LISA: Learning Implicit Shape and Appearance of Hands

Enric Corona[1†] Tomas Hodan[2] Minh Vo[2] Francesc Moreno-Noguer[1]
Chris Sweeney[2] Richard Newcombe[2] Lingni Ma[2]
[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain [2]Reality Labs, Meta

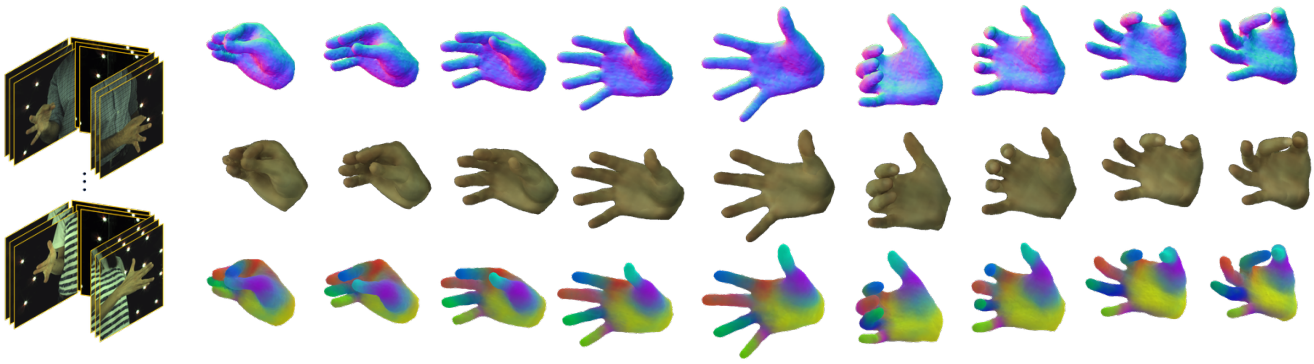

Figure 1. The LISA hand model is defined by an articulated implicit representation learned from multi-view RGB videos annotated with coarse 3D hand poses (left). The shape, color and pose parameters are disentangled in the model by design, enabling fine control of selected aspects of the model. In this figure, we animate the pose of a learned LISA model, while keeping the shape and color parameters fixed. The three rows show the shape shaded by surface normal, the appearance, and the color-coded skinning weights, respectively. The skinning weights are explicitly predicted and used to combine the per-bone predictions of the Signed Distance Field (SDF) and the surface color.

Abstract

This paper proposes a do-it-all neural model of human hands, named LISA. The model can capture accurate hand shape and appearance, generalize to arbitrary hand subjects, provide dense surface correspondences, be reconstructed from images in the wild, and can be easily animated. We train LISA by minimizing the shape and appearance losses on a large set of multi-view RGB image sequences annotated with coarse 3D poses of the hand skeleton. For a 3D point in the local hand coordinates, our model predicts the color and the signed distance with respect to each hand bone independently, and then combines the per-bone predictions using the predicted skinning weights. The shape, color, and pose representations are disentangled by design, enabling fine control of the selected hand parameters. We experimentally demonstrate that LISA can accurately reconstruct a dynamic hand from monocular or multi-view sequences, achieving a noticeably higher quality of reconstructed hand shapes compared to baseline approaches. Project page: https://www.iri.upc.edu/people/ecorona/lisa/.

可以从单目或者多视角序列中重建手

† Work performed during internship with Reality Labs, Meta.

1. Introduction

Since the thumb opposition enabled grasping around 2 million years ago [25], humans interact with the physical world mainly with hands. The problems of modeling and tracking human hands have therefore naturally received a considerable attention in computer vision [43]. Accurate and robust solutions to these problems would unlock a wide range of applications in, *e.g.*, human-robot interaction, prosthetic design, or virtual and augmented reality.

Most research efforts related to modeling and tracking human hands, *e.g.*, [8, 20, 21, 29, 38, 70], rely on the MANO hand model [53], which is defined by a polygon mesh that can be controlled by a set of shape and pose parameters. Despite being widely used, the MANO model has a low resolution and does not come with texture coordinates, which makes representing the surface color difficult.

The related field of modeling and tracking human bodies has been relying on parametric meshes as well, with the most popular model being SMPL [31] which suffers from similar limitations as the MANO model. Recent approaches for modeling human bodies, *e.g.*, [1, 4, 10, 15, 34, 54, 61], rely on articulated models based on implicit representations, such as Neural Radiance Field [35] or Signed Distance Field (SDF) [46]. Such representations are capable of representing both shape and appearance and able to capture

finer geometry compared to approaches based on parametric meshes. However, it is yet to be explored how well implicit representations apply to articulated objects such as the human hand and how they generalize to unseen poses.

We explore articulated implicit representations for modeling human hands and make the following contributions:

1. We introduce LISA, the first neural model of human hands that can capture accurate hand shape and appearance, generalize to arbitrary hand subjects, provide dense surface correspondences (via predicted skinning weights), be reconstructed from images in the wild, and easily animated.
2. We show how to train LISA by minimizing shape and appearance losses on a large set of multi-view RGB image sequences annotated with coarse 3D poses of the hand skeleton.
3. The shape, color and pose representations in LISA are disentangled by design, enabling fine control of selected aspects of the model.
4. Our experimental evaluation shows that LISA surpasses baselines in hand reconstruction from 3D point clouds and hand reconstruction from RGB images.

2. Related work

Parametric meshes. Thanks to their simplicity and efficiency, parametric meshes gained great popularity for modeling articulated objects such as bodies [24, 31, 44, 49], hands [53], faces [28] and animals [71]. The MANO hand model [53] is learned from a large set of carefully registered hand scans and captures shape-dependent and pose-dependent blend shapes for hand personalization. Despite widely adopted in hand tracking and shape estimation [6, 8, 13, 20–22, 27, 29, 36, 38, 68, 70], the MANO mesh is limited by a low resolution rooted from solving a large optimization problem with classical techniques. To reconstruct finer hand geometry, graph convolutional networks are explored in [12, 16, 57] and spiral filters in [26]. Based on a professionally designed mesh template, DeepHandMesh [37] learns the pose and shape corrective parameters by a neural network. Chen et. al., [9] refined MANO by developing a UV-based representation. GHUM [64] introduces a generative parametric mesh where the shape corrective parameters, skeleton and blend skinning weights are predicted by a neural network.

Implicit shape representations. Many works adopt neural networks to model geometry by learning an implicit function, which is continuous and differentiable, such as the signed distance field (SDF) [2, 3, 11, 14, 19, 46] or the occupancy field [33]. To improve learning efficiency, [7, 17, 18, 59] studied part-based implicit templates to model mid-level object-agnostic shape features. Implicit representations were extended to articulated deformation, in LoopReg [4] with a weakly-supervised training using cycle consistency by learning inverse skinning, which maps surface points to the SMPL human body model [31]. Based on SMPL, NASA [15] trains one OccNet [33] per skeleton bone to approximate the shape blend shapes and pose blend shapes. PTF [61] extends NASA and registers point clouds to SMPL. In a similar spirit, imGHUM [1] trains four DeepSDF networks [46] whose predictions are fused by an additional lightweight network. To eliminate the need of having the ground-truth SMPL in NASA training, SNARF [10] utilizes an iterative root finding technique to link every query point in the posed space to the corresponding point in the canonical space, which enables differentiable forward skinning. LEAP [34] and SCANimate [54] additionally model both forward and inverse skinning by a neural network, and use cycle consistency to supervise training of transformation to the canonical space. LEAP also extends the framework to multi-subject learning by mapping the bone transformation to a shape feature, and SCANimate builds animatable customized clothed avatars. We take inspiration from NASA to constrain hand deformation, but explicitly model the skinning weights for blending shape and color.

Implicit appearance representations. A number of approaches have been proposed to learn appearance of a scene from multi-view images. The idea is to model the image formation process by rendering a neural volume with ray-casting [30, 40, 55, 66]. Particularly, NeRF [35] gains popularity with an efficient formulation of modeling the radiance field. Follow-up studies show that geometry can be improved if the density is regulated by occupancy [42] or SDF [60, 65]. In this work, we use VolSDF [65] as a backbone renderer. For dynamic scenes, [47, 52, 58] combine NeRF with learning a deformation field. For modeling dynamic human bodies, Neural Body [51] attaches learnable vertex features to SMPL, and diffuses the features with sparse convolution for volumetric rendering. A-NeRF [56] conditions NeRF with SMPL bone transformations to learn an animatable avatar. Similar ideas are proposed in [50] and NARF [41]. H-NeRF [63] combines imGHUM with NeRF to enable appearance learning and train a separate network to predict SDF. In our work, the prediction of appearance and SDF is independent within each bone and later weighted by the corresponding skinning weights.

Disentangled representations. Disentangling parameters of certain properties such as pose, shape or color is desireable as it allows treating (*e.g.*, estimating or animating) these properties independently. Inspired by the parametric mesh models, Zhou et al., [69] trained a mesh auto-encoder to disentangle shape and pose of humans and animals. They developed an unsupervised learning technique

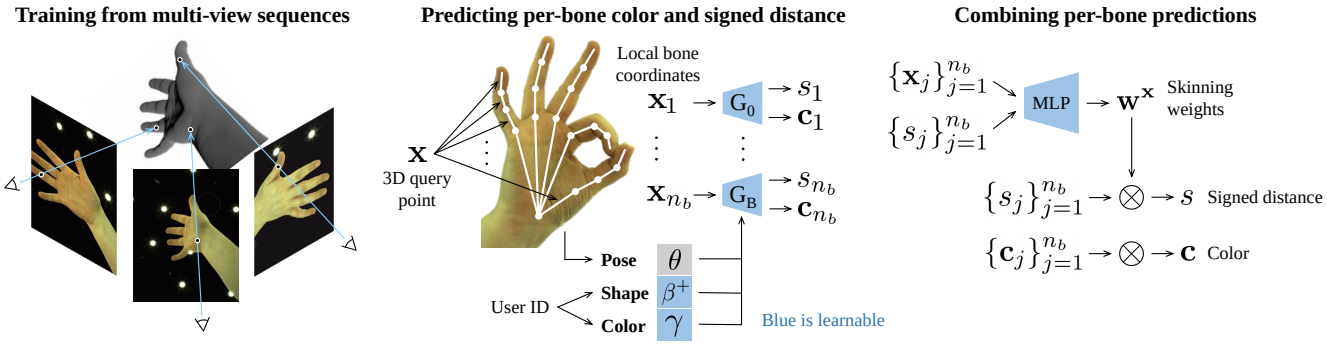

Figure 2. **Training and architecture of the LISA hand model.** *Left:* LISA is trained by minimizing shape and appearance losses from a dataset of multi-view RGB image sequences. The sequences are assumed annotated with coarse 3D poses of the hand skeleton that are refined during training. The training sequences show hands of multiple people and are used to learn disentangled representations of pose, shape and color. *Middle:* LISA approximates the hand by a collection of rigid parts defined by the hand bones. A 3D query point is transformed to the local coordinate systems of the bones associated with independent neural networks, which predict the signed distance to the hand surface and the color. Note that G_j is realized by two independent MLPs, one predicting the signed distance and one predicting the color (see Section 4.1). *Right:* The per-bone predictions are combined using skinning weights predicted by an additional network.

based on a cross-consistency loss. DiForm [62] adopted a decoder network to disentangle identity and deformation in learning an SDF-based shape embedding. A-SDF [39] factored out shape embedding and joint angles to model articulated objects. NPM [45] proposed to train shape embedding on canonically posed scans, followed by another network to learn the deformation field with dense supervision. A similar idea to the deformation field was adopted by i3DMM [67] to learn a human head model. The method disentangles identity, hairstyle, and expression and is trained with dense colored SDF supervision. In this work, we propose a generative hand representation with disentangled shape, pose and appearance parameters.

3. Background

MANO [53] represents the human hand as a function of the pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$:

$$M : (\boldsymbol{\theta}, \boldsymbol{\beta}) \mapsto \mathbf{V} \ , \tag{1}$$

where the hand is defined by a skeleton rig with $n_j = 16$ joints, and the pose parameters $\boldsymbol{\theta} \in \mathbb{R}^{n_j \times 3}$ represent the axis-angle representation of the relative rotation between bones of the skeleton. $\boldsymbol{\beta}$ is a 10-dimensional vector and $\mathbf{V} \in \mathbb{R}^{n_v \times 3}$ are the vertices of a triangular mesh. The mapping M is estimated by deforming a canonical hand $\mathbf{V}^r$ by a Linear Blend Skinning (LBS) transformation, with weights $\mathbf{W} \in \mathbb{R}^{n_b \times n_v}$, where n_b is the number of bones. Concretely, given a vertex $\mathbf{v}_i^r$ on the canonical shape, LBS transforms the vertex as follows:

$$\mathbf{v}_i = \sum_{j=1}^{n_b} w_{i,j} \mathbf{T}_j \bar{\mathbf{v}}_i^r \ , \tag{2}$$

where $\mathbf{T}_j \in \mathbb{R}^{3 \times 4}$ is the rigid transformation applied on the rest pose of bone j, $w_{i,j}$ is the (i, j) entry of $\mathbf{W}$ and $\bar{\mathbf{v}}$ denotes the homogeneous coordinates of $\mathbf{v}$. LISA builds on MANO's definition of the skeleton by using the same pose parameters $\boldsymbol{\theta}$ and bone transformations.

NeRF/VolSDF. NeRF [35] is a state-of-the-art rendering algorithm for novel view synthesis. The algorithm models the continuous radiance field of a static scene by learning the following function:

$$F : (\mathbf{x}, \mathbf{d}) \mapsto (\mathbf{c}, \sigma) \ , \tag{3}$$

which maps a 3D location $\mathbf{x} \in \mathbb{R}^3$ and the viewing direction $\mathbf{d} \in \mathbb{R}^2$ passing through $\mathbf{x}$ to the color value $\mathbf{c} \in \mathbb{R}^3$ and its density $\sigma \in \mathbb{R}$. The function F is modeled by an Multilayer Perceptron (MLP) network, which is trained from a set of dense multi-view posed RGB images of a single static scene. While NeRF has shown impressive novel view synthesis results, the estimated volume density is however not effective to infer accurate geometry. A number of recent works studied this problem [42, 60, 65] and propose to extended NeRF by incorporating SDF [46]. In this paper we adapt the formulation of VolSDF [65], which defines the volume density as a Laplace's Cumulative Distribution Function (CDF) applied to a SDF representation. VolSDF also disentangles the geometry and appearance learning using two MLPs for SDF and color estimation, respectively.

4. LISA: The proposed hand model

This section provides a detailed description of the proposed hand model, which we dub LISA for Learning Implicit Shape and Appearance model.

Problem settings. Consider a dataset of multi-view RGB video sequences with known camera calibration. Each sequence captures a single hand from a random person posing random motion. The objective is to learn a hand model,

which reconstructs the hand geometry, the deformation and the appearance, while also generalizes to reconstruct unseen hands and motion from test images. In contrast to prior hand modeling works, which often require a large collection of high-quality 3D hand scans, we consider a setup that lowers the requirement for data collection but adds the challenge to the algorithm. Inspired by classical hand modeling approaches, we assume that a kinematic skeleton is associated with the hand, where the coarse 3D poses are produced by pre-processing the training sequences with the state-of-art hand tracking. The motivation of using a skeleton is to regulate the hand deformation with articulation and to enable animation for the obtained model. To focus the deep network on the hands, we further simplify the input by assuming the foreground masks are known.

4.1. Model definition

Our goal is to learn a mapping function from the parametric skeleton to a full hand model of the shape and appearance. In this work, we choose the skeleton to be parameterized by MANO and formulate the learning as:

$$M^{+} : (\boldsymbol{\theta}, \boldsymbol{\beta}^{+}, \boldsymbol{\gamma}) \mapsto \boldsymbol{\psi} \ , \tag{4}$$

which maps the pose parameter $\boldsymbol{\theta}$, shape parameter $\boldsymbol{\beta}^{+}$ and the color parameter $\boldsymbol{\gamma}$ to an implicit representation $\boldsymbol{\psi}$. Here, we indicate the shape parameter is different from that of MANO using the superscript $^{+}$. The implicit representation $\boldsymbol{\psi}$ is a continuous function for the geometry and appearance. Similar to radiance field definition, this is defined as:

$$\boldsymbol{\psi} : (\mathbf{x}, \mathbf{d}) \mapsto (s, \mathbf{c}) \ , \tag{5}$$

which returns the SDF value $s \in \mathbb{R}$ and the color value $\mathbf{c}$ for a query 3D point $\mathbf{x}$ and the view direction $\mathbf{d}$. Using the implicit representation, the learned model is not tied to template meshes with fixed resolution, and therefore can encode detailed deformation more efficiently. The hand surface is represented by the 0-level set of s, where the 3D mesh $\mathbf{V}$ can be extracted by uniformly sampling the 3D space and applying Marching Cubes [32]. Putting Eq. (4) and Eq. (5) together, and with removing the viewing direction for simplified notation, yield the mapping we aim to learn:

$$G : (\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta}^{+}, \boldsymbol{\gamma}) \mapsto (s, \mathbf{c}) \ . \tag{6}$$

In the remainder of the section, we explain how to model Eq. (6) with network training.

Independent per-bone predictions with skinning. Following [15, 48], we approximate the overall hand shape by a collection of rigid parts, which are in our case defined by n_b bones. Specifically, the network G is split into n_b MLPs predicting the signed distance, $\{G_j^s\}_{j=1}^{n_b}$, and n_b MLPs predicting the color, $\{G_j^c\}_{j=1}^{n_b}$, with each MLP making an independent prediction with respect to one bone. As the input images correspond to posed hands, the point $\mathbf{x}$ is first unposed (*i.e.*, transformed to the coordinate space of the hand in the rest pose) using the kinematic transformations of bones, $\{\mathbf{T}_j\}_{j=1}^{n_b}$: $\mathbf{x}_j = \mathbf{R}_j^{-1}(\mathbf{x} - \mathbf{t}_j)$, where $\mathbf{R}_j$ and $\mathbf{t}_j$ are the rotation and translation components of the transformation $\mathbf{T}_j$. With this formulation, we collect a set of independent SDF predictions and color predictions for each query point $\{s_j, \mathbf{c}_j\}_{j=1}^{n_b}$, where:

$$G_j^s : (\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\beta}^{+}, \boldsymbol{\gamma}) \mapsto s_j \ , \tag{7}$$

$$G_j^c : (\mathbf{x}_j, \boldsymbol{\theta}, \boldsymbol{\beta}^{+}, \boldsymbol{\gamma}) \mapsto \mathbf{c}_j \ . \tag{8}$$

To combine the per-bone output into a single SDF and color addition, we introduce an additional MLP to learn the weights. The weight MLP takes the input as the concatenation of n_b unposed $\mathbf{x}_j$ and the predicted SDF s_j per MLP , to output the weighting vector $\mathbf{w} = [w_1, \ldots, w_{n_b}]$. A softmax layer is used to constrain the value of $\mathbf{w}$ to be probability-alike, i.e., $w_i \geq 0 \ , \forall i$ and $\sum_i w_i = 1$. The final output for a query point is then computed by:

$$s = \sum_{j=1}^{n_b} w_j s_j \ , \quad \mathbf{c} = \sum_{j=1}^{n_b} w_j \mathbf{c}_j \ . \tag{9}$$

Note the weight vector $\mathbf{w}$ is an analogy to skinning weights in classic LBS-based models. The similar design has also been explored by NASA [15] and NARF [41]. The difference is that NASA selects one MLP output, which is determined by the maximum of the predicted occupancies. NARF proposes to learn the weights with an MLP, but only uses the canonicalized points to train this module. In our design, the network sees both canonicalized points and the per-bone SDF. The SDF serves as a valuable guide in learning skinning weight. More importantly, the gradients can now back-propagate via the weights to train the per-bone MLPs. This means MLPs can leverage $\mathbf{w}$ to avoid learning SDF for far-away points. We show in experiments that this design greatly improves geometry.

Model rendering. As in [65], we first need to obtain the volume densities from the predicted signed distance field before rendering. We infer it indirectly from the predicted signed distances:

$$\sigma(\mathbf{x}) = \alpha \Psi_\beta(-s) \ , \tag{10}$$

where s is the signed distance of $\mathbf{x}$, $\Psi_\beta(\cdot)$ is the CDF of the Laplace distribution, and α and β are two learnable parameters (see [65] for further details).

The color of a specific image pixel is then estimated via the volume rendering integral, by accumulating colors and volume densities along its corresponding camera ray $\mathbf{d}$. In particular, the color of the pixel $\hat{\mathbf{c}}_k$ is approximated by a discrete integration between near and far bounds t_n and t_f of a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with origin $\mathbf{o}$:

$$\mathbf{c}_k = \int_{t_n}^{t_f} T(t)\sigma(r(t))\mathbf{c}(\mathbf{r}(t),)dt \, , \tag{11}$$

将手分为骨节，每个骨节2个MLP网络，预测符号距离和颜色

where:

$$T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right) . \quad (12)$$

4.2. Training

As shown in Fig. 2, the parameters of LISA that need to be learned are: (1) the MLPs for predicting signed distance and color for the n_b bones, (2) the MLP that estimates the skinning weights, and (3) the shape $\boldsymbol{\beta}^+$ and color $\boldsymbol{\gamma}$ latent codes to control the generation process. Note that the pose $\boldsymbol{\theta}$ is not learned and assumed given during training. We next explain how we learn these parameters from the multi-view image sequences from the InterHand2.6M dataset [38].

Disentangling shape, color and pose. LISA is designed to completely disentangle the representations of pose, shape and color. The shape $\boldsymbol{\beta}^+$ and color $\boldsymbol{\gamma}$ parameters are fully learnable latent vectors. Since both are user specific, we assign the same latent code for all images of the same person. In both cases, they are represented as 128-dimensional vectors, initialized from a zero-mean multivariate-Gaussian distribution with a spherical covariance, and optimized during training following the auto-decoder formulation of [46].

The pose parameters $\boldsymbol{\theta}$ are defined by the 48-dimensional representation of MANO. When training on InterHand2.6M, we kept the provided ground-truth pose parameters fixed for the initial 10% of training steps, then we started optimizing the parameters to account for errors in the ground-truth annotations.

Color calibration. In order to allow for slight differences in the intensity of the training images, we follow Neural Volumes [30] and introduce a per-camera and per-channel gain g and bias b that is applied to the rendered images at training time. At inference, we use the average of these calibration parameters.

Loss functions. To learn LISA, we minimize a combination of losses that aim to ensure accurate representation of the hand color while properly regularizing the learned geometry. Specifically, we optimize the network by randomly sampling a batch of viewing directions $\mathbf{d}_k$ and estimating the corresponding pixel color via volume rendering. Let $\mathbf{c}_k$ be the estimated pixel color and $\hat{\mathbf{c}}_k$ the ground truth value. The first loss we consider is:

$$\mathcal{L}_{\text{col}} = \|\mathbf{c}_k - \hat{\mathbf{c}}_k\|_1 , \quad (13)$$

where $\|\cdot\|_j$ denotes the j-norm. We also regularize the SDF of $G(\cdot)$ with the Eikonal loss [19] to ensure it approximates a signed distance function:

$$\mathcal{L}_{\text{Eik}} = \sum_{\mathbf{x}\in\Omega} (\|\nabla_{\mathbf{x}} G(\mathbf{x})\|_2 - 1)^2 , \quad (14)$$

where Ω is a set of points sampled both on the surface and uniformly taken from the entire scene. In order to prevent local minima in regions relying only on one or a few bones, We use the pseudo-ground truth pose and shape parameters to obtain an approximate 3D mesh and its corresponding skinning weights $\hat{\mathbf{w}}$, which we use to supervise the predicted skinning weights $\mathbf{w}$:

$$\mathcal{L}_{\text{w}} = \|\mathbf{w} - \hat{\mathbf{w}}\|_1 . \quad (15)$$

Finally, we also regularize the latent vectors $\boldsymbol{\beta}^+$ and $\boldsymbol{\gamma}$:

$$\mathcal{L}_{\text{reg}} = \|\boldsymbol{\beta}^+\|_2 + \|\boldsymbol{\gamma}\|_2 . \quad (16)$$

The full loss is a linear combination of the four previous loss terms (with hyperparameters λ_{col}, λ_{Eik}, λ_{w} and λ_{reg}):

$$\mathcal{L} = \lambda_{\text{col}}\mathcal{L}_{\text{col}} + \lambda_{\text{Eik}}\mathcal{L}_{\text{Eik}} + \lambda_{\text{w}}\mathcal{L}_{\text{w}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} . \quad (17)$$

Learning a prior for human hand SDFs. When minimizing Eq. (17), we face two main challenges. First, since we only supervise on images, the simultaneous optimization of shape and texture parameters may lead to local minima with good renders but wrong geometries. Second, the InterHand2.6M dataset [38] we use for training has a large number of images ($\sim$130k) but they only correspond to 27 different users, compromising the generalization of the model.

To alleviate these problems, we build a shape prior using the 3DH dataset [62], which contains $\sim$13k 3D posed hand scans of 183 different users. The scans are used to pre-train the geometry MLPs in $G(\cdot)$, which we denote $G_{\beta^+}(\cdot)$, and which are responsible for predicting the signed distance s:

$$G_{\beta^+} : (\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta}^+) \mapsto s . \quad (18)$$

We pre-train G_{β^+} with two additional losses. First, assuming $\mathbf{x}_{\text{surf}}$ to be a point of a 3D scan, we enforce G_{β^+} to predict a 0 distance on that point:

$$\mathcal{L}_{\text{surf}} = \|G_{\beta^+}(\mathbf{x}_{\text{surf}}, \boldsymbol{\theta}, \boldsymbol{\beta}^+)\|_1 . \quad (19)$$

We also supervise the gradient of the signed distance with the ground truth normal $N(\mathbf{x}_{\text{surf}})$ at $\mathbf{x}_{\text{surf}}$:

$$\mathcal{L}_{\text{N}} = \|\nabla_{\mathbf{x}_{\text{surf}}} G(\mathbf{x}_{\text{surf}}) - N(\mathbf{x}_{\text{surf}})\|_1 , \quad (20)$$

where $N(\mathbf{x})$ is the 3D normal direction at $\mathbf{x}$.

With these two losses, jointly with losses $\mathcal{L}_{\text{Eik}}$, $\mathcal{L}_{\text{w}}$ and the regularization $\|\boldsymbol{\beta}^+\|_2$, we learn a prior on $\boldsymbol{\beta}^+$ which is used to initialize the full optimization of the model in Eq. (17). As we show in the experimental section, this prior allows to significantly boost the performance of LISA.

4.3. Inference

In the experimental section we apply the learned model to 3D reconstruction from point-clouds and to 3D reconstruction from images. Both of these applications involve an optimization scheme which we describe below.

Reconstruction from point clouds. Let $\mathcal{P} = \{\mathbf{x}_i\}_{i=1}^n$ be a point-cloud with n 3D points. To fit our trained model to this data, we follow a very similar pipeline as the one used to learn the prior. Specifically, we minimize the following objective function:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}^+) = \sum_{\mathbf{x}\in\mathcal{P}} \|G_{\boldsymbol{\beta}^+}(\mathbf{x}, \boldsymbol{\theta}, \boldsymbol{\beta}^+)\|_1 + \|\boldsymbol{\beta}^+\|_2 \,. \quad (21)$$

Reconstruction from monocular or multiview images. Given an input image $\mathcal{I}$, we assume we have a coarse foreground mask and that the 2D locations of n_j hand joints, denoted as $\hat{\mathbf{J}}^{\text{2D}}$, are available. These locations can be detected using, *e.g.*, OpenPose [5]. To fit LISA to this data, we minimize the following objective:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\beta}^+) = \sum_{\mathbf{d}\in\mathcal{I}} \mathcal{L}_{\text{col}}(\mathbf{d}) + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{joints}} \,, \quad (22)$$

where the first two terms correspond to the color loss of Eq. (13) (expanded to all viewing directions intersecting the pixels of the input image), and the shape and pose regularization loss in Eq. (16). The last term is a joint-based data term that penalizes the 2D distance between the estimated 2D joints and the projected 3D joints $\mathbf{J}^{\text{3D}}$ computed from the estimated pose parameters $\boldsymbol{\theta}$:

$$\mathcal{L}_{\text{joints}} = \|\hat{\mathbf{J}}^{\text{2D}} - \pi(\mathbf{J}^{\text{3D}})\|_1 \,, \quad (23)$$

where $\pi(\cdot)$ is the 3D-to-2D projection. We also use extrinsic camera parameters in case of multi-view reconstruction.

5. Experiments

In this section, we evaluate LISA on the tasks of hand reconstruction from point clouds and hand reconstruction from RGB images, and demonstrate that it outperforms the state of the art by a considerable margin.

5.1. Datasets and baselines

Datasets. We train LISA on a non-released version of the InterHand2.6M dataset [38], which contains multi-view sequences showing hands of 27 users. In total, there are 5804 multi-view frames and 131k images with the resolution of 1024×667 px. Every frame has $\sim$22 views on average, two of which were not used for training and left for validation. The dataset also provides a pseudo ground truth of the 3D joints, and we remove background in all images using hand masks obtained by a Mask R-CNN model [23] provided by the authors of the dataset. The geometry prior is learned on the 3DH dataset [62] which contains sequences of 3D scans of 183 users (we use the same training/test split of 150/33 users proposed by the authors). For evaluating hand reconstruction from point clouds, we use the test split of the MANO dataset [53], which includes 50 3D scans of a single user, and the test set of 3DH, which includes scans of 33 users. For hand reconstruction from images, we use the DeepHandMesh dataset [37], which is annotated with ground-truth 3D hand scans.

	Reconstruction to scan		Scan to reconstruction	
Method	V2V [mm]	V2S [mm]	V2V [mm]	V2S [mm]
3DH dataset [62]:				
MANO [53]	3.27	2.11	3.44	3.23
VolSDF [46]	3.69	1.26	5.33	5.23
NASA [15]	3.05	1.14	3.69	3.66
NARF [41]	4.69	2.19	2.05	2.01
LISA-im	**2.93**	**0.93**	**1.90**	**1.87**
LISA-geom	0.83	0.43	0.63	0.54
LISA-full	1.93	0.63	1.50	1.43
MANO dataset [53]:				
MANO [53]	3.14	2.92	3.90	1.57
VolSDF [46]	3.69	2.22	2.37	2.23
NASA [15]	5.31	3.80	2.57	2.33
NARF [41]	4.02	2.69	2.11	2.06
LISA-im	**3.09**	**1.96**	**1.19**	**1.13**
LISA-geom	0.36	0.16	0.81	0.26
LISA-full	1.45	0.64	0.64	0.58

Table 1. **Shape reconstruction from point clouds.** The 3D shape reconstructions are evaluated by the vertex-to-vertex and vertex-to-surface distances (in mm). LISA-im is consistently superior among methods trained on images only. Using the geometric prior (LISA-geom, LISA-full) yields a significant boost in performance.

Evaluated hand models. As LISA is the first neural model able to simultaneously represent hand geometry and texture, there are no published methods that would be directly comparable. To define baselines, we have therefore re-implemented several recent methods based on articulated implicit representations from the related field of human body modeling. We adapt NASA [15] and NARF [41] to our setup by changing their geometry representation to signed distance fields, adding a positional encoding to NASA, and duplicating their geometry MLPs to predict also color. We train these methods on the InterHand2.6M dataset [38] with supervision on the skinning weights. We did not manage to extend SNARF [10], as it relies on an intermediate non-differentiable optimization during the forward pass that impedes calculating the output gradient with respect to the input points, which is necessary for applying the Eikonal loss. We also compare to the original MANO model and to our implementation of VolSDF parameterized by the pose, shape and color vectors, but which does not consider a per-bone reasoning. Besides, we ablate the following versions of the proposed model: the full model when trained with images and the geometric prior (LISA-full), a version trained solely with images (LISA-im), and a version trained only with the geometric prior (LISA-geom).

Method	1 view			2 views			4 views		
	V2V	V2S	PSNR	V2V	V2S	PSNR	V2V	V2S	PSNR
MANO [53]	13.81	8.93	-	-	-	-	-	-	-
DHM [37]	9.86	6.55	-	-	-	-	-	-	-
VolSDF [65]	7.15	7.06	23.19	7.15	7.10	22.63	7.27	7.18	25.05
NASA [15]	5.89	5.79	**25.20**	5.11	4.99	25.17	5.04	4.91	25.18
NARF [41]	7.44	7.35	24.11	7.45	7.36	28.48	7.93	7.85	29.89
LISA-im	**5.48**	**5.36**	25.04	**3.86**	**3.72**	**29.84**	**3.62**	**3.47**	**30.21**
LISA-full	3.84	3.68	25.43	3.70	3.56	29.40	3.53	3.38	29.69

Table 2. **Shape and color reconstruction of DHM [37] images.** The 3D shape reconstructions are evaluated by the vertex-to-vertex and vertex-to-surface distances (in mm) and color renderings of the hand models in novel views are evaluated by the PSNR metric [35]. Scores for MANO and DeepHandMesh (DHM) were taken from [37]. We also report metrics for 1, 2 or 4 views, out of the 5 available images in [37]. In the same conditions, LISA-im outperforms all other methods trained on images only. When trained also with the geometry prior (LISA-full), it achieves an additional boost that is most noticeable in the single-view setup.

5.2. Shape reconstruction from point clouds

Table 1 summarizes the results of hand reconstruction from point clouds from the 3DH and MANO datasets. As the evaluation metrics, we report the vertex-to-vertex (V2V) and vertex-to-surface (V2S) distances (in millimeters). We compute these metrics in both directions, *i.e.* from the reconstruction to the scan and the other way around. For a fair comparison, all reconstructions from all methods based on implicit representations are obtained with the same Marching Cubes resolution ($256\times256\times256$). Since MANO uses a mesh with only 778 vertices, we subdivide its reconstructed surface into $\sim$100k vertices.

The results show that LISA-im consistently outperforms the other methods when only images are used for training. Adding the geometric prior (LISA-full) yields a significant boost in performance. When the model is trained solely with the geometric prior (LISA-geom), it yields even lower errors than when trained using both the geometric prior and images (LISA-full). This is because we segmented out the hand in the training images and LISA-full and LISA-im therefore learned to close the surface right after the wrist. This spurious surface increases the measured error.

Figure 3 visualizes examples of the reconstructions. Clear artifacts can be seen in most implicit models, except of LISA-full and the parametric MANO model.

5.3. Shape and color reconstruction from images

Table 2 evaluates the hand models on the task of 3D reconstruction from single and multiple views on the DeepHandMesh dataset [37]. Among methods trained on images only, LISA-im is consistently superior in 3D shape reconstruction, and its performance is further boosted when the geometric prior is employed (LISA-full).

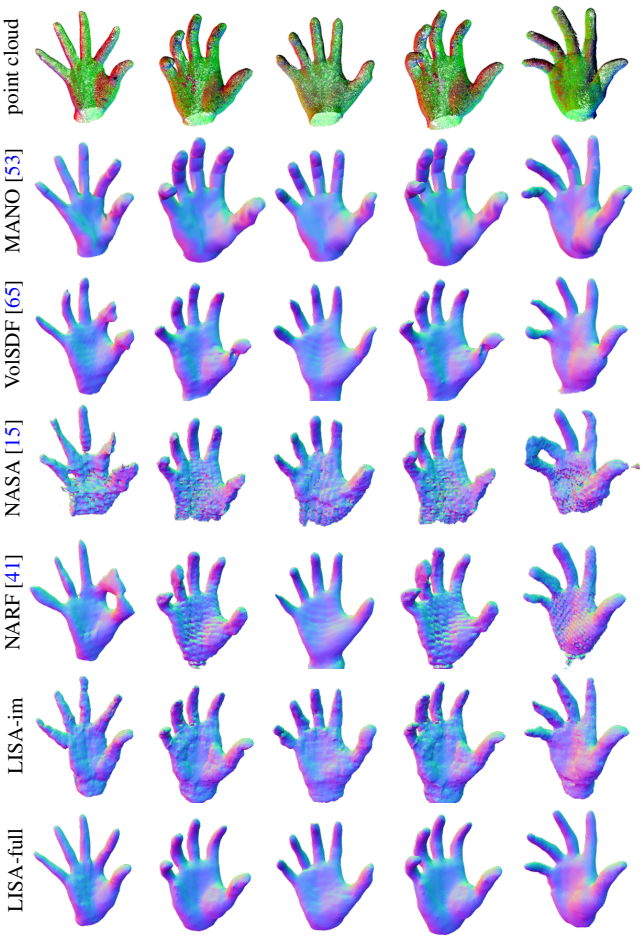

Figure 3. **Shape reconstruction from MANO points clouds [53].** VolSDF, NASA, NARF and LISA-im are trained only on InterHand2.6M [38] and tested on MANO. Implicit models without skinning-based regularization (VolSDF, NASA, NARF) often generate connected regions. LISA-full pre-learns the geometry from 3DH [62] and achieves smoother reconstructions.

	1 view			2 views			4 views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
VolSDF	23.01	0.92	0.12	23.36	0.87	0.11	23.89	0.93	0.11
NASA	26.90	0.95	0.07	28.16	0.96	0.06	28.44	0.96	0.06
NARF	28.42	0.95	0.09	28.49	0.96	0.72	28.59	0.96	0.08
LISA-im	27.45	0.95	0.08	28.40	0.95	0.07	28.27	0.95	0.07
LISA-full	27.07	0.95	0.06	27.69	0.96	0.06	28.27	0.96	0.05

Table 3. **Color reconstruct. from InterHand2.6M images [38].** All models achieve comparable performance in terms of PSNR and SSIM (measuring the pixel error; higher is better) and LPIPS (measuring the overall perceptual similarity; lower is better) [35].

Color reconstruction from DeepHandMesh images [37] is evaluated in Table 2 by the PSNR metric [35] calculated on renderings of the hand model from novel views. LISA-im is slightly superior in this metric, with exception of the case when a single image is used for the reconstruction,

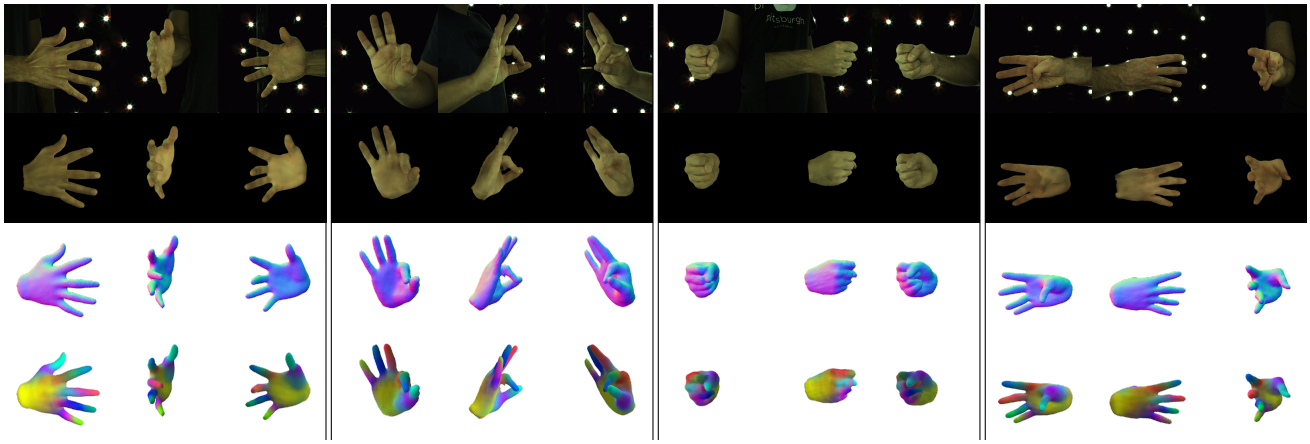

Figure 4. **Reconstructions from InterHand2.6M [38] images produced by LISA-full.** The top row shows the ground truth RGB images. We only use the first RGB for monocular reconstruction, while the other two images serve as the ground truth for novel view rendering. The other three rows show the color reconstruction, the shape reconstruction, and the predicted skinning weights, respectively.

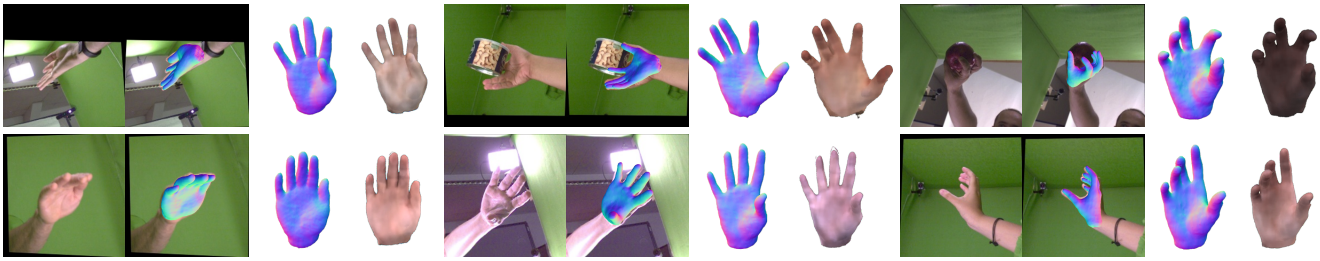

Figure 5. **In-the-wild reconstructions produced by LISA-full.** For each example, we show from left to right: (a) the input RGB image, (b) the reconstructed shape rendered to the input image, (c) the reconstructed shape in a reference view, (d) the reconstructed color.

where the performance of LISA-im is on par with NASA. Color reconstruction from InterHand2.6M images [38] is evaluated in Table 3 by the PSNR, SSIM and LPIPS [35] metrics. In this case, all methods are fairly comparable in terms of rendering quality, which we suspect is likely due to noisy hand masks used in training.

Qualitative results are shown in Figure 4. Additionally, we demonstrate in Figure 5 that LISA can reconstruct hands from images in the wild, even in cases where the hand is partially occluded by an object. We refer the reader to the supplementary material for additional qualitative results.

5.4. Inference speed

To reconstruct the LISA hand model from one or multiple views, we first optimize the pose parameters for 1k iterations, after which we jointly optimize shape, pose and color parameters for additional 5k iterations. This process takes approximately 5 minutes. After converging, we reconstruct meshes at the resolution of 128^3, which takes around 5 seconds, or render novel views in approximately one minute. These measurements were made on 1024×667 px images with a single Nvidia Tesla P100 GPU. The inference speed is similar for the NASA and NARF models, which also perform per-bone predictions. VolSDF is $\sim$2 faster due to the fact that it only uses a single MLP.

6. Conclusion

We have introduced LISA, a novel neural representation of textured hands, which we learn by combining volume rendering approaches with a hand geometric prior. The resulting model is the first one to allow full and independent control of pose, shape and color domains. We show the utility of LISA in two challenging problems, hand reconstruction from point-clouds and hand reconstruction from images. In both of these applications we obtain highly accurate 3D shape reconstructions, achieving a sub-millimeter error in point-cloud fitting and surpassing the evaluated baselines by large margins. This level of accuracy is not possible to achieve with low-resolution parametric meshes such as MANO [53] or with models representing a single person such as DeepHandMesh [37]. Future research directions include exploring temporal consistency for tracking applications, eliminating the need of rough 2D/3D pose of the hand skeleton and foreground mask at inference, improving the run-time efficiency, or enhancing the expressiveness in terms of high-frequency textural details while maintaining the generalization capability.

Acknowledgements: This work is supported in part by the Spanish government with the project MoHuCo PID2020-120049RB-I00.

References

[1] Thiemo Alldieck, Hongyi Xu, and Cristian Sminchisescu. imGHUM: Implicit generative models of 3d human shape and articulated pose. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2

[2] Matan Atzmon and Yaron Lipman. SAL: Sign agnostic learning of shapes from raw data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[3] Matan Atzmon and Yaron Lipman. SALD: Sign agnostic learning with derivatives. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. 2

[4] Bharat Lal Bhatnagar, Cristian Sminchisescu, Christian Theobalt, and Gerard Pons-Moll. Loopreg: Self-supervised learning of implicit surface correspondences, pose and shape for 3d human mesh registration. In *Neural Information Processing Systems (NeurIPS)*, December 2020. 1, 2

[5] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: realtime multi-person 2d pose estimation using part affinity fields. *PAMI*, 43(1):172–186, 2019. 6

[6] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 12417–12426, October 2021. 2

[7] Rohan Chabra, Jan E. Lenssen, Eddy Ilg, Tanner Schmidt, Julian Straub, Steven Lovegrove, and Richard Newcombe. Deep local shapes: Learning local sdf priors for detailed 3d reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 608–625, 2020. 2

[8] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. DexYCB: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9044–9053, 2021. 1, 2

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

[10] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 1, 2, 6

[11] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5932–5941, 2019. 2

[12] Hongsuk Choi, Gyeongsik Moon, and Kyoung Mu Lee. Pose2Mesh: Graph convolutional network for 3d human pose and mesh recovery from a 2d human pose. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

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

[14] Enric Corona, Albert Pumarola, Guillem Alenya, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In *CVPR*, pages 11875–11885, 2021. 2

[15] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 612–628. Springer, 2020. 1, 2, 4, 6, 7

[16] 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 (CVPR)*, pages 10833–10842, 2019. 2

[17] Kyle Genova, Forrester Cole, Avneesh Sud, Aaron Sarna, and Thomas Funkhouser. Local deep implicit functions for 3d shape. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[18] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T. Freeman, and Thomas A. Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 7153–7163, 2019. 2

[19] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2020. 2, 5

[20] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3196–3206, 2020. 1, 2

[21] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2

[22] Yana Hasson, Gül Varol, Dimitris 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 (CVPR)*, 2019. 2

[23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, pages 2961–2969, 2017. 6

[24] Hanbyul Joo, Tomas Simon, and Yaser Sheikh. Total capture: A 3d deformation model for tracking faces, hands, and bodies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2

[25] Fotios Alexandros Karakostis, Daniel Haeufle, Ioanna Anastopoulou, Konstantinos Moraitis, Gerhard Hotz, Vangelis Tourloukis, and Katerina Harvati. Biomechanics of the

human thumb and the evolution of dexterity. *Current Biology*, 31(6):1317–1325, 2021. 1

[26] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4990–5000, 2020. 2

[27] Taein Kwon, Bugra Tekin, Jan Stuhmer, Federica Bogo, and Marc Pollefeys. H2O: Two hands manipulating objects for first person interaction recognition. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[28] Tianye Li, Timo Bolkart, Michael. J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *SIGGRAPH*, 36(6):194:1–194:17, 2017. 2

[29] 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 Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[30] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *ACM Transactions on Graphics (TOG)*, 38(4), July 2019. 2, 5

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

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

[33] 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 (CVPR)*, 2019. 2

[34] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. LEAP: Learning articulated occupancy of people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10461–10471, 2021. 1, 2

[35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 1, 2, 3, 7, 8

[36] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[37] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. DeepHandMesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 440–455. Springer, 2020. 2, 6, 7, 8

[38] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2, 5, 6, 7, 8

[39] Jiteng Mu, Weichao Qiu, Adam Kortylewski, Alan L. Yuille, Nuno Vasconcelos, and Xiaolong Wang. A-SDF: learning disentangled signed distance functions for articulated shape representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[40] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[41] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 4, 6, 7

[42] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2, 3

[43] Iason Oikonomidis, Guillermo Garcia-Hernando, Angela Yao, Antonis Argyros, Vincent Lepetit, and Tae-Kyun Kim. Hands18: Methods, techniques and applications for hand observation. In *ECCV Workshops*, pages 0–0, 2018. 1

[44] Ahmed A A Osman, Timo Bolkart, and Michael J. Black. STAR: A spare trained articulated human body regressor. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[45] Pablo Palafox, Aljaž Božič, Justus Thies, Matthias Nießner, and Angela Dai. NPMs: Neural parametric models for 3d deformable shapes. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[46] 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 (CVPR)*, 2019. 1, 2, 3, 5, 6

[47] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[48] Despoina Paschalidou, Angelos Katharopoulos, Andreas Geiger, and Sanja Fidler. Neural parts: Learning expressive 3d shape abstractions with invertible neural networks. In *CVPR*, pages 3204–3215, 2021. 4

[49] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[50] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[51] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural Body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9054–9063, 2021. 2

[52] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-NeRF: Neural Radiance Fields for Dynamic Scenes. In *CVPR*, 2020. 2

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

[54] Shunsuke Saito, Jinlong Yang, Qianli Ma, and Michael J Black. SCANimate: Weakly supervised learning of skinned clothed avatar networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[55] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Neural Information Processing Systems (NeurIPS)*, 2019. 2

[56] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. In *Neural Information Processing Systems (NeurIPS)*, 2021. 2

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

[58] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 2

[59] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Carsten Stoll, and Christian Theobalt. PatchNets: Patch-based generalizable deep implicit 3d shape representations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2

[60] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Neural Information Processing Systems (NeurIPS)*, 2021. 2, 3

[61] Shaofei Wang, Andreas Geiger, and Siyu Tang. Locally aware piecewise transformation fields for 3d human mesh registration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2

[62] Binbin Xu, Lingni Ma, Yuting Ye, Tanner Schmidt, Christopher D. Twigg, and S. Lovegrove. Identity-disentangled neural deformation model for dynamic meshes. *ArXiv*, abs/2109.15299, 2021. 3, 5, 6, 7

[63] Hongyi Xu, Thiemo Alldieck, and Cristian Sminchisescu. H-NeRF: Neural radiance fields for rendering and temporal reconstruction of humans in motion. In *Neural Information Processing Systems (NeurIPS)*, 2021. 2

[64] Hongyi Xu, Eduard Gabriel Bazavan, Andrei Zanfir, William Freeman, Rahul Sukthankar, and Cristian Sminchisescu. GHUM & GHUML: Generative 3d human shape and articulated pose models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2

[65] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *ArXiv*, abs/2106.12052, 2021. 2, 3, 4, 7

[66] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Neural Information Processing Systems (NeurIPS)*, 33, 2020. 2

[67] Tarun Yenamandra, Ayush Tewari, Florian Bernard, Hans-Peter Seidel, Mohamed Elgharib, Daniel Cremers, and Christian Theobalt. i3DMM: Deep implicit 3d morphable model of human heads. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12803–12813, 2021. 3

[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 International Conference on Computer Vision (ICCV)*, pages 2354–2364, 10 2019. 2

[69] Keyang Zhou, Bharat Lal Bhatnagar, and Gerard Pons-Moll. Unsupervised shape and pose disentanglement for 3d meshes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020. 2

[70] 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 *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 813–822, 2019. 1, 2

[71] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5524–5532, Los Alamitos, CA, USA, Jul 2017. 2