DOI: 10.1111/cgf.70012
EUROGRAPHICS 2025 / A. Bousseau and A. Dai
(Guest Editors)

COMPUTER GRAPHICS *forum*
Volume 44 (2025), Number 2

NoPe-NeRF++: Local-to-Global Optimization of NeRF with No Pose Prior

Dongbo Shi[†,1] Shen Cao[†,2] Bojian Wu[2] Jinhui Guo[2] Lubin Fan[‡,2] Renjie Chen[‡,1] Ligang Liu[1] Jieping Ye[2]

[1]University of Science and Technology of China, [2]Independent Researcher

Figure 1: *In the context of unknown camera poses, accurately training high-quality NeRF while estimating the camera pose is a challenging task. The previous state-of-the-art method, NoPe-NeRF, focuses on local constraints between adjacent images and can only handle simple camera motions; it completely fails in more complex scenarios. To tackle this issue, we introduce a local-to-global training strategy, termed as NoPe-NeRF++, which considers both local position estimation and global geometric consistency, achieving not only stable estimates but also improved rendering quality and more accurate pose estimation, as demonstrated in our results.*

Abstract

In this paper, we introduce NoPe-NeRF++, a novel local-to-global optimization algorithm for training Neural Radiance Fields (NeRF) without requiring pose priors. Existing methods, particularly NoPe-NeRF, which focus solely on the local relationships within images, often struggle to recover accurate camera poses in complex scenarios. To overcome the challenges, our approach begins with a relative pose initialization with explicit feature matching, followed by a local joint optimization to enhance the pose estimation for training a more robust NeRF representation. This method significantly improves the quality of initial poses. Additionally, we introduce global optimization phase that incorporates geometric consistency constraints through bundle adjustment, which integrates feature trajectories to further refine poses and collectively boost the quality of NeRF. Notably, our method is the first work that seamlessly combines the local and global cues with NeRF, and outperforms state-of-the-art methods in both pose estimation accuracy and novel view synthesis. Extensive evaluations on benchmark datasets demonstrate our superior performance and robustness, even in challenging scenes, thus validating our design choices.

CCS Concepts

• ***Computing methodologies*** → ***Image-based rendering;***

1. Introduction

Given a set of RGB images of a 3D scene and the corresponding camera parameters, Neural Radiance Fields (NeRF) [MST*20] can

† These authors contributed equally to this work.
‡ Corresponding authors

© 2025 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

delivered by EUROGRAPHICS DIGITAL LIBRARY www.eg.org diglib.eg.org

represent the 3D scene using an implicit neural representation and synthesize high-quality images from novel viewpoints. Due to its strong capability for restoring geometry and appearance, NeRF has already been widely used in various tasks, such as 3D reconstruction[OPG21; TCY*22], editing[YSL*22; WCH*22], appearance modeling[ZSD*21; ZSH*22], SLAM[ZPL*22; MYZ*24], etc. Despite this, it is well-known that before training NeRF, we typically need to obtain highly accurate camera parameters for each input viewpoint, which are usually obtained from COLMAP [SF16]. This approach of providing camera parameters through preprocessing presents several issues. For example, the dependency on externally supplied camera parameters limits its real scalability and applicability. Moreover, the poses remain fixed during training, making the introduced pose noise unavoidable and impossible to mitigate. This negatively impacts the training results, especially in scenes with substantial camera motions. To address these issues, recent works such as NeRFmm [WWX*21], BARF [LMTL21], SC-NeRF [JAC*21], and NoPe-NeRF [BWL*23] focus on simultaneously optimizing camera parameters and the neural implicit representation of the scene. However, these methods still make strong assumptions, including the requirement for a forward-facing scene, the ordered images, and minimal changes between adjacent frames. Among these, NoPe-NeRF stands out as one of the most notable recent contributions. Similarly, without relying on any prior poses, the algorithm introduces monocular depth and relative poses to create *local* loss constraints, and train a more accurate neural implicit representation of the scene. Nevertheless, NoPe-NeRF still struggles with complex scenarios, such as significant frame-to-frame camera motion changes and long-term variations. Traditionally, to simultaneously obtain camera poses and geometry from unposed images, the Structure from Motion (SfM) [SSS06; ASS*09; SF16] is used, which typically resorts to a local-to-global strategy to achieve the robust results. Basically, the local matching algorithms first roughly estimate the relative poses. The bundle adjustment is then applied to optimize and obtain a globally optimal solution. Generally, local estimation is a crucial step as it provides reasonable initial values for the subsequent global optimization. In particularly challenging cases, poor initialization can cause global optimization to fail entirely. Similarly, global optimization is also essential because relying solely on local optimization can lead to cumulative errors that cannot be corrected. Therefore, we claim that both local and global components are indispensable for this task. This is why NoPe-NeRF performs poorly in challenging cases, as it only considers local information and neglects global constraints. To address the aforementioned problems, following the previous works, we took a step further and proposed a novel joint optimization algorithm to improve the quality of NeRF and accuracy of camera parameters, term as **NoPe-NeRF++**. Concretely, we adopted a local-to-global strategy to achieve the goal. Before that, we firstly created a maximum spanning tree to explicitly determine the association relationships among all the input images with feature matching, where the local matching pairs and the global tracks can be extracted by traversing the tree. Next, during the *local joint optimization* phase, by leveraging the local matching information, we initialized the camera poses similar with NoPe-NeRF, but enhanced the estimation with random matching schemes in order to reduce accumulated errors. Additionally, inspired by bundle adjustment [DRMS07], we introduced a *global joint optimization* to further improve the results, by combining the track trajectories and multi-view geometric consistency constraints with NeRF training, which enforces the network to reach a global optimal result. Compared to previous methods, our algorithm achieves the state-of-the-art performance on both camera pose estimation results and the photo-realistic rendering quality of NeRF, under the same input conditions. Even with more complex camera motions, our proposed method still demonstrate the strongest robustness. The contributions can be summarized as follows. Codes and datasets will be made public upon acceptance.

- We proposed a local-to-global NeRF model optimization method without pose priors, the seamless combination of the two phases demonstrate strong stability and interpretability. To the best of our knowledge, this is the first work for the task that simultaneously considers the integration of both local and global cues.
- Compared to NoPe-NeRF, during local optimization phase, we enhanced the estimated poses with random matching schemes to improve the stability of training.
- Moreover, we incorporated the global image feature-based trajectories into joint optimization to further mitigate the accumulated noises and errors.
- Our algorithm achieved SOTA results in both camera pose estimation and synthetic image quality across public datasets. Furthermore, extensive quantitative and qualitative analyses demonstrate that our algorithm exhibits superior robustness.

2. Related Work

2.1. Novel View Synthesis

The goal of novel view synthesis is to generate unseen views based on the images taken from the existing views. It is usually achieved by interpolating or warping images from the neighboring point of views. Traditionally, this problem is posed as image-based rendering [CW93; GGSC96; SD96] or 3D geometry based rendering, where different representations of the 3D scene include multi-plane images [LFS*21; TS20; ZTF*18], planes [HEH05; HAA97], point clouds [ZBRH22; XXP*22] and mesh [HRBP21; RK20; RK21], etc.

Recently, neural implicit representations have revolutionized this field. The pioneering work, NeRF [MST*20], models a scene as radiance field parameterized by a deep neural network, and predicts volume density and view-dependent radiance based on spatial location and viewing direction. To enhance the quality and efficiency, significant advancements have been made, such as reducing aliasing effects [BMT*21], boosting training efficiency [MESK22], and enabling sparse viewpoint synthesis [KSH22]. Additionally, 3D Gaussian Splatting (3DGS) [KKLD23] has dramatically sped up rendering during both the training and inference phases by using explicit Gaussian kernels to represent 3D scenes.

Despite the impressive advancements in novel view synthesis brought by NeRF and 3DGS, these methods still depend on offline COLMAP [SF16] to obtain precise intrinsic and extrinsic camera parameters. This reliance significantly limits their practical use in real-world applications.

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

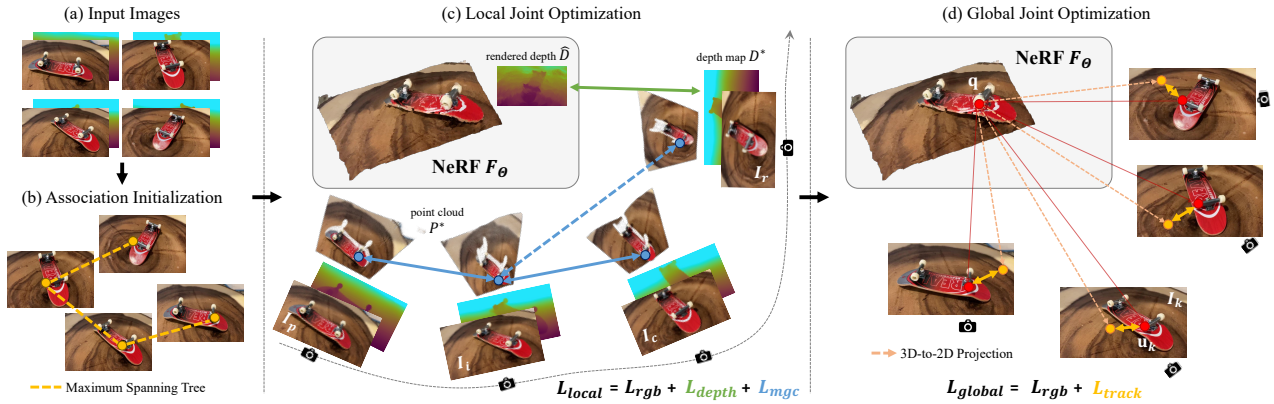

Figure 2: ***Overview of NoPe-NeRF++.*** *Given a set of images and their corresponding camera intrinsics (a), our method first establishes the association relationships among the images using a maximum spanning tree (b). Next, we apply local optimization with photometric, depth, and multiview geometric constraints to estimate the camera poses and train the NeRF model jointly (c). Finally, these estimates are further refined globally by considering photometric differences and geometric consistency in feature matching trajectories (d).*

2.2. NeRF with Pose Optimization

NeRF can deliver impressive scene representation when prior camera parameters are accurate. However, when this information is inaccurate or unavailable, estimating camera poses and neural representations becomes crucial. iNeRF [YFB*21] is one of the early works that uses NeRF for pose estimation, treating it as an inverse NeRF process. Given initial camera poses, it fixes NeRF parameters and minimizes the photometric loss function to achieve its goal. Following this, BARF [LMTL21] applies a coarse-to-fine strategy, yielding promising results. In addition to optimizing camera poses, NeRFmm [WWX*21] also estimates camera intrinsics during the optimization process. However, both BARF and NeRFmm are limited to achieving satisfactory optimization results only within a small range of motion, which restricts their broader applicability. L2G-NeRF [CCW*23] employs a local-to-global registration strategy to relax the accuracy requirements of input poses during initialization. It calculates the local pose for each image sample and uses least squares to determine the globally optimal poses from each viewpoint. LU-NeRF [CEJ*23] initially applies BARF to a selected set of 3-5 frames to develop local models with relative poses, followed by an offline pose synchronization process for global alignment. SAMURAI [BEK*22] and SHINOBI [ERB*24] are also joint optimizations of NeRF and poses that focus on improving BARF in challenging cases. LocalRF [MLG*23] provides a robust solution for large-scale scenes from a single casually captured video, by progressively optimizing multiple overlapping local NeRF and poses. CamP [PHM*23] and INN [CGSL24] improve the performance of NeRF with pose estimation by changing the representations of camera parameters in NeRF. To incorporate additional global constraints, TrackNeRF [MZR*24] simultaneously optimizes sparse NeRF and camera poses based on feature tracks, achieving promising results in noisy and sparse views.

To tackle the challenges of NeRF training when camera poses are unknown, SC-NeRF [JAC*21] introduces a projected ray distance loss based on correspondences, optimizing it alongside photometric errors. This approach allows SC-NeRF to learn scene representations while simultaneously self-calibrating camera intrinsic and extrinsic parameters, as well as distortion coefficients, from scratch. However, SC-NeRF relies on pair matches to construct the projected ray distance, resulting in a lack of global constraints. NoPe-NeRF [BWL*23] leverages monocular depth estimation priors provided by DPT [RBK21], using undistorted depth constraints to define the relative poses between consecutive frames. This enables the joint optimization of camera parameters and NeRF. Combined with 3D Gaussian Splatting, CF-3DGS [FLK*24] processes input frames sequentially, progressively constructing the 3D Gaussian set by analyzing one frame at a time, where the camera parameters and 3D Gaussians are jointly optimized. This method achieves realistic scene representation and precise camera pose estimation under large camera motion changes. However, both NoPe-NeRF and CF-3DGS lack support for processing arbitrary unordered images, limiting their broader applications.

3. Preliminaries

Neural Radiance Fields (NeRF) [MST*20] have been proven to be an effective representation to learn dense 3D geometry of a scene, which use a neural network $F_\Theta : (\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ to map a 3D point $\mathbf{x} \in \mathbb{R}^3$ and a viewing direction $\mathbf{d} \in \mathbb{R}^3$ to a radiance color $\mathbf{c} \in \mathbb{R}^3$ and a volume density σ. Later, to render images of specified camera poses, we obtain the color $\hat{C}$ by aggregating the radiance colors on camera ray $\mathbf{r}(s) = \mathbf{o} + s\mathbf{d}$ shooting from the camera position $\mathbf{o}$ along the direction $\mathbf{d}$ between near and far bound s_n and s_f. This volume rendering equation can be formulated as:

$$\hat{C}(\mathbf{r}) = \int_{s_n}^{s_f} T(s)\sigma(\mathbf{r}(s))\mathbf{c}(\mathbf{r}(s), \mathbf{d})ds, \tag{1}$$

where $T(t) = \exp(-\int_{s_n}^{t} \sigma(\mathbf{r}(s))ds)$ is the accumulated transmittance along ray $\mathbf{r}$.

Given N images $\mathcal{I} = \{I_i\}_{i=0}^{N-1}$ and corresponding camera parameters $\Pi = \{\pi_i\}_{i=0}^{N-1}$, traditional NeRF is trained by minimizing the photometric error $L_{rgb} = \sum_{i=0}^{N-1} \|I_i - \hat{I}_i\|$, where $\hat{\mathcal{I}} = \{\hat{I}_i\}_{i=0}^{N-1}$ are the synthesised ones from above volume rendering. Typically, camera

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

parameters Π are precomputed and fixed during the whole optimizing. Furthermore, prior works [BWL*23; FLK*24; JAC*21] show that Π could be estimated and learnt together with NeRF parameters at the same time, as the ray r is a function of camera parameters Π. This joint optimization is formulated as:

$$\Theta^*, \Pi^* = \arg\min_{\Theta,\Pi} L_{rgb}(\hat{\mathcal{I}}, \hat{\Pi}|\mathcal{I}), \tag{2}$$

where $\hat{\Pi}$ denotes the learnable camera parameters of the optimization. Note that the camera parameters $\Pi = \{\pi_i\}_{i=0}^{N-1}$ are composed of camera extrinsics $\mathrm{T} = \{\mathrm{T}_i\}_{i=0}^{N-1}$ and intrinsics K and other distortions, where $\mathrm{T}_i = [\mathrm{R}_i|\mathrm{t}_i]$ is a rigid transformation with a rotation matrix $\mathrm{R}_i \in \mathrm{SO}(3)$ and a translation vector $\mathrm{t}_i \in \mathbb{R}^3$. we only focus on the extrinsics T in our work, while the other parameters are settled during the optimization.

4. Method

Overview. Given a set of images and their corresponding camera intrinsics, our method uses a local-to-global strategy to jointly estimate camera poses and the NeRF, as illustrated in Fig. 2. To effectively handle large view changes at the start, we introduce an association relationships initialization. Specifically, we construct a Maximum Spanning Tree (MST) to establish adjacency relationship among all input images, where each node represents an image, and the edge weights are determined by the maximum feature matches. This approach allows us to extract local matching pairs and global tracks from the tree, which proves effective in subsequent optimization. During the local joint optimization phase, we begin with the tree's root node, as it represents the image with the most matching pixels. We then progressively determine the relative pose between adjacent nodes based on the tree's adjacency structure to complete the initialization. These initial estimations serve as the starting poses for NeRF. Throughout training, we enforce photometric and depth consistency within a single view and geometric consistency across multiple views. Importantly, when introducing multi-view geometric consistency constraints, we also introduce a random matching scheme and incorporate non-adjacent random images as supervision. This inclusion allows global information to be integrated into local optimization, reducing accumulated errors in pose estimation and yielding more stable results. In the global joint optimization phase, we further refine the camera poses and NeRF by applying global geometric consistency constraints through the tracking trajectory. This involves supervising all images within the trajectory, similar to bundle adjustment, to enhance camera poses and the NeRF on a global scale.

4.1. Association Initialization

In order to achieve more stable camera pose estimation during joint optimization and to handle large view changes, we propose leveraging the explicit image feature matching relationship to re-establish the association relationships among input images. This approach takes advantage of the robustness and stability of feature matching, using the number of matched pixels as a critical indicator of image view similarity, thereby enabling accurate estimation of the relative pose transformation between image pairs.

Specifically, we construct a Maximum Spanning Tree (MST) based on feature matching pairs between images, see Fig. 2(b). The process involves the following steps. Firstly, we use the deep feature point extraction method, SuperPoint [DMR18], and the deep feature matching algorithm, SuperGlue [SDMR20], to compute the feature matches among all images. We define the weight between two images as the number of matching pairs. Then we construct the MST using Kruskal's algorithm, which sorts edges by weight from highest to lowest, ensuring no loops in the tree. By traversing this MST, we can extract the local matching pairs of images connected by an edge, and the global feature matching trajectories or track by union-find structure. These are used in subsequent local and global joint optimization to maintain geometric consistency and guide the optimization of camera relative pose estimation.

4.2. Local Joint Optimization

Inspired by NoPe-NeRF [BWL*23], we adopt a similar joint optimization strategy to simultaneously estimate camera poses and train the NeRF. For constraints on a single view, our approach considers both photometric and monocular depth loss. In addition, geometric constraints are introduced to improve the robustness. However, unlike NoPe-NeRF, it relies on the construction of dense point clouds and uses the chamfer distance to measure discrepancies, while this loss term struggles to capture the relative transformation accurately when camera motions are complex. In our procedure, we employ explicit feature matching to ensure robust geometric consistency across multiple associated views, as in Eq. 4 and 5, which is primarily derived from adjacent nodes obtained through the pre-constructed MST, focusing on local consistency. Therefore, we refer to this process as local joint optimization.

Mono-depth distortion constraint. As evidenced by previous works, monocular depth from the input image can provide a strong geometric prior for NeRF training. It helps reduce shape ambiguity in addition to the photometric constraint. In our approach, we leverage the estimated depth D from DPT [RBK21] for supervision. However, the absolute scale of depth map is unknown. To address this, we apply two linear transformations to map the depth to the real scale. Specifically, we define the transformation as $\Psi = \{\alpha_i, \beta_i\}$, where α and β denote a scale and a shift, respectively. With the transformed depth map $D^* = \alpha D + \beta$, we integrate this additional depth supervision into our optimization process,

$$L_{depth} = \sum_{i=0}^{N-1} \|D_i^* - \hat{D}_i\|, \tag{3}$$

where D_i^* and $\hat{D}_i$ denote the transformed depth map and the NeRF rendered depth map of image I_i, respectively.

Multiview geometric consistency constraint. Since the camera poses can be estimated by multi-view geometry, we also incorporate this constraint into our joint optimization framework. Taking two images I_i and I_j as an example, we back-project the corresponding depth maps to obtain the point sets P_i^* and P_j^*. We then measure the geometric consistency as follows,

$$l_{gc}(P_i^*, P_j^*) = \sum_k \|\mathbf{p}_{j,k}^* - T_{i,j} \cdot \mathbf{p}_{i,k}^*\|, \tag{4}$$

where $T_{i,j}$ denotes the relative pose that transforms from P_i^* to P_j^*, $\mathbf{p}_{i,k}^*$ and $\mathbf{p}_{j,k}^*$ represent the corresponding matching point pair in

© 2025 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

P_i^* and P_j^*, respectively, where the correspondence originates from pixel level matching. Concretely, for each input image I_i, we take its parent node I_p and child node I_c from the MST as references for multi-view geometric consistency. Since this constraint is local, it may be affected by noise and accumulated errors. Therefore, we incorporate a random matching scheme, which randomly introduces another non-adjacent image I_r with matching points for constraint. The matching number is at least 100 pairs for I_r to be considered for random selection. Since the image I_r is randomly selected, it can be regarded as integrating a weak global information into the constraint, which strengthens the constraint of multi-view geometric consistency. Therefore, the loss function can be defined as:

$$L_{mgc} = \sum_i l_{gc}(P_i^*, P_p^*) + l_{gc}(P_i^*, P_c^*) + l_{gc}(P_i^*, P_r^*). \tag{5}$$

Optimization formulation. We obtain the loss function of our local joint optimization as:

$$L_{local} = L_{rgb} + \lambda_{depth} L_{depth} + \lambda_{mgc} L_{mgc}, \tag{6}$$

where λ_{depth} and λ_{mgc} are the weighting factors for corresponding loss terms. Then the local joint model can be formulated as:

$$\Theta^*, \Pi^*, \Psi^* = \arg\min_{\Theta,\Pi,\Psi} L_{local}(\hat{\mathcal{I}}, \hat{\Pi}, \hat{\Psi} | \mathcal{I}, D). \tag{7}$$

Initialization of the optimization. Since NeRF training relies on camera poses, and to enhance joint optimization, some methods like [YFB*21; WWX*21; LMTL21] require initial poses that are very close to the ground truth or derived from COLMAP. Meanwhile, approaches like InstantSplat [FCW*24] employ 3D reconstruction (DUSt3R [WLC*24]) to initialize camera parameters. A better initialization consistently yields better results. However, we choose a rough initialization using monocular depth which is regarded as geometric prior in our local optimization. Experiments prove that this simple initialization is sufficient for delivering stable results. First, we use the monocular depth map D_i with the default scale and the camera intrinsics to back-project and restore the point cloud P_i^* for each image I_i. Then we use the association relationships from the MST to estimate the relative pose transformation $T_{i,j}$ for each child node I_j in the tree. The relative rigid transformation $T_{i,j}$ is obtained by minimizing the distance between P_j^* and $T_{i,j} \cdot P_i^*$. We treat this as a point cloud registration problem and solve it using the Iterative Closest Point (ICP) algorithm.

4.3. Global Joint Optimization

After local joint optimization, we obtain a set of camera poses that roughly align with the ground truth, as evident from Fig. 6. However, several issues remain. While the introduction of the mono-depth provides a strong geometric prior for NeRF training and helps reduce shape ambiguity that would otherwise rely solely on photometric constraints, the estimated monocular depth is often inaccurate, hindering optimal NeRF training. Additionally, the geometric consistency constraint is only applied to limited number of adjacent images, leading to accumulated errors that further impact the NeRF training. To address this, we aim to further improve the accuracy of camera pose estimation and the performance of NeRF by focusing on two main aspects: reducing the camera's accumulated errors and providing more accurate geometric priors for NeRF training. Inspired by the bundle adjustment algorithm, we introduce global geometric consistency constraints to achieve these goals.

Specifically, we use image matching and tracking in the association initialization phase to add geometric consistency constraints for more images. A track consists of a set of corresponding feature-matched pixels across a sequence of images, all representing the same 3D point in the scene. Using the camera poses obtained through local joint optimization, we can calculate the reprojection error. This error measures the distance between the reprojection of the point $\mathbf{q}$ and the corresponding image feature pixel $\mathbf{u}$, allowing us to evaluate the accuracy of the camera pose.

$$l_{proj}(\mathbf{q}) = \sum_k \|\mathbf{u}_k - K \cdot T_k \cdot \mathbf{q}\|, \tag{8}$$

where K is the camera intrinsic parameter, T_k is the camera pose of the k-th image in the track, and $\mathbf{u}_k$ represents the pixel coordinates of the feature point in the k-th image. It is worth noting that for the initialization of $\mathbf{q}$, we randomly select a pixel from one frame of a track, such as pixel $\mathbf{u}_k$ from the k-th image. Using the initial NeRF obtained from the local optimization phase, we render the depth value of this pixel and calculate the corresponding 3D coordinates. This 3D point is then treated as a 3D optimization variable and further optimized during the global optimization phase. Finally, we define the reprojection error calculated for all tracks within the association relationships of the MST as the track constraint:

$$L_{track} = \sum_{\mathbf{q} \in Q} l_{proj}(\mathbf{q}), \tag{9}$$

where Q is the set of tracking 3D points, whose average size is 10k in our experiments.

Compared to the multiview geometric consistency constraint in the local joint optimization, the track constraint simultaneously constrains a set of image poses across all trajectories. This results in a broader impact range than L_{mgc} and is beneficial for reducing accumulated errors and obtaining stable results. Additionally, for each track, the same 3D point is used as the reference and further optimized, leading to more accurate geometric priors. Consequently, L_{track} serves as a stronger global geometric constraint.

Then the overall loss function of our global joint optimization is defined as:

$$L_{global} = L_{rgb} + \lambda_{track} L_{track}, \tag{10}$$

where λ_{track} is the weighting factor. Our global joint model can be formulated as:

$$\Theta^*, \Pi^*, Q^* = \arg\min_{\Theta,\Pi,Q} L_{global}(\hat{\mathcal{I}}, \hat{\Pi}, \hat{Q} | \mathcal{I}). \tag{11}$$

5. Experiments

5.1. Experimental Setup

Datasets. We conduct experiments on two benchmark datasets, *Tanks and Temples* [KPZK17] and *CO3D V2* [RSH*21], to evaluate our performance in novel view synthesis and pose estimation. Specifically, for **Tanks and Temples**, we create two categories of scenes. The first is directly adapted from NoPe-NeRF [BWL*23] and includes 8 indoor and outdoor scenes with maximum relative rotation angles between any two frames ranging from 30° to 70°.

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

Figure 3: ***Qualitative comparison for novel view synthesis on Tanks and Temples.*** *Our method can convey richer details, while NoPe-NeRF appears more blurred. The red boxes indicate the regions to focus on.*

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

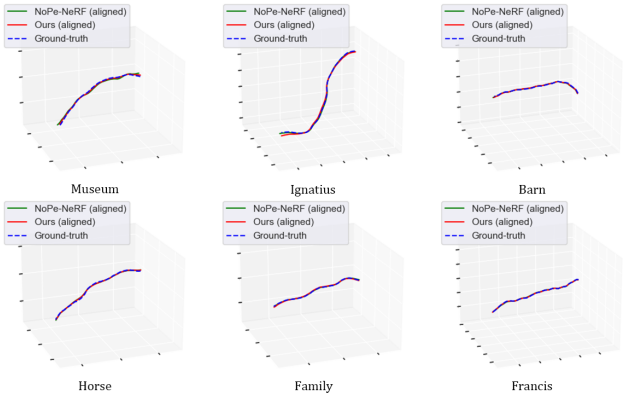

Figure 4: ***Comparison of pose estimation on Tanks and Temples.***

Scenes	Ours			NoPe-NeRF			BARF		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Church	25.08	**0.76**	**0.33**	**25.17**	0.73	0.39	23.17	0.62	0.52
Barn	**26.69**	**0.73**	**0.39**	26.35	0.69	0.44	25.28	0.64	0.48
Museum	**27.55**	**0.81**	**0.29**	26.77	0.76	0.35	23.58	0.61	0.55
Family	**26.66**	**0.78**	**0.36**	26.01	0.74	0.41	23.04	0.61	0.56
Horse	27.08	**0.84**	**0.25**	**27.64**	**0.84**	0.26	24.09	0.72	0.41
Ballroom	**26.65**	**0.81**	**0.27**	25.33	0.72	0.38	20.66	0.50	0.60
Francis	**29.92**	**0.82**	**0.36**	29.48	0.80	0.38	25.85	0.69	0.57
Ignatius	**24.33**	**0.63**	**0.44**	23.96	0.61	0.47	21.78	0.47	0.60
mean	**26.74**	**0.77**	**0.34**	26.34	0.74	0.39	23.42	0.61	0.54

Table 1: ***Quantitative comparison of novel view synthesis on Tanks and Temples.*** *Each baseline method is trained with its public code under the original settings and evaluated with the same evaluation protocol. The best results are highlighted in bold.*

Scenes	Ours			NoPe-NeRF			BARF		
	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE
Church	**0.012**	0.011	**0.001**	0.034	**0.008**	0.008	0.114	0.038	0.052
Barn	**0.010**	**0.024**	**0.001**	0.046	0.032	0.004	0.314	0.265	0.050
Museum	**0.032**	**0.133**	**0.002**	0.207	0.202	0.020	3.442	1.128	0.263
Family	**0.023**	0.029	0.002	0.047	**0.015**	**0.001**	1.371	0.591	0.115
Horse	**0.086**	0.050	**0.002**	0.179	**0.017**	0.003	1.333	0.394	0.014
Ballroom	**0.019**	**0.015**	**0.001**	0.041	0.018	0.002	0.531	0.228	0.018
Francis	**0.012**	0.036	**0.001**	0.057	**0.009**	0.005	1.321	0.558	0.082
Ignatius	**0.024**	0.032	0.003	0.026	**0.005**	**0.002**	0.736	0.324	0.029
mean	**0.027**	0.041	**0.002**	0.080	**0.038**	0.006	1.046	0.441	0.078

Table 2: ***Quantitative comparison of pose accuracy on Tanks and Temples.*** *Note that, we use COLMAP poses as the ground truth. The unit of RPE_r is in degrees, ATE is in the ground truth scale and RPE_t is scaled by 100. The best results are highlighted in bold.*

The second, more challenging dataset is created by us from the original videos of [KPZK17]. We choose *'Family'* video as our experimental scene and randomly extract 100 images from it. This results in significant transformations between adjacent frames, allowing us to demonstrate the stability of our method in non-sequential cases. For **CO3D V2**, borrowed from CF-3DGS[FLK*24], we select 5 scenes of different categories of objects, all involving large camera motions. We follow their experimental setup on these object-centric scenes to show the robustness of our method. For all mentioned scenes, we sample 1 from every 8 frames for novel view synthesis, and the rest of the images are used for training.

Metrics. We use standard evaluation metrics, including PSNR, SSIM [WBSS04] and LPIPS [ZIE*18] to evaluate the quality of novel view synthesis. For pose estimation, we rely on the Absolute Trajectory Error (ATE) and Relative Pose Error (RPE) [LMTL21; BWL*23; FLK*24]. RPE_r and RPE_t are utilized to measure the accuracy of rotation and translation, respectively. To ensure the metrics are comparable on the same scale, we align the camera poses by Umeyama's method [Ume91] for both estimation and evaluation. Since all the mentioned datasets lack ground truth camera poses, we use COLMAP [SF16] to estimate the poses as reference.

Implementation details. Our implementation is primarily based on NoPe-NeRF [BWL*23]. We optimize the mono-depth transformation (Ψ) parameters only during the local optimization phase, and track points (Q) only during global phase. NeRF (F_Θ) and camera poses (Π) parameters are refined in both phases. Moreover, each pose is formatted as a combination of an axis-angle representation $q \in SO(3)$ and a translation vector $t \in \mathbb{R}^3$. The initial learning rate for NeRF is set at 0.001 and the rates for other parameters are set at 0.0005. During the local phase, we train the joint optimization (Eq. 7) for 40k steps with $\lambda_{depth} = 0.04$, $\lambda_{mgc} = 1.0$ of loss terms in Eq. 6. For global optimization, the model (Eq. 11) is first trained for 10k steps with $\lambda_{track} = 10.0$ in Eq. 10, followed by 5k steps with $\lambda_{track} = 1.0$. For other training parameters, we use the same settings as NoPe-NeRF, such as learning rate decreasing strategy.

5.2. Comparing with Pose-Unknown Methods

We compare our method with state-of-the-art pose-unknown methods, including NoPe-NeRF [BWL*23], BARF [LMTL21], and CF-3DGS [FLK*24]. Notably, for novel view synthesis, following NeRFmm [WWX*21], we obtain the camera poses of test views by minimizing the photometric errors between the synthesized and test images, while keeping the trained NeRF parameters fixed. The test camera poses are initialized to the poses of most geometry-related training images to expedite convergence. All baselines use the same pre-processing method to ensure a fair comparison.

5.2.1. With NeRF-based Methods

We first compare our results with NoPe-NeRF and BARF, both of which use the NeRF model as the scene representation.

We first present the qualitative comparisons of novel view synthesis and pose estimation on Tanks and Temples in Fig. 3 and Fig. 4. Notably, our method obviously enhances the rendering quality, as highlighted in the image. Our results convey richer details, while NoPe-NeRF appears more blurred. This improvement is attributed to the accurate camera poses aligned with ground truth, which is also directly validated in Fig. 4. Following this, we summarize the numerical accuracy in Tab. 1 and Tab. 2. On average, our method outperforms existing state-of-the-art approaches across all metrics.

Notably, in terms of pose accuracy, our method performs better in RPE_t and ATE, but is slightly worse on RPE_r. We claim that the calculation for RPE_r is directly related to the accuracy of adjacent frame poses, which is where local optimization excels; hence, NoPe-NeRF achieves better results there. However, we focus more on global optimization, leading to higher performance in the other

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

Figure 5: ***Qualitative comparison of novel view synthesis on CO3D V2.*** *We also visualize the rendered depth maps of NeRF models (bottom right of each image) for both methods. Notice that our geometric quality is much better than NoPe-NeRF.*

Methods	110_13051_23361			415_57112_110099			106_12648_23157			245_26182_52130			34_1403_4393		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
NoPe-NeRF	26.86	0.73	0.47	24.78	**0.64**	0.55	20.41	0.46	0.58	25.05	0.80	0.49	28.62	0.80	0.35
Ours	**30.26**	**0.83**	**0.31**	**25.40**	**0.64**	**0.53**	**20.91**	**0.48**	**0.52**	**27.18**	**0.83**	**0.43**	**32.00**	**0.87**	**0.29**

Table 3: ***Comparison of novel view synthesis on CO3D V2.***

Methods	110_13051_23361			415_57112_110099			106_12648_23157			245_26182_52130			34_1403_4393		
	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE
NoPe-NeRF	0.400	1.966	0.046	0.326	1.919	0.054	0.387	1.312	0.049	0.587	1.867	0.038	0.591	1.313	0.053
Ours	**0.015**	**0.065**	**0.001**	**0.004**	**0.020**	**0.001**	**0.010**	**0.116**	**0.001**	**0.011**	**0.059**	**0.001**	**0.030**	**0.110**	**0.001**

Table 4: ***Comparison of pose estimation on CO3D V2.***

two metrics. Additionally, improved global consistency contributes to higher rendering quality, as demonstrated above.

Notice that, in the Tanks and Temples datasets, the camera movements are relatively small, and estimating relative poses between adjacent frames during the local phase is quite easy. To better highlight the advantages of our algorithm, we choose the more challenging CO3D V2 dataset for validation, where camera pose changes are more obvious and the range of camera movements between frames is larger. Local pose estimation struggles to achieve good results in this case. As shown in Fig. 5, Tab. 3 and 4, our algorithm still delivers better rendering quality and more accurate camera poses, while NoPe-NeRF may completely fail. In Fig. 5, we also present a comparison of depth maps rendered by the trained NeRF, illustrating that our improved geometry results from optimizing both NeRF parameters and track points during the global optimization phase, allowing us to learn better geometry.

On this more challenging dataset, we further analyze the performance of our algorithm at each phase, as shown in Fig. 6. It is clear that during the initialization, we can only obtain a very rough estimate of the camera poses. However, in the local optimization phase, observing that the relationships between adjacent frames are well constrained, allows the optimized poses to get closer to the ground truth, but the noticeable accumulated errors are still present, as seen in the examples of the second and fourth rows. After applying global optimization, the final poses converge accurately.

In terms of training time, our method requires approximately

	scenes	Ours			CF-3DGS		
		RPE_t ↓	RPE_r ↓	ATE↓	RPE_t	RPE_r	ATE
Tanks and Temples	Church	0.012	**0.011**	**0.001**	**0.008**	0.018	0.002
	Barn	**0.010**	**0.024**	**0.001**	0.034	0.034	0.003
	Museum	**0.032**	**0.133**	**0.002**	0.052	0.215	0.005
	Family	0.023	0.029	**0.002**	**0.022**	**0.024**	**0.002**
	Horse	**0.086**	**0.050**	**0.002**	0.122	0.057	0.003
	Ballroom	**0.019**	**0.015**	**0.001**	0.037	0.024	0.003
	Francis	**0.012**	**0.036**	**0.001**	0.029	0.154	0.006
	Ignatius	**0.024**	**0.032**	**0.003**	0.033	**0.032**	0.005
CO3D V2	110_13051_23361	**0.015**	**0.065**	**0.001**	0.061	0.253	0.005
	415_57112_110099	**0.004**	**0.020**	**0.001**	0.149	1.412	0.016
	106_12648_23157	**0.010**	**0.116**	**0.001**	0.094	0.360	0.008
	245_26182_52130	**0.011**	**0.059**	**0.001**	0.056	0.324	0.006
	34_1403_4393	**0.030**	**0.110**	**0.001**	0.108	0.390	0.010

Table 5: ***Comparison to CF-3DGS on pose accuracy.*** *The best results are highlighted in bold.*

30 hours on average for a single scene, compared to 12 hours for Vanilla-NeRF and 20 hours for NoPe-NeRF. This total includes about 20 hours for the local optimization phase and around 10 hours for the global refinement. The increased training time is primarily due to the use of NeRF as a back-end. We acknowledge this trade-off in time efficiency for significantly improved accuracy and robustness, which we believe is worthwhile.

5.2.2. With CF-3DGS

In Tab. 5, we compare pose estimation accuracy with CF-3DGS. Overall, our method significantly outperforms CF-3DGS across

© 2025 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

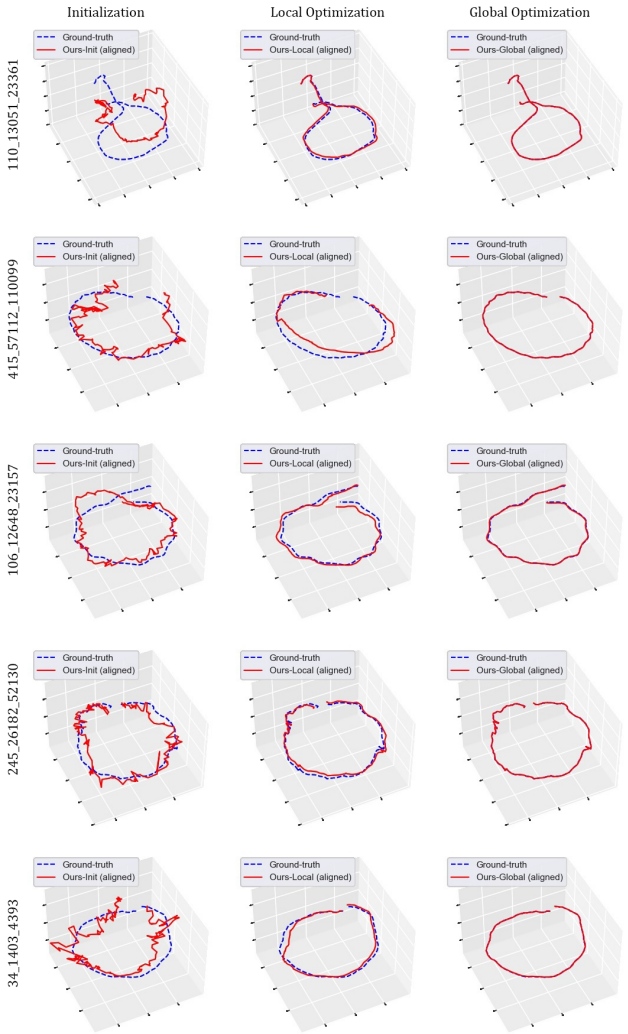

Figure 6: ***Pose estimation results for each phase on CO3D V2.***

all test datasets. Notably, while CF-3DGS also considers local information during initialization, we introduce a random matching scheme during the local optimization phase to incorporate weak global information which results in better initialization. In other words, although CF-3DGS also utilizes global optimization, in certain challenging scenarios, relatively poor initial values for camera poses may not be fully corrected through global optimization. This example underscores the interplay between local and global optimization; their collaboration is crucial for achieving optimal results. Additionally, we do not compare the rendering quality here, as it is well-known that the rendering quality of the CF-3DGS base model is obviously better than that of NeRF. We focus solely on pose accuracy to validate the rationale of our algorithm. We firmly believe that using 3D Gaussian Splatting as the base model would yield even better rendering results in our case.

Figure 7: ***Results of unordered sequence of images.*** *We extract these 100-frames from the scene of Family in Tanks and Temples dataset. The top left shows the pose distribution of the ground truth (COLMAP) and our optimized results. The translational errors are quite small for all trained frames. The rest of the images reveal the synthesis images and corresponding depth maps rendered from the trained NeRF model.*

5.2.3. Results on More Challenging Scene

In Tanks and Temples and CO3D V2 datasets, all frames are sampled from a continuous video, ensuring a strong spatial correlation between consecutive frames, and making it easier to determine their relative poses. Algorithms like NoPe-NeRF and CF-3DGS are designed based on this assumption, estimating relative poses under local adjacency or performing global incremental optimization. If the adjacency relationship in the input data is disrupted, these methods fail. Our method, however, is robust against such data, utilizing MST for association determination, a random matching mechanism in local optimization, and global optimization to achieve better results in these challenging scenarios. In the example shown in Fig. 7, our method still demonstrates robust results.

5.3. Ablation Study

Effectiveness of local/global joint optimization. To assess the effectiveness of our local joint phase, we remove it from the pipeline and begin global joint optimization directly from the training data preparation. Since we do not have a pre-trained rough NeRF model for evaluating the track points, we use projected 3D points from the mono-depth DPT as the initial track points. The results for novel view synthesis and pose accuracy are shown as *w.o. local* in Tab. 6. We also analyze the metrics of local-only results to highlight the need for global joint optimization, reporting these as *w.o. global*. Our observations indicate that omitting either phase results in poorer metrics for view synthesis and pose estimation. Thanks to the estimated camera poses and the rough NeRF model from the local phase, along with pose refinement from the global phase, our joint optimization demonstrates outstanding performance in novel view synthesis and pose estimation.

Comparison of NeRF with COLMAP poses. We further analyze the novel view synthesis quality of the NeRF model trained with our learned poses compared to COLMAP poses on the Tanks

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

scenes	w.o. local			w.o. global			Ours		
	RPE_t ↓	RPE_r ↓	ATE↓	RPE_t	RPE_r	ATE	RPE_t	RPE_r	ATE
110_13051_23361	0.016	0.069	**0.001**	0.061	0.253	0.005	**0.015**	**0.065**	**0.001**
415_57112_110099	0.021	0.066	**0.001**	0.149	1.412	0.016	**0.004**	**0.020**	**0.001**
106_12648_23157	0.034	0.325	0.003	0.074	0.525	0.008	**0.010**	**0.116**	**0.001**
245_26182_52130	0.012	0.066	**0.001**	0.056	0.324	0.006	**0.011**	**0.059**	**0.001**
34_1403_4393	0.089	0.330	0.002	0.108	0.390	0.010	**0.030**	**0.110**	**0.001**

Table 6: ***Effectiveness of local and global joint optimization phase.***

scenes	Ours			NeRF+COLMAP		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS
Church	25.08	**0.76**	**0.33**	**25.72**	0.75	0.37
Barn	26.69	**0.73**	**0.39**	**26.72**	0.71	0.42
Museum	**27.55**	**0.81**	**0.29**	27.21	0.78	0.34
Family	**26.66**	**0.78**	**0.36**	26.61	0.77	0.39
Horse	**27.08**	**0.84**	**0.25**	27.02	0.82	0.29
Ballroom	**26.65**	**0.81**	**0.27**	25.47	0.73	0.38
Francis	29.92	**0.82**	**0.36**	**30.05**	0.81	0.38
Ignatius	**24.33**	**0.63**	**0.44**	24.08	0.61	0.47

Table 7: ***Comparison to NeRF trained with COLMAP poses.***

scenes	3DGS+Ours			CF-3DGS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR	SSIM	LPIPS
110_13051_23361	**31.60**	**0.93**	**0.09**	29.69	0.89	0.29
415_57112_110099	**29.53**	**0.84**	**0.13**	26.21	0.73	0.32
106_12648_23157	**26.13**	**0.83**	**0.11**	22.14	0.64	0.34
245_26182_52130	**33.12**	**0.93**	**0.14**	27.24	0.85	0.30
34_1403_4393	**31.33**	**0.94**	**0.11**	27.75	0.86	0.20

Table 8: ***Comparison to CF-3DGS with original 3DGS and our poses.***

and Temples dataset. This evaluation allows us to indirectly compare the accuracy of the estimated poses based on their impact on novel view synthesis. Specifically, in Tab. 7, we achieve results that are comparable to COLMAP across all scenes, and in some cases, even outperform COLMAP. The main reason for this, as we analyzed, is that COLMAP uses traditional algorithms for pose estimation, which do not perform well in challenging scenarios with weak texture and large camera motions. In contrast, our proposed joint learning approach, through the design of various loss functions, not only ensures better rendering result quality but also enhances the accuracy of camera parameter estimation.

Comparison to CF-3DGS with original 3DGS and our poses. To make the comparison of novel view synthesis with CF-3DGS reasonable, we also extract the estimated poses from the results of our final joint optimization and train an original 3DGS model with our poses. Tab. 8 shows the quantitative comparisons of 3DGS with our estimated camera poses and CF-3DGS on CO3D V2 dataset. Notice that based on the same 3DGS model, our results of novel view synthesis are much better than CF-3DGS, mainly benefiting from the more accurate pose estimation of our joint optimization, as exhibited in Tab. 5.

6. Conclusions

In this work, we present NoPe-NeRF++, a novel local-to-global algorithm for joint camera pose estimation and NeRF training. Compared to previous methods, our approach can handle complex camera trajectories, dramatic camera viewpoint changes, and non-sequential image inputs. To tackle these challenges, our method begins by re-establishing the association relationships among input images through explicit feature matching for optimization. During the local joint optimization phase, constraints from local views, combined with a randomly selected non-adjacent reference image, enhance the robustness of camera poses and the NeRF model. Additionally, we introduce a global optimization phase that incorporates geometric consistency constraints via bundle adjustment, integrating feature trajectories to further improve pose estimation and the NeRF model. Extensive evaluations on benchmark datasets demonstrate our superior performance and robustness, even in challenging scenes. On the other hand, due to the limited expressive capability of NeRF itself, our method also has limitations in handling very large scenes, managing complex lighting conditions, and training/testing efficiency. In the future, we will consider more advanced representation methods, such as 3DGS [KKLD23], to address these issues. Because we only consider color and depth values of samples during the whole procedure, which are fundamental outputs of other representations. Therefore, it is intuitive to seamlessly integrate those models with our framework. Nevertheless, we believe our method represents an important step towards training NeRF models without pose priors for complex applications in the future.

7. Acknowledgements

We thank the anonymous reviewers for their constructive comments that help improve this paper.

References

[ASS*09] AGARWAL, SAMEER, SNAVELY, NOAH, SIMON, IAN, et al. "Building Rome in a day". *2009 IEEE 12th International Conference on Computer Vision*. 2009, 72–79 2.

[BEK*22] BOSS, MARK, ENGELHARDT, ANDREAS, KAR, ABHISHEK, et al. "SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections". *Advances in Neural Information Processing Systems (NeurIPS)*. 2022 3.

[BMT*21] BARRON, JONATHAN T., MILDENHALL, BEN, TANCIK, MATTHEW, et al. *Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields*. 2021. arXiv: 2103.13415 [cs.CV] 2.

[BWL*23] BIAN, WENJING, WANG, ZIRUI, LI, KEJIE, et al. "NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior". *CVPR*. 2023 2–5, 7.

[CCW*23] CHEN, YUE, CHEN, XINGYU, WANG, XUAN, et al. "Local-to-global registration for bundle-adjusting neural radiance fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, 8264–8273 3.

[CEJ*23] CHENG, ZEZHOU, ESTEVES, CARLOS, JAMPANI, VARUN, et al. "LU-NeRF: Scene and Pose Estimation by Synchronizing Local Unposed NeRFs". *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2023 3.

[CGSL24] CHNG, SHIN-FANG, GARG, RAVI, SARATCHANDRAN, HEMANTH, and LUCEY, SIMON. "Invertible Neural Warp for NeRF". *ECCV*. 2024 3.

[CW93] CHEN, SHENCHANG ERIC and WILLIAMS, LANCE. "View interpolation for image synthesis". *Proc. of SIGGRAPH* (1993), 279–288 2.

[DMR18] DETONE, DANIEL, MALISIEWICZ, TOMASZ, and RABINOVICH, ANDREW. "Superpoint: Self-supervised interest point detection and description". *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. 2018, 224–236 4.

© 2025 Eurographics - The European Association for Computer Graphics and John Wiley & Sons Ltd.

[DRMS07] DAVISON, ANDREW J., REID, IAN D., MOLTON, NICHOLAS D., and STASSE, OLIVIER. "MonoSLAM: Real-Time Single Camera SLAM". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29.6 (2007), 1052–1067 2.

[ERB*24] ENGELHARDT, ANDREAS, RAJ, AMIT, BOSS, MARK, et al. "SHINOBI: Shape and Illumination using Neural Object Decomposition via BRDF Optimization In-the-wild". *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, 19636–19646. DOI: 10.1109/CVPR52733.2024.01857 3.

[FCW*24] FAN, ZHIWEN, CONG, WENYAN, WEN, KAIRUN, et al. *InstantSplat: Unbounded Sparse-view Pose-free Gaussian Splatting in 40 Seconds*. 2024. arXiv: 2403.20309 [cs.CV] 5.

[FLK*24] FU, YANG, LIU, SIFEI, KULKARNI, AMEY, et al. "COLMAP-Free 3D Gaussian Splatting". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, 20796–20805 3, 4, 7.

[GGSC96] GORTLER, STEVEN J, GRZESZCZUK, RADEK, SZELISKI, RICHARD, and COHEN, MICHAEL F. "The lumigraph". *Proc. of SIGGRAPH* (1996), 43–54 2.

[HAA97] HORRY, YOUICHI, ANJYO, KEN-ICHI, and ARAI, KIYOSHI. "Tour into the picture: using a spidery mesh interface to make animation from a single image". *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*. SIGGRAPH '97. 1997, 225–232 2.

[HEH05] HOIEM, DEREK, EFROS, ALEXEI A., and HEBERT, MARTIAL. "Automatic photo pop-up". *ACM SIGGRAPH 2005 Papers*. SIGGRAPH '05. 2005, 577–584. ISBN: 9781450378253 2.

[HRBP21] HU, RONGHANG, RAVI, NIKHILA, BERG, ALEXANDER C., and PATHAK, DEEPAK. "Worldsheet: Wrapping the World in a 3D Sheet for View Synthesis from a Single Image". *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2021 2.

[JAC*21] JEONG, YOONWOO, AHN, SEOKJUN, CHOY, CHRISTOPHER, et al. "Self-Calibrating Neural Radiance Fields". *ICCV*. 2021 2–4.

[KKLD23] KERBL, BERNHARD, KOPANAS, GEORGIOS, LEIMKÜHLER, THOMAS, and DRETTAKIS, GEORGE. "3D Gaussian Splatting for Real-Time Radiance Field Rendering". *ACM Transactions on Graphics* 42.4 (July 2023) 2, 10.

[KPZK17] KNAPITSCH, ARNO, PARK, JAESIK, ZHOU, QIAN-YI, and KOLTUN, VLADLEN. "Tanks and temples: benchmarking large-scale scene reconstruction". *ACM Trans. Graph.* 36.4 (July 2017) 5, 7.

[KSH22] KIM, MIJEONG, SEO, SEONGUK, and HAN, BOHYUNG. "InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering". *CVPR*. 2022 2.

[LFS*21] LI, JIAXIN, FENG, ZIJIAN, SHE, QI, et al. "MINE: Towards Continuous Depth MPI with NeRF for Novel View Synthesis". *ICCV*. 2021 2.

[LMTL21] LIN, CHEN-HSUAN, MA, WEI-CHIU, TORRALBA, ANTONIO, and LUCEY, SIMON. "BARF: Bundle-Adjusting Neural Radiance Fields". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 5721–5731 2, 3, 5, 7.

[MESK22] MÜLLER, THOMAS, EVANS, ALEX, SCHIED, CHRISTOPH, and KELLER, ALEXANDER. "Instant Neural Graphics Primitives with a Multiresolution Hash Encoding". *ACM Trans. Graph.* 41.4 (July 2022), 102:1–102:15 2.

[MLG*23] MEULEMAN, ANDREAS, LIU, YU-LUN, GAO, CHEN, et al. "Progressively Optimized Local Radiance Fields for Robust View Synthesis". *CVPR*. 2023 3.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P., TANCIK, MATTHEW, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis". (2020) 1–3.

[MYZ*24] MAO, YUNXUAN, YU, XUAN, ZHANG, ZHUQING, et al. "Ngel-slam: Neural implicit representation-based global consistent low-latency slam system". *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2024, 6952–6958 2.

[MZR*24] MAI, JINJIE, ZHU, WENXUAN, ROJAS, SARA, et al. "TrackNeRF: Bundle Adjusting NeRF from Sparse and Noisy Views via Feature Tracks". *European Conference on Computer Vision*. 2024 3.

[OPG21] OECHSLE, MICHAEL, PENG, SONGYOU, and GEIGER, ANDREAS. "UNISURF: Unifying Neural Implicit Surfaces and Radiance Fields for Multi-View Reconstruction". *International Conference on Computer Vision (ICCV)*. 2021 2.

[PHM*23] PARK, KEUNHONG, HENZLER, PHILIPP, MILDENHALL, BEN, et al. "CamP: Camera Preconditioning for Neural Radiance Fields". *ACM Trans. Graph.* 42.6 (Dec. 2023). ISSN: 0730-0301. DOI: 10.1145/3618321. URL: https://doi.org/10.1145/3618321 3.

[RBK21] RANFTL, RENÉ, BOCHKOVSKIY, ALEXEY, and KOLTUN, VLADLEN. "Vision Transformers for Dense Prediction". *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. Oct. 2021, 12179–12188 3, 4.

[RK20] RIEGLER, GERNOT and KOLTUN, VLADLEN. "Free View Synthesis". *European Conference on Computer Vision*. 2020 2.

[RK21] RIEGLER, GERNOT and KOLTUN, VLADLEN. "Stable View Synthesis". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2021 2.

[RSH*21] REIZENSTEIN, JEREMY, SHAPOVALOV, ROMAN, HENZLER, PHILIPP, et al. "Common Objects in 3D: Large-Scale Learning and Evaluation of Real-life 3D Category Reconstruction". *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2021, 10881–10891 5.

[SD96] SEITZ, STEVEN M and DYER, CHARLES R. "View morphing". *Proc. of SIGGRAPH* (1996), 21–30 2.

[SDMR20] SARLIN, PAUL-EDOUARD, DETONE, DANIEL, MALISIEWICZ, TOMASZ, and RABINOVICH, ANDREW. "SuperGlue: Learning Feature Matching with Graph Neural Networks". *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, 4938–4947 4.

[SF16] SCHONBERGER, JOHANNES L. and FRAHM, JAN-MICHAEL. "Structure-From-Motion Revisited". *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2016 2, 7.

[SSS06] SNAVELY, NOAH, SEITZ, STEVEN M., and SZELISKI, RICHARD. "Photo tourism: Exploring photo collections in 3D". *SIGGRAPH Conference Proceedings*. ACM Press, 2006, 835–846 2.

[TCY*22] TANCIK, MATTHEW, CASSER, VINCENT, YAN, XINCHEN, et al. "Block-NeRF: Scalable Large Scene Neural View Synthesis". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2022, 8248–8258 2.

[TS20] TUCKER, RICHARD and SNAVELY, NOAH. "Single-view View Synthesis with Multiplane Images". *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2020 2.

[Ume91] UMEYAMA, S. "Least-squares estimation of transformation parameters between two point patterns". *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13.4 (1991), 376–380 7.

[WBSS04] WANG, ZHOU, BOVIK, A.C., SHEIKH, H.R., and SIMONCELLI, E.P. "Image quality assessment: from error visibility to structural similarity". *IEEE Transactions on Image Processing* 13.4 (2004), 600–612 7.

[WCH*22] WANG, CAN, CHAI, MENGLEI, HE, MINGMING, et al. "Clip-nerf: Text-and-image driven manipulation of neural radiance fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 3835–3844 2.

[WLC*24] WANG, SHUZHE, LEROY, VINCENT, CABON, YOHANN, et al. "DUSt3R: Geometric 3D Vision Made Easy". *CVPR*. 2024 5.

[WWX*21] WANG, ZIRUI, WU, SHANGZHE, XIE, WEIDI, et al. "NeRF−−: Neural Radiance Fields Without Known Camera Parameters". *arXiv preprint arXiv:2102.07064* (2021) 2, 3, 5, 7.

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.

[XXP*22] XU, QIANGENG, XU, ZEXIANG, PHILIP, JULIEN, et al. "Point-nerf: Point-based neural radiance fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 5438–5448 2.

[YFB*21] YEN-CHEN, LIN, FLORENCE, PETE, BARRON, JONATHAN T., et al. "iNeRF: Inverting Neural Radiance Fields for Pose Estimation". *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. 2021, 1323–1330 3, 5.

[YSL*22] YUAN, YU-JIE, SUN, YANG-TIAN, LAI, YU-KUN, et al. "NeRF-editing: geometry editing of neural radiance fields". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, 18353–18364 2.

[ZBRH22] ZHANG, QIANG, BAEK, SEUNG-HWAN, RUSINKIEWICZ, SZYMON, and HEIDE, FELIX. "Differentiable Point-Based Radiance Fields for Efficient View Synthesis". *arXiv preprint arXiv:2205.14330* (2022) 2.

[ZIE*18] ZHANG, RICHARD, ISOLA, PHILLIP, EFROS, ALEXEI A., et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric". *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018, 586–595 7.

[ZPL*22] ZHU, ZIHAN, PENG, SONGYOU, LARSSON, VIKTOR, et al. "NICE-SLAM: Neural Implicit Scalable Encoding for SLAM". *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022 2.

[ZSD*21] ZHANG, XIUMING, SRINIVASAN, PRATUL, DENG, BOYANG, et al. "NeRFactor: Neural Factorization of Shape and Reflectance Under an Unknown Illumination". *ACM Transactions on Graphics* 40.6 (Dec. 2021), 1–18 2.

[ZSH*22] ZHANG, YUANQING, SUN, JIAMING, HE, XINGYI, et al. "Modeling Indirect Illumination for Inverse Rendering". *CVPR*. 2022 2.

[ZTF*18] ZHOU, TINGHUI, TUCKER, RICHARD, FLYNN, JOHN, et al. "Stereo Magnification: Learning View Synthesis using Multiplane Images". *SIGGRAPH*. 2018 2.

© 2025 Eurographics - The European Association
for Computer Graphics and John Wiley & Sons Ltd.