Supplementary Material for “Local-to-Global Registration for Bundle-Adjusting Neural Radiance Fields”

Yue Chen[1,2*] Xingyu Chen[1,2*] Xuan Wang[3†] Qi Zhang[4]
Yu Guo[1,2†] Ying Shan[4] Fei Wang[1,2]
[1] National Key Laboratory of Human-Machine Hybrid Augmented Intelligence
[2] IAIR, Xi’an Jiaotong University [3]Ant Group [4]Tencent AI Lab

Scene	BARF	Ours	ref. NeRF
Synthetic objects	08:18	04:35	04:30
Real-World scenes	10:38	07:42	07:25

Table 1. Average training time (hh:mm).

Scene	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
n_r	0.01	0.05	0.03	0.04	0.07	0.04	0.04	0.09
n_t	0.4	0.5	0.3	0.4	0.5	0.3	0.5	0.7

Table 2. Multiplier of pose perturbation for synthetic scenes.

Scene	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	T-rex
λ	$1{\times}10^2$	$1{\times}10^3$	$1{\times}10^5$	$1{\times}10^5$	$1{\times}10^2$	$1{\times}10^2$	$1{\times}10^5$	$1{\times}10^5$

Table 3. Multiplier λ of global alignment objective.

Here we provide more implementation details and experimental results. We encourage readers to view the supplementary video for an intuitive experience about different types of bundle-adjusting neural radiance fields.

A. Additional Details

A.1. Time Consumption

We implement all experiments on a single NVIDIA GeForce RTX 2080 Ti GPU. As shown in Table 1, L2G-NeRF takes about 4.5 and 8 hours for training in synthetic objects and real-world scenes, respectively, while training BARF [1] takes about 8 and 10.5 hours. As a reference, we also compare time consumption against the ref. NeRF [3] trained under ground-truth poses (without the requirement of optimizing poses), showing that L2G-NeRF can achieve comparable time consumption. The time analysis indicates that calculating the gradient w.r.t. local pose (local-to-global registration) is more efficient than calculating the gradient w.r.t. global pose (global registration).

Scene	Camera pose registration: Rotation (°) ↓				Camera pose registration: Translation ↓				View synthesis quality: PSNR ↑				View synthesis quality: LPIPS ↓			
	$1{\times}10^2$	$1{\times}10^3$	$1{\times}10^4$	$1{\times}10^5$	$1{\times}10^2$	$1{\times}10^3$	$1{\times}10^4$	$1{\times}10^5$	$1{\times}10^2$	$1{\times}10^3$	$1{\times}10^4$	$1{\times}10^5$	$1{\times}10^2$	$1{\times}10^3$	$1{\times}10^4$	$1{\times}10^5$
Flower	0.44	**0.33**	\	\	0.30	**0.24**	\	\	24.59	**24.90**	\	\	0.18	**0.17**	\	\
Horns	0.36	0.24	0.23	**0.22**	0.80	0.57	0.32	**0.27**	22.51	22.84	22.82	**23.12**	0.28	0.28	0.27	**0.26**

Table 4. Ablation on the global alignment objective multiplier λ.

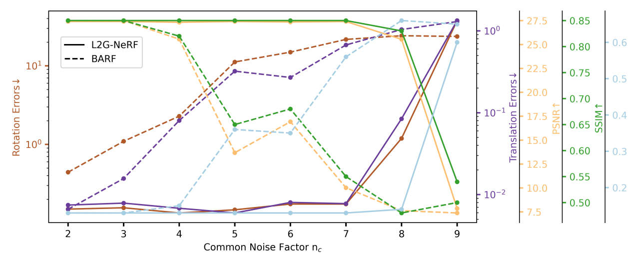

Figure 1. Convergence w.r.t. camera pose perturbation.

A.2. Camera Pose Perturbation

In all experiments, we always use the same initial conditions for all methods (fixed random seeds). For each object of synthetic scenes, we perturb the camera poses with additive noise as initial poses. Note that the way we add noise differs from [1], which perturbs ground-truth camera poses using left multiplication (transform cameras around the object’s center). Transformed cameras almost still face the object’s center, and the distances between the cameras and the object are almost unchanged. In contrast, we perturb ground-truth camera poses using right multiplication (transform cameras around themselves), thereby perturbing camera viewing directions (not always toward the object’s center) and camera positions (including the distances from them to the object), respectively.

The 6-DoF perturbation is parametrized by $\mathbf{T} = [\mathbf{R}|\mathbf{t}] \in \mathrm{SE}(3)$, where $\mathbf{R} \in \mathrm{SO}(3)$, $\mathbf{t} \in \mathbb{R}^3$, and $\mathbf{R}$ is generated by exponential map $\exp(\mathbf{r})$ from the Lie algebra $\mathfrak{so}(3)$ to the Lie group $\mathrm{SO}(3)$. The additive rotation noise $\mathbf{r} \in \mathfrak{so}(3)$ and translation noise $\mathbf{t} \in \mathbb{R}^3$ are distributed as $\mathbf{r} \sim \mathcal{N}(\mathbf{0}, n_r\mathbf{I})$ and $\mathbf{t} \sim \mathcal{N}(\mathbf{0}, n_t\mathbf{I})$, where the multiplier n_r and n_t are scene-dependent and given in Table 2.

*Authors contributed equally to this work.
†Corresponding Author.

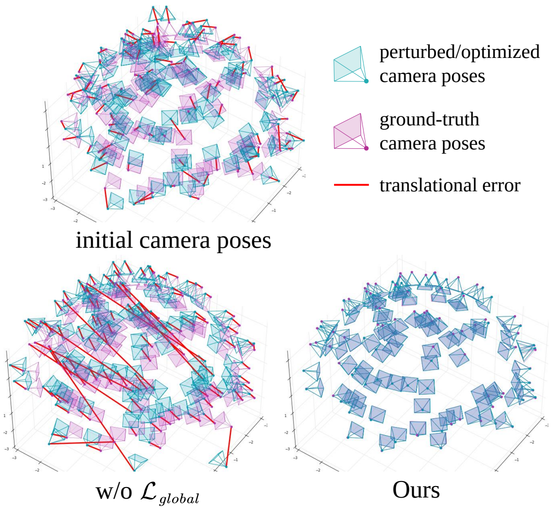

Figure 2. Visual comparison of ablation study about optimized camera poses (Procrustes aligned) for *hotdog* object. Full L2G-NeRF successfully aligns camera frames while w/o $\mathcal{L}_{global}$ gets stuck at suboptimal poses.

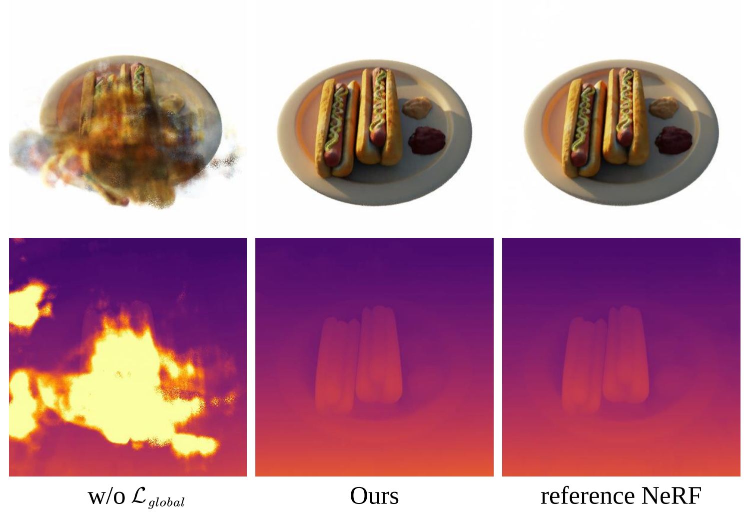

Figure 3. Ablation study of NeRF on *hotdog* synthetic object. The image synthesis and the expected depth are visualized with ray compositing in the top and bottom rows, respectively. Full L2G-NeRF achieves comparable rendering quality to the reference NeRF (trained using ground-truth poses), while ablation w/o $\mathcal{L}_{global}$ renders artifacts due to suboptimal registration.

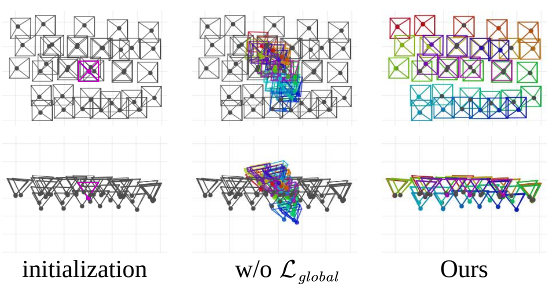

Figure 4. Visualization of ablation study about registration for *room* scene (Procrustes aligned). Results from L2G-NeRF highly agree with S*f*M [4] (colored in black), whereas w/o $\mathcal{L}_{global}$ results in suboptimal alignment.

Figure 5. Ablation study of NeRF on *room* real-world scenes from *unknown* camera poses. While L2G-NeRF can jointly optimize poses and scenes, L2G-NeRF produces high fidelity results, which is competitive to reference NeRF trained using S*f*M poses. Ablation w/o $\mathcal{L}_{global}$ diverges to wrong poses and hence produces ghosting artifacts.

Scene	Camera pose registration						View synthesis quality											
	Rotation (°) ↓			Translation ↓			PSNR ↑				SSIM ↑				LPIPS ↓			
	Global BARF	Local w/o $\mathcal{L}_g$	L2G Ours	Global BARF	Local w/o $\mathcal{L}_g$	L2G Ours	Global BARF	Local w/o $\mathcal{L}_g$	L2G Ours	ref. NeRF	Global BARF	Local w/o $\mathcal{L}_g$	L2G Ours	ref. NeRF	Global BARF	Local w/o $\mathcal{L}_g$	L2G Ours	ref. NeRF
Synthetic objects	7.02	3.63	**0.15**	29.84	14.34	**0.61**	20.51	22.70	**28.62**	29.42	0.82	0.85	**0.93**	0.94	0.22	0.14	**0.07**	0.06
Real-World scenes	0.55	23.82	**0.46**	0.33	10.66	**0.32**	24.23	20.71	**24.54**	22.44	0.73	0.64	**0.75**	0.65	0.23	0.33	**0.20**	0.29

Table 5. Quantitative results of ablation study about bundle-adjusting neural radiance fields. L2G-NeRF outperforms the local registration method (ablation w/o $\mathcal{L}_{global}$) and global registration method (BARF) on the average evaluation criteria of both synthetic objects and real-world scenes, which reveals the advantage of our local-to-global registration process. Translation errors are scaled by 100.

A.3. Convergence

We analyze the convergence of joint optimization on the *Ship* scene. We first set the base rotation noise multiplier n_r as 0.01 and the base translation noise multiplier n_t as 0.1, then linearly increased them by a common factor of $\{n_c\}_{n_c=2}^{9}$. As shown in Fig. 1, BARF fails to converge with n_c=4 (n_r=0.04,n_t=0.4) while L2G-NeRF fails to converge with n_c=8 (n_r=0.08,n_t=0.8). Moreover, we also analyze the influence of individual noise. Let n_r=0, BARF and L2G-NeRF can handle the largest n_t of 0.6 and 1.1, respectively. Let n_t=0, BARF and L2G-NeRF can handle the largest n_r of 0.16 and 0.25, respectively. In more noisy cases (such as random init), all methods cannot converge.

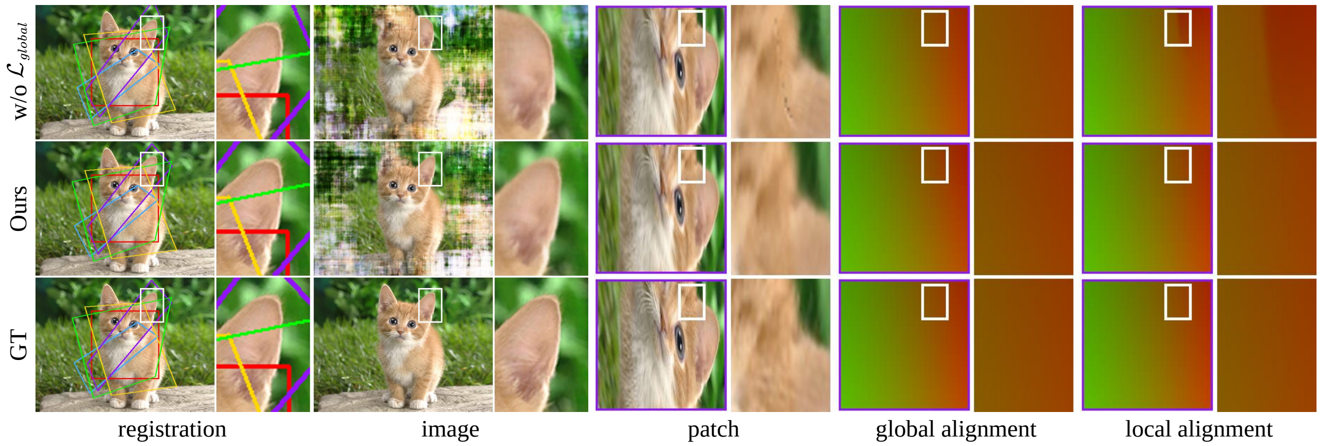

Figure 6. Ablation study of the neural image alignment experiment. Given color-coded image patches, we aim to recover the alignment and the neural field of the entire image. L2G-NeRF is able to find proper alignment and reconstruct high-fidelity neural image, while w/o $\mathcal{L}_{global}$ falls into false local alignments that do not obey the geometric constraint (global alignments), which results in ambiguous registration and distorted reconstruction (cat ears).

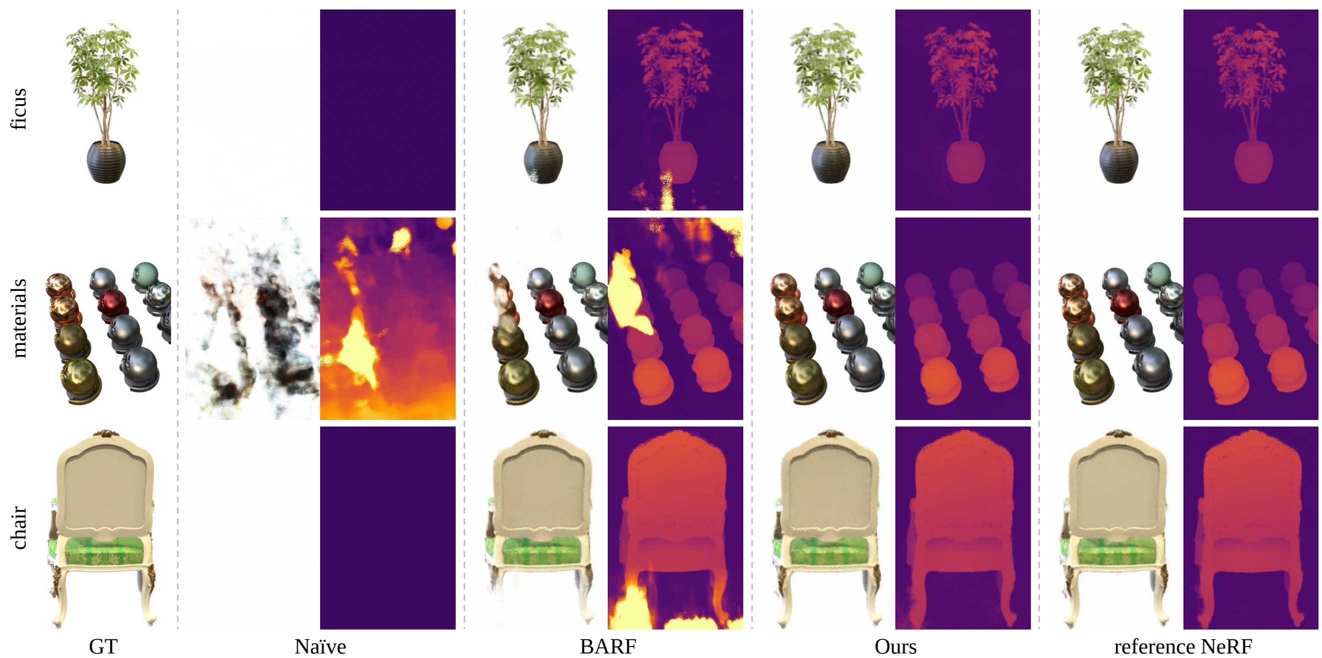

Figure 7. Additional qualitative results of bundle-adjusting neural radiance fields on synthetic scenes. The image synthesis and the expected depth are visualized with ray compositing in the left and right columns, respectively. While baselines render artifacts due to suboptimal solutions, L2G-NeRF achieves qualified visual quality, which is comparable to the reference NeRF trained using ground-truth poses.

A.4. Tuning Parameters

We set the multiplier λ of the global alignment objective to 1×10^2 for both the neural image alignment experiment and learning NeRF from imperfect camera poses with synthetic object-centric scenes. To further solve the challenging problem of learning NeRF in forward-facing LLFF scenes from *unknown* poses, we float the multiplier λ between 1×10^2 and 1×10^5 (summarized in Table 3) to achieve preferable results for specific scenes. As shown in Table 4, a larger λ encourages the model to emphasize geometric constraints more, achieving better accuracy but worse robustness (fails to converge on the *Flower* scene).

B. Ablation Studies

We propose a local-to-global registration method that combines the benefits of parametric and non-parametric methods. The key idea is to apply a pixel-wise alignment that optimizes photometric reconstruction errors

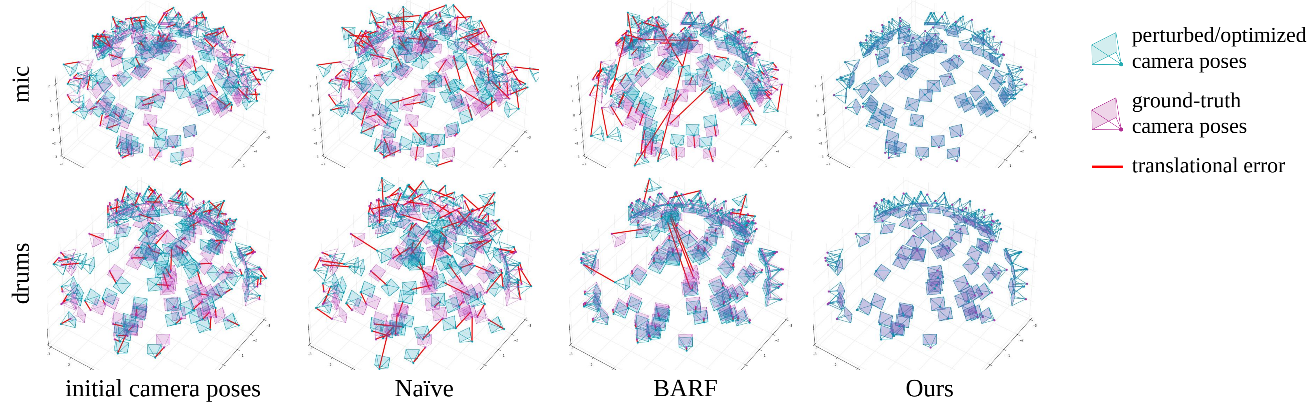

Figure 8. Additional visual comparison of the optimized camera poses (Procrustes aligned) for the *mic* and *drums* objects. L2G-NeRF successfully aligns all the camera frames while baselines get stuck at suboptimal solutions.

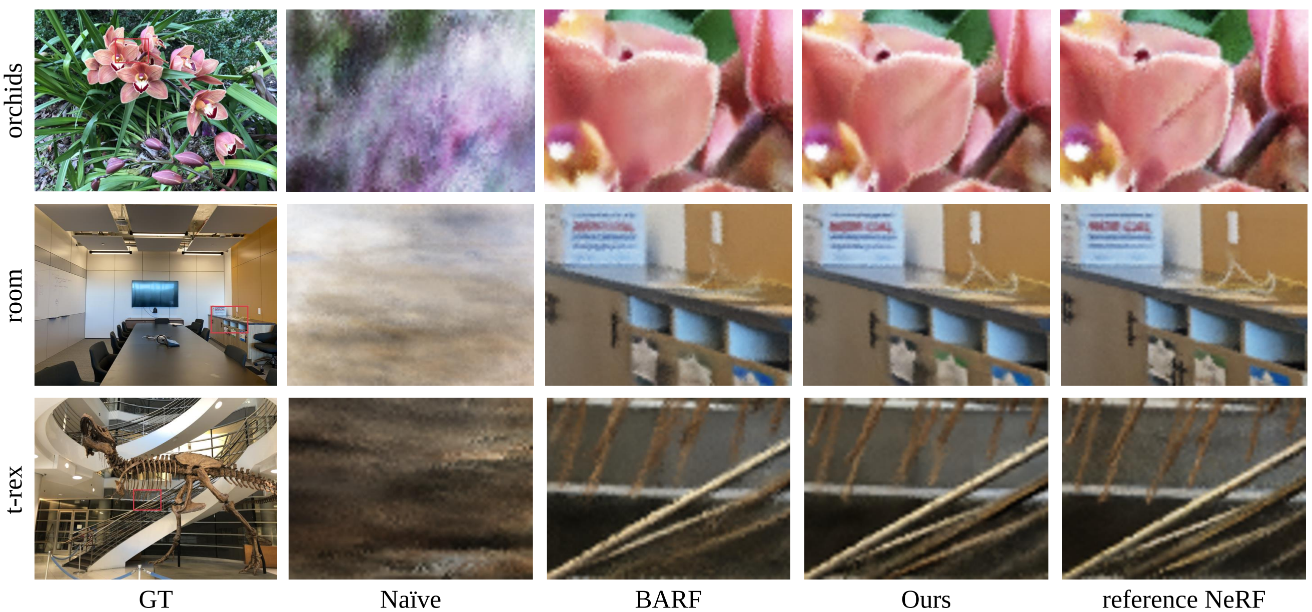

Figure 9. Additional novel view synthesis results of NeRF on real-world scenes (LLFF dataset) from *unknown* camera poses. L2G-NeRF can optimize for neural fields of higher quality than baselines, while achieving the comparable quality of the reference NeRF model that is trained under the camera poses provided by S*f*M [4].

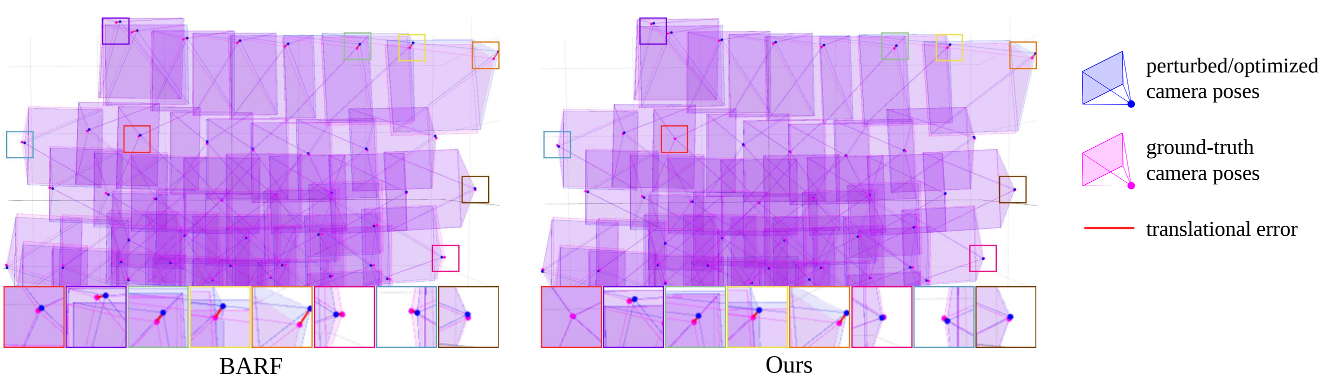

Figure 10. Visual comparison of the optimized camera poses (Procrustes aligned) for the *t-rex* real-world scene. L2G-NeRF successfully recovers the camera poses from *identity* transformation, which achieves fewer errors than BARF.

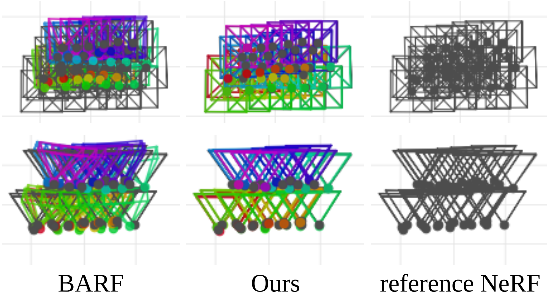

Figure 11. Visual comparison of optimized camera poses (Procrustes aligned) for the challenging *toys* scene captured under large displacements (hierarchical camera poses). L2G-NeRF successfully aligns all camera frames, which highly agrees with S*f*M [4] camera poses (colored in black), while BARF gets stuck at suboptimal solutions.

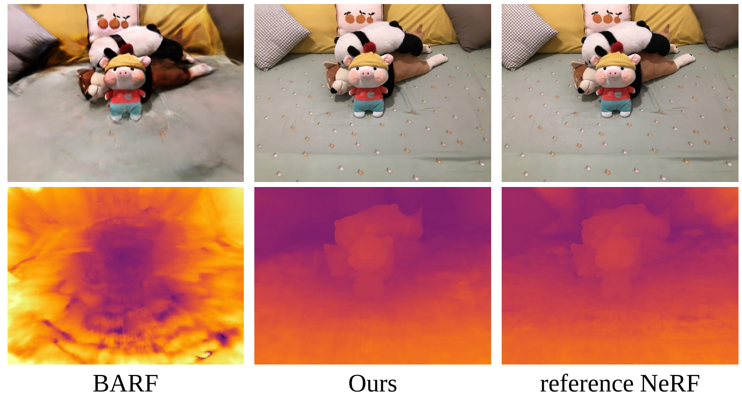

Figure 12. Results of NeRF on *toys* scene. L2G-NeRF achieves comparable synthesis quality to the reference NeRF (trained under S*f*M camera poses). But BARF fails to recover the proper geometry, which results in artifacts.

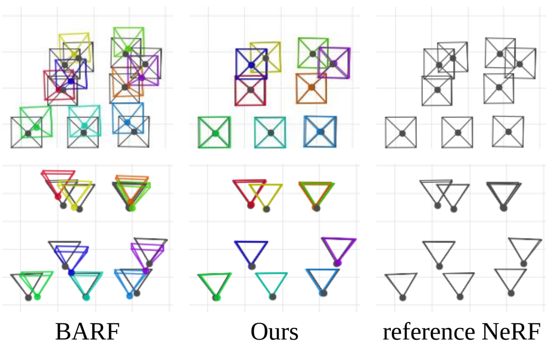

Figure 13. Visual comparison of optimized camera poses for the challenging *foods* scene captured under sparse views. Results from L2G-NeRF highly agree with S*f*M, whereas BARF results in suboptimal alignment.

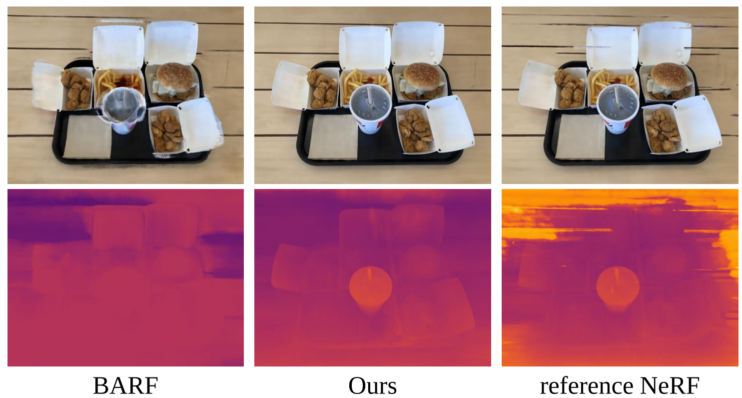

Figure 14. Results of NeRF on *foods* scene. L2G-NeRF outperforms BARF and even achieves better performance than reference NeRF in the scene where S*f*M [4] struggles with finding accurate registration from sparse views.

Scene	Camera pose registration: Rotation (°) ↓			Translation ↓			View synthesis quality: PSNR ↑				SSIM ↑				LPIPS ↓			
	Naïve	BARF	Ours	Naïve	BARF	Ours	Naïve	BARF	Ours	ref. NeRF	Naïve	BARF	Ours	ref. NeRF	Naïve	BARF	Ours	ref. NeRF
Toys	14.22	179.73	**0.42**	6.14	24.84	**0.33**	15.55	11.29	**29.58**	32.90	0.57	0.49	**0.94**	0.96	0.50	0.77	**0.06**	0.04
Foods	5.30	10.99	**0.31**	7.76	10.15	**0.62**	19.11	18.02	**31.83**	24.58	0.71	0.68	**0.95**	0.89	0.23	0.26	**0.05**	0.13

Table 6. Quantitative results of bundle-adjusting neural radiance fields on real-world scenes captured using an iPhone under large displacements (*toys*) or sparse views (*foods*). L2G-NeRF outperforms baselines and even achieves better performance than reference NeRF that trained under S*f*M poses in the Foods scene, which is hard for S*f*M to find accurate camera poses. Translation errors are scaled by 100.

$\sum_{i=1}^{M}\sum_{j=1}^{N}\left\|\mathcal{R}(\mathbf{T}_i^j\mathbf{x}^j;\mathbf{\Theta}) - \mathcal{I}_i(\mathbf{x}^j)\right\|_2^2$, followed by a frame-wise alignment $\sum_{i=1}^{M}\sum_{j=1}^{N}\lambda\left\|\mathbf{T}_i^j\mathbf{x}^j - \mathbf{T}_i^*\mathbf{x}^j\right\|_2^2$ to globally constrain the geometric transformations. We evaluate our proposed L2G-NeRF against an ablation (w/o $\mathcal{L}_{global}$), which builds upon our full model by eliminating the global alignment objective, *i.e.*, $\lambda = 0$. The ablation is equivalent to a local registration method, while BARF is the chosen representative global registration method, such that we unfold the comparison with both of them.

B.1. Ablation on NeRF (3D): Synthetic Objects

We first investigate the ablation study of learning NeRF from imperfect camera poses. We experiment with 8 synthetic object-centric scenes [3]. The results in Fig. 3 and Table 5 show that L2G-NeRF achieves better performance than the ablation w/o $\mathcal{L}_{global}$. Fig. 2 further illustrates that L2G-NeRF can achieve near-perfect registration while the ablation w/o $\mathcal{L}_{global}$ suffers from suboptimal solutions.

Ours BARF Ours BARF

Figure 15. Results of NeRF on reflective scenes.

B.2. Ablation on NeRF (3D): Real-World Scenes

We further explore the ablation study of employing NeRF to learn 3D neural fields in real-world scenes with *unknown* camera poses. We evaluate on the standard benchmark LLFF dataset [2]. Quantitative results are summarized in Table 5. The ablation w/o $\mathcal{L}_{global}$ diverges to wrong poses (visualized in Fig. 4), producing ghosting artifacts (shown in Fig. 5). L2G-NeRF outperforms the ablation w/o $\mathcal{L}_{global}$ and achieves high-quality view synthesis that is competitive to the reference NeRF.

B.3. Ablation on Neural Image Alignment (2D)

We further concrete analysis on the homography image alignment experiment and visualize the results in Fig. 6. Alignment with w/o $\mathcal{L}_{global}$ results in distorted artifacts (cat ears) in the recovered neural image due to ambiguous registration. This is the consequence of w/o $\mathcal{L}_{global}$'s attempt to directly optimize the pixel agreement metric, which minimizes photometric errors but does not obey the geometric constraint (global alignments). As L2G-NeRF discovers precise warps, it optimizes neural image with high fidelity.

C. Additional Results

C.1. NeRF (3D): Synthetic Objects

We report additional qualitative results of learning 3D NeRF from noisy camera poses for synthetic objects in Fig. 7. The baselines still perform poorly, while L2G-NeRF can achieve near-perfect registration (reflected in Fig. 8) and render images with comparable visual quality against reference NeRF that trained under ground-truth poses.

C.2. NeRF (3D): Real-World Scenes (LLFF)

We report additional qualitative results of learning NeRF for the standard LLFF dataset in Fig. 9, where camera poses are *unknown*. L2G-NeRF successfully recovers the 3D scene with higher fidelity than baselines. Fig. 10 shows that the recovered camera poses from L2G-NeRF agree more with those estimated from S*f*M methods than BARF.

C.3. NeRF (3D): Real-World Scenes (Ours)

We take one step further to experiment with images captured using an iPhone under challenging camera pose distribution. Fig. 11 and Fig. 13 indicate the advantage of L2G-NeRF in registering images captured under large displacements and sparse views, while baselines exhibit artifacts (Fig. 12 and Fig. 14) due to unreliable registration, which is reflected in Table 6. Moreover, the difficulty of registering from sparse views prevents S*f*M from finding accurate poses, which results in broken stripes on the synthesis of reference NeRF trained under S*f*M poses in *foods* scene. This further demonstrates the effectiveness of removing the requirement of pre-computed S*f*M poses. Fig. 11 and Fig. 13 show the largest displacements (hierarchical but adjacent camera poses) and the sparsest camera setting (9 views) of L2G-NeRF to register images in these scenes successfully, than which we can not handle a more challenging camera pose distribution.

C.4. NeRF (3D): Real-World Scenes (Shiny)

To analyze the influence of reflective surfaces, We present an example in Fig. 15 that models scenes [6] with reflections from identity initialization (L2G-NeRF converges, BARF fails in the *guitars* scene). Interestingly, global alignment loss increases by 4 to 10 times w.r.t. other datasets. This may be caused by inaccurate local registration in specular regions, and our convergence benefits from the global registration constraint. Specific methods (e.g., [5]) could be employed to handle reflective surfaces better.

References

[1] 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] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 6

[3] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf:

Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, 2020. 1, 5

[4] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 2, 4, 5

[5] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5481–5490. IEEE, 2022. 6

[6] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2021. 6