HandNeRF: Neural Radiance Fields for Animatable Interacting Hands

Zhiyang Guo[1] Wengang Zhou[1,2*] Min Wang[2] Li Li[1] Houqiang Li[1,2*]
[1]CAS Key Laboratory of Technology in GIPAS, EEIS Department,
University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
guozhiyang@mail.ustc.edu.cn, {zhwg, lil1, lihq}@ustc.edu.cn, wangmin@iai.ustc.edu.cn

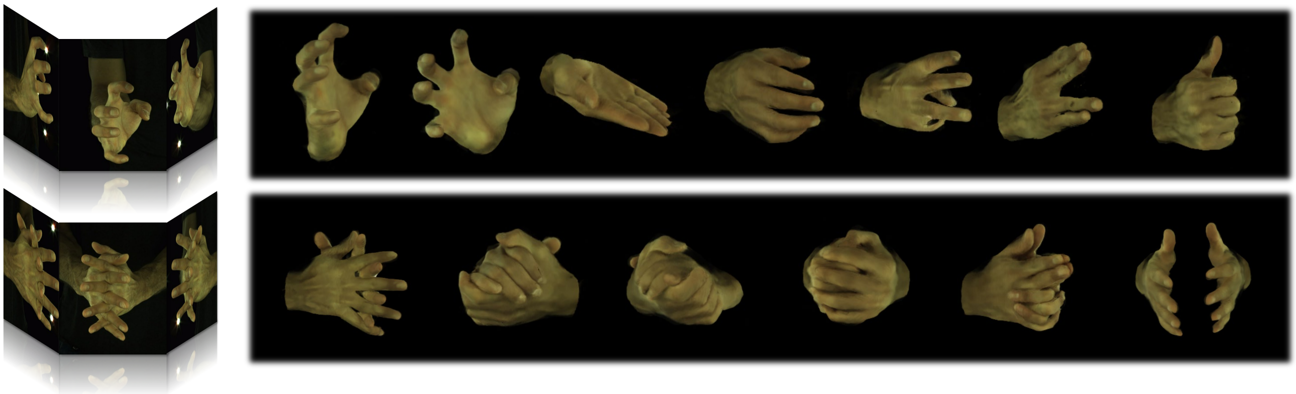

Figure 1. Given a set of multi-view images capturing a pose sequence of a single hand or two interacting hands (left), HandNeRF models the scene in a unified manner with neural radiance fields, enabling rendering of novel hand poses from arbitrary viewing directions (right).

Abstract

We propose a novel framework to reconstruct accurate appearance and geometry with neural radiance fields (NeRF) for interacting hands, enabling the rendering of photo-realistic images and videos for gesture animation from arbitrary views. Given multi-view images of a single hand or interacting hands, an off-the-shelf skeleton estimator is first employed to parameterize the hand poses. Then we design a pose-driven deformation field to establish correspondence from those different poses to a shared canonical space, where a pose-disentangled NeRF for one hand is optimized. Such unified modeling efficiently complements the geometry and texture cues in rarely-observed areas for both hands. Meanwhile, we further leverage the pose priors to generate pseudo depth maps as guidance for occlusion-aware density learning. Moreover, a neural feature distillation method is proposed to achieve cross-domain alignment for color optimization. We conduct extensive experiments to verify the merits of our proposed HandNeRF and report a series of state-of-the-art results both qualitatively and quantitatively on the large-scale InterHand2.6M dataset.

1. Introduction

As a dexterous tool to interact with the physical world and convey rich semantic information, the modeling and reconstruction of human hands have attracted substantial attention from the research community. Typically, the synthesis of realistic hand images or videos with different postures in motion has a wide range of applications, *e.g.*, human-computer interaction, sign language production, virtual and augmented reality technologies such as telepresence, *etc.*

Classic hand-modeling works are mainly built upon parameterized mesh models such as MANO [31]. They fit the geometry of hands to polygon meshes manipulated by shape and pose parameters, and then complete coloring via texture mapping. Despite being widely adopted, those models have the following limitations. On the one hand, high-frequency details are hard to present on low-resolution meshes, hindering the production of photo-realistic images. On the other hand, no special design is developed for interacting hands, which is a non-trivial scenario involving complex postures with self-occlusion.

To address the above issues and push the boundary of realistic human hand modeling, motivated by the recent success of NeRF [17] in modeling human body [11, 25, 26], we propose **HandNeRF**, a novel framework that unifiedly models the geometry and texture of animatable interacting

[*]Corresponding Authors.

hands with neural radiance fields (NeRF). Specifically, a pose-conditioned deformation field is introduced to warp the sampled observing ray into a canonical space, guided by the prior-based blend skinning transformation and a learnable error-correction network dealing with non-rigid deformations. The different input postures are thereby mapped to a common mean pose, where a canonical NeRF is competent at modeling. Thanks to the continuous implicit representation of NeRF and the multi-view-consistent volume rendering, we are able to produce high-fidelity images of posed hands from arbitrary viewing directions. This can not only be applied in the synthesis of free-viewpoint videos, but also help to perform data augmentation for multi-view detection and recognition tasks in computer vision, *e.g.*, sign language recognition.

Meanwhile, modeling one single hand is nowhere near enough from an application perspective. The semantics expressed by single-hand movements is quite limited. Many practical scenarios such as sign language conversations require complex interacting postures of both hands. However, handling interaction scenarios is far from trivial and still lacks exploration. Interacting hands exhibit fine-grained texture in small areas, while incompleteness of visible texture permeates the image samples due to self-occlusion and limited viewpoints. To this end, we extend the aforementioned model into a unified framework for both hands. By introducing the hand mapping and ray composition strategy into the pose-deformable NeRF, we make it possible to naturally handle interaction contacts and complement the geometry and texture in rarely-observed areas for both hands. Note that with such a design, HandNeRF is compatible with both single hand and two interacting hands.

Moreover, to ensure a correct depth relationship when rendering the hand interactions, we re-exploit the human priors and propose a low-cost depth supervision for occlusion-robust density optimization. Such strong constraint guides the model to extract accurate geometry from sparse-view training samples. Additionally, a neural feature distillation branch is designed to achieve feature alignment between a pre-trained 2D teacher and the 3D color field. By implicitly leveraging spatial contextual cues for color learning, this cross-domain distillation effectively alleviates the artifacts on the target shape and further improves the quality of the learned texture.

Our main contributions are summarized as follows:

- To the best of our knowledge, we are the first to develop a unified framework to model photo-realistic interacting hands with deformable neural radiance fields.
- We propose several elaborate strategies, including the depth-guided density optimization and the neural feature distillation, in order to effectively address practical challenges in interacting hands training and ensure high-fidelity results for novel view/pose synthesis.
- Extensive experiments on the large-scale dataset InterHand2.6M [18] show that our HandNeRF outperforms the baselines both qualitatively and quantitatively.

2. Related Work

2.1. Neural Radiance Fields (NeRF)

Recent years have witnessed the rapid development of neural implicit representations [16, 17, 22] in 3D modeling and image synthesis. Compared with classic discrete counterparts such as meshes, point clouds, and voxels, neural implicit representations model the scene with neural networks, which are spatially continuous and indicate a higher fidelity and flexibility. As the most popular implicit representation in neural rendering, Neural Radiance Fields (NeRF) [17] has exhibited stunning results in various tasks since its first introduction. The original NeRF overfits on one static scene by design, therefore it cannot model time-varying contents.

Many efforts have been made to adapt NeRF to dynamic scenes. Some works condition the NeRF with local [36, 39] or global scene representations [7, 8, 13] to implicitly provide generalizability for it. As the pioneers, D-NeRF [27] and Nerfies [23] use an explicit deformation field to bend straight rays passing through varying targets into a common canonical scene, where a conventional NeRF is optimized. Such a pipeline is adopted by many follow-up works [24, 33]. These methods provide hard constraints by sharing geometry and appearance information across time, while presenting a relatively harder optimization problem.

2.2. Neural Rendering of Articulated Objects

The image rendering of animatable articulated objects, *i.e.*, human bodies, hands, *etc*., can be regarded as a special case of modeling dynamic scenes. Most early works [15, 31] complete reconstruction using skeleton-based meshes, which generally rely on expensive calibration and massive samples to produce high-quality results.

Neural Body [26] signifies a breakthrough in low-cost human rendering by combining NeRF with the mesh-based SMPL model [15]. Neural Actor [14] optimizes the human model in a canonical space along with a volume deformation based on the linear blending skinning (LBS) algorithm of SMPL mesh. Similar LBS-based pipelines are adopted by a lot of works [11,12,25,32,37,40]. Since the LBS deformation cannot handle non-rigid transformation, other strategies have to be introduced for better rendering quality. Most methods [11, 37] regress an extra point-wise offset for samples, while some works like Animatable-NeRF [25] try to jointly optimize NeRF with the LBS weights for deformation. To this end, a forward and a backward skinning field are introduced to save LBS weights for the bidirectional mapping between the posed and canonical shapes. The main limitation here is the poor generalizability of inverse LBS since the weights vary when the pose changes [28].

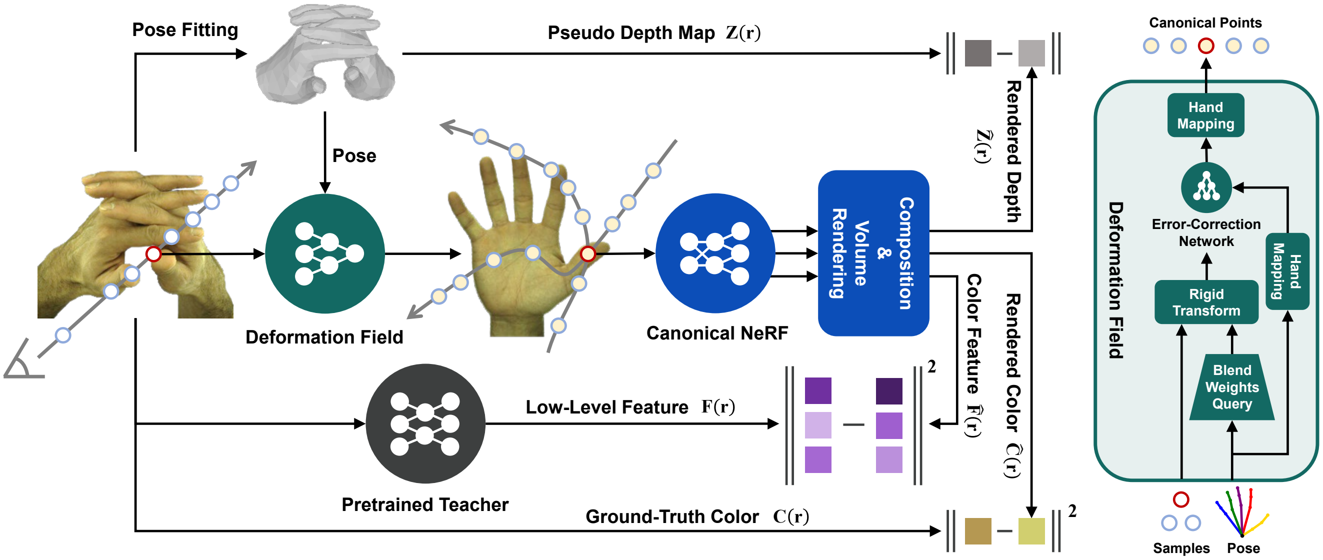
Pose Fitting
Pseudo Depth Map $\mathbf{Z}(\mathbf{r})$
Pose
Rendered Depth $\hat{\mathbf{Z}}(\mathbf{r})$
Canonical Points
Hand Mapping
Error-Correction Network
Hand Mapping
Rigid Transform
Blend Weights Query
Deformation Field
Deformation Field
Canonical NeRF
Composition & Volume Rendering
Color Feature $\hat{\mathbf{F}}(\mathbf{r})$
Rendered Color $\hat{\mathbf{C}}(\mathbf{r})$
Low-Level Feature $\mathbf{F}(\mathbf{r})$
Pretrained Teacher
Ground-Truth Color $\mathbf{C}(\mathbf{r})$
2
2
Samples
Pose

Figure 2. **Overview of HandNeRF.** A straight observing ray is warped to a canonical space by the deformation field, depending on the different poses of two hands. Colors and densities of the two sets of samples are then produced by the shared NeRF. We establish supervision for the integrated colors, color features and depth values, to help reconstruct fine-grained details of both texture and geometry.

Another series of methods [5, 20] model the human body with separate parts. They decompose an articulated object into several rigid bones, and then perform per-bone prediction with separated NeRFs. Although being good at maintaining partial rigidity, those methods struggle to merge different parts. They inevitably produce overlap or breakage between bones, and are consequently inferior to the overall modeling approaches in terms of pose generalizability. As a result, LBS-based methods are still the mainstream practice for human modeling. Aside from NeRF, some works [28] adopt neural implicit surfaces such as the signed distance field (SDF) [21, 35, 38] to better model the geometry of human body. Those methods can produce relatively smoother surface predictions, but are not good at rendering appearance with high-frequency details, unlike NeRF.

Compared with human body, the neural rendering of human hands still lacks exploration. Recently, LISA [4] is proposed as the first neural implicit model of textured hands. It is focused on the reconstruction of hand geometry using separately-optimized SDF, while the color results are barely satisfactory. Meanwhile, it suffers from similar limitations as faced by the aforementioned SDF-based and per-bone optimizing methods. Moreover, it only supports one single hand and cannot be applied in the interacting-hand scenarios that are common in practice.

3. Method

Given a set of multi-view RGB videos capturing a short pose sequence of a single hand or two interacting hands, we propose a novel framework named HandNeRF, which is intended to model the dynamic scene, enabling image rendering of novel hand poses from arbitrary viewing directions. The overview of HandNeRF is shown in Fig. 2. We disentangle the pose of both hands using a deformation field and optimize a shared canonical hand with NeRF (Sec. 3.2). To ensure the correct depth relationship when compositing two hands, we further establish depth supervision for density optimization (Sec. 3.3). Moreover, to mine useful cues from RGB images for better texture learning, we propose a feature distillation framework compatible with our efficient sampling strategy (Sec. 3.4). We will elaborate our method in the following subsections.

3.1. Preliminary: Neural Radiance Fields

We first quickly review the standard NeRF model [17] for a self-contained interpretation. Given a 3D coordinate $\mathbf{x}$ and a viewing direction $\mathbf{d}$, NeRF queries the view-dependent emitted color $\mathbf{c}$ and density σ of that 3D location using a multi-layer perceptron (MLP). A pixel color $\hat{\mathbf{C}}(\mathbf{r})$ can then be obtained by integrating the colors of N samples along a ray $\mathbf{r}$ in the viewing direction $\mathbf{d}$ using the differentiable discrete volume rendering function [17]:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) \mathbf{c}_i, \tag{1}$$

where δ_i is the distance between adjacent samples, and $T_i = \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j)$. To further obtain the multi-scale representation of a scene, Mip-NeRF [1] extends NeRF to represent the samples along each ray as conical frustums, which can be modeled by multivariate Gaussians $(\mathbf{x}, \mathbf{\Sigma})$, with $\mathbf{x}$ as the mean and $\mathbf{\Sigma} \in \mathbb{R}^{3\times 3}$ as the covariance. Thus, the density and emitted color for a sample can be given by the NeRF MLP: $(\mathbf{x}, \mathbf{\Sigma}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$.

3.2. Modeling Pose-Driven Interacting Hands

The conventional NeRF is optimized on a static scene and lacks the ability to model hands with different poses. Therefore, for pose-driven hands modeling, we introduce a pose-conditioned deformation field that warps the observing rays passing through both hands to a shared space, where a static NeRF is established for one canonical hand.

Canonical hand representation. We model the geometry and texture of hands with a neural radiance field in a pose-independent canonical space. Considering the multi-scale distribution of observers in practice, a cone-tracing architecture similar to Mip-NeRF [1] is adopted. To be specific, two MLPs denoted by $F_{\mathbf{\Theta}_\sigma}$ and $F_{\mathbf{\Theta}_c}$ output the density σ and emitted color $\mathbf{c}$ of the queried 3D sample, respectively:

$$\sigma = F_{\mathbf{\Theta}_\sigma}\left(\mathrm{IPE}\left(\mathbf{x}_{can}, \mathbf{\Sigma}\right)\right) = F_{\mathbf{\Theta}_\sigma}\left(\mathbf{f}_\sigma\right), \tag{2}$$

$$\mathbf{c} = F_{\mathbf{\Theta}_c}\left(\mathrm{PE}\left(\mathbf{d}\right), \mathbf{f}_\sigma, \ell_c\right), \tag{3}$$

where $\mathbf{x}_{can}$ is the sample coordinate in the canonical space, $\mathrm{PE}(\cdot)$ is the sinusoidal positional encoding in [17], $\mathrm{IPE}(\cdot)$ is the anti-aliased integrated positional encoding proposed by [1], and ℓ_c is a per-frame latent code to model subtle texture differences between frames. Definitions of other notations are consistent with those in Sec. 3.1.

Deformation field. Given an arbitrary hand pose, the deformation field is intended to learn a mapping from that observation space to a canonical space shared by all posed hands. Without any motion priors, it is an extremely under-constrained problem to model the deformation field as a trainable pose-conditioned coordinate transformation jointly-optimized with NeRF [13, 27]. Therefore, we follow previous works on NeRF for dynamic human body [11, 25, 37, 40] to leverage the parameterized human priors. Specifically, to establish a pose-driven deformation field, HandNeRF follows the settings of MANO [31] with the 16 hand joints, the pose parameters $\mathbf{p} \in \mathbb{R}^{16\times3}$ (axis angles at each joint), the canonical (mean/rest) pose $\overline{\mathbf{p}}$, and the blend skinning weight $\mathbf{w}_b \in \mathbb{R}^{16}$. Similar to many classic mesh-based methods, MANO uses linear blend skinning (LBS) to accomplish skeleton-driven deformation for mesh vertices. It models the coordinate transformation between poses as the accumulation of joints' rigid transformations weighted by the blend weight $\mathbf{w}_b$.

HandNeRF employs such skeleton-driven transformation as a strong prior for the deformation field. Given a pose $\mathbf{p}$ and a 3D sample $\mathbf{x}_{ob}$ from the observation space, we obtain the posed MANO mesh and query the nearest mesh facet for $\mathbf{x}_{ob}$. The queried blend weight $\mathbf{w}_b = [w_{b,1}, \dots, w_{b,16}]$ is then calculated by barycentric interpolating those of corresponding facet vertices. Thus, a coarse deformation can be expressed by

$$\hat{\mathbf{x}}_{can} = T\left(\mathbf{x}_{ob}, \mathbf{p}\right) = (\sum_{j=1}^{16} w_{b,j}\mathbf{T}_j)\mathbf{x}_{ob}, \tag{4}$$

where $\mathbf{T}_j \in \mathrm{SE}(3)$ is the observation-to-canonical rigid transformation matrix of each joint.

Due to the inevitable errors caused by the interpolation and the parameterized model itself, we introduce an additional pose-conditioned error-correction network denoted by $F_{\mathbf{\Theta}_e}$ to model the non-linear deformation as a residual term for $\hat{\mathbf{x}}_{can}$. In this way, the deformation field can capture pose-specific details beyond the mesh estimation while preserving the generalizability of the canonical hand.

To enable the complementation of geometry and texture for left and right hands in textureless or rarely-observed areas during training, we propose a unified modeling of canonical space for both hands. Since the pose parameters and canonical pose of two hands are defined differently in MANO, we introduce a hand mapping module denoted by $\psi(\cdot)$ in practice to align the left hand with the right one. Formally, the deformation field (illustrated in Fig. 2, right) can be expressed by

$$\mathbf{x}_{can} = \psi\left(\hat{\mathbf{x}}_{can} + F_{\mathbf{\Theta}_e}\left(\psi\left(\hat{\mathbf{x}}_{can}\right), \psi\left(\mathbf{p}\right)\right)\right). \tag{5}$$

Note that different from previous works [25, 32] relying on per-pose latent code to guide the deformation, we use pose representation instead, ensuring robustness to unseen poses.

Sampling and composition strategy. Based on the estimated parameterized hand mesh, it is convenient to obtain the coarse scene bounds of both 3D space and 2D image. The 2D image bounds serve as a pseudo label of the foreground mask, which guides the pixel (ray) sampling. For a high-resolution training image, we perform ray-tracing on only 1% of the pixels, mainly focusing on the foreground. Since the target hand covers only a small area of a typical image, such an unbalanced pixel sampling strategy ensures that more importance is attached to the texture of the target hands, and also significantly speeds up the training.

Meanwhile, the 3D scene bounds help to determine the near and far bounds for a camera ray, along which N 3D samples are evenly selected. In order to render two interacting hands while the canonical NeRF only models a single hand, we have to perform object composition before volume rendering. Instead of introducing an extra composition operator (*e.g.*, density-weighted mean of colors [19]), we argue that for each pixel, sampling twice within both hands' own bounds is more reasonable for non-transparent targets without clipping. Specifically, a straight observing ray is warped with two different solutions produced by the deformation field, depending on the corresponding poses of two hands. The colors and densities of the two sets of deformed samples are produced by the shared canonical NeRF, and then re-sorted based on their depth values. Finally, we integrate over all the samples belonging to the same ray using Eq. (1) and obtain the final pixel color.

3.3. Depth-Guided Density Optimization

The conventional NeRF is susceptible to visual overfitting when given insufficient training views [6]. That is, even if the scene geometry (density) fails to be correctly extracted, the rendered images from specific camera views can still be fine. However, these seemingly fine color results occur only on training views and will collapse for novel view synthesis. This will become a catastrophe in our task with sparse training views. Worse still, our composition strategy for interacting hands will exhibit poor performance without a relatively accurate geometry prediction. Obviously, the rendering quality of complex poses such as interlocking hands relies highly on a correct depth relationship.

To address this issue, we establish 2D depth supervision on the optimization of 3D density. Recent works [6, 30] introduce depth constraints to NeRF by running structure-from-motion (SFM) preprocessing to produce sparse 3D point clouds that function as depth labels. Unlike those works, we leverage the parameterized hand model estimated in Sec. 3.2 at a lower cost. Once the posed hand mesh is obtained, the depth of each pixel from a specific view is freely available as a byproduct. We then use it to build a pseudo depth map as the ground truth for the training view. Meanwhile, the pixel-wise depth estimated by NeRF can be derived with volume rendering. For N samples along a ray $\mathbf{r}$, we denote their depth values as $\{t_1, t_2, \ldots, t_N\}$. Then we integrate these values with the same weights as Eq. (1):

$$\hat{Z}(\mathbf{r}) = \sum_{i=1}^{N} T_i \left(1 - \exp\left(-\sigma_i \delta_i\right)\right) t_i, \tag{6}$$

where $\hat{Z}(\mathbf{r})$ is the estimated depth value of a specified ray.

Our objective is to minimize the difference between $\hat{Z}(\mathbf{r})$ and the target depth map $Z(\mathbf{r})$. While SFM-based depth-supervised methods aim at minimizing the KL divergence [6] or a Gaussian negative log likelihood term [30] on the depth, we deem it more reasonable to regularize the pixel-wise smooth L_1 distance in HandNeRF. That is because unlike sparse point clouds with noise, our mesh-based pseudo depth naturally maintains the surface consistency.

3.4. Neural Feature Distillation

In a conventional NeRF pipeline, the multi-view training images are only used for independent pixel-wise supervision. However, with such a vanilla training framework, artifacts and blurs can often be observed in our task for unseen views or poses with sparse training views. Besides, the model is prone to local optimum on some training sequences due to the miniature visible hand in specific views. All these phenomena call for the re-usage of training images to give attention to the spatial context of individual pixel and impose more constraints on color learning.

Unlike image-based extensions [36, 39] for NeRF that directly feed pixel features learned with a jointly-optimized feature extractor into the color fields, we adopt a more efficient and general method — neural feature distillation. Our objective is to align the 2D image features, produced by a pre-trained extractor, with the corresponding sample features defined in 3D space. Therefore, contextual cues can be implicitly introduced to the optimization of color field, owing to the receptive field of the feature extractor.

Specifically, we adopt a cross-domain student-teacher paradigm, where features of a 2D teacher network are distilled into a 3D student network. Instead of learning an extra neural feature field as in N3F [34], we make the NeRF output a color feature $\mathbf{f}_c \in \mathbb{R}^D$, where D is the number of feature channels. $\mathbf{f}_c$ is derived from the viewing direction $\mathbf{d}$ and the density feature $\mathbf{f}_\sigma$, as an intermediate product of the color field in Eq. (3). Then $\mathbf{f}_c$ of all samples along the sampled ray is integrated using volume rendering (Eq. (1)) to produce a pixel-wise feature $\hat{\mathbf{F}}(\mathbf{r})$. As for the 2D teacher network, we choose the self-supervised extractor DINO [2] built based on vision transformer. Note that other popular image feature extractors [3, 9] can also be applied in our framework. The target image feature is extracted from the second layer of the pre-trained DINO using the publicly available weights, which is meant to focus on texture details rather than high-level semantics. It is then L_2-normalized and reduced to D dimensions with PCA before distillation, yielding the target pixel feature $\mathbf{F}(\mathbf{r})$.

3.5. Training

Loss function. Following [17], the main loss for optimizing the NeRF network parameters $\mathbf{\Theta}_\sigma$ and $\mathbf{\Theta}_c$ is applied directly between the rendered pixel color $\hat{\mathbf{C}}(\mathbf{r})$ and the ground truth $\mathbf{C}(\mathbf{r})$:

$$\mathcal{L}_{rgb} = \sum_{\mathbf{r}} \|\mathbf{C}(\mathbf{r}) - \hat{\mathbf{C}}(\mathbf{r})\|_2^2. \tag{7}$$

As mentioned in Sec. 3.3, we propose an extra constraint on $\mathbf{\Theta}_\sigma$, regularizing the pixel-wise distance between the rendered depth $\hat{Z}(\mathbf{r})$ and the target pseudo depth $Z(\mathbf{r})$:

$$\mathcal{L}_{depth} = \sum_{\mathbf{r}} \mathrm{SLL}(Z(\mathbf{r}) - \hat{Z}(\mathbf{r})), \tag{8}$$

where $\mathrm{SLL}(\cdot)$ is the smooth L_1 loss.

As interpreted in Sec. 3.4, the neural distillation is performed on the color feature $\hat{\mathbf{F}}(\mathbf{r})$ to achieve cross-domain alignment:

$$\mathcal{L}_{dst} = \sum_{\mathbf{r}} \|\mathbf{F}(\mathbf{r}) - \hat{\mathbf{F}}(\mathbf{r})\|_2^2. \tag{9}$$

Note that $\mathcal{L}_{color}$, $\mathcal{L}_{depth}$ and $\mathcal{L}_{dst}$ are also back propagated to update parameters of the deformation field, $\mathbf{\Theta}_e$.

Besides, we add a regularizer for the error-correction term of each sample $\mathbf{x}$ in the deformation field (Eq. (5)), so that the non-linear deformation is minor and does not degrade the generalizability for unseen poses:

$$\mathcal{L}_{dfm} = \sum_{\mathbf{x}} \|F_{\mathbf{\Theta}_e}\left(\psi\left(\hat{\mathbf{x}}_{can}\right), \psi\left(\mathbf{p}\right)\right)\|_2. \tag{10}$$

Additionally, to mitigate the semi-transparent geometry and the misty halo around the target hand, we apply the hard surface loss similar to [29], encouraging the weight of each sample in volume rendering to be either 1 or 0:

$$\mathcal{L}_{hs} = \sum_{\mathbf{x}} -\log(e^{-|w_v|} + e^{-|1-w_v|}), \tag{11}$$

where $w_v = T_i\left(1 - \exp\left(-\sigma_i \delta_i\right)\right)$ is the weight in Eq. (1).

Moreover, we observe that on some sequences, our model is prone to a local optimum where all pixels on the target hands are converged to the same mean color. We have to impose stronger regularization for samples that are closer to the mean. To this end, a color variance loss is proposed:

$$\mathcal{L}_{cvar} = \mathrm{SLL}(\mathrm{Var}(\{\mathbf{C}\}) - \mathrm{Var}(\{\hat{\mathbf{C}}\})), \tag{12}$$

where $\mathrm{Var}(\cdot)$ calculates the biased sample variance.

Overall, the final loss is given by

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{rgb} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{dst}\mathcal{L}_{dst} + \\ & \lambda_{dfm}\mathcal{L}_{dfm} + \lambda_{hs}\mathcal{L}_{hs} + \lambda_{cvar}\mathcal{L}_{cvar}.\end{aligned} \tag{13}$$

Pose generalization and adaptation. Once trained, our model is able to produce full-resolution images for novel poses as well as novel views. Due to our design of fully manipulable pose input, rendering animatable interacting hands is as simple as feeding the desired pose parameters into HandNeRF. The mesh priors and our canonical hand model ensure the generalizability for out-of-distribution poses. Nevertheless, if the training pose sequences are too homogeneous, HandNeRF may still fail to disentangle pose-specific shapes (*e.g.*, the tense muscles) from the canonical geometry, resulting in conspicuous artifacts for novel poses. Fortunately, our framework can be conveniently modified into a fine-tuning pipeline for pose adaptation. Specifically, we disable the feature distillation branch, freeze the parameters of NeRF, and fine-tune the deformation field. Only the depth and the deformation loss are used in this stage. No ground-truth RGB image is needed, as the depth map can be derived directly from pose parameters. After pose adaptation on a few samples, the rendered images will have much fewer artifacts and geometric errors.

4. Experiments

4.1. Experimental Settings

Dataset and preprocessing. HandNeRF is trained on the 30FPS version of Interhand2.6M [18] that contains large-scale multi-view sequences of various hand poses. Each sequence contains images (512 × 334 px) of a single hand or interacting hands from dozens of views. 18 common views are selected as the test views. Intra-sequence test is to evaluate the novel view synthesis quality, while cross-sequence test is to evaluate the novel pose rendering quality.

Baselines. HandNeRF is the first NeRF model designed for photo-realistic novel view/pose image synthesis of interacting hands, thus no method is available for direct comparison. Therefore, we develop three baselines inspired by works that explore NeRF for human body. 1) Pose-NeRF: we modify Mip-NeRF [1] to learn a NeRF conditioned on pose; 2) Ani-NeRF: we adapt [25] to the setup of human hands; 3) NeuMan: we re-implement the "Human NeRF" module of [11] on the settings of hands while preserving its various training losses. We do not include LISA [4] because its source code and customized datasets are unavailable for a fair comparison. Since all the above baselines are for one single articulated object only, we extend them with the proposed composition strategy in HandNeRF to integrate two independent canonical models for both hands.

Metrics. Following previous works, we evaluate the synthesized results with peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS). To show the effect of our proposed depth supervision, we additionally provide the average L_1 error for the rendered depth map (DE), representing the quality of geometric reconstruction to some extent.

4.2. Comparison Results

Tab. 1 and 2 summarize the performance of HandNeRF and the baselines. Qualitative results are exhibited in Fig. 3.

Novel view synthesis. We train the model on a single sequence with 4, 7, or 10 views to show the effect of view quantity. As presented in Tab. 1, our method outperforms all the baselines across all metrics. Notably, training a model only with interacting hands samples is a non-trivial task, since it involves self-occlusion, incompleteness of visible texture, and subtle contacts during interaction. Therefore, the superiority of our proposed unified modeling can be evidently observed from the results. Even trained with extremely sparse views, HandNeRF can still achieve 29dB for interacting hands in terms of PSNR. For comparison, NeuMan [11] fails to converge properly on interacting hands, rendering mask-like textureless images. Similar issue also arises during the training of HandNeRF, but we manage to resolve it with the proposed color variance loss. Besides, we can observe some semi-transparent mist floating around the rendered hand in some methods' results (Fig. 3), which proves the effectiveness of the proposed depth-guided density optimization in HandNeRF.

Novel pose synthesis. We first directly test the learned model on an unseen sequence. Then we apply pose adaptation for those novel poses and re-test the performance. Three different tasks are included, where the learned model of single hand or interacting hands is generalized or adapted to both hand types. Obviously, it is most challenging to render novel poses for interacting hands when the learned model is also trained on interacting hands. As shown in

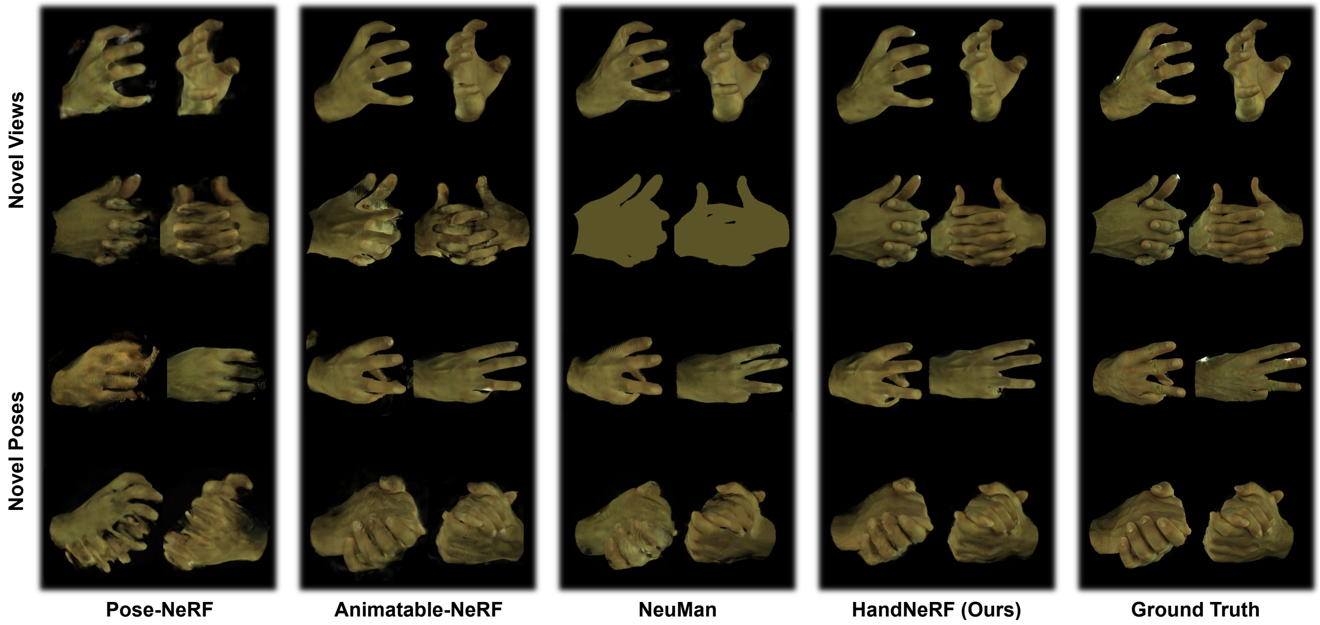

Figure 3. **Qualitative performance comparison.** We present the results of both novel view rendering (first two rows) and novel pose adaptation (last two rows). All models are trained with 10 different camera views. Pose adaptation results for interacting hands (last row) are produced with models pre-trained on one single hand.

	4 views				7 views				10 views			
	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
						Single hand						
Pose-NeRF	27.0855	0.9355	0.0921	0.1517	29.2643	0.9301	0.0704	0.1855	29.2126	0.9397	0.0739	0.1910
Ani-NeRF	30.2606	0.9589	0.0704	0.1700	31.6422	0.9632	0.0581	0.1623	31.7784	0.9684	0.0621	0.1582
NeuMan	30.3428	0.9596	0.0691	0.1685	31.2364	0.9623	0.0573	0.1617	31.8419	0.9702	0.0552	0.1507
Ours	**31.0493**	**0.9655**	**0.0588**	**0.1278**	**31.8556**	**0.9691**	**0.0459**	**0.1238**	**32.7036**	**0.9742**	**0.0375**	**0.1210**
						Interacting hands						
Pose-NeRF	25.0193	0.8745	0.1873	0.2604	27.2416	0.9014	0.1381	0.2464	27.6461	0.9162	0.1071	0.2312
Ani-NeRF	28.0323	0.9414	0.0865	0.2260	28.8543	0.9440	0.0841	0.2187	29.3577	0.9491	0.0798	0.2118
NeuMan	×	×	×	×	×	×	×	×	×	×	×	×
Ours	**29.0351**	**0.9555**	**0.0841**	**0.1861**	**30.0691**	**0.9624**	**0.0818**	**0.1863**	**30.7571**	**0.9568**	**0.0724**	**0.1864**

Table 1. **Performance comparison on novel view synthesis.** "×" means the model does not converge properly on one or more training sequences. Our method achieves the best rendering quality across all scenes, even only trained with extremely sparse views.

Tab. 2, HandNeRF gives the best performance in both pose generalization and further adaptation for all tasks. Note that the training texture details (usually different from test hands) are preserved in novel pose synthesis, and we do not further optimize color in pose adaptation. Therefore, it is not surprising that pixel-wise metrics like PSNR drop significantly for novel poses. As a perceptual metric, LPIPS is considered to be more meaningful here.

4.3. Ablation Study

We conduct ablative experiments to validate the effectiveness of two essential components in our HandNeRF: the depth-guided density optimization and the neural feature distillation. The results are listed in Tab. 3. We also provide more details of the synthesized images in Fig. 4.

Depth supervision. We ablate our depth supervision and compare it with GNLL (Gaussian negative log likelihood) [30]. We observe dramatic performance degradation and blurred interfacial areas without depth guidance. Although overfitting to training views, the model fails to infer correct depth when rendering from novel views, let alone compositing both hands for unseen poses. This can be further proved by its noisy depth map. Besides, as mentioned in Sec. 3.3, GNLL is more suitable for depth values produced by noisy point clouds. When applied to our mesh-based depth map, it leads to artifacts near hand geometry.

Neural distillation. We replace the image features produced by our pre-trained teacher with the same-shaped random vectors. The results in Tab. 3 show that HandNeRF actually exploits the texture information from the low-level features. The performance improvement does not come from the effect of regularization.

	PSNR ↑	SSIM ↑	LPIPS ↓	DE ↓
Single hand → Single hand				
Pose-NeRF	23.0118 / —	0.8959 / —	0.1454 / —	0.1985 / —
Ani-NeRF	— / 25.0533	— / 0.9317	— / 0.0742	— / 0.1596
NeuMan	25.0254 / 25.8456	0.9258 / 0.9324	0.0955 / 0.0608	0.1605 / 0.1321
Ours	**26.5088 / 27.9717**	**0.9345 / 0.9532**	**0.0911 / 0.0576**	**0.1435 / 0.1279**
Single hand → Interacting hands				
Pose-NeRF	21.1971 / —	0.8344 / —	0.1959 / —	0.2137 / —
Ani-NeRF	— / 23.9512	— / 0.9218	— / 0.0934	— / 0.1800
NeuMan	24.0815 / 24.9451	0.9104 / 0.9283	0.1203 / 0.0951	0.1766 / 0.1626
Ours	**25.4666 / 26.5207**	**0.9180 / 0.9348**	**0.1162 / 0.0897**	**0.1652 / 0.1601**
Interacting hands → Interacting hands				
Pose-NeRF	19.8561 / —	0.8468 / —	0.2321 / —	0.1954 / —
Ani-NeRF	— / 23.0223	— / 0.8928	— / 0.1465	— / 0.2221
NeuMan	× / ×	× / ×	× / ×	× / ×
Ours	**23.6411 / 24.8599**	**0.8945 / 0.9152**	**0.1315 / 0.0858**	**0.1835 / 0.1802**

Table 2. **Performance (generalization / adaptation) comparison on novel pose synthesis.** "—" means the method is inapplicable for that setup. "×" means the model does not converge properly on previous training. The hand types on both sides of "→" indicate the NeRF training samples and novel pose samples. Since Ani-NeRF [25] cannot directly generalize to unseen poses, we report its pose adaptation performance after re-training with blend weight consistency. Due to the local optimum results of NeuMan [11] on interacting hands, we exclude it in those comparisons.

	PSNR ↑	SSIM ↑	LPIPS↓	DE ↓
w/o $\mathcal{L}_{depth}$	30.1057	0.9528	0.0755	0.2106
w/ GNLL	30.4304	0.9552	0.0845	0.1852
Ours	**30.9256**	**0.9570**	**0.0700**	**0.1840**
w/o distillation	32.8421	0.9720	0.0488	0.1361
random distillation	32.7892	0.9712	0.0506	0.1361
w/ CNNRenderer	31.6366	0.9703	0.0493	0.1362
w/ TransRenderer	31.3229	0.9680	0.0479	0.1364
Ours	**33.0204**	**0.9737**	**0.0475**	**0.1360**

Table 3. **Ablation study.** Experiments about depth supervision are performed on interacting hands, while the results of neural distillation and neural renderer are produced on one single hand.

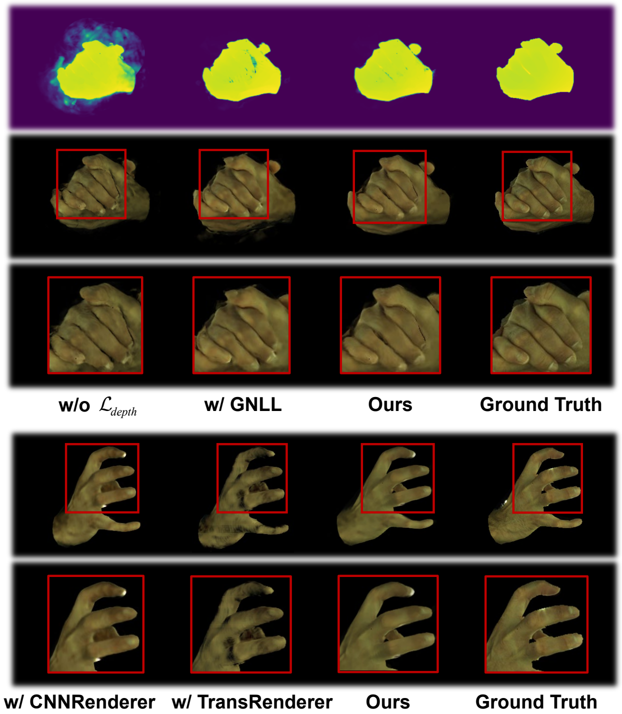

Figure 4. **Visualization of ablation study.** We exhibit rendering results and zoomed-in details of ablations for depth supervision (upper) and neural renderer (lower). Additionally, the rendered depth maps are shown in the first row.

Neural renderer. To get better rendering results, some works [10, 19] propose a neural renderer alongside the conventional volume rendering. Technically, they increase the number of channels of emitted color to model the more expressive color features with NeRF, which are integrated using volume rendering to produce a feature map. Then 2D neural networks are adopted to render the final RGB image. However, due to our efficient ray sampling strategy (Sec. 3.2), only a few sampled pixels are available during training, resulting in an incomplete feature map. Since full-resolution ray-tracing has an unacceptable overhead, a compromised solution is to produce a much smaller feature map and perform upsampling in the neural renderer, at a cost of NeRF's expressive power, especially for high-frequency details on small targets like hands. To compare with those neural renderers, we follow [10] to feed a low-resolution 2D feature map into a modified version of its neural rendering module composed of CNN and upsampling layers. We also develop a TransRenderer that uses the transformer for pixel-wise neighborhood attention. It can be observed in Fig. 4 that the CNNRenderer produces a visually smoother image but degrades the quantitative results, while the TransRenderer tends to fuse hand skin with background noise.

5. Conclusion

In this paper, we propose HandNeRF, a novel framework that reconstructs photo-realistic appearance and geometry of single or interacting hands with pose-deformable neural radiance fields. By developing several elaborate strategies including depth-guided density optimization and neural feature distillation, our method can effectively handle non-trivial challenges in complex hand interactions (*e.g.*, self-occlusion and invisible texture). We thereby enable the rendering of high-fidelity images and videos for gesture animation from arbitrary views. Comprehensive experiments on the large-scale InterHand2.6M dataset demonstrate the superiority of our approach.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Contract U20A20183 and 61836011, and in part by the Fundamental Research Funds for the Central Universities under contract WK3490000007. It was also supported by the GPU cluster built by MCC Lab of Information Science and Technology Institution and the Supercomputing Center of the USTC.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. In *ICCV*, pages 5855–5864, 2021. 3, 4, 6

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 5

[3] Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *ICCV*, pages 9640–9649, 2021. 5

[4] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. LISA: Learning implicit shape and appearance of hands. In *CVPR*, pages 20533–20543, 2022. 3, 6

[5] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. NASA neural articulated shape approximation. In *ECCV*, pages 612–628. Springer, 2020. 3

[6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2022. 5

[7] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *ICCV*, pages 14304–14314, 2021. 2

[8] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, pages 5712–5721, 2021. 2

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pages 16000–16009, 2021. 5

[10] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. HeadNeRF: A real-time NeRF-based parametric head model. In *CVPR*, pages 20374–20384, 2022. 8

[11] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. In *ECCV*, 2022. 1, 2, 4, 6, 8

[12] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. TAVA: Template-free animatable volumetric actors. In *ECCV*, 2022. 2

[13] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, pages 6498–6508, 2021. 2, 4

[14] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG*, 40(6):1–16, 2021. 2

[15] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 34(6):1–16, 2015. 2

[16] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 2

[17] 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 *ECCV*, pages 405–421. Springer, 2020. 1, 2, 3, 4, 5

[18] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. InterHand2.6M: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *ECCV*, 2020. 2, 6

[19] Michael Niemeyer and Andreas Geiger. GIRAFFE: Representing scenes as compositional generative neural feature fields. In *CVPR*, pages 11453–11464, 2021. 4, 8

[20] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pages 5762–5772, 2021. 3

[21] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 3

[22] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2

[23] 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 *ICCV*, pages 5865–5874, 2021. 2

[24] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG*, 40, 2021. 2

[25] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *ICCV*, pages 14314–14323, 2021. 1, 2, 4, 6, 8

[26] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, pages 9054–9063, 2021. 1, 2

[27] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural radiance fields for dynamic scenes. In *CVPR*, pages 10318–10327, 2021. 2, 4

[28] Shenhan Qian, Jiale Xu, Ziwei Liu, Liqian Ma, and Shenghua Gao. UNIF: United neural implicit functions for clothed human reconstruction and animation. In *ECCV*, 2022. 2, 3

[29] Daniel Rebain, Mark Matthews, Kwang Moo Yi, Dmitry Lagun, and Andrea Tagliasacchi. LOLNeRF: Learn from one look. In *CVPR*, pages 1558–1567, 2022. 6

[30] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, pages 12892–12901, 2022. 5, 7

[31] Javier Romero, Dimitris Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 36(6), 2017. 1, 2, 4

[32] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, volume 34, pages 12278–12291, 2021. 2, 4

[33] 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 *ICCV*, pages 12959–12970, 2021. 2

[34] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. In *3DV*, 2022. 5

[35] 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. In *NeurIPS*, 2021. 3

[36] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. IBRNet: Learning multi-view image-based rendering. In *CVPR*, pages 4690–4699, 2021. 2, 5

[37] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, pages 16210–16220, 2022. 2, 4

[38] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 3

[39] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 2, 5

[40] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Efficiently generated human radiance field from sparse inputs. In *CVPR*, pages 7743–7753, 2021. 2, 4