TrackNeRF: Bundle Adjusting NeRF from Sparse and Noisy Views via Feature Tracks

Jinjie Mai[1], Wenxuan Zhu[1], Sara Rojas[1], Jesus Zarzar[1], Abdullah Hamdi[2], Guocheng Qian[3], Bing Li[1], Silvio Giancola[1], and Bernard Ghanem[1]

[1] King Abdullah University of Science and Technology
[2] Visual Geometry Group, University of Oxford
[3] Snap Inc.
{jinjie.mai,bernard.ghanem}@kaust.edu.sa

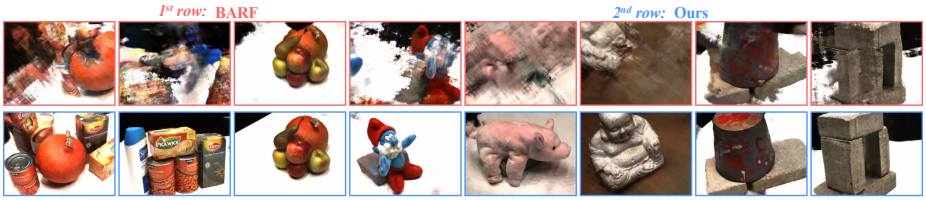


Fig. 1: **Novel View Synthesis from Sparse and Noisy Views.** TrackNeRF achieves high-quality novel view synthesis through *bundle adjusting feature tracks*.

Abstract. Neural radiance fields (NeRFs) generally require many images with accurate poses for accurate novel view synthesis, which does not reflect realistic setups where views can be sparse and poses can be noisy. Previous solutions for learning NeRFs with sparse views and noisy poses only consider local geometry consistency with pairs of views. Closely following *bundle adjustment* in Structure-from-Motion (SfM), we introduce TrackNeRF for more globally consistent geometry reconstruction and more accurate pose optimization. TrackNeRF introduces *feature tracks*, *i.e.* connected pixel trajectories across *all* visible views that correspond to the *same* 3D points. By enforcing reprojection consistency among feature tracks, TrackNeRF encourages holistic 3D consistency explicitly. Through extensive experiments, TrackNeRF sets a new benchmark in noisy and sparse view reconstruction. In particular, TrackNeRF shows significant improvements over the state-of-the-art BARF and SPARF by ~ 8 and ~ 1 in terms of PSNR on DTU under various sparse and noisy view setups. The code is available at https://tracknerf.github.io/.

Keywords: NeRF, Sparse views, Camera pose optimization

1 Introduction

The pursuit of reconstructing and creating immersive virtual environments, such as the metaverse, has accelerated significantly with the advent of 3D vision and

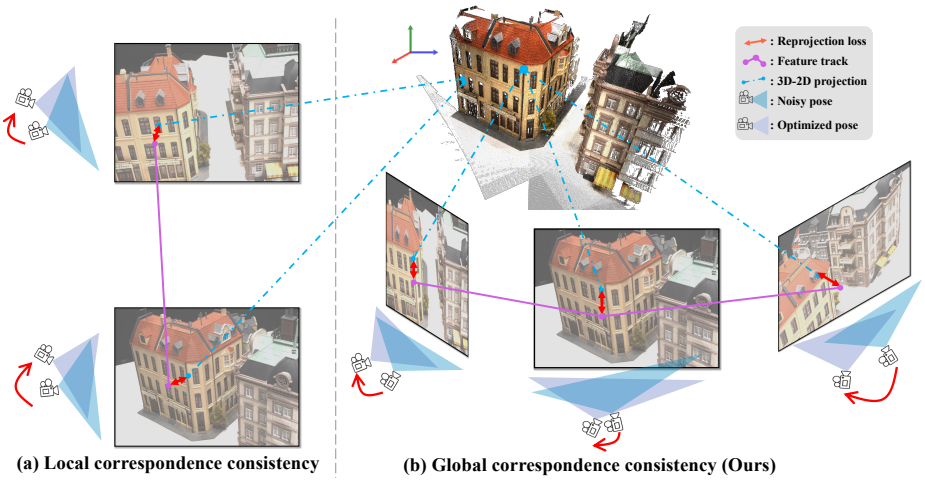


Fig. 2: Illustration of Track Reprojection Loss. *Left:* Pairwise correspondence objective employed by CorresNeRF [39] and SPARF [70]. *Right:* Feature tracks objective proposed by **TrackNeRF**. TrackNeRF minimizes the reprojection loss across all visible views for feature tracks corresponding to the same landmarks.

AR/VR devices. Creating such realistic virtual environments typically requires intricate manual design by artists. Concurrently, the evolution of novel-view synthesis has been notably influenced by the emergence of Neural Radiance Fields (NeRFs) [52]. NeRFs have revolutionized this domain by parameterizing the volume of scenes through neural networks. They enable rendering novel views of a scene with unparalleled precision and realism, thus emulating a virtual environment. This foundational shift towards leveraging neural networks lays the groundwork for the next generation of immersive digital experiences, driving forward the boundaries of what is possible in virtual world creation.

Neural Radiance Fields and their derivatives, including prominent methods like InstantNGP [53] and Plenoxels [19], along with recent advancements in Gaussian Splatting [35], fundamentally rely on the availability of a substantial corpus of posed multiview images to attain photorealistic renderings. This assumption, however, starkly contrasts with the typical conditions encountered in real-world scenarios. In most cases, available visual data is comprised of sparse image collections or unposed videos with standard Structure-from-Motion (SfM) tools such as COLMAP [49,62] employed to derive camera poses. Both scenarios inevitably result in estimations tarnished by noise and inaccuracies. This discrepancy between the ideal conditions presumed by NeRF-based methodologies and the reality of data acquisition presents a significant challenge. Novel approaches that efficiently work with less-than-ideal datasets are necessary to ensure robust 3D reconstruction and rendering without the luxury of extensive posed imagery.

Previous works [14, 30, 39, 55, 75, 76, 79, 83] have attempted to address this more realistic setup. BARF [43] first proposes adopting coarse-to-fine frequency

encoding to ease NeRF's backward optimization for pose estimation. However, BARF does not perform de facto *"bundle adjustment"* since it does not use any multiview constraints. Other methods tackle noisy poses as well by assuming dense views [43, 55, 75] or address the few-view limitation by assuming perfect poses [76,83], but don't tackle both issues jointly. Recently, SPARF [70] proposes to tackle sparse views and noisy poses simultaneously, achieving a remarkable milestone in the development of realistic setup novel views synthesis. However, SPARF adopts two-view correspondence-based reprojection loss as the core of optimization without considering long-term consistency.

To overcome the local consistency limitation of previous works, we propose TrackNeRF. TrackNeRF is inspired by the fact that all views are taken from *a single holistic 3D scene*, thus the corresponding pixels from *all* sparse views rendered from NeRF should ideally be back-projected to the *same* 3D landmark. Following this motivation, TrackNeRF extracts *feature tracks* for optimization, *i.e.* connected pixel trajectories across all visible views corresponding to the same 3D points. TrackNeRF enforces reprojection consistency among each track and thus introduces a holistic geometry consistency into NeRF. As a result, TrackNeRF achieves the most robust and highest reconstruction fidelity with more precise camera poses among all sparse and noisy NeRF solutions. Fig. 1 illustrates the difference between TrackNeRF and the pioneering methods.

Contributions. We summarize our contributions as follows:

(i) We introduce TrackNeRF, which utilizes feature tracks that closely follow the *bundle adjustment* literature. TrackNeRF can reconstruct a more geometry-consistent volumetric representation and recover more accurate camera poses.

(ii) Our TrackNeRF achieves $\sim$ 1 PSNR boost against previous SOTA and halves pose errors on the challenging DTU [33] dataset for all 3-view, 6-view, and 9-view setups with noisy poses. Under 3 views with ground truth poses, TrackNeRF also outperforms SOTA diffusion and regularization-based methods.

(iii) In practice, we demonstrate that TrackNeRF can tolerate greater pose noise, perform faster pose optimization, and synthesize high-quality novel views aligned with correct and smooth depth.

2 Related Work

Multi-View Reconstruction. The multi-view 3D reconstruction field is dedicated to the restoration of a scene's three-dimensional structure from its two-dimensional RGB images, obtained from various camera perspectives [1, 18]. Historically, these approaches have focused on generating a point cloud representation of a scene's geometry through the utilization of SIFT-based point matching techniques [48, 62, 63]. Advances in this domain have seen a shift towards the use of neural networks to improve feature extraction, as evidenced by several studies (*e.g.* [25, 31, 46, 80, 81, 84]). The introduction of Neural Radiance Fields (NeRF) [47, 52] has marked a significant transition toward the volumetric radiance reconstruction of 3D spaces [67], facilitating the creation of photorealistic novel views [2, 3, 71]. However, as a common constraint, all of

these approaches assume *sufficient overlapped views*, typically around 100 images, with *precise camera poses*. In many real applications, the captured views are sparse and noisy, consisting of only two to four views with inaccurate poses. Our aim in this work is to robustly reconstruct 3D geometry from these sparse and noisy poses.

Few-View Reconstruction. Further research has delved into optimizing NeRF (*e.g.* [10,16,23,24,30,32,37,54,68,88]) and Gaussian Splatting (*e.g.* [11,26,35,40, 78, 89]) for scenarios with limited shots and even single-shot contexts (*e.g.* PixelNeRF [83]), focusing on density fields without explicit 3D geometrical storage. DS-NeRF [14] regulates the NeRF rendering with monocular depth estimation, enhancing the quality and speed of few-view optimization. SfMNeRF [6] optimizes the left-right reprojection loss in 3D and also applies the depth smoothness term. FreeNeRF [79] regularizes the input frequency of the NeRFs and the occluded regions to improve the performance at the few-views setup. The evolution of zero-shot single-view 3D reconstruction has been significantly propelled by advances in multi-modal diffusion models and zero-shot 3D synthesis technologies [4, 5, 42, 45, 50, 57, 58, 64, 66]. Recently, ReconFusion [76] and ZeroMVS [60] utilized the 2D diffusion prior to greatly enhance the quality of novel view synthesis with very few views, achieving state-of-the-art in this domain. Our TrackNeRF does not use any generative priors to enhance the sparse view reconstruction but rather relies on generic geometric cues and feature tracks from the same scene that are more generalizable and tackle both sparse *and* noisy poses.

NeRFs with no Pose Requirements. Many works [3, 7–9, 12, 13, 20, 21, 29, 34, 36, 41, 54, 55, 59, 65, 72, 90] have tried to optimize NeRF from noisy pose or without pose. NeRF−− [75] is a pioneering work to investigate fitting a NeRF jointly with the camera poses for novel view synthesis. BARF [43] adds frequency modulation coordinate embedding in NeRF to greatly enhance the optimization of the cameras in noisy pose cases. SPARF [70] introduces a pairwise correspondence loss to the NeRF formulation and shows great performance in the case of *jointly* noisy and sparse views, a combined setup that was not properly performed previously. Our TrackNeRF is different because it assumes a global geometry loss that constrains *all* the cameras jointly based on the feature tracks.

3 Preliminaries

NeRF. Given a set of sparse views and associated camera parameters, NeRF [52] learns an implicit neural representation to represent the 3D scene. Let $R_i \in SO(3)$ and $\mathbf{t}_i \in \mathbb{R}^3$ denote the rotation and translation of camera pose i, respectively. The camera-to-world transformation of the i-th camera is denoted as $P_i = [R_i|\mathbf{t}_i] \in SE(3)$.

To render a pixel $\mathbf{p} \in \mathbb{R}^2$ from a given camera with pose P_i, NeRF traces a ray from the projection of the camera center $\mathbf{t}_i$ along the direction defined by $\mathbf{d}_{i,\mathrm{p}} = R_i K_i^{-1} \bar{\mathbf{p}}$ in the world coordinate system where $K \in \mathbb{R}^{3\times3}$ is the camera's intrinsic matrix and $\bar{\mathbf{p}}$ is homogeneous representation of $\mathbf{p}$. We then discretely

sample M points along the ray, bounded by the near and far planes, to predict the color $\hat{\mathbf{I}}_{i,\mathrm{p}}$ of a pixel from the radiance field as:

$$\hat{\mathbf{I}}_{i,\mathrm{p}} = \hat{I}(\mathbf{p}; \theta, P_i) = \sum_{m=1}^{M} \alpha_m \mathbf{c}_m \,, \tag{1}$$

where $\hat{I}$ is the RGB rendering function, $\{(\mathbf{c}_m, \sigma_m)\}_{m=1}^{M}$ are the color and volume density of sampled points predicted by a radiance field parameterized by θ. Let $\hat{z}(\cdot)$ be the depth rendering functions and z_m be the ray depth at sampled point m. We approximately estimate the depth of the scene perceived from $\mathbf{p}$ as,

$$\hat{z}_{i,\mathrm{p}} = \hat{z}(\mathbf{p}; \theta, P_i) = \sum_{m=1}^{M} \alpha_m z_m \,. \tag{2}$$

Photometric loss. NeRF approaches [13, 39, 43, 70, 75, 77] typically use a photometric loss to optimize radiance field parameters θ as well as camera poses $\hat{\mathcal{P}}$. Let $\hat{P}_i$ denote the pose estimate for the i-th training image. The photometric loss is defined as follows:

$$\mathcal{L}_{\text{Photometric}}(\theta, \hat{\mathcal{P}}) = \frac{1}{n} \sum_{i=1}^{n} \sum_{\mathrm{p}} \left\| I_i(\mathbf{p}) - \hat{I}(\mathbf{p}; \theta, \hat{P}_i) \right\|_2^2 . \tag{3}$$

where n is the number of training images. Different from these approaches, we propose a track reprojection loss that effectively reduces the negative effects of noisy camera poses on NeRF results.

4 TrackNeRF

Given the fact that all sparse views are shot from a single holistic 3D scene, corresponding points from *all* views rendered from the 3D model should ideally be projected back to the same 3D landmark. Unfortunately, even the most recent state-of-the-art methods SPARF [70] and CorresNeRF [30,39] only consider local matching consistency from a pair of renderings to train NeRF from sparse and noisy views. These works fail to exploit holistic consistency across all views. Inspired by the bundle adjustment from Structure-from-Motion [44,62], our work follows the track-wise objective of BA instead:

$$E_{BA} = \sum_{k} \sum_{(\mathbf{u}_i, \mathbf{v}_j) \in \mathcal{T}_{(k)}} \| h(\mathbf{u}_i) - \mathbf{v}_j \| \tag{4}$$

where $(\mathbf{u}_i, \mathbf{v}_j)$ is a correspondence inside feature track $\mathcal{T}_k$ between pixels $\mathbf{u}_i$ and $\mathbf{v}_j$ from views i and j respectively. The function h lifts a pixel onto its 3D location and projects it to a different view. Note that $\mathcal{T}_k$ considers correspondences between points across all visible views rather than pairs of views as in [39, 70]. This track-wise loss encourages all pixels in a feature track to correspond to the

same 3D landmark and enforces holistic consistency. In our paper, we propose to jointly optimize the matching correspondence for the whole feature track that aligns with the concept of *bundle adjustment* (BA) in Eq. 4. This differs our work from BARF [43], which only considers position encoding strategy without multiview correspondences and objective of BA. We elaborate on the formalization of this track-wise objective in the context of NeRF and details of our method next.

4.1 Track Adjustment

Track Extraction. Given a set of images $\{\mathcal{I}\}$ of size N and a feature matcher [69] $\mathcal{F}$, we can extract feature correspondences for every image pair. If a correspondence $(\mathbf{u}, \mathbf{v})$ between $\mathcal{I}_i$ and $\mathcal{I}_j$ is found, and there is also a correspondence $(\mathbf{v}, \mathbf{q})$ between $\mathcal{I}_j$ and $\mathcal{I}_k$, then it implies a transitive relationship that can be extended to form a continuous feature track. By connecting all such correspondences, we can get feature tracks $\{\mathcal{T}\}$, with $\mathcal{T}_k = \{\mathbf{u}, \mathbf{v}, \mathbf{q}, ...\}$, where $\mathbf{u}, \mathbf{v}, \mathbf{q}$ are pixel coordinates from different images corresponding to the same 3D point. Note that although we are using a dense correspondence model [69], we adopt the term *"feature"* in our paper to align with the concept of *"track"*.

Track Keypoint Adjustment. Recent advances [44,86] in SfM [62] have shown significant benefits by optimizing feature tracks before triangulation and bundle adjustment. To obtain more accurate feature tracks, for each track $\mathcal{T}_k$, we adhere to PixSfM [44] by doing track-wise keypoint adjustment to encourage multiview consistency:

$$E^k_{TKA} = \sum_{(\mathbf{u}_i, \mathbf{v}_i) \in \mathcal{T}_{(k)}} w_{\mathbf{u}_i \mathbf{v}_j} \|\mathbf{F}(\mathbf{u}_i) - \mathbf{F}(\mathbf{v}_j)\| \tag{5}$$

where $\mathbf{F}$ is the function that extracts pixel-wise features for keypoints, and $w_{\mathbf{u}_i \mathbf{v}_j}$ is the matching confidence from our matcher [69]. The objective of Eq. 5 will refine the location of matched keypoints by minimizing the difference of pixel-wise features through traceable numeric gradients. Since the refinement optimizes the feature-metric consistency inside the whole feature track, we obtain more accurate keypoints for the supervision of NeRF.

4.2 Track Reprojection Loss

Previous methods [6,30,39,70] only exploit optimizing NeRF with pairwise correspondence from two views independently during each iteration. However, feature extractors [15, 48, 69] working on single view are easily perturbed by appearance variations [61] thus may introduce multiview inconsistency for matching and reconstruction [17]. To alleviate this, we propose our track reprojection loss to enforce global geometric consistency for the radiance field, where multiview geometry constraints for the feature tracks are optimized simultaneously.

The track projection loss is based on the principle of bundle adjustment in Eq. 4, *i.e.*, all correspondences in a feature track $\mathcal{T}_k$ should correspond to the same 3D landmark (see Fig. 2). In particular, let $(\mathbf{u}_i, \mathbf{v}_j)$ be a pair of matching pixels between image $\mathcal{I}_i$ and $\mathcal{I}_j$ sampled from a feature track $\mathcal{T}_k$, where $\mathbf{v}_j$ is a pixel in $\mathcal{I}_j$. For $\mathbf{v}_i$ and its depth estimated by Eq. 2, we can obtain its corresponding 3D point in the world coordinate system using camera projection and camera-to-world transformation. We then reproject the corresponding 3D point of $\mathbf{v}_j$ to the image plane of image $\mathcal{I}_i$. Since $(\mathbf{u}_i, \mathbf{v}_j)$ corresponds to the same 3D point in the world coordinate system, the reprojection of $\mathbf{v}_j$ should overlap with pixel $\mathbf{u}_i$. Therefore, given the sampled feature track $\mathcal{T}_k$ optimized by Eq. 5 and consisting of *all-to-all* correspondences like $(\mathbf{u}_i, \mathbf{v}_j)$, our track reprojection loss minimizes the overall distance between all such $\mathbf{u}_i$ and the reprojection of $\mathbf{v}_j$ inside $\mathcal{T}_k$ as follows:

$$\mathcal{L}_{\text{Track}} = \sum_{k} \sum_{(\mathbf{u}_i, \mathbf{v}_j) \in \mathcal{T}_k} \frac{1}{|\mathcal{T}_k|} \rho \Big(\mathbf{u}_i - \pi \left(\hat{P}_i^{-1} \hat{P}_j \;\; \pi^{-1}(\mathbf{v}_j, \hat{z}(\mathbf{v}_j; \theta, \hat{P}_i)) \right) \Big) \tag{6}$$

where ρ is the Huber loss function [27], π is the camera projection operator which maps a 3D point in the camera coordinate system to the image plane, π^{-1} is the backprojection operator which projects a pixel v_j back to the camera coordinate system using the pixel's depth $\hat{z}(\mathbf{v}_j; \theta, \hat{P}_i)$ estimated by Eq. 2, and $\hat{P}_i^{-1}$ projects a 3D point in the world coordinate system to the camera coordinate system of image $\mathcal{I}_i$.

4.3 Depth Regularization

NeRF [52] is trained only using photometric loss, so it suffers from poor geometry, especially in sparse view settings. As reported by previous methods [54, 56, 79], depth ambiguity results in many floaters in the trained NeRF and significantly degrades the quality of novel view synthesis. To regularize the internal geometry of the radiance field, we also introduce a depth regularization loss to encourage depth gradients to align with rendered image gradients following the practice of monocular depth estimation [22, 28, 82]:

$$\mathcal{L}_{\text{Depth}} = \sum_{i,j}^{\Psi} |\nabla_x D_{ij}^{\Psi}| e^{-\|\nabla_x I_{ij}^{\Psi}\|} + |\nabla_y D_{ij}^{\Psi}| e^{-\|\nabla_y I_{ij}^{\Psi}\|} \tag{7}$$

where ∇ is the vector differential operator, D is the disparity map by taking the reciprocal of Eq. 2, I is the RGB image, and Ψ represents the sampled image patch for regularization.

4.4 NeRF training

Combining the L2 photometric loss, the depth regularization loss, and the proposed track reprojection loss, we optimize our TrackNeRF from sparse views

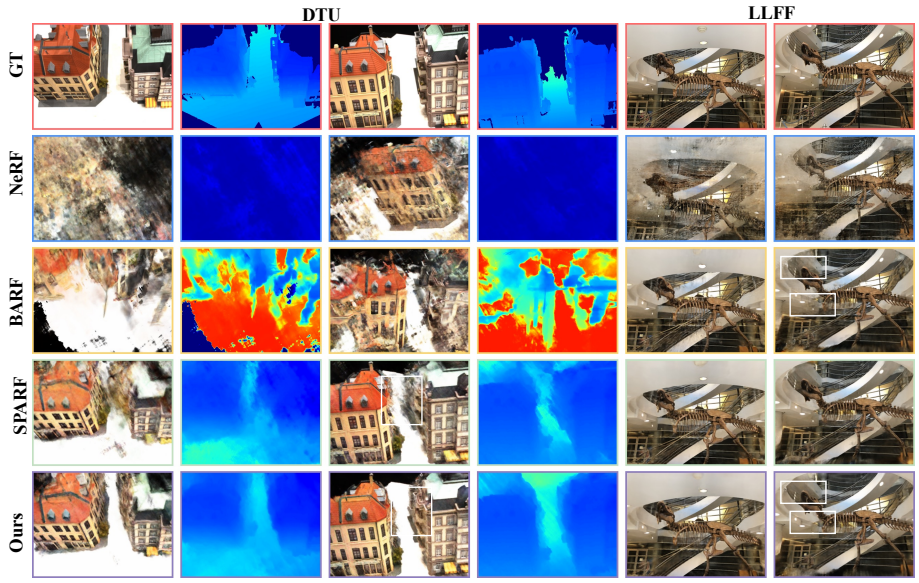


Fig. 3: Quanlitative Comparison on DTU [33] and LLFF [51]. We show views from the test view split of both datasets to visually compare our TrackNeRF renderings to the baselines. For DTU dataset where GT depth maps are available, we additionally visualize the rendered depth by Eq. 2 to compare the learned geometry.

with possibly noisy poses as follows:

$$\mathcal{L} = \mathcal{L}_{\text{Photometric}} + \lambda_{\text{Depth}}\mathcal{L}_{\text{Depth}} + \lambda_{\text{Track}}\mathcal{L}_{\text{Track}} \tag{8}$$

where λ_{Depth} and λ_{Track} are weighting factors of $\mathcal{L}_{\text{Depth}}$ and $\mathcal{L}_{\text{Track}}$, respectively.

5 Experiments and Results

5.1 Experiment Setups

Datasets and Metrics. We extensively evaluate our proposed TrackNeRF on **DTU** [33] and **LLFF** [51] datasets under various settings. DTU is a challenging benchmark as there are usually wide baselines between different views. We follow the split of PixelNeRF [83] to conduct our evaluation on the test split of 15 scenes, while we also report the results with masking background as done by [70, 76,79]. For the LLFF dataset, we select every 8^{th} image for testing as NeRF [52]. Regarding the metrics, we adopt PSNR, SSIM [74], and LPIPS [38,85] following the community standard. Since ground truth depth maps are available on DTU, we also report the mean depth absolute error (DE) like SPARF [70].
Implementation details. We follow the training strategy of SPARF [70] for a fair comparison, where we jointly optimize NeRF with poses at the first stage

Table 1: DTU Evaluation (3, 6 & 9 Noisy Views). We evaluate methods for unseen view rendering and camera extrinsic recovery on the DTU dataset [33], using initial poses that are noisy and vary in terms of the number of input views (3, 6, or 9). We introduce noise to these poses by adding 15% Gaussian noise to the true poses. The rotation errors are measured in degrees, while the translation errors are scaled by a factor of 100. The results in parentheses (·) are obtained after applying a mask to the background. Our TrackNeRF achieves the best performance in all setups.

	Method	Rot. ↓	Trans. ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
3 input views	BARF [43]	10.33	51.5	10.71 (9.76)	0.43 (0.62)	0.59 (0.36)	1.90
	RegBARF [43, 54]	11.20	52.8	10.38 (9.20)	0.45 (0.62)	0.61 (0.38)	2.33
	DistBARF [3, 43]	11.69	55.7	9.50 (9.15)	0.34 (0.76)	0.67 (0.36)	1.90
	SCNeRF [34]	3.44	16.4	12.04 (11.71)	0.45 (0.66)	0.52 (0.30)	0.85
	SPARF [70]	*1.81*	*5.0*	*17.74 (18.92)*	*0.71 (0.83)*	*0.26 (0.13)*	*0.12*
	TrackNeRF (Ours)	**1.12**	**2.48**	**18.53 (19.65)**	**0.73 (0.83)**	**0.25 (0.13)**	**0.11**
6 input views	BARF [43]	9.20	31.1	14.02 (14.22)	0.54 (0.69)	0.46 (0.27)	0.49
	RegBARF [43, 54]	9.19	26.63	14.59 (14.58)	0.57 (0.70)	0.44 (0.27)	0.32
	DistBARF [3, 43]	8.96	28.85	14.31 (14.60)	0.55 (0.70)	0.43 (0.26)	0.53
	SCNeRF [34]	4.10	12.80	17.76 (18.16)	0.70 (0.80)	0.31 (0.18)	0.28
	SPARF [70]	*1.31*	*2.7*	*21.39 (22.01)*	*0.81 (0.88)*	*0.18 (0.10)*	*0.09*
	TrackNeRF (Ours)	**0.24**	**0.65**	**22.78 (23.66)**	**0.84 (0.89)**	**0.14 (0.08)**	**0.06**
9 input views	BARF [43]	8.34	26.72	16.20 (16.38)	0.60 (0.73)	0.38 (0.22)	0.35
	RegBARF [43, 54]	5.28	18.51	18.98 (19.08)	0.67 (0.77)	0.29 (0.18)	0.23
	DistBARF [3, 43]	7.00	26.42	16.18 (16.27)	0.58 (0.71)	0.37 (0.22)	0.29
	SCNeRF [34]	4.76	16.25	18.19 (18.01)	0.69 (0.81)	0.31 (0.17)	0.31
	SPARF [70]	*1.15*	*2.55*	*24.69 (25.05)*	*0.88 (0.92)*	*0.12(0.06)*	*0.06*
	TrackNeRF (Ours)	**0.25**	**0.70**	**25.57 (26.03)**	**0.89 (0.92)**	**0.11 (0.06)**	**0.05**

and then only finetune NeRF at the second stage. We also utilize the same correspondence network, PDCNet++ [69], as SPARF for fairness. We adopt 6-DoF camera pose representation in [87]. We sample random rays for photometric loss, random feature tracks for track loss, and random pixel patches for depth loss. The depth regularization loss is only enabled in the second stage after pose optimization.

5.2 Main Results

DTU noisy 3 views. We keep the pose setting from previous methods [43, 70] by adding 15% additive gaussian noise to the groundtruth camera poses. As shown in Tab. 1, TrackNeRF achieves state-of-the-art performance across all the metrics. Specifically, for pose optimization, TrackNeRF nearly halves both the rotation and translation errors. It also improves PSNR by a remarkable margin of ~ 0.8.

DTU noisy 6 views and 9 views. We extend similar settings to 6-view and 9-view cases, the results of which are presented in Tab. 1. Surprisingly, the biggest improvement is observed when there are *moderate sparse* views, where a 1.65

Table 2: DTU Evaluation with Ground Truth Poses (3 Views). We show the evaluation on DTU [33] with three input views and *ground truth poses*, where the first column lists the schemes of methods and their architectures (**VN**: Vanilla NeRF , **MN**: MipNeRF, **PN**: PixelNeRF , **D**: Diffusion).

Archit.	Method	Settings	PSNR ↑ (masked)	SSIM ↑ (masked)	LPIPS ↓ (masked)
PN	PixelNeRF [83]	Trained	**19.36** (18.00)	0.70 (0.77)	0.32 (0.23)
PN+D	ReconFusion [76]	on	- (20.74)	- (0.88)	- (0.12)
D	ZeroNVS [60]	datasets	- (16.71)	- (0.72)	- (0.22)
MN	MipNeRF [2]		- (16.11)	- (0.40)	- (0.46)
MN	RegNeRF [54]		- (18.84)	- (0.57)	- (0.36)
MN	FreeNeRF [79]	Optimized	- (20.46)	- (0.83)	- (0.17)
VN	DS-NeRF [14]	per	16.52 (-)	0.54 (-)	0.48 (-)
VN	CorresNeRF [39]	scene	18.23 (20.58)	0.76 (0.77)	0.33 (0.13)
VN	SPARF [70]		18.30 *(21.01)*	*0.78 (0.87)*	*0.21 (0.10)*
VN	**TrackNeRF (ours)**		*18.78* **(21.45)**	**0.79 (0.88)**	**0.20 (0.10)**

boost on PSNR and huge reduction of pose error can be observed in the 6-view case. We believe such improvements benefit from multiview consistency from longer feature tracks. When the views become denser, we can still achieve 1 PSNR boost and smaller pose drift in the 9-view case.

DTU 3 views with ground-truth (GT) poses. Many approaches [39, 76, 79] assume precise GT camera poses available as the sparse view setup, although popular SfM methods like COLMAP [62] often fail in sparse-view scenario [70, 73]. But we also compare our method with representative approaches using different backbone architecture and training settings, to show the effectiveness of our method even using GT poses. To realize this setup, we fix camera poses to be GT ones and train the NeRF with Eq. 8. As shown in Tab. 2, despite vanilla NeRF being used, our method achieves the best performance in PSNR and SSIM, compared with methods using vanilla NeRF or more advanced MipNeRF. It is worth noting that our method also outperforms diffusion methods like ReconFusion [76] and ZeroNVS [60] trained on large-scale scene data. These results show TrackNeRF improves sparse view rendering quality besides camera pose optimization.

LLFF 3 views without pose. For the forward-facing dataset LLFF, we start optimization by identical camera poses. As we can see in Tab. 3, the improvement is not that many compared to DTU, with a slightly better performance on PSNR over SPARF [70]. We think that the reason is that LLFF is a simple dataset without large camera translation and rotation, and thereby the improvement from our feature track consistency and camera pose accuracy are somehow saturated, as previous works [43, 75] have shown pose recovery can be done solely from the photometric loss. Therefore, we conduct our ablation studies mainly on DTU dataset in the following.

Table 3: LLFF Evaluation with No Poses (3 Views) We show the evaluation on the forward-facing dataset LLFF [51] (3 views) with initial identity poses. Due to the simple camera motion in this dataset, the usage of feature tracks in our method does not lead to significant improvements over SPARF here.

Method	Rot. ↓	Trans. ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BARF [43]	2.04	11.6	17.47	0.48	0.37
RegBARF [43, 54]	1.52	5.0	18.57	0.52	0.36
DistBARF [3, 43]	5.59	26.5	14.69	0.34	0.49
SCNeRF [34]	1.93	11.4	17.10	0.45	0.40
SPARF [70]	**0.53**	*2.8*	*19.58*	**0.61**	**0.31**
TrackNeRF (Ours)	*0.77*	**2.8**	**19.60**	*0.59*	*0.35*

Visualization. We visualize the test view renderings from these two datasets to compare the quality of novel view synthesis in Fig. 3. For the DTU dataset with large camera motions, NeRF [52, 75] completely fails while BARF [43] cannot recover reasonable poses. On the contrary, TrackNeRF can still preserve a sharp and high-quality layout compared to SPARF and GT. Also, we can find clear geometry structure from the rendered depth maps, which indicates that the introduced track optimization effectively helps the radiance field to learn multiview geometry consistency. Moreover, the floaters and artifacts near the camera or background are also significantly reduced by the introduced depth smoothness loss in Eq. 7. For the forward-facing LLFF dataset, NeRF [52, 75] is still struggling, while BARF is unable to generate meaningful results. In contrast, TrackNeRF clearly outperforms these methods in generating views with sharp details, especially in the regions we have highlighted.

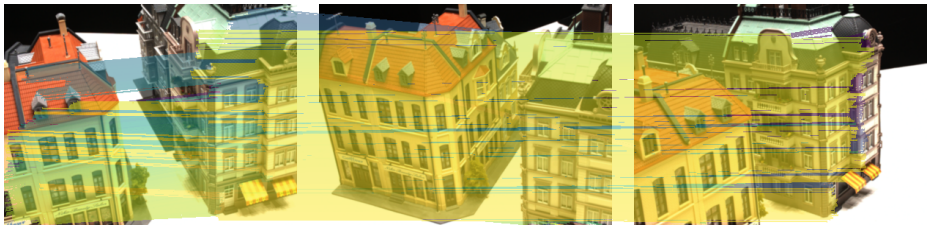
6 Ablation Study and Analysis

6.1 Ablation on components

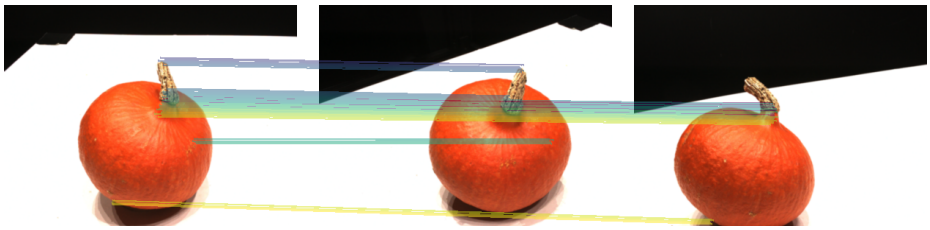
Starting from BARF [43]'s coarse-2-fine encoding inherited by other methods [70, 79], we ablate the effectiveness of the key components of our method in Tab. 4. The proposed feature track reprojection loss increases the performance significantly, especially on DTU dataset, where the camera motion is large and sparse views are far away from each other with wide baselines. We achieve +8 PSNR boost and also recover much more precise camera poses on DTU. Since LLFF is a simple forward-facing scene without much camera motion, our improvement is not that much, which aligns with our visual comparison in Fig. 3. Notably, track keypoint adjustment and depth regularization introduced by us also improve the performance further.

6.2 Ablation on robustness and effectiveness

Advantage on tolerance of noise level. We follow SPARF's [70] settings to certify the robustness of TrackNeRF on pose noise for fair comparison. As shown



(a) DTU scan 21. An example of a successful case for the correspondence.



(b) DTU scan 30, The only failed case for the correspondence we find.

Fig. 4: Visualization of Feature Tracks. In almost all scenes, we can always find dense enough and accurate correspondence like the one example from DTU scan 21 shown in Fig. 4a. We provide a rare case (scan 30) in Fig. 4b where the correspondence network [69] fails to find enough reliable correspondences, in which cases we uses a lower λ_{Track} for better performance.

Table 4: Ablation Study on Proposed Components. We conduct our ablation study on DTU and LLFF datasets for all scenes. **I**: Coarse to fine frequency encoding introduced by BARF [43]; **II**: Track reprojection loss proposed by us in Eq. 6; **III**: Track keypoints refinement in Eq. 5; **IV**: Depth regularization loss in Eq. 7.

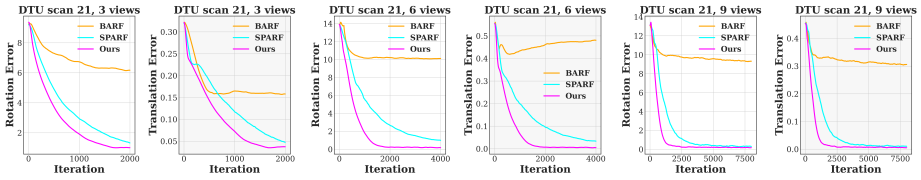
	Method	I	II	III	IV	Rot. ↓	Trans. ↓	PSNR ↑ (masked)	SSIM ↑ (masked)	LPIPS ↓ (masked)	DE ↓
DTU	BARF	✓				10.33	51.5	10.70 (9.8)	0.43 (0.62)	0.59 (0.36)	1.9
	Ours	✓	✓			1.40	2.61	18.07 (19.19)	0.72 (0.81)	0.27 (0.14)	0.11
	Ours	✓	✓	✓		1.12	2.48	18.39 (19.50)	0.73 (0.83)	0.25 (0.13)	0.11
	Ours	✓	✓	✓	✓	1.12	2.48	18.53 (19.65)	0.73 (0.83)	0.25 (0.13)	0.11
LLFF	BARF	✓				2.04	11.6	17.47	0.48	0.37	-
	Ours	✓	✓			0.80	2.96	19.34	0.58	0.36	-
	Ours	✓	✓	✓		0.77	2.80	19.51	0.58	0.35	-
	Ours	✓	✓	✓	✓	0.77	2.80	19.60	0.59	0.35	-

in Tab. 5, impressively, TrackNeRF is able to converge under 35% of Gaussian noise. While SPARF [70] reports failure on 20% noise, TrackNeRF shows stronger robustness to pose noise. As standard benchmarks always adopt 15% noise as the default scenario, our results show that exploring more challenging pose noise with wider baselines could be more interesting in future directions.

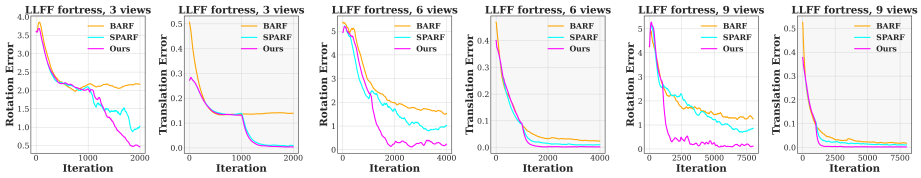
Table 5: Ablation study for noise level on DTU. We show an ablation study on the noise level for novel view synthesis and camera pose estimation with 3 views in DTU [33]. The initial rotation and translation error is obtained by multiplying the noise level with random samples from $\mathcal{N}(\mathbf{0}, \mathbf{I}_6)$ on $\mathfrak{se}(3)$, and then transferring it back to SE(3).

Noise	Rot. ↓	Trans.↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE↓
0.05	0.24	0.61	15.83	0.67	0.21	0.06
0.15	0.22	0.60	16.14	0.67	0.21	0.07
0.25	0.31	0.59	15.32	0.62	0.27	0.06
0.35	0.51	1.59	14.88	0.60	0.28	0.06
0.45	28.28	145.07	8.34	0.32	0.54	1.84

Advantage on convergence speed. As shown in Fig. 5, TrackNeRF (purple curve) clearly converge faster than BARF and SPARF, especially under 6-view and 9-view settings, where we have longer feature tracks to guide the pose and underly geometry optimization. Note for a fair comparison, we force all three methods to sample the same number of rays during each batch.



(a) DTU, scan 21



(b) LLFF, fortress scene

Fig. 5: Comparison on the Convergence of Pose Optimization. We show convergence plots of BARF [43], SPARF [70] and our TrackNeRF on DTU and LLFF datasets. For a fair comparison, we keep sampling the same number of rays for each iteration as SPARF [70]. Plots with white background and gray background represent rotation and translation errors, respectively. Our TrackNeRF converges faster and to a lower loss than the state-of-the-art.

6.3 Analysis of feature tracks adequacy

Due to the sparse view settings for NeRF optimization, concerns may arise regarding the adequacy of feature tracks, both in terms of track number and track length. To address these potential concerns, we provide the track statistics as depicted in Fig. 4. Considering the most extreme DTU 3-view case, without generality, NeRF typically samples 2,048 rays per batch for an image with a typical resolution of $(300, 400)$. We find an average of $\sim$ 20k tracks of length 3. Given these statistics, it is evident that the density of feature tracks is sufficient to facilitate robust NeRF optimization without the risk of excessive sampling redundancy. This surfeit of feature tracks ensures that the optimization process is well-supported by adequate data, even in the most extreme sparse view conditions.

7 Limitations

A noteworthy scenario arises when only two views are available. Under these rather rare circumstances, our track loss reverts to an optimization based on pairwise correspondences. However, we demonstrate the superiority of our proposed track optimization approach, particularly within the context of a *moderate sparse* setting, such as the case with six noisy views in Tab. 1. Another limitation occurs when the matching network fails to get correct and sufficient correspondence as shown in Fig. 4b, which is also shared by [39, 70]. We believe this can be addressed by leveraging more advanced feature matchers, which we leave for future exploration.

8 Conclusions and Future Work

We propose TrackNeRF for novel view synthesis under sparse and noisy views. Our method introduces feature track optimization for joint learning of camera poses and neural radiance field. Our joint optmization seamlessly aligns with the objective of bundle adjustment (BA) that has been widely adopted in SfM as a golden practice for decades. TrackNeRF can tolerate greater pose noise and converge faster, benefiting from global BA. We also show that TrackNeRF can achieve higher rendering quality, restore more accurate poses, and even outperform advanced diffusion-based methods trained on large-scale datasets. Overall, we show that multiview correspondence is critical in sparse view settings, especially with inaccurate or without poses. Moreover, our contribution is orthogonal to those learning-based methods and can be easily integrated into them without bells and whistles. We believe TrackNeRF can inspire future interesting research and have further impacts on the community.

Acknowledgement. The research reported in this publication was supported by funding from KAUST Center of Excellence on GenAI under award number 5940, as well as, the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). Part of the support is also coming from the KAUST Ibn Rushd Postdoc Fellowship program.

References

1. Agarwal, S., Furukawa, Y., Snavely, N., Simon, I., Curless, B., Seitz, S.M., Szeliski, R.: Building rome in a day. Communications of the ACM **54**(10), 105–112 (2011)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: CVPR. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5470–5479 (June 2022)
4. Charatan, D., Li, S.L., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19457–19467 (2024)
5. Chen, R., Chen, Y., Jiao, N., Jia, K.: Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. arXiv preprint arXiv:2303.13873 (2023)
6. Chen, S., Li, J., Zhang, Y., Zou, B.: Improving neural radiance fields with depth-aware optimization for novel view synthesis. arXiv preprint arXiv:2304.05218 (2023)
7. Chen, S., Zhang, Y., Xu, Y., Zou, B.: Structure-aware nerf without posed camera via epipolar constraint. arXiv preprint arXiv:2210.00183 (2022)
8. Chen, Y., Lee, G.H.: Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 24–34 (2023)
9. Chen, Y., Chen, X., Wang, X., Zhang, Q., Guo, Y., Shan, Y., Wang, F.: Local-to-global registration for bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8264–8273 (2023)
10. Chen, Y., Xu, H., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Explicit correspondence matching for generalizable neural radiance fields. arXiv preprint arXiv:2304.12294 (2023)
11. Chen, Y., Xu, H., Zheng, C., Zhuang, B., Pollefeys, M., Geiger, A., Cham, T.J., Cai, J.: Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images. arXiv preprint arXiv:2403.14627 (2024)
12. Cheng, K., Long, X., Yin, W., Wang, J., Wu, Z., Ma, Y., Wang, K., Chen, X., Chen, X.: Uc-nerf: Neural radiance field for under-calibrated multi-view cameras in autonomous driving. arXiv preprint arXiv:2311.16945 (2023)
13. Chng, S.F., Ramasinghe, S., Sherrah, J., Lucey, S.: Gaussian activated neural radiance fields for high fidelity reconstruction and pose estimation. In: European Conference on Computer Vision. pp. 264–280. Springer (2022)
14. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised NeRF: Fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
15. DeTone, D., Malisiewicz, T., Rabinovich, A.: Superpoint: Self-supervised interest point detection and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 224–236 (2018)
16. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: CVPR (2023)

17. Dusmanu, M., Schönberger, J.L., Pollefeys, M.: Multi-view optimization of local feature geometry. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 670–686. Springer (2020)
18. Faugeras, O.D.: What can be seen in three dimensions with an uncalibrated stereo rig? In: Computer Vision—ECCV'92: Second European Conference on Computer Vision Santa Margherita Ligure, Italy, May 19–22, 1992 Proceedings 2. pp. 563–578. Springer (1992)
19. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5501–5510 (June 2022)
20. Fu, H., Yu, X., Li, L., Zhang, L.: Cbarf: Cascaded bundle-adjusting neural radiance fields from imperfect camera poses. arXiv preprint arXiv:2310.09776 (2023)
21. Fu, Y., Liu, S., Kulkarni, A., Kautz, J., Efros, A.A., Wang, X.: Colmap-free 3d gaussian splatting (2023)
22. Godard, C., Mac Aodha, O., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 270–279 (2017)
23. Guangcong, Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. IEEE/CVF International Conference on Computer Vision (ICCV) (2023)
24. Hamdi, A., Ghanem, B., Nießner, M.: Sparf: Large-scale learning of 3d sparse radiance fields from few input images. arxiv (2022)
25. Hamdi, A., Giancola, S., Ghanem, B.: Voint cloud: Multi-view point cloud representation for 3d understanding. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=IpGgfpMucHj
26. Hamdi, A., Melas-Kyriazi, L., Mai, J., Qian, G., Liu, R., Vondrick, C., Ghanem, B., Vedaldi, A.: Ges: Generalized exponential splatting for efficient radiance field rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19812–19822 (June 2024)
27. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics, Springer (2009), https://books.google.com.sa/books?id=eBSgoAEACAAJ
28. Heise, P., Klose, S., Jensen, B., Knoll, A.: Pm-huber: Patchmatch with huber regularization for stereo matching. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2360–2367 (2013)
29. Hong, S., Jung, J., Shin, H., Yang, J., Kim, S., Luo, C.: Unifying correspondence, pose and nerf for pose-free novel view synthesis from stereo pairs. arXiv preprint arXiv:2312.07246 (2023)
30. Hu, S., Zhou, K., Li, K., Yu, L., Hong, L., Hu, T., Li, Z., Lee, G.H., Liu, Z.: Consistentnerf: Enhancing neural radiance fields with 3d consistency for sparse view synthesis. arXiv preprint arXiv:2305.11031 (2023)
31. Huang, P.H., Matzen, K., Kopf, J., Ahuja, N., Huang, J.B.: Deepmvs: Learning multi-view stereopsis. In: CVPR. pp. 2821–2830 (2018)
32. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: CVPR. pp. 5885–5894 (2021)
33. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: CVPR. pp. 406–413 (2014)

34. Jeong, Y., Ahn, S., Choy, C., Anandkumar, A., Cho, M., Park, J.: Self-calibrating neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5846–5854 (2021)
35. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023), `https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/`
36. Kim, I., Choi, M., Kim, H.J.: Up-nerf: Unconstrained pose prior-free neural radiance field. Advances in Neural Information Processing Systems **36** (2024)
37. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: CVPR. pp. 12912–12921 (2022)
38. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems **25** (2012)
39. Lao, Y., Xu, X., Cai, Z., Liu, X., Zhao, H.: Corresnerf: Image correspondence priors for neural radiance fields. arXiv preprint arXiv:2312.06642 (2023)
40. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20775–20785 (2024)
41. Li, M., Wang, P., Zhao, L., Liao, B., Liu, P.: Usb-nerf: Unrolling shutter bundle adjusted neural radiance fields. arXiv preprint arXiv:2310.02687 (2023)
42. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR (2023)
43. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: IEEE International Conference on Computer Vision (ICCV) (2021)
44. Lindenberger, P., Sarlin, P.E., Larsson, V., Pollefeys, M.: Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. In: ICCV (2021)
45. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. arXiv preprint arXiv:2303.11328 (2023)
46. Liu, T., Ye, X., Zhao, W., Pan, Z., Shi, M., Cao, Z.: When epipolar constraint meets non-local operators in multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 18088–18097 (October 2023)
47. Lombardi, S., Simon, T., Saragih, J., Schwartz, G., Lehrmann, A., Sheikh, Y.: Neural volumes: Learning dynamic renderable volumes from images. arXiv preprint arXiv:1906.07751 (2019)
48. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV **60**, 91–110 (2004)
49. Mai, J., Hamdi, A., Giancola, S., Zhao, C., Ghanem, B.: Egoloc: Revisiting 3d object localization from egocentric videos with visual queries. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 45–57 (October 2023)
50. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Realfusion: 360{\deg} reconstruction of any object from a single image. In: CVPR (2023)
51. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)

52. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV. pp. 405–421. Springer (2020)
53. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graph. **41**(4), 102:1–102:15 (Jul 2022). https://doi.org/10.1145/3528223.3530127, https://doi.org/10.1145/3528223.3530127
54. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
55. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. ACM Trans. Graph. (2023)
56. Philip, J., Deschaintre, V.: Floaters No More: Radiance Field Gradient Scaling for Improved Near-Camera Training. In: Ritschel, T., Weidlich, A. (eds.) Eurographics Symposium on Rendering. The Eurographics Association (2023). https://doi.org/10.2312/sr.20231122
57. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. ICLR (2022)
58. Qian, G., Mai, J., Hamdi, A., Ren, J., Siarohin, A., Li, B., Lee, H.Y., Skorokhodov, I., Wonka, P., Tulyakov, S., et al.: Magic123: One image to high-quality 3d object generation using both 2d and 3d diffusion priors. In: ICLR (2024)
59. Rosinol, A., Leonard, J.J., Carlone, L.: Nerf-slam: Real-time dense monocular slam with neural radiance fields. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3437–3444. IEEE (2023)
60. Sargent, K., Li, Z., Shah, T., Herrmann, C., Yu, H.X., Zhang, Y., Chan, E.R., Lagun, D., Fei-Fei, L., Sun, D., Wu, J.: ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. CVPR, 2024 (2023)
61. Sattler, T., Maddern, W., Toft, C., Torii, A., Hammarstrand, L., Stenborg, E., Safari, D., Okutomi, M., Pollefeys, M., Sivic, J., et al.: Benchmarking 6dof outdoor visual localization in changing conditions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8601–8610 (2018)
62. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
63. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV) (2016)
64. Seo, H., Kim, H., Kim, G., Chun, S.Y.: Ditto-nerf: Diffusion-based iterative text to omni-directional 3d model. arXiv preprint arXiv:2304.02827 (2023)
65. Sun, Y., Wang, X., Zhang, Y., Zhang, J., Jiang, C., Guo, Y., Wang, F.: icomma: Inverting 3d gaussians splatting for camera pose estimation via comparing and matching. arXiv preprint arXiv:2312.09031 (2023)
66. Szymanowicz, S., Rupprecht, C., Vedaldi, A.: Splatter image: Ultra-fast single-view 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10208–10217 (2024)
67. Tagliasacchi, A., Mildenhall, B.: Volume rendering digest (for nerf). arXiv preprint arXiv:2209.02417 (2022)
68. Tristram, F., Gasperini, S., Tombari, F., Navab, N., Busam, B.: Re-nerfing: Enforcing geometric constraints on neural radiance fields through novel views synthesis. In: arXiv preprint, under review (2023)

69. Truong, P., Danelljan, M., Timofte, R., Van Gool, L.: Pdc-net+: Enhanced probabilistic dense correspondence network. IEEE Transactions on Pattern Analysis and Machine Intelligence (2023)
70. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR (2023)
71. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. In: CVPR. pp. 5481–5490. IEEE (2022)
72. Wang, P., Zhao, L., Ma, R., Liu, P.: Bad-nerf: Bundle adjusted deblur neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4170–4179 (2023)
73. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy (2023)
74. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
75. Wang, Z., Wu, S., Xie, W., Chen, M., Prisacariu, V.A.: NeRF−−: Neural radiance fields without known camera parameters. arXiv preprint arXiv:2102.07064 (2021)
76. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., Holynski, A.: Reconfusion: 3d reconstruction with diffusion priors. arXiv (2023)
77. Xia, Y., Tang, H., Timofte, R., Van Gool, L.: Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. arXiv preprint arXiv:2210.04553 (2022)
78. Xiong, H., Muttukuru, S., Upadhyay, R., Chari, P., Kadambi, A.: Sparsegs: Real-time 360 {\deg} sparse view synthesis using gaussian splatting. arXiv preprint arXiv:2312.00206 (2023)
79. Yang, J., Pavone, M., Wang, Y.: Freenerf: Improving few-shot neural rendering with free frequency regularization. Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR) (2023)
80. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: Mvsnet: Depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 767–783 (2018)
81. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent mvsnet for high-resolution multi-view stereo depth inference. In: CVPR. pp. 5525–5534 (2019)
82. Yin, Z., Shi, J.: Geonet: Unsupervised learning of dense depth, optical flow and camera pose. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1983–1992 (2018)
83. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: CVPR. pp. 4578–4587 (2021)
84. Yu, Z., Gao, S.: Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: CVPR. pp. 1949–1958 (2020)
85. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR (2018)
86. Zhao, W., Liu, S., Guo, H., Wang, W., Liu, Y.J.: Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In: European Conference on Computer Vision. pp. 523–542. Springer (2022)
87. Zhou, Y., Barnes, C., Lu, J., Yang, J., Li, H.: On the continuity of rotation representations in neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5745–5753 (2019)

88. Zhu, H., He, T., Chen, Z.: Cmc: Few-shot novel view synthesis via cross-view multiplane consistency. arXiv preprint arXiv:2402.16407 (2024)
89. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. arXiv preprint arXiv:2312.00451 (2023)
90. Zhu, Z., Peng, S., Larsson, V., Xu, W., Bao, H., Cui, Z., Oswald, M.R., Pollefeys, M.: Nice-slam: Neural implicit scalable encoding for slam. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12786–12796 (2022)