PHRIT: Parametric Hand Representation with Implicit Template

Zhisheng Huang[1†] Yujin Chen[2†] Di Kang[3] Jinlu Zhang[1] Zhigang Tu[1*]
[1]Wuhan University [2]Technical University of Munich [3]Tencent AI Lab

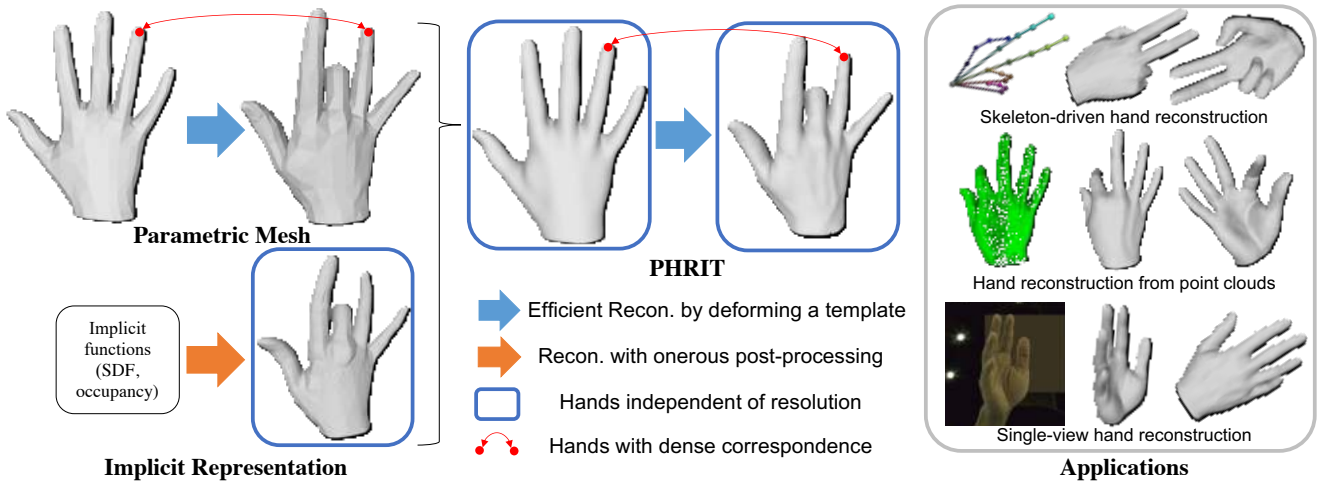

Figure 1: We propose a novel hand representation PHRIT, combing advantages of previous methods on hand geometry modeling (namely parametric mesh and implicit representation). PHRIT can serve as a learned fully differentiable layer driven by input hand skeleton and shape latent code, making it easily applicable to downstream tasks.

Abstract

We propose PHRIT, a novel approach for parametric hand mesh modeling with an implicit template that combines the advantages of both parametric meshes and implicit representations. Our method represents deformable hand shapes using signed distance fields (SDFs) with part-based shape priors, utilizing a deformation field to execute the deformation. The model offers efficient high-fidelity hand reconstruction by deforming the canonical template at infinite resolution. Additionally, it is fully differentiable and can be easily used in hand modeling since it can be driven by the skeleton and shape latent codes. We evaluate PHRIT on multiple downstream tasks, including skeleton-driven hand reconstruction, shapes from point clouds, and single-view 3D reconstruction, demonstrating that our approach achieves realistic and immersive hand modeling with state-of-the-art performance.

1. Introduction

The human hand plays a vital role in communication and interaction, making high-fidelity hand modeling crucial for immersive applications, especially in the era of digital twins and metaverses. High-fidelity hand modeling can facilitate immersive applications such as virtual meetings and video games. However, to drive these real-time applications, it's essential to achieve realistic reconstruction results, ensure stable cross-user generalization, and optimize the reconstruction process for efficiency.

Previous research on hand geometry modeling can be divided into two broad categories: parametric meshes and implicit representations. Parametric meshes rely on a predefined mesh template that is deformed to match posed hands [32, 44, 61]. While this approach is efficient and provides useful dense correspondence between the reconstructed hand and the canonical template [21, 60], it requires careful supervision to learn the deformation of each vertex. This can be difficult and expensive, as noted in [32]. This challenge has led previous work to either sacrifice resolution [61] or resort to leveraging weak supervisions [44], which may limit the model's generalization to personalized settings. In contrast, recent implicit representations [13, 26] take a different approach to hand geometry modeling by focusing on the continuous representation of static shapes. By learning implicit functions such as signed distance and occupancy fields, they can represent high-fidelity, resolution-

[†]Equal contributions.
[*]Corresponding author.

Method	*Diff.*	*Effi.*	*Corres.*	*Conti.*
Parametric Hand Meshes	✔	✔	✔	✘
Implicit Hand Representation	✔	✘	✘	✔
PHRIT (Ours)	✔	✔	✔	✔

Table 1: A comparison of PHRIT with parametric hand meshes and implicit hand representation in terms of some key properties. *Diff.*: whether the reconstruction is differentiable. *Effi.*: whether reconstruction is efficient without onerous pose-processing such as Marching Cubes [36]. *Corres.*: whether dense correspondences are maintained during reconstruction. *Conti.*: whether this representation is continuous in the output space.

independent shapes. However, the method requires time-consuming post-processing to obtain reconstructions and lacks dense correspondence between the reconstructions. Overall, both approaches have their strengths and weaknesses, and neither fully meets the requirements for high fidelity, generalization ability, and efficiency.

To combine the advantages of both paradigms, we propose PHRIT (as shown in Fig. 1), a parametric hand model with an implicit template that can generate high-fidelity hand reconstructions for various poses and identities (i.e., generalization ability) with favorable properties such as differentiation, correspondence, efficiency, and continuity. As summarized in Table 1, our proposed method combines some key strengths of existing paradigms. We argue that our method achieves continuity to parameterized hand representations by introducing an implicit representation (i.e., using SDF to represent the hand) while maintaining efficiency during inference and maintaining the dense correspondences through our novel deformation field to learn per-vertex deformation implicitly.

Specifically, PHRIT learns to deform a canonical hand with theoretically unlimited resolution based on implicit representation. To achieve efficiency and continuity (i.e., high fidelity), we represent the canonical hand with an SDF using a neural network. This means that the canonical hand mesh only needs to be extracted once for all inferences. Along with the learning of implicit canonical hand, MLPs are utilized to retrieve per-vertex deformation of the canonical hand conditioned on both poses and shapes. To learn such deformation, we use real-world 3D hand scans [61] and develop a novel deformation field that bridges the SDF of deformed and canonical hand space to build dense correspondences (i.e., per-vertex deformation) implicitly. To improve the generalization towards unseen poses and identities, we adopt a part-based design [15] on deformation learning, but rather eliminate the requirement of ground-truth bone transformation by deriving local coordinate systems directly upon the hand skeleton based on [26]. Additionally, we adopt the locally pose-aware design similar to [6] on deformation learning to boost generalization. Moreover, we propose a skip-connection structure, which is experimentally proven to be more effective in capturing the non-rigidity of the deformation.

In summary, our contributions are:

- We introduce a neural hand model combining parametric meshes and implicit representations to efficiently produce high-fidelity hand reconstruction with full differentiability to the skeleton and shape latent codes.
- We develop a novel deformation field to learn pointwise deformation and implicitly establish a dense correspondence between the canonical hand and its deformed shape.
- Experiments demonstrate our method achieves state-of-the-art performance in multiple hand modeling tasks, including reconstruction from skeleton, point clouds, and images, resulting in immersive results.

2. Related Work

Parametric hand mesh. Parametric meshes are widely used to model deformable shapes such as faces [2, 31], bodies [25, 35, 50, 56], hands [7, 32, 44, 71], feet [37], and animals [79]. While MANO [61] is the most commonly used hand model [3, 7–9, 14, 20, 22, 45, 53, 66, 73–74], it has limited resolution and a linear representation of non-rigid deformation, leading to unrealistic hand reconstructions. To overcome these limitations, previous works have explored graph networks [12, 17, 30, 70], transformers [11, 33], and UV maps [5]. DHM [44] proposes to model high-fidelity hand meshes by deforming a professionally designed template and predicting pose- and shape-dependent correctives through neural networks, but with a personalized setup. Inspired by MANO and DHM, our generalizable canonical hand template mesh addresses resolution limitations by using an implicit representation (i.e., SDF) and allows for non-rigid deformation of vertices.

Implicit hand representation. Implicit representations, like SDF [54], occupancy [40], and neural radiance fields [42], are widely used in 3D modeling because they are continuous and differentiable [10, 23, 26–27, 38, 41–42, 47, 49, 58–59, 62, 67–68, 77]. Recently, implicit representations have been applied to articulated shapes in [1, 4, 16, 34, 46, 48, 52, 57, 69]. These works use implicit skinning fields [63] or piecewise deformable models [15] to represent articulated objects. Our work draws inspiration from these previous works and proposes a novel solution for modeling high-fidelity articulated hands. In contrast to Mehta *et al.* [39], who maintain differentiability with Marching Cubes to enable supervision for explicit representation and update the geometry of the implicit surface, we directly impose supervision on the implicit representation. Unlike previous methods using implicit skinning functions to blend a fixed shape [6, 24, 63], our model is conditioned on pose and shape variation. Compared to methods [51–52, 76] using implicit functions to encode shape variation and pose-dependent deformation separately, our model

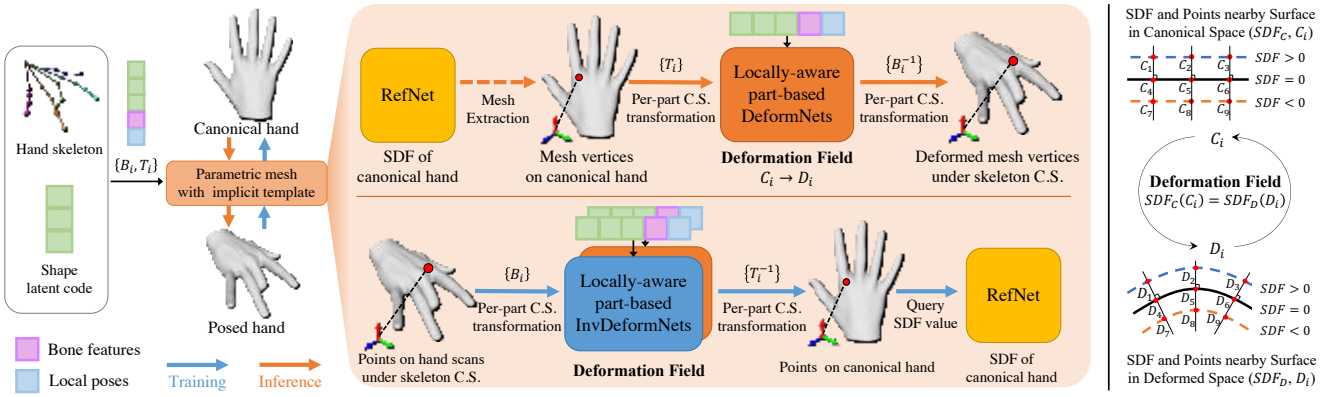

Figure 2: Overiew of PHRIT (Section 3.1). During training, PHRIT learns a one-to-one mapping of query points between deformed space and canonical space based on the proposed deformation field (Section 3.2), given the hand skeleton and shape latent code. During inference (Section 3.3), PHRIT deforms the high-resolution canonical hand mesh extracted from the implicit representation to obtain a differentiable, high-fidelity hand reconstruction. We use a dashed line to denote that mesh extraction is only necessary once.

is more efficient and shares correspondences across identities. Furthermore, we address the multiple correspondence problem in SNARF [6] by developing a deformation field and introducing a deformation skip-connection structure to learn it.

Skeleton-driven articulated hand. Although previous studies [12, 17] have demonstrated the feasibility of skeleton-driven shape reconstruction, recent research [13, 15, 48] on articulated shape modeling still relies on ground truth bone transformations. This is mainly because bone transformations provided by iterative optimization procedures like inverse kinematics prevent gradient flow and introduce ambiguity in the twist angle of bones. To address this issue, [26] exploits biomechanical constraints [65] to define the local coordinate systems of hand skeletons and derive their bone transformation accordingly. However, their proposed approach involves learning deformations in canonical space and requires additional canonicalization of the hand skeleton. In contrast, we directly learn the deformations in each hand's local coordinate system via more readily available transformation matrices.

3. Approach

We present a high-fidelity hand model PHRIT (Section 3.1). As shown in Fig. 2, PHRIT consists of RefNet and DeformNets, which are jointly learned through the proposed deformation field (Section 3.2) with an inverse counterpart of DeformNets, denoted as InvDeformNets. The resulting PHRIT allows for efficient and high-fidelity hand reconstruction that is fully differentiable with respect to both the input hand skeleton and shape latent code (as explained in Section 3.3).

3.1. Parametric Mesh with Implicit Template

Our PHRIT comprises an implicit template (RefNet) and an associated deformation function (DeformNets).

3.1.1 RefNet

RefNet f encodes a canonical hand mesh M, which serves as the basis for all hand reconstructions generated by our model. We define M to be compatible with the MANO [61] template mesh $\bar{M}$ (of zero pose and mean shape). That is, M and $\bar{M}$ have the same skeleton $\bar{K}$ and shape, but M has an infinite number of mesh vertices. RefNet f maps a query point x to its signed distance d to the hand surface: $f(x) = d \in \mathbb{R}$.

3.1.2 DeformNets

Our deformation function utilizes a part-based design, where we decompose the human hand into 16 rigid parts $\{P_i \mid 0 \leq i \leq 15\}$ based on the MANO [61] skinning weights. For each part P_i, we employ an independent DeformNet g_i conditioned on a local pose θ_i and a shape code β_i, transforming the query points x_i on P_i of canonical hand to corresponding points y_i on P_i of deformed hand: $g_i(x_i, \theta_i, \beta_i) = y_i \in \mathbb{R}^3$.

To define the query points x_i in a local coordinate system LCS_i of the part P_i, we follow NASA's approach [15]. However, we derive the LCS_i from the hand skeleton rather than using ground truth bone transformations.

Local coordinate system LCS_i. Given a hand skeleton consisting of 21 3D keypoints and 20 bones $\{b_j \mid 1 \leq j \leq 20\}$, we first define a transformation B_j for each bone b_j and then define the LCS_i for each of the 16 rigid hand parts P_i. For B_j of bone b_j, the translation is set to the middle point of b_j, and the rotation (orientation) is set following HALO [26]. Details can be found in the Appendix. Based on B_j, we construct the local coordinate system LCS_i for each part P_i as follows: (1) For palm part P_1, LCS_1 is defined by B_9 (i.e., the transformation of the middle palmar bone), where the origin is set to the translation of B_9 and the axes are set to the orientation of B_9. (2) For the remaining finger parts P_i, LCS_i is defined by corresponding finger bone transformation B_j, in the same way

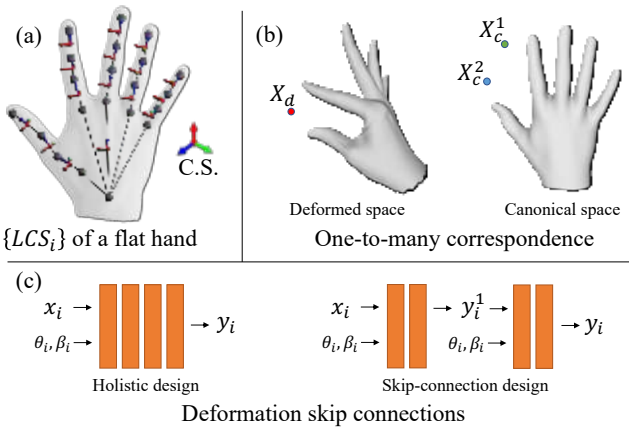

Figure 3: (a) Local coordinate systems derived from hand skeleton. (b) Multi-correspondence (one-to-many) between X_d and X_c (Section 3.2). (c) Deformation skip-connection architecture with comparison to the holistic design.

as LCS_1 is determined by B_9. An illustration of LCS_i is provided in Fig. 3 (a).

Local pose θ_i. The local pose θ_i of P_i is defined differently for the palm part P_1 and the remaining finger parts P_i. (1) For palm part P_1, the $\theta_1 \in \mathbb{R}^{17}$ consists of a palmar bone configuration $\theta_1^p \in \mathbb{R}^7$ and palmar joint configuration $\theta_1^f \in \mathbb{R}^{10}$. Following HALO [26], θ_1^p is decoupled into a finger spreading $\theta_s \in \mathbb{R}^4$ and a palm arching $\theta_a \in \mathbb{R}^3$. θ_s is defined by angles between adjacent palmar bones, and θ_a is defined by angles between the two planes spanned by three adjacent palmar bones. The θ_1^f are joint angles (abduction and flexion) of each five palmar joints. (2) For the remaining finger parts P_i, θ_i are joint angles from the corresponding finger bones. Each finger bone has two joints with two joint angles, namely abduction and flexion, except for the tips of level-2 finger bones, which have no joint angles.

Shape code β_i. Similar to SNARF [6], we split shape space into surface properties and bone features. Formally, the shape code $\beta_i = \gamma \oplus F_i$, where i indexes hand parts and $\oplus$ denotes feature concatenation. $\gamma \in \mathbb{R}^{128}$ is a trainable shape latent code shared across all hand parts, which is optimized per hand ID (subscript omitted), and F_i is bone features from the hand part P_i. Following HALO [26], F_i is a concatenation of global bone features $F_g \in \mathbb{R}^{16}$ obtained from a global bone encoder and local bone lengths L_i of P_i. Except for $L_1 \in \mathbb{R}^5$, which is set to the lengths of five palmar bones, the remaining $L_i \in \mathbb{R}$ is the length of the corresponding finger bone.

Deformation skip connections. We propose a skip connection design for learning our DeformNets g_i instead of the holistic design depicted in Fig.3. The g_i is decomposed as $g_i = g_i^0 \circ g_i^1 \cdots \circ g_i^N$, where g_i^{n+1} takes the deformation results y_i^n from g_i^n to predict further deformation: $g_i^{n+1}(y_i^n, \theta_i, \beta_i) = y_i^{n+1} \in \mathbb{R}^3$. N is the number of skip connections and θ_i and β_i are the local pose and shape code as before. We empirically find that the skip-connection design, with the same total number of layers, better captures the details in hand reconstruction. (See Section 4.4 for an ablation study on this design choice.)

3.2. Learning with Deformation Field

To learn the RefNet and DeformNets introduced in the previous section, we propose a novel deformation field and derive training objectives based on it.

Deformation field ϕ. The proposed deformation field is a one-to-one mapping function that transforms a 3D point X_c in the canonical hand space to X_d in deformed hand space: $\phi : X_c \mapsto X_d$. To exclude potential ambiguity arising from multi-correspondence between X_d and X_c, as pointed by SNARF [6] (Refer to Fig. 3 (b)), we define ϕ as folows: (1) ϕ has a constrained domain nearby the hand surface. (2) For X_c on the hand surface, ϕ is based on the correspondence between canonical hand surface and the deformed hand surface. (3) For X_d off the hand surface but within the domain, ϕ is further based on the signed distance to the hand surface. A 2D illustration is provided on the right side of Fig. 2. The resulting ϕ satisfies the following equation:

$$SDF_d(X_d) = SDF_d(\phi(X_c)) = SDF_c(\phi^{-1}(X_d)) \quad (1)$$

Here, SDF_c and SDF_d are signed distance fields of the hand in canonical space and deformed space, respectively. A more concrete mathematical definition of ϕ and its constrained domain can be found in Appendix.

Training objectives. Our training objectives are based on the deformation field ϕ. Specifically, our proposed DeformNets g_i follow ϕ as an invertible mapping function, and we introduce an inverse counterpart of DeformNets called InvDeformNets g_i^{-1}. The g_i and g_i^{-1}, along with the RefNet f, are jointly learned using the following loss function:

$$E = \sum_{s=1}^{B} \sum_{i=1}^{P} \sum_{x_i^s \in X_i^s} (w_S E_{SDF} + w_n E_{norm} + w_c E_{cycle} + w_m E_{mano} + w_r E_{regu}) + w_O E_O + w_I E_{IGR} \quad (2)$$

Here, B is the number of scans in the batch, $P = 16$ is the number of hand parts, and $\{X_i^s \mid x_i^s \in \mathbb{R}^3\}$ are hand surface points on the part P_i in the s-th scan. The individual loss terms are defined as follows:

$$E_{SDF} = \left\| f(T_i^{-1} \cdot g_i^{-1}(x_i^s, \theta_i^s, \beta_i^s)) \right\| \quad (3)$$

$$E_{norm} = \left\| \nabla_{x_i^s} f(T_i^{-1} \cdot g_i^{-1}(x_i^s, \theta_i^s, \beta_i^s)) - N(x_i^s) \right\| \quad (4)$$

$$E_{cycle} = \left\| g_i(g_i^{-1}(x_i^s, \theta_i^s, \beta_i^s), \theta_i^s, \beta_i^s) - x_i^s \right\| \quad (5)$$

$$E_{mano} = \left\| g_i^{-1}(\bar{x}_i^s, \theta_i^s, \beta_i^s) - \bar{y}_i^s \right\| \tag{6}$$

$$E_{regu} = \|\gamma^s\| \tag{7}$$

$$E_O = \mathbb{E}_{x \in \Omega \setminus \Omega_0}(\exp(-\alpha \cdot |f(x)|)), \quad \alpha \gg 1 \tag{8}$$

$$E_{IGR} = \mathbb{E}_{x \in \Omega}(\|\nabla f(x)\| - 1)^2 \tag{9}$$

E_{SDF} ensures the points x_i^s on deformed hand surface have zero signed distance. We query the SDF of x_i^s through the deformation field ϕ. Specifically, we first deform x_i^s using g_i^{-1}, then transform it back from the LCS_i of canonical hand using T_i^{-1}, and finally query the signed distance using f.

E_{norm} further constrains the normal of the points x_i^s on deformed hand surface based on ϕ, which satisfies Eq.1. This constraint is based on the following two observations: (1) The derivative of SDF is the corresponding gradient field, and (2) The SDF of deformed hand space is associated with the SDF of canonical hand space by Eq.1. We denote the surface normal of x_i^s with $N(x_i^s)$.

E_{cycle} facilitates that g_i and g_i^{-1} are reciprocal functions, ensuring a one-to-one mapping between the canonical hand and deformed hand surfaces.

E_{mano} enforces f to learn a canonical hand compatible with the MANO template. To achieve this, we use MANO annotations to provide sparse correspondence supervision. We denote MANO vertices of the canonical hand and deformed hand under LCS_i as $\bar{x}_i^s$ and $\bar{y}_i^s$ respectively.

E_{regu} regularizes the trainable shape latent code γ^s.

E_O and E_{IGR} are utilized to regularize f to learn a standard SDF. E_{IGR} is the Eikonal regularization term proposed in [19], while E_O ensures off-surface points do not have zero signed distances, as previously introduced in [64].

3.3. Inference with Skinning Weights

During inference, our fully differentiable parametric hand mesh model is driven by an input skeleton K and a shape latent code γ. The inference pipeline consists of the following steps:

(1) We extract the canonical hand mesh M from RefNet f using Marching Cubes and assign skinning weights to its vertices by upsampling the skinning weights of the MANO template to ensure compatibility with M.

(2) Using K and γ, we compute LCS_i, θ_i, and β_i.

(3) DeformNets deforms the vertices of M based on θ_i and β_i, and then transforms the deformed vertices to the coordinate system where the skeleton K is located using LCS_i. Finally, we combine the transformed vertices based on their skinning weights to obtain the final reconstruction result.

As our canonical hand mesh represented by RefNet f is static, step (1) is only needed once for all inferences.

4. Experiments

4.1. Datasets and Evaluation Metrics

Youtube3D (YT3D) [29]: The YT3D dataset has 50,175 hand meshes from 109 videos, with 47,125 meshes from 102 videos in the training set and 1,525 meshes from 7 videos in the test set. As the YT3D annotations are fitted from video data, we use its MANO pose and mesh for skeleton-driven hand shape reconstruction.

MANO dataset [61]: The training set contains 1,554 scans with MANO annotations of 31 subjects performing 51 poses, and the test set contains 50 scans of unseen subjects performing unseen poses to the training set. We use the real scans from the MANO trainset to learn our high-fidelity hand mesh model and use the MANO test set to evaluate point cloud 3D hand reconstruction.

DeepHandMesh (DHM) dataset [44]: The DHM dataset contains multi-view observations including RGB images and depth maps. 3D shape reconstructions and 3D hand keypoints are also provided. Since only one subject data is publically accessible, we use this available subset in our experiments, where the released data is split into a training set with 4,550 samples and a test set with 915 samples. We use DHM to learn hand reconstruction from images.

FreiHAND [78]: FreiHAND has 130,240 training and 3,960 evaluation samples, all with MANO annotations. Although the MANO annotations are not as accurate as real scans, the dataset is used to showcase how our model can learn high-fidelity single-view hand reconstruction using common yet coarse MANO-level annotations.

Evaluation metrics. For skeleton-driven hand reconstruction, we report the mean Intersection over Union (**IoU**), Chamfer-L1 distance in mm (**Cham.**) and normal consistency score (**Norm.**) in [40]. For hand reconstruction from point clouds, we compute vertex-to-vertex (**V2V** in mm) and vertex-to-surface (**V2S**) distances between the reconstruction and scan in both directions. For hand reconstruction from images, the evaluation metrics include 3D joint distance error ($\mathbf{P_{err}}$ in mm), mesh vertex error ($\mathbf{M_{err}}$), mean per-joint position error and per-vertex position error with Procrustes analysis [18] (**PA MPJPE** and **PA MPVPE** in mm), and F-score evaluated at thresholds of 5mm and 15mm (**F@5 mm** and **F@15 mm**).

Method	IoU↑	Cham.↓	Norm.↑
NASA [15]	0.896	1.057	0.955
HALO [26]	0.932	0.719	0.959
HALO(Keypoints)	0.930	0.740	0.959
Ours	**0.949**	**0.432**	**0.979**

Table 2: Results of skeleton-driven hand reconstruction on YT3D. NASA [15] and HALO [26] use GT bone transformation, while HALO(Keypoints) and ours take keypoints as input only.

Method	Recon. to scan		Scan to recon.	
	V2V↓	V2S↓	V2V ↓	V2S↓
MANO [61]	3.14	2.92	3.90	1.57
VolSDF [75]	3.69	2.22	2.37	2.23
NASA [15]	5.31	3.80	2.57	2.33
NARF [48]	4.02	2.69	2.11	2.06
LISA-im [13]	3.09	1.96	1.19	1.13
LISA-geom [13]	**0.36**	**0.16**	0.81	**0.26**
LISA-full [13]	1.45	0.64	0.64	0.58
Ours	0.37	0.27	**0.37**	0.31

Table 3: Results of shape reconstruction from point clouds on the MANO test set. We achieve comparable results to LISA using much fewer training scans.

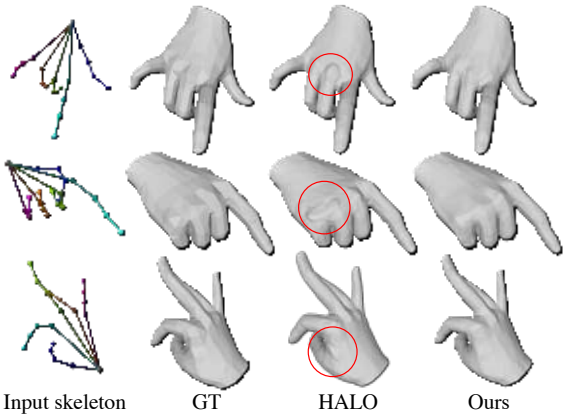

Figure 4: Skeleton-driven hand reconstruction results on YT3D. Both HALO and ours maintain the keypoint locations faithfully, while ours achieve more accurate and smoother reconstructions.

4.2. Implementation Details

Our model is implemented within Pytorch [55], where the Adam solver is used with mini-batches of size 32. The initial learning rate is set to 0.0005, and decayed by a factor of 0.5 after 250 epochs and 500 epochs. We train for 1000 epochs on 2 NVIDIA RTX 2080 Ti GPUs, which takes around 16 hours for training on the MANO dataset. For weighting factors, we set $w_S = 0.1$, $w_n = 1$, $w_c = 0.1$, $w_m = 0.1$, $w_r = 0.0001$, $w_O = 0.1$ and $w_I = 1$. We refer to the appendix for more details.

4.3. Results

We validated our method on three tasks using different datasets and compared it with state-of-the-art baselines.

Skeleton-driven hand reconstruction. In this task, we set NASA [15] and HALO [26] as our baselines. NASA represents an articulated shape with part-based implicit functions, and HALO builds upon NASA and specializes in neural hand representation. HALO proposes to eliminate demands for GT bone transformation by using differentiable skeleton canonicalization with 3D keypoints. Following HALO, we conduct experiments on YT3D [29]. As shown in Table 2, our method consistently outperforms HALO and NASA in all metrics. Note that we fixed the latent shape codes γ (Section 3.1) to be fully skeleton-driven, as HALO, to ensure a fair comparison. As shown in Fig. 4, our method achieves more accurate and smoother outputs in articulated hand modeling than HALO.We attribute the artifacts of HALO to their rigid skeleton canonicalization, which can not handle non-rigid deformation at connecting spots of parts effectively.

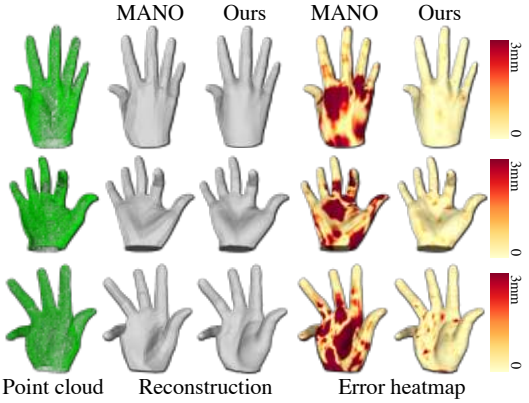

Figure 5: Results of hand reconstruction from point clouds. Comparisons between MANO and ours are shown.

Hand reconstruction from point clouds. To evaluate the accuracy of our 3D hand modeling, we test our model on high-quality point cloud data using the MANO test set [61]. This dataset includes poses and identities that differ from the training data used to develop our model. In Table 3, we compare our results with several state-of-the-art approaches, including MANO [61], image-learned methods (such as VolSDF [75], NASA [15], NARF [48], and LISA-im [13]) that are trained with multi-view images with the help of volume rendering, and another scan-learned method (LISA [13]). Our method outperforms the MANO model, which uses the same 1,544 scans for training, as well as all image-learned methods. We also outperform LISA-full and achieve comparable results to LISA-geom [13]. However, it is worth noting that LISA-geom and LISA-full are trained on a private dataset (3DH) consisting of 13k scans with optional multi-view images, while our method uses nearly 10 times fewer data. This demonstrates our method's superior ability to model high-fidelity hand meshes and generalize well. We provide visualizations of some reconstructions and error heatmaps in Fig. 5. Our method produces accurate and high-fidelity hand reconstructions.

Single-view hand reconstruction. Many existing methods for reconstructing 3D hands from single-view images only evaluate their performance on benchmarks with coarse MANO mesh annotations, which fail to capture the fine-detailed 3D reconstructions achievable with more advanced techniques such as DHM [44] and our approach. To address this limitation, we evaluate our method on the DHM dataset, which provides 3D reconstructions from depth observations

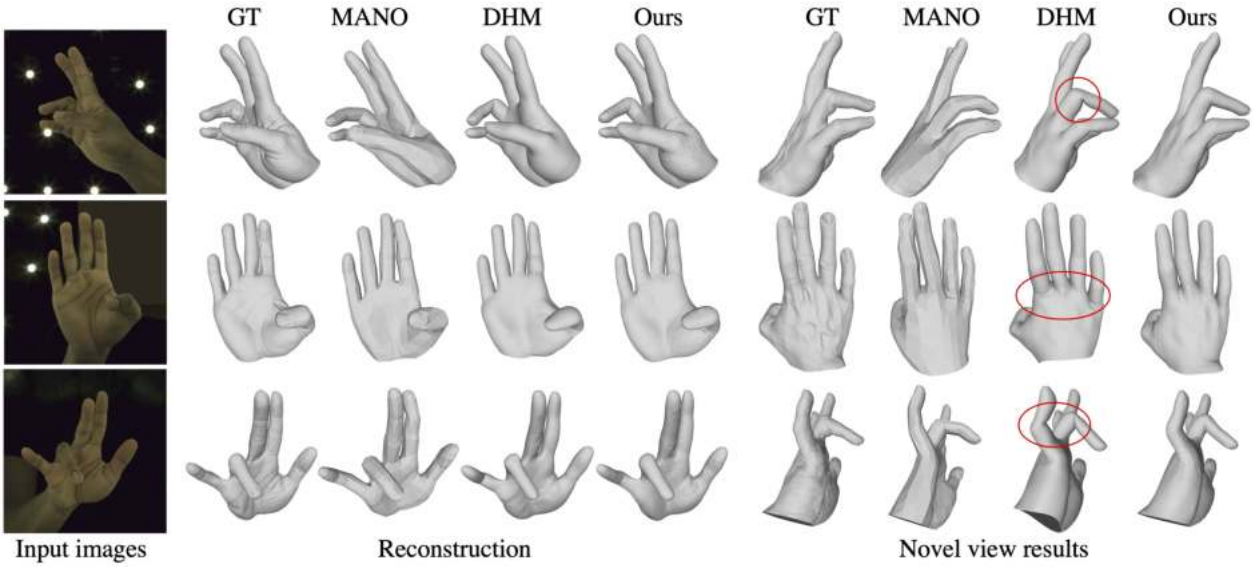

Figure 6: 3D reconstruction results from images on DHM. Both DHM [44] and ours reconstruct high-fidelity hands, while ours achieve more natural and accurate results.

Method	MPJPE↓	MPVPE↓	F@5 mm↑	F@15 mm↑
[22]	-	13.2	0.426	0.908
[3]	-	13.0	0.435	0.935
[78]	-	10.7	0.529	0.935
I2L-MeshNet [43]	**7.4**	**7.6**	**0.681**	0.973
MANO* [61]	10.8	12.4	0.485	0.947
NIMBLE* [32]	-	9.4	0.547	0.955
Ours-low*	7.5	7.8	0.660	**0.974**
Ours-high*	7.7	8.0	0.654	**0.974**

Table 4: Results of shape reconstruction from images on FreiHAND. *To be consistent with the evaluation in NIMBLE [32], MANO and ours results are also obtained upon an I2L-MeshNet pipeline.

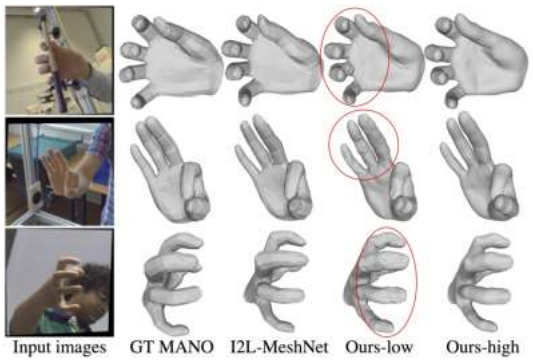

Figure 7: Results of shape reconstruction from images on FreiHAND.

for evaluating meticulous 3D shape predictions. Using the publicly available data and official implementation of multi-view training, we compare our model's single-view reconstruction performance to DHM and its variant with our hand mesh model in Table 5. The result of DHM is different from that of [44] since all the methods in the Table 5 are trained and tested on the publicly available portion of DHM, as the entire dataset is not accessible. Our results show that our model significantly improves the accuracy of both shape and pose predictions, demonstrating its superior ability to represent articulated hand shapes. Moreover, our model's clear-cut shapes can also aid in accurate pose estimation, as observed by [28]. Qualitative results are presented in Fig. 6, confirming the efficacy of our approach.

We also evaluate the proposed method on FreiHAND [78] with MANO pose and mesh annotations. We use our high-fidelity parametric hand model, which was trained on the MANO dataset scans, as a differentiable layer in the I2L-MeshNet pipeline. We use the predicted hand skeleton output and a two-layer MLP to estimate our shape latent code. We refer to this setting as "Ours-low" as we directly replace our template from the RefNet with the MANO template taking advantage of their compatibility (Section 3.1.1), and use the annotated MANO mesh to supervise the predicted shape with corresponding per-vertex constraints. We compare our results with two other parametric hand models, MANO [61] and NIMBLE [32], using the I2L-MeshNet pipeline. As shown in Table 4, Ours-low outperforms both MANO and NIMBLE by a significant margin. Despite using a lightweight MLP to predict parametric factors, our approach achieves comparable results to the state-of-the-art method [43], with significantly reduced network memory (by 78%) and computation costs (by 95%). Additionally, differentiable parametric models, such as MANO, NIMBLE, and our proposed method, can serve as a differentiable layer, enabling more related tasks, such as weakly-supervised learning [7, 29] and explicit control of hand motion in videos [72]. Visualizations of our results are shown in Fig. 7, where I2L-MeshNet fits MANO-level resolution GT meshes, while Ours-low demonstrates the ability to obtain finer surfaces.

However, using coarse MANO meshes in our training

Method	$P_{err}\downarrow$	$M_{err}\downarrow$
MANO [61]	8.51	3.75
DHM [44]	3.45	1.98
Ours	**3.14**	**1.50**

Table 5: Results of shape reconstruction from images on the publically released part of the DHM dataset.

Configs	Recon. to scan		Scan to recon.	
	V2V↓	V2S↓	V2V↓	V2S↓
w/o E_{norm}	0.51	0.40	0.51	0.44
w/o D.S.C.	0.47	0.36	0.46	0.39
w/ E_{mano^+}	0.45	0.34	0.44	0.38
Ours	**0.37**	**0.27**	**0.37**	**0.31**

Table 6: Result of ablation study on MANO test set.

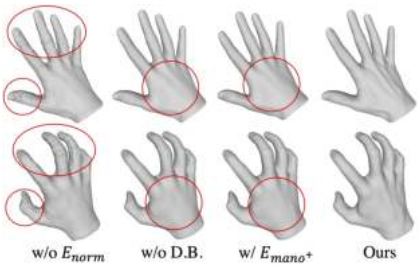

Figure 8: Qualitative results of ablation study.

resulted in rough and non-smooth hand reconstructions. To address this, we introduced looser supervision by using the chamfer distance between the GT MANO vertex coordinates and our hand reconstruction to supervise the network, rather than tight per-vertex constraints. This approach was applied to our high-resolution parametric template (Ours-high), resulting in higher-fidelity hand reconstructions from input single-view images (as shown in Fig. 7). However, while Ours-high achieves state-of-the-art qualitative results, it performs slightly worse in terms of quantitative evaluation metrics in Table 4. We suspect that the vertex-wise correspondence of MANO annotations using linear blend skinning and linear deformation correction may not accurately reflect real-world deformations, which are non-rigid and non-linear. Overall, our model can be easily integrated into existing methods to learn high-quality hand reconstructions from MANO-level annotated datasets. We achieved state-of-the-art qualitative results and competitive quantitative results under the MANO metric.

4.4. Ablation Study

We conduct ablation studies on our high-fidelity hand model by exploring various configurations.

Effect of the deformation field. The proposed deformation field bridges the SDF of the deformed hand and the canonical hand. On this basis, we derive E_{norm} in our training objectives. We ablate whether E_{norm} is used during training (Ours vs. w/o E_{norm} in Fig. 8 and Table 6), and the results show that E_{norm} is crucial to guarantee a smooth and immersive hand shape and brings key improvements to the quantitative results of point cloud 3D hand reconstruction on MANO test set.

Effect of deformation skip connections. We report the results of the holistic architecture in Fig. 3 in terms of whether the deformation skip connections are used (Ours vs. w/o D.S.C.). As presented in Fig. 8, the proposed skip-connection design more effectively represents the details of the hand surface and achieves better overall reconstruction performance (see Table 6).

Effect of dense correspondence supervision. To compare the impact of explicit dense correspondence supervision versus implicit deformation field, we replace E_{mano} in our training objectives (Section 3.2) with E_{mano^+}, which provides dense correspondence through barycentric coordinates by upsampling MANO mesh annotations by a factor of 10. As shown in Fig. 8 and Table 6, using E_{mano^+} resulted in degraded performance, demonstrating the superiority of implicitly building dense correspondence over pseudo supervision.

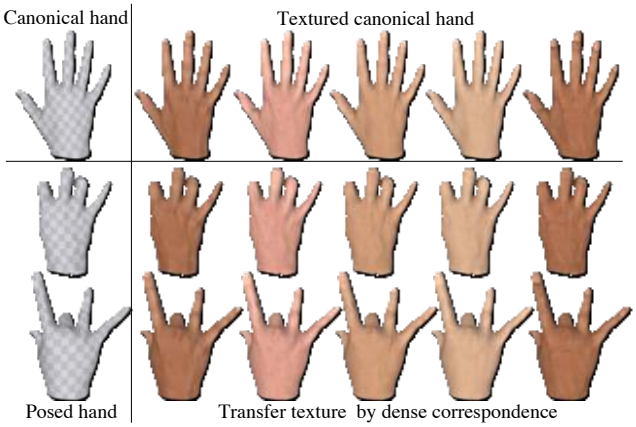

Figure 9: Results of transferring arbitrary hand textures through dense correspondence. We utilize HTML [60] to obtain textured canonical hands.

Method	MANO [61]	DHM [44]	LISA [13]	HALO [26]	Ours
Speed (s)	0.003	0.005	>5	16.3	0.099

Table 7: Comparison of inference speed. The time consumed for a feed-forward hand reconstruction is reported.

4.5. Discussion

Dense correspondence for texture modeling. Our model deforms a canonical hand mesh shared by all poses and identities to obtain dense correspondences among the hand reconstructions. We explore an application of per-vertex correspondences in transferring arbitrary textures from the texture model, such as HTML [60], to our hand reconstructions. Fig. 9 shows that our model can seamlessly combine with the HTML texture model to produce vividly textured hands with accurate geometry.

Inference speed. We compare the proposed method's inference speed with some baselines in Table 7. We implement all methods on the same consumer device: NVIDIA RTX2080Ti GPU, Intel Xeon E5-2620 v4@2.10GHz, 11 GB main memory. Our method achieved significantly faster inference speed than other implicit models like LISA [13] and HALO [26], by avoiding the time-consuming Marching Cubes process. However, our part-based design, which

improves accuracy and generalization, resulted in slightly slower inference times than other parametric mesh methods like MANO [61] and DHM [44]. We believe that improving the efficiency of our part-based design can be a future research direction.

Limitation. While our method achieves state-of-the-art performances in multiple tasks, it currently lacks pose priors, real-time inference capabilities, and support for queries beyond geometry modeling, such as intersection detection. Future research directions could include parameterizing hand skeletons, refining the part-based design to reduce computation cost and latency, modeling temporal smoothness and consistency and extending the deformation field to enable potential one-to-many correspondence.

5. Conclusion

We introduce PHRIT, a parametric hand model featuring an implicit template that generates high-fidelity hand reconstructions for various poses and identities. Our model offers efficient inference and dense correspondence inherited from traditional parametric meshes, as well as infinite-resolution reconstructions derived from implicit representations. By utilizing the proposed deformation field, we establish a dense correspondence between canonical and deformed spaces. Our model is fully differentiable with respect to both the skeleton and shape latent code, making it easy to incorporate into existing methods for versatile applications. Our experiments demonstrate that PHRIT outperforms previous methods in multiple downstream tasks.

Acknowledgements: This work was supported by the National Natural Science Foundation of China (NSFC) under Grant 62106177, and the Natural Science Fund for Distinguished Young Scholars of Hubei Province under Grant 2022CFA075. The numerical calculation was supported by the super-computing system in the Super-computing Center of Wuhan University.

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imghum: Implicit generative models of 3d human shape and articulated pose. In *ICCV*, pages 5461–5470, 2021. 2

[2] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194, 1999. 2

[3] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *ICCV*, pages 10843–10852, 2019. 2, 7

[4] Aljaz Bozic, Michael Zollhofer, Christian Theobalt, and Matthias Nießner. Deepdeform: Learning non-rigid rgb-d reconstruction with semi-supervised data. In *CVPR*, pages 7002–7012, 2020. 2

[5] 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 *ICCV*, pages 12929–12938, 2021. 2

[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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11594–11604, 2021. 2, 3, 4

[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*, pages 10451–10460, 2021. 2, 7

[8] Yujin Chen, Zhigang Tu, Di Kang, Ruizhi Chen, Linchao Bao, Zhengyou Zhang, and Junsong Yuan. Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. *IEEE Transactions on Image Processing*, 30:4008–4021, 2021.

[9] Zheng Chen, Sihan Wang, Yi Sun, and Xiaohong Ma. Self-supervised transfer learning for hand mesh recovery from binocular images. In *ICCV*, pages 11626–11634, 2021. 2

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5939–5948, 2019. 2

[11] Junhyeong Cho, Kim Youwang, and Tae-Hyun Oh. Cross-attention of disentangled modalities for 3d human mesh recovery with transformers. In *ECCV*, pages 342–359. Springer, 2022. 2

[12] 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 *ECCV*, pages 769–787. Springer, 2020. 2, 3

[13] 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. 1, 3, 6, 8

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, pages 5031–5041, 2020. 2

[15] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *ECCV*, pages 612–628. Springer, 2020. 2, 3, 5, 6

[16] Yao Feng, Jinlong Yang, Marc Pollefeys, Michael J Black, and Timo Bolkart. Capturing and animation of body and clothing from monocular video. *arXiv preprint arXiv:2210.01868*, 2022. 2

[17] 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 *CVPR*, pages 10833–10842, 2019. 2, 3

[18] John C Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975. 5

[19] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[20] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 2

[21] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *2021 International Conference on 3D Vision (3DV)*, pages 659–668. IEEE, 2021. 1

[22] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, pages 11807–11816, 2019. 2, 7

[23] Fangzhou Hong, Mingyuan Zhang, Liang Pan, Zhongang Cai, Lei Yang, and Ziwei Liu. Avatarclip: Zero-shot text-driven generation and animation of 3d avatars. *arXiv preprint arXiv:2205.08535*, 2022. 2

[24] Timothy Jeruzalski, David IW Levin, Alec Jacobson, Paul Lalonde, Mohammad Norouzi, and Andrea Tagliasacchi. Nilbs: Neural inverse linear blend skinning. *arXiv preprint arXiv:2004.05980*, 2020. 2

[25] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *CVPR*, pages 8320–8329, 2018. 2

[26] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *3DV*, pages 11–21. IEEE, 2021. 1, 2, 3, 4, 5, 6, 8

[27] 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

[28] Deying Kong, Linguang Zhang, Liangjian Chen, Haoyu Ma, Xiangyi Yan, Shanlin Sun, Xingwei Liu, Kun Han, and Xiaohui Xie. Identity-aware hand mesh estimation and personalization from rgb images. In *ECCV*, pages 536–553. Springer, 2022. 7

[29] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, pages 4990–5000, 2020. 5, 6, 7

[30] Mengcheng Li, Liang An, Hongwen Zhang, Lianpeng Wu, Feng Chen, Tao Yu, and Yebin Liu. Interacting attention graph for single image two-hand reconstruction. In *ICCV*, pages 2761–2770, 2022. 2

[31] 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):194–1, 2017. 2

[32] Yuwei Li, Longwen Zhang, Zesong Qiu, Yingwenqi Jiang, Nianyi Li, Yuexin Ma, Yuyao Zhang, Lan Xu, and Jingyi Yu. Nimble: a non-rigid hand model with bones and muscles. *ACM Transactions on Graphics (TOG)*, 41(4):1–16, 2022. 1, 2, 7

[33] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers. In *CVPR*, pages 1954–1963, 2021. 2

[34] Sandro Lombardi, Bangbang Yang, Tianxing Fan, Hujun Bao, Guofeng Zhang, Marc Pollefeys, and Zhaopeng Cui. Latenthuman: Shape-and-pose disentangled latent representation for human bodies. In *3DV*, pages 278–288. IEEE, 2021. 2

[35] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16, 2015. 2

[36] 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. 2

[37] Ameersing Luximon and Ravindra S Goonetilleke. Foot shape modeling. *Human Factors*, 46(2):304–315, 2004. 2

[38] Qianli Ma, Jinlong Yang, Michael J Black, and Siyu Tang. Neural point-based shape modeling of humans in challenging clothing. *arXiv preprint arXiv:2209.06814*, 2022. 2

[39] Ishit Mehta, Manmohan Chandraker, and Ravi Ramamoorthi. A level set theory for neural implicit evolution under explicit flows. In *ECCV*, pages 711–729. Springer, 2022. 2

[40] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, pages 4460–4470, 2019. 2, 5

[41] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *CVPR*, pages 10461–10471, 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[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 *ECCV*, pages 752–768. Springer, 2020. 7

[44] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, pages 440–455. Springer, 2020. 1, 2, 5, 6, 7, 8, 9

[45] 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*, pages 548–564. Springer, 2020. 2

[46] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *ICCV*, pages 5379–5389, 2019. 2

[47] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, pages 3504–3515, 2020. 2

[48] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *ICCV*, pages 5762–5772, 2021. 2, 3, 6

[49] Michael Oechsle, Michael Niemeyer, Christian Reiser, Lars Mescheder, Thilo Strauss, and Andreas Geiger. Learning implicit surface light fields. In *2020 International Conference on 3D Vision (3DV)*, pages 452–462. IEEE, 2020. 2

[50] Ahmed AA Osman, Timo Bolkart, and Michael J Black. Star: Sparse trained articulated human body regressor. In *ECCV*, pages 598–613. Springer, 2020. 2

[51] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. Npms: Neural parametric models for 3d deformable shapes. In *ICCV*, pages 12695–12705, 2021. 2

[52] Pablo Palafox, Nikolaos Sarafianos, Tony Tung, and Angela

Dai. Spams: Structured implicit parametric models. In *CVPR*, pages 12851–12860, 2022. 2

[53] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, pages 1496–1505, 2022. 2

[54] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2

[55] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019. 6

[56] 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. 2

[57] 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*, pages 14314–14323, 2021. 2

[58] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *ECCV*, pages 523–540. Springer, 2020. 2

[59] 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*, pages 9054–9063, 2021. 2

[60] 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*, pages 54–71. Springer, 2020. 1, 8

[61] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36(6):1–17, 2017. 1, 2, 3, 5, 6, 7, 8, 9

[62] 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*, pages 2304–2314, 2019. 2

[63] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. Scanimate: Weakly supervised learning of skinned clothed avatar networks. In *CVPR*, pages 2886–2897, 2021. 2

[64] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 5

[65] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, pages 211–228. Springer, 2020. 3

[66] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 2

[67] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6454–6471, 2021. 2

[68] Jiapeng Tang, Jiabao Lei, Dan Xu, Feiying Ma, Kui Jia, and Lei Zhang. Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6504–6513, 2021. 2

[69] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. *Advances in Neural Information Processing Systems*, 35:17117–17132, 2022. 2

[70] Xiao Tang, Tianyu Wang, and Chi-Wing Fu. Towards accurate alignment in real-time 3d hand-mesh reconstruction. In *ICCV*, pages 11698–11707, 2021. 2

[71] Zhigang Tu, Zhisheng Huang, Yujin Chen, Di Kang, Linchao Bao, Bisheng Yang, and Junsong Yuan. Consistent 3d hand reconstruction in video via self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023. 2

[72] John Yang, Hyung Jin Chang, Seungeui Lee, and Nojun Kwak. Seqhand: Rgb-sequence-based 3d hand pose and shape estimation. In *ECCV*, pages 122–139. Springer, 2020. 7

[73] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, pages 20953–20962, 2022. 2

[74] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *ICCV*, pages 11097–11106, 2021. 2

[75] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 6

[76] 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*, pages 12803–12813, 2021. 2

[77] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[78] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 5, 7

[79] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *ICCV*, pages 6365–6373, 2017. 2