arXiv:2211.12782v2 [cs.CV] 22 Mar 2023

Hand Avatar: Free-Pose Hand Animation and Rendering from Monocular Video

Xingyu Chen Baoyuan Wang Heung-Yeung Shum
Xiaobing.AI

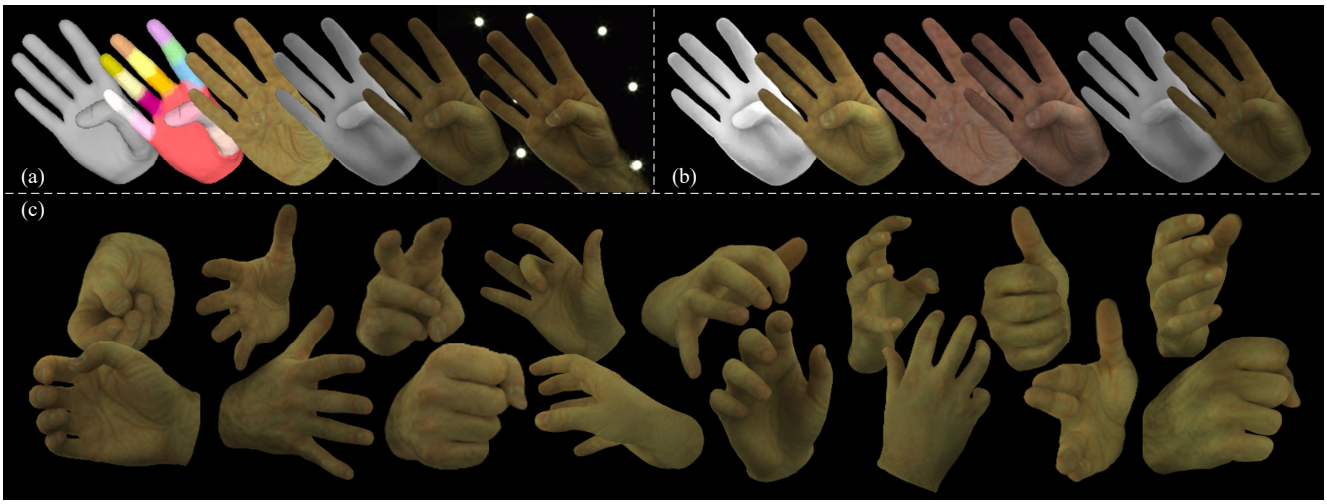

Figure 1. Demonstration of HandAvatar. (a) Personalized hand rendering. From left to right: hand mesh, compositional occupancy, albedo, illumination, shaded appearance, and ground truth; (b) three groups of texture editing in terms of lighting, albedo, and shadow (by altering the self-occlusion effect of thumb); (c) free-pose hand animation and rendering.

Abstract

We present HandAvatar, a novel representation for hand animation and rendering, which can generate smoothly compositional geometry and self-occlusion-aware texture. Specifically, we first develop a MANO-HD model as a high-resolution mesh topology to fit personalized hand shapes. Sequentially, we decompose hand geometry into per-bone rigid parts, and then re-compose paired geometry encodings to derive an across-part consistent occupancy field. As for texture modeling, we propose a self-occlusion-aware shading field (SelF). In SelF, drivable anchors are paved on the MANO-HD surface to record albedo information under a wide variety of hand poses. Moreover, directed soft occupancy is designed to describe the ray-to-surface relation, which is leveraged to generate an illumination field for the disentanglement of pose-independent albedo and pose-dependent illumination. Trained from monocular video data, our HandAvatar can perform free-pose hand animation and rendering while at the same time achieving superior appearance fidelity. We also demonstrate that HandAvatar provides a route for hand appearance editing. Project website: `https://seanchenxy.github.io/HandAvatarWeb`.

1. Introduction

Human avatars [5,16,20,21,28,75] have been vigorously studied for years. However, there has been limited research that particularly focuses on hand avatars [9]. Due to the nature of distinctive properties (*e.g.*, serious self-occlusion and contact) between the hand and the rest of the human parts (*i.e.*, face, head, and body), it is essential to investigate a specialized representation tailored for modeling both the hand geometry and texture.

Traditional pipeline tends to adopt texture maps and colored mesh for hand appearance modeling [7,11,12,25,44], but developing an elaborate personalized hand mesh and texture map usually requires expensive scan data [55] and artistic knowledge. Recently, the neural rendering technique has gained raising attention, where neural radiance field (NeRF) [33] has been adapted to represent humans by predicting geometry and texture properties for an arbitrary

3D point query [9, 10, 21, 26, 38, 40, 41, 49, 52, 59, 63–65, 70, 73, 76] . Compared to the conventional mesh-texture pipeline, NeRF is cheap in training data collection and superior in rendering fidelity. Despite the huge success of human body and face modeling, neural rendering-based hand representation [9] remains much less explored. The hand is highly articulated such that the complex hand motion brings difficulties for neural rendering. Firstly, the deformation of hand geometry is hard to model. When coping with large and complex hand deformations (*e.g.*, self-contact), previous skinning-based methods can hardly find accurate skinning weights for an arbitrary query [3, 6, 19, 20, 32, 36, 40, 48, 63, 75], while part-aware methods usually suffer from across-part inconsistency issue [18, 22, 31, 61]. Secondly, hand texture is hard to model because of the highly articulated structure. For example, articulated hand motion induces serious self-occlusion so that different hand poses lead to noticeable variations in illumination and shadow patterns. Illumination is important for realistic rendering, but we are not aware of any prior work in estimating illumination caused by articulated self-occlusion.

Motivated by the above challenges, we propose HandAvatar for animatable realistic hand rendering. Considering different difficulties in geometry and texture modeling, we follow the idea of inverse graphics [77] to disentangle hand geometry, albedo, and illumination. At first, we employ explicit mesh to depict hand shapes. However, the popular hand mesh model, *i.e.*, MANO [45], only provides a coarse mesh with 778 vertices, whose shape fitting capacity is limited. Therefore, we design a super-resolution version of MANO with 12,337 vertices and 24,608 faces, namely MANO-HD, which can fit personalized hand shapes with per-vertex displacements. Additionally, massive existing MANO-annotated data can be seamlessly represented by MANO-HD. For introducing mesh-based hand shape to the volume rendering pipeline [33], we propose a local-pair occupancy field (PairOF), where every two part-level geometry encodings are reassembled according to physical connections to yield an across-part consistent field. As for hand texture, we propose a self-occlusion-aware shading field (SelF). SelF is comprised of an albedo field and an illumination field. The albedo field resorts to anchors that are uniformly paved on MANO-HD surfaces, each of which holds positional and albedo encodings to model a small hand region. The illumination field is to cope with articulated self-occlusion, where directed soft occupancy is designed to estimate illumination and shadow patterns.

MANO-HD and PairOF are pre-trained with MANO parameter annotations, then they cooperate with SelF in end-to-end training on monocular video data. Finally, with hand pose as the input, our HandAvatar can perform hand animation and rendering. We evaluate our approach on the InterHand2.6M dataset [35] and achieve high-fidelity geometry and texture for free-pose hand animation. We also demonstrate that it is convenient to edit hand appearance in HandAvatar as shown in Fig. 1. Therefore, our main contributions are summarized as follows:

- We propose a HandAvatar framework, the first method for neural hand rendering with self-occluded illumination.
- We develop MANO-HD and a local-pair occupancy field that fit hand geometry with personalized shape details.
- We propose a self-occlusion-aware shading field that can render hand texture with faithful shadow patterns.
- Our framework is end-to-end developed for free-pose realistic hand avatars. Extensive evaluations indicate our method outperforms prior arts by a large margin.

2. Related Work

Articulated Human Geometry. Parametric human models [24, 29, 39, 45] have developed for years, where mesh can be inferred given pose and shape parameters. Specifically, as a common-used hand model, MANO [45] can produce a hand mesh with 778 vertices and 1,538 faces. This mesh template is too coarse so its representation capacity is largely limited. Gyeongsik *et al.* [34] proposed DeepHandMesh to generate dense and high-fidelity hand mesh, but brought restricted generalization as multi-view depth data was required for training. In contrast, our MANO-HD is a general high-resolution hand mesh model so that all existing MANO-annotated data can be seamlessly represented using MANO-HD. Meanwhile, MANO-HD can fit personalized hand shapes with monocular RGB video data.

Mesh suffers drawbacks of discontinuity and unalterable topology structure. To remedy this issue, recent research tends to explore implicit human geometry [2, 23, 32, 47, 68, 74], which has the advantages of flexibility and continuity. For example, GraspField [23] leveraged the signed distance field (SDF) to describe hand-object contact. However, implicit geometry is poor in free-pose animation when compared to explicit mesh, so the articulated driving of implicit human geometry is widely studied. As reported in [3, 36, 40, 63], a posed-space query can be transformed back to canonical space with linear blend skinning and inverse skinning weights. The inverse skinning paradigm fails to deal with self-contact, where a query can match multiple canonical-space points. Then, forward skinning deformation is designed to transform canonical-space points to posed space with an iterative root finding method [6, 20, 32, 48, 75], but the iterative optimization algorithm could hurt end-to-end network training. By and large, per-bone rigid transformations can compose a large motion space with the difficulty of optimizing accurate skinning weights for an arbitrary 3D point query. With the aid of parametric models [29, 45], another idea of deformation between posed and canonical spaces is to leverage the surface

motion [26, 54, 65]. For a query, a mesh-surface point is found according to Euclidean distance as the reference, then the query deformation is set the same as that of the reference point. Yuan *et al.* [69] and Garbin *et al.* [13] pointed out that the reference from triangular mesh is not accurate enough and proposed to use tetrahedral mesh. However, the deformation of tetrahedral mesh is hard to cooperate with popular human priors [24,29,45] and could be potentially slow [50]. Without requiring motion approximation, part-aware methods [18, 22, 31, 61] are developed by fusing part-wise geometries. NASA [18] divided the body into per-bone parts, then a query was deformed into each part space with unambiguous rigid transformation for decoding of part-level occupancy. NASA can describe complex deformations owing to accurate query motion, but information between body parts is ignored. To relieve this issue, COAP [31] encoded connected parts with PointNet [43], but it still incurred non-smooth part connections. Instead of part-wise modeling, we propose a part-pair-wise decoder to generate across-part consistent geometry.

Human Texture. Previously, the primary focus on hand texture is the texture map and colored mesh [7, 11, 12, 25, 44]. Although many high-quality texture maps are explored, the design of personalized texture maps usually requires expensive scan data [55] and artistic knowledge [11]. In contrast, LISA [9] employed radiance field [66] to learn hand appearance from multi-view images and introduced color parameters for texture generalization. Different from LISA, we design a monocular method for the convenience of data collection. Moreover, we pay attention to detailed personalized textures including albedo and illumination. Because of the aforementioned difficulty in implicit deformation, the learned texture on the human surface could be blurred [20, 63]. To enhance surface texture representation, local representations are developed with explicit mesh as the guidance. NeuralBody [41] attached latent codes to mesh vertices, which can diffuse into space with sparse convolution [15]. NeuMesh [8] also put color features on mesh vertices, and achieved an editable radiance field. Furthermore, mesh-guided local volume [28] and local radiance field [76] were designed. We follow the local modeling paradigm and uniformly place anchors on MANO-HD surface using barycentric sampling to trace local information.

Human Inverse Rendering. Most methods model human appearance with entangled geometry, albedo, and illumination [63]. Meanwhile, there has been a surge of interest in human inverse rendering, the purpose of which is to extract intrinsic components (*i.e.*, geometry, material, and illumination) from RGB data [37,56–58,77]. For example, GAN2X [37] designed an unsupervised framework to model albedo and specular properties of non-Lambertian material, then rendered face with Phong shading [42]. With a similar purpose, HyFRIS-Net [77] disentangled albedo and illumination with an inverse 3DMM model to achieve a considerably improved quality of face rendering. S2HAND [7] simultaneously estimated camera pose, colored mesh, and lighting condition to form a photometric loss for hand pose estimation, but its rendering quality was coarse without a detailed appearance. Although the inverse rendering technique on the human face has been becoming a well-studied issue, the knowledge cannot be trivially transferred to hand tasks. Different from the face, the hand is characterized by articulated self-occlusion. Illumination and shadow caused by self-occlusion have not yet been discussed in prior works, and thus we fill this gap for hand inverse graphics.

Illumination in Radiance Field. The existing literature on NeRF-based illumination technique [4, 51, 53, 72] is to estimate source light condition or surface reflection property (*i.e.*, bidirectional reflectance distribution function, BRDF). For example, NeRV [51] took as input a set of images under known lighting to predict BRDF, and achieved novel-view rendering with arbitrary lighting conditions. NeLF [53] designed a lighting estimation module, and then performed face relighting. NeRF-OSR [46] collected multi-view outdoor images to predict albedo and shadow maps. In contrast to the prior art, we dedicate modeling illumination under the condition of articulated self-occlusion.

3. Method

Fig. 2 illustrates the overview pipeline of our HandAvatar system, including MANO-HD (Sec. 3.1), PairOF (Sec. 3.2), and SelF (Sec. 3.3). Table 1 also provides the list of symbol notations and their definitions used in this paper.

3.1. MANO-HD

Mesh Subdivision. MANO [45] deforms hand mesh with shape parameter $\boldsymbol{\beta}$ and pose parameter $\boldsymbol{\theta} \in \mathbb{R}^{B\times 3}$ ($B = 16$ indicates the number of per-bone parts). For lifting mesh resolution, we uniformly subdivide MANO template mesh by adding new vertices on edge middle points [16]. This operation increases the vertex number to 12,337 and the face amount to 24,608 (see Fig. 3). Then, the skinning weights of added vertices are given with the average of their seminal vertices. To eliminate artifacts during skinning, we optimize upsampled skinning weights for better dynamic performance. Please see the *suppl. material* for details.

Shape Fitting. Although MANO-HD has a high-resolution template, its shape is still bounded by $\boldsymbol{\beta}$. Hence, when modeling personalized hand mesh, we get rid of $\boldsymbol{\beta}$ and use a multi-layer perceptron (MLP) to derive a refined shape $\hat{\mathbf{V}} = \bar{\mathbf{V}} + \mathcal{M}_{shape}([\mathcal{P}(\bar{\mathbf{V}}), \boldsymbol{\theta}])$, where $\bar{\mathbf{V}}, \mathcal{P}(\cdot), [\cdot]$ denote the MANO-HD template vertices, positional encoding, and concatenation. The MLP can be trained with monocular

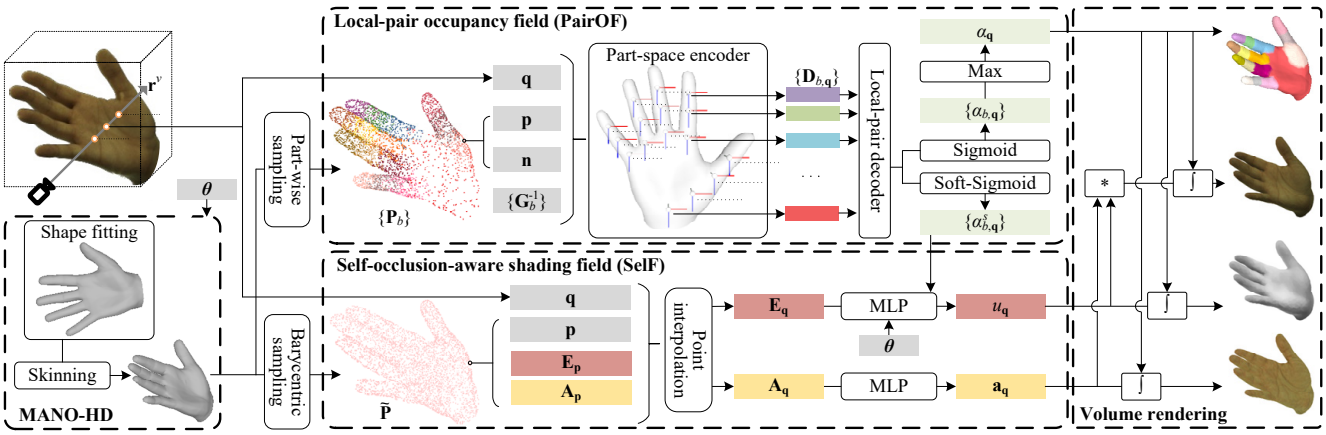

Figure 2. HandAvatar overview. Given hand pose, MANO-HD produces personalized mesh, while PariOF yields accordingly occupancy field. SelF estimates albedo and illumination fields under self-occlusion. Then, hand appearance is synthesized by volume rendering.

$\boldsymbol{\theta}$	hand pose parameter	$\boldsymbol{\beta}$	hand shape parameter
$\mathbf{P}$	point cloud	$\mathbf{r}$	ray direction
b	index of per-bone parts	$\mathbf{n}$	sampled point normal
$\mathbf{q}$	query point position	$\mathbf{p}$	sampled point position
$\mathbf{G}_b$	bone transformation matrix	$\mathbf{D}_b$	part geometry encoding
$\mathbf{E}$	positional encoding	$\mathbf{A}$	albedo encoding
u	illumination value	$\mathbf{a}$	albedo value in RGB
α	occupancy value	α^s	soft occupancy value

Table 1. Symbolical notations.

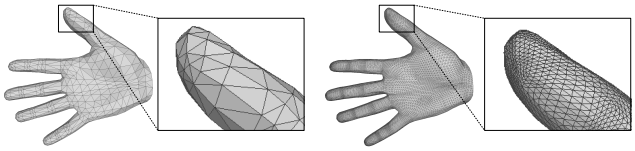

Figure 3. Templates of MANO (left) and MANO-HD (right).

video data and IoU loss $\mathcal{L}_{shape} = 1 - \text{IoU}(\mathcal{D}(\mathcal{S}(\tilde{\mathbf{V}})), \mathbf{S}^*)$, where $\mathcal{D}, \mathcal{S}, \mathbf{S}^*$ are silhouette rendering [27], linear blend skinning, and ground-truth silhouette.

3.2. Local-Pair Occupancy Field

Given the query point $\mathbf{q} \in \mathbb{R}^3$, PairOF predicts the occupancy value $\alpha_{\mathbf{q}}$ to describe whether it locates in ($\alpha_{\mathbf{q}} > 0.5$) or out of ($\alpha_{\mathbf{q}} < 0.5$) the surface. Hence, the hand surface can be formulated as $\{\mathbf{q} | \alpha_{\mathbf{q}} = 0.5\}$.

Part-Space Encoder. Following NASA [18], we divide hand mesh into per-bone part meshes and uniformly sample N^p points on part mesh faces to obtain point clouds $\mathbf{P}_b = \{\mathbf{p} \in \mathbb{R}^3\}$ and normals $\mathbf{N}_b = \{\mathbf{n} \in \mathbb{R}^3\}$. Points, normals, and query are transferred back to part canonical spaces with $\hat{\mathbf{P}}_b = \mathbf{G}_b^{-1}\mathbf{P}_b, \hat{\mathbf{N}}_b = \mathbf{R}_b^{-1}\mathbf{N}_b, \hat{\mathbf{q}}_b = \mathbf{G}_b^{-1}\mathbf{q}$, where $\mathbf{G}$ is bone transformation matrix and $\mathbf{R}$ is the rotation component of $\mathbf{G}$. Following COAP [31], $\hat{\mathbf{P}}_b, \hat{\mathbf{N}}_b$ are fed to a PointNet $\mathcal{Q}_{part}$ to extract latent geometry features. The part geometry encoding is ultimately formulated by concatenating PointNet representation and canonicalized query, *i.e.*, $\mathbf{D}_{b,\mathbf{q}} = [\mathcal{Q}_{part}([\hat{\mathbf{P}}_b, \hat{\mathbf{N}}_b]), \hat{\mathbf{q}}_b]$.

Local-Pair Decoder. To explore inter-part relations, some previous works fuse per-bone features according to the kinematic tree to form structured representations [1, 14, 32]. Our intuition is that the local part shape is not related to the kinematic tree, instead, there is a strong relationship between parts. Hence, we define local pair as two parts that are physically connected. Then, we propose local-pair decoder $\mathcal{Q}_{pair}$ based on PointNet to fuse each paired encodings and predict occupancy value:

$$\alpha_{b,\mathbf{q}} = \sigma(\max\{\mathcal{Q}_{pair}(\{\mathbf{D}_{b,\mathbf{q}}, \mathbf{D}_{b',\mathbf{q}}\}) | b' \in \mathbb{P}(b)\}) \quad (1)$$

where $\mathbb{P}(b)$ selects locally paired parts that has a physical connection with part b; σ is the sigmoid function; $\alpha_{b,\mathbf{q}}$ is part-level occupancy value. Through the part-pair-wise decoding, the part boundaries become blurred. Therefore, the intuition behind the maximum operator in Eq. (1) is to yield a union of part-level geometries that extend to the connections. Finally, The global occupancy value is given by fusing part-level values, *i.e.*, $\alpha_{\mathbf{q}} = \max\{\alpha_{b,\mathbf{q}}\}_{b=1}^{B}$.

Pre-Training. With free MANO parameter annotations, we pre-train PairOF as a prior model to endow PairOF with prior knowledge of 3D hands. With a hand mesh inferred by MANO-HD, we sample point clouds with N^t points as training data [31]. The objective is to minimize the mean squared error between ground truth α^* and predicted occupancy values, *i.e.*, $\mathcal{L}_{PairOF} = \frac{1}{N^t}\sum_{\mathbf{q}}(\alpha_{\mathbf{q}} - \alpha^*_{\mathbf{q}})^2$.

3.3. Self-Occlusion-Aware Shading Field

Given a query $\mathbf{q}$, SelF predicts its albedo and illumination values, which are rendered with the volumetric method.

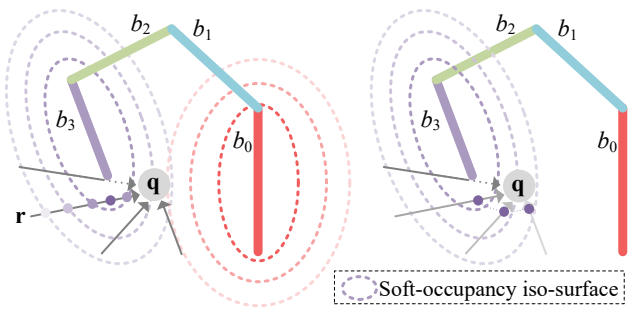

Figure 4. Illustration of self-occluded illumination along the kinematic tree of the forefinger. The deepest purple positions contribute to directed soft occupancy value.

Volume Rendering. For a ray casting on view direction $\mathbf{r}^v$, we uniformly sample N^q queries $\{\mathbf{q}_i\}_{i=1}^{N^q}$, each of which has occupancy α, albedo $\mathbf{a}$, and illumination value u. We render neural fields with the volumetric method [33]:

$$\mathbf{X}_{\mathbf{r}^v} = \sum_{i=1}^{N^q}(\prod_{j=1}^{i-1}(1-\alpha_{\mathbf{q}_j}))\alpha_{\mathbf{q}_i}\mathbf{X}_{\mathbf{q}_i}. \quad (2)$$

When $\mathbf{X}$ equates to $\mathbf{a}$, u, or $u\mathbf{a}$, we obtain the albedo value, illumination value, or shaded RGB color of a pixel.

Albedo Field. Albedo describes the intrinsic color of the material, which is invariant *w.r.t.* hand pose, illumination, *etc.* Motivated by the invariant property, we fix anchors on the MANO-HD surface, whose relative geodesic locations are independent from varying poses. To this end, we uniformly sample point clouds $\bar{\mathbf{P}}$ with N^a points on MANO-HD template mesh and represent them with barycentric coordinates. Compared to directly using vertex as the anchor [8,28,41], our barycentric anchors are more uniform to cover the hand surface. Then, we develop albedo encodings $\mathbf{A} \in \mathbb{R}^{N^a \times D^a}$ with random initialization and attach them to anchors. Given hand pose, anchors can be re-sampled based on deformed vertices and fixed barycentric coordinates to form deformed points clouds $\tilde{\mathbf{P}}$. For a query $\mathbf{q}$, we find N^n nearest points in $\tilde{\mathbf{P}}$ and interpolate $\mathbf{A}$ using inverse Euclidean distances as the weights. Thereby, we obtain the albedo encoding $\mathbf{A}_\mathbf{q} \in \mathbb{R}^{D^a}$ and then fed it to an MLP to predict the albedo value $\mathbf{a}_\mathbf{q} \in \mathbb{R}^3$, *i.e.*, $\mathbf{a}_\mathbf{q} = \mathcal{M}_{albedo}(\mathbf{A}_\mathbf{q})$.

Directed Soft Occupancy. For self-occluded illumination estimation, we require the near-far relationship for a bone part. That is, the illumination of a query $\mathbf{q}$ is affected by self-occlusion when $\mathbf{q}$ is close to multiple parts. Although occupancy value can describe the relation between $\mathbf{q}$ and parts, the value is nearly binary so can only depict inside-outside relations. Hence, a soft factor τ is introduced to the sigmoid function to soften the occupancy value:

$$\sigma^s(x) = \frac{1}{1+e^{-\tau x}}, \quad 0 < \tau < 1. \quad (3)$$

Soft occupancy $\alpha^s_{b,\mathbf{q}}$ is derived by replacing σ in Eq. (1) with σ^s. Further, instead of modeling a single query, we design *directed soft occupancy* to reflect the near-far relation between a ray casting and an articulated part. For a ray casting $\mathbf{r}$ that can reach $\mathbf{q}$, the directed soft occupancy $\alpha^s_{b,\mathbf{q},\mathbf{r}}$ is defined as the maximal value on $\mathbf{r}$ before the ray hits $\mathbf{q}$. For discretization, we uniformly sample queries $\{\mathbf{q}_i\}_{i=1}^{N^q}$ on a ray casting $\mathbf{r}$, and compute directed soft occupancy as

$$\alpha^s_{b,\mathbf{q},\mathbf{r}} = \max\{\alpha^s_{b,\mathbf{q}_i} | \mathbf{q}_i \leq \mathbf{q}\}, \quad (4)$$

where $\mathbf{q}_i \leq \mathbf{q}$ selects queries that the ray traverses before reaching $\mathbf{q}$. For example, $\alpha^s_{b_3,\mathbf{q},\mathbf{r}}$ equals to $\alpha^s_{b_3}$ of the deepest purple query in Fig. 4.

Illumination Field. It is well known that the illumination effects come with light-source distribution, irradiance, and reflectance. Independent from self-occlusion, reflectance is the material property, which is not our focus. Affected by self-occlusion, some ambient lighting rays could be occluded such that the irradiance could be changed. Thereby, the problem is formulated as estimating irradiance of an outside query $\mathbf{q}$ ($\alpha^s_{b,\mathbf{q}} < 0.5$), which indicates the energy amount that can reach $\mathbf{q}$. To this end, we use the hand pose $\boldsymbol{\theta}$ and query location as cues. Similar to the albedo encodings, positional encodings $\mathbf{E} = \mathcal{P}(\bar{\mathbf{P}})$ are attached to anchors, and we obtain $\mathbf{E}_\mathbf{q}$ with interpolation as the surface-calibrated location of $\mathbf{q}$. Nevertheless, self-occlusion is quite intractable for $\boldsymbol{\theta}$ and $\mathbf{E}_\mathbf{q}$, so we leverage directed soft occupancy to enhance the awareness of self-occlusion.

As shown in Fig. 4, the articulated structure prohibits a portion of energy from arriving $\mathbf{q}$. Apparently, the situation of energy occlusion around a ray direction is implied in a set of directed soft occupancy $\{\alpha^s_{b,\mathbf{q},\mathbf{r}}\}_{b=1}^B$. That is, if a ray casting is close to multiple parts before hitting $\mathbf{q}$, the illumination of $\mathbf{q}$ shall be impacted by self-occlusion. Prohibited by a limited memory budget, we cannot consider spherically distributed ray directions, and thus the number of ray castings is imperative to be reduced. Our institution is that (1) the selected ray should be able to arrive $\mathbf{q}$ (*i.e.*, $\alpha^s_{b,\mathbf{q},\mathbf{r}} < 0.5$) such that can elaborate the near-far relations for all articulated parts; (2) an articulated part can only affect the illumination around it, where the query is close to the part (*i.e.*, $\alpha^s_{b,\mathbf{q}} \to 0.5$). Meanwhile, we have $\alpha^s_{b,\mathbf{q},\mathbf{r}} \geq \alpha^s_{b,\mathbf{q}}$ from Eq. (4). Thereby, the variation caused by ray directions is minor, and we use $\{\alpha^s_{b,\mathbf{q},\mathbf{r}^v}\}_{b=1}^B$ as the guidance to estimate irradiance of $\mathbf{q}$, where $\mathbf{r}^v$ is the view direction. Without introducing extra ray castings, we significantly reduce computational costs by leveraging the ray casting and queries on the view direction.

Finally, we use an MLP to predict the illumination value, *i.e.*, $u_\mathbf{q} = \mathcal{M}_{illum}([\boldsymbol{\theta}, \mathbf{E}_\mathbf{q}, [\alpha^s_{b,\mathbf{q},\mathbf{r}^v}]_{b=1}^B])$.

Optimization. The training of SelF is based on reconstruction loss functions, including LPIPS [71] loss and l_1 error between the rendered image $\mathbf{C}$ and the ground truth $\mathbf{C}^*$, *i.e.*, $\mathcal{L}_{SelF} = \mathcal{L}_{LPIPS}(\mathbf{C}, \mathbf{C}^*) + ||\mathbf{C} - \mathbf{C}^*||_1$.

Method	#Param	Guided mesh	IoU (%) ↑	*Lap.* ↓	*Cham.* ↓
COAP [31]	138K	MANO	94.08	2.371	8.564
COAP	138K	MANO-HD	94.01	2.348	7.694
COAP*	287K	MANO-HD	95.08	2.339	7.405
PairOF †	256K	MANO-HD	96.06	2.288	**6.998**
PairOF (ours)	237K	MANO-HD	**96.32**	**2.281**	7.151

Table 2. Effects of MANO-HD and PairOF. *: w/ wider-MLP decoder; †: w/ Transformer-based decoder.

Figure 5. Effects of MANO-HD and PariOF. (a) COAP w/ MANO. (b) COAP w/ MANO-HD. (c) PariOF w/ MANO-HD.

4. Experiments

4.1. Implementation Details and Metrics

Pre-Training of PairOF. We adopt all right-hand annotations in InterHand2.6M [35] for pre-training, whose training/test set contains 875,530/565,611 samples. Learnable parameters includes $\mathcal{Q}_{part}, \mathcal{Q}_{pair}$. We set $N^t = N^p = 256$, and the training process is to minimize $\mathcal{L}_{PairOF}$.

End-to-End Training. With personalized monocular video, we optimize $\mathcal{M}_{shape}, \mathcal{Q}_{pair}, \mathcal{M}_{albedo}, \mathcal{M}_{illum}$ and $\mathbf{A}$ in an end-to-end manner. Video data are selected from InterHand2.6M dataset [35]. The objective is to minimize $\mathcal{L}_{shape} + \mathcal{L}_{PairOF} + \mathcal{L}_{SelF}$. Hyperparameters in SelF are set as $N^q = 64, N^a = 4096, N^n = 4, D^a = 128, \tau = 0.05$. The rendering resolution is 256×256. Please see the *suppl. material* for data selection, pre-processing, and more training details.

Metrics Following COAP [31], IoU is used to evaluate the occupancy field. We also employ Laplacian smooth (*Lap.*) and Chamfer distance (*Cham.*) to evaluate the mesh quality, the latter of which is formulated as the minimal distance between the vertices extracted from occupancy field [30] and the guided mesh faces. *Lap.* and *Cham.* are presented in 10^{-4}m. Consistent with HumanNeRF [63], we report LPIPS [71], PSNR, and SSIM [62] to reflect image similarity as the metrics of rendering quality. All evaluation data are with novel poses that are unseen in training.

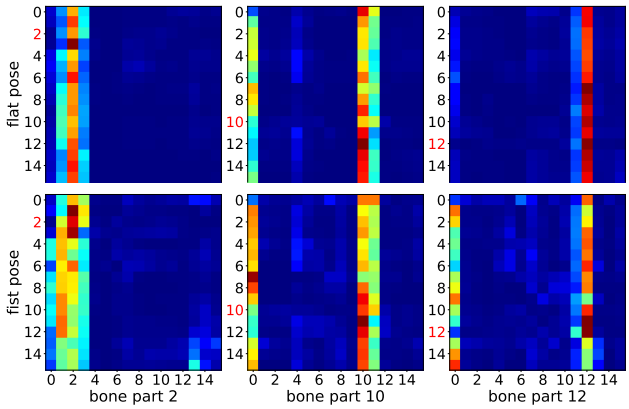

Figure 6. Attention map for fusing part-wise geometry encodings. The rows with red indices contribute to global occupancy value.

4.2. Evaluation on Geometry Performance

Comparison with Prior Arts PairOF and COAP [31] use the same encoder but different decoders, so their comparison can reveal the effect of our part-pair-wise decoding. The results of COAP are from the officially released code, and we re-train models on InterHand2.6M dataset. Because our local-pair decoder is larger than that of COAP, we enlarge MLP width for comparable model size (denoted as COAP*). Referring to Table 2, an occupancy field guided by MANO-HD has a smoother surface (lower *Lap.*) and higher fidelity (lower *Cham.*), so there are benefits of MANO-HD over MANO in guiding an implicit function. Moreover, PairOF can improve all metrics by a large margin. As shown in Fig. 5, MANO-HD can improve overall smoothness, while PairOF exhaustively eliminates non-smooth part connections to achieve across-part consistency.

Comparison with Transformer-Based Decoder To verify the local-pair prior knowledge in PairOF, the Transformer [60] technique is employed as the decoder, where self-attention can adaptively fuse part-wise geometry without inductive prior. As shown in Table 2, the local-pair decoder performs on par with the Transformer-based decoder. To unveil the effect of self-attention, we delve deep into attention-based feature fusion based on two representative hand poses (*i.e.*, flat and fist poses). At first, we extract mesh vertices from the occupancy field, each of which comes with respective attention maps. Then, we find the part that has the maximal occupancy value for each vertex and gather vertices into groups accordingly. Each group of vertices can reflect the property of a bone part, and we show their average attention maps in Fig. 6. Because of the maximum operator in part-wise geometry fusion, only one part contributes

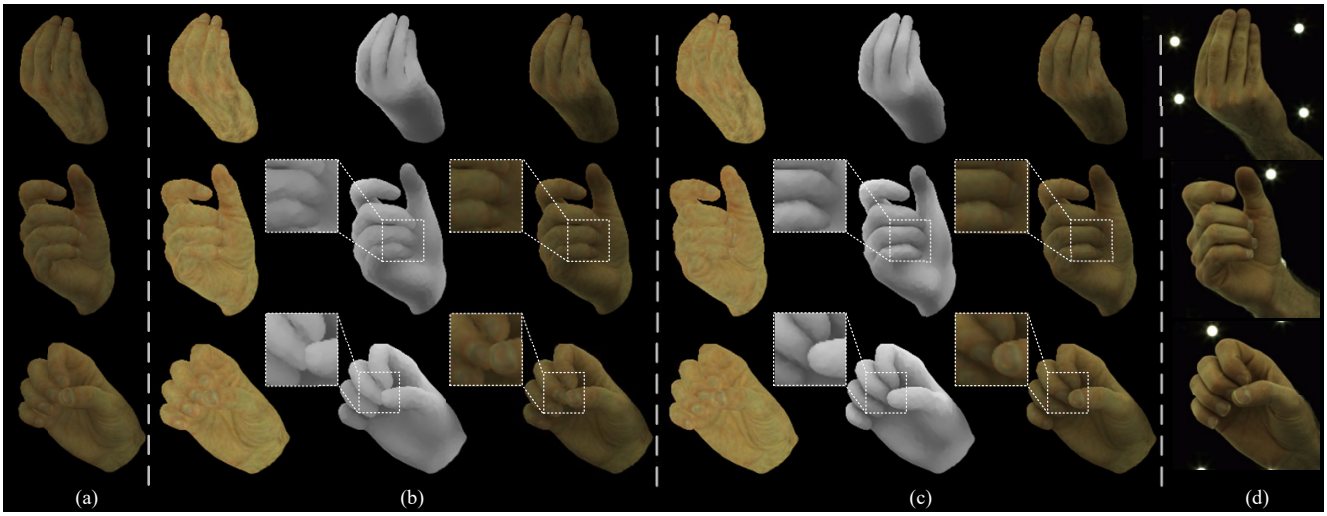

Figure 7. Effects of the disentangled albedo and illumination fields in SelF. (a) Coupled albedo and illumination. (b,c) Disentangled albedo and illumination; directed soft occupancy is not involved in (b); from left to right: albedo, illumination, shaded image. (d) Ground truth.

to the global occupancy value. That is, for part b, we should focus on the bth row (red number in Fig. 6). Consequently, attention-based fusion is consistent with our local-pair design. For example, the attention map fuses parts b_0, b_{10}, b_{11} to evolve the encoding of part b_{10}. Therefore, the design concept of the local-pair decoder is evident. Nevertheless, the Transformer-based decoder is not efficient enough because the attention map contains *vertical* patterns instead of *diagonal* ones. That is, meaningless computations are introduced by the Transformer, despite they do not contribute to global occupancy. For example, referring to "fist pose" and "bone part 12" in Fig. 6, the attention map integrates encodings of parts b_0, b_{12} (instead of b_1) for the prediction of part b_1. The reason behind this is that the *inside* property is exclusively enhanced, and vertices belonging to part b_{12} also have inside properties to part b_0 under the fist pose. Refer *suppl. material* for more details and part indices.

4.3. Evaluation on Rendering Quality

Ablation Study on Shape Fitting We fit personalized hand shape with $\mathcal{M}_{shape}$. For comparison, $\mathcal{M}_{shape}$ is replaced with $\boldsymbol{\beta}$-based shape fitting, where $\boldsymbol{\beta}$ is the annotation in InterHand2.6M dataset. As shown in Table 3, $\boldsymbol{\beta}$-based shape induces poor rendering quality, while our method brings a significant improvement. Therefore, the shape fitting capacity of our proposed MANO-HD is confirmed.

Ablation Study on SelF Referring to Fig. 7(a) and (b), the disentanglement of albedo and illumination fields can improve the rendering reality by introducing shadow patterns. Moreover, directed soft occupancy can further elevate the illumination representation through ray-based occlusion estimation. Referring to Fig. 7(c), it is remarkable that the shadow patterns on the palm and fingers are more faithful with fewer artifacts when compared to Fig. 7(b). In Table 3, the illumination field and directed soft occupancy lead to quantitative improvements in rendering metrics, indicating our SelF is conducive to realistic rendering.

Shape fit.	Illum.	Dir. occ.	LPIPS ↓	PSNR ↑	SSIM ↑
			0.1268	26.53	0.8692
✓			0.1113	27.32	0.8830
✓	✓		0.1063	28.02	0.8903
✓	✓	✓	**0.1035**	**28.23**	**0.8941**

Table 3. Effects of shape fitting (Shape fit.), illumination field (Illum.), and directed soft occupancy (Dir. occ.) on rendering quality. Results are from Interhand2.6M *test/Capture0*.

Comparison with Prior Arts We compare HandAvatar with previous monocular methods HumanNeRF [63] and SelfRecon [20], both of which are from officially released codes and re-trained on InterHand2.6M dataset. SelfRecon uses the surface-based rendering [67] method, and its representation of texture detail is not good enough, as shown in Fig. 8. HumanNeRF and our method leverage the volume rendering method [33], which can produce realistic hand texture. However, limited by inverse skinning deformation, HumanNeRF cannot cope with self-contact that commonly occurs in hand animation. As shown in Fig. 8, HumanNeRF has corrupted geometry when fingers contact with each other. In contrast, benefiting from our PairOF, HandAvatar has the advantage of free-pose animation while at the same time maintaining geometry fidelity. In addition, both SelfRecon and HumanNeRF employ entangled albedo and illumination for color prediction, so the shadow on hand is hard to be aware of, as shown in the 3rd-8th

Method	*test/Capture0*			*test/Capture1*			*val/Capture0*		
	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑
SelfRecon [20]	0.1421	26.38	0.8786	0.1389	25.18	0.8758	0.1490	25.78	0.8687
HumanNeRF [63]	0.1145	27.64	0.8836	0.1177	26.31	0.8803	0.1192	27.80	0.8816
ours	**0.1035**	**28.23**	**0.8941**	**0.1076**	**26.56**	**0.8902**	**0.1062**	**28.04**	**0.8900**

Table 4. Rendering quality comparison among our HandAvatar and prior arts on the InterHand2.6M dataset.

Figure 8. Visualization results of SelfRecon [20], HumanNeRF [63], and our HandAvatar on free-pose animation and rendering.

rows of Fig. 8. On the contrary, our HandAvatar leads to a superiority in modeling illumination and shadow patterns caused by articulated self-occlusion. This is attributed to the disentanglement of albedo and illumination brought by our proposed SelF. Further, referring to the white arrows in Fig. 8, we can present varying shadow intensities for different self-occlusion situations, thanks to our directed soft occupancy. For quantitative comparison, we train models on three video sequences and achieve the best results in all metrics, as shown in Table 4.

Figure 9. Visualization results of LISA [9] and our HandAvatar. The result of LISA is cropped from their original paper.

As the most related work to this paper, LISA [9] is trained on non-released multi-view data, and its models/codes remain unavailable. Thereby, we compare LISA based on the result reported in their original paper. As shown in Fig. 9, LISA has difficulty in capturing accurate hand pose with a learnable skinning-based deformation. Besides the faithful shape and pose reconstruction, our rendered texture details are more realistic than that of LISA.

In addition to rendering fidelity, HandAvatar also provides a route for appearance editing as shown in Fig. 1.

5. Conclusions

In this work, we present a novel hand representation called HandAvatar for free-pose animation and rendering. First, we extend MANO to MANO-HD as a high-resolution topology structure to improve the shape-fitting capacity of hand mesh. Subsequently, PairOF with a local-pair decoder is developed, which can generate an across-part consistent occupancy field. Furthermore, we propose SelF, the first approach to model hand texture under articulated self-occlusion, to disentangle hand albedo and illumination. Extensive experiments demonstrate our superior results on free-pose hands animation and rendering. We believe our method paves a new way for dynamic hand representation.

Limitation and future works For affordable computational costs, we use the directed soft occupancy on view direction to estimate the irradiance. This could lead to view-direction-dependent shadow patterns. Thus, the improved illumination field is worthy of ongoing exploration.

References

[1] Emre Aksan, Manuel Kaufmann, and Otmar Hilliges. Structured prediction helps 3D human motion modelling. In *ICCV*, 2019.

[2] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, 2021.

[3] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. LoopReg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *NeurIPS*, 2020.

[4] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. NeRD: Neural reflectance decomposition from image collections. In *ICCV*, 2021.

[5] Chen Cao, Tomas Simon, Jin Kyu Kim, Gabe Schwartz, Michael Zollhoefer, Shun-Suke Saito, Stephen Lombardi, Shih-En Wei, Danielle Belko, Shoou-I Yu, et al. Authentic volumetric avatars from a phone scan. *ACM TOG*, 2022.

[6] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021.

[7] 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 *CVPR*, 2021.

[8] Chong Bao and Bangbang Yang, Zeng Junyi, Bao Hujun, Zhang Yinda, Cui Zhaopeng, and Zhang Guofeng. NeuMesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *ECCV*, 2022.

[9] 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 *CVPR*, 2022.

[10] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. GRAM: Generative radiance manifolds for 3D-aware image generation. In *CVPR*, 2022.

[11] Daiheng Gao, Yuliang Xiu, Kailin Li, Lixin Yang, Feng Wang, Peng Zhang, Bang Zhang, Cewu Lu, and Ping Tan. DART: Articulated hand model with diverse accessories and rich textures. In *NeurIPS*, 2022.

[12] Daiheng Gao, Xindi Zhang, Xingyu Chen, Andong Tan, Bang Zhang, Pan Pan, and Ping Tan. CycleHand: Increasing 3D pose estimation ability on in-the-wild monocular image through cyclic flow. In *ACM MM*, 2022.

[13] Stephan J Garbin, Marek Kowalski, Virginia Estellers, Stanislaw Szymanowicz, Shideh Rezaeifar, Jingjing Shen, Matthew Johnson, and Julien Valentin. VolTeMorph: Real-time, controllable and generalisable animation of volumetric representations. *arXiv:2208.00949*, 2022.

[14] Georgios Georgakis, Ren Li, Srikrishna Karanam, Terrence Chen, Jana Košecká, and Ziyan Wu. Hierarchical kinematic human mesh recovery. In *ECCV*, 2020.

[15] Benjamin Graham, Martin Engelcke, and Laurens Van Der Maaten. 3D semantic segmentation with submanifold sparse convolutional networks. In *CVPR*, 2018.

[16] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. Neural head avatars from monocular RGB videos. In *CVPR*, 2022.

[17] Yuantao Gu, Jian Jin, and Shunliang Mei. l_0 norm constraint lms algorithm for sparse system identification. *IEEE Signal Processing Letters*, 16(9):774–777, 2009.

[18] Timothy Jeruzalski, Boyang Deng, Mohammad Norouzi, John P Lewis, Geoffrey Hinton, and Andrea Tagliasacchi. NASA: Neural articulated shape approximation. In *ECCV*, 2020.

[19] Timothy Jeruzalski, David IW Levin, Alec Jacobson, Paul Lalonde, Mohammad Norouzi, and Andrea Tagliasacchi. NiLBS: Neural inverse linear blend skinning. *arXiv:2004.05980*, 2020.

[20] Boyi Jiang, Yang Hong, Hujun Bao, and Juyong Zhang. SelfRecon: Self reconstruction your digital avatar from monocular video. In *CVPR*, 2022.

[21] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. NeuMan: Neural human radiance field from a single video. *arXiv:2203.12575*, 2022.

[22] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. *arXiv:2109.11399*, 2021.

[23] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *3DV*, 2020.

[24] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM Trans. Graph.*, 36(6), 2017.

[25] 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 TOG*, 2022.

[26] 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 TOG*, 2021.

[27] Shichen Liu, Tianye Li, Weikai Chen, and Hao Li. Soft rasterizer: A differentiable renderer for image-based 3d reasoning. In *ICCV*, 2019.

[28] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM TOG*, 2021.

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. SMPL: A skinned multi-person linear model. *ACM TOG*, 2015.

[30] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987.

[31] Marko Mihajlovic, Shunsuke Saito, Aayush Bansal, Michael Zollhoefer, and Siyu Tang. COAP: Compositional articulated occupancy of people. In *CVPR*, 2022.

[32] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *CVPR*, 2021.

[33] 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.

[34] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, 2020.

[35] 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 *ECCV*, 2020.

[36] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, 2021.

[37] Xingang Pan, Ayush Tewari, Lingjie Liu, and Christian Theobalt. GAN2X: Non-lambertian inverse rendering of image gans. In *3DV*, 2022.

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, pages 5865–5874, 2021.

[39] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed AA Osman, Dimitrios Tzionas, and Michael J Black. Expressive body capture: 3D hands, face, and body from a single image. In *CVPR*, pages 10975–10985, 2019.

[40] 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 *ICCV*, 2021.

[41] 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.

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

[43] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep learning on point sets for 3D classification and segmentation. In *CVPR*, 2017.

[44] 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 *ECCV*, 2020.

[45] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG*, 2017.

[46] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. NeRF for outdoor scene relighting. In *ECCV*, 2022.

[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019.

[48] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, 2021.

[49] Ruizhi Shao, Hongwen Zhang, He Zhang, Mingjia Chen, Yan-Pei Cao, Tao Yu, and Yebin Liu. DoubleField: Bridging the neural surface and radiance fields for high-fidelity human reconstruction and rendering. In *CVPR*, 2022.

[50] Olga Sorkine and Marc Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, 2007.

[51] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. NeRV: Neural reflectance and visibility fields for relighting and view synthesis. In *CVPR*, 2021.

[52] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-NeRF: Articulated neural radiance fields for learning human shape, appearance, and pose. In *NeurIPS*, 2021.

[53] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. NeLF: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021.

[54] Gusi Te, Xiu Li, Xiao Li, Jinglu Wang, Wei Hu, and Yan Lu. Neural capture of animatable 3d human from monocular video. In *ECCV*, 2022.

[55] Ten24. https://ten24.info.

[56] Ayush Tewari, Florian Bernard, Pablo Garrido, Gaurav Bharaj, Mohamed Elgharib, Hans-Peter Seidel, Patrick Pérez, Michael Zollhofer, and Christian Theobalt. FML: Face model learning from videos. In *CVPR*, 2019.

[57] Ayush Tewari, Hans-Peter Seidel, Mohamed Elgharib, Christian Theobalt, et al. Learning complete 3D morphable face models from images and videos. In *CVPR*, 2021.

[58] Ayush Tewari, Michael Zollhofer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Christian Theobalt. MoFA: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *ICCV*, 2017.

[59] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 2021.

[60] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[61] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *CVPR*, 2021.

[62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004.

[63] 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 *CVPR*, 2022.

[64] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. *NeurIPS*, 2021.

[65] Tianhan Xu, Yasuhiro Fujita, and Eiichi Matsumoto. Surface-aligned neural radiance fields for controllable 3D human synthesis. In *CVPR*, 2022.

[66] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *NeurIPS*, 2021.

[67] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *NeurIPS*, 2020.

[68] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3DMM: Deep implicit 3D morphable model of human heads. In *CVPR*, 2021.

[69] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. NeRF-Editing: geometry editing of neural radiance fields. In *CVPR*, 2022.

[70] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM TOG*, 2021.

[71] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[72] 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 TOG*, 2021.

[73] Fuqiang Zhao, Wei Yang, Jiakai Zhang, Pei Lin, Yingliang Zhang, Jingyi Yu, and Lan Xu. HumanNeRF: Efficiently generated human radiance field from sparse inputs. In *CVPR*, 2022.

[74] Mingwu Zheng, Hongyu Yang, Di Huang, and Liming Chen. Imface: A nonlinear 3D morphable face model with implicit neural representations. In *CVPR*, 2022.

[75] 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 *CVPR*, 2022.

[76] Zerong Zheng, Han Huang, Tao Yu, Hongwen Zhang, Yandong Guo, and Yebin Liu. Structured local radiance fields for human avatar modeling. In *CVPR*, 2022.

[77] Yongjie Zhu, Chen Li, Si Li, Boxin Shi, and Yu-Wing Tai. Hybrid face reflectance, illumination, and shape from a single image. *IEEE TPAMI*, pages 5002–5015, 2021.

A. MANO-HD

MANO. MANO can be driven with parameters $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and $\boldsymbol{\theta} \in \mathbb{R}^{B\times 3}$ ($B = 16$ indicates the number of per-bone parts), where $\boldsymbol{\beta}$ is the coefficients of a shape PCA bases while $\boldsymbol{\theta}$ represents joint rotations in axis-angle form. Mean template mesh is deformed to match different shapes:

$$\begin{aligned} \tilde{\mathbf{V}} &= \bar{\mathbf{V}} + \mathcal{B}_s(\boldsymbol{\beta}) + \mathcal{B}_p(\boldsymbol{\theta}) \\ \mathbf{J} &= \mathcal{J}(\bar{\mathbf{V}} + \mathcal{B}_s(\boldsymbol{\beta})), \end{aligned} \tag{5}$$

where $\bar{\mathbf{V}}, \mathcal{B}_s, \mathcal{B}_p$ are template vertices and shape/pose blendshapes. Canonical joint locations $\mathbf{J} \in \mathbb{R}^{B\times 3}$ are given with the regressor $\mathcal{J}$.

Then, bone transformation matrix $\mathbf{G}_b \in \mathbb{R}^{4\times 4}$ is computed along the kinematic chain $\mathcal{K}$ with the Rodriguez formula $\mathcal{R}$:

$$\mathbf{G}_b(\boldsymbol{\theta}, \mathbf{J}) = \prod_{j\in\mathcal{K}(b)} \left[\begin{array}{c|c} \mathcal{R}(\boldsymbol{\theta}_j) & \mathbf{J}_j \\ \hline \mathbf{0} & 1 \end{array}\right] \tag{6}$$

Finally, linear blend skinning is used to pose vertices with skinning weights $\mathbf{W} \in \mathbb{R}^{V\times B}$ (V denotes the number of vertices) as follows,

$$\mathbf{V}_i = \sum_{b=1}^{B} \mathbf{W}_{b,i} \mathbf{G}_b(\boldsymbol{\theta}, \mathbf{J}) G_k(\mathbf{0}, \mathbf{J})^{-1} \tilde{\mathbf{V}}_i. \tag{7}$$

Optimization of MANO-HD. Following [16], we subdivide the MANO template mesh to obtain a high-resolution version with 12,337 vertices and 24,608 faces. Nevertheless, articulated dynamic mesh subdivision is a non-trivial task, and mesh skinning operation is likely to introduce artifacts to deformed mesh. Thus, we optimize upsampled skinning weights $\mathbf{W}^{HD} \in \mathbb{R}^{V^{HD}\times B}$ to eliminate dynamic artifacts under various hand poses using energy functions as follows,

$$\begin{aligned} \mathcal{L}_{l_0} &= \sum_{i=1}^{B\cdot V^{HD}} (1 - e^{-\eta \mathbf{W}_i^{HD}}) \\ \mathcal{L}_{lap} &= \frac{1}{V^{HD}} \sum_{i=1}^{V^{HD}} \sum_{j\in\mathbb{N}(i)} \frac{1}{\omega} ||\mathbf{V}_i^{HD} - \mathbf{V}_j^{HD}||_2 \\ \mathcal{L}_{surf} &= \mathrm{Cham}(\mathbf{V}^{HD}, \mathbf{F}), \end{aligned} \tag{8}$$

where $\mathcal{L}_{l_0}$ is approximated l_0 norm constraint [17] to produce sparse skinning weights. $\mathcal{L}_{lap}$ is the Laplacian term for mesh smoothness, where $\mathbb{N}(\cdot)$ represents vertex neighborhood and ω is the normalization factor. The function $\mathrm{Cham}(\cdot,\cdot)$ computes the chamfer distance between MANO-HD mesh vertices $\mathbf{V}^{HD}$ and the MANO mesh faces $\mathbf{F}$. The overall energy function is given as $\mathcal{L}_{HD} = \lambda_{l_0}\mathcal{L}_{l_0} + \lambda_{lap}\mathcal{L}_{lap} + \lambda_{surf}\mathcal{L}_{surf}$ with balance term λ.

Implement Details of MANO-HD For $\mathbf{W}^{HD}$ optimization, we adopt the original MANO dataset [45] with 1,554 pose parameters for training and evaluation. We randomly compose and interpolate finger-level rotations for data augmentation. The training process lasts 3,000 steps with a batch size of 1,024. The learning rate begins at 10^{-5} and decreases with exponential decay. We use $\mathcal{L}_{HD}$ as the objective and adjust hyperparameters to balance multiple energy terms as $\eta = 100, \lambda_{l_0} = 0.01, \lambda_{lap} = 1, \lambda_{surf} = 10$.

Method	*Lap.*	*Cham.*	l_0 norm (%)
MANO	23.31	-	16.29
MANO-HD w/o $\mathbf{W}^{HD}$ opt.	1.923	7.014	17.97
MANO-HD w/o $\mathcal{L}_{l_0}$	1.576	7.170	44.15
MANO-HD	1.753	7.039	16.89

Table 5. Effects of the $\mathbf{W}^{HD}$ optimization (opt.).

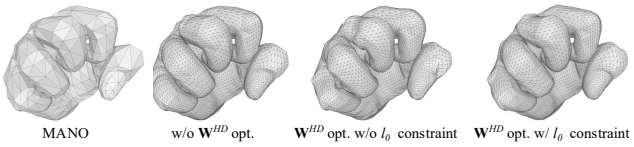

Figure 10. Hand mesh comparison under a large deformation.

Name	Depth	Width	Input size	Output size
$\mathcal{M}_{shape}$	4	128	84	3
MLP in $\mathcal{Q}_{pair}$	4	128	64	1
$\mathcal{M}_{albedo}$	4	256	128	3
$\mathcal{M}_{illum}$	4	256	85	1

Table 6. The details of MLPs.

Effects of MANO-HD We use Laplacian smoothing *Lap.* and chamfer distance *Cham.* to reflect the smoothness and accuracy of MANO-HD, whose definition is the same as $\mathcal{L}_{lap}$ and $\mathcal{L}_{surf}$ in Eq. 8. Both *Lap.* and *Cham.* are presented in 10^{-4}m. As the MANO surface is not smooth enough, *Lap.* and *Cham.* cannot be simultaneously improved. Besides, we introduce l_0 norm as the metric, which is defined as the proportion of non-zero elements in the skinning weights.

Referring to Tab. 5, mesh subdivision can improve *Lap.*, yet incurs artifacts during skinning as shown in the part-connection regions in Fig. 10. Moreover, $\mathbf{W}^{HD}$ optimization without $\mathcal{L}_0$ cannot counteract the issue despite inducing lower *Lap.* value. The reason behind this is the poor sparsity of $\mathbf{W}^{HD}$. After $\mathcal{L}_{l_0}$-based optimization, the l_0 norm of $\mathbf{W}^{HD}$ is on par with that of MANO, and the skinning performance is improved (see Fig. 10).

B. Network Structures

Kinematic Tree and Local Pair. Following the definition of MANO [45], Fig. 11 shows the bone indices and connections for the hand. In addition, we demonstrate our defined local pairs.

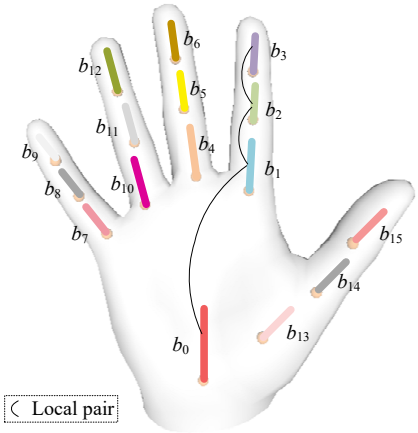

Figure 11. Kinematic tree and part indices of the hand. We also show local pairs along the kinematic chain of forefinger.

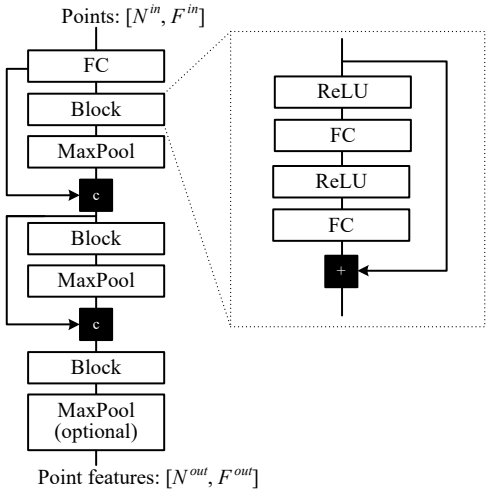

Figure 12. PointNet structure. "c,+" indicate concatenation and element-wise sum; "FC" denotes fully connected layer.

PointNet Structure. The PointNet used in part-space encoder $\mathcal{Q}_{part}$ and local pair decoder $\mathcal{Q}_{pair}$ is shown in Fig 12, where N, F denote point amount and feature size. For $\mathcal{Q}_{part}$, $N^{in} = 256, F^{in} = 6, N^{out} = 1, F^{out} = 64$. For $\mathcal{Q}_{pair}$, we get rid of the last MaxPool, leading to $N^{in} = 2, F^{in} = 64, N^{out} = 2, F^{out} = 64$.

MLP Details. Referring to Table 6, we list MLP details used in this paper.

Local-Pair Decoder. The local pair decoder $\mathcal{Q}_{pair}$ contains a PointNet and an MLP, whose detailed structures are introduced before. Furthermore, we illustrate how the $\mathcal{Q}_{pair}$ processes the geometry encodings along the kinematic chain of forefinger, as shown in Fig. 13(a). First, locally paired encodings are treated as two *points* and fused

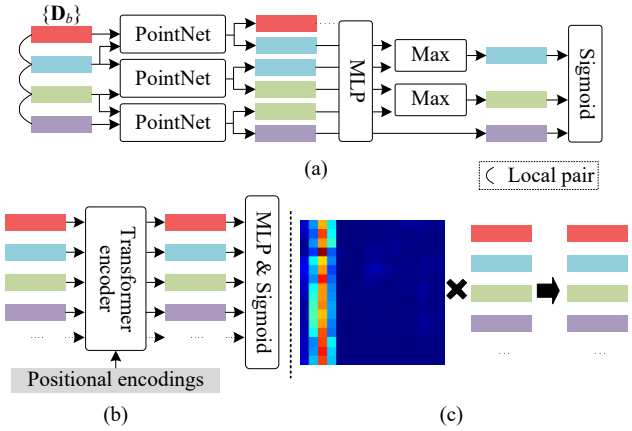

Figure 13. (a) Our proposed local-pair decoder. Multiple PointNets/MLPs share weights. (b) Transformer-based decoder for comparison. (c) Attention-based fusion of part-level encodings in Transformer for bone part b_2. Consistent with Fig.11, colors distinguish part-level geometry encodings. For visual conciseness, we only show the process of forefinger.

Capture	Training set	Validation set
test/Capture0	11,757	194
test/Capture1	18,474	232
val/Capture0	18,340	184

Table 7. Data amount for training and validation sets.

by the PointNet. Without across-point maxpooling, the shape of PointNet output remains the same as its input. Then, the MLP maps PointNet outputs to the occupancy domain. As a result, a bone part could have multiple occupancy values from multiple local-pair predictions, which are fused by a maximum operator. Because of the feature fusion by the PointNet, part boundaries are blurred and extend to the connection direction. Hence, the maximum operator is used to produce a union of boundary-extended part geometries. Finally, the sigmoid function is employed for occupancy normalization.

Transformer-Based Decoder. To validate the local-pair design, we develop a learning-based method with Transformer. As shown in Fig. 13(b), all part-level encodings are fed to the Transformer encoder without any inductive prior. Through 4 self-attention blocks, the Transformer can perform adaptive feature fusion. Referring to Fig. 13(c), the attention map determines the way to select important features, where the effect is consistent with our local-pair decoder. That is, bone parts b_1, b_2, b_3 are connected in the attention map for evolving the encoding of part b_2.

C. Training Details

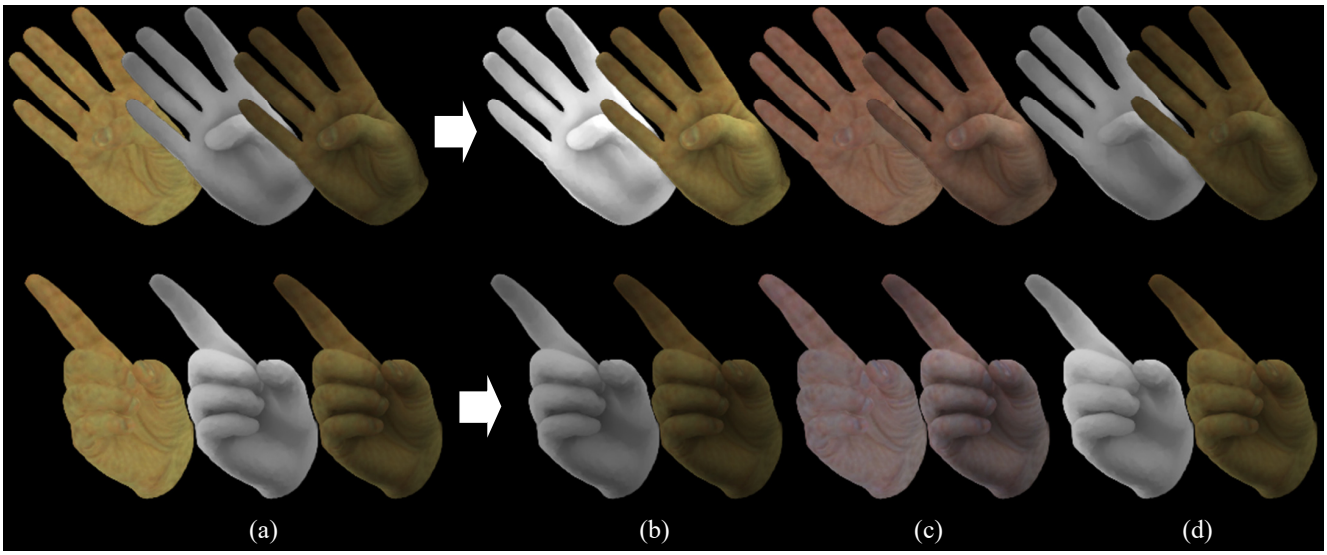

Figure 14. Three groups of texture editing. (a) Reconstruction results. From left to right: albedo, illumination, and shaded image. (b) Lighting editing. (c) Albedo editing. (d) Shadow editing.

Video Data. For quantitative evaluation, we select three sequences from the InterHand2.6M dataset [35]. Data amount is shown in Table 7, where training data are from the *ROM04_RT_Occlusion* sequence and validation data are from the *ROM03_RT_No_Occlusion* sequence. Because video frames are highly redundant, validation data are selected by fixed skip steps, and we adjust the steps to assure various hand poses and global rotations can be covered.

For each frame, we crop the hand region with annotated detection boxes as the ground truth. First, the box is regulated as a square box with 1.3 times expansion. Then, the hand region is cropped and resized to 256×256 resolution.

Training Settings. For PairOF pre-training, the learning rate begins at 5×10^{-4} and decreases with exponential decay. The training process has 270K steps with a batch size of 32.

For end-to-end training, the learning rate begins at 5×10^{-4} and decreases with exponential decay. Due to the pre-training, the learning rate of $\mathcal{Q}_{pair}$ is set to be 10 times smaller. Following [63], we use a patch strategy for training with a patch size of 32×32. The training process has 50K steps with a batch size of 16.

D. Texture Editing

Referring to Fig 14, our HandAvatar supports hand texture editing. Firstly, we change the illumination field by a 1.5- or 0.8-times multiplication, as shown in Fig 14(b). Then, we shift the mean RGB value of the albedo field, leading to the results of Fig 14(c). Besides, we edit shadow in Fig 14(d). In the top row, the texture is induced by letting thumb-related directed soft occupancy values equal 0. As a result, the self-occluded shadow patterns on the palm become weakened. In the bottom row, we remove the shadow patterns between the middle and ring fingers by setting the directed soft occupancy values as 0 for bone parts b_5 and b_{11}. The shadow editing results also validate the effect of directed soft occupancy.

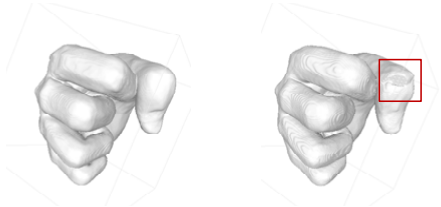

Figure 15. PairOF (left) *vs*. Transformer-based method (right)

E. Discussion

The Disentanglement of Albedo and Illumination. The disentanglement of albedo and illumination does not require extra regularization. Albedo is known to be independent of hand pose, while illumination depends on hand pose. In SelF, the input of the albedo field is unrelated to the hand pose, while pose-relevant elements are fed into the illumination field. Therefore, with various hand poses as training data, the optimization process would ensure that the illumination is free from the albedo field. In addition, the illumination field outputs a scalar, which cannot model RGB-based (3-channel) albedo.

PairOF *vs*. Transformer-Based Method The motivation of PairOF is to fuse part-level geometry encodings to elim-

inate the shape inconsistency in the area of part connections. Despite similar numerical results in Tab. 2 of the main text, our PairOF is more effective in feature fusion than the Transformer-based method. That is, hand bone connections are unchangeable, and PairOF can use this prior knowledge, leading to a more effective feature fusion than self-attention. As a result, self-attention cannot visually achieve our motivation as shown in Fig. 15. Also, we are more efficient in terms of learning, *i.e.*, the convergence of the Transformer-based method is slower and more data-hungry.

F. Limitations and Future Works

Besides the limitation discussed in the main text, HandAvatar can be further improved from the following perspectives in the future. First, full lighting editing is worthy of future research. For example, despite the lighting editing as shown in Fig 14(b), it is hard to edit or add a point light in the illumination field. Second, the representation of specular effects in hand appearance is another interesting topic. To achieve this goal, hand surface properties with BRDF should be explored. Third, the hand geometry demonstrated in Fig. 5 in the main text is the result of PairOF pre-training. After end-to-end training with texture losses, the PairOF could produce a non-smooth surface with geometry wrinkles. This is caused by the hand-pose annotation error of InterHand2.6M. As a result, the PairOF could produce fragile hand geometry to compensate for this error.