Fine-Grained Multi-View Hand Reconstruction Using Inverse Rendering

Qijun Gan, Wentong Li, Jinwei Ren, Jianke Zhu*

College of Computer Science and Technology, Zhejiang University, China
{ganqijun,liwentong,zijinxuxu,jkzhu}@zju.edu.cn

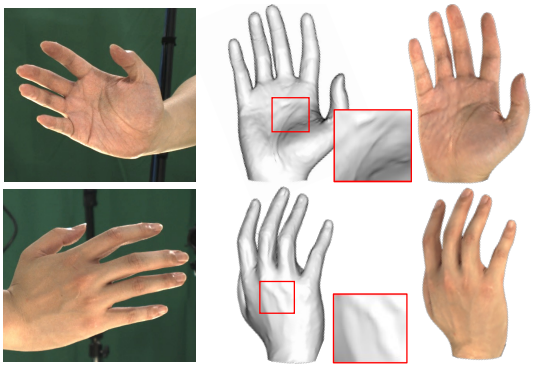
Abstract

Reconstructing high-fidelity hand models with intricate textures plays a crucial role in enhancing human-object interaction and advancing real-world applications. Despite the state-of-the-art methods excelling in texture generation and image rendering, they often face challenges in accurately capturing geometric details. Learning-based approaches usually offer better robustness and faster inference, which tend to produce smoother results and require substantial amounts of training data. To address these issues, we present a novel fine-grained multi-view hand mesh reconstruction method that leverages inverse rendering to restore hand poses and intricate details. Firstly, our approach predicts a parametric hand mesh model through Graph Convolutional Networks (GCN) based method from multi-view images. We further introduce a novel Hand Albedo and Mesh (HAM) optimization module to refine both the hand mesh and textures, which is capable of preserving the mesh topology. In addition, we suggest an effective mesh-based neural rendering scheme to simultaneously generate photo-realistic image and optimize mesh geometry by fusing the pre-trained rendering network with vertex features. We conduct the comprehensive experiments on InterHand2.6M, DeepHandMesh and dataset collected by ourself, whose promising results show that our proposed approach outperforms the state-of-the-art methods on both reconstruction accuracy and rendering quality. Code and dataset are publicly available at https://github.com/agnJason/FMHR.

Introduction

3D human reconstruction has attracted considerable research attentions (Saito et al. 2019; Peng et al. 2021a; Weng et al. 2022; Chen et al. 2021; Noguchi et al. 2021; Bhatnagar et al. 2020) in recent years. While there have been promising advancements in reconstructing the human body and face (Lei et al. 2023; Grassal et al. 2022; Peng et al. 2021b; Xiu et al. 2022), it still remains a formidable challenge to achieve highly accurate hand reconstruction due to the inherent complexity of joint variations. By taking consideration of the distinctive nature for hands, it is essential to investigate the hand geometry and rendering for obtaining the realistic and fine-grained representations.

*Corresponding author.



a) Input Images b) Hyper-realistic Hand Mesh

Figure 1: Our proposed approach focuses on reconstructing hands from multi-view images, allowing for the generation of precise poses, geometry, and photo-realistic rendering.

Conventional model-based methods, such as MANO (Romero, Tzionas, and Black 2017) and Nimble (Li et al. 2022b), often rely on smoothing meshes and texture maps for hand representation. Nevertheless, it typically requires costly scanning data and artistic expertise in order to achieve intricate and personalized hand meshes with texture maps. Moreover, the complex nature of hand movements and challenges posed by occlusions hinder the faithful restoration of hand mesh. Model-free approaches like LISA (Corona et al. 2022) aim to address these challenges by reconstructing coherent hands from image sequences, while HandAvatar (Chen, Wang, and Shum 2023) focuses on reconstructing and rendering hands in arbitrary poses by disentangling reflectance and lighting. Nonetheless, the resulting mesh often exhibits smoothness due to large pose variations across different frames.

Neural rendering-based methods, such as NeRF (Mildenhall et al. 2021) and NeuS (Wang et al. 2021; Fu et al. 2022), have been widely used in synthesizing static objects from multi-view images. As for the dynamic objects like hands, these methods have the difficulties in obtaining a fixed topological structure through implicit surface representation. Despite their impressive rendering results, NeuS (Wang et al. 2021) requires a large amount of training time due to sampling along the ray. Meanwhile, NeuralBody (Peng et al.

2021b) attaches latent codes to SMPL (Loper et al. 2015) vertices, which enables to diffuse into space through sparse convolution. However, this may lead to some artifacts on the predicted mesh. To address these limitations, HandNeRF (Guo et al. 2023) proposes a pose-driven deformation field to render photo-realistic hands from various views and poses, while it does not explicitly provide a hand mesh for animation and real-time rendering.

To overcome the above challenges, we propose an effective coarse-to-fine approach to hand mesh reconstruction from multi-view images. By synergizing the benefits of parametric models and mesh-based rendering, our method achieves high-fidelity reconstruction results while maintaining fast training time with only 1.5 minutes. By incorporating multi-layer perceptrons into Graph Convolutional Networks (GCN) (Kipf and Welling 2017), we can simultaneously recover hand mesh and MANO parameters. Unlike learning-based methods that demand extensive training data, we introduce a Hand Albedo and Mesh (HAM) optimization module by leveraging inverse rendering to enhance level of details. The resulting fine-grained mesh preserves the flexibility provided by the parametric model MANO (Romero, Tzionas, and Black 2017), which enables to repose it into novel configurations. To further enhance rendering quality and refine the mesh, we suggest a mesh-based neural rendering scheme by fusing a pre-trained rendering network with vertex features. Our proposed method is evaluated on the InterHand2.6M, DeepHandMesh and dataset collected by ourself. The promising experimental results demonstrate its efficacy in reconstructing high-fidelity hands from multi-view images, as illustrated in Fig. 1.

Our main contributions are summarized as below.

- We propose a coarse-to-fine approach to accurately recover the fine-grained hand mesh model from multi-view images by taking advantage of inverse rendering.
- A novel HAM optimization module is presented to refine the over-smoothing results of parametric hand models.
- We devise an effective mesh-based neural rendering scheme to simultaneously generate photo-realistic image and optimize mesh geometry by fusing the pre-trained rendering network with vertex features.

Related Work

Model-Based Hand Reconstruction

Parametric models have been widely used in representing objects with the fixed typologies, such as the human body (Loper et al. 2015; Pavlakos et al. 2019; Osman, Bolkart, and Black 2020; Chen et al. 2022), face (Li et al. 2017; Hong et al. 2022), hands (Romero, Tzionas, and Black 2017; Li et al. 2022b), and animals (Zuffi et al. 2017). These models enable the transformation of mesh geometry by adjusting model parameters corresponding to pose and shape variations. Moreover, hand pose can be effectively estimated by images (Li, Gao, and Sang 2021) or point clouds (Cheng et al. 2022; Ren et al. 2023). In hand reconstruction, the parametric models like MANO (Romero, Tzionas, and Black 2017), NIMBLE (Li et al. 2022b), have been used to recover the hand in the input image (Boukhayma, de Bem, and Torr 2019; Hasson et al. 2019; Kong et al. 2022; Cao et al. 2021; Doosti et al. 2020; Hasson et al. 2020). In (Fan et al. 2021; Ren, Zhu, and Zhang 2023; Kim, Kim, and Baek 2021; Zhang et al. 2021; Li et al. 2022a), the parametric models are employed to reconstruct two hands, where hand interactions and gestures could be simulated. Recently, (Chen et al. 2023) employ the HandTrackNet to track the variations of MANO parameters, which is utilized for hand-object interactions. While parametric methods are able to recover hand poses and shapes, the resulting meshes lack the capability to represent geometric textures.

Model-Free Hand Reconstruction

Parametric models (Romero, Tzionas, and Black 2017) are valuable for incorporating the prior knowledge of pose and shape, while their representation power is constrained by imposing the template shape and details. To overcome this limitation, various approaches have been explored. Instead of directly regressing the MANO parameters, I2L-MeshNet (Moon and Lee 2020) predicts a 1D heatmap for each vertex coordinate, and (Ge et al. 2019; Kulon et al. 2020) employ GCN-based methods to recover hand meshes. On the other hand, DeepHandMesh (Moon, Shiratori, and Lee 2020) make use of an encoder-decoder framework to generate highly detailed hand meshes. To achieve photo-realistic hand rendering, HARP (Karunratanakul et al. 2023) suggest an optimization-based approach to recover both normal and albedo maps. Recently, (Luan et al. 2023) introduce a frequency decomposition loss to capture personalized hand from a single image, which address the problem of data scarcity through multi-view reshaping. HandAvatar (Chen, Wang, and Shum 2023) yields occupancy and illumination field to generate free-pose photo-realistic hand avatar.

Neural Rendering-Based Reconstruction

In past few years, rapid progress in 3D modeling and image synthesis has been obtained through neural implicit representations (Mescheder et al. 2019; Mildenhall et al. 2021). In contrast to the classical discrete representations like meshes, point clouds, and voxels, neural implicit representations leverage neural networks to model scenes, which offer continuous results with higher fidelity and flexibility. Among the various methods, Neural Radiance Fields (NeRF) (Mildenhall et al. 2021) gains great popularity and demonstrates impressive performance across various tasks. Nevertheless, the traditional NeRF-based methods face challenges in dealing with objects having temporal changes (Fu et al. 2022; Wang et al. 2021).

Neural Body (Peng et al. 2021b) combines NeRF with the parametric SMPL body model (Loper et al. 2015), which is able to recover dynamic objects. Moreover, (Liu et al. 2021) leverage NeRF to learn pose-related geometric deformations and textures in canonical space from multi-view videos. Recently, LISA (Corona et al. 2022) fuses volumetric rendering with hand geometric priors to capture animatable hand appearances. HandNeRF (Guo et al. 2023) represents the interactive hands by deformable neural radiance fields to generate photo-realistic images.

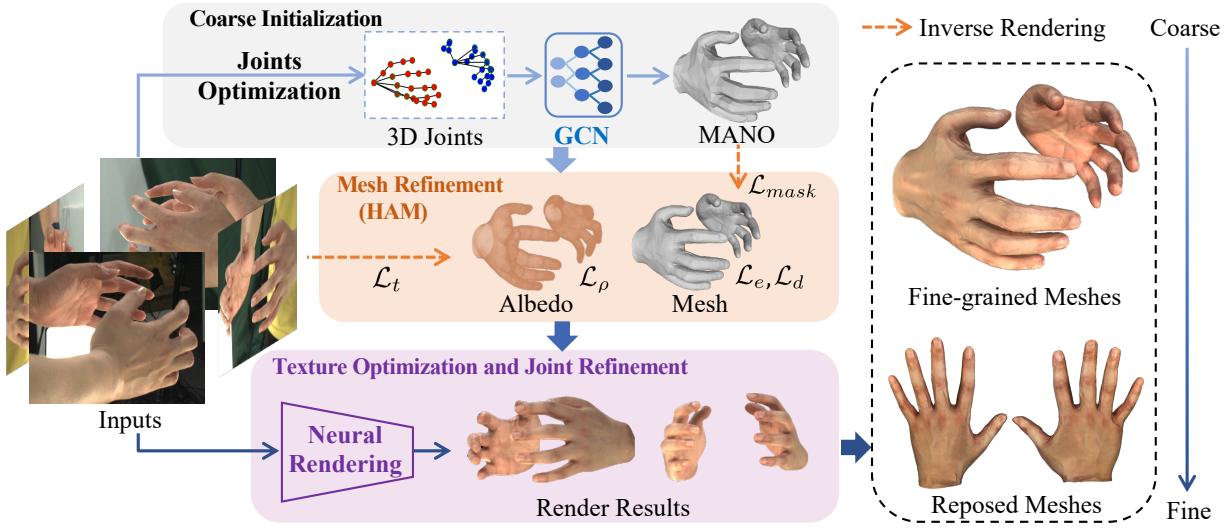


Figure 2: Overview of our coarse-to-fine framework. Given a set of calibrated images, we initialize MANO parameters and refine the mesh using our proposed HAM module and inverse rendering to achieve geometric details. By jointly optimizing the mesh using a model-based neural rendering, a fine-grained mesh can be obtained along with its hyper-realistic rendered images.

Method

Our objective is to achieve fine-grained 3D hand mesh reconstruction and synthesize photo-realistic novel views. To tackle this challenge, our method consists of three key steps. Firstly, we estimate an initial coarse hand mesh and the parameters of MANO model (Romero, Tzionas, and Black 2017), which are recovered from multi-view images through incorporating multi-layer perceptrons (MLPs) into GCN. Secondly, an HAM optimization module is introduced to restore a fine-grained mesh with the albedo map and surface details by leveraging the power of inverse rendering. Finally, we suggest a mesh-based neural rendering scheme to efficiently generate photo-realistic images and refine the mesh through joint optimization. Fig. 2 illustrates the entire pipeline of our presented method.

Coarse Initialization

In hand reconstruction, it is crucial to accurately predict the 3D pose and shape due to the high flexibility of the hand. Since generating 3D poses is typically challenging in the wild cases, some approaches like HandAvatar (Chen, Wang, and Shum 2023) employ the annotated hand poses and shapes as initialization. Instead, we aim to estimate the consistent 3D hand poses from multi-view images.

Given a set of images $\mathcal{I} = \{I^1, \cdots, I^n\}$ captured by the calibrated cameras and their corresponding 2D joints $\mathcal{J}_{2D} = \{J^1_{2D}, \cdots, J^n_{2D}\}$, 3D hand joints $J_{3D} \in \mathbb{R}^{B\times 3}$ in world space with B per-bone parts can be estimated. To address the limitations of representing joint Euler angles directly with 3D joints information, we introduce a GCN-based network $\mathcal{G}$ to recover the MANO model as in (Choi, Moon, and Lee 2020).

$$\mathcal{M}(\hat{\theta}, \hat{\beta}) = \mathcal{G}(J_{3D}), \tag{1}$$

where $\hat{\theta} \in \mathbb{R}^{B\times 3}$ and $\hat{\beta} \in \mathbb{R}^{10}$ represent pose and shape parameters of MANO model, respectively. The network $\mathcal{G}$ is designed as a four-layer GCN with a MANO head. The MANO head consists of MLPs to obtain the corresponding MANO parameters via the features of the first three GCN layers, as shown in Fig. 3.

Given n-view images I^n along with their corresponding camera parameters π^n and the positions of the hand joints J^n_{2D} in the images, J_{3D} can be obtained by minimizing the reprojection error across different views. The camera parameters include the intrinsic matrix K and extrinsic matrix T. In the i-th view, J^i_{2D} can be calculated by $J^i_{2D} = \pi^i(J_{3D})$. Since it is difficult to circumvent the issue of imperfect estimation of J_{2D}, we recover J_{3D} by minimizing the following objective

$$\mathcal{L}_{joints} = \sum_{i=1}^{n} \frac{1}{n} ||J^i_{2D} - \pi^i(\hat{J}_{3D})||^2. \tag{2}$$

Once the 3D joints $\hat{J}_{3D}$ with multi-view consistency is obtained, the GCN-based network $\mathcal{G}$ can be used to recover the MANO parameters. $\mathcal{G}$ is trained using the annotated datasets by minimizing the following energy function

$$\mathcal{L}_{\mathcal{G}} = \mathcal{L}_v + \mathcal{L}_n + \mathcal{L}_j + \mathcal{L}_{MANO}. \tag{3}$$

Specifically, $\mathcal{L}_v$ measures the L_1 distance between the predicted vertices $\hat{V}$ and annotated MANO vertices V as below

$$\mathcal{L}_v = \sum |\hat{V} - V|. \tag{4}$$

$\mathcal{L}_n$ ensures the alignment of mesh normals $\mathbf{n}$ with MANO, which is defined as

$$\mathcal{L}_n = \sum |\hat{\mathbf{n}} - \mathbf{n}|. \tag{5}$$

$\mathcal{L}_j$ constrains the discrepancy between the generated MANO joints $\hat{J_{3D}}$ and the input joints J_{3D} as follows

$$\mathcal{L}_j = \sum |\hat{J_{3D}} - J_{3D}|. \tag{6}$$

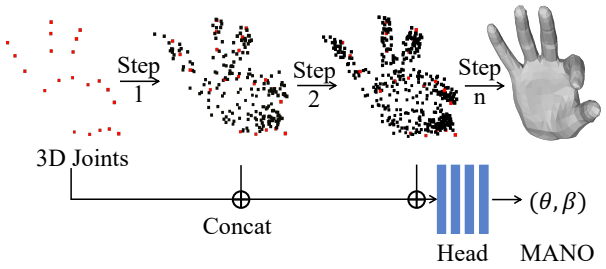


Figure 3: Our GCN-based network. The four-layer GCN progressively doubles the number of vertices and MANO head outputs the corresponding MANO parameters.

Additionally, $\mathcal{L}_{MANO}$ is used to supervise the generation of MANO parameters, which is formulated as below

$$\mathcal{L}_{MANO} = \sum |(\hat{\theta}, \hat{\beta}) - (\theta, \beta))|. \tag{7}$$

While the GCN-based network $\mathcal{G}$ excels at accurately recovering the pose by utilizing 3D joints as input, it lacks information about hand fatness or leanness. To overcome this limitation, we incorporate segmentation map obtained by promptable SAM (Kirillov et al. 2023) in order to optimize the shape parameters $\hat{\beta}$. This extra optimization step further enhances the accuracy of the MANO parameters.

Mesh Refinement

The MANO model generated by $\mathcal{G}$ represents a smoothing mesh without geometric textures. Inspired by the Shape from Shading (SFS) algorithm (Horn 1970) and PatchShading (Lin et al. 2024; Lin, Zhu, and Zhang 2022), we accurately capture the folds and textures on the coarse hand mesh. Assuming that the diffuse reflection of human skin follows Lambertian reflectance, we propose a Hand Albedo and Mesh (HAM) optimization module to refine the coarse mesh, which takes advantage of both SFS and inverse rendering. To ensure the coherence in the mesh optimization process, we incorporate four effective regularization terms described in the following.

We utilize $\mathcal{M}_r$ as the initial model that is obtained from the MANO subdivision with 49,281 vertices and 98,432 faces. The albedo values ρ are assigned to each vertex and an inverse renderer ζ is employed to generate albedo map $\mathbf{C}_\rho$ and normal map $\mathbf{C}_N$. The inverse renderer ζ is represented as follows

$$\mathbf{C} = \zeta(c, V, F, \pi), \tag{8}$$

where V and F denote the vertices and faces of mesh $\mathcal{M}_r$, respectively. The feature at each vertex, such as normal and albedo, is represented by c. The rendering map $\mathbf{C}$ is obtained by the inverse renderer ζ. With the calibrated images I^i, the HAM optimizes the vertex positions V and vertex albedo ρ by minimizing the texture loss $\mathcal{L}_t$ in the following equation

$$\mathcal{L}_t = \sum_n |B(\pi^i) - I^i|, \tag{9}$$

where $B(\pi^i)$ represents the rendered image under camera parameters π^i. It is obtained by computing the illumination matrix $\mathbf{G}$, the normal map $\mathbf{C}_N$ and the albedo map $\mathbf{C}_\rho$ as below

$$B(\pi^i) = \mathbf{C}_\rho \cdot SH(\mathbf{G}, \mathbf{C}_N), \tag{10}$$

where $SH(\cdot)$ represents sphere harmonic (SH) function with the third order. Considering the variations in lighting across different views, we optimize the lighting matrix $\mathbf{G}$ during the process. We compute the texture loss between $B(\pi^i)$ and the original image I^i as the loss $\mathcal{L}_t$.

To enhance the efficacy of extracting geometric information from shadows, the regularization term $\mathcal{L}_\rho$ is introduced. This term is especially designed to align with the observation that human skin tends to exhibit the consistent color. $\mathcal{L}_\rho$ is defined as follows

$$\mathcal{L}_\rho = \lambda_1 L * \rho, \tag{11}$$

where L denotes the Laplacian matrix. λ is the balanced weight. To ensure the optimized vertices to be smoothing, we introduce the following term for conforming to the geometric characteristics of the hand

$$\begin{aligned} \mathcal{L}_r &= \lambda_2 L * V + \lambda_3 \mathcal{L}_{mask} + \lambda_4 \mathcal{L}_e + \lambda_5 \mathcal{L}_d \\ &= \lambda_2 L * V + \lambda_3 \sum |\hat{M} - M_{MANO}| \\ &+ \lambda_4 \sum_{i,j} ||E_{ij}||^2 + \lambda_5 \sum_i ||\triangle V_i||^2. \end{aligned} \tag{12}$$

Specifically, $\mathcal{L}_{mask}$ represents the L_1 loss between the rendered mask and the original MANO mask. E_{ij} is obtained by calculating the Euclidean distance $||\cdot||^2$ between adjacent vertices V_i and V_j on the mesh edges. $\mathcal{L}_e$ is employed to restrict the length of E_{ij}. Let $\triangle V_i$ represent the displacement distance of vertices V_i. $\mathcal{L}_d$ is utilized to ensure that the optimized hand remains close to the MANO model. Each term is assigned with a constant coefficient denoted by λ. The overall loss function $\mathcal{L}_{total}$ is defined as below

$$\mathcal{L}_{total} = \mathcal{L}_t + \mathcal{L}_\rho + \mathcal{L}_r. \tag{13}$$

The HAM module facilitates the refinement of the hand model by jointly optimizing both the mesh vertices $\hat{V}$ and albedo $\hat{\rho}$. This results in a high-quality output $\mathcal{M}_f$ that retains the consistent topology.

Texture Optimization and Joint Refinement

While having achieved geometric alignment of the fine mesh, there are still some limitations in image rendering and mesh refinement. To alleviate this issue, we adopt a mesh-based neural rendering method. Leveraging the preserved topology of our refined hands, the neural rendering model can be pre-trained with diverse hand data, thereby reducing the training time for each individual hand. We propose an efficient strategy that involves the pre-training a neural rendering network using a large amount of hand data, followed by fine-tuning on individual data, and ultimately conducting joint optimization with the mesh to achieve hyper-realistic rendering and accurate geometry.

Having acquired a refined mesh denoted as $\mathcal{M}_f(\hat{V}, F)$ along with its corresponding vertex albedo $\hat{\rho}$, we aim to enhance the accuracy of the rendered texture. To this end, we design a neural renderer $\mathcal{T}$ defined as below

$$\mathbf{t}(\mathbf{r}) = \mathcal{T}(\mathbf{x}, \mathbf{f}, \rho, \mathbf{n}), \tag{14}$$

where the output pixel value $\mathbf{t}$ is determined by the direction of the light ray $\mathbf{r}$. At a given pixel position, the texture field t is adjusted with respect to the position $\mathbf{x}$, feature vector $\mathbf{f}$, albedo ρ, and normal $\mathbf{n}$, which are obtained by inverse rendering ζ. For the neural renderer, we employ the MLPs with four layers, each of which consists of 256 dimensions. The length of vertex features $\mathbf{f}$ is set to 20. The training loss is defined as follows

$$\mathcal{L}_{tex} = \sum |\hat{\mathbf{t}}_i - I_i|. \tag{15}$$

Pre-training. During the pre-training, we conduct sampling from the subdivided MANO mesh $\mathcal{M}_r$ instead of utilizing $\mathcal{M}_f$. Through the pre-training, we obtain a well-trained neural renderer that unifies the vertices features $\mathbf{f}$ independent from position $\mathbf{x}$, albedo ρ and normal $\mathbf{n}$.

Fine-tuning. Given the diversity of data and potential overfitting during training, the pre-trained neural rendering model often yields average results, which is unable to capture intricate texture details. To synthesize high-fidelity hand images and achieve promising rendering quality, it is necessary to fine-tune the model for each specific dataset. The advantage of the pre-trained model lies in its capacity to accelerate neural rendering, allowing for efficient completion of fine-tuning within minutes. During the fine-tuning process, we employ the vertices of mesh $\mathcal{M}_f$ obtained from Mesh Refinement along with its corresponding albedo for training. Both the vertex features $\mathbf{f}$ and the neural renderer $\mathcal{T}$ are set to be learnable.

Joint Optimization. After achieving fine geometric structures and realistic image rendering, it becomes necessary to perform joint optimization on both mesh and texture to further enhance the overall quality. Drawing inspiration from (Walker et al. 2023), we adopt a geometry-based shader $\tilde{\mathbf{t}}$ with the detached output $\hat{\mathbf{t}}$ of $\mathbf{t}(\mathbf{r})$, which is illustrated as follows

$$\tilde{\mathbf{t}}(\mathbf{r}) = \tilde{\mathcal{T}}(\hat{\mathbf{t}}, f, \rho, \mathbf{n}). \tag{16}$$

To fine-tune the geometry and train the geometry-based shader, the loss functions $\mathcal{L}_e$ and $\mathcal{L}_d$ depicted in Eq. 12 are employed to ensure surface smoothness. Moreover, the vertices are designated to be learnable. The shader loss is formulated as follows

$$\mathcal{L}_{geo} = \sum |\tilde{\mathbf{t}}_i - I_i| + \gamma_1 \mathcal{L}_e + \gamma_2 \mathcal{L}_d. \tag{17}$$

where γ_1 and γ_2 are weights for L_e and L_d, respectively.

Experiments

Datasets

InterHand2.6M. InterHand2.6M (Moon et al. 2020) is a large-scale dataset, comprising images of size 512×334 pixels and associated MANO annotations. The dataset contains multi-view temporal sequences of single hand as well as interacting hands. Our experiments primarily focus on the 5 FPS version of the InterHand2.6M dataset. In all experiments, both the GCN-based network and pre-trained neural rendering network are trained using the data from the training set.

DeepHandMesh. The DeepHandMesh dataset (Moon, Shiratori, and Lee 2020) consists of images captured from five different views with the same size as the images in InterHand2.6M dataset. Additionally, this dataset provides the corresponding 3D hand scans, which enables the validation of the mesh reconstruction quality against 3D ground truth.

Our Dataset. Due to the restricted resolution of the above datasets, the attainment of higher geometric precision and color fidelity requires high-resolution hand images. To address this issue, we collect a dataset using 16 calibrated cameras, which captures the synchronized images at a resolution of 1280×1024 pixels at 15 FPS. The cameras are distributed mainly in a semi-circle and placed at various heights to ensure a comprehensive visual coverage.

Implementation Details and Metrics

Optimization. To achieve the fine-grained mesh, we adopt a subdivision technique inspired by (Chen, Wang, and Shum 2023), which expands the original 778 vertices in the MANO model to a total of 49,281 vertices. During the optimization process, we utilize the Adam optimizer (Kingma and Ba 2014) with the balanced weights of $\lambda_1 = 20$, $\lambda_2 = 40$, $\lambda_3 = 20$, $\lambda_4 = 100$, and $\lambda_5 = 2$ to jointly optimize the vertices, vertex albedo, and lighting coefficients over 100 iterations. This optimization process takes approximately 20 seconds. Additionally, the neural renderer is pre-trained on InterHand2.6M dataset for 20 epochs. Subsequently, for fine-tuning and joint optimization, each process requires 100 epochs of training with $\gamma_1 = 100$ and $\gamma_2 = 2$, respectively. Notably, the entire optimization pipeline is computationally efficient, which takes approximately 90 seconds on a single NVIDIA 3090Ti GPU.

Evaluation Metrics. We evaluate the accuracy of the reconstructed 3D surface by computing the average point-to-surface Euclidean distance (P2S) between the vertices of the recovered surface and their corresponding ground truth, which are measured in millimetre. Due to the disparate size ranging between the generated hand mesh and the 3D scans in the DeepHandMesh data, the Chamfer distance metric is considered unsuitable. In line with prior research in neural rendering, peak signal-to-noise ratio (PSNR), structural similarity index (SSIM), and learned perceptual image patch similarity (LPIPS) are adopted as evaluation metrics to gauge the fidelity of the synthesized image results.

Results on Hand Reconstruction

Results of InterHand2.6M. To evaluate the quality of novel view synthesis, we conduct experiments on the InterHand2.6M dataset using 10 views and evaluate on the rest views. Table 1 summarizes the evaluation results on rendering quality. Fig. 4 illustrates the visual comparisons against HandNeRF (Guo et al. 2023), HandAvatar (Chen, Wang, and Shum 2023) and NeuS (Wang et al. 2021). The metrics of Ani-NeRF are extracted from the data presented in (Guo et al. 2023). It is important to note that HandAvatar lacks support for interactive hands, while HandNeRF is not able to directly predict geometry. Both HandNeRF and HandAvatar rely on learning from video sequence for voxel rendering with large pose variations, which may result in smoothing texture. By taking advantage of the design of our topology-consistent hand mesh and the mesh-based neural

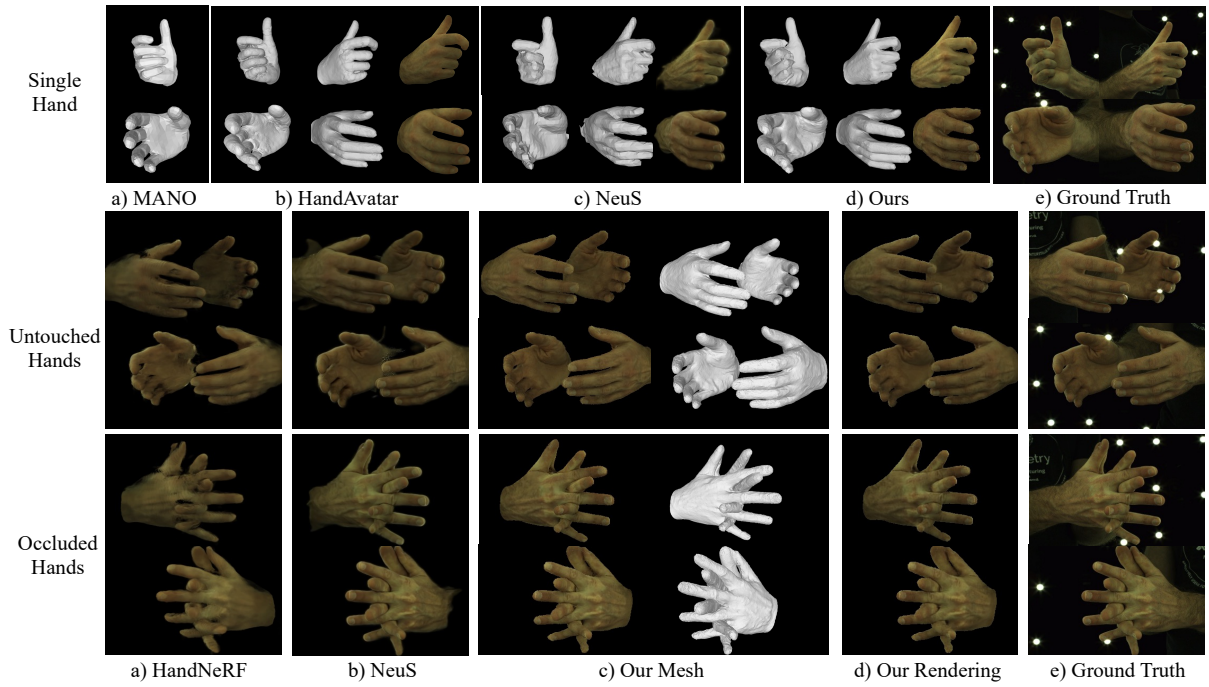


Figure 4: Qualitative performance comparison. We show the rendering results of single hand (first two rows) and dual hands (last four rows), which are optimized and trained from 10-view images. The hands rendered with pure white color represent the shading in order to highlight the level of mesh detail.

Method	*test/Capture0-Single*			*test/Capture1-Single*			*test/Capture0-Interacting*		
	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
Ani-NeRF	0.0621	31.78	0.968	-	-	-	0.0798	29.35	0.949
HandAvatar	0.0504	33.01	0.933	0.0425	32.12	0.938	-	-	-
HandNerf	0.0375	32.70	0.974	-	-	-	0.0367	30.62	0.958
NeuS	0.0197	35.34	0.986	0.0211	34.98	0.984	0.0743	31.46	0.927
Ours	**0.0069**	**37.26**	**0.992**	**0.0098**	**37.91**	**0.991**	**0.0126**	**37.08**	**0.986**

Table 1: Rendering quality comparisons among our method and prior arts on the InterHand2.6M dataset.

P2S ↓	*Two occlusions*	*Thumb tuck*	*Shake-speare*	*Total Average*
DHM	3.467	2.624	1.773	2.937
w/o.HAM	1.637	2.317	1.668	1.873
Ours	**1.532**	**1.281**	**1.271**	**1.456**

Table 2: Mesh reconstruction quality comparison among ours and DHM (Moon, Shiratori, and Lee 2020) on DeepHandMesh dataset with 5 views.

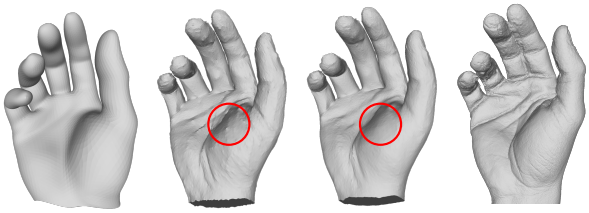
Init	*HAM*	*Joint refine*	P2S ↓	PSNR ↑
✓			1.873	-
	✓	✓	5.506	31.96
✓	✓		1.457	35.68
✓	✓	✓	**1.456**	**37.91**

Table 3: The impact of coarse initialization, HAM and joint refinement on rendering quality and mesh geometry.

rendering network, our presented method achieves PSNR of 37dB with just about one minute of fine-tuning. Comparing to NeuS, some semi-transparent mist-like artifacts are observed around the rendered hand, as shown in Fig. 4.

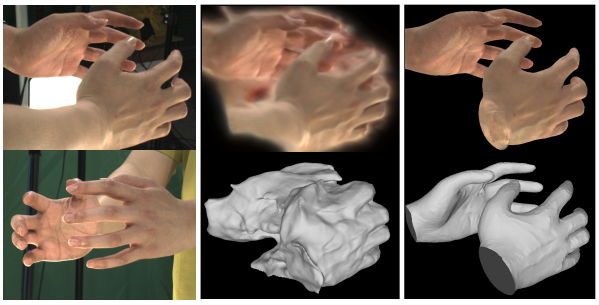
Results of DeepHandMesh. In the DeepHandMesh dataset, five views are utilized to estimate MANO parameters and optimize the mesh. 3D scans of DeepHandMesh is employed to assess the geometric quality. Table 2 reports the mesh geometry quality, and Fig. 5 shows the generated mesh. Unlike DeepHandMesh relying on weakly supervised learning from depth maps, our method leverages the diffuse reflection assumption and incorporates the HAM module to achieve fine-grained hand reconstruction with sparse views. It is challenging for learning-based methods to capture the personalized details on DeepHandMesh, which leads to smoothing results due to their generalization capability. In contrast, the multi-view reconstruction method, NeuS (Wang et al. 2021), fails due to the insufficient number of views.

Results of Our Dataset. Fig. 6 shows the results of our



a) DHM b) Ours w/o. JR c) Ours d) 3D scans

Figure 5: Comparison on mesh quality. The generated meshes are compared in terms of geometric quality using 5 different views on the DeepHandMesh (Moon, Shiratori, and Lee 2020) dataset. JR represents the Joint Refinement.



Images a) NeuS b) Ours

Figure 6: Results of our dataset. To evaluate both geometric accuracy and rendering quality, we compare our framework with NeuS (Wang et al. 2021) on our dataset.

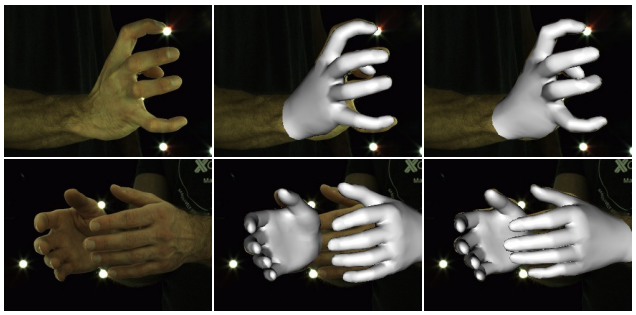
dataset. Our dataset exhibits a more authentic color quality. The experimental results indicate that our proposed approach is effective for various datasets with different capture devices. Due to the absence of hand priors, the mesh generated by NeuS (Wang et al. 2021) may not conform to the characteristics typically associated with hands.

Ablation Study

Coarse Initialization. We directly compare our results with the annotations in the InterHand2.6M dataset. As illustrated in Fig. 7, the MANO model generated by our GCN-based network exhibits a closer alignment with the hands in the images. Moreover, Table 3 demonstrates that directly using the MANO parameters provided by the dataset may introduce errors, which will subsequently affect the rendering quality and reconstruction results.

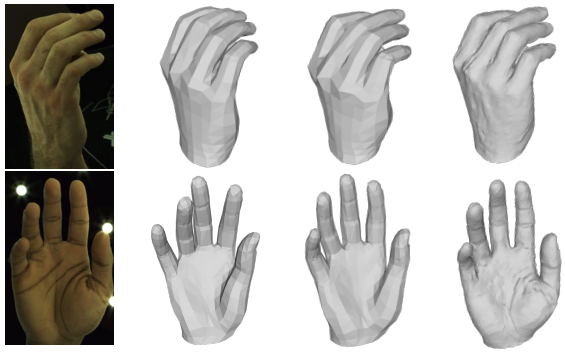
Effects of HAM. The results reported in Fig. 8 provide clear evidence of the substantial impact of mesh subdivision and the HAM module in achieving fine-detailed hand reconstructions. The mesh subdivision process effectively increases the number of vertices in MANO model, while the HAM module plays a crucial role in capturing the surface wrinkles and intricate details of the reconstructed hand. As indicated in Table 3, the inclusion of HAM module greatly contributes to generating the surface reconstructions that closely resemble the real 3D scans.

Texture Optimization and Joint Refinement. Due to variations in viewpoints, the mesh generated by HAM



a) Images b) InterHand2.6M c) Ours

Figure 7: Comparison on MANO meshes. We compare our proposed GCN-based network with the annotations from the InterHand2.6M dataset.



Images a) MANO b) Our Init. c) HAM

Figure 8: The visual results of different methods on mesh geometry quality.

may exhibit artifacts resulting from shading inconsistencies and ambiguous reflectance. Comparison results in Table 3 clearly demonstrate the significant improvement in image quality achieved through joint refinement. The visualizations presented in Fig. 4 further illustrate that neural rendering excels in reconstructing accurate lighting conditions compared to the results of HAM module. Although the geometric optimization achieves the subtle metrics changes in terms of rendering results, it effectively eliminates the non-smoothing singularities, as shown in Fig. 5.

Conclusion

In this paper, we introduced a novel fine-grained multi-view hand mesh reconstruction method by leveraging effective inverse rendering to restore hand poses and intricate details. Our approach predicted a parametric hand mesh model by a GCN-based network while refining both the hand mesh and textures through the Hand Albedo and Mesh (HAM) optimization module. To generate photo-realistic image, we suggested an effective mesh-based neural rendering scheme by fusing the pre-trained rendering network with vertex features. Through extensive experiments on diverse datasets, the promising results demonstrated the efficacy of our proposed approach.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants (62376244, 61831015). It is also supported by the Information Technology Center and State Key Lab of CAD&CG, Zhejiang University.

References

Bhatnagar, B. L.; Sminchisescu, C.; Theobalt, C.; and Pons-Moll, G. 2020. LoopReg: Self-supervised Learning of Implicit Surface Correspondences, Pose and Shape for 3D Human Mesh Registration. In *NeurIPS*, 12909–12922.

Boukhayma, A.; de Bem, R.; and Torr, P. H. 2019. 3D Hand Shape and Pose From Images in the Wild. In *CVPR*, 10835–10844.

Cao, Z.; Radosavovic, I.; Kanazawa, A.; and Malik, J. 2021. Reconstructing Hand-Object Interactions in the Wild. In *ICCV*, 12397–12406.

Chen, J.; Yan, M.; Zhang, J.; Xu, Y.; Li, X.; Weng, Y.; Yi, L.; Song, S.; and Wang, H. 2023. Tracking and Reconstructing Hand Object Interactions from Point Cloud Sequences in the Wild. In *AAAI*, 304–312.

Chen, X.; Jiang, T.; Song, J.; Yang, J.; Black, M. J.; Geiger, A.; and Hilliges, O. 2022. gDNA: Towards Generative Detailed Neural Avatars. In *CVPR*, 20427–20437.

Chen, X.; Wang, B.; and Shum, H.-Y. 2023. Hand Avatar: Free-Pose Hand Animation and Rendering From Monocular Video. In *CVPR*, 8683–8693.

Chen, X.; Zheng, Y.; Black, M. J.; Hilliges, O.; and Geiger, A. 2021. SNARF: Differentiable Forward Skinning for Animating Non-Rigid Neural Implicit Shapes. In *ICCV*, 11594–11604.

Cheng, J.; Wan, Y.; Zuo, D.; Ma, C.; Gu, J.; Tan, P.; Wang, H.; Deng, X.; and Zhang, Y. 2022. Efficient Virtual View Selection for 3D Hand Pose Estimation. In *AAAI*, 419–426.

Choi, H.; Moon, G.; and Lee, K. M. 2020. Pose2Mesh: Graph Convolutional Network for 3D Human Pose and Mesh Recovery from a 2D Human Pose. In *ECCV*, 769–787.

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 *CVPR*, 20501–20511.

Doosti, B.; Naha, S.; Mirbagheri, M.; and Crandall, D. J. 2020. HOPE-Net: A Graph-Based Model for Hand-Object Pose Estimation. In *CVPR*, 6607–6616.

Fan, Z.; Spurr, A.; Kocabas, M.; Tang, S.; Black, M. J.; and Hilliges, O. 2021. Learning to Disambiguate Strongly Interacting Hands via Probabilistic Per-pixel Part Segmentation. In *3DV*, 1–10.

Fu, Q.; Xu, Q.; Ong, Y. S.; and Tao, W. 2022. Geo-neus: Geometry-consistent Neural Implicit Surfaces Learning for Multi-view Reconstruction. *NeurIPS*, 3403–3416.

Ge, L.; Ren, Z.; Li, Y.; Xue, Z.; Wang, Y.; Cai, J.; and Yuan, J. 2019. 3D Hand Shape and Pose Estimation from a Single RGB Image. In *CVPR*, 10833–10842.

Grassal, P.-W.; Prinzler, M.; Leistner, T.; Rother, C.; Nießner, M.; and Thies, J. 2022. Neural Head Avatars from Monocular RGB Videos. In *CVPR*, 18653–18664.

Guo, Z.; Zhou, W.; Wang, M.; Li, L.; and Li, H. 2023. HandNeRF: Neural Radiance Fields for Animatable Interacting Hands. In *CVPR*, 21078–21087.

Hasson, Y.; Tekin, B.; Bogo, F.; Laptev, I.; Pollefeys, M.; and Schmid, C. 2020. Leveraging Photometric Consistency Over Time for Sparsely Supervised Hand-object Reconstruction. In *CVPR*, 571–580.

Hasson, Y.; Varol, G.; Tzionas, D.; Kalevatykh, I.; Black, M. J.; Laptev, I.; and Schmid, C. 2019. Learning Joint Reconstruction of Hands and Manipulated Objects. In *CVPR*, 11807–11816.

Hong, Y.; Peng, B.; Xiao, H.; Liu, L.; and Zhang, J. 2022. HeadNeRF: A Real-time NeRF-based Parametric Head Model. In *CVPR*, 20374–20384.

Horn, B. K. P. 1970. *Shape from shading; a method for obtaining the shape of a smooth opaque object from one view*. Ph.D. thesis, Massachusetts Institute of Technology, USA.

Karunratanakul, K.; Prokudin, S.; Hilliges, O.; and Tang, S. 2023. HARP: Personalized Hand Reconstruction from a Monocular RGB Video. In *CVPR*, 12802–12813.

Kim, D. U.; Kim, K. I.; and Baek, S. 2021. End-to-end Detection and Pose Estimation of Two Interacting Hands. In *ICCV*, 11189–11198.

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

Kipf, T. N.; and Welling, M. 2017. Semi-Supervised Classification with Graph Convolutional Networks. In *ICLR*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollar, P.; and Girshick, R. 2023. Segment Anything. In *ICCV*, 4015–4026.

Kong, D.; Zhang, L.; Chen, L.; Ma, H.; Yan, X.; Sun, S.; Liu, X.; Han, K.; and Xie, X. 2022. Identity-aware Hand Mesh Estimation and Personalization from RGB Images. In *ECCV*, 536–553.

Kulon, D.; Guler, R. A.; Kokkinos, I.; Bronstein, M. M.; and Zafeiriou, S. 2020. Weakly-supervised Mesh-convolutional Hand Reconstruction in the Wild. In *CVPR*, 4990–5000.

Lei, B.; Ren, J.; Feng, M.; Cui, M.; and Xie, X. 2023. A Hierarchical Representation Network for Accurate and Detailed Face Reconstruction from In-The-Wild Images. In *CVPR*, 394–403.

Li, M.; An, L.; Zhang, H.; Wu, L.; Chen, F.; Yu, T.; and Liu, Y. 2022a. Interacting Attention Graph for Single Image Two-hand Reconstruction. In *CVPR*, 2761–2770.

Li, M.; Gao, Y.; and Sang, N. 2021. Exploiting Learnable Joint Groups for Hand Pose Estimation. In *AAAI*, 1921–1929.

Li, T.; Bolkart, T.; Black, M. J.; Li, H.; and Romero, J. 2017. Learning a Model of Facial Shape and Expression from 4D Scans. *ACM TOG*, 194–1.

Li, Y.; Zhang, L.; Qiu, Z.; Jiang, Y.; Li, N.; Ma, Y.; Zhang, Y.; Xu, L.; and Yu, J. 2022b. NIMBLE: a Non-rigid Hand Model with Bones and Muscles. *ACM TOG*, 1–16.

Lin, L.; Peng, S.; Gan, Q.; and Zhu, J. 2024. FastHuman: Reconstructing High-Quality Clothed Human in Minutes. In *3DV*.

Lin, L.; Zhu, J.; and Zhang, Y. 2022. Multiview textured mesh recovery by differentiable rendering. *TCSVT*, 1684–1696.

Liu, L.; Habermann, M.; Rudnev, V.; Sarkar, K.; Gu, J.; and Theobalt, C. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM TOG*, 1–16.

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

Luan, T.; Zhai, Y.; Meng, J.; Li, Z.; Chen, Z.; Xu, Y.; and Yuan, J. 2023. High Fidelity 3D Hand Shape Reconstruction via Scalable Graph Frequency Decomposition. In *CVPR*, 16795–16804.

Mescheder, L.; Oechsle, M.; Niemeyer, M.; Nowozin, S.; and Geiger, A. 2019. Occupancy Networks: Learning 3D Reconstruction in Function Space. In *CVPR*, 4460–4470.

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. *CACM*, 99–106.

Moon, G.; and Lee, K. M. 2020. I2l-meshnet: Image-to-lixel Prediction Network for Accurate 3D Human Pose and Pesh Estimation from a Single RGB Image. In *ECCV*, 752–768.

Moon, G.; Shiratori, T.; and Lee, K. M. 2020. Deephandmesh: A Weakly-supervised Deep Encoder-decoder Framework for High-fidelity Hand Mesh Modeling. In *ECCV*, 440–455.

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 *ECCV*, 548–564.

Noguchi, A.; Sun, X.; Lin, S.; and Harada, T. 2021. Neural Articulated Radiance Field. In *ICCV*, 5762–5772.

Osman, A. A.; Bolkart, T.; and Black, M. J. 2020. STAR: Sparse Trained Articulated Human Body Regressor. In *ECCV*, 598–613.

Pavlakos, G.; Choutas, V.; Ghorbani, N.; Bolkart, T.; Osman, A. A.; Tzionas, D.; and Black, M. J. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *CVPR*, 10975–10985.

Peng, S.; Dong, J.; Wang, Q.; Zhang, S.; Shuai, Q.; Zhou, X.; and Bao, H. 2021a. Animatable Neural Radiance Fields for Modeling Dynamic Human Bodies. In *ICCV*, 14314–14323.

Peng, S.; Zhang, Y.; Xu, Y.; Wang, Q.; Shuai, Q.; Bao, H.; and Zhou, X. 2021b. Neural Body: Implicit Neural Representations with Structured Latent Codes for Novel View Synthesis of Dynamic Humans. In *CVPR*, 9054–9063.

Ren, J.; Zhu, J.; and Zhang, J. 2023. End-to-end weakly-supervised single-stage multiple 3D hand mesh reconstruction from a single RGB image. *CVIU*, 103706.

Ren, P.; Chen, Y.; Hao, J.; Sun, H.; Qi, Q.; Wang, J.; and Liao, J. 2023. Two Heads Are Better than One: Image-Point Cloud Network for Depth-Based 3D Hand Pose Estimation. In *AAAI*, 2163–2171.

Romero, J.; Tzionas, D.; and Black, M. J. 2017. Embodied hands: modeling and capturing hands and bodies together. *ACM TOG*, 245:1–245:17.

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 *ICCV*, 2304–2314.

Walker, T.; Mariotti, O.; Vaxman, A.; and Bilen, H. 2023. Explicit Neural Surfaces: Learning Continuous Geometry With Deformation Fields. *arXiv preprint arXiv:2306.02956*.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS*, 34: 27171–27183.

Weng, C.-Y.; Curless, B.; Srinivasan, P. P.; Barron, J. T.; and Kemelmacher-Shlizerman, I. 2022. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 16210–16220.

Xiu, Y.; Yang, J.; Tzionas, D.; and Black, M. J. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *CVPR*, 13286–13296.

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 *ICCV*, 11354–11363.

Zuffi, S.; Kanazawa, A.; Jacobs, D. W.; and Black, M. J. 2017. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *CVPR*, 6365–6373.