SPARF: Neural Radiance Fields from Sparse and Noisy Poses

Prune Truong[1,2*] Marie-Julie Rakotosaona[2] Fabian Manhardt[2] Federico Tombari[2,3]
[1]ETH Zurich [2]Google [3]Technical University of Munich
prune.truong@vision.ee.ethz.ch {mrakotosaona, fabianmanhardt, tombari}@google.com

Abstract

Neural Radiance Field (NeRF) has recently emerged as a powerful representation to synthesize photorealistic novel views. While showing impressive performance, it relies on the availability of dense input views with highly accurate camera poses, thus limiting its application in real-world scenarios. In this work, we introduce Sparse Pose Adjusting Radiance Field (SPARF), to address the challenge of novel-view synthesis given only few wide-baseline input images (as low as 3) with noisy camera poses. Our approach exploits multi-view geometry constraints in order to jointly learn the NeRF and refine the camera poses. By relying on pixel matches extracted between the input views, our multi-view correspondence objective enforces the optimized scene and camera poses to converge to a global and geometrically accurate solution. Our depth consistency loss further encourages the reconstructed scene to be consistent from any viewpoint. Our approach sets a new state of the art in the sparse-view regime on multiple challenging datasets.

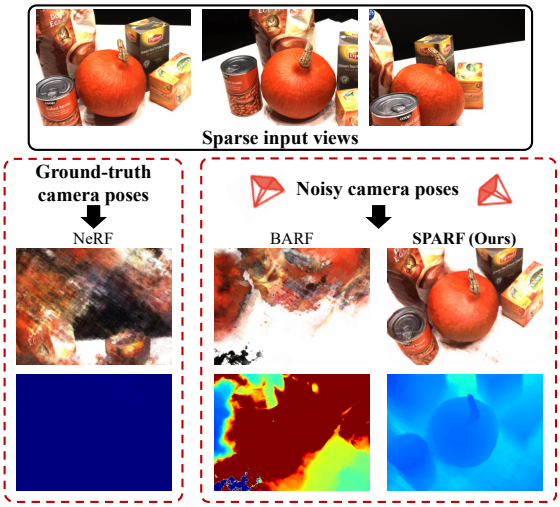

Figure 1. **Novel-view rendering from sparse images**. We show the RGB (second row) and depth (last row) renderings from an unseen viewpoint under sparse settings (3 input views only). Even with ground-truth camera poses, NeRF [30] overfits to the training images, leading to degenerate geometry (almost constant depth). BARF [24], which can successfully handle noisy poses when dense views are available, struggles in the sparse regime. Our approach SPARF instead produces realistic novel-view renderings with accurate geometry, given only 3 input views with noisy poses.

1. Introduction

Novel-view synthesis (NVS) has long been one of the most essential goals in computer vision. It refers to the task of rendering unseen viewpoints of a scene given a particular set of input images. NVS has recently gained tremendous popularity, in part due to the success of Neural Radiance Fields (NeRFs) [30]. NeRF encodes 3D scenes with a multi-layer perceptron (MLP) mapping 3D point locations to color and volume density and uses volume rendering to synthesize images. It has demonstrated remarkable abilities for high-fidelity view synthesis under two conditions: dense input views and highly accurate camera poses.

Both these requirements however severely impede the usability of NeRFs in real-world applications. For instance, in AR/VR or autonomous driving, the input is inevitably much sparser, with only few images of any particular object or region available per scene. In such sparse-view scenario, NeRF rapidly overfits to the input views [11, 22, 32], leading to inconsistent reconstructions at best, and degenerate solutions at worst (Fig. 1 left). Moreover, the de-facto standard to estimate per-scene poses is to use an off-the-shelf Structure-from-Motion approach, such as COLMAP [37]. When provided with many input views, COLMAP can generally estimate accurate camera poses. Its performance nevertheless rapidly degrades when reducing the number of views, or increasing the baseline between the images [55].

Multiple works focus on improving NeRF's performance in the sparse-view setting. One line of research [6, 53] trains conditional neural field models on large-scale datasets. Alternative approaches instead propose various regularization on color and geometry for per-scene training [11, 19, 22, 32, 34]. Despite showing impressive results in the sparse scenario, all these approaches assume *perfect camera poses* as a pre-requisite. Unfortunately, estimating accurate camera poses for few wide-baseline images is challenging [55] and has spawned its own research direction [1, 7, 14–16, 28, 60],

*This work was conducted during an internship at Google.

hence making this assumption unrealistic.

Recently, multiple approaches attempt to reduce the dependency of NeRFs on highly accurate input camera poses. They rely on per-image training signals, such as a photometric [9, 24, 29, 48, 50] or silhouette loss [5, 23, 56], to jointly optimize the NeRF and the poses. However, in the sparse-view scenario where the 3D space is under-constrained, we observe that it is crucial to explicitly *exploit the relation between* the different training images and their underlying scene geometry, to enforce learning a *global and geometrically accurate solution*. This is not the case of previous works [5, 23, 24, 48, 50, 56], which hence fail to register the poses in the sparse regime. As shown in Fig. 1, middle for BARF [24], it leads to poor novel-view synthesis quality.

We propose Sparse Pose Adjusting Radiance Field (SPARF), a joint pose-NeRF training strategy. Our approach produces realistic novel-view renderings given only *few wide-baseline input images* (as low as 3) with *noisy camera poses* (see Fig. 1 right). Crucially, it does not assume any prior on the scene or object shape. We introduce novel constraints derived from multi-view geometry [17] to drive and bound the NeRF-pose optimization. We first infer pixel correspondences relating the input views with a pre-trained matching model [43]. These pixel matches are utilized in our multi-view correspondence objective, which minimizes the re-projection error using the depth rendered by the NeRF and the current pose estimates. Through the explicit connection between the training views, the loss enforces convergence to a global and geometrically accurate pose/scene solution, consistent across all training views. We also propose the depth consistency loss to boost the rendering quality from novel viewpoints. By using the depth rendered from the training views to create pseudo-ground-truth depth for unseen viewing directions, it encourages the reconstructed scene to be consistent *from any viewpoint*. We extensively evaluate and compare our approach on the challenging DTU [20], LLFF [38], and Replica [39] datasets, setting a new state of the art on all three benchmarks.

2. Related Work

We review approaches focusing on few-shot novel view rendering as well as joint pose-NeRF refinement.

Sparse input novel-view rendering: To circumvent the requirement of dense input views, a line of works [6, 8, 25, 41, 46, 53] incorporates prior knowledge by pre-training conditional models of radiance fields on large posed multi-view datasets. Despite showing promising results on sparse input images, their generalization to out-of-distribution novel views remains a challenge. Multiple works [11, 19, 22, 32, 34] follow a different direction, focusing on per-scene training for few-shot novel view rendering. DietNeRF [19] compares CLIP [33] embeddings of rendered and training views. InfoNeRF [22] penalizes the NeRF overfitting to limited input views with a ray entropy regularization. Similarly, Barron *et al*. [4] introduce a distortion loss, which encourages sparsity of the density in each ray. In RegNeRF, Niemeyer *et al*. [32] propose to regularize the geometry and appearance of rendered patches with a depth smoothness and normalizing flow objectives. Recently, a number of works [11, 34, 49, 54] incorporate depth priors to constraint the NeRF optimization. Notably, DS-NeRF [11] improves reconstruction accuracy by including additional sparse depth supervision. Related are also approaches that learn a signed distance function (SDF), aiming for accurate 3D reconstruction in the sparse-view scenario [26, 51]. However, all these works assume perfect poses as a prerequisite. We instead propose a novel training strategy leading to accurate geometry and novel-view renderings in the sparse regime, *even when facing imperfect input poses*.

Joint NeRF and pose refinement: Several approaches attempt to reduce NeRF's reliance on highly accurate input camera poses [9, 24, 29, 48, 50]. BARF [24] and NeRF-- [48] jointly optimize the radiance field and camera parameters of initial noisy poses, relying on the photometric loss as the only training signal. SiNeRF [50] and GARF [9] propose different activation functions, easing the pose optimization. GNeRF [29] introduces a sequential training approach including a rough initial pose network that uses GAN-style training, thereby circumventing the need for initial pose estimates. SCNeRF [21] proposes a geometric loss minimizing the ray intersection re-projection error at previously extracted sparse correspondences to optimize over camera extrinsics and intrinsics. A number of works [5, 23, 52] also combine the photometric objective with a silhouette or mask loss, requiring accurate foreground segmentation, and limiting their applicability to objects. Related are also implicit SLAM systems [2, 40, 59], which progressively optimize over the geometry and camera estimates of an input RGB-D sequence. While previous works assume a dense coverage of the 3D space, Zhang *et al*. propose NeRS [56], which tackles the task of single object reconstruction by deforming a unit sphere over time while refining poses of few input views. However, NeRS is restricted to simple objects with a known shape prior. We instead assume access to only few wide-baseline RGB images with noisy pose estimates, without any prior on the scene or object shape.

3. Preliminaries

We first briefly introduce notation, the basics of NeRF representation, and camera operations.

Camera pose: Let $P_i^{c2w} = \left[R_i^{c2w}|\mathbf{t}_i^{c2w}\right] \in SE(3)$ be the camera-to-world transform of camera i, where $R_i^{c2w} \in SO(3)$ and $\mathbf{t}_i^{c2w} \in \mathbb{R}^3$ are the rotation and translation, respectively. We denote as $K \in \mathbb{R}^{3\times3}$ the intrinsic matrix.

For the rest of the manuscript, we drop the superscript ^{c2w}. As a result, unless otherwise stated, $P = P^{c2w}$ and all 3D quantities are defined in the world coordinate system.

Camera projection: For any vector $\mathbf{x} \in \mathbb{R}^l$ of dimension l, $\bar{\mathbf{x}} \in \mathbb{R}^{l+1}$ corresponds to its homogeneous representation, *i.e.* $\bar{\mathbf{x}} = \left[\mathbf{x}^T, 1\right]$. We additionally define π to be the camera projection operator, which maps a 3D point in the camera coordinate frame $\mathbf{x}^c \in \mathbb{R}^3$ to a pixel coordinate $\mathbf{p} \in \mathbb{R}^2$. Likewise, π^{-1} is defined to be the backprojection operator, which maps a pixel $\mathbf{p}$ and depth z to a 3D point $\mathbf{x}^c$.

$$\pi(\mathbf{x}^c) \cong K\mathbf{x}^c, \qquad \pi^{-1}(\mathbf{p}, z) = zK^{-1}\bar{\mathbf{p}}\,. \tag{1}$$

Scene representation: We adopt the NeRF [30] framework to represent the underlying 3D scene and image formation. A neural radiance field is a continuous function that maps a 3D location $\mathbf{x} \in \mathbb{R}^3$ and a unit-norm ray viewing direction $\mathbf{d} \in \mathbb{S}^2$ to an RGB color $\mathbf{c} \in [0,1]^3$ and volume density $\sigma \in \mathbb{R}^+$. It can be formulated as

$$[\mathbf{c}, \sigma] = F_\theta\left(\gamma_x(\mathbf{x}), \gamma_d(\mathbf{d})\right)\,. \tag{2}$$

Here, F is an MLP with parameters θ, and $\gamma : \mathbb{R}^3 \rightarrow \mathbb{R}^{3+6L}$ is a positional encoding function with L frequency bases.

Volume rendering: Given a camera pose P_i, each pixel coordinate $\mathbf{p} \in \mathbb{R}^2$ determines a ray in the world coordinate system, whose origin is the camera center of projection $\mathbf{o}_i = \mathbf{t}_i$ and whose direction is defined as $\mathbf{d}_{i,\mathrm{p}} = R_i K_i^{-1}\bar{\mathbf{p}}$. We can express a 3D point along the viewing ray associated with $\mathbf{p}$ at depth t as $\mathbf{r}_{i,\mathrm{p}}(t) = \mathbf{o}_i + t\mathbf{d}_{i,\mathrm{p}}$. To render the color $\hat{\mathbf{I}}_{i,\mathrm{p}} \in [0,1]^3$ at pixel $\mathbf{p}$, we sample M discrete depth values t_m along the ray within the near and far plane $[t_n, t_f]$, and query the radiance field F_θ (2) at the underlying 3D points. The corresponding predicted color and volume density values $\{(\mathbf{c}_m, \sigma_m)\}_{m=1}^M$ are then composited as,

$$\hat{\mathbf{I}}_{i,\mathrm{p}} = \hat{I}(\mathbf{p}; \theta, P_i) = \sum_{m=1}^{M} \alpha_m \mathbf{c}_m\,, \tag{3}$$

$$\text{where} \quad \alpha_m = T_m\left(1 - \exp(-\sigma_m \delta_m)\right)\,, \tag{4}$$

$$T_m = \exp\left(-\sum_{m'=1}^{m} \sigma_{m'}\delta_{m'}\right)\,. \tag{5}$$

T_m denotes the accumulated transmittance along the ray from t_n to t_m, and $\delta_m = t_{m+1} - t_m$ is the distance between adjacent samples. Similarly, the approximate depth of the scene viewed from pixel $\mathbf{p}$ is obtained as,

$$\hat{z}_{i,\mathrm{p}} = \hat{z}(\mathbf{p}; \theta, P_i) = \sum_{m=1}^{M} \alpha_m t_m\,. \tag{6}$$

Here, $\hat{I}$ and $\hat{z}$ denote the RGB and depth rendering functions. In practice, NeRF [30] trains two MLPs, a coarse network F_θ^c and a fine network F_θ^f, where the former is used to guide sampling along the ray for the latter, thereby enabling more accurate estimation of (3)-(6).

Photometric loss: Given a dataset of n RGB images $\mathcal{I} = \{I_1, I_2, ..., I_n\}$ of a scene associated with initial noisy poses $\hat{\mathcal{P}} = \left\{\hat{P}_1, \hat{P}_2, ..., \hat{P}_n\right\}$, previous approaches [9, 24, 48, 50] optimize the radiance field function F_θ along with the camera pose estimates $\hat{\mathcal{P}}$ using a photometric loss as follows,

$$\mathcal{L}_{\text{photo}}(\theta, \hat{\mathcal{P}}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{\mathrm{p}} \left\| I_i(\mathbf{p}) - \hat{I}(\mathbf{p}; \theta, \hat{P}_i)\right\|_2^2\,. \tag{7}$$

While this works well with dense views, it fails in the sparse regime. We propose an approach to effectively refine the poses and train the neural field for this challenging scenario.

4. Method

This work addresses the challenge of novel view synthesis based on neural implicit representations, in the sparse-view regime. In particular, we assume access to only *sparse input views* with *noisy camera pose estimates*. The training image collection contains few images (as low as 3) and they present large viewpoint variations.

This leads to two major challenges: (i) given only few input images, the NeRF model [30] instantly overfits to the training views without learning a meaningful 3D geometry, even with perfect input camera poses [19, 22, 32]. As shown in Fig. 1, this leads to degenerate novel view renderings, including for similar train/test viewing directions. The problem becomes amplified when considering noisy input camera poses. (ii) Previous pose-NeRF refinement approaches [5,9,24,48,50] were designed considering a dense coverage of the 3D space, *i.e.* many input views. They apply their training objectives, *e.g.* the photometric loss (7), on each training image *independently*. However, in the sparse-view regime, *i.e.* where the 3D space is under-constrained, such supervision is often too weak for the pose/NeRF system to converge to a *globally consistent* geometric solution. Failure to correctly register the training poses also leads to poor novel view rendering quality (see Fig. 1, 4).

We propose SPARF, a simple, yet effective training strategy to jointly learn the scene representation and refine the initial training poses, tailored for the sparse-view scenario. As the prominent source of inspiration, we draw from well-established principles of multi-view geometry [17], which we adapt to the NeRF framework. In Sec. 4.1, we introduce our *multi-view correspondence objective* as the main driving signal for the joint pose-NeRF training. By relying on pixel correspondences between the training views, the loss enforces convergence to a global and accurate geometric solution consistent across all training views, thereby solving both challenges (i) and (ii). Moreover, in Sec. 4.2 we propose an additional term, *i.e.* the *depth consistency loss*,

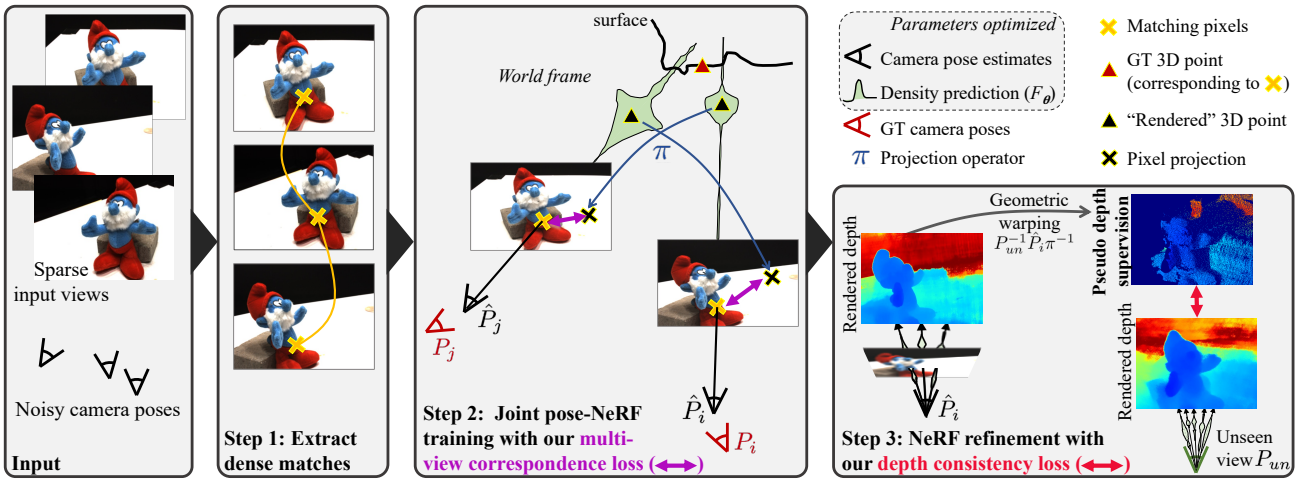

Figure 2. Our approach **SPARF** for joint pose-NeRF training given only *few input images* with *noisy camera pose* estimates. We first rely on a pre-trained dense correspondence network [43] to extract matches between the training views. Our multi-view correspondence loss (Sec. 4.1) minimizes the re-projection error between matches, *i.e.* it enforces each pixel of a particular training view to project to its matching pixel in another training view. We use the rendered NeRF depth (6) and the current pose estimates $\hat{\mathcal{P}}$ to backproject each pixel in 3D space. This constraint hence encourages the learned scene and pose estimates to converge to a global and accurate geometric solution, consistent across all training views. Our depth consistency loss (Sec. 4.2) further uses the rendered depths from the training viewpoints to create pseudo-depth supervision for unseen viewpoints, thereby encouraging the reconstructed scene to be consistent from any direction.

which encourages the learned scene geometry to be consistent across *all viewpoints*, including those for which no RGB supervision is available. In doing so, it boosts novel-view rendering quality, further tackling the overfitting problem (i). We present our final training strategy in Sec. 4.3 and visualize our approach in Fig. 2.

4.1. Multi-View Correspondence Loss

Directly overfitting on the training images leads to a corrupted neural radiance field collapsing towards the provided views, even when assuming perfect camera poses [11, 19, 32]. With noisy input poses, the problem becomes amplified, making it impossible to use the photometric loss (7) as the main signal for the joint pose-NeRF training. We propose a training objective, the multi-view correspondence loss, to enforce learning a *globally consistent 3D solution* over the optimized scene geometry and camera poses.

Multi-view geometry constraint: We draw inspiration from principles of multi-view geometry [17]. We assume that given an image pair (I_i, I_j), we can obtain pairs of matching pixels $\mathbf{p} \in I_i$ and $\mathbf{q} \in I_j$. We then compute estimates of the depth at both pixels $\hat{z}_{i,\text{p}} = \hat{z}(\mathbf{p}; \theta, \hat{P}_i)$ and $\hat{z}_{j,\text{q}} = \hat{z}(\mathbf{q}; \theta, \hat{P}_j)$ according to eq. (6). Principles of multi-view geometry dictate that both pixels must backproject to the same 3D point in the world coordinate system. This is formulated as $\hat{P}_j \pi^{-1}(\mathbf{q}, \hat{z}_{j,\text{q}}) = \hat{P}_i \pi^{-1}(\mathbf{p}, \hat{z}_{i,\text{p}})$. However, when translating this constraint into a training objective, the magnitude of the loss is subject to large variations depending on the scene scale and the initial camera poses, requiring a tedious tuning of the loss weighting.

Training objective: We instead project the 3D points back to image space, therefore minimizing the distance between pixels rather than directly in 3D space. We illustrate this objective in Fig. 2 (steps 1-2). For a randomly sampled training image pair (I_i, I_j), our multi-view correspondence objective is formulated as $\mathcal{L}_{\text{MVCorr}}(\theta, \hat{\mathcal{P}}) = \sum_{\text{p} \in \mathcal{V}} \mathcal{L}_\text{p}$, where

$$\mathcal{L}_\text{p} = w_\text{p} \rho \left(\mathbf{q} - \pi \left(\hat{P}_j^{-1} \hat{P}_i \ \pi^{-1}(\mathbf{p}, \hat{z}(\mathbf{p}; \theta, \hat{P}_i)) \right) \right) . \quad (8)$$

Here ρ denotes the Huber loss function [18] and $w_\text{p} \in [0, 1]$ is the confidence associated with the correspondence $(\mathbf{p}, \mathbf{q})$, which we obtain as detailed below. We additionally define the set $\mathcal{V} = \{\mathbf{p} : w_\text{p} \geq \kappa\}$, where $\kappa = 0.95$. The homogenization operations were omitted for clarity.

Our loss serves two purposes. By connecting the training images through correspondences, our multi-view correspondence objective enforces the learned geometry and camera poses to converge to a solution geometrically consistent across all training images. This is unlike the photometric loss (7) which applies supervision on each training image independently. Moreover, the underlying constraint is only satisfied if the learned 3D points converge to the true reconstructed scene (up to a similarity). As such, the objective (8) provides direct supervision on the rendered depth (6), implicitly enforcing it to be close to the surface.

Correspondence prediction: Any classical [27, 35] or learned [13, 36, 42, 44, 45] matching approach could be used to obtain the matches relating pairs of training views. We rely on a pre-trained dense correspondence regression network, in particular PDC-Net [43]. It predicts a match $\mathbf{q}$ for each pixel $\mathbf{p}$, along with a confidence w_p. We found

the high number of accurate matches to be beneficial for our joint pose-NeRF refinement. Similar conclusions were derived in the context of dense versus sparse depth supervision [11,34]. The dense correspondence map also implicitly imposes a smoothness prior to the rendered depth. In suppl., we present results using a sparse matcher [12, 36] instead.

4.2. Improving Geometry at Unobserved Views

The multi-view correspondence loss favors a global and geometrically accurate solution, consistent across all training images. Nevertheless, the reconstructed scene often still suffers from inconsistencies when seen from novel viewpoints. For those, no RGB supervision is available during training. We propose an additional training objective, the depth consistency loss, which encourages the learned geometry to be consistent from any viewing direction.

Depth consistency loss: The main idea is to use the depth maps rendered from the training viewpoints to create pseudo-depth supervision for novel, unseen, viewpoints (Fig. 2, step 3). We sample a virtual pose P_{un}, in practice obtained as an interpolation between the poses of two close-by training views. For a pixel $\mathbf{p}$ in a sampled training image I_i, $\mathbf{r}_{\text{p}}^{un} = P_{un}^{-1}\hat{P}_i\ \pi^{-1}(\mathbf{p}, \hat{z}(\mathbf{p};\theta,\hat{P}_i))$ is the corresponding 3D point in the coordinate system of the unseen view P_{un}. $\mathbf{y} \in \mathbb{R}^2$ denotes its pixel projection in view P_{un} as $\mathbf{y} = \pi\left(\mathbf{r}_{\text{p}}^{un}\right)$, and z_{y} is its projected depth in P_{un}, *i.e.* $z_{\text{y}} = \left[\mathbf{r}_{\text{p}}^{un}\right]_3$, where $[\cdot]_3$ refers to taking the third coordinate of the vector. We formulate our depth consistency loss as,

$$\mathcal{L}_{\text{DCons}}(\theta) = \sum_{\text{p}} \gamma_{\text{y}}\, \rho\left(z_{\text{y}} - \hat{z}(\mathbf{y};\theta,P_{un})\right)\,. \tag{9}$$

To account for occlusion and out-of-view projections in which (9) is invalid, we have included a visibility mask $\gamma_{\text{y}} \in [0,1]$. We explain its definition in the section below.

Since the pseudo-depth supervision z_{y} is created from renderings, it is subject to errors. For this reason, we find it important to backpropagate through the pseudo-supervision $\mathbf{y}$ and z_{y}. Note that we however do not backpropagate through the pose estimate $\hat{P}_i$. Moreover, as verified experimentally in Tab. 2, our depth consistency objective (9) is complementary to our multi-view correspondence loss (8), the latter enforcing an *accurate* reconstructed geometry while the former ensures it is *consistent* from any viewpoint.

Visibility mask γ_{y}: We first exclude points if their pixel projections $\mathbf{y}$ is outside of the virtual view, by setting the mask as $\gamma_{\text{y}} = 0$. The depth consistency loss is also invalid for pixels that are occluded by the reconstructed scene in the virtual view. To identify these occluded pixels, we follow the strategy of [10]. In particular, we check whether there are occupied regions on the ray between the camera center $\mathbf{o}_{un}$ of P_{un} and the 3D point $\mathbf{r}_{un,\text{y}}(z_{\text{y}})$ at depth z_{y}. We compute how occluded a 3D point is with its transmittance (5) in the unseen view, as $\gamma_{\text{y}} = T_{un,z_{\text{y}}}$. Intuitively, γ_{y} is close to 1 if there is no point with a large density between the camera center $\mathbf{o}_{un}$ and $\mathbf{r}_{un,\text{y}}(z_{\text{y}})$, otherwise it is close to 0. Next, we present our overall training framework.

4.3. Training Framework

Staged training: Our final training objective is formulated as $\mathcal{L}(\theta,\hat{\mathcal{P}}) = \mathcal{L}_{\text{photo}}(\theta,\hat{\mathcal{P}}) + \lambda_{\text{c}}\mathcal{L}_{\text{MVCorr}}(\theta,\hat{\mathcal{P}}) + \lambda_{\text{d}}\mathcal{L}_{\text{DCons}}(\theta)$, where λ_{c} and λ_{d} are predefined weighting factors. The training is split into two stages. In the first part, the pose estimates are trained jointly with the coarse MLP F_θ^c. However, due to the exploration of the pose space at the early stages of training, the learned scene tends to showcase blurry surfaces. As a result, in the second training stage, we freeze the pose estimates and train both the coarse and fine networks F_θ^c and F_θ^f. This ensures that the fine network learns a sharp geometry, benefiting from the pre-trained coarse network. From a practical perspective, our training objectives can be integrated at a low computational cost, since the RGB or depth pixel renderings (3)-(6) can be shared between the three loss terms.

Coarse-to-fine positional encoding: In BARF [24], Lin *et al*. propose to gradually activate the high-frequency components of the positional encodings (2) over the course of the optimization. We refer the reader to [24] for the exact formulation. While originally proposed in the context of pose refinement, we found that this strategy is also extremely beneficial in the sparse-view setting, even when the poses are fixed. It prevents the network from immediately overfitting to the training images, thereby avoiding the worst degenerate geometries. We therefore adopt this coarse-to-fine positional encoding approach as default.

5. Experimental Results

We evaluate the proposed SPARF for novel-view rendering in the few-view setting, in particular when only three input views are available. Results with different numbers of views are provided in suppl. We extensively analyze our method and compare it to earlier approaches, setting a new state of the art on multiple datasets. Further results, visualizations, and implementation details are provided in suppl.

5.1. Experimental Settings

Datasets and metrics: We report results on the **DTU** [20], **LLFF** [38] and **Replica** [39] datasets, for the challenging scenario of 3 input views. DTU is composed of complex object-level scenes with wide-baseline views spanning a half hemisphere. We adhere to the protocol of [53] and evaluate on their reported test split of 15 scenes. Following [32], we additionally evaluate all methods with the object masks applied to the rendered images, to avoid penalizing methods for incorrect background predictions. On LLFF, we follow community standards [30] and use every 8$^{\text{th}}$ image as

	Method	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
I	Full PE [30]	8.41 (9.34)	0.31 (0.63)	0.71 (0.36)	0.87
II	Smaller MLP model	9.03 (10.06)	0.34 (0.65)	0.68 (0.34)	0.79
III	No PE	16.11 (18.40)	0.68 (0.80)	0.37 (0.24)	**0.30**
IV	CF PE [24] (Sec. 4.3)	**16.27** (**18.41**)	**0.69** (**0.81**)	**0.29** (**0.14**)	0.39

Table 1. Comparison of different positional encoding strategies applied to NeRF [30] on DTU (3 views), using ground-truth poses. Results in (·) are computed by masking the background.

the test set. We sample the training views evenly from the remaining images. For the Replica dataset, which depicts videos of room-scale indoor scenes, we subsample every k^{th} frame, from which we randomly select a triplet of consecutive training images. As metrics, we report the average rotation and translation errors for pose registration, and PSNR, SSIM [47] and LPIPS [57] for view synthesis. On the DTU and Replica datasets, we additionally compare the rendered depth with the available ground-truth depth and compute the mean depth absolute error (DE).

Implementation details: We train our approach for 100K iterations, which takes about 10 hours on a single A100 GPU. As pose parametrization, we adopt the continuous 6-vector representation [58] for the rotation and directly optimize the translation vector. We provide all training hyperparameters in the supplementary.

5.2. Method Analysis

We first perform a comprehensive analysis of our approach, on DTU [20], considering only 3 input views.

Impact of positional encoding: Training on sparse input views using the standard NeRF [30] immediately overfits to the provided images, even with perfect poses. We noticed that the overfitting is largely due to the high-frequency positional encodings (PE), and thus experimented with different PE strategies. We present the results in Tab. 1. The standard NeRF (I) with high-frequency PE [30] leads to degenerate geometry and novel view renderings. In (II), using a simplified MLP makes little difference. While training without PE (III) largely prevents overfitting, the coarse-to-fine PE strategy [24] leads to the best result, as shown in (IV).

Ablation study: In Tab. 2, we ablate the key components of our approach, here assuming fixed ground-truth poses and starting from NeRF with coarse-to-fine PE. Adding our multi-view correspondence loss (8) results in drastically better performance on all metrics. Including our depth-consistency module (9) further leads to a small improvement, achieving the best performance overall. Also note that our depth-consistency module (9) works best in collaboration with our multi-view correspondences loss (8) since the latter is needed to learn an accurate geometry.

MV-Corr (8)	DCons (9)	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
✗	✗	16.27 (18.41)	0.69 (0.81)	0.29 (0.14)	0.39
✗	✓	15.86 (18.91)	0.71 (0.82)	0.28 (0.14)	0.20
✓	✗	18.13 (20.81)	0.77 (**0.87**)	0.22 (**0.10**)	0.10
✓	✓	**18.30** (**21.01**)	**0.78** (**0.87**)	**0.21** (**0.10**)	**0.08**

Table 2. Ablation study on the DTU dataset (3 views), with fixed ground-truth poses. Results in (·) are computed by masking the background. All networks use the coarse-to-fine PE [24].

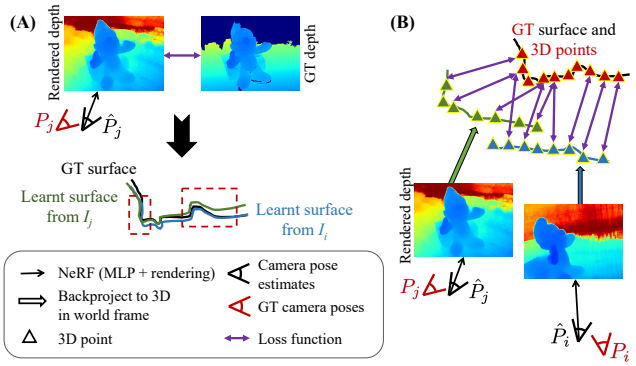

Figure 3. We compare two training objectives using *ground-truth depth* for pose-NeRF training in the sparse-view regime. In (A), a loss comparing for each training image the rendered depth (6) with the ground truth one, can learn locally perfect geometry (as highlighted by the dashed red rectangles). However, the NeRF/poses do not converge to a global solution, because the optimized poses and geometry of the different images are disjoint. Instead, supervising the learned 3D points of each training image to be equal to the ground-truth 3D points in (B) solves this issue, by enforcing the system to converge to a global (unique) geometric solution.

Intuition on pose-NeRF training losses: We first want to build an intuition on what loss might be suitable for joint pose-NeRF training in the *sparse regime*. To do so, we use ground-truth depth or 3D data in two alternative training losses, which we compare here. We illustrate this experiment in Fig. 3 and present results in Tab. 3, top part. As in previous work [24], for each scene of DTU [20], we synthetically perturb the ground-truth camera poses with 15% of additive gaussian noise. In (I), we train with an L1 loss comparing the rendered depth (6) with the ground-truth depth (Fig. 3A). Surprisingly, this loss struggles to refine the poses. Instead, in (II) we minimize the distance between the *learned 3D points* (rendered depth (6) backprojected to world frame) and the ground-truth 3D points, as illustrated in Fig. 3B. This training loss successfully registers the poses, resulting in drastically better novel-view rendering quality. As the main insight from this experiment, we hypothesize that, in the sparse-view regime, it is crucial to enforce an explicit geometric connection between the different training images and their underlying scene geometry. This is not the case in (I), where the depth loss favors per-image locally accurate geometry, but the NeRF/poses can converge to disconnected solutions for each training image.

Comparison of losses for pose-NeRF training: In Tab. 3 bottom part, we then compare our loss (8) to objectives commonly used for joint pose-NeRF training. The photometric loss (7) (III), even associated with a mask/silhouette

	Losses	Rot. ↓	Trans. ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE↓
I	Photo. + L1 **GT** depth	7.3	28.9	13.8 (14.0)	0.54 (0.70)	0.46 (0.27)	0.17
II	Photo. + L1 **GT** 3D points	**0.4**	**1.5**	**18.8 (20.3)**	**0.75 (0.84)**	**0.21 (0.11)**	**0.07**
III	Photo. (7)	10.3	51.5	10.7 (9.8)	0.43 (0.62)	0.59 (0.36)	1.9
IV	Photo. + mask loss [23,56]	13.2	57.7	-	-	-	-
V	MVCorr (8)	1.98	6.6	-	-	-	0.19
VI	**Photo. + MVCorr** ((7)-(8))	**1.85**	**5.5**	**16.0 (17.8)**	**0.68 (0.81)**	**0.28 (0.14)**	**0.13**

Table 3. Comparison of training objectives for joint pose-NeRF refinement on DTU [20] with initial noisy poses (3 views). Rotation errors are in degree and translation errors are multiplied by 100. Results in (·) are computed by masking the background.

loss [5,23,56] in (IV), completely fails to register the poses, thus leading to poor novel-view synthesis performance. This is in line with our hypothesis that it is important to explicitly exploit the *geometric relation* between the training views for successful registration. Moreover, because the 3D space is under-constrained in the sparse-view regime, multiple neighboring poses can lead to similar mask losses. While our multi-view correspondence loss (8) alone (V) already drastically outperforms the photometric loss (III) in terms of pose and learned geometry (depth error), combining the two in (VI) leads to the best performance. This is because, through the correspondences, our approach favors a NeRF/pose solution consistent across all training images. Note that this version neither includes our depth consistency loss (9) (Sec. 4.2) nor our staged training (Sec. 4.3).

5.3. Comparison to SOTA with Noisy Poses

Here, we evaluate SPARF, our joint pose and NeRF training approach. Results with different pose initialization schemes are presented in the supplementary.

Baselines: We compare to BARF [24], the state-of-the-art in pose-NeRF refinement when assuming dense input views. It is representative of a line of approaches [9, 24, 29, 48, 50] using the photometric loss (7) as the main signal. We also experiment with adding the depth regularization loss of [32] or the ray sparsity loss of [4] to BARF, which we denote as RegBARF and DistBARF respectively. We additionally compare to SCNeRF [21], which uses a geometric loss based on correspondences, minimizing the rays' intersection re-projection error. For a fair comparison, we integrate coarse-to-fine PE [24] (Sec. 4.3) in all methods.

Results on DTU: Following [9, 24, 50], for each scene, we synthetically perturb the ground-truth camera poses with 15% of additive gaussian noise. The initial poses thus have

Method	Rot. ↓	Trans. ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
BARF [24]	10.33	51.5	10.71 (9.76)	0.43 (0.62)	0.59 (0.36)	1.90
RegBARF [24,32]	11.20	52.8	10.38 (9.20)	0.45 (0.62)	0.61 (0.38)	2.33
DistBARF [4,24]	11.69	55.7	9.50 (9.15)	0.34 (0.76)	0.67 (0.36)	1.90
SCNeRF [21]	3.44	16.4	12.04 (11.71)	0.45 (0.66)	0.52 (0.30)	0.85
SPARF (Ours)	**1.81**	**5.0**	**17.74 (18.92)**	**0.71 (0.83)**	**0.26 (0.13)**	**0.12**

Table 4. Evaluation on DTU [20] (3 views) with noisy initial poses. Rotation errors are in ° and translation errors are multiplied by 100. Results in (·) are computed by masking the background.

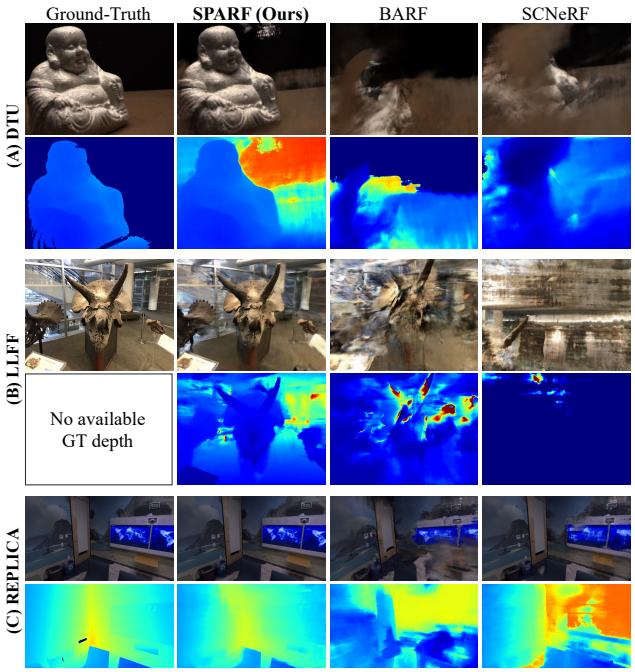

Figure 4. Novel-view rendering (RGB and depth). The input (not shown here) contains 3 images with initial noisy camera poses.

an average rotation and translation error of 15° and 70 respectively. We show initial and optimized poses in Fig. 5. From the results in Tab. 4 and Fig. 4A, we observe that BARF, RegBARF, and DistBARF completely fail to register the poses, leading to poor view-synthesis quality. SCNeRF's geometric loss performs better at registering the poses but the learned scene still suffers from many inconsistencies. This is because SCNeRF's loss [21] does not influence the learned radiance field function, and thus, cannot prevent the NeRF model from overfitting to the sparse input views. Since our multi-view correspondence loss (8) acts on *both* the camera pose estimates and the learned neural field by enforcing them to fit the correspondence constraint, it leads to an accurate reconstructed scene. Our approach SPARF hence significantly outperforms all others both in novel-view rendering quality and pose registration.

Results on LLFF: The LLFF dataset consists of 8 complex forward-facing scenes. Following [24], we initialize all camera poses with the *identity* transformation and present results in Tab. 5. In [24], Lin *et al*. show that BARF almost perfectly registers the camera poses given *dense input views*. However, we show here that it strug-

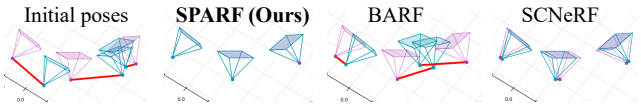

Figure 5. Optimized poses on DTU with 3 input views. We compare the ground-truth poses (in pink) with the optimized ones (in blue). In the first column, the initial noisy poses are in blue.

	Rot. (°) ↓	Trans. (×100) ↓	PSNR ↑	SSIM ↑	LPIPS ↓
BARF [24]	2.04	11.6	17.47	0.48	0.37
RegBARF [24, 32]	1.52	5.0	18.57	0.52	0.36
DistBARF [4, 24]	5.59	26.5	14.69	0.34	0.49
SCNeRF [21]	1.93	11.4	17.10	0.45	0.40
SPARF (Ours)	**0.53**	**2.8**	**19.58**	**0.61**	**0.31**

Table 5. Evaluation on the forward-facing dataset LLFF [38] (3 views) starting from initial identity poses.

	Method	Rot (°) ↓	Trans (×100) ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
G	**SPARF (Ours)**	Fixed GT poses		*26.43*	**0.88**	**0.13**	*0.39*
F	NeRF [30]	Fixed poses obtained		20.99	0.73	0.32	1.33
	DS-NeRF [11]	from COLMAP (run w.		23.52	0.81	0.20	0.99
	SPARF (Ours)	PDC-Net [43] matches)		25.03	*0.84*	*0.15*	0.66
R	BARF [24]	3.35	16.96	20.73	0.72	0.30	0.84
	RegBARF [24, 32]	3.66	20.87	20.00	0.70	0.32	1.00
	DistBARF [4, 24]	2.36	7.73	22.46	0.77	0.23	0.47
	SCNeRF [21]	*0.65*	*4.12*	22.54	0.79	0.24	0.73
	DS-NeRF [11]	1.30	5.04	24.75	0.83	0.20	0.69
	SPARF (Ours)	**0.15**	**0.76**	**26.98**	**0.88**	**0.13**	**0.36**

Table 6. Evaluation on Replica [39] (3 views) with initial poses obtained by COLMAP [37, 43]. The initial rotation and translation errors are 0.39° and 3.01 respectively. In the middle part (F), these initial poses are fixed and used as "pseudo-gt". In the bottom part (R), the poses are refined along with training the NeRF. For comparison, in the top part (G), we use fixed ground-truth poses. The best and second-best results are in red and blue respectively.

gles in the sparse-view setting, thereby severely impacting the accuracy of novel view synthesis. While adding the depth smoothness loss (RegBARF) improves results, our approach SPARF outperforms all previous works. A qualitative comparison is shown in Fig. 4B.

Results on Replica: To demonstrate that our approach is also applicable to non-forward-facing indoor scenes, we evaluate on the Replica dataset in Tab. 6 and Fig. 4C. As pose initialization, we use COLMAP [37] with improved matches, *i.e.* using PDC-Net [43]. The initial pose estimates thus have an average rotation and translation error of respectively 0.39° and 3.01. Comparing the top (G) and middle part (F) of Tab. 6, we show that even such a low initial error impacts the novel-view rendering quality when using fixed poses. In the bottom part (R), our pose-NeRF training strategy leads to the best results, matching the accuracy obtained by our approach with perfect poses (top row, G).

5.4. Comparison to SOTA with Ground-Truth Poses

Finally, we show that our approach brings significant improvement in novel view rendering quality even when considering *fixed ground-truth* poses.

Baselines: We compare to works specifically designed to tackle per-scene few-shot novel view rendering, namely DietNeRF [19], DS-NeRF [11], InfoNeRF [22] and RegNeRF [32], along with the standard NeRF [30] and MipNeRF [3]. For completeness, we also compare against a state-of-the-art conditional model, PixelNeRF [53], trained on DTU [20] and further finetuned per-scene on LLFF [38].

Method	DTU PSNR ↑	DTU SSIM ↑	DTU LPIPS ↓	LLFF PSNR ↑	LLFF SSIM ↑	LLFF LPIPS ↓
PixelNeRF [53]	**19.36** (18.00)	*0.70* (*0.77*)	*0.32* (0.23)	7.93	0.27	0.68
PixelNeRF-ft [53]	-	-	-	16.17	0.44	0.51
MipNeRF [3]	7.64 (8.68)	0.23 (0.57)	0.66 (0.35)	14.62	0.35	0.50
NeRF [30]	8.41 (9.34)	0.31 (0.63)	0.71 (0.36)	13.61	0.28	0.56
DietNeRF [19]	10.01 (11.85)	0.35 (0.63)	0.57 (0.31)	14.94	0.37	0.5
InfoNeRF [22]	11.23 (-)	0.44 (-)	0.54 (-)	-	-	-
RegNeRF [32]	15.33 (*18.89*)	0.62 (0.75)	0.34 (*0.19*)	*19.08*	*0.59*	0.34
DS-NeRF [11]	16.52 (-)	0.54 (-)	0.48 (-)	18.00	0.55	*0.27*
SPARF (Ours)	*18.30* (**21.01**)	**0.78** (**0.87**)	**0.21** (**0.10**)	**20.20**	**0.63**	**0.24**

Table 7. Evaluation on DTU [20] and LLFF [38] (3 views), with fixed ground-truth poses. Results in (·) are computed by masking the background. Results of [3, 19, 32] are taken from [32]. The best and second-best results are in red and blue respectively.

Results: We present results on DTU and LLFF in Tab. 7. Compared to previous per-scene approaches [19, 22, 32] that only apply different regularization to the learned scene, our multi-view correspondence loss (8) provides a strong supervision on the rendered depth, implicitly encouraging it to be close to the true surface. Our depth consistency objective (9) further boosts the performance, by directly enforcing the learned scene to be consistent from any viewpoint. As a result, our approach SPARF performs best compared to all baselines on both datasets and for all metrics. The only exception is PSNR on the whole image compared to conditional model PixelNeRF [53]. This is because DTU has black backgrounds, where a wrong color prediction (like in Fig. 4A for SPARF) has a large impact on the PSNR value. For conditional models which rely on feature projections, it is easier to predict a correct background color. However, most real-world applications are more interested in accurately reconstructing the object of interest than the background. When evaluated only in the object region, our SPARF obtains 3.24dB higher PSNR than PixelNeRF.

6. Conclusion

We propose SPARF, a joint pose-NeRF training strategy capable of producing realistic novel-view renderings given few wide-baseline input images with noisy camera pose estimates. By integrating two novel objectives inspired by multi-view geometry principles, we set a new state of the art on three challenging datasets.

Limitations and future work: Our approach is only applicable to input image collections where each image has covisible regions with at least one other. Moreover, the performance of our method depends on the quality of the matching network. Filtering strategies or per-scene online refinement of the correspondence network thus appear as promising future directions. An interesting direction is also to refine the camera intrinsics and distortion parameters along with the extrinsics. Finally, using voxel grids to encode the radiance field [31] instead of an MLP could lead to faster convergence, and potentially even better results.

References

[1] Wei Wang 0108, Zheng Dang, Yinlin Hu, Pascal Fua, and Mathieu Salzmann. Backpropagation-friendly eigendecomposition. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché Buc, Edward A. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, pages 3156–3164, 2019. 1

[2] Dejan Azinovic, Ricardo Martin-Brualla, Dan B. Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6280–6291. IEEE, 2022. 2

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5835–5844. IEEE, 2021. 8

[4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5460–5469. IEEE, 2022. 2, 7, 8

[5] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2, 3, 7

[6] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 14104–14113. IEEE, 2021. 1, 2

[7] Kefan Chen, Noah Snavely, and Ameesh Makadia. Wide-baseline relative camera pose estimation with directional learning. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3257–3267, 2021. 1

[8] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7907–7916, 2021. 2

[9] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. GARF: gaussian activated radiance fields for high fidelity reconstruction and pose estimation. *CoRR*, abs/2204.05735, 2022. 2, 3, 7

[10] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 6250–6259. IEEE, 2022. 5

[11] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 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. 1, 2, 4, 5, 8

[12] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *2018 IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 224–236, 2018. 5

[13] Mihai Dusmanu, Ignacio Rocco, Tomas Pajdla, Marc Pollefeys, Josef Sivic, Akihiko Torii, and Torsten Sattler. D2-Net: A Trainable CNN for Joint Detection and Description of Local Features. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 4

[14] Sovann En, Alexis Lechervy, and Frédéric Jurie. Rpnet: An end-to-end network for relative camera pose estimation. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part I*, pages 738–745, 2018. 1

[15] Antoine Fond, Luca Del Pero, Nikola Sivacki, and Marco Paladini. End-to-end learning of keypoint detection and matching for relative pose estimation. *CoRR*, abs/2104.01085, 2021. 1

[16] Johan Fredriksson, Viktor Larsson, Carl Olsson, and Fredrik Kahl. Optimal relative pose with unknown correspondences. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1728–1736. IEEE Computer Society, 2016. 1

[17] Richard Hartley and Andrew Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, USA, 2 edition, 2003. 2, 3, 4

[18] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA, 2001. 4

[19] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021. 1, 2, 3, 4, 8

[20] Rasmus Ramsbøl Jensen, Anders Lindbjerg Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 406–413. IEEE Computer Society, 2014. 2, 5, 6, 7, 8

[21] Yoonwoo Jeong, Seokjun Ahn, Christopher B. Choy, Animashree Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5826–5834, 2021. 2, 7, 8

[22] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 12902–12911. IEEE, 2022. 1, 2, 3, 8

[23] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: neural rendering of objects from online image collections. *ACM Trans. Graph.*, 41(4):56:1–56:12, 2022. 2, 7

[24] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 3, 5, 6, 7, 8

[25] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 7814–7823, 2022. 2

[26] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. *ECCV*, 2022. 2

[27] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, Nov. 2004. 4

[28] Iaroslav Melekhov, Juha Ylioinas, Juho Kannala, and Esa Rahtu. Relative camera pose estimation using convolutional neural networks. In *Advanced Concepts for Intelligent Vision Systems - 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18-21, 2017, Proceedings*, pages 675–687, 2017. 1

[29] Quan Meng, Anpei Chen, Haimin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. Gnerf: Gan-based neural radiance field without posed camera. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6331–6341. IEEE, 2021. 2, 7

[30] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 65(1):99–106, 2022. 1, 3, 5, 6, 8

[31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 8

[32] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3, 4, 5, 7, 8

[33] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[34] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. 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. 1, 2, 5

[35] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary R. Bradski. ORB: an efficient alternative to SIFT or SURF. In *IEEE International Conference on Computer Vision, ICCV 2011, Barcelona, Spain, November 6-13, 2011*, pages 2564–2571, 2011. 4

[36] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 4937–4946, 2020. 4, 5

[37] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR 2016, Las Vegas, NV, USA*, pages 4104–4113, 2016. 1, 8

[38] Mohammad Shafiei, Sai Bi, Zhengqin Li, Aidas Liaudanskas, Rodrigo Ortiz Cayon, and Ravi Ramamoorthi. Learning neural transmittance for efficient rendering of reflectance fields. In *32nd British Machine Vision Conference 2021, BMVC 2021, Online, November 22-25, 2021*, page 45. BMVA Press, 2021. 2, 5, 8

[39] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard A. Newcombe. The replica dataset: A digital replica of indoor spaces. *CoRR*, abs/1906.05797, 2019. 2, 5, 8

[40] Edgar Sucar, Shikun Liu, Joseph Ortiz, and Andrew J. Davison. imap: Implicit mapping and positioning in real-time. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 6209–6218. IEEE, 2021. 2

[41] Alex Trevithick and Bo Yang. GRF: learning a general radiance field for 3d representation and rendering. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 15162–15172, 2021. 2

[42] Prune Truong, Stefanos Apostolopoulos, Agata Mosinska, Samuel Stucky, Carlos Ciller, and Sandro De Zanet. Glampoints: Greedily learned accurate match points. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10731–10740, 2019. 4

[43] Prune Truong, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning accurate dense correspondences and when

to trust them. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 5714–5724. Computer Vision Foundation / IEEE, 2021. 2, 4, 8

[44] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 2020. 4

[45] Prune Truong, Martin Danelljan, Radu Timofte, and Luc Van Gool. Pdc-net+: Enhanced probabilistic dense correspondence network. In *Preprint*, 2021. 4

[46] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 2021. 2

[47] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 6

[48] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF−−: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2, 3, 7

[49] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 5590–5599. IEEE, 2021. 2

[50] Yitong Xia, Hao Tang, Radu Timofte, and Luc Van Gool. Sinerf: Sinusoidal neural radiance fields for joint pose estimation and scene reconstruction. *CoRR*, abs/2210.04553, 2022. 2, 3, 7

[51] Wenqi Yang, Guanying Chen, Chaofeng Chen, Zhenfang Chen, and Kwan-Yee K. Wong. Ps-nerf: Neural inverse rendering for multi-view photometric stereo. *CoRR*, abs/2207.11406, 2022. 2

[52] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 2

[53] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 1, 2, 5, 8

[54] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[55] Jason Y. Zhang, Deva Ramanan, and Shubham Tulsiani. Relpose: Predicting probabilistic relative rotation for single objects in the wild. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXI*, volume 13691 of *Lecture Notes in Computer Science*, pages 592–611. Springer, 2022. 1

[56] Jason Y. Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. NeRS: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. In *Conference on Neural Information Processing Systems*, 2021. 2, 7

[57] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 6

[58] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5745–5753. Computer Vision Foundation / IEEE, 2019. 6

[59] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[60] Bingbing Zhuang and Manmohan Chandraker. Fusing the old with the new: Learning relative camera pose with geometry-guided uncertainty. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 32–42, 2021. 1