Appendix of
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

In this appendix, we provide various visual illustrations, extended evaluations, and other details.The contents of this appendix are summarized below:

- Visualization of hand part partitions and hand skeleton (Section A).
- Details on the rotation of each bone transformation B_j (Section B).
- Extended evaluations of shape reconstruction from images on FreiHAND using chamfer distance (Section C); see main manuscripts for Table 4.
- Determination and sensitivity of the loss weights (Section D).
- Mathematical definition of the proposed deformation field and its constrained domain (Section E).
- Details on network architecture and training (Section F).
- Additional qualitative results (Section G).

A. Visualization of hand part partitions and hand skeleton.

In Fig. 10, we visualize a canonical skeleton of the canonical hand mesh, with all the bones $\{b_j \mid 1 \leq j \leq 20\}$ and the local coordinate systems $\{LCS_i \mid 1 \leq i \leq 16\}$.

In Fig. 11, we show part segments $\{P_i \mid 1 \leq i \leq 16\}$ on a colored canonical hand. For each part, both the palm view and back view are shown.

B. Details on the rotation of each bone transformation B_j.

For bone transformation B_j of bone b_j, the rotation (orientation) is set following HALO [2]: (1) For 5 palmar bones, z-axes are defined by normalized bone vectors, x-axes are defined by the normal of palmar planes (same as HALO [2]), and y-axes are obtained by a cross product. (2) For 15-finger bones, axes are defined by rotating the axes of parent bones along the kinematic chain using computed abduction and flexion angles.

C. Extended evaluations of shape reconstruction from images on FreiHAND using chamfer distance.

Table 4 of the main paper compares our reconstruction results with baselines on FreiHAND using per-vertex metrics. However, since the vertex-wise correspondence of MANO annotations may not reflect real-world deformation, we also evaluate shape reconstruction based on Chamfer distance. As shown in Table 8, Ours-high achieves state-of-the-art performance, surpassing MANO [5] and I2L-MeshNet [3].

Method	Chamfer Distance (mm)↓
I2L-MeshNet	<u>3.46</u>
MANO*	4.95
Ours-low*	3.58
Ours-high*	**3.43**

Table 8: Comparison with baselines of shape reconstruction from images on FreiHAND [8] based on the evaluation metric of Chamfer Distance.

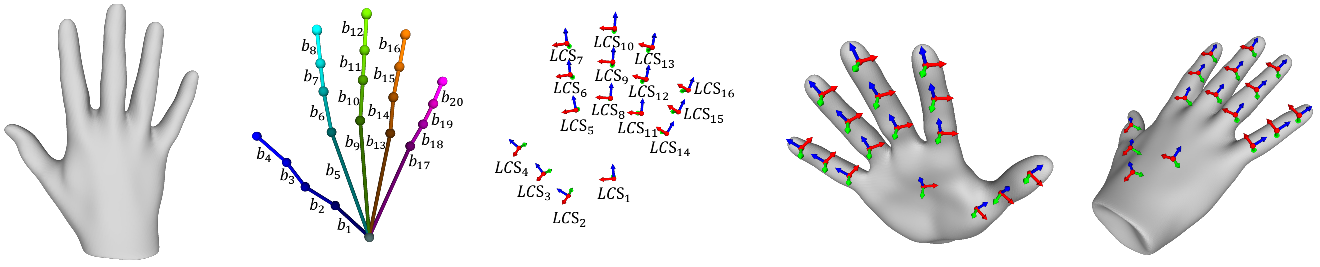
D. Determination and sensitivity of the loss weights

We use consistent weighting factors in our experiments. The $w_r = 0.0001$, $w_O = 0.1$, and $w_I = 1$ follow previous works [6–7]; the remaining loss terms are empirically selected. We find our model is not that sensitive to these weights: using $w_S = 1$ rather than $w_S = 0.1$ provides V2V 0.39 in Tab.3, using $w_n = 10$ than $w_n = 1$ provides V2V 0.38 in Tab.3.

E. Mathematical definition of the proposed deformation and its constrained domain

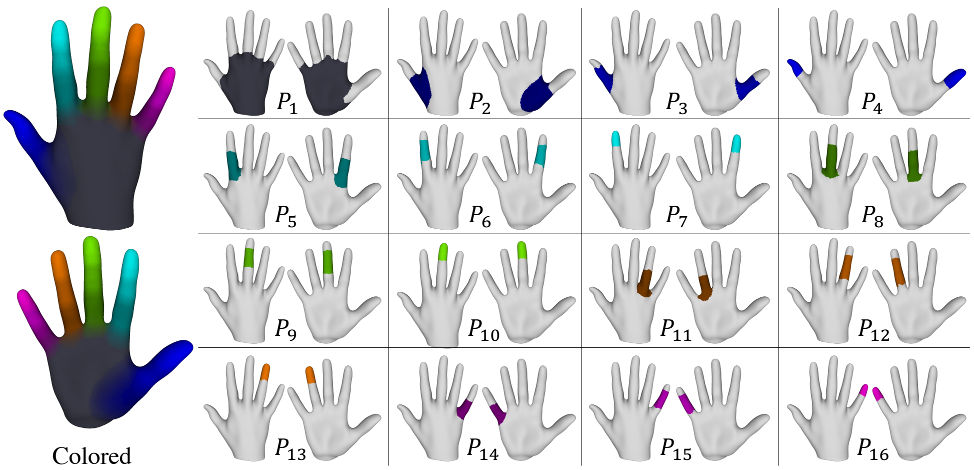
The proposed deformation field ϕ is a one-to-one mapping function between canonical and deformed space. We constrain the domain of ϕ nearby the hand surface (such as within 3mm from the surface). Using this deformation

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



Canonical hand　　Canonical skeleton　　$\{LCS_i\}$ of canonical hand　　$\{LCS_i\}$ of canonical hand in other views

Figure 10: Visualization of the canonical hand learned by PHRIT with its canonical hand skeleton and local coordinated systems.



Colored canonical hand　　$\{P_i\}$ of canonical hand

Figure 11: Visualization of our decomposition of the human hand, where the hand is partitioned into 16 rigid parts.

field, we design our DeformNets as an invertible mapping function, introducing an inverse counterpart called InvDeformNets. With these networks, we derive our training objectives to learn per-vertex deformation (i.e., dense correspondence) between canonical and deformed hands effectively, without requiring dense supervision. Further details are provided in Section 3.2 of the main manuscript.

To validate our approach theoretically, we provide a more concrete mathematical definition of the deformation field ϕ and its constrained domain in this appendix. For simplicity, we present the content in 2D space, but it can be easily generalized to 3D space.

In this context, we denote the hand surface in canonical space and deformed space as two smoothing curves, L_c and L_d, respectively. These curves are without self-interaction, and each point on the curve has its normal, as shown in the top of Fig. 12. It is important to note that there are multiple feasible one-to-one continuous correspondences between L_c and L_d, as demonstrated in Fig. 12. We argue that such a one-to-one correspondence can be arbitrarily assigned, but it will correspond to a unique ϕ based on our definition. In other words, once our ϕ is established, a one-to-one correspondence is implicitly built.

To define our deformation field ϕ based on the correspondence between L_c and L_d, we introduce some important concepts and notations (shown in Fig. 13). The curve SDF extends SDF to 2D space, and the point projection and the maximum projection radius are the keys to defining the ϕ as a one-to-one mapping function.

Curve SDF. SDF (Signed Distance Field) can be extended to a curve L in 2D space by assigning the sign of the distance. In 2D space, points on the right side of the curve L have negative distances, while those on the left side have positive distances. The right and left sides of a curve can be determined by the direction of the curve (see Fig. 13(a)).

Point projection. Given a query point X and a curve L, the point projection finds the closest point $P(X)$ on L to X. Since there may exist multiple projections, we use $P(X)$ to indicate that X has only one projection, and $\mathbb{P}(X)$ to indicate that X has multiple projections, where $\mathbb{P}(X)$ is corresponding projection set of X (see Fig. 13(b)).

Maximum projection radius. Given a point X on a curve L, we observe that all the points that have X as their only projection on the L lie on the normal at X: $\forall Y \in \{Y \mid P(Y) = X\}, Y \in \mathcal{N}(X)$, where $\mathcal{N}(X)$ denotes the normal at X. The maximum projection radius further narrows down this range by finding a symmetric area nearby the surface. The maximum projection radius gives the maximum radius $M(X)$ at X, that satisfies $\forall Y \in \mathcal{N}(X) \cap \mathcal{O}, P(Y) = X$, where $\mathcal{O}$ is the area with the center X and radius $M(X)$: $\mathcal{O} = \{Y \mid \|XY\| < M(X)\}$, $\|\cdot\|$ computes the length of a line. As shown in Fig. 13 (c), $M(X)$ is no larger than the well-known curvature radius at X.

Notations. We denote SDFs for curve L_c and L_d as SDF_c and SDF_d, point projections for curve L_c and L_d as $P_c(\cdot)$ ($\mathbb{P}_c(\cdot)$) and $P_d(\cdot)$ ($\mathbb{P}_d(\cdot)$), and the maximum projection radius for L_c and L_d as $M_c(\cdot)$ and $M_d(\cdot)$. We use $\|\cdot\|$ to compute the length of a line.

Now, we can define our deformation field ϕ as depicted in Fig. 14. For a point X_c in canonical space, ϕ maps it to deformed space $\phi(X_c) = X_d$ by first finding X_c 's projection on L_c denoted as X_c^*, then finding X_c^* 's correspondence on L_d denoted as X_d^*, and finally, along the normal of X_d^* (i.e., $\mathcal{N}(X_d^*)$), finding X_d such that $\|X_c X_c^*\| = \|X_d X_d^*\|$ and X_c and X_d are on the same side of the L_c and L_d (i.e., $SDF_c(X_c) = SDF_C(X_d)$).

As shown in Fig. 14, this definition immediately requires a constrained domain to be a one-to-one mapping. Specifically, the constraint domain for ϕ in canonical space can be:

$$\{X_c \mid |SDF_c(X_c)| < \min(M_c(X_c^*), M_d(X_d^*))\} \quad (1)$$

where $X_c^* = P_c(X_c), X_d^* = \phi(X_c^*)$ as illustrated in Fig. 14. And the constraint domain for ϕ^{-1} in deformed space can be:

$$\{X_d \mid |SDF_d(X_d)| < \min(M_c(X_c^*), M_d(X_d^*))\} \quad (2)$$

F. Details on network architecture and training.

Network architecture. To train PHRIT with scans, we use SIREN MLPs with frequency $\omega = 30$ in the sinus activation (as described in [7]) for our RefNet, DeformNets and InvDeformNets. RefNet is composed of 2 layers with 256 hidden dimensions. DeformNets and InvDeformNets have the same structure. Take DeformNets as an example: (1) For the palm part P_1, the DeformNet g_1 has 6 layers with 512 hidden dimensions. (2) For other part P_i, the DeformNet g_i has 4 layers with 256 hidden dimensions. (3) For conditioning each DeformNets g_i on pose code θ_i and shape code β_i, we employ FiLM conditioning proposed by [1], as they have demonstrated conditioning-by-concatenation is sub-optimal for period activations. FiLm conditioning involves a mapping network that takes in a latent code z (which in our case is the concatenation of θ_i and β_i) and outputs frequencies and phases to condition each layer of SIREN MLPs. The mapping network is a 4-layer LeakyReLU MLP with hidden dimension of 128. (4) For deformation skip connections, we set the number of skip connections (See N in the main manuscript) to 2 for each part model g_i.

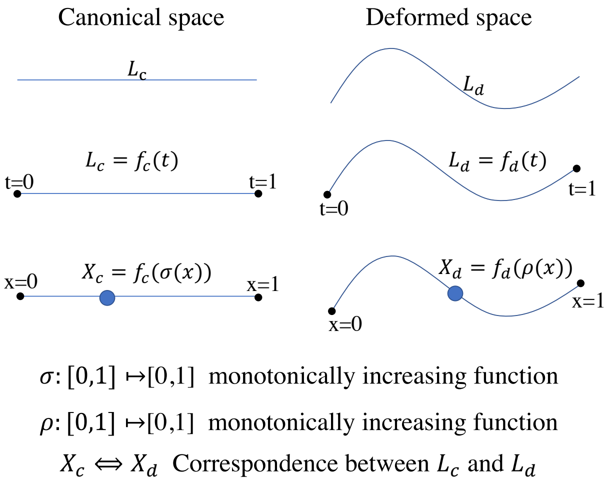

Canonical space
Deformed space
L_c
L_d
$L_c = f_c(t)$
t=0
t=1
$L_d = f_d(t)$
t=0
t=1
$X_c = f_c(\sigma(x))$
x=0
x=1
$X_d = f_d(\rho(x))$
x=0
x=1
σ: [0,1] ↦[0,1] monotonically increasing function
ρ: [0,1] ↦[0,1] monotonically increasing function
$X_c \Leftrightarrow X_d$ Correspondence between L_c and L_d

Figure 12: Canonical hand surface and deformed hand surface are denoted as 2D curves L_c and L_d in respective spaces. L_c and L_d can be parameterized by $f_c(t)$ and $f_d(t)$, where $t \in [0, 1]$), and $f_c(0)$ and $f_d(0)$ are the starts of the curves, $f_c(1)$ and $f_d(1)$ are the ends of the curves. By defining monotonically increasing functions ρ and σ, multiple continuous correspondence between L_c and L_d can be assigned.

For fair comparison with HALO [2] in skeleton-driven hand reconstruction with MANO meshes, we obtain the results in Table 2 of the main papaer by using MLP w/o PE architecture (no SIREN included), which is consistent with [2]. We further compare using SIREN vs. MLP+PE learning with real-world scans (on reconstruction from point clouds or images), resulting in 3.14 vs. 3.29 on P_{err}, 1.50 vs. 1.58 on M_{err} in Table 5 of the main paper, and 0.37 vs. 0.45 on V2V in Table 3 of the main paper, showing SIREN could slightly advance the performance.

Traning data. To train PHRIT with scans, we preprocess each scan to obtain query points with their normals for each hand part, where mano annotations are utilized to label the scan points with skinning weights, and then scan points are gathered for each hand part based on these skinning weights. During training, for each scan in a training batch, the query points consist of: (1) 2000 points with their part labels that are sampled from the preprocessed scan, (2) 778 MANO vertices from the MANO annotation of the scan, and (3) an additional 1000 points uniformly sampled in the

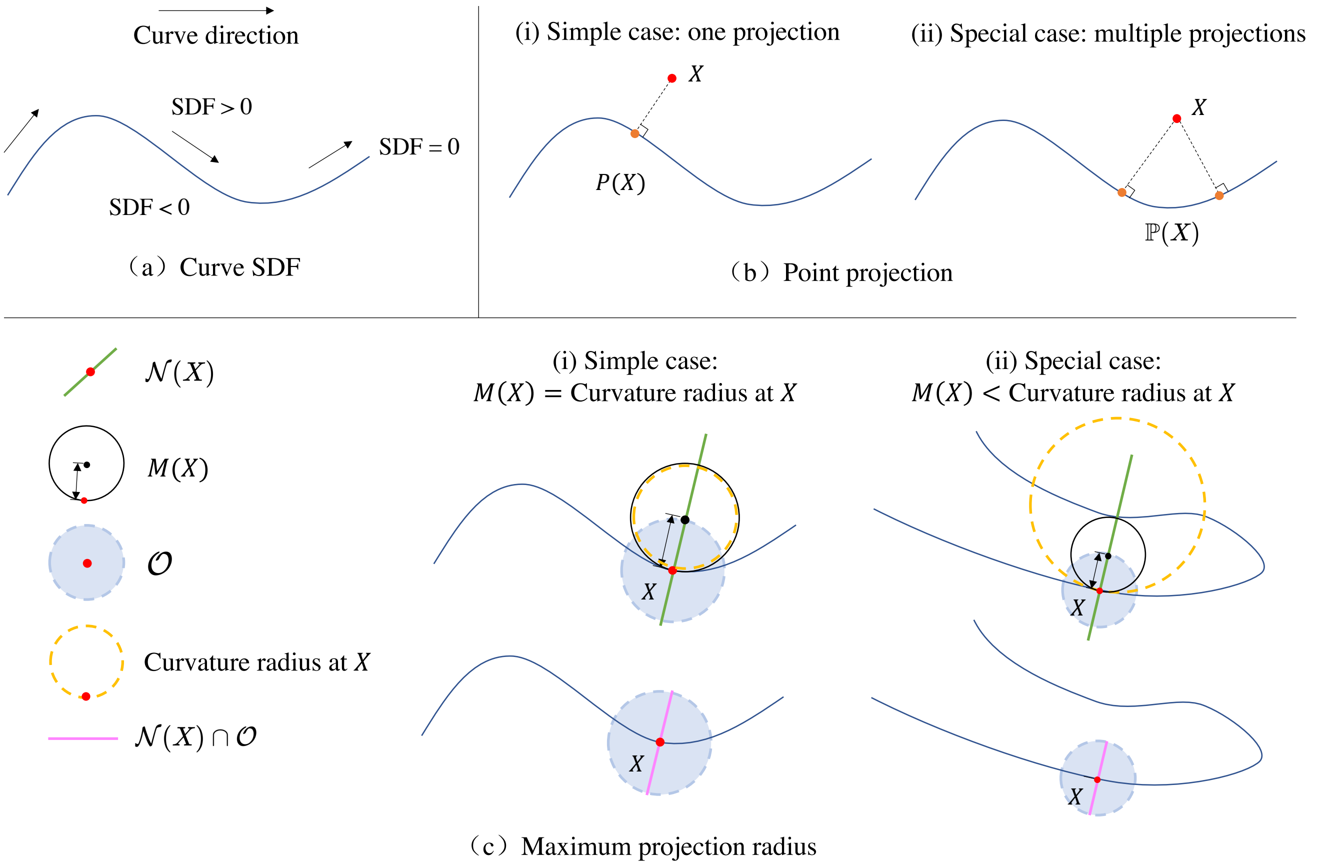


Figure 13: (a) SDF (Signed Distance Field) for a curve in 2D space. (b) Point projection, where a query point is projected onto the curve. (c) Maximum projection radius for a query point on the curve.

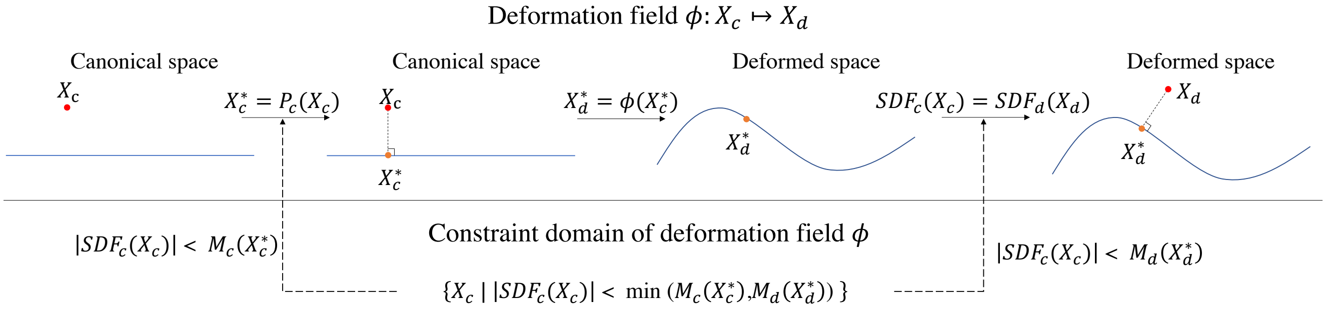

Deformation field $\phi: X_c \mapsto X_d$
Canonical space
X_c
$X_c^* = P_c(X_c)$
Canonical space
X_c
X_c^*
$X_d^* = \phi(X_c^*)$
Deformed space
X_d^*
$SDF_c(X_c) = SDF_d(X_d)$
Deformed space
X_d
X_d^*
$|SDF_c(X_c)| < M_c(X_c^*)$
Constraint domain of deformation field ϕ
$\{X_c \mid |SDF_c(X_c)| < \min(M_c(X_c^*), M_d(X_d^*))\}$
$|SDF_c(X_c)| < M_d(X_d^*)$

Figure 14: Deformation field $\phi : X_c \mapsto X_d$ with its constraint domain. ϕ is essentially a one-to-one mapping function nearby the hand surface, build upon the correspondence between the canonical hand and deformed hand ($X_d^* = \phi(X_c^*)$).

canonical hand space to train RefNet.

G. Additional qualitative results.

We present additional results on DHM [4] and FreiHAND [8] in Fig. 15 and Fig. 16, respectively. We also provide visualizations of the effects of learned shape latent code and bone lengths on hand shape reconstruction, in Fig. 17. Furthermore, we demonstrate that our method, PHRIT, can reconstruct hands from point clouds (as described in Section 4.3 of the main manuscript) and drive the reconstruction to previously unseen poses with impressive results, as shown in Fig. 18.

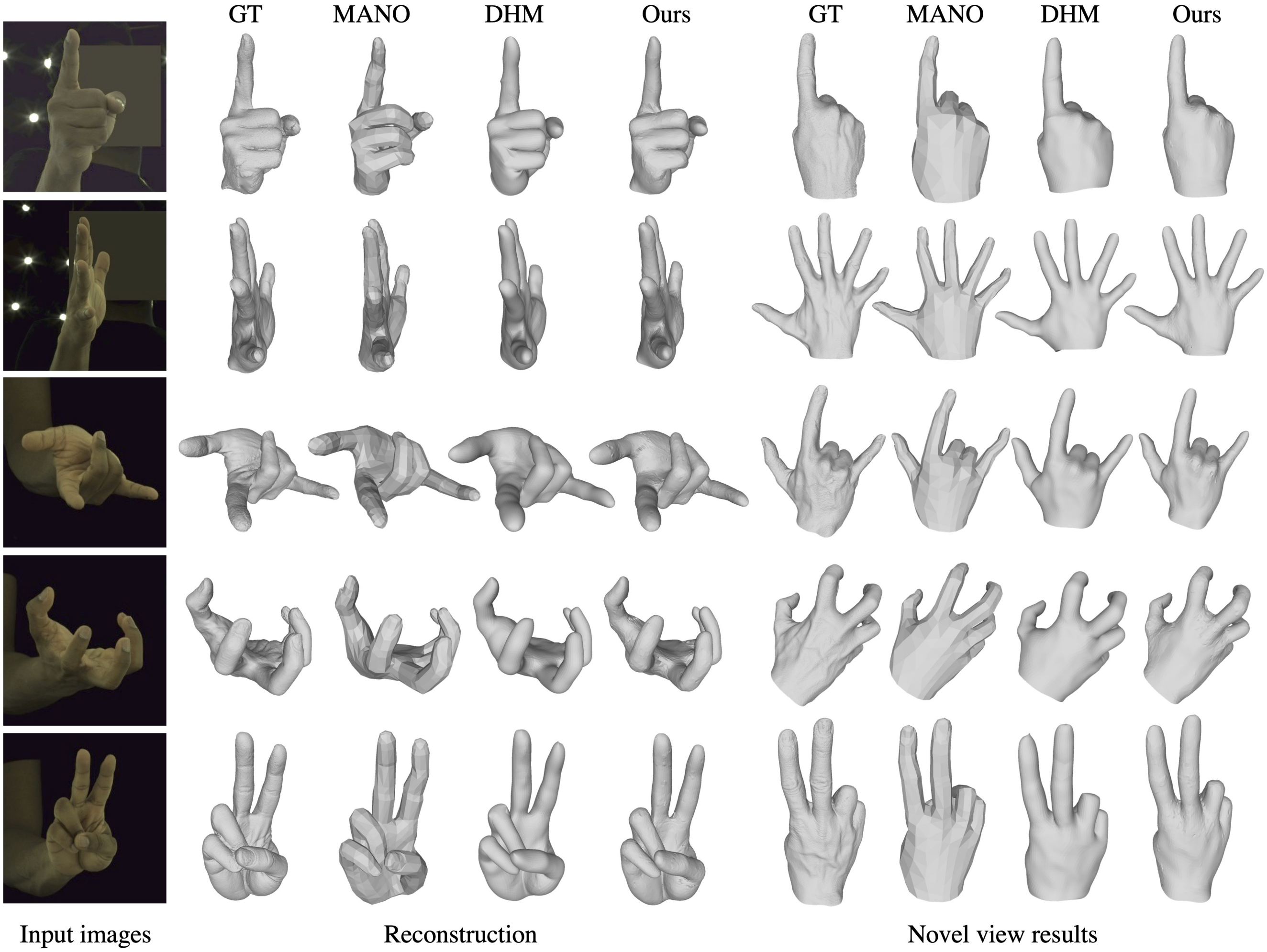


Figure 15: Additional results of reconstruction from images on DHM as a supplement to Fig6 and Table 5 in the main manuscript.

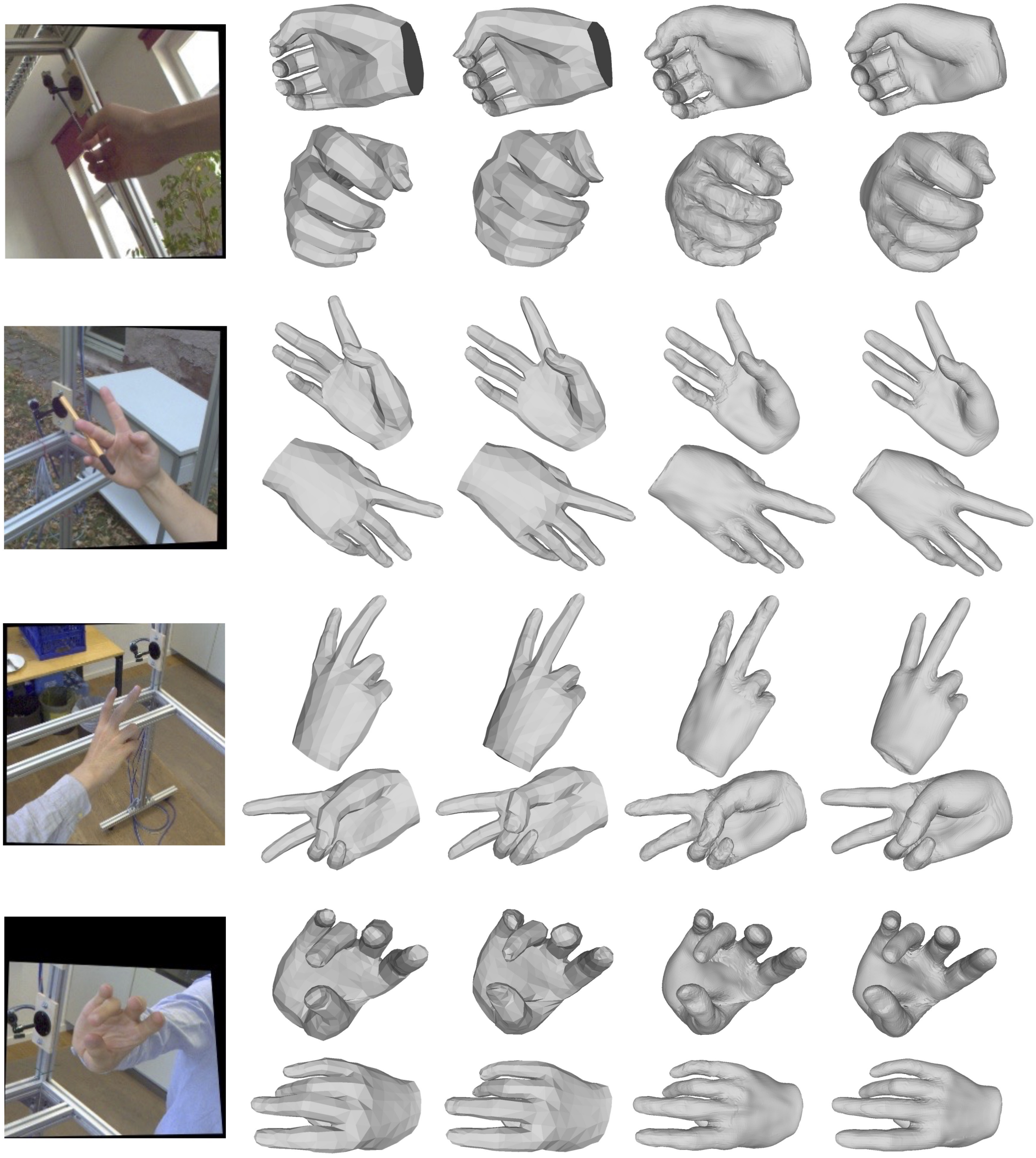


Figure 16: Additional results of reconstruction from images on FreiHAND as a supplement to Fig.7 and Table 4 in the main manuscript.

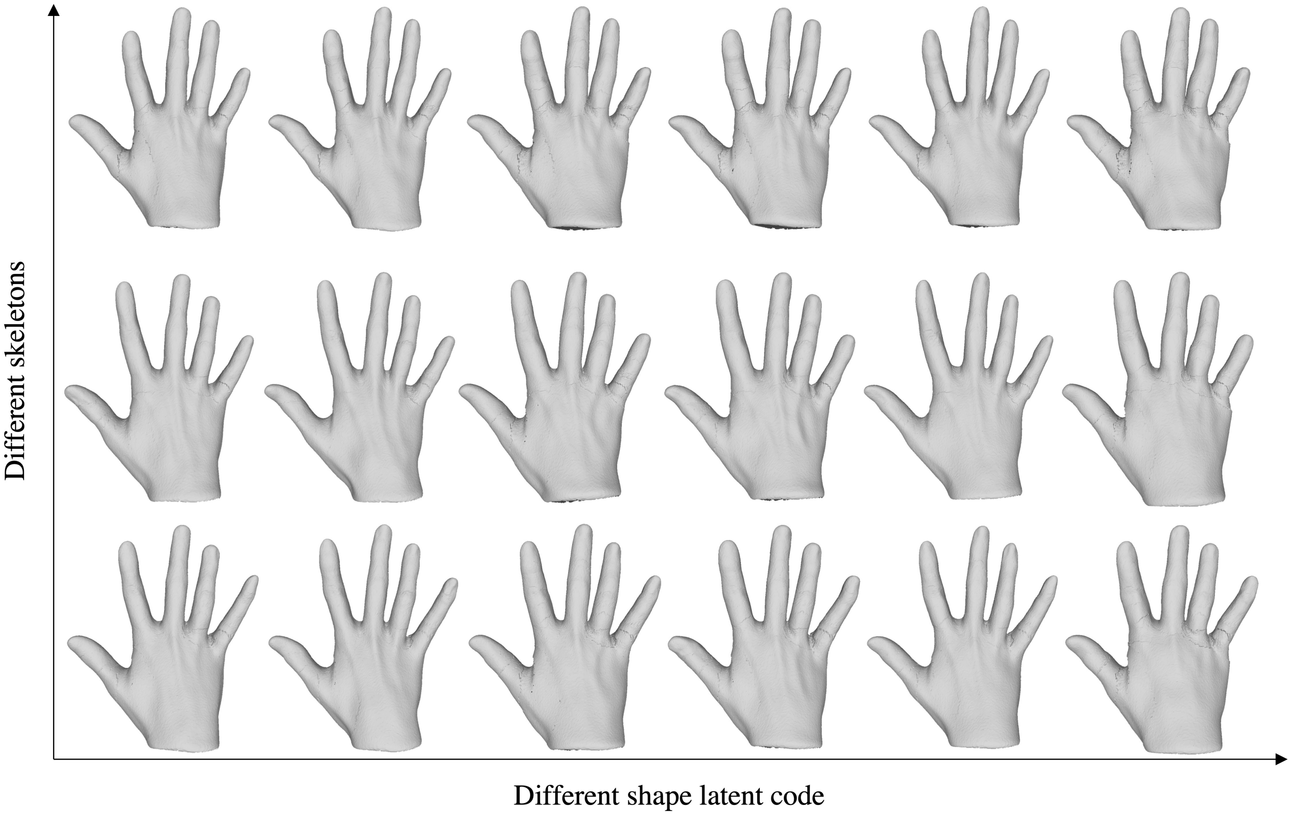


Figure 17: The visualization in the figure demonstrates the effects of changing the learned shape latent code and bone lengths on hand shape reconstruction. Horizontally, we show the results of changing the input shape latent code. Vertically, we show the results of changing the input skeleton, which corresponds to the same poses but with different bone lengths. As can be seen, the shape latent code controls surface properties such as hand thickness and veins, while the bone lengths control joint positions. For example, in the first column, the ring finger is longer than the index finger, while in the second row, the index finger is longer than the ring finger. In the third row, the little finger has the longest relative length compared to other rows.

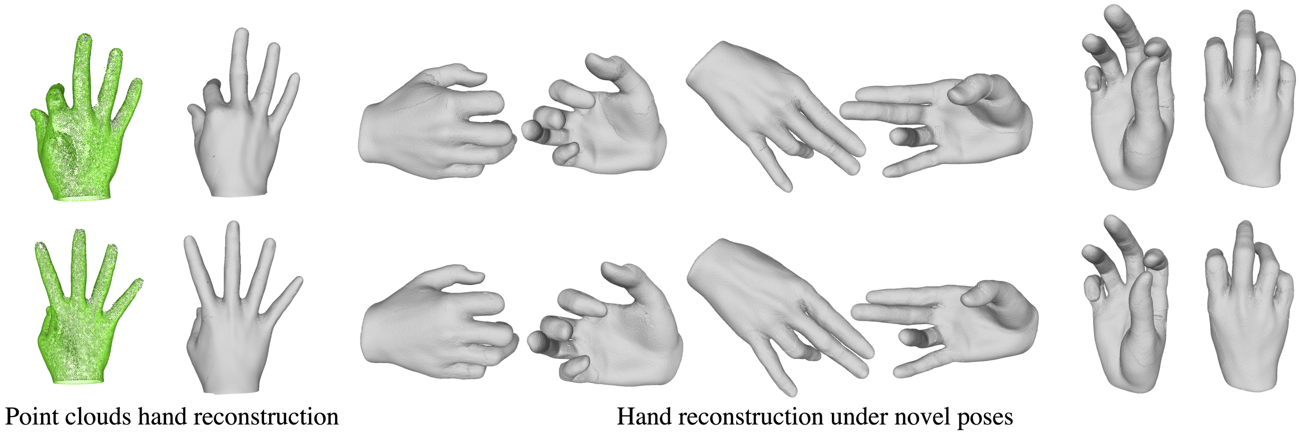

Point clouds hand reconstruction
Hand reconstruction under novel poses

Figure 18: Results of our hand reconstruction from point clouds under novel poses. The poses are randomly generated using the MANO model. PHRIT is able to drive the hand reconstruction to previously unseen poses with realistic results.

References

[1] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *CVPR*, pages 5799–5809, 2021. 3

[2] 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, 3

[3] 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. 1

[4] 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. 4

[5] 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

[6] 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. 1

[7] 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. 1, 3

[8] 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. 1, 4