Adaptive Positional Encoding for Bundle-Adjusting Neural Radiance Fields

Zelin Gao[1], Weichen Dai[3], and Yu Zhang *[1,2]

[1]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University
[2]Key Laboratory of Collaborative Sensing and Autonomous Unmanned Systems of Zhejiang Province
[3]Key Laboratory of Brain Machine Collaborative Intelligence of Zhejiang Province and School of Computer Science, Hangzhou Dianzi University

Abstract

Neural Radiance Fields have shown great potential to synthesize novel views with only a few discrete image observations of the world. However, the requirement of accurate camera parameters to learn scene representations limits its further application. In this paper, we present adaptive positional encoding (APE) for bundle-adjusting neural radiance fields to reconstruct the neural radiance fields from unknown camera poses (or even intrinsics). Inspired by Fourier series regression, we investigate its relationship with the positional encoding method and therefore propose APE where all frequency bands are trainable. Furthermore, we introduce period-activated multilayer perceptrons (PMLPs) to construct the implicit network for the high-order scene representations and fine-grained gradients during backpropagation. Experimental results on public datasets demonstrate that the proposed method with APE and PMLPs can outperform the state-of-the-art methods in accurate camera poses and high-fidelity view synthesis.

1. Introduction

Simultaneously localizing from the given camera frames and reconstructing the scene from multi-view 2D images is one of the crucial tasks in computer vision, which can carry out self-localization as well as sense surroundings through visual information as human beings do. Most classic methods, such as Structure from Motion (SfM) [35, 1] and Simultaneously Localization and Mapping (SLAM) [30, 10, 11], leverage the similarity of texture patterns to find the correspondences. Then, all states are estimated, including camera parameters (intrinsic and extrinsic parameters) and a map in conventional representations (*e.g.*, sparse 3D point cloud [17]). However, the quality of the created map representations depends on the correspondences, and the sparse nature of common 3D point clouds limits downstream vision tasks that require dense geometric reasoning.

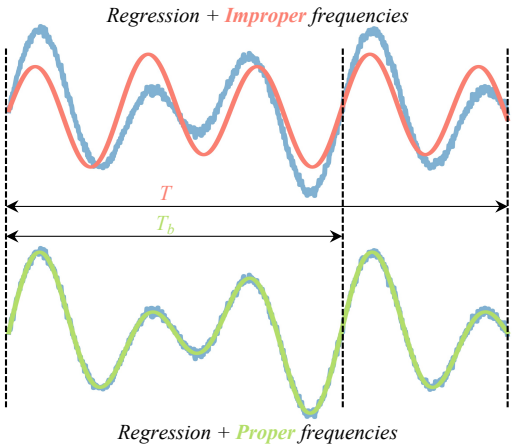


Figure 1. **Fourier Series Regression.** Fitting 1D signal (blue curve, duration T, period T_b) requires a Fourier series with proper frequencies (bottom, green curve). Conversely, Fourier series regression with improper frequencies (top, red curve) cannot fit well.

Different from explicitly reconstructing scenes from images, Neural Radiance Fields (NeRF) [28] present a continuous function to parameterize the scene through an implicit neural network [38, 25] mapping 3D point positions and direction to color and volume density. By volume rendering theorem [43, 19], NeRF can realize high-fidelity dense representation results with a few observed images of the scene. Although many follow-up works [2, 15, 26, 18, 32, 47, 49] have improved the performance of NeRF in various aspects, most require pre-provided accurate camera parameters, typically obtained through off-the-shelf SfM tools, limiting this technique to realize simultaneous localization and reconstruction in more practice scenarios.

There are some methods [6, 22, 40] jointly optimizing implicit network parameters, camera poses (extrinsic parameters), and even camera intrinsic parameters to address the solution of training radiance field representations with-

*Corresponding author

out accurate camera parameters. They additionally optimize the camera parameters with frequency encoding via back-propagation of the NeRF model [45]. However, the second-order derivatives of ReLU-activated MLPs, which make up classic NeRF, are zero and thus incapable of modeling the fine-grained information contained in higher-order derivatives of radiance fields [37]. Therefore, it is difficult for ReLU-activated MLPs to derive more fine-grained gradient information to update camera parameters.

It should be noted that positional encoding (PE) [41] plays a critical role in neural radiance field methods. PE can embed the input (spatial position and direction in NeRF) in a high-dimensional frequency space, similar to the Fourier transform. However, since the information of different scenes is distributed in different frequencies, different scenes require different Fourier series to better fit. Therefore, PE in most methods with fixed handcraft frequency parameters limits the performance of joint optimization, and the frequency parameters of PE should be adaptive to the scene and also optimized during simultaneous localization and radiance field reconstruction.

In this paper, we address the problem of training neural radiance fields from unknown camera parameters (intrinsic and extrinsic) — the joint problem of reconstructing the 3D scene, registering the camera poses (extrinsics), and updating the camera intrinsics. We propose adaptive positional encoding (APE) for bundle-adjusting neural radiance fields that can simultaneously optimize implicit network parameters and camera parameters. Inspired by the Fourier series, which shows that regression with proper frequency parameters can produce better performance in fitting the reference curve, as shown in Fig. 1, we propose APE that can better fit the distribution of radiance information in the field. After encoding the input, period-activated multilayer perceptrons (PMLPs) are introduced to implicitly reconstruct the scene since PMLPs can represent complex higher-order information hidden in nature signals [31] and can yield more effective and fine-grained gradients for updating camera and APE parameters. In the training step, in addition to the photometric loss, a frequency diversity loss function is designed for APE, preventing the adaptive frequency parameters from converging to a single band and thereby losing fine representation ability. In experiments, the proposed method is compared to state-of-the-art methods to prove that the proposed method can learn the higher-order representation of neural radiance fields and estimate camera parameters. Finally, a comprehensive ablation study is carried out to show the effectiveness of APE and PMLPs.

The main contributions of this work are as follows:

- We propose adaptive positional encoding for bundle-adjusting neural radiance fields that can jointly reconstruct the 3D scene, register the camera poses, and update the camera intrinsics.
- We propose adaptive positional encoding that can adaptively fit radiance fields with proper frequency parameters to realize a high-synthesis quality.
- We reconstruct the radiance fields using adaptive positional encoding and period-activated multilayer perceptrons, that can represent complex scenes and provide more fine-grained gradients for updating APE and camera parameters.

2. Related Work

SfM and SLAM. The main goal of SfM [52, 7] and SLAM [48, 51] systems is to recover the 3D structure of the scene and estimate the camera poses from the given image sequences. Most of these systems can be classified as feature-based methods or direct methods [3].

Feature-based methods associate images with visual features [33, 24] to construct the cost function based on multi-view geometry theory and then calculate the loss of projection to optimize the pose of the camera. Many feature-based methods [5, 8] achieve great success under various conditions, though they often suffer from texture-less, where features are hard to detect and establish the correct data association. To overcome these effects, some studies [23, 34, 9] propose to employ convolutional neural networks to associate data between two-view images.

Direct methods define an objective function on the assumption of photometric consistency [12]. Minimizing this photometric consistency function contributed by each pixel to estimate camera poses and the 3D structure, direct methods [36, 10] can utilize raw image information. Taking this photometry consistency function as the reconstruction loss, some direct methods [53, 42, 40] can thus integrate the neural network into the system and realize dense real-time reconstruction. Similar to direct methods, the proposed method also supervises all parameters with the photometric loss. It is worth noting that some methods [53, 40] can represent the scene as a continuous function via a neural network, while explicit 3D representations of traditional methods are typically discrete and sparse.

Neural Radiance Fields. Novel view synthesis (NVS) [16, 20] is a long-standing issue in computer graphics. Although many works [14, 4] indicate that the pixel colors of images can be synthesized through geometric information or interpolation, there are still numerous restrictions involved since NVS is inherently an elusive problem [46]. Recently, neural radiance field (NeRF) methods [26, 2, 29, 28] are presented to learn scene radiance fields using the neural network, showing their impressive abilities to generate realistic images. However, these works require accurate camera parameters obtained by SFM or SLAM techniques (*e.g.*, COLMAP[35]).

By removing the requirement of the prior camera param-

eters, NeRF-- [45] and BARF [22] propose a joint optimization issue of camera parameters and the implicit network. However, the positional encoding in most methods utilizes fixed handcraft frequency parameters so that the performance of the joint optimization is degraded. To solve this problem, the proposed method proposes APE can dynamically fit the proper frequency bands of the scene for better simultaneously estimating accurate camera parameters and radiance field reconstruction.

3. Method

The overview of our method is illustrated in Fig. 2. We represent the neural radiance fields using adaptive positional encoding (Sec. 3.1) and an implicit network composed of period-activated multilayer perceptrons (Sec. 3.2). By sampling points along the ray generated by the estimated camera parameters (Sec. 3.3) and putting them into the neural radiance fields, we can get predicted colors through volume rendering theorem (Sec. 3.4). To solve this joint optimization problem that involves the APE parameters $\mathbf{\Omega}$, implicit network parameters $\mathbf{\Phi}$, and camera parameters $\mathbf{\Theta}$, we supervise these parameters with observed colors $\mathbf{C}$ by both types of the loss of photometric and frequency diversity (Sec. 3.5).

$$\mathbf{\Omega}^*, \mathbf{\Phi}^*, \mathbf{\Theta}^* = \underset{\mathbf{\Omega},\mathbf{\Phi},\mathbf{\Theta}}{\arg\min}\, \mathcal{L}(\mathbf{\Omega}, \mathbf{\Phi}, \mathbf{\Theta} \mid \mathbf{C}) \tag{1}$$

3.1. Adaptive Positional Encoding

We propose adaptive positional encoding (APE) to learn the proper frequency bands corresponding to the scene representation during the joint optimization.

For a clearer description of the frequency role in the positional encoding method, we first discuss the Fourier series regression of a 1D signal $f(t)$ lasting T seconds, which can be defined as:

$$f(t) = [m_1, n_1, \cdots, m_L, n_L] \begin{bmatrix} \sin(\omega t) \\ \cos(\omega t) \\ \vdots \\ \sin\left(2^{L-1}\omega t\right) \\ \cos\left(2^{L-1}\omega t\right) \end{bmatrix} + k \tag{2}$$

where, L denotes the number of harmonics (sin and cos), $[m_1, n_1, \cdots, m_L, n_L]$ denotes the amplitude coefficients of Fourier series, and k denotes the bias. Note that the Fourier series regression usually considers the duration T as the period and calculates $\omega = 2\pi f_0$ with the frequency $f_0 = \frac{1}{T}$. As shown in Fig. 1, if the duration T does not match the true base period T_b, it leads to improper frequencies and degrades the regression performance. To overcome this problem, additional Fourier series are required, but this may also cause over-fitting.

The positional encoding method used in NeRF has a similar formulation to the Fourier series, which can also refer to a kind of Fourier encoding [13]. Supposing that activation functions are linear, the transformation of the network can be represented using weights $\mathbf{W}$ and biases $\mathbf{b}$, respectively. Based on this assumption, the positional encoding and neural network of NeRF can be simplified by

$$NeRF(\mathbf{x}) = \mathbf{W} \begin{bmatrix} \sin(\pi\mathbf{x}) \\ \cos(\pi\mathbf{x}) \\ \vdots \\ \sin\left(2^{L-1}\pi\mathbf{x}\right) \\ \cos\left(2^{L-1}\pi\mathbf{x}\right) \end{bmatrix} + \mathbf{b} \tag{3}$$

By comparing Eq. (2) and Eq. (3), it can be observed that the two calculations are similar. The difference between the Fourier series regression and NeRF is the input dimension, resulting in different dimensions of the amplitude coefficients and biases. Therefore, the detailed variation of the signal can be represented by the Fourier series at high frequencies, indicating that additional high-frequency features of positional encoding can contribute to fine representations of NeRF. Meanwhile, the positional encoding also requires proper frequencies to fit the scene radiance distribution, similar to Fourier series regression in Fig. 1.

To address this issue, the proposed APE consists of trainable frequency parameters $(pe_1, pe_2, \cdots, pe_L)^\top$, which can be adjusted during training and converge to the proper frequency bands of the scene representations. Assuming that the input is a three-dimensional vector $\mathbf{x} = (x, y, z)^\top$, the process of applying APE to $\mathbf{x}$ can be defined as

$$\boldsymbol{\gamma}(\mathbf{x}) = [\boldsymbol{\gamma}_0(\mathbf{x}), \boldsymbol{\gamma}_1(\mathbf{x}), \cdots, \boldsymbol{\gamma}_{L-1}(\mathbf{x})]^\top \tag{4}$$

and the k-th APE $\boldsymbol{\gamma}_k(\mathbf{x})$ is

$$\boldsymbol{\gamma}_k(\mathbf{x}) = [\sin(pe_k\mathbf{x}), \cos(pe_k\mathbf{x})]^\top \tag{5}$$

APE can project the input into a proper frequency space where the appropriate frequency band combination can be searched for the current scene.

Moreover, the Jacobian of the k-th APE $\boldsymbol{\gamma}_k(\mathbf{x})$ is:

$$\frac{\partial \boldsymbol{\gamma}_k(\mathbf{x})}{\partial \mathbf{x}} = pe_k \cdot [\sin(pe_k\mathbf{x}), \cos(pe_k\mathbf{x})]^\top \tag{6}$$

APE can adjust the gradients of frequency parameters to predict effective updates for optimizing camera parameters.

3.2. Period-Activated Multilayer Perceptrons

Motivated by SIREN [37], we construct the implicit neural network with the period-activated multilayer perceptron (PMLP) for its capability of modeling information contained in higher-order derivatives of natural signals and retrieve effective gradients for updating all parameters, which

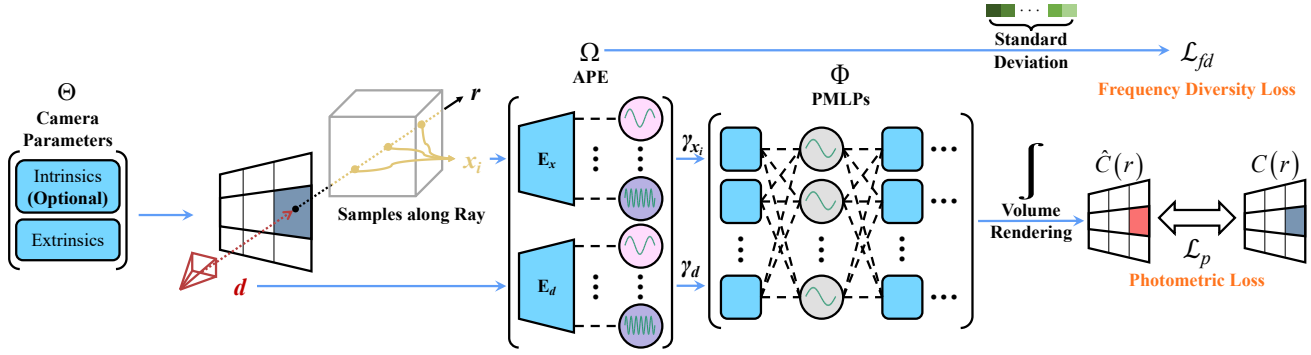


Figure 2. **Overview of Our Method.** The estimated states include APE parameters $\mathbf{\Omega}$, implicit network parameters $\mathbf{\Phi}$, and camera parameters $\mathbf{\Theta}$. The 3D location $\mathbf{x}_i$ is uniformly sampled along the ray $\mathbf{r}$ generated by the estimated camera parameters. Then, $\mathbf{x}_i$ and the 2D viewing direction $\mathbf{d}$ are converted to $\boldsymbol{\gamma}_{\mathbf{x}_i}$ and $\boldsymbol{\gamma}_{\mathbf{d}}$ by APE ($E_{\mathbf{x}}$ for location and $E_{\mathbf{d}}$ for direction) and then fed into the implicit network composed of PMLPs, respectively. The predicted outputs are integrated into color values $\hat{\mathbf{C}}(\mathbf{r})$ via volume rendering theorem. Finally, we supervise all parameters with the photometric loss $\mathcal{L}_p$ between observed colors $\mathbf{C}(\mathbf{r})$ and predicted colors $\hat{\mathbf{C}}(\mathbf{r})$. Additionally, we add the frequency diversity loss $\mathcal{L}_{fd}$ to prevent Ω from degrading all to a single frequency band.

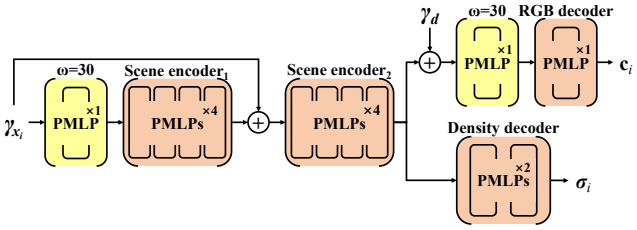


Figure 3. **Implicit Network Architecture.** We set the frequency coefficient $\omega = 30$ for the PMLPs marked by yellow and $\omega = 1$ for the others. Our implicit network takes $\boldsymbol{\gamma}_{\mathbf{x}_i}$ and $\boldsymbol{\gamma}_{\mathbf{d}}$ as input and predicts the color $\mathbf{c}_i$ and density σ_i.

can be formulated as follows:

$$\mathbf{x}_i \mapsto \phi_i\left(\mathbf{x}_i\right) = \sin\left(\omega_i \cdot \mathbf{W}_i \mathbf{x}_i + \mathbf{b}_i\right) \tag{7}$$

where, $\phi_i : \mathbb{R}^{M_i} \mapsto \mathbb{R}^{N_i}$ denotes i^{th} PMLP layer. It applies the sine function to each component of the output vector after using the affine transform defined by the weight matrix $\mathbf{W}_i \in \mathbb{R}^{N_i \times M_i}$, the frequency coefficient $\omega_i \in \mathbb{R}$, and the biases $\mathbf{b}_i \in \mathbb{R}^{N_i}$ applied to the input $\mathbf{x}_i \in \mathbb{R}^{M_i}$.

As shown in Fig. 3, the proposed implicit network takes PMLPs as the basic unit to carry out the radiance field representation. Since the derivative of the sine is a phase-shifted sine, it is allowed for the derivative of a PMLP to inherit the properties of PMLP [37]. Moreover, the implicit representation fitted by PMLPs is a higher-order complex function that can contain richer information due to their derivative properties, whereas ReLU-activated MLPs are piecewise linear and can only learn incomplete features hidden in the first-order derivatives. Therefore, PMLP can yield more effective and fine-grained gradients for updating camera and APE parameters. We also insert the input as complementary features into the deeper layer with the skip connection, which is beneficial for further acceleration of convergence [21].

SIREN demonstrates that setting the first layer $\omega_0 = 30$ and the other $\omega_i = 1$ helps to perform implicit representations. Here we give the explanation that $\omega_0 = 30$ PMLP can first map the input into a high-dimensional space containing rich features since higher-frequency signals are more capable of distinguishing features. Then, the following $\omega_i = 1$ PMLPs can gradually extract detailed information from these high-frequency features.

3.3. Camera Parameters

Setting the pinhole camera as the default model, it is essential to confirm the variables of camera intrinsics and extrinsics to be optimized along with their specific formats.

Camera Intrinsics. We consider the image center as the camera principle point, i.e., $cx \approx \frac{W}{2}$ and $cy \approx \frac{H}{2}$, where H and W denote the height and width of the image. Therefore, the focal f is the only variable in the intrinsics of the camera to be estimated.

Camera Extrinsics. The camera extrinsics contain rotation $\mathbf{R} \in SO(3)$ and translation $\mathbf{t} \in \mathbb{R}^3$. Since it is challenging to directly optimize a rotation matrix $\mathbf{R}$ with 9 elements, we choose to optimize translation $\mathbf{t}$ as well as rotation matrix $\mathbf{R}$ of the axial-angle format represented by a 3D vector $\boldsymbol{\psi}$ which can be recovered to a rotation matrix $\mathbf{R}$ through Rodrigues' rotation formula.

Therefore, there are three types of variables requiring to be optimized in the camera parameters $\mathbf{\Theta}$ including f, $\mathbf{t}$, and $\boldsymbol{\psi}$.

3.4. Volume Render

In this work, we follow the volume rendering theorem in NeRF to integrate the predicted color and density in

Sec. 3.2. By confirming camera parameters in Sec. 3.3, we can generate the ray $\mathbf{r}$ of a pixel. Without prior camera parameters, the volume distribution varies with each update of the camera parameters. Therefore, we uniformly sample N points along the ray r and feed these samples into the neural radiance fields.

Given the predicted color c_i and its density σ_i of samples along the ray r, the color $\hat{\mathbf{C}}(\mathbf{r})$ can be approximated with the volume rendering theorem as

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{c}_i \tag{7}$$

where $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma_j \delta_j\right)$ denotes how much light is transmitted on ray $\mathbf{r}$ up to sample i and $\left(1 - \exp\left(-\sigma_i \delta_i\right)\right)$ denotes how much light is contributed by sample i.

3.5. Reconstruction Loss

To solve the joint optimization problem involving the parameters mentioned in Sec. 3.1, Sec. 3.2, and Sec. 3.3, we uniformly sample M rays from the image sets. Then, we jointly optimize these parameters with photometric loss and frequency diversity loss.

Photometric Loss. The photometric loss $\mathcal{L}_p$ is L_2 loss between the approximated color $\hat{\mathbf{C}}(\mathbf{r})$ and observed color $\mathbf{C}(\mathbf{r})$ of sampled rays,

$$\mathcal{L}_p = \sum_{\mathbf{r} \in \mathcal{R}_M} \left\| \mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}\left(\mathbf{r}\right) \right\|_2^2 \tag{8}$$

where $\mathcal{R}_M$ denotes the set of M sampled rays.

Frequency Diversity Loss. APE allows frequency bands to be learned with the scene representation in parallel. However, the neural network always tends to recover the low-frequency information at first [22], which may lead to a trend of degradation toward a single frequency band. Therefore, we propose the loss of frequency diversity to maintain the standard deviation of APE at diversity,

$$\mathcal{L}_{fd} = 1 - \frac{1}{1 + \beta \cdot e^{-\Sigma_{fd}}} \tag{9}$$

where β denotes the coefficient to shift the slope of the loss function and Σ_{fd} denotes the standard deviation of APE, which can be defined as

$$\Sigma_{fd} = \sqrt{\frac{1}{n}\sum_{k=1}^{n}(pe_k - \frac{1}{n}\sum_{k=1}^{n} pe_k)^2} \tag{10}$$

Since Σ_{fd} characterizes the frequency dispersion, we maintain it at a high level by applying frequency diversity loss, thus ensuring that the high-frequency weights are not lost during training. Though the simpler function (e.g. minimizing $-\Sigma_{fd}$ to $-\infty$) has the similar effect on maintaining frequency diversity, converging to $-\infty$ makes $\mathcal{L}_{fd}$ consistently increase frequency dispersion throughout training, leading to unstable frequency diversity. However, Eq 9 can drop rapidly to a relatively low value during training as compared to photometric loss and has a less impact on frequency dispersion in the later stage of training, thus achieving stable frequency diversity.

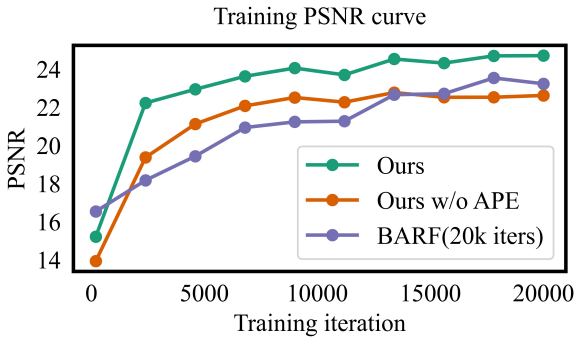


Figure 4. **Comparison of Optimization Efficiency.** We visualize the training PSNR curves including our method (green), our method without the positional encoding (orange), and BARF with 20k iterations (purple) in **Fern** scene.

Finally, we optimize APE $\mathbf{\Omega}$, implicit network $\mathbf{\Phi}$ of PMLPs, and camera parameters $\mathbf{\Theta}$ using these two reconstruction losses, which can be formulated as the following minimization problem,

$$\min_{\mathbf{\Omega},\mathbf{\Phi},\mathbf{\Theta}} (\mathcal{L}_p + \lambda_{fd}(\mathcal{L}_{fd}^x + \mathcal{L}_{fd}^d)) \tag{11}$$

where $\mathcal{L}_{fd}^x$ and $\mathcal{L}_{fd}^d$ denote the frequency diversity loss of $E_{\mathbf{x}}$ and $E_{\mathbf{d}}$. λ_{fd} denotes the weighting factor between photographic loss and frequency diversity loss.

4. Experiments

We mainly validate the proposed method using the challenging real-world LLFF dataset [27], which consists of 8 forward-facing scenes with sequentially captured images by hand-held cameras. We also use nerf_real_360 dataset [28] to evalute the performance of our method on a more complex scene with a wider range of camera viewpoints. Experiments are conducted under different prior situations to illustrate that the proposed method can solve the joint optimization issue of training neural radiance fields from unknown camera parameters. Moreover, we investigate the effectiveness of each part of our method through the ablation study.

4.1. Experiment Setup

Evaluation Metrics. We evaluate the performance in two aspects: camera parameter errors and view synthesis qual-

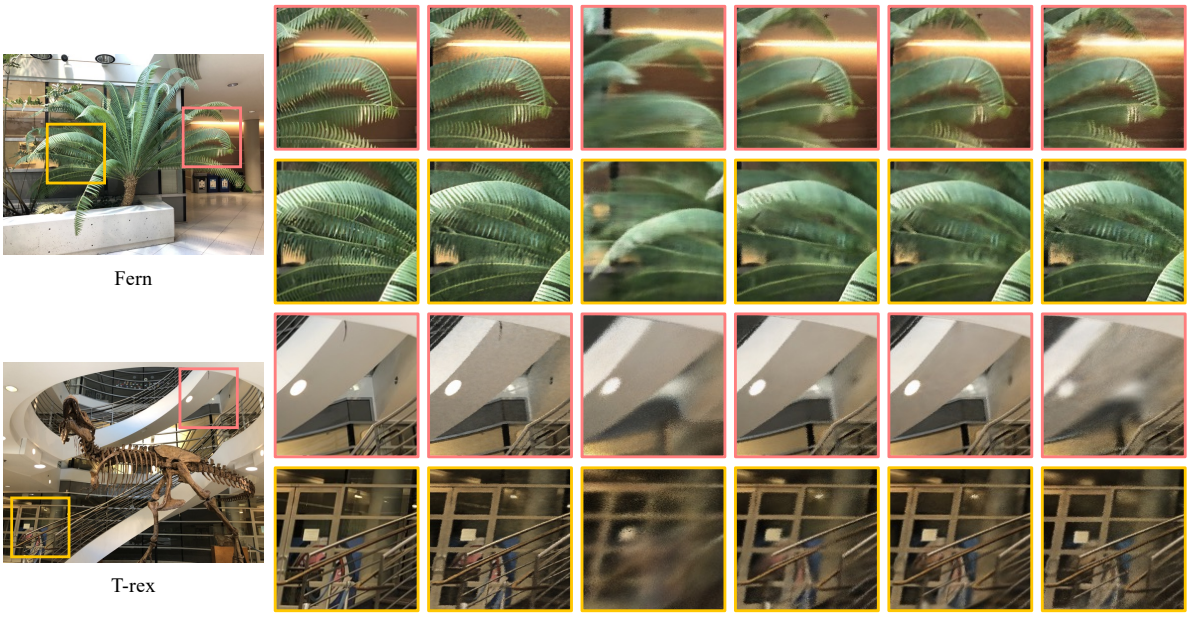


Figure 5. **Qualitative Comparison of Evaluation w/o Camera Extrinsics.** We compare the fine details of our method (20k iterations, unknown camera poses) with other methods including BARF (20k and 200k iterations, unknown camera poses), GARF (200k iterations, unknown camera poses) and ref.NeRF (200k iterations, known camera poses). It should be noted that the qualitative results of GARF are taken directly from their paper.

Scene	Camera parameter estimation						View synthesis quality											
	Δrot(°) ↓			Δtrans(m)(10^{-2}) ↓			PSNR ↑				SSIM ↑				LPIPS ↓			
	Ours (20k iters)	GARF (200k iters)	BARF (200k iters)	Ours (20k iters)	GARF (200k iters)	BARF (200k iters)	Ours (20k iters)	GARF (200k iters)	BARF (200k iters)	ref.NeRF* (200k iters)	Ours (20k iters)	GARF (200k iters)	BARF (200k iters)	ref.NeRF* (200k iters)	Ours (20k iters)	GARF (200k iters)	BARF (200k iters)	ref.NeRF* (200k iters)
Fern	**0.05**	0.47	0.19	**0.17**	0.25	0.19	**24.63**	24.51	23.79	23.72	**0.75**	0.74	0.71	0.73	**0.27**	0.29	0.31	0.26
Flower	**0.21**	0.46	0.25	**0.19**	0.22	0.22	26.32	**26.40**	23.37	23.24	0.73	**0.79**	0.70	0.67	0.15	**0.11**	0.21	0.24
Fortress	0.17	**0.03**	0.48	**0.21**	0.27	0.36	**29.16**	29.09	29.08	25.97	**0.84**	0.82	0.82	0.78	**0.13**	0.15	0.13	0.19
Horns	0.08	**0.03**	0.30	**0.09**	0.21	0.22	**24.27**	22.54	22.78	20.35	**0.76**	0.69	0.73	0.62	**0.29**	0.33	0.30	0.42
Leaves	0.24	**0.13**	1.27	0.28	**0.23**	0.25	19.42	**19.72**	18.78	15.33	0.59	**0.61**	0.54	0.31	0.28	**0.27**	0.35	0.53
Orchids	**0.26**	0.43	0.63	**0.32**	0.41	0.40	**20.02**	19.37	19.45	17.34	**0.60**	0.57	0.57	0.52	**0.25**	0.26	0.29	0.31
Room	**0.12**	0.42	0.32	**0.10**	0.32	0.27	**32.09**	31.90	31.95	32.42	**0.95**	0.94	0.94	0.95	**0.10**	0.13	0.10	0.08
T-rex	**0.12**	0.66	1.14	**0.18**	0.48	0.72	**23.88**	22.86	22.55	22.12	**0.82**	0.80	0.77	0.74	**0.18**	0.19	0.21	0.24
Mean	**0.16**	0.33	0.57	**0.19**	0.30	0.33	**24.95**	24.54	23.97	22.56	**0.76**	0.75	0.72	0.67	**0.21**	0.22	0.24	0.28

Table 1. **Quantitative Comparison of Evaluation w/o Camera Extrinsics.** The best is in bold. 10^{-2} denotes Δtrans are scaled by 100. $*$ denotes that ref.NeRF is trained with known camera poses. Conducted with unknown camera poses and fewer iterations, the proposed method can achieve better view synthesis quality and camera parameter estimation in most scenes even compared to ref.NeRF.

Scene	Camera parameter estimation				View synthesis quality		
	Δrot(°) ↓		Δtrans(m)(10^{-2}) ↓		PSNR ↑		
	Ours (20k iters)	BARF (200k iters)	Ours (20k iters)	BARF (200k iters)	Ours (20k iters)	BARF (200k iters)	ref.NeRF* (200k iters)
vasedeck	**0.64**	87.23	**0.23**	152.41	**20.85**	14.74	18.81
pincone	**0.25**	114.47	**0.74**	161.39	**22.56**	13.17	19.76
Mean	**0.45**	100.85	**0.49**	156.90	**21.71**	13.96	19.29

Table 2. **Quantitative Comparison over nerf_real_360 Dataset.** The best is in bold. 10^{-2} denotes Δtrans are scaled by 100. Conducted with unkown camera poses with a wider range of viewpoints, the proposed method can reconstruct radiance fields and estimate camera poses but BARF fails.

ity. For evaluating camera parameter errors, we take camera poses and intrinsic parameters estimated by COLMAP as the ground truth (provided by dataset). We measure the absolute error in the metric of pixels on focal length (Δfocal). We compute the rotation angle error (Δrot) and the absolute translation error (Δtrans) by aligning the optimized camera poses with the ground truth with a similarity transformation $Sim(3)$ [39]. To evaluate the quality of the view synthesis, we perform an additional pose-only optimization step on the trained model in case misaligned poses may contaminate the novel view synthesis results [45, 22]. PSNR, SSIM [44] and LPIPS [50] are used to measure view synthesis quality.

Implement Details. We set the dimensions of $E_{\mathbf{x}}$ and $E_{\mathbf{d}}$ to 10 and 4, respectively. All axis angles and translations are initialized with zero vectors. We use the Adam optimizer with a learning rate of 1×10^{-3} exponentially decaying to 1×10^{-4} for $E_{\mathbf{x}}$ as well as $E_{\mathbf{d}}$, a learning rate of 1×10^{-3} exponentially decaying to 1×10^{-5} for the implicit network, and a learning rate of 1×10^{-3} exponentially decaying to

1×10^{-6} for all camera parameters. We uniformly sample $N = 128$ points for numerical integration along each ray. We set $\beta = 0.01$ and the loss weighting factor $\lambda_{fd} = 0.005$.

For fair evaluation without camera extrinsics, we resize the images to 480×640 pixels and assume known camera intrinsics. We train the framework for 20k iterations and uniformly sample $M = 2048$ rays from the whole scene in each iteration. We follow the same train/test splits as BARF, i.e., the last 10% images are used as test images. All these configurations except the training iterations are consistent with BARF and GARF. Note that there is no official source code for GARF, so only part of its results are available.

For fair evaluation without camera extrinsics and intrinsics, we resize the images to 756×1008 pixels. We initial the focal length f with image width (provided by dataset) and add it to the optimization as well. The hidden units of all PMLPs in the implicit network are cut in half to check the limits of our method under challenging conditions. We train the framework for 10k epochs and uniformly sample $M = 1024$ rays from every image during one epoch. We follow the same train/test splits as NeRF--, i.e., every 8-th image is used as a test image.

For the ablation study on APE and β, we use ref.NeRF as the baseline method. For the ablation study on APE and PMLPs, we follow the same implementation details as evaluation without camera extrinsics and intrinsics to investigate the performance under full-size and half-size network.

4.2. Evaluation w/o Camera Extrinsics

We investigate the joint optimization problem involving camera extrinsics (poses) and neural radiance field parameters to evaluate the proposed method, similar to the bundle adjustment. We mainly compare the performance of the proposed method with the current state-of-the-art method, BARF [22] and GARF [6]. Reference NeRF (ref.NeRF) [22] is also introduced to show the limits of view synthetic quality by training the neural radiance field with the given camera poses. Furthermore, we remove the APE (Ours w/o APE) and directly train the implicit network to represent the neural radiance fields to demonstrate the optimization efficiency of PMLP.

Fig. 4 shows the optimization efficiency in **Fern** scene. Training these frameworks for 20k iterations, it can be observed that the implicit network consisting of PMLPs proposed in our method converges faster than BARF consisting of ReLU-activated MLPs. However, it is difficult for the implicit network to learn fine details with the lack of APE, thus leading to a lower PSNR than BARF. By adding APE to the framework, the proposed method can better represent the fine details of the scene and achieve a higher PSNR during the training accelerated by the implicit network.

Fig. 5 presents the qualitative results of the synthetic images. Both trained for 20k iterations, the proposed method can reach a better neural radiance field representation than the results of BARF. Even compared with baselines trained for 200k iterations, the proposed methods still preserve more fine details in the neural radiance field representation. Importantly, the adjusted high-frequency terms of APE can realize more fine detail representations than ref.NeRF with fixed high-frequency items. These results prove that the proposed method can perform better with fewer iterations.

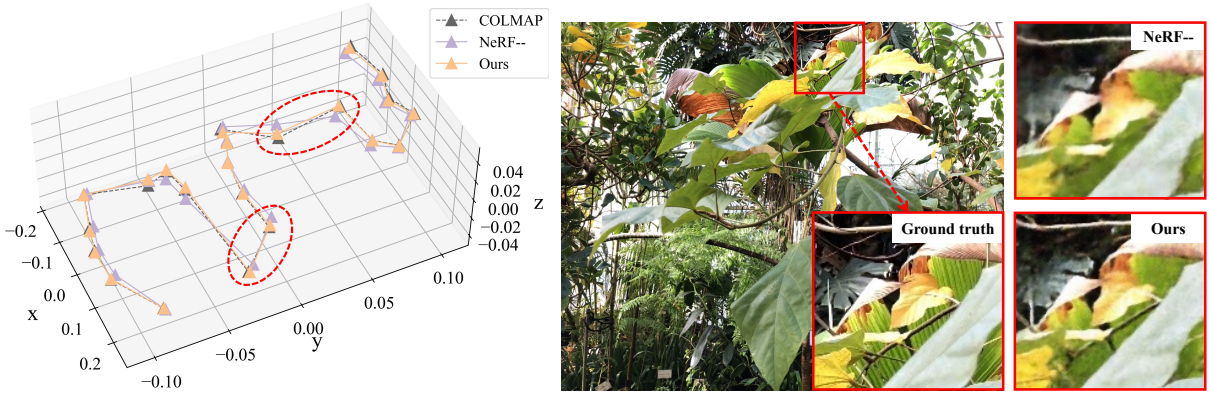
Tab. 1 shows the quantitative results of camera parameter estimation and view synthesis quality. The proposed method outperforms the other methods on average for solving the joint optimization problem of training neural radiance fields from unknown camera poses (extrinsics). As we discuss in Sec. 3.1 and Sec. 3.2, the APE parameters can be adjusted with the fine-grained gradients of PMLPSs to effectively predict accurate camera parameters (with 0.16° rotation error and $0.0019(m)$ translation error on average). Moreover, the different scenes should be represented by proper frequency bands. APE can search the proper frequency bands during training and synthesize images of more complete and fine details (with PSNR=24.95 on average).

Tab. 2 shows the quantitative results over a wider range of camera viewpoints. The camera poses estimated by our method highly agree with those estimated by COLMAP, indicating that the proposed APE can improve the performance of the implicit representation to better identify the image details in different camera viewpoints. Therefore, the proposed method can train the scene implicit representation with the unknown camera poses even in real-world 360° camera viewpoints.

4.3. Evalution w/o Camera Extrinsics and Intrinsics

We investigate the challenging problem of training the neural radiance fields with both unknown camera extrinsics and intrinsics. To increase the experimental difficulty and test the limits of the proposed method, the hidden units of all PMLPs are cut in half. We mainly compare the performance of the proposed method with NeRF--, which can also estimate the extrinsics and intrinsics of the camera. The baseline NeRF (colmap.NeRF) [45] is introduced to show the view synthetic quality for training the neural radiance field representation on the given known camera extrinsics and intrinsics (provided by the LLFF dataset) in the case where the size of original NeRF network is also reduced.

Fig. 6 shows some qualitative comparison of **Leaves** scene. Fig. 6(a) shows the comparison of 3D camera trajectories after training. Both trained for 10k epochs, the camera trajectory of our method more closely matches the ground truth (estimated by COLMAP) rather than NeRF--, especially at the corners (marked with red ellipses). Fig. 6(b) shows the comparison of synthetic details. In the same challenging situation where the hidden units of the



(a) **Comparison of Trajectories.** (b) **Comparison of Synthetic Details.**

Figure 6. **Qualitative Comparison of Evaluation w/o Camera Extrinsics and Intrinsics (Scene: Leaves).** We compare the 3D trajectories and fine details with the ground truth and NeRF--. Both results show that the proposed method can provide better performance.

Scene	Camera parameter estimation						View synthesis quality								
	Δrot(°) ↓		Δtrans(m) ↓		Δfocal(pixel) ↓		PSNR ↑			SSIM ↑			LPIPS ↓		
	NeRF--	Ours	NeRF--	Ours	NeRF--	Ours	NeRF--	Ours	colmap. NeRF*	NeRF--	Ours	colmap. NeRF*	NeRF--	Ours	colmap. NeRF*
Fern	1.78	**0.72**	0.009	**0.003**	153.5	**135.3**	21.67	**22.94**	22.22	0.61	**0.66**	0.64	0.50	**0.45**	0.47
Flower	4.84	**0.91**	0.005	**0.004**	13.2	**6.8**	25.34	**26.66**	25.25	0.71	**0.78**	0.71	0.37	**0.31**	0.36
Fortress	1.36	**1.17**	0.013	**0.011**	144.1	**130.6**	26.20	**28.14**	27.60	0.63	**0.73**	0.73	0.49	**0.38**	0.38
Horns	5.55	**2.21**	0.031	**0.010**	156.2	**132.1**	22.53	**24.20**	24.25	0.61	**0.67**	0.68	0.50	**0.45**	0.44
Leaves	3.90	**0.47**	0.003	**0.002**	59.0	**39.9**	18.88	**19.77**	18.81	0.53	**0.59**	0.52	0.47	**0.46**	0.47
Orchids	4.96	**2.38**	0.020	**0.009**	199.3	**153.3**	16.73	**19.32**	19.09	0.39	**0.49**	0.51	0.55	**0.50**	0.46
Room	2.77	**1.25**	0.019	**0.013**	331.8	**309.9**	25.84	**27.42**	27.77	0.84	**0.85**	0.87	0.44	**0.41**	0.40
T-rex	4.66	**1.63**	0.015	**0.003**	89.3	**86.2**	22.67	**22.83**	23.19	0.72	**0.74**	0.74	0.44	**0.40**	0.41
Mean	3.73	**1.34**	0.014	**0.007**	143.3	**124.3**	22.48	**23.91**	23.52	0.63	**0.69**	0.68	0.47	**0.42**	0.42

Table 3. **Quantitative Comparison of Evaluation w/o Camera Extrinsics and Intrinsics.** The best are in bold. $*$ denotes that colmap.NeRF is conducted under known camera parameters. Without known camera extrinsics and intrinics, the proposed method outperforms other methods in all scenes and the view synthesis quality is comparable to the reference results from colmap.NeRF.

implicit network are cut in half, the proposed method can still preserve some fine details, while the rendering results of NeRF-- are coarse and incomplete.

Tab. 3 shows the quantitative results of camera parameter estimation and view synthesis quality. The proposed approach outperforms NeRF-- or even colmap.NeRF in all scenes for training neural radiance fields from unknown camera extrinsics and intrinsics. Shrunking the size of the network degrades its performance in presenting a scene, and its corresponding frequency bands also shift. APE can adjust the frequency bands to fit the current scene, which the implicit network can represent and improve the view synthesis quality (with an average PSNR=23.91). The excellent scene representation can also additionally improve the camera parameter estimation performance (with 1.34° rotation error and $0.007(m)$ translation error on average).

4.4. Ablation Study

To investigate the effectiveness of each part of the proposed method, we conduct the ablation study on APE and β using ref.NeRF as the baseline method, as well as the ablation study on APE and PMLPs by following the same implementation details as the experiment without camera extrinsics and intrinsics. Note that all metrics in all tables are mean values from the evaluation on the LLFF dataset.

Configures	Camera parameter estimation	View synthesis quality PSNR
ref.NeRF	Known (Estimated from SfM)	22.56
ref.NeRF w APE w $\mathcal{L}_{fd}$ ($\beta = 1$)		23.65
ref.NeRF w APE w $\mathcal{L}_{fd}$ ($\beta = 0.1$)		23.70
ref.NeRF w APE w $\mathcal{L}_{fd}$ ($\beta = 0.01$)		**23.72**

Table 4. **Ablation Study on APE and β.** The best is in bold. ref.NeRF is served as a baseline to evaluate the effect of APE and $\mathcal{L}_{fd}$ under known camera parameters estimated from SfM. ref.NeRF with APE and lower β can further improve the view synthesis quality as compared to ref.NeRF.

Ablation study on APE and β. Since the scene is jointly represented with APE and network, it is hard to compute the accuracy of APE parameters with frequency values. However, as presented in Fig. 1, where the influence of improper frequencies can be verified from the regression accuracy, we can validate the effectiveness of APE through the improvement on view synthesis quality. Tab. 4 illustrates that ref.NeRF with APE can learn more proper frequency parameters for the scene representations and thus achieve bet-

Configures	Camera parameter estimation						View synthesis quality	
	Δrot(°)		Δtrans(m)		Δfocal(pixel)		PSNR	
	half-size	full-size	half-size	full-size	half-size	full-size	half-size	full-size
Ours w/o APE w/o PMLP	3.33	4.23	0.015	0.019	149.9	177.0	22.68	22.93
Ours w/o APE w PMLP	2.43	2.16	0.013	0.013	145.9	138.5	23.63	23.91
Ours w APE w/o PMLP	1.75	1.03	0.010	0.006	131.5	120.1	23.08	23.97
Ours w APE w PMLP	**1.34**	**0.86**	**0.007**	**0.004**	**124.3**	**109.8**	**23.91**	**24.52**

Table 5. **Ablation Study on APE and PMLPs.** The best is in bold. The study is conducted under unknown camera parameters with half-size (the numbers of hidden units in all layers are reduced to half) and full-size (unchanged) network. We investigate the contribution of the proposed APE and PMLPs. We compare the performance in the four cases of using APE with ReLU-activated MLPs (Ours w APE w/o PMLP), PE with PMLPs (Ours w/o APE w PMLP), PE with ReLU-activated MLPs (Ours w/o APE w/o PMLP) or APE with PMLPs (w APE w PMLP).

ter synthesis quality than ref.NeRF with PE. Moreover, β denotes a factor that can adjust the dropping rate of $\mathcal{L}_{fd}$ and achieve stable frequency diversity. Though all $\mathcal{L}_{fd}$ with various β converge to a low value, lower β can make $\mathcal{L}_{fd}$ drop faster, leading to a slight improvement in synthesis quality.

Ablation study on APE and PMLPs. Tab. 5 shows the quantitative comparison results. As compared with the positional encoding (PE) of handcraft frequency bands, our proposed APE can adjust the frequency bands to make these frequency bands proper to the current scene, improving the performance of scene representations and the camera pose estimation accuracy. As also shown in the results, the method with PMLPs can provide fine-grained gradients for reconstructing the neural radiance fields and thus improve both camera pose estimation accuracy and view synthesis quality. We also note that APE can reduce about 47% and 75% rotation error while PMLP can reduce 27% and 47% rotation error in half-size and full-size network, respectively. Furthermore, APE can exploit the richer representation capability of the full-size network to learn more proper frequency parameters and thereby achieves better view synthesis quality (PNSR=23.97) as compared to PMLP (PNSR=23.91).

5. Conclusion

We present adaptive positional encoding for bundle-adjusting neural radiance fields to train the scene radiance field representation from unknown camera parameters. The proposed APE can fit the proper frequency bands for better scene representation. Meanwhile, the PMLPs can reconstruct neural radiance fields as a high-order implicit function and yield fine-grained gradients. The experimental results demonstrate that the proposed method can solve the proposed joint optimization problem involving unknown camera parameters, frequency bands of APE, and implicit network parameters with the results of high-fidelity synthetic images and accurate camera parameters.

It should be noted that proper number of frequencies can also improve the performance of Fourier series regression. Since different scenes are distributed in different frequency bands, we should also represent different scenes with frequency parameters of proper values and numbers. However, the number of frequencies is not adaptive in our method and the performance can be further improved by implementing it. In the future work, we will follow these ideas and focus on implementing visual SLAM systems with neural radiance fields.

6. Acknowledge

This work was supported by STI 2030-Major Projects 2021ZD0201403, NSFC 62088101 Autonomous Intelligent Unmanned Systems, Zhejiang Provincial Natural Science Foundation of China under Grant No.LQ22F030022.

References

[1] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Building rome in a day. *ACM Communications*, 2011. 1

[2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 1, 2

[3] Cesar Cadena, Luca Carlone, Henry Carrillo, Yasir Latif, Davide Scaramuzza, José Neira, Ian Reid, and John J Leonard. Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age. *IEEE Transactions on robotics*, 32(6):1309–1332, 2016. 2

[4] Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32(3):1–12, 2013. 2

[5] Yisong Chen, Antoni B Chan, Zhouchen Lin, Kenji Suzuki, and Guoping Wang. Efficient tree-structured sfm by ransac generalized procrustes analysis. *Computer Vision and Image Understanding*, 157:179–189, 2017. 2

[6] S Chng, S Ramasinghe, J Sherrah, and S Lucey. Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In *ECCV*, 2022. 1, 7

[7] Hainan Cui, Shuhan Shen, Wei Gao, and Zhiheng Wang. Progressive large-scale structure-from-motion with orthog-

onal msts. In *2018 International Conference on 3D Vision (3DV)*, pages 79–88. IEEE, 2018. 2

[8] Weichen Dai, Yu Zhang, Ping Li, Zheng Fang, and Sebastian Scherer. Rgb-d slam in dynamic environments using point correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(1):373–389, 2020. 2

[9] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018. 2

[10] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. *IEEE transactions on pattern analysis and machine intelligence*, 40(3):611–625, 2017. 1, 2

[11] Christian Forster, Zichao Zhang, Michael Gassner, Manuel Werlberger, and Davide Scaramuzza. Svo: Semidirect visual odometry for monocular and multicamera systems. *IEEE Transactions on Robotics*, 33(2):249–265, 2016. 1

[12] Yasutaka Furukawa, Carlos Hernández, et al. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2):1–148, 2015. 2

[13] Kyle Gao, Yina Gao, Hongjie He, Denning Lu, Linlin Xu, and Jonathan Li. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*, 2022. 3

[14] Marcel Germann, Tiberiu Popa, Richard Keiser, Remo Ziegler, and Markus Gross. Novel-view synthesis of outdoor sport events using an adaptive view-dependent geometry. In *Computer graphics forum*, volume 31, pages 325–333. Wiley Online Library, 2012. 2

[15] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022. 1

[16] Benno Heigl, Reinhard Koch, Marc Pollefeys, Joachim Denzler, and L Van Gool. Plenoptic modeling and rendering from image sequences taken by a hand-held camera. In *Mustererkennung 1999*, pages 94–101. Springer, 1999. 2

[17] Fang Huang, Hao Yang, Xicheng Tan, Shuying Peng, Jian Tao, and Siyuan Peng. Fast reconstruction of 3d point cloud model using visual slam on embedded uav development platform. *Remote Sensing*, 12(20):3308, 2020. 1

[18] 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. 1

[19] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH computer graphics*, 18(3):165–174, 1984. 1

[20] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 2

[21] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems*, 31, 2018. 4

[22] 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. 1, 3, 5, 6, 7

[23] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5987–5997, 2021. 2

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[25] Mateusz Michalkiewicz, Jhony K Pontes, Dominic Jack, Mahsa Baktashmotlagh, and Anders Eriksson. Implicit surface representations as layers in neural networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4743–4752, 2019. 1

[26] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. 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*, pages 16190–16199, 2022. 1, 2

[27] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 5

[28] 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. 1, 2, 5

[29] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2

[30] Raul Mur-Artal and Juan D Tardós. Orb-slam2: An open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE transactions on robotics*, 33(5):1255–1262, 2017. 1

[31] Stanley Osher and Ronald Fedkiw. Signed distance functions. In *Level set methods and dynamic implicit surfaces*, pages 17–22. Springer, 2003. 2

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1

[33] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[34] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[35] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[36] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 134–144, 2019. 2

[37] 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. 2, 3, 4

[38] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 1

[39] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 6

[40] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6229–6238, 2021. 1, 2

[41] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2

[42] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in Neural Information Processing Systems*, 34:16558–16569, 2021. 2

[43] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Nießner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, volume 39, pages 701–727. Wiley Online Library, 2020. 1

[44] 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. 6

[45] 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. 2, 3, 6, 7

[46] Jimei Yang, Scott E Reed, Ming-Hsuan Yang, and Honglak Lee. Weakly-supervised disentangling with recurrent transformations for 3d view synthesis. *Advances in neural information processing systems*, 28, 2015. 2

[47] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1

[48] Raza Yunus, Yanyan Li, and Federico Tombari. Manhattanslam: Robust planar tracking and mapping leveraging mixture of manhattan frames. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6687–6693. IEEE, 2021. 2

[49] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[50] 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. 6

[51] Yipu Zhao and Patricio A Vela. Good feature matching: toward accurate, robust vo/vslam with low latency. *IEEE Transactions on Robotics*, 36(3):657–675, 2020. 2

[52] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4568–4577, 2018. 2

[53] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 2