
NeRF—: Neural Radiance Fields Without Known Camera Parameters

Zirui Wang
Active Vision Lab
University of Oxford
ryan@robots.ox.ac.uk

Shangzhe Wu
Visual Geometry Group
University of Oxford
szwu@robots.ox.ac.uk

Weidi Xie
Visual Geometry Group
University of Oxford
weidi@robots.ox.ac.uk

Min Chen
e-Research Centre
University of Oxford
min.chen@oerc.ox.ac.uk

Victor Adrian Prisacariu
Active Vision Lab
University of Oxford
victor@robots.ox.ac.uk

Abstract

Considering the problem of novel view synthesis (NVS) from only a set of 2D images, we simplify the training process of Neural Radiance Field (NeRF) on forward-facing scenes by removing the requirement of known or pre-computed camera parameters, including both intrinsics and 6DoF poses. To this end, we propose *NeRF—*, with three contributions: *First*, we show that the camera parameters can be jointly optimised as learnable parameters with NeRF training, through a photometric reconstruction; *Second*, to benchmark the camera parameter estimation and the quality of novel view renderings, we introduce a new dataset of path-traced synthetic scenes, termed as Blender Forward-Facing Dataset (BLEFF); *Third*, we conduct extensive analyses to understand the training behaviours under various camera motions, and show that in most scenarios, the joint optimisation pipeline can recover accurate camera parameters and achieve comparable novel view synthesis quality as those trained with COLMAP pre-computed camera parameters. Our code and data is available at <https://nerfmm.active.vision>.

1 Introduction

Generating photo-realistic images from arbitrary viewpoints not only requires to understand the 3D scene geometry but also complex viewpoint-dependent appearance resulted from sophisticated light transport phenomena. One way to achieve this is by constructing a 5D plenoptic function that directly models the light passing through each point in space [1] (or a 4D light field [2, 3] if we restrict ourselves outside the convex hull of the objects of interest). Unfortunately, physically measuring such a plenoptic function is usually not feasible in practice. As an alternative, Novel View Synthesis (NVS) aims to generate the unseen views from a small set of images captured from diverse viewpoints.

Recently, Neural Radiance Fields (NeRF) [4, 5, 6, 7] and Multi-Plane Images (MPI) [8, 9, 10, 11, 12, 13] have demonstrated that, by optimising a volumetric scene representation on a sparse set of images, one can re-render the scene from novel viewpoints with impressive visual quality, including sophisticated view-dependent effects, such as specularities and transparency.

One limitation for these methods is the requirement of the camera parameters of the images at the training time, which is rarely accessible in real scenarios, *e.g.*, images taken with a mobile phone. In practice, these methods usually pre-compute the camera parameters via conventional techniques. For example, NeRF and its variants [4, 5, 6, 7] adopt COLMAP [14] to estimate the camera parameters (both intrinsics and extrinsics) for each input image. This pre-processing step,

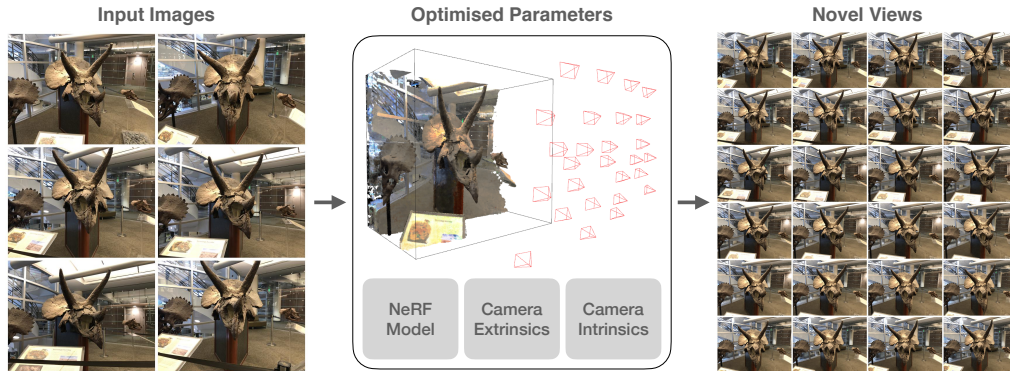


Figure 1: We propose *NeRF--*, a NeRF-based framework for novel view synthesis without pose supervision. Specifically, our method jointly optimises camera poses and intrinsics of a set of forward-facing input images while training a NeRF model.

apart from introducing additional complexity, can be potentially erroneous or simply fail, due to homogeneous regions, fast-changing view-dependent appearance, or degenerate solutions in linear equations from highly ambiguous camera trajectories [15, 16, 17].

Our goal in this paper is to investigate novel view synthesis for forward-facing scenes with unknown camera parameters. We introduce *NeRF--*, a framework that treats camera poses and intrinsics as learnable parameters, and jointly optimises them with a 3D scene representation (as shown in Fig. 1). Given only a sparse set of images captured in a forward-facing setup, our system can be trained 1) to estimate camera parameters for each image and 2) to learn a volumetric scene representation through a photometric reconstruction loss, achieving high fidelity NVS results comparable to existing two-stage COLMAP-NeRF pipelines while removing the pre-processing step for camera parameters.

In order to establish a comprehensive evaluation of the method, we make two additional contributions: 1) we introduce a high-quality path-traced synthetic dataset, named Blender Forward Facing (BLEFF), to benchmark camera parameter estimation and novel view synthesis quality, which is publicly accessible for future research; 2) we conduct thorough analyses on our method and COLMAP-NeRF under various camera motions, showing that both systems can tolerate up to $\pm 20^\circ$ of rotation and $\pm 20\%$ translation perturbations, and that the joint optimisation is more favourable than COLMAP in translational perturbations, but less competitive in rotational perturbations.

2 Related Work

We divide the related work roughly into two categories, one assuming camera parameters are input to NVS representations, and the other jointly estimating camera parameters with an NVS representation from uncalibrated images.

With Known Camera Parameters Images and their camera parameters are usually required by many novel view synthesis systems. In this line of work, classical methods Light Field Rendering [3] and Lumigraph [2] model a simplified plenoptic function. [18] integrates sophisticated hand-crafted material BRDF models into Signed Distance Functions (SDF). FVS[19] and SVS[20] combine meshes and deep features from images. Recently, dense volumetric representations have been proposed to enable smooth gradients for photometry-based optimisation and has shown to be promising for photo-realistic novel view synthesis of highly complex shapes and view-dependent appearance. These representations include Soft3D[21], Multi-Plane Images (MPI)[8, 9, 10, 11, 12, 13], Scene Representation Networks (SRN)[22], implicit surface [23, 24], and Neural Radiance Fields (NeRF)[4, 6, 7, 5]. Despite these approaches differ in the way of scene representations, they all require input images to be associated with known camera parameters.

With Unknown Camera Parameters The second line of work, mostly consisting of SLAM and SfM systems, aims to reconstruct the scene directly from RGB images by jointly estimating camera parameters and 3D geometries. For example, MonoSLAM[25] and ORB-SLAM [26] reconstruct a point cloud map and estimate camera poses in real-time by associating feature correspondences. Similarly, DTAM [27], SVO[28], and LSD-SLAM [29] approach the SLAM task by minimising a photometric loss. SfM systems Bundler[30] and COLMAP [14] can reconstruct a global consistent

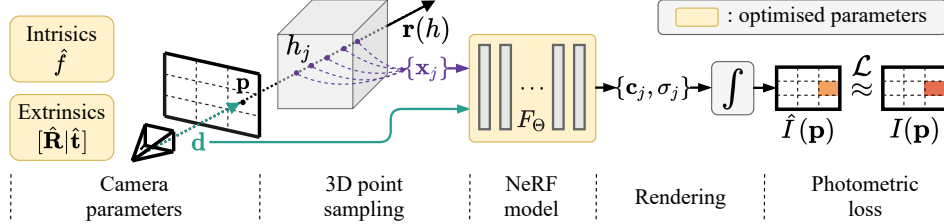


Figure 2: Our pipeline jointly optimises a NeRF model and the camera parameters of the input images by minimising the photometric reconstruction errors. To render a pixel \mathbf{p} from NeRF, given the optimised camera parameters $(\hat{f}, \hat{\mathbf{R}}, \hat{\mathbf{t}})$, we feed 3D points \mathbf{x}_j sampled along the camera ray together with the viewing direction \mathbf{d} into NeRF F_Θ , and aggregate the output radiance \mathbf{c}_j and densities σ_j to obtain its colour $\hat{I}(\mathbf{p})$. The entire pipeline can be trained end-to-end using only RGB images with unknown cameras as input.

map and estimate camera parameters for large image sets, but sensitive to initial image pairs. Although these methods only require RGB images as input, they often assume diffuse surface appearance in order to establish correspondences and cannot recover view-dependent appearance, hence resulting in unrealistic novel view rendering.

Concurrent to our work, several methods [31, 32, 33] have developed similar ideas on estimating camera parameters with a NeRF model. These methods either requires intrinsics as input [32, 33], or assumes a pre-trained NeRF is available [31]. In contrast, we propose to jointly optimise camera poses, intrinsics and a NeRF in an end-to-end manner on forward-facing scenes. As a result, we show the proposed framework is able to produce photo-realistic view synthesis that are comparable to two-stage NVS systems.

3 Preliminary

Given a set of images $\mathcal{I} = \{I_1, I_2, \dots, I_N\}$ captured from N sparse viewpoints of a scene, with their associated camera parameters $\Pi = \{\pi_1, \pi_2, \dots, \pi_N\}$, including both intrinsics and 6DoF poses, the goal of novel view synthesis is to come up with a scene representation that enables the generation of realistic images from novel and unseen viewpoints.

In this paper, we follow the approach proposed in NeRF [4]. Specifically, NeRF adopts an neural representation to construct a radiance field from sparse input views, where the view-dependent appearance is modelled by a continuous function $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$, which maps a 3D location $\mathbf{x} = (x, y, z)$ and a viewing direction $\mathbf{d} = (\theta, \phi)$ to a radiance colour $\mathbf{c} = (r, g, b)$ and a density σ .

To render an image from the NeRF model, the colour at each pixel $\mathbf{p} = (u, v)$ on the image plane \hat{I}_i is obtained by a rendering function \mathcal{R} , aggregating the radiance along a ray shooting from the camera position \mathbf{o}_i , passing through the pixel \mathbf{p} into the volume [34, 2]:

$$\hat{I}_i(\mathbf{p}) = \mathcal{R}(\mathbf{p}, \pi_i | \Theta) = \int_{h_n}^{h_f} T(h) \sigma(\mathbf{r}(h)) \mathbf{c}(\mathbf{r}(h), \mathbf{d}) dh, \quad (1)$$

where $T(h) = \exp(-\int_{h_n}^h \sigma(\mathbf{r}(s)) ds)$ denotes the accumulated transmittance along the ray, *i.e.*, the probability of the ray travelling from h_n to h without hitting any other particle, and $\mathbf{r}(h) = \mathbf{o} + h\mathbf{d}$ denotes the camera ray that starts from camera origin \mathbf{o} and passes through \mathbf{p} , controlled by the camera parameter π_i , with near and far bounds h_n and h_f . In practice, the integral in Eq. (1) is approximated by accumulating radiance and densities of a set of sampled points along a ray. With this implicit scene representation $F_\Theta(\mathbf{x}, \mathbf{d})$ and a differentiable renderer \mathcal{R} , NeRF can be trained by minimising the photometric error $\mathcal{L} = \sum_i^N \|I_i - \hat{I}_i\|_2^2$ between the observed views and synthesised ones $\hat{\mathcal{I}} = \{\hat{I}_1, \dots, \hat{I}_N\}$ under known camera parameters:

$$\Theta^* = \arg \min_{\Theta} \mathcal{L}(\hat{\mathcal{I}} | \mathcal{I}, \Pi). \quad (2)$$

To summarise, NeRF represents a 3D scene as a radiance field parameterised by MLPs, which is trained by minimising the discrepancy between the observed and rendered images. Note that, the camera parameters π_i for these given images are required for training, which are usually estimated by SfM packages, such as COLMAP [14]. For more details of NeRF, we refer the readers to [4].

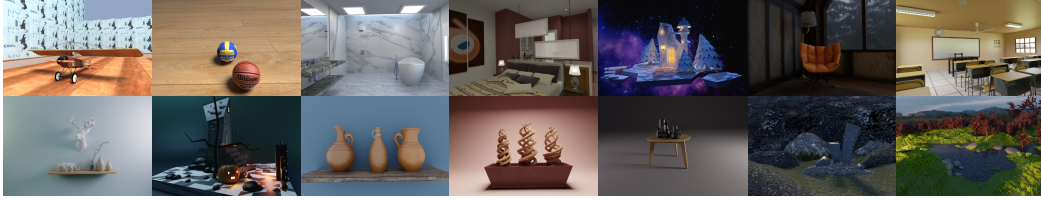


Figure 3: Thumbnails of BLEFF dataset.

4 Method

In this paper, we investigate the training framework for novel view synthesis without known camera parameters. Our framework jointly optimises the camera parameters and a scene representation from a set of input RGB images \mathcal{I} . We make two assumptions: *First*, all images are captured in a forward-facing setup with certain amount of rotation and translation flexibility, as described in LLFF [10] and original NeRF[4]. *Second*, all images are captured with same intrinsic parameters. Mathematically, our proposed framework *NeRF--* can be formulated as:

$$\Theta^*, \Pi^* = \arg \min_{\Theta, \Pi} \mathcal{L}(\hat{\mathcal{I}}, \hat{\Pi} | \mathcal{I}), \quad (3)$$

where Π refers to the camera parameters, including both intrinsics and 6DoF poses. In addition to simplifying the original two-stage approach, this joint optimisation approach enables globally consistent reconstruction results, similar to the bundle adjustment used in conventional SfM pipelines [35] and SLAM systems [25, 27, 29].

In the following sections, we first introduce the representations for the camera parameters and then describe the process of the joint optimisation.

4.1 Camera Parameters

Camera Intrinsics can be expressed as a focal length f and principle points c_x and c_y for a pinhole camera model. Without losing generality, we consider the centre of sensor as the camera principle points, *i.e.*, $c_x \approx W/2$ and $c_y \approx H/2$, where H and W denote the height and the width of the image, and all input images are taken by the same camera. Thus, estimating the camera intrinsics only refers to finding the focal length f in this case.

Camera Poses can be expressed as a camera-to-world transformation matrix $\mathbf{T}_{wc} = [\mathbf{R} | \mathbf{t}]$ in $\text{SE}(3)$, where $\mathbf{R} \in \text{SO}(3)$ and $\mathbf{t} \in \mathbb{R}^3$ denote the camera rotation and translation respectively. Specifically, optimising the translation vector \mathbf{t} can be as simple as setting it as trainable parameters as it is defined in the Euclidean space. Whereas for the camera rotation, which is in $\text{SO}(3)$ space, we adopt the axis-angle representation: $\phi := \alpha\omega$, $\phi \in \mathbb{R}^3$, where ω and α denote a normalised rotation axis and a rotation angle α respectively. A rotation matrix \mathbf{R} can be recovered from the Rodrigues' formula $\mathbf{R} = \mathbf{I} + \frac{\sin(\alpha)}{\alpha} \phi^\wedge + \frac{1 - \cos(\alpha)}{\alpha^2} (\phi^\wedge)^2$, where $(\cdot)^\wedge$ is the skew operator that converts a vector ϕ to a skew matrix. Till here, we can optimise the camera poses for each input image I_i , with trainable parameters ϕ_i and \mathbf{t}_i .

4.2 Joint Optimisation of Camera Parameters and Scene Representation

Now we introduce the details for training *NeRF--*, which jointly optimises camera parameters along with a NeRF model on a sparse set of forward-facing views.

Specifically, to render a pixel p on a given image I_i , we shoot a ray $\hat{\mathbf{r}}_{i,p}(h) = \hat{\mathbf{o}}_i + h\hat{\mathbf{d}}_{i,p}$, from the camera position through the pixel p into the radiance field, with the current estimates of the camera parameters $\hat{\pi}_i = (\hat{f}, \hat{\phi}_i, \hat{\mathbf{t}}_i)$, and the ray direction

$$\hat{\mathbf{d}}_{i,p} = \hat{\mathbf{R}}_i \begin{pmatrix} (u - W/2)/\hat{f} \\ -(v - H/2)/\hat{f} \\ -1 \end{pmatrix}, \quad (4)$$

where $\hat{\mathbf{o}}_i = \hat{\mathbf{t}}_i$ and $\hat{\mathbf{R}}_i$ is computed from $\hat{\phi}_i$ using the Rodrigues' formula. Along the ray, we sample a number of 3D points $\{\mathbf{x}_j\}$ and evaluate the radiance colours $\{c_j\}$ and the density values $\{\sigma_j\}$ of

these points from the NeRF model F_{Θ} . To get the colour for that pixel, the rendering function Eq. (1) is applied to aggregate the predicted radiance and densities along the ray.

During training, we randomly render M pixels for each input image, and minimise the reconstruction loss between the rendered colours against the ground-truth colours on these pixels. Note that, the entire pipeline is fully differentiable, which enables the jointly optimisation of NeRF and the camera parameters $\{\pi_i\}$ by minimising the reconstruction loss.

5 Blender Forward Facing Dataset

Evaluating the performance of a joint optimisation system like ours requires both images and their ground truth camera parameters, which are not easy to acquire for real captured images. To that end, we introduce a synthetic dataset, Blender Forward Facing (BLEFF), containing 14 path-traced scenes¹, with each rendered in multiple levels of rotation and translation perturbations. We will make this dataset publicly available, expecting to foster future research on the joint optimisation of camera parameters and scene representations.

We now detail how we create the dataset. Starting with a perfect forward-facing camera trajectory (top row in Fig. 4), *i.e.*, all cameras looking forward and moving in xy -plane ($z=0$), we generate 16 trajectory variants for each scene by adding 5/5/6 levels of rotation/translation/full-6DoF perturbations on the perfect trajectory. We denote these trajectory variants via a $t_{aaa}r_{bbb}$ notation, where t_{aaa} and r_{bbb} denote $\pm aaa\%$ of translation and $\pm bbb$ degree of rotation perturbed, respectively. Since each synthetic scene is in different scale, we add translation perturbation proportional to the maximum dimension of a trajectory, hence the % notation. Fig. 3 shows image examples of BLEFF and Fig. 4 presents a t_{010} trajectory example.

Background It is surprisingly difficult to find suitable datasets to jointly evaluate the performance of camera parameters and a NVS model at the same time. In general, three types of dataset are available, but cannot meet our requirements. The first type focuses on NVS rendering quality with their camera parameters estimated from SfM and SLAM packages[14, 26]. These datasets include LLFF-NeRF[10], RealEstate10K[8], Spaces[12], Shiny[13], and Tanks & Temples[36]. Without ground truth camera parameters, it is impossible to evaluate the camera pose estimation accurately. The second type is designed to benchmark SLAM systems, *e.g.*, TUM-RGBD[37] and ETH3D[38], with “ground truth” camera parameters being estimated from an external motion capture system. However, we find the time lags between motion capture system and the captured images are unignorable in NVS tasks and often lead to blurry renderings. The third type, *e.g.*, SceneNet RGB-D[39] and InteriorNet[40], offers large-scale simulated indoor images but lacks of scene diversities.

6 Experiments

6.1 Setups

Dataset Including our BLEFF dataset, we evaluate *NeRF*— on three datasets: **LLFF**: We first conduct experiments on the same forward-facing dataset as that in NeRF, namely, LLFF-NeRF [10], which has 8 forward-facing scenes captured by mobile phones or consumer cameras, each containing 20-62 images. In all experiments, we follow the official pre-processing procedures and train/test splits, *i.e.*, the resolution of the training images is 756×1008 , and every 8-th image is used as a test image. **RealEstate10K**: To understand the behaviour of NVS in real-life capture scenarios, we additionally collected a number of diverse scenes extracted from the short video segments in RealEstate10K [8]. In particular, the image resolution in these sequences is 1080×1920 and the frame rate is 30 fps. We sub-sample the frames and reduce the frame rates to 5 fps, with each sequence containing 14-43 images. **BLEFF**: We use BLEFF to evaluate camera parameter estimation accuracy

¹We modified open sourced blender files downloaded from <https://www.blendswap.com/>. Most blender files come with CC-BY-3.0 licence. Details see our supplementary.

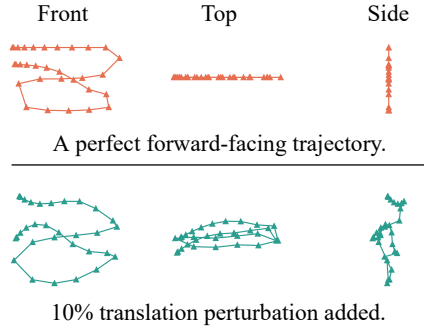


Figure 4: Dataset trajectory generation.

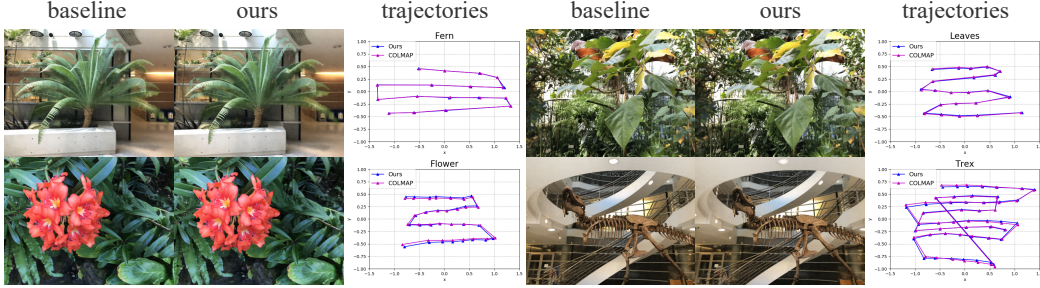


Figure 5: Qualitative comparison between our *NeRF*— model with unknown cameras and the *baseline NeRF* on LLFF-NeRF dataset. For each example, on the left, we show the synthesised novel views from the *baseline NeRF* and from our model; on the right, we compare our optimised camera trajectories with the ones estimated from COLMAP, aligned with a Sim(3) transformation. Our proposed *NeRF*— recovers similar camera poses with COLMAP and produces high quality novel views comparable to the *baseline NeRF*. Full results can be found in the supplementary.

and NVS rendering quality, as well as to explore the breaking point of the *NeRF*— and *baseline NeRF*. All scenes in BLEFF come with 31 images in 1040×1560 resolution, with 27 for training and 4 for testing, and focal lengths set to 1040 pixels in both horizontal and vertical directions. In the experiments below, all BLEFF images are downsized by half, *i.e.*, images are in 520×780 and the ground truth focal length f_{gt} is 520 for training and testing.

Metrics We evaluate the proposed framework from two aspects: First, to measure the quality of novel view rendering, we adopt common metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [41] and Learned Perceptual Image Patch Similarity (LPIPS) [42]; Second, we evaluate the accuracy of the optimised camera parameters, including focal length, rotation and translation. For focal length evaluation, we report the absolute error in the metric of pixels. For the camera poses, we follow the evaluation protocol of Absolute Trajectory Error (ATE) [43, 37], aligning two sets of pose trajectories globally with a similarity transformation Sim(3) and reporting the rotation angle between two rotations and the absolute distance between two translation vectors.

Implementation Details We implement our framework in PyTorch following the same architecture as original NeRF, except that, for computation efficiency, we: (a) do not use the hierarchical sampling strategy; (b) reduce the hidden layer dimension from 256 to 128; and (c) sample only 1024 pixels from every input image and 128 points along each ray. We use Kaiming initialisation [44] for the NeRF model, and initialise all cameras to be at origin looking at $-z$ direction, with focal length f to be the image width. We use three separate Adam optimisers for NeRF, camera poses and focal lengths respectively, all with an initial learning rate of 0.001. The learning rate of the NeRF model is decayed every 10 epochs by multiplying with 0.9954 (equivalent to a stair-cased exponential decay), and learning rates of the pose and focal length parameters are decayed every 100 epochs with a multiplier of 0.9. All models are trained for 10000 epochs unless otherwise specified. More technical details are included in the supplementary material.

6.2 Results

6.2.1 On Novel View Synthesis Quality

We compare the novel view rendering results from *baseline NeRF*, where camera parameters are estimated from COLMAP, and our proposed model *NeRF*—, which jointly optimises the camera parameters and the 3D scene representation. We report the quantitative evaluations in Table 1 and visual results in Fig. 5 and Fig. 6. Overall, our joint optimisation model, which does not require camera parameters as inputs, achieves similar NVS quality compared to the *baseline NeRF* model (perceptual quality metric Δ SSIM and Δ LPIPS = 0.05). This confirms



Figure 6: Qualitative comparison of NVS quality between our model and the baseline on RealEstate10K. Our model produces better results on the first examples but slightly worse results on the second example, indicating comparable performance overall compared to baseline.

| Scene | SSIM \uparrow | | LPIPS \downarrow | | PSNR \uparrow | | Camera Parameters Difference | | |
|----------|-----------------|------|--------------------|------|-----------------|-------|------------------------------|---------------|------------------------|
| | colmap | ours | colmap | ours | colmap | ours | Δ rot (deg) | Δ tran | Δ focal (pixel) |
| Fern | 0.64 | 0.61 | 0.47 | 0.50 | 22.22 | 21.67 | 1.78 | 0.029 | 153.5 |
| Flower | 0.71 | 0.71 | 0.36 | 0.37 | 25.25 | 25.34 | 4.84 | 0.016 | 13.2 |
| Fortress | 0.73 | 0.63 | 0.38 | 0.49 | 27.60 | 26.20 | 1.36 | 0.025 | 144.1 |
| Horns | 0.68 | 0.61 | 0.44 | 0.50 | 24.25 | 22.53 | 5.55 | 0.044 | 156.2 |
| Leaves | 0.52 | 0.53 | 0.47 | 0.47 | 18.81 | 18.88 | 3.90 | 0.016 | 59.0 |
| Orchids | 0.51 | 0.39 | 0.46 | 0.55 | 19.09 | 16.73 | 4.96 | 0.051 | 199.3 |
| Room | 0.87 | 0.84 | 0.40 | 0.44 | 27.77 | 25.84 | 2.77 | 0.030 | 331.8 |
| Trex | 0.74 | 0.72 | 0.41 | 0.44 | 23.19 | 22.67 | 4.67 | 0.036 | 89.3 |
| Mean | 0.68 | 0.63 | 0.42 | 0.47 | 23.52 | 22.48 | 3.73 | 0.031 | 143.3 |

Table 1: Quantitative comparison between our model and the *baseline NeRF* on LLFF-NeRF dataset. On the left, we show the NVS quality of our method with unknown cameras, achieving comparable results to *baseline NeRF*: Δ SSIM and Δ LPIPS = 0.05, Δ PSNR = 1.0. Note that the *baseline NeRF* numbers in this table is lower than the numbers reported in original paper, as we employed a smaller NeRF model for both *baseline NeRF* and *NeRF--* for computation efficiency. On the right, we report the difference between our optimised camera parameters and ones computed from COLMAP. The result suggests our optimised camera poses are close to COLMAP estimations.

that jointly optimising the camera parameters and 3D scene representation is indeed possible. Nevertheless, we observe that for both the *Orchids* and the *Room*, our *NeRF--* model produces slightly worse results compared to the *baseline NeRF*. One reason can be our joint optimisation for these two scenes might have trapped into local minima, ending up with inaccurate intrinsics, as we notice from Table 1 that the difference between optimised camera focal lengths and COLMAP estimation are most noticeable for these two scenes (199.3 and 331.8).

6.2.2 On Camera Parameter Estimation

Considering a forward-facing image capturing process in real life, where a user is likely to introduce unstable camera motions, to simulate this effect, we select $t_{010}r_{010}$ subset in BLEFF². Table 2 lists the L1 difference on the estimated focal length, and metrics on camera rotation and translation computed with the ATE toolbox [43]. In summary, our rotation estimation error is about 5°, and focal length error is about 25 pixels, where as COLMAP has a slightly larger rotation error but lower focal length error. As a result, the NVS quality of our method is comparable with *baseline NeRF*.

| Method | Camera Parameters Error (vs. GT) | | | NVS Quality | |
|--------|----------------------------------|--------------------|---------------|-----------------|-----------------|
| | Δ focal (pixel) | Δ rot (deg) | Δ tran | SSIM \uparrow | PSNR \uparrow |
| colmap | 14.89 | 13.65 | 0.0127 | 0.90 | 33.92 |
| ours | 20.55 | 4.45 | 0.0654 | 0.90 | 33.24 |

Table 2: Quantitative evaluation of camera parameters estimation for both COLMAP and our method. We show that both the camera parameters estimation and NVS performance are comparable to COLMAP in BLEFF $t_{010}r_{010}$. A comprehensive list of results for each scene can be found in the supplementary.

Visualisation of Pose Optimisation For a better understanding of the optimisation process, we visualize the camera poses at various training epochs for the scene *Flower* from LLFF-NeRF dataset (Fig. 7). Starting from identity, our optimised camera poses gradually converge towards COLMAP estimations after about 1000 epochs, subject to a similarity transformation between optimised camera parameters and those estimated from COLMAP.

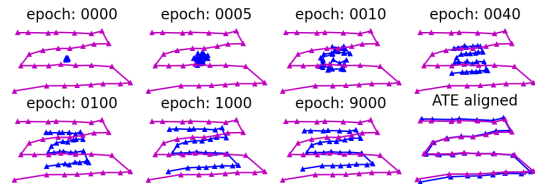


Figure 7: Pose optimisation during training, visualised on xy -plane (purple: COLMAP, blue: ours).

6.3 Discussion: Effects of Camera Motions

Considering *baseline NeRF* and *NeRF--* as two alternatives, this section demystifies when and why one method are more favourable than the other under various camera motion patterns. We expand

²Suppose a trajectory of interest is bounded by a $1m \times 1m$ square, $\pm 10\%$ translation perturbation offers about $\pm 10\% \times 1.414m \Rightarrow 28.3cm$ flexibility in z direction during capturing.

this section from two aspects: First, we study the performance of our system and COLMAP under multiple levels of camera trajectory perturbations and explore the breaking point of both systems; Second, we examine system behaviours when a camera follows certain motion patterns during image capturing. As this section unfolds, we can see the necessity of discussing these common aspects in real-life capturing.

6.3.1 Breaking Point in Forward-Facing Scenes

The forward-facing assumption we made, which in an ideal case, assumes all cameras looking to $-z$ direction and moving within xy -plane. However, this assumption is often over optimistic in real-life image capturing, it is thus interesting to study the robustness of both COLMAP and *NeRF*— under multiple levels of trajectory perturbations (Fig. 8).

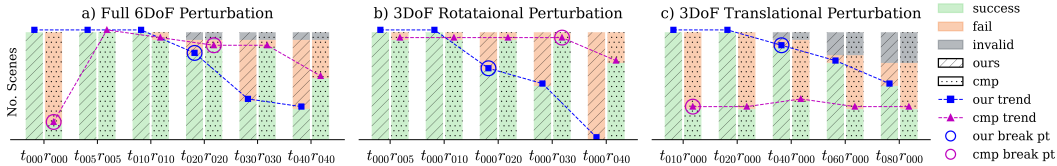


Figure 8: Breaking point analysis for camera parameter estimation in three groups of perturbation experiments: a) full 6DoF, b) 3DoF rotation, and c) 3DoF translation. COLMAP is shorten to "cmp" in the graph. We define a breaking point to the trajectory variant where a method starts failing in an experiment group. In group a) with full 6DoF noise, both methods start failing at $t_{020}r_{020}$. In 3DoF perturbations, our method performs more stable in translation perturbations but less stable in rotation perturbations. Note that COLMAP also faces degenerate issues with $t_{000}r_{000}$. Check Section 6.3.2 for more details on degenerate cases.

Setup Three groups of perturbation experiments are conducted in this section (Fig. 8): a) full 6DoF, b) 3DoF rotation, and c) 3DoF translation. We train all models for 3K epochs (not full 10K) and compare camera parameter estimations with ground truth with a criteria of mean rotation error $< 20^\circ$ and focal length error $< 50\%$ ground truth focal, as we observe severe NVS performance drop when the criteria is violated.

Result With full 6DoF perturbation added, both *NeRF*— and COLMAP start breaking when translation noise $> \pm 20\%$ and rotation noise $> \pm 20^\circ$. Our method is favourable than COLMAP in large translation perturbations, but underperforms COLMAP in large rotation cases. Two reasons account for this result. First, it is well known that feature descriptors can handle appearance changes better than photometric methods, hence the COLMAP’s superior performance in large rotation perturbation. However, feature-based method can also be vulnerable to degenerate cases, resulting in the failure cases in translation perturbation experiments, as will be discussed in Section 6.3.2.

6.3.2 Controlled Motion Patterns

In this section, we study three controlled camera motion patterns (Fig. 9): a) Rotational, where a camera motion contains mostly rotation with little translation. Note this is different from rotation perturbation experiment in Section 6.3.1, where the camera follows a planar trajectory but with perturbed viewing directions; b) Translational, in which a camera moves parallel to the camera sensor; c) Track-to-object, when a camera tracks an object. At the first glance, one may doubt if these controlled motions only exist in the synthetic world. In fact, they can often happen in real-life image capturing, especially when a camera is attached to a gimbal, dolly, or a drone.

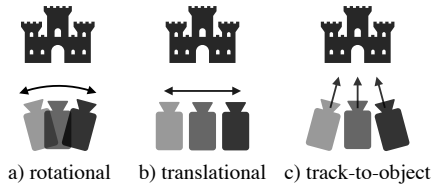


Figure 9: Three common camera motions in real-life capturing.

In short, our method outperforms COLMAP in rotational and translation camera motions, but underperforms COLMAP when a camera tracks an object.

Rotational and Translational As shown in Fig. 8c) and $t_{000}r_{000}$ column, COLMAP gives inaccurate camera parameter estimations in all translational experiments. Similar results are also observed in rotational motion. Fig. 10 presents two examples from the RealEstate10K dataset, where COLMAP fails to initialise or produces inaccurate estimations. For explanation, these results are known as degenerate cases while solving fundamental matrices from the point correspondences [15, 17]. Moreover, in [16], Torr et al. further point out that these degenerate cases are even more difficult to identify when matching outliers exist. Consequently, feature-matching-based SfM and SLAM methods tend to be vulnerable to these cases, whereas our photometry-based method handles them well.

Track-to-object We observe an inaccurate joint optimisation from our model, when a camera is set to track to an object during capturing, even when the camera rotation is well below our breaking point ($\pm 20^\circ$). Fig. 11 illustrates such a situation. One hypothesis is, such track-to-object motion results in an effect that the object position and the image composition remain almost unchanged during capturing, which is not a problem for systems that work with high-res images directly, e.g., COLMAP and SLAM systems. However, our method is supervised from the photometric loss between captured images and synthesised images. When the model is trained from scratch, synthesised images are blurry during most of the training time, producing little supervision for camera parameter optimisation.

6.4 Limitations and Future Work

Our system faces two limitations. *Firstly*, as motioned in section Section 6.3.2, our method does not estimate meaningful camera parameters when images are captured in a track-to-object motion, as there is not enough supervision to guide the optimisation of camera parameters in such motions. *Secondly*, we design our system to handle forward-facing scenes, which allows us initialising camera poses from identity matrices and optimising camera poses via gradient descent. However, this assumption restricts our pose estimation capability when the camera undergoes large rotations. As a result, camera motions with rotation perturbations larger than $\pm 20^\circ$ might cause a failure (Fig. 8), and our method certainly cannot handle 360° scenes (Fig. 12). As for future work, exploring relative pose estimation between image pairs and temporal relationships in sequences might enable the processing of such 360° scenes.

7 Conclusions

Our contribution in this work is threefold. *First*, we present *NeRF--*, a framework that jointly optimises camera parameters and scene representation for forward-facing scenes, eliminating the need of pre-computing camera parameters, while achieving view synthesis results on par with the COLMAP-enabled NeRF baseline. *Second*, we introduce the BLEFF datasets for evaluating such jointly optimised novel view synthesis system with ground-truth camera parameters and high-quality path-traced images. *Third*, we present extensive experimental results and demonstrate the viability of this joint optimisation framework under different camera trajectory patterns, even when COLMAP fails to estimate the camera parameters. Despite its current limitations discussed above, our proposed joint optimisation pipeline has demonstrated promising results on this highly challenging task, and advanced one step towards end-to-end novel view synthesis on more general scenes.

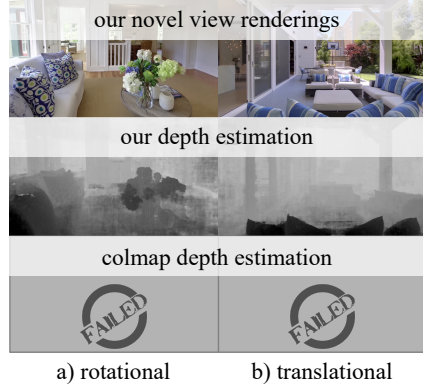


Figure 10: Rotational and translational scenes that are problematic to COLMAP in real image dataset RealEstate10K.

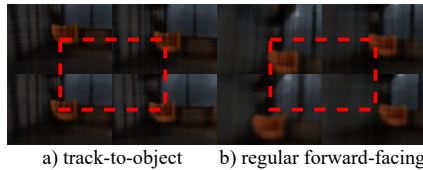


Figure 11: Rendered training views in two types of scenes at epoch 20. Unlike regular forward-facing scenes, where blurry rendered images are distinguishable even at early training stage, track-to-object scenes cannot provide enough photometric supervision from blurry rendered images, as the object position remains mostly unchanged (hinted by red dashed line) in all views.

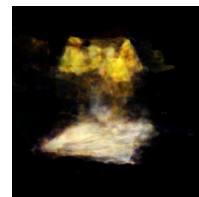


Figure 12: Failed *Lego* 360° scene.

Acknowledgement

Shangzhe Wu is supported by Facebook Research. Weidi Xie is supported by Visual AI (EP/T028572/1). The authors would also like to thank Tim Yuqing Tang for insightful discussions and proofreading.

Appendix A Additional Results

In this section, we provide additional results from 4 directions: 1) NVS rendering quality; 2) Camera parameter estimation accuracy; 3) Full breaking point analysis results; and 4) Results of an optional refinement step.

A.1 On NVS Quality

Fig. 13 presents additional qualitative results on LLFF-NeRF dataset. Our method achieves comparable novel view synthesis quality to the *baseline NeRF* while taking RGB images as the only inputs.

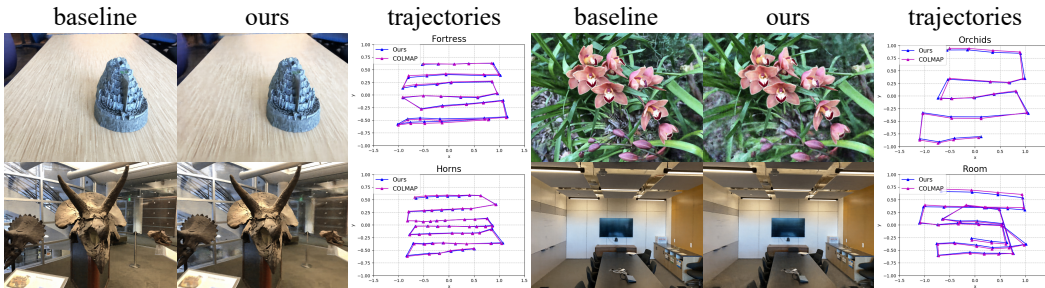


Figure 13: Additional Qualitatively Results on LLFF-NeRF Dataset. Our method achieves comparable NVS quality to the COLMAP-enabled NeRF while taking RGB images as the only input.

A.2 On Camera Parameter Estimation

We report the full pose estimation results for Section 6.2.2. In short, our method produces accurate pose estimations and comparable NVS rendering quality to COLMAP-enabled NeRF ($\Delta\text{SSIM} = 0$).

Pose Ambiguity We notice a special case *Roundtable*, where there is a 179.10° rotation error between COLMAP predicted poses and ground truth poses but the COLMAP-NeRF model still produces high quality NVS renderings. This is a case of ambiguous pose estimation as stated in [45, 46], where two sets of pose estimations are mirror images with respect to a plane. Nevertheless, as the estimated mirrored poses and the estimated 3D geometries from COLMAP are self-consistent, a NeRF model trained in this case can still provide plausible novel view synthesis, hence the high PSNR and SSIM in evaluation.

A.3 Breaking Point Analysis

In Fig. 14 we present a per-scene breakdown for the breaking point analysis in Section 6.3.1. Our method performs better than COLMAP in translation perturbation but worse in rotation perturbation.

A.4 Optional Refinement

We propose an optional refinement step to further improve the quality of the synthesised images. Specifically, after the first training process is completed, we drop the trained NeRF model and re-initialise it with random parameters while keeping the pre-trained camera parameters. We then repeat the joint optimisation using the pre-trained camera parameters as initialisation. We find this additional refinement step generally leads to sharper images and improves the synthesis results, as evidenced by the comparison in Table 4.

| | Traj \ S-ID | Ours | | | | | | | | | | | | | COLMAP | | | | | | | | | | | | | Legend | | |
|------------------|-------------|------|---|---|---|---|---|---|---|---|---|----|----|----|--------|---|---|---|---|---|---|---|---|---|---|----|----|--------|------------|----------|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | S-ID |
| 6DoF | t000r000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 0 | Airplane |
| | t005r005 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 1 | Balls |
| | t010r010 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 2 | Bathroom |
| | t020r020 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 3 | Bed |
| | t030r030 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 4 | Castle |
| | t040r040 | | | | | | | | | | | | | | | | | | | | | | | | | | | | 5 | Chair |
| Rot. Pert. Only | t000r005 | | | | | | | | | | | | | | | | | | | | | | | | | | | 6 | Classroom | |
| | t000r010 | | | | | | | | | | | | | | | | | | | | | | | | | | | 7 | Deer | |
| | t000r020 | | | | | | | | | | | | | | | | | | | | | | | | | | | 8 | Halloween | |
| | t000r030 | | | | | | | | | | | | | | | | | | | | | | | | | | | 9 | Jugs | |
| | t000r040 | | | | | | | | | | | | | | | | | | | | | | | | | | | 10 | Root | |
| | t010r000 | | | | | | | | | | | | | | | | | | | | | | | | | | | 11 | Roundtable | |
| Tran. Pert. Only | t020r000 | | | | | | | | | | | | | | | | | | | | | | | | | | 12 | Stone | | |
| | t040r000 | | | | | | | | | | | | | | | | | | | | | | | | | | 13 | Valley | | |
| | t060r000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| | t080r000 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |

Figure 14: Per-scene camera parameter estimation results for breaking point analysis. Left: ours. Right: COLMAP. For each row (trajectory variant), the larger green area is, the better performance a method achieves in that trajectory. *Invalid* (grey) denotes that a part of a camera trajectory is out of a blender scene after perturbed, resulting in meaningless image renderings, from which it is impossible for any method to recover camera parameters or scene geometries. We set a criteria of $\pm 20^\circ$ rotation error and $\pm 0.5 f_{gt}$ focal length error to separate *successful* (green) and *failed* (red) camera parameter estimations.

Additionally, the camera parameters can also be initialised with estimated values from external toolboxes, where they are available, and jointly refined during the training of the NeRF model. We conduct experiments to refine the camera parameters estimated using COLMAP during NeRF training, and find the novel view results slightly improved through the joint refinement, as shown in Table 4.

| Scene | Focal Err. (pixel) | | Rot. Err. (deg) | | Tran. Err. | | SSIM \uparrow | | PSNR \uparrow | |
|-------------|--------------------|--------|-----------------|-------|------------|-------|-----------------|------|-----------------|-------|
| | colmap | ours | colmap | ours | colmap | ours | colmap | ours | colmap | ours |
| Airplane | 0.09 | 0.87 | 0.02 | 0.61 | 0.000 | 0.003 | 0.83 | 0.87 | 28.19 | 30.57 |
| Balls | 0.21 | 15.44 | 0.06 | 13.43 | 0.001 | 0.285 | 0.89 | 0.81 | 34.56 | 32.12 |
| Bathroom | 0.02 | 0.52 | 1.49 | 1.50 | 0.001 | 0.004 | 0.93 | 0.94 | 32.64 | 31.58 |
| Bed | 0.26 | 0.39 | 2.21 | 2.21 | 0.003 | 0.004 | 0.94 | 0.94 | 33.28 | 32.41 |
| Castle | 0.04 | 3.23 | 0.04 | 3.17 | 0.001 | 0.020 | 0.88 | 0.89 | 32.44 | 32.74 |
| Chair | 0.01 | 6.12 | 0.03 | 3.52 | 0.000 | 0.078 | 0.84 | 0.81 | 33.42 | 32.24 |
| Classroom | 0.15 | 2.29 | 0.35 | 8.14 | 0.002 | 0.032 | 0.90 | 0.86 | 27.49 | 25.14 |
| Deer | 1.93 | 6.29 | 0.42 | 6.17 | 0.015 | 0.166 | 0.99 | 0.99 | 41.44 | 42.01 |
| Halloween | 0.69 | 9.77 | 0.13 | 6.74 | 0.003 | 0.064 | 0.94 | 0.91 | 31.59 | 29.30 |
| Jugs | 0.05 | 0.16 | 0.22 | 2.30 | 0.004 | 0.065 | 0.98 | 0.99 | 40.76 | 42.53 |
| Root | 15.27 | 36.42 | 6.95 | 4.51 | 0.078 | 0.054 | 0.97 | 0.97 | 36.49 | 35.45 |
| Roundtable* | 189.64 | 206.10 | 179.10 | 9.68 | 0.069 | 0.139 | 0.99 | 0.99 | 45.13 | 39.88 |
| Stone | 0.04 | 0.11 | 0.04 | 0.12 | 0.000 | 0.001 | 0.79 | 0.86 | 29.95 | 31.74 |
| Valley | 0.02 | 0.05 | 0.04 | 0.25 | 0.000 | 0.001 | 0.73 | 0.75 | 27.53 | 27.66 |
| Mean | 14.89 | 20.55 | 13.65 | 4.45 | 0.013 | 0.065 | 0.90 | 0.90 | 33.92 | 33.24 |

Table 3: Detailed quantitative camera parameter estimation results for Section 6.2.2. Our method produces accurate camera parameter estimations and comparable NVS quality to COLMAP-NeRF ($\Delta SSIM = 0$). *Roundtable** is a special case where COLMAP estimated poses are mirrored with the ground truth. We tried 10 random seeds for COLMAP on this scene, 9/10 trials produce 179.1° rotation error and 1/10 trials produce 0.71° rotation error. Hence we report the 179.1° here. More details see the text above.

Appendix B Additional Details

B.1 Pseudo Code

We provide a pseudo code to show the key steps in our pipeline. In addition to set `require_grad` to `True`, the online ray construction step is another key that enables the direct joint optimisation via back-propagation.

ALGORITHM 1: NeRF— Pipeline

```

Input:  $N$  images  $\mathcal{I} = \{I\}_{i=1}^N$ 
Output: NeRF model  $F_{\Theta}$ , camera parameters  $[\hat{\phi}_i], [\hat{\mathbf{t}}_i]$ , and  $\hat{f}$ 

import torch.nn as nn

// Initialisation
 $[\hat{\phi}_i] = \text{nn.Parameter}(\text{shape}=(N, 3), \text{require\_grad}=\text{True})$ 
 $[\hat{\mathbf{t}}_i] = \text{nn.Parameter}(\text{shape}=(N, 3), \text{require\_grad}=\text{True})$ 
 $\hat{f} = \text{nn.Parameter}(\text{shape}=(1,), \text{require\_grad}=\text{True})$ 
// NeRF structure see our supp.
 $F_{\Theta} = \text{NeRF\_Module}(\text{require\_grad}=\text{True})$ 

// Training
for  $i$  in range ( $N$ ) do
  for  $m$  in range ( $M$ ) do
     $\hat{\mathbf{d}}_{i,m} = \text{construct\_ray}(\hat{\phi}_i, \hat{\mathbf{t}}_i, \hat{f}, \mathbf{p}_{i,m})$  // Eq. 4
    for  $h$  from  $h_n$  to  $h_f$  do
       $\mathbf{x}_j = \text{sample\_point}(\hat{\mathbf{d}}_{i,m}, \hat{\mathbf{t}}_i, h)$ 
       $\mathbf{c}_h, \sigma_j = F_{\Theta}(\mathbf{x}_h, \hat{\mathbf{d}}_{i,m})$  // forward NeRF
    end
     $\hat{I}_{i,m} = \text{render\_ray}([\mathbf{c}_h], [\sigma_h])$ 
  end
   $L = \text{loss}(\hat{I}_i, I_i)$  // Eq. 3
  L.backward()
  optimiser.update( $[\hat{\phi}_i], [\hat{\mathbf{t}}_i], \hat{f}, \hat{\Theta}$ )
end

```

B.2 NeRF Architecture

We employ a smaller NeRF [4] network than the original NeRF paper proposed without a hierarchical structure. Specifically, our NeRF implementation shrinks all hidden layer dimensions by half and follows the same positional encoding and skip connections as implemented in the original NeRF. The network architecture is presented in Fig. 15.

B.3 Training Details

Apart from the implementation details we mentioned in Section 6.1, we report more training details here. Specifically, a BLEFF scene with 27 training images can be trained in 5.5 hours on a single

| Scene | SSIM \uparrow | | | | LPIPS \downarrow | | | | PSNR \uparrow | | | |
|----------|-----------------|------|------------|----------|--------------------|------|------------|----------|-----------------|-------|------------|----------|
| | colmap | ours | colmap + r | ours + r | colmap | ours | colmap + r | ours + r | colmap | ours | colmap + r | ours + r |
| Fern | 0.64 | 0.61 | 0.64 | 0.63 | 0.47 | 0.50 | 0.47 | 0.48 | 22.22 | 21.67 | 22.28 | 21.93 |
| Flower | 0.71 | 0.71 | 0.72 | 0.72 | 0.36 | 0.37 | 0.36 | 0.35 | 25.25 | 25.34 | 25.53 | 25.58 |
| Fortress | 0.73 | 0.63 | 0.72 | 0.71 | 0.38 | 0.49 | 0.38 | 0.38 | 27.60 | 26.20 | 27.13 | 27.37 |
| Horns | 0.68 | 0.61 | 0.68 | 0.64 | 0.44 | 0.50 | 0.44 | 0.48 | 24.25 | 22.53 | 24.44 | 23.18 |
| Leaves | 0.52 | 0.53 | 0.53 | 0.53 | 0.47 | 0.47 | 0.47 | 0.46 | 18.81 | 18.88 | 18.94 | 18.96 |
| Orchids | 0.51 | 0.39 | 0.52 | 0.41 | 0.46 | 0.55 | 0.45 | 0.53 | 19.09 | 16.73 | 19.31 | 17.03 |
| Room | 0.87 | 0.84 | 0.87 | 0.85 | 0.40 | 0.44 | 0.39 | 0.42 | 27.77 | 25.84 | 27.96 | 26.28 |
| Trex | 0.74 | 0.72 | 0.74 | 0.73 | 0.41 | 0.44 | 0.41 | 0.43 | 23.19 | 22.67 | 23.13 | 22.95 |
| Mean | 0.68 | 0.63 | 0.68 | 0.65 | 0.42 | 0.47 | 0.42 | 0.44 | 23.52 | 22.48 | 23.59 | 22.91 |

Table 4: An optional camera parameter refinement (denote by ‘+r’) can further improve NVS quality. During the refinement, camera parameters are first initialised from COLMAP estimations or from previous joint optimisations, and jointly optimised with a randomly initialised NeRF model.

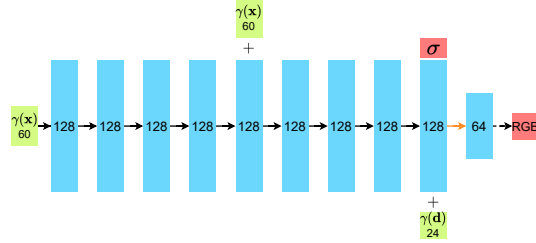


Figure 15: The NeRF implementation used in our paper. Green: position-encoded input. Blue: MLP hidden layers, with channel dimension shown inside. Red: radiance and density outputs. + denotes feature concatenation. Solid black arrow denotes layers with ReLU activation. Dashed black arrow denotes layers with Sigmoid activation. Orange arrow denotes layers without activation. We shrink all hidden layer dimensions in the original implementation by half, i.e. from 256 to 128 for the first 9 layers, and from 128 to 64 for the last layer. We employ the same number of positional encoding frequencies and the same skip links as in original NeRF. The figure style is borrowed from the original NeRF paper so readers can easily make comparisons.

1080Ti GPU. This training time is slightly longer than the training time (5 hours) of the original NeRF with our simplifications. The extra half hour training time mainly comes from the online ray direction construction from the learned camera poses and focal length.

B.4 Evaluation Details

Since our optimised camera parameters might lie in different spaces from the ones estimated using COLMAP, for evaluation, we first align the two trajectories globally with a Sim(3) transformation using Umeyama algorithm[47] in an ATE toolbox [43], followed by a more fine-grained gradient-driven camera pose alignment by minimising the photometric error on the synthesised image, while keeping the NeRF model *fixed*. Finally, we compute the metrics between the test image and our synthesised image rendered from the best possible viewpoint. Simply put, all the above mentioned processing aims to eliminate the effect from camera mis-alignment and make a fair comparison on quality of the 3D scene representation.

B.5 Focal Length Parameterisation

As mentioned in Section 6.1, we initialise our focal length f to be input image width W . In this section, we discuss the focal length parameterisation details, and compare three types of parameterisations.

In short, we choose to parameterise focal length as a single focal length f with a so-called 2nd-order trick. The details are elaborated in the following subsections.

B.5.1 f vs. $f_x f_y$

We choose to optimise a focal length f instead of a horizontal f_x and vertical f_y , as we found no performance difference between these choices (Table 5).

B.5.2 1st-order vs. 2nd-order

A naive way to optimise focal length f is to initialise it with image width W directly, but this poses difficulties to the direct optimisation as W is usually in a much larger numerical magnitude than other learnable parameters.

A better way to parameterise f is through a scale factor s :

$$f = sW, \quad (5)$$

and initialise with $s = 1.0$. By parameterising focal length with s , the network avoids optimising f in pixel unit directly, whose value are often large and pose numerical difficulties in optimisation.

In practice, we found that optimising the square root of s , denoted by \tilde{s} , leads to slightly better results, *i.e.*,

$$f = \tilde{s}^2 W, \quad (6)$$

| Scene | PSNR \uparrow | | SSIM \uparrow | | LPIPS \downarrow | |
|----------|-----------------|-----------|-----------------|-----------|--------------------|-----------|
| | f | $f_x f_y$ | f | $f_x f_y$ | f | $f_x f_y$ |
| Fern | 21.67 | 21.79 | 0.61 | 0.62 | 0.50 | 0.50 |
| Flower | 25.34 | 25.24 | 0.71 | 0.71 | 0.37 | 0.37 |
| Fortress | 26.20 | 26.24 | 0.63 | 0.65 | 0.49 | 0.46 |
| Horns | 22.53 | 23.08 | 0.61 | 0.63 | 0.50 | 0.50 |
| Leaves | 18.88 | 18.79 | 0.53 | 0.52 | 0.47 | 0.47 |
| Orchids | 16.73 | 16.48 | 0.39 | 0.37 | 0.55 | 0.56 |
| Room | 25.84 | 25.72 | 0.84 | 0.84 | 0.44 | 0.44 |
| Trex | 22.67 | 22.53 | 0.72 | 0.72 | 0.44 | 0.44 |
| Mean | 22.48 | 22.48 | 0.63 | 0.63 | 0.47 | 0.47 |

Table 5: NVS quality on LLFF-NeRF dataset when using f and $f_x f_y$. We choose to optimise a single focal length f for simplicity, as no performance gain acquired by optimising two focal lengths separately.

where \tilde{s} is initialised to 1.0 too. We refer this parameterisation as the 2nd-order trick. Table 6 shows a quantitative comparison of the novel view rendering quality using these two parameterisations.

| Scene | PSNR \uparrow | | SSIM \uparrow | | LPIPS \downarrow | |
|----------|-----------------|--------------|-----------------|-------------|--------------------|------|
| | \tilde{s}^2 | s | \tilde{s}^2 | s | \tilde{s}^2 | s |
| Fern | 21.67 | 21.48 | 0.61 | 0.60 | 0.50 | 0.53 |
| Flower | 25.34 | 25.44 | 0.71 | 0.71 | 0.37 | 0.37 |
| Fortress | 26.20 | 24.92 | 0.63 | 0.58 | 0.49 | 0.58 |
| Horns | 22.53 | 22.53 | 0.61 | 0.61 | 0.50 | 0.50 |
| Leaves | 18.88 | 18.85 | 0.53 | 0.52 | 0.47 | 0.47 |
| Orchids | 16.73 | 16.67 | 0.39 | 0.39 | 0.55 | 0.55 |
| Room | 25.84 | 25.74 | 0.84 | 0.83 | 0.44 | 0.44 |
| Trex | 22.67 | 22.45 | 0.72 | 0.71 | 0.44 | 0.44 |
| Mean | 22.48 | 22.26 | 0.63 | 0.62 | 0.47 | 0.49 |

Table 6: Quantitative comparison (PSNR on novel views) of two different focal length parameterisations on LLFF-NeRF dataset. The 2nd-order parameterisation \tilde{s}^2 produces slightly better results than the 1st-order s .

B.6 RealEstate10K Details

We use two sequences from RealEstate10K [8] in Section 6.2.1 and Section 6.3.2. The details of these videos and frame rate down-sampling procedures are listed in Table 7.

| YouTube video ID | Original fps | Original res. | Training fps | Training res. |
|------------------|--------------|---------------|--------------|---------------|
| MVVJodQ50HQ | 30 | 1920x1080 | 5 | 1920x1080 |
| OT04jHhqYyw | 30 | 1920x1080 | 5 | 1920x1080 |

Table 7: Details for selected sequences in RealEstate10K.

B.7 BLEFF Asset Licences

Our BLEFF dataset is made of 14 blender scenes, which we modified and downloaded from <https://www.blendswap.com>. The detailed licence info is listed in Table 8.

| Scene | Licence | Author | Link |
|------------------|--------------|---------------|---|
| Airplane | CC-BY-3.0 | fabien | https://blendswap.com/blend/15016 |
| Balls-basketball | CC-BY-3.0 | CGMasters | https://blendswap.com/blend/14758 |
| Balls-volleyball | CC-BY-3.0 | Shri | https://blendswap.com/blend/16958 |
| Bathroom | CC-BY-SA-3.0 | ORBANGeoffrey | https://blendswap.com/blend/20604 |
| Bed | CC-BY-SA-3.0 | Warcos | https://blendswap.com/blend/20975 |
| Castle | CC-BY-0 | BySky | https://blendswap.com/blend/27390 |

| | | | |
|------------|-----------|---------------|---|
| Chair | CC-BY-0 | akashdlfps | https://blendswap.com/blend/25057 |
| Classroom | CC-BY-3.0 | SwastikDas | https://blendswap.com/blend/21410 |
| Deer | CC-BY-0 | Spine69 | https://blendswap.com/blend/26863 |
| Halloween | CC-BY-0 | Yash Deokar | https://blendswap.com/blend/26793 |
| Jugs | CC-BY-0 | BigBadCat | https://blendswap.com/blend/23234 |
| Root | CC-BY-0 | AceTop | https://blendswap.com/blend/24472 |
| Roundtable | CC-BY-0 | gandre82 | https://blendswap.com/blend/26159 |
| Stone | CC-BY-0 | rajatg8008 | https://blendswap.com/blend/22490 |
| Valley | CC-BY-0 | ShadowCrystal | https://blendswap.com/blend/27325 |

Table 8: BLEFF licence info.

Appendix C Broader Impact

Joint Optimisation Method Our method focuses on simplifying the classical two-stage NVS systems, by jointly optimising a NeRF model and camera parameters for forward-facing images. We expect our method to be applied in a wide ranged of applications, for example, AR and VR content creation, 3D modelling, camera pose estimation and etc. As this method is designed to be trained on a per-scene basis and is designed for static scenes and cannot handle dynamic objects such as animals or humans, we expect the potential negative social impact (*i.e.*, bias in models and human privacy) that can be introduced by our method is very limited, as long as the image capturing procedure confronts local legal requirements.

BLEFF Dataset We also introduce a synthetic dataset BLEFF, which contains high-quality path-traced images rendered from 3D Blender models. The BLEFF scenes are selected to cover various common real-life subjects, such as furniture, decorations, toys, rooms, or outdoor landscapes and etc. There are two main purposes for this dataset: 1) to evaluate pose estimation and NVS rendering quality as the same time; and 2) to evaluate the robustness of a joint optimisation system similar to ours. Therefore, we expect our dataset to be beneficial to other joint optimisation systems (NVS and camera parameters) in the future.

References

- [1] Edward H. Adelson and James R. Bergen. The plenoptic function and the elements of early vision. In *Computational Models of Visual Processing*, 1991. 1
- [2] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. In *SIGGRAPH*, 1996. 1, 2, 3
- [3] Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH*, 1996. 1, 2
- [4] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 4, 12
- [5] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *arXiv preprint arXiv:2103.13415*, 2021. 1, 2
- [6] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. *arXiv preprint arXiv:2008.02268*, 2020. 1, 2
- [7] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 1, 2
- [8] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018. 1, 2, 5, 14
- [9] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *CVPR*, 2020. 1, 2
- [10] 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. *ToG*, 2019. 1, 2, 4, 5

- [11] Inchang Choi, Orazio Gallo, Alejandro Troccoli, Min H Kim, and Jan Kautz. Extreme view synthesis. In *CVPR*, 2019. 1, 2
- [12] 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 *CVPR*, 2019. 1, 2, 5
- [13] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *CVPR*, 2021. 1, 2, 5
- [14] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 1, 2, 3, 5
- [15] Q-T Luong and Olivier D Faugeras. A stability analysis of the fundamental matrix. In *ECCV*, 1994. 2, 9
- [16] Philip HS Torr, Andrew Zisserman, and Stephen J Maybank. Robust detection of degenerate configurations while estimating the fundamental matrix. *Computer Vision and Image Understanding*, 1998. 2, 9
- [17] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. 2003. 2, 9
- [18] Zhenglong Zhou, Zhe Wu, and Ping Tan. Multi-view photometric stereo with spatially varying isotropic materials. In *CVPR*, 2013. 2
- [19] Gernot Riegler and Vladlen Koltun. Free view synthesis. In *ECCV*, 2020. 2
- [20] Gernot Riegler and Vladlen Koltun. Stable view synthesis. *arXiv preprint arXiv:2011.07233*, 2020. 2
- [21] Eric Penner and Li Zhang. Soft 3d reconstruction for view synthesis. In *SIGGRAPH Asia*, 2017. 2
- [22] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2
- [23] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2
- [24] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020. 2
- [25] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. *TPAMI*, 2007. 2, 4
- [26] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. Orb-slam: a versatile and accurate monocular slam system. *TRO*, 2015. 2, 5
- [27] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *ICCV*, 2011. 2, 4
- [28] Christian Forster, Matia Pizzoli, and Davide Scaramuzza. Svo: Fast semi-direct monocular visual odometry. In *ICRA*, 2014. 2
- [29] Jakob Engel, Thomas Schöps, and Daniel Cremers. Lsd-slam: Large-scale direct monocular slam. In *ECCV*, 2014. 2, 4
- [30] Noah Snavely, Steven M Seitz, and Richard Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH*, 2006. 2
- [31] Lin Yen-Chen, Pete Florence, Jonathan T. Barron, Alberto Rodriguez, Phillip Isola, and Tsung-Yi Lin. inerf: Inverting neural radiance fields for pose estimation. *arXiv preprint arXiv:2012.05877*, 2020. 3
- [32] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. *arXiv preprint arXiv:2103.15606*, 2021. 3
- [33] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405*, 2021. 3
- [34] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1995. 3
- [35] Bill Triggs, Philip F. McLauchlan, Richard I. Hartley, and Andrew W. Fitzgibbon. Bundle adjustment — a modern synthesis. In *Vision Algorithms: Theory and Practice*, 2000. 4
- [36] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ToG*, 2017. 5
- [37] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *IROS*, 2012. 5, 6
- [38] Thomas Schops, Torsten Sattler, and Marc Pollefeys. Bad slam: Bundle adjusted direct rgb-d slam. In *CVPR*, 2019. 5
- [39] John McCormac, Ankur Handa, Stefan Leutenegger, and Andrew J Davison. Scenenet rgb-d: 5m photorealistic images of synthetic indoor trajectories with ground truth. In *ICCV*, 2016. 5

- [40] Wenbin Li, Sajad Saeedi, John McCormac, Ronald Clark, Dimos Tzoumanikas, Qing Ye, Yuzhong Huang, Rui Tang, and Stefan Leutenegger. Interiornet: Mega-scale multi-sensor photo-realistic indoor scenes dataset. In *BMVC*, 2018. [5](#)
- [41] 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*, 2004. [6](#)
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. [6](#)
- [43] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual(-inertial) odometry. In *IROS*, 2018. [6](#), [7](#), [13](#)
- [44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*, 2015. [6](#)
- [45] Denis Oberkampf, Daniel F DeMenthon, and Larry S Davis. Iterative pose estimation using coplanar feature points. *Computer Vision and Image Understanding*, 1996. [10](#)
- [46] Gerald Schweighofer and Axel Pinz. Robust pose estimation from a planar target. *TPAMI*, 2006. [10](#)
- [47] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 1991. [13](#)