arXiv:2109.11399v1 [cs.CV] 23 Sep 2021

A Skeleton-Driven Neural Occupancy Representation for Articulated Hands

Korrawe Karunratanakul Adrian Spurr Zicong Fan Otmar Hilliges Siyu Tang
ETH Zürich
{korrawe.karunratanakul,adrian.spurr,zicong.fan,otmar.hilliges,siyu.tang}@inf.ethz.ch

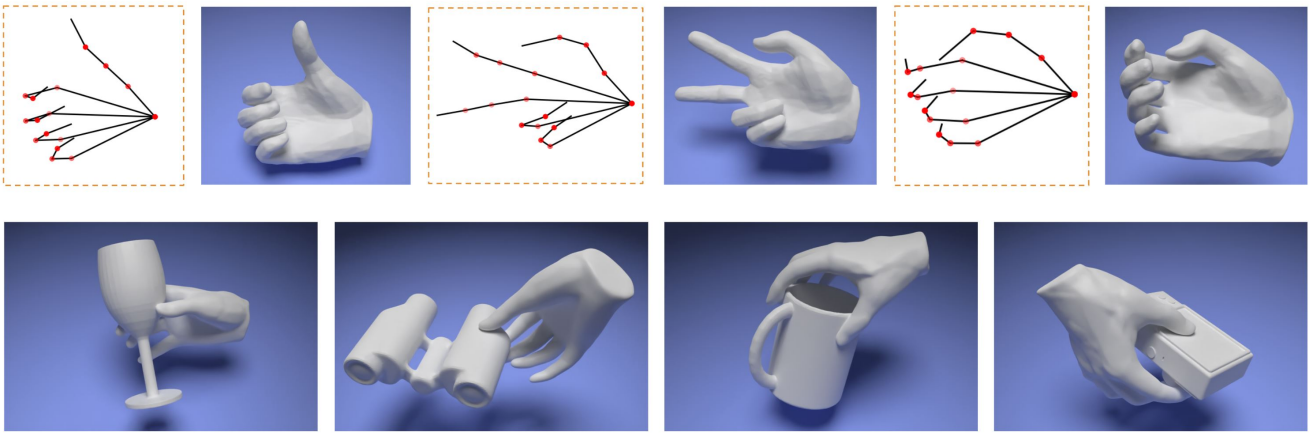

Figure 1. We introduce a novel neural implicit surface representation of human hands (HALO) that is fully driven by keypoint-based skeleton articulation. Taking 3D keypoints as input, a fully differentiable implicit occupancy representation produces high-fidelity reconstructions of the hand surface (top row). We show that HALO facilitates the conditional generation of articulated hands that grasp 3D objects in a realistic and physically plausible manner (bottom row).

Abstract

We present Hand ArticuLated Occupancy (HALO), a novel representation of articulated hands that bridges the advantages of 3D keypoints and neural implicit surfaces and can be used in end-to-end trainable architectures. Unlike existing statistical parametric hand models (e.g. MANO), HALO directly leverages 3D joint skeleton as input and produces a neural occupancy volume representing the posed hand surface. The key benefits of HALO are (1) it is driven by 3D keypoints, which have benefits in terms of accuracy and are easier to learn for neural networks than the latent hand-model parameters; (2) it provides a differentiable volumetric occupancy representation of the posed hand; (3) it can be trained end-to-end, allowing the formulation of losses on the hand surface that benefit the learning of 3D keypoints. We demonstrate the applicability of HALO to the task of conditional generation of hands that grasp 3D objects. The differentiable nature of HALO is shown to improve the quality of the synthesized hands both in terms of physical plausibility and user preference.

1. Introduction

Humans grasp and manipulate objects with their hands. Modeling 3D poses and surfaces of human hands is important for numerous applications such as animation, games, augmented and virtual reality. Existing hand representations in the literature can be categorized into two paradigms: skeleton representations [19, 28, 40, 41, 71] and mesh-based representations [39, 51]. Even though 3D skeletons are defined in the Euclidean space and are easy to interface with deep neural networks, they lack surface information and therefore not suitable for reasoning about hand-object interaction. In contrast, mesh-based hand models provide surfaces and can thus explicitly reason about physical interactions, such as hand-object manipulation [25, 59]. However, their pose and shape parametrizations are often hard to directly interpret and more difficult to learn in an end-to-end fashion than 3D keypoints.

In this work, we aim to bridge the gap between 3D keypoints and dense surface models. To this end, we propose **H**and **A**rticu**L**ated **O**ccupancy (HALO), a novel hand representation that is driven by keypoint-based skeleton artic-

ulation and provides a high-fidelity, resolution-independent neural implicit surface. Importantly, the proposed representation is fully differentiable and can be trained end to end such that volume-based losses can back-propagate gradients to the 3D keypoints. Specifically, HALO makes two key innovations: (1) a fully differentiable skeleton canonicalization algorithm and (2) a shape-aware neural articulated hand representation that is driven directly by skeleton and naturally affords differentiable reasoning about surface and volumetric occupancy.

Differentiable skeleton canonicalization. The core advantage of the HALO model is that its surface representation is *fully differentiable* and driven by the *3D joint skeleton*. Even though a naive iterative fitting procedure such as inverse kinematics could provide the transformations needed for changing from the canonical pose to the target skeleton, it prevents gradient flow from the surface to the keypoints. In addition, due to ambiguities of the twist angle around the finger bones, the optimization procedure may lead to unnatural surfaces. By leveraging bio-mechanical constraints that ensure plausible poses, we propose a novel differentiable layer that converts 3D keypoint locations to the corresponding bone rotations and translations with respect to a canonicalized pose space in a single forward pass. We call this operation a canonicalization layer. Importantly, this layer does not rely on iterative fitting and can therefore be effectively used in back-propagation-based learning. Furthermore, this layer enables the learning of implicit hand shape representations in the canonical space. This significantly improves the generalization capability of the learned representations across different hand poses and shapes, as shown in Sec. 5.

Skeleton-driven neural articulated hand. Recently, several neural occupancy networks for human body modelling, e.g. [15, 36, 53], have been proposed. Despite demonstrating the feasibility of parameterizing articulated deformations via neural implicit surfaces, all these methods require ground-truth bone transformations as input, making them unable to interface with keypoints. In addition, it is uncertain and has not been demonstrated how these models will function under noisy predictions without the ground truth transformations. Note that making implicit representations applicable to articulated hands driven by skeleton is non-trivial. The key challenge is how to infer realistic shapes of unseen hands from the skeletons with highly articulated hand poses. Our solution includes two steps. First, by leveraging bio-mechanical constraints, we ensure bijective mappings between 3D skeletons and hand surfaces using the canonicalization layer, which greatly simplifies the learning of the implicit surface for highly articulated hand skeletons; Second, in the canonical space, we learn both identity-dependent and pose-dependent deformations of the hand surface with a set of multi-layer perceptrons that generalize well across different hand shapes and poses. We systematically compare HALO with several baseline methods. The results show that HALO significantly improves the accuracy, generality and visual fidelity over the baselines.

Hand grasps synthesis. To demonstrate the utility of the keypoint-driven implicit surface representation, we deploy HALO for conditional generation of hands grabbing 3D objects. We propose a novel generation pipeline that synthesizes hand keypoints and yields the neural occupancy of the hand via HALO. By exploiting the differentiable nature of HALO, we design a hand-object interpenetration loss to guide the training of the 3D keypoint generator in an end-to-end fashion. Our experiments show that this loss leads to convincing hand-object contact of the generated hands. Furthermore, compared to a MANO-based method (GrabNet [59]), HALO produces more physically plausible and visually convincing grasps even before refinement, suggesting that the volume-based losses are effective for learning the 3D keypoint generator.

Overall, the contributions include (1) HALO, the first neural implicit hand model that is driven by keypoint-based skeleton articulation, provides smooth neural implicit surfaces, and enables differentiable reasoning about volumetric occupancy; (2) A differentiable skeleton canonicalization layer that maps any skeleton to the canonical pose with a unique bio-mechanically valid transformations; (3) A realistic human grasp generation framework that leverages the efficient volumetric occupancy checks enabled by HALO.

2. Related Work

Hand pose and mesh estimation. Hand pose estimation is a long standing question and several learning-based approaches have been introduced. These approaches generally involve predicting 3D key point locations [8, 13, 16, 18, 19, 28, 40, 41, 43, 54, 57, 58, 60, 64, 66, 71], regressing MANO [51] parameters [1, 2, 4, 25, 26, 68], or directly predicting the full dense surface of the hand [20, 30, 39, 61]. The methods that directly predict 3D key points usually achieve better performance, however, they do not yield dense surface which is crucial for hand interaction. Iteratively fitting a templated-based model such as MANO to the key points could recover the dense surface but also make the process non-differentiable [44, 46, 62]. Alternatively, the dense surface can also be estimated from 3D or 2D keypoints [11, 65, 70]. However, such estimation could result in a change of hand pose from the input keypoints. In contrast, our model produce hand surface that faithfully respects the input pose and allows surface or volumetric losses to back-propagate directly to the keypoints.

Hand representation. The surface of 3D hands can be represented explicitly or implicitly. The most common template-based approaches such as MANO [51] induce a prior of poses and shapes over its learned parameter space

for regularization. However, using the learned parameters also increases the learning complexity as these features do not correspond directly to features in the inputs such as visible joints. In [1, 4, 25, 26, 68], the MANO parameters are predicted directly using additional weak supervision such as hand masks [1, 68] or 2D annotations [4, 68]. Another way of representing explicit 3D hand is to directly store the dense vertex locations of the MANO template [20, 30, 38]. While being more generalizable by avoiding constraints on the parameter space, these approaches require the corresponding, dense 3D annotations, which might be difficult to acquire. Our work differs in that we can recover dense hand surfaces from 3D keypoints, eliminating the need for learning model parameters or predicting dense surface points.

Implicit representation. Several works represent object shapes by learning an implicit function using neural networks [10, 14, 21, 22, 35, 37, 47, 55], which allows for the modeling of arbitrary object topologies with dynamic resolution. Many approaches for learning such implicit function from various input types were also proposed [32, 33, 45, 48, 52, 56]. These works focus on rigid objects and do not permit shape deformation. Recently, the interest is also on learning an articulated implicit function for human body [3, 15, 27, 36, 63]. NASA [15] represents human bodies using a set of implicit functions, but the model is limited to a specific body shape. LEAP [36] proposes to learn inverse linear blend skinning functions for multiple body shapes, however, it relies on ground truth bone transformation matrices instead of 3D joint locations. To the best of our knowledge, there are no implicit hand representations that can generalize well to various shapes. Grasping Field [29] learns an implicit function for hand and objects together to represent contact but treats every posed hand as a rigid object. As a result, the complexity of learning a wide range of poses increases significantly. In this work, we leverages bio-mechanical constraints of human hand to learn a novel hand model that only takes a skeleton as input and generalizes to different hand shapes and poses.

Hand-object interaction. There has been many studies into hand interacting with object in various settings [5, 7, 9, 12, 17, 18, 23, 24, 29, 31, 42, 59]. Recently, the community has begun exploring the task of generating plausible hand grasps given an object with notable studies including [12], [29], and [59]. GanHand [12] generates grasps suitable for each object in a given RGB image by predicting a grasp type from grasp taxonomy [17] and its initial orientation, then optimize for a better contact with the object. GrabNet [59] uses Basis Point Set [49] to represent 3D objects as input to generate MANO parameters. The predicted hand is then fed to a refinement model to improve the contact. Grasping Field [29] learns a signed distance field for both hand surface and object surface in one space, allowing the contact to be learned as regions where distances to both surfaces are zeros. However, the output surface cannot be articulated and requires hand model fitting. Our work differs from others in the way that we use the proposed hand representation to model the contact while keeping the synthesis task as simple as generating 3D keypoints.

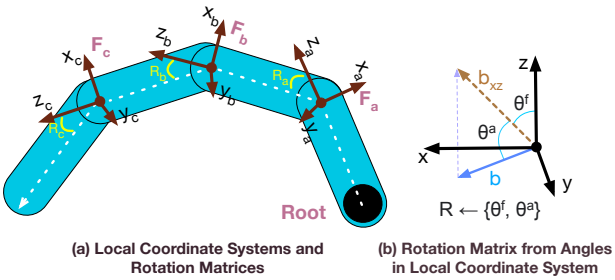

Figure 2. **Local coordinate systems and rotation matrices defined in the systems.** (a) We adopt the technique from [57] to construct local coordinate systems $\boldsymbol{F}_i$ for each segment of the kinematic chains of the hand. (b) Each $\boldsymbol{F}_i$ is constructed by measuring the flexion angle θ^f and the abduction angle θ^a relative to the parent bone. See Supp. Mat. for further details.

3. HALO: Hand ArticuLated Occupancy

The HALO model is a skeleton-driven neural occupancy function, formally defined as $\mathcal{O}_w(x|\boldsymbol{J}) \rightarrow \{0, 1\}$. Parameterized by neural network weights w, it maps a 3D point x to its occupancy value given the hand skeleton represented by a set of 3D keypoint locations $\boldsymbol{J}$. In this section, we first describe how to convert an arbitrary 3D joint skeleton to the reference canonical pose in a differentiable and consistent manner, then we introduce our simple yet effective neural occupancy networks for hands.

Notations. Given a hand skeleton represented by 3D key points $\boldsymbol{J} : \mathbb{R}^{21\times3}$, we denote θ_i^f and θ_i^a to be the flexion/extension and abduction/adduction angles of bone i relative to its parent, respectively. For simplicity, we refer to them as flexion and abduction angle. The angle between a palmar bone i and its adjacent palmar bone $i+1$ is denoted as θ_i^p. Lastly, θ_i^n is the palmar plane angle between plane n_i and n_{i+1} spanned by the palmar bone $i-1, i, i+1$ respectively. We denote the properties of the reference canonical hand with ${}^c(.)$. We refer to Supp. Mat. for further details.

3.1. Canonicalization of 3D Hand Skeleton

Our goal is to learn a neural representation of the surface of human hands in a canonical space. Furthermore, we want to deform this shape based on the spatial configuration of the underlying skeleton, represented by 3D keypoints. To do so, we require a mechanism that allows us to convert the 3D keypoints into valid skeletons in the canonical pose in terms of joint angles. As the keypoints have no notion of the surface, naively converting them to axis-angles does not

work due to the unconstrained twist of bones. While twist does not affect keypoints, they affect the surface.

We take inspiration from Spurr et al. [57] which defines a consistent local coordinate system for each bone to measure the bone angles for semi-supervised learning. Our objective is to derive a differentiable mapping layer that i) provides means to convert predicted keypoints to the rest pose and back, and ii) ensures that the skeleton is free of implausible twist that would influence the surface.

Building on [57], we represent each finger bone by two rotation angles, flexion and abduction, relative to its parent bone (Fig. 2). Each bone cannot rotate about itself, thus, no twist. However, such formulation ignores the palm configuration, which is needed for defining the canonical pose. In this work, we propose a method to parameterize the pose of a palm in order to define a consistent canonical pose. We decouple the palmar bone configuration into 1) finger spreading and 2) palm arching. The spreading of fingers is captured via the angles between two adjacent palmar bones. The arching of the palm is defined by the angle between the two planes spanned by three adjacent palmar bones. The resulting palmar region then serves as a frame of reference for the remaining fingers. Please refer to Supp. Mat. Fig. B.1 for better visualization.

Converting 3D Keypoints to Bone Transformations. Formally, we seek the unique set of transformations $\{\boldsymbol{B}^{-1}\}$ that maps the skeleton $\boldsymbol{J} : \mathbb{R}^{21\times 3}$ to the canonical pose ${}^{c}\boldsymbol{J}$. Given a skeleton, we obtain the set $\{\boldsymbol{B}^{-1}\}$ by sequentially performing the following operations: First, we rotate each finger to match the description of our canonical palm pose, which we define as a flat hand with fixed angles between palmar bones; Second, we compute joint angles and local coordinate systems following [57] (Fig. 2b), which we use to iteratively undo the angles along the kinematic chain (Fig. 2a) to acquire the canonical pose. By combining the transformations from both steps, then adjust for the conversion from keypoints to bone vectors, we could obtain a set of transformation matrices $\{\boldsymbol{B}^{-1}\}$ that maps the given skeleton $\boldsymbol{J}_i$ to our canonical pose ${}^{c}\boldsymbol{J}_i$. Formally,

$$\begin{aligned} {}^{c}\boldsymbol{J}_i &= \boldsymbol{B_i}^{-1}\boldsymbol{J_i} \quad , \\ \text{where } \boldsymbol{B}^{-1} &= \boldsymbol{TSF'R}^{c}\boldsymbol{FPK} \quad . \end{aligned} \tag{1}$$

Here $\boldsymbol{K} : \mathbb{R}^{3\times 21} \mapsto \mathbb{R}^{3\times 20}$ is a function that maps the keypoints $\boldsymbol{J} \in \mathbb{R}^{21\times 3}$ to bone vectors by translating them to the local origin and scaling to unit norm; $\boldsymbol{P}$ normalizes the palmar bone and palmar planes angles; $\boldsymbol{F}$ then maps the bones to their local coordinate frames; $\boldsymbol{R^c}$ rotates each bone to have the same flexion and abduction angles as the canonical pose. Finally, $\boldsymbol{F'}$ maps each coordinate frame back to the global coordinate system; $\boldsymbol{S}$ reverts bones back to their original length and $\boldsymbol{T}$ translates the bones to the tip of their parent bones.

This set of transformations is unique for each skeleton pose and only allows biomechanically valid transformation. For details, we kindly refer the reader to our Supp.Mat.

3.2. Neural Occupancy Networks for Hands

Here, we describe how to leverage the unique mapping $\{\boldsymbol{B}^{-1}\}$ between the posed skeleton and the canonical skeleton to learn the neural hand representation that generalizes to different shapes with highly articulated poses. We draw inspirations from NASA [15] and explore similar neural network structure due to its simplicity and efficacy.

NASA [15]. NASA learns an implicit representation of a human body $\mathcal{O}_w(x|\theta) \rightarrow \{0,1\}$, conditioned on the pose descriptor θ. Specifically, it defines the implicit surface for each body part separately. Let $\boldsymbol{B}_b^{-1}$ be the transformation to the canonical pose for bone b, NASA can be denoted by:

$$\mathcal{O}_w(x|\theta) = \max_b\{\bar{\mathcal{O}}_w^b(\boldsymbol{B}_b^{-1}x, \Pi_w^b[\{\boldsymbol{B}^{-1}t_0\}])\}. \tag{2}$$

where the pose descriptor θ is defined by a collection of transformation matrices $\{\boldsymbol{B}_b\}_{b=1}^{B}$, and the probability of $\mathcal{O}_w(x|\theta)$ is derived from the maximum occupancy probability across B child occupancy functions $\bar{\mathcal{O}}_w^b(\cdot,\cdot)$, where each represents the body part of the bone b. For a query point x, each child function $\bar{\mathcal{O}}_w^b(\cdot,\cdot)$ maps x to its local coordinate system by the transformation matrix $\boldsymbol{B}_b^{-1}$, so that the local shape of each body part can be learned. The term $\Pi_w^b[\cdot]$ is used to provide global pose information to each child function. Essentially, by querying the occupancy value x using $\boldsymbol{B}_b^{-1}x$, the NASA model learns a template shape and the correction based on the global pose with $\Pi_w^b[\cdot]$. Note that, the bone transformation $\{\boldsymbol{B}_b\}_{b=1}^{B}$ is assumed to be given. For more details, we kindly refer the reader to [15].

Neural Occupancy Networks for Hands. A naive adaptation of the NASA model for human hand results in erroneous surface reconstructions as shown in Sec. 5.1 (Fig. 5). In order to represent hands with highly articulated **poses** and diverse **shapes**, we propose to learn the child occupancy functions by conditioning on a shape descriptor β, effectively learning $\mathcal{O}_w(x|\theta,\beta)$. We assume that the identify-dependent deformations of the hand are highly correlated to the bones, hence, we propose the use of a collection of bone lengths as the shape descriptor. In particular, we propose a simple yet effective bone length encoder $f^b(\boldsymbol{D})$ that takes $\boldsymbol{D} = [d_1, d_2, \cdots, d_B] \in \mathbb{R}^B$ the bone length of individual bones as inputs. We emphasize that under the proposed formulation, we could learn the hand surface using only the key points $\boldsymbol{J}$, as the pose descriptor $\boldsymbol{B}_b^{-1}$ is derived from $\boldsymbol{J}$ by Eq. 1. Our final occupancy function is given by:

$$\mathcal{O}_w(x|\theta,\beta) = \max_b\{\bar{\mathcal{O}}_w^b(\boldsymbol{B}_b^{-1}x, \Pi_w^b[\{\boldsymbol{B}^{-1}t_0\}], f^b(\boldsymbol{D}))\} \tag{3}$$

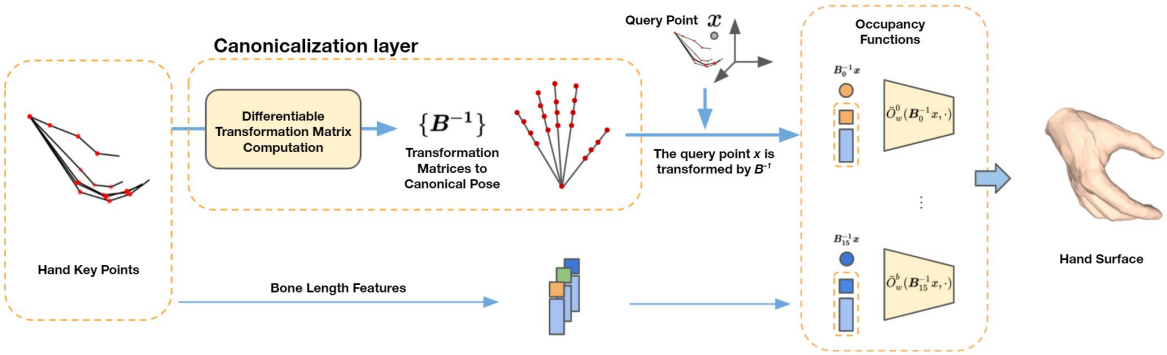

Figure 3. **Overview of HALO**. Given a hand skeleton, HALO derives bone transformations $\{\boldsymbol{B}^{-1}\}$ that map the bones to the canonical pose using the canonicalization layer. The query point x is then transformed into the canonical space as $\boldsymbol{B_i}^{-1}x$ for the occupancy check.

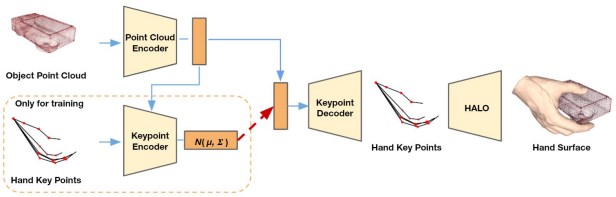

Figure 4. **HALO-VAE Architecture**. Given object point cloud, the VAE model synthesizes a grasping hand represented by a set of 3D keypoints. The input hand keypoints and the hand encoder are only used during the training. From the synthesized hand skeleton, HALO then produces the hand surface. The whole pipeline is end-to-end trainable, therefore, the VAE can leverage volume-based losses to improve the generation quality of the 3D keypoints.

where each implicit function $\bar{\mathcal{O}}_w^b$ learns the corresponding part shape based on the hand pose descriptor $\Pi_w^b[\{\boldsymbol{B}^{-1}t_0\}]$ and our bone length descriptor $f^b(\boldsymbol{D})$.

Shape descriptor variations. We investigate two versions of bone length encoders $f^b(\boldsymbol{D})$: the local bone encoder $f_l^b(\boldsymbol{D}) = d_b$ and the global encoder $f_g^b(\boldsymbol{D}) = [d_b; MLP(\boldsymbol{D})]$, where the MLP for the global encoder has two linear layers. We follow a similar training strategy as [15], for more training details, please refer to the Supp. Mat.

3.3. Skeleton-driven Articulated Hand Model

To build a skeleton-driven articulated hand model, we combine the previously described canonicalization layer and the neural hand surface together. Specifically, HALO takes the input 3D keypoints to compute bone transformations $\{\boldsymbol{B}^{-1}\}$ for the occupancy networks using the canonicalization layer. As the canonicalization layer is differentiable, the model can be trained end to end and allows volume-based losses from the surface to back-propagate to the keypoints. The overview of HALO is shown in Fig. 3. Note that the bone lengths $\boldsymbol{D}$ can also be computed from the keypoints. During inference, only 3D keypoints are needed as input to reconstruct hand surface.

4. Human Grasps Generation

We show the applicability of the HALO model in the challenging task of grasps generation. Given an object, we aim to generate diverse grasps with natural and plausible hand-object interaction. Our grasp generation pipeline consists of two parts: a 3D keypoints generator based on a variational autoencoder (VAE) and the HALO model for obtaining the hand surface.

HALO-VAE Architecture. The architecture of the HALO-VAE model is illustrated in Fig. 4. During training, the object point cloud is first passed to the object encoder, which is a modification of PointNet [50] with residual connection [35], to obtained an object latent code. The object latent code is then concatenated to the 3D hand joint location, $\boldsymbol{J} \in \mathbb{R}^{21\times3}$, and passed to the VAE encoder. The decoder reconstructs the 3D hand joint positions conditioned on the hand and object latent representation. From the key points, the surface is obtained using HALO through the skeleton canonicalization layer.

The advantages of using HALO are two-fold. First, we decouple the complexity of learning the pose, represented by the skeleton, from that of learning the surface that corresponds to the pose; Second, the implicit model enables fast intersection tests between hand and object, which can be used to efficiently compute an interpenetration loss. Combined with the differentiable skeleton canonicalization layer, the interpenetration loss can be used to improve the keypoints generator in both end-to-end training and post-optimization refinement.

Our grasp generation pipeline is similar to [59] and [29], but with the following key differences. First, in [29], the output is a rigid implicit surface that cannot be articulated. To obtain an animatable hand for downstream tasks, additional MANO model fitting is required. Second, in [59], the grasps generator is trained to produce the MANO parameters which is not directly related to the Euclidean space where the hand and the object live in. The challenge of in-

terfacing the MANO parameters with deep neural networks is reflected in the GrabNet (CoarseNet) [59] results which will be discussed in the experiments section.

4.1. Learning and Losses

To train the VAE model, we use the following losses: the KL-divergence loss on the hand latent Z, L2 loss on the predicted key points, L1 bone lengths loss, and the bone angle losses. The bone angle losses are used to provide additional supervisions for learning the hand structure which consists of 1) flexion angles θ_i^f and abduction angles θ_i^a, 2) angles between adjacent palmar bones θ_i^p and 3) angles between palmar planes θ_i^n. The bone angles are the same as used in Sec. 3. The losses are defined as L1 angle difference between the prediction and the ground truth.

Interpenetration loss. In addition to the losses on the keypoints, we also use the interpenetration loss on the hand surface to avoid collision between hand and object. The key idea is to penalize every points inside the object that is also occupied by the hand. Concretely, for a set of points sampled inside the object P_o and the predicted key points $\bar{\boldsymbol{J}}$, the interpenetration loss is defined as:

$$\mathcal{L}_{in}(\bar{\boldsymbol{J}}) = \sum_{p \in P_o} \mathcal{O}_w(p|\Theta(\bar{\boldsymbol{J}}), f_g(D_{\bar{\boldsymbol{J}}})), \text{where } \mathcal{O}_w(\cdot) > 0.5 \tag{4}$$

where $D_{\bar{\boldsymbol{J}}}$ is the bone length vector for $\bar{\boldsymbol{J}}$ and $\Theta(\bar{\boldsymbol{J}})$ maps the predicted key points to the HALO pose vector θ using the differentiable transformation matrices in Eq. 1.

4.2. Optimization-based Refinement

To demonstrate that the efficient intersection tests enabled by HALO can be used for optimization, we refine the sampled hands by changing the global translation t to avoid collision with the object. The refinement is run for 10 steps with the interpenetration loss term in Eq. 4. The optimization objective is:

$$\min_t \mathcal{L}_{in}(\bar{\boldsymbol{J}} + t) \tag{5}$$

This simple optimization step aims at refining the contact after the initial prediction of HALO-VAE. It is analogous to the RefineNet in [59], but with an explicit objective to avoid collision instead of being a neural network denoiser.

5. Experiments

In this section, we assess our skeleton-driven hand model and the grasp synthesis pipeline. First, in Sec. 5.1, we validate the efficacy of HALO as a neural implicit hand model and compare it to the surface baseline [15] and keypoints-to-surface baselines [11, 70]. Second, we show in Sec. 5.2 that HALO can be used effectively in generative tasks which require surface-based reasoning in form of grasp synthesis. For more experiments, please see supplementary materials.

	IoU ↑	Cham. (mm) ↓	Norm. ↑
NASA [15]	0.896	1.057	0.955
NASA+surf.	0.883	1.177	0.944
NASA+surf.+local b.	0.913	0.884	0.950
HALO (ours)	**0.932**	**0.719**	**0.959**
HALO keypoints (ours)	0.930	0.740	**0.959**

Table 1. **Evaluation on IoU, Chamfer-distance (L1), and normal consistency score (Norm.) between NASA [15] and HALO**. All models are trained using groundtruth bone transformations except for *HALO keypoints* where only the 3D key points are given. *NASA+surf.* indicates a NASA model with resampled surface points for the skinning loss $\mathcal{L}_s$ and *local. b* indicates that a corresponding bone length is given to each occupancy function. The results show that HALO outperforms NASA [15] on IoU and Chamfer-L1 by large margins, suggesting the superior performance of HALO in terms of fidelity and generality.

Methods	IOU ↑	Cham. (mm) ↓	MPJPE (mm) ↓
Choi et al. [11]	0.43	4.651	14.1
Zhou et al. [70]	0.54	2.811	7.95
HALO keypoints (ours)	**0.93**	**0.740**	**0**

Table 2. **Comparison between the estimated, root-aligned surfaces when only 3D keypoints are given as input.**

5.1. Neural Hand Model

We first evaluate the performance of the proposed implicit surface representation and analyze the effect of the keypoint-to-transformation mapping layer.

Training data. To train our neural occupancy hand model, we utilize MANO [51] hand meshes. Following [15], for each mesh we sample points with two strategies: 1) uniformly sampling in the hand bounding box, 2) sampling on the surface with additional isotropic Gaussian noise. Only the uniformly sampled points are used for evaluation. The associated occupancy value of each query point is computed by casting a ray from the sampled point and counting the number of intersections along the ray. The ground truth bone transformation matrices are computed along the kinematic chain to transform the template MANO hand into the target pose. The skinning weights are taken from the skinning weights of MANO. We use the Youtube3D (YT3D) hands dataset [30] in all our experiments. The YT3D training set contains 50,175 hand meshes of hundreds of subjects performing a wide variety of tasks in 102 videos. The test set covers 1,525 meshes from 7 videos.

Evaluation metrics. For 3D surface reconstruction evaluation, we compute the mean Intersection over Union (IoU), Chamfer-L1 distance, and normal consistency score [35].

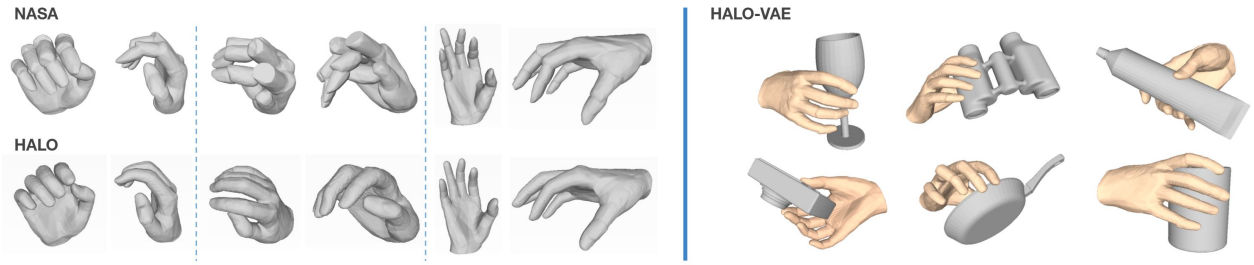

Figure 5. **Qualitative results of HALO and HALO-VAE.** (Left) Comparison between the NASA and HALO. Two views are shown for the same pose for comparison. (Right) Hands sampled from HALO-VAE.

	HALO-VAE w/o $\mathcal{L}_{in}$	HALO-VAE (ours)	[59]-coarse	HALO-VAE + Optim	[59]-refine
Int. vol. (cm3) ↓ / Cont. ra. (%) ↑					
Binoculars	9.19 / 1.00	**6.35** / 1.00	8.97 / 0.95	4.47 / 1.00	**3.24** / 1.00
Camera	3.99 / 1.00	4.47 / 1.00	**3.31** / 0.75	1.66 / 1.00	**1.46** / 1.00
Frying pan	0.25 / **0.85**	**0.22** / 0.65	0.94 / 0.80	**0.05** / 0.35	1.10 / **0.85**
Mug	**3.38** / 1.00	6.41 / 1.00	5.72 / 1.00	**3.48** / 1.00	4.48 / 1.00
Toothpaste	6.25 / 1.00	**1.85** / 1.00	6.05 / 1.00	**1.11** / 1.00	2.28 / 1.00
Wineglass	**1.41** / 0.90	2.35 / 1.00	2.62 / 1.00	**1.56** / 0.80	1.60 / **0.95**
Average	4.08 / **0.96**	**3.61** / 0.94	4.60 / 0.92	**2.06** / 0.85	2.36 / **0.97**
Diversity					
Entropy ↑	**2.88**	**2.88**	2.84	**2.88**	2.83
Cluster size ↑	**2.25**	2.15	1.65	**2.15**	1.70

Table 3. **Evaluation on interpenetration volume (Int. vol.), contact ratio (Cont. ra.) and diversity.** HALO-VAE w/o $\mathcal{L}_{in}$ denotes the HALO-VAE model *without* the interpenetration loss. Best numbers are in bold, except for the contact ratios when more than one model have the same best result.

% users rated as 'more realistic'	
HALO-VAE w/o $\mathcal{L}_{in}$	HALO-VAE
32.77 %	**67.23** %
HALO-VAE	GrabNet - coarse
55.75 %	44.25 %
HALO-VAE + Optim	GrabNet - refine
51.29 %	48.71 %

Table 4. **Binary choice user study.** The number show the percentage of users who rate the corresponding method as more realistic.

5.1.1 Comparison to implicit surface baseline

Here we investigate the generalization ability of the proposed HALO model to represent articulated hands with various **poses** and **shapes**. The results are summarized in Tab. 1

Baseline. We use the NASA model [15] as our baseline. The NASA model is designed to represent an implicit function of an articulated body. However, by changing the input dimension and the number of part-models to match the number of hand parts, it can also be used to represent an articulated hand. We trained the baseline model using the bone transformation matrices taken from MANO and the sampled query points. For details on implementation and network architecture we refer to the supplementary.

Surface vertex re-sampling. In [15], the surface vertices v used for enforcing the part models in the skinning loss $\mathcal{L}_s$ are the mesh vertices of SMPL [34]. Similarly, we use MANO surface vertices during training. However, we notice that the human-designed mesh often has many more vertices in the area around the joints which could cause the part models to bias toward the bone endpoints. Thus, we propose to re-sample the surface vertices uniformly on the mesh surface. This result is performance degradation but the bone connections are more natural with less artifact.

Local and global bone encoders. The bone lengths of a human hand greatly influence the hand shape. Therefore, for the local bone encoder, we add the bone length d_i to the back-projected query point as input to the part model $[\boldsymbol{B}_i^{-1}x; d_i]$. As shown in Tab. 1, the local bone encoder improves the reconstruction quality both in terms of IoU and Chamfer-L1 distance. We further extend the local bone encoder by considering all the bone lengths as input. A concatenated vector of bone lengths is first fed into a small feed-forward neural network to get the global bone feature $f_g(\boldsymbol{D})$, which is then concatenated with the query point x and the local bone length d_i as input to the part model.

Results. By combining the local and global bone encoders, HALO significantly improves the 3D surface reconstruction quality compared to NASA. As shown in Tab. 1, the IoU is increased from 0.896 to 0.932 and the Chamfer-L1 distance is decreased from $1.057mm$ to $0.719mm$.

We provide a qualitative comparison between NASA, and HALO in Fig. 5, confirming the quantitative results. The proposed HALO representation generalizes well for highly articulated poses, whereas the NASA model produces severe artifacts at the connection between parts.

5.1.2 3D keypoints to hand surface

Tab. 1 also shows the result from HALO that only takes 3D keypoints as input. The keypoint model achieves comparable surface reconstruction performance as when the ground truth transformation are given, showing the effectiveness of our method. We show the qualitative results in Fig. 1 and 5.

In addition, to evaluate the keypoint-to-surface pipeline, we then compare HALO to the equivalent component in [70] and [11] which estimates hand surface from 3D keypoints. The evaluation is done on same the Youtube3D test set where the ground truth 3D keypoints are given as input. As [11] requires both 2D and 3D coordinates as inputs, the 2D keypoints is obtain by projecting the 3D keypoints perpendicular to the palm. For evaluation, we also report the 3D joint error between the predicted hand and the input joints. This metric measures if the input keyoints are faithfully respected by the models. By design, the HALO model does not change the keypoint locations from input to output, thus does not have this error. The comparison in Tab. 2 shows that [70] and [11] change the hand pose and shape in the prediction while HALO faithfully reconstructs the hand surface according to the given keypoints.

5.2. Grasp Synthesis

To assess the utility of HALO in downstream tasks we demonstrate our grasp generative model, HALO-VAE.

Dataset. We leverage the recently introduced GRAB dataset [6, 59] and compare our results to GrabNet [59]. We compare both to the initial (coarse) predictions of GrabNet [59] and the refined results which matches with our own two-stage generation process. The test set contains 6 unseen objects. For each object, we fix the object orientation and sample 20 hand proposals from each model.

Physics Metrics. Following [67, 69], we evaluate the physical plausibility (interpenetration volume and contact ratio) and diversity, and provide results from a perceptual study. To evaluate the interpenetration and contact, we measure the ratio of frames in which the hand is in contact with the object and average the interpenetration volume. The volume is calculated by voxelizing hand and object mesh with 1mm cubes and counting the number of intersecting cubes.

User study. We asked 75 participants in a forced-alternative-choice perceptual study to 'select the grasp that is more realistic'. For each question, the user is shown 4 views per grasp and forced to select one. We compare all possible combinations on the same object. Each question is assigned to at least 2 participants, totaling 4,800 data points per pair of model comparison. To ensure that the grasps from HALO-VAE and GrabNet have the exact same texture, we fit MANO to our generated key points for rendering.

Diversity. Following [69], we compute the diversity of the sampled grasps by performing k-means with 20 clusters on all samples, then evaluate the entropy of the cluster assignment and the average cluster size. More diversity results in higher value for both metrics. We use the flatten key point locations of the hands after aligning the root joint and the plane spanned by middle and index palmar bone as features.

Results. We first validate the efficacy of the interpenetration loss (Eq. 4). We compare the HALO-VAE models with and without the interpenetration loss. The results show that the interpenetration loss helps in: 1) reducing the collision between the objects and the generated hands (Tab. 3, col.1-2), and 2) largely improves the user preference of the corresponding model (Tab. 4, first row), demonstrating the efficacy of the proposed neural occupancy representation of articulated hand for reasoning about hand-object interaction.

Next, we compare HALO-VAE with GrabNet [59]. Both HALO-VAE and GrabNet-coarse are CVAE based generative models and end-to-end trainable, the key difference is that GrabNet-coarse generates MANO model parameters whereas HALO-VAE generates 3D keypoints. As shown in Tab. 3, HALO-VAE outperforms GrabNet-coarse by a large margin for interpenetration volume and sample diversity. Moreover, the HALO-VAE model without interpenetration also compares favorably to GrabNet-coarse, suggesting that the 3D keypoints based representation is well suited to interface with deep neural networks.

Finally, we compare our optimization-based refinement with GrabNet-refine. To the best of our knowledge, the RefineNet is not trained end-to-end with GrabNet-coarse and used for three steps during the inference. As shown in Tab. 4, our refined grasps attain a higher user score, suggesting they are more realistic and natural compared to the grasps refined by GrabNet-refine.

6. Discussion and Conclusion

In this work, we introduce HALO, a novel surface representation for articulated hands that can generalize to different hand poses and shapes. We address the issue of the transformation matrix requirement for inferring the 3D occupancy hand by proposing a skeleton canonicalization algorithm that computes valid transformations from 3D keypoints. The experiments show that our proposed hand model outperforms the baseline and can represent a wide range of hand poses and shapes. Finally, we demonstrate the HALO can be used to train an end-to-end grasp generator conditioned on an object and produces hand grasps with natural and realistic interaction. We believe that HALO can be useful in future work attempting to reconstruct the surface of articulated hands directly from images via differentiable rendering and for several downstream tasks that need to perform surface-based computation such as collision detection and response.

7. Acknowledgement

We sincerely acknowledge Shaofei Wang and Marko Mihajlovic insightful discussions and help with baselines.

References

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019. 2, 3

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Weakly-supervised domain adaptation via gan and mesh model for estimating 3d hand poses interacting objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6121–6131, 2020. 2

[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. *Advances in Neural Information Processing Systems*, 33, 2020. 3

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10843–10852, 2019. 2, 3

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8709–8719, 2019. 3

[6] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 8, 6

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *The European Conference on Computer Vision (ECCV)*, August 2020. 3

[8] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 666–682, 2018. 2

[9] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. DexYCB: A benchmark for capturing hand grasping of objects. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[11] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 6, 8

[12] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[13] Martin de La Gorce, David J Fleet, and Nikos Para-

gios. Model-based 3d hand pose estimation from monocular video. *IEEE transactions on pattern analysis and machine intelligence*, 33(9):1793–1805, 2011. 2

[14] Boyang Deng, Kyle Genova, Soroosh Yazdani, Sofien Bouaziz, Geoffrey Hinton, and Andrea Tagliasacchi. Cvxnet: Learnable convex decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–44, 2020. 3

[15] Boyang Deng, JP Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Neural articulated shape approximation. *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4, 5, 6, 7

[16] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6608–6617, 2020. 2

[17] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *IEEE Transactions on human-machine systems*, 46(1):66–77, 2015. 3

[18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 409–419, 2018. 2, 3

[19] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. Robust 3d hand pose estimation in single depth images: from single-view cnn to multi-view cnns. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3593–3601, 2016. 1, 2

[20] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10833–10842, 2019. 2, 3

[21] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[22] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7154–7164, 2019. 3

[23] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1471–1481, 2021. 3

[24] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3196–3206, 2020. 3, 1

[25] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 571–580, 2020. 1, 2, 3

[26] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11807–11816, 2019. 2, 3, 1

[27] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3093–3102, 2020. 3

[28] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018. 1, 2

[29] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *arXiv preprint arXiv:2008.04451*, 2020. 3, 5, 1, 2

[30] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M. Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 3, 6

[31] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14687–14697, 2021. 3

[32] Shichen Liu, Shunsuke Saito, Weikai Chen, and Hao Li. Learning to infer implicit surfaces without 3d supervision. In *Advances in Neural Information Processing Systems*, pages 8295–8306, 2019. 3

[33] Shaohui Liu, Yinda Zhang, Songyou Peng, Boxin Shi, Marc Pollefeys, and Zhaopeng Cui. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2019–2028, 2020. 3

[34] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, Oct. 2015. 7

[35] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4460–4470, 2019. 3, 5, 6

[36] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10461–10471, 2021. 2, 3

[37] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view syn-

thesis. In *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 3

[38] 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 *European Conference on Computer Vision (ECCV)*, 2020. 3

[39] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[40] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE conference on computer vision and pattern Recognition*, pages 5079–5088, 2018. 1, 2

[41] Gyeongsik Moon, Shoou-I Yu, He Wen, Takaaki Shiratori, and Kyoung Mu Lee. Interhand2.6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 7

[42] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019. 3

[43] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018. 2

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

[45] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. 3

[46] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018. 2

[47] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 165–174, 2019. 3

[48] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. *arXiv preprint arXiv:2003.04618*, 2020. 3

[49] Sergey Prokudin, Christoph Lassner, and Javier Romero. Efficient learning on point clouds with basis point sets. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 1

[50] Charles Ruizhongtai Qi, Hao Su, Kaichun Mo, and Leonidas J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. *CoRR*, abs/1612.00593, 2016. 5, 1

[51] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), 2017. 1, 2, 6, 5

[52] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 2304–2314, 2019. 3

[53] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J. Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, June 2021. 2

[54] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2

[55] 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, 2020. 3

[56] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, pages 1121–1132, 2019. 3

[57] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 4

[58] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 89–98, 2018. 2

[59] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 3, 5, 6, 7, 8

[60] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4511–4520, 2019. 2

[61] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Dual grid net: Hand mesh vertex regression from single depth maps. In *European Conference on Computer Vision*, pages 442–459. Springer, 2020. 2

[62] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A. Otaduy, Dan Casas, and Christian Theobalt. RGB2Hands: Real-Time Tracking of 3D Hand Interactions from Monocular RGB Video. *ACM Transactions on Graphics (TOG)*, 39(6), 12 2020. 2

[63] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7639–7648, 2021. 3

[64] Ying Wu, John Lin, and Thomas S Huang. Analyzing and capturing articulated hand motion in image sequences. *IEEE transactions on pattern analysis and machine intelligence*, 27(12):1910–1922, 2005. 2

[65] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In *BMVC*, 2020. 2

[66] Linlin Yang and Angela Yao. Disentangling latent hands for image synthesis and pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9877–9886, 2019. 2

[67] Siwei Zhang, Yan Zhang, Qianli Ma, Michael J Black, and Siyu Tang. Place: Proximity learning of articulation and contact in 3d environments. In *8th international conference on 3D Vision (3DV 2020)(virtual)*, 2020. 8

[68] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2354–2364, 2019. 2, 3

[69] Yan Zhang, Mohamed Hassan, Heiko Neumann, Michael J Black, and Siyu Tang. Generating 3d people in scenes without people. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6194–6204, 2020. 8

[70] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 6, 8

[71] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *Proceedings of the IEEE international conference on computer vision*, pages 4903–4911, 2017. 1, 2

A Skeleton-Driven Neural Occupancy Representation for Articulated Hands **Appendix**

A. More Experimental Analysis

A.1. MANO Parameters vs 3D Keypoints

To demonstrate the applicability of the HALO hand model, we introduce the HALO-VAE model (Sec. 4) for the conditional human grasps generation task. Our model learns to generate the 3D keypoints of a hand that grasps a given object, whereas our baseline model GrabNet [59] generates MANO parameters that represent the grasping hand. As shown in the Sec. 5.2, the proposed HALO-VAE model largely outperforms the GrabNet in terms of physical plausibility and naturalness of the generated human grasps.

Here, we provide more details and analysis on the object encoding schemes. The GrabNet [59] encodes the object using the BPS features [49]. Specifically, the BPS encoder is a 4-layers feed-forward network with residual connections between each layer. In our trials, we have experimented with this BPS encoder instead of our PointNet encoder. We used the same 4096 basis points as provided in the GRAB dataset. The rest of the architecture is the same to HALO-VAE.

However, with the near-zero keypoint reconstruction loss on the validation set, the generated grasps for a given object are always the same, suggesting that the information from the sampled Gaussian is not used. We suspect that during training, the hand keypoints can be inferred using only the object BPS, as the 3D keypoints and the BPS features are highly related. Consequently, the decoder can entirely ignore the features from the hand encoder, producing the same grasp for different samples.

In order to directly compare the keypoint-based and the MANO parameter based grasps generation frameworks, here we use the same object encoding scheme that employs the PointNet architecture [50]. Specifically, we change the last layer of HALO-VAE (Fig. 4) from predicting the hand keypoints (63 dimensions, including 3×21 keypoints) to predicting MANO parameters (61 dimentions, including 3 global translation, 3 global rotation, 10 shape parameters and 45 pose parameters). Both models are trained without the interpenetration loss. The results are shown in Tab A.1. The keypoint-based generative model produces grasps with better contact and interpenetration while also being more diverse than those generated from the MANO parameters based model, demonstrating the efficacy of the proposed HALO-VAE model.

HALO-VAE w/o $\mathcal{L}_{in}$	Keypoint prediction	MANO parameter prediction
Interpenetration volume (cm3) ↓ / Contaxt ratio (%) ↑		
Binoculars	9.19 / **1.00**	**4.88** / 0.95
Camera	**3.99** / 1.00	4.00 / 1.00
Frying pan	0.25 / **0.85**	**0.21** / 0.50
Mug	**3.38** / 1.00	7.22 / 1.00
Toothpaste	**6.25** / 1.00	6.94 / 1.00
Wineglass	**1.41** / 0.90	3.35 / **1.00**
Average	**4.08 / 0.96**	4.43 / 0.91
Diversity		
Entropy ↑	**2.88**	2.85
Cluster size ↑	**2.25**	1.13

Table A.1. **Comparison between the HALO models predicting key points and predicting MANO parameters.** The results from the keypoint based model are the same as those reported in the main paper.

A.2. Comparison with Grasping Field [29] and GrabNet [59] on other datasets

In this section, we show the comparison between the generated grasps from the Grasping Field (GF) model, Grabnet, and HALO-VAE on the ObMan [26] and HO3D [24] test objects. Note that the HALO-VAE and Grasping Field are not directly comparable as the meshes produced by Grasping Field do not guarantee to be a valid human hand and require MANO fitting, while our HALO-VAE produces articulated implicit hand surfaces.

Nevertheless, we show qualitative and quantitative comparisons between the GF meshes after MANO fitting and the HALO hand surfaces in Fig. A.1 and Tab A.2, A.2, respectively. Due to the artifacts in the hand-designed objects used in

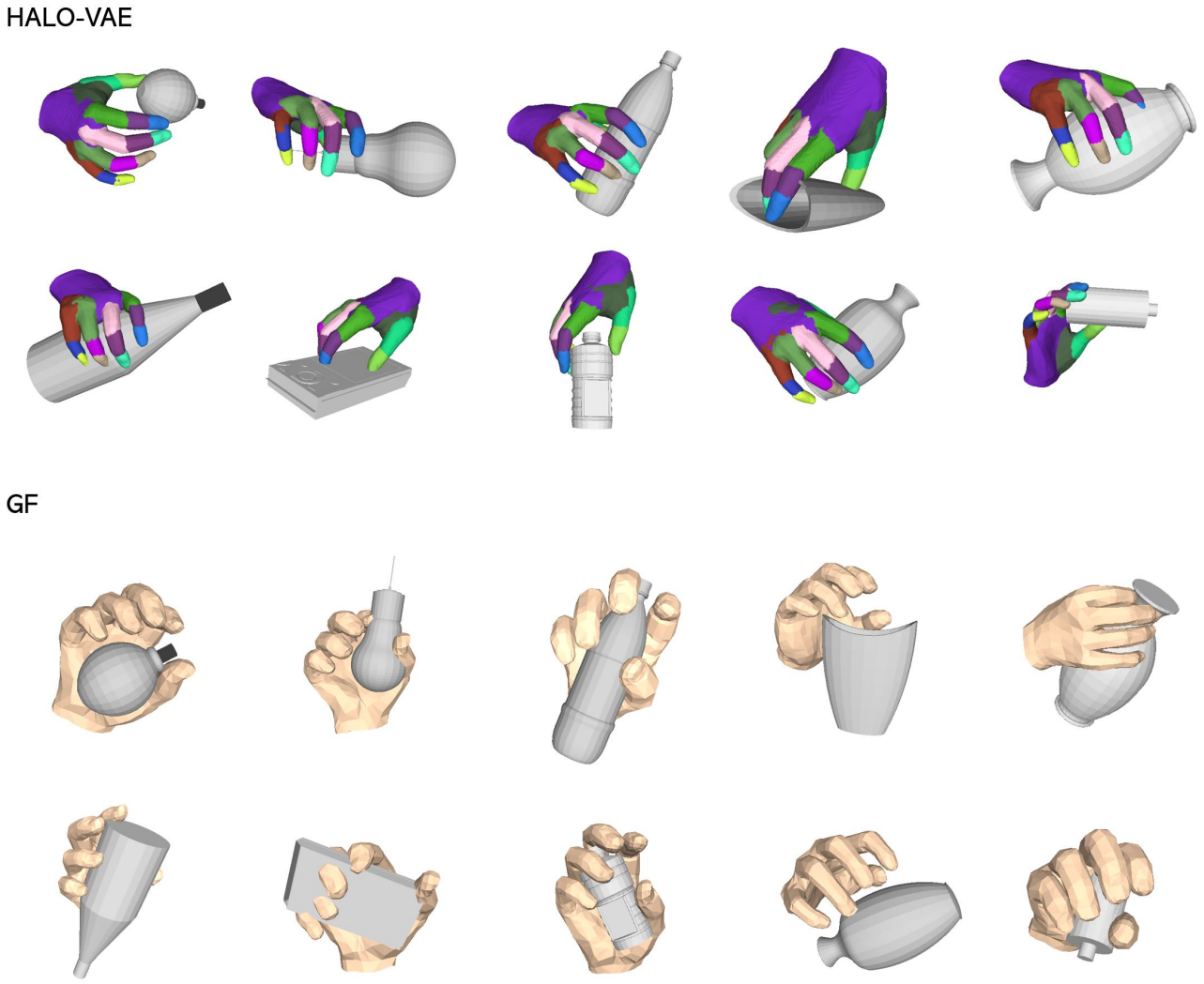

Figure A.1. **Qualitative comparison between HALO-VAE and Grasping Field [29] on ObMan test objects.** The colors indicate the hand parts based on the child occupancy functions. Our HALO-VAE shows more reasonable grasp and less interpenetration with the object.

the ObMan dataset that interfere with the interpenetration evaluation, e.g non-watertight meshes, surface with holes, internal structure with wrong winding number and zero-volume meshes, we perform the evaluation using the object convex hull instead. The evaluation is performed by generating 5 grasps per object on 30 randomly chosen test objects from the ObMan dataset. In total, we evaluate 150 generated grasps from each model.

The results in Tab. A.2 and A.3 show that our HALO-VAE model produces comparable physically-plausible human grasps than Grasping Field [29] and GrabNet [59] with more diversity.

B. Differentiable Bio-mechanical Canonicalization Layer

In this section, we elaborate on the method for converting 3D keypoints to bone transformation matrices. We closely follow the formulation in Spurr *et al*. [57] to construct the local coordinate systems $\boldsymbol{F}$. Here we provide a brief summary of the method. For more details on $\boldsymbol{F}$, we refer the readers to the supplementary material of [57].

Recall that we seek to compute the set of matrices $\boldsymbol{B}^{-1}$, which, in details, is obtained by sequentially performing the following operations: 1) normalizing palmar plane angles, 2) normalizing palmar bone angles, 3) constructing local coordinate frames $\boldsymbol{F}_i$ for each bone with respect to its parent along the kinematic chain, 4) undoing the rotation in the local frames $\boldsymbol{F}_i$,

	HALO-VAE w/ $\mathcal{L}_{in}$ (ours)	Grasping Field	GrabNet - Corase
Interpenetration volume* (cm3) ↓	**19.95**	21.93	21.82
Contact ratio* (%) ↑	0.98	0.90	**1.00**
Diversity			
Entropy ↑	2.76	2.83	**2.88**
Cluster size ↑	**3.34**	2.87	2.68

Table A.2. **Comparison with Grasping Field [29] and GrabNet [59] on randomly chosen ObMan test objects.** *Due to the non-watertight object meshes, the interpenetration volume and the contact ratio is approximated using the convex hull of the objects.

	HALO-VAE w/ $\mathcal{L}_{in}$ (ours)	Grasping Field	GrabNet - Corase
Interpenetration volume (cm3) ↓	25.84	93.01	**24.62**
Contact ratio (%) ↑	0.97	**1.00**	**1.00**
Diversity			
Entropy ↑	**2.81**	2.75	2.79
Cluster size ↑	**4.87**	3.44	3.23

Table A.3. **Comparison with Grasping Field [29] and GrabNet [59] on HO3D test objects.**

5) reverting back to the global coordinate frames. Formally,

$$
\begin{aligned}
{}^{c}\boldsymbol{J}_i &= \boldsymbol{B_i}^{-1}\boldsymbol{J_i} \quad , \\
\text{where } \boldsymbol{B}^{-1} &= \boldsymbol{TSF'R^cFPK} \quad .
\end{aligned}
$$

Here $\boldsymbol{K} : \mathbb{R}^{3\times 21} \mapsto \mathbb{R}^{3\times 20}$ is a function that maps the keypoints $\boldsymbol{J} \in \mathbb{R}^{21\times 3}$ to bone vectors by translating them to the local origin and scaling to unit norm; $\boldsymbol{P}$ normalizes the palmar bone and palmar planes angles; $\boldsymbol{F}$ then maps the bones to their local coordinate frames; $\boldsymbol{R^c}$ rotates each bone to have the same flexion and abduction angles as the canonical pose. Finally, $\boldsymbol{F'}$ maps each coordinate frame back to the global coordinate system; $\boldsymbol{S}$ reverts bones back to their original length and $\boldsymbol{T}$ translates the bones to the tip of their parent bones.

In the following, we define the notations needed and describe the methods for constructing the local coordinate system $\boldsymbol{F}_i$ and each matrix in $\boldsymbol{B}^{-1}$.

B.1. Notation

We define all the notations with respect to the right hand. The same procedure could also be applied to the left hand by flipping the x-axis of all joints without loss of generality. We denote 3D root-aligned joint locations of a posed hand as $\boldsymbol{J}$ where j_0 is the root joint. A bone is defined as a vector pointing from the parent joint to its child $\widetilde{b}_i = j_i - j_{p(i)}$ where $p(i)$ denotes the parent of joint i in the kinematic tree (see Fig. B.1). We define b_i as a normalized bone of $\widetilde{b}_i$ and call $\boldsymbol{K}$ the mapping from joints $\boldsymbol{J}$ to normalized bones. As a shorthand, we call the palmar bones that are connected to the root joint j_0 as the level-0 bones (bones $b_1, \cdots, b_5$). We call a bone with k bone segments in between itself and the root joint a k-level bone. The bone level from 0 to 3 are denoted by the color black, blue, dark purple, and orange respectively in Fig. B.1(b).

Palmar bone rotations. Given a hand skeleton in global coordinate frame, we denote θ_i^p to be the angle between a palmar bone i and its adjacent palmar bone $i+1$; θ_i^n to be the plane angle between plane n_i spanned by the palmar bone $i, i+1$ and plane n_{i+1} spanned by the palmar bone $i+1, i+2$. We denote the properties of the reference canonical hand with ${}^{c}(.)$

Non-palmar bone rotations. Given a local coordinate system $\boldsymbol{F}_i = (x_i, y_i, z_i)$ where (x_i, y_i, z_i) are the orientation of the coordinate system, for a bone b_i, here we define the flexion θ_i^f angle and the abduction θ_i^a angle. Since each bone b_i is defined the same way in a local coordinate system $\boldsymbol{F}_i$, we drop the subscript i for brevity. The flexion angle θ^f of a bone b is the angle between b_{xz} (the projection of b on the xz plane) and the z axis in a coordinate frame $\boldsymbol{F}$. The abduction angle θ^a is defined by the angle between the bone b and the projection b_{xz} (see Fig. B.2(b)).

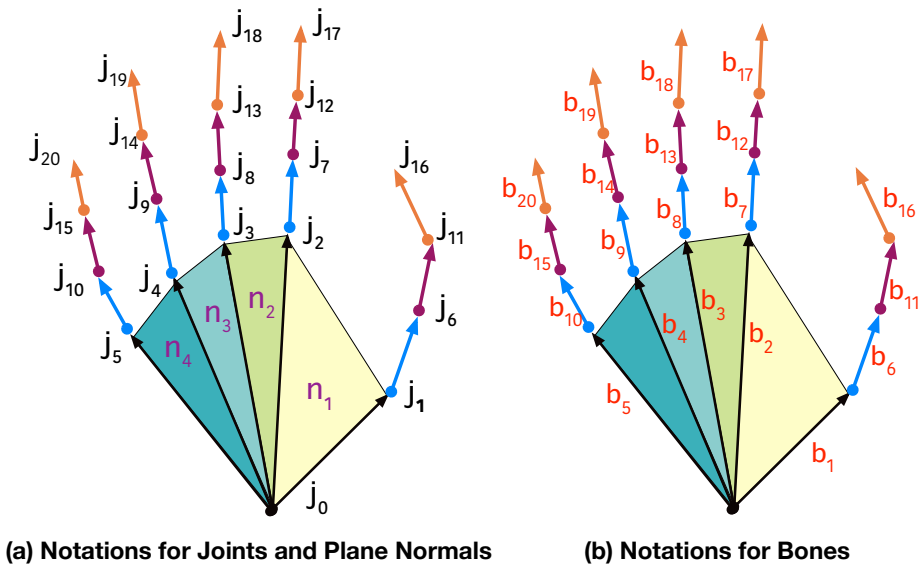

Figure B.1. Notations for joints, plane normals and bones for **right** hand facing up.

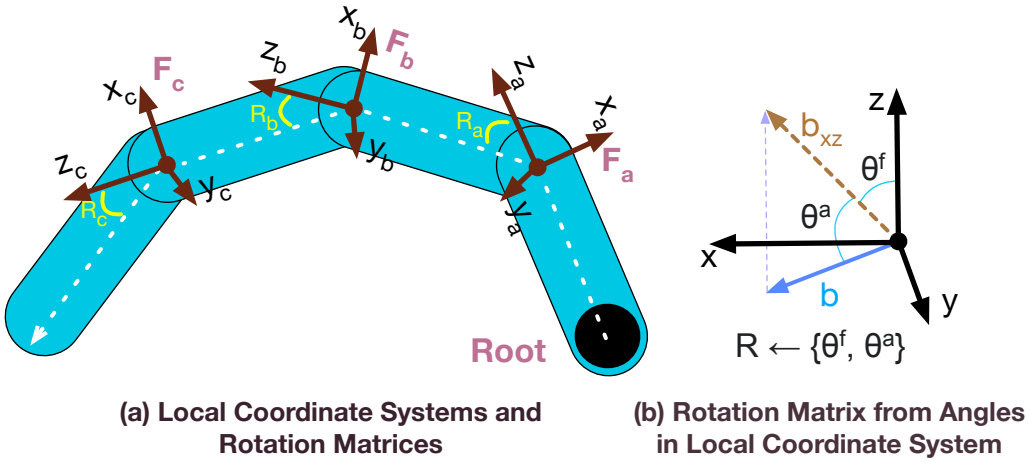

Figure B.2. **(Same as Fig. 2) Local coordinate systems and rotation matrices defined in the systems.** (a) The local coordinate systems $\boldsymbol{F}_a, \boldsymbol{F}_b, \boldsymbol{F}_c$ are constructed based on the kinematic chain. For brevity, we show the kinematic chain of one finger. (b) The flexion angle θ^f and the abduction angle θ^a that are constructed based on a bone b in the local coordinate system $\boldsymbol{F}$. The bone b_{xz} is a projection of b in the xz plane.

B.2. Palmar Bone Normalization ($\boldsymbol{P}$)

Given a globally normalized hand keypoints, we first compute the transformation matrices $\boldsymbol{P}$ the rotate the palmar bones to match the canonical pose. The palmar bone transformation normalization $\boldsymbol{P}$ is a combination of the palmar plane angle normalization $\boldsymbol{P^p}$ and the palmar bone angle normalization $\boldsymbol{P^a}$, with $\boldsymbol{P} = \boldsymbol{P^a}\boldsymbol{P^p}$.

Palmar Plane Angle ($\boldsymbol{P^p}$). To change the bone angle, we rotate the outer bone (with middle finger being the center) about the shared bone until the plane angle is equal to the canonical angle ${}^c\theta_i^p$, which we set to 0.8, 0.2, 0.2 radian for ${}^c\theta_1^p$, ${}^c\theta_2^p$, ${}^c\theta_3^p$, respectively. The plane between the index and middle finger (n_2) is fixed as reference. The rotation applied on the ring finger bone b_4 is also propagated to the pinky finger bone b_5.

Palmar Bone Angle ($\boldsymbol{P^a}$). Secondly, we normalize the spread of the fingers by rotating the bones on the plane two adjacent bones. Concretely, we use the middle finger as reference then rotate b_2 on plane n_2, b_1 on plane n_1, b_4 on plane n_3, and b_5 on plane n_4. The transformation applied on b_2 and b_4 are also applied to b_1 and b_5 respectively. We set the canonical angle to to 0.4, 0.2, 0.2, 0.2 radian for ${}^c\theta_1^n$, ${}^c\theta_2^n$, ${}^c\theta_3^n$, and ${}^c\theta_4^n$ respectively.

B.3. Local Coordinate System ($\boldsymbol{F}$)

Now we define the local coordinate system $\boldsymbol{F}_i$ for each bone b_i. Note that since in our formulation level-0 bones are always fixed, the only bones that characterize the hand pose are bones at level 1 to 3. Thus, we only describe the coordinate systems for non-zero-level bones. For each non-zero level bone b_i ($i > 5$), its coordinate frame $\boldsymbol{F}_i$ is defined by three normalized vectors x_i, y_i, z_i. To construct the coordinate system for level-1 to level-3 bones, the z-axis of $\boldsymbol{F}_i$ for non-zero level bones are always defined by the normalized bone vector of its parent $z_i = \text{norm}(b_{p(i)})$. We then describe how to define the x-axis. Afterwards, each local coordinate system is defined as y_i can be obtained by a cross product: $y_i = \text{norm}(z_i \times x_i)$. Note that the coordinate frame $\boldsymbol{F}_i$ does not have a position component because we obtain the bone vectors by subtracting the child joint. Thus, all bones are aligned to the origin. The translation components will be added in the final step of our formulation. However, for illustration purposes, we present each coordinate frame $\boldsymbol{F}_i$ with the translation in mind for our figures.

Coordinate Systems for Level-1 Bones Formally, we denote n_i as the normal of a plane spanned by two adjacent level-0 bones where

$$n_i = \text{norm}(b_{i+1} \times b_i), \text{ for } i \in \{1, 2, 3, 4\}. \tag{6}$$

For better illustration, Fig. B.1(a) demonstrates the normal vectors n_i defining each plane. Using the normal vectors above, we define the x-axis for level-1 frames $\boldsymbol{F}_i$ for $i \in \{6, 7, 8, 9, 10\}$ as follows:

$$\begin{aligned} x_6 &= -n_1 \\ x_7 &= -n_2 \\ x_8 &= -\text{norm}(n_2 + n_3) \\ x_9 &= -\text{norm}(n_3 + n_4) \\ x_{10} &= -n_4 \end{aligned} \tag{7}$$

In other words, for bone b_6 and b_{10}, we define the x-axis for their coordinate system $\boldsymbol{F}_6$ and $\boldsymbol{F}_{10}$ by $-n_1$ and $-n_4$ because they are on the edge of the palm. For bones b_8, and b_9, the x-axis for their coordinate systems are defined by the average normal around the bones.

Coordinate Systems for Level-2 and 3 Bones Given the rotation angles $\boldsymbol{R}_{p(i)}$ of the level-1 bones $b_{p(i)} (6 \leq p(i) \leq 10)$) and the corresponding coordinate systems $\boldsymbol{F}_{p(i)}$, we can construct coordinate frame $\boldsymbol{F}_i$ ($11 \leq i \leq 15$) for the second-level bones by rotating the coordinate frame $\boldsymbol{F}_{p(i)}$ along the kinematic chain using $\boldsymbol{R}_p(i)$. Concretely, the new coordinate frame is given by

$$\boldsymbol{F}_i = \boldsymbol{R}_{p(i)} \boldsymbol{F}_{p(i)}. \tag{8}$$

Similarly, the coordinate systems for level-3 bones can be obtained by rotating the coordinate systems of level 2 bones using the rotation angles on level 2.

Rotations to the Canonical Pose ($\boldsymbol{R}^c$) Given a local coordinate system $\boldsymbol{F}_i = (x_i, y_i, z_i)$ where (x_i, y_i, z_i) are the axis of the coordinate system, for a bone b_i, we can measure the flexion angle θ_i^f and the abduction angle θ_i^a that characterize bone i with respect to $\boldsymbol{F}_i$. Fig. 2 visualizes the local coordinate system $\boldsymbol{F}_i$ and how rotation angles are measured. Then, given a canonical pose bone b_i^c, we compute the rotation matrix $\boldsymbol{R}_i^c$ to transform b_i to its canonical pose b_i^c based on the angles relative to the canonical pose ($\bar{\theta}_i^f$, $\bar{\theta}_i^a$) in $\boldsymbol{F}_i$.

B.4. Angles to the Canonical Pose ($\bar{\theta}_i^f$, $\bar{\theta}_i^a$). ($\boldsymbol{R}^c$)

Since the angles are measured in a consistent coordinate frame, to obtain the angle needed to rotate to the canonical pose, we simply offset the angles of b_i by the angle of b_i^c in the canonical pose:

$$\begin{aligned} \bar{\theta}_i^f &:= -\theta_i^f + \theta_{i_c}^f \\ \bar{\theta}_i^a &:= -\theta_i^a + \theta_{i_c}^a \end{aligned} \tag{9}$$

where $\theta_{i_c}^f$ and $\theta_{i_c}^a$ are the flexion and abduction angle of the canonical pose with respect to $\boldsymbol{F}_i$. In our experiments, we use the canonical pose identical to that of MANO [51].

B.5. Local Coordinate Frame to Global Coordinate Frame($\boldsymbol{F'}$).

Each rotation matrix $\boldsymbol{R}_i^c$ is defined locally with respect to a coordinate frame $\boldsymbol{F}_i$. The accumulated rotations $\boldsymbol{G}_i$ with respect to the root of the hand for a bone b_i is then the product of the rotation matrices along the kinematic chain:

$$\boldsymbol{G_i} = \begin{cases} \boldsymbol{G}_{p(i)}\boldsymbol{R}_i^c & \text{if } b_i \text{ is not a root bone} \\ \boldsymbol{I}(4) & \text{otherwise.} \end{cases} \tag{10}$$

Where $\boldsymbol{G}_{p(i)}$ are the rotations up to the parent of b_i. With the accumulated rotation matrices encoding the joint-angles relative to the palm of the hand, we need to map all matrices to global coordinates by multiplying with $\boldsymbol{F}_i^{-1}$. Recall that $\boldsymbol{F}_i$ encodes the mapping from the global coordinate system to the local frame. Thus, its inverse brings the angles back to the global coordinate frame. We summarize the accumulation of angles and the mapping to the global coordinate with the matrix $\boldsymbol{F}_i'$:

$$\boldsymbol{F}_i' = \boldsymbol{F}_i^{-1}\boldsymbol{G}_{p(i)}. \tag{11}$$

With all the necessary components in place, we can compute the transformation matrices $\boldsymbol{B}^{-1}$ for all bones using only the 3D keypoints $\boldsymbol{J}$ in a differentiable manner. In Sec. 5.1.2 we show that the matrices derived using this formulation can be uses to recover hand surfaces that are very close to those attained by the ground truth transformation matrices from MANO.

C. Implementation details

C.1. HALO

Training loss. The loss used to train our articulated hand model can be written as:

$$\mathcal{L} = \mathcal{L}_o + \lambda_s \mathcal{L}_s \tag{12}$$

where λ determines the weight of the skinning loss. It is set to 0.5 in all experiments. We turn off the skinning loss after the validation IoU reach 80% as we observe it allows smoother transition between hand parts.

Network architecture. For a fair comparison with the baseline [15], we use the same network architecture with 4 layers of size 40 for each part model in the ablation study. For the final model used in HALO-VAE, the size is increased to 64 as we observed a small surface quality improvement. The LeakyRelu with factor 0.1 is used as the activation function. All layers have a residual connection and a dropout of 0.2. The subspace projection layers Π map the input of size $B \times 3$ to a vector of size 8. When used, the global bone length encoder is a 2-layer feed-forward network of size 40 that maps 16 bone lengths to an encoded vector of size 16.

We define the number of parts as $B = 16$, with one part responsible for the palm and three parts for each finger. When using the 20 transformation matrices obtained from our formulation as our pose descriptor, we disregard the transformations of the root bones, and add an identity transformation for the palm, resulting in 16 transformation matrices.

Training data. To train our neural occupancy hand model, we utilize MANO [51] hand meshes. The query points for each mesh are selected using two strategies: 1) uniformly sampling points in the bounding box of the hand mesh where the root joint is at the origin, 2) sampling on the surface with additional isotropic Gaussian noise. For each strategy, we sample 100,000 points. The associated occupancy value of each query point is computed by casting a ray from the sampled point and counting the number of intersections along the ray. For evaluation, following [15], we use uniformly sampled points. The bone transformation matrices are computed along the kinematic chain to transform the template hand into the target pose. The shape descriptor β is based on bone length, defined as the Euclidean distance between adjacent joints. The skinning weights are taken from the skinning weights for posing the template mesh in MANO. We use the Youtube3D (YT3D) hands dataset [30] in all our experiments. The YT3D training set contains 50,175 hand meshes of hundreds of subjects performing a wide variety of tasks in 102 videos. The test set covers 1,525 meshes from 7 videos.

Training. We used the Adam optimizer with a learning rate $1e-4$ and a batch size of 64 in all experiments. For each mesh at each training step, we sample 2048 points from the 200K pre-sampled query points for the occupancy loss and 2,000 out of 6,000 surface points for the skinning loss. When surface point re-sampling is not used, we sampled 200 out of 778 mesh vertices for the skinning loss.

C.2. HALO-VAE

Pre-processing. For training, we use GRAB dataset [6, 59] with the default train/test split. The objects are centered at the origin and 600 points are sampled from the surface. The keypoints are obtained by projecting the surface point using MANO.

Network architecture. Our keypoint VAE consists of an object encoder, keypoint encoder, and a decoder. The object encoder is a 4-layers PointNet encoder with a residual connection between each layer. The hand encoder and the decoder are 4-layers MLP networks with residuals connections. The hand encoder that takes hand key points and the object latent code then produces mean and standard variation of a 32-dimension Gaussian distribution. The decoder takes as inputs noise sample from the Gaussian distribution and the object latent vector to predict the hand key point locations. All layers have size 256. From key points to hand mesh, we use the final HALO model with differentiable canonicalization layer that takes 3D key points as inputs.

D. Limitation

HALO relies on biomechanically plausible 3D keypoints. Training HALO with the model end-to-end with angle losses alleviate this problem and results in a more robust surface prediction, as highlighted in HALO-VAE. This indicates that the inductive bias of our model helps encourage a biomechanically plausible 3D hand surface. However, a severely physically implausible hand skeleton could still produce artifacts on the hand surface.

Tolerance to Noisy Keypoints. We further analyse the impact of the biomechanical violations on the reconstruction quality of hand surfaces. We uniformly sample noises with the amplitude $[-x, +x]$ and add them to every dimension of every joints of a *valid* hand. As shown in Fig. D.1, with the amplitudes are within $2mm$, HALO produces reasonable hand surface. When x is increased to $5mm$, the reconstructed hand surface starts to show visible artifacts. Nevertheless, we reiterate that this problem can be mitigated by encouraging a bio-mechanically valid skeleton output from the estimator or generator, as in HALO-VAE.

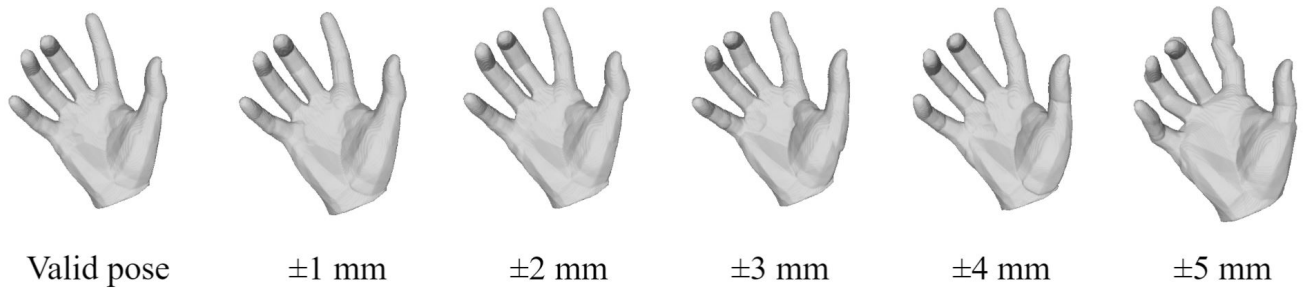

Figure D.1. HALO from noisy 3D keypoints.

E. Qualitative Results

E.1. HALO from Keypoints

Figure E.1 shows the HALO hand surfaces driven by the keypoint-based skeleton articulations.

In addition, to further demonstrate the generalisability of HALO, we show HALO surface driven by ground truth skeletons from the unseen Interhand2.6M [41] dataset in Figure E.2.

E.2. Generative Results

Figure E.3 shows the grasps **randomly** sampled from HALO-VAE conditioned on the objects from the test set of the GRAB dataset.

Figure E.1. Visualization of the HALO hand surfaces driven by the keypoint-based skeleton articulations. Note that the hand poses (articulated skeletons) are **randomly** sampled from the test set of Youtube3D.

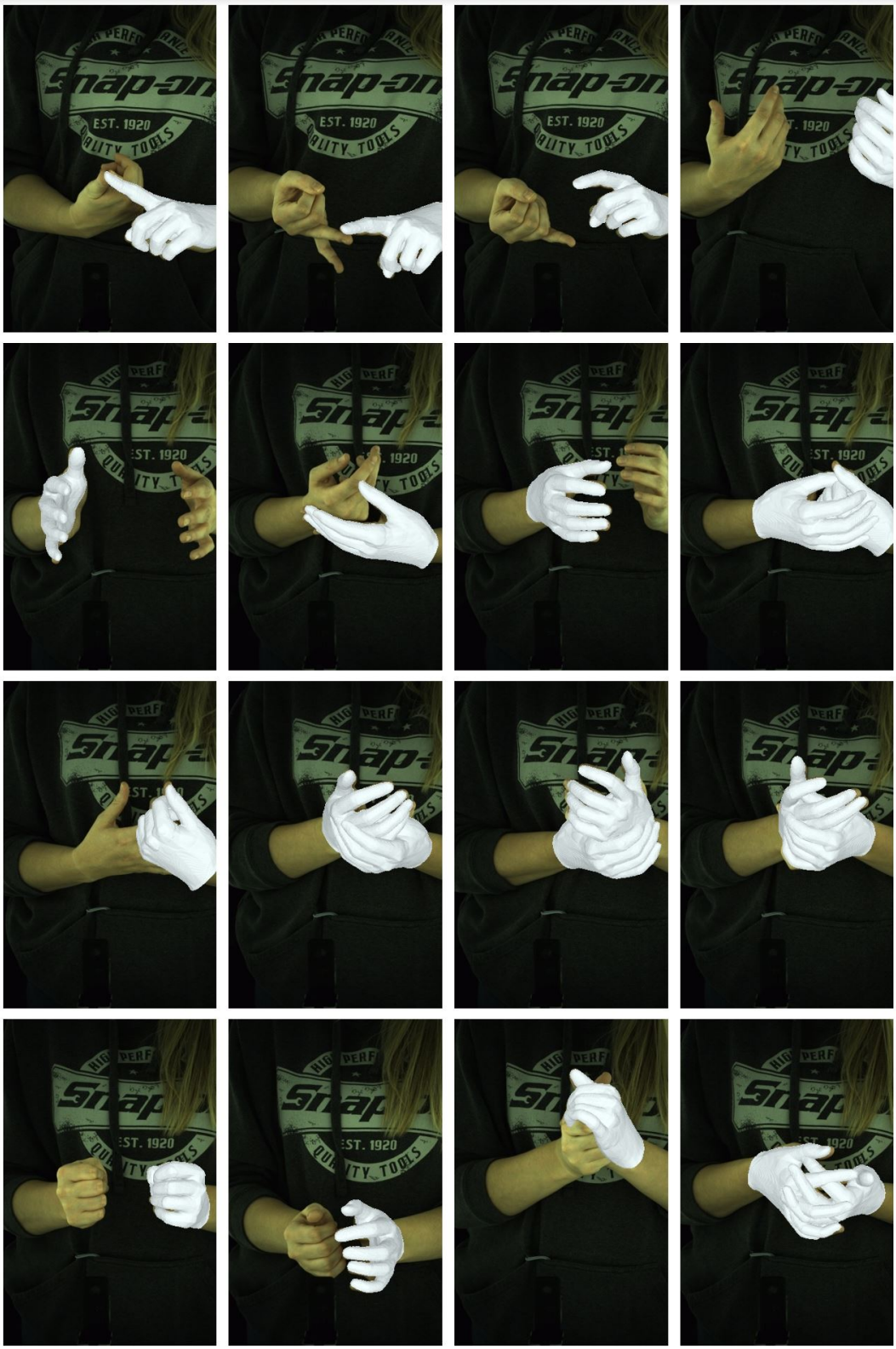

Figure E.2. Visualization of the HALO hand surfaces driven by the keypoint-based skeleton from the unseen Interhand2.6M dataset. The ground truth hand keypoints are given as input to reconstruct the hand surfaces. The RGB images are for visualization purpose only, and the results are not from pose estimation.

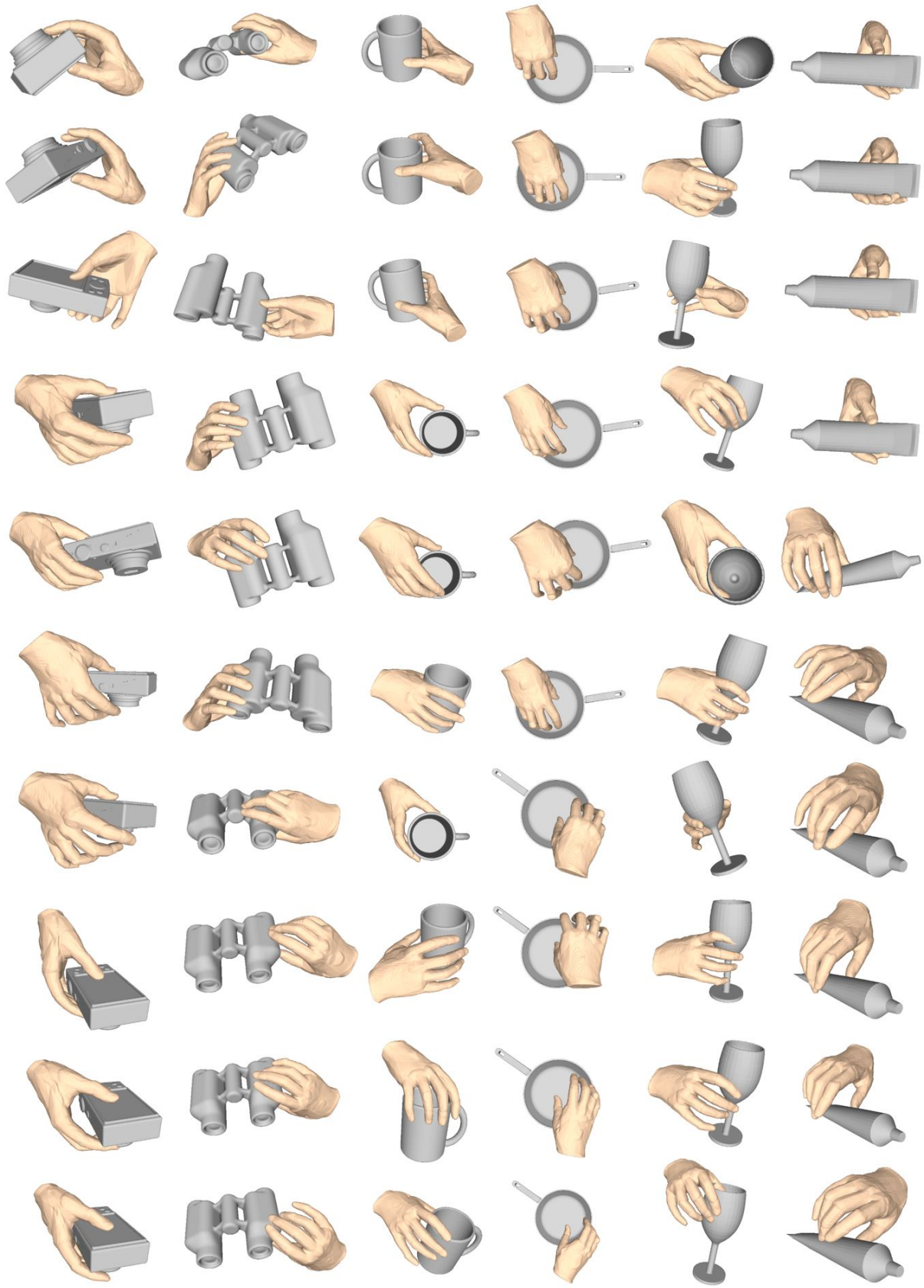

Figure E.3. Visualization of the grasps **randomly** sampled from HALO-VAE conditioned on the 6 unseen test objects from the GRAB dataset, 10 grasps per object.