HARP: Personalized Hand Reconstruction from a Monocular RGB Video

Korrawe Karunratanakul   Sergey Prokudin   Otmar Hilliges   Siyu Tang
ETH Zürich, Switzerland

{korrawe.karunratanakul,sergey.prokudin,otmar.hilliges,siyu.tang}@inf.ethz.ch
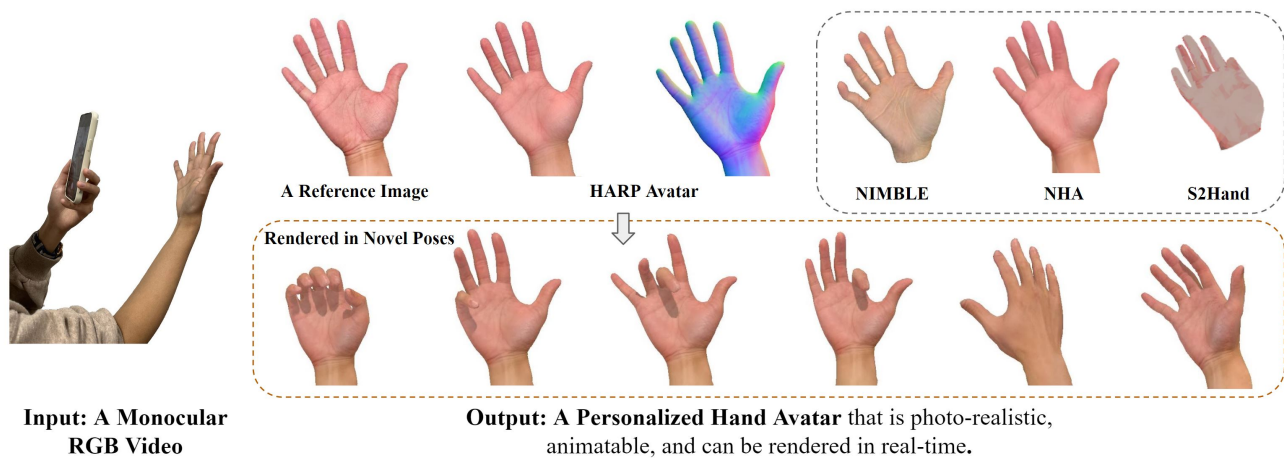https://korrawe.github.io/harp-project/

Figure 1. Given a short monocular video of a hand and a coarse hand pose and shape estimation for initialization, we reconstruct a photo-realistic hand avatar exhibiting faithful personalized appearance and geometry using standard explicit representations and a differentiable renderer *without any neural networks*. Compared to the baselines, our hand avatar demonstrates better-reconstructed geometry and appearance. The hand avatar can be used to render high-fidelity hand images in novel views and poses in real-time, which serves as a foundation for many AR/VR applications.

Abstract

We present HARP (HAnd Reconstruction and Personalization), a personalized hand avatar creation approach that takes a short monocular RGB video of a human hand as input and reconstructs a faithful hand avatar exhibiting a high-fidelity appearance and geometry. In contrast to the major trend of neural implicit representations, HARP models a hand with a mesh-based parametric hand model, a vertex displacement map, a normal map, and an albedo without any neural components. The explicit nature of our representation enables a truly scalable, robust, and efficient approach to hand avatar creation as validated by our experiments. HARP is optimized via gradient descent from a short sequence captured by a hand-held mobile phone and can be directly used in AR/VR applications with real-time rendering capability. To enable this, we carefully design and implement a shadow-aware differentiable rendering scheme that is robust to high degree articulations and self-shadowing regularly present in hand motions, as well

as challenging lighting conditions. It also generalizes to unseen poses and novel viewpoints, producing photo-realistic renderings of hand animations. Furthermore, the learned HARP representation can be used for improving 3D hand pose estimation quality in challenging viewpoints. The key advantages of HARP are validated by the in-depth analyses on appearance reconstruction, novel view and novel pose synthesis, and 3D hand pose refinement. It is an AR/VR-ready personalized hand representation that shows superior fidelity and scalability.

1. Introduction

Advancements in AR/VR devices are introducing a new reality in which the physical and digital worlds merge. The human hand is a crucial element for an intimate and interactive experience in these environments, serving as the primary interface between humans and the digital world. Therefore, it is essential to capture, reconstruct, and animate life-like digital hands for AR and VR applications. Without

this capability, the authenticity and practicality of AR/VR consumer products will always be limited.

Despite its importance, the research into hand avatar creation has so far been limited. Most works [8, 35, 59] focus on creating an appearance space on top of a parametric hand model such as MANO [62]. Such an appearance space provides a compact way to represent hand texture but is rather limited in expressivity to handle non-standard textures. The recent LISA [12] model has emerged as an alternative, using an implicit function to represent hand geometry and texture color fields. Training a new identity in LISA, however, requires a multi-view capturing setup as well as a large amount of data and computing power. In the nearby fields of face and body avatar creation, many works that leverage an implicit function [19,40,41,74] or NeRF-based [42] volume rendering [39,53,77] have also been recently explored. The NeRF-based method such as HumanNeRF [77] produces a convincing novel view synthesis but still shows blurry artifacts around highly articulated parts and cannot be easily exported to other applications.

We argue that democratizing hand avatar creation for AR/VR users requires a method that is *(1) accurate*: so that personalized hand appearance and geometry can be faithfully reconstructed; *(2) scalable*: allowing hand avatars to be obtained using a commodity camera; *(3) robust*: capable of handling out-of-distribution appearance and self-shadows between fingers and palm; and *(4) efficient*: with real-time rendering capability.

To this end, we propose HARP, a personalized hand reconstruction method that can create a faithful hand avatar from a short RGB video captured by a hand-held mobile phone. HARP leverages a parametric hand model, an explicit appearance, and a differentiable rasterizer and shader to reconstruct a hand avatar and environment lighting in an analysis-by-synthesis manner, *without any neural network component*. Our observation is that human hands are highly articulated. The appearance changes of observed hands in a captured sequence can be dramatic and largely attributed to articulations and light interaction. Learning neural representations, such as implicit texture fields [12] or volume-based representations like NeRF [70], is vulnerable to the over-fitting to a short monocular training sequence and can hardly generalize well to sophisticated and dexterous hand movements. By properly disentangling geometry, appearance, and self-shadow with explicit representations, HARP can significantly improve the reconstruction quality and generate life-like renderings on novel views and novel animations performing highly articulated motions. Furthermore, the nature of the explicit representation allows the results from HARP to be conveniently exported to standard graphics applications.

In summary, the key advantages of HARP are: (1) HARP is a simple personalized hand avatar creation method

that reconstructs high-fidelity appearance and geometry using only a short monocular video. HARP demonstrates that an explicit representation with a differentiable rasterizer and shader is enough to obtain life-like hand avatars. (2) The hand avatar from HARP is controllable and compatible with standard rasterization graphics pipelines allowing for photo-realistic rendering in AR/VR applications. (3) Moreover, HARP can be used to improve 3D hand pose estimation in challenging viewpoints. We perform extensive experiments on the tasks of appearance reconstruction, novel-view-and-pose synthesis, and 3D hand poses refinement. Compared to existing approaches, HARP is more accurate, robust, and generalizable with superior scalability.

2. Related Work

Hand Models. Hand models are crucial for compactly representing a hand surface and using it in downstream applications. Many models rely on an explicit mesh surface [34, 35, 44, 62, 80], while others implicitly represent the surface with neural networks [12, 25, 26]. The widely used MANO model [62] represents a hand with pose and shape vectors, generating hand meshes using a PCA model and linear blend skinning. While parametric models like MANO can handle various hand geometries, their expressivity is limited by the learned parameter spaces. An alternative approach is to store vertex locations directly to broaden geometry representation [18, 36, 43, 67], but this prevents re-animation. Conversely, a limited amount of work has been explored for hand appearance modeling. Notable works include HTML [59], a linear appearance model for inferring the UV texture on top of MANO, and NIMBLE [35], an extension of the hand skeleton model [34], which can infer the surface appearance, including an albedo, specular, and normal map. Despite their realistic texture, these model suffers from their linear nature of appearance space and limited appearance data, making them unsuitable for adapting to a novel identity. In this work, we enhance the parametric model with personalized geometry adjustment, explicit albedo, and normal maps, enabling animatable personalized geometry and stronger texture representation than PCA-based textures.

Geometry Reconstruction. To build a hand avatar, one must first derive the hand geometry from the input images which has been a long-studied problem [1, 4, 6, 48, 49, 63, 71, 82]. To estimate the hand surface, these methods generally leverage the statistical prior in the MANO model by predicting its pose and shape parameters. One advantage of such a method is that it could prevent geometry from collapsing by encouraging the hand shape to be close to the mean shape [21, 26]. To overcome this limited expressiveness of the MANO shape space, a convolution neural network (CNN) can be used to directly estimate the vertex lo-

cations of the mesh topology [10, 18, 30, 31, 43]. Lin *et al.* [36] replace the convolution operation with a Transformer model [73] and achieve state-of-the-art results on various datasets. Alternatively, the hand surface can be represented by an implicit function [12, 25, 26]. With HALO [25], the hand geometry can be estimated by any key point estimation methods [3, 14, 17, 23, 45–47, 69, 72, 81, 86]. Given ground truth hand poses in multi-view images, differentiable ray tracing can also be used to refine the annotations [27]. In this work, we leverage the prediction by METRO [36] as initialization to refine a personalized geometry.

Appearance Representations. Numerous methods have been proposed to learn and estimate appearance from images or videos for bodies with clothing [22,33,51,56,58,65] and faces [2, 15, 16, 19, 29, 53], including those that use NeRF [42] to implicitly represent appearance [53, 54]. To model a human, pose information can also be used as conditions to transform the radiance field of a human avatar [9,52,70,77,79]. However, these models are not suitable for hand appearance due to the high degree of articulation. Furthermore, mesh extraction is required to make these methods compatible with traditional graphics applications. In contrast, less research has been done on hand appearance representation [64]. While LISA [12] and HandAvatar [7] both learn implicit color fields together with implicit surface representations, they require a large number of images to train a subject-specific shape and appearance. Given an image, S2hand [8] employs an MLP to estimate the vertex colors and lighting together with MANO parameters. Concurrently, Wang *et al.* proposed SunStage [75], an outdoor face reconstruction method that shares similarities to our method, in particular, the use of explicit mesh representation and differentiable rendering. In this work, we present an explicit hand appearance model and the optimization pipeline that can capture detailed hand texture while also taking lighting and shadowing into account.

3. Method

Overview. The overview of HARP is illustrated in Fig. 2. Given a short monocular RGB video of a hand consisting of N frames, we reconstruct a realistic hand avatar with *personalized* shape and texture. Specifically, our method outputs a triangle mesh M, containing vertices V and faces F, and its UV texture, which is decomposed into albedo and normal maps. We optimize the personalized hand mesh M, the albedo, and the normal map using an analysis-by-synthesis approach by comparing the input images to the images of our reconstruction M rendered using a differentiable rendering framework.
Our approach focuses on *efficiency*, *robustness*, and *exportability*, while also maintaining a *high-quality hand appearance*. Given the focus, we employ (1) explicit represen-

tations (mesh, normal map, and albedo) which can be easily exported to any graphics application; (2) direct optimization of the explicit appearance without relying on a learned appearance space, such that we do not require pre-training nor a large number of training images. (3) the explicit and efficient rasterizing and shading which provides a good balance between rendering quality and computation cost.

3.1. Hand Representation

Template Model. Our hand template model is built upon the MANO [62] model, which we extend to a higher mesh resolution and allow surface vertex deformation from the template. Concretely, we perform a linear subdivision on the MANO template to increase the number of vertices from 778 to 3093, which in turn allows our template to capture more surface details. The subdivision process is differentiable, thus enabling gradient propagation back to the MANO pose parameter γ and shape parameter β. Additionally, the MANO hand is truncated at the wrist, which does not reflect the reality where hands are attached to the arms and the foreground mask does not separate the wrist. Therefore, we built another template from SMPLX [55] by truncating the arm at the elbow to facilitate the hand-and-arm fitting depending on the available mask. Their interactions with the rest of the system remain identical.

Geometry Refinement. To utilize the higher mesh resolution for finer geometry details, we allow each posed vertex to additionally deform based on a personalized vertex displacement D along the vertex normal. The posed hand vertex locations without arm could be obtained with:

$$V = \mathcal{S}(\mathcal{M}(\gamma, \beta)) + D, \tag{1}$$

$$\mathcal{S} : \mathbb{R}^{778 \times 3} \to \mathbb{R}^{3093 \times 3}, \tag{2}$$

where $\mathcal{M}$ is the MANO function which takes pose γ and shape β as inputs and returns posed mesh vertex locations, $\mathcal{S}$ is the mesh subdivision function. The vertex displacement vector D is subject-specific, which we obtain by optimizing it along with other parameters.

Texture. We model the hand skin as a Lambertian surface with an albedo map a, which is defined per subject in a UV space. Additionally, to add high-frequency details without upsampling the mesh, we utilize a UV-space normal map which can be combined with the surface normal $\hat{N}$ when computing the illumination at the 3D surface point x.

3.2. Shadow-aware Differentiable Rendering

There are multiple options for differentiable rendering such as a differentiable path tracer [50], differentiable rasterizer [38], neural renderer [28] or the recently developed NeRF-based volume rendering [42]. The path tracer is known for its realistic rendering and explicit factorization of
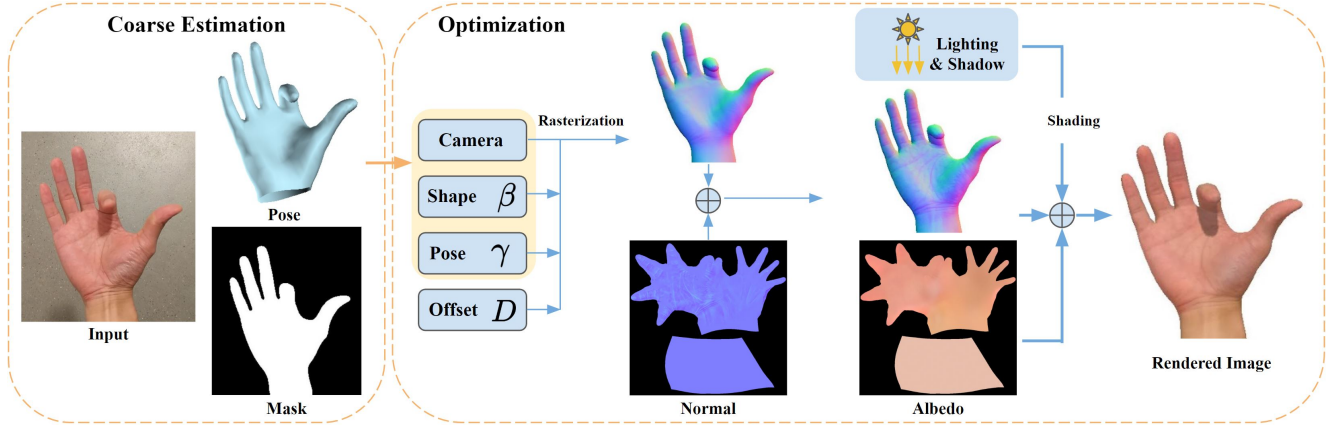
Figure 2. **Method Overview.** Given a short monocular RGB video of a hand, our hand avatar creation method includes two steps: (1) coarse hand pose and shape estimation for initialization; (2) an optimization framework to reconstruct the *personalized* hand geometry and appearance. The hand geometry is first rasterized and combined with a normal map. Then, the shader combines the albedo, geometry, and lighting to render the personalized hand. The optimization solves the hand and scene parameters using only the input images.

material properties. However, this comes at a high computation cost and has not been shown to work with highly articulated objects [20, 50]. NeRF-based methods [70, 77, 84] and implicit texture-based methods [12] are often computationally expensive, and as they leverage MLPs to approximate the light interaction with the articulated parts; the result can be blurry and have self-shadows baked into the texture representation.

In this work, we demonstrate that, to create a realistic hand avatar from a short video sequence, it is not necessary to rely on neural implicit representations, such as neural volume rendering. Standard explicit geometry and appearance representations together with a carefully implemented differentiable rendering scheme are able to provide the right balance between quality, speed, and compatibility with other graphics applications.

To render the hand in the camera view, we first utilize the differentiable rasterizer [60] to determine the visible surfaces from the camera views. We then use the Phong model [57] without the specular component to compute the illumination, i.e. color, at the surface point x:

$$I_x = k_a i_a + \sum_{m \in \text{lights}} k_d i_{m,d} (\hat{L}_m \cdot \hat{N}) \quad (3)$$

where k_a is an ambient reflection constant, k_d is a diffuse reflection constant, i_a and $i_{m,d}$ are intensities of the light sources for diffuse surface, $\hat{N}$ is the normal at point x, and $\hat{L}_m$ is a ray from the surface point toward each light source. We observe that having just one dominant light source ($|m| = 1$) provides a good balance between computation cost and the rendering quality, under the assumption that the hard shadow is usually produced by the closest light source when indoors and by the sun outdoors.

Self-shadowing. Notably, the Phong model does not account for self-shadowing which often occurs when a finger

casts a shadow onto other fingers and the palm. To accommodate this scenario, we add the visibility term $V(x, m)$ at x with respect to the light m to the diffuse component, making it $k_d(\hat{L}_m \cdot \hat{N})V(x, m)i_{m,d}$. We compute the visibility V by performing a two-step rasterization of the mesh. We integrate the classic idea of shadow mapping in computer graphics [78] into our differentiable pipeline.

The step-by-step computation is shown in Alg 1. The z-buffer Z_m is a depth image when seen from the camera cam_m. If the light position is more than 1 m away from the hand, we project it to 1 m distance. The Sigmoid function is used to ensure smooth gradients, with a bias term $b = 0.005$ and scale $s = 1000$. To produce a softer shadow, we use percentage-closer filtering [61], which averages the visibility values of the nearby points. These visibility values then allow us to integrate the self-shadow into the pipeline. We note that our self-shadowing implementation is differentiable with respect to both geometry and appearance. Our implementation is compatible with the pyTorch3D [60] package, and will be made publicly available.

Algorithm 1 Visibility $V(x, m)$ from light m

1: Place virtual camera cam_m at light m pointing at hand
2: Compute z-buffer $Z_m(\cdot)$ from cam_m
3: Get 3D points $\{X_{hit}\}$ seen from the actual camera
4: **for** $x \in X_{hit}$ **do**
5:     Transform x to cam_m coordinate
       $x^m = T_m(x)$
6:     Get 2D pixel coordinate of x^m
       $x^{2d} = \pi(x^m)$
7:     Compute distance to the light
       $d_{m \to x} = ||x - m||$
8:     $V(x, m) = Sigmoid(s(Z_m(x^{2d}) - d_{m \to x} + b))$
9: **end for**

3.3. Optimization

To find the parameters that describe the personalized hand, we optimize the parameters using short RGB videos. The optimization objective is to minimize the difference between the input and the rendered hand images using the proposed differentiable rendering pipeline. For each subject, we optimize for a joint objective consisting of a geometry alignment term E_{geo} and an appearance term E_{app}:

$$E \quad = E_{geo} + E_{app}, \qquad (4)$$

where E_{geo} focuses on mesh configuration and E_{app} encourages the same appearance as in the input images.

Geometry Objective. For the geometry, the goal is to match the rendered silhouette with the hand mask while also satisfying 3D mesh constraints. The geometry objective is defined independently from the appearance as follows:

$$E_{geo} \quad = w_{sil} \cdot E_{sil} + E_{reg}, \qquad (5)$$

where E_{sil} is the silhouette difference term, E_{reg} is the mesh regularization term, and w are the weights.

The silhouette difference term is the l_1-difference between the hand mask and the rendered silhouette S_{render}:

$$E_{sil} = |S_{in} - S_{render}|, \qquad (6)$$

where $S_{in} \in \{0,1\}^{H \times W}$ is an input binary hand mask that can be obtained from off-the-shelf segmentation tool [24].

To prevent the optimized mesh from collapsing, we employ a combination of 3D mesh regularizations defined as:

$$E_{reg} \quad = E_{init} + E_{verts} + E_{lap} + E_{norm} + E_{arap}, \qquad (7)$$

where each term is accompanied by its weight.

The term E_{init} penalizes key points deviation from the initial pose with l_1 distance. The vertex offset regularization E_{verts} controls the deviation from the MANO mesh to be small using an l_2-norm: $E_{verts} = \|D\|^2$.

The mesh vertices V are regularized by the Laplacian mesh regularizer [13] E_{lap} and the normal consistency regularizer E_{norm} defined on the posed mesh.

The term E_{arap} is the as-rigid-as-possible term [68] that encourages the 3D mesh to be more rigid and distributes the changes in length among multiple edges. The edge length difference is defined with respect to the MANO template as:

$$E_{arap} \quad = \sum_v^V \sum_{u \in \mathcal{N}(v)} \left\| \|v_t - u_t\| - \|v^* - u^*\| \right\|^2, \qquad (8)$$

where $\mathcal{N}(v)$ are the adjacent vertices of v, v_t is a vertex from frame t, and v^* is a vertex from the MANO template.

Appearance Objective. The appearance term E_{app} measures the similarity between the input and the rendered image. Note that as the texture is mapped to the triangle mesh for rendering, the appearance term is also affected by the geometry change. We define the appearance term E_{app} as:

$$E_{app} \quad = w_{photo} \cdot E_{photo} + w_{vgg} \cdot E_{vgg} + E_{app\_reg} \qquad (9)$$

where E_{photo} is a per-pixel l_1 color difference between the input images and the predicted images, E_{vgg} is the VGG loss [32] that captures the perceptual difference between the two images by comparing the features extracted using the VGG model [66], and E_{app_reg} is a regularization term that encourages both albedo and normal map to be locally smooth [84] (details in the Appendix).

Initialization. For optimization, we initialize the hand parameters with per-frame predictions from the hand pose estimator METRO [36]. As METRO directly predicts the mesh coordinate without using the MANO pose and shape space, we obtain the equivalent MANO pose γ and shape β parameters by minimizing the l_2-distances between corresponding vertices from the prediction and the MANO mesh.

Optimization. In summary, we optimize: **hand geometry parameters**: (1) β, the global MANO shape parameter, (2) D, the per-vertex displacement, (3) γ, the per-frame MANO pose parameter, including (4) the per-frame translation, **global appearance parameters**: (5) a, the UV-space albedo, and (6) the UV-space normal map, **lighting parameters**: (7) x_{lights}, the light positions, (8) k, the global reflection constant. Fig. 2 shows the overview of our optimization process. More details are in Appendix.

4. Experiments

Datasets. There is a rich literature on datasets for hand pose estimation and geometry reconstruction [5,46,85], but less on hand appearance reconstruction. For our personalized hand avatar reconstruction task, there is no suitable dataset captured in the out-of-lab environment with monocular commodity equipment. Therefore, apart from evaluating on the existing InterHand2.6M [46] dataset, we create our own datasets for hand avatar creation, including a hand appearance dataset and a synthetic dataset (Fig. 3).

Hand Appearance Dataset. To simulate less-constrained capture settings similar to what the end-users of AR/VR applications typically have, all of our videos are captured by holding a phone camera pointing at a right hand in normal office lighting (Fig.1). Our hand appearance dataset consists of three parts: **(1) Multi-subject single-view hand sequences.** The captures contain four subjects, three male subjects and one female subject, in motions ranging from flipping the hand to more complex interactions between fingers. In total, there are 750 training and 600 testing frames for each subject. **(2) Out-of-distribution hand appearance.** We captured 2 additional subjects with tattoos for testing the out-of-distribution appearance, each containing
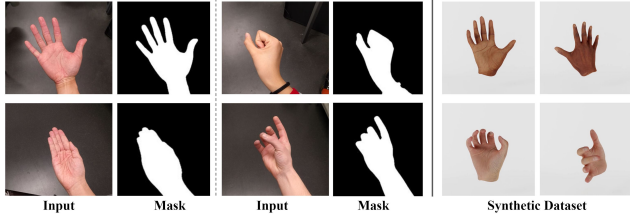
Figure 3. **Sample images.** (left) Each subject in the hand appearance dataset. (right) The synthetic dataset. The hand segmentation is obtained using an off-the-shelf segmentation tool [24].

	Appearance Rep.	Out-of-dist.	Compatibility
w/ shared appearance space			
S2Hand [8]	vertex color	✗	✓
HTML [59]	UV map	✗	✓
NIMBLE [35]	UV+spec+normal	✗	✓
w/o shared appearance space			
LISA [12]	3D Implicit	✓	✓[1]
NHA* [19]	2D Implicit	✓	✗[2]
HARP	UV + normal	✓	✓

Table 1. **Overview of hand appearance models**. Appearance Rep indicates the method used to represent texture information in the model. Out-of-dist refers to its ability to represent arbitrary non-standard hand appearance, including hands with scars or tattoos. Compatibility indicates if the method could export the 3D model with appearance for use in other applications. *For Neural Head Avatar (NHA), we adapt the official code to hand domain. To make it compatible with other methods, we dropped the face-specific components from the model, including face landmarks, segmentation, and normals. [1]Can be extracted as a mesh by Marching Cubes. [2]Geometry is compatible but colors cannot be extracted.

4 sequences. Otherwise, the setting is the same as the first part. **(3) Lighting and shadow variation.** We selected a subject from the previous part to capture 6 sequences of simple hand motions under varying directions of a single dominant light source. The shadow is highly pronounced in this portion. For all parts, we ensure that both sides of the hand are visible. It is still possible that parts that are usually occluded, such as areas between fingers, are not visible. We obtain a binary hand mask for each frame using an online segmentation tool [24] that considers the hand and the visible part of the arm as foreground.

Interhand2.6M Dataset. [46] The data is captured in a capturing dome with controlled lighting which is more restrictive than our primary goal of casually captured video. The foreground masks are obtained using RVM [37], which are sometimes noisy. We select a single-view sequence of length 500 frames from the test set where most of the surface is visible for appearance evaluation.

Synthetic Dataset. As the 3D annotations of real datasets often contain fitting error (reported at 2-3mm for Inter-Hand2.6M [46]), we opt to use a synthetic dataset with perfect ground truth to evaluate hand pose estimation task. We rendered images of two subjects, each with two sequences, using a ray tracing engine Cycles in Blender [11]. More details about the datasets can be found in the Appendix.

Baselines. We summarize the overview of the available baselines for hand avatar creation from RGB images in Tab. 1. Our main advantages compared to the baselines are: (i) by relying on explicit UV and normal maps, our output is directly compatible with standard graphics pipelines; (ii) we can represent non-standard hand textures which are often not captured well by the models that rely on a PCA space for appearance (Sec 4.1); (iii) HARP is more efficient in term of optimization time and memory than MLP-based methods due to the lower number of parameters.

For the methods which are a variation of a hand model, namely S2Hand [8], HTML [59], and NIMBLE [35], the appearance is often predicted by a regressor, which can be inaccurate. Nevertheless, such models also allow test time optimization to refine the appearance according to the input images. For a fair comparison, we **optimize** the hand appearances in all of these baselines at test time.

We provide details on how we adapt each work for the hand avatar creation in the Appendix. For the methods that require pose and shape initialization–all except S2Hand–we fit the hand model to the vertex predictions by METRO [36]. The comparison to LISA [12] is omitted as there is no released code, model, or result that we could compare to. Nevertheless, our results on InterHand2.6M (Fig. 6) show a superior qualitative appearance than those presented in [12].

Evaluation Metrics. For appearance evaluation, we report metrics that focus on the rendered image quality including the pixel-wise L1, the silhouette intersection-over-union (IoU), the learned perceptual image patch similarity (LPIPS) [83], and the multi-scale structural similarity metric (MS-SSIM) [76]. The rendered image with white background is compared to the masked input. Importantly, the pixel-wise L1 difference, LPIPS, and the intersection-over-union evaluation are not directly comparable between methods that produce different hand geometry due to the missing appearance of the truncated wrist in MANO and NIMBLE. Therefore they should be considered as references rather than direct comparisons. For pose evaluation, we report Procrustes-aligned vertex error (PA-MPVPE) in mm compared to the MANO ground truth.

4.1. Hand Appearance Reconstruction

To demonstrate the realism and robustness of HARP, we first evaluate its ability to reconstruct and re-render hand appearance from RGB sequences. We use the first part of our appearance dataset which reflects a more common environment for hand-related applications, e.g., personalized hand reconstruction and rendering for AR/VR. We show the qualitative texture and geometry in Fig. 4 and the quantitative evaluation in Tab. 2 averaged over all subjects. The results

	IoU ↑	L1 ↓	LPIPS ↓	MS-SSIM ↑
METRO [36]	0.651	-	-	-
S2Hand [8]	0.430	0.058	0.270	0.595
HTML [59]	0.778	0.033	0.136	0.791
NHA [19]	0.860	0.025	0.114	0.878
NIMBLE [35]	0.641	0.048	0.204	0.691
HARP	**0.929**	**0.018**	**0.071**	**0.902**

Table 2. Quantitative evaluation of the appearance reconstruction task on the train split of our captured sequences.

	IoU ↑	L1 ↓	LPIPS ↓	MS-SSIM ↑
METRO [36]	0.561	-	-	-
HTML [59]	0.571	0.091	0.203	0.822
NHA [19]	0.651	0.084	0.229	0.819
NIMBLE [35]	0.621	0.083	0.229	0.775
HARP	**0.779**	**0.051**	**0.173**	**0.876**

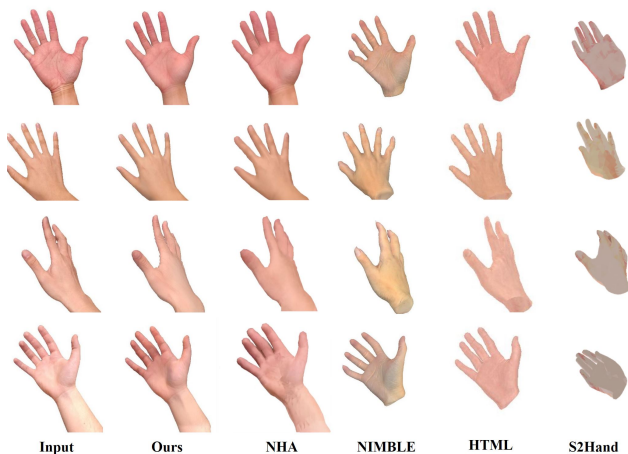Table 3. Appearance reconstruction results on InterHand2.6M.

Figure 4. Qualitative comparison to the baselines on our captured sequences. Zoom in for details.

suggest that HARP can faithfully reconstruct hand appearance with much higher details than the baselines. In addition, we demonstrate the robustness of HARP with appearance evaluation on the InterHand2.6M sequence in Tab. 3 and avatar reconstruction on various datasets in Fig. 6.

Out-of-distribution Hand Appearance. To demonstrate the advantage of our method to capture out-of-distribution appearance, we compare the optimized results to those of the PCA-based HTML and NIMBLE and the MLP-based NHA on the videos of hands with tattoos, using the same optimization pipeline and objectives. The results are shown in Tab. 4. By not being constrained by the PCA space, we can reasonably capture such out-of-distribution appearance. At the same time, the tattoos are completely discarded by HTML and NIMBLE as they are not in the training set of those models. The qualitative results Fig. 5 demonstrate the robustness of our system to capture the non-standard ap-

	L1 ↓	LPIPS ↓	MS-SSIM ↑
HTML [59]	0.018	0.121	0.836
NHA [19]	0.017	0.131	0.891
NIMBLE [35]	0.029	0.178	0.736
HARP	**0.012**	**0.080**	**0.897**

Table 4. Evaluation on out-of-distribution appearance.

Figure 5. Qualitative comparison between our method and baselines on hands with out-of-distribution appearance.

	L1 ↓	LPIPS ↓	MS-SSIM ↑
HARP w/o shadow	0.0129	0.054	0.940
HARP w/ shadow	**0.0123**	**0.051**	**0.943**

Table 5. Comparison between HARP with and without self-shadow modeling on the shadow portion of our captured dataset. Qualitative comparison in the Appendix.

pearance. Note that we only demonstrate with the tattoo but such appearance deviation can also be scars or nail coloring (please see the Appendix).

Differentiable Self-shadow Modeling. Self-shadowing between fingers and palm is almost inevitable due to highly articulated and dexterous hand movements. However, none of the baselines can properly capture and model self-shadowing. To validate the effectiveness of our shadow-aware differentiable shader, we compare the appearance reconstruction quality with and without shadow modeling. The quantitative comparison is shown in Tab. 5. Without properly modeling shadow, the optimized color at each pixel averages the color when that pixel is in and out of the shadow, resulting in dark patches baked into the texture.

4.2. Hand Pose Reconstruction via Appearance Modeling

With a realistic hand appearance obtained from a video using HARP, we demonstrate that such appearance information can improve the RGB hand pose estimation from the same identity using differentiable rendering if the appearance is known. Intuitively, with differentiable rendering, the optimization should be able to improve the initial prediction using only the hand mask to refine the pose. However, we observe that such masks are not always useful depending on the poses and viewpoints (see Appendix). In such cases, knowing the hand appearance supplements the mask in guiding the optimization toward the correct pose.

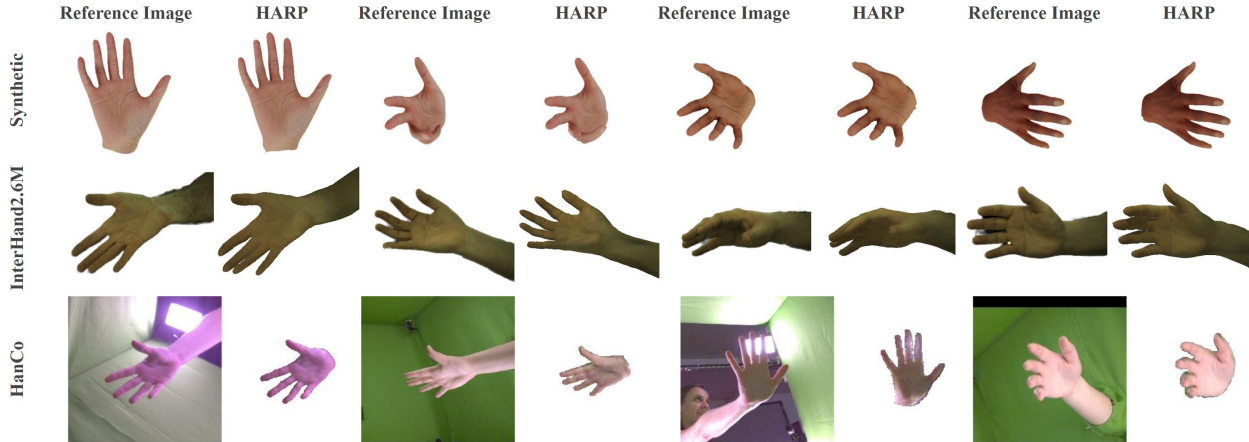To the best of our knowledge, we are the first to qual-

Figure 6. **Results on various datasets** (our Synthetic, InterHand2.6M [46], and HanCo [85]). HARP is very robust, it can obtain faithful hand avatars on different datasets that have a large variety of capture setups, lighting and image conditions, and textures.

Dataset	METRO	HARP-sil	HARP-full	HARP-known
Synthetic	6.12	6.16	6.04	**5.65**

Table 6. Pose error on synthetic sequences (PA-MPVPE in mm).

	IoU ↑	L1 ↓	LPIPS ↓	MS-SSIM ↑
METRO [36]	0.651	-	-	-
S2Hand [8]	0.448	0.059	0.271	0.598
HTML [59]	0.754	0.040	0.144	0.771
NHA [19]	0.826	0.034	0.134	**0.850**
NIMBLE [35]	0.752	0.038	0.157	0.789
HARP	**0.870**	**0.029**	**0.105**	0.831

Table 7. Appearance synthesis evaluation on the test sequences.

itatively and quantitatively demonstrate such improvement for hand pose estimation. We believe that the main component for enabling such improvement lies in the design of the rendering process such as shading and visibility check. However, such topics did not receive much attention in the hand community especially in the context of pose estimation prior to this work. We analyze the scenario in which such improvement is possible in the Appendix.

In Tab. 6, we compare the hand pose error between METRO [36], HARP with silhouette loss only (HARP-sil), a normal HARP (HARP-full), and HARP with a known appearance that is frozen during optimization (HARP-known). The known appearance is obtained from running HARP on a simple video of the same identity. The results indicate that HARP can leverage the differentiable rendering of appearance to improve poses. When the appearance is known in advance (HARP-known), the optimization can perform better as it avoids the shadow and lighting effect that will be baked into the texture during the optimization.

4.3. Novel View and Pose Synthesis

The explicit geometry obtained using HARP also enables consistent 3D geometry and appearance rendering across different poses and viewpoints. We evaluate this feature by rendering the known appearance onto the hands in novel views and novel poses, using the test sequences of our dataset. To match the pose in the test sequences, we optimize only the pose parameter γ (see Sec. 3.1) with respect to the hand masks while freezing all other components. The quantitative results shown in Tab. 7 suggest that HARP can produce consistent appearances in novel views and poses.

5. Conclusion and Limitation

In conclusion, we present HARP, a method for reconstructing personalized hand geometry and appearance from monocular RGB sequences. Starting from a parametric hand model as a geometry backbone, HARP refines the surface to a personalized hand shape. The texture is factorized into an albedo and a normal map. The resulting hand model is robust when rendered in novel views and novel poses, outperforming existing baselines both qualitatively and quantitatively. Furthermore, HARP is efficient, scalable, and compatible with traditional rendering pipelines. It provides a foundation for the realistic experience of personalized hands in AR/VR applications.

Limitation. As our system assumes only a single light source and ambient light, with no specular effect, its ability to replicate the appearance under other lighting assumptions might still be limited. Incorporating an environment map, modeling bounced light, as well as increasing the resolution of the rendered texture are all interesting steps for future works toward a more photorealistic rendering.

References

[1] L. Ballan, A. Taneja, J. Gall, L. van Gool, and M. Pollefeys. Motion capture of hands in action using discriminative salient points. In *European Conference on Computer Vision (ECCV)*, volume 7577 of *LNCS*, pages 640–653. Springer, 2012. 2

[2] Marcel C. Buehler, Abhimitra Meka, Gengyan Li, Thabo Beeler, and Otmar Hilliges. Varitex: Variational neural face textures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 3

[4] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. *ICCV*, 2021. 2

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[6] Ping Chen, Yujin Chen, Dong Yang, Fangyin Wu, Qin Li, Qingpei Xia, and Yong Tan. I2uv-handnet: Image-to-uv prediction network for accurate and high-fidelity 3d hand mesh modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12929–12938, 2021. 2

[7] Xingyu Chen, Baoyuan Wang, and Heung-Yeung Shum. Hand avatar: Free-pose hand animation and rendering from monocular video. *arXiv:2211.12782*, 2022. 3

[8] Yujin Chen, Zhigang Tu, Di Kang, Linchao Bao, Ying Zhang, Xuefei Zhe, Ruizhi Chen, and Junsong Yuan. Model-based 3d hand reconstruction via self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10451–10460, 2021. 2, 3, 6, 7, 8

[9] Zhaoxi Chen and Ziwei Liu. Relighting4d: Neural relightable human from videos. In *ECCV*, 2022. 3

[10] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[11] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 6

[12] 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

[13] Mathieu Desbrun, Mark Meyer, Peter Schröder, and Alan H Barr. Implicit fairing of irregular meshes using diffusion and curvature flow. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 317–324, 1999. 5

[14] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020. 3

[15] Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 40(8), 2021. 3

[16] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 3

[17] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 3

[18] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10833–10842, 2019. 2, 3

[19] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular rgb videos. *arXiv preprint arXiv:2112.01554*, 2021. 2, 3, 6, 7, 8

[20] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, Light, and Material Decomposition from Images using Monte Carlo Rendering and Denoising. *arXiv:2206.03380*, 2022. 4

[21] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[22] Tao Hu, Tao Yu, Zerong Zheng, He Zhang, Yebin Liu, and Matthias Zwicker. Hvtr: Hybrid volumetric-textural rendering for human avatars. In *2022 International Conference on 3D Vision (3DV)*, 2022. 3

[23] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 3

[24] Kaleido AI GmbH. Unscreen. February 2021. 5, 6

[25] 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, 3

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

[27] Giorgos Karvounas, Nikolaos Kyriazis, Iason Oikonomidis, Aggeliki Tsoli, and Antonis A Argyros. Multi-view image-based hand geometry refinement using differentiable monte carlo ray tracing. *BMVC*, 2021. 3

[28] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3907–3916, 2018. 3

[29] Hyeongwoo Kim, Pablo Garrido, Ayush Tewari, Weipeng Xu, Justus Thies, Matthias Niessner, Patrick Pérez, Christian Richardt, Michael Zollhöfer, and Christian Theobalt. Deep video portraits. *ACM Transactions on Graphics (TOG)*, 37(4):1–14, 2018. 3

[30] Nikos Kolotouros, Georgios Pavlakos, and Kostas Daniilidis. Convolutional mesh regression for single-image human shape reconstruction. In *CVPR*, 2019. 3

[31] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[32] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 5

[33] Ruilong Li, Julian Tanke, Minh Vo, Michael Zollhofer, Jurgen Gall, Angjoo Kanazawa, and Christoph Lassner. Tava: Template-free animatable volumetric actors. 2022. 3

[34] 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. International Joint Conferences on Artificial Intelligence Organization, 8 2021. 2

[35] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Yuyao Zhang, Nianyi Li, Yuexin Ma, Lan Xu, and Jingyi Yu. Nimble: A non-rigid hand model with bones and muscles. *arXiv preprint arXiv:2202.04533*, 2022. 2, 6, 7, 8

[36] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, 2021. 2, 3, 5, 6, 7, 8

[37] Shanchuan Lin, Linjie Yang, Imran Saleemi, and Soumyadip Sengupta. Robust high-resolution video matting with temporal guidance, 2021. 6

[38] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2019. 3

[39] Marko Mihajlovic, Aayush Bansal, Michael Zollhoefer, Siyu Tang, and Shunsuke Saito. KeypointNeRF: Generalizing image-based volumetric avatars using relative spatial encoding of keypoints. In *European conference on computer vision*, 2022. 2

[40] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2022. 2

[41] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[42] 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

[43] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 752–768. Springer, 2020. 2, 3

[44] 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 (ECCV)*, 2020. 2

[45] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 3

[46] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 5, 6, 8

[47] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 3

[48] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. Real-time Pose and Shape Reconstruction of Two Interacting Hands With a Single Depth Camera. *ACM Transactions on Graphics (TOG)*, 38(4), 2019. 2

[49] Evonne Ng, Shiry Ginosar, Trevor Darrell, and Hanbyul Joo. Body2hands: Learning to infer 3d hands from conversational gesture body dynamics. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11865–11874, 2021. 2

[50] Merlin Nimier-David, Delio Vicini, Tizian Zeltner, and Wenzel Jakob. Mitsuba 2: A retargetable forward and inverse renderer. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 38(6), Dec. 2019. 3, 4

[51] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 3

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

[53] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021. 2, 3

[54] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 3

[55] 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)*, 2019. 3

[56] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3

[57] Bui Tuong Phong. Illumination for computer generated pictures. *Communications of the ACM*, 18(6):311–317, 1975. 4

[58] 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

[59] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 2, 6, 7, 8

[60] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 4

[61] William T Reeves, David H Salesin, and Robert L Cook. Rendering antialiased shadows with depth maps. In *Proceedings of the 14th annual conference on Computer graphics and interactive techniques*, pages 283–291, 1987. 4

[62] 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), 2017. 2, 3

[63] Viktor Rudnev, Vladislav Golyanik, Jiayi Wang, Hans-Peter Seidel, Franziska Mueller, Mohamed Elgharib, and Christian Theobalt. Eventhands: Real-time neural 3d hand pose estimation from an event stream. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[64] Michael Seeber, Roi Poranne, Marc Polleyfeys, and Martin R Oswald. Realistichands: A hybrid model for 3d hand reconstruction. In *2021 International Conference on 3D Vision (3DV)*, pages 22–31. IEEE, 2021. 3

[65] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. First order motion model for image animation. In *Conference on Neural Information Processing Systems (NeurIPS)*, December 2019. 3

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

[67] B. Smith, Chenglei Wu, He Wen, Patrick Peluse, Yaser Sheikh, J. Hodgins, and Takaaki Shiratori. Constraining dense hand surface tracking with elasticity. *ACM Transactions on Graphics (TOG)*, 39:1 – 14, 2020. 2

[68] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Proceedings of EUROGRAPHICS/ACM SIGGRAPH Symposium on Geometry Processing*, pages 109–116, 2007. 5

[69] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision (ECCV)*, 2020. 3

[70] Shih-Yang Su, Frank Yu, Michael Zollhoefer, and Helge Rhodin. A-nerf: Surface-free human 3d pose refinement via neural rendering. *arXiv preprint arXiv:2102.06199*, 2021. 2, 3, 4

[71] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *International Conference on Computer Vision (ICCV)*, pages 11698–11707, 2021. 2

[72] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019. 3

[73] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3

[74] Shaofei Wang, Marko Mihajlovic, Qianli Ma, Andreas Geiger, and Siyu Tang. Metaavatar: Learning animatable clothed human models from few depth images. In *Advances in Neural Information Processing Systems*, 2021. 2

[75] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Cecilia Zhang. Sunstage: Portrait reconstruction and relighting using the sun as a light stage. *arXiv preprint arXiv:2204.03648*, 2022. 3

[76] Zhou Wang, Eero P Simoncelli, and Alan C Bovik. Multi-scale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 6

[77] 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. 2, 3, 4

[78] Lance Williams. Casting curved shadows on curved surfaces. In *Proceedings of the 5th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '78, page 270–274, New York, NY, USA, 1978. Association for Computing Machinery. 4

[79] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-nerf: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *Advances in Neural Information Processing Systems*, 34, 2021. 3

[80] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William T Freeman, Rahul Sukthankar, and Cristian Sminchisescu. Ghum & ghuml: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6184–6193, 2020. 2

[81] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 3

[82] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 40(3):1–12, 2021. 2

[83] 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*, pages 586–595, 2018. 6

[84] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 4, 5

[85] Christian Zimmermann, Max Argus, and Thomas Brox. Contrastive representation learning for hand shape estimation. In *arxive*, 2021. 5, 8

[86] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 3