

VMRF: View Matching Neural Radiance Fields

Jiahui Zhang jiahui003@e.ntu.edu.sg Nanyang Technological University Singapore	Fangneng Zhan fzhan@mpi-inf.mpg.de Max Planck Institute for Informatics Saarbrücken, Saarland, Germany	Rongliang Wu ronglian001@e.ntu.edu.sg Nanyang Technological University Singapore
Yingchen Yu yingchen001@e.ntu.edu.sg Nanyang Technological University & Alibaba Group Singapore	Wenqing Zhang wenqingzhang@bytedance.com ByteDance Singapore	Bai Song songbai.site@gmail.com ByteDance Singapore
Xiaoqin Zhang zhangxiaoqinnan@gmail.com Wenzhou University China		Shijian Lu* shijian.lu@ntu.edu.sg Nanyang Technological University Singapore

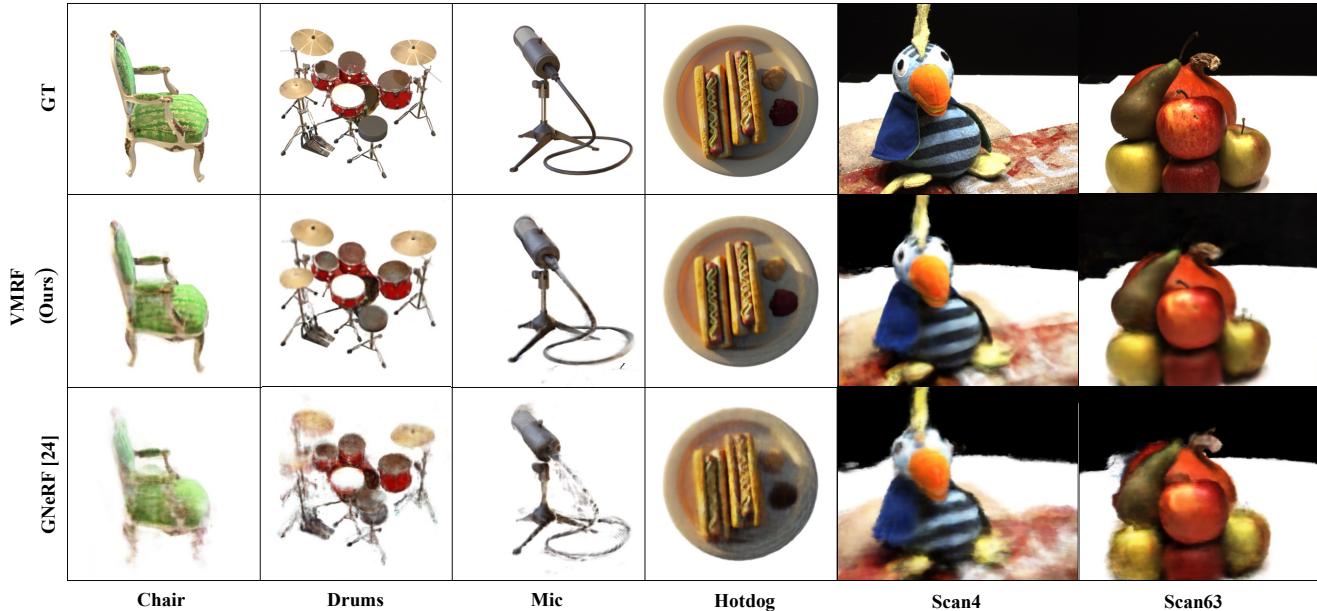


Figure 1: Illustration of novel view synthesis by NeRFs (without prior knowledge in camera poses or pose distributions in NeRF training): The state-of-the-art GNeRF [21] directly estimates *absolute pose* from image, which is susceptible to out-of-dataset views in training. The proposed VMRF instead estimates *relative pose* (with respect to the initially randomized pose) based on view matching results, which is robust to out-of-dataset views and learns better NeRF representations. The samples are from Synthetic-NeRF [22] and DTU [12].

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '22, October 10–14, 2022, Lisbon, Portugal

ABSTRACT

Neural Radiance Fields (NeRF) has demonstrated very impressive performance in novel view synthesis via implicitly modelling 3D representations from multi-view 2D images. However, most existing studies train NeRF models with either reasonable camera pose initialization or manually-crafted camera pose distributions which

are often unavailable or hard to acquire in various real-world data. We design *VMRF*, an innovative view matching NeRF that enables effective NeRF training without requiring prior knowledge in camera poses or camera pose distributions. VMRF introduces a view matching scheme, which exploits unbalanced optimal transport to produce a feature transport plan for mapping a rendered image with randomly initialized camera pose to the corresponding real image. With the feature transport plan as the guidance, a novel pose calibration technique is designed which rectifies the initially randomized camera poses by predicting relative pose transformations between the pair of rendered and real images. Extensive experiments over a number of synthetic and real datasets show that the proposed VMRF outperforms the state-of-the-art qualitatively and quantitatively by large margins.

CCS CONCEPTS

- Computing methodologies → Artificial intelligence.

KEYWORDS

computer vision, deep learning, neural radiance field, view matching, optimal transport, pose calibration

ACM Reference Format:

Jiahui Zhang, Fangneng Zhan, Rongliang Wu, Yingchen Yu, Wenqing Zhang, Bai Song, Xiaoqin Zhang, and Shijian Lu. 2022. VMRF: View Matching Neural Radiance Fields. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisbon, Portugal*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

1 INTRODUCTION

Neural Radiance Fields (NeRF) [22] has recently been proposed to learn 3D scene representations from multi-view 2D images. By representing a static scene as a continuous 5D function, it can achieve novel view synthesis via volume rendering. However, the NeRF training in most existing methods requires knowledge of accurate camera poses associated with input images, which is often either unavailable or hard to collect in real-world data. Although classical methods such as Structure-from-Motion [11, 27] can be adopted for pre-computing camera poses, they rely on keypoint detection and cannot recover view-dependent appearance which directly leads to unrealistic novel view rendering especially for scenes with low texture and repetitive patterns [21, 33].

To reduce the reliance on accurate camera poses, Wang et al. [33] modify the original NeRF [22] for joint optimization of 3D scene representations and camera poses. In addition, Lin et al. [16] design bundle-adjusting neural radiance fields for training NeRF from imperfect camera poses. However, both the two methods require camera pose initialization with reasonable accuracy. Meng et al. [21] propose GNeRF that optimizes NeRF representations with an inversion network for direct pose estimation. It entails an accurate pose distribution with which random camera poses can be sampled for training the inversion network for directly generating poses from input images. However, the accurate pose distribution is often unavailable and needs to be handcrafted for specific datasets. Without this accurate pose distribution, views that are beyond the training dataset would disturb the training of

inversion network, which further compromise the optimization of NeRF representations [21].

In this paper, we propose an innovative **View Matching NeRF (VMRF)** that learns NeRF representations without prior knowledge in camera poses or hand-crafted camera pose distributions. Instead of estimating camera poses from images directly, we design a novel relative pose estimation strategy that mitigates the dependence on hand-crafted camera pose distributions effectively. Specifically, with random initialization of camera poses, we design a novel view matching scheme that leverages optimal transport [26] to generate a transport plan that represents the feature matching between a real image and the corresponding rendering with the initial camera poses. To tackle non-surjective matching between different views, we introduce unbalanced optimal transport that achieves effective transport in the presence of unequal distribution masses. With feature correspondences from the obtained transport plan, a matching-based pose calibration technique with a relative transformation predictor is designed which learns *relative* pose transformation to calibrate the initially randomized poses towards the corresponding real-image poses. Thanks to the *relative* pose estimation strategy, VMRF learns accurate NeRF representations and achieves superior novel view synthesis as illustrated in Fig. 1.

The contributions of this work are threefold. *First*, we propose VMRF, a novel view matching NeRF that achieves superior NeRF representations without prior knowledge in camera poses or hand-crafted camera pose distributions. *Second*, we design a view matching scheme that exploits unbalanced optimal transport to build feature correspondences between cross-view images effectively. *Third*, we design a matching-based pose calibration technique that yields accurate camera poses by predicting relative pose transformation between real and randomly rendered images.

2 RELATED WORK

Neural Radiance Fields (NeRF). Recently, Mildenhall et al. [22] propose NeRF to achieve novel view synthesis by learning 3D scene representations from a set of posed input images. The key idea is to optimize a 5D scene function (represented by multi-layer perceptron (MLP)), which first maps 3D location and 2D viewing direction to RGB color and volume density and then utilizes differentiable volume rendering to accumulate colors and densities into 2D images from arbitrary viewpoints. Due to the multi-view consistency of the generated images, NeRF as well as its variants [1, 3, 9, 17, 20, 28, 42] has attracted increasing attention in different tasks such as dynamic scene representation [7, 8, 10, 24, 31], color and shape editing [18, 35], compositional scene modeling [23], scene relighting [2, 30] and skeleton-driven synthesis [25]. However, most existing NeRF-related designs require accurate camera poses for reliable NeRF training.

Several attempts have been made for relaxing the requirement for accurate camera poses. For example, Wang et al. [33] firstly propose to optimize camera poses and NeRF representations jointly. Lin et al. [16] design a coarse-to-fine bundle adjustment technique to learn NeRF representations from imperfect camera poses. Jeong et al. [13] propose a camera self-calibration algorithm for fine-tuning the initialized camera poses for improving NeRF representations.

However, these methods tend to suffer severe performance degradation while working with rough pose initialization. Meng et al. [21] propose to learn NeRF representations with random camera pose initialization, where an inversion network is introduced to directly estimate camera poses from images. However, it relies on accurate camera pose distributions which are dataset-specific and hand-crafted. Differently, the proposed VMRF allows to optimize NeRF representations with no knowledge of camera poses or hand-crafted camera pose distributions, thanks to our designed view matching scheme and matching-based camera pose calibration technique.

Camera Pose Estimation for NeRF. As reliable NeRF training requires accurate camera poses that are associated with the input images, most existing NeRF-based methods exploit classical Structured from Motion (SfM) techniques [11, 27, 34]) to build up and estimate per-scene camera poses based on keypoints detection and matching. However, SfM tends to fail in scenes with low texture and repetitive patterns, where few keypoints can be detected. With the recent advance of pose-free NeRF (i.e., NeRF trained without accurate camera poses), learning-based camera pose estimation [16, 21] has attracted increasing attention for estimating camera poses of input real images for NeRF training. A representative work is inversion network in GNeRF [21], which is trained to invert the input image to the corresponding camera pose. However, the training of the inversion network is susceptible to new views (beyond the training dataset) when hand-crafted camera pose distributions of reasonable accuracy are unavailable for camera pose initialization [21]. Instead of predicting camera poses directly from images, VMRF exploits view matching results to learn relative pose transformation which calibrates randomly initialized camera poses towards accurate camera pose with no prerequisite for either camera poses or hand-crafted camera pose distributions.

Optimal Transport. Optimal transport (OT) [32] is proposed to compare different distributions and yield an optimal transport plan for distribution matching. In recent years, optimal transport has been widely explored in the computer vision community [5, 14, 38–40]. For example, Liu et al. [19] exploit optimal transport in a unified framework to model semantic correspondences. For tackling unbalanced distributions with different masses and deviations, unbalanced optimal transport [4, 15] has also been explored for different tasks, such as image translation [36, 37, 41]. To the best of our knowledge, the proposed VMRF is the first that adapt unbalanced optimal transport for optimizing NeRF representations.

3 PRELIMINARY

This section presents the preliminary knowledge in camera pose distributions and NeRF 3D representations.

3.1 Camera Pose Distribution

Camera pose distribution imposes a range on random initialization of camera poses, which can be determined by five parameters [21]: the radius of the sphere where cameras are distributed, the range of camera elevation, the range of camera azimuth, camera lookat points and camera up vector. Well aligned with the finding in GNeRF [21], we observe that novel view synthesis is largely affected by

the variation in the range of camera azimuth and elevation (more details provided in the supplementary material). We therefore focus on studying the range of camera azimuth and elevation (via random selection with no handcrafted priors) in camera pose distribution in this work.

3.2 NeRF 3D Representation

We adopt the NeRF in [22] to represent 3D scenes. A neural radiance field is a continuous 5D function that maps an input 3D point $x \in \mathbb{R}^3$ and a 2D viewing direction $d \in \mathbb{S}^2$ to an RGB color $c \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}$. It can be formulated as $F_\Theta(x, b) \rightarrow (c, \sigma)$, where F is parameterized with MLP and Θ are network parameters. To render a 2D image from 3D NeRF representations, a differentiable volume renderer is adopted with a numerical integrator [22] for approximating the intractable volumetric projection integral. By taking N random samples along a camera ray r and representing their color and volume density values as $\{(c_r^i, \sigma_r^i)\}_{i=1}^N$, the RGB color c_r in 2D image can be rendered by:

$$c_r = \sum_{i=1}^N T_r^i \alpha_r^i c_r^i \quad T_r^i = \prod_{j=1}^{i-1} (1 - \alpha_r^j) \quad \alpha_r^i = 1 - \exp(\sigma_r^i \delta_r^i) \quad (1)$$

where T_r^i and α_r^i represent the transmittance and alpha value of i -th sample along camera ray r , respectively, and $\delta_r^i = x_r^{i+1} - x_r^i$ is the distance between adjacent samples. Given a dataset of RGB images $\mathcal{I} = \{I_1, I_2, \dots, I_n\}$ of a static scene with their corresponding camera poses $\Phi = \{\phi_1, \phi_2, \dots, \phi_n\}$, NeRF optimizes parameters Θ by a photometric loss (in the form of the sum of squared differences) between the real images \mathcal{I} and the corresponding rendered images.

4 PROPOSED METHOD

4.1 Pipeline

VMRF learns effective NeRF representations without prior knowledge in camera poses or camera pose distributions. Similar to GNeRF [21], it first learns a coarse NeRF with randomly initialized camera poses $\Phi' = \{\phi'_i, i \in [1, n]\}$ as illustrated in Fig. 2. With the coarse NeRF, camera poses $\hat{\Phi} = \{\hat{\phi}_i, i \in [1, n]\}$ of real images $\mathcal{I} = \{I_i, i \in [1, n]\}$ are then estimated via view matching and matching-based pose calibration. Specifically, the view matching yields an optimal transport plan by matching the views of a real image and the corresponding rendering with the randomly initialized camera pose. The relative pose transformation between the real and the rendered images are then derived from the optimal transport plan by the *Relative Transformation Predictor*. The camera pose of the real image (i.e., $\hat{\phi}_i$) can thus be determined for NeRF training. Note that we also employ a photometric loss to jointly refine the NeRF training as well as the estimated camera poses as illustrated in *Joint Refinement*. VMRF utilizes a hybrid and iterative optimization scheme as in GNeRF [21] to achieve end-to-end training.

4.2 View Matching Scheme

The view matching scheme aims to match features between rendered images with randomly initialized camera poses and real images, thus providing accurate guidance for the subsequent camera pose calibration. To avoid many-to-one matching and ensure end-to-end optimization, we introduce optimal transport with entropic

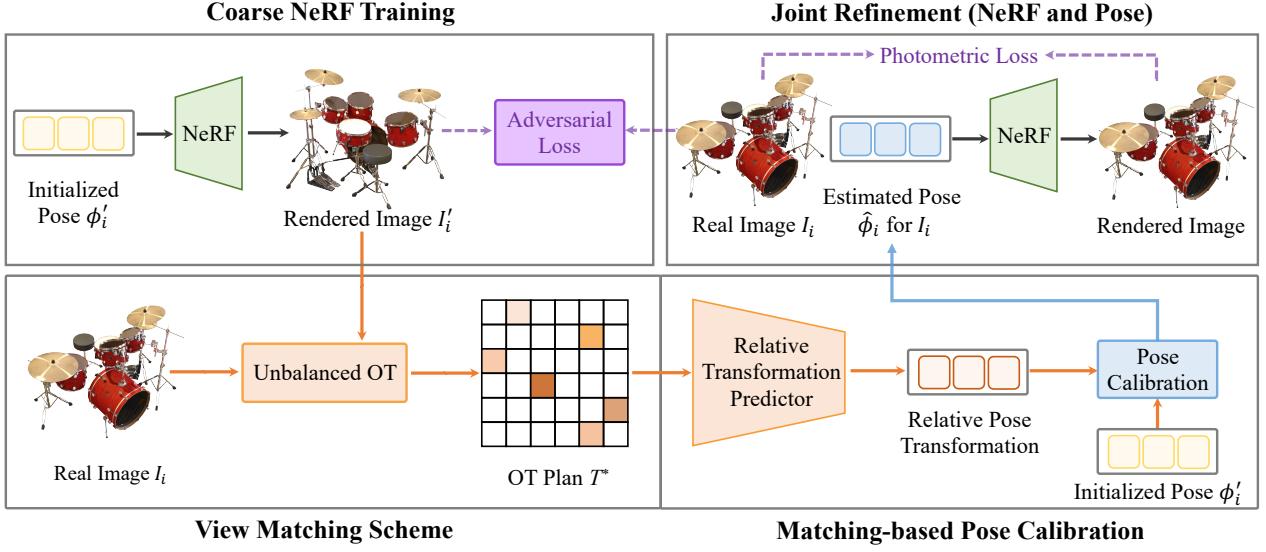


Figure 2: The framework of VMRF: A coarse NeRF is first trained with randomly *Initialized Poses* as in [21]. Camera poses of real images are then estimated via the proposed view matching scheme and matching-based pose calibration. Specifically, view matching learns to build an *Optimal Transport (OT) Plan* T^* between the image I'_i rendered by the initial pose ϕ'_i and the corresponding real image I_i . The relative pose transformation between the two images is then estimated from T^* , with which the pose $\hat{\phi}_i$ of I_i can be computed by calibrating the initial camera pose ϕ'_i . Thus, camera pose $\hat{\phi}_i$ and real image I_i can be paired for NeRF training. Meanwhile, $\hat{\phi}_i$ and the coarse NeRF are jointly refined with a photometric loss. Similar to GNeRF [21], VMRF is end-to-end trainable under a hybrid and iterative optimization scheme.

regularization that can achieve image feature matching in a differentiable way.

Standard optimal transport aims at computing a minimal cost transportation (i.e., optimal transport plan) to transform a source distribution into a target distribution with the same total mass. In our work, we model feature matching between images from different views as an optimal transport problem. The image rendered with randomly initialized pose and the real image are thus treated as the source image and the target image, respectively. Specifically, we first adopt two pre-trained VGG-19 [29] to extract source feature set $f^s \in \mathbb{R}^{l \times d}$ and target feature set $f^t \in \mathbb{R}^{l \times d}$ from the source and target images, respectively, where each feature set involves l feature vectors and d denotes the dimension of feature vectors. Then, we encode the two feature sets to build discrete source and target distributions (μ_s and μ_t) as follows:

$$\mu_s = \sum_{i=1}^l p_i^s \delta(f_i^s), \quad \mu_t = \sum_{j=1}^l p_j^t \delta(f_j^t), \quad (2)$$

where $\delta(\cdot)$ denotes the Dirac function, and p_i^s and p_j^t represent the probability mass of the i -th source feature vector f_i^s and the probability mass of the j -th target feature vector f_j^t , respectively. Then, we define a cost matrix $M \in \mathbb{R}^{l \times l}$ based on matching similarities between the source and target feature vectors, where each element M_{ij} denotes the moving cost from probability mass p_i^s to p_j^t . The cost matrix can be computed by: $M = 1 - \frac{f^s f^t^\top}{\|f^s\| \|f^t\|}$. The optimal

transport problem can thus be formulated by:

$$\begin{aligned} T^* &= \arg \min_T \sum_{i,j=1}^l T_{ij} M_{ij} \\ \text{s.t. } & T \mathbf{1}_l = \mu_s, \quad T^\top \mathbf{1}_l = \mu_t \end{aligned} \quad (3)$$

where T^* is defined as the optimal transport plan and T_{ij} represents the optimal amount of mass moved from f_i^s to f_j^t in order to minimize the total moving cost. Many-to-one matching is effectively circumvented by constraining μ_s to be the row sum of T and constraining μ_t to be the column sum of T .

A key constraint of the standard optimal transport is that the total masses of the source and target distributions need to be equal. Due to the occlusion of certain features caused by camera poses, feature matching between images from different views is not surjective and so the source and target distributions have inconsistent total mass. Directly applying the standard optimal transport will thus match all features including outlier samples (i.e., features that exist only in one view), which leads to inaccurate matching and false matching. We introduce unbalanced optimal transport (UOT) [4] for accurate matching distributions with arbitrary masses. Specifically, we replace the ‘strict’ constraints on the marginals of plan T (in Eq.3) by ‘soft’ penalties with Kullback-Leibler divergence. The unbalanced optimal transport can be formulated as follows:

$$T^* = \arg \min_T \left[\sum_{i,j=1}^l T_{ij} M_{ij} + \epsilon \text{KL}(T \mathbf{1}_l || \mu_s) + \epsilon \text{KL}(T^\top \mathbf{1}_l || \mu_t) \right] \quad (4)$$

where $\epsilon > 0$ denotes the regularization parameter.

To implement a differentiable UOT, we introduce entropic regularization for approximating the UOT as follows:

$$T^* = \arg \min_T \left[\sum_{i,j=1}^l T_{ij} M_{ij} - \eta H(T) + \epsilon KL(T\mathbf{1}_l || \mu_s) + \epsilon KL(T^T \mathbf{1}_l || \mu_t) \right], \quad (5)$$

where $H(T)$ is an entropic regularization term which is defined by: $H(T) = -\sum_{i,j=1}^l T_{ij} \log(T_{ij} - 1)$, and $\eta > 0$ is a regularization parameter.

To obtain optimal transport plan of the entropic regularized UOT, we represent the entropic regularized UOT in the Fenchel-Legendre dual form and introduce the Sinkhorn algorithm [4] to compute the optimal solution (u^*, v^*) .

$$B(u^*, v^*) = T^* = \arg \min_{u, v \in \mathbb{R}^l} \left[\eta \sum_{i,j=1}^l \exp\left(\frac{u^i + v^j - M_{ij}}{\eta}\right) + \epsilon \langle e^{-u/\epsilon}, \mu_s \rangle + \epsilon \langle e^{-v/\epsilon}, \mu_t \rangle \right], \quad (6)$$

where $u, v \in \mathbb{R}^l$ denote dual vectors. The optimal transport plan T^* in the dual form can thus be represented by:

$$B(u^*, v^*) = \text{diag}(e^{u^*/\eta}) e^{M/\eta} \text{diag}(e^{v^*/\eta}) \quad (7)$$

4.3 Matching-based Pose Calibration

With optimal transport plans between real images and images rendered with randomly initialized poses, we design a matching-based pose calibration technique that estimates the pose of the real image by predicting a relative pose transformation. It involves two major steps as illustrated in Fig 2.

First, with feature-level matching (from the obtained transport plan) between a real image and a rendered image, we introduce a relative transformation predictor that predicts relative rotation and relative translation for pose calibration. This process can be formulated as follows:

$$[t_x, t_y, t_z, \theta_x, \theta_y, \theta_z] = RTP(T^*), \quad (8)$$

where $RTP(\cdot)$ denotes the relative transformation predictor, t_x, t_y, t_z represent relative translations along the x, y , and z axes, respectively, and $\theta_x, \theta_y, \theta_z$ represent relative rotations about the x, y and z axes, respectively.

We then estimate the camera pose $\hat{\phi}_i$ of the real image by calibrating the randomly initialized camera pose ϕ'_i of the rendered image with the predicted relative rotation and relative translation. Specifically, we construct a relative rotation matrix ΔR and a translation vector Δt as follows:

$$\Delta R = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta_x & \sin\theta_x \\ 0 & -\sin\theta_x & \cos\theta_x \end{bmatrix} \begin{bmatrix} \cos\theta_y & 0 & -\sin\theta_y \\ 0 & 1 & 0 \\ \sin\theta_y & 0 & \cos\theta_y \end{bmatrix} \begin{bmatrix} \cos\theta_z & \sin\theta_z & 0 \\ -\sin\theta_z & \cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (9)$$

$$\Delta t = [t_x, t_y, t_z] \quad (10)$$

Given the randomly initialized pose ϕ'_i , the estimated camera pose $\hat{\phi}_i$ associated with the real image can be obtained by:

$$\hat{\phi}_i = \begin{bmatrix} \Delta R & \Delta t^T \\ 0^3 & 1 \end{bmatrix} \phi'_i \quad (11)$$

4.4 Loss Function

In this section, we detail loss functions that are employed in the training process as below.

Adversarial loss. Due to the lack of known camera poses, we follow [21] and first train a coarse NeRF with randomly initialized poses and an adversarial loss. The images generated by the coarse NeRF have similar distribution as that of real images in training dataset. The adversarial loss \mathcal{L}_{adv} can be defined as follows:

$$\mathcal{L}_{adv}(F, D) = \mathbb{E}_{I \sim P_{data}} [\log(D(I))] + \mathbb{E}_{F(\Phi') \sim P_g} [\log(1 - D(F(\Phi')))], \quad (12)$$

where F and D represent the NeRF model and the discriminator, P_{data} and P_g denote the distribution of real images and generated images, and Φ' represents a set of randomly initialized poses.

Calibration loss. We introduce a calibration loss \mathcal{L}_{ca} to train the relative transformation predictor to estimate relative pose transformation from the obtained transport plan. This is performed in an unsupervised manner as follows:

$$\mathcal{L}_{ca}(RTP) = \mathbb{E} \left[\|F_{ca}(RTP(T_{ab}^*), \phi'_a) - \phi'_b\|_2^2 \right] = \mathbb{E} \left[\|F_{ca}(RTP(F_{uot}(F(\phi'_a), F(\phi'_b))), \phi'_a) - \phi'_b\|_2^2 \right] \quad (13)$$

where (ϕ'_a, ϕ'_b) is a pair of poses with random initialization, T_{ab}^* denotes optimal transport plan between images rendered with ϕ'_a and ϕ'_b , $F_{ca}(\cdot)$ represents the calibration process described in Eq. 11, F_{uot} denotes the unbalanced optimal transport, and $RTP(\cdot)$ denotes our relative transformation predictor.

Photometric loss. For the joint refinement of NeRF model F and the estimated camera poses $\hat{\Phi}$, a photometric loss \mathcal{L}_P is employed which is formulated as follows:

$$\mathcal{L}_P(F, \hat{\Phi}) = \frac{1}{n} \sum_{i=1}^n \|I_i - F(\hat{\phi}_i)\|_2^2 \quad (14)$$

Note, the VMRF is trained end-to-end following the hybrid and iterative optimization scheme in GNeRF [21].

4.5 Network Details

In the view matching scheme, The number of feature vector in each feature set is set to $l = 64$ and the dimension of feature vector is set to $d = 256$. The regularization parameter η that controls the smoothness of UOT is set to 0.005, parameter study of η is presented in Section 5.5. For the matching-based pose calibration, we utilize the vision transformer [6] as relative transformation predictor, where the last layer is modified to generate relative transformation values. For camera pose, we employ a 3D embedding in Euclidean space as camera position. Following [44], we represent camera rotation as a continuous 6D embedding. In term of NeRF in VMRF, we utilize the original NeRF architecture [22].

5 EXPERIMENT

5.1 Datasets and Implementation Details

Datasets. For training and testing, we follow the dataset setting of GNeRF[21] and conduct experiments on both synthetic and real scenes. For synthetic scenes, we employ the dataset Synthetic-NeRF

Table 1: Quantitative comparisons of novel view synthesis on the dataset Synthetic-NeRF: VMRF outperforms the state-of-the-art GNeRF [21] consistently in PSNR, SSIM and LPIPS for different synthetic scenes with no hand-crafted pose distributions. D_{s1} , D_{s2} and D_{s3} denote randomly selected pose distributions, where the ranges of azimuth and elevation (R_{azi}, R_{ele}) are $([0^\circ, 360^\circ], [-30^\circ, 90^\circ])$, $([0^\circ, 360^\circ], [-60^\circ, 90^\circ])$ and $([0^\circ, 360^\circ], [-90^\circ, 90^\circ])$, respectively. GNeRF* denote a GNeRF variant that has the same NeRF settings and similar number of parameters as the proposed VMRF. All methods are trained with the same training data and batch size.

Scene	PSNR↑		SSIM↑		LPIPS↓	
	GNeRF*	VMRF	GNeRF*	VMRF	GNeRF*	VMRF
	25.01	26.05	0.8940	0.9083	0.1529	0.1397
D_{s1}	20.63	23.07	0.8628	0.8917	0.2019	0.1605
	22.95	25.23	0.8493	0.8865	0.1630	0.1215
	23.68	27.63	0.9332	0.9483	0.1095	0.0803
	17.91	21.39	0.7626	0.7998	0.3628	0.2933
	22.18	24.53	0.8753	0.8987	0.2006	0.1695
D_{s2}	19.05	21.25	0.8312	0.8652	0.2670	0.2098
	21.39	23.51	0.8412	0.8618	0.1794	0.1429
	23.22	24.39	0.9216	0.9392	0.1301	0.1021
	15.89	20.35	0.7416	0.7895	0.4034	0.3502
	21.27	23.18	0.8665	0.8886	0.2326	0.1593
D_{s3}	18.08	20.01	0.8132	0.8371	0.3314	0.2877
	18.22	21.59	0.8176	0.8320	0.2363	0.1762
	17.22	20.29	0.8592	0.8971	0.3154	0.2208
	14.31	17.58	0.7295	0.7602	0.4673	0.3913

[22] which consists of objects with complicated geometry. We adopt the original split of the dataset where 100 images are used for training and eight images are randomly sampled from the test set for evaluations. The input image is resized to 400×400 for both training and test. For real scenes, we select three representatives in the dataset DTU [12] including ‘scan4’ and ‘scan63’ that are characterized by rich textures as well as ‘scan48’ that has very sparse texture. For a total of 49 images per scene, we perform the training on 43 images and use the rest 6 images for evaluation. For both training and test, the image size is fixed at 500×400 .

Implementation Details. For various parameters in camera pose distributions (detailed in Section 3.1), we focus on the range of camera elevation and azimuth (via random selection with no handcrafted priors) as the two parameters are decisive factors that affect novel view synthesis (which has been verified in [21] as well as our studies as detailed in the supplementary material). For other insignificant parameters, we follow GNeRF [21] and set the sphere radius at 4.0 for both synthetic and real datasets, and set the camera lookat points at $(0, 0, 0)$ and $\mathcal{N}(0, 0.01^2)$, respectively, for the synthetic dataset and real dataset. We adopt Adam optimizer to train the NeRF, discriminator, relative transformation predictor and camera poses with initial learning rates of 0.0005, 0.0001, 0.0001 and 0.005, respectively. The mini-batch size is set at 12 for both synthetic and real scenes in training. We use the Pytorch framework in implementation and employ one NVIDIA RTX 3090ti GPU for both training and inference.

Table 2: Quantitative comparisons of novel view synthesis on the dataset DTU: VMRF outperforms the state-of-the-art GNeRF* [21] with no hand-crafted pose distributions. D_{r1} and D_{r2} denote randomly selected pose distributions, where the range of azimuth and elevation (R_{azi}, R_{ele}) are $([0^\circ, 180^\circ], [0^\circ, 90^\circ])$ and $([0^\circ, 180^\circ], [-30^\circ, 90^\circ])$, respectively. All methods are trained with the same training data and batch size.

Scene	↑ PSNR		↑ SSIM		↓ LPIPS	
	GNeRF*	VMRF	GNeRF*	VMRF	GNeRF*	VMRF
D_{r1}	17.04	19.51	0.6124	0.6503	0.4676	0.4205
	18.24	21.30	0.7743	0.7986	0.3459	0.3063
	19.43	22.52	0.8217	0.8636	0.3041	0.2551
D_{r2}	16.22	18.38	0.6079	0.6297	0.4955	0.4707
	15.81	20.61	0.7520	0.7913	0.4207	0.3185
	17.30	20.59	0.8152	0.8538	0.3202	0.2839

Table 3: Quantitative comparisons of camera pose estimation for NeRF training without hand-crafted pose distributions (on Synthetic-NeRF): Rot and Trans denote mean camera rotation differences and mean camera translation differences, respectively.

Scene	GNeRF* [21]		Ours	
	Rot($^\circ$)↓	Trans↓	Rot($^\circ$)↓	Trans↓
Chair	6.078	0.339	4.853	0.283
Drums	2.769	0.137	1.284	0.081
Lego	3.315	0.251	2.160	0.164
Mic	3.021	0.161	1.394	0.069
Ship	31.56	1.379	16.89	0.714

5.2 Comparisons with the State-of-the-Art

Novel View Synthesis. We compare VMRF with the state-of-the-art GNeRF [21] over different synthetic scenes and real scenes. We did not compare with NeRF— [33], BARF [16] and SCNeRF [13] as the three methods require reasonable (instead of random) camera pose initialization. In addition, we employ randomly selected pose distributions to verify the effectiveness of the proposed VMRF. Specifically, we set the range of azimuth and elevation (R_{azi}, R_{ele}) at $([0^\circ, 360^\circ], [-30^\circ, 90^\circ])$ to $([0^\circ, 360^\circ], [-90^\circ, 90^\circ])$ for the dataset Synthetic-NeRF with synthetic scenes. For the dataset DTU with real scenes, (R_{azi}, R_{ele}) is set at $([0^\circ, 180^\circ], [0^\circ, 90^\circ])$ to $([0^\circ, 180^\circ], [-30^\circ, 90^\circ])$ instead. For fair comparisons, we re-train GNeRF*¹ by using its official codes, and both GNeRF* and VMRF are trained with the same training dataset and the same batch size.

Table 1 and Table 2 show experimental results over datasets Synthetic-NeRF and DTU, respectively, where all evaluations were conducted on the same test images as described in section 5.1. We can observe that VMRF outperforms GNeRF* consistently across PSNR, SSIM and LPIPS as well as all evaluated scenes. The superior performance of VMRF is largely attributed to our designed view matching and pose calibration techniques that conduct effective

¹* indicates a version of GNeRF that has the same NeRF settings and similar number of parameters with VMRF. Details are provided in the supplementary material.

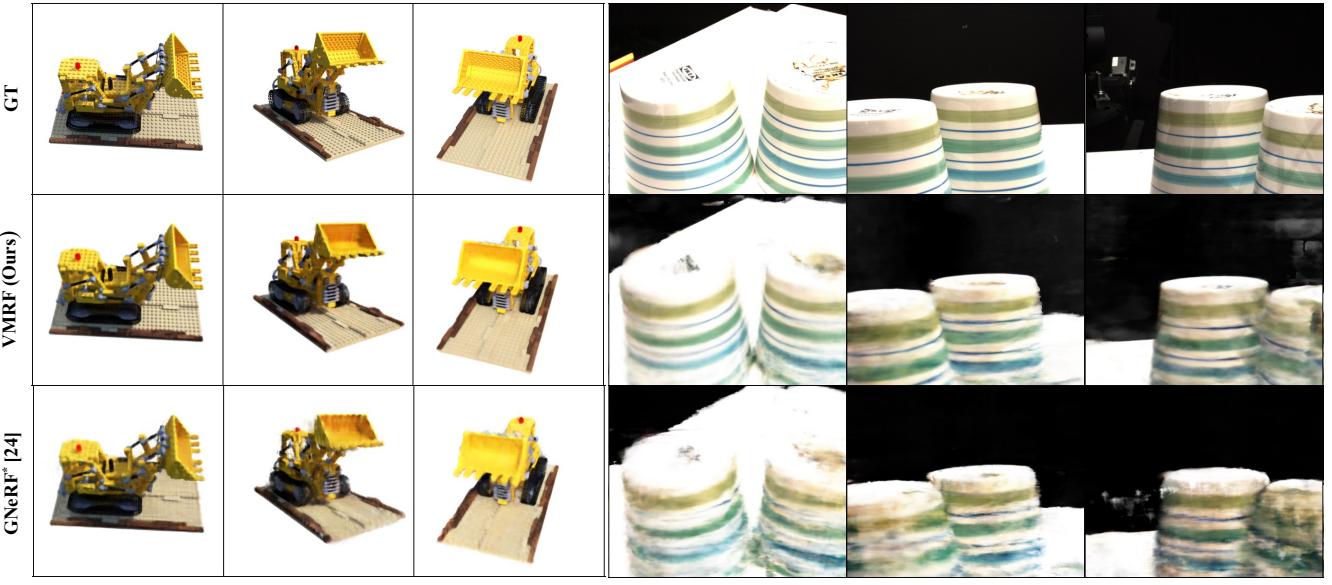


Figure 3: Qualitative comparisons of VMRF with GNeRF* in novel view synthesis: The comparisons are conducted over different views of the scene ‘lego’ in Synthetic-NeRF and the scene ‘scan48’ in DTU, where ‘GT’ denotes the ground-truth images. It is clear that VMRF synthesizes high-fidelity images with less artifacts and finer details compared with GNeRF*.

NeRF training with relative pose estimation without requiring hand-crafted pose distributions. The relative pose estimation is more tolerant to the disturbance resulting from out-of-dataset views which leads to robust training and effective NeRF representations. The quantitative experiments are well aligned with the qualitative experiments as illustrated in Fig. 3 where VMRF synthesizes high-fidelity images with less artifacts and finer details.

Camera Poses Estimation. We also compare the accuracy of the estimated camera poses used for NeRF training. The comparison is under the setup of random pose distribution selection where the range of azimuth and elevation are randomly set to $[0^\circ, 360^\circ]$ and $[-30^\circ, 90^\circ]$, respectively. Following [21], we perform this evaluation on Synthetic-NeRF with the metric of mean camera rotation difference (Rot) and mean translation difference (Trans) that are computed with the toolbox [43] on the training set. As Table 3 shows, VMRF outperforms GNeRF* clearly and consistently across all evaluated scenes. The superior accuracy is largely attributed to the relative pose estimation in VMRF which is more independent of hand-crafted pose distributions as compared with the direct pose estimation in GNeRF*.

5.3 Ablation Studies

Effect of relative pose estimation. To examine the effectiveness of the proposed relative pose estimation, we perform ablation experiments to study how it affects view synthesis in PSNR, SSIM, LPIPS, Rot and Trans. As shown in Table 4, the model *Base* works as a baseline [21] which directly predicts camera poses from images. The complete model *Base+VMS+MPC* with both view

Table 4: Ablation studies of the proposed VMRF on the scene ‘mic’ of Synthetic-NeRF: The baseline model *Base* [21] estimates camera poses from images directly. *Base+VMS* incorporates the proposed view matching scheme (VMS) above *Base*, where real-image pose is estimated from the concatenation of a rendered image with randomly initialized camera pose and a transport plan between the real and rendered images. *Base+MPC* includes the proposed pose calibration (MPC) above *Base*, where MPC receives the concatenation of a real image and a rendered image with randomly initialized camera pose. *Base+VMS+MPC* is equivalent to the proposed VMRF. All models are trained without known camera poses or hand-crafted pose distributions (the range of azimuth and elevation are randomly set to $[0^\circ, 360^\circ]$ and $[-30^\circ, 90^\circ]$).

Models	Evaluation Metrics				
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	Rot($^\circ$) \downarrow	Trans \downarrow
Base	23.15	0.9270	0.1263	3.582	0.187
Base+VMS	25.03	0.9337	0.1089	2.735	0.153
Base+MPC	25.21	0.9366	0.1007	2.651	0.144
Base+VMS+MPC	27.63	0.9483	0.0803	1.394	0.069

matching scheme and matching-based pose calibration is equivalent to the proposed VMRF which performs relative pose estimation that utilizes view matching results to model relative transformation for pose calibration. Quantitative experiments show that *Base+VMS+MPC* performs clearly better than the *Base* in novel view synthesis, demonstrating that the proposed relative pose

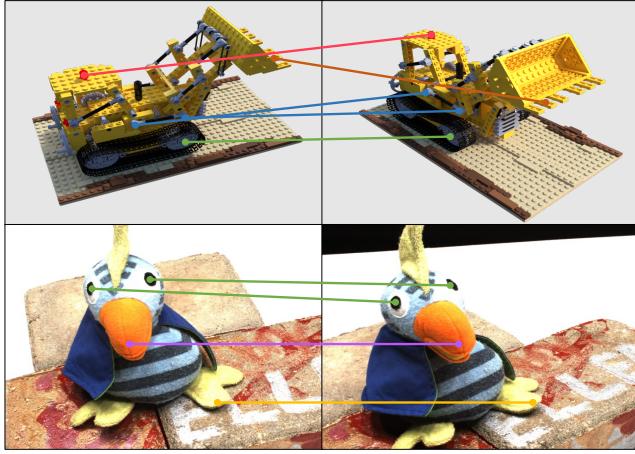


Figure 4: Visualization of feature matching in synthetic and real scenes: Links are drawn based on the feature matching in the derived transport plan. The upper sample is the synthetic scene ‘lego’ from Synthetic-NeRF and the lower sample is the real scene ‘Scan4’ from DTU. It shows that the proposed view matching with unbalanced optimal transport achieves feature-level matching effectively.

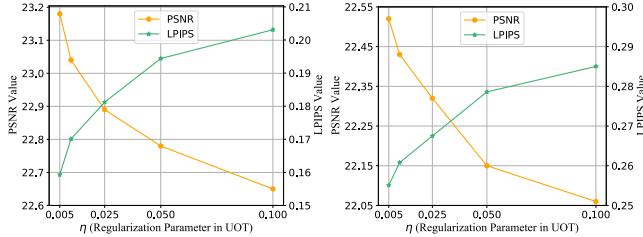


Figure 5: Parameter η in VMRF: The performance of VMRF varies with η (regularization parameter of UOT), and decreasing η improves novel view synthesis consistently. The two graphs from left to right are computed for the scene ‘chair’ of Synthetic-NeRF and the scene ‘scan63’ of DTU, respectively.

estimation alleviates the dependence of NeRF training on hand-crafted pose distribution effectively. The experimental results are also well aligned with qualitative experimental results in Fig. ?? where *Base+VMS+MPC* with the proposed relative pose estimation generates new views with clearly less artifacts and finer details.

Effect of view matching scheme. We also conduct experiments to examine the contribution of our proposed view matching scheme. As shown in Table 4, we train the model *Base+VMS* that includes the view matching scheme above the model *Base*. Compared with *Base*, *Base+VMS* performs clearly better in PSNR, SSIM, LPIPS, Rot and Trans which demonstrates the effectiveness of the proposed view matching scheme. The quantitative results are well aligned with the qualitative experiments in Fig. ?? as well. Due to space limit,

ablation experiments about the unbalanced optimal transport in view matching scheme are presented in the supplementary material.

Effect of matching-based pose calibration. We further examine how the proposed matching-based pose calibration contributes to NeRF training. As Table 4 shows, we train the model *Base+MPC* which incorporates matching-based pose calibration above the model *Base*. It can be observed that *Base+MPC* improves PSNR, SSIM, LPIPS, Rot and Trans consistently as compared with *Base*. The better performance is largely attributed to the included relative transformation predictor that leads to robust training (with no hand-crafted pose distributions) by modeling relative pose transformation. The effectiveness of the proposed matching-based pose calibration can be observed in Fig. ?? as well where the model *Base+MPC* can produce clearer visual results than the model *Base*.

5.4 Visualization

Based on the transport plan obtained by the proposed matching scheme, we visualize feature matching between images of different views in both synthetic and real scenes as illustrated in Fig. 4. It can be observed that the features on the left image can be accurately matched to those on the right image. This shows that the proposed view matching with unbalanced optimal transport can produce accurate feature matching which provides critical guidance for the ensuing pose calibration.

5.5 Parameters Investigation

We also conduct experiments on the scene ‘chair’ of dataset Synthetic-NeRF and the scene ‘scan63’ of dataset DTU for studying how the regularization parameter η affects the VMRF performance. As described in Section 4.2, a larger η tends to increase the smoothness of the unbalanced optimal transport. It can be observed in Fig. 5 that with the decrease of η , the performance (in PSNR \uparrow and LPIPS \downarrow scores) of VMRF is improved consistently. On the other hand, the stability of the model training is impaired while η becomes too small. Therefore, we set the regularization parameter η at 0.005 in our implemented system.

6 CONCLUSION

This paper presents VMRF, a view matching NeRF that aims to achieve superior NeRF representations without prior knowledge in camera poses or hand-crafted pose distributions. Specifically, a view matching scheme with unbalanced optimal transport is designed to build an optimal transport plan to guide camera pose calibration. In addition, we design a matching-based pose calibration technique with a relative transformation predictor that utilizes the obtained optimal transport plan to predict relative pose transformation for camera pose calibration. With the view matching and matching-based pose calibration, VMRF achieves a novel relative pose estimation strategy that allows to reduce the dependence on hand-crafted pose distributions. Extensive quantitative and qualitative experiments over synthetic and real scenes show the superiority of proposed VMRF without prior knowledge in camera poses or hand-crafted pose distributions. Moving forward, we will focus on studying NeRF training without prior knowledge in camera poses or hand-crafted pose distributions in challenging multi-object scenes.

ACKNOWLEDGMENTS

This work was supported by the Ministry of Education of Singapore (Project number: MOE-T2EP20220-0003) and in part by the National Natural Science Foundation of China [grant nos. 61922064, U2033210].

REFERENCES

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5855–5864.
- [2] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. 2021. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12684–12694.
- [3] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyu Yu, and Hao Su. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14124–14133.
- [4] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. 2018. Scaling algorithms for unbalanced optimal transport problems. *Math. Comp.* 87, 314 (2018), 2563–2609.
- [5] Nicolas Courty, Rémi Flamary, and Devis Tuia. 2014. Domain adaptation with regularized optimal transport. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 274–289.
- [6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [7] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. 2021. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14324–14334.
- [8] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. 2021. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5712–5721.
- [9] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. 2021. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985* (2021).
- [10] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. 2021. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5784–5794.
- [11] Richard Hartley and Andrew Zisserman. 2003. *Multiple view geometry in computer vision*. Cambridge university press.
- [12] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 406–413.
- [13] Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. 2021. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5846–5854.
- [14] Nicholas Kolkin, Jason Salavon, and Gregory Shakhnarovich. 2019. Style transfer by relaxed optimal transport and self-similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10051–10060.
- [15] Matthias Liero, Alexander Mielke, and Giuseppe Savaré. 2018. Optimal entropy-transport problems and a new Hellinger–Kantorovich distance between positive measures. *Inventiones mathematicae* 211, 3 (2018), 969–1117.
- [16] Chen-Hsuan Lin, Wei-Chiu Ma, Antonia Torralba, and Simon Lucey. 2021. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5741–5751.
- [17] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural sparse voxel fields. *Advances in Neural Information Processing Systems* 33 (2020), 15651–15663.
- [18] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. 2021. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5773–5783.
- [19] Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. 2020. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4463–4472.
- [20] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7210–7219.
- [21] Quan Meng, Anpei Chen, Haomin Luo, Minye Wu, Hao Su, Lan Xu, Xuming He, and Jingyi Yu. 2021. Gnerf: Gan-based neural radiance field without posed camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 6351–6361.
- [22] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision*. Springer, 405–421.
- [23] Michael Niemeyer and Andreas Geiger. 2021. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11453–11464.
- [24] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. 2021. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5865–5874.
- [25] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.
- [26] Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning* 11, 5–6 (2019), 355–607.
- [27] Johannes L Schonberger and Jan-Michael Frahm. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4104–4113.
- [28] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems* 33 (2020), 20154–20166.
- [29] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [30] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. 2021. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7495–7504.
- [31] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. 2021. 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*. 12959–12970.
- [32] Cédric Villani. 2009. *Optimal transport: old and new*. Vol. 338. Springer.
- [33] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. 2021. NeRF+: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064* (2021).
- [34] Changchang Wu. 2013. Towards linear-time incremental structure from motion. In *2013 International Conference on 3D Vision-3DV 2013*. IEEE, 127–134.
- [35] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. 2021. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13779–13788.
- [36] Fangneng Zhan, Yingchen Yu, Kaiwen Cui, Gongjie Zhang, Shijian Lu, Jianxiang Pan, Changgong Zhang, Feiyi Ma, XuanSong Xie, and Chunyan Miao. 2021. Unbalanced feature transport for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15028–15038.
- [37] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Kaiwen Cui, Aoran Xiao, Shijian Lu, and Ling Shao. 2021. Bi-level feature alignment for versatile image translation and manipulation. *arXiv preprint arXiv:2107.03021* (2021).
- [38] Fangneng Zhan, Yingchen Yu, Changgong Zhang, Rongliang Wu, Wenbo Hu, Shijian Lu, Feiyi Ma, XuanSong Xie, and Ling Shao. 2022. Gmlight: Lighting estimation via geometric distribution approximation. *IEEE Transactions on Image Processing* 31 (2022), 2268–2278.
- [39] Fangneng Zhan, Changgong Zhang, Wenbo Hu, Shijian Lu, Feiyi Ma, XuanSong Xie, and Ling Shao. 2021. Sparse needlets for lighting estimation with spherical transport loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12830–12839.
- [40] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiyi Ma, and XuanSong Xie. 2021. Emlight: Lighting estimation via spherical distribution approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3287–3295.
- [41] Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Rongliang Wu, and Shijian Lu. 2022. Modulated contrast for versatile image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18280–18290.
- [42] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492* (2020).
- [43] Zichao Zhang and Davide Scaramuzza. 2018. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 7244–7251.
- [44] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. 2019. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5745–5753.