Direct Alignment for Robust NeRF Learning

Ravi Garg[1,2], Shin-Fang Chng[1,2], and Simon Lucey[1,2]

[1] Adelaide University
[2] Australian Institute of Machine Learning
ravi.garg@adelaide.edu.au

Abstract. Differentiable volume rendering has evolved to be the prevalent optimization technique for creating implicit and explicit maps. Numerous efforts have explored the role of camera pose optimization and non-rigid tracking within Neural Radiance Fields (NeRFs). However, the relation between differentiable volume rendering and classical multi-view geometry remains under explored. In this work, we investigate the role of direct alignment in radiance field estimation by incorporating a simple but effective loss while training NeRFs. Armed with good practices for direct alignment while leveraging the effectiveness of volumetric representation in occlusion handling, our proposed framework is able to reconstruct real scenes from sparse or dense views at a much higher accuracy. We show despite relying on the photometric consistency, incorporating direct alignment improves view synthesis accuracy of NeRFs by 12% with known poses on LLFF dataset whereas joint optimization of pose and radiance field gets a boost in view synthesis accuracy of over 18% with rotation and translation errors going down by 64% and 57% respectively.

Keywords: Neural Radiance Field, Direct Alignment, Volume Rendering

1 Introduction

Since the advent of Neural Radiance Fields (NeRFs) for novel-view synthesis, myriads of developments have taken place in adapting the volume rendering frameworks to solve the classical problem reconstructing scene from multiple images. Originally introduced for view synthesis, NeRF [18] offered some very obvious advantages over the classic reconstruction methods. The first was compact, learnable scene representation, in the from of coordinate MLPs. These allowed for continuous volumetric representation to be distilled into a MLP weights, compressing maps to a few MegaBytes instead of typically a few GigaBytes required by the explicit counterparts (with reasonable volumetric resolutions). Second, it facilitated view-dependent rendering of surfaces without explicitly modeling the material properties and scene lighting which traditional reconstruction methods struggled with. Finally, optimizing the reconstruction error of a single pixel (working on a single ray) at a time, allowed volume rendering frameworks to bypass handling occlusions and multi-view matching which was known to be one of the biggest hindrance for direct alignment based reconstruction and flow estimation methods.

Leveraging these advantages, a plethora of work adapted volume rendering frameworks for rigid and non-rigid reconstructions in the recent past [4, 8, 23, 28, 30, 35, 36]. In this work, we try to understand some of the key advantages offered by volume rendering and implicit representations better, by comparing the radiance field based multi-view reconstruction with that of traditional structure-from-motion approaches to answer the question:

> *What concepts from classic structure-from-motion are still relevant in the world of NeRFs?*

In particular, we explore the utility of direct-alignment – that dominated the dense multi-view reconstruction for decades – while deploying differentiable volume rendering. Our view is that while it seems advantageous to bypass explicit occlusion reasoning, direct matching and triangulation of multiple-rays at the first glance, ignoring explicit alignment leads to over-fitting in learning NeRFs. This was first observed in [9] where the authors show that NeRFs learned undesirable multi-modal ray termination distributions when trained with small number of training views [3]. This practically means that a single opaque 3D surface element in the scene can be divided into multiple semi-transparent surfaces scattered along the viewing ray leading to ghosting artifacts. [9] proposed to utilize the sparse depths estimated by COLMAP [25, 26] to partially mitigate this over-fitting by explicitly enforcing unimodal ray terminations. This was achieved by aligning sparse depths with peaks on the ray termination distributions. Note that with known camera poses, depth supervision is equivalent to enforcing sparse feature matching across multiple-views which came as the byproduct of COLMAP.

Similar ideas of utilizing large and small baseline correspondences established by off-the-shelf optic-flow estimations, filtered by enforcing rigidity [9, 27] or with other heuristics such as enforcing cycle consistency of estimated pairwise correspondences have been used in literature to stabilize training of static and dynamic NeRFs [5,28,30]. In this work, we extend the idea of enforcing this multi-view matching to all the rays in the dataset without requiring accurate camera poses, approximate depths, optic-flow or sparse reconstructions from an off-the-shelf method. To achieve this, we advocate on-the-fly direct dense-alignment of multiple views using intermediate structure and motion estimates coming out of a framework like BARF [16] at all iterations.

We postulate that direct alignment and volume rendering frameworks have complementary advantages for dense reconstruction. In particular, volume rendering frameworks represent depths and colors of a pixel as occlusion aware alpha compositions of these quantities sampled on corresponding viewing ray. This provides a much desired tool to explicitly handle occlusions. Further, while dense alignment approaches work best with small baselines, a common template

[3] Notice that the multi-modality of ray terminations are important to model non-opaque surfaces

(in the form of a 3D radiance field) to align all input images allows for seamless integration of optic-flow that traditional dense reconstruction method struggle with. For example dense SLAM frameworks such as [20] utilize depthmap of the keyframe as a template to anchor multi-frame correspondences and rely on robust loss functions to model occlusions, thereby are restricted to match frames in a narrow window around keyframes. Frameworks such as Omnimotion [30] have already showcased the effectiveness of even 'quasi-3D' radiance fields to act as a suitable template to anchor multi-frame correspondences while facilitating occlusion reasoning and integrating sparse matching across large number of views.

In this work we propose to marry the impressive capabilities of radiance field in occlusion reasoning and flow-integration with vastly studied literature of direct alignment to facilitate multi-view consistency. We do so by introducing occlusion aware direct alignment loss to jointly optimize for radiance field and camera poses. We show that despite using simple color as the features for multi-view matching, i.e. making Lambertian surface assumption, the direct alignment assisted NeRFs outperforms the counterparts by a substantial margin on real sequences. Further, the direct alignment facilitates more accurate pose estimation with no initialization requirement when baseline for capture is small such as in forward facing LLFF dataset. In the standard setup of large baseline captures such as DTU, our method requires noisy camera pose estimate like most other joint radiance field and pose estimation methods but outperforms comparative baselines in both pose and structure estimation. While we deploy direct alignment loss on implicit radiance field estimation, the loss is generic to be used with any volumetric rendering pipeline including semi-implicit or explicit representation. Our work provides a fresh perspective on the efficacy of feature matching to assist volume rendering.

2 Literature Review

A full review of advances in neural radiance field is out of scope of this paper, below we present thematic review a subset of related approaches that informs the presented approach.

NeRF with Unknown Camera Poses: While classic gradient based optimization for per frame pose estimation was the natural choice to extend NeRFs, the naive attempt in doing so failed catastrophically with camera poses estimations diverging quickly to bad local minima. BARF [16] presented first framework to jointly optimize radiance field and poses and remain to date a go to approach which modern rigid NeRF learning that we base our work on. They proposed to gradually introduce positional encoding used in [18] for high frequency signal reconstructions. This gradual introduction of the high frequency positional encoding mimicked coarse-to-fine reconstruction leading to accurate camera pose estimations. A range of alternatives [4, 7, 8, 10, 13, 22, 31, 32] have been proposed to optimize camera pose while learning NeRFs. [8] proposes to use Gaussian Activation in place of ReLU to facilitate high fequency image synthesis without the positional encoding, [22] proposes to precondition cameras, [4, 27] uses off the shelf single view depth or optic flow predictions to assist camera pose estimation. The contributions of these works over [16] are orthogonal to the ours and can be used with direct matching.

NeRF over-fitting and additional supervision Many works have been proposed to address NeRF's overfitting problem, most of them focus on learning radiance field with sparse viewpoints. Solutions in this domain range from adding regularizations [3, 11, 14, 15, 21], providing additional supervision in the form of ground truth or estimated depth maps [5,9,24,34] or utilizing off the shelf sparse or dense feature correspondences [15, 27, 29, 30].

Methods using sparse depths/flow fields in many ways have embodied the popular approach or track first and reconstruct later from traditional large baseline multi-view reconstruction methods such as [1, 19, 25, 26]. These approaches however heavily relied upon powerful model fitting frameworks like RANSAC that are non-trivial to integrate with NeRF frameworks. Direct alignment [2] based structure and motion estimation systems such as [20] however use dense maps and can be integrated with NeRFs relatively easily as shown in this work.

In particular closest to our approach is [4] which propose depth guided direct alignment loss minimization similar to ours albeit it uses scaled and translated single-view depth predictions to align neighbouring images in a sequence. The alignment loss thus does not depend on the radiance field at all but provide meaningful gradients for camera poses directly. [4] In contrast, we use intermediate depth estimations from the radiance field for the image alignment and thus require no external depth estimator.

NeRF with Learned Features Another loosely connected line of work to ours focuses on pre-training viewpoint invariant features on large datasets, often combining information from multiple images, to form a proxy to matching loss [6,33]. These learned features are than used to inform coordinate networks (often coded as MLPs) about multi-view information implicitly to accurately learn radiance fields. In principle, while these methods do not explicitly match features across multiple test-views, the coordinate networks implicitly takes the role of learning per pixel cost volume optimization.

3 Methodology

In this section we set the mathematical notations and introduce the volume rendering framework introduced in [18] and joint coarse-to-fine camera pose estimation with radiance field learning. Further, we outline the direct matching loss introduced in this work followed by the simple occlusion reasoning mechanism that forms proposed approach.

[4] This enables [4] to estimate large baseline motions. [4] advocate annealing the alignment loss over the course of training to avoid baking in errors in single view depth predictions and can be seen as alternative initialization scheme to pose via depth guided image alignment loss instead of using COLMAP.

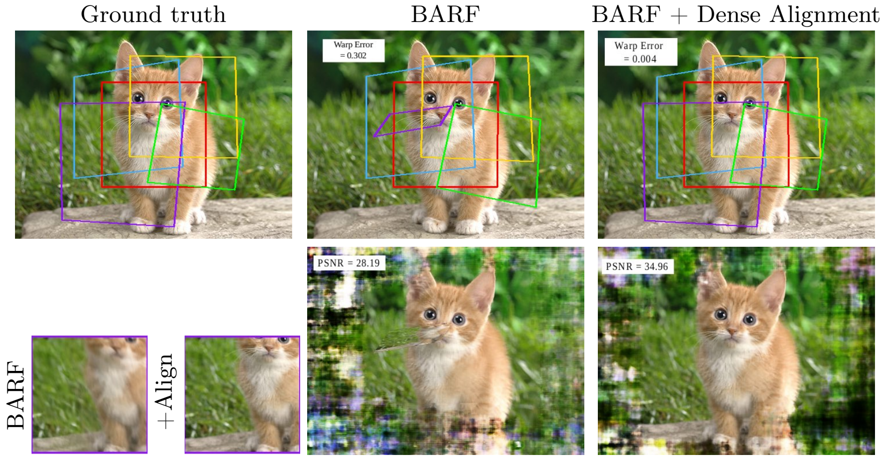


Fig. 1: Comparative analysis of image field learning and homography estimation with seed 9 and $\mathcal{H} = 0.1$. First row consists for homography visualisation for Ground truth, BARF without and with direct alignment respectively. Second row has selected crop reconstructions for both approaches followed by learned 2D field by BARF without and with alignment respectively.

3.1 Preliminary: Learning NeRF with camera pose estimation

In this section we layout a general problem of learning radiance field from a set of given images and notations used in the rest of the paper. We recommend BARF for details to the reader which our approach closely follows. Let us assume that we are given with F images $\{\mathcal{I}_f\}_{f=1}^{F}$ captured by a camera with intrinsic matrix $K \in \mathbb{R}^{3\times3}$ from unknown locations $\mathbf{P}_f = (\mathbf{R}_f, \mathbf{t}_f) \in SE(3)$. Our goal is to learn a continuous radiance field function in the form of the coordinate network $f(\mathbf{\Theta}, \mathbf{x}, \mathbf{v}) \rightarrow (\mathbf{c}, \sigma)$ that corresponds to the given images alongside camera locations $\mathbf{P}_f$'s.

The coordinate network is represented as a learnable multi-layer perceptron (we use the same architecture as BARF see supplementary material for details)

	Mean Warp Err. ($\Delta\mathcal{H}$)			Mean PSNR			Acc. $\Delta\mathcal{H}$ <0.025			Acc. $\Delta\mathcal{H}$ <0.05		
$\mathcal{H}$	BARF	+ Align	+SS	BARF	+Align	+SS	BARF	+Align	+SS	BARF	Align	+SS
0.05	0.018	**0.005**	0.007	35.17	34.94	**35.98**	6/9	**9/9**	**9/9**	9/9	**9/9**	**9/9**
0.1	0.22	0.06	**0.01**	29.63	33.83	**35.83**	1/9	8/9	**9/9**	3/9	8/9	9/9
0.15	0.38	0.14	**0.13**	25.60	30.67	**31.79**	0/9	4/9	**5/9**	0/9	**4/9**	**4/9**

Table 1: Panoramic stitching performance. We report homography estimation errors and patch reconstruction PSNR of vanilla BARF, RGB Aligned BARF without and RGB Aligned BARF with explicit Gaussian scale-space coarse to fine (depicted as SS) deployed while training NeRFs. We define successful homographies at different error thresholds an present accuracies to compare the basin on convergence for different methods.

with parameters $\mathbf{\Theta}$ that predicts the volume density σ and color $\mathbf{c}$ of the scene coordinate $\mathbf{x}$ as observed form the viewing direction $\mathbf{v} \in \mathbb{R}^3$. A pixel $\mathbf{u} \in \mathcal{R}^2$ in the image $\mathcal{I}_f$ thus can be rendered by integrating the implicit radiance field along a ray $\mathbf{r}(\mathbf{u}, d) = \mathbf{t}_f + d\mathbf{v}$ passing through the camera center $\mathbf{t}_f$ and $\mathbf{u}$ by:

$$\hat{\mathcal{I}}_f(\mathbf{u}) = \int_{d_n}^{d_f} \underbrace{T(\mathbf{u}, d)\sigma(\mathbf{r}(\mathbf{u}, d))}_{h(\mathbf{u},\mathbf{d})} \mathbf{c}(\mathbf{r}(\mathbf{u}, d))\delta d, \tag{1}$$

where viewing direction $\mathbf{v}$ is a normalized vector along the line that joins camera center t_f and pixel $\mathbf{u}$. Transmittance $T(\mathbf{u}, d)$ of a point on this ray at depth d is defined as $\exp(-\int_{d_n}^{d} \sigma(\mathbf{r}(u, d))\delta d)$.

Note that the ray termination distribution $h(\mathbf{u}, d) = T(\mathbf{u}, d)\sigma(\mathbf{r}(\mathbf{u}, d))$ for sampled depths quantify the probability that the ray hits the scene first at depth d and is desired to be unimodal for opaque surfaces. NeRFs are trained to minimize the following rendering loss summed over all pixels in the dataset:

$$L_{render} = \sum_{f=1}^{F} \sum_{\mathbf{u} \in \mathit{\Omega}(I_f)} \|\hat{\mathcal{I}}_f(\mathbf{u}, P_f; \mathbf{\Theta}) - \mathcal{I}_f(\mathbf{u})\|_2^2, \tag{2}$$

3.2 Direct Matching

We propose to align a reference image $\mathcal{I}_r$ with a set of surrogate views $s \in \mathcal{N}_r$ while training the NeRF. The alignment loss L_{align} is defined as:

$$\begin{aligned} L_{align} &= \sum_{r=1}^{F} \sum_{s \in \mathcal{N}_r} \sum_{\mathbf{u} \in \mathit{\Omega}(I_r)} \mathcal{V}(\mathbf{u_s}, \mathbf{u}) \|\mathcal{I}_s(\mathbf{u_s}) - \mathcal{I}_r(\mathbf{u})\|_\epsilon, && (3) \\ \mathbf{u}_s &= \mathit{\Pi}(KR_s(R_r^{-1}K^{-1}[\mathbf{u}, 1]^T \bar{d}(\mathbf{u}) - \mathbf{t}_r) + \mathbf{t}_s), && (4) \\ \bar{d}(\mathbf{u}) &= \sum_{d_{sample}} h(\mathbf{u}, d_{sample}) d_{sample} && (5) \\ \text{and } & \mathit{\Pi}([X, Y, Z]^T) = [X/Z, Y/Z]^T && (6) \end{aligned}$$

where $\mathcal{V}(p, q)$ is an indicator function depicting co-visibility of argument pixels in two given views, $\bar{d}\mathbf{u}$ is the estimated depth of the pixel $\mathbf{u}$ in the reference frame r obtained using alpha composition of the sample points d_{sample} on the ray $\mathbf{r}(\mathbf{u}, .)$ and $\|.\|_h$ represents huber penalty.

In practice, the L_{align} gets minimized using SGD where reference frame r and point samples $\mathbf{u}$ on the reference frame are selected randomly in every iteration. Unless specified, in this work we select a single reference image at random from the training set and consider the entire training set as surrogate views for the reference image.

Occlusion Reasoning We adapt a simplest strategy to model occlusion function $\mathbf{V}(\mathbf{u_s}, \mathbf{u})$ described in eq. 3. We estimate the depth $\bar{d}(\mathbf{u}_s)$ of the pixel $\bar(u_s)$ that correspond to the $\mathbf{u}$ in reference image. If the 3D point at depth $\bar{d}\mathbf{u}_s$ in front of the one we projected from the reference image, the $\mathbf{u}$ is considered occluded in frame s [5]. Formally:

$$\mathcal{V}(\mathbf{u}_s, \mathbf{u}) = \mathbb{1}[\bar{d}(\mathbf{u}_s) < \mathcal{T}_s(\mathbf{u}, \bar{d}(\mathbf{u})) - \delta_d] \tag{7}$$

$$\text{where } \mathcal{T}_s(\mathbf{u}, \bar{d}(\mathbf{u})) = \omega(K(R_s(R_r^{-1}K^{-1}[\mathbf{u}, 1]^T\bar{d}(\mathbf{u}) - \mathbf{t}_r) + \mathbf{t}_s)) \tag{8}$$

$$\text{and } \omega([X, Y, Z]^T) = Z \tag{9}$$

4 Experiments

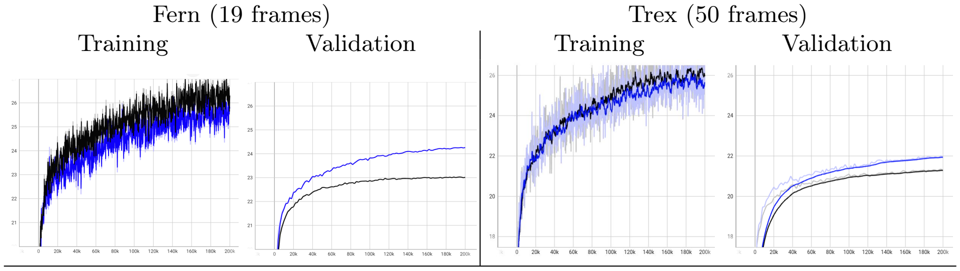
In this section, we present experimental results along the relevant ablation studies to justify the utility of the direct alignment in the radiance field learning. Section 4.1 compares the effectiveness of direct alignment for homography induced 2D image field learning. Section 4.2 focuses on how direct alignment reduces the overfitting of NeRFs. Finally in Section 4.3 tests our full model for multi-view 3D Reconstruction without on multiple real sequences.

4.1 2D image field and homography estimation

We follow BARF [16] to learn the 2D neural image field corresponding to a homography based panaroma stitching of given crops. This simple 2D variant of radiance field estimation with unknown cameras allow us to discard challenges in aligning non-lambartian surfaces and occlusion reasoning involved in the 3D counterpart but establish utility of direct alignment. Following BARF, a 2D coordinate network is designed to render the stitched image while pixel p_i in the training patch i is warped with the estimated homography $\mathcal{H}_i$ to rendered image. We add the homography induced direct image-alignment loss in this setup and compare the results with BARF baseline. For efficiency propose, at every iteration, we randomly select a reference crop r to align other crops to instead of considering all crop-pairs. To this end, pixel p_i from any crop i is warped by estimating composed homography $\mathcal{H}_{ri} = \mathcal{H}_r^{-1}.\mathcal{H}_i$ to its location $p_i^{'} = \mathcal{H}_{ri} \cdot p_i$ in the reference crop. Color constancy of the backward warps to reference crop for all surrogate views is penalized with huber function and minimized alongside the rendering MSE.

We generate nine random homographies each, with three scale-noise parameters $\mathcal{H} = \{0.05, 0.1, 0.15\}$ while using fix translation noise of 0.2 to analyse the efficacy of different approach with varying level of difficulty. Table 1 shows the crop reconstruction and homography estimation errors of different experiments and Figure 1 presents visual analysis for a single seed highlighting the key differences. It can be seen that image alignment helps in estimating accurate

[5] Note that as $\mathbf{u_s}$ is the projection of the point that is intersection of the two rays by definition it lies on the ray corresponding to pixel $\mathbf{u_s}$.



Fern (19 frames)

Training

Validation

Trex (50 frames)

Training

Validation

Fig. 2: NeRF training and validation PSNR with and **without** dense alignment. One can clearly observe that NeRFs overfit to training data despite having 50 frames to train on in Trex Sequence.

	NERF			**RGB Aligned**		
Seq	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fern	23.06	0.72	0.28	**24.50**	**0.75**	**0.25**
Leaves	13.60	0.21	0.59	**17.45**	**0.45**	**0.42**
Orchid	17.38	0.51	0.31	**17.85**	**0.54**	**0.30**
T-rex	21.82	0.76	0.22	**21.90**	**0.78**	**0.19**
Flower	23.10	0.67	0.24	**24.37**	**0.72**	**0.20**
Fortress	26.12	0.79	0.18	**28.74**	**0.84**	**0.10**
Horns	19.18	0.57	0.46	**22.45**	**0.72**	**0.31**
Room	**33.30**	**0.95**	**0.06**	32.85	0.95	0.08
Mean	22.20	0.65	0.29	**23.76**	**0.72**	**0.23**

Table 2: View synthesis accuracy of NeRF with Direct Dense Alignment of Images.

homographies for all difficulty levels and converges more frequently when compared to BARF. As the homography estimation becomes more challenging, the 2D field reconstruction without alignment loss has significantly lower PSNRs. In the interesting case of smaller homography perturbations the mean PSNR without alignment looks marginally better. However we observe that these differences are often within range of the *jitter* in the estimated 2D fields observed in the final few iterations. Further we observe in Table 1 that incorporating coarse-to-fine scale space estimations helps in minimizing the reconstruction loss as-well-as in image alignment to a large extent when the homographies are difficult. As expected you loose on precise homography estimation in simpler cases but the optimization has higher basin of convergence. These findings translates in training NeRFs with camera pose estimations presented below as well.

4.2 Nerf With Given Pose

Next we evaluate the efficiency of the proposed approach in learning radiance field on LLFF dataset [17]. We use COLMAP camera poses as fixed ground truth and evaluate the radiance field learning in isolation on real-scenes. Like section 4.1, for direct alignment we sample a random image from the training data as

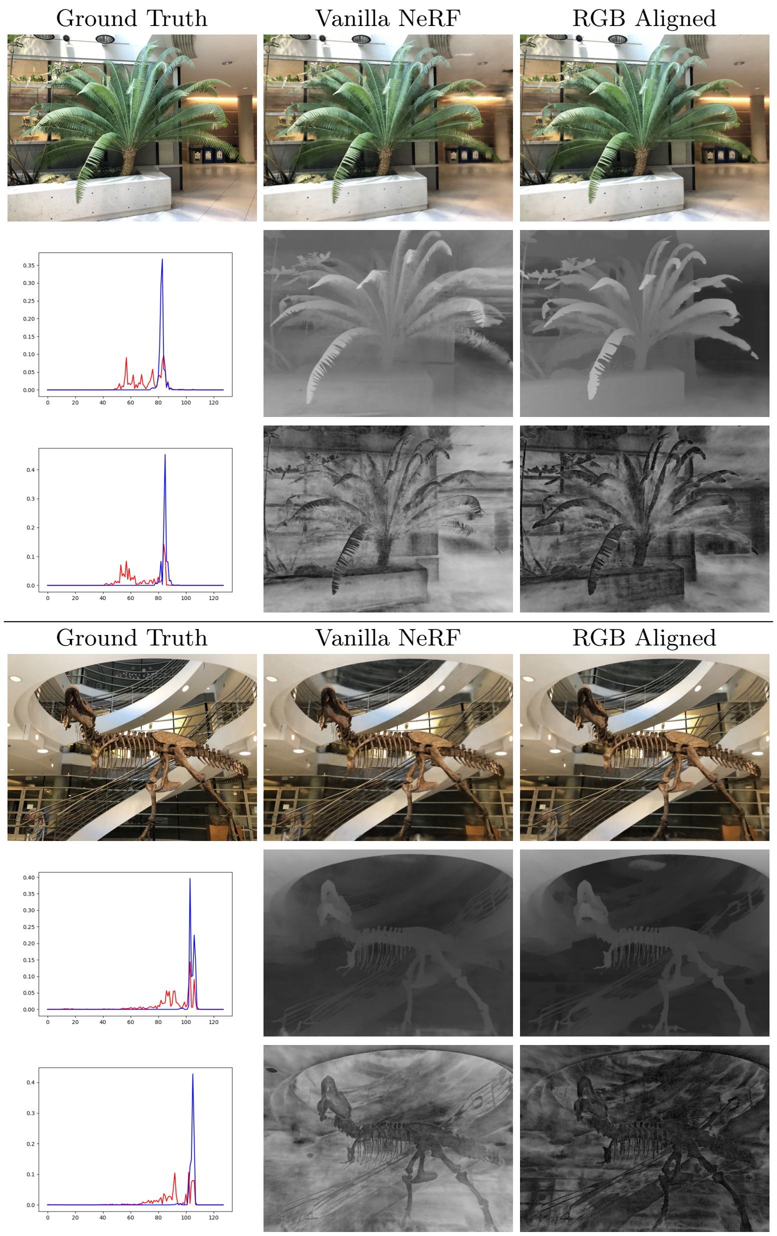


Fig. 3: Radiance field learned with and without matching given COLMAP poses. Top row from left to right show Test image, and renderings using estimated radiance field without and with direct alignment loss. column 1's, row2 and row3 show the ray termination distributions, with following columns showing estimated depth and entropy of ray distributions.

reference and minimize the dense alignment loss alongside the per-ray rendering loss for all the views in the training set. We use the BARF implementation without normalized device coordinates or coarse probability density function based fine samplings along the rays for simplicity in these experiments. In all LLFF experiments with known pose, network architecture, and training hyper-parameters are kept exactly same as BARF. Our method incorporating RGB alignment consistently boost radiance field reconstruction accuracies, with novel-view synthesis PSNR going up by 7% on average as shown in Table 2. The exception is Room sequence where the proposed method suffers due to large reflective regions present in the images. [6]

Overfitting Figure 2 presents training and validation PSNRs observed for the LLFF dataset on two selected sequences Fern with sparse and T-rex with large number of views. It is clear to see that NeRF overfits to training views and this problem can be mitigated by using direct alignment loss. We visualize these ray termination distributions for vanilla NeRF and presented approach for a few selected points in Figure 3 to confirm that the overfitting relates to the multi-modal ray terminations as hypothesised in [16]. To analyse this over-fitting more closely, we also visualize the normalized entropy of all the pixels for a validation image as a joint indicator of non-peaky ray terminations. It is important to note that our experimental setup does not use finer depth sample informed by the PDF predicted by coarse NeRF which naturally leads to overall flatter distributions when compared to that reported in [9].

It is important to note though that the over-fitting happens despite having large number of training views (50+ views in T-rex) and systematic mitigation of the over-fitting should help NeRF training for both sparse and dense view regime. This is a contrary observation to the conventional literature working on regularizing NeRF where [9, 27] show diminishing returns of using depth/flow regularization in scenarios with dense captures for LLFF dataset with [27].

4.3 Joint Camera Pose and Radiance Field Estimation

In this section we present the results on joint camera pose optimization and radiance field learning on LLFF and DTU sequences. We follow the same training and test split as used in BARF and similar evaluation protocol with the exception of not using run-time optimization for view-synthesis. While run-time optimization in BARF was used to isolate pose estimation noise to bring PSNR closer to be a measure of geometric error, the optimization ends up *peaking* into answers (images which are to be reconstructed) for view synthesis evaluations. Further, changing the pose of the query image to optimize rendering loss distorts relative camera locations between query and training images – i.e. changing the nature of view synthesis problem altogether. Thus, in absence of direct geometric evaluations such as via depthmaps, we advocate evaluating PSNR without

[6] We see small boast in performance by linearly decreasing the weight of alignment loss to zero and the results in 2 uses this scheduler. Ablation for scheduling can be find in supplementary material. Please note that all other reported results and visualization (except table 2) use no scheduling.

Seq.	**BARF**		**+Alignment**		**BARF**			**+Alignment**		
	E Rot	E Trans	E Rot	E Trans	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fern	0.197	1.82	**0.172**	**1.78**	21.12	0.63	**0.32**	**21.66**	**0.64**	0.33
Leaves	1.311	2.65	**1.019**	**2.29**	11.74	0.13	0.46	**12.18**	**0.14**	**0.42**
Orchid	0.580	3.94	**0.508**	**3.42**	13.50	0.19	**0.34**	**13.71**	**0.19**	0.35
T-rex	1.198	7.51	**0.185**	**2.37**	14.61	0.33	0.33	**18.26**	**0.53**	**0.25**
Flower	0.212	2.26	**0.202**	**1.88**	19.96	0.51	0.24	**21.47**	**0.59**	**0.21**
Fortress	0.372	3.14	**0.253**	**1.97**	23.25	0.49	0.14	**23.27**	**0.55**	0.14
Horns	2.950	13.97	**0.120**	**1.41**	11.43	0.30	0.57	**21.54**	**0.67**	**0.35**
Room	0.375	2.93	**0.071**	**1.07**	20.70	0.75	0.15	**29.19**	**0.91**	**0.13**
Mean	0.90	4.78	**0.32**	**2.02**	17.04	0.42	0.32	**20.16**	**0.53**	**0.27**

Table 3: BARF and BARF with Matching. Rotation error are in degrees where as the Translation errors are in millimeters. Note that reported reconstruction results do not use run time optimization as is done in [16].

Seq.	**BARF**		**+Alignment**		**BARF**			**+Alignment**		
	E Rot	E Trans	E Rot	E Trans	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
Fern	4.10	8.70	**3.13**	**6.65**	11.86	0.33	0.67	12.42	0.32	0.63
Leaves	2.00	4.30	**1.13**	**2.41**	10.88	0.11	0.48	11.27	0.10	0.42
Orchid	0.74	5.58	**0.38**	**2.15**	13.43	0.23	0.32	14.45	0.27	0.29
T-rex	8.22	40.82	**0.34**	**1.71**	10.47	0.34	0.67	17.04	0.49	0.31
Flower	**0.55**	2.40	0.60	**1.14**	17.83	0.44	0.31	16.02	0.30	0.34
Fortress	11.84	63.84	**2.19**	**13.31**	11.48	0.30	0.62	16.33	0.35	0.39
Horns	3.39	28.26	**1.71**	**13.96**	11.90	0.27	0.62	13.37	0.30	0.54
Room	0.65	4.49	**0.15**	**1.17**	16.91	0.65	0.25	21.19	0.77	0.21
Mean	3.99	19.93	**1.20**	**5.31**	13.09	0.33	0.49	**15.26**	**0.36**	**0.39**

Table 4: Results on LLLF sequence with first five images used for training with and without RGB Alignment. Rotation error are in degrees where as the Translation errors are in millimeters. Note that this experiment do not correspond to the once reported in DS-NeRF and other literature where training views are sampled uniformly to create with large baselines.

run-time optimization as an accurate measure of view-synthesis errors. Table 3 shows the quantitative results of jointly optimizing the radiance field and camera poses on standard train and test splits used in [16] on LLFF dataset. One can clearly see the advantage of using direct matching which boost BARF camera pose estimation on all sequences with an average of 60% in rotation and 53% for translation. In particular, Trex and Horns despite having more than 50 image for training sees maximum boost in performance. This improved camera pose accuracy translates directly to the view synthesis performance. Figure 4 shows the qualitative comparison for camera pose estimation and reconstruction for selected sequences. It can be seen easily that the BARF rendered images does not align with the ground truth and is shifted upwards while the generated depth maps have more ghosting artifacts and structural inaccuracies. NeRFs trained with direct alignment of RGB image mitigate both these problems and learns low entropy ray termination distributions with better poses. One can see the limitation of direct RGB alignment where violation in photo-consistency hampers reconstructions. However if the RGB alignment loss is replaced with alignment loss of DINO features (capturing non-localized information about local patches) and gradient images (capturing localizes features), these artifact can be avoided to an extent while semi transparent object such as the glass support for the tiny skull on the right can be recovered.

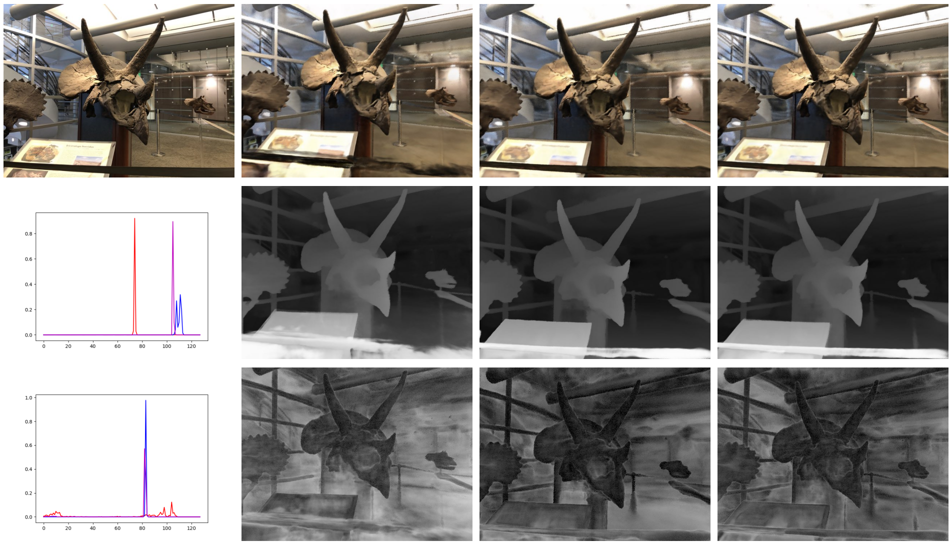


Fig. 4: Qualitative comparison on joint motion and structure estimation on Horns sequence. Top row shows test image follow by view synthesis from BARF, BARF with direct RGB alignment and BARF with alignment of DINO features and image gradients respectively. Next two row show the depthmaps and ray termination entropies for different methods in column 2-4 while column 1 show ray termination distributions for Vanilla BARF, and barf with RGB Alignment and DINO + GRAD Alignment.

		Scan IDs														
Methods	Metric	24	37	40	55	63	65	69	83	97	105	106	110	114	118	**mean**
BARF		0.80	2.08	0.47	0.62	**0.25**	0.65	3.02	0.27	0.38	**0.16**	0.32	**0.27**	0.40	0.54	0.73
+RGB	rotation	0.99	1.68	**0.18**	0.36	0.27	**0.37**	**0.24**	**0.23**	**0.26**	0.17	0.31	**0.27**	0.32	0.48	**0.44**
+RGB+SS		**0.76**	**0.35**	0.21	**0.25**	0.34	0.52	**0.21**	0.24	0.34	3.79	**0.28**	0.34	**0.28**	**0.44**	0.59
BARF	translation	**2.27**	3.36	1.59	2.21	**0.70**	2.41	7.01	0.60	1.10	**0.38**	1.02	**0.52**	1.19	**0.62**	1.78
+RGB	(×100)	2.84	3.27	**0.47**	1.17	0.95	**1.26**	0.54	**0.58**	**0.57**	0.44	0.88	0.69	0.74	1.01	**1.10**
+RGB+SS		2.5	**0.91**	0.63	**0.78**	1.25	1.45	**0.41**	0.70	1.06	11.22	**0.67**	0.97	**0.53**	**0.99**	1.72
BARF		9.34	8.46	12.92	20.24	9.85	13.16	17.79	12.65	8.93	11.34	13.35	15.19	13.77	15.46	13.03
+RGB	abs. depth ↓	5.01	6.86	2.30	2.80	4.64	3.77	1.95	4.09	1.85	2.47	2.24	3.18	1.16	2.55	**3.20**
+RGB+SS x100		4.65	3.77	5.79	2.12	5.03	3.37	1.69	3.51	2.36	9.64	1.81	3.03	1.61	1.78	3.58

Table 5: Absolute pose and depth accuracy evaluation on the DTU dataset [12] **before** test-time optimization, using 15% noisy pose initialization. Row blocks represents rotation errors in degrees, translation errors in **cm**s and absolute depth errors in **cm**

To show the effect of sparse views, we repeat the aforementioned ablation with using first five images in each sequence and report results in Table 4 and Figure 5. Note that this is a particularly challenging setup due to small baseline, less number of views and the test image being further away from training set requiring view-extrapolation. It can be seen that while BARF starts failing completely in estimating camera pose for Fortress, Trex and Horns, the performance of presented approach degrades gracefully with sparse view.

Finally, we evaluate effectiveness of proposed approach on large baseline dataset DTU [12]. We use same architecture and training setup as used for LLFF dataset with a couple of exceptions. For every frame, a surrogate view is chosen from a neighbourhood starting from 2 frame till 5 frames on either side of keyframe for direct alignment. We train BARF and our direct alignment assisted version (called +RGB in Table 5) for 100k iterations and report the results. Additionally we experimented with deploying traditional Gaussian scale-space based coarse-to-fine (dubbed +RGB+SS) procedure which is vastly used in variational optic-flow estimation. For this we smooth input images while training NeRFs with a Gaussian kernel with standard deviation $4*(1-\gamma_i)$. We use a linear schedule to increase γ_i from 0 to 1 in first half of training. Quantitative result in table 5 show that the RGB image alignments immensely help both camera pose and structure estimation despite lack of texture and change in exposure between frames of the sequences. alignment with / without scale

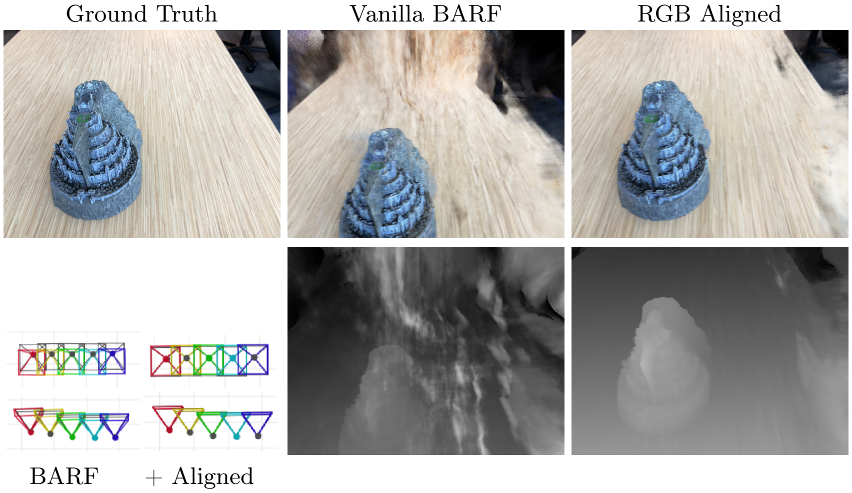


Fig. 5: Qualitative comparison on joint motion and structure estimation on sparse fortress sequence. Top row shows test image follow by view synthesis from BARF without and with direct RGB alignment respectively. Next two row show from left to right the estimated poses and depth-maps for different methods in the same order as row 1. It can be seen that while BARF fails to reconstruct a test-frame due to large pose error as well as outfitted Nerf's with ghosting artefacts, our approach provide significantly better results.

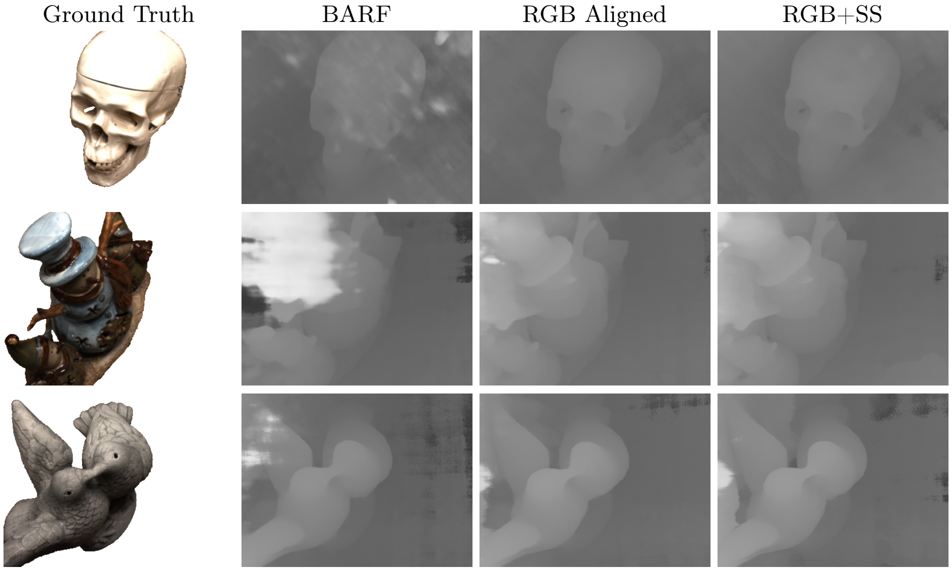


Fig. 6: Visual comparison of depth estimates on DTU sequences.

space coarse to fine outperform baseline 12/14 and 11/14 cases for rotation and translation estimation. We notice that scale space coarse to fine helps in pose estimation for reasonably textured scenes but gets stuck into local minima for pose estimation in very homogeneous scene. Absolute depth errors for proposed work consistently outperform baseline by large margin despite the noisy pose estimation. Figure 6 visualize selected depthmaps where one can clearly see the ghosting artifacts as observed in many LLFF sequences which the direct alignment removes. Additional results, visualizations and implementation details can be found in the supplementary material.

5 Conclusion

We present a simple but effective occlusion aware direct alignment loss and show its effectiveness in learning radiance field. Our paper critically analyse the key advantages of volume rendering and complement them with well studied direct alignment. We show that, despite violation of the Lambertian assumption, direct multi-view alignment of images helps circumvent the over-fitting problem in training NERFs for most real scenes. Further, direct alignment helps in avoiding local minima in joint camera pose estimation and reconstruction, without requiring expansive optic-flow depth or other foundation models to regularize NeRF. Our approach fails to reconstruct severely reflective or semi-transparent surfaces. Like many NeRF and pose estimation methods ours too struggles with large baselines without coarse pose initialization.

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**, 105–112 (2011)
2. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework part 1: The quantity approximated, the warp update rule, and the gradient descent approximation. International Journal of Computer Vision - IJCV (01 2004)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 5460–5469 (2022). `https://doi.org/10.1109/CVPR52688.2022.00539`
4. Bian, W., Wang, Z., Li, K., Bian, J.W., Prisacariu, V.A.: Nope-nerf: Optimising neural radiance field with no pose prior. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4160–4169 (2023)
5. Cai, H., Feng, W., Feng, X., Wang, Y., Zhang, J.: Neural surface reconstruction of dynamic scenes with monocular rgb-d camera. Advances in Neural Information Processing Systems **35**, 967–981 (2022)
6. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
7. Chng, S.F., Garg, R., Saratchandran, H., Lucey, S.: Invertible neural warp for nerf. arXiv preprint arXiv:2407.12354 (2024)
8. 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)
9. 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. pp. 12882–12891 (2022)
10. Gao, Z., Dai, W., Zhang, Y.: Adaptive positional encoding for bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3284–3294 (2023)
11. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis (2021)
12. Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanæs, H.: Large scale multi-view stereopsis evaluation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition. pp. 406–413. IEEE (2014)
13. 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)
14. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: CVPR (2022)
15. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2021)
16. Lin, C.H., Ma, W.C., Torralba, A., Lucey, S.: Barf: Bundle-adjusting neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5741–5751 (2021)

17. 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)
18. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
19. Mur-Artal, R., Montiel, J.M.M., Tardós, J.D.: ORB-SLAM: a versatile and accurate monocular SLAM system. IEEE Transactions on Robotics **31**(5), 1147–1163 (2015). https://doi.org/10.1109/TRO.2015.2463671
20. Newcombe, R.A., Lovegrove, S.J., Davison, A.J.: Dtam: Dense tracking and mapping in real-time. In: 2011 international conference on computer vision. pp. 2320–2327. IEEE (2011)
21. 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)
22. Park, K., Henzler, P., Mildenhall, B., Barron, J.T., Martin-Brualla, R.: Camp: Camera preconditioning for neural radiance fields. ACM Transactions on Graphics (TOG) **42**(6), 1–11 (2023)
23. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
24. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2022)
25. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
26. 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)
27. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: Sparf: Neural radiance fields from sparse and noisy poses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4190–4200 (2023)
28. Wang, C., MacDonald, L.E., Jeni, L.A., Lucey, S.: Flow supervision for deformable nerf. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21128–21137 (2023)
29. Wang, C., MacDonald, L.E., Jeni, L.A., Lucey, S.: Flow supervision for deformable nerf. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 21128–21137 (June 2023)
30. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. arXiv preprint arXiv:2306.05422 (2023)
31. 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)
32. 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)
33. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: CVPR (2021)

34. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in Neural Information Processing Systems (NeurIPS) (2022)
35. Zhan, H., Zheng, J., Xu, Y., Reid, I., Rezatofighi, H.: Activermap: Radiance field for active mapping and planning. arXiv preprint arXiv:2211.12656 (2022)
36. Zhao, F., Yang, W., Zhang, J., Lin, P., Zhang, Y., Yu, J., Xu, L.: Humannerf: Efficiently generated human radiance field from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7743–7753 (2022)