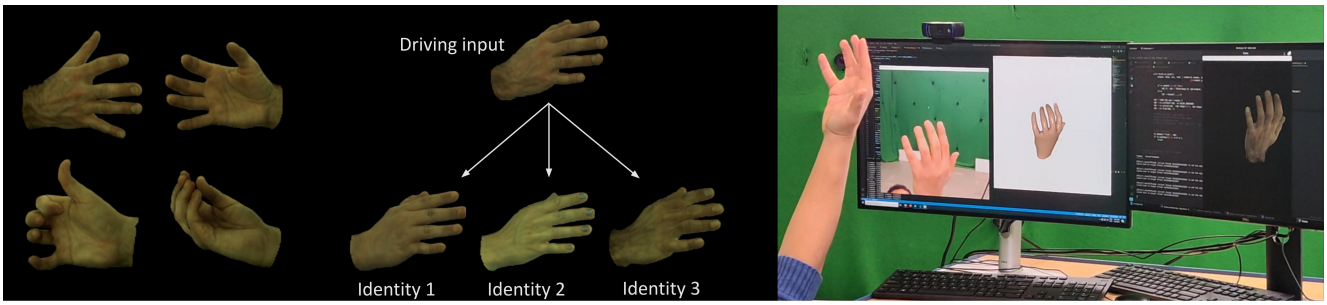
arXiv:2302.07672v3 [cs.GR] 20 Aug 2023

LiveHand: Real-time and Photorealistic Neural Hand Rendering

Akshay Mundra [1,2], Mallikarjun B R [1], Jiayi Wang [1],
Marc Habermann [1], Christian Theobalt [1,2], Mohamed Elgharib [1]
[1] Max Planck Institute for Informatics [2] Saarland University
mundra.akshay15@gmail.com, {mbr,jwang,mhaberma,theobalt,elgharib}@mpi-inf.mpg.de



(a) Pose variations (b) Re-enactment (c) Real-time demo

Figure 1: We present LiveHand, the first neural implicit approach for rendering articulated hands in real-time. (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 identities. (c) Our method is designed to optimize rendering speed and quality – we show a live demo where we track the 3D hand-pose and render a photo-realistic hand avatar, all in real-time.

Abstract

The human hand is the main medium through which we interact with our surroundings, making its digitization an important problem. While there are several works modeling the geometry of hands, little attention has been paid to capturing photo-realistic appearance. Moreover, for applications in extended reality and gaming, real-time rendering is critical. 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 pose-dependent effects. However, we show that this aim is achievable 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 sampling and space canonicalization. We demonstrate a novel application of perceptual loss on the image space, which is critical for learning details accurately. We also show a live demo where we photo-realistically render the human hand in real-time for the first time, while also modeling pose- and view-dependent appearance effects. We ablate all our design choices and show that they optimize for rendering speed and quality. Video results and our code can be accessed from https://vcai.mpi-inf.mpg.de/projects/LiveHand/

1. Introduction

As the popularity of VR/AR technology rises, providing a natural interface with these digital contents becomes vital. Undoubtedly, 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 varying with both pose and viewing direction. Moreover, ensuring 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 [35, 32, 24, 16, 17]. 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 [23] 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 [38, 30, 26, 18, 28, 42, 10, 9]. 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 representation that is consistent across different articulations. One particular work of interest is LISA [6], 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 *photorealistic* neural hand avatar, which achieves *real-time* performance while being solely learned from segmented multi-view videos 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 photorealistic neural hand rendering.

Methods	Real-time	Photo-real	Pose-dep. app.	View-dep. app.
HTML [32]	✓	✗	✗	✗
NIMBLE [17]	✓	✓	✗	✗
LISA [6]	✗	✗	✓	✗
Ours	✓	✓	✓	✓

Table 1: Conceptual comparison of our method with other hand-modeling approaches.

- The real-time performance is achieved with our careful combination of design choices, namely, a mesh-guided 3D sampling strategy, a low-resolution neural radiance field, 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 and that it out-performs the commonly used patch-based loss.

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.

2. Related Works

Geometry Modeling. Parametric 3D morphable models map low-dimensional control variables to deforming meshes enabling easy and efficient control of the generated geometry [15, 27, 35, 1]. Relevant to our work, MANO [35] learns a parametric hand model using high-resolution 3D scans, parametrizing the mesh as a function of the hand shape and pose. 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 [1] builds a parametric full-body model comprising of detailed body, face, and hand geometry. GraspingField [14] learns a signed distance function (SDF) of hand-object interaction, which fits the MANO model onto the SDF to recover the final pose estimate. However, none of the existing works [35, 16, 24, 14, 13] include a component for the hand texture. In contrast, our goal is to model the photorealistic hand appearance in real-time.

Geometry and Appearance Modeling. A few approaches extend parametric mesh by complementing it with a texture map. HTML [32] builds a low-dimensional hand appearance model by applying principal component analysis (PCA) to texture maps of 51 subjects. NIMBLE [17] 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 gives 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 details such as veins and colored fingernails since these are person-specific attributes. Closest to our approach is LISA [6], which models the hand shape and appearance using a neural implicit field. The underlying MLPs are conditioned on pose and appearance parameters, 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 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.

Other Animatable Objects. The literature contains works for modeling other animatable objects such as the human face [20, 4, 44, 8, 7, 22], human body [10, 41, 3, 18, 42, 30, 38], and animals [21]. The face related methods can not handle large deformations [20, 4, 7, 22] and/or are not real-time [8, 44, 7], while [21] does not model pose-dependent appearance effects. To handle more articulated motions that occur in the human body, two classes of body-specific methods have been proposed. The explicit mesh-based methods [41, 10, 2] rely on a template mesh obtained from a static scene and then learn appearance in the mesh space either by retrieval [41] or by using a CNN to directly regress the texture map [10, 2]. However, due to the strong reliance on a template mesh, the learned appearance becomes blurry if the deformed template mesh does not match the real deformation of the surface. In contrast, neural implicit models have the capacity to learn more fine-grained deformations at much higher resolution. For example, it has been used to model the geometry and appearance of clothed humans [38, 30, 26, 18, 28, 42, 9, 11, 29, 12]. However, these can not operate in real-time. Some efficient approaches [31, 33] formulate the rendering task as an image-translation problem, but suffer from inaccuracies in parametric model fitting. Yet another line of implicit body-modeling approaches [34, 39, 36] require RGB images from multiple cameras at test time, and thus can not be controlled with arbitrary poses. Extending the body modeling methods to human hands is not trivial, as hands exhibit even stronger articulation, which in turn results in severe self-occlusion and other pose-dependent effects. Our proposed method tackles this setting by utilizing elements from both mesh-based and neural implicit modeling to create a detailed model that runs 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 the corresponding coarse parametric hand meshes $\{\mathcal{M}(\psi^p) | p = 1 \dots P\}$, our method creates a photo-realistic hand avatar that can accurately model hand-pose and view-dependent appearance effects, and can be rendered 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 3D 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 image-aligned feature maps using volumetric rendering, which is then up-sampled 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 [35] 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 [35] 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 [23], 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:

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

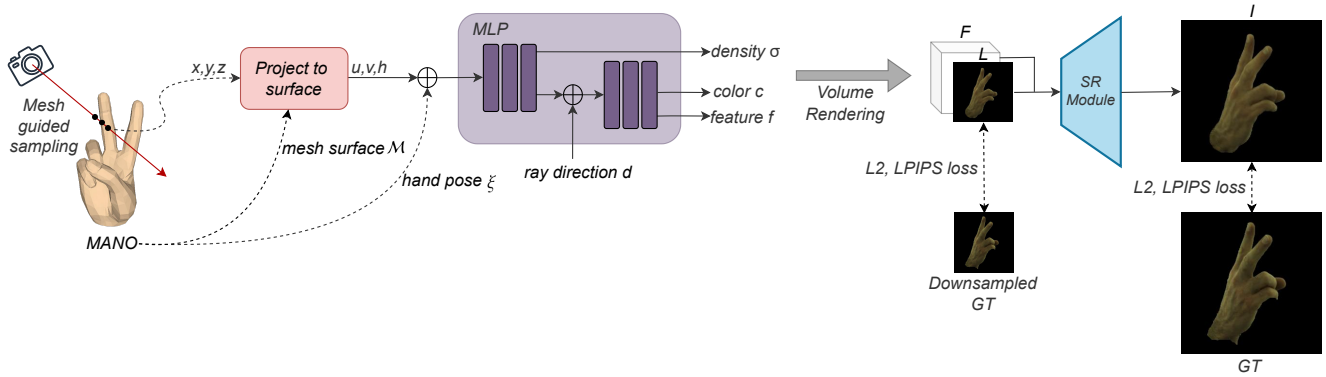


Figure 2: **Overview of our approach.** Given a hand pose and camera view, our method renders a photorealistic image of the hand in real-time. To this end, we employ an efficient MANO mesh-guided sampling and canonicalization strategy. The hand appearance is captured by an MLP that maps points from the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

where x is 3D point, 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 any point on the hand surface gets mapped to completely different world coordinates based on the hand pose.

Per-bone Canonicalization. One way to overcome this problem in the literature [6] 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 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 Canoncialization. For a more efficient representation, we take inspiration from mesh-based texturing which associates each point on the mesh surface with a 2D texture coordinate $(u, v) \in [0, 1] \times [0, 1]$ from which a color value can be obtained using a texture image. We extend this surface representation to 3D volumes by introducing a signed distance h to support volume rendering and to account for the coarseness of the MANO-based geometry approximation. More concretely, for a given point x in 3D, we first find its projection on the given MANO surface. The (u, v) co-ordinate of this projected point can be estimated by performing barycentric interpolation on the (u, v) coordinates of the corresponding mesh-triangle vertices. The signed distance h of the sampling point to its projection on the mesh is used to disambiguate points orthogonal to the mesh surface [18]. With this canonicalization, we can formulate the radiance field mapping as,

$$H_\alpha : (u, v, h, d) \rightarrow (\mathbf{c}, \sigma) \tag{4}$$

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 [23] to all inputs in Eq. 4.

This canonicalized uvh space does not contain any pose information. Since a point on the hand surface could have a different appearance based on the hand pose, we also explicitly condition our model with the hand pose ξ after canonicalization. This leads to the modified representation:

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

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 $\mathbf{c}$ for each point in 3D space. For a ray with origin $\mathbf{o}$ and direction $\mathbf{d}$, volumetric integration - as proposed in NeRF [23] - can be used to obtain the integrated color $\mathbf{C}$ for the ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$:

$$\mathbf{C}(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t))dt$$
$$\text{where} \quad T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds), \tag{6}$$

and t_n and t_f are near and far bounds. 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 [23] 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.

Mesh-Guided Sampling. To make the rendering faster, we utilize the coarse MANO geometry to efficiently sample points around the approximate hand surface. 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 [9]. This eliminates the two-pass approach needed for hierarchical sampling.

Super-resolution. Although this efficient sampling strategy improves the run-time, it still can not achieve real-time rendering speeds. We introduce a super-resolution network [5] 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. We accomplish this by extending Eq. 4 with:

$$H_\alpha : (u, v, h, d, \xi) \rightarrow (\mathbf{c}, \mathbf{f}, \sigma) \tag{7}$$

We then apply volumetric integration as done in Eq. 6 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

$$S_\phi : (L, F) \rightarrow I \tag{8}$$

to recover a high-resolution image I_j^p that preserves the details. To ensure efficiency, 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 multi-view images, in general, are not color corrected to be consistent across views, we compensate for this, as done in Neural Volumes [19], 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

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_{perc} \tag{9}$$

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:

$$\mathcal{L}_{rec} = ||G_j^p - I_j^p(\alpha, \phi)||_2 \tag{10}$$

To capture the perceptual difference in the image, we apply $\mathcal{L}_{perc}$ as suggested in [43]

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

Where $f(\cdot)$ is the activation of the $conv1$-$conv5$ layers in pre-trained VGG network [37]. Thanks to our efficient design choices, we can apply the perceptual loss on the full image, as opposed to the traditional approach of applying it on smaller patches [40]. We show later that the perceptual loss plays a vital role in recovering high-frequency details, and our image-based approach improves photorealism over using the patch-based strategy (see Tab. 4, Fig. 7 and the supplementary video). We employ the above loss functions to both low-resolution volumetrically rendered images and high-resolution super-resolved images.

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 5 FPS and 512×334 pixels 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 demo for real-time photorealistic neural hand reenactment. The details of this demo 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, LPIPS, and FID metrics for numerical evaluation. Following the conventions of [43], LPIPS score is calculated using AlexNet backbone. For rendering speed, we report the time it takes to render an image on an NVIDIA GeForce RTX 3090 at the training resolution (i.e. 512×334 pixels) in frames per second (FPS). For super-resolution experiments, volumetric integration produces a rendering at 256×167 pixels which are then super-resolved to 512×334 pixels. More implementation details and results can be found in the supplementary material.

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. Note how the avatar of each identity captures high-frequency skin texture as well as hand-pose dependent illumination, which contributes to the photo-realism of our renderings.

To show that this method can work in real applications, we also implement 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. We estimate the pose using [45] and pass it to our method for rendering. 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 State of the Art

The only other neural implicit hand model that exists in the literature is LISA [6]. 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 approach for a fair comparison. As an additional baseline, we use the body modeling method A-NeRF [38] and adapt it for hand modeling. We also compare against SMPLpix [31] because of its real-time performance. We adapt it to hands by changing the conditioning input from SMPL to MANO renderings. 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.

As shown in Table 2, our method outperforms other neural implicit baselines while also being real-time. These improvements in the metrics also translate to significant improvements in perceptual quality on the test set, which can be seen in Fig. 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

	PSNR ↑	LPIPS(x1000) ↓	FID ↓	FPS ↑
Mesh wrapping	28.28	49.44	298.28	**82.33**
SMPLpix [31]	**32.37**	26.57	202.99	58.82
A-NeRF* [38]	28.07	94.41	318.61	0.83
LISA* [6]	29.36	78.46	255.43	3.70
Ours	32.04	**25.73**	**197.39**	45.45

Table 2: **Comparison on InterHand2.6M [25].** * indicates we use our implementation of the approach.

	PSNR ↑	LPIPS(x1000) ↓	FID ↓	#parameters ↓	FPS ↑
xyz	29.31	42.50	247.77	0.95M	43.03
per-bone xyz	**32.51**	**23.82**	198.95	1.14M	27.04
uvh w.o. pose cond.	30.33	32.36	204.24	**0.40M**	**45.73**
Ours (uvh w. pose cond.)	32.04	25.73	**197.39**	0.41M	45.45

Table 3: **Ablation on various canonicalization strategies.** Our approach optimizes for both quality and speed.

the ambiguities in case of severe articulations. Moreover, as we will show in Sec. 4.3, our addition of a perceptual loss drastically improves the level of detail the model can capture over those obtained from simple per-pixel loss used in A-NeRF and LISA.

SMPLpix comes close to our method quantitatively, but fails to capture the details, as shown in Fig. 5. This is because, unlike our method, SMPLpix can not account for person-specific geometric changes as it strictly relies on coarse MANO geometry.

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. But by no means can such a simple rendering 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.

We show additional comparisons using a synthetic dataset in the supplementary document.

4.3. Ablation

Our design choices are crucial for optimizing both the rendering quality and processing speed. To evaluate their significance, we perform ablation studies of various components. 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 as it is vital for capturing pose-dependent

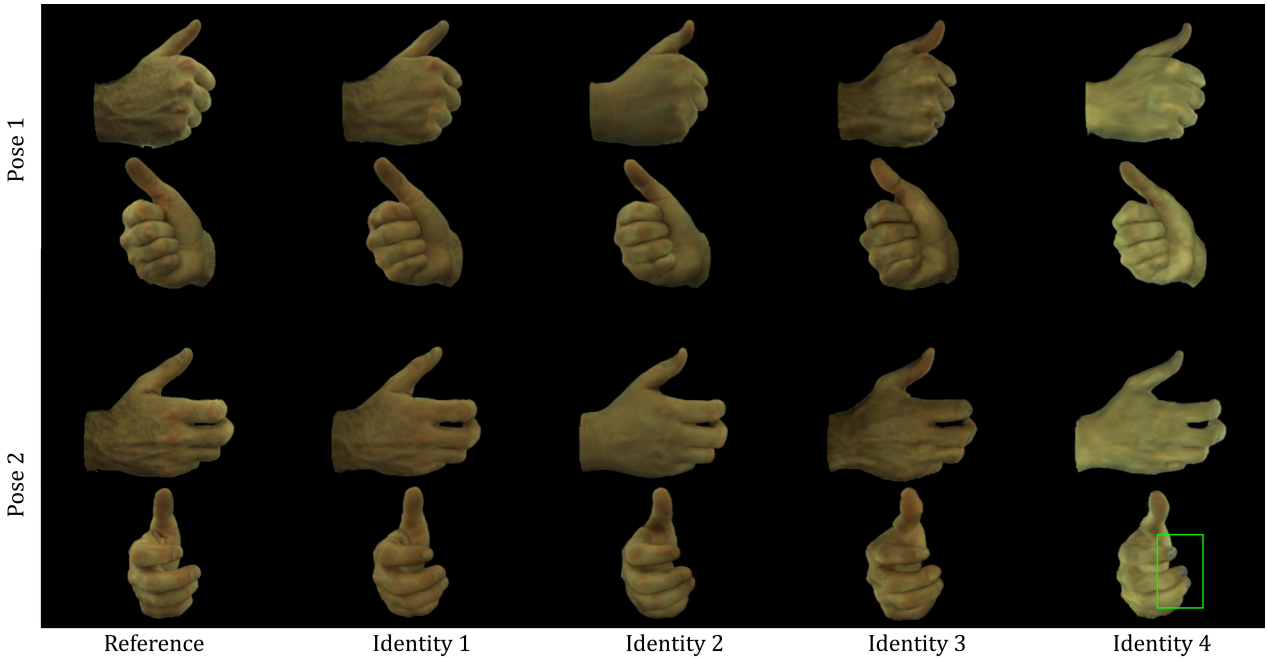
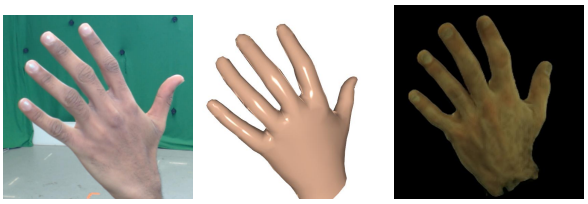


Figure 3: **Hand Reenactment.** Our method can transfer the pose of a reference actor (Reference) to new identities (Identity 1-4). Note that our model captures pose-dependent changes, which is especially apparent for veins and in the knuckle region. It also captures view-dependent shading and self-shadowing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



(a) Driving input (b) Estimated pose (c) Our rendering

Figure 4: **Demo Visualization.** The real-time demo takes in a monocular RGB input (left) to estimate the MANO parameters (center). The MANO pose is then transferred to the target identity using our method (right).

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

Table 4: **Ablation study on model components.** All design choices consistently improve the accuracy and runtime.

effects such as self-shadowing and skin wrinkles, and this can be seen in Fig. 6.

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 our method is 5 times faster. We also evaluated the impact of the superresolution module 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 ('w.o. $\mathcal{L}_{perc}$'); we implement the commonly-used patch perceptual loss [40] where random crops of 64×64 pixels are used for the perceptual loss instead ('patch $\mathcal{L}_{perc}$'); finally, we use the perceptual loss on the full images ('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, adding any form of $\mathcal{L}_{perc}$ greatly increases the level of details (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 quantitatively and qualitatively. Finally, our full method ('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. Moreover, in the supplementary material, we use synthetic data to show our method's robustness to MANO fitting inaccuracies. We also present an additional

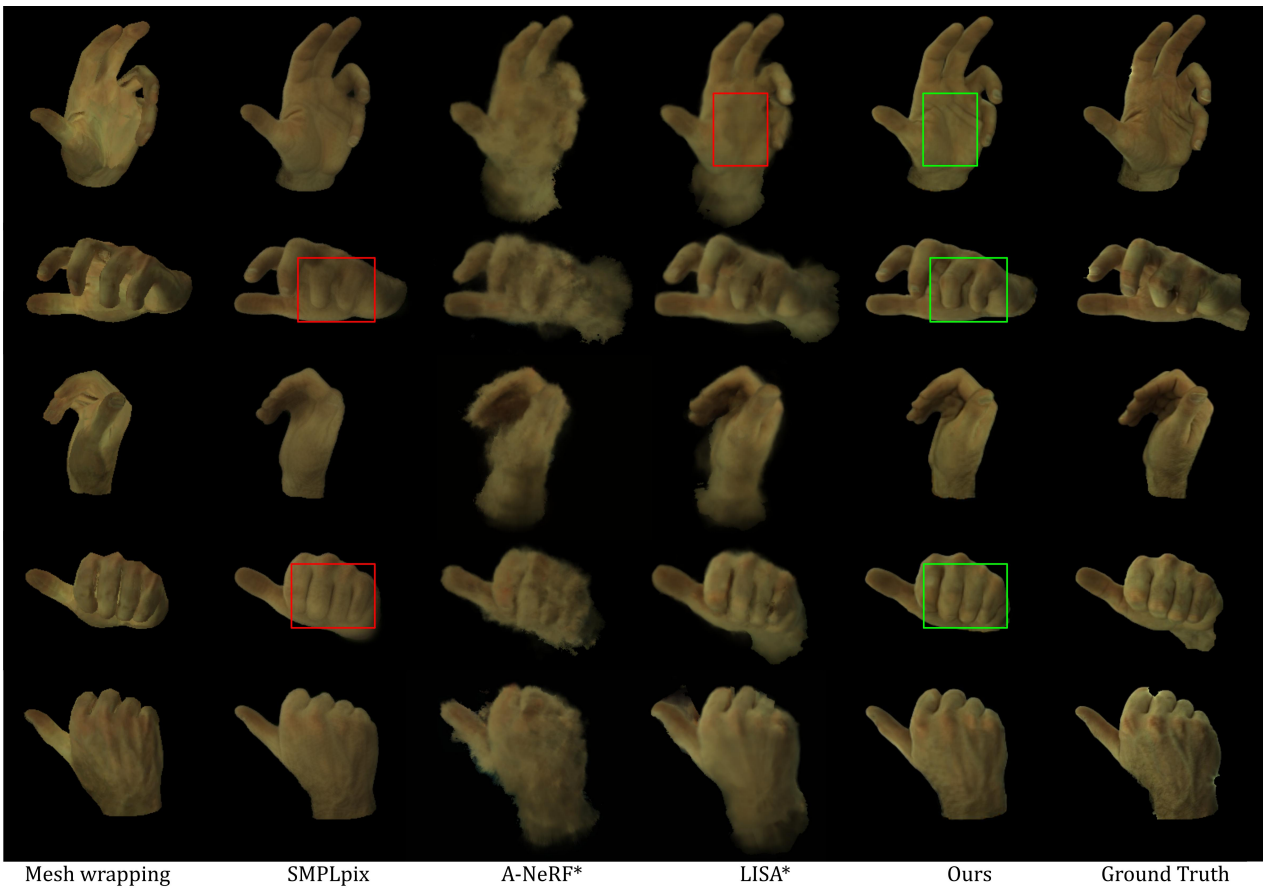


Figure 5: **Comparison to SoTA on unseen hand poses.** A-NeRF and Mesh wrapping produce artifacts while SMPLpix and LISA do not capture high-frequency details. Our method outperforms these approaches and captures high-frequency details.

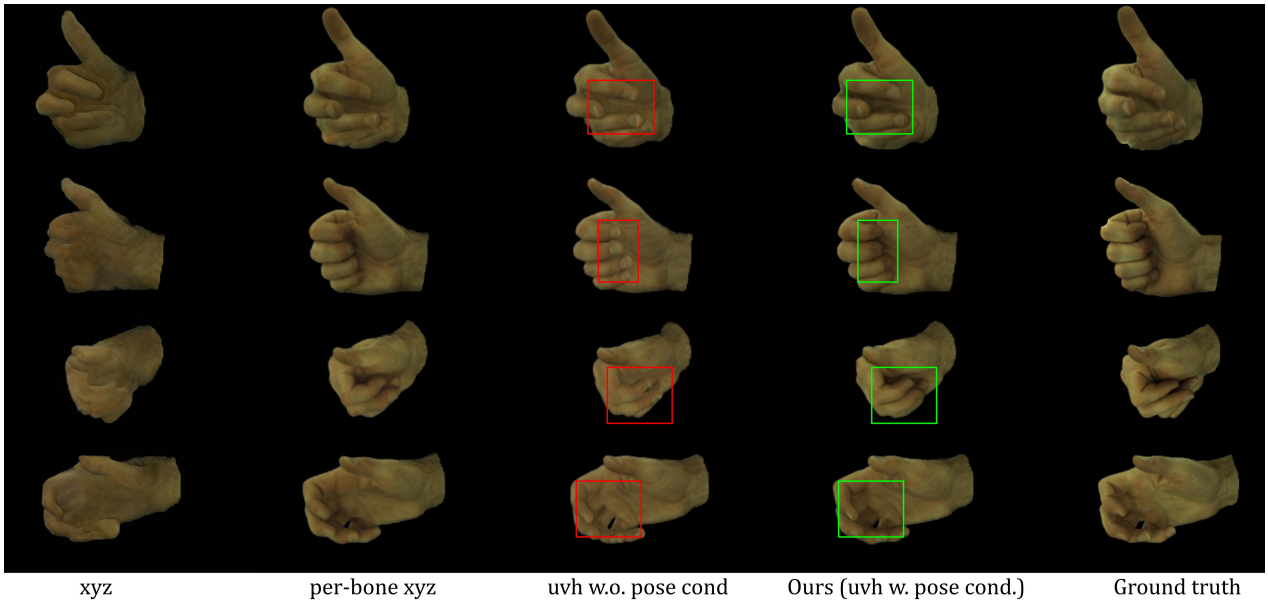


Figure 6: **Canonicalization Ablation.** Global xyz coordinates with naive conditioning fails to generalize to novel poses. Our proposed uvh canonicalization achieves similar visual results to per-bone xyz canonicalization while being much faster. Note that hand pose conditioning is vital for capturing pose-dependent effects such as self-shadowing (see red and green regions).

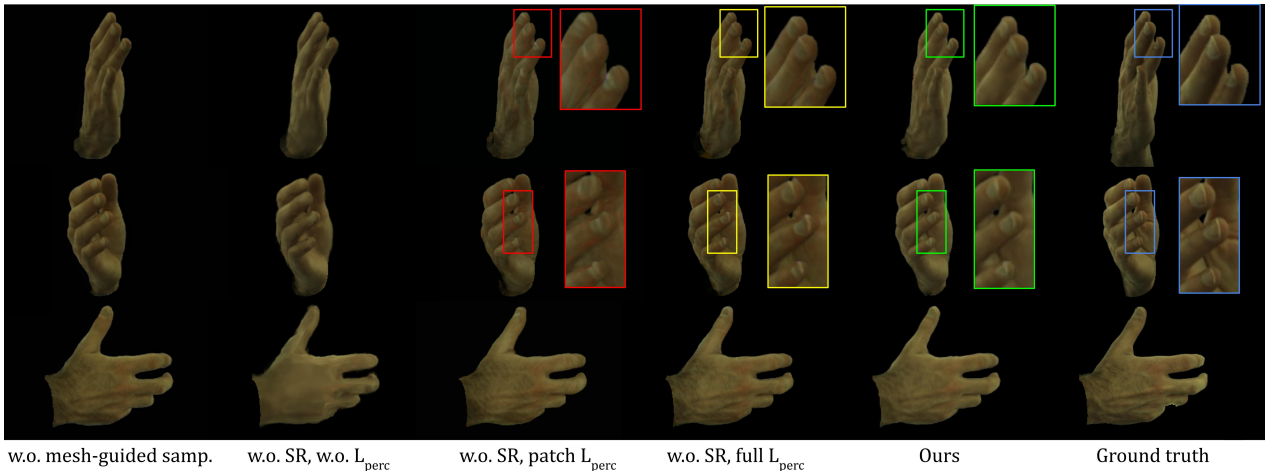


Figure 7: **Model Ablation.** Left to right: without the mesh-guided sampling, the visual quality is good but the inference is slow (see Tab. 4); without any perceptual loss, the reconstructions lack details; with patch-based perceptual loss, subtle artifacts appear in the details (as highlighted in red); with full-image perceptual loss, these details are captured correctly (as highlighted in yellow); finally, by using a super-resolution module, the rendering speed is further improved without compromising the details (as highlighted in green).

application where the hand geometry can be edited at inference time, without any additional retraining of the model.

5. Discussion

5.1. Limitations and Future Work

While our work is an important milestone for the full digitization of human hands, there are still several avenues for future work. Since our approach depends on the MANO mesh, future work could look into improving the quality of such a mesh. 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 hand semantics. While our approach models hand-pose dependant illumination effects, it can not model shadow as a function of any random illumination condition other than the one the training set was captured under. We leave this modeling for future works. We hope our work encourages research into the important problem of photorealistic rendering of the human hands.

5.2. Societal Impact

Alongside its immense applications, human modeling also presents challenging societal problems. A digital avatar of an individual has the potential of being misused by bad actors. Though detecting real vs. fake images is a possibility, a more strategic approach would be watermarking the generative models. This way, a generated image can always be attributed back to the model it was generated from. This is an active area of research, and we hope the community adopts it in their body modeling works.

6. Conclusion

We presented the first neural implicit approach that can render human hands in a photorealistic manner in real-time. Our approach is carefully designed to optimize the rendering quality and speed. At the heart of our method is a low-resolution NeRF rendering and a super-resolution module that produces 3D-consistent results. We show that a novel application of the perceptual loss on the full image space is important for generating accurate details. 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, high-frequency texture details, and pose-dependent effects. Comparison with related methods clearly shows that our approach outperforms the baselines by a significant margin. We also demonstrate editing the hand geometry while keeping the texture fixed. Future work could investigate learning a generalized implicit 3D morphable model of the human hands that is photoreal.

Acknowledgements

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

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5461–5470, 2021. 2

[2] Timur Bagautdinov, Chenglei Wu, Tomas Simon, Fabián Prada, Takaaki Shiratori, Shih-En Wei, Weipeng Xu, Yaser Sheikh, and Jason Saragih. Driving-signal aware full-body avatars. *ACM Trans. Graph.*, 40(4), jul 2021. 3

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Combining implicit function learning and parametric models for 3d human reconstruction. In *European Conference on Computer Vision (ECCV)*. Springer, August 2020. 3

[4] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, Yaser Sheikh, and Jason Saragih. Authentic volumetric avatars from a phone scan. *ACM Trans. Graph.*, 41(4), jul 2022. 3

[5] 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. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 5

[6] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20533–20543, 2022. 2, 3, 4, 6

[7] Xuan Gao, Chenglai Zhong, Jun Xiang, Yang Hong, Yudong Guo, and Juyong Zhang. Reconstructing personalized semantic facial nerf models from monocular video. *ACM Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 41(6), 2022. 3

[8] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18653–18664, 2022. 3

[9] Marc Habermann, Lingjie Liu, Weipeng Xu, Gerard Pons-Moll, Michael Zollhoefer, and Christian Theobalt. Hdhumans: A hybrid approach for high-fidelity digital humans. In *SCA '23: SIGGRAPH/Eurographics Symposium on Computer Animation*, 2023. 2, 3, 5

[10] Marc Habermann, Lingjie Liu, Weipeng Xu, Michael Zollhoefer, Gerard Pons-Moll, and Christian Theobalt. Real-time deep dynamic characters. *ACM Transactions on Graphics*, 40(4), aug 2021. 2, 3

[11] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. *arXiv preprint arXiv:2112.10203*, 2021. 3

[12] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. Selfrecon: Self reconstruction your digital avatar from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[13] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, 2021. 2

[14] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *2020 International Conference on 3D Vision (3DV 2020)*, pages 333–344, Piscataway, NJ, Nov. 2020. IEEE. 2

[15] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):194:1–194:17, 2017. 2

[16] Yuwei Li, Minye Wu, Yuyao Zhang, Lan Xu, and Jingyi Yu. Piano: A parametric hand bone model from magnetic resonance imaging. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 816–822, 8 2021. 2

[17] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 2

[18] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 2, 3, 4

[19] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Trans. Graph.*, 38(4):65:1–65:14, July 2019. 5

[20] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Trans. Graph.*, 40(4), jul 2021. 3

[21] Haimin Luo, Teng Xu, Yuheng Jiang, Chenglin Zhou, Qiwei Qiu, Yingliang Zhang, Wei Yang, Lan Xu, and Jingyi Yu. Artemis: Articulated neural pets with appearance and motion synthesis. *ACM Trans. Graph.*, 41(4), jul 2022. 3

[22] S. Ma, T. Simon, J. Saragih, D. Wang, Y. Li, F. La Torre, and Y. Sheikh. Pixel codec avatars. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 64–73, jun 2021. 3

[23] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2, 3, 4, 5

[24] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *European Conference on Computer Vision*, pages 440–455. Springer, 2020. 2

[25] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb im-

age. In *European Conference on Computer Vision (ECCV)*, 2020. 6

[26] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *International Conference on Computer Vision*, 2021. 2, 3

[27] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019. 2

[28] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2, 3

[29] Sida Peng, Shangzhan Zhang, Zhen Xu, Chen Geng, Boyi Jiang, Hujun Bao, and Xiaowei Zhou. Animatable neural implicit surfaces for creating avatars from videos. *arXiv preprint arXiv:2203.08133*, 2022. 3

[30] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9054–9063, 2021. 2, 3

[31] Sergey Prokudin, Michael J Black, and Javier Romero. Smplpix: Neural avatars from 3d human models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1810–1819, 2021. 3, 6

[32] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. Html: A parametric hand texture model for 3d hand reconstruction and personalization. In *European Conference on Computer Vision, (ECCV)*, pages 54–71. Springer, 2020. 2

[33] Amit Raj, Julian Tanke, James Hays, Minh Vo, Carsten Stoll, and Christoph Lassner. Anr: Articulated neural rendering for virtual avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3722–3731, June 2021. 3

[34] Edoardo Remelli, Timur Bagautdinov, Shunsuke Saito, Chenglei Wu, Tomas Simon, Shih-En Wei, Kaiwen Guo, Zhe Cao, Fabian Prada, Jason Saragih, and Yaser Sheikh. Drivable volumetric avatars using texel-aligned features. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. 3

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6):245:1–245:17, Nov. 2017. 2, 3

[36] Ruizhi Shao, Liliang Chen, Zerong Zheng, Hongwen Zhang, Yuxiang Zhang, Han Huang, Yandong Guo, and Yebin Liu. Floren: Real-time high-quality human performance rendering via appearance flow using sparse rgb cameras. In *SIGGRAPH Asia Conference Papers*, 2022. 3

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[38] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021. 2, 3, 6

[39] Xin Suo, Yuheng Jiang, Pei Lin, Yingliang Zhang, Minye Wu, Kaiwen Guo, and Lan Xu. Neuralhumanfvv: Real-time neural volumetric human performance rendering using rgb cameras. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6222–6233, 2021. 3

[40] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16210–16220, June 2022. 5, 7

[41] Feng Xu, Yebin Liu, Carsten Stoll, James Tompkin, Gaurav Bharaj, Qionghai Dai, Hans-Peter Seidel, Jan Kautz, and Christian Theobalt. Video-based characters – creating new human performances from a multi-view video database. *ACM Transactions on Graphics*, 30:32, 07 2011. 3

[42] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2863–2873, 2022. 2, 3

[43] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[44] Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 13545–13555, June 2022. 3

[45] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 6