NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior

Wenjing Bian Zirui Wang Kejie Li Jia-Wang Bian
Victor Adrian Prisacariu
Active Vision Lab, University of Oxford
{wenjing, ryan, kejie, jiawang, victor}@robots.ox.ac.uk

Figure 1. Novel view synthesis comparison. We propose NoPe-NeRF for joint pose estimation and novel view synthesis. Our method enables more robust pose estimation and renders better novel view synthesis than previous state-of-the-art methods.

Abstract

Training a Neural Radiance Field (NeRF) without pre-computed camera poses is challenging. Recent advances in this direction demonstrate the possibility of jointly optimising a NeRF and camera poses in forward-facing scenes. However, these methods still face difficulties during dramatic camera movement. We tackle this challenging problem by incorporating undistorted monocular depth priors. These priors are generated by correcting scale and shift parameters during training, with which we are then able to constrain the relative poses between consecutive frames. This constraint is achieved using our proposed novel loss functions. Experiments on real-world indoor and outdoor scenes show that our method can handle challenging camera trajectories and outperforms existing methods in terms of novel view rendering quality and pose estimation accuracy. Our project page is https://nope-nerf.active.vision.

1. Introduction

The photo-realistic reconstruction of a scene from a stream of RGB images requires both accurate 3D geometry reconstruction and view-dependent appearance modelling. Recently, Neural Radiance Fields (NeRF) [24] have demonstrated the ability to build high-quality results for generating photo-realistic images from novel viewpoints given a sparse set of images.

An important preparation step for NeRF training is the estimation of camera parameters for the input images. A current go-to option is the popular Structure-from-Motion (SfM) library COLMAP [35]. Whilst easy to use, this pre-processing step could be an obstacle to NeRF research and real-world deployments in the long term due to its long processing time and its lack of differentiability. Recent works such as NeRFmm [46], BARF [18] and SC-NeRF [12] propose to simultaneously optimise camera poses and the neural implicit representation to address these issues. Nevertheless, these methods can only handle forward-facing scenes when no initial parameters are supplied, and fail in dramatic camera motions, e.g., a casual handheld captured video.

This limitation has two key causes. First, all these methods estimate a camera pose for each input image individually without considering relative poses between images. Looking back to the literature of Simultaneous localisation and mapping (SLAM) and visual odometry, pose estimation can significantly benefit from estimating relative poses be-
between adjacent input frames. Second, the radiance field is known to suffer from shape-radiance ambiguity [55]. Estimating camera parameters jointly with NeRF adds another degree of ambiguity, resulting in slow convergence and unstable optimisation.

To handle the limitation of large camera motion, we seek help from monocular depth estimation [22, 28, 29, 51]. Our motivation is threefold: First, monocular depth provides strong geometry cues that are beneficial to constraint shape-radiance ambiguity. Second, relative poses between adjacent depth maps can be easily injected into the training pipeline via Chamfer Distance. Third, monocular depth is lightweight to run and does not require camera parameters as input, in contrast to multi-view stereo depth estimation. For simplicity, we use the term mono-depth from now on.

Utilising mono-depth effectively is not straightforward with the presence of scale and shift distortions. In other words, mono-depth maps are not multi-view consistent. Previous works [9, 17, 47] simply take mono-depth into a depth-wise loss along with NeRF training. Instead, we propose a novel and effective way to thoroughly integrate mono-depth into our system. First, we explicitly optimise scale and shift parameters for each mono-depth map during NeRF training by penalising the difference between rendered depth and mono-depth. Since NeRF by itself is trained based on multiview consistency, this step transforms mono-depth maps to undistorted multiview consistent depth maps. We further leverage these multiview consistent depth maps in two loss terms: a) a Chamfer Distance loss between two depth maps of adjacent images, which injects relative pose to our system; and b) a depth-based surface rendering loss, which further improves relative pose estimation.

In summary, we propose a method to jointly optimise camera poses and a NeRF from a sequence of images with large camera motion. Our system is enabled by three contributions. First, we propose a novel way to integrate mono-depth into unposed-NeRF training by explicitly modelling scale and shift distortions. Second, we supply relative poses to the camera-NeRF joint optimisation via an inter-frame loss using undistorted mono-depth maps. Third, we further regularise our relative pose estimation with a depth-based surface rendering loss.

As a result, our method is able to handle large camera motion, and outperforms state-of-the-art methods by a significant margin in terms of novel view synthesis quality and camera trajectory accuracy.

2. Related Work

Novel View Synthesis. While early Novel View Synthesis (NVS) approaches applied interpolations between pixels [3], later works often rendered images from 3D reconstructions [1, 6]. In recent years, different representations of the 3D scene are used, e.g., meshes [30, 31]. Multi-Plane Images [41, 59], layered depth [42] etc. Among them, NeRF [24] has become a popular scene representation for its photorealistic rendering.

A number of techniques are proposed to improve NeRF’s performance with additional regularisation [13, 26, 54], depth priors [7, 32, 47, 53], surface enhancements [27, 43, 49] or latent codes [40, 44, 52]. Other works [2, 10, 25, 34] have also accelerated NeRF training and rendering. However, most of these approaches require pre-computed camera parameters obtained from SfM algorithms [11, 35].

NeRF With Pose Optimisation. Removing camera parameter preprocessing is an active line of research recently. One category of the methods [33, 38, 60] use a SLAM-style pipeline, that either requires RGB-D inputs or relies on accurate camera poses generated from the SLAM tracking system. Another category of works optimises camera poses with the NeRF model directly. We term this type of method as unposed-NeRF in this paper. iNeRF [50] shows that poses for novel view images can be estimated using a reconstructed NeRF model. GNeRF [21] combines Generative Adversarial Networks with NeRF to estimate camera poses but requires a known sampling distribution for poses. More relevant to our work, NeRFmm [46] jointly optimises both camera intrinsics and extrinsics alongside NeRF training. BARF [18] proposes a coarse-to-fine positional encoding strategy for camera poses and NeRF joint optimisation. SC-NeRF [12] further parameterises camera distortion and employs a geometric loss to regularise rays. GARF [4] shows that using Gaussian-MLPs makes joint pose and scene optimisation easier and more accurate. Recently, SiNeRF [48] uses SIREN [36] layers and a novel sampling strategy to alleviate the sub-optimality of joint optimisation in NeRFmm. Although showing promising results on the forward-facing dataset like LLFF [23], these approaches face difficulties when handling challenging camera trajectories with large camera motion. We address this issue by closely integrating mono-depth maps with the joint optimisation of camera parameters and NeRF.

3. Method

We tackle the challenge of handling large camera motions in unposed-NeRF training. Given a sequence of images, camera intrinsics, and their mono-depth estimations, our method recovers camera poses and optimises a NeRF simultaneously. We assume camera intrinsics are available in the image meta block, and we run an off-the-shelf mono-depth network DPT [7] to acquire mono-depth estimations. Without repeating the benefit of mono-depth, we unroll this section around the effective integration of monocular depth into unposed-NeRF training.

The training is a joint optimisation of the NeRF, camera poses, and distortion parameters of each mono-depth
map. The distortion parameters are supervised by minimising the discrepancies between the mono-depth maps and depth maps rendered from the NeRF, which are multiview consistent. The undistorted depth maps in return effectively mediate the shape-radiance ambiguity, which eases the training of NeRF and camera poses.

Specifically, the undistorted depth maps enable two constraints. We constrain global pose estimation by supplying relative pose between adjacent images. This is achieved via a Chamfer-Distance-based correspondence between two point clouds, back-projected from undistorted depth maps. Further, we regularise relative pose estimation with a surface-based photometric consistency where we treat undistorted depth as surface.

We detail our method in the following sections, starting from NeRF in Sec. 3.1 and unposed-NeRF training in Sec. 3.2, looking into mono-depth distortions in Sec. 3.3, followed by our mono-depth enabled loss terms in Sec. 3.4, and finishing with an overall training pipeline Sec. 3.5.

3.1. NeRF

Neural Radiance Field (NeRF) [24] represents a scene as a mapping function \( F_\Theta : (x, d) \rightarrow (c, \sigma) \) that maps a 3D location \( x \in \mathbb{R}^3 \) and a viewing direction \( d \in \mathbb{R}^3 \) to a radiance colour \( c \in \mathbb{R}^3 \) and a volume density value \( \sigma \). This mapping is usually implemented with a neural network parameterised by \( F_\Theta \). Given \( N \) images \( \mathcal{I} = \{ I_i \mid i = 0 \ldots N - 1 \} \) with their camera poses \( \Pi = \{ \pi_i \mid i = 0 \ldots N - 1 \} \), NeRF can be optimised by minimising photometric error

\[
\mathcal{L}_{\text{rgb}} = \sum_i^N \| I_i - \hat{I}_i \|^2 \quad \text{between synthesised images } \hat{I} \text{ and captured images } I:
\]

\[
\Theta^* = \arg \min_{\Theta} \mathcal{L}_{\text{rgb}}(\hat{I} \mid \mathcal{I}, \Pi),
\]

where \( \hat{I}_i \) is rendered by aggregating radiance colour on camera rays \( r(h) = o + hd \) between near and far bound \( h_n \) and \( h_f \). More concretely, we synthesise \( \hat{I}_i \) with a volumetric rendering function

\[
\hat{I}_i(r) = \int_{h_n}^{h_f} T(h)\sigma(r(h))c(r(h), d)dh,
\]

where \( T(h) = \exp(-\int_{h_n}^{h_f} \sigma(r(s))ds) \) is the accumulated transmittance along a ray. We refer to [24] for further details.

3.2. Joint Optimisation of Poses and NeRF

Prior works [12, 18, 46] show that it is possible to estimate both camera parameters and a NeRF at the same time by minimising the above photometric error \( \mathcal{L}_{\text{rgb}} \) under the same volumetric rendering process in Eq. (2).

The key lies in conditioning camera ray casting on variable camera parameters \( \Pi \), as the camera ray \( r \) is a function of camera pose. Mathematically, this joint optimisation can be formulated as:

\[
\Theta^*, \Pi^* = \arg \min_{\Theta, \Pi} \mathcal{L}_{\text{rgb}}(\hat{I}, \Pi \mid \mathcal{I}),
\]

where \( \Pi \) denotes camera parameters that are updated during optimising. Note that the only difference between Eq. (1)
and Eq. (3) is that Eq. (3) considers camera parameters as variables.

In general, the camera parameters $\Pi$ include camera intrinsics, poses, and lens distortions. We only consider estimating camera poses in this work, e.g., camera pose for frame $I_i$ is a transformation $T_i = [R_i \mid t_i]$ with a rotation $R_i \in \text{SO}(3)$ and a translation $t_i \in \mathbb{R}^3$.

3.3. Undistortion of Monocular Depth

With an off-the-shelf monocular depth network, e.g., DPT [28], we generate mono-depth sequence $D = \{D_i \mid i = 0 \ldots N - 1\}$ from input images. Without surprise, mono-depth maps are not multi-view consistent so we aim to recover a sequence of multi-view consistent depth maps, which are further leveraged in our relative pose loss terms.

Specifically, we consider two linear transformation parameters for each mono-depth map, resulting in a sequence of transformation parameters for all frames $\Psi = \{(\alpha_i, \beta_i) \mid i = 0 \ldots N - 1\}$, where $\alpha_i$ and $\beta_i$ denote a scale and a shift factor. With multi-view consistent constraint from NeRF, we aim to recover a multi-view consistent depth map $D^*_i$ for $D_i$:

$$D^*_i = \alpha_i D_i + \beta_i,$$

by joint optimising $\alpha_i$ and $\beta_i$ along with a NeRF. This joint optimisation is mostly achieved by enforcing the consistency between an undistorted depth map $D^*_i$ and a NeRF rendered depth map $\hat{D}_i$ via a depth loss:

$$L_{\text{depth}} = \sum_i \left\| D^*_i - \hat{D}_i \right\|,$$  

where

$$\hat{D}_i(r) = \int_{h_i}^{h_j} T(h) \sigma(r(h)) dh.$$

denotes a volumetric rendered depth map from NeRF.

It is important to note that both NeRF and mono-depth benefit from Eq. (5). On the one hand, mono-depth provides strong geometry prior for NeRF training, reducing shape-radiance ambiguity. On the other hand, NeRF provides multi-view consistency so we can recover a set of multi-view consistent depth maps for relative pose estimations.

3.4. Relative Pose Constraint

Aforementioned unposed-NeRF methods [12, 18, 46] optimise each camera pose independently, resulting in an overfit to target images with incorrect poses. Penalising incorrect relative poses between frames can help to regularise the joint optimisation towards smooth convergence, especially in a complex camera trajectory. Therefore, we propose two losses that constrain relative poses.

Point Cloud Loss. We back-project the undistorted depth maps $D^*$ using the known camera intrinsics, to point clouds $P^* = \{P^*_i \mid i = 0 \ldots N - 1\}$ and optimise the relative pose between consecutive point clouds by minimising a point cloud loss $L_{\text{pc}}$:

$$L_{\text{pc}} = \sum_{(i,j)} l_{cd}(P^*_i, T_{ji} P^*_j),$$

where $T_{ji} = T_j T_i^{-1}$ represents the related pose that transforms point cloud $P^*_i$ to $P^*_j$, tuple $(i, j)$ denotes indices of a consecutive pair of instances, and $l_{cd}$ denotes Chamfer Distance:

$$l_{cd}(P^*_i, P^*_j) = \sum_{p_i \in P^*_i} \min_{p_j \in P^*_j} \| p_i - p_j \|_2 + \sum_{p_j \in P^*_j} \min_{p_i \in P^*_i} \| p_i - p_j \|_2.$$

Surface-based Photometric Loss. While the point cloud loss $L_{\text{pc}}$ offers supervision in terms of 3D-3D matching, we observe that a surface-based photometric error can alleviate incorrect matching. With the photometric consistency assumption, this photometric error penalises the differences in appearance between associated pixels. The association is established by projecting the point cloud $P^*_i$ onto images $I_i$ and $I_j$.

The surface-based photometric loss can then be defined as:

$$L_{\text{rgb-s}} = \sum_{(i,j)} \left\| I_i(\hat{K}_i P^*_i) - I_j(\hat{K}_j T_j T_i^{-1} P^*_i) \right\|,$$

where $\langle \cdot \rangle$ denotes the sampling operation on the image and $\hat{K}_i$ denotes a projection matrix for $i$th camera.

3.5. Overall Training Pipeline

Assembling all loss terms, we get the overall loss function:

$$L = L_{\text{rgb}} + \lambda_1 L_{\text{depth}} + \lambda_2 L_{\text{pc}} + \lambda_3 L_{\text{rgb-s}},$$

where $\lambda_1, \lambda_2, \lambda_3$ are the weighting factors for respective loss terms. By minimising the combined of loss $L$:

$$\Theta^*, \Pi^*, \Psi^* = \arg \min_{\Theta, \Pi, \Psi} L(\hat{\Theta}, \hat{\Pi}, \hat{\Psi} \mid \mathcal{I}, \mathcal{D}),$$

our method returns the optimised NeRF parameters $\Theta$, camera poses $\Pi$, and distortion parameters $\Psi$.

4. Experiments

We begin with a description of our experimental setup in Sec. 4.1. In Sec. 4.2, we compare our method with pose-unknown methods. Next, we compare our method with the COLMAP-assisted NeRF baseline in Sec. 4.3. Lastly, we conduct ablation studies in Sec. 4.4.
4.1. Experimental Setup

Datasets. We conduct experiments on two datasets Tanks and Temples [15] and ScanNet [5]. Tanks and Temples: we use 8 scenes to evaluate pose accuracy and novel view synthesis quality. We chose scenes captured at both indoor and outdoor locations, with different frame sampling rates and lengths. All images are down-sampled to a resolution of $960 \times 540$. For the family scene, we sample 200 images and take 100 frames with odd frame ids as training images and the remaining 100 frames for novel view synthesis, in order to analyse the performance under smooth motion. For the remaining scenes, following NeRF [24], 1/8 of the images in each sequence are held out for novel view synthesis, unless otherwise specified. ScanNet: we select 4 scenes for evaluating pose accuracy, depth accuracy, and novel view synthesis quality. For each scene, we take 80-100 consecutive images and use 1/8 of these images for novel view synthesis. For evaluation, we employ depth maps and poses provided by ScanNet as ground truth. ScanNet images are down-sampled to $648 \times 484$. We crop images with dark orders during preprocessing.

Metrics. We evaluate our proposed method in three aspects. For novel view synthesis, we follow previous methods [12,18,46], and use standard evaluation metrics, including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [45] and Learned Perceptual Image Patch Similarity (LPIPS) [56]. For pose evaluation, We use standard visual odometry metrics [16,37,57], including the Absolute Trajectory Error (ATE) and Relative Pose Error (RPE). ATE measures the difference between the estimated camera positions and the ground truth positions. RPE measures the relative pose errors between pairs of images, which consists of relative rotation error (RPE$_r$) and relative translation error (RPE$_t$). The estimated trajectory is aligned with the ground truth using Sim(3) with 7 degrees of freedom. We use standard depth metrics [8,19,20,39] (Abs Rel, Sq Rel, RMSE, RMSE log, $\delta_1$, $\delta_2$ and $\delta_3$) for depth evaluation. For further detail, please refer to the supplementary material. To recover the metric scale, we follow Zhou et al. [58] and match the median value between rendered and ground truth depth maps.

 Implementation Details. Our model architecture is based on NeRF [24] with a few modifications: a) replacing ReLU activation function with Softplus and b) sampling 128 points along each ray uniformly with noise, between a predefined range (0.1, 10). We use 2 separate Adam optimisers [14] for NeRF and other parameters. The initial learning rate for NeRF is 0.001 and for the pose and distortion is 0.0005. Camera rotations are optimised in axis-angle representation $\phi_i \in \text{so}(3)$. We first train the model with all losses with constant learning rates until the inter-frame losses converge. Then, we remove the inter-frame losses and depth loss to refine the model with the RGB loss only. We decay the learning rates with different schedulers to refine for 10,000 epochs. We balance the loss terms with $\lambda_1 = 0.04$, $\lambda_2 = 1.0$ and $\lambda_3 = 1.0$. For each training step, we randomly sample 1024 pixels (rays) from each input image and 128 samples per ray. More details are provided in the supplementary material.

4.2. Comparing With Pose-Unknown Methods

We compare our method with pose-unknown baselines, including BARF [18], NeRFmm [46] and SC-NeRF [12].

View Synthesis Quality. To obtain the camera poses of
We visualise the trajectory (3D plot) and relative rotation errors RPE \( r \) (bottom colour bar) of each method on Ballroom and Museum. The colour bar on the right shows the relative scaling of colour. More results are in the supplementary.

Table 1. Novel view synthesis results on ScanNet and Tanks and Temples. Each baseline method is trained with its public code under the original settings and evaluated with the same evaluation protocol.

<table>
<thead>
<tr>
<th>Scenarios</th>
<th>Ours</th>
<th>BARF</th>
<th>NeRFmm</th>
<th>SC-NeRF</th>
</tr>
</thead>
<tbody>
<tr>
<td>ScanNet</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0079_00</td>
<td>32.47</td>
<td>0.84</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>0418_00</td>
<td>31.33</td>
<td>0.79</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>0301_00</td>
<td>29.83</td>
<td>0.77</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>0431_00</td>
<td>33.83</td>
<td>0.91</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>31.86</td>
<td>0.83</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>Tanks and Temples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Church</td>
<td>25.17</td>
<td>0.73</td>
<td>0.39</td>
<td></td>
</tr>
<tr>
<td>Barn</td>
<td>26.35</td>
<td>0.69</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>Museum</td>
<td>24.77</td>
<td>0.76</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Family</td>
<td>26.01</td>
<td>0.74</td>
<td>0.41</td>
<td></td>
</tr>
<tr>
<td>Horse</td>
<td>27.64</td>
<td>0.84</td>
<td>0.26</td>
<td></td>
</tr>
<tr>
<td>Ballroom</td>
<td>25.33</td>
<td>0.72</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>Francis</td>
<td>29.48</td>
<td>0.80</td>
<td>0.38</td>
<td></td>
</tr>
<tr>
<td>Ignatius</td>
<td>23.96</td>
<td>0.61</td>
<td>0.47</td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>26.34</td>
<td>0.74</td>
<td>0.39</td>
<td></td>
</tr>
</tbody>
</table>

We recognised that because the test views, which are sampled from videos, are close to the training views, good results may be obtained due to overfitting to the training images. Therefore, we conduct an additional qualitative evaluation on more novel views. Specifically, we fit a bezier curve from the estimated training poses and sample interpolated poses for each method to render novel view videos. Sampled results are shown in Fig. 5, and the rendered videos are in the supplementary material. These results show that our method renders photo-realistic images consistently, while other methods generate visible artifacts.

Camera Pose. Our method significantly outperforms other baselines in all metrics. The quantitative pose evaluation results are shown in Tab. 2. For ScanNet, we use the camera poses provided by the dataset as ground truth. For Tanks and Temples, not every video comes with ground truth poses, so we use COLMAP estimations for reference. Our estimated trajectory is better aligned with the ground truth views.
then we fix the optimised poses to COLMAP poses on ScanNet and Tanks and Temples. The original NeRF training contains two stages, finding poses using COLMAP and optimising the scene representation. In order to make our comparison fairer, in this section only, we mimic a similar two-stage training as the original NeRF [24]. In the first stage, we train our method with all losses for camera pose estimation, i.e., mimicking the COLMAP processing. Then, we fix the optimised poses and train a NeRF model from scratch, using the same settings and loss as the original NeRF. This evaluation enables us to compare our estimated poses to the COLMAP poses indirectly, i.e., in terms of contribution to view synthesis.

Our two-stage method outperforms the COLMAP-assisted NeRF baseline, which indicates a better pose estimation for novel view synthesis. The results are summarised in Tab. 5.

As is commonly known, COLMAP performs poorly in low-texture scenes and sometimes fails to find accurate camera poses. Fig. 6 shows an example of a low-texture scene where COLMAP provides inaccurate pose estimation that causes NeRF to render images with visible artifacts. The full rendered videos and details about generating novel views are provided in the supplementary.

Figure 5. Sampled frames from rendered novel view videos. For each method, we fit the learned trajectory with a bezier curve and uniformly sample new viewpoints for rendering. Our method generates significantly better results than previous methods, which show visible artifacts. The full rendered videos and details about generating novel views are provided in the supplementary.

4.3. Comparing With COLMAP Assisted NeRF

We make a comparison of pose estimation accuracy between our method and COLMAP against ground truth poses in ScanNet. We achieve on-par accuracy with COLMAP, as shown in Tab. 4. We further analyse the novel view synthesis quality of the NeRF model trained with our learned poses to COLMAP poses on ScanNet and Tanks and Temples. The original NeRF training contains two stages, finding poses using COLMAP and optimising the scene representation. In order to make our comparison fairer, in this section only, we mimic a similar two-stage training as the original NeRF [24]. In the first stage, we train our method with all losses for camera pose estimation, i.e., mimicking the COLMAP processing.
4.4. Ablation Study

In this section, we analyse the effectiveness of the parameters and components that have been added to our model. The results of ablation studies are shown in Tab. 6.

Effect of Distortion Parameters. We find that ignoring depth distortions (i.e., setting scales to 1 and shifts to 0 as constants) leads to a degradation in pose accuracy, as it introduces errors to the estimation of relative poses and confuse NeRF for geometry reconstruction.

Effect of Inter-frame Losses. We observe that the inter-frame losses are the major contributor to improving relative poses. When removing the pairwise point cloud loss \(L_{pc}\) or the surface-based photometric loss \(L_{rgb-s}\), there is less constraint between frames, and thus the pose accuracy becomes lower.

Effect of NeRF Losses. When the depth loss \(L_{depth}\) is removed, the distortions of input depth maps are only optimised locally through the inter-frame losses. We find that this can lead to drift and degradation in pose accuracy.

4.5. Limitations

Our proposed method optimises camera pose and the NeRF model jointly and works on challenging scenes where other baselines fail. However, the optimisation of the model is also affected by non-linear distortions and the accuracy of the mono-depth estimation, which we did not consider.

5. Conclusion

In this work, we present NoPe-NeRF, an end-to-end differentiable model for joint camera pose estimation and novel view synthesis from a sequence of images. We demonstrate that previous approaches have difficulty with complex trajectories. To tackle this challenge, we use mono-depth maps to constrain the relative poses between frames and regularise the geometry of NeRF, which leads to better pose estimation. We show the effectiveness and robustness of NoPe-NeRF on challenging scenes. The improved pose estimation leads to better novel view synthesis quality and geometry reconstruction compared with other approaches. We believe our method is an important step towards applying the unknown-pose NeRF models to large-scale scenes in the future.

Acknowledgements

We thank Theo Costain, Michael Hobley, Shuai Chen and Xinghui Li for their helpful proofreading and discussions. Wenjing Bian is supported by the China Scholarship Council - University of Oxford Scholarship.
References

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph.  
[42] Xuan Luo, Jia-Bin Huang, Richard Szeliski, Kevin Matzen, and Johannes Kopf. Consistent video depth estimation. ToG.  
[46] Thomas Müller, Alex Evans, ChristophSchied, and AlexanderKeller. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph.  
