LiveHand: Real-time and Photorealistic Neural Hand Rendering

AKSHAY MUNDRA, Max Planck Institute for Informatics and Saarland University
MALLIKARJUN B R, Max Planck Institute for Informatics
JIAYI WANG, Max Planck Institute for Informatics
MARC HABERMANN, Max Planck Institute for Informatics
CHRISTIAN THEOBALT, Max Planck Institute for Informatics and Saarland University
MOHAMED ELGHARIB, Max Planck Institute for Informatics

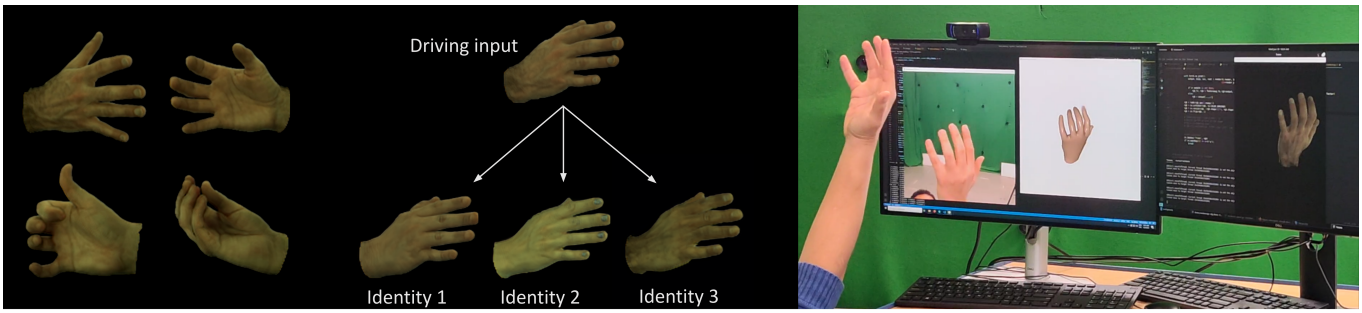

Fig. 1. We present LiveHand, the first neural implicit approach for rendering articulated hands in real-time at unprecedented photorealism. (a) Our method captures pose-dependent effects such as hand shadows, popping veins, and skin wrinkles. (b) We can use the hand-pose obtained from an input sequence to re-enact different learned identities. (c) Our method is designed for optimal rendering speed and quality - we demonstrate this with a live demo where we track the 3D hand-pose and render a photo-realistic hand avatar, all in real-time.

The human hand is the main medium through which we interact with our surroundings. Hence, its digitization is of uttermost importance, with direct applications in VR/AR, gaming, and media production amongst other areas. While there are several works for modeling the geometry and articulations of hands, little attention has been dedicated to capturing photo-realistic appearance. In addition, for applications in extended reality and gaming, real-time rendering is critical. In this work, we present the first neural-implicit approach to photo-realistically render hands in real-time. This is a challenging problem as hands are textured and undergo strong articulations with various pose-dependent effects. However, we show that this can be achieved through our carefully designed method. This includes training on a low-resolution rendering of a neural radiance field, together with a 3D-consistent super-resolution module and mesh-guided space canonicalization and sampling. In addition, we show the novel application of a perceptual loss on the image space is critical for achieving photorealism. We show rendering results for several identities, and demonstrate that our method captures pose- and view-dependent appearance effects. We also show a live demo of our method where we photo-realistically render the human hand in real-time for the first time in literature. We ablate all our design choices and show that our design optimizes for both photorealism and rendering speed. Our code will be released to encourage further research in this area.

Additional Key Words and Phrases: Human digitalization, neural rendering, hand modeling

1 INTRODUCTION

As the popularity of VR/AR technology rises, providing a natural interface with these digital contents becomes vital. Without a doubt, hands are the most intuitive mode of interaction for users in a 3D environment. Therefore, it is quintessential to digitize the users' hands to render their personalized, controllable, and photorealistic counterparts in the virtual world. Achieving this is a challenging task since hand appearance is a complex function that varies with both pose and viewing direction. Moreover, ensuring the real-time performance of such a system is key to enabling applications such as telepresence, teleoperation, and computer-aided design.

While the creation of photorealistic hand models is possible to some extent using traditional computer graphics techniques, it typically requires extensive manual efforts from experienced artists. Therefore, recent research has started to investigate whether hand models can be directly derived from 2D imagery. Here, most existing methods use some data-driven explicit model to constrain the hand geometry and appearance to a low dimensional space for the sake of tractability and robustness to occlusions [Li et al. 2021, 2022; Moon et al. 2020a; Qian et al. 2020; Romero et al. 2017]. Reconstruction is then formulated as a search in this space for the best fitting parameters. Although these approaches can rapidly provide plausible results, the reconstruction is constrained to the space spanned by the registered hand mesh data used to create the model, thus limiting the visual quality and level of personalization.

More recently, neural implicit representations [Mildenhall et al. 2020] have shown impressive results on static scenes for novel-view synthesis. Some works have extended these formulations beyond static scenes to enable photorealistic renderings of articulated objects such as the human body [Habermann et al. 2022, 2021; Liu et al. 2021; Noguchi et al. 2021; Peng et al. 2021a,b; Su et al. 2021a; Yang et al. 2022]. Despite their successes, very little work has been done applying these ideas to hands. In contrast to bodies, hand motions

exhibit more severe self-occlusions and more self-contact, which hinders the learning of scene representations that are consistent across different articulations. One particular work of interest is LISA [Corona et al. 2022], which proposed a method to create neural hand avatars. Although their approach shows promising results, it does not support real-time rendering during inference and the results lack high-frequency details.

In this paper, we propose the first method for creating a *photo-realistic* neural hand avatar, which achieves *real-time* performance while being solely learned from segmented multi-view video of an articulated hand and respective hand pose annotations (see Fig. 1). To this end, we introduce a hybrid hand model representation using the MANO hand model as a coarse proxy, which is surrounded by a neural radiance field. The idea is to simplify the learning problem by bounding the learnable volume through the canonicalization of global coordinates into a texture cube. These normalized coordinates can then be fed into a shallow coordinate-based MLP to regress the scene color and density. This formulation can also leverage the coarse mesh proxy for more efficient sampling of a low-resolution NeRF representation of the scene; we show that this, when combined with a CNN-based super-resolution module carefully designed for efficient upsampling, can achieve real-time performance. Moreover, we found that our highly efficient representation allows training not only on a few ray samples per iteration but on full images. Therefore, we can for the first time supervise an implicit scene representation using a perceptual loss on *full images* during training. Again our experiments show that this greatly improves our results over the baseline, which runs perceptual supervision on a patch basis. Together, these design choices allow us to render and re-enact photo-realistic hands in real-time detailed enough to capture even pose- and view-dependent appearance changes. In summary, our contributions are:

- We propose LiveHand, the first method for real-time neural hand rendering with unprecedented photorealism.
- The real-time performance is achieved with a combination of design choices, namely, training a neural radiance field to capture the scene in low resolution, a mesh-guided 3D sampling strategy, and a 3D-consistent super-resolution module.
- With these computationally-efficient design choices, we for the first time demonstrate that a perceptual loss on the full image can be effectively used for supervising implicit representations.

Our results demonstrate that we clearly outperform the state of the art in terms of visual quality and runtime performance. Moreover, we show a live demo of our approach, which convincingly shows the straightforward use of our method in daily life scenarios. We will release our code and data for future research.

2 RELATED WORKS

In the following, we discuss works on modeling the geometry of hands, as well as works that jointly model the geometry and appearance of hands. Lastly, we review methods of neural rendering for human bodies as the underlying concepts are related to our approach.

Table 1. Conceptual comparison of our method to the previous works HTML [Qian et al. 2020], NIMBLE [Li et al. 2022], and LISA [Corona et al. 2022] in terms of real-time capability, photorealism, pose-dependent appearance modeling, and view-dependent appearance modeling. Note that all other methods fall short in either one or multiple desired aspects while our method supports all of them.

Comparison with other hand-modeling approaches				
Methods	Real-time	Photo-real	Pose-dep. app.	View-dep. app.
HTML [Qian et al.]	✓	✗	✗	✗
NIMBLE [Li et al.]	✓	✓	✗	✗
LISA [Corona et al.]	✗	✗	✓	✗
Ours	✓	✓	✓	✓

Geometry Modeling. Parametric 3D morphable models map low-dimensional control variables to deforming meshes accounting for shape and pose variations, which enables easy and efficient control of the generated geometry. They have been used to model various articulated objects like faces [Li et al. 2017], bodies [Joo et al. 2018; Loper et al. 2015; Osman et al. 2020, 2022; Pavlakos et al. 2019], hands [Romero et al. 2017], and others [Zuffi et al. 2017]. Relevant to our work, MANO [Romero et al. 2017] learns a parametric hand model from high-resolution 3D scans, parametrizing the mesh as a function of the hand shape and pose. PIANO [Li et al. 2021] uses MRI data to develop a parametric bone model of the human hands. Despite their ease of use, these models can not model fine-scale geometry details, owing to the low-dimensional representation. DeepHandMesh [Moon et al. 2020a] addresses this limitation by learning a pose and identity-dependent delta space to overcome the coarse MANO representation to obtain fine-scale details in geometry. In stark contrast to our proposed approach, none of the above methods targets high-quality appearance modeling.

Implicit geometry modeling uses a neural network to encode the geometry as an isosurface. Since the learned representation is resolution-independent, it can - in theory - be used to retrieve meshes at arbitrarily-high resolution at inference time. imGHUM [Alldieck et al. 2021] builds a parametric full-body model comprising of detailed body, face, and hand geometry. GraspingField [Karunratanakul et al. 2020] learns a signed distance function (SDF) of hand-object interaction, which fits the MANO model onto the SDF to recover the final pose estimate. HALO [Karunratanakul et al. 2021] builds a skeleton-driven occupancy model, taking 3D joint locations as input and producing a neural occupancy field. THOR-Net [Aboukhadra et al. 2022] takes a very different approach to modeling the hands with Graph Convolutional Networks (GCNs), allowing them to preserve the topologies of hand poses and shapes. They show the reconstruction of two hands and an object from a single RGB image. None of these geometry modeling methods, however, include a component for the hand texture. In contrast, our goal is to model the photorealistic appearance of the hand at real-time framerates.

Geometry and Appearance Modeling. A few approaches extend parametric meshes by complementing them with a texture map. HTML [Qian et al. 2020] builds a low-dimensional hand appearance model by applying principal component analysis (PCA) to texture

maps of 51 subjects. NIMBLE [Li et al. 2022] uses MRI data to learn a parametric mesh model based on the bones and muscles, and uses light-stage captures to obtain the appearance maps (including albedo, normal maps, and specular maps). A PCA on the various components of appearance maps give them an appearance model. Since both HTML and NIMBLE use a linear model to compress the appearance variations to a low-dimensional space, their expressivity is severely limited. For example, they lack high-frequency details such as veins and colored fingernails since these are person-specific details. DART [Gao et al. 2022] provides a way to overcome this limitation by releasing 325 hand-crafted 2D texture maps for blemishes, make-up, and accessories, which can be added to the hand texture. However, since the texture map does not change based on the hand pose, these approaches can not model hand appearance as a function of hand pose. In contrast, our neural hand model learns pose- and view-dependent appearance effects leading to increased immersiveness and photorealism.

Closest to our approach is LISA [Corona et al. 2022], which models the hand shape and appearance using a neural implicit field. They decompose the hand into individual bones, each of which is encoded with a separate MLP. The MLPs are conditioned on pose and appearance parameters, thus allowing pose and appearance changes at inference. However, the reconstructions lack high-frequency details, and the approach takes about one minute to render an image at a resolution of 1024×667 pixels. On the other hand, we focus on creating a digital hand avatar in a person-specific setup and show photorealistic results in real-time. Please refer to Tab. 1 for a conceptual comparison of the existing hand modeling methods.

Body Modeling. The literature on human body modeling has a similar history to hands. The existing explicit mesh-based methods [Habermann et al. 2021; Xu et al. 2011] rely on a template mesh obtained from a static scene and then learn appearance in the mesh space either by retrieval [Xu et al. 2011] or by using a CNN to directly regress the texture map [Habermann et al. 2021]. The strong reliance on the template mesh has two limitations. First, if the deformed template mesh does not match the real deformation of the surface, the learned appearance will be oversmoothed and blurry. Secondly, the mesh representation is limited by its resolution, which may prevent the capture of high-quality details. In addition to the aforementioned work on parametric body meshes, many approaches have studied implicit geometry modeling of the clothed human body using neural networks [Palafox et al. 2021; Saito et al. 2021; Tiwari et al. 2021; Xiu et al. 2022]. Some approaches decompose the body into individual body parts [Bhatnagar et al. 2020; Biswas et al. 2021; Mihajlovic et al. 2022], allowing a better generalization to out-of-distribution poses. Due to their capabilities of modeling non-rigid clothing deformations well, implicit models are also used to model the geometry and appearance of clothed humans [Habermann et al. 2022; Liu et al. 2021; Noguchi et al. 2021; Peng et al. 2021a,b; Su et al. 2021a; Yang et al. 2022]. However, none of the existing implicit-based body reenactment methods can operate in real-time.

3 METHODOLOGY

Given multi-view images $\{\mathbf{G}_j^p | j = 1 \dots N, p = 1 \dots P\}$ for P frames captured from N viewpoints and coarse estimates of the corresponding posed parametric hand mesh $\{\mathcal{M}(\psi^p) | p = 1 \dots P\}$, our method creates a photo-realistic hand avatar that can accurately capture pose-dependent changes in both geometry and texture, while still modeling view-dependent appearance effects, all in real-time.

An overview of our method is shown in Fig. 2. Given the hand parameters ψ, we can canonicalize every point in the scene based on the point's projection onto the posed mesh $\mathcal{M}(\psi)$. The point coordinates are then re-parameterized in terms of the corresponding texture coordinates after projection. A multi-layer perception (MLP) H_α is then trained to map the re-parameterized coordinates to a radiance field, conditioned on articulation parameters. For the given camera extrinsics and intrinsics, we render low-resolution images and feature maps using volumetric rendering, which is then upsampled using a super-resolution network S_ϕ to obtain the final rendering. In this section, we initially describe the hand model required to build the neural hand representation in Sec. 3.1, the scene representation in Sec. 3.2, and its efficient $2D$ rendering in Sec. 3.3. Finally, in Sec. 3.4, we describe how our neural hand model can be effectively trained.

3.1 MANO Model

We leverage the MANO [Romero et al. 2017] model to parameterize the approximate hand geometry. MANO maps the model parameter ψ to a posed mesh $\mathcal{M}$ using its Linear Blend Skinning (LBS) weights W and a canonical hand mesh $\overline{\mathcal{M}}$.

$$\mathcal{M}(\psi) = MANO(\overline{\mathcal{M}}, \psi, W) \tag{1}$$

$\psi : \{\theta, \beta, t, R\} \in \mathbb{R}^{61}$ consists of the articulation parameters $\theta \in \mathbb{R}^{45}$, shape parameters $\beta \in \mathbb{R}^{10}$, and the global translation $t \in \mathbb{R}^3$ and rotation in axis-angle format $R \in \mathbb{R}^3$. We refer the readers to [Romero et al. 2017] for more details. For convenience, we also define hand pose as $\xi : \{\theta, R\} \in \mathbb{R}^{48}$ here. ξ encodes only the articulation and orientation of the hand, and is, thus, independent of identity and position in global 3D space.

3.2 Implicit Hand Representation

Inspired by the state-of-the-art implicit novel view synthesis method, NeRF [Mildenhall et al. 2020], we model our hand avatar with a view-dependent implicit representation. Since NeRF can only capture static scenes, we must extend the radiance field to account for deformations. In this section, we systematically motivate and describe our chosen representation.

Naive Conditioning. One way to formulate the hand radiance field H_α is by naively conditioning it as follows:

$$H_\alpha : (x, d, \xi) \rightarrow (\mathbf{c}, \sigma) \tag{2}$$

where, x is a point in $3D$ space, d is the viewing direction, ξ is the hand pose, $\mathbf{c}$ is the color and σ is the density. The trainable radiance field H_α is parameterized by an MLP with parameters α.

However, this leads to poor generalization to novel test hand poses as will be shown in Sec. 4. This is because a point on the surface of the hand gets mapped to completely different world coordinates based on the hand pose.

Per-bone Canonicalization. One way to overcome this problem in the literature [Corona et al. 2022] is to canonicalize the scene with respect to the hand pose. Specifically, a point in world space is transformed into each bone's local coordinate systems obtained

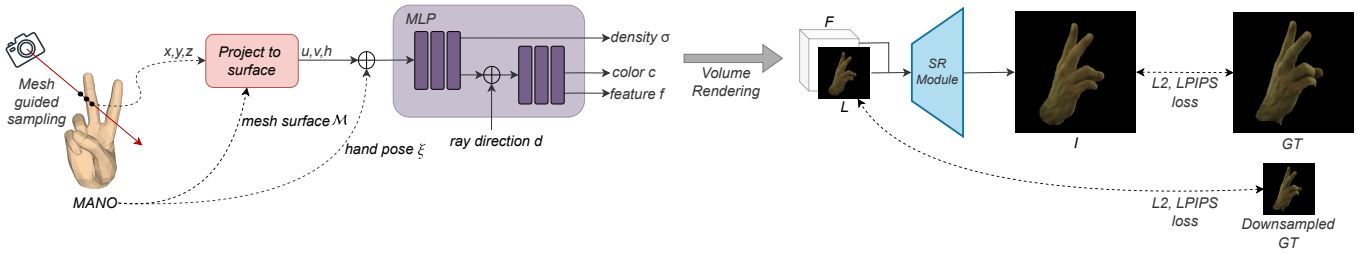

Fig. 2. Overview of our approach. Given a hand pose and camera view, our method renders a photorealistic image of the hand in real-time. To this end, we introduce a mesh-based canonicalization process that transforms the global points into a texture space. The hand appearance is captured by an MLP that maps points from this texture space to radiance values. We then leverage volume rendering to obtain an image-aligned feature tensor where the first three channels contain a low-resolution image of the hand. Finally, an efficient super-resolution module up-samples the low-resolution tensor to obtain the final full-resolution image. Since our method achieves very fast inference speeds, we can supervise it with a perceptual loss on the full image resolution.

from a skeleton pose estimate. Separate implicit fields are learnt in the local coordinate systems, which are combined as follows:

$$\sigma = \sum_{k=1}^{n_b} w_k \sigma_k \ , \quad \mathbf{c} = \sum_{k=1}^{n_b} w_k \mathbf{c}_k \tag{3}$$

where w is analogous to LBS weights. We evaluate such a canonicalization approach in Sec. 4. Such a per-bone canonicalization requires inferring multiple MLPs for each $3D$ point, making it slower for both training and inference.

Mesh-based Canonicalization. To consolidate the multiple MLPs into a more efficient representation, we take inspiration from Neural Actor [Liu et al. 2021] and canonicalize the $3D$ points using the texture space of a mesh surface. Instead of using pre-trained features in the texture space to add details to the radiance field, we use the texture coordinates directly for training. More concretely, for a given point x in $3D$, we find the nearest point on the MANO surface to get its uv co-ordinate as follows:

$$(u, v, l) = \arg\min_{u,v,l} \left\| x - B_{u,v}\left(\mathcal{V}_{[\mathcal{F}(l,\mathcal{M})]}\right) \right\|_2^2 \tag{4}$$

where $l \in \{1 \dots N_T\}$ is the index of mesh triangle, $\mathcal{V}_{[\mathcal{F}(l)]}$ are the three vertex position of the posed triangle $\mathcal{F}(l, \mathcal{M})$, $(u, v) : u, v, u + v \in [0, 1]$ are the barycentric coordinates on the face, and $B_{u,v}(.)$ is the barycentric interpolation function. Additionally, we use the signed distance h of the sampling point to its projection on the mesh to disambiguate points in 3D space. With this canonicalization, we can formulate the radiance field mapping as,

$$H_\alpha : (u, v, h, d) \rightarrow (\mathbf{c}, \sigma) \tag{5}$$

This allows us to canonicalize the world coordinates to a representation that stays consistent with respect to hand surface irrespective of hand pose ξ, thus, preventing the dispersion of learned features in the input space. In practice, we apply positional encoding [Mildenhall et al. 2020] to all inputs described in Eq. 5.

Although Eq. 5 is implicitly a function of ξ due to the dependency of (u, v, h) on the posed mesh, we find that explicit pose conditioning ξ is still needed in order to capture pose-dependent appearance changes. This leads to the modified representation:

$$H_\alpha : (u, v, h, d, \xi) \rightarrow (\mathbf{c}, \sigma) \tag{6}$$

Note that although we rely on the coarse hand mesh for canonicalization, the implicit representation H_α can learn fine-scale details that are hard to model using MANO mesh alone. We show this later in Sec. 4 where our method significantly outperforms a baseline that naively textures the coarse MANO mesh using ground truth images.

3.3 Efficient Rendering

Since H_α is parameterized with an MLP, it can be queried to regress the density σ and color c for each point in $3D$ space. For a ray with origin $\mathbf{o}$ and direction $\mathbf{d}$, volumetric integration - as proposed in NeRF [Mildenhall et al. 2020] - can be used to obtain the integrated color $\mathbf{C}$ for the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, with near and far bounds t_n and t_f as follow:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))dt$$

$$\text{where} \quad T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds). \tag{7}$$

This integral can be approximated through stratified sampling within the bounds. However, such a strategy will waste samples on regions that do not contain useful features. Hierarchical sampling was introduced in NeRF [Mildenhall et al. 2020] to address this inefficiency. However, this involves the use of two MLPs to encode both the coarse and detailed scene, and sampling the scene twice.

3.3.1 Mesh-Guided Sampling. To make the rendering faster, we adopt HDHumans's [Habermann et al. 2022] strategy of using the fitted MANO mesh to efficiently sample points in 3D space. Specifically, to define the bounds of each ray, we use the depth rendering of the coarse mesh to constrain the samples to lie close to the approximate surface. This eliminates coarse-scale MLP needed for hierarchical sampling. We empirically chose 16 as the number of samples to draw per ray as it best trades off image quality and rendering speed.

3.3.2 Super-resolution. Although this efficient sampling strategy improved the run-time, it still could not achieve real-time rendering speeds. We introduce a super-resolution network [Chan et al. 2022] S_ϕ that can super-resolve the rendered output in a $3D$ consistent manner. To do so, we first modify the H_α to additionally predict a

29-channel $\mathbf{f}$, which encodes scene features alongside the color to capture additional details for super-resolution. We accomplish this by extending Eq. 5 with:

$$H_\alpha : (u, v, h, d, \xi) \rightarrow (\mathbf{c}, \mathbf{f}, \sigma) \tag{8}$$

We then apply volumetric integration as done in Eq. 7 to obtain low-resolution renderings of color L_j^p and features F_j^p for each viewpoint j and hand pose p.

These low-resolution encodings are used in a super-resolution module

$$S_\phi : (L, F) \rightarrow I \tag{9}$$

to recover a high-resolution image I_j^p that preserves the details. To ensure this module is efficient, we parameterize S_ϕ using a CNN-based network with the trainable parameters ϕ.

3.4 Training

As described in the previous section, we need to learn the parameters of the MLP H_α and super-resolution module S_ϕ using the multi-view image sequence.

Color Calibration. As most multi-view images in general are not color corrected to be consistent across views, we compensate for this, as done in Neural Volumes [Lombardi et al. 2019], by learning separate per-camera gain and bias parameters g_j and b_j.

Objective Function. We train the parameters of our modules H_α and S_ϕ in a supervised manner using the following loss functions

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{perc} \tag{10}$$

between ground truth target image $\mathbf{G}_j^p$ and rendering image I_j^p using gradient descent. Here $\mathcal{L}_{rec}$ is the L2 reconstruction loss given by:

$$\mathcal{L}_{rec} = ||G_j^p - I_j^p(\alpha, \phi)||_2 \tag{11}$$

To capture the perceptual difference in the image, we apply $\mathcal{L}_{perc}$ as suggested in [Zhang et al. 2018]

$$\mathcal{L}_{perc} = ||f(G_j^p) - f(I_j^p(\alpha, \phi))||_2 \tag{12}$$

Where $f(\cdot)$ is the activation of the $conv1$-$conv5$ layers in pre-trained VGG network [Simonyan and Zisserman 2014]. We show later in the results section that the perceptual loss plays a vital role in recovering high-frequency details. Here, thanks to our design choices, we were able to apply the perceptual loss in a computationally efficient manner on the full image resolution. We will show later that this novel application of the perceptual loss improves photorealism over using a traditional patch-based strategy.

We employ the above loss functions to both low-resolution volumetrically rendered images and super-resolved high-resolution images. Note that although EG3D [Chan et al. 2022] also uses a 3D-consistent super-resolution module, it is trained in an adversarial manner using a very large training dataset. In contrast, we show that this module can also be trained in absence of a large amount of training data by combining the above losses for supervised training. We also show that with our carefully designed solution, we can achieve photorealistic rendering in real-time for the first time in literature.

3.5 Implementation Details

Here we provide more details about the implementation. Our density field network is parameterized by a 4 layer-deep MLP and the color and feature network is parameterized by a 2 layer-deep MLP. We use a similar CNN network architecture as in EG3D [Chan et al. 2022] as our super-resolution network with an upsampling factor of 2. All our experiments use 16 samples per ray for volumetric integration. We learn our radiance field module H_α, super-resolution module S_ϕ, and color calibration parameters g_j, b_j with learning rates of 0.0025, 0.0025, and 0.0001 respectively using Adam optimizer [Kingma and Ba 2015] with a decay rate of 0.1.

4 EXPERIMENTS

We use the publicly released version of the InterHand2.6M benchmark for our experiments. The dataset contains multi-view sequences of different users performing a wide range of actions at 5fps and 512×334 resolution. To test our method, we select the right-hand sequences from four users in the "train/capture0", "train/capture5", "test/capture0", and "test/capture1" subsets. We reserve the last 50 frames of each capture for evaluation and use the rest for training.

We show that the advantages of our proposed model work synergistically together to enable the first system that can perform real-time photorealistic neural hand reenactment. The details of this demo application and its results are presented in Section 4.1. We additionally provide quantitative and qualitative evaluations of our method on the established benchmark in Section 4.2 and Section 4.3. For this, we used PSNR, SSIM, LPIPS, and FID metrics for numerical evaluation. Note that following the conventions of Zhang et al. [2018], LPIPS score is calculated using AlexNet backbone. For rendering speed, we report the time it takes to render an image with the same resolution as the training data (512×334) in frames per second (FPS). All FPS experiments are performed on an NVIDIA GeForce RTX 3090. Results show that our method outperforms existing state-of-the-art methods by a large margin and that each component is necessary to achieve this result.

4.1 Application: Real-time Hand Reenactment

We carefully design our method specifically for real-time hand reenactment applications. After training our neural implicit representation H_α and the super-resolution module S_ϕ to create a user's hand avatar, we can drive the articulation of that hand using new motion. Fig. 3 show this transfer of hand performance from a reference user ('Reference') to 4 learned identities. Note that our approach is able to generalize well across identities even when the driving poses were not seen during training. It is also interesting to see that the avatar of each identity is able to capture high-frequency details such as the veins and the bone structure, as well as person-specific pose-dependent changes such as skin wrinkles. All these intricate details contribute to the photo-realism of our rendering.

To show that this method can work in real applications, we also implemented a live demo. This application consists of two parts: a hand tracker which estimates a posed MANO mesh, and a hand avatar trained using our methods on InterHand2.6M. For the pose estimator, we used the work of Zhou et al. [2020]. This method takes a stream of the monocular image from a webcam and estimates the

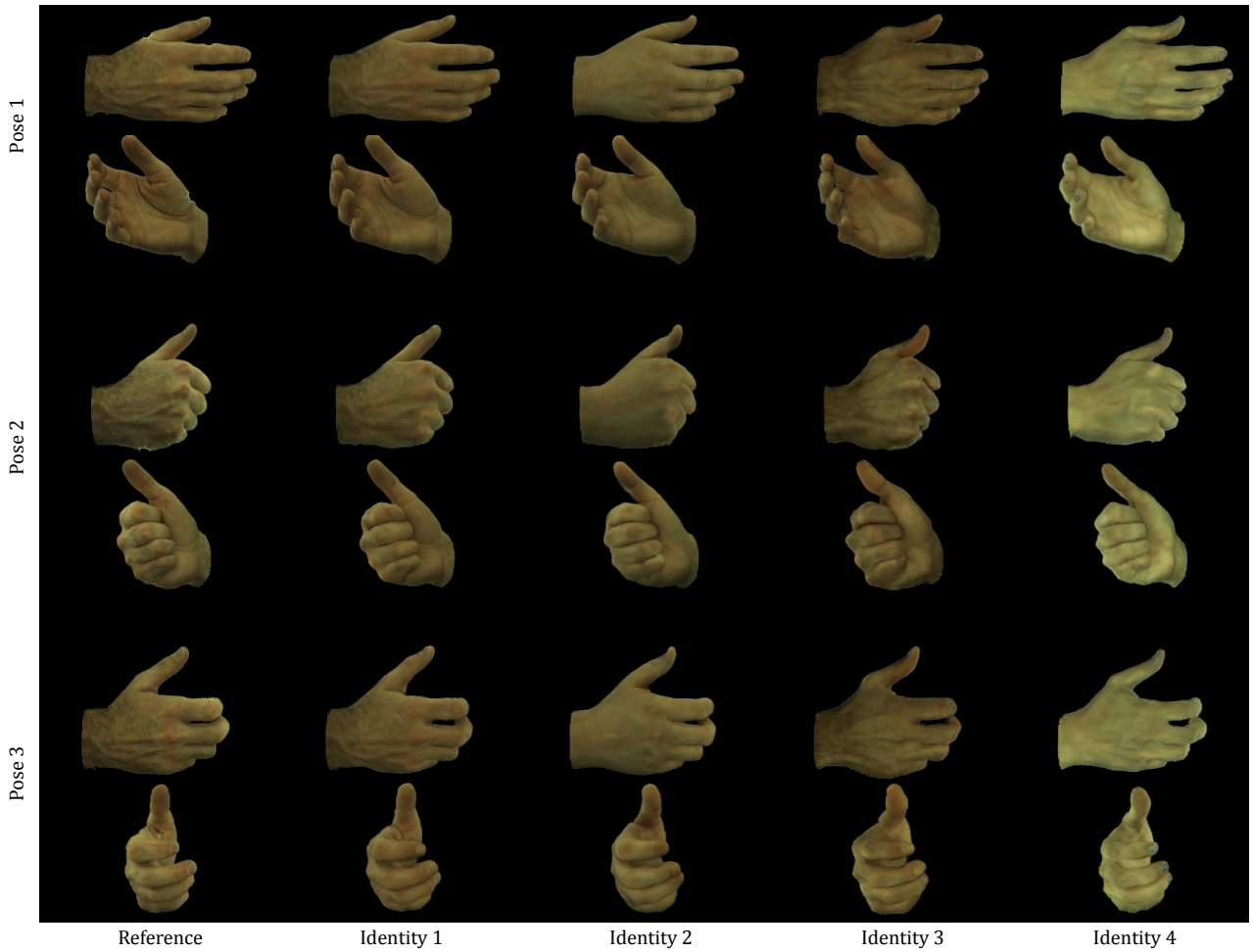

Fig. 3. Hand Reenactment: Our method can transfer the pose of a reference actor (Reference) to new identities (Identity 1-4). Note that our model captures pose-dependent changes, which is especially apparent for veins and in the knuckle region. It also captures view-dependent shading and self-shadowing effects.

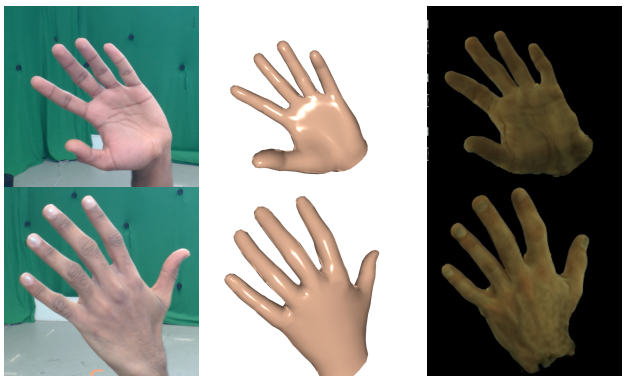

Fig. 4. Demo Visualization: The demo takes in a monocular RGB input (Left) to estimate the shape and pose of the hand (Center). This coarse mesh is then used in our method to transfer the pose to the target identity (Right).

corresponding MANO hand parameters. The estimated parameters are then used to pose and render our hand avatar at 512×334 pixels resolution. The pose estimator takes 10 milliseconds while rendering our hand avatar takes 20 milliseconds on average, giving our system an effective speed of 33 FPS.

We show the qualitative results of this demo in Fig. 4. Note the plausible high-frequency details of the rendered hand avatar driven by new poses captured live in the monocular RGB stream. We encourage the readers to check the supplementary video for the demo, as well as $3D$ consistent rendering sequences with view-dependent effects.

4.2 Comparison to SoTA

The only other neural implicit hand model that exists in the literature is the work of Corona et al. [2022] (LISA). As their method is trained and evaluated on an unreleased high-resolution version of the Interhand2.6M dataset and the code is not publicly available, we re-implemented their method for a fair comparison. As an additional

baseline, we use the body modeling method of Su et al. [2021b] (A-NeRF) and adapt it for hand modeling. Because our method requires a coarse hand mesh for canonicalization, we also compare against a baseline explicit method that re-textures this mesh using a pre-estimated texture map ('Mesh wrapping'). For this, we extract the texture from a flat-hand pose and wrap it to the target poses.

Table 2. Comparison on InterHand2.6M [Moon et al. 2020b]. * indicates we use our implementation of the approach. Note how our method reaches the best balance between rendering quality and speed.

	PSNR ↑	SSIM ↑	LPIPS(x1000) ↓	FID ↓	FPS ↑
Mesh wrapping	28.28	0.94	49.44	298.28	**82.33**
A-NeRF* [Su et al. 2021b]	28.07	0.76	94.41	318.61	0.83
LISA* [Corona et al. 2022]	29.36	0.82	78.46	255.43	3.70
Ours	**32.04**	**0.95**	**25.73**	**197.39**	45.45

As shown in Table 2, our method outperforms all other neural implicit baselines while being much faster. These improvements in the metrics also translate to significant improvements in perceptual quality on the test set, which can be seen in Figure 5. We hypothesize that this is owing to our improved canonicalization strategy and our use of perceptual loss. Both A-NeRF and LISA use per-part canonicalization similar to the one described in Eq.3. However, learning to combine per-part output is not trivial, and could lead the ambiguities in case of severe articulations. Also as we will show in Sec. 4.3, our addition of a perceptual loss drastically improved the level of detail the model can capture over those obtained from simple per-pixel loss used in A-NeRF and LISA.

Our method also significantly outperforms the mesh wrapping baseline, quantitatively and qualitatively. Note that modern graphics pipelines can achieve much higher frame rates for mesh rendering based on their implementation, and we only benchmark ours. However, by no means such a simple rendering can achieve the complex appearance effects and photorealism as our method can. This demonstrates that our model can learn improvements upon what is possible using only the coarse geometric initialization.

4.3 Ablation

Table 3. Ablation study on various canonicalization strategies.

	PSNR ↑	SSIM↑	LPIPS(x1000) ↓	FID ↓	#parameters ↓	FPS ↑
xyz	29.31	0.80	42.50	247.77	0.95M	43.03
per-bone xyz	**32.51**	**0.95**	**23.82**	198.95	1.14M	27.04
uvh w.o. pose cond.	30.33	0.86	32.36	204.24	**0.41M**	**45.73**
Ours (uvh w. pose cond.)	32.04	**0.95**	25.73	**197.39**	**0.41M**	45.45

Our design choices are crucial for optimizing both the rendering quality and processing speed. To evaluate their significance, we performed an ablation study of the different components of our method. First, we report the impact of different canonicalization strategies on the metrics in Tab. 3 and on visual quality in Fig. 6. We see that naive pose conditioning ('xyz') performs the worse in all metrics, and the results are blurry and indistinct. While per-bone canonicalization ('per-bone xyz') produces high-quality renderings, our formulation is 1.7 times faster as it does not rely on the evaluation of multiple MLPs. Finally, our experiments show that without pose conditioning ('uvh w.o. pose cond.'), the performance of our method drops. Pose conditioning is vital for capturing pose-dependent effects such as self-shadowing and skin wrinkles, and this can be seen in Fig. 6.

Table 4. Ablation study on model components.

		PSNR ↑	SSIM ↑	LPIPS(x1000) ↓	FID ↓	FPS ↑
w.o. mesh-guided samp.		31.25	0.72	25.95	202.40	9.07
w.o. SR	w.o. $\mathcal{L}_{perc}$	**32.69**	**0.96**	38.45	226.78	19.64
	patch $\mathcal{L}_{perc}$	30.52	0.50	31.13	197.70	19.52
	full $\mathcal{L}_{perc}$	31.61	0.93	26.63	**197.35**	19.37
Ours (full $\mathcal{L}_{perc}$)		32.04	0.95	**25.73**	197.39	**45.45**

We evaluated the impact of mesh-guided sampling by defaulting to hierarchical sampling instead ('w.o mesh-guided samp.'). While this produces similar rendering quality, it can be seen in Tab. 4 that mesh-guided sampling is 5 times faster. We also evaluated the impact of the superresolution module (SR) by training our method to directly render the full-resolution image instead (w.o. SR). For this experiment, we investigated 3 different settings: we remove $\mathcal{L}_{perc}$ entirely and train with only $\mathcal{L}_{rec}$ ('w.o. $\mathcal{L}_{perc}$'); we implemented the existing patch perceptual loss (e.g. used in Weng et al. [2022]) where randomized crops of 64×64 resolution are used in the perceptual loss function instead ('patch $\mathcal{L}_{perc}$'); finally, we used the perceptual loss on the full images to train the neural representation (full $\mathcal{L}_{perc}$). Tab. 4 shows that the SR module makes our method 2.4 times faster for all variants. Although the method w.o. $\mathcal{L}_{perc}$ achieved the highest PSNR and SSIM, adding any form of $\mathcal{L}_{perc}$ greatly increases the level of details captured in the model (see Fig. 7). This increase in realism is captured quantitatively by the lower LPIPS and FID in Tab. 4, which better reflects human preference. Furthermore, we show our novel application of the perceptual loss on the full image enabled by our efficient formulation ('full $\mathcal{L}_{perc}$') greatly improves the rendering quality on all metrics (see Tab. 4). This improvement is also captured in the qualitative results in Fig. 7. Finally, our full method with the SR module ('Ours') achieves superior or comparable rendering quality while being significantly faster.

Overall, it is clear that our design choices optimize both rendering quality and speed, thus enabling us to photo-realistically render human hands in real-time for the first time in literature.

4.4 Additional Application: Shape Editing

Our UVD encoding and the mesh-guided sampling formulation are not only advantageous in terms of rendering speed and quality, but they also enables easy editing of the hand-avatar geometry. Given the original hand parameter $\psi_{\text{init}} : \{\theta, \beta_{init}, t, R\}$, we can modify the shape parameter to obtain $\psi_{\text{new}} : \{\theta, \beta_{new}, t, R\}$. By using the corresponding mesh $\mathcal{M}(\psi_{\text{new}})$ in the canonicalization procedure as described in Eq. 5, the rendered hand appearance will change accordingly. This allows the geometry of the hand avatar to be modified without retraining. We show the results of this application in Fig. 8, where we modified the first principal component of the MANO shape parameter β.

Fig. 5. Comparison to SoTA on unseen hand poses. A-NeRF and Mesh wrapping produce clear artifacts while LISA does not capture high-frequency details. Our method clearly outperforms these approaches and captures high-frequency details.

5 CONCLUSION AND FUTURE WORK

We presented the first neural implicit approach that can render human hands in a photorealistic manner in real-time. Our approach is carefully designed in a way to optimize the rendering quality and speed. At the heart of our method is a low-resolution NeRF rendering together with a super-resolution module that produces 3D-consistent results. Here, we showed that a novel application of the perceptual loss on the full image space is important for high-quality photorealistic rendering, which was only possible because of our design choices which made the model efficient. To ensure our model can capture pose-dependent effects, the representation is also conditioned on the MANO pose parameters. We also utilize the MANO hand mesh to guide the sampling of points in 3D space to better improve the rendering speed. Results show that our method generates a wide variety of hand articulations, capturing high-frequency skin textural changes arising from skin wrinkles and hand veins. Comparison with related methods clearly shows that our approach outperforms the baselines by a significant margin. In addition, we show the novel application of editing the hand geometry while keeping the texture fixed.

While our work is an important milestone for the full digitization of human hands, there are still several avenues for future work. Since our work requires a MANO mesh, future work could look into improving the quality of such a mesh and corresponding pose parameters. This could include refining the geometry, possibly in an end-to-end manner. Another more strategic direction moving forward is to learn a generalizable implicit 3D morphable model of the human hands that is photoreal. This will give full access to all the hand semantics. While our approach models hand-pose dependant shadow effects, it can not model shadow as a function of any random illumination condition other than the one under the training set that was captured. We leave this modeling for future works. We hope our work encourages research into the important

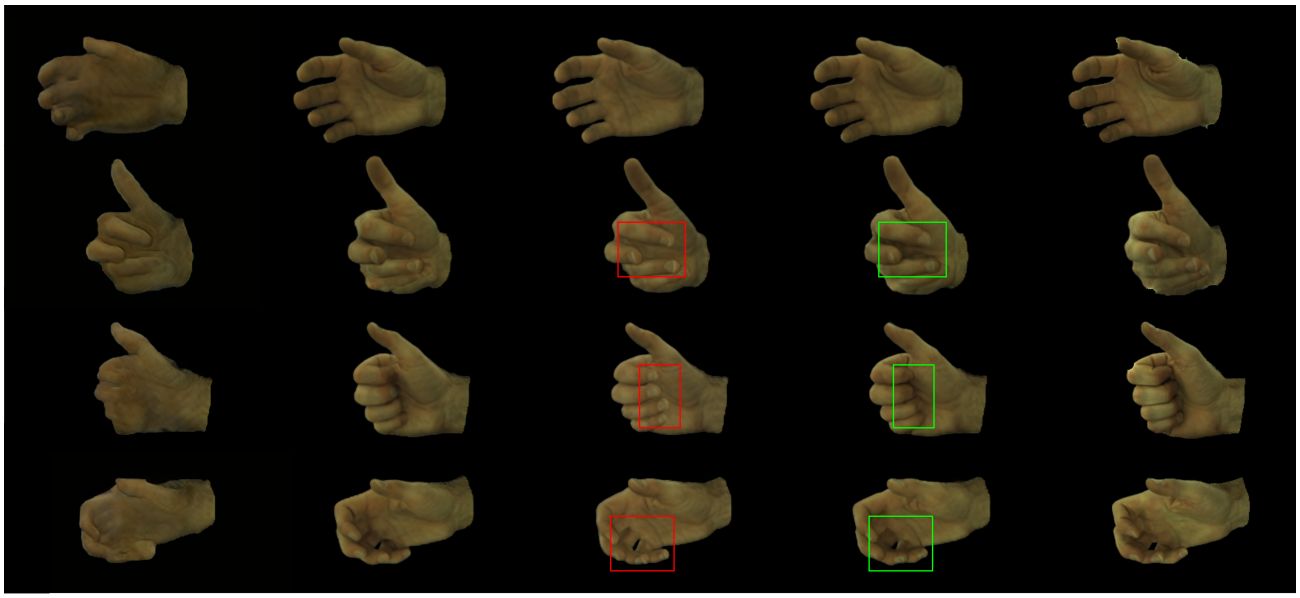

Fig. 6. Canonicalization Ablation. Global xyz coordinates with naive conditioning does not generalize well to novel poses. Our proposed uvh canonicalization achieves similar visual results to per-bone xyz canonicalization while being much faster. It is interesting to note that hand pose conditioning is vital for capturing pose-dependent effects such as self-shadowing (see red and green regions).

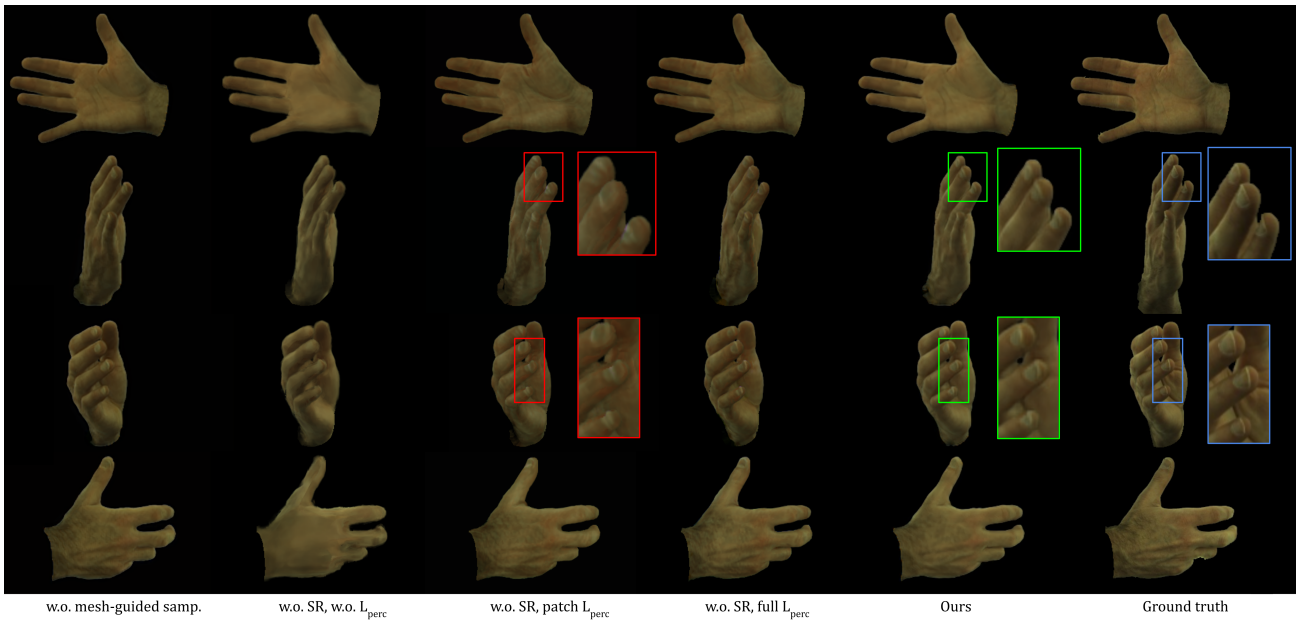

Fig. 7. Model Ablation. Mesh-guided sampling helps improve the inference speed without sacrificing the quality ('w.o. mesh-guided samp.' vs 'Ours'). Without any perceptual loss ('w.o. SR, w.o. L_{perc}'), the reconstructed images are overly smooth and lack details. Although patch-based perceptual loss ('w.o. SR, patch L_{perc}') helps, the details are often incorrect, as highlighted in the figure with red. In contrast, the model trained with full-image perceptual loss ('w.o. SR, full L_{perc}') captures this correctly. Our final method makes use of a super-resolution module to further improve the rendering speed while retaining fine details (notice how the green highlights match the ground truth blue highlights).

Fig. 8. Application: Shape Editing. The hand geometry can be edited without any additional retraining of the model.

problem of photorealistic rendering of the human hands. For this, we will release our code.

Acknowledgements

We thank Ashwath Shetty, Yiming Wang, Oleksandr Sotnychenko, and Basavaraj Sunagad for their help; the MPII IST department for the technical support. This work was supported by the ERC Consolidator Grant 4DRepLy (770784).

REFERENCES

Ahmed Tawfik Aboukhadra, Jameel Malik, Ahmed Elhayek, Nadia Robertini, and Didier Stricker. 2022. THOR-Net: End-to-end Graformer-based Realistic Two Hands and Object Reconstruction with Self-supervision. *arXiv e-prints* (2022), arXiv–2210.

Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. 2021. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5461–5470.

Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. 2020. Combining Implicit Function Learning and Parametric Models for 3D Human Reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer.

Sourav Biswas, Kangxue Yin, Maria Shugrina, Sanja Fidler, and Sameh Khamis. 2021. Hierarchical Neural Implicit Pose Network for Animation and Motion Retargeting. arXiv:2112.00958 [cs.CV]

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. In *CVPR*.

Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. 2022. LISA: Learning Implicit Shape and Appearance of Hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20533–20543.

Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. 2022. DART: Articulated Hand Model with Diverse Accessories and Rich Textures. *arXiv preprint arXiv:2210.07650* (2022).

Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. 2022. HDHumans: A Hybrid Approach for High-fidelity Digital Humans. *arXiv preprint arXiv:2210.12003* (2022).

Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. 2021. Real-time Deep Dynamic Characters. *ACM Transactions on Graphics* 40, 4, Article 94 (aug 2021).

Hanbyul Joo, Tomas Simon, and Yaser Sheikh. 2018. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8320–8329.

Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. 2021. A Skeleton-Driven Neural Occupancy Representation for Articulated Hands. In *International Conference on 3D Vision (3DV)*.

Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. 2020. Grasping Field: Learning Implicit Representations for Human Grasps. In *2020 International Conference on 3D Vision (3DV 2020)*. IEEE, Piscataway, NJ, 333–344. https://doi.org/10.1109/3DV50981.2020.00043

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (2017), 194:1–194:17. https://doi.org/10.1145/3130800.3130813

Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. 2021. PIANO: A Parametric Hand Bone Model from Magnetic Resonance Imaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. 816–822. https://doi.org/10.24963/ijcai.2021/113

Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. 2022. NIMBLE: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–16.

Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. 2021. Neural Actor: Neural Free-view Synthesis of Human Actors with Pose Control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)* (2021).

Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. 2019. Neural Volumes: Learning Dynamic Renderable Volumes from Images. *ACM Trans. Graph.* 38, 4, Article 65 (July 2019), 14 pages.

Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)* 34, 6 (Oct. 2015), 248:1–248:16.

Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. 2022. COAP: Compositional Articulated Occupancy of People. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13201–13210.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *ECCV*.

Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. 2020a. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*. Springer, 440–455.

Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. 2020b. InterHand2.6M: A Dataset and Baseline for 3D Interacting Hand Pose Estimation from a Single RGB Image. In *European Conference on Computer Vision (ECCV)*.

Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. 2021. Neural Articulated Radiance Field. In *International Conference on Computer Vision*.

Ahmed A A Osman, Timo Bolkart, and Michael J. Black. 2020. STAR: A Sparse Trained Articulated Human Body Regressor. In *European Conference on Computer Vision (ECCV)*. 598–613. https://star.is.tue.mpg.de

Ahmed A A Osman, Timo Bolkart, Dimitrios Tzionas, and Michael J. Black. 2022. SUPR: A Sparse Unified Part-Based Human Body Model. In *European Conference on Computer Vision (ECCV)*. https://supr.is.tue.mpg.de

Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. 2021. Npms: Neural parametric models for 3d deformable shapes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12695–12705.

Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. 2019. Expressive Body Capture: 3D Hands, Face, and Body from a Single Image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. 10975–10985.

Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. 2021a. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14314–14323.

Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. 2021b. 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.

Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. 2020. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision, (ECCV)*. Springer, 54–71.

Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* 36, 6 (Nov. 2017), 245:1–245:17. https://doi.org/10.1145/3130800.3130883

Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. 2021. SCANimate: Weakly Supervised Learning of Skinned Clothed Avatar Networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021a. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems* 34 (2021), 12278–12291.

Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. 2021b. A-NeRF: Articulated Neural Radiance Fields for Learning Human Shape, Appearance, and Pose. In *Advances in Neural Information Processing Systems*.

Garvita Tiwari, Nikolaos Sarafianos, Tony Tung, and Gerard Pons-Moll. 2021. Neural-GIF: Neural Generalized Implicit Functions for Animating People in Clothing. In *International Conference on Computer Vision (ICCV)*.

Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. 2022. HumanNeRF: Free-Viewpoint Rendering of Moving People From Monocular Video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16210–16220.

Yuliang Xiu, Jinlong Yang, Dimitrios Tzionas, and Michael J. Black. 2022. ICON: Implicit Clothed humans Obtained from Normals. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 13296–13306.

Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. 2011. Video-based Characters – Creating New Human Performances from a Multi-View Video Database. *ACM Transactions on Graphics* 30 (07 2011), 32. https://doi.org/10.1145/2010324.1964927

Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. 2022. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2863–2873.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Monocular Real-Time Hand Shape and Motion Capture Using Multi-Modal Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 2017. 3D Menagerie: Modeling the 3D Shape and Pose of Animals. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2017*. IEEE, 5524–5532.