Local-to-Global Registration for Bundle-Adjusting Neural Radiance Fields

Yue Chen\textsuperscript{1,2*}, Xingyu Chen\textsuperscript{1,2*}, Xuan Wang\textsuperscript{3†}, Qi Zhang\textsuperscript{4}
Yu Guo\textsuperscript{1,2†}, Ying Shan\textsuperscript{4} Fei Wang\textsuperscript{1,2}

\textsuperscript{1} National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
\textsuperscript{2} IAIR, Xi’an Jiaotong University \textsuperscript{3} Ant Group \textsuperscript{4} Tencent AI Lab

Abstract

Neural Radiance Fields (NeRF) have achieved photorealistic novel views synthesis; however, the requirement of accurate camera poses limits its application. Despite analysis-by-synthesis extensions for jointly learning neural 3D representations and registering camera frames exist, they are susceptible to suboptimal solutions if poorly initialized. We propose L2G-NeRF, a Local-to-Global registration method for bundle-adjusting Neural Radiance Fields: first, a pixel-wise flexible alignment, followed by a frame-wise constrained parametric alignment. Pixel-wise local alignment is learned in an unsupervised way via a deep network which optimizes photometric reconstruction errors. Frame-wise global alignment is performed using differentiable parameter estimation solvers on the pixel-wise correspondences to find a global transformation. Experiments on synthetic and real-world data show that our method outperforms the current state-of-the-art in terms of high-fidelity reconstruction and resolving large camera pose misalignment. Our module is an easy-to-use plugin that can be applied to NeRF variants and other neural field applications. The Code and supplementary materials are available at https://rover-xingyu.github.io/L2G-NeRF/.

1. Introduction

Recent success with neural fields [47] has caused a resurgence of interest in visual computing problems, where coordinate-based neural networks that represent a field gain traction as a useful parameterization of 2D images [4,7,40], and 3D scenes [27,29,34]. Commonly, these coordinates are warped to a global coordinate system by camera parameters obtained via computing homography, structure from motion (SfM), or simultaneous localization and mapping (SLAM) [17] with off-the-shelf tools like COLMAP [39], before being fed to the neural fields.

This paper considers the generic problem of simultaneously reconstructing the neural fields from RGB images and registering the given camera frames, which is known as a long-standing chicken-and-egg problem — registration is needed to reconstruct the fields, and reconstruction is needed to register the cameras.

One straightforward way to solve this problem is to jointly optimize the camera parameters with the neural fields via backpropagation. Recent work can be broadly placed into two camps: parametric and non-parametric. Parametric methods [10,20,24,44] directly optimize global geometric transformations (e.g. rigid, homography). Non-parametric methods [22,31] do not make any assumptions...
on the type of transformation, and attempt to directly optimize some pixel agreement metric (e.g. brightness constancy constraint in optical flow and stereo).

However, both approaches have flaws: parametric methods fail to minimize the photometric errors (falling into the suboptimal solutions) if poorly initialized, as shown in Fig. 1, while non-parametric methods have trouble dealing with large displacements (e.g. although the photometric errors are minimized, the alignments do not obey the geometric constraint). It is natural, therefore, to consider a hybrid approach, combining the benefits of parametric and non-parametric methods together.

In this paper, we propose L2G-NeRF, a local-to-global process integrating parametric and non-parametric methods for bundle-adjusting neural radiance fields — the joint problem of reconstructing the neural fields and registering the camera parameters, which can be regarded as a type of classic photometric bundle adjustment (BA) \cite{3,12,25}. Fig. 2 shows an overview. In the first non-parametric stage, we initialize the alignment by predicting a local transformation field for each pixel of the camera frames. This is achieved by self-supervised training of a deep network to optimize standard photometric reconstruction errors. In the second stage, differentiable parameter estimation solvers are applied to a set of pixel-wise correspondences to obtain a global alignment, which is then used to apply a soft constraint to the local alignment. In summary, we present the following contributions:

- We show that the optimization of bundle-adjusting neural fields is sensitive to initialization, and we present a simple yet effective strategy for local-to-global registration on neural fields.
- We introduce two differentiable parameter estimation solvers for rigid and homography transformation respectively, which play a crucial role in calculating the gradient flow from the global alignment to the local alignment.
- Our method is agnostic to the particular type of neural fields, specifically, we show that the local-to-global process works quite well in 2D neural images and 3D Neural Radiance Fields (NeRF) \cite{29}, allowing for applications such as image reconstruction and novel view synthesis.

### 2. Related Work

**SfM and SLAM.** SfM \cite{2,36,37,41,42} and SLAM \cite{16,30,32,48} systems attempt to simultaneously recover the 3D structure and the sensor poses from a set of input images. They reconstruct an explicit geometry (e.g. point clouds) and estimate camera poses through image registration via associating feature correspondences \cite{11,36} or minimizing photometric errors \cite{3,15}, followed by BA \cite{3,12,25}.

However, the explicit point clouds assume a diffuse surface, hence cannot model view-dependent appearance. And the sparse nature of point clouds also limits downstream vision tasks, such as photorealistic rendering. In contrast, L2G-NeRF encodes the scenes as coordinate-based neural fields, which is qualified for solving the high-fidelity visual computing problems.

**Neural Fields.** Recent advances in neural fields \cite{47}, which employ coordinate-based neural networks to parameterize physical properties of scenes or objects across space and time, have led to increased interest in solving visual computing problems, causing more accurate, higher fidelity, more expressive, and memory-efficient solutions. They have seen widespread success in problems such as image synthesis \cite{4,7,40}, 3D shape \cite{9,27,34}, view-dependent appearance \cite{6,18,29,33}, and animation of humans \cite{8,35,45}.

While these neural fields have achieved impressive results, the requirement of camera parameters limits its application. We are able to get around the requirement with our proposed L2G-NeRF.

**Bundle-Adjusting Neural Fields.** Since neural fields are end-to-end differentiable, camera parameters can be jointly estimated with the neural fields. The optimization problem is known to be non-convex, and is reflected by NeRF– \cite{44}, in which the authors jointly optimize the scene and cameras for forward-facing scenes. Adversarial objective is utilized \cite{26} to relax forward-facing assumption and supports inward-facing 360° scenes. SCNeRF \cite{20} is further developed to learn the camera intrinsics. BARF \cite{24} shows that bundle-adjusting neural fields could benefit from coarse-to-fine registration. Recent approaches employ Gaussian activations \cite{10} or Sinusoidal activations \cite{46} to overcome local minima in optimization.

Nevertheless, these parametric methods directly optimize global geometric transformations, which are prone to falling into suboptimal solutions if poorly initialized. Non-parametric methods \cite{22,31} directly optimize decent local transformations based on brightness constancy constraints, whereas they can not handle large displacements. We show that by combining the parametric and non-parametric methods together with a simple local-to-global process, we can achieve surprising anti-noise ability, allowing utilities for various NeRF extensions and other neural field applications.

### 3. Approach

We first present the formulation of recovering the neural field jointly with camera parameters. Given a collection of images \( \{ I_i \}_{i=1}^M \), we aim to jointly find the parameters \( \Theta \) of the neural field \( R \) and the camera parameters (geometric transformation matrices) \( \{ T_i \}_{i=1}^M \) that minimize the photometric error between renderings and images. Let \( \{ x_j \}_{j=1}^N \) be the query coordinates and \( I \) be the imaging function, we
formulate the problem as:

\[
\min_{\{T_i\}_{i=1}^M, \Theta} \sum_{i=1}^M \sum_{j=1}^N \left( \left\| \mathcal{R}(T_i; x^j; \Theta) - I_i(x^j) \right\|_2^2 \right).
\] (1)

Gradient-based optimization is the preferred strategy to solve this nonlinear problem. Nevertheless, gradient-based registration is prone to finding suboptimal poses. Therefore, we propose a simple yet effective strategy for local-to-global registration. The key idea is to apply a pixel-wise flexible alignment that optimizes photometric reconstruction errors individually, followed by a frame-wise alignment to globally constrain the local geometric transformations, which acts like a soft extension of Eq. (1):

\[
\min_{\{T_i\}_{i=1}^M, \Theta} \sum_{i=1}^M \sum_{j=1}^N \left( \left\| \mathcal{R}(T_i; x^j; \Theta) - I_i(x^j) \right\|_2^2 \right) + \lambda \left\| T_i^i x^j - T_i^j x^j \right\|_2^2,
\] (2)

where the pixel-wise local transformations \{T_i\}_{i=1}^M are modeled by a warp neural field \(\mathcal{W}\) parametrized by \(\Phi\), along with frame-dependent embeddings \{\ell_i\}_{i=1}^M:

\[
T_i^i = \mathcal{W}(x^j; \ell_i, \Phi),
\] (3)

and the frame-wise global transformations \{\(T_i^j\)\}_{i=1}^M are solved by using differentiable parameter estimation solvers on the pixel-wise correspondences:

\[
T_i^j = \arg\min_{T_i} \sum_{j=1}^N \left\| T_i^j x^j - T_i x^j \right\|_2^2.
\] (4)

**3.1. Neural Image Alignment (2D)**

To develop intuition, we first consider the case of a 2D neural image alignment problem. More specifically, let \(x \in \mathbb{R}^2\) be the 2D pixel coordinates and \(I : \mathbb{R}^2 \to \mathbb{R}^3\), we aim to optimize a 2D neural field parameterized as the weights \(\Theta\) of a multilayer perceptron (MLP) \(f_R : \mathbb{R}^2 \to \mathbb{R}^3\):

\[
\mathcal{R}(T; x; \Theta) = f_R(Tx; \Theta),
\] (5)

while also solving for geometric transformation parameters as \(T = [R | t] \in SE(2)\) or \(T \in SL(3)\), where \(R \in SO(2)\) and \(t \in \mathbb{R}^2\) denote the rigid rotation and translation, and \(T \in SL(3)\) denotes the homography transformation matrix, respectively. We use another MLP with weights \(\Phi\) to model the coordinate-based warp neural field \(f_W : \mathbb{R}^2 \to \mathbb{R}^3\) condition on the frame-dependent embedding \(\ell\):

\[
W(x; \ell, \Phi) = \exp \left( f_W(x; \ell, \Phi) \right),
\] (6)

where the operator \(\exp(\cdot)\) denotes the exponential map from Lie algebra \(se(2)\) or \(sl(3)\) to the Lie group \(SE(2)\) or \(SL(3)\), which ensures that the optimized transformation matrices \(T\) lie on the Lie group manifold during the gradient-based optimization.

**3.2. Bundle-Adjusting Neural Radiance Fields (3D)**

We then discuss the problem of simultaneously recovering the 3D Neural Radiance Fields (NeRF) [29] and the camera poses. Given an 3D point, we predict the RGB color \(c \in \mathbb{R}^3\) and volume density \(\sigma \in \mathbb{R}\) via an MLP \(f_R : \mathbb{R}^3 \to \mathbb{R}^4\), which encodes the 3D scene using network parameters. We begin by formulating NeRF’s rendering process in the space of the camera view. Denoting

---

8266
the homogeneous coordinates of pixel coordinates \( u \in \mathbb{R}^2 \) as \( x = [u; 1]^\top \in \mathbb{R}^3 \), the 3D point along the viewing ray at depth \( z \) can be expressed as \( zx \), thus the query quantity \( y = [c; \sigma]^\top = f_R(zx; \Theta) \), where \( \Theta \) is the parameters of \( f_R \). Then the rendering color \( R \) at pixel location \( x \) can be composed by volume rendering

\[
R(x) = \int_{z_{\text{near}}}^{z_{\text{far}}} T(z, x) \sigma(zx) e(zx) dz , \tag{7}
\]

where \( T(z, x) = \exp(-f_{z_{\text{near}}}^z \sigma(z'x) dz') \), and \( z_{\text{near}} \) and \( z_{\text{far}} \) are the near and far depth bounds of the scene. Numerically, the integral formulation is discretely approximated using \( K \) points sampled along a ray at depth \( \{z_1, \ldots, z_K\} \). The network \( f_R \) is evaluated \( K \) times, and the outputs \( \{y_1, \ldots, y_K\} \) are then composited via volume rendering.

Denoting the differentiable and deterministic compositing function as \( g : \mathbb{R}^{4K} \rightarrow \mathbb{R}^3 \), such that \( R(x) \) can be expressed as \( R(x) = g(y_1, \ldots, y_K) \).

Here the camera poses are parametrized by \( T = [R(t) \in \text{SE}(3), \text{where } R \in \text{SO}(3) \text{ and } t \in \mathbb{R}^3] \). Next, we use a 3D rigid transformation \( T \) to transform the 3D point \( z \) from camera view space to world coordinates, and formulate the rendering color at pixel \( x \) as

\[
R(Tx; \Theta) = g\left(f_R(Tz_1x; \Theta), \ldots, f_R(Tz_Kx; \Theta)\right). \tag{8}
\]

Similar to neural image alignment, we use another MLP with weights \( \Phi \) to model the coordinate-based warp neural field \( f_W : \mathbb{R}^3 \rightarrow \mathbb{R}^6 \) condition on the frame-dependent embedding \( \ell \):

\[
W(x; \ell, \Phi) = \exp\left(f_W(x; \ell, \Phi)\right) , \tag{9}
\]

where the operator \( \exp(\cdot) \) denotes the exponential map from Lie algebra \( \mathfrak{s}\mathfrak{e}(3) \) to the Lie group \( \text{SE}(3) \).

### 3.3. Differentiable Parameter Estimation

The local-to-global process allows L2G-NeRF to discover the correct registration with an initially flexible pixelwise alignment and later shift focus to constrained parametric alignment. We derive the gradient flow of global alignment objective \( \mathcal{L}_g^i = \|T_i^j x^j - T_i^j x^i\|_2 \) w.r.t. the parameters \( \Phi \) of warp neural field \( W \) as

\[
\frac{\partial \mathcal{L}_g^i}{\partial \Phi} = \frac{\partial \mathcal{L}_g^i}{\partial T_i^j} \frac{\partial T_i^j}{\partial \Phi} + \frac{\partial \mathcal{L}_g^i}{\partial T_i^j} \sum_{j=1}^{N} \frac{\partial T_i^j}{\partial T_j^j} \frac{\partial T_j^j}{\partial \Phi} . \tag{10}
\]

Such that a differentiable solver is of critical importance to calculating the gradient of \( T_i^j \) w.r.t. \( T_i^j \), then backpropagated to update the parameters \( \Phi \). We use Romain [5] and Korinia [38] for the differentiable rigid and homography parametric alignment. Next, we expound the two differentiable solvers, respectively.

1For the sake of simplicity, the viewing direction is omitted here.

#### Rigid parametric alignment.

In the rigid parametric alignment problem, we assume \( \{T_i^j x^j\}^{N}_{j=1} \) is transformed from \( \{x^j\}^{N}_{j=1} \) by an unknown global rigid transformation \( T = [R[t] \in \text{SE}(2) \text{ or } T = [R[t] \in \text{SE}(3)] \). To solve this classic orthogonal Procrustes problem [19], we define centroids of \( \{x^j\}^{N}_{j=1} \) and \( \{T_i^j x^j\}^{N}_{j=1} \) as

\[
x = \frac{1}{N} \sum_{j=1}^{N} (x^j) \quad \text{and} \quad \bar{T}x = \frac{1}{N} \sum_{j=1}^{N} (T_i^j x^j) . \tag{11}
\]

Then the cross-covariance matrix \( H \) is given by

\[
H = \sum_{j=1}^{N} (x^j - \bar{x})(T_i^j x^j - \bar{T}x)^\top . \tag{12}
\]

We use Singular Value Decomposition (SVD) to decompose \( H \) as introduced in [21, 43]:

\[
H = USV^\top . \tag{13}
\]

Thus the optimal transformation minimizing Eq. (4) is given in closed form by

\[
R = VU^\top \quad \text{and} \quad t = -Rx + \bar{T}x . \tag{14}
\]

#### Homography parametric alignment.

In the homography parametric alignment problem, we assume \( \{x^j\}^{N}_{j=1} \) is transformed from \( \{x^j\}^{N}_{j=1} \) by an unknown homography transformation \( T \in \text{SE}(3) \). Written element by element, in homogenous coordinates, we get the following constraint:

\[
\begin{bmatrix}
x_1^i \\
x_2^i \\
1
\end{bmatrix} = \begin{bmatrix}
T_{11} & T_{12} & T_{13} \\
T_{21} & T_{22} & T_{23} \\
T_{31} & T_{32} & T_{33}
\end{bmatrix} \begin{bmatrix}
x_1^j \\
x_2^j \\
1
\end{bmatrix} . \tag{15}
\]

Rearranging Eq. (15) as [1], we get \( A^i h = 0 \), where

\[
A^i = \begin{bmatrix}
0 & 0 & 0 & -x_1^i & -x_2^i & -1 & x_1^j x_1^j & x_1^j x_2^j & x_2^j x_2^j \\
x_1^j & x_2^j & 1 & 0 & 0 & -x_1^i & x_2^j x_2^j & x_1^j x_1^j & x_2^j x_2^j
\end{bmatrix} \begin{bmatrix}
0 \\
0 \\
0 \\
x_1^i \\
x_1^i \\
x_2^i \\
x_2^i \\
x_1^j \\
x_2^j
\end{bmatrix} . \tag{16}
\]

Given the set of correspondences, we can form the linear system of equations \( Ah = 0 \), where \( A = (A^1 \ldots A^N)^\top \). Thus we can solve the Homogeneous Linear Least Squares problem and calculate the non-trivial solution by SVD de-composition:

\[
A = USV^\top = \sum_{i=1}^{9} \sigma_i u_i v_i^\top , \tag{17}
\]

where singular value \( \sigma_i \) represents the reprojection error. Then we take the singular vector \( \sigma_0 \) that corresponds to the smallest singular value \( \sigma_0 \) as the solution of \( h \), and reshape it into the homography transformation matrix \( T \).
4. Experiments

We first unfold the validation of L2G-NeRF and baselines on a 2D neural image alignment experiment, and then show that the local-to-global registration strategy can also be generalized to learn 3D neural fields (NeRF [29]) from both synthetic data and photo collections.

4.1. Neural image Alignment (2D)

We choose two representative images of “Girl With a Pearl Earring” renovation ©Koorosh Orooj (CC BY-SA 4.0) and “cat” from ImageNet [13] for rigid and homography image alignment experiments, respectively. As shown in Fig. 3 and Fig. 4, given $M = 5$ patches sampled from the original image with rigid or homography perturbations, we optimize Eq. (2) to find the rigid transformation $T \in \text{SE}(2)$ or homography transformation $T \in \text{SL}(3)$ for each patch with network $f_W$, and learn the neural field of the entire image (Fig. 5 and Fig. 6) with network $f_R$ at the same time. We follow [24] to initialize patch warps as identity and anchor the first warp to align the neural image to the raw image.

**Experimental settings.** We evaluate our proposed method against a bundle-adjusting extension of the naive 2D neural field, dubbed as Naïve, and the current state-of-the-art BARF [24], which employs a coarse-to-fine strategy for registration. We use the default coarse-to-fine scheduling, architecture and training procedure of neural field $f_R$ for both BARF and L2G-NeRF. For L2G-NeRF, We use a ReLU MLP for $f_W$ with six 256-dimensional hidden units, and use the embedding with 128 dimensions for each image to model the frame-dependent embeddings $\{\ell_i\}_{i=1}^M$. We set multiplier $\lambda$ of the global alignment objective to $1 \times 10^2$.

| Method  | Rigid perturbations | Homography perturbations |
|---------|---------------------|--------------------------|
|         | Corner error (pixels) | Patch PSNR ↑ | Corner error (pixels) | Patch PSNR ↑ |
| Naïve   | 120.00              | 14.83 | 55.80              | 21.79 |
| BARF [24] | 110.20             | 17.78 | 30.21              | 23.24 |
| Ours    | 0.31               | 29.25 | 0.76               | 31.93 |

Table 1. Quantitative results of neural image alignment experiment under rigid and homography perturbations. L2G-NeRF optimizes for high-quality alignment and patch reconstruction, while baselines exhibit large errors.
Results. We visualize the rigid and homography registration results in Fig. 5 and Fig. 6. Alignment with Naïve results in ghosting artifacts in the recovered neural image due to large misalignment. On the other hand, alignment with BARF improves registration results but still falls into the suboptimal solutions, and struggles with image reconstruction. As L2G-NeRF discovers the precise geometric warps of all patches, it can optimize the neural image with high fidelity. We report the quantitative results in Table 1, where we use the mean average corner error (L2 distance between the ground truth corner position and the estimated corner position) \( \xi \in \mathbb{se}(3) \) and \( \xi \sim \mathcal{N}(0, nI) \) as initial poses, where the multiplier \( n \) is scene-dependent and given in the supplementary materials. We assume known camera intrinsics and minimize the objective in Eq. (2) for optimizing the 3D neural fields \( f_R \) and the warp field \( f_W \) that finds rigid transformations relative to the initial poses. We evaluate L2G-NeRF against a naïve extension of the original NeRF model that jointly optimizes poses, dubbed as Naïve, and the coarse-to-fine bundle-adjusting neural radiance fields (BARF) [24].

Implementation details. Our implementation of NeRF and BARF follows [24]. For L2G-NeRF, We use a 6-layer ReLU MLP for \( f_W \) with 256-dimensional hidden units. We set multiplier \( \lambda \) of the global alignment objective to \( 1 \times 10^2 \) and employ the Adam optimizer to train all models for 200K iterations with a learning rate that begins at \( 5 \times 10^{-4} \) for the 3D neural field \( f_R \), and \( 1 \times 10^{-3} \) for the warp field \( f_W \), and exponentially decays to \( 1 \times 10^{-4} \) and \( 1 \times 10^{-8} \), respectively. We follow the default coarse-to-fine scheduling for both BARF and L2G-NeRF.

Evaluation criteria. Following BARF [24], we use Procrustes analysis to find a 3D similarity transformation that aligns the optimized poses to the ground truth before evaluating registration quality (quantitative results based on average translation and rotation errors), and perform test-time photometric pose optimization [24,25,49] before evaluating view synthesis quality (quantitative results based on PSNR, SSIM and LPIPS [50]).
Table 2. Quantitative results of bundle-adjusting neural radiance fields on synthetic scenes. L2G-NeRF successfully optimizes camera poses, thus rendering high-quality images comparable to the reference NeRF model (trained using ground-truth camera poses), outperforming the baselines on all evaluation criteria. Translation errors are scaled by 100.

![Visual comparison of the initial and optimized camera poses (Procrustes aligned) for the lego scene. L2G-NeRF properly aligns all of the camera frames while baselines get stuck at suboptimal poses.](image-url)

**Results.** We visualize the results in Fig. 7, which are quantitatively reflected in Table 2. On both sides of reconstruction and registration, L2G-NeRF achieves the best performance. Fig. 8 shows that L2G-NeRF can achieve near-perfect registration for the synthetic scenes. Naïve NeRF suffers from suboptimal registration and ghosting artifacts. BARF is able to recover a part of the pose misalignment and produce plausible reconstructions. However, it still suffers from blur artifacts like the fog effect around the objects. This fog effect is the consequence of BARF’s attempt to reconstruct the scenes with half-baked registration. We then compare the rendering quality to the reference standard NeRF (ref. NeRF), which is trained using ground truth poses, demonstrating that L2G-NeRF can achieve comparable image quality, despite being initialized from a significant camera pose misalignment.

**4.3. NeRF (3D): Real-World Scenes**

We further explore the challenge of employing NeRF to learn 3D neural fields in real-world scenes with unknown camera poses. We evaluate our method and baselines on the standard benchmark LLFF dataset [28], which is captured by hand-held cameras that record 8 forward-facing scenes in the real world.

**Experimental settings.** We initialize all cameras with the identity transformation, i.e., \( T_i = I \) \( \forall i \), and use the camera intrinsics provided by LLFF scene. L2G-NeRF properly aligns all of the camera frames while baselines get stuck at suboptimal poses.

---

**Table 2.** Quantitative results of bundle-adjusting neural radiance fields on synthetic scenes. L2G-NeRF successfully optimizes camera poses, thus rendering high-quality images comparable to the reference NeRF model (trained using ground-truth camera poses), outperforming the baselines on all evaluation criteria. Translation errors are scaled by 100.

| Scene      | Camera pose registration | View synthesis quality |
|------------|--------------------------|------------------------|
|            | Rotation (°) ↓ | Translation ↓ | PSNR ↑ | SSIM ↑ | LPIPS ↓ |
|            | Naïve | BARF | Ours | Naïve | BARF | Ours | ref. NeRF | Naïve | BARF | Ours | ref. NeRF | Naïve | BARF | Ours | ref. NeRF |
| Chair      | 1.39 | 2.58 | **0.14** | 60.32 | 10.43 | **0.28** | 14.13 | 27.84 | **30.99** | 31.93 | 0.83 | 0.92 | **0.95** | 0.96 | 0.39 | 0.06 | **0.05** | 0.04 |
| Drums      | 7.99 | 4.54 | **0.06** | 78.20 | 19.19 | **0.40** | 11.63 | 21.92 | **23.75** | 23.98 | 0.61 | 0.87 | **0.90** | 0.90 | 0.62 | 0.14 | **0.10** | 0.10 |
| Ficus      | 3.13 | 1.65 | **0.26** | 48.78 | 5.46 | **1.11** | 14.30 | 25.85 | **26.11** | 26.66 | 0.83 | 0.93 | **0.93** | 0.94 | 0.33 | 0.07 | **0.06** | 0.05 |
| Hotdog     | 7.04 | 2.42 | **0.27** | 58.37 | 14.98 | **1.42** | 15.10 | 27.34 | **34.56** | 34.90 | 0.74 | 0.93 | **0.97** | 0.97 | 0.42 | 0.06 | **0.03** | 0.03 |
| Lego       | 7.82 | 9.93 | **0.09** | 81.93 | 47.42 | **0.37** | 11.36 | 14.48 | **27.91** | 29.29 | 0.61 | 0.69 | **0.91** | 0.94 | 0.56 | 0.29 | **0.06** | 0.04 |
| Materials  | 5.57 | 0.68 | **0.06** | 47.56 | 4.97 | **0.28** | 11.51 | 26.29 | **27.60** | 28.54 | 0.64 | 0.92 | **0.93** | 0.94 | 0.49 | 0.08 | **0.06** | 0.05 |
| Mic        | 4.43 | 10.44 | **0.10** | 77.47 | 45.66 | **0.44** | 13.14 | 12.20 | **30.91** | 31.96 | 0.85 | 0.76 | **0.97** | 0.97 | 0.43 | 0.41 | **0.05** | 0.04 |
| Ship       | 11.10 | 23.90 | **0.19** | 112.01 | 90.62 | **0.61** | 9.41 | 8.19 | **27.31** | 28.06 | 0.50 | 0.50 | **0.85** | 0.86 | 0.64 | 0.63 | **0.13** | 0.12 |
| Mean       | 6.06 | 7.02 | **0.15** | 70.58 | 29.84 | **0.61** | 12.57 | 20.51 | **28.62** | 29.42 | 0.70 | 0.82 | **0.93** | 0.94 | 0.49 | 0.22 | **0.07** | 0.06 |
Figure 9. Qualitative results of bundle-adjusting neural radiance fields on real-world scenes. While BARF and L2G-NeRF can jointly optimize poses and scenes, L2G-NeRF produces higher fidelity results, which is competitive to reference NeRF trained under SfM poses.

Table 3. Quantitative results of bundle-adjusting neural radiance fields on real-world scenes. L2G-NeRF outperforms baselines and achieves high-quality view synthesis that is competitive to reference NeRF trained under SfM poses. Translation errors are scaled by 100.

5. Conclusion

We present Local-to-Global Registration for Bundle-Adjusting Neural Radiance Fields (L2G-NeRF), which is demonstrated by extensive experiments that can effectively learn the neural fields of scenes and resolve large camera pose misalignment at the same time. By establishing a unified formulation of bundle-adjusting neural fields, we demonstrate that local-to-global registration is beneficial for both 2D and 3D neural fields, allowing for various applications of diverse neural fields. Code and models will be made available to the research community to facilitate reproducible research.

Although local-to-global registration is much more robust than current state-of-the-art [24], L2G-NeRF still cannot recover camera poses from scratch (identity transformation) for inward-facing 360° scenes, where large displacements of rotation exist. Specific methods such as epipolar geometry and graph optimization could be employed to handle these issues.
References

[1] Youset I Abdel-Aziz, Hauck Michael Karara, and Michael Hauck. Direct linear transformation from comparator coordinates into object space coordinates in close-range photogrammetry. Photogrammetric engineering & remote sensing, 81(2):103–107, 2015. 4

[2] Sameer Agarwal, Yasutaka Furukawa, Noah Snavely, Ian Simon, Brian Curless, Steven M Seitz, and Richard Szeliski. Buildingrome in a day. ACM Transactions, 2011. 2

[3] Hater Aliasmia, Brett Browning, and Simon Lucey. Photometric bundle adjustment for vision-based slam. In Asian Conference on Computer Vision, pages 324–341. Springer, 2016. 2

[4] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. ACM Transactions on Graphics (TOG), 39(6):1–15, 2020. 1, 2

[5] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. 2021. 4

[6] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Ying Feng, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 12943–12952, 2022. 2

[7] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. In Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition, pages 8628–8638, 2021. 1, 2

[8] Yue Chen, Xuan Wang, Qi Zhang, Xiaoyu Li, Xingyu Chen, Yu Guo, Jue Wang, and Fei Wang. UV Volumes for real-time rendering of editable free-view human performance. arXiv preprint arXiv:2203.14402, 2022. 2

[9] Zhiqin Chen and Hao Zhang. Learning implicit fields for high fidelity reconstruction and pose estimation. In European Conference on Computer Vision, pages 264–280. Springer, 2022. 1, 2

[10] Andrew J Davison, Ian D Reid, Nicholas D Molton, and Olivier Stasse. Monoslam: Real-time single camera slam. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007. 2

[11] Amael Delaunoy and Marc Pollefeys. Photometric bundle adjustment for dense multi-view 3d modeling. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1486–1493, 2014. 2

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 248–255. Ieee, 2009. 5

[13] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. arXiv preprint arXiv:1606.03798, 2016. 6

[14] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. 2

[15] Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1

[16] John R Hurley and Raymond B Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. Behavioral science, 7(2):258, 1962. 4

[17] Yu Guo, Jue Wang, and Fei Wang. UV Volumes for real-time rendering of editable free-view human performance. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6351–6361, 2021. 2

[18] Chun-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5846–5854, 2021. 1, 2

[19] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Deep image homography estimation. arXiv preprint arXiv:1606.03798, 2016. 6

[20] Jakob Engel, Vladlen Koltun, and Daniel Cremers. Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017. 2

[21] Richard Hartley and Andrew Zisserman. Multiple View Geometry in Computer Vision. Cambridge University Press, ISBN: 0521540518, second edition, 2004. 1

[22] Yoni Kasten, Dolev Ofri, Oliver Wang, and Tali Dekel. Layered neural atlases for consistent video editing. ACM Transactions on Graphics (TOG), 40(6):1–12, 2021. 1, 2

[23] Hoang Le, Feng Liu, Shu Zhang, and Aseem Agarwala. Deep homography estimation for dynamic scenes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7652–7661, 2020. 6

[24] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5741–5751, 2021. 1, 2, 5, 6, 7, 8

[25] Stephen Lucey. Gaussian activated neural radiance fields for generative shape modeling. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021. 4

[26] Christopher Olah, John Shlens, and Alexander Mordvintsev. A solution for the best rotation to relate two sets of vectors. Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography, 32(5):922–923, 1976. 4

[27] John R Hurley and Raymond B Cattell. The procrustes program: Producing direct rotation to test a hypothesized factor structure. Behavioral science, 7(2):258, 1962. 4

[28] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and
