LU-NeRF: Scene and Pose Estimation by Synchronizing Local Unposed NeRFs

Zezhou Cheng[1,2*] Carlos Esteves[2] Varun Jampani[2] Abhishek Kar[2]
Subhransu Maji[1] Ameesh Makadia[2]
[1]University of Massachusetts, Amherst [2]Google Research

Abstract

A critical obstacle preventing NeRF models from being deployed broadly in the wild is their reliance on accurate camera poses. Consequently, there is growing interest in extending NeRF models to jointly optimize camera poses and scene representation, which offers an alternative to off-the-shelf SfM pipelines which have well-understood failure modes. Existing approaches for unposed NeRF operate under limiting assumptions, such as a prior pose distribution or coarse pose initialization, making them less effective in a general setting. In this work, we propose a novel approach, LU-NeRF, that jointly estimates camera poses and neural radiance fields with relaxed assumptions on pose configuration. Our approach operates in a local-to-global manner, where we first optimize over local subsets of the data, dubbed "mini-scenes." LU-NeRF estimates local pose and geometry for this challenging few-shot task. The mini-scene poses are brought into a global reference frame through a robust pose synchronization step, where a final global optimization of pose and scene can be performed. We show our LU-NeRF pipeline outperforms prior attempts at unposed NeRF without making restrictive assumptions on the pose prior. This allows us to operate in the general SE(3) pose setting, unlike the baselines. Our results also indicate our model can be complementary to feature-based SfM pipelines as it compares favorably to COLMAP on low-texture and low-resolution images.

1. Introduction

NeRF [34] was introduced as a powerful method to tackle the problem of learning neural scene representations and photorealistic view synthesis, and subsequent research has focused on addressing its limitations to extend its applicability to a wider range of use cases (see [54, 59] for surveys). One of the few remaining hurdles for view synthesis in the wild is the need for accurate localization. As images captured in the wild have unknown poses, these approaches often use structure-from-motion (SfM) [48, 40] to determine the camera poses. There is often no recourse when SfM fails (see Fig. 7 for an example), and in fact, even small inaccuracies in camera pose estimation can have a dramatic impact on photorealism.

Few prior attempts have been made to reduce the reliance on SfM by integrating pose estimation directly within the NeRF framework. However, the problem is severely underconstrained (see Fig. 1) and current approaches make additional assumptions to make the problem tractable. For example, NeRf−− [56] focuses on pose estimation in forward-facing configurations, BARF [29] initialization must be close to the true poses, and GNeRF [32] assumes a 2D camera model (upright cameras on a hemisphere).

We propose an approach for jointly estimating the camera pose and scene representation from images from a single scene while allowing for a more general camera configuration than previously possible. Conceptually, our approach is organized in a local-to-global learning framework using NeRFs. In the *local* processing stage we partition the scene into overlapping subsets, each containing only a few images (we call these subsets *mini-scenes*). Knowing images in a mini-scene are mostly nearby is what makes the joint estimation of pose and scene better conditioned than performing the same task globally. In the *global* stage, the overlapping mini-scenes are registered in a common reference frame through pose synchronization, followed by jointly refining all poses and learning the global scene representation.

This organization into mini-scenes requires learning from a few local unposed images. Although methods exist for few-shot novel view synthesis [60, 27, 38, 20, 12, 11], and separately for optimizing unknown poses [29, 32, 56], the combined setting presents new challenges. Our model must reconcile the ambiguities prevalent in the local unposed setting – in particular the mirror symmetry ambiguity [39], where two distinct 3D scenes and camera configurations produce similar images under affine projection.

We introduce a Local Unposed NeRF (LU-NeRF) model to address these challenges in a principled way. The information from the LU-NeRFs (estimated poses, confidences, and mirror symmetry analysis) is used to register all cam-

*Work done during an internship at Google.
Website: https://zezhoucheng.github.io/lu-nerf/

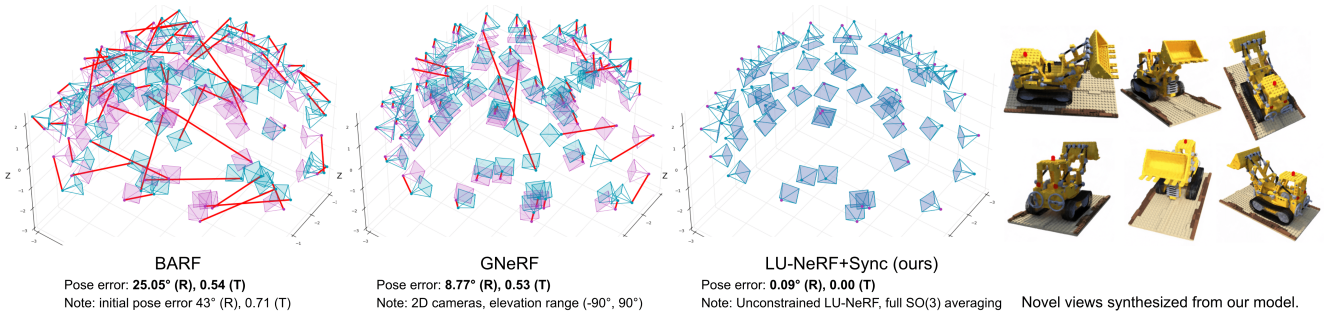

Figure 1. Jointly optimizing camera poses and scene representation over a full scene is difficult and underconstrained. This example is the Lego scene with 100 images from the Blender dataset. **Left**: When provided noisy observations of the true camera locations, BARF [29] cannot converge to the correct poses. **Middle**: GNeRF [32] assumes a 2D camera representation (azimuth, elevation) which is accurate for the Blender dataset which has that exact configuration (upright cameras on a sphere). However, GNeRF also requires an accurate prior distribution on poses for sampling. The Lego images live on one hemisphere, but when GNeRF's prior distribution is the full sphere it also fails to localize the images accurately. **Right**: Our full model, LU-NeRF+Sync, is able to recover poses almost perfectly in this particular example. By taking a local-to-global approach, we avoid having strong assumptions about camera representation or pose priors. Following [29, 32] pose errors for each method are reported after optimal global alignment of estimated poses to ground truth poses. To put the translation errors in context, the Blender cameras are on a sphere of radius 4.03.

eras in a common reference frame through pose synchronization [19, 42, 23], after which we refine the poses and optimize the neural scene representations using all images. In summary, our key contributions are:

- A local-to-global pipeline that learns both the camera poses in a general configuration and a neural scene representation from only an unposed image set.
- LU-NeRF, a novel model for few-shot local unposed NeRF. LU-NeRF is tailored to the unique challenges we have identified in this setting, such as reconciling mirror-symmetric configurations.

Each phase along our local-to-global process is designed with robustness in mind, and the consequence is that our pipeline can be successful even when the initial mini-scenes contain frequent outliers (see Sec 4 for a discussion on different mini-scene construction techniques). The performance of our method surpasses prior works that jointly optimize camera poses and scene representation, while also being flexible enough to operate in the general SE(3) pose setting unlike prior techniques. Our experiments indicate that our pipeline is complementary to the feature-based SfM pipelines used to initialize NeRF models, and is more reliable in low-texture or low-resolution settings.

2. Related work

Structure from motion (SfM). Jointly recovering 3D scenes and estimating camera poses from multiple views of a scene is the classic problem in Computer Vision [24]. Numerous techniques have been proposed for SfM [40, 48] with unordered image collections and visual-SLAM for sequential data [53, 37]. These techniques are largely built upon local features [31, 44, 21, 51] and require accurate detection and matching across images. The success of these techniques has led to their widespread adoption, and existing deep-learning approaches for scene representation and novel view synthesis are designed with the implicit assumption that the SfM techniques provide accurate poses in the wild. For example, NeRF [34] and its many successors (*e.g.* [4, 5, 36]) utilize poses estimated offline with COLMAP [48, 30]. However, COLMAP can fail on textureless regions and low-resolution images.

The local-to-global framework proposed in this work is inspired by the "divide-and-conquer" SfM and SLAM methods [7, 64, 22, 14, 18, 63, 17].

Neural scene representation with unknown poses. BARF [29] and GARF [15] jointly optimize neural scene and camera poses, but require good initialization (*e.g.* within 15° of the groundtruth). NeRF−− [56], X-NeRF [41], SiNeRF [58], and SaNeRF [13] only work on forward-facing scenes; SAMURAI [9] aims to handle coarsely specified poses (octant on a sphere) using a pose multiplexing strategy during training; GNeRF [32] and VMRF [61] are closest to our problem setting. They do not require accurate initialization and work on 360° scenes. However, they make strong assumptions about the pose distribution, assuming 2DoF and a limited elevation range. Performance degrades when the constraints are relaxed.

Approaches that combine visual SLAM with neural scene representations [65, 50, 43] typically rely on RGB-D streams and are exclusively designed for video sequences. The use of depth data significantly simplifies both scene and pose estimation processes. There are several parallel efforts to ours in this field. For instance, NoPe-NeRF [8] trains a NeRF without depending on pose priors; however, it relies on monocular depth priors. In a manner akin to our approach, LocalRF [33] progressively refines camera poses

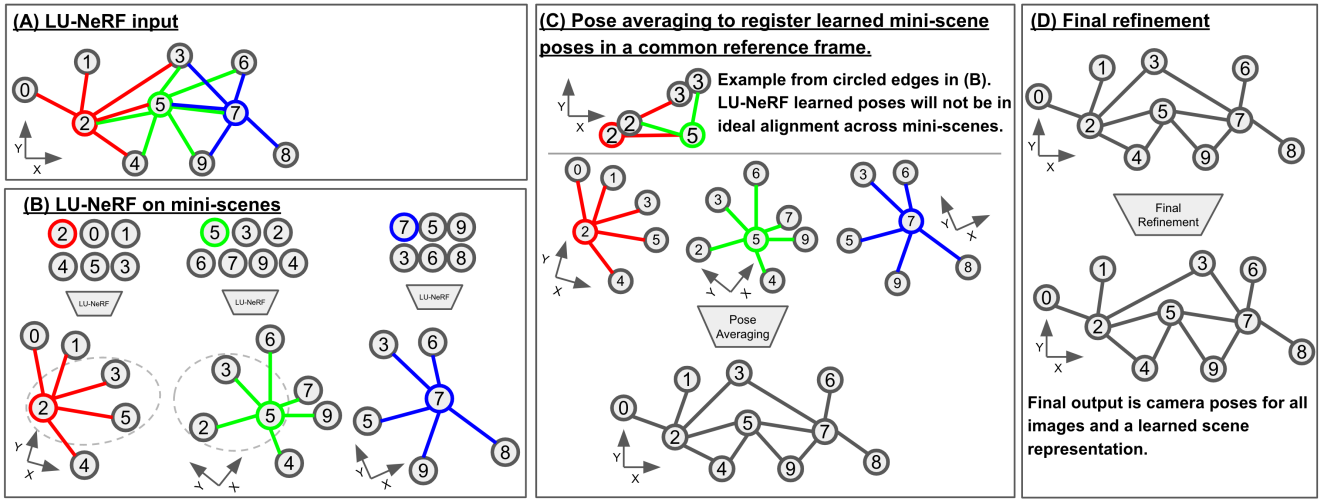

Figure 2. **Proposed method.** (A) shows the ground truth locations of each image (we show this only for visualization). Edge colors show the grouping within mini-scenes. We create a mini-scene for each image, though here only three mini-scenes are highlighted; the ones centered at image 2 (red edges), image 5 (green edges), and image 7 (blue edges). Depending on the strategy used to create mini-scenes, the grouped images can contain outlier images far from the others. (B) LU-NeRF takes unposed images from a single mini-scene and optimizes poses without any constraints on the pose representation. (C) The reference frame and scene scale learned by LU-NeRF is unique to each mini-scene. This, plus estimation errors, means the relative poses between images in overlapping mini-scenes will not perfectly agree. To register the cameras in a common reference frame, we utilize pose synchronization which seeks a globally optimal positioning of all cameras from noisy relative pose measurements – this is possible since we have multiple relative pose estimations for many pairs of images. (D) Lastly, we jointly refine the synchronized camera poses and learn a scene representation.

and radiance fields within local scenes. Despite this similarity, it presumes monocular depth and optical flow as supervision, and its application is limited to ordered image collections; MELON [28] optimizes NeRF with unposed images using equivalence class estimation, yet it is limited to $\mathbf{SO}(3)$; RUST [45] and FlowCam [49] learn a generalizable neural scene representation from unposed videos.

In summary, prior work on neural scene representation with unknown poses assumes either small perturbations [29, 15, 56, 58], a narrow distribution of camera poses [32, 61], or depth priors [8, 33]. To the best of our knowledge, we are the first to address the problem of neural rendering with unconstrained unknown poses for both ordered and unordered image collections.

Few-shot scene estimation. Learning scene representations from a few images has been studied in [60, 20, 12, 11, 27, 38]. PixelNeRF [60] uses deep CNN features to construct NeRFs from few or even a single image. MVSNeRF [11] leverages cost-volumes typically applied in multi-view stereo for the same task, while DS-NeRF [20] assumes depth supervision is available to enable training with fewer views. Our approach to handle the few-shot case relies on a standard neural field optimization with strong regularization, similar to RegNeRF [38].

Unsupervised pose estimation. There are a number of techniques that can learn to predict object pose from categorized image collections without explicit pose supervision. Multiple views of the same object instance are used in [55, 25] to predict the shape and pose while training is self-supervised through shape rendering. RotationNet [26] uses multiple views of an object instance to predict both poses and class labels but is limited to a small set of discrete uniformly spaced camera viewpoints. The multi-view input is relaxed in [35, 57] which operates on single image collections for a single category. UNICORN [35] learns a disentangled representation that includes pose and utilizes cross-instance consistency at training, while an assumption about object symmetry guides the training in [57].

3. Methodology

An illustration of our approach is shown in Figure 2. At the core of our method is the idea of breaking up a large scene into mini-scenes to overcome the non-convexity of global pose optimization without accurate initialization. When the camera poses in the mini-scene are close to one another, we are able to initialize the optimization with all poses close to the identity and optimize for relative poses. In Sec. 4, we describe how we construct mini-scenes, and below we describe the process of local shape estimation followed by global synchronization.

3.1. Local pose estimation

The local pose estimation step takes in mini-scenes of typically three to five images and returns the relative poses

between the images. The model, denoted LU-NeRF-1, is a small NeRF [34] that jointly optimizes the camera poses as extra parameters as in BARF [29]. In contrast with BARF, in this stage, we are only interested in a rough pose estimation that will be improved upon later, so we aim for a lightweight model with faster convergence by using small MLPs and eliminating positional encoding and view dependency. As we only need to recover relative poses, without loss of generality, we freeze one of the poses at identity and optimize all the others.

Few-shot radiance field optimization is notoriously difficult and requires strong regularization [38]. Besides the photometric ℓ_2 loss proposed in NeRF, we found that adding a loss term for the total variation of the predicted depths over small patches is crucial for the convergence of both camera pose and scene representation:

$$\frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}\in\mathcal{R}}\sum_{i,j=1}^{K}\big(d_\theta(\mathbf{r}_{i,j})-d_\theta(\mathbf{r}_{i,j+1})\big)^2+\big(d_\theta(\mathbf{r}_{i,j})-d_\theta(\mathbf{r}_{i+1,j})\big)^2$$

where $\mathcal{R}$ is a set of ray samples, $d_\theta(\mathbf{r})$ is the depth rendering function for a ray $\mathbf{r}$, θ are the model parameters and camera poses, K is the patch size, and (i, j) is the pixel index.

3.2. Mirror-symmetry ambiguity

The ambiguities and degeneracies encountered when estimating 3D structure have been extensively studied [52, 6, 16]. One particularly relevant failure mode of SfM is distant small objects, where the perspective effects are small and can be approximated by an affine transform, and one cannot differentiate between reflections of the object around planes parallel to the image plane [39]. When enforcing multi-view consistency, this effect, known as mirror-symmetry ambiguity, can result in two different configurations of structure and motion that cannot be told apart (see Fig. 3). We notice, perhaps for the first time, that neural radiance fields with unknown poses can degenerate in the same way.

One potential solution to this problem would be to keep the two possible solutions and drop one of them when new observations arrive. This is not applicable to our case since at this stage the only information available is the few images of the mini-scene.

To mitigate the issue, we introduce a second stage for the training, denoted LU-NeRF-2. We take the estimated poses in world-to-camera frame $\{R_i\}$ from LU-NeRF-1, and the reflected cameras $\{R_\pi R_i\}$, where R_π is a rotation around the optical axis. Note that this is different than post-multiplying by R_π, which would correspond to a global rotation that wouldn't change the relative poses that we are interested in at this stage. We then train two new models, with the scene representation started from scratch and poses initialized as the original and reflected sets, and resolve the ambiguity by picking the one with the smallest photometric training loss. The rationale is that while the issue is caused by LU-NeRF-1 ignoring small perspective distortions, the distortions can be captured on the second round of training, which is easier since one of the initial sets of poses is expected to be reasonable.

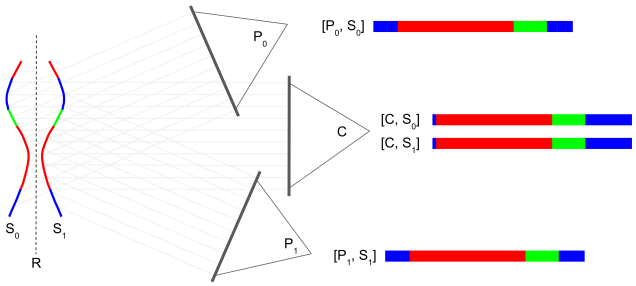

Figure 3. **Mirror symmetry ambiguity.** Under affine projection, a 3D scene (S_0) and its reflection (S_1) across a plane (R) will produce the same image viewed from affine camera C. The consequence of this is that two distinct 3D scenes and camera poses will produce similar images. In this illustration, scene S_0 viewed from camera P_0 will produce the same image as the reflected scene S_1 viewed from P_1. While this relationship is exact in the affine model, we observe that the mini-scene configuration with respect to the scene structure is often well-approximated as affine and training can converge to the near-symmetric solutions. Our LU-NeRF model is explicitly designed to anticipate this failure mode. This illustration is inspired by a similar diagram in [39].

3.3. Local to global pose estimation

After training LU-NeRF-2, we have sets of relative poses for each mini-scene in some local frame. The problem of finding a global alignment given a set of noisy relative poses is known as pose synchronization or pose averaging. It is formalized as optimizing the set of N global poses $\{P_i\}$ given relative pose observations R_{ij},

$$\underset{P\in\mathbf{SE}(3)^N}{\operatorname{argmin}}\; d(P_{ij}, P_jP_i^\top), \tag{1}$$

for some metric $d\colon \mathbf{SE}(3)\times\mathbf{SE}(3)\mapsto\mathbb{R}$. The problem is challenging due to non-convexity and is an active subject of research [3, 42, 19]. We use the Shonan rotation method [19] to estimate the camera rotations, followed by a least-squares optimization of the translations.

Global pose and scene refinement. After pose averaging, the global pose estimates are expected to be good enough such that any method that requires cameras initialized close to the ground truth should work (*e.g.* BARF [29], GARF [15]). We apply BARF [29] at this step, which results in both accurate poses and a scene representation accurate enough for realistic novel view synthesis. We refer to the full pipeline as LU-NeRF+Sync.

	Chair		Hotdog		Lego		Mic		Drums		Ship	
	rot	trans	rot	trans	rot	trans	rot	trans	rot	trans	rot	trans
COLMAP	0.12	0.01	1.24	0.04	2.29	0.10	8.37	0.18	5.91	0.28	0.17	0.01
+BARF	0.14	0.01	1.20	0.01	1.88	0.09	3.73	0.15	8.71	0.54	0.15	0.01
VMRF 120°	4.85	0.28	–	–	2.16	0.16	1.39	0.07	1.28	0.08	16.89	0.71
GNeRF 90°	0.36	0.02	2.35	0.12	0.43	0.02	1.87	0.03	0.20	0.01	3.72	0.18
GNeRF 120°	4.60	0.16	17.19	0.74	4.00	0.20	2.44	0.08	2.51	0.11	31.56	1.38
GNeRF 150°	16.10	0.76	23.53	0.92	4.17	0.36	3.65	0.26	5.01	0.18	–	–
GNeRF 180° (2DOF)	24.46	1.22	36.74	1.46	8.77	0.53	12.96	0.66	**9.01**	0.49	–	–
Ours (3DOF)	**2.64**	**0.09**	**0.24**	**0.01**	**0.09**	**0.00**	**6.68**	**0.10**	12.39	**0.23**	–	–

Table 1. **Camera pose estimation on unordered image collection.** GNeRF [32] and VMRF [61] constrain the elevation range, where the maximum elevation is always 90°. For example, GNeRF 120° only samples elevations in $[-30°, 90°]$. The 180° variations don't constrain elevation and are closest to our method, but they are still limited to 2 degrees of freedom for assuming upright cameras. Bold numbers indicate superior performance between the bottom two rows, which are the fairest comparison among NeRF-based methods, although our method is still solving a harder 3DOF problem versus 2DOF of GNeRF. We outperform GNeRF in all but one scene in this comparison. COLMAP [48] results in its best possible scenario are shown for reference (higher resolution images and assuming optimal graph to set unregistered poses to the closest registered pose). COLMAP+BARF runs a BARF refinement on top of these initial results, and even in this best-case scenario, our method still outperforms it in some scenes, which shows that LU-NeRF can complement COLMAP and work in scenes COLMAP fails. Our model fails on the Ship scene due to outliers in the connected graph; GNeRF with fewer constraints also fails on it. We provide a detailed error analysis on the Drums scene in the supplementary material.

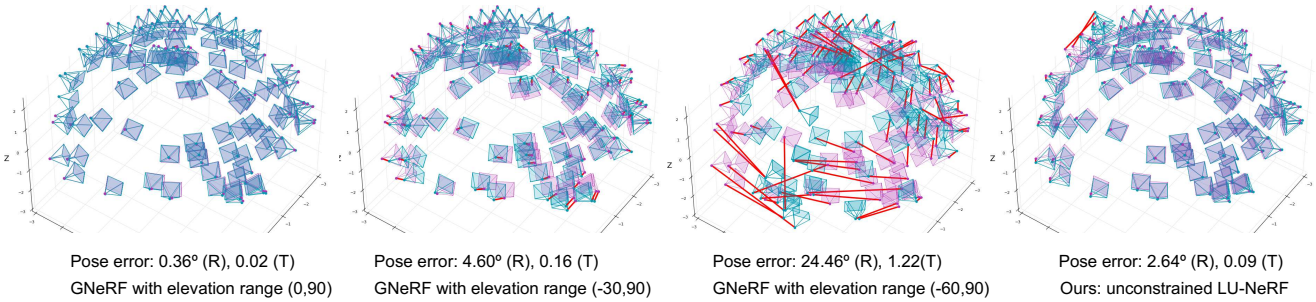

Figure 4. **Camera pose estimation on unordered image collections.** The performance of GNeRF drops dramatically when the pose prior is expanded beyond the true distribution. In comparison, our method does not rely on any prior knowledge of pose distribution.

4. Experiments

Our method as described in Sec. 3 starts from a set of mini-scenes that covers the input scene. We evaluate different approaches to constructing mini-scenes, each with different assumptions on the input.

The most strict assumption is that we have an *optimal graph* connecting each image to its nearest neighbors in camera pose space. While this seems unfeasible in practice, some real-life settings approximate this, for example, when images are deliberately captured in a pattern such as a grid, or if they are captured with camera arrays.

In a less constrained version of the problem, we assume an *ordered image collection*, where the images form a sequence, from where a line graph is trivially built. This is a mild assumption that is satisfied by video data, as well as the common setting of a camera physically moving around a scene sequentially capturing images.

In the most challenging setting, we assume nothing about the scene and only take an *unordered image collection*.

Building graphs from unordered image collections. We evaluate two simple ways of building graphs from unordered image collections. The first is to use deep features from a self-supervised model trained on large image collections. We use the off-the-shelf DINO model [10, 2] to extract image features and build the graph based on the cosine distance between these features. The second is to simply use the ℓ_1 distance in pixel space against slightly shifted and rotated versions of the images. Neither of these approaches is ideal. The deep features are typically coarse and too general, failing to detect specific subtle changes on the scene. The ℓ_1 distance has the opposite issue, where small changes can result in large distances. We provide a detailed analysis in the supplementary material. Exploring other methods for finding a proxy metric for the relative pose in image space is a direction for future work.

Datasets. We compare with existing published results on the synthetic-NeRF dataset [34]. We use the training split of the original dataset as our *unordered image collection* which consists of 100 unordered images per 3D scene. We use the

	Chair			Drums			Lego			Mic		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GNeRF 90°	31.30	0.95	0.08	24.30	0.90	0.13	28.52	0.91	0.09	31.07	0.96	0.06
GNeRF 120°	25.01	0.89	0.15	20.63	0.86	0.20	22.95	0.85	0.16	23.68	0.93	0.11
GNeRF 150°	22.18	0.88	0.20	19.05	0.83	0.27	21.39	0.84	0.18	23.22	0.92	0.13
VMRF 120°	26.05	0.90	0.14	23.07	0.89	0.16	25.23	0.89	0.12	27.63	0.95	0.08
VMRF 150°	24.53	0.90	0.17	21.25	0.87	0.21	23.51	0.86	0.14	24.39	0.94	0.10
GNeRF 180° (2DOF)	21.27	0.87	0.23	18.08	0.81	0.33	18.22	0.82	0.24	17.22	0.86	0.32
VMRF 180° (2DOF)	23.18	0.89	0.16	20.01	0.84	0.29	21.59	0.83	0.18	20.29	0.90	0.22
Ours (3DOF)	**30.57**	**0.95**	**0.05**	**23.53**	**0.89**	**0.12**	**28.29**	**0.92**	**0.06**	**22.58**	**0.91**	**0.08**

Table 2. **Novel view synthesis on unordered collections.** Our method outperforms the baselines on most scenes while being more general for considering arbitrary rotations with 3 degrees-of-freedom. Here we quote the baseline results from VMRF [61], where *hotdog* is not available. We provided the results on all scenes (including *hotdog*) using the public source code of GNeRF in the supplementary material.

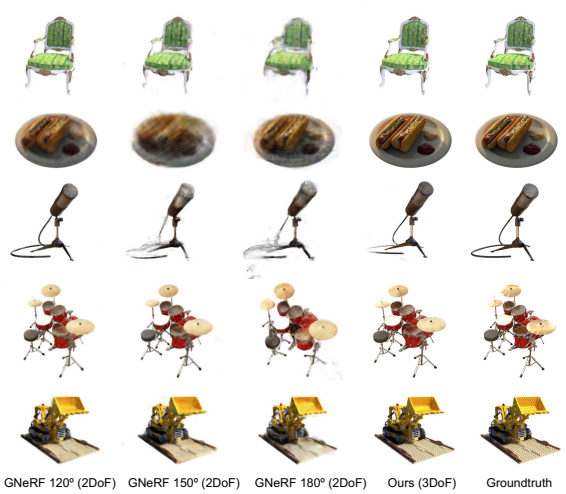

Figure 5. **Novel view synthesis on unordered image collections**. GNeRF makes assumptions on the elevation range, where the maximum elevation is always 90°. For instance, GNeRF 150° only samples elevations in [-60°, 90°]. The 180° variations don't constrain elevation and are closest to our method, but they are still limited to 2 degrees of freedom for assuming upright cameras. The performance of GNeRF drops as prior poses are less constrained. Please zoom into the figure to see the details in the renderings.

Image size	Chair	Hotdog	Lego	Mic	Drums	Ship
400×400	100	88	100	15	74	45
800×800	100	98	100	80	84	100

Table 3. **Number of images registered by COLMAP on Blender.**

first 8 images from the validation set as our test set for the novel view synthesis task, following prior works [32, 61]

To evaluate on image sequences, where the order of images is known, we further render a Blender *ordered image collection* with 100 images along a spiral path per scene. The images are resized to 400×400 in our experiments.

We also evaluate on real images from the object-centric videos in Objectron [1]. The dataset provides ground truth poses computed using AR solutions at 30fps, and we construct a wider-baseline dataset by subsampling every 15th frame and selecting videos with limited texture (Fig. 7).

Evaluation metrics. We evaluate the tasks of camera pose estimation and novel view synthesis. For camera pose estimation, we report the camera rotation and translation error using Procrustes analysis as in BARF [29]. For novel view synthesis, we report the PSNR, SSIM, and LPIPS [62].

Baseline methods. We compare with GNeRF [32], VMRF [61], and COLMAP [48] throughout our experiments. GNeRF samples camera poses from a predefined prior pose distribution and trains a GAN-based neural rendering model to build the correspondence between the sampled camera poses and 2D renderings. The method provides accurate pose estimation under *proper* prior pose distribution. However, its performance degrades significantly when the prior pose distribution doesn't match the groundtruth. VMRF attempts to relieve the reliance of GNeRF on the prior pose distribution but still inherits its limitations. In our experiments, we evaluate with the default pose priors of GNeRF on the NeRF-synthetic dataset, *i.e.*, azimuth $\in [0°, 360°]$ and elevation $\in [0°, 90°]$, and also on less constrained cases. COLMAP works reliably in texture-rich scenes but may fail dramatically on texture-less surfaces.

Implementation details. We use a compact network for LU-NeRF to speed up the training and minimize the memory cost. Specifically, we use a 4-layer MLP without positional encoding and conditioning on the view directions. We stop the training early when the change of camera poses on mini-scenes is under a predefined threshold. To resolve the mirror symmetry ambiguity (Sec. 3.2), we train two additional LU-NeRFs for a fixed number of training iterations (50k by default). The weight of the depth regularization is 10 times larger than the photometric ℓ_2 loss throughout our experiments. More details are in the supplementary material.

4.1. Unordered Image Collections

Camera pose estimation. Tab. 1 compares our method to GNeRF, VMRF, and COLMAP in the camera pose estimation task. GNeRF achieves high pose estimation accuracy

	Chair		Drums		Lego		Materials		Mean	
	rot	trans	rot	trans	rot	trans	rot	trans	rot	trans
GNeRF 90°	11.6	0.49	8.03	0.29	7.89	0.19	6.80	0.12	8.91	0.30
GNeRF 180°	27.7	1.17	130	6.23	123	4.31	30.9	1.40	94.9	3.27
Ours (3DOF)	**0.72**	**0.03**	**0.07**	**0.08**	**1.96**	**0.00**	**0.31**	**0.00**	**0.76**	**0.03**

Table 4. **Pose estimation on the Blender *ordered image collections*.** We report rotation errors in degrees and translation at the input scene scale. Our method can be more easily applied to ordered image collections since the graph-building step becomes trivial. In this case, we outperform GNeRF even when it is aided by known and constrained pose distributions.

	Chair			Drums		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GNeRF 90°	27.22	0.93	0.17	20.88	0.84	0.29
GNeRF 180° (2DOF)	23.50	0.91	0.26	11.01	0.81	0.56
Ours (3DOF)	**33.94**	**0.98**	**0.03**	**25.29**	**0.91**	**0.08**
	Lego			Materials		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GNeRF 90°	22.83	0.83	0.25	22.58	0.85	0.20
GNeRF 180° (2DOF)	9.78	**0.78**	0.53	9.48	0.65	0.50
Ours (3DOF)	**15.90**	0.72	**0.20**	**29.73**	**0.96**	**0.03**

Table 5. **Novel view synthesis on Blender *ordered image collections*.** The relative improvement of our method with respect to GNeRF is larger with an ordered image collection, since we avoid the difficult step of building the initial graph.

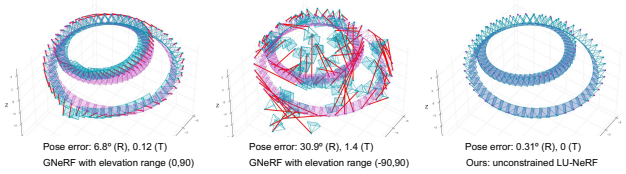

Figure 6. **Pose estimation on the Blender Materials *ordered image collection*.**

when the elevation angles are uniformly sampled from a 90° interval; however, its performance drops significantly when the range of elevation is enlarged. Our method outperforms GNeRF in most scenes when the prior pose distribution is unknown, since we do not require any prior knowledge of the camera poses. Fig. 4 provides the visualization of the estimated camera poses from GNeRF under different prior pose distributions and our method.

Tab. 3 shows the number of images COLMAP registers out of 100 in each scene. COLMAP is sensitive to image resolution, and its performance drops significantly on low-resolution images. For instance, COLMAP only registers 15 images out of 100 on the Mic scene when the image size is 400×400. Our method provides accurate pose estimation for all cameras given 400×400 images. Tab. 1 also reports how COLMAP performs in the pose estimation task on the Blender scenes. We use the most favorable settings for COLMAP – 800×800 images and set the poses of unregistered cameras to the poses of the nearest registered camera, assuming the *optimal graph* is known, while our method makes no such assumption. Nevertheless, our

	Bike	Chair	Cup	Laptop	Shoe	Book
Rotation:						
COLMAP	–	17.2	–	–	14.1	–
COLMAP-SPSG	129	28.3	–	–	**8.3**	–
COLMAP-LoFTR	**1.1**	6.7	6.3	**9.5**	14.5	83.4
Ours	15.6	**2.6**	**6.1**	17.8	8.8	**3.2**
Translation:						
COLMAP	–	0.04	–	–	**0.03**	–
COLMAP-SPSG	1.71	0.12	–	–	0.04	–
COLMAP-LoFTR	**0.10**	0.07	**0.03**	0.34	0.14	0.67
Ours	0.13	**0.03**	0.11	**0.16**	0.20	**0.03**

Table 6. **Comparison with COLMAP on Objectron [1].** We report rotation (°) and translation errors on select scenes from Objectron that are challenging to COLMAP. **"–"** denotes failure to estimate any camera poses. **COLMAP-SPSG** is an improved version [46] with SuperPoint [21] and SuperGLUE [47] as descriptor and matcher, respectively. **COLMAP-LoFTR** improves COLMAP with LoFTR [51], a detector-free feature matcher. Translation errors are in the scale of the ground truth scene.

model achieves better performance than COLMAP in some scenes, even when a BARF refinement is applied to initial COLMAP results. This shows that LU-NeRF complements COLMAP by working in scenes where COLMAP fails.

Novel view synthesis. Fig. 5 and Tab. 2 show our results in the task of novel view synthesis on unordered image collections. The results are consistent with the quantitative pose evaluation – our model outperforms both VMRF and GNeRF when no priors on pose distribution are assumed.

4.2. Ordered Image Collections

4.3. Blender

Tab. 4, Tab. 5, and Fig. 6 summarize the results on the Blender *ordered image collection*. Our method outperforms GNeRF with both constrained and unconstrained pose distributions even though the elevation of the cameras in this dataset is constrained. Our method utilizes the image order to build a connected graph and does not make any assumptions about the camera distribution. Results in Tab. 5 show that the view synthesis results are in sync with the pose estimation results. GNeRF degrades significantly under unconstrained pose priors, while our method outperforms GNeRF consistently across different scenes.

4.4. Objectron

We further compare with COLMAP on real images from the Objectron dataset. COLMAP can be improved with modern feature extraction and matching algorithms [46] such as SuperPoint [21] and SuperGLUE [47] (denoted COLMAP-SPSG), or LoFTR [51] (denoted COLMAP-LoFTR), but these still struggle in scenes with little or repeated texture. Tab. 6 and Fig. 7 show our results *without BARF refinement* on difficult scenes from Objectron.

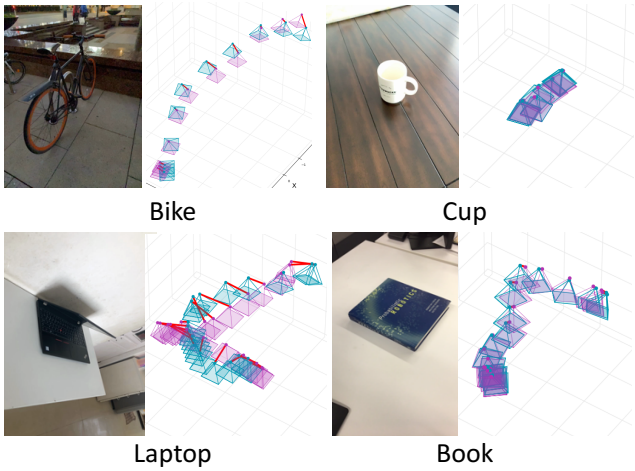

Figure 7. **Camera pose estimation on textureless scenes.** COLMAP fails to register any cameras in these Objectron scenes. Ground truth cameras are in purple, our predictions in blue.

Ambiguity	Chair	Hotdog	Lego	Mic	Drums
w/o resolution	39.14	138.9	0.48	107.9	11.35
w/ resolution	**4.24**	**0.23**	**0.07**	**0.84**	**0.05**

Table 7. **Mirror symmetry ambiguity.** The mean rotation error in degrees for our pipeline (starting with the optimal graph), with and without the proposed strategy to resolve the ambiguity.

4.5. Analysis

This section provides additional analysis of our approach. All the experiments discussed below were conducted on the unordered image collection. See the supplementary material for an extended discussion.

Mirror symmetry ambiguity. Tab. 7 shows the performance of our full method with and without the proposed solution to the mirror-symmetry ambiguity (Sec. 3.2). Resolving the ambiguity improves performance consistently, confirming the importance of this component to our pipeline. For closer inspection, we present qualitative results for LU-NeRF with and without ambiguity resolution for select mini-scenes in Fig. 8. Fig. 8 presents a visual comparison between LU-NeRF with and without the proposed solution to the mirror-symmetry ambiguity. Without the ambiguity resolution, the predicted depths are reflected across a plane parallel to the image plane (having the effect of inverted disparity maps), and the poses are reflected across the center camera of a mini-scene. Our LU-NeRF-2 rectifies the predicted geometry and local camera poses, which effectively resolves the ambiguity.

5. Discussion

In this work, we propose to estimate the neural scene representation and camera poses jointly from an unposed image collection through a process of synchronizing local unposed NeRFs. Unlike prior works, our method does not rely on a proper prior pose distribution and is flexible enough to operate in general $\mathbf{SE}(3)$ pose settings. Our framework works reliably in low-texture or low-resolution images and thus complements the feature-based SfM algorithms. Our pipeline also naturally exploits sequential image data, which is easy to acquire in practice.

One limitation of our method is the computational cost, which can be relieved by recent advances in neural rendering [54]. Another limitation is the difficulty in building graphs for unordered scenes, which is a promising direction for future work.

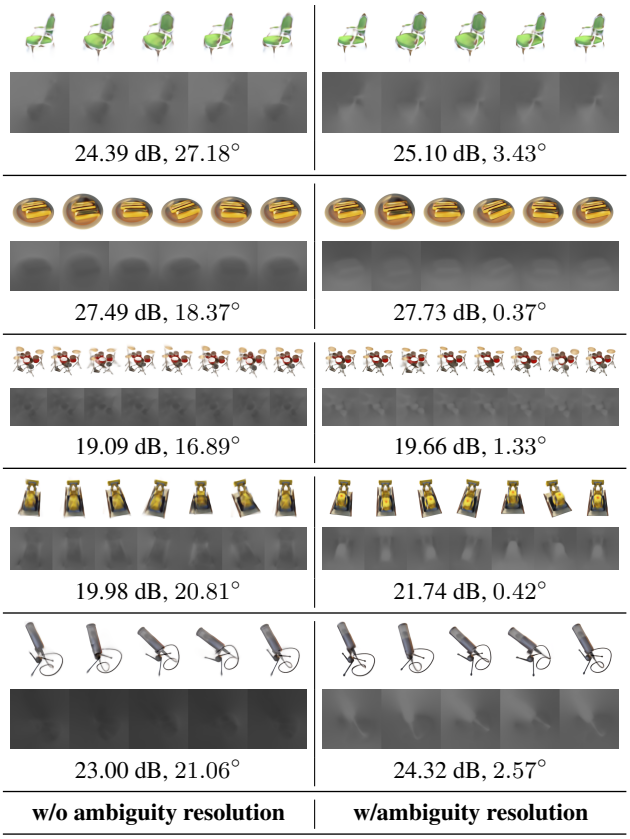

Figure 8. **Mirror symmetry ambiguity.** For specific mini-scenes, we present renderings, disparity maps, PSNRs between the renderings and the groundtruth, and relative rotation errors (*lower is better*) for LU-NeRF with and without the proposed solution to the mirror-symmetry ambiguity. Brightness is inversely related to depth in the disparity map. The groundtruth depth maps are not available with the dataset.

6. Acknowledgements

We thank Zhengqi Li and Mehdi S. M. Sajjadi for fruitful discussions. The research is supported in part by NSF grants #1749833 and #1908669. Our experiments were partially performed on the University of Massachusetts GPU cluster funded by the Mass. Technology Collaborative.

References

[1] Adel Ahmadyan, Liangkai Zhang, Artsiom Ablavatski, Jianing Wei, and Matthias Grundmann. Objectron: A large scale dataset of object-centric videos in the wild with pose annotations. In *CVPR*, 2021. 6, 7

[2] Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep vit features as dense visual descriptors. *arXiv preprint arXiv:2112.05814*, 2021. 5

[3] Federica Arrigoni, Beatrice Rossi, and Andrea Fusiello. Spectral synchronization of multiple views in se(3). *SIAM Journal on Imaging Sciences*, 2016. 4

[4] 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. In *ICCV*, 2021. 2

[5] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 2

[6] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *IJCV*. 4

[7] Brojeshwar Bhowmick, Suvam Patra, Avishek Chatterjee, Venu Madhav Govindu, and Subhashis Banerjee. Divide and conquer: Efficient large-scale structure from motion using graph partitioning. In *ACCV*, 2015. 2

[8] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. *arXiv preprint arXiv:2212.07388*, 2022. 2, 3

[9] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *NeurIPS*, 2022. 2

[10] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 5

[11] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *CVPR*, 2021. 1, 3

[12] Di Chen, Yu Liu, Lianghua Huang, Bin Wang, and Pan Pan. GeoAug: Data augmentation for few-shot nerf with geometry constraints. In *ECCV*, 2022. 1, 3

[13] Shu Chen, Yang Zhang, Yaxin Xu, and Beiji Zou. Structure-aware nerf without posed camera via epipolar constraint. *CoRR*, abs/2210.00183, 2022. 2

[14] Yu Chen, Shuhan Shen, Yisong Chen, and Guoping Wang. Graph-based parallel large scale structure from motion. *Pattern Recognition*, 2020. 2

[15] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. GARF: Gaussian activated radiance fields for high fidelity reconstruction and pose estimation. In *ICCV*, 2021. 2, 3, 4

[16] Ondrej Chum, Tomás Werner, and Jiri Matas. Two-view geometry estimation unaffected by a dominant plane. In *CVPR*, 2005. 4

[17] Andrea Porfiri Dal Cin, Luca Magri, Federica Arrigoni, Andrea Fusiello, and Giacomo Boracchi. Synchronization of group-labelled multi-graphs. In *ICCV*, 2021. 2

[18] Mihai Cucuringu, Yaron Lipman, and Amit Singer. Sensor network localization by eigenvector synchronization over the euclidean group. *TOSN*, 2012. 2

[19] Frank Dellaert, David M. Rosen, Jing Wu, Robert Mahony, and Luca Carlone. Shonan rotation averaging: Global optimality by surfing $so(p)^n$. In *ECCV*, 2020. 2, 4

[20] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, June 2022. 1, 3

[21] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description, 2018. 2, 7

[22] Meiling Fang, Thomas Pollok, and Chengchao Qu. Merge-sfm: Merging partial reconstructions. In *BMVC*, 2019. 2

[23] Richard Hartley, Jochen Trumpf, Yuchao Dai, and Hongdong Li. Rotation Averaging. *IJCV*, 101(2), 2013. 2

[24] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. 2003. 2

[25] Eldar Insafutdinov and Alexey Dosovitskiy. Unsupervised learning of shape and pose with differentiable point clouds. In *NeurIPS*, 2018. 3

[26] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. Rotationnet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *CVPR*, 2018. 3

[27] Jonáš Kulhánek, Erik Derner, Torsten Sattler, and Robert Babuška. ViewFormer: NeRF-free neural rendering from few images using transformers. In *ECCV*, 2022. 1, 3

[28] Axel Levy, Mark Matthews, Matan Sela, Gordon Wetzstein, and Dmitry Lagun. MELON: Nerf with unposed images using equivalence class estimation. *arXiv:preprint*, 2023. 3

[29] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-adjusting neural radiance fields. In *ECCV*, 2022. 1, 2, 3, 4, 6

[30] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *ICCV*, 2021. 2

[31] David G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 2

[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. In *ICCV*, 2021. 1, 2, 3, 5, 6

[33] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. *arXiv preprint arXiv:2303.13791*, 2023. 2, 3

[34] 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, 4, 5

[35] Tom Monnier, Matthew Fisher, Alexei A. Efros, and Mathieu Aubry. Share With Thy Neighbors: Single-View Reconstruction by Cross-Instance Consistency. In *ECCV*, 2022. 3

[36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 2022. 2

[37] Raúl Mur-Artal, J. M. M. Montiel, and Juan D. Tardós. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015. 2

[38] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 1, 3, 4

[39] Kemal Egemen Ozden, Konrad Schindler, and Luc Van Gool. Multibody structure-from-motion in practice. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010. 1, 4

[40] Onur Özyeşil, Vladislav Voroninski, Ronen Basri, and Amit Singer. A survey of structure from motion. *Acta Numerica*, 26, 2017. 1, 2

[41] Matteo Poggi, Pierluigi Zama Ramirez, Fabio Tosi, Samuele Salti, Stefano Mattoccia, and Luigi Di Stefano. Cross-spectral neural radiance fields. *arXiv preprint arXiv:2209.00648*, 2022. 2

[42] David M Rosen, Luca Carlone, Afonso S Bandeira, and John J Leonard. SE-Sync: A certifiably correct algorithm for synchronization over the special euclidean group. *IJRR*, 2019. 2, 4

[43] Antoni Rosinol, John J Leonard, and Luca Carlone. Nerf-slam: Real-time dense monocular slam with neural radiance fields. *arXiv preprint arXiv:2210.13641*, 2022. 2

[44] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *ICCV*, 2011. 2

[45] Mehdi S. M. Sajjadi, Aravindh Mahendran, Thomas Kipf, Etienne Pot, Daniel Duckworth, Mario Lučić, and Klaus Greff. RUST: Latent Neural Scene Representations from Unposed Imagery. *CVPR*, 2023. 3

[46] Paul-Edouard Sarlin, Cesar Cadena, Roland Siegwart, and Marcin Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *CVPR*, 2019. 7

[47] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 7

[48] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 1, 2, 5, 6

[49] Cameron Smith, Yilun Du, Ayush Tewari, and Vincent Sitzmann. Flowcam:training generalizable 3d radiance fields without camera poses via pixel-aligned scene flow. *arXiv preprint arXiv:2306.00180*, 2023. 3

[50] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J Davison. imap: Implicit mapping and positioning in real-time. In *ICCV*, 2021. 2

[51] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. In *CVPR*, 2021. 2, 7

[52] Rick Szeliski and Sing Bing Kang. Shape ambiguities in structure from motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1997. 4

[53] Takafumi Taketomi, Hideaki Uchiyama, and Sei Ikeda. Visual slam algorithms: A survey from 2010 to 2016. *IPSJ Transactions on Computer Vision and Applications*, 2017. 2

[54] A. Tewari, J. Thies, B. Mildenhall, P. Srinivasan, E. Tretschk, W. Yifan, C. Lassner, V. Sitzmann, R. Martin-Brualla, S. Lombardi, T. Simon, C. Theobalt, M. Nießner, J. T. Barron, G. Wetzstein, M. Zollhöfer, and V. Golyanik. Advances in neural rendering. *CGF*, 2022. 1, 8

[55] Shubham Tulsiani, Alexei A. Efros, and Jitendra Malik. Multi-view consistency as supervisory signal for learning shape and pose prediction. In *CVPR*, 2018. 3

[56] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 1, 2, 3

[57] Shangzhe Wu, Christian Rupprecht, and Andrea Vedaldi. Unsupervised learning of probably symmetric deformable 3d objects from images in the wild. In *CVPR*, 2020. 3

[58] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. *arXiv preprint arXiv:2210.04553*, 2022. 2, 3

[59] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *CGF*, 2022. 1

[60] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1, 3

[61] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. Vmrf: View matching neural radiance fields. In *ACM MM*, 2022. 2, 3, 5, 6

[62] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, June 2018. 6

[63] Lei Zhou, Zixin Luo, Mingmin Zhen, Tianwei Shen, Shiwei Li, Zhuofei Huang, Tian Fang, and Long Quan. Stochastic bundle adjustment for efficient and scalable 3d reconstruction. In *ECCV*, 2020. 2

[64] Siyu Zhu, Runze Zhang, Lei Zhou, Tianwei Shen, Tian Fang, Ping Tan, and Long Quan. Very large-scale global sfm by distributed motion averaging. In *CVPR*, 2018. 2

[65] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *CVPR*, 2022. 2