SiNeRF: Sinusoidal Neural Radiance Fields for Joint Pose Estimation and Scene Reconstruction

Yitong Xia[1]
yitongxia@student.ethz.ch
Hao Tang[1]
hao.tang@vision.ee.ethz.ch
Radu Timofte[1, 2]
radu.timofte@vision.ee.ethz.ch
Luc Van Gool[1]
vangool@vision.ee.ethz.ch

[1] Computer Vision Laboratory
D-ITET
ETH Zürich
Switzerland

[2] Computer Vision Laboratory
CAIDAS, Institute of Computer Science
University of Würzburg
Germany

Abstract

NeRFmm [25] is the Neural Radiance Fields (NeRF) that deal with *Joint Optimization* tasks, i.e., reconstructing real-world scenes and registering camera parameters simultaneously. Despite NeRFmm producing precise scene synthesis and pose estimations, it still struggles to outperform the full-annotated baseline on challenging scenes. In this work, we identify that there exists a systematic sub-optimality in joint optimization and further identify multiple potential sources for it. To diminish the impacts of potential sources, we propose *Sinusoidal Neural Radiance Fields* (SiNeRF) that leverage sinusoidal activations for radiance mapping and a novel *Mixed Region Sampling* (MRS) for selecting ray batch efficiently. Quantitative and qualitative results show that compared to NeRFmm, SiNeRF achieves comprehensive significant improvements in image synthesis quality and pose estimation accuracy. Codes are available at https://github.com/yitongx/sinerf.

1 Introduction

Adopting neural networks for Novel View Synthesis (NVS) has gained popularity. The community has achieved progress on various of representation forms, including multi-layer image plane [20, 23], distance-based representation [2, 16, 18], volume-based representation [9, 11], etc. Among all, Neural Radiance Fields (NeRF) [11] are receiving growing attention for their concise structure and compelling synthesis image quality. NeRF implicitly represents scene space with a continuous radiance mapping function parameterized by a multi-layer perceptron (MLP), followed by volume rendering [5] to composite intermediate color and density outputs into final synthesis.

Despite the compelling performances on scene reconstruction, NeRF-based methods are all trained on images with annotated camera parameters. Yet real-world scene images with

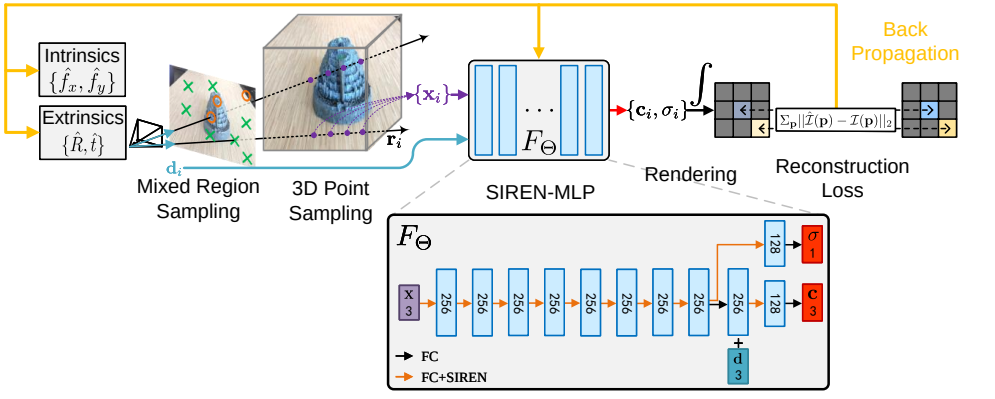

Figure 1: General overview of SiNeRF. Our proposed *Mixed Region Sampling* contains both key point ray candidates (in orange circles) and random ray candidates (in green crosses). The reconstruction loss updates both SiNeRF and camera parameters. We empirically scale σ by 25 to avoid faded synthesis.

precisely annotated camera parameters are always expensive and are not accessible all the time. NeRFmm [25] proposes an end-to-end NVS framework without camera annotations, reconstructs high-fidelity real-world scenes, and estimates accurate poses comparable to a fully-annotated baseline. But NeRFmm is reported to struggle on textureless scenes where joint optimization can easily fall into local minima.

In this work, we aim to improve NeRFmm by alleviating its systematic sub-optimality. Inspired by the smooth nature and powerful expressiveness in complex signals of periodic activations, we design a novel **Si**nusoidal **Ne**ural **R**adiance **F**ields (SiNeRF) for joint optimization. We further reveal the inefficiency of conventional random sampling and propose a novel named *Mixed Region Sampling* that allocates different weights to each pixel and samples from candidates strategically.

To conclude our contributions, in this work

- we propose a novel neural radiance field named *SiNeRF* for alleviating the systematic sub-optimality of joint optimization in NeRFmm.
- we reveal the inefficiency of *Random Sampling* and propose a novel *Mixed Region Sampling* strategy that proved to be beneficial for tasks on challenging scenes. We prove that its combination with *SiNeRF* provides the best performances.
- comprehensive quantitative, qualitative results, and ablation study on real-world scene dataset show our method's **comprehensive improvements** on camera pose estimation accuracy and novel view synthesis quality compared to NeRFmm.

2 Related Work

Neural Scene Representation. Mildenhall *et al.* [11] propose to encode scene representation inside a multi-layer perceptron (MLP) that directly regresses raw color and density. The final synthesis is composited by volume rendering [5, 10] which is differentiable for

backpropagating reconstruction loss. NeRF's success on high-fidelity NVS tasks has expanded its applications on series of vision tasks, *e.g.*, scene relighting [21], dynamic scene reconstruction [7, 13, 14], real-time scene synthesis [3, 12], etc.

Scene Reconstruction with Imperfect Camera Annotations. Recently some NeRF-related works tackle scene reconstruction tasks without accurately annotated camera parameters. Our work is improved upon NeRFmm [25], which proposes an end-to-end framework that achieves compelling NVS performances without both camera intrinsics and extrinsics. A similar pipeline is proposed by iNeRF [26], yet it only estimates poses of unknown images, and its *Interest Region Sampling* inspires us to improve sampling strategy for joint optimization. BARF [8] builds the connection between 2D image alignment and 3D scene reconstruction and uses a coarse-to-fine encoding adjustment for efficient training. Yet BARF is equipped with known intrinsic, and it initializes extrinsic with priors, whereas our work estimates both intrinsics and extrinsics with non-prior initializations. SCNeRF [4] focuses on self-calibrating image distortions and does not output pose estimations, whose task concentrations are different from ours.

Sinusoidal-Activated Multi-Layer Perceptron. SIREN [19] is the first to use sinusoidal-activated MLP for implicit neural representation. A SIREN-MLP is found to have rich expressiveness for representing zero- and first-order complex signals and thus relieving the network's hard prerequisite on Fourier-based input encoding. π-GAN [1] applies sinusoidal activations for Generative Adversarial Networks and achieves disentanglements on viewing-angle control and implicit 3D scene representation. Inspired by prior works, to alleviate the sub-optimality of joint optimization in NeRFmm, we adopt sinusoidal activations into our SiNeRF for scene reconstruction and camera parameter estimations and further stabilize the training with our novel sampling strategy.

3 Methods

3.1 NeRFmm Preliminary

NeRFmm [25] reconstructs 3D scenes from sparse scene images $\mathbf{I} = \{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_N\}$ **without** annotated camera extrinsics $\mathbf{T} = \{\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_N\}$ and intrinsics $\mathbf{f} = \{f_x, f_y\}$. A continuous function, parameterized by a multilayer perceptron (MLP) Θ, is used for view-dependent radiance mapping: $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, where $\mathbf{x}$ is the point location in the implicit scene space, $\mathbf{d}$ is the corresponding unit-length viewing direction, $\mathbf{c} \in \mathbb{R}^3$ and $\sigma \in \mathbb{R}$ are the raw color and density values, respectively.

The volume rendering [5, 10], denoted as operator $\mathcal{R}$, is used for compositing raw color and density values into final RGB pixels. Given a 2D pixel location $\mathbf{p} \in \mathbb{R}^2$ of the i-th image and a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ that meets $\mathbf{p}$ on the image plane, the final color outputs would be:

$$\hat{\mathcal{I}}_i(\mathbf{p}) = \mathcal{R}(\mathcal{T}_i, \mathbf{p}; \Theta) = \sum_{i=1}^{n} W_i \left(1 - \exp(-\sigma_i \delta_i)\right) \mathbf{c}_i, \; W_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \, \delta_j\right), \tag{1}$$

where W_i is the accummulated transmittance of the ray $\mathbf{r}(t)$. $\mathbf{o} \in \mathbb{R}^3$ denotes the camera origin, $t \in [t_n, t_f]$ denotes the sampling point location within the nearest and farest field. $(\mathbf{c}_i, \sigma_i)$ are the raw color-density values of the i-th sampling point. $\delta_i = t_{i+1} - t_i$ is the interval between two sampling points along the ray. And we acquire a set of synthesized scene images $\hat{\mathbf{I}} = \{\hat{\mathcal{I}}_1, \hat{\mathcal{I}}_2, \ldots, \hat{\mathcal{I}}_N\}$.

Now that the implicit mapping F_Θ and volume rendering operator $\mathcal{R}$ are differentiable, the pipeline is trained in a supervised learning fashion:

$$\Theta^*, \mathbf{T}^*, \mathbf{f}^* = \arg\min_{\Theta,\hat{\mathbf{T}},\hat{\mathbf{f}}} L(\hat{\mathbf{I}}, \hat{\mathbf{T}}, \hat{\mathbf{f}} | \mathbf{I}) = \arg\min_{\Theta,\hat{\mathbf{T}},\hat{\mathbf{f}}} \sum_{i=1}^{N} \sum_{\mathbf{p}} \|\hat{\mathcal{I}}_i(\mathbf{p}) - \mathcal{I}_i(\mathbf{p})\|_2^2. \tag{2}$$

3.2 Camera Parameters Formation

Camera Intrinsics. Using a pinhole camera model, the camera intrinsics are represented in a calibration matrix:

$$K = \begin{bmatrix} \hat{f}_x & 0 & p_x \\ 0 & \hat{f}_y & p_y \\ 0 & 0 & 1 \end{bmatrix}, \tag{3}$$

where $\hat{f}_x$ and $\hat{f}_y$ are optimizable focal length along horizontal and vertical axis respectively, p_x and p_y are the known coordinates of the image plane's origin. All intrinsics are shared across all input images.

Camera Extrinsics. Following prior works [25, 26], the camera extrinsics in our work are represented by a rigid transform $T_{world} = [\mathbf{R}|\mathbf{t}] \in \mathrm{SE}(3)$, where $\mathbf{R} \in \mathrm{SO}(3)$ denotes the camera rotation and $\mathbf{t} \in \mathbb{R}^3$ denotes the camera translation. A vector $\mathbf{x} \in \mathbb{R}^3$ after rigid transform would be $\mathbf{x}' = \mathbf{R}\mathbf{x} + \mathbf{t}$.

To make the rotation parameters optimizable, by using Rodrigues' formula, we do axis-angle decomposition:

$$\mathbf{R} \equiv e^{[\mathbf{r}]_\times} = \mathbf{I} + \frac{\sin\theta}{\theta}[\mathbf{r}]_\times + \frac{1-\cos\theta}{\theta^2}([\mathbf{r}]_\times)^2, \tag{4}$$

where $\mathbf{r} \in \mathbb{R}^3$, $\theta = \|\mathbf{r}\|$ denotes the angle of rotation, $\bar{\mathbf{r}} = \mathbf{r}/\|\mathbf{r}\|$ denotes the axis of rotation, and $[\cdot]_\times$ denotes the skew operator that converts a vector into a cross-product matrix. For each input image $\mathcal{I}_i$ we define its extrinsics $\{\hat{\mathbf{r}}_i, \hat{\mathbf{t}}_i\}$, which can be directly optimized during training.

In our work we focus on jointly optimize $\{\hat{f}_x, \hat{f}_y, \hat{\mathbf{r}}_i, \hat{\mathbf{t}}_i\}$ without using any two-stage refinements in NeRFmm [25] or prior knowledge on intrinsics or extrinsics in BARF [8].

3.3 Improve Joint Optimization with SiNeRF

3.3.1 Potential Sub-Optimality of NeRFmm

Although NeRFmm is capable of producing compelling camera parameter estimation and scene reconstruction, it suffers from falling into minima on specific scenes, *e.g.*, textureless *Room* scene and inconsistent *Orchids* in the LLFF dataset. Moreover, in those scenes, there exist obvious gaps between COLMAP estimated intrinsics and estimated intrinsics as well as degeneration on NVS quality compared to ground truth, as reported in [25].

It has been a convention in NeRF-related tasks to use a 256-width ReLU-MLP for radiance mapping, while NeRFmm only adopts a 128-width ReLU-MLP for joint optimizing. Thus, it seems natural to blame the weak expressiveness of a narrow MLP for causing the potential sub-optimality. However, our experiment results show that simply increasing the width of MLP does not always improve NeRFmm's performances, *e.g.*, there exists performance degenerations on *Fortress* and *Orchids* scenes, as shown in the *ref-128* and *ref-256* columns in Table 1 and 2.

To conclude, NeRFmm's joint optimization suffers from **a systematic sub-optimality** that cannot be solved by simply adopting a larger MLP.

In the following sections, we identify that, the absence of a better radiance mapping network and the inefficient ray sampling techniques, would be two of the potential sources for such sub-optimality and they are the very focuses of our work.

3.3.2 SiNeRF Architecture

SiNeRF, our proposed radiance mapping network, consists of a SIREN-MLP [19] head followed by a color branch and a density branch. We denote a L-layer SIREN-MLP head as:

$$\Phi(\mathbf{x}) = \phi_L \circ \phi_{L-1} \circ \cdots \circ \phi_1(\mathbf{x}), \tag{5}$$

where $\phi_l : \mathbb{R}^{d_{l-1}} \mapsto \mathbb{R}^{d_l}$ is the l-th fully-connected SIREN layer

$$\phi_l(\mathbf{x}) = \sin(\alpha_l (\mathbf{W}_l \mathbf{x} + \mathbf{b}_l) + \beta_l), \tag{6}$$

defined by a weight $\mathbf{W}_l \in \mathbb{R}^{d_l \times d_{l-1}}$, a bias $\mathbf{b}_l \in \mathbb{R}^{d_l}$, a frequency scaling factor $\alpha_l \in \mathbb{R}$, and a phase shift factor $\beta_l \in \mathbb{R}^{d_l}$.

We follow the initialization scheme of [19]. We set frequency scaling factors $\alpha_1 = 30$ and $\alpha_l = 1$ for $l \in \{2, 3, \cdots, L\}$. We set all phase shift factors $\beta_l = \mathbf{0}$ for $l \in \{1, \cdots, L\}$. We initialize weight matrices $\mathbf{W}_1 \sim \mathcal{U}(-1/d_1, 1/d_1)$ and $\mathbf{W}_l \sim \mathcal{U}(-\sqrt{6/d_l}, \sqrt{6/d_l})$ for $l \in \{2, 3, \cdots, L\}$. For SiNeRF in experiments, we keep layer number $L = 8$ and hidden unit number $d_l = 256$ for $l \in \{1, 2, \cdots, L\}$. No positional encoding is used for input.

After the SIREN-MLP head, we append two branches for outputting raw color $\mathbf{c} \in \mathbb{R}^3$ and raw density $\sigma \in \mathbb{R}$ respectively. Please see Figure 1 for more details about SiNeRF architecture.

3.3.3 Partial Benefits For the Joint Optimization

We observe that simply replacing ReLU-MLP with SiNeRF for radiance mapping can **only** improve performances on a limited number of scenes. The reason for the **partially** improvements can be that: the input signal is clamped by sinusoidal activation's amplitude and limiting the output value range may bring stability for optimization while also may reduce the signal's expressiveness. Thus, to achieve the general alleviation of sub-optimality in joint optimization, we propose improvements on the ray sampling strategy in the following section.

3.4 Mixed Region Sampling (MRS)

Random Sampling is to randomly choose M rays from random candidate set $\mathcal{P}^{(i)}_{random} = \{\mathbf{p} \,|\, \forall \mathbf{p} \in \mathcal{I}_i\}$. Along each ray, we then apply 3D point sampling to select several spatial points for radiance mapping. Recall that NeRF's reconstruction loss computes the average pixel difference between rendered pixels and corresponding ground truth. A batch of randomly sampled rays would have different colors. **It guarantees the ray batch's diversity and is a critical condition for efficient training.**

To prove that, we set a simple experiment NeRFmm on *Flower* scene and fix all settings unchanged, except that every ray batch is selected from a randomly-placed 32×32 image patch instead of randomly-selected 1024 rays. Such sampling strategy is the most extreme

case where the ray batch has minimum diversity. As for results, the PSNR score halves and pose errors increase by 20 times, compared to baseline.

Random Sampling is straightforward and has been widely adopted by NeRFmm and other NeRF-related works [8, 11, 25]. Yet this strategy still has its limitation. On a textureless scene like *Fortress* in the LLFF dataset, even a batch of randomly sampled pixels may have homogenized colors, which will result in an iteration of "*poor supervision*", i.e., the network is not forced to produce discriminative outputs at different pixel locations.

We argue that, on the one hand, there are fewer constraints for joint optimization tasks. The optimizing direction provided by such poor supervision might potentially make models easier to fall into local minima. On the other hand, inspired by offline SfM methods [17, 22] that leverage key points for efficient inter-image matching, we believe that, compared to Random Sampling with equal weights, treating pixels with different importances and sampling them strategically would help alleviating the sub-optimality in joint optimization.

Thus, we propose a novel sampling strategy designed for joint optimization tasks, named *Mixed Region Sampling* (MRS):

- For the i-th image, we use a SIFT detector to find a set of keypoints $\mathcal{P}_0^{(i)} = \{\bar{\mathbf{p}}_1, \bar{\mathbf{p}}_2, \cdots, \bar{\mathbf{p}}_K\}$. For each keypoint $\bar{\mathbf{p}}_j$ we form a local region set $\mathcal{P}_j^{(i)} = \{\mathbf{p} \,|\, \forall \mathbf{p} \in \overline{\mathcal{N}}_{5\times5}(j)\}$ where $\overline{\mathcal{N}}_{5\times5}(j)$ denotes a set of neighbour points of keypoint $\bar{\mathbf{p}}_j$ within a 5×5 region. The region candidate set is defined as $\mathcal{P}_{region}^{(i)} = \mathcal{P}_0^{(i)} \cup \mathcal{P}_1^{(i)} \cup \cdots \cup \mathcal{P}_K^{(i)}$.
- Then, we define a time-variant weight for region sampling:

$$w(t) = \begin{cases} 1 - t/t_r & 0 \le t \le t_r \\ 0 & t > t_r \end{cases}, \tag{7}$$

that linearly decreases within the range $[0, t_r]$, where t_r represents the end time step of MRS. For the M-ray sampling at time t, $w(t)M$ rays are sampled from region set $\mathcal{P}_{region}^{(i)}$ and $(1 - w(t))M$ rays are sampled from random set $\mathcal{P}_{random}^{(i)}$. After t_r, MRS falls back to random sampling.

Our *Mixed Region Sampling* (MRS) leverages both region candidates for efficient region matching at the early training stage and random candidates for discriminative learning at the late training stage, which is improved upon the *Interst Region Sampling* [26] that only samples candidates from the regions of interest for pose optimization. We show in the ablation study in Section 4.4 that adopting MRS is critical for improving both image quality and pose accuracy, and its combination with SIREN-MLP provides more general alleviation on sub-optimality in joint optimization.

3.5 Testing

Because of the ambiguity between camera translation scale and camera intrinsics [15], the learned poses may not be in the same pose space with COLMAP annotated poses. Thus, *pose alignment* is required for a valid test. Following [25], firstly, we use ATE toolbox [28] to compute a Sim(3) transformation that aligns the COLMAP testing trajectory with the testing trajectory in the SiNeRF pose space. Secondly, we optimize the roughly-aligned testing trajectories by minimizing reconstruction loss while keeping intrinsics and network parameters fixed. This step is to provide precise alignments on trajectories for testing. Lastly, we compute the image quality and pose metrics on test images.

4 Experiments

4.1 Settings

Dataset. We experiment on the LLFF dataset with 8 forward-facing real-world scenes. Every 8-th image in the image sequences is selected for testing. All image resolution is set to 756×1008. Camera annoations are estimated by COLMAP [17].

Training Details. For intrinsics, we initialize $\hat{f}_x$ and $\hat{f}_y$ to be the width and height of the image. For extrinsics, for each image $\mathcal{I}_i$ we initialize the translation $\hat{\mathbf{t}}_i$ to be a zero vector, and rotation matrix $\mathbf{R}_i$ to be an identity matrix, i.e., $\hat{\mathbf{r}}_i$ to be a zero vector. We initialize an 8-layer 256-width SIREN-MLP with methods mentioned in Section 3.3.2. For each iteration, 1024 rays are selected by *Mixed Region Sampling* (MRS), where t_r is set to 500 epochs for *Fortress* and *Trex* scenes and 50 epochs for the rest. Along each ray 128 coordinates are uniformly selected **without** using hierarchical sampling [11].

We train the model for 10k epochs for each scene with three Adam [6] optimizers for intrinsics, extrinsics, and SIREN-MLP, respectively. Extrinsics' and intrinsics' learning rates are initialized to 1e-3 and exponentially decay by 0.9 every 100 epochs. SIREN-MLP's learning rate is initialized to 1e-3 for all scenes except that we lower it to 5e-4 for *Fortress* and 1e-4 for *Orchids* scene, and exponentially decay by 0.9954 every 10 epochs.

4.2 Quantitative Evaluations

We compare the performances between NeRFmm baselines and SiNeRF. Mean pose translation and rotation errors are shown in Table 1. NVS image qualities on three metrics PSNR, SSIM [24] and LPIPS [27] are shown in Table 2.

As shown in the results, adopting a wider network does not necessarily improve the pose accuracy (*e.g.*, *Fern* scene) and image quality (*e.g.*, *Fortress* scene), indicating that joint optimization may exist a systematic sub-optimality. Meanwhile, SiNeRF improves image qualities significantly while achieving pose estimations closed to the ground truth provided by COLMAP.

We mention that the pose errors only indicate how well our pose estimations match the COLMAP estimations. Small pose errors do not guarantee good NVS image qualities.

Our method does not outperform NeRFmm256 baseline on *Leaves* scenes, which indicates that joint optimization is sensitive to scene content. The systematic sub-optimality can only be alleviated instead of completely solved by SiNeRF.

4.3 Qualitative Results

In Figure 2 we display the comparisons on image synthesis between NeRFmm baselines and SiNeRF. Our method is able to reconstruct fine details in the real-world scene with high fidelity.

We also display the comparisons on pose trajectories between SiNeRF estimations and COLMAP estimations. The highly overlapped trajectories show that our method can learn accurate pose estimations closed to classical SfM estimations.

4.4 Ablation Study

To exclude the influence of adopting a wider MLP, we list 256-width ReLU-MLP results in the *NeRFmm256* columns in Table 1 and 2.

Scene	Pose Error					
	Translation($\times 10^{-2}$) ↓			Rotation(°) ↓		
	NeRFmm128	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**
Fern	0.514	0.765	**0.438**	0.957	1.566	**0.743**
Flower	1.039	1.200	**0.796**	3.657	3.211	**0.506**
Fortress	6.463	6.046	**4.068**	2.590	2.410	**1.772**
Horns	1.607	**1.476**	2.153	3.806	3.044	**2.662**
Leaves	0.676	**0.608**	0.831	8.248	**6.782**	8.762
Orchids	1.627	2.243	**1.257**	4.140	5.459	**3.244**
Room	**1.315**	2.148	2.145	3.357	3.745	**2.075**
Trex	1.213	1.467	**0.462**	4.953	6.339	**0.856**
Mean	1.807	1.994	**1.519**	3.964	4.070	**2.578**

Table 1: Quantitative results of pose estimation on LLFF dataset. *NeRFmm128* and *NeRFmm256* denote the NeRFmm baseline with MLP width to be 128 and 256 respectively. Best results are **bolded**.

Scene	Image Quality								
	PSNR ↑			SSIM ↑			LPIPS ↓		
	NeRFmm128	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**	*NeRFmm128*	*NeRFmm256*	**SiNeRF**
Fern	21.811	22.154	**22.482**	0.631	0.648	**0.665**	0.479	0.459	**0.437**
Flower	25.430	26.606	**27.229**	0.714	0.772	**0.798**	0.366	0.296	**0.295**
Fortress	26.173	25.596	**27.465**	0.653	0.602	**0.722**	0.438	0.538	**0.393**
Horns	22.949	23.174	**24.142**	0.626	0.635	**0.684**	0.492	0.506	**0.431**
Leaves	18.647	**19.741**	19.152	0.512	**0.609**	0.571	0.476	**0.385**	0.392
Orchids	16.695	15.858	**16.922**	0.391	0.350	**0.408**	0.540	0.550	**0.529**
Room	25.623	25.675	**26.101**	0.831	0.836	**0.844**	0.450	**0.411**	0.426
Trex	22.551	23.376	**24.939**	0.719	0.759	**0.816**	0.438	0.390	**0.356**
Mean	22.485	22.773	**23.554**	0.635	0.651	**0.689**	0.460	0.442	**0.407**

Table 2: Quantitative results of novel view synthesis on LLFF dataset. *NeRFmm128* and *NeRFmm256* denote the NeRFmm baseline with MLP width to be 128 and 256 respectively. Best results are **bolded**.

Scene	Items	Pose Error		Image Quality		
		Translation($\times 10^{-2}$) ↓	Rotation(°) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Fortress	**SiNeRF**	**4.068**	**1.772**	**27.465**	**0.722**	**0.393**
	w/o SIREN	18.005	155.553	18.593	0.492	0.484
	w/o MRS	6.242	1.797	25.542	0.605	0.532
	w/o SIREN and MRS	6.046	2.410	25.596	0.602	0.538
Trex	**SiNeRF**	**0.462**	**0.856**	**24.939**	**0.816**	**0.356**
	w/o SIREN	1.891	7.958	22.478	0.719	0.430
	w/o MRS	17.755	133.462	14.984	0.452	0.659
	w/o SIREN and MRS	1.467	6.339	23.376	0.759	0.390

Table 3: Quantitative results of ablation study. (1) *w/o SIREN* denotes NeRFmm with 256-width ReLU-MLP and MRS. (2) *w/o MRS* denotes SIREN-MLP with Random Sampling. (3) *w/o SIREN and MRS* denotes NeRFmm with 256-with ReLU-MLP and Random Sampling, which is the baseline. For all MRS in the table we set $t_r = 500$. Best results are **bolded**.

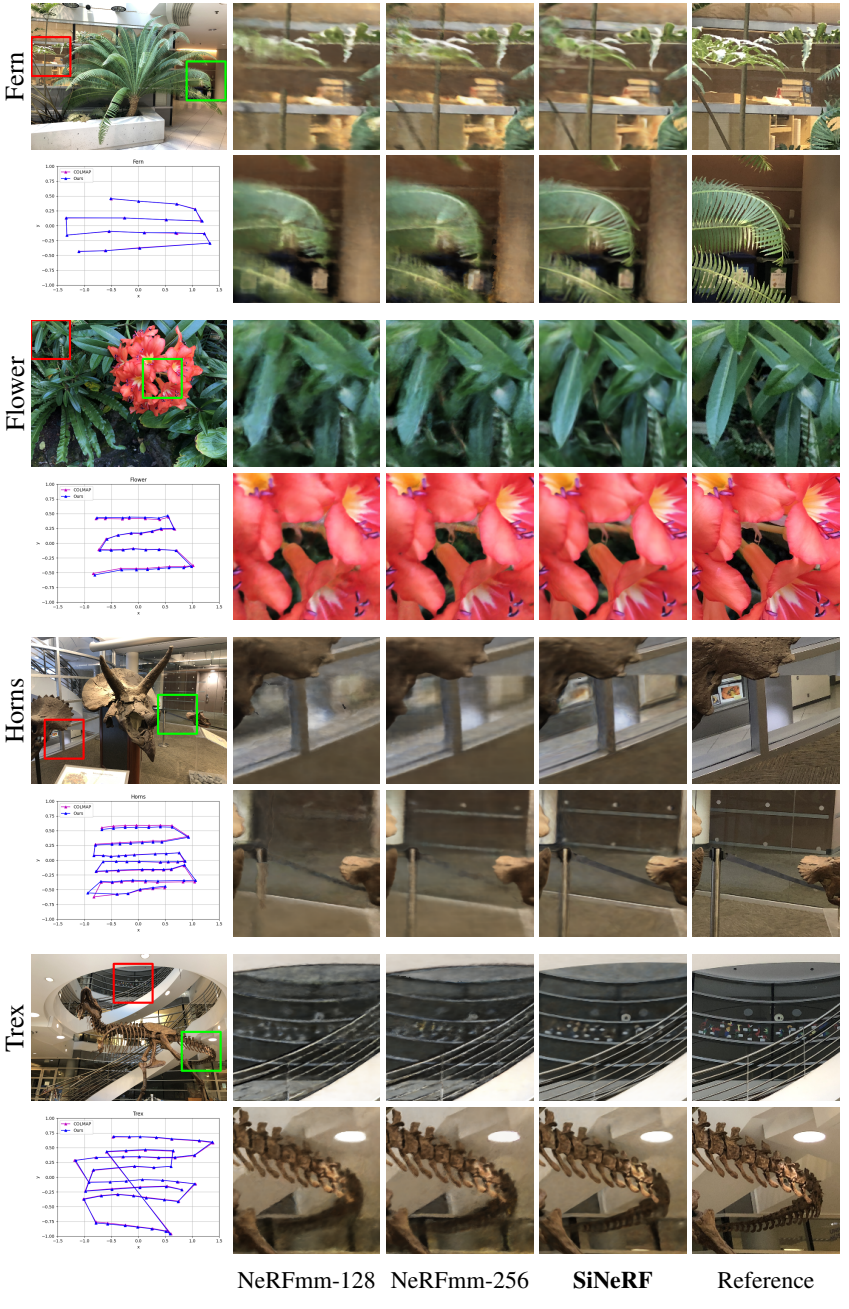

Figure 2: Qualitative results of our method on the LLFF dataset. Comparisons on pose trajectories between SiNeRF and COLMAP are displayed in the bottom-left corner for each scene.

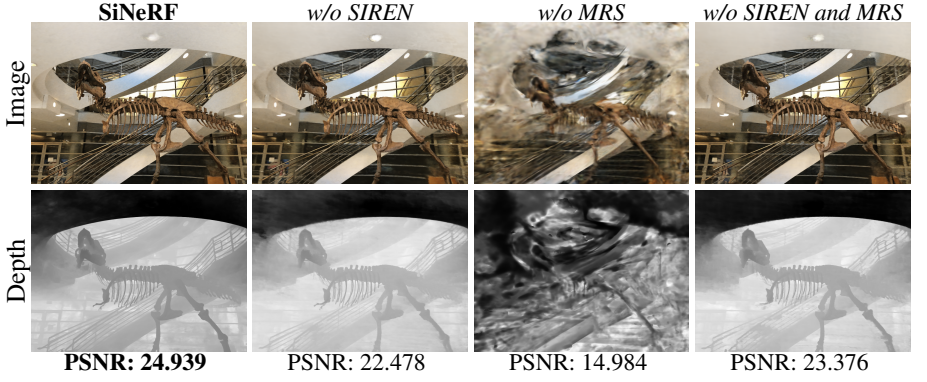

Figure 3: Qualitative results of ablation study on *Trex* scene.

To prove the effectiveness of MRS, we conduct the ablation study on two challenging and representative scenes: *Fortress*, on which NeRFmm is reported to have difficulties converging, and *Trex*, which has lots of fine details and causes relative high pose errors in baselines. Quantitative metrics are shown in Table 3. Qualitative results are shown in Figure 3.

Analysis on results on *Fortress*. As mentioned in Section 3.4, the sampling strategy's diversity is critical for efficient training on textureless scenes like *Fortress*, where Random Sampling may happen to produce a homogenized ray batch that leads to poor supervision. MRS, which selects ray batches within a limited amount of key-point regions, would be even more likely to produce undiversified ray batches. Using SIREN-MLP can smoothen the optimization surface and help escape from early local minima. But SIREN-MLP doesn't always guarantee comprehensive improvements. This explains performances of "SiNeRF" >"*w/o SIREN and MRS*" or "*w/o MRS*" >"*w/o SIREN*" in Table 3.

Analysis on results on *Trex*. SIREN-MLP favors scenes with large consistent patterns like *Flowers* and *Fortress*. Yet on fine-detailed scenes like *Trex*, SIREN-MLP will struggle to distinguish close-by pixels with various colors and produces blurred consistent patterns, as shown in Figure 3. Besides, MRS provides the best performances in combination with SIREN-MLP. This explains performances of "SiNeRF" >"*w/o SIREN and MRS*" >"*w/o MRS*" or "*w/o SIREN*" in Table 3.

5 Conclusion

In this work, we identify the potential sources of the systematic sub-optimality of joint optimization. We propose *SiNeRF* architecture and *Mixed Region Sampling* for alleviating such sub-optimality. Experiments and ablation studies show comprehensive improvements in both image synthesis quality and pose accuracy compared to NeRFmm baselines and prove the effectiveness of our designs.

Acknowledgements

This work was partly supported by ETH General Fund, the Alexander von Humboldt Foundation and a Huawei research project.

References

[1] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021.

[2] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019.

[3] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021.

[4] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021.

[5] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[7] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021.

[8] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021.

[9] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019.

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

[11] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020.

[12] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

[13] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021.

[14] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[15] Marc Pollefeys and Luc Van Gool. A stratified approach to metric self-calibration. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition*, pages 407–412. IEEE, 1997.

[16] Gernot Riegler and Vladlen Koltun. Stable view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12216–12225, 2021.

[17] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[18] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020.

[19] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020.

[20] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019.

[21] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021.

[22] Bill Triggs, Philip F McLauchlan, Richard I Hartley, and Andrew W Fitzgibbon. Bundle adjustment—a modern synthesis. In *International workshop on vision algorithms*, pages 298–372. Springer, 1999.

[23] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020.

[24] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[25] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021.

[26] Lin Yen-Chen, Pete Florence, Jonathan T Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1323–1330. IEEE, 2021.

[27] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[28] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018.