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

Kun Wang[1], Zhiqiang Yan[1], Huang Tian[1], Zhenyu Zhang[2], Xiang Li[3], Jun Li[1*] and Jian Yang[1*]

[1]PCA Lab, Nanjing University of Science and Technology, China
[2]Nanjing University, Suzhou Campus, China
[3]Nankai University, China
{kunwang, Yanzq, tianhuang, xiang.li.implus, junli, csjyang}@njust.edu.cn, zhangjesse@foxmail.com

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. Extensive experiments showcase the compelling capabilities of AltNeRF in generating high-fidelity and robust novel views that closely resemble reality.

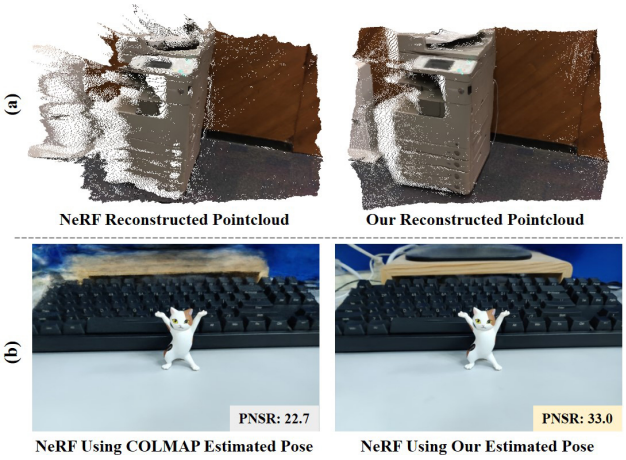

Figure 1: (a) We showcase that NeRF is prone to fitting incorrect geometry, by back-projecting a training view to point cloud using estimated radiance and density. (b) We also show how NeRF creation is affected by inaccurate poses, by optimizing NeRF with COLMAP and our pose estimation.

Introduction

Neural rendering has achieved unprecedented progress on the long-standing view synthesis task in computer vision research (Zhou et al. 2016; Kellnhofer et al. 2021; Yu et al. 2022; Li, Li, and Zhu 2023). A prominent exemplar of this task is NeRF (Mildenhall et al. 2020), which captures the continuous volumetric essence of real-world scenes using multi-view images and precise camera poses, thereby generating lifelike new perspectives. However, NeRF often struggles with suboptimal outcomes that compromise novel view synthesis and distort scene geometry. In this paper, we identify and address two primary causes for this issue. First, NeRF solely hinges on 2D image supervision, which may provide inadequate geometric constraints for textureless or view-limited scenes. As evidenced in Fig. 1 (a), NeRF is trapped into a suboptimal solution that manifests itself as incorrect scene geometry. Second, NeRF's reliance on precise camera poses for constructing proper volumetric representations is a stumbling block in the face of pose inaccuracies or noise. Such errors in the camera poses compound the optimization challenges for NeRF, as illustrated in Fig. 1 (b).

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; Wang et al. 2023) leverage depth priors as an explicit 3D supervision to prevent NeRF from fitting incorrect scene geometry. These methods derive depth priors from structure-from-motion (SFM) methodologies or depth estimation techniques, employing them as fixed 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 joint optimization of NeRF and camera poses to remove the requirement for accurate camera poses. Nevertheless, this combined task encompasses a non-convex optimization conundrum that is acutely sensitive to the initialization of cam-

*Corresponding authors

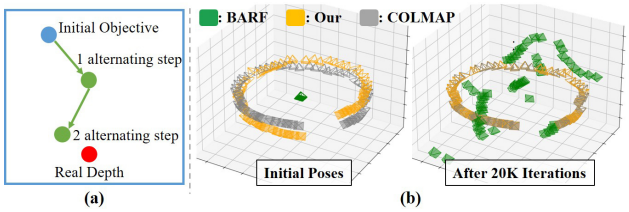

Figure 2: (a) Existing methods establish a fixed target (the blue dot) using inaccurate depth prior, whereas we leverage valuable intermediate results from NeRF to dynamically adjust the objective (the green dots) towards the real depth (the red dot). (b) Pose refinement starting from different initial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

era 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 address the previously mentioned problems, we propose AltNeRF—a novel framework designed to generate high-fidelity neural radiance fields from monocular videos. The core concept of AltNeRF is a synergistic process of self-supervised monocular depth estimation (SMDE) and NeRF optimization. SMDE can learn from accessible video data and provide robust depth and pose estimation, which can serve as depth-pose priors for NeRF. Meanwhile, NeRF maintains a continuous volume field, which can represent 3D scenes more accurately than warping-based SMDE. We propose an alternating algorithm that leverages the complementary strengths of SMDE and NeRF to progressively enhance both methodologies. Specifically, we use the pose estimated by SMDE as an effective initialization for NeRF, facilitating smoother optimization akin to the orange poses depicted in Fig. 2 (b). Furthermore, we use the estimated depth as an initial objective for NeRF, steering it away from reconstructing inaccurate scene geometries. After optimizing NeRF for many iterations, we can obtain improved pose and depth from it, which can further improve the depth estimation of SMDE. This alternation continuously updates the depth objective to converge towards actual scene depths, as illustrated in Fig. 2 (a). By harnessing the complementary strengths of SMDE and NeRF, AltNeRF can achieve more reliable 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 existing NeRF methods.

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 decomposing depth scale (Wang et al. 2021b; Yan et al. 2023), introducing 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 (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; 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. NeRF is 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; 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), 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; Chen and Lee 2023), bundle sampling (Kurz et al. 2022), initialization (Bergman, Kellnhofer, and Wetzstein 2021; Tancik et al. 2021) and data structure (Yu et al. 2021a; 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) 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 introduce depth priors (Deng et al. 2022; Roessle et al. 2022) from external sources, which may be noisy or inaccurate and result in suboptimal NeRF outcomes. In contrast, we introduce SMDE to estimate the depth-pose priors to assist NeRF optimization, and devise an alternating algorithm to harnesses the complementary strengths of SMDE and NeRF for robust NeRF creation.

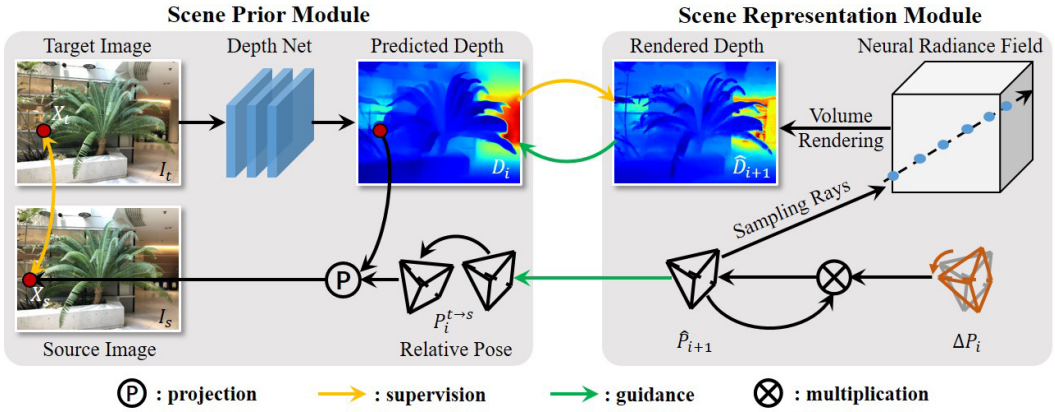

Figure 3: The overall pipeline of our AltNeRF. The scene prior module estimates depth and pose, which serves as the depth reference and initial poses, respectively. The scene representation module simultaneously refines the initial poses with ΔP_i and learns 3D scene representation, which is regularized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

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 warping 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. It 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, via

$$\hat{D}(r) = \int_{t_n}^{t_f} T(t)\sigma(t)t dt, \tag{4}$$

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:

$$L_c = \|\hat{C}(r) - C(r)\|_2. \tag{5}$$

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.

Scene Prior Module

Pretraining NeRF is optimized individually for each scene, and thus lacks the prior knowledge for scene understanding. To address this limitation, we pretrain SPM on a large dataset to accumulate the prior knowledge for 3D recovering. SPM is built on SMDE, which only requires accessible video data for training. However, SMDE is susceptible to dynamic objects and view-dependent appearances, which can degrade its performance. Therefore, we employ the distilling strategy introduced in (Wu et al. 2022) to mitigate these disadvantages:

$$L_r = 1 - SSIM(D, D_r) + 0.1 \times (E_r \oplus E/\,size(E)), \tag{6}$$

where D_r is the reference depth map produced by an off-the-shelf relative depth estimator, DPT(Ranftl, Bochkovskiy,

and Koltun 2021), $\oplus$ 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. The final loss of this stage is:

$$L_{pt} = L_p + L_r + 1.0e^{-3} \times L_s. \tag{7}$$

Test-time Adaptation The input for AltNeRF is also video data. To adapt SPM to the target video, we fine-tune it for a little iterations using the input video data. This helps close the domain gap between the target and the training data. However, SMDE produces relative depths defined up to an unknown scale factor, which can cause potential inconsistency across frames. Therefore, we introduce the geometry consistency loss from (Bian et al. 2019) to ensure the scale consistency:

$$L_g = \frac{\|D_s(x_s) - D_t(x_t)\|_1}{D_s(x_s) + D_t(x_t)}, \tag{8}$$

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

$$L_{ad} = L_{pt} + 0.5 \times L_g. \tag{9}$$

Pose Conversion To register each camera to an unified world coordinate, we use the following procedure. We establish a world coordinate system that aligns with the camera coordinate system of the first frame I_0, whose pose matrix is an identity matrix. SPM predicts the relative 3D transformations $P_{i-1\to i}$ between adjacent frames, which we use to calculate the camera poses P_i of subsequent frames. We apply the chain rule to compute the camera poses via $P_i = P_{i-1\to i} \times P_{i-1}$.

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 details below.

Depth Regularization Recovering 3D geometry from the view-limited scenes (*e.g.* forward-facing scenes) or textureless scenes (*e.g.* indoor scenes) is an ill-posed problem, since there are numerous incorrect shapes that can also explain the input images. To address this problem, we introduce the depth prior estimated by SPM as an explicit 3D supervision for NeRF. This helps mitigate the shape ambiguity that can mislead the NeRF optimization to a degenerate solution. Specifically, we enforce the consistency between depth prior D and NeRF rendered depth $\hat{D}$. However, we do not strictly align the rendered depth with the depth prior as previous works did (Deng et al. 2022; Roessle et al. 2022), since the depth prior itself is not precise either. Instead, we propose an error-tolerant depth regularization that enforces the rendered depth to fit a possible depth range:

$$L_e = H\left(\max\left(\frac{\|\hat{D} - D\|_1}{\hat{D} + D} - \epsilon, 0\right)\right), \tag{10}$$

where $H(\cdot)$ denotes the Huber loss (Huber 1964), and ϵ is a tolerance coefficient controlling the length of the possible depth range.

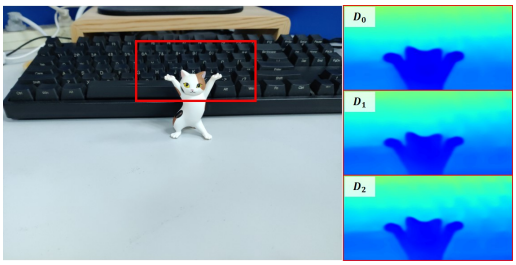

Figure 4: We illustrate that the depth estimation of SPM is improved by visualizing the initial depth estimate D_0, and depth estimates D_1 and D_2 after 1 and 2 alternating steps.

Improved Pose Initialization To improve the joint optimization of camera pose and NeRF representation, which is a highly non-convex problem that is prone to converging to a suboptimal solution when the initial pose is far from the actual pose, we use a pose prior estimated by SPM as initialization for SRM. This pose prior, denoted as P_0, is closer to the actual pose, which helps it converge to the global minimum. We then refine it by optimizing a residual pose ΔP that represents the difference between them. The refined pose, denoted by $\hat{P}$, is calculated by combining them through $\hat{P} = \Delta P \times P$.

Warmup Learning To improve the joint optimization of camera pose and scene representation, we propose a warmup learning strategy that synchronizes the learning process for these two tasks. SRM learns the scene representation from scratch, but refines the camera pose using a good pose prior. This asynchrony can result in an incorrect update direction for the camera pose. Therefore, we set the learning rate of ΔP to a low value at the beginning of the training and linearly increase it to the original learning rate after 1K iterations. This allows SRM to learn a prototype of the scene representation before updating the camera pose.

The final loss for optimizing SRM consists of both reconstruction loss and depth regularization:

$$L_{sr} = L_c + \gamma \cdot L_e, \tag{11}$$

where γ is a scalar hyper-parameter that balances these two terms of losses.

Alternating Algorithm

To achieve robust NeRF creation, we propose an alternating algorithm that synergistically boosts SPM and SRM based on their complementary advantages. SPM can produce robust depth and pose estimates as a good scene prior for SRM, but these estimates are not accurate enough to recover precise 3D scenes. On the other hands, SRM learns a continuous volumetric field that can accurately represent 3D space with fine geometries and improves the initial pose through bundle adjustment. Therefore, SRM can overcome the shortcomings of SPM, while an improved SPM can provide more accurate depth regularization which helps SRM achieve better results. The overall framework is illustrated in Fig. 3. In the following sections, we will introduce the workflow of

the alternating algorithm, the multi-view consistency check, and discuss its potential extension.

Workflow We denote the alternating step as i, SPM at step i as Φ_i, and SRM as Ψ_i. The alternation is formulated as:

$$\begin{aligned} &\Psi_i : (D_i, P_i) \rightarrow (\hat{D}_{i+1}, \hat{P}_{i+1}), \\ &\Phi_i : (\hat{D}_{i+1}, \hat{P}_{i+1}) \rightarrow (D_{i+1}, P_{i+1}). \end{aligned} \tag{12}$$

In each cycle, SRM takes P_i as initial pose and simultaneously learns scene representation and pose refinement under the regularization of D_i. This process is supervised by Eq. (11). After Sr iterations, SRM outputs the estimated scene depth $\hat{D}_{i+1}$ and improved camera pose $\hat{P}_{i+1}$. These improved outcomes are then used to fine-tune the depth estimation network f_d of SPM. Specifically, we use Eq. (6) to distill structure information from SRM by replacing the reference depth D_r with rendered depth $\hat{D}_{i+1}$. The final loss that supervises the fine-tuning process is Eq. (7). We fine-tune SPM for S_p iterations. Through these cyclist steps, both SPM and SRM can be improved. We showcase the improved depth estimation of SPM in Fig. 4.

Multi-view Consistency Check To assess the possible unreliability of the depth estimates from SPM and SRM, we use a multi-view consistency check to measure the uncertainty of the predicted depth maps. We denote a depth map of a target image I_t as D_t, and 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 $\hat{P}$ from SRM. We expect D_t and $D_{s\rightarrow t}$ to be identical, except for occlusions. Therefore, we define an uncertainty U_t of depth map D_t as the difference between D_t and $D_{s\rightarrow t}$: $U_t = \|D_t - D_{s\rightarrow t}\|_1$. To account for occlusions, we compute the average difference from four views with the smallest differences. We incorporate this depth uncertainty into our loss functions (*i.e*. Eq. (6) and Eq. (10)) by weighting them with the Softmin$(\cdot)$ function. This helps to mitigate the affect of unreliable depth estimates.

Discussion Our alternating algorithm is a general method that can actually leverage any depth-pose priors, not just those learned from SMDE. 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 for robust NeRF creation.

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 introduce the datasets and implementation details, and then report the experiment results.

Dataset

We evaluate AltNeRF on four datasets: LLFF (Mildenhall et al. 2019), CO3D (Reizenstein et al. 2021), ScanNet (Dai et al. 2017) and our collected dataset, named Captures. Each dataset contains different scenes with varying levels of complexity and camera motion. We employ the same train/test data division as BARF, which uses the first 90% of frames for training and the remaining 10% for testing. **LLFF**: we select five scenes from LLFF for evaluation: Fern, Flower, Fortress, Orchids and Room. We also incorporate Vasedeck from nerf_llff_360 into this set. **CO3D**: we randomly select three scenes from the Couch category of this dataset: *193_20797_40499*, *349_36504_68102* and *415_57184_110444*. These three scenes have more than 80 frames per scene and exhibit complex camera motions with simultaneous panning and rotation. **ScanNet**: we randomly select 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 reduce each scene to 20 frames to enlarge the viewpoint difference of adjacent frames. **Captures**: Captures consists of four scenes that we collect with our smartphone. Two scenes are forward-facing (Scene_01 and Scene_02) and two scenes are inner-facing with semicircular camera trajectory (Scene_03 and Scene_04). This dataset is challenging as it contains many textureless views and its viewpoint diversity is also limited.

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. SPM is pretrained on around 300K images collected from NYUv2 (Nathan Silberman and Fergus 2012) and VOID (Wong et al. 2020). The scene representation function $f_n(\cdot, \cdot)$ in SRM uses the same network structure as NeRF, *i.e*. eight fully-connected layers with skip connections for density output, and one linear layer for color.

The γ in Eq. (11) 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 10^{-4}, and fine-tune it with 5.0×10^{-5}. For SRM, the initial learning rate for NeRF learning is set to 10^{-3}, and exponentially decays to 10^{-4} throughout the training process. The initial learning rate for pose refinement is set to 10^{-5}, and linearly increases to 2.0×10^{-3} after 1K iterations before exponentially decaying to 10^{-5}. The number of iterations S_r and S_p are set to 50K and 500, respectively, and we perform two alternations 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.

Comparing with Existing Method

Here, we evaluate AltNeRF on novel view synthesis, depth estimation and camera pose estimation tasks, and compare it with existing methods to showcase its SOTA performance.

Evaluation on LLFF and Captures We compare AltNeRF with existing methods on novel view synthesis task. The compared methods are NeRF (Mildenhall et al. 2020),

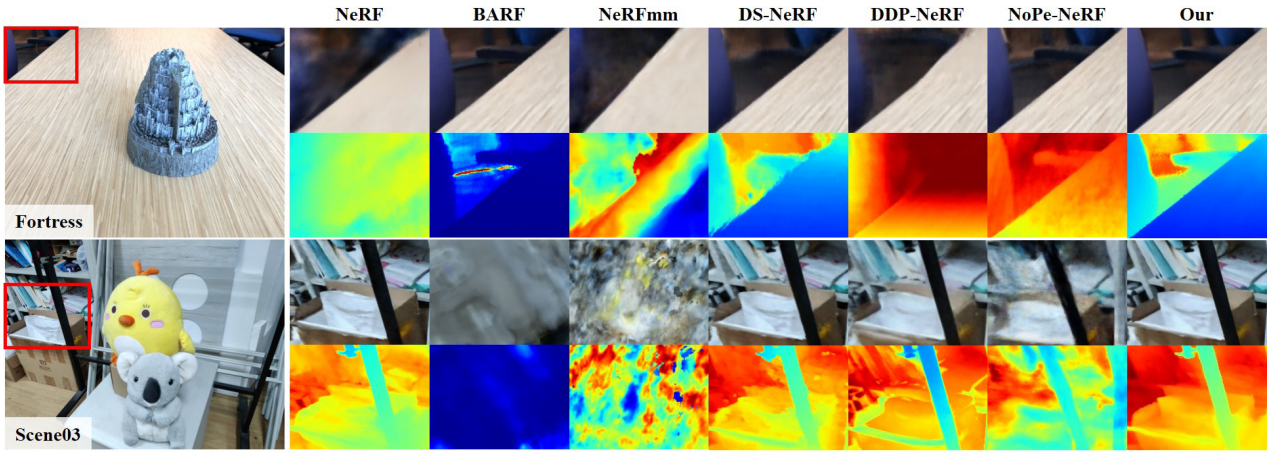

Figure 5: Qualitative comparisons of novel view synthesis and depth estimation on LLFF and Captures datasets.

Method	*LLFF*			*Captures*		
	PSNR↑	SSIM↑	LPIPS ↓	PSNR↑	SSIM↑	LPIPS↓
NeRF	26.36	0.778	0.147	25.48	0.886	0.098
BARF	24.80	0.722	0.262	19.33	0.705	0.425
NeRFmm	20.40	0.557	0.451	19.61	0.670	0.402
DS-NeRF	26.11	0.773	0.174	27.35	0.893	0.109
DDP-NeRF	22.33	0.690	0.242	24.20	0.848	0.148
NoPe-NeRF	24.77	0.714	0.265	23.50	0.786	0.232
Our	**27.41**	**0.803**	**0.139**	**29.72**	**0.922**	**0.067**

Table 1: Quantitative comparison on novel view synthesis. The best result is in bold, and the second is underlined.

Method	*Fortress*		*Orchids*		*Vasedeck*	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
DS-BARF	30.61	0.883	19.46	0.582	21.83	0.641
Our	**30.98**	**0.899**	**20.33**	**0.648**	**23.34**	**0.701**

Table 2: Quantitative comparison on novel view synthesis with COLMAP assisted baseline, DS-BARF.

Method	Abs Rel↓	Sq Rel↓	RMSE↓	σ_1↑	σ_2↑	σ_3↑
NeRF	0.143	0.072	0.312	80.5	95.8	96.7
DS-NeRF	0.075	0.025	0.169	90.4	95.6	99.5
NerfingMVS	0.075	0.025	0.164	93.8	98.9	99.8
Our	**0.051**	**0.008**	**0.106**	**98.7**	**99.8**	**99.9**

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

Method	*Couch_193*		*Couch_349*		*Couch_415*	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
NeRF	34.56	0.929	32.46	0.906	36.07	0.948
BARF	9.42	0.384	12.89	0.547	15.78	0.717
DS-NeRF	32.48	**0.939**	31.19	0.887	34.97	0.928
NoPe-NeRF	15.21	0.495	13.09	0.593	16.35	0.698
Our	**34.69**	0.930	**33.08**	**0.912**	**37.08**	**0.949**

Table 4: Quantitative comparison on novel view synthesis task. The experiments are conducted on CO3D.

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). These methods aim to address the shape ambiguity or camera pose requirement of NeRF. Tab. 1 shows the mean quantitative results. We use PSNR, SSIM (Wang et al. 2004) and LPIPS (Zhang et al. 2018) metrics to evaluate the image synthesis performance. We employ identity matrices to initialize the camera pose of BARF and NeRFmm. The LLFF and Captures datasets do not have dense depth ground truth, thus we directly use the pretrained depth completion model provided by DDP-NeRF, as its dense depth prior. In general, our method outperforms the competitors on all metrics. On LLFF, it outperforms the second best method, NeRF, by 3.98%, 3.21% and 5.44% on PSNR, SSIM and LPIPS, respectively. The improvement on Captures is more significant since the competitors perform pool on textureless and view-limited scenes. Specifically, AltNeRF improves the second best method, DS-NeRF, by 8.66%, 3.25% and 38.53% on these three metrics.

Existing methods focus on either addressing the shape ambiguity or the pose requirement of NeRF, while our method addresses these two problems simultaneously. To further demonstrate the advantages of our method, we combine BARF with DS-NeRF as a new baseline, which can also optimize pose and regularize NeRF learning. We denote this new baseline as *DS-BARF*. It initializes the camera pose using COLMAP (Schönberger and Frahm 2016) estimated pose, and supervises NeRF learning using the same method as DS-NeRF, with the depth estimated by COLMAP. We freeze its camera pose for the first 1K iterations to align the learning process of camera pose and NeRF. We compare with this baseline on three scenes of LLFF, and report the quantitative results in Tab. 2. The results demonstrates that simply combining existing methods is insufficient for high-quality NeRF creation.

Fig. 5 shows the qualitative comparisons on novel view synthesis and depth estimation tasks. Each method is eval-

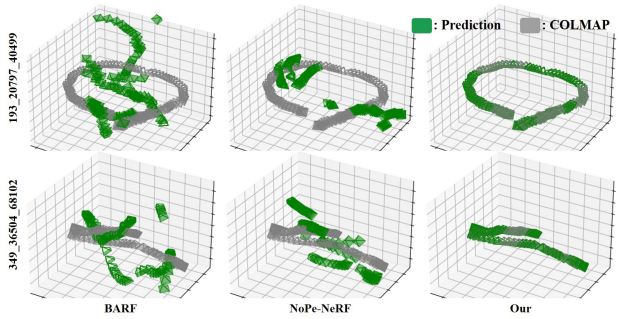

Figure 6: Qualitative results of pose estimation on CO3D.

Method	*Flower*			*Scene_02*		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
BARF	23.99	0.744	0.183	29.68	0.932	0.053
+ pose prior	25.21	0.752	0.154	31.31	0.962	0.038
+ depth prior	24.99	0.757	0.141	34.09	0.974	0.030
1 alternation	25.96	0.786	0.117	35.05	0.976	0.030
2 alternations	26.07	**0.794**	0.114	35.05	0.978	0.029
4 alternations	**26.13**	0.793	**0.112**	**35.05**	**0.978**	**0.028**

Table 5: Quantitative results for ablation study. BARF is the baseline method and we gradually enable each component to demonstrate their effectiveness.

uated on two scenes from LLFF and Captures datasets. The results show that AltNeRF can synthesize realistic novel views and reasonable depth maps for both scenes. 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.

Evaluation on ScanNet ScanNet has high-quality depth ground truth, therefore we use it to demonstrate the superior geometry reconstruction ability of AltNeRF. Specifically, we compare AltNeRF with three existing methods, namely NeRF, DS-NeRF, and NerfingMVS, and report the quantitative results in Tab. 3. We use three error metrics, Abs Rel, Sq Rel and RMSE, and three accuracy metrics, σ_1, σ_2 and σ_3 (%), to measure the quality of the estimated depth maps. Our method outperforms the competitors by a large margin on each scene. It reduces the Sq Rel and RMSE metrics by 68.0% and 35.37%, respectively, compared to the second best method, NerfingMVS. It also achieves a near-perfect performance on the σ_3 metric, which indicates a high accuracy of depth estimation. This demonstrates that AltNeRF can perform well on the depth estimation task, and also shows that it can learn a more accurate scene representation than the existing methods.

Evaluation on CO3D CO3D dataset contains long-length videos with more complicated room-level camera motion. We use this dataset to demonstrate that AltNeRF is also applicable in these challenging scenes. First, we report the quantitative results on novel view synthesis task in Tab. 4. Note that, SFM methods (*e.g.* COLMAP) usually work well when input images are abundant. However, AltNeRF still

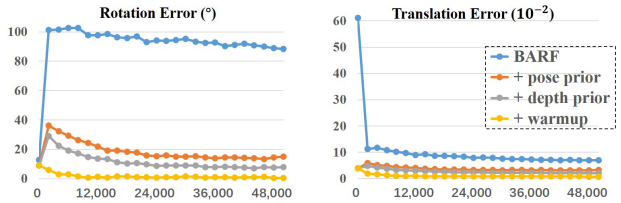

Figure 7: Ablation study on pose estimation on Scene_04. The x-axis indicates the number of iterations. BARF is the baseline method, and we gradually enable each component to test their performance.

outperforms NeRF and DS-NeRF, which employ the camera pose and depth estimated by COLMAP. This demonstrates the superiority of our method over the COLMAP assisted approaches. To evaluate AltNeRF on the camera pose estimation task, we use the camera pose estimated by COLMAP as pseudo ground truth and report the qualitative comparison with BARF and NoPe-NeRF in Fig. 6. Our predictions are closely coincide with those of COLMAP, while BARF and NoPe-NeRF fail to produce meaningful pose outputs.

Ablation Study

We demonstrate the effectiveness of each component of AltNeRF through ablation study reported in Tab. 5. We use BARF with identity matrices initialization as the baseline. First, we introduce the pose prior by initializing the camera poses with SPM estimates. This improves the performance on all metrics. Then, we introduce the depth prior and regularize NeRF with the proposed error-tolerant loss. This improves most metrics except for PSNR of Flower. We attribute this to the inaccurate depth prior, which impairs the learning of scene representation. The inaccurate depth prior can be improved via alternation. We report the results using 1, 2 and 4 alternations in the last three rows of this table. More alternating steps can consistently improve the performance, although the improvement is decreasing.

We evaluate the performance of each component on pose estimation and report the results in Fig. 7. We use BARF as baseline and gradually enable the pose priors, the error-tolerant depth loss 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. The error-tolerant loss 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.

Conclusion

NeRf creation often suffers from suboptimal solutions due to the lack of explicit 3D supervision and imprecise camera poses. In this paper, we propose a alternation-based framework that harmoniously melds self-supervised depth estimation and neural rendering to address these problems. Our method can produce high-quality NeRF representations and accurate camera poses only from monocular videos.

Acknowledgments

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was partially supported by the National Natural Science Foundation of China under Grant 62361166670, 62072242 and 62376121, the Fundamental Research Funds for the Central Universities under Grant 070-63233084, the Young Scientists Fund of the National Natural Science Foundation of China under Grant 62206134 and the Tianjin Key Laboratory of Visual Computing and Intelligent Perception. Note that the PCA Lab is associated with Key Lab of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, and Jiangsu Key Lab of Image and Video Understanding for Social Security, School of Computer Science and Engineering, Nanjing University of Science and Technology.

References

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 *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 *CVPR*, 5470–5479.

Bergman, A.; Kellnhofer, P.; and Wetzstein, G. 2021. Fast training of neural lumigraph representations using meta learning. *NIPS*, 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. *NIPS*, 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 *CVPR*.

Chen, Y.; and Lee, G. H. 2023. DBARF: Deep Bundle-Adjusting Generalizable Neural Radiance Fields. In *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 *CVPR*, 5828–5839.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 12882–12891.

Godard, C.; Mac Aodha, O.; Firman, M.; and Brostow, G. J. 2019. Digging into self-supervised monocular depth estimation. In *ICCV*, 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 *ICCV*.

Guizilini, V.; Ambrus, R.; Pillai, S.; Raventos, A.; and Gaidon, A. 2020. 3d packing for self-supervised monocular depth estimation. In *CVPR*, 2485–2494.

Guizilini, V.; Ambruș, R.; Chen, D.; Zakharov, S.; and Gaidon, A. 2022. Multi-Frame Self-Supervised Depth With Transformers. In *CVPR*, 160–170.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*, 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 *ECCV*, 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 *ICCV*, 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 *ICCV*, 12787–12796.

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

Kellnhofer, P.; Jebe, L. C.; Jones, A.; Spicer, R.; Pulli, K.; and Wetzstein, G. 2021. Neural lumigraph rendering. In *CVPR*, 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 *ECCV*, 254–270. Springer.

Li, Z.; Li, L.; and Zhu, J. 2023. Read: Large-scale neural scene rendering for autonomous driving. In *AAAI*, 2.

Lin, C.-H.; Ma, W.-C.; Torralba, A.; and Lucey, S. 2021. Barf: Bundle-adjusting neural radiance fields. In *ICCV*.

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 *ICCV*, 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 *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 *AAAI*, 3.

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 *CVPR*, 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 *ICCV*, 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 *CVPR*.

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. *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 *ECCV*, 405–421.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *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*.

Petrovai, A.; and Nedevschi, S. 2022. Exploiting Pseudo Labels in a Self-Supervised Learning Framework for Improved Monocular Depth Estimation. In *CVPR*, 1578–1588.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 10318–10327.

Ranftl, R.; Bochkovskiy, A.; and Koltun, V. 2021. Vision transformers for dense prediction. In *ICCV*, 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 *CVPR*, 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 *ICCV*.

Ren, W.; Wang, L.; Piao, Y.; Zhang, M.; Lu, H.; and Liu, T. 2022. Adaptive Co-teaching for Unsupervised Monocular Depth Estimation. In *ECCV*, 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 *CVPR*.

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 *CVPR*.

Shu, C.; Yu, K.; Duan, Z.; and Yang, K. 2020. Feature-metric loss for self-supervised learning of depth and ego-motion. In *ECCV*, 572–588. Springer.

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 *CVPR*, 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 *CVPR*, 2846–2855.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-NeRF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. In *CVPR*, 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 *ECCV*, 443–459. Springer.

Wang, C.; Sun, J.; Liu, L.; Wu, C.; Shen, Z.; Wu, D.; Dai, Y.; and Zhang, L. 2023. Digging into Depth Priors for Outdoor Neural Radiance Fields. In *ACM MM*, 1221–1230.

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 *ICCV*, 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 *ICCV*, 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 *CVPR*, 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 *ICCV*, 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.

Wu, C.-Y.; Wang, J.; Hall, M.; Neumann, U.; and Su, S. 2022. Toward Practical Monocular Indoor Depth Estimation. In *CVPR*, 3814–3824.

Yan, Z.; Wang, K.; Li, X.; Zhang, Z.; Li, J.; and Yang, J. 2023. Desnet: Decomposed scale-consistent network for unsupervised depth completion. In *AAAI*, 3, 3109–3117.

Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; and Kanazawa, A. 2021a. Plenoctrees for real-time rendering of neural radiance fields. In *ICCV*, 5752–5761.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021b. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 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 *AAAI*, 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, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 586–595.

Zhou, T.; Brown, M.; Snavely, N.; and Lowe, D. G. 2017. Unsupervised learning of depth and ego-motion from video. In *CVPR*, 1851–1858.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *ECCV*.

Zhou, Y.; Barnes, C.; Lu, J.; Yang, J.; and Li, H. 2019. On the continuity of rotation representations in neural networks. In *CVPR*, 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 *ECCV*, 36–53.