arXiv:2308.10001v1 [cs.CV] 19 Aug 2023

AltNeRF: Learning Robust Neural Radiance Field via Alternating Depth-Pose Optimization

Kun Wang, Zhiqiang Yan, Huang Tian, Zhenyu Zhang, Xiang Li, Jun Li and Jian Yang

PCA Lab, Nanjing University of Science and Technology, China

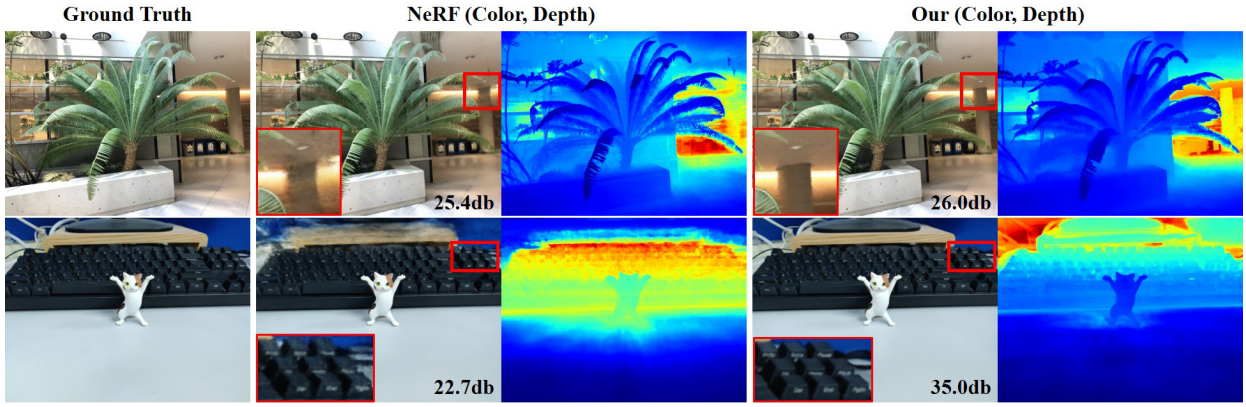

Figure 1: NeRF (Mildenhall et al. 2020) often suffers from degenerate solutions due to the lack of explicit 3D supervision and imprecise camera poses. To address these problem, we propose AltNeRF, a novel framework that can create robust NeRF representations from monocular videos obviating the need for known camera poses.

Abstract

Neural Radiance Fields (NeRF) have shown promise in generating realistic novel views from sparse scene images. However, existing NeRF approaches often encounter challenges due to the lack of explicit 3D supervision and imprecise camera poses, resulting in suboptimal outcomes. To tackle these issues, we propose AltNeRF—a novel framework designed to create resilient NeRF representations using self-supervised monocular depth estimation (SMDE) from monocular videos, without relying on known camera poses. SMDE in AltNeRF masterfully learns depth and pose priors to regulate NeRF training. The depth prior enriches NeRF's capacity for precise scene geometry depiction, while the pose prior provides a robust starting point for subsequent pose refinement. Moreover, we introduce an alternating algorithm that harmoniously melds NeRF outputs into SMDE through a consistence-driven mechanism, thus enhancing the integrity of depth priors. This alternation empowers AltNeRF to progressively refine NeRF representations, yielding the synthesis of realistic novel views. Additionally, we curate a distinctive dataset comprising indoor videos captured via mobile devices. Extensive experiments showcase the compelling capabilities of AltNeRF in generating high-fidelity and robust novel views that closely resemble reality.

1 Introduction

Neural rendering has achieved unprecedented progress on the long-standing view synthesis task in computer vision communities (Zhou et al. 2016; Kellnhofer et al. 2021; Yu et al. 2022; Li, Li, and Zhu 2023). One prominent exemplar of this task is NeRF (Mildenhall et al. 2020), which comprehensively captures the continuous volumetric essence of real-world scenes using multi-view images alongside precise camera poses, consequently generating lifelike new perspectives. Nonetheless, NeRF often grapples with suboptimal outcomes that compromise novel view synthesis and distort scene geometry, as evidenced in Fig. 1. We identify two primary catalysts for this issue: 1) *The lack of explicit 3D supervision.* NeRF solely hinges on 2D image supervision, which may furnish inadequate geometric constraints for textureless or view-limited scenes. Introducing explicit 3D supervision holds potential to shepherd NeRF towards superior convergence. 2) *Inaccurate camera poses.*NeRF's reliance on precise camera poses for constructing accurate volumetric scenes becomes a stumbling block in the face of pose inaccuracies or noise. Such errors in camera poses compound the optimization challenges for NeRF.

Although existing methods have endeavored to tackle either of these issues, they remain encumbered by certain limitations. Firstly, some methods (Deng et al. 2022; Roessle et al. 2022) leverage depth priors to facilitate NeRF's convergence towards improved solutions. These methods derive depth priors from structure-from-motion methodologies or depth completion techniques, employing them as rigid constraints for NeRF. However, these depth priors might not attain the requisite accuracy, potentially skewing NeRF's optimization trajectory and yielding deteriorated performance. Secondly, some methods (Wang et al. 2021d; Lin et al. 2021; Jeong et al. 2021) alternative strategies undertake the

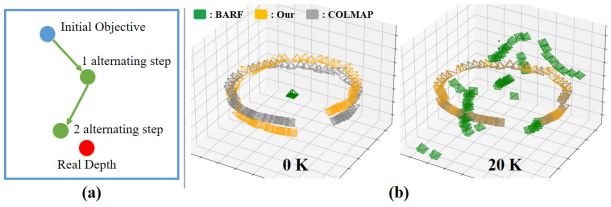

Figure 2: (a) Existing methods establish a rigid target (the blue dot) using inaccurate depth prior, whereas we leverage valuable intermediate results from NeRF to dynamically adjust the objective (the green dots) towards the real depth (the red dot). (b) Pose refinement starting from different initial poses. The experiment is conducted on Vasedeck scene.

joint optimization of NeRF and camera poses to mitigate the ramifications of imprecise camera poses. Nevertheless, this combined task encompasses a non-convex optimization conundrum that is acutely sensitive to the initialization of camera poses. Consequently, these approaches necessitate initial camera poses that closely approximate the optimal values; otherwise, they frequently converge towards unfavorable local minima. Illustratively, Fig. 2 (b) depicts the scenario where BARF (Lin et al. 2021) employs identity matrices as green-hued initializations, eventually converging to nonsensical poses after numerous iterations.

To addresses the above problems, we propose AltNeRF—a novel framework designed to generate robust neural radiance fields from unposed images. The core concept involves a cyclic process of self-supervised monocular depth estimation (SMDE) and NeRF optimization, synergistically enhancing both methodologies. Leveraging SMDE from monocular videos (as described in (Zhou et al. 2022; Zhang et al. 2023)), we infer depth and pose for each frame without the need for manual annotations. The estimated pose serves as an effective initialization, facilitating smoother optimization akin to the orange poses depicted in Fig. 2 (b). Furthermore, the estimated depth provides an initial objective that steers NeRF away from optimizing inaccurate scene geometries. After further optimizing NeRF, we can obtain better pose and depth to refine the depth of SMDE. This alternation continually updates the depth objective to converge towards actual scene depths, as illustrated in Fig. 2 (a). Our AltNeRF harnesses the complementary strengths of SMDE and NeRF, leading to more robust scene representations. Overall, our contributions can be summarized as:

- We introduce depth-pose priors learned from monocular videos to simultaneously regularize the scene geometries and initialize the camera poses to enhance the novel view synthesis of NeRF.
- To the best of our knowledge, we are the first to propose AltNeRF—a novel framework that alternately optimizes self-supervised monocular depth estimation and NeRF, synergistically boosting both components.
- We also collect a new dataset of indoor videos captured with a cellphone. Extensive experiments on LLFF, ScanNet, CO3D and our dataset demonstrate that our AltNeRF can synthesize realistic novel views with high fidelity and robustness, and outperforms the realted NeRF methods.

2 Related Work

Self-supervised Monocular Depth Estimation. The learning of SMDE is an image reconstruction problem. It is supervised by the photometric loss that measures the difference between a target frame and frames warped from nearby views. SfM-Learner (Zhou et al. 2017) is a seminal work that proposed to jointly predict scene depth and relative camera poses. Follow-up works enhanced SfM-Learner by enforcing depth scale consistency (Bian et al. 2019; Wang et al. 2021b), introducing more powerful neural networks (Guizilini et al. 2020; Lyu et al. 2021; Guizilini et al. 2022), and applying iterative refinement (Bangunharcana, Magd, and Kim 2023). Furthermore, MonoDepth2 (Godard et al. 2019) proposed a minimum reprojection loss to handle occlusions, and some works addressed the dynamic object problem by compensating and masking pixels within dynamic areas using optical flow (Zou, Luo, and Huang 2018; Ranjan et al. 2019) and pretrained segmentation models (Casser et al. 2019; Gordon et al. 2019). Some other works boosted the performance of self-supervised depth estimation by introducing a feature-metric loss (Shu et al. 2020), proposing a resolution adaptive framework (He et al. 2022), and exploring the knowledge distilling approaches (Petrovai and Nedevschi 2022; Ren et al. 2022). Recently, some works have focused on challenging environments, such as indoor (Ji et al. 2021; Li et al. 2021; Wu et al. 2022) and nighttime (Vankadari et al. 2020; Wang et al. 2021a; Liu et al. 2021) scenes and shown impressive performance.

View Synthesis with NeRF. NeRFs (Mildenhall et al. 2020) are a powerful technique for novel view synthesis, but they face several challenges in different scenarios. Many works have extended NeRFs to handle dynamic (Pumarola et al. 2021; Li et al. 2022; Liu et al. 2023), unbounded (Zhang et al. 2020; Barron et al. 2022; Reiser et al. 2023), and large-scale scenes (Tancik et al. 2022; Turki, Ramanan, and Satyanarayanan 2022; Zhang et al. 2022), as well as to optimize NeRFs from in-the-wild (Martin-Brualla et al. 2021) and dark images (Mildenhall et al. 2022). Some works have also improved the generalization (Yu et al. 2021b; Wang et al. 2021c; Johari, Lepoittevin, and Fleuret 2022; Chen and Lee 2023; Smith et al. 2023), bundle sampling (Kurz et al. 2022; Arandjelović and Zisserman 2021), initialization (Bergman, Kellnhofer, and Wetzstein 2021; Tancik et al. 2021) and data structure (Yu et al. 2021a; Xu et al. 2022; Müller et al. 2022) of NeRFs. However, these methods still rely on accurate camera poses, which are not always available or realistic. To address this problem, recent works (Wang et al. 2021d; Jeong et al. 2021; Meng et al. 2021; Lin et al. 2021; Bian et al. 2023) have studied the joint task of optimizing NeRF model and camera poses. However, they are restricted to simple or known pose distribution. Moreover, some methods use depth priors (Deng et al. 2022; Roessle et al. 2022) from external sources, which may be noisy or inaccurate and lead to suboptimal NeRF representations. In contrast, we propose a novel framework that can learn robust NeRF representations from monocular videos. Our framework leverages self-supervised depth estimation to obtain depth and pose priors that regularize NeRF learning. Fur-

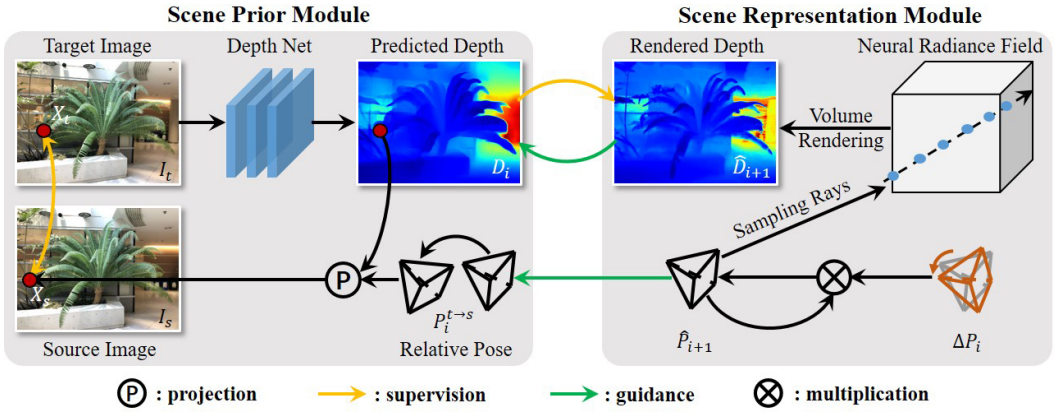

Figure 3: The overall pipeline of our AltNeRF. The scene prior module learns depth and pose priors, which serves as the depth reference and initial poses, respectively. The scene representation module simultaneously refines the initial poses with ΔP_i and learns 3D scene representation, which is regularized by D_i, and produces more accurate poses $\hat{P}_{i+1}$ and finer depth maps $\hat{D}_{i+1}$. These materials are then fed back to the scene prior module as guidance to improve its performance.

thermore, we devise an alternating algorithm that refines the depth and pose priors with consistent NeRF outputs, leading to better 3D geometry and camera poses.

3 Preliminary

In this section, we review the key concepts and techniques of Self-supervised Monocular Depth Estimation (SMDE) and Neural Radiance Field (NeRF) to provide the necessary background for our method.

Self-supervised Monocular Depth Estimation. SMDE is a training method that only requires monocular videos $\mathcal{V}$ and known camera intrinsic K. It employs two neural networks, $f_d : I \rightarrow D$ and $f_p : (I_t, I_s) \rightarrow P_{t\rightarrow s}$, to predict the depth map D of an input image I and relative camera pose $P_{t\rightarrow s}$ between frames I_t and I_s. The training objective is to reconstruct the target frame I_t from nearby views I_s by mapping pixels x_s from the source image to the target image x_t based on the predicted depth and camera pose: $x_s \sim KP_{t\rightarrow s}D(x_t)K^{-1}x_t$. The photometric loss is used to supervise this process, which consists of the structural similarity term and the ℓ_1 term:

$$L_p(I_t, \hat{I}_t) = \frac{\alpha}{2}(1 - SSIM(I_t, \hat{I}_t)) + (1-\alpha)\|I_t - \hat{I}_t\|_1, \tag{1}$$

where α is often set to 0.85. An edge-aware smoothness loss is also added to ensure smoothness in predicted depth maps. This loss is based on the image gradients ∂_x and ∂_y along the horizontal and vertical axes, and is weighted by an exponential function of the image gradients to preserve edges:

$$L_s = |\partial_x D|e^{-|\partial_x I|} + |\partial_y D|e^{-|\partial_y I|}, \tag{2}$$

where $|\cdot|$ returns the absolute value.

Neural Radiance Field. NeRF represents a scene as a continuous volumetric field. NeRF takes in a 3D point $\mathrm{p} \in \mathbb{R}^3$ and a unit viewing direction $\mathrm{d} \in \mathbb{R}^3$, and returns the corresponding density σ and color c: $f_n : (\mathrm{p}, \mathrm{d}) \rightarrow (\sigma, c)$. The volumetric field can be rendered to 2D images using volume rendering techniques (Kajiya and Von Herzen 1984):

$$\hat{C}(r) = \int_{t_n}^{t_f} T(t)\sigma(t)c(t)dt. \tag{3}$$

Similarly, the scene depths are created by computing the mean terminating distance of a ray $r = \mathrm{o} + t\mathrm{d}$ parameterized by camera origin o and viewing direction d:$\hat{D}(r) = \int_{t_n}^{t_f} T(t)\sigma(t)tdt$, where $T(t) = exp(-\int_{t_n}^{t}\sigma(s)ds)$ handles occlusions, and t_n and t_f are near and far depth bounds, respectively. The optimization objective of NeRF is to minimize the reconstruction loss, which is computed as the squared differences between the rendered and ground truth colors for all rays:

$$L_c = \|\hat{C}(r) - C(r)\|_2. \tag{4}$$

4 AltNeRF Framework

In this section, we introduce our AltNeRF framework, which comprises two components: the Scene Prior Module (SPM) and the Scene Representation Module (SRM). These modules work together under an alternating algorithm. In the following sections, we will delve into the details.

4.1 Scene Prior Module

Pretraining SPM leverages SMDE to provide initial scene depths and camera poses. For better robustness, it is first pretrained on many monocular videos to learn prior knowledge. To improve the generalization ability of the model on unseen scenes, we employ the knowledge distilling strategy introduced in (Wu et al. 2022) and distill knowledge from an off-the-shelf relative depth estimator, DPT (Ranftl, Bochkovskiy, and Koltun 2021), via

$$L_r = 1 - SSIM(D, D_r) + 0.1 \times (E_r \oplus E/\, size(E)), \tag{5}$$

where D_r is the reference depth map produced by DPT, $\oplus$ denotes XOR operation, $size(\cdot)$ returns the size of a set, and

E_r and E are occluding boundary maps of D_r and D, respectively. Overall, the loss function for pretraining is:

$$L_{pt} = L_p + L_r + 1.0e^{-3} \times L_s. \tag{6}$$

Test-time Adaptation SPM predicts relative depths, which are defined up to an unknown scale factor, leading to potential inconsistencies across frames. Moreover, the data distribution of the target video often differs from that of the training data. To mitigate these challenges, we employ self-supervised finetuning to adapt SPM to the video before generating predictions. To ensure scale-consistent depth-pose estimates, we additionally introduce the geometry consistency loss from (Bian et al. 2019):

$$L_g = \frac{\|D_s(x_s) - D_t(x_t)\|_1}{D_s(x_s) + D_t(x_t)}, \tag{7}$$

where D_s and D_t are predicted depth map of I_s and I_t, respectively. Finally, the loss used in adaptation step is

$$L_{ad} = L_{pt} + 0.5 \times L_g. \tag{8}$$

Pose Conversion SPM predicts relative 3D transformations between frames, while SRM requires absolute camera poses. To reconcile these requirements, we establish a world coordinate system that aligns with the camera coordinate system of the first frame I_0, whose pose matrix is a identity matrix. We then use the chain rule to calculate the camera poses P_i of subsequent frames based on their relative pose $P_{i\to i+1}$ predicted by SPM: $P_{i+1} = P_{i\to i+1} \times P_i$.

4.2 Scene Representation Module

SRM serves a dual purpose of learning 3D scene representation and refining camera poses simultaneously. It extends the BARF approach (Lin et al. 2021) by introducing three improvements: depth regularization, improved pose initialization, and warmup learning. These enhancements will be discussed in more detail below.

Depth Regularization Learning the NeRF representation only from 2D images is intrinsically a non-convex problem, which can result in a multitude of incorrect solutions that fit the training images well but fail to generate plausible novel views. These degenerate solutions are more likely to occur when the training image set is small (Deng et al. 2022; Roessle et al. 2022) or the image texture is weak. Typically, such solutions show up as inaccurate scene depths, as shown in Fig. 1. To overcome this issue, we propose introducing depth prior from SPM as explicit 3D supervision to regularize the learned depth $\hat{D}$. However, the scene depths produced by SPM are inaccurate and may provide incorrect guidance. To address this, we introduce an error-tolerant depth loss that specifies the possible depth range and uses Huber loss (Huber 1964) $H(\cdot)$ to prevent the model from being significantly affected by large gradients resulting from SPM's inaccurate predictions:

$$L_e = H\left(\max\left(\frac{\|\hat{D}(r) - D(r)\|_1}{\hat{D}(r) + D(r)} - \epsilon, 0\right)\right), \tag{9}$$

where ϵ is a tolerance coefficient and D is the depth prior.

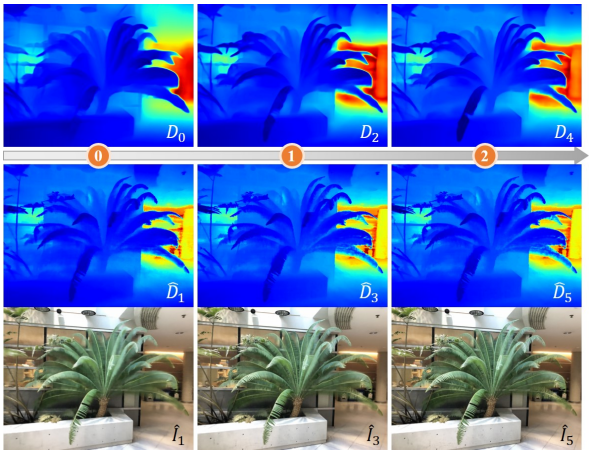

Figure 4: The intermediate results of depth maps and color images from SPM (the first row) and SRM (the last two rows) after 0, 1 and 2 alternating steps.

Improved Pose Initialization When without pose prior, BARF initializes the camera poses with identity matrices and refines them via bundle adjustment. But this fails to capture the complex camera motions, as shown in Fig. 2 (b). In our framework, we use SPM to obtain a better initial pose P for each camera, which is closer to the true pose. Then, we optimize a residual pose ΔP that represents the difference between the initial and refined poses. We update the camera poses as $\hat{P} = \Delta P \times P$. This way, our framework can efficiently estimate plausible camera poses for scenes with complex camera poses.

Warmup Learning While SRM learns the scene representation from scratch, it refines camera poses using a good initialization that is already close to the ideal ones. We discovered that this asynchrony in the learning process results in an incorrect update direction for camera poses. To address this issue, we propose a warmup learning strategy that synchronizes the learning process for these two tasks. Specifically, we set the learning rate of ΔP to a small value l_s at the beginning of training and gradually increase it to the original learning rate l_t after 1K iterations.

Overall Loss The learning of SRM is supervised by both reconstruction loss and depth regularization, with a scalar hyper-parameter γ that balances these two terms of losses:

$$L_{sr} = L_c + \gamma \cdot L_t. \tag{10}$$

4.3 Alternating Algorithm

Optimizing the scene representations and camera poses jointly is highly underdetermined, so the model can easily converge to a "bad" local optimum. However, we show that our model can converge to a reasonable solution by combining the prior knowledge of SPM and the scene-dependent optimization of SRM. We propose a novel alternating algorithm for this, as shown in Fig. 3. Next, we explain the workflow and introduce the multi-view consistency check that can extract confident scene depths from SPM and SRM.

Workflow We denote the alternating step as i, SPM at step i as Φ_i, and SRM as Ψ_i. The alternating process when $i > 0$ can be formulated as:

$$\begin{aligned} \Psi_i &: D_i \rightarrow (\hat{D}_{i+1}, \hat{P}_{i+1}), \\ \Phi_i &: (\hat{D}_{i+1}, \hat{P}_{i+1}) \rightarrow D_{i+2}, \end{aligned} \tag{11}$$

The process begins at $i = 0$, when SPM generates the initial depth maps D_0 and camera poses P_0 using Eq. 8: $\Phi_0 : \mathcal{V} \rightarrow (D_0, P_0)$. Next, Ψ_i takes the depth maps D_i predicted by Φ_{i-1} as input to regularize the scene representation learning via Eq. 9, while simultaneously optimizing the residual poses ΔP_i. After S_r iterations, Ψ_i produces finer depth maps $\hat{D}_{i+1}$ and more accurate camera poses $\hat{P}_{i+1} = \Delta P_i \times P_0$, which are fed back to Φ_i. Since the poses $\hat{P}_{i+1}$ are relatively accurate after refinement, Φ_i directly uses them instead of predicting new ones. The relative camera pose $P_{i+1}^{t \rightarrow s}$ is calculated by converting P_{i+1}^{t} and P_{i+1}^{s} through $P_{i+1}^{t \rightarrow s} = (P_{i+1}^{s})^{-1} P_{i+1}^{t}$. Furthermore, we apply Eq. 5 to distill knowledge from the finer depth maps by treating $\hat{D}_{i+1}$ as the reference. As a result, Φ_i is fine-tuned for S_p iterations with Eq. 6, producing more accurate depth maps D_{i+2}, which are fed into the next alternating step. By repeating these steps, the performance of both SPM and SRM are improved, as shown in Fig. 4.

Multi-view Consistency Check To account for the potential unreliability of the depth predictions from SPM and SRM, we use a multi-view consistency check to assess the uncertainty of the predicted depth. Specifically, we denote the depth map of a target image I_t as D_t, then we compute the depth maps $D_{s \rightarrow t}$ warped from nearby source views I_s using camera poses $P_{t \rightarrow s}$ from SPM or P from SRM. Since D_t and $D_{s \rightarrow t}$ are expected to be identical without considering the occlusions, we define the uncertainty U_t of depth map D_t as the difference between D_t and $D_{s \rightarrow t}$: $U_t = \|D_t - D_{s \rightarrow t}\|_1$. In practice, we compute the mean from the four views with the smallest differences to account for occlusions. To incorporate this depth uncertainty into our loss functions (*i.e.* Eq. 5 and Eq. 9), we weight them with the $Softmin(\cdot)$ function. This helps to mitigate the impact of unreliable depth predictions on our optimization process.

Discussion Our alternating algorithm is a general method that can actually leverage any depth-pose priors, not just those learned from SMDE. By using the valuable intermediate results of SPM and SRM, the algorithm can tolerate imprecise priors and still create high-quality NeRF representations, which helps reduce the cost to create NeRF models.

5 Experiment

In this section, we evaluate AltNeRF on 16 scenes of four datasets and compare it with existing methods to demonstrate its state-of-the-art (SOTA) performance. We first describe the datasets and implementation details, and then report the experiment results.

5.1 Dataset

We evaluate AltNeRF on LLFF (Mildenhall et al. 2019), CO3D (Reizenstein et al. 2021), ScanNet (Dai et al. 2017) and our collected dataset, Captures. **LLFF**: we include five scenes from original LLFF: Fern, Flower, Fortress, Orchids, and Room. We also use the Vasedeck scene from the NeRF dataset. We follow the data split strategy of BARF (Lin et al. 2021), which uses the first 90% of frames for training and the remaining 10% for testing. **CO3D**: we select three scenes from the Couch category: *193_20797_40499*, *349_36504_68102* and *415_57184_110444*. These scenes have more than 80 frames per scene and exhibit complex camera motions with simultaneous panning and rotation. **ScanNet**: we choose three scenes, *scene0079_00*, *scene0553_00* and *scene0653_00*, to evaluate the depth estimation performance of AltNeRF. We use the data processed by NerfingMVS (Wei et al. 2021) and downsample each scene to 20 frames. **Captures**: we collect four scenes using a cellphone, which form our Captures dataset. The captured data contains two types of challenging scenes: 1) scenes with few frames and weak textures (Scene_01 and Scene_02), and 2) scenes with complex camera motions (Scene_03 and Scene_04). Please refer to the supplement for more details.

Method	Rot (°) ↓	Trans (10^{-2}) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	-	-	26.009	0.821	0.127
BARF	33.082	21.311	22.613	0.715	0.327
NeRFmm	65.254	28.625	20.083	0.602	0.432
DS-NeRF	-	-	26.605	0.821	0.148
DDP-NeRF	-	-	23.081	0.753	0.205
NoPe-NeRF	32.197	19.343	24.262	0.743	0.252
Our	1.317	0.725	28.333	0.851	0.110

Table 1: Quantitative results of camera pose estimation (middle) and novel view synthesis (right) tasks. The best is in red, and the second is in orange. The reported results are average over ten scenes of LLFF and Captures.

5.2 Implementation Detail

The depth estimation network $f_d(\cdot)$ in SPM is based on the U-Net (Ronneberger, Fischer, and Brox 2015) architecture. The encoder is a ResNet-50 (He et al. 2016) with the fully-connected layer removed, and the decoder consists of ten 3×3 convolutional layers, two for each scale, and uses bilinear up-sampling. The pose estimation network $f_p(\cdot, \cdot)$ is structured with a ResNet-34 and outputs a vector of nine element length, where the first six elements are continuous rotation representation (Zhou et al. 2019) and the last three elements denote translations. The scene representation function $f_n(\cdot, \cdot)$ in SRM employs the same network structure as NeRF, *i.e.* eight fully-connected layers with skip connections for density output, and one linear layer for color output. The γ in Eq. 10 is set to 0.08 for LLFF and CO3D, and 0.15 for ScanNet and Captures. We pretrain the SPM with a learning rate of $1.0e-4$, and finetune it with $5.0e-5$. For SRM, the initial learning rate to learn NeRF model is set to $1.0e-3$, and exponentially decays to $1.0e-4$ throughout the training process. The initial learning rate for pose refinement is set to $1.0e-5$, and linearly increases to $2.0e-3$ after 1K iterations before exponentially decaying to $1.0e-5$. We employ a hierarchical sampling strategy similar to NeRF, with 64 coarse samples and 64 fine samples, but we do not add coarse samples to the fine pass to save training time. The number of

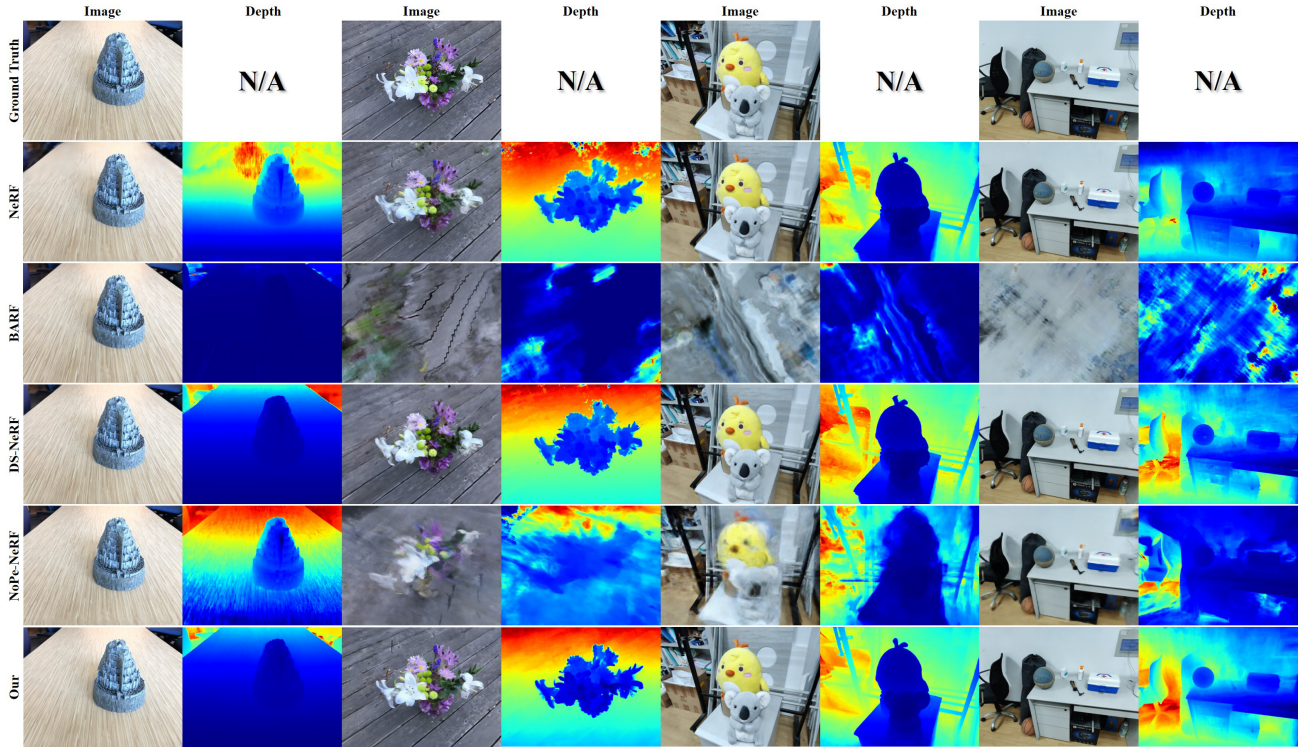

Figure 5: Qualitative comparisons of novel view synthesis and depth estimation on Fortress, Vasedeck, Scene_03 and Scene_04.

Method	Rot (°) ↓	Trans (10^{-2}) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	-	-	34.363	0.928	0.132
BARF	106.58	140.38	12.697	0.549	0.762
NoPe-NeRF	102.99	116.33	14.885	0.595	0.667
Our	2.29	0.89	34.951	0.930	0.135

Table 2: Quantitative results of camera pose estimation (middle) and novel view synthesis (right) tasks. The reported results are average over three scenes of CO3D Couch.

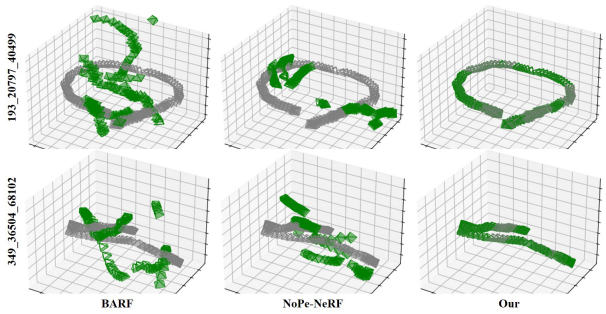

Figure 6: Quantitative comparison of pose estimation on CO3D Couch. The estimated poses are in green and the COLMAP poses are in gray.

iterations S_r and S_p is set to 50K and 500, respectively, and we perform two alternating steps in all experiments unless otherwise specified. Our method is trained for 150K-200K iterations according to the number of frames, which costs around 4.0-6.4 hours totally on single RTX 3090.

5.3 Comparing with Existing Method

Here, we evaluate AltNeRF on novel view synthesis, camera pose estimation, and depth estimation tasks, and compare it with seven existing methods.

Evaluation on LLFF and Captures We evaluate AltNeRF and six SOTA methods from related fields on camera pose estimation and novel view synthesis tasks. The compared methods are NeRF (Mildenhall et al. 2020), BARF (Lin et al. 2021), NeRFmm (Wang et al. 2021d), DS-NeRF (Deng et al. 2022), DDP-NeRF (Roessle et al. 2022) and NoPe-NeRF (Bian et al. 2023). We use ten scenes from LLFF and Captures datasets for this comparison. Tab. 1 shows the mean quantitative results for each method and task. We use *Rot* and *Trans* to present the rotation and translation error between the estimated camera poses and the pseudo ground truth poses from COLMAP (Schönberger and Frahm 2016), and PSNR, SSIM (Wang et al. 2004), and LPIPS (Zhang et al. 2018) to measure the quality of the synthesized images. We initialize the camera poses of BARF and NeRFmm with identity matrices. AltNeRF significantly outperforms the methods that do not use pose priors, *i.e.* BARF, NeRFmm and NoPe-NeRF, on camera pose estimation task. For example, it reduces the Rot and Trans by 95.91% and 96.15%, respectively, compared to NoPe-NeRF. This demonstrates the importance of pose priors for accurate camera pose estimation. AltNeRF also surpasses the COLMAP assisted methods, *i.e.* NeRF, DS-NeRF, and DDP-NeRF, on novel view synthesis task. For example, it improves DS-NeRF by 6.50%, 3.65%, and 13.36%, respectively, on PSNR, SSIM, and LPIPS metrics.

Fig. 5 shows the qualitative comparisons of AltNeRF and

Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑
NeRF	0.143	0.072	0.312	0.805	0.958	0.967
DS-NeRF	0.075	0.025	0.169	0.904	0.956	0.995
NerfingMVS	0.075	0.025	0.164	0.938	0.989	0.998
Our	0.051	0.008	0.106	0.987	0.998	0.999

Table 3: Quantitative results of depth estimation. The reported results are average over three scenes of ScanNet.

four methods on novel view synthesis and depth estimation tasks. It uses four scenes: Fortress and Vasedeck from LLFF, and Scene_03 and Scene_04 from Captures. AltNeRF can synthesize realistic novel views and more accurate depth maps than the competitors. For example, it estimates the depth of the distant chairs in the Fortress scene more accurately, while the other methods underestimate their depth or fail to capture their details. BARF completely fails on the last three scenes, which contain complex camera motions. In contrast, our method can still synthesize realistic novel views and estimate reasonable depth maps with the help of the depth-pose priors and our alternating strategy.

Evaluation on CO3D We evaluate AltNeRF and three existing methods, namely NeRF, BARF, and NoPe-NeRF, on CO3D dataset. We report the mean quantitative results in Tab. 2, and the qualitative results of pose estimation in Fig. 6, respectively. We use the pseudo ground truth poses from COLMAP as the reference to measure the pose error. AltNeRF significantly outperforms BARF and NoPe-NeRF on camera pose estimation task. Our predictions are very close to those of COLMAP, while BARF and NoPe-NeRF fail to produce meaningful pose outputs. AltNeRF also surpasses NeRF on novel view synthesis task. It improves the PSNR metric of NeRF by 1.71%. This demonstrates the effectiveness of our method on challenging scenes.

Evaluation on ScanNet We evaluate AltNeRF and three existing methods, namely NeRF, DS-NeRF, and NerfingMVS (Wei et al. 2021), on depth estimation task. We use the ScanNet dataset for this comparison. Tab. 3 shows the quantitative results for each method. AltNeRF outperforms the existing methods by a large margin on depth estimation task. It reduces the Sq Real and RMSE metrics by 68.0% and 35.37%, respectively, compared to the second best method, NerfingMVS. It also achieves a performance very close to 1.0 on the σ_3 metric, which indicates a high accuracy of depth estimation. This demonstrates the superior performance of AltNeRF on depth estimation task, and also shows that it can learn a more reasonable scene representation than the existing methods.

5.4 Ablation Study

Here, we demonstrate the effectiveness of each component through ablation study. Tab. 4 shows the quantitative results on Flower and Scene_02. We use BARF with identity matrices as initial camera poses as baseline. First, we introduce the pose prior by initializing the camera poses of BARF with SPM estimates. This improves the performance on all metrics, which indicates the importance of the pose prior. Second, we introduce the depth prior and regularize NeRF with

Method	Flower			Scene_02		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BARF	23.988	0.744	0.183	29.682	0.932	0.053
+ pose prior	25.213	0.752	0.154	31.311	0.962	0.038
+ depth prior	24.989	0.757	0.141	34.093	0.974	0.03
+ one alternating step	25.955	0.786	0.117	35.049	0.976	0.03
+ two alternating step	26.073	0.794	0.114	35.049	0.978	0.029
+ four alternating step	26.127	0.793	0.112	35.051	0.978	0.028

Table 4: Quantitative results of ablation study. BARF is the baseline method and we gradually enable each component to demonstrate their effectiveness.

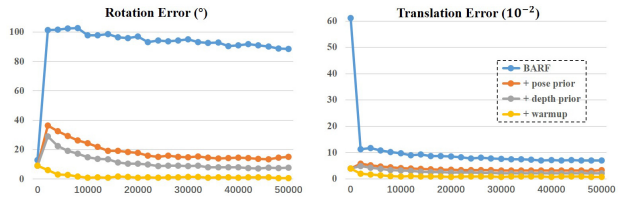

Figure 7: Ablation study on pose estimation of Scene_04. BARF is the baseline method, and we gradually enable the pose prior, the depth supervision L_e and the warmup learning to evaluate their effectiveness.

the proposed error-tolerant loss L_e. This improves the PSNR metric of Scene_02 by 8.89%, which indicates the importance of the depth prior. Third, we alternate between SPM and SRM for one, two and four times. The first alternating step significantly improves the PSNR, SSIM and LPIPS metrics of Flower by 3.87%, 3.83% and 17.02%, respectively, which indicates the effectiveness of our alternating algorithm. The gain of more alternating steps is not as significant as the first one, but can still improve performance. We think this is because the model learning converges fast at the beginning of the training and slows down afterwards, thus most of the useful information is already exchanged in the first alternation step.

We evaluate the performance of each component on pose estimation and show the results in Fig. 7. We use BARF as baseline and gradually enable the pose priors, the error-tolerant depth loss L_e and the warmup learning strategy to test their effects. The results show that the camera pose errors decrease as more components are enabled. In particular, enabling the pose priors significantly reduces the pose error, which indicates the importance of pose priors. The error-tolerant loss L_e also improves the performance over + *pose prior*, which verifies its effectiveness. With the warmup learning strategy, the errors are further reduced, leading to the most accurate pose estimation. This justifies the necessity of the warmup learning strategy for pose estimation.

6 Conclusion

Robust high-quality NeRFs require accurate camera pose and scene depth, which are hard and expensive to obtain, especially for non-technical users. In this paper, we propose a more practical approach that uses inaccurate depth-pose priors from self-supervised depth estimation to address this problem. By combining our proposed contributions, we can generate robust and high-quality NeRF models and estimate accurate camera poses at low cost.

References

Arandjelović, R.; and Zisserman, A. 2021. Nerf in detail: Learning to sample for view synthesis. *arXiv preprint arXiv:2106.05264*.

Bangunharcana, A.; Magd, A.; and Kim, K.-S. 2023. DualRefine: Self-Supervised Depth and Pose Estimation Through Iterative Epipolar Sampling and Refinement Toward Equilibrium. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Bergman, A.; Kellnhofer, P.; and Wetzstein, G. 2021. Fast training of neural lumigraph representations using meta learning. *Advances in Neural Information Processing Systems*, 34: 172–186.

Bian, J.; Li, Z.; Wang, N.; Zhan, H.; Shen, C.; Cheng, M.-M.; and Reid, I. 2019. Unsupervised scale-consistent depth and ego-motion learning from monocular video. *Advances in neural information processing systems*, 32.

Bian, W.; Wang, Z.; Li, K.; Bian, J.-W.; and Prisacariu, V. A. 2023. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Casser, V.; Pirk, S.; Mahjourian, R.; and Angelova, A. 2019. Depth prediction without the sensors: Leveraging structure for unsupervised learning from monocular videos. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, 8001–8008.

Chen, Y.; and Lee, G. H. 2023. DBARF: Deep Bundle-Adjusting Generalizable Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24–34.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3828–3838.

Gordon, A.; Li, H.; Jonschkowski, R.; and Angelova, A. 2019. Depth from videos in the wild: Unsupervised monocular depth learning from unknown cameras. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8977–8986.

Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2020. 3d packing for self-supervised monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2485–2494.

Guizilini, V.; Ambruș, R.; Chen, D.; Zakharov, S.; and Gaidon, A. 2022. Multi-Frame Self-Supervised Depth With Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 160–170.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

He, M.; Hui, L.; Bian, Y.; Ren, J.; Xie, J.; and Yang, J. 2022. RA-Depth: Resolution Adaptive Self-Supervised Monocular Depth Estimation. In *European Conference on Computer Vision*, 565–581. Springer.

Huber, P. J. 1964. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1): 73 – 101.

Jeong, Y.; Ahn, S.; Choy, C.; Anandkumar, A.; Cho, M.; and Park, J. 2021. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5846–5854.

Ji, P.; Li, R.; Bhanu, B.; and Xu, Y. 2021. Monoindoor: Towards good practice of self-supervised monocular depth estimation for indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12787–12796.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18365–18375.

Kajiya, J. T.; and Von Herzen, B. P. 1984. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3): 165–174.

Kellnhofer, P.; Jebe, L. C.; Jones, A.; Spicer, R.; Pulli, K.; and Wetzstein, G. 2021. Neural lumigraph rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4287–4297.

Kurz, A.; Neff, T.; Lv, Z.; Zollhöfer, M.; and Steinberger, M. 2022. AdaNeRF: Adaptive Sampling for Real-Time Rendering of Neural Radiance Fields. In *European Conference on Computer Vision*, 254–270. Springer.

Li, B.; Huang, Y.; Liu, Z.; Zou, D.; and Yu, W. 2021. StructDepth: Leveraging the structural regularities for self-supervised indoor depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12663–12673.

Li, T.; Slavcheva, M.; Zollhoefer, M.; Green, S.; Lassner, C.; Kim, C.; Schmidt, T.; Lovegrove, S.; Goesele, M.; Newcombe, R.; et al. 2022. Neural 3D Video Synthesis From Multi-View Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5521–5531.

Li, Z.; Li, L.; and Zhu, J. 2023. Read: Large-scale neural scene rendering for autonomous driving. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 1522–1529.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5741–5751.

Liu, L.; Song, X.; Wang, M.; Liu, Y.; and Zhang, L. 2021. Self-supervised monocular depth estimation for all day images using domain separation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12737–12746.

Liu, Y.-L.; Gao, C.; Meuleman, A.; Tseng, H.-Y.; Saraf, A.; Kim, C.; Chuang, Y.-Y.; Kopf, J.; and Huang, J.-B. 2023. Robust Dynamic Radiance Fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lyu, X.; Liu, L.; Wang, M.; Kong, X.; Liu, L.; Liu, Y.; Chen, X.; and Yuan, Y. 2021. Hr-depth: High resolution self-supervised monocular depth estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 2294–2301.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7210–7219.

Meng, Q.; Chen, A.; Luo, H.; Wu, M.; Su, H.; Xu, L.; He, X.; and Yu, J. 2021. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6351–6361.

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16190–16199.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 405–421. Cham: Springer International Publishing. ISBN 978-3-030-58452-8.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Petrovai, A.; and Nedevschi, S. 2022. Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1578–1588.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10318–10327.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12179–12188.

Ranjan, A.; Jampani, V.; Balles, L.; Kim, K.; Sun, D.; Wulff, J.; and Black, M. J. 2019. Competitive collaboration: Joint unsupervised learning of depth, camera motion, optical flow and motion segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12240–12249.

Reiser, C.; Szeliski, R.; Verbin, D.; Srinivasan, P. P.; Mildenhall, B.; Geiger, A.; Barron, J. T.; and Hedman, P. 2023. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *arXiv preprint arXiv:2302.12249*.

Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; and Novotny, D. 2021. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *International Conference on Computer Vision*.

Ren, W.; Wang, L.; Piao, Y.; Zhang, M.; Lu, H.; and Liu, T. 2022. Adaptive Co-teaching for Unsupervised Monocular Depth Estimation. In *European Conference on Computer Vision*, 89–105. Springer.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shu, C.; Yu, K.; Duan, Z.; and Yang, K. 2020. Feature-metric loss for self-supervised learning of depth and egomotion. In *European Conference on Computer Vision*, 572–588. Springer.

Smith, C.; Du, Y.; Tewari, A.; and Sitzmann, V. 2023. FlowCam: Training Generalizable 3D Radiance Fields without Camera Poses via Pixel-Aligned Scene Flow. *arXiv preprint arXiv:2306.00180*.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Tancik, M.; Mildenhall, B.; Wang, T.; Schmidt, D.; Srinivasan, P. P.; Barron, J. T.; and Ng, R. 2021. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2846–2855.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-NeRF: Scalable Construction of Large-Scale NeRFs

for Virtual Fly-Throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12922–12931.

Vankadari, M.; Garg, S.; Majumder, A.; Kumar, S.; and Behera, A. 2020. Unsupervised monocular depth estimation for night-time images using adversarial domain feature adaptation. In *European Conference on Computer Vision*, 443–459. Springer.

Wang, K.; Zhang, Z.; Yan, Z.; Li, X.; Xu, B.; Li, J.; and Yang, J. 2021a. Regularizing nighttime weirdness: Efficient self-supervised monocular depth estimation in the dark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16055–16064.

Wang, L.; Wang, Y.; Wang, L.; Zhan, Y.; Wang, Y.; and Lu, H. 2021b. Can Scale-Consistent Monocular Depth Be Learned in a Self-Supervised Scale-Invariant Manner? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12727–12736.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021c. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *TIP*, 13(4): 600–612.

Wang, Z.; Wu, S.; Xie, W.; Chen, M.; and Prisacariu, V. A. 2021d. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*.

Wei, Y.; Liu, S.; Rao, Y.; Zhao, W.; Lu, J.; and Zhou, J. 2021. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5610–5619.

Wu, C.-Y.; Wang, J.; Hall, M.; Neumann, U.; and Su, S. 2022. Toward Practical Monocular Indoor Depth Estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3814–3824.

Xu, Q.; Xu, Z.; Philip, J.; Bi, S.; Shu, Z.; Sunkavalli, K.; and Neumann, U. 2022. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5438–5448.

Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; and Kanazawa, A. 2021a. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5752–5761.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021b. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Yu, H.; Chen, A.; Chen, X.; Xu, L.; Shao, Z.; and Yu, J. 2022. Anisotropic fourier features for neural image-based rendering and relighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 3152–3160.

Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.

Zhang, N.; Nex, F.; Vosselman, G.; and Kerle, N. 2023. Lite-Mono: A Lightweight CNN and Transformer Architecture for Self-Supervised Monocular Depth Estimation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, X.; Bi, S.; Sunkavalli, K.; Su, H.; and Xu, Z. 2022. NeRFusion: Fusing Radiance Fields for Large-Scale Scene Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5449–5458.

Zhou, K.; Hong, L.; Chen, C.; Xu, H.; Ye, C.; Hu, Q.; and Li, Z. 2022. Devnet: Self-supervised monocular depth learning via density volume construction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXIX*, 125–142. Springer.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1851–1858.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, 286–301. Springer.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5745–5753.

Zou, Y.; Luo, Z.; and Huang, J.-B. 2018. Df-net: Unsupervised joint learning of depth and flow using cross-task consistency. In *Proceedings of the European conference on computer vision (ECCV)*, 36–53.

Supplementary Material of AltNeRF

In this supplementary material, we supplement the main text by introducing additional implementation details, reporting additional experimental results and analyzing the limitations and possible improvement of our methods. In the next sections, we will expand the above contents in turn.

1 Additional Implementation Detail

In this section, we provide some implementation details that are omitted in the main text. We use COLMAP (Schönberger and Frahm 2016; Schönberger et al. 2016) to obtain the camera intrinsic K and poses P for each scene from all datasets. NeRFmm (Wang et al. 2021), BARF (Lin et al. 2021), NoPe-NeRF (Bian et al. 2023), and our method only use K as input and estimate the camera poses. NeRF (Mildenhall et al. 2020), DS-NeRF (Deng et al. 2022), DDP-NeRF (Roessle et al. 2022), and NerfingMVS (Wei et al. 2021) use both K and P as input, since they do not estimate the camera poses. AltNeRF takes consecutive video frames as input, but Room and Vasedeck from LLFF dataset (Mildenhall et al. 2019, 2020) have large viewpoint changes between adjacent frames. Therefore, we only use the first 32 and 58 frames of Room and Vasedeck, respectively, for our experiments. We set the image resolution to 640×480 for all datasets.

The camera poses we learned is variable up to a 3D similarity transformation, which causes misalignment with the COLMAP poses. We use several approaches to remove the impact of this misalignment on metric evaluation. To evaluate the pose estimation performance, we follow the criteria of BARF to pre-align the learned poses with the COLMAP poses using Procrustes analysis on the camera locations. For novel view synthesis evaluation, we run an additional step of test-time photometric optimization on the learned NeRF model to reduce the pose error that may affect the image quality. For depth estimation evaluation, we scale the predicted depth maps $\hat{D}$ with a factor $s = median(D)/median(\hat{D})$ to address the scale inconsistency between $\hat{D}$ and ground-truth depth maps D.

2 Additional Experimental Result

In this section, we report additional quantitative and qualitative results that are not shown in the main text to further demonstrate the effectiveness of AltNeRF.

Scene	Rotation Error (°) ↓				Translation Error (10^{-2}) ↓			
	BARF	NeRFmm	NoPe-NeRF	Our	BARF	NeRFmm	NoPe-NeRF	Our
Fern	**0.175**	16.326	3.101	0.264	0.176	3.351	0.821	**0.171**
Flower	0.884	3.058	1.582	**0.306**	0.217	2.783	0.243	**0.202**
Fortress	**0.622**	163.741	1.060	0.919	**0.390**	28.486	0.767	0.709
Orchids	0.341	8.704	2.604	**0.340**	0.348	5.047	1.108	**0.345**
Room	0.329	6.583	3.549	**0.086**	0.277	8.519	3.163	**0.070**
Vasedeck	106.804	136.129	173.343	**0.415**	121.099	136.282	130.091	**0.770**
Scene_01	9.998	149.902	42.820	**6.654**	1.578	9.764	3.914	**1.466**
Scene_02	5.975	12.664	6.287	**2.904**	2.421	1.537	3.564	**1.397**
Scene_03	112.274	98.311	14.284	**0.946**	79.966	67.714	38.281	**1.447**
Scene_04	93.420	57.124	73.344	**0.339**	6.634	22.770	11.475	**0.668**

Table A: Quantitative comparison on camera pose estimation task. Rotation Error and Translation Error measure the difference from the COLMAP poses, for reference only.

2.1 Detailed Results on LLFF and Captures

We provide the quantitative results of each scene to complement the Tab. 1 of the main text. Tab. A shows the detailed quantitative comparison with BARF (Lin et al. 2021), NeRFmm (Wang et al. 2021) and NoPe-NeRF (Bian et al. 2023) on pose estimation task and Tab. B shows the detailed quantitative results on novel view synthesis task. The best result is **bold**, and the second is underlined.

2.2 Detailed Results on CO3D

We provide the detailed results of each scene to complement the Tab. 2 of the main text. The experiments are conducted on CO3D (Reizenstein et al. 2021) dataset. The quantitative results of novel view synthesis are reported in Tab. C, and the qualitative results of novel view synthesis and depth estimation are reported in Fig. A.

2.3 Detailed Results on ScanNet

We provide the detailed results of each scene to complement the Tab. 3 of the main text. The experiments are conducted on ScanNet (Dai et al. 2017) dataset. The quantitative and qualitative results of depth estimation task are reported in Tab. D and Fig. B, respectively.

2.4 Comparison with SC-NeRF and GNeRF

We compare AltNeRF with SC-NeRF (Jeong et al. 2021) and GNeRF (Meng et al. 2021) on three challenging real-world scenes, including Vasedeck from LLFF, and Scene_03 and Scene_04 from Captures, and report the quantitative

	Method	Fern	Flower	Fortress	Orchids	Room	Vasedeck	Scene_01	Scene_02	Scene_03	Scene_04	Mean
PSNR ↑	NeRF	25.938	25.339	28.124	19.167	**38.085**	21.523	26.204	22.657	26.698	26.350	26.009
	BARF	25.490	23.988	30.435	20.295	34.362	14.241	25.377	29.682	11.441	10.819	22.613
	NeRFmm	20.735	24.053	18.769	16.030	28.027	14.792	18.506	31.376	11.127	17.416	20.083
	DS-NeRF	25.446	25.950	30.299	20.000	33.939	21.026	26.333	29.238	27.310	26.509	26.605
	DDP-NeRF	22.562	21.244	25.228	18.788	26.641	19.534	22.879	26.979	22.093	24.861	23.081
	NoPe-NeRF	23.904	**26.888**	29.912	18.566	31.618	17.754	24.710	28.196	16.691	24.384	24.262
	Our	**26.377**	26.073	**30.977**	**20.334**	37.371	**23.340**	**28.142**	**35.049**	**28.056**	**27.612**	**28.333**
SSIM ↑	NeRF	**0.804**	0.762	0.838	0.610	**0.978**	0.673	0.897	0.870	0.866	0.912	0.821
	BARF	0.756	0.744	0.857	0.623	0.955	0.398	0.890	0.932	0.436	0.561	0.715
	NeRFmm	0.571	0.708	0.495	0.353	0.861	0.354	0.737	0.948	0.334	0.661	0.602
	DS-NeRF	0.776	0.788	0.886	0.628	0.955	0.605	0.894	0.907	0.864	0.908	0.821
	DDP-NeRF	0.751	0.607	0.857	0.580	0.875	0.467	0.836	0.929	0.746	0.881	0.753
	NoPe-NeRF	0.696	**0.817**	0.848	0.530	0.927	0.467	0.854	0.908	0.547	0.834	0.743
	Our	0.801	0.794	**0.899**	**0.648**	0.977	**0.701**	**0.910**	**0.978**	**0.886**	**0.914**	**0.851**
LPIPS ↓	NeRF	**0.168**	0.142	0.147	**0.203**	**0.027**	**0.196**	0.091	0.109	0.128	0.062	0.127
	BARF	0.294	0.183	0.112	0.266	0.073	0.643	0.136	0.053	0.766	0.745	0.327
	NeRFmm	0.492	0.187	0.588	0.436	0.285	0.719	0.380	0.037	0.792	0.399	0.432
	DS-NeRF	0.226	0.135	0.067	0.240	0.064	0.309	0.114	0.108	0.147	0.065	0.148
	DDP-NeRF	0.224	0.216	0.115	0.274	0.182	0.441	0.149	0.073	0.275	0.096	0.205
	NoPe-NeRF	0.342	**0.107**	0.105	0.244	0.136	0.656	0.180	0.077	0.461	0.210	0.252
	Our	0.196	0.114	**0.067**	0.210	0.029	0.215	**0.060**	**0.029**	**0.124**	**0.054**	**0.110**

Table B: Detailed quantitative results of novel view synthesis on LLFF and Captures.

Method	193_20797_40499			349_36504_68102			415_57184_110444		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF	34.556	0.929	**0.105**	32.459	0.906	0.154	36.073	0.948	**0.137**
BARF	9.421	0.384	0.735	12.894	0.547	0.839	15.776	0.717	0.712
NoPe-NeRF	15.214	0.495	0.615	13.088	0.593	0.739	16.353	0.698	0.647
Our	**34.687**	**0.93**	0.113	**33.083**	**0.912**	**0.145**	**37.082**	**0.949**	0.146

Table C: Detailed quantitative results of novel view synthesis on CO3D dataset.

Scene	Method	Abs Rel ↓	Sq Rel ↓	RMSE ↓	RMSE log ↓	σ_1 ↑	σ_2 ↑	σ_3 ↑
Scene0079	NeRF	0.087	0.024	0.212	0.112	0.953	0.998	1.000
	DS-NeRF	0.040	0.007	0.120	0.061	0.987	1.000	1.000
	NerfingMVS	0.040	0.005	0.099	0.055	0.998	1.000	1.000
	Our	**0.039**	**0.005**	**0.094**	**0.048**	**0.998**	**1.000**	**1.000**
Scene0553	NeRF	0.100	0.024	0.185	0.130	0.905	0.997	1.000
	DS-NeRF	0.046	0.005	0.090	0.064	0.994	1.000	1.000
	NerfingMVS	0.048	0.008	0.109	0.074	0.972	1.000	1.000
	Our	**0.041**	**0.004**	**0.075**	**0.051**	**0.998**	**1.000**	**1.000**
Scene0653	NeRF	0.241	0.167	0.540	0.370	0.557	0.878	0.901
	DS-NeRF	0.138	0.062	0.297	0.250	0.731	0.867	0.986
	NerfingMVS	0.138	0.061	0.285	0.186	0.843	0.966	0.993
	Our	**0.075**	**0.016**	**0.150**	**0.111**	**0.963**	**0.993**	**0.997**

Table D: Detailed quantitative results of depth estimation on ScanNet dataset.

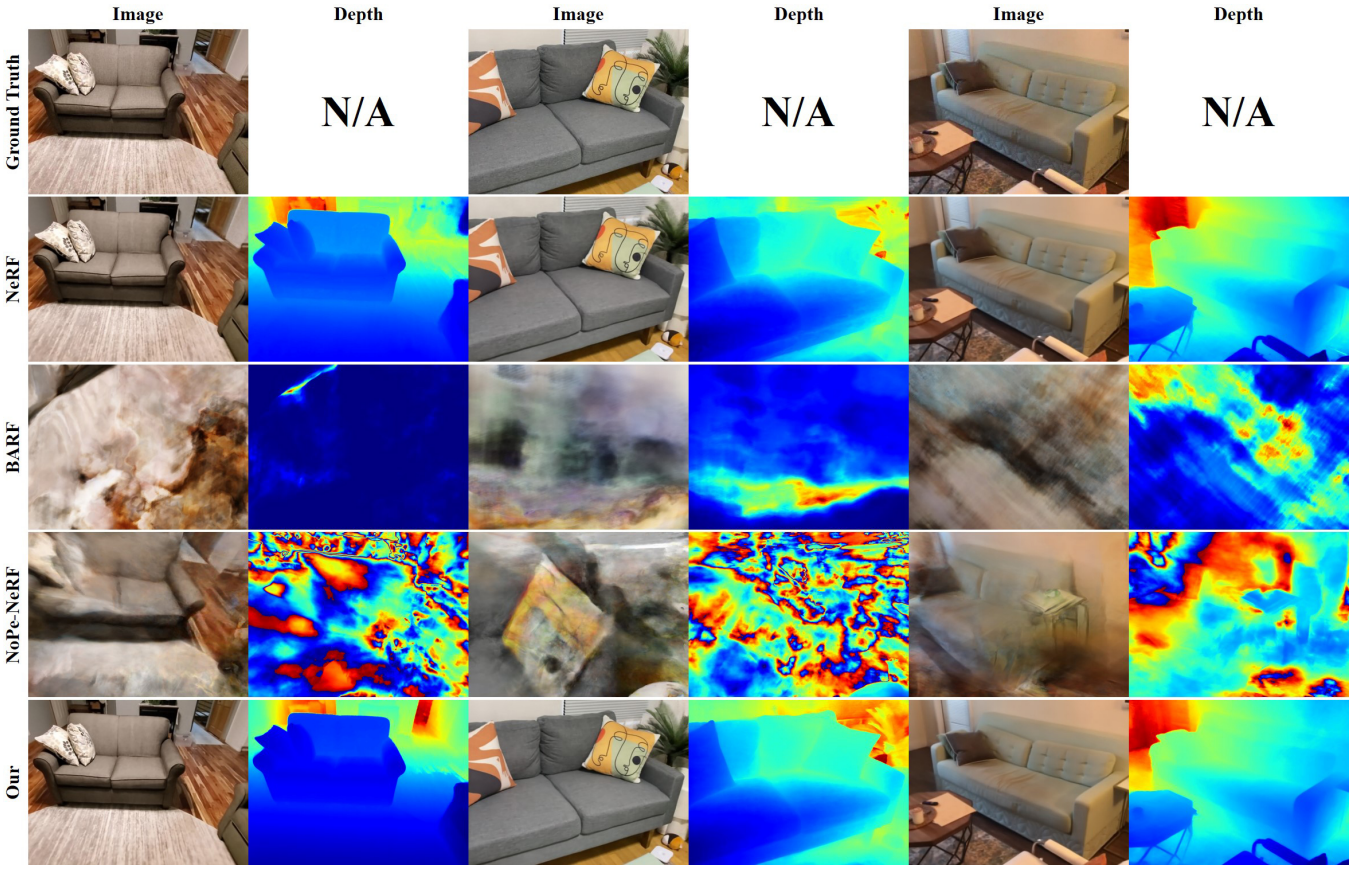

Figure A: Qualitative comparison with NeRF, BARF and NoPe-NeRF on novel view synthesis and depth estimation tasks. The reported scenes are *193_20797_40499*, *349_36504_68102* and *415_57184_110444*, respectively.

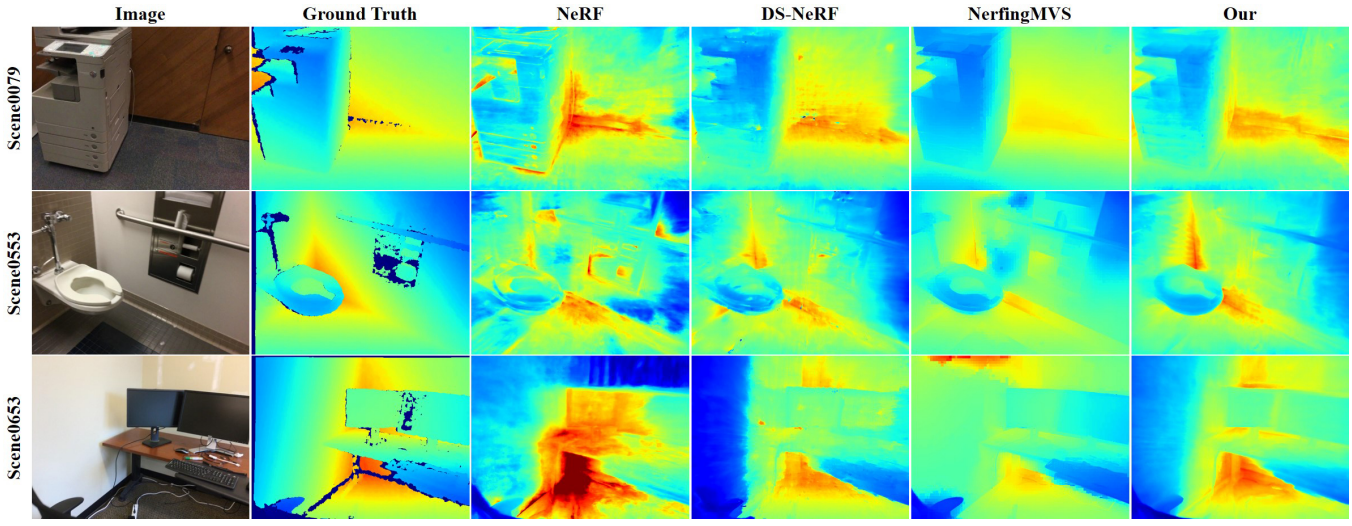

Figure B: Qualitative results of depth estimation on three scenes of ScanNet dataset. We use the same colormap for each scene to visualize the depth maps.

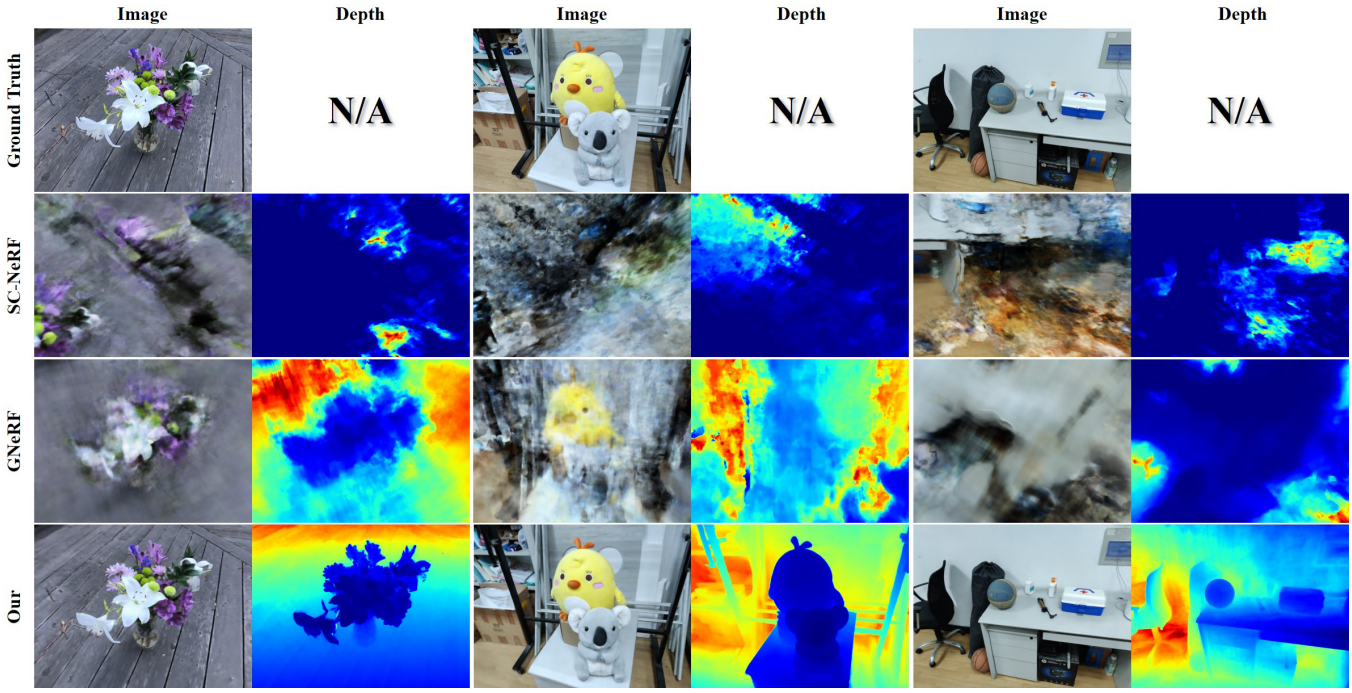

Figure C: Qualitative comparison with SC-NeRF and GNeRF on novel view synthesis and depth estimation tasks. The reported scenes are Vasedeck, Scene_03 and Scene_04.

	Scene	Vasedeck	Scene_03	Scene_04
PSNR ↑	SC-NeRF	13.108	8.583	9.304
	GNeRF	16.762	13.514	11.862
	Our	**23.340**	**28.056**	**27.612**
SSIM ↑	SC-NeRF	0.373	0.264	0.367
	GNeRF	0.447	0.415	0.55
	Our	**0.701**	**0.886**	**0.914**
LPIPS ↓	SC-NeRF	0.753	0.857	0.838
	GNeRF	0.732	0.697	0.804
	Our	**0.215**	**0.124**	**0.054**

Table E: Quantitative comparison with SC-NeRF and GNeRF on Vasedeck, Scene_03 and Scene_04.

and qualitative results in Tab. E and Fig. C, respectively. SC-NeRF and GNeRF have recently shown impressive performance to create NeRF from unposed images of facing-forward and synthetic scenes. However, they fail to synthesize plausible novel views of these challenging scenes. In contrast, AltNeRF can still render realistic novel views and reasonable depth maps.

2.5 Comparison with COLMAP Priors

We evaluate AltNeRF and a simple combination of existing methods that use depth and pose priors from COLMAP to address the challenges of 3D supervision and camera poses. We use BARF as the baseline method and introduce the depth-pose priors from COLMAP. We call this method *w/ COLMAP Priors*. The pose prior is used to initialize the camera poses of BARF and we use the same method as DS-NeRF to regularize NeRF learning with the depth prior. To synchronize the learning progress of NeRF representation and poses, we start to refine poses from the 1000th iteration. Tab. F shows the quantitative results on Fortress, Orchids, and Vasedeck from LLFF dataset. AltNeRF outperforms *w/ COLMAP Priors* on all metrics by a large margin. For example, it improves the PSNR, SSIM, and LPIPS metrics by 6.81%, 9.36%, and 15.69%, respectively, on Vasedeck. This indicates the advantages of our framework over the simple combination of existing methods. However, we note that *w/ COLMAP Priors* performs worse than the baseline method, BARF, on Orchids scene. This suggests that directly combining the depth and pose priors is not sufficient to create robust NeRF representations.

2.6 Accuracy of Initial Camera Pose

Our framework uses Self-supervised Monocular Depth Estimation (SMDE) to provide the initial camera poses. To measure its accuracy, we compare the predicted camera poses with the pseudo ground truth poses from COLMAP on three challenging scenes from CO3D dataset (Reizenstein et al. 2021). We report the quantitative results in Tab. G. Since the translation scale is variable, we also report the results of initializing the camera pose with the identity matrix as a reference for comparison.

3 Limitation and Future Work

In this section, we discuss some limitations of AltNeRF and outline some possible directions for future work.

First, the depth and pose priors are provided by SPM. To improve its generalization, we pretrain SPM on NYU-V2

Method	Fortress			Orchids			Vasedeck		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BARF	30.435	0.857	0.112	<u>20.295</u>	<u>0.623</u>	<u>0.266</u>	14.241	0.398	0.643
w/ COLMAP Priors	<u>30.605</u>	<u>0.883</u>	<u>0.085</u>	19.459	0.582	0.273	<u>21.834</u>	<u>0.641</u>	<u>0.255</u>
Our	**30.977**	**0.899**	**0.067**	**20.334**	**0.648**	**0.210**	**23.34**	**0.701**	**0.215**

Table F: Quantitative comparison with method using COLMAP priors.

Method	Metric	193_20797_40499	349_36504_68102	415_57184_110444
Identity	Rot (°)	94.622	26.845	30.100
	Trans (10^{-2})	209.417	123.619	151.640
SMDE	Rot (°)	18.592	5.972	6.356
	Trans (10^{-2})	49.489	16.794	3.621

Table G: Quantitative comparison of camera pose initialization on CO3D dataset. *Rot* and *Trans* measures the difference to COLMAP poses, respectively.

(Nathan Silberman and Fergus 2012) and VOID (Wong et al. 2020) datasets, which contain around 300K video frames in total. Although SPM performs well on current data, training it on more and diverse data could further improve its performance. Since a good scene prior helps to better learn the NeRF representations, we plan to train SPM on larger datasets in the future.

Second, AltNeRF currently requires known camera intrinsics as input. Although coarse camera intrinsics can be obtained in several simple ways, such as using the image's EXIF information or Zhang's camera calibration algorithm (Zhang 1999), removing the requirement for camera intrinsics would make AltNeRF more applicable. Therefore, we plan to explore methods for simultaneously estimating camera intrinsics and poses in the future.

Third, the current implementation of AltNeRF is based on the original NeRF (Mildenhall et al. 2020), which has some limitations in representing unbounded scenes and rendering speed. Recently, some works, such as NeRF++ (Zhang et al. 2020), Mip-NeRF 360 (Barron et al. 2022) and instant-ngp (Müller et al. 2022) have addressed these problems and achieved great progress. Therefore, we plan to incorporate these works into AltNeRF in the future.

References

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Bian, W.; Wang, Z.; Li, K.; Bian, J.-W.; and Prisacariu, V. A. 2023. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Jeong, Y.; Ahn, S.; Choy, C.; Anandkumar, A.; Cho, M.; and Park, J. 2021. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5846–5854.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5741–5751.

Meng, Q.; Chen, A.; Luo, H.; Wu, M.; Su, H.; Xu, L.; He, X.; and Yu, J. 2021. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6351–6361.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 405–421. Cham: Springer International Publishing. ISBN 978-3-030-58452-8.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4): 1–15.

Nathan Silberman, P. K., Derek Hoiem; and Fergus, R. 2012. Indoor Segmentation and Support Inference from RGBD Images. In *ECCV*.

Reizenstein, J.; Shapovalov, R.; Henzler, P.; Sbordone, L.; Labatut, P.; and Novotny, D. 2021. Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction. In *International Conference on Computer Vision*.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Schönberger, J. L.; and Frahm, J.-M. 2016. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*.

Wang, Z.; Wu, S.; Xie, W.; Chen, M.; and Prisacariu, V. A. 2021. NeRF–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*.

Wei, Y.; Liu, S.; Rao, Y.; Zhao, W.; Lu, J.; and Zhou, J. 2021. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5610–5619.

Wong, A.; Fei, X.; Tsuei, S.; and Soatto, S. 2020. Unsupervised Depth Completion From Visual Inertial Odometry. *IEEE Robotics and Automation Letters*, 5(2): 1899–1906.

Zhang, K.; Riegler, G.; Snavely, N.; and Koltun, V. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*.

Zhang, Z. 1999. Flexible camera calibration by viewing a plane from unknown orientations. In *Proceedings of the seventh ieee international conference on computer vision*, volume 1, 666–673. Ieee.