-resolutions for each plane: 128, 256, and 512 in both height and width, and 32 in feature depth. This approach also incorporates small MLP decoders and a two-stage proposal sampler. It should be noted that our framework is not restricted to the K -planes baseline, but can be incorporated into any NeRF backbone models [31, 32, 28]. In our experiments, we implemented our framework on top of the K -planes hybrid version codebase due to its quality, reasonable optimization speed, and model size. For the monocular depth estimation (MDE) module, we choose the pre-trained DPT [38] as our base MDE model due to its powerful generalization ability in a zero-shot setting. Trained on very large datasets, DPT demonstrates impressive prediction quality and generalizes well to novel scenes. However, any MDE model can be utilized within our framework [58, 39, 38].

A.2 Training details

We use the Adam optimizer [21] and a cosine annealing with warm-up scheduler for NeRF optimization. The learning rate is set to $1 \cdot 10^{-2}$, and we perform 512 warm-up steps. For MDE adaptation, we also employ the Adam optimizer [21] with a learning rate of $1 \cdot 10^{-5}$. NeRF optimization is performed with a pixel batch size of 4,096, totaling 20K iterations. For $\mathcal{L}_{\text{seen}}$, we render a 64×64 patch, while for $\mathcal{L}_{\text{unseen}}$, we render a 128×128 patch with a stride of 3.

For the loss functions, we set the coefficients of $\mathcal{L}_{\text{seen}}$, \mathcal{L}_{MDE} , and \mathcal{L}_{reg} as 0.01, 0.01, and 0.1, respectively. During the warm-up stage of 5,000 steps, the coefficient of $\mathcal{L}_{\text{unseen}}$ is initially set to 0 and then increased to 0.01 after 5,000 warm-up steps. For the first 1,000 steps, we employ the ranking loss [58] with a coefficient of 0.1, in addition to $\mathcal{L}_{\text{seen}}$. All experiments were conducted using a single NVIDIA GeForce RTX 3090. The training process takes approximately 3 hours.

A.3 Training loss details

In the following, we describe a least-square alignment [53] used in loss functions for MDE prior distillation in detail. As described in the main paper, we use a scale-shift invariant loss [39] with patch-wise adjustment for depth consistency as follows:

$$\mathcal{L} = \sum_{I_i \in \mathcal{S}} \sum_{\mathbf{p} \in \mathcal{P}} \| (w_i D_i^*(\mathbf{p}) + q_i) - \bar{D}_i(\mathbf{p}) \|, \quad (1)$$

where w_i and q_i are scale and shift values that align $D_i^*(\mathbf{p})$ to the absolute locations of $\bar{D}_i(\mathbf{p})$. In this loss function, to calculate w_i and q_i , we follow least-squares criterion [39]:

$$(w_i, q_i) = \arg \min_{w_i, q_i} \sum_{\mathbf{p} \in \mathcal{P}} \| (w_i D_i^*(\mathbf{p}) + q_i) - \bar{D}_i(\mathbf{p}) \| \quad (2)$$

In other words, we can rewrite the above scheme as a closed problem. Let $\mathbf{h}_i = [s_i, t_i]^T$ and $\vec{D}_i(\mathbf{p}) = [D_i^*(\mathbf{p}), 1]^T$, then we can modify our problem as

$$\mathbf{h}_i^{\text{opt}} = \arg \min_{\mathbf{h}_i} \sum_{\mathbf{p} \in \mathcal{P}} (\vec{D}_i(\mathbf{p})^T \mathbf{h}_i - \bar{D}_i(\mathbf{p}))^2, \quad (3)$$

which can be solved as follows:

$$\mathbf{h}_i^{\text{opt}} = \left(\sum_{\mathbf{p} \in \mathcal{P}} \vec{D}_i(\mathbf{p}) \vec{D}_i(\mathbf{p})^T \right)^{-1} \left(\sum_{\mathbf{p} \in \mathcal{P}} \vec{D}_i(\mathbf{p}) \bar{D}_i(\mathbf{p}) \right) \quad (4)$$

A.4 Baseline implementations

We directly use quantitative results reported in prior literature [50] for the comparison of Nerfing-MVS [52], DS-NeRF [14] and DDP-NeRF [41]. As the setting [50] requires out-of-domain priors, it

should be noted that the results for DDP-NeRF are with out-of-domain priors. The results of DDP-NeRF with in-domain priors are 20.96, 0.737, and 0.236 for PSNR, SSIM, and LPIPS, respectively. However, we were unable to evaluate DDP-NeRF in the extreme settings of ScanNet and Tanks and Temples, as reliable COLMAP 3D points could not be obtained.

We utilized the authors' provided official implementations of RegNeRF [33] and K -planes [16], training one model for each scene using two different scenarios on the ScanNet [13] and Tanks and Temples [22] datasets. However, since there is no official code available for SCADE [50], we are unable to provide performance comparisons for this method.

B Datasets and Metrics

B.1 Datasets

ScanNet [13]. We adhere to the few-shot protocol provided by DDP-NeRF [41] in our experimental setup. We noticed that the split contained major overlaps across the train and test sets, which makes the task easier compared to realistic few-shot settings where images exhibit minimal overlap. For this reason, we construct an extreme few-shot scenario, using only half of the training images while maintaining the same test set.

Tanks and Temples [22]. To test the robustness of our method in challenging real-world outdoor environments, we conduct further experiments on Tanks and Temples dataset, an real-world outdoor dataset acquired under drastic lighting effects and reflectances. As no existing protocols exist for a few-shot scenario for this dataset, we introduce a new split for the few-shot setting. We carefully selected 5 object-centric scenes —truck, francis, family, lighthouse, and ignatius— with inward-facing cameras. From each scene, we sample 10 training images that capture the overall geometry of the whole scene. For testing, we use one-eighth of the dataset as a test set, consisting every 8th repeating image from the entire image set. We run COLMAP [43] on all images to obtain camera poses for NeRF training. However, for the lighthouse scene, which exhibits highly sensitive lighting and specular effects dependent on view pose, we manually preprocess the parts that contain these effects.

B.2 Evaluation metrics

To evaluate the quality of novel view synthesis, following previous works [31], we measure PSNR, SSIM, and LPIPS. It is mentioned in K -planes that an implementation of SSIM from mip-NeRF [4] results in lower values than standard scikit-image implementation. For a fair comparison per dataset, we use the latter scikit-image SSIM implementation following the relevant prior work.

For the evaluation of the MDE module, we use 4 depth estimation metrics as follows:

- AbsRel: $\frac{1}{|\mathcal{I}|} \sum_{\mathbf{p} \in \mathcal{I}} \|\bar{D}(\mathbf{p}) - D^{\text{GT}}(\mathbf{p})\| / D^{\text{GT}}(\mathbf{p})$;
- SqRel: $\frac{1}{|\mathcal{I}|} \sum_{\mathbf{p} \in \mathcal{I}} \|\bar{D}(\mathbf{p}) - D^{\text{GT}}(\mathbf{p})\|^2 / D^{\text{GT}}(\mathbf{p})$;
- RMSE: $\sqrt{\frac{1}{|\mathcal{I}|} \sum_{\mathbf{p} \in \mathcal{I}} \|\bar{D}(\mathbf{p}) - D^{\text{GT}}(\mathbf{p})\|^2}$;
- RMSE log: $\sqrt{\frac{1}{|\mathcal{I}|} \sum_{\mathbf{p} \in \mathcal{I}} \|\log \bar{D}(\mathbf{p}) - \log D^{\text{GT}}(\mathbf{p})\|^2}$;

where \mathbf{p} is a pixel in the image \mathcal{I} and D^{GT} is ground truth depth map. In addition, following [64], we use single scaling factor s for each scene which is obtained by

$$s = \frac{1}{N} \sum_{I_i \in \mathcal{S}} (\text{median}(D_i^{\text{GT}} / \bar{D}_i)), \quad (5)$$

rather than fit each frame to ground truth, to evaluate view consistency of MDE models. Here, \mathcal{S} denotes set of images from single scene.

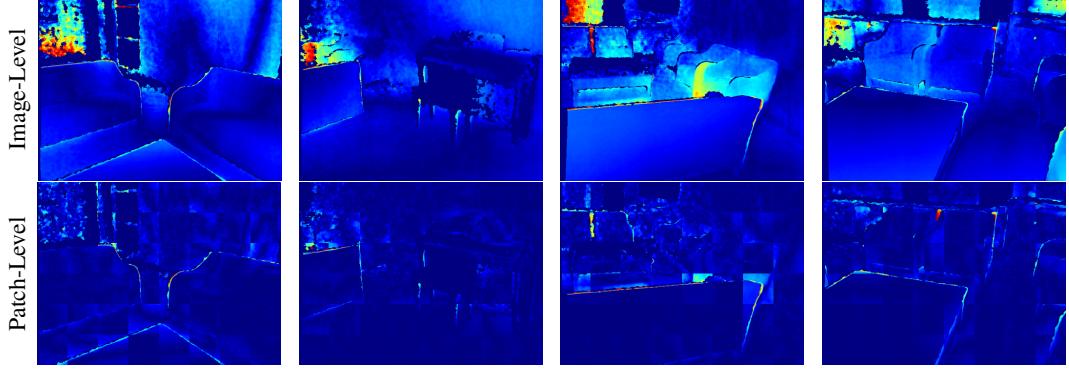


Figure 1: **Error map visualization of image-level and patch-wise scale and shift adjustment:** relative depth map in various viewpoints is fitted in two ways, image-level fitting (first row) and patch-level fitting (second row).

C Additional Analysis

C.1 Comparison of patch- and image-level scale-shift adjustment

We provide additional analysis and visualization results regarding the patch-wise scale and shift adjustment. In Fig.1, we present error maps showing the discrepancies between the ground truth sensor depth and the predicted depth. Additionally, in Fig.2, we present qualitative results of rendered color and depth using each fitting method. It is important to note that in the image-level fitting scheme, a single set of scale and shift values is computed for an entire depth map. Conversely, in our patch-level fitting method, scale and shift values are calculated individually for each 80×80 patch within the depth map. The error map clearly demonstrates the significant reduction in misalignment errors achieved by our patch-level fitting method compared to the image-level fitting approach. For

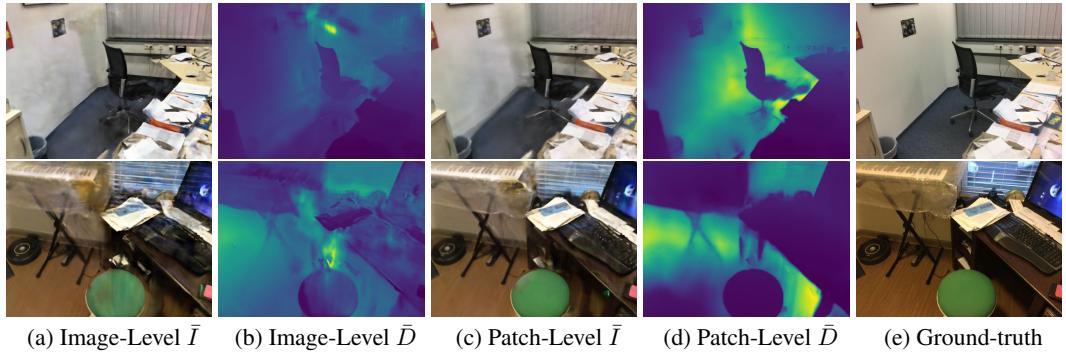


Figure 2: **Comparison of patch- and image-level scale-shift adjustment.** Rendered color and depth from NeRF with (a-b) image-level scale and shift adjustment and (c-d) patch-level scale and shift adjustment.

the comparison of image-level and patch-level fitting provided in the Fig. 2 and Tab. 4 of the main paper, we set the scale and shift as learnable parameters per image for image-level fitting and conduct patch-wise scale-shift invariant loss for patch-level fitting. This comparison is conducted only with $\mathcal{L}_{\text{seen}}$ given and results with patch-level fitting show better performance compared to image-level fitting. The difference between the two methods is especially distinguished in rendered depth maps of these two settings, in that patch-level fitting lets NeRF learn depth more accurately.

C.2 Confidence modeling

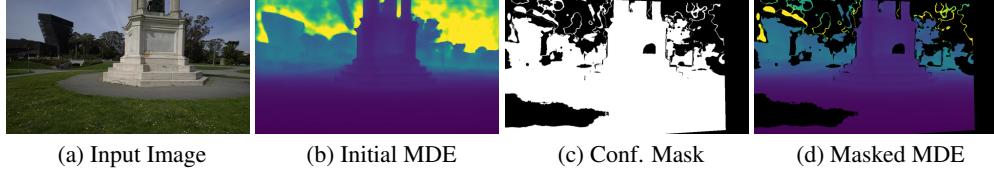


Figure 3: **Comparisons on MDE depth map with and without confidence masking.** the initial MDE depth map predicted is filtered through mask from our confidence modeling.

In Fig. 3, we demonstrate the effectiveness of our confidence modeling which effectively eliminates inaccurate information present in depth maps from both NeRF and the MDE network through leveraging multi-view consistency of NeRF. MDE depth from the input image contains errors, which can be filtered out by verifying consistency with depth from NeRF’s other viewpoint. Likewise, the error of MDE depth from unseen viewpoint can be filtered through consistency check with MDE depth from the seen viewpoint.

C.3 Ablation of MDE baselines

Table 1: **Ablation study on MDE baseline.**

Components	PSNR↑	SSIM↑	LPIPS↓
DäRF with LeReS [58]	21.31	0.757	0.343
DäRF with MiDaS [39]	21.48	0.758	0.337
DäRF with DPT [38]	21.58	0.765	0.325

We conduct an ablation on the Monocular Depth Estimation (MDE) network to assess its impact on our methodology. Considering the recent advancements [39, 38] in MDE models that shows strong generalization power for depth estimation in unseen images, we replace our MDE network with state-of-the-art models such as LeReS, MiDaS, and DPT. The results in Tab. 1 show that our method shows consistent performance across different baselines.

D Additional Qualitative Results

In this section, we show additional qualitative comparisons in Fig. 4, Fig. 5, Fig. 6, Fig. 7, Fig. 8, and Fig. 9 for ScanNet [13] dataset in two different settings and in Fig. 10, Fig. 11, Fig. 12, Fig. 13, and Fig. 14 for Tanks and Temples [22] dataset.

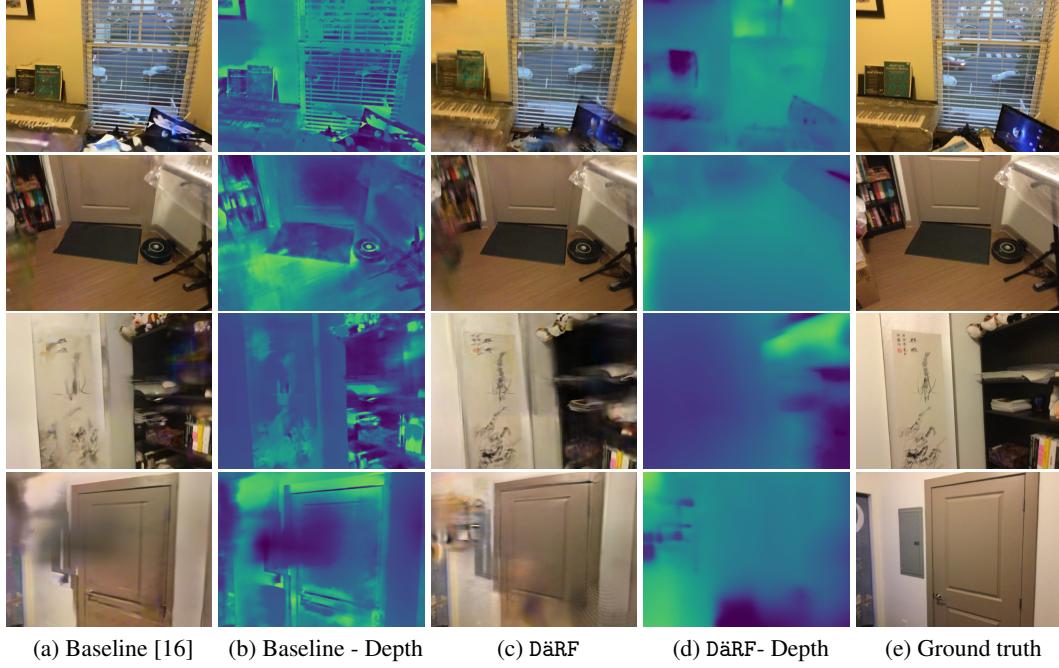


Figure 4: Qualitative results on Scan 0710 of ScanNet [13] with 9 - 10 input views.

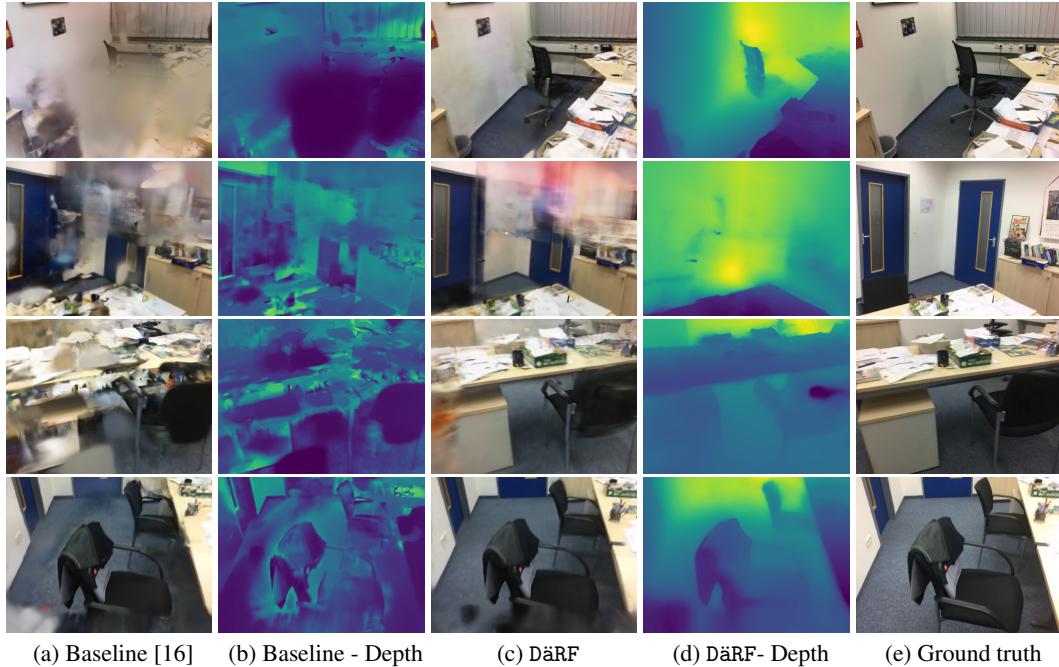
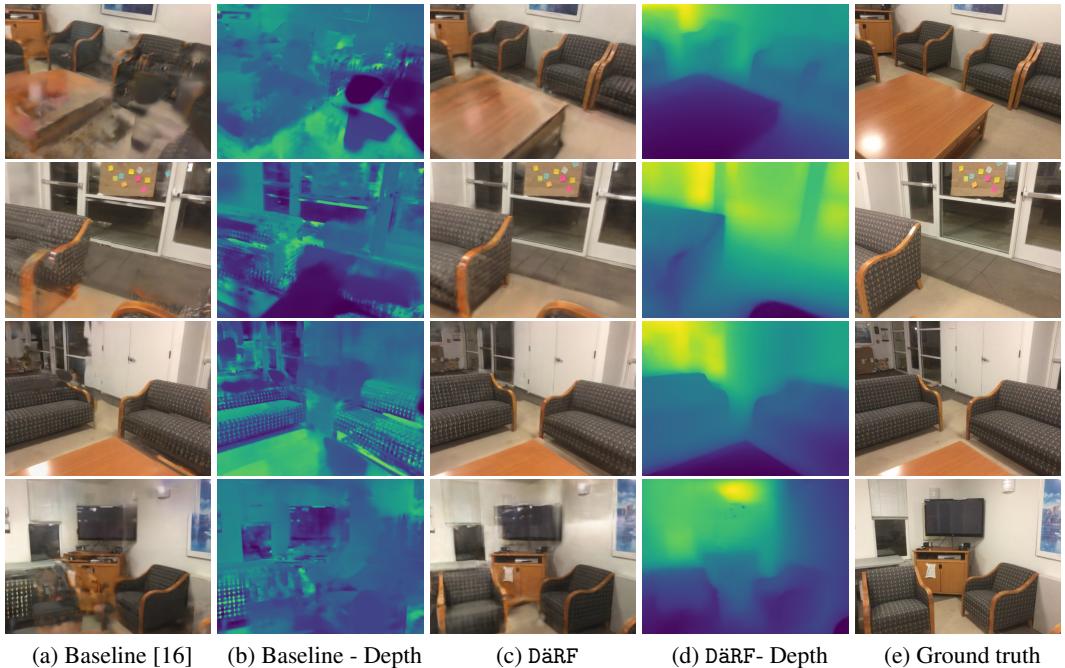
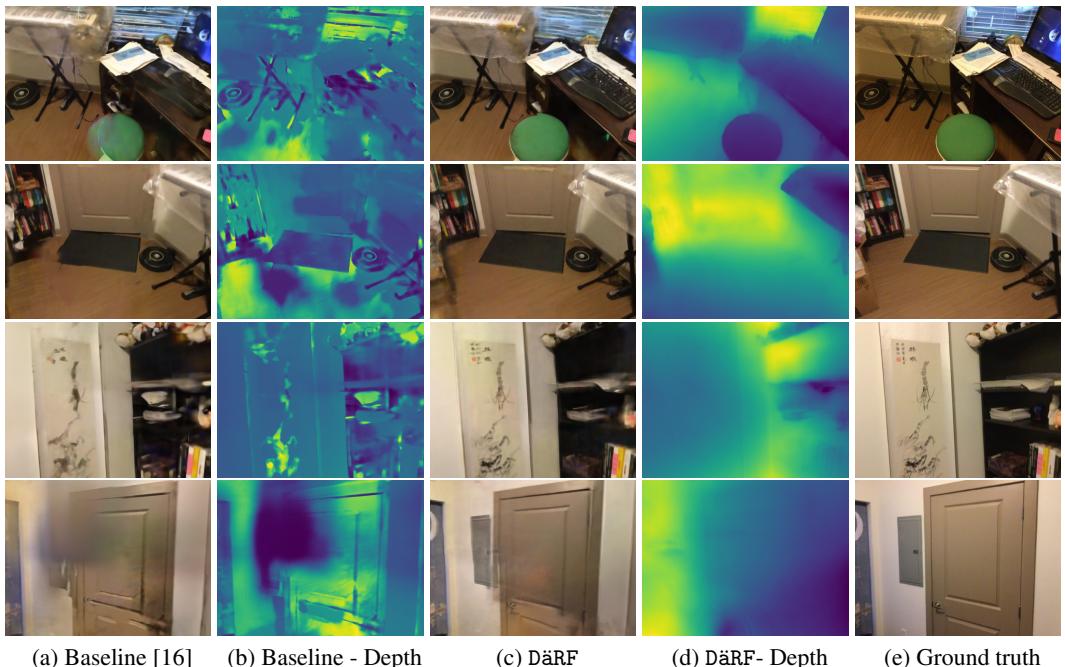


Figure 5: Qualitative results on Scan 0758 of ScanNet [13] with 9 - 10 input views.



(a) Baseline [16] (b) Baseline - Depth (c) DäRF (d) DäRF- Depth (e) Ground truth

Figure 6: Qualitative results on Scan 0781 of ScanNet [13] with 9 - 10 input views.



(a) Baseline [16] (b) Baseline - Depth (c) DäRF (d) DäRF- Depth (e) Ground truth

Figure 7: Qualitative results on Scan 0710 of ScanNet [13] with 18 - 20 input views.

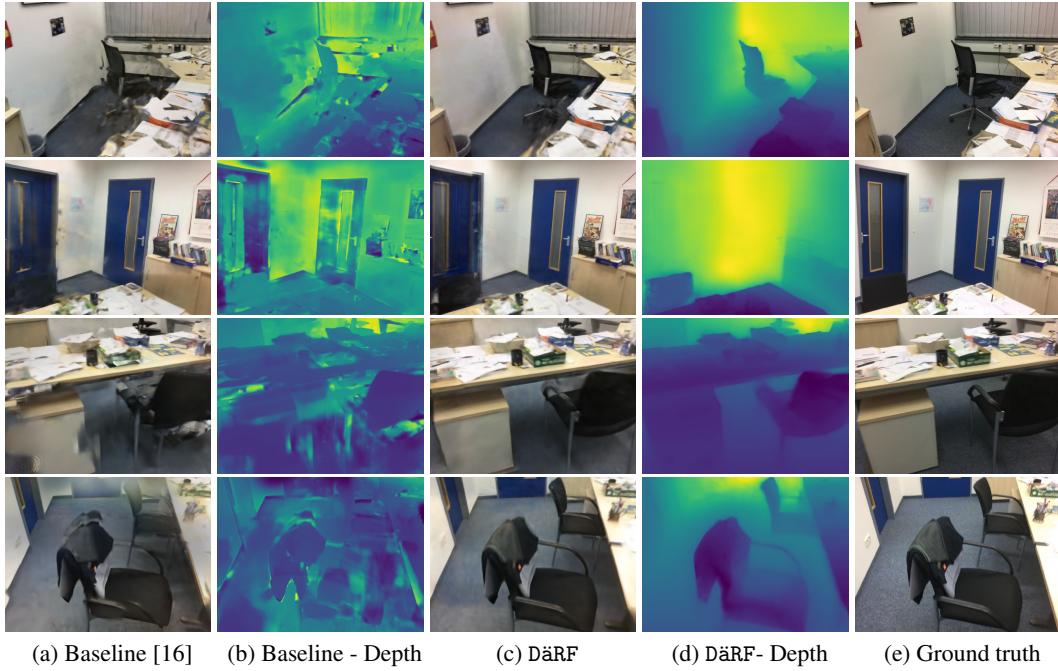


Figure 8: Qualitative results on Scan 0758 of ScanNet [13] with 18 - 20 input views.

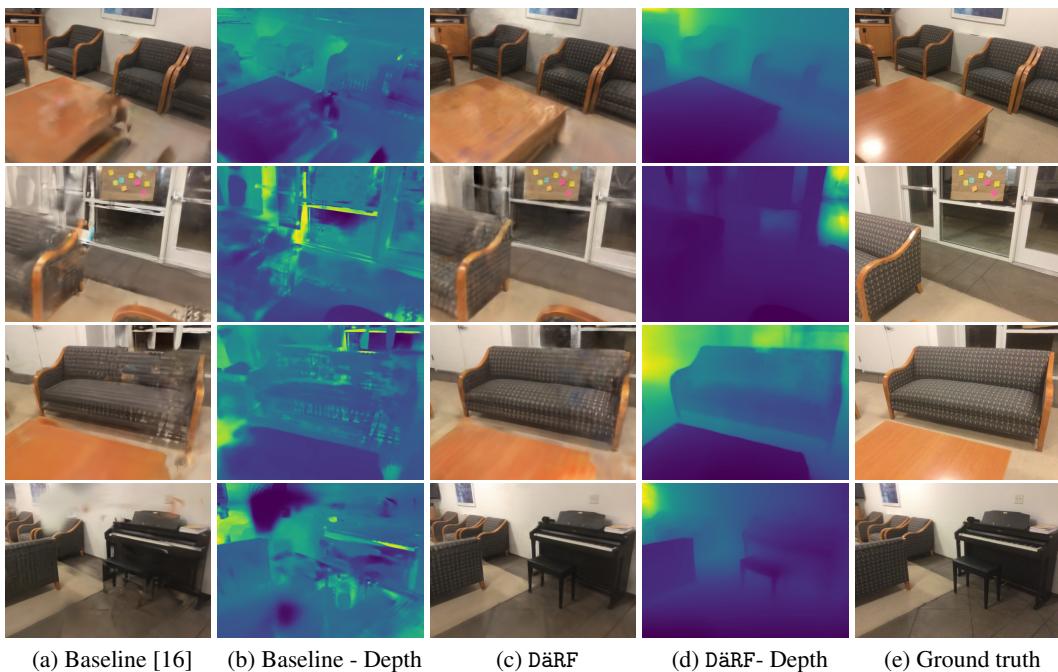


Figure 9: Qualitative results on Scan 0781 of ScanNet [13] with 18 - 20 input views.

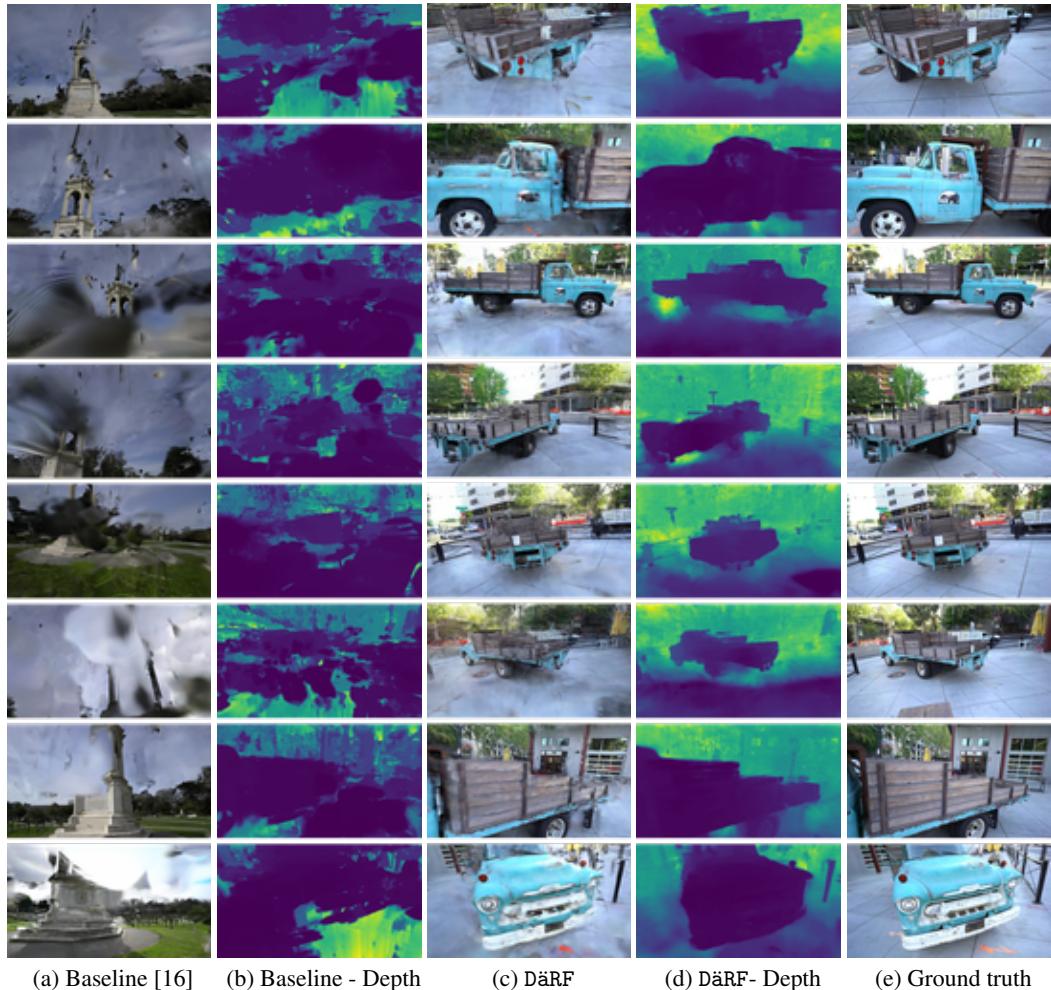


Figure 10: Qualitative results on truck scene of Tanks and Temples [22] with 10 input views.

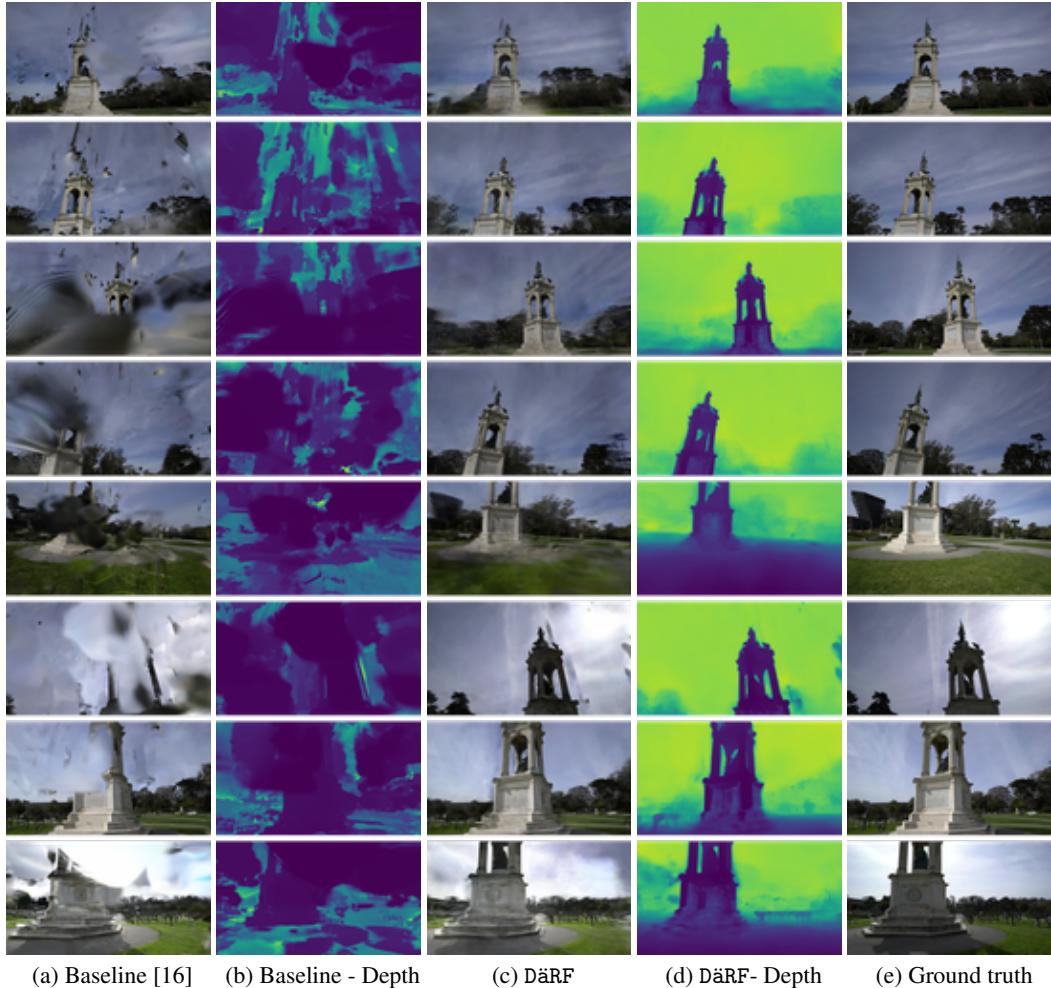


Figure 11: **Qualitative results on francis scene of Tanks and Temples [22]** with 10 input views.

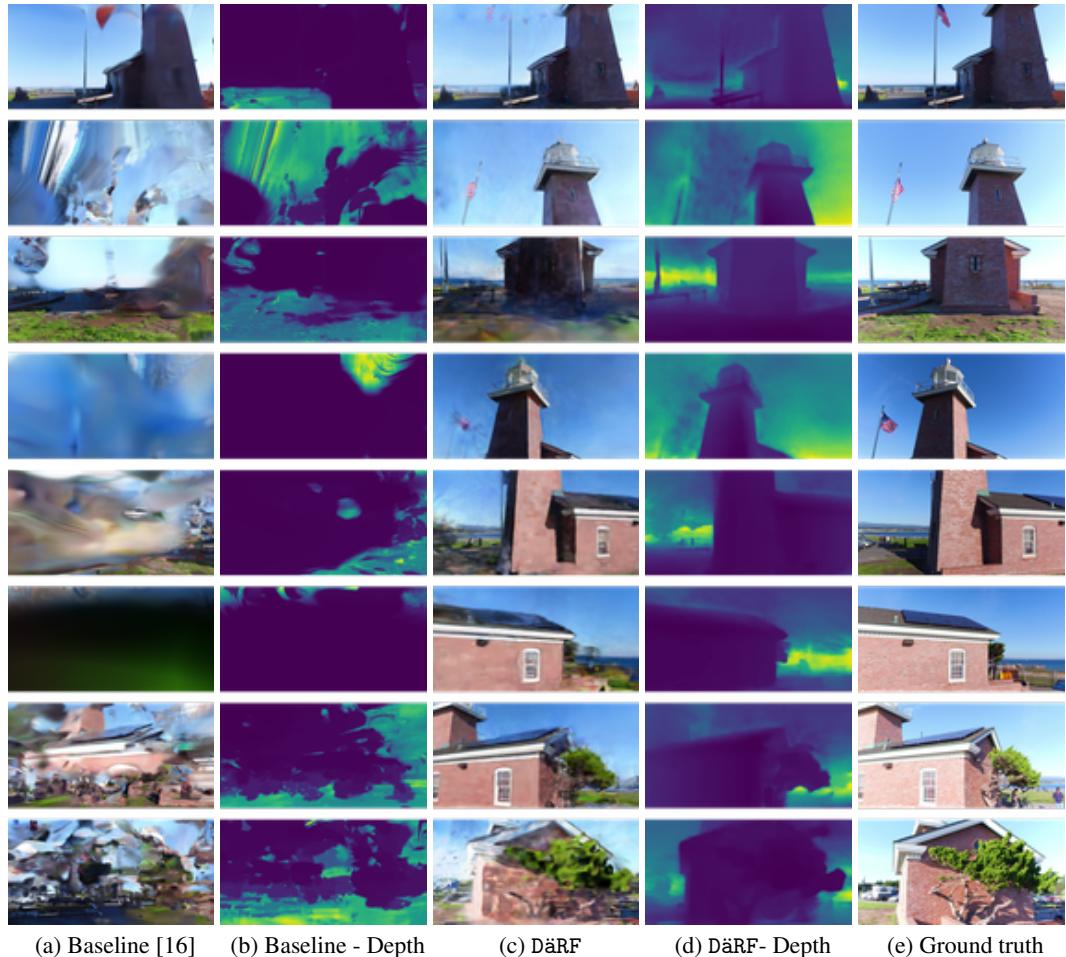


Figure 12: Qualitative results on lighthouse scene of Tanks and Temples [22] with 10 input views.

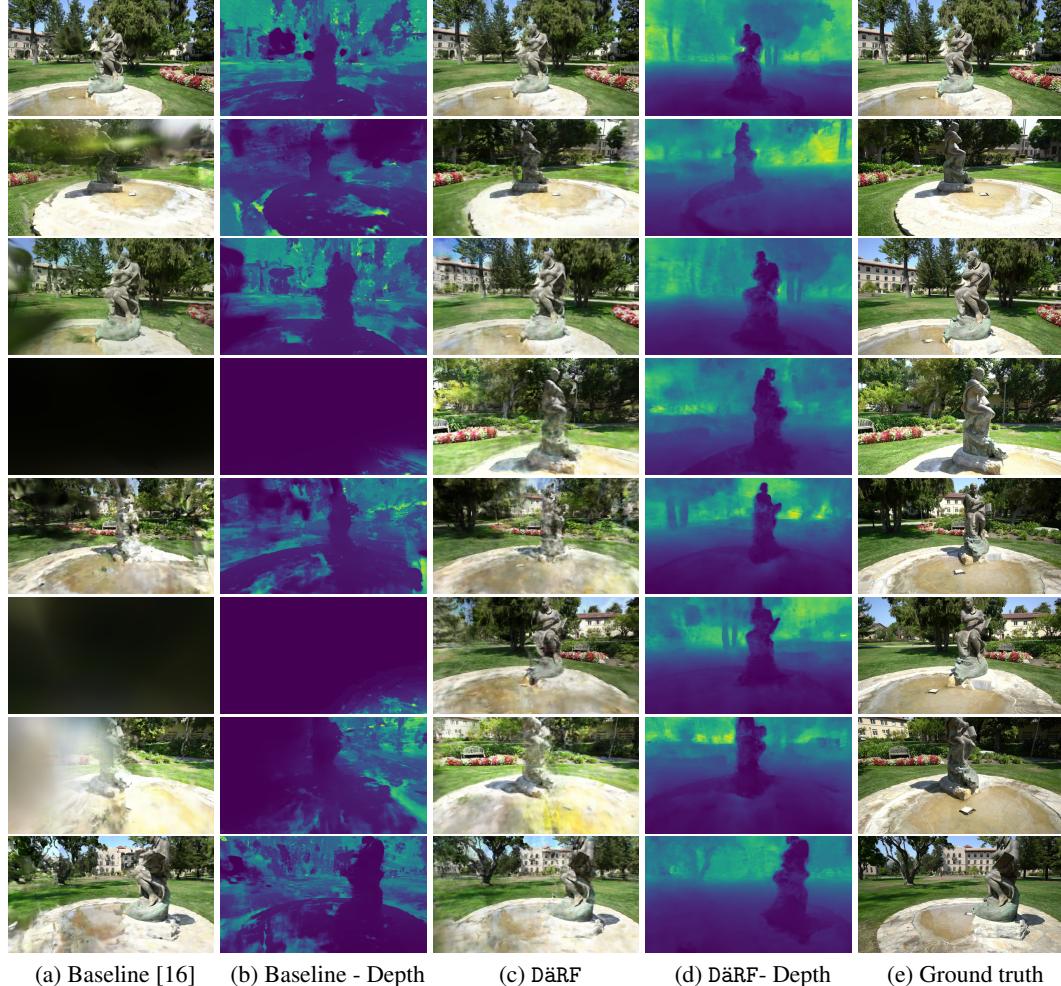


Figure 13: Qualitative results on ignatius scene of Tanks and Temples [22] with 10 input views.

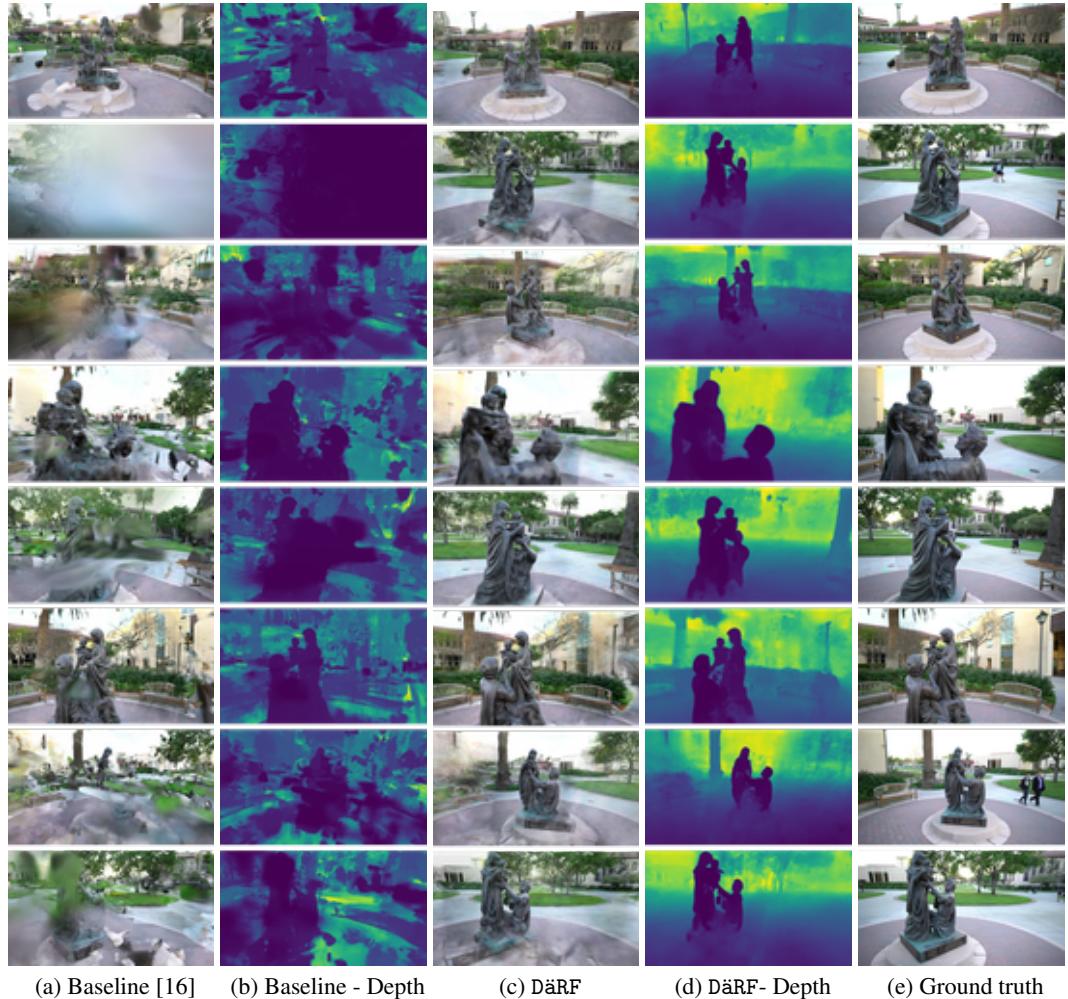


Figure 14: **Qualitative results on family scene of Tanks and Temples [22] with 10 input views.**

E Limitations and Future Works

While our method shows powerful performance quantitatively, its limitations can be noticed in its qualitative results above, where it struggles to reconstruct the fine-grained details present in ground truth images. Also, our usage of depth supervision from various viewpoints does not get rid of the artifacts completely: some artifacts that cloud the space between objects and the camera, are reduced yet still visible in rendering of unseen viewpoints.

These limitations may be attributed to fundamental limitations in the few-shot NeRF setting [18], where fine-grained details are often occluded from one viewpoint to another due to an extreme lack of input images, preventing faithful geometric reconstruction of details. Also, since the seen viewpoints view a comparatively small portion of the entire scene, there inevitably occur artifacts in the unseen viewpoint as some depths cannot be perfectly determined from given input information.

F Broader Impacts

Our work achieves robust optimization and rendering of NeRF under sparse view scenarios, drastically reducing the number of viewpoints required for NeRF and bringing NeRF closer to real-life applications such as augmented reality, 3D reconstruction, and robotics. Our extension of few-shot NeRF to a real-world setting with the usage of monocular depth estimation networks also would enable NeRF optimization under various real-life lighting conditions and specular surfaces due to its increased robustness and generalization power.