3D Visibility-Aware Generalizable Neural Radiance Fields for Interacting Hands

Xuan Huang[1*], Hanhui Li[1*], Zejun Yang[2], Zhisheng Wang[2], Xiaodan Liang[1, 3†]
[1]Shenzhen Campus of Sun Yat-sen University, Shenzhen, China
[2]Tencent, Shenzhen, China
[3]DarkMatter AI Research, Guangzhou, China
lihh77@mail.sysu.edu.cn, xdliang328@gmail.com

Abstract

Neural radiance fields (NeRFs) are promising 3D representations for scenes, objects, and humans. However, most existing methods require multi-view inputs and per-scene training, which limits their real-life applications. Moreover, current methods focus on single-subject cases, leaving scenes of interacting hands that involve severe inter-hand occlusions and challenging view variations remain unsolved. To tackle these issues, this paper proposes a generalizable visibility-aware NeRF (VA-NeRF) framework for interacting hands. Specifically, given an image of interacting hands as input, our VA-NeRF first obtains a mesh-based representation of hands and extracts their corresponding geometric and textural features. Subsequently, a feature fusion module that exploits the visibility of query points and mesh vertices is introduced to adaptively merge features of both hands, enabling the recovery of features in unseen areas. Additionally, our VA-NeRF is optimized together with a novel discriminator within an adversarial learning paradigm. In contrast to conventional discriminators that predict a single real/fake label for the synthesized image, the proposed discriminator generates a pixel-wise visibility map, providing fine-grained supervision for unseen areas and encouraging the VA-NeRF to improve the visual quality of synthesized images. Experiments on the Interhand2.6M dataset demonstrate that our proposed VA-NeRF outperforms conventional NeRFs significantly. Project Page: https://github.com/XuanHuang0/VANeRF.

Introduction

Recent progress in neural radiance fields (NeRFs) (Mildenhall et al. 2021; Gao et al. 2022b; Niemeyer et al. 2022; Johari, Lepoittevin, and Fleuret 2022a) is promising, as the continuous implicit representation of NeRFs can be disentangled with spatial volume resolution and generate high-fidelity results. This facilitates interesting research such as human avatars (Jiang et al. 2022), text-to-3d generation (Poole et al. 2022), and large-scale 3D urban scene modeling (Turki, Ramanan, and Satyanarayanan 2022).

However, to synthesize high-quality images, current NeRFs require dozens of well-calibrated multi-view images and hours of per-scene optimization. Although several methods (Müller et al. 2022; Chen et al. 2022a) have been proposed to reduce the computational cost, it is still expensive and difficult to apply NeRFs in real-life applications where only single-view inputs are available.

Generalizable NeRFs (Mihajlovic et al. 2022; Kwon et al. 2021) that represent geometry and textures separately seem to be a potential solution for the above issues. Nevertheless, these methods are designed for human bodies and cannot be applied to interacting hands directly. This is because, unlike single-subject human avatars, interacting hands involve challenging factors like severe self/inter-hand occlusions and large view variations (Park et al. 2022; Deng et al. 2022b). These factors make it difficult to exploit reliable features and cause artifacts that cannot be ignored.

Therefore, to fulfill application needs and tackle the above challenging factors, this paper aims to design a single-image generalizable NeRF model for interacting hands. In our early exploration, we find that (Mihajlovic et al. 2022) retains textures and details well but is sensitive to view variations and occlusions, since it adopts the pixel-aligned feature representation (Saito et al. 2019). (Kwon et al. 2021) uses a global feature representation that maintains the overall hand structures but is hard to generate fine-grained textures. This indicates that self-adaptive and robust features are the key to constructing feasible NeRFs for interacting hands.

Particularly, we propose a visibility-aware NeRF framework (denoted as VA-NeRF), of which the core is to leverage the visibility of 3D points. Such an idea is natural, because if a 3D point is visible, then its corresponding feature is more reliable. Otherwise, we should select other related points or global features for reference and information complement. Formally, the proposed method achieves this via a visibility-aware feature fusion module that determines the feature of a 3D query point not only by its visibility, but also by vertices selected from hand meshes and by global features. In this way, we can overcome the limitations of feature representation in previous methods.

Moreover, we also propose a visibility-guided adversarial learning strategy to further enhance the synthesized images of VA-NeRF. This is motivated by our observation that the quality of invisible areas in synthesized images is usually worse than that of visible areas. Hence, we propose to encourage the NeRF model to refine invisible areas. However,

[*]These authors contributed equally.

[†]Xiandan Liang is the corresponding author.

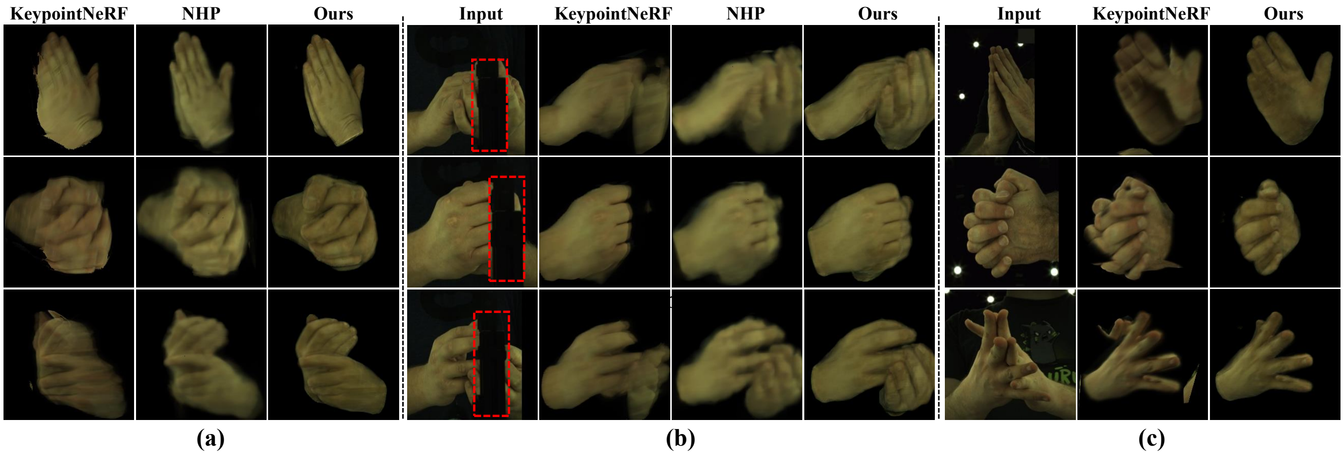

Figure 1: Compared with previous generalizable NeRFs, our visibility-aware NeRF not only (a) generates images of better quality, but also tackles challenging tasks such as (b) inpainting obstructed areas and (c) removing hands in interacting scenes.

conventional binary-class discriminators can only provide global supervision by classifying an image as real or fake. Therefore, we design a discriminator that predicts pixel-wise conditional visibility maps, so that it can provide localized and fine-grained supervision for our NeRF model.

With the above feature fusion module and adversarial learning strategy, the proposed VA-NeRF can synthesize images of interacting hands effectively, and achieve state-of-the-art performance that is validated by experiments on the Interhand2.6M dataset (Moon et al. 2020). In addition, as shown in Figure 1, our VA-NeRF can accomplish tasks that are challenging for conventional methods, such as removing hands and recovering invisible areas, and consequently benefit downstream applications like hand pose estimation (Meng et al. 2022).

In summary, the contributions of this paper can be listed as follows:

- To the best of our knowledge, this paper proposes the first single-image generalizable neural radiance field model for interacting hands.
- A visibility-aware feature fusion module is proposed, which adaptively leverages various visual features (global features, pixel-wise aligned features, and symmetric hand features) to tackle challenging occlusions and view variations.
- An adversarial learning strategy guided by visibility maps is introduced, which further improves the visual quality of synthesized two-hand images.

Related Work

Neural radiance fields. In recent years, NeRF (Mildenhall et al. 2021) has been widely studied in the area of 3D human reconstruction due to its stunning results. Many efforts have been made to adapt NeRFs to high-fidelity novel-view synthesis of human performers/avatars (Deng et al. 2022a; Mildenhall et al. 2022; Johari, Lepoittevin, and Fleuret 2022b; Niemeyer et al. 2022; Johari, Lepoittevin, and Fleuret 2022a; Martin-Brualla et al. 2021). (Raj et al. 2021; Wang et al. 2021; Yu et al. 2021) propose to utilize pixel-aligned features to learn generalized models from sparse views. Besides, KeypointNeRF (Mihajlovic et al. 2022) proposes to encode relative spatial 3D information with sparse 3D key points as references. Recent approaches (Peng et al. 2021; Kwon et al. 2021) have incorporated parametric hand meshes as geometry priors to reduce the dependence on multi-view captures. NHP (Kwon et al. 2021) applies a temporal transformer and a multi-view transformer to fuse visual features conditioned on SMPL (Loper et al. 2015). For a smoother surface prediction and better 3D consistency, (Or-El et al. 2022; Hong et al. 2023; Corona et al. 2022) merge implicit representations into the density prediction procedure of NeRF. Especially, (Corona et al. 2022) is designed for single-hand images. A concurrent approach that aims at NeRF for hands is proposed in (Guo et al. 2023), yet it requires multi-view inputs while we focus on monocular single-image scenes. SHERF (Hu et al. 2023) recovers animatable 3D humans from a single input image by extracting 3D-aware hierarchical features, including global, point-level, and pixel-aligned features, to facilitate informative encoding. The proposed method differs from SHERF in the aspect of tasks and feature fusion modules. SHERF combines features via self-attention while our module is conditioned on visibility.

Unlike previous methods, our VA-NeRF is designed for single-view generalizable interacting-hand image synthesis. Moreover, conventional generalizable methods rely on the single local/global representation and hence suffer from occlusions and view variations in our task. On the contrary, our VA-NeRF exploits visibility in both feature fusion and adversarial learning, which helps to improve the visual quality of synthesized images significantly.

3D hand reconstruction. Parametric hand models such as MANO (Romero, Tzionas, and Black 2022) have enabled hand mesh to be reconstructed via inferring a set of pose and shape parameters. Prior approaches belonging to this type (Zhang et al. 2021; Chen et al. 2021; Kulon et al. 2020; Zhou

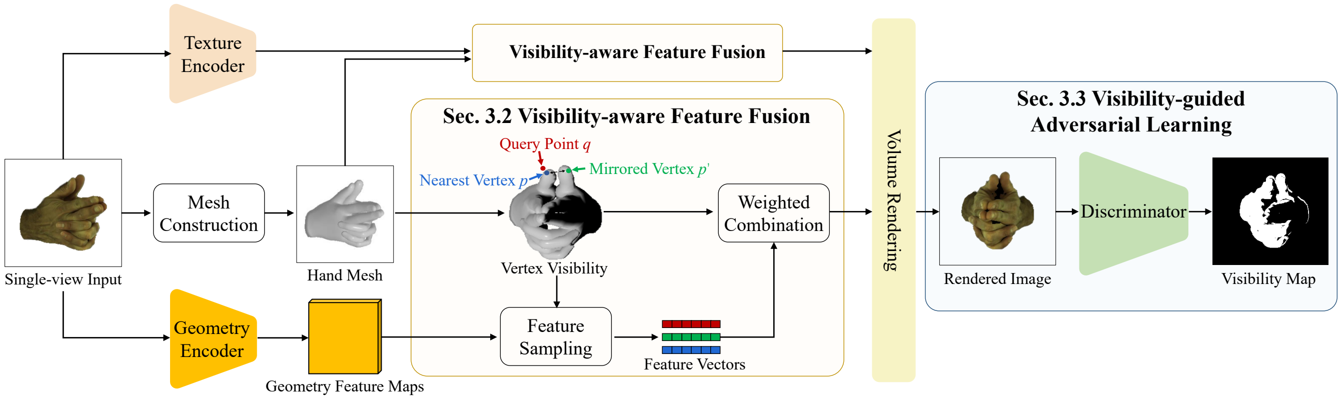

Figure 2: The framework of VA-NeRF. It consists of two key components and both of them are designed to leverage the visibility of 3D points. The first one is the visibility-aware feature fusion module that estimates appropriate features for query points, while the second one is the visibility-guided adversarial learning strategy that is used to enhance synthesized results.

et al. 2020) can achieve parameter fitting on single images. Moreover, the MANO-HD model (Chen, Wang, and Shum 2023) is developed as a high-resolution mesh topology to fit personalized hand shapes and generate smooth geometry. Implicit representations (Chen et al. 2022b; Karunratanakul et al. 2021), such as signed distance fields, have received considerable attention recently, as theoretically, they can approximate any geometry details. A dataset that consists of diverse textures and hand accessories is introduced in (Gao et al. 2022a).

3D-aware adversarial learning. Recent years have witnessed the great success of generative adversarial networks (GANs) (Goodfellow et al. 2020) in photorealistic image generation. While GANs based on 2D latent space lack 3D understanding, 3D generative models (Schwarz et al. 2020; Chan et al. 2021; Or-El et al. 2022) enable explicit camera control and render more realistic images from random viewpoints. (Hong et al. 2023; Deng et al. 2022c; Niemeyer and Geiger 2021; Xu et al. 2021) combine implicit NeRF with GAN for better view-consistency and more detailed 3d shape. To reduce the computational cost, EG3D (Chan et al. 2022) proposes an efficient tri-plane structure while EVA3D (Hong et al. 2023) divides the human body into local parts and omits unnecessary computations in blank space.

Methodology

The goal of this paper is to construct a single-view generalizable NeRF for interacting hands and to tackle occlusions and view variations. To this end, we propose a visibility-aware NeRF framework that leverages the visibility of query points and hand mesh vertices through a feature fusion module and an adversarial learning strategy. These two approaches serve distinct purposes: the former aims to infer complementary features for occluded regions, while the latter seeks to enhance the quality of rendered results.

VA-NeRF Framework

Problem formulation. Our task is to construct a generalizable NeRF that can render novel views of an arbitrary pair of interacting hands given a single input image. Specifically, the NeRF can be formulated as a function $f : (q, d, I) \rightarrow (c, \sigma)$, where $q \in \mathbb{R}^3$ is a query point, $d \in \mathbb{R}^3$ denotes a viewing direction, and I is the input image. The output $c \in \mathbb{R}^3$ and $\sigma \in \mathbb{R}$ are the color and volume density of the query point, respectively. With query points densely sampled in a 3D volume, we infer their colors and densities and synthesize a target-view image by volume rendering (Mildenhall et al. 2021).

To complete the above task, we propose the VA-NeRF framework as shown in Fig. 2. As the core of VA-NeRF is to leverage the visibility of query points and mesh vertices, we first construct hand meshes by fitting the MANO parametric model (Romero, Tzionas, and Black 2017) to the input image. Following (Mihajlovic et al. 2022), we employ two encoder branches to disentangle geometric and textual features: an hourglass network (Newell, Yang, and Deng 2016) serves as the geometry encoder and a convolutional neural network (CNN) with residual connections (Johnson, Alahi, and Fei-Fei 2016) serves as the texture encoder. Subsequently, we introduce a visibility-aware feature fusion (VAFF) module in each feature branch to obtain the visibility-enhanced feature of the query point. We utilize a multilayer perception (MLP) to infer the color c from the texture feature. For the density σ, we follow (Hong et al. 2023) to estimate a deviated signed distance field (SDF) with respect to the hand meshes:

$$\sigma(q) = w^{-1}\mathrm{sig}(-(s(q) + \delta(q))/w), \tag{1}$$

where $w \in \mathbb{R}$ is a weighting parameter optimized along with the network. $\mathrm{sig}(\cdot)$ represents the sigmoid function. $s(q) \in \mathbb{R}$ is the explicit SDF value of the query point calculated with the mesh surfaces as the zero level-set, while $\delta(q) \in \mathbb{R}$ is the deviation inferred by another MLP that takes the geometry feature of q as input. The target-view image is then generated by using an off-the-shelf differentiable renderer (Wang et al. 2021).

In addition, our VA-NeRF network is optimized using an adversarial learning strategy that exploits visibility to provide additional supervision. Our adversarial learning strat-

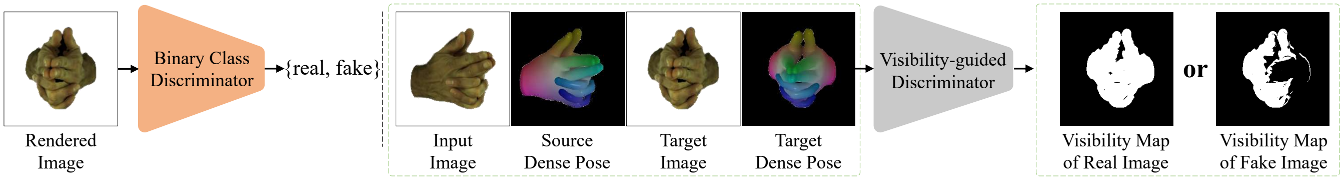
Rendered Image
Binary Class Discriminator
{real, fake}
Input Image
Source Dense Pose
Target Image
Target Dense Pose
Visibility-guided Discriminator
Visibility Map of Real Image
or
Visibility Map of Fake Image

Figure 3: Comparison between the traditional binary-class discriminator (left) and the proposed visibility-guided discriminator (right). Note that the visibility map is conditioned on the input view and whether the target image is real or synthesized.

egy relies on a discriminator that learns not only to distinguish between real and synthesized images but also to predict pixel-wise visibility maps conditioned on input and target views. Correspondingly, through the competition between the VA-NeRF network and the discriminator, we not only require the VA-NeRF network to synthesize high-fidelity images, but also encourage it to improve the quality of invisible areas in results (in order to "deceive" the discriminator). Details of our network architecture are available in the supplemental material on our project page.

Visibility-aware Feature Fusion

Traditional approaches that rely solely on either pixel-aligned or global features are ineffective in addressing the challenges caused by heavy occlusions and complex interacting poses, as these features are hard to remain reliable under such circumstances. To overcome this limitation, we introduce the VAFF module, which adaptively selects and combines a set of features for each query point.

To elaborate, we take the VAFF module in the texture feature branch as an example. For the sake of presentation conciseness, we assume that the dimensions of all features for fusion are D. Given a query point q, we locate its nearest neighboring mesh vertex p on one hand. We assume that both hands have the same topology (i.e., vertices are ordered following the MANO model) and consider the vertex with the same order on the other hand as the mirrored point p'. Note that the above topological assumption does not impose the constraint that both hands are of the same pose. q, p, and p' are projected onto the 2D image plane, and their corresponding feature vectors are retrieved via bilinear interpolation on the texture feature maps. Let $k(q), m(p), n(p') \in \mathbb{R}^D$ denote the retrieved feature vectors of q, p, and p', respectively. We consider the following weighted concatenation $t(q) \in \mathbb{R}^{6D}$ as the enhanced feature of q:

$$\begin{aligned} t(q) = [a_\varphi \varphi(q),\ a_k k(q),\ a_m m(p), \\ a_n n(p'),\ a_g^l \mathbf{g}^l,\ a_g^r \mathbf{g}^r], \end{aligned} \tag{2}$$

where $a_\varphi, a_k, a_m, a_n, a_g^l, a_g^r \in [0, 1]$ are feature weights. $\varphi(q) \in \mathbb{R}^D$ is the spatial feature of q obtained by positional encoding. $\mathbf{g}^l, \mathbf{g}^r \in \mathbb{R}^D$ are the global average texture features of the left hand and right hand.

To ensure that $t(q)$ can select and fuse appropriate features, we define a weighting function governed by 3D point visibility. Particularly, let $a \in \{a_\varphi, a_k, a_m, a_n, a_g^l, a_g^r\}$ be an arbitrary weight and $v(p, d) \in \{0, 1\}$ denote the visibility of point p from the viewing direction d. $v(p, d) = 1$ if p is visible, otherwise $v(p, d) = 0$. We calculate a via a function $\mu : \mathbb{R}^{6D+3} \to [0, 1]$ as follows:

$$\begin{aligned} a = \mu(v(q, d),\ v(p, d),\ v(p', d),\ \varphi(q), \\ k(q),\ m(p),\ n(p'),\ \mathbf{g}^l,\ \mathbf{g}^r). \end{aligned} \tag{3}$$

We implement μ simply by a MLP. The architecture of the VAFF module in the geometry feature branch is similar, except that it does not contain the global average of geometry feature maps (since geometry information is rather local).

Discussion. Our VAFF is based on the assumption that the feature vector of a visible point should be assigned a high weight. Additionally, we also consider p because spatially closed points tend to share similar features. As for p', it is included due to the structural symmetry of human hands. p and p' together facilitate the exploitation of both local and long-range feature dependencies, so to tackle pose and view variations. In cases where q, p, and p' are all invisible, the global average feature vectors can still serve as a coarse approximation to the texture feature of q. This is feasible as most areas of the hands have similar textures. Consequently, the VAFF module effectively leverages the strengths of both local pixel-aligned features for detail preservation and global features for hand structure preservation.

Visibility-guided Adversarial Learning

Due to the lack of information, the visual quality of invisible areas in synthesized images tends to be lower than that of visible areas. Therefore, here we propose the visibility-guided adversarial learning (VGAL) strategy to facilitate quality improvement in invisible areas.

To begin with, a discriminator that can identify invisible areas is necessary. This can be achieved by conducting hand mesh rasterization to generate ground-truth visibility maps for supervised learning. Specifically, given an input-view image I and a target-view image I_t, where I and I_t are of size $H \times W$, the discriminator can be formulated as $\Phi(I, I_t) : \mathbb{R}^{H \times W} \times \mathbb{R}^{H \times W} \to \{0, 1\}^{H \times W}$. It is important to note that I_t can be either real or synthesized and the ground-truth visibility map V_t should be conditioned on this. We define the following two criteria for generating V_t:

(i) If I_t is real, we consider the foreground (hand areas) of I_t to be visible,

(ii) otherwise V_t is rendered with vertex visibility computed in the input view.

An example of V_t is shown in Figure 3. With these criteria, the discriminator needs to recognize invisible areas in

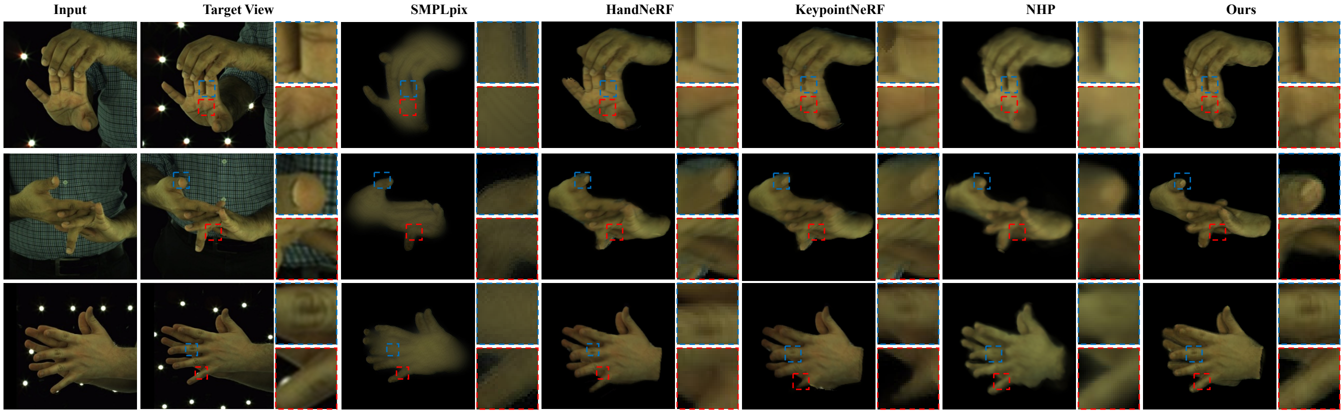

Figure 4: Visual comparison of the proposed method against state-of-the-art methods. Results of the proposed method better preserve hand structures and textures.

synthesized images while the VA-NeRF network is encouraged to generate results that have visibility maps similar to those of real target-view images.

We implement Φ by a CNN with the sigmoid function as its last activation function. Except I and I_t, we also generate dense correspondence maps (Güler, Neverova, and Kokkinos 2018) as the auxiliary inputs to Φ, which are used to provide structural priors of human hands. Finally, the objective functions of our VGAL are defined as follows:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{VGG}\mathcal{L}_{VGG} + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{vis}\mathcal{L}_{vis}, \quad (4)$$

where λ_{rgb}, λ_{VGG}, λ_{adv} and λ_{vis} are user-defined loss weights. $\mathcal{L}_{rgb}$ and $\mathcal{L}_{VGG}$ are the $l1$ loss and the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) between the target image and the synthesized image. $\mathcal{L}_{adv}$ is the non-saturating GAN loss (Mescheder, Geiger, and Nowozin 2018; Hong et al. 2023) widely used in adversarial learning. $\mathcal{L}_{vis}$ is introduced to supervise visibility learning, which is formulated as the pixel-wise binary cross entropy between the predicted visibility map V and V_t as follows:

$$\mathcal{L}_{vis}(V, V_t) = -(V_t \odot \log V + (1 - V_t) \odot \log(1 - V)), \quad (5)$$

where $\odot$ denotes the element-wise dot product. We follow the common practice of adversarial learning (Mescheder, Geiger, and Nowozin 2018) to optimize the VA-NeRF network and the discriminator alternately with their corresponding loss terms. Take $\mathcal{L}_{vis}$ as an example, during the training phase of the VA-NeRF network, $\mathcal{L}_{vis}$ is calculated with V_t conditioned on the real target image, and its gradients are propagated backward from the discriminator to the VA-NeRF network; while at the training phase of the discriminator, $\mathcal{L}_{vis}$ is calculated with V_t determined by whether the target image is real or rendered.

Experiments

In this section, we validate the effectiveness of the proposed VA-NeRF method via extensive experiments. Due to the page limitation, interested readers can refer to the supplemental material for more implementation details, experimental results, and discussions.

Setup

Dataset. Our experiments are conducted on the large-scale Interhand2.6M (Moon et al. 2020) dataset that consists of single and interacting hand images with various subjects, poses, and views. As the scope of this paper is to construct NeRFs for interacting hands, we select a subset of images on Interhand2.6M, which contains 143,893 training images and 9,475 test images in total. We further cropped out all hand regions based on the bounding boxes provided by the dataset and resize all images to 256×256.

Implementation details. Our network is implemented using PyTorch and trained with the Adam optimizer (Kingma and Ba 2014) with a batch size of 4. For both the VA-NeRF and the discriminator, their initial learning rates are set to 1×10^{-3} and decay by half four times (at the 2nd, 5th, 10th, and 20th epoch respectively) during training. The whole training process takes about 40 hours on four NVIDIA RTX 3090 GPUs. Loss weights in Eq. (4) are set as $\lambda_{rgb} = 10.0, \lambda_{VGG} = 1.0, \lambda_{adv} = 0.1, \lambda_{vis} = 0.1$. The total number of training epochs is 30. As in (Mihajlovic et al. 2022), we adopt a coarse-to-fine rendering strategy during training that first renders patches by accumulating color and density values of 64 sampled points along a camera ray, and then 128 sampled points for fine-grained rendering.

Baselines. As there is no open-source baseline for interacting hands, we adopt two state-of-the-art generalizable NeRFs designed for humans, including NHP (Kwon et al. 2021) and KeypointNeRF (Mihajlovic et al. 2022). Besides, although HandNeRF (Guo et al. 2023) is non-generalizable, we still combine its core module, i.e., the depth-guided density optimization strategy with KeypointNeRF for comparison. We also choose SMPLpix (Prokudin, Black, and Romero 2021) as an image-space baseline. All networks are trained with the same experimental setting for fair comparison. We select three widely-used evaluation metrics, including peak signal-to-noise ratio (PSNR) (Sara, Akter, and Uddin 2019), structural similarity index (SSIM) (Wang et al. 2004), and learned perceptual image patch similarity (LPIPS) (Zhang et al. 2018).

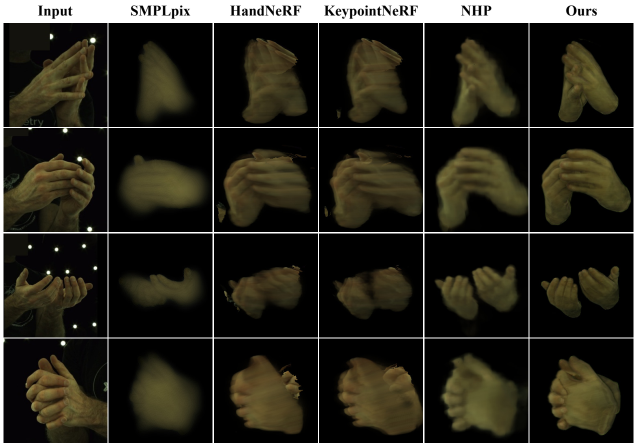

Figure 5: Qualitative examples of novel-view rendering with large view variations (rotation angles > 30 degrees).

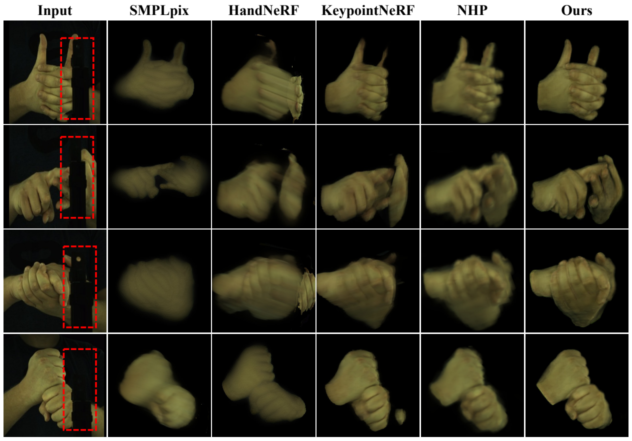

Figure 6: Qualitative comparison of synthesized images in scenes involving severe occlusions.

Method	PSNR↑	SSIM↑	LPIPS↓
SMPLpix	22.49	0.82	0.33
KeypointNeRF	23.49	0.82	0.27
NHP	23.63	0.83	0.33
HandNeRF	23.68	0.83	0.27
Ours	**25.01**	**0.86**	**0.21**

Table 1: Comparison with state-of-the-art methods on Interhand2.6M.

Method	PSNR↑	SSIM↑	LPIPS↓
KeypointNeRF	22.35	0.77	0.34
NHP	22.98	0.80	0.36
Ours	**24.23**	**0.84**	**0.22**

Table 2: Performance comparison among generalizable NeRFs under large view variations (> 30 degrees).

Method	Mask Ratio	PSNR↑	SSIM↑	LPIPS↓
KeypointNeRF	0.1	24.23	0.84	0.23
NHP		23.67	0.83	0.32
Ours		**25.61**	**0.86**	**0.19**
KeypointNeRF	0.2	22.88	0.81	0.25
NHP		23.60	0.83	0.33
Ours		**25.50**	**0.86**	**0.19**
KeypointNeRF	0.3	21.38	0.78	0.29
NHP		23.50	0.82	0.33
Ours		**24.93**	**0.85**	**0.21**

Table 3: Performance comparison among generalizable NeRFs under occlusions.

Comparison with State-of-the-arts

Quantitative comparison. Table 1 reports the quantitative results of our VA-NeRF against the baselines on Interhand2.6M. We can see that VA-NeRF raises the PSNR and the SSIM of state-of-the-art methods from 22.49 to 25.01 and 0.82 to 0.86, and also reduces the LPIPS from 0.33 to 0.21. These performance gains indicate that the synthesized results of VA-NeRF better preserve the hand structures and details (PSNR and SSIM) and are more realistic (LPIPS).

Qualitative comparison. Figure 4 provides the visual comparison between our VA-NeRF and the baselines. Compared with the baselines, the results of VA-NeRF are of better quality and have fewer artifacts. SMPLpix relies on image-space transfer only and hence its results are over-smoothed. The positional encoding in KeypointNeRF relies on MANO joints only and hence it is affected by joint estimation errors severely. HandNeRF alleviates depth ambiguities but is still hard to maintain hand shapes under complex interacting cases. NHP adopts global features and hence the hand structures are preserved well in its results. However, it omits details like wrinkles and nails. Our method adopts the VAFF module to select and merge features, hence it alleviates the disadvantages of baselines successfully.

Moreover, since the major scope of this paper is to tackle large view variations and heavy occlusions, we also evaluate the proposed method in these two cases.

Robustness to large view variations. Novel-view synthesis with large view variations (rotation angles > 30 degrees) requires NeRFs to model long-range feature dependencies. Figure 5 shows the results of the proposed VA-NeRF against baselines in this case. We can see that the results of baselines are severely blurred. On the contrary, the results of our VA-NeRF are still satisfying. We also provide the quantitative comparison between VA-NeRF and the two generalizable baselines in Table 2. Thanks to the VAFF module, our VA-NeRF can model the long-range dependencies among symmetric mesh vertices, and combine global features adaptively to address large view variations.

Robustness to heavy occlusions. The task of image synthesis under heavy occlusions is difficult as well, since certain regions may remain obscured from multiple viewing angles. Figure 6 demonstrates visual examples of VA-NeRF in this case. It is clear that the proposed method successfully recov-

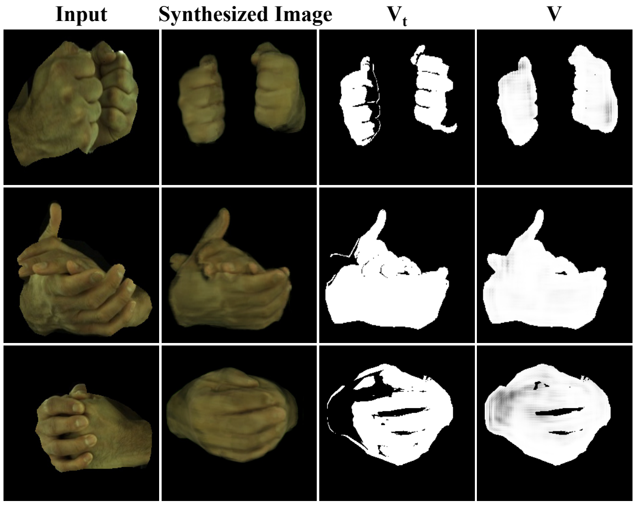

Figure 7: Visualization of visibility maps. V_t/V are target-view/predicted visibility maps. Our NeRF successfully fools the discriminator as most areas in its synthesized images are recognized as visible.

Feature	PSNR↑	SSIM↑	LPIPS↓
q	24.86	0.87	0.18
$q+p$	25.06	0.86	0.19
$q+p'$	25.47	0.86	0.18
$q+p+p'$	**25.74**	**0.87**	**0.18**

Table 4: Ablation study on selected features.

ers unseen areas with realistic textures and structures, while some results of the baselines are distorted and details are not preserved well. Moreover, we generate occluded test images by adding masks centered at images with different ratios (i.e., 0.1, 0.2, and 0.3 of the image size) to conduct quantitative analysis. From the results in Table 3, we can see that our VA-NeRF better maintains its performance compared with the baselines under occlusions.

Ablation Study

Feature selection. To verify that each selected feature (q, p, and p') in our VA-NeRF is necessary, we evaluate the performance of all possible feature combinations. The results are shown in Table 4 and they reflect that each feature does bring performance gains. The best performance is obtained by using all three features.

Effectiveness of VAFF. To validate the effectiveness of visibility in our feature fusion module, we implement a variant that learns the weights of features without the guidance of visibility. Table 5 reports the performance of VAFF and the variant (denoted as Attn. w/o Vis.). We can see that, with the help of visibility, our feature fusion module obtains significantly better performance.

Effectiveness of VGAL. We also evaluate the effectiveness of visibility maps in adversarial learning. We implement a conventional binary discriminator (denoted as Bi. Dis.) and compare it with our visibility-guided discriminator (Vis. Dis.). The results are shown in Table 6 and it is clear that the proposed discriminator outperforms the binary one by large margins in scenes with view variations and occlusions.

Method	PSNR↑	SSIM↑	LPIPS↓
KeyPointNeRF	24.37	0.85	0.24
Attn. w/o Vis.	24.74	0.86	0.20
Attn. w/ Vis.	**25.74**	**0.87**	**0.18**

Table 5: Ablation study on attention strategies.

Scene	Method	PSNR↑	SSIM↑	LPIPS↓
View Variation	Bi. Dis.	23.77	0.83	0.24
	Vis. Dis.	**24.34**	**0.84**	**0.21**
Occlusion	Bi. Dis.	24.57	0.85	0.20
	Vis. Dis.	**25.43**	**0.86**	**0.19**

Table 6: Ablation study on discriminators.

Visualization of visibility maps. The predictions of the proposed discriminator are shown in Figure 7. We can see that the VA-NeRF network does learn to compete with the discriminator effectively, as most regions are recognized as visible by the discriminator. Hence, the proposed VAGL strategy has achieved our goals successfully.

Conclusion

In this paper, we introduce a single-image generalizable visibility-aware neural radiance field framework for image synthesis of interacting hands. The proposed framework leverages the visibility of 3D points for feature fusion and adversarial learning. Our feature fusion is achieved by fusing features of reference vertices closely related to query points, with fusion weights determined by point visibility. Our adversarial learning is accomplished through the training of a pixel-wise discriminator capable of estimating visibility maps. With these two components cooperating together, the proposed method can obtain reliable features and high-quality results, even in challenging scenarios involving heavy occlusions and large view variations. The proposed method is evaluated on Interhand2.6M and obtains performance superior to state-of-the-art generalizable models.

Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), National Natural Science Foundation of China (NSFC) under Grant No. 61976233, No. 92270122, No. 62372482 and No. 61936002, Mobility Grant Award under Grant No. M-0461, Shenzhen Science and Technology Program (Grant No. RCYX20200714114642083), Shenzhen Science and Technology Program (Grant No. GJHZ20220913142600001), Nansha Key R&D Program under Grant No.2022ZD014 and Sun Yat-sen University under Grant No. 22lgqb38 and 76160-12220011.

References

Chan, E. R.; Lin, C. Z.; Chan, M. A.; Nagano, K.; Pan, B.; De Mello, S.; Gallo, O.; Guibas, L. J.; Tremblay, J.; Khamis, S.; et al. 2022. Efficient geometry-aware 3D generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 16123–16133.

Chan, E. R.; Monteiro, M.; Kellnhofer, P.; Wu, J.; and Wetzstein, G. 2021. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 5799–5809.

Chen, A.; Xu, Z.; Geiger, A.; Yu, J.; and Su, H. 2022a. Tensorf: Tensorial radiance fields. In *Proceedings of the European Conference on Computer Vision*, 333–350.

Chen, X.; Liu, Y.; Ma, C.; Chang, J.; Wang, H.; Chen, T.; Guo, X.; Wan, P.; and Zheng, W. 2021. Camera-space hand mesh recovery via semantic aggregation and adaptive 2d-1d registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13274–13283.

Chen, X.; Wang, B.; and Shum, H.-Y. 2023. Hand Avatar: Free-Pose Hand Animation and Rendering from Monocular Video. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Chen, Z.; Hasson, Y.; Schmid, C.; and Laptev, I. 2022b. AlignSDF: Pose-Aligned Signed Distance Fields for Hand-Object Reconstruction. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, 231–248. Springer.

Corona, E.; Hodan, T.; Vo, M.; Moreno-Noguer, F.; Sweeney, C.; Newcombe, R.; and Ma, L. 2022. LISA: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 20533–20543.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022a. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Deng, X.; Zuo, D.; Zhang, Y.; Cui, Z.; Cheng, J.; Tan, P.; Chang, L.; Pollefeys, M.; Fanello, S.; and Wang, H. 2022b. Recurrent 3D Hand Pose Estimation Using Cascaded Pose-guided 3D Alignments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1): 932–945.

Deng, Y.; Yang, J.; Xiang, J.; and Tong, X. 2022c. Gram: Generative radiance manifolds for 3d-aware image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10673–10683.

Gao, D.; Xiu, Y.; Li, K.; Yang, L.; Wang, F.; Zhang, P.; Zhang, B.; Lu, C.; and Tan, P. 2022a. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. In *Advances in Neural Information Processing Systems (Datasets and Benchmarks Track)*.

Gao, K.; Gao, Y.; He, H.; Lu, D.; Xu, L.; and Li, J. 2022b. Nerf: Neural radiance field in 3d vision, a comprehensive review. *arXiv preprint arXiv:2210.00379*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2020. Generative adversarial networks. *Communications of the ACM*, 63(11): 139–144.

Güler, R. A.; Neverova, N.; and Kokkinos, I. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7297–7306.

Guo, Z.; Zhou, W.; Wang, M.; Li, L.; and Li, H. 2023. HandNeRF: Neural Radiance Fields for Animatable Interacting Hands. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Hong, F.; Chen, Z.; Lan, Y.; Pan, L.; and Liu, Z. 2023. Eva3d: Compositional 3d human generation from 2d image collections. *International Conference on Learning Representations*.

Hu, S.; Hong, F.; Pan, L.; Mei, H.; Yang, L.; and Liu, Z. 2023. SHERF: Generalizable Human NeRF from a Single Image. *arXiv preprint arXiv:2303.12791*.

Jiang, T.; Chen, X.; Song, J.; and Hilliges, O. 2022. InstantAvatar: Learning Avatars from Monocular Video in 60 Seconds. *arXiv preprint arXiv:2212.10550*.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022a. GeoNeRF: Generalizing NeRF With Geometry Priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 18365–18375.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022b. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18365–18375.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, 694–711.

Karunratanakul, K.; Spurr, A.; Fan, Z.; Hilliges, O.; and Tang, S. 2021. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision*, 11–21. IEEE.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kulon, D.; Guler, R. A.; Kokkinos, I.; Bronstein, M. M.; and Zafeiriou, S. 2020. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4990–5000.

Kwon, Y.; Kim, D.; Ceylan, D.; and Fuchs, H. 2021. Neural human performer: Learning generalizable radiance fields for human performance rendering. *Advances in Neural Information Processing Systems*, 34: 24741–24752.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2015. SMPL: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6): 1–16.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 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.

Meng, H.; Jin, S.; Liu, W.; Qian, C.; Lin, M.; Ouyang, W.; and Luo, P. 2022. 3d interacting hand pose estimation by hand de-occlusion and removal. In *Proceedings of the European Conference on Computer Vision*, 380–397.

Mescheder, L.; Geiger, A.; and Nowozin, S. 2018. Which training methods for GANs do actually converge? In *International conference on machine learning*, 3481–3490.

Mihajlovic, M.; Bansal, A.; Zollhoefer, M.; Tang, S.; and Saito, S. 2022. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *Proceedings of the European Conference on Computer Vision*, 179–197.

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16190–16199.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Moon, G.; Yu, S.-I.; Wen, H.; Shiratori, T.; and Lee, K. M. 2020. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *Proceedings of the European Conference on Computer Vision*, 548–564.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4): 1–15.

Newell, A.; Yang, K.; and Deng, J. 2016. Stacked hourglass networks for human pose estimation. In *Proceedings of the European Conference on Computer Vision*, 483–499.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 5480–5490.

Niemeyer, M.; and Geiger, A. 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.

Or-El, R.; Luo, X.; Shan, M.; Shechtman, E.; Park, J. J.; and Kemelmacher-Shlizerman, I. 2022. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13503–13513.

Park, J.; Oh, Y.; Moon, G.; Choi, H.; and Lee, K. M. 2022. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1496–1505.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9054–9063.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Prokudin, S.; Black, M. J.; and Romero, J. 2021. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1810–1819.

Raj, A.; Zollhofer, M.; Simon, T.; Saragih, J.; Saito, S.; Hays, J.; and Lombardi, S. 2021. Pixel-aligned volumetric avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11733–11742.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics*, 36(6): 1–17.

Romero, J.; Tzionas, D.; and Black, M. J. 2022. Embodied hands: Modeling and capturing hands and bodies together. *arXiv preprint arXiv:2201.02610*.

Saito, S.; Huang, Z.; Natsume, R.; Morishima, S.; Kanazawa, A.; and Li, H. 2019. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE International Conference on Computer Vision*, 2304–2314.

Sara, U.; Akter, M.; and Uddin, M. S. 2019. Image quality assessment through FSIM, SSIM, MSE and PSNR—a comparative study. *Journal of Computer and Communications*, 7(3): 8–18.

Schwarz, K.; Liao, Y.; Niemeyer, M.; and Geiger, A. 2020. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33: 20154–20166.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 12922–12931.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4690–4699.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Xu, X.; Pan, X.; Lin, D.; and Dai, B. 2021. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34: 20683–20695.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zhang, B.; Wang, Y.; Deng, X.; Zhang, Y.; Tan, P.; Ma, C.; and Wang, H. 2021. Interacting two-hand 3d pose and shape reconstruction from single color image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 11354–11363.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 586–595.

Zhou, Y.; Habermann, M.; Xu, W.; Habibie, I.; Theobalt, C.; and Xu, F. 2020. Monocular real-time hand shape and motion capture using multi-modal data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5346–5355.