This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Pose-Free Neural Radiance Fields via Implicit Pose Regularization

Jiahui Zhang[1] Fangneng Zhan[2] Yingchen Yu[1] Kunhao Liu[1]
Rongliang Wu[1] Xiaoqin Zhang[3] Ling Shao[4] Shijian Lu *[1]

[1]Nanyang Technological University [2]Max Planck Institute for Informatics
[3]Wenzhou University [4]UCAS-Terminus AI Lab, UCAS

Abstract

Pose-free neural radiance fields (NeRF) aim to train NeRF with unposed multi-view images and it has achieved very impressive success in recent years. Most existing works share the pipeline of training a coarse pose estimator with rendered images at first, followed by a joint optimization of estimated poses and neural radiance field. However, as the pose estimator is trained with only rendered images, the pose estimation is usually biased or inaccurate for real images due to the domain gap between real images and rendered images, leading to poor robustness for the pose estimation of real images and further local minima in joint optimization. We design IR-NeRF, an innovative pose-free NeRF that introduces implicit pose regularization to refine pose estimator with unposed real images and improve the robustness of the pose estimation for real images. With a collection of 2D images of a specific scene, IR-NeRF constructs a scene codebook that stores scene features and captures the scene-specific pose distribution implicitly as priors. Thus, the robustness of pose estimation can be promoted with the scene priors according to the rationale that a 2D real image can be well reconstructed from the scene codebook only when its estimated pose lies within the pose distribution. Extensive experiments show that IR-NeRF achieves superior novel view synthesis and outperforms the state-of-the-art consistently across multiple synthetic and real datasets.

1. Introduction

Novel view synthesis has recently achieved remarkable progress, largely driven by the development of neural radiance fields (NeRF) [21] that learns 3D scene representations from multi-view 2D images and can generate novel views with superb multi-view consistency. However, most existing works rely heavily on accurate camera poses of the multi-view 2D images which are complicated to collect and not available in many existing image datasets. The camera pose constraint can be mitigated by leveraging structure-from-motion (SfM) [14, 26] that allows estimating camera poses from multi-view 2D images. On the other hand, SfM requires keypoint detection and is prone to errors while handling objects and scenes with low texture or repeated visual patterns. How to train effective NeRF with unposed multi-view images has become one bottleneck for the wide adoption of NeRF in various 3D synthesis tasks.

Several studies attempt pose-free NeRF by training NeRF with unposed multi-view images. One approach is to train NeRF with certain inaccurate camera poses or prior knowledge about camera pose distributions. For example, [33] jointly optimizes NeRF and camera poses to alleviate the requirement for accurate camera poses. BARF [18] exploits bundle adjusting to train NeRF with imperfect camera poses. [3] introduces Gaussian activated radiance field that employs Gaussian activation to avoid falling into local minima. Nevertheless, this approach still requires reasonable camera pose initialization that is often not easy to obtain. Another approach does not require any pose information in training. For example, GNeRF [20] first trains coarse NeRF with randomly initialized camera poses and predicts coarse camera poses, and then jointly refines them with the NeRF training process. However, the pose estimator in GNeRF is trained only with images rendered by the coarse NeRF. The pose prediction for real images is biased or inaccurate due to the domain gap between rendered images and real images, leading to poor robustness of pose estimation for real images and local minima [20] while jointly refining NeRF and camera poses.

We propose IR-NeRF, an innovative pose-free NeRF that introduces Implicit Regularization to promote the robustness of pose estimator for real images. Specifically, given a set of multi-view images of a scene, a scene codebook is first constructed which stores the scene features and encodes scene-specific pose distribution implicitly as priors. With that, a pose-guided view reconstruction scheme is then

[*] Corresponding author, E-mail: shijian.lu@ntu.edu.sg

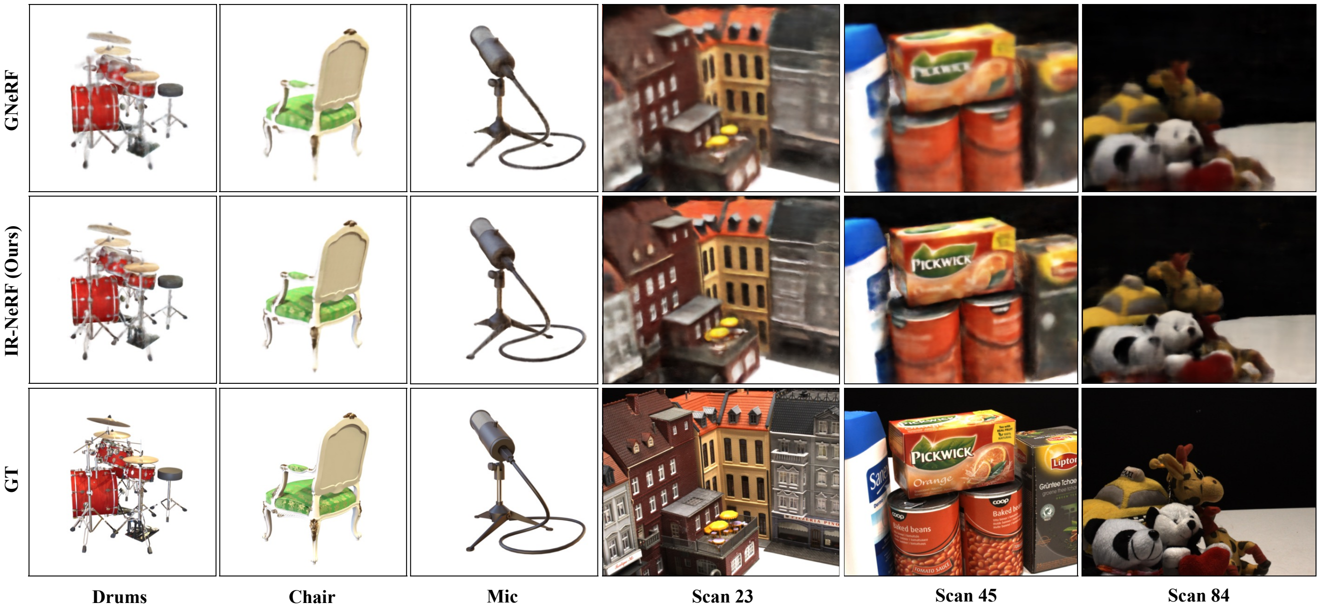


Figure 1: **Examples of novel view synthesis by GNeRF and our IR-NeRF.** The samples are from Synthetic-NeRF [21] and DTU [16]. It can be observed that the IR-NeRF synthesized novel views have less artifacts and finer details than GNeRF.

designed to refine the pose estimator with unposed real images, based on the rationale that a real image can be reconstructed well from the codebook only when its estimated camera pose lies within the scene-specific pose distribution. With the accurate camera poses predicted by the refined pose estimator, IR-NeRF can jointly optimize NeRF and estimated camera poses without getting stuck in local minima, yielding accurate NeRF representations with superior novel view synthesis as illustrated in Fig. 1.

The contributions of this work are threefold. *First*, we propose IR-NeRF, a novel pose-free NeRF that introduces implicit pose regularization that enables effective NeRF training with unposed multi-view images. *Second*, with a set of multi-view 2D images of a scene, we construct a scene codebook that encodes scene features and implicitly captures scene-specific camera pose distribution as priors. *Third*, we design a pose-guided view reconstruction scheme that utilizes the scene priors to refine pose estimator with unposed real images, which allows to promote the robustness of the pose estimator.

2. Related Work

Neural Radiance Fields. NeRF [21] encodes 3D locations and 2D viewing directions into RGB colour and volume density, and it has demonstrated very impressive performance in novel view synthesis. With implicit scene representation and differentiable volume rendering, NeRF has been developing quickly recently with a number of variants and extensions, including generative radiance fields [27, 23, 11], generalizable radiance fields [2, 35], dynamic scene representations [24, 6, 9, 13, 30], fast scene representations [22, 10, 25], neural surface representations [32, 37] and unbounded scene representations [39, 1]. However, most existing work requires accurate camera poses of 2D training images for proper NeRF training, whereas camera pose collection is often complicated and prone to errors which impairs the scalability of NeRF greatly. As a comparison, our proposed IR-NeRF can train effective NeRF with a set of unposed multi-view images.

Pose-Free NeRF Pose-free NeRF has attracted increasing attention recently for training effective NeRF with unposed images. Most existing methods manage to estimate the camera pose of training images, and they can be broadly grouped into two categories depending on whether they involve learning in camera pose estimation. Most non-learning methods [39, 19] exploit conventional techniques such as Structured-from-Motion (SfM) [14, 8, 34, 26]) for camera pose estimation. However, conventional methods often have limited robustness and accuracy. For example, SfM estimates camera poses from key-point correspondences across images which does not work well for scenes with very sparse textures or repeating visual patterns.

Methods in the second category estimate camera poses via learning. One typical approach trains pose-free NeRF with certain roughly initialized camera poses. For example, [33] jointly optimizes initialized camera poses and NeRF model. [18] exploits bundle adjusting for coarse-to-fine camera pose registration and joint optimization of camera poses and NeRF. [3] employs Gaussian activation for pose estimation and NeRF optimization. [17] refines the initialized camera poses via self-calibration. However, this approach often produces degraded NeRF models when the initialized camera pose does not have reasonable accuracy. Another approach [20, 38] learns NeRF with randomly initialized camera poses. For example, GNeRF [20] introduces a pose estimator to directly estimate camera poses from images. However, the pose estimator is trained only with rendered images, leading to inaccurate or biased predictions on real images used in NeRF training due to the domain gap between rendered images and real images. The poor robustness of pose estimator for real images tends to result in local minima in NeRF training. Our IR-NeRF introduces implicit pose regularization to refine pose estimator training with unposed real images, which enhances the robustness of pose estimation for real images, leading to superior pose-free NeRF.

Visual Codebook Standard visual codebook [31] aims to learn a discrete and compressed image representation via vector quantization, and it has been widely explored in the computer vision community. For example, [7] constructs a rich visual codebook to achieve high-resolution image synthesis with transformer. [12] combines the visual codebook with diffusion model [29, 15, 4] for text-to-image generation. [36] proposes multiple improvements over vanilla VQ-GAN [7] for improving vector quantized image modeling tasks. In IR-NeRF, we design a novel scene codebook construction technique that adopts linear combination instead of vector quantization for implicit pose regularization. To the best of our knowledge, IR-NeRF is the first work that adapts visual codebook for pose-free NeRF optimization.

3. Preliminary

Camera Pose Estimation Camera pose distribution is determined by camera poses of multi-view images of a specific 3D scene [20]. Specifically, camera positions are distributed on the surface of partial sphere which is determined by the radius of sphere and camera elevation range and camera azimuth range. Camera rotation depends on camera position, camera lookat points and camera lookup vector. As greater viewpoint uncertainty tends to lead to local minima while jointly optimizing camera poses and NeRF model [20], it is critical to ensure that estimated camera poses are located within scene-specific camera pose distribution.

Neural Radiance Field NeRF [21] is proposed to represent a 3D scene as a 5D function that is parameterized with MLP. It takes a 3D location $x \in \mathbb{R}^3$ and a 2D viewing direction $d \in \mathbb{S}^2$ as input and generates a RGB color $[r, g, b]$ and volume density σ for this location. This process can be formulated by $F_\Theta : (x, d) \to ([r, g, b], \sigma)$, where F and Θ denote MLP network and its parameters, respectively. Volume rendering is then adopted to render 2D images from NeRF scene representation by the accumulation of colors and densities at camera rays. To ensure the differentiability of the volume rendering, numerical quadrature is adopted to approximate the continuous integral by stratified sampling from depth bounds. Additionally, NeRF models are optimized by a photometric loss between the real and corresponding rendered pixel colors, which is formulated by sum of squared differences.

4. Proposed Method

4.1. Overall Framework

With initial camera poses $\Phi = \{\phi_i, i \in [1, T]\}$ randomly sampled from a predefined pose distribution following the settings in GNeRF [20], the proposed IR-NeRF first learns a coarse NeRF with an adversarial loss, and then utilizes the trained NeRF to render images with Φ. A pose estimator P is trained in two steps to predict camera poses. First, it is trained by regressing initialized camera poses with rendered images as in [20]. Second, IR-NeRF introduces an implicit pose regularization to refine the pose estimator with unposed real images. The implicit pose regularization for real images leads to robust pose estimation, as pose estimator trained with only rendered images is inaccurate for real images due to the domain gap between real images and rendered images.

As shown in Fig.2, the key components in the implicit pose regularization are scene codebook construction and pose-guided view reconstruction with view consistency loss. Specifically, a scene codebook C with scene features and scene-specific pose distribution is first learned by the reconstruction of the unposed real images in training dataset $\mathcal{I}$ used in NeRF training. Then, given real image I, pose-guided view reconstruction exploits pose estimator P to predict camera pose ϕ' of image I, and further utilizes ϕ' to guide linear combination of feature embeddings in the learned scene codebook to reconstruct the corresponding image I'. As the trained C and G ensure that an image can be reconstructed well only when its estimated camera pose lies within accurate pose distribution, implicit pose regularization can be achieved with a view consistency loss $\mathcal{L}_c$ between I' and $\hat{I}$. We also jointly refine the learned coarse NeRF and predicted camera poses. With the implicit pose regularization, the joint refinement can effectively avoid falling into local minima. Details of the designed scene codebook construction and pose-guided view reconstruction will be discussed in the ensuing section 4.2 and section 4.3, respectively.

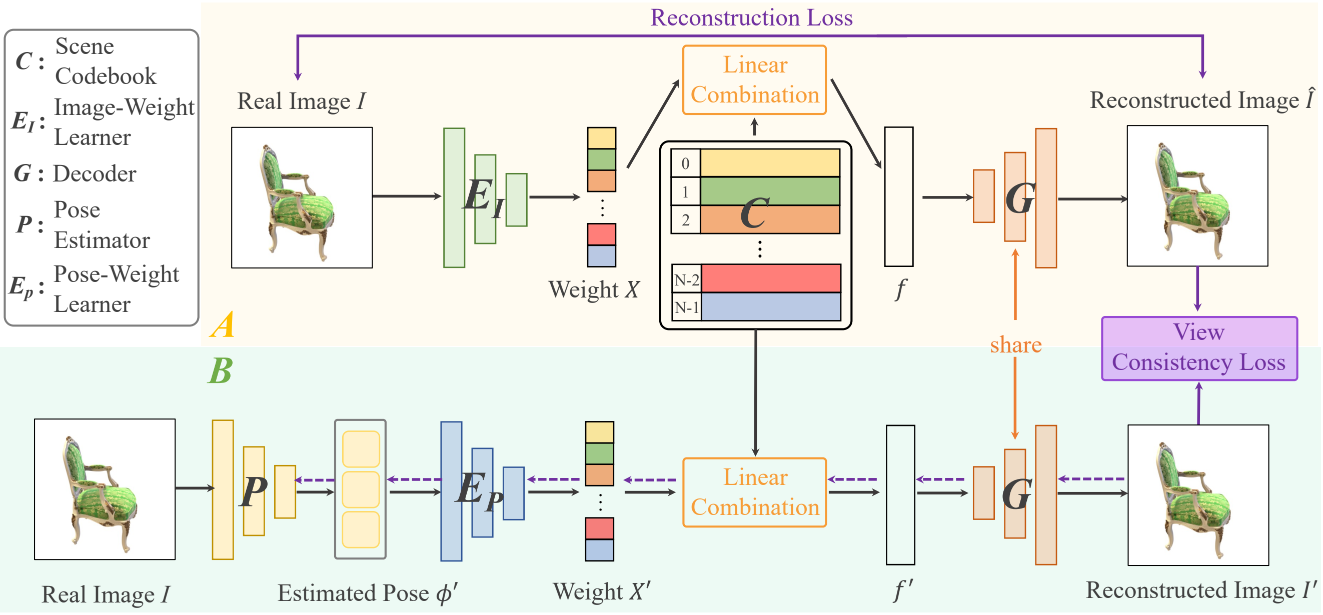


Figure 2: **Overview of the proposed implicit pose regularization.** Part 'A' in yellow and part 'B' in green represent scene codebook construction and pose-guided view reconstruction, respectively. Leveraging image-weight learner E_I, scene codebook C and decoder G, the real image I can be reconstructed from a feature embedding f which is constructed by linear combination of feature embeddings in the codebook. E_I, C and G are trained simultaneously via the image reconstruction process. Pose estimator P predicts the camera pose ϕ' of the real image I in training dataset. With the learned C and G, image I' corresponding to ϕ' is reconstructed by linear combination of learned feature embeddings in C, where combination weights X' are derived from ϕ' through a pose-weight learner E_P. The view consistency loss between I' and $\hat{I}$ regularizes the pose estimation. The purple dashed line highlights the regularization process for pose estimation.

4.2. Scene Codebook Construction

The scene codebook construction allows to learn scene-specific pose distribution implicitly as priors which lays a base for the subsequent pose-guided view reconstruction. Instead of naively encoding input images to latent representations which fails to capture overall pose distribution, we design a novel scene codebook construction scheme with a linear combination which can serve as implicit distribution prior to achieve robust pose estimation.

As shown in Fig. 2, the scene codebook construction consists of an image-weight learner E_I, a scene codebook $C = \{c_n\}_{n=1}^N \in \mathbb{R}^{N\times D}$ and a decoder G. The scene codebook is learned via the reconstruction of unposed real images. The image-weight learner E_I is utilized to yield a collection of combination weights $X = \{x_n\}_{n=1}^N \in \mathbb{R}^N$ based on the real image I:

$$X = Softmax(E_I(I)), \quad x_n = \frac{e^{l_n}}{\sum_{j=1}^N e^{l_j}}, \tag{1}$$

where $Softmax(\cdot)$ denotes the softmax function and l_n represents the Logits output of E_I (before Softmax). The feature embedding f of the real image I is then constructed by linear combination of feature embeddings in codebook, which can be formulated as follows:

$$f = \sum_{n=1}^{N} c_n x_n \tag{2}$$

With the feature embedding f, the real image I can be reconstructed via the decoder G by:

$$I \approx \hat{I} = G(f), \tag{3}$$

where $\hat{I}$ denotes the reconstructed image. With an image reconstruction loss $\mathcal{L}_{rec}$, the scene codebook can be learned with the image-weight learner E_I and the decoder G:

$$\mathcal{L}_{rec}(E_I, C, G) = \|I - \hat{I}\|^2, \tag{4}$$

To reduce the difficulty of joint training of E_I, C and G and improve the training stability, we employ the pre-trained VGG19 [28] to initialize the scene codebook C by encoding a set of real images $[I_0, I_1, ..., I_T]$, where T represents the number of real images. This process can be formulated as follows:

$$C_{ini} = VGG([I_0, I_1, ..., I_T]), \tag{5}$$

where $VGG(\cdot)$ represents the VGG19 network, C_{ini} denotes the initialized scene codebook, which will be further optimized by image reconstruction loss $\mathcal{L}_{rec}$.

4.3. Pose-Guided View Reconstruction

With the learned scene codebook C and decoder G, it can be guaranteed that only images with camera poses within scene-specific pose distribution can be well-reconstructed. Under this rationale, we design pose-guided view reconstruction with view consistency loss to refine pose estimation with unposed real images. Based on the estimated camera pose ϕ' of real image I, the image I' corresponding to ϕ' is constructed by linear combination of the learned feature embeddings in scene codebook. Specifically, a pose-weight learner E_P is first utilized to produce a set of combination weights $X' = \{x'_n\}_{n=1}^{N} \in \mathbb{R}^N$ based on the estimated camera pose ϕ', which can be formulated as follows:

$$X' = Softmax(E_P(\phi')), \quad x'_n = \frac{e^{l'_n}}{\sum_{j=1}^{N} e^{l'_j}}, \tag{6}$$

where l'_n represents the Logits output of E_P (before Softmax). The construction of feature embedding f' corresponding to ϕ' can then be represented as $f' = \sum_{n=1}^{N} c_n x'_n$, where c_n and x'_n denote the n-th feature embedding in scene codebook and n-th combination weight, respectively. Finally, the corresponding image I' can be reconstructed via the frozen decoder G which focuses on decoding the feature embedding generated by linear combination of feature embeddings in scene codebook.

Leveraging the image $\hat{I}$ reconstructed from the shared decoder G as pseudo ground truth, a view consistency loss $\mathcal{L}_c$ between the reconstructed image I' and the pseudo ground truth can be formulated as below:

$$\mathcal{L}_c(P, E_P) = \frac{1}{i} \sum_{i=1}^{S} \left\| I'_i - \hat{I}_i \right\|_2^2. \tag{7}$$

If the camera pose ϕ' estimated by P deviates from scene-specific pose distribution, the corresponding view I' reconstructed by the learned C and G will not be aligned with the pseudo ground truth $\hat{I}$. Thus, the robustness of pose estimation can be promoted as the out-of-distribution pose estimation are suppressed.

4.4. Training Process

The training process of the proposed IR-NeRF includes coarse NeRF learning, camera pose estimation, and joint refinement of NeRF and predicted camera poses. For the coarse NeRF training, we introduce an adversarial loss to train a coarse NeRF F with randomly initialized poses Φ due to lack of known camera poses. The adversarial loss $\mathcal{L}_{adv}$ can be defined as follows:

$$\begin{aligned}\mathcal{L}_{adv}(F, D) = \ & \mathbb{E}_{I \sim P_{data}}[log(D(I))] \\ & + \mathbb{E}_{F(\Phi) \sim P_g}[log(1 - D(F(\Phi)))],\end{aligned} \tag{8}$$

where D denotes the discriminator, P_{data} and P_g represent the distribution of images generated by NeRF and real images in training dataset, respectively.

For the camera pose estimation, we first employ MSE loss to optimize a coarse pose estimator P with images rendered by the trained coarse NeRF as in GNeRF [20]. The pose estimator is then refined for real images via implicit pose regularization. Specifically, the scene codebook construction is performed with unposed real images under the supervision of the image reconstruction loss $\mathcal{L}_{rec}$. With the learned scene codebook and decoder, the pose estimator can be optimized to predict the camera poses of real images driven by the view consistency loss $\mathcal{L}_c$. With coarse NeRF and predicted camera poses, we also employ a photometric loss for joint optimization of the NeRF and camera poses. Specifically, we leverage the hybrid and iterative optimization scheme [20] for end-to-end training of the proposed IR-NeRF, where the pose estimation and joint optimization are interleaved in the training. Note that NeRF is frozen during camera pose estimation but is trainable during joint refinement.

5. Experiment

5.1. Datasets and Implementation Details

Datasets Following GNeRF [20], we conduct experiments on synthetic and real-world scenes with the same split of training and evaluation sets. For synthetic scenes, we use NeRF-Synthetic dataset [21] which consists of object-centric scenes with complex geometry. For each scene, we train with 100 multi-view training images which are resized to 400 by 400 pixels. The evaluation is conducted on eight images that are randomly selected from the test set. For real-world scenes, we employ six representative scenes in the DTU dataset [16]. We randomly split the 49 images of each scene into training and test sets, where the training set includes 43 images of resolution 500×400 and the test set consists of the remaining 6 images.

Implementation Details For predefined camera pose distribution, we follow the settings in GNeRF [20]. Specifically, the range of azimuth, elevation, sphere radius and camera lookat point are set at $[0°, 360°]$, $[0°, 90°]$, 4.0 and $(0, 0, 0)$, and $[0°, 150°]$, $[0°, 80°]$, 4.0 and $\mathcal{N}(0, 0.01^2)$ for both synthetic and real-world datasets, respectively. For camera poses, camera position and camera rotation are represented by a 3D embedding in Euclidean space and a continuous 6D embedding [41], respectively. The camera pose

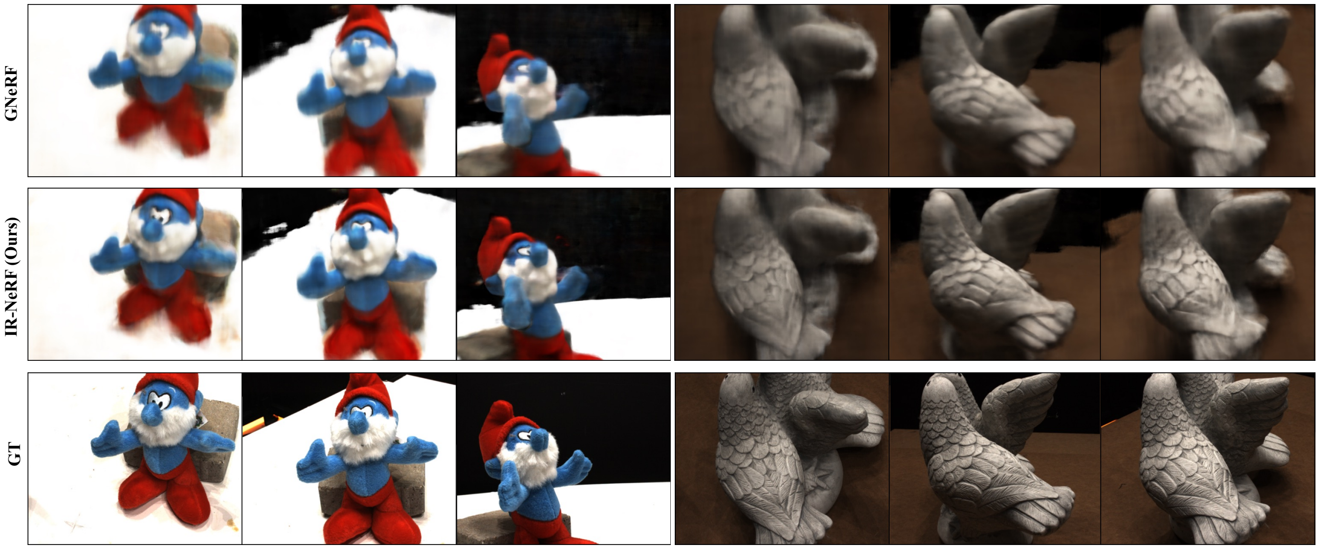


Figure 3: **Qualitative comparisons of IR-NeRF with GNeRF in novel view synthesis:** The comparisons are conducted over different views of scenes 'scan82' and 'scan109' in DTU, where 'GT' denotes the ground-truth image. It is clear that IR-NeRF synthesizes high-fidelity images with less artifacts and finer details compared with GNeRF. **Zoom in for best view.**

embedding can be recovered to a transformation matrix by a Gram-Schmidt-like process [41]. For the architecture of IR-NeRF, the image-weight learner E_I, pose-weight learner E_P and decoder G are CNN-based, MLP-based and CNN-based networks, respectively. For the pose estimator P, we leverage the vision transformer [5] where the output of last layer is modified to an estimated camera pose. The number of feature embeddings in the scene codebook is set to 1024 and the dimension of each feature embedding is set to 512. The dimensions of obtained weights X and X' are the same as the number of feature embeddings in the scene codebook. In term of NeRF in IR-NeRF, we adopt the hierarchical volume sampling strategy [21] to simultaneously optimize 'coarse' and 'fine' networks to represent scenes. The MLPs in 'coarse' and 'fine' networks are shared and the dimension of MLPs is set to 360 [20]. We sample 64 locations along each camera ray in both stratified sampling and inverse transform sampling [21]. The Adam optimizer is adopted to train our IR-NeRF and the mini-batch size is set to 12 for both synthetic and real scenes. We use the Pytorch framework in implementation and employ one NVIDIA RTX 3090ti GPU for both training and inference.

5.2. Comparisons with the State-of-the-Art

Novel View Synthesis. We compare IR-NeRF with the most related work GNeRF [20] over different synthetic and real scenes. We did not compare with NeRF−− [33], BARF [18], SCNeRF [17] and GARF [3] as the four methods require reasonable camera pose initialization and are not applicable to random camera pose initialization. As there is no available pretrained models, we train GNeRF based on its official codes and all methods (including IR-NeRF) are trained with the same training dataset and training setting in experiments. Table 1 shows experimental results over

Scenes	PSNR↑		SSIM↑		LPIPS↓	
	GNeRF	**Ours**	GNeRF	**Ours**	GNeRF	**Ours**
Chair	31.30	**32.87**	0.94	**0.96**	0.08	**0.07**
Drums	24.30	**25.98**	0.90	**0.91**	0.13	**0.11**
Hotdog	32.00	**33.52**	0.96	**0.97**	0.07	**0.06**
Lego	28.52	**30.07**	0.91	**0.93**	0.09	**0.07**
Mic	31.07	**32.33**	0.96	**0.97**	0.06	**0.04**
Ship	26.51	**27.96**	0.85	**0.87**	0.21	**0.18**
Scan23	17.89	**19.96**	0.55	**0.59**	0.54	**0.45**
Scan45	18.06	**20.19**	0.68	**0.73**	0.48	**0.41**
Scan58	21.83	**24.02**	0.62	**0.67**	0.67	**0.55**
Scan82	19.91	**21.55**	0.77	**0.85**	0.33	**0.27**
Scan103	22.67	**24.72**	0.74	**0.82**	0.44	**0.37**
Scan109	22.88	**25.36**	0.71	**0.75**	0.54	**0.44**

Table 1: **Quantitative comparisons of novel view synthesis** on the dataset Synthetic-NeRF and DTU. The proposed IR-NeRF outperforms the state-of-the-art GNeRF consistently in PSNR, SSIM and LPIPS under different synthetic and real scenes. All methods are trained with the same training data and batch size.

Scenes	GNeRF [20]		IR-NeRF	
	Rot(°)↓	Trans↓	Rot(°)↓	Trans↓
Chair	0.363	0.018	**0.251**	**0.013**
Drums	0.204	0.010	**0.185**	**0.008**
Hotdog	2.349	0.122	**1.932**	**0.098**
Lego	0.430	0.023	**0.371**	**0.015**
Mic	1.865	0.031	**1.598**	**0.019**
Ship	3.721	0.176	**3.253**	**0.125**

Table 2: **Quantitative comparisons of the accuracy of camera pose estimation** (on Synthetic-NeRF): Rot and Trans represent mean camera rotation differences and mean camera translation differences, respectively. IR-NeRF outperforms the state-of-the-art GNeRF consistently in Rot and Trans in all studied scenes.

the same test images as described in section 5.1. We can observe that IR-NeRF outperforms the state-of-the-art GNeRF consistently in PSNR, SSIM and LPIPS across all synthetic and real scenes. The superior performance is largely attributed to our proposed implicit pose regularization that allows to refine pose estimator with unposed real images which further improves the robustness of pose estimation for real images. The quantitative experimental results are well aligned with the qualitative results in Figs. 3 where IR-NeRF produces superior multi-view images with less artifacts and finer details.

Camera Pose Estimation. We also compare the accuracy of the estimated camera poses of real images as used in NeRF training. The evaluation is performed over the dataset Synthetic-NeRF. For the evaluation metric, we adopt mean camera rotation difference (Rot) and mean translation difference (Trans) that are computed with the toolbox [40] on the training set. As Table 2 shows, IR-NeRF outperforms GNeRF clearly and consistently across all evaluated scenes. The superior estimation accuracy is largely attributed to our designed implicit pose regularization. The robustness of camera pose estimation for real images can be improved with this pose regularization, further leading to superior joint refinement of camera poses and NeRF without falling into local minima.

5.3. Ablation Studies

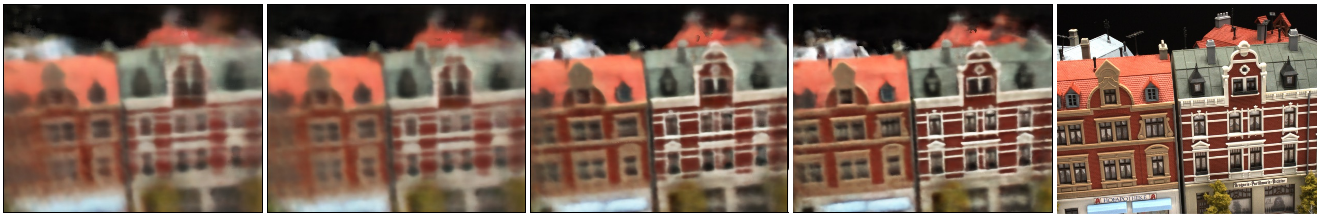
Effect of Implicit Pose Regularization . We examine the contribution of our proposed implicit pose regularization. As Table 3 shows, we train the model *IR-NeRF (w/o REG)* by removing the implicit pose regularization from the *IR-NeRF*. The IR-NeRF (w/o REG) does not involve the two key components so it can be regarded as a baseline that trains the pose estimator in the similar way as GNeRF. It can be seen that *IR-NeRF (w/o REG)* degrades PSNR, SSIM and LPIPS significantly as compared with *IR-NeRF*, indicating that the proposed implicit pose regularization can effectively improve the robustness of pose estimation and further achieve superior novel view synthesis for IR-NeRF. The effectiveness of the proposed implicit pose regularization can be observed in Fig. 4 as well where the model *IR-NeRF* can produce clearer visual results than the model *IR-NeRF (w/o REG)*.

Models	**Evaluation Metrics**		
	PSNR ↑	SSIM ↑	LPIPS ↓
IR-NeRF (w/o REG)	17.05	0.53	0.65
IR-NeRF (w/o SCC)	18.23	0.55	0.54
IR-NeRF (w/o VCL)	17.38	0.54	0.64
IR-NeRF	**19.88**	**0.59**	**0.47**

Table 3: **Ablation studies of the proposed IR-NeRF** on the scene 'scan23' of DTU. *IR-NeRF (w/o REG)* removes the implicit pose regularization (REG) from IR-NeRF, which is equivalent to baseline. *IR-NeRF (w/o SCC)* removes the scene codebook construction (SCC), where input image is naively encoded to latent features. *IR-NeRF (w/o VCL)* removes the view consistency loss (VCL) in the pose-guided view reconstruction, where a reconstruction loss is introduced between pose-guided reconstructed image and real image. All models are trained with same training settings.

Effect of Scene Codebook Construction. To examine the effectiveness of the designed scene codebook construction, we study how it affects view synthesis in PSNR, SSIM and LPIPS. As shown in Table 3, we train *IR-NeRF (w/o SCC)* that removes the designed scene codebook construction from the complete model *IR-NeRF*. Quantitative experiments show that *IR-NeRF* performs clearly better than the *IR-NeRF (w/o SCC)* in novel view synthesis, demonstrating the effectiveness of pre-learning a scene codebook for subsequent pose-guided view reconstruction. The experimental results are also well aligned with qualitative experimental results in Fig. 4 where *IR-NeRF* with the designed scene codebook construction synthesizes novel views with less artifacts and finer details compared with the synthesis by *IR-NeRF (w/o SCC)*.

Effect of View Consistency Loss. We further examine the view consistency loss in the pose-guided view reconstruction by comparing the model *IR-NeRF (w/o VCL)* and *IR-NeRF*. As Table 3 shows, adopting view consistency loss improves PSNR, SSIM and LPIPS consistently, indicating the effectiveness of our designed view consistency loss in pose-guided view reconstruction.



IR-NeRF (w/o REG) IR-NeRF (w/o VCL) IR-NeRF (w/o SCC) IR-NeRF GT

Figure 4: **Qualitative ablation studies of the proposed IR-NeRF:** IR-NeRF and its variants (including IR-NeRF (w/o REG), IR-NeRF (w/o VCL), and IR-NeRF (w/o SCC) that remove the proposed implicit pose regularization, view consistency loss, and scene codebook construction, respectively) are trained on the scan15 of DTU dataset. **Zoom in for best view.**

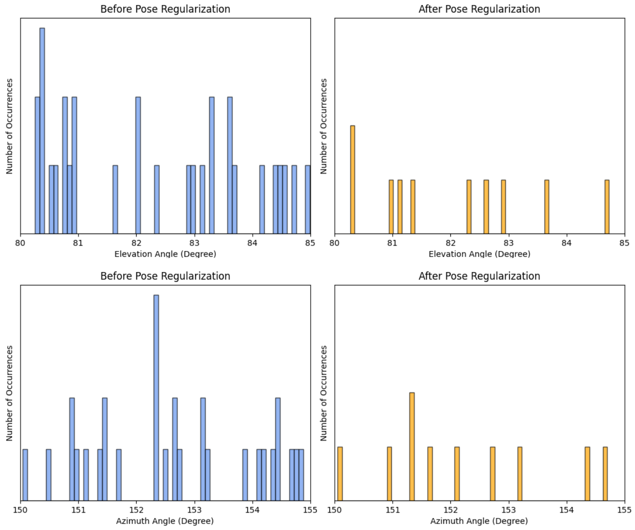


Figure 5: **Visualization of estimated out-of-distribution camera poses**: Camera poses are estimated from real images of the DTU dataset. We focus on the elevation and azimuth angles of the estimated camera poses, which are two key parameters in camera pose distribution. The range ([150°, 155°]) of the elevation angle and the range ([80°, 85°]) of the azimuth angle are out of the camera pose distribution for the DTU scenes. The y-axis represents the number of occurrences. It can be observed that after applying the proposed implicit pose regularization, the estimated out-of-distribution camera poses (in azimuth angle and elevation angle) are significantly reduced.

The quantitative results are well aligned with the qualitative experiments in Fig. 4 as well.

5.4. Visualization

We visualize out-of-distribution camera poses estimated before and after the proposed implicit pose regularization by using histograms. As Fig. 5 shows, much less out-of-distribution camera poses are predicted after applying the proposed implicit pose regularization. This shows that the proposed implicit pose regularization can effectively refine the pose estimation and improve the robustness of pose estimation for real images, which greatly helps to mitigate local minima in the subsequent joint refinement of camera poses and NeRF.

6. Limitation

Although the proposed IR-NeRF achieves superior NeRF training by implicit pose regularization as compared with state-of-the-art GNeRF, it still has one major limitation. Specifically, the training process of IR-NeRF includes coarse NeRF learning, coarse camera pose estimation, and joint refinement of camera poses and NeRF, which requires a long training time. Moving forward, we will focus on pose-free NeRF training at much higher speed. The training speed could potentially be improved by introducing more efficient representation, such as triplane and tensor decomposition.

7. Conclusion

This paper presents IR-NeRF, a pose-free NeRF with implicit pose regularization that promotes the robustness of pose estimation for real images, thus preventing the joint refinement of NeRF and predicted camera poses from falling into local minima. Given a set of multi-view images of a scene, we construct a scene codebook to encode scene features and capture scene-specific pose distribution as priors. In addition, we design pose-guided view reconstruction with view consistency loss which refines pose estimation for real images with the scene priors based on the rationale that a real image can be reconstructed well from the learned scene codebook only when its estimated camera pose lies within the scene-specific pose distribution. Extensive experiments over synthetic and real scenes demonstrate the superiority of IR-NeRF.

8. Acknowledgements

This project is funded by the Ministry of Education Singapore, under the Tier-2 project scheme with a project number MOE-T2EP20220-0003. Fangneng Zhan is supported by the ERC Consolidator Grant 4DReply (770784).

References

[1] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[2] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 2

[3] Shin-Fang Chng, Sameera Ramasinghe, Jamie Sherrah, and Simon Lucey. Garf: Gaussian activated radiance fields for high fidelity reconstruction and pose estimation. *arXiv e-prints*, pages arXiv–2204, 2022. 1, 3, 6

[4] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 6

[6] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14324–14334, 2021. 2

[7] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 3

[8] Olivier Faugeras and Quang-Tuan Luong. *The geometry of multiple images: the laws that govern the formation of multiple images of a scene and some of their applications*. MIT press, 2001. 2

[9] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2

[10] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 2

[11] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 2

[12] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10696–10706, 2022. 3

[13] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021. 2

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1, 2

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 3

[16] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 2, 5

[17] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5846–5854, 2021. 3, 6

[18] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 1, 2, 6

[19] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[20] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6351–6361, 2021. 1, 3, 5, 6, 7

[21] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 5, 6

[22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2

[23] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2

[24] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[25] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 2

[26] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2

[27] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2

[28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[29] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[30] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2

[31] Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017. 3

[32] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 2

[33] 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. 1, 2, 6

[34] Changchang Wu. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*, pages 127–134. IEEE, 2013. 2

[35] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022. 2

[36] Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, and Yonghui Wu. Vector-quantized image modeling with improved vqgan. *arXiv preprint arXiv:2110.04627*, 2021. 3

[37] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: Neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 2

[38] Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. Vmrf: View matching neural radiance fields. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6579–6587, 2022. 3

[39] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[40] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. 7

[41] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019. 5, 6