gSDF: Geometry-Driven Signed Distance Functions for 3D Hand-Object Reconstruction

Zerui Chen Shizhe Chen Cordelia Schmid Ivan Laptev
Inria, École normale supérieure, CNRS, PSL Research Univ., 75005 Paris, France
firstname.lastname@inria.fr

Abstract

Signed distance functions (SDFs) is an attractive framework that has recently shown promising results for 3D shape reconstruction from images. SDFs seamlessly generalize to different shape resolutions and topologies but lack explicit modelling of the underlying 3D object geometry. In this work we exploit the object structure and use it as guidance for SDF-based shape reconstruction. In particular, we address reconstruction of hands and manipulated objects from monocular RGB images. To this end, we estimate poses of hands and objects and use them to guide SDF shape reconstruction. More specifically, we predict kinematic chains of pose transformations from images and align SDFs with highly-articulated hand poses. We improve the visual features of 3D points with geometry alignment and further leverage temporal information to enhance the robustness to occlusion and motion blurs. We conduct extensive experiments on the challenging ObMan and DexYCB benchmarks and demonstrate significant improvements of the proposed method over the state of the art.

1. Introduction

Understanding how hands interact with objects is becoming increasingly important for widespread applications, including virtual reality, robotic manipulation and human-computer interaction. Compared to 3D estimation of sparse hand joints [24, 38, 51, 53, 67], joint reconstruction of hand and object meshes [11, 18, 21, 26, 62] provides rich information about hand-object interactions and has received increased attention in recent years.

To reconstruct high-quality meshes, some recent works [9, 17, 61] explore multi-view image inputs. Multi-view images, however, are less common both for training and testing scenarios. In this work, we focus on a more practical and user-friendly setting where we aim to reconstruct hand and object meshes from monocular RGB images. Given the ill-posed nature of the problem, many existing methods [7, 19, 21, 54, 62] employ parametric mesh models (*e.g.*, MANO [46]) to encode prior knowledge on hands and to reduce ambiguities in 3D reconstruction. MANO hand meshes, however, have relatively limited resolution and can be suboptimal for the precise capture of hand-object interactions.

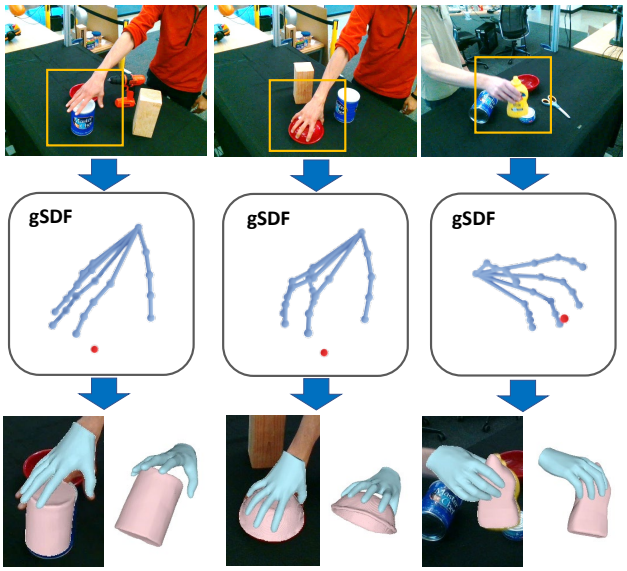


Figure 1. We aim to reconstruct 3D hand and object meshes from monocular images *(top)*. Our method gSDF *(middle)* first predicts 3D hand joints (blue) and object locations (red) from input images. We use estimated hand poses and object locations to incorporate strong geometric priors into SDF by generating hand- and object-aware kinematic features for each SDF query point. Our resulting gSDF model generates accurate results for real images with various objects and grasping hand poses *(bottom)*. We illustrate reconstructed hand and object meshes in both the camera viewpoint and a rotated viewpoint.

To reconstruct detailed hand and object meshes, another line of efforts [11, 26] employ signed distance functions (SDFs). Grasping Field [26] makes the first attempt to model hand and object surfaces using SDFs. However, it does not explicitly associate 3D geometry with image cues and has

no prior knowledge incorporated in SDFs, leading to unrealistic meshes. AlignSDF [11] proposes to align SDFs with respect to global poses (*i.e.*, the hand wrist transformation and the object translation) and produces improved results. However, it is still challenging to capture geometric details for dexterous hand motions with more degrees of freedom (DoF) on diverse objects.

To address limitations of prior works, we propose a geometry-driven SDF (gSDF) method that encodes strong pose priors and improves reconstruction by disentangling pose and shape estimation. (see Figure 1). To this end, we first predict sparse 3D hand joints from images and derive full kinematic chains of local pose transformations from joint locations using inverse kinematics. Instead of only using the global pose as in [11] to learn SDFs, we optimize SDFs with respect to poses of all the hand joints, which leads to a more fine-grained alignment between reconstructed 3D shapes and articulated hand poses. In addition, we project 3D points into the image plane to extract more geometry-aligned local visual features for signed distance prediction. The visual features are further refined with spatio-temporal contexts using a transformer model to enhance the robustness to occlusions and motion blurs.

We conduct extensive ablation experiments to show the effectiveness of different components in our approach. The proposed gSDF model greatly advances state-of-the-art accuracy on the challenging ObMan and DexYCB benchmarks. Our contributions can be summarized in three-fold: *(i)* To embed strong pose priors into SDFs, we propose to align the SDF shape with its underlying kinematic chains of pose transformations, which reduces ambiguities in 3D reconstruction. *(ii)* To further reduce the misalignment induced by inaccurate pose estimations, we propose to extract geometry-aligned local visual features and enhance the robustness with spatio-temporal contexts. *(iii)* We conduct comprehensive experiments to show that our approach outperforms state-of-the-art results by a significant margin.

2. Related Work

This paper focuses on jointly reconstructing hands and hand-held objects from RGB images. In this section, we first review prior work on the 3D hand pose and shape estimation. We then discuss relevant work on the reconstruction of hands and objects.

3D hand pose and shape estimation. The topic of 3D hand pose estimation has received widespread attention since the 90s [23, 45] and has seen significant progress in recent years [31, 65]. Methods which take RGB images as input [24, 36, 38, 39, 48, 50, 51, 53, 59, 67] often estimate sparse 3D hand joint locations from visual data using well-designed deep neural networks. Though these methods can achieve high estimation accuracy, their 3D sparse joints outputs provide limited information about the 3D hand surface, which is critical in AR/VR applications. Following the introduction of the anthropomorphic parametric hand mesh model MANO [46], several works [2, 5, 10, 18, 29, 30, 32, 34, 40, 57] estimate the MANO hand shape and pose parameters to recover the full hand surface. However, MANO has a limited mesh resolution and cannot produce fine surface details. Neural implicit functions [13, 25] have the potential to reconstruct more realistic high resolution hand surfaces [12, 37, 42]. In this work, we combine the advantages of sparse, parametric and implicit modelling. We predict sparse 3D joints accurately from images and estimate the MANO parameters using inverse kinematics. We then optimize neural implicit functions with respect to underlying kinematic structures and reconstruct realistic meshes.

3D hand and object reconstruction. Joint reconstruction of hand and object meshes provides a more comprehensive view about how hands interact with manipulated objects in the 3D space and has received more attention in the past few years. Previous works often rely on multiple view correspondence [3, 9, 17, 41, 58, 61] or additional depth information [15, 16, 49, 55, 56] to approach this task. In this work, we focus on a more challenging setting and perform joint reconstruction from monocular RGB images. Given the ill-posed nature of the problem, many works [7, 18–21, 54, 60, 62] deploy MANO, which encodes hand prior knowledge learned from hand scans, to reconstruct hand meshes. To further simplify the object reconstruction task, several works [18, 60, 62] make a strong assumption that the ground-truth object model is available at test time. Our work and some previous efforts [11, 21, 26] relax this assumption and assume unknown object models. Hasson *et al.* [21] employ a differentiable MANO layer to estimate the hand shape and AtlasNet [14] to reconstruct the manipulated object. However, both MANO and AtlasNet can only produce meshes of limited resolution, which prevents the modelling of detailed contacts between hands and objects. To generate more detailed surfaces, Karunratanakul *et al.* [26] introduce grasping fields and propose to use SDFs to reconstruct both hand and object meshes. However, such a model-free approach does not capture any prior knowledge about hands or objects, which can lead to predicting unrealistic 3D geometry. To mitigate this, Ye *et al.* [63] propose to use hand poses estimated from an off-the-shelf model to help reconstruct the hand-held object mesh. The main difference with our work is that we jointly reconstruct hand meshes and object meshes using our proposed model, which is more challenging. Also, in addition to using hand poses to help capture the object shapes, we predict object poses and show their benefits for SDF-based object reconstruction. Another work AlignSDF [11] optimizes SDFs with respect to estimated hand-object global poses and encodes pose priors into SDFs. In addition to using global poses as a guide for SDFs, we propose to learn SDFs from the full kinematic chains of local pose transformations, and

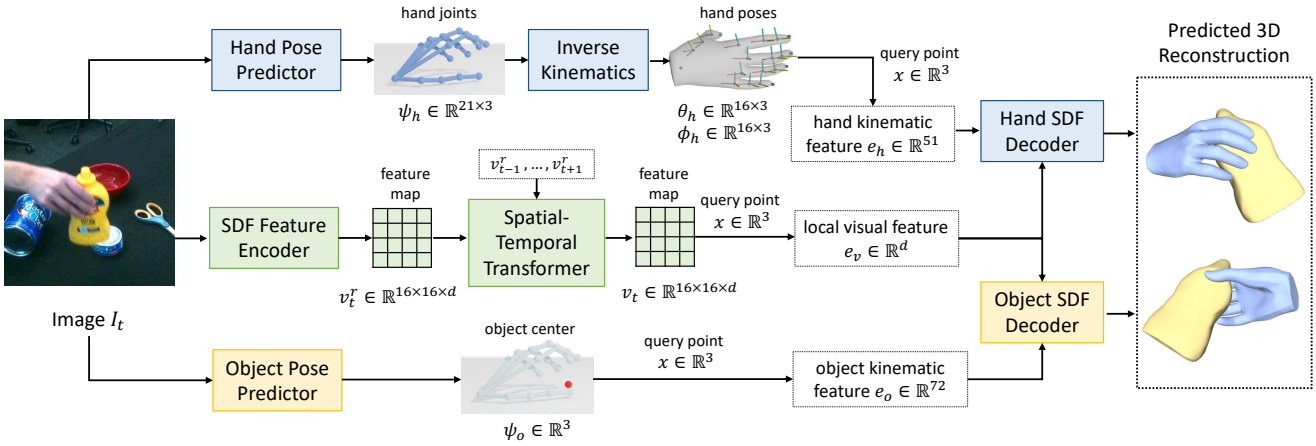


Figure 2. The overview of our proposed single-frame model. It can reconstruct realistic hand and object meshes from a single RGB image. The dashed arrows denote the Marching Cubes algorithm [33] used at test time.

achieve a more precise alignment between the 3D shape and the underlying poses. To further handle hard cases induced by occlusion or motion blur where pose estimations are inaccurate, we leverage a transformer to accumulate corresponding image features from multiple frames and benefit the geometry recovery.

3. Method

This section presents our geometry-driven SDF (gSDF) method for 3D hand and object reconstruction from monocular RGB images. We aim to learn two signed distance functions SDF_{hand} and SDF_{obj} to implicitly represent 3D shapes for the hand and the object. The SDF_{hand} and SDF_{obj} map a query 3D point $x \in \mathbb{R}^3$ to a signed distance from the hand surface and object surface, respectively. The Marching Cubes algorithm [33] can thus be employed to reconstruct hand and object from SDF_{hand} and SDF_{obj}.

3.1. Overview of gSDF

Figure 2 illustrates the overview of our gSDF reconstruction approach. Given an image I_t, we extract two types of features to predict the signed distance for each query point x, namely kinematic features and visual features.

The kinematic feature encodes the position of x under the coordinate system of hand or object, which can provide strong pose priors to assist SDFs learning. Since the feature is based on canonical hand and object poses, it also *disentangles* the shape learning from the pose learning.

The existing work [63] proposes to use hand poses to help reconstruct object meshes but does not consider using pose priors to reconstruct hand meshes. Another work [11] only deploys coarse geometry in terms of the hand wrist object locations, which fails to capture fine-grained details. In this work, we aim to strengthen the kinematic feature with geometry transformation of x to poses of all the hand joints (see Figure 3) for both hand and object reconstruction. However, it is challenging to directly predict hand pose parameters [6, 28, 66]. To improve the hand pose estimation, we propose to first predict sparse 3D joint locations j_h from the image and then use inverse kinematics to derive pose transformations θ_h from the predicted joints. In this way, we are able to obtain kinematic features e_h and e_o for hand and object respectively.

The visual feature encodes the visual appearance for the point x to provide more shape details. Prior works [11, 26] mainly use the same global visual feature for all the points, *e.g.*, averaging the feature map of a SDF feature encoder on the spatial dimension. Such global visual features suffers from imprecise geometry alignment between a point and its visual appearance. To alleviate the limitation, we propose to use the geometry information to extract aligned local visual features. Moreover, to address hard cases where there are occlusion and motion blur in a single image I_t, we further propose to enhance the local visual feature with its temporal contexts from videos using a spatio-temporal transformer. We denote the local visual feature of a point as e_v.

Finally, we concatenate the kinematic feature and local visual feature to predict the signed distance for x:

$$\begin{aligned} \text{SDF}_{hand}(x) &= f_h([e_v; e_h]), \\ \text{SDF}_{object}(x) &= f_o([e_v; e_o]), \end{aligned} \tag{1}$$

where f_h and f_o are the hand SDF decoder and the object SDF decoder respectively.

In the following, we first present the proposed geometry-driven kinematic feature and visual feature encodings in Section 3.2 and 3.3 respectively. Then, in Section 3.4 we introduce different strategies of sharing image backbones for hand and object pose predictors as well as the SDF feature

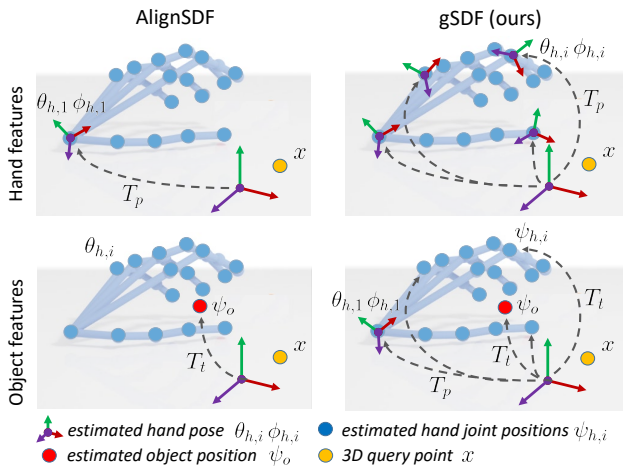


Figure 3. We define hand and object features by transforming queries x into hand- and object-centered coordinate systems. Compared to AlignSDF [11] (left), each hand joint in our method defines its own coordinate frame.

encoder. Finally, the training of our model is described in Section 3.5.

3.2. Kinematic Feature Encoding

Hand and object pose estimation. Directly regressing hand pose parameters of MANO from image features [11, 19, 21] has proved to be difficult [6, 28, 66]. In contrast, predicting sparse 3D joint locations is easier and can achieve higher accuracy. Therefore, we first train a 3D hand joint prediction model which produces volumetric heatmaps [38, 44] for 21 hand joints. We use a differentiable soft-argmax operator [50] to extract 3D coordinates $\psi_h \in \mathbb{R}^{21\times 3}$ of hand joints from the heatmaps. We then obtain an analytic solution for hand poses $\theta_h \in \mathbb{R}^{16\times 3}, \phi_h \in \mathbb{R}^{16\times 3}$ from estimated 3D joints ψ_h using inverse kinematics, where each $\theta_{h,i} \in \mathbb{R}^3$ and $\theta_{h,i} \in \mathbb{R}^3$ denote the relative pose of i_{th} joint in terms of rotation and translation with respect to its ancestor joint. Here, we only calculate the rotation and use the default limb lengths provided by the MANO model. Specifically, we first compute the pose of the hand wrist using the template pose defined in MANO, and then follow the hand kinematic chain to solve the pose of other finger joints recursively. More details are presented in the supplementary material.

For object pose estimation it is often difficult to accurately estimate the rotation of the object since many objects have a high degree of symmetry and are often occluded by the hand. We therefore follow [11] and only estimate the center position of the object $\psi_o \in \mathbb{R}^3$ relative to the hand wrist.

Hand kinematic feature. Given the 3D point x, we generate the hand kinematic feature $e_h \in \mathbb{R}^{51}$ by transforming x into canonical coordinate frames defined by hand joints. Figure 3(top,right) illustrates the proposed geometry transformation for the hand. For the i_{th} hand joint pose $\theta_{h,i}, \phi_{h,i}$, the pose transformation $T_p(x, \theta_{h,i}, \phi_{h,i})$ to obtain the local hand kinematic feature $e_{h,i} \in \mathbb{R}^3$ is defined as

$$G_{h,i} = \prod_{j\in A(i)} \left[\begin{array}{c|c} \exp(\theta_{h,j}) & \phi_{h,j} \\ \hline 0 & 1 \end{array}\right], \qquad (2)$$
$$e_{h,i} = T_p(x, \theta_{h,i}, \phi_{h,i}) = \widetilde{H}(G_{h,i}^{-1} \cdot H(x)),$$

where $A(i)$ denotes the ordered set of ancestors of the i_{th} joint. We use *Rodrigues formula* $\exp(\cdot)$ to convert $\theta_{h,i}$ into the form of a rotation matrix. By traversing the hand kinematic chain, we obtain the global transformation $G_{h,i} \in \mathbb{R}^{4\times 4}$ for the i_{th} joint. Then, we take the inverse of $G_{h,i}$ to transform x into the i_{th} hand joint canonical coordinates. $H(\cdot)$ transforms x into homogeneous coordinates while $\widetilde{H}(\cdot)$ transforms homogeneous coordinates back to Euclidean coordinates. Given local kinematic features $e_{h,i}$, the hand kinematic feature $e_h \in \mathbb{R}^{51}$ is defined as:

$$e_h = [x, e_{h,1}, \cdots, e_{h,16}]. \qquad (3)$$

Object kinematic feature. To obtain geometry-aware SDF for object recoinstruction we propose object kinematic feature $e_o \in \mathbb{R}^{72}$. Following [11], we use estimated object center ψ_o to transform x into the object canonical coordinate frame by the translation transformation $x_{oc} = T_t(x, \psi_o) = x - \psi_o$. As the grasping hand pose also gives hints about the shape of the manipulated object, similar to [63] we incorporate the knowledge of hand poses into object reconstruction. To this end, for each joint i and its estimated 3D location $\psi_{h,i}$ we transform x by translation as

$$e_{o,i} = T_t(x, \psi_{h,i}) = x - \psi_{j,i}. \qquad (4)$$

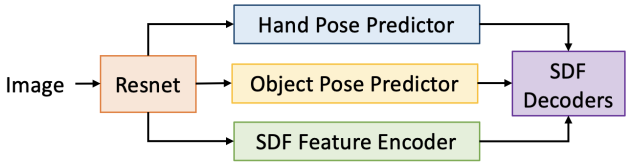
Given the importance of the wrist motion for object grasping, we also transform x into the canonical coordinate system of the hand wrist $x_{ow} = T_p(x, \theta_{h,1}, \phi_{h,1}) = \widetilde{H}(G_{h,1}^{-1} \cdot H(x))$, which normalizes the orientation of the grasping and further simplifies the task for the SDF object decoder. The object kinematic feature is then defined by $e_o \in \mathbb{R}^{72}$ as

$$e_o = [x, x_{oc}, e_{o,1}, \cdots, e_{o,21}, x_{ow}]. \qquad (5)$$

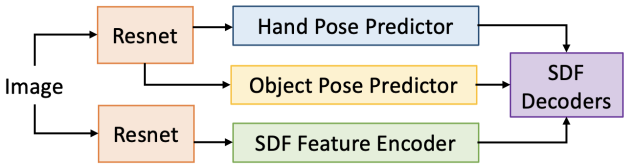
Figure 3(bottom,right) illustrates the proposed geometry transformation for the object kinematic feature.

3.3. Visual Feature Encoding

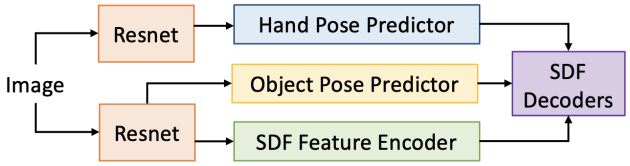
Geometry-aligned visual feature. Previous works [11, 26] typically predict signed distances from global image features that lack spatial resolution. Motivated by [47], we aim to generate geometry-aligned local image features for each input point x. Assume $v_t^r \in \mathbb{R}^{16\times 16\times d}$ is the feature map generated from the SDF feature encoder, *e.g.* a ResNet model [22],



(a) Single backbone.



(b) Symmetric backbone.



(c) Asymmetric backbone.

Figure 4. Illustrations of three image backbone sharing strategies.

where 16×16 is the spatial feature resolution and d is the feature dimension. We project the 3D input point x to $\hat{x}$ on the image plane with the camera projection matrix and use bilinear sampling to obtain a local feature e_v from the location on the feature map corresponding to $\hat{x}$.

Temporaly-enhanced visual feature. To improve the robustness of visual features in a single frame I_t from occlusion or motion blur, we propose to exploit temporal information from videos to refine v_t^r. We make use of the spatial-temporal transformer architecture [1, 4] to efficiently propagate image features across frames. Assume $v_{t-1}^r, \cdots v_{t+1}^r$ are the feature maps from neighboring frames of I_t in a video. We flatten all the feature maps as a sequence in the spatial-temporal dimension leading to $3 \times 16 \times 16$ tokens fed into the transformer model. We reshape the output features of the transformer into a feature map again for I_t, denoted as $v_t \in \mathbb{R}^{16\times16\times d}$. By aggregating spatial and temporal information from multiple frames, v_t becomes more robust to noise and can potentially produce more stable reconstruction results compared to v_t^r. Our full gSDF model relies on the feature map v_t to compute the local visual feature e_v for the given input point x.

3.4. Image Backbone Sharing Strategy

As shown in Figure 2, we have three branches for hand and object pose estimations and SDF feature encoding. These different branches can share the image backbones which might be beneficial with the multi-task learning. In this section, we describe three image backbone sharing strategies for the three branches.

Single image backbone (Figure 4a). We only employ one single image backbone for both pose and shape predictions. This is the strategy used in AlignSDF [11].

Symmetric image backbone (Figure 4b). To disentangle pose and shape learning, we share the image backbone for hand and object pose estimation, but use a different backbone to extract visual features for SDFs learning.

Asymmetric image backbone (Figure 4c). Since hand pose estimation plays an critical role in the task, we use a separate backbone to predict the hand pose, while share the image backbone for object pose predictor and SDF feature encoder.

3.5. Training

We apply a two-stage training strategy. In the first stage, we train the hand pose predictor to predict hand joint coordinates ψ_h with $\ell 2$ loss $\mathcal{L}_{hp}$ and an ordinal loss [43] $\mathcal{L}_{ord}$ to penalize the case if the predicted depth order between the i_{th} joint and the j_{th} joint is misaligned with the ground-truth relation $\mathbb{1}_{i,j}^{ord}$, which are:

$$\mathcal{L}_{hp} = \frac{1}{21}\sum_{i=1}^{21}\left\|\psi_{h,i} - \hat{\psi}_{h,i}\right\|_2^2, \tag{6}$$

$$\mathcal{L}_{ord} = \sum_{j=2}^{21}\sum_{i=1}^{j-1}\mathbb{1}_{i,j}^{ord} \times \left|(\psi_{h,i} - \psi_{h,j}) \cdot \vec{n}\right|, \tag{7}$$

where $\vec{n} \in \mathbb{R}^3$ denotes the viewpoint direction. We randomly sample twenty virtual views to optimize $\mathcal{L}_{ord}$. Since the proposed kinematic features are based on the predicted hand joints ψ_h, we empirically find that pretraining the hand joint predictor in the first stage and then freezing its weights can achieve better performance.

In the second training stage, we learn all the modules except the hand joint predictor in an end-to-end manner. We use the $\ell 2$ loss $\mathcal{L}_{op}$ to predict the object pose ψ_o as follows:

$$\mathcal{L}_{op} = \left\|\psi_o - \hat{\psi}_o\right\|_2^2 \tag{8}$$

where $\hat{\psi}_o$ denote the ground-truth location for the object center. To train the SDFs, we sample many 3D points around the hand-object surface and calculate their ground-truth signed distances to the hand mesh and the object mesh. We use $\ell 1$ loss to optimize the SDF decoders:

$$\begin{aligned}\mathcal{L}_{hsdf} &= \left\|\mathrm{SDF}_{hand} - \hat{\mathrm{SDF}}_{hand}\right\|_1^1, \\ \mathcal{L}_{osdf} &= \left\|\mathrm{SDF}_{obj} - \hat{\mathrm{SDF}}_{obj}\right\|_1^1,\end{aligned} \tag{9}$$

where $\hat{\mathrm{SDF}}_{hand}$ and $\hat{\mathrm{SDF}}_{obj}$ denote ground-truth signed distances to the hand and the object, respectively. The overall training objective $\mathcal{L}_{shape}$ in the second training stage is:

$$\mathcal{L}_{shape} = \mathcal{L}_{op} + 0.5 \times \mathcal{L}_{hsdf} + 0.5 \times \mathcal{L}_{osdf}. \tag{10}$$

4. Experiments

We conduct extensive experiments on two 3D hand-object reconstruction benchmarks to evaluate the effectiveness of our proposed gSDF model.

4.1. Datasets

ObMan [21] is a large-scale synthetic dataset that contains diverse hand grasping poses on a wide range of objects imported from ShapeNet [8]. We follow previous methods [11, 26, 42, 63] to generate data for SDFs training. First, we remove meshes that contain too many double-sided triangles, which results in 87,190 hand-object meshes. Then, we fit the hand-object mesh into a unit cube and sample 40,000 points inside the cube. For each sampled point, we compute its signed distance to the ground-truth hand mesh and object mesh, respectively. At test time, we report the performance on the whole ObMan test set of 6,285 testing samples.

DexYCB [9] is currently the largest real dataset that captures hand and object interactions in videos. Following [11,60], we focus on right-hand samples and use the official s0 split. We follow the same steps as in ObMan to obtain SDF training samples. To reduce the temporal redundancy, we downsample the video data to 6 frames per second, which results in 29,656 training samples and 5,928 testing samples.

4.2. Evaluation metrics

We follow prior works to comprehensively evaluate the 3D reconstructions with multiple metrics as below.

Hand Chamfer Distance ($\mathbf{CD_h}$). We evaluate Chamfer distance (cm^2) between our reconstructed hand mesh and the ground-truth hand mesh. We follow previous works [11, 26] to optimize the scale and translation to align the reconstructed mesh with the ground truth and sample 30,000 points on both meshes to compute Chamfer distance. We report the median Chamfer distance on the test set to reflect the quality of our reconstructed hand mesh.

Hand F-score ($\mathbf{FS_h}$). Since Chamfer distance is vulnerable to outliers [52, 63], we also report the F-score to evaluate the predicted hand mesh. After aligning the hand mesh with its ground truth, we report F-score at 1 mm ($\mathrm{FS_h}@1$) and 5 mm ($\mathrm{FS_h}@5$) thresholds.

Object Chamfer Distance ($\mathbf{CD_o}$). Following [11, 26], we first use the optimized hand scale and translation to transform the reconstructed object mesh. Then, we follow the same process as $\mathrm{CD_h}$ to compute $\mathrm{CD_o}$ (cm^2) and evaluate the quality of our reconstructed object mesh.

Object F-score ($\mathbf{FS_o}$). We follow the previous work [63] to evaluate the reconstructed object mesh using F-score at 5 mm ($\mathrm{FS_o}@5$) and 10 mm ($\mathrm{FS_o}@10$) thresholds.

Hand Joint Error ($\mathbf{E_h}$). To measure the hand pose estimation accuracy, we compute the mean joint error (cm) relative to the hand wrist over all 21 joints in the form of $\ell 2$ distance.

Table 1. Hand reconstruction performance with different hand kinematic features K_*^h and visual feature V_1 on DexYCB dataset.

	Wrist only	All joints	$\mathrm{CD_h}\downarrow$	$\mathrm{FS_h}@1\uparrow$	$\mathrm{FS_h}@5\uparrow$
K_1^h	×	×	0.364	0.154	0.764
K_2^h	✓	×	0.344	0.167	0.776
K_3^h	×	✓	**0.317**	**0.171**	**0.788**

Table 2. Object reconstruction performance with different object kinematic features K_*^o and visual feature V_1 on DexYCB dataset.

	Obj pose	Hand pose	$\mathrm{CD_o}\downarrow$	$\mathrm{FS_o}@5\uparrow$	$\mathrm{FS_o}@10\uparrow$
K_1^o	×	×	2.06	0.392	0.660
K_2^o	✓	×	1.93	0.396	0.668
K_3^o	✓	✓	**1.71**	**0.418**	**0.689**

Object Center Error ($\mathbf{E_o}$). To evaluate the accuracy of our predicted object translation, we report the $\ell 2$ distance (cm) between the prediction and its ground truth.

Additionally, we report Contact ratio (C_r), Penetration depth (P_d) and Intersection volume (I_v) [11, 21, 26, 60, 62] to present more details about the interaction between the hand mesh and the object mesh. Please see supplementary material for more details.

4.3. Implementation details

Model architecture. We use ResNet-18 [22] as our image backbone. For hand and object pose estimation, we adopt volumetric heatmaps of spatial resolution $64 \times 64 \times 64$ to localize hand joints and the object center in 3D space. For the spatial-temporal transformer, we use 16 transformer layers with 4 attention heads. We present more details about our model architecture in supplementary material.

Training details. We take the image crop of the hand-object region according to their bounding boxes for DexYCB benchmark. Then, we modify camera intrinsic and extrinsic parameters [35, 64] accordingly and take the cropped image as the input to our model. The spatial size of input images is 256×256 for all our models. We perform data augmentation including rotation ($\pm 45°$) and color jittering. During SDF training, we randomly sample 1000 points (500 points inside the mesh and 500 points outside the mesh) for the hand and the object, respectively. We train our model with a batch size of 256 for 1600 epochs on both ObMan and DexYCB using the Adam optimizer [27] with 4 NVIDIA RTX 3090 GPUs. We use an initial learning rate of 1×10^{-4} and decay it by half every 600 epochs. It takes 22 hours for training on DexYCB and 60 hours on ObMan dataset.

4.4. Ablation studies

We carry out ablations on the DexYCB dataset to validate different components in our gSDF model. We evaluate different settings of hand kinematic features (K_*^h in Table 1),

Table 3. Hand-object reconstruction performance with different visual features on DexYCB dataset. The visual features are combined with the best kinematic features K_3^h (Table 1) and K_3^o (Table 2) to reconstruct hand and object respectively.

	Global	Local	Transformer Spatial	Transformer Temp.	$CD_h \downarrow$	$FS_h@1 \uparrow$	$FS_h@5 \uparrow$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$E_h \downarrow$	$E_o \downarrow$
V_1	✓	×	×	×	0.317	0.171	0.788	1.71	0.418	0.689	1.44	**1.91**
V_2	×	✓	×	×	0.310	0.172	0.795	1.71	0.426	0.694	1.44	1.98
V_3	×	✓	✓	×	0.304	0.174	0.797	1.60	0.434	0.703	1.44	1.94
V_4	×	✓	✓	✓	**0.302**	**0.177**	**0.801**	**1.55**	**0.437**	**0.709**	**1.44**	1.96

Table 4. Hand-object reconstruction performance using different image backbone sharing strategies on DexYCB dataset. The ablation is carried out with visual features V_1 and kinematic features K_3^h and K_3^o.

Backbone	$CD_h \downarrow$	$FS_h@1 \uparrow$	$FS_h@5 \uparrow$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$E_h \downarrow$	$E_o \downarrow$
Single	0.401	0.112	0.560	2.36	0.307	0.521	2.01	2.16
Symmetric	0.324	0.168	0.779	1.84	0.405	0.672	1.46	1.93
Asymmetric	**0.317**	**0.171**	**0.788**	**1.71**	**0.418**	**0.689**	**1.44**	**1.91**

object kinematic features (K_*^o in Table 2), and visual features (V_* in Table 3). We use the asymmetric image backbone if not otherwise mentioned.

Hand kinematic feature. In Table 1, we evaluate the contribution of the proposed hand kinematic features for 3D hand reconstruction. The model in K_1^h does not use any pose priors to transform the 3D point. The model in K_2^h only uses the hand wrist pose to transform the 3D point as AlignSDF [11]. Our model in K_3^h computes the transformations to all the hand joints, which achieves the best performance on all the evaluation metrics. Compared to K_1^h without any pose priors, our model achieves more than 12% and 9% improvement on CD_h and $FS_h@1$ respectively. Compared to K_2^h with only hand wrist, our model greatly reduces the hand Chamfer distance from 0.344 cm^2 to 0.317 cm^2, leading to 7.8% relative gains. These results demonstrate the significance of pose priors and the advantage of gSDF for 3D hand reconstruction.

Object kinematic feature. In Table 2, we validate the effectiveness of our proposed object kinematic feature. The model in K_1^o does not contain any pose priors, while the model in K_2^o alignes query points to the object center as in [11]. Our model in K_3^o further employs the hand pose to produce the object kinematic feature, which significantly boosts the performance for the object reconstruction on different metrics. Compared to K_2^o, our proposed object kinematic feature achieves more than 11% and 5.5% improvement on CD_o and $FS_o@5$ respectively.

Visual features. We compare different visual features for SDF prediction in Table 3. V_1 uses the global visual feature *e.g.* the average pooling of ResNet feature map as in previous works [11,26]. Our local visual features with geometry alignment with the query point in V_2 reduces the hand Chamfer distance from 0.317 cm^2 to 0.310 cm^2. However, it shows less improvement on the object shape accuracy. In

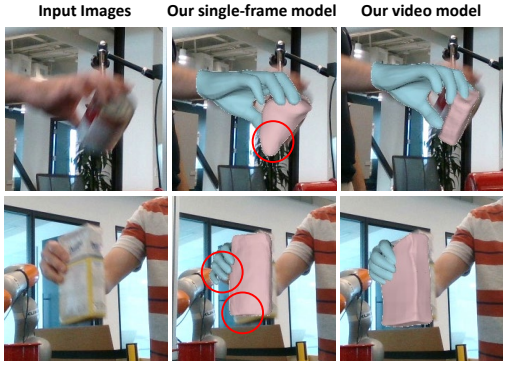


Figure 5. The qualitative comparsion between our single-frame model built with the transformer and our video model. By exploiting temporal information, our model can produce more robust 3D hand-object reconstruction results.

V_3 and V_4, we use the transformer model to refine the feature maps. To ablate the improvement from the transformer architecture and from the temporal information in videos, we only use transformer for each single frame in V_3 while use it for multiple frames in V_4. We can see that the transformer architecture alone is beneficial for the reconstruction. Enhancing the visual features with temporal contexts further improves the performance in terms of all the evaluation metrics especially for the objects. In Figure 5, compared with our single-frame model built with the transformer, our video model can make more robust predictions under some hard cases (*e.g.,* occlusion and motion blur).

Image backbone sharing strategy. The results of different image backbone sharing strategies are presented in Table 4. The model with one single backbone achieves the worst performance both in pose estimation accuracy and shape reconstruction accuracy. This is because the pose learning

Table 5. Comparison with state-of-the-art methods on the image ObMan dataset.

Methods	$CD_h \downarrow$	$FS_h@1 \uparrow$	$FS_h@5 \uparrow$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$E_h \downarrow$	$E_o \downarrow$
Hasson *et al.* [21]	0.415	0.138	0.751	3.60	0.359	0.590	1.13	-
Karunratanakul *et al.* [26]	0.261	-	-	6.80	-	-	-	-
Ye *et al.* [63]	-	-	-	-	0.420	0.630	-	-
Chen *et al.* [11]	0.136	0.302	0.913	3.38	0.404	0.636	1.27	**3.29**
gSDF (Ours)	**0.112**	**0.332**	**0.935**	**3.14**	**0.438**	**0.660**	**0.93**	3.43

Table 6. Comparison with state-of-the-art methods on the video DexYCB dataset.

Methods	$CD_h \downarrow$	$FS_h@1 \uparrow$	$FS_h@5 \uparrow$	$CD_o \downarrow$	$FS_o@5 \uparrow$	$FS_o@10 \uparrow$	$E_h \downarrow$	$E_o \downarrow$
Hasson *et al.* [21]	0.537	0.115	0.647	1.94	0.383	0.642	1.67	-
Karunratanakul *et al.* [26]	0.364	0.154	0.764	2.06	0.392	0.660	-	-
Chen *et al.* [11]	0.358	0.162	0.767	1.83	0.410	0.679	1.58	**1.78**
gSDF (Ours)	**0.302**	**0.177**	**0.801**	**1.55**	**0.437**	**0.709**	**1.44**	1.96

and shape learning compete with each other during training. The symmetric strategy to separate backbones for pose and SDFs performs better than the single backbone model. Our asymmetric strategy with a separate backbone for hand pose estimation but a shared backbone for object pose and SDF feature encoder achieves the best performance. We also empirically find that learning the object pose and SDFs together promotes both the pose accuracy and the shape accuracy. The possible reason is that estimating object pose also helps our model to focus on regions of the hand-object interaction and boosts the 3D reconstruction accuracy.

4.5. Comparison with state of the art

We compare our gSDF model with state-of-the-art methods on ObMan and DexYCB benchmarks. In Figure 6, we qualitatively demonstrate our approach can produce convincing 3D hand-object reconstruction results.

ObMan. Table 5 shows the comparison of hand and object reconstruction results on the synthetic ObMan dataset. Since ObMan does not contain video data, we do not use the spatial-temporal transformer in this model. The proposed gSDF outperforms previous methods by a significant margin. Compared with the recent method [63] that only reconstructs hand-held objects, our joint method produces more accurate object meshes. gSDF achieves a 17.6% improvement on CD_h and a 7.1% improvement on CD_o over the state-of-the-art accuracy, which indicates that our model can better reconstruct both hand meshes and diverse object meshes.

DexYCB. Table 6 presents results on the DexYCB benchmark. Our model demonstrates a large improvement over recent methods. In particular, it advances the state-of-the-art accuracy on CD_h and CD_o by 15.6% and 15.3%, respectively. The high accuracy of gSDF on DexYCB demonstrates that our approach generalizes well to real images.

Figure 6. Qualitative results of our model on test images from the ObMan and DexYCB benchmarks. Our model produces convincing results for different grasping poses and diverse objects.

5. Conclusion

In this work, we propose a geometry-driven SDF (gSDF) approach for 3D hand and object reconstruction. We explicitly model the underlying 3D geometry to guide the SDF learning. We first estimate poses of hands and objects according to kinematic chains of pose transformations, and then derive kinematic features and local visual features using the geometry information for signed distance prediction. Extensive experiments on ObMan and DexYCB datasets

demonstrate the effectiveness of our proposed method. In the future, our approach could incorporate stronger object priors to facilitate reconstruction of particular object categories.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, 2021. 5

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for RGB-based dense 3D hand pose estimation via neural rendering. In *CVPR*, 2019. 2

[3] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021. 5

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3D hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[6] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. In *3DV*, 2021. 3, 4

[7] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 1, 2

[8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. ShapeNet: An information-rich 3D model repository. *arXiv preprint arXiv:1512.03012*, 2015. 6

[9] 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 *CVPR*, 2021. 1, 2, 6

[10] Xingyu Chen, Yufeng Liu, Chongyang Ma, Jianlong Chang, Huayan Wang, Tian Chen, Xiaoyan Guo, Pengfei Wan, and Wen Zheng. Camera-space hand mesh recovery via semantic aggregation and adaptive 2D-1D registration. In *CVPR*, 2021. 2

[11] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. AlignSDF: Pose-Aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[12] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 2

[13] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. LISA: Learning implicit shape and appearance of hands. In *CVPR*, 2022. 2

[14] Thibault Groueix, Matthew Fisher, Vladimir G Kim, Bryan C Russell, and Mathieu Aubry. A papier-mâché approach to learning 3D surface generation. In *CVPR*, 2018. 2

[15] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 2

[16] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009. 2

[17] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. HOnnotate: A method for 3D annotation of hand and object poses. In *CVPR*, 2020. 1, 2

[18] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Keypoint Transformer: Solving joint identification in challenging hands and object interactions for accurate 3D pose estimation. In *CVPR*, 2022. 1, 2

[19] 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 *CVPR*, 2020. 1, 2, 4

[20] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from RGB videos. In *3DV*, 2021. 2

[21] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 4, 6, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[23] Tony Heap and David Hogg. Towards 3D hand tracking using a deformable model. In *FG*, 1996. 2

[24] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5D heatmap regression. In *ECCV*, 2018. 1, 2

[25] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *3DV*, 2021. 2

[26] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping Field: Learning implicit representations for human grasps. In *3DV*, 2020. 1, 2, 3, 4, 6, 7, 8

[27] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[28] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3D human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3, 4

[29] Dominik Kulon, Riza Alp Güler, I. Kokkinos, M. Bronstein, and S. Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 2

[30] Dominik Kulon, Haoyang Wang, Riza Alp Güler, Michael M. Bronstein, and Stefanos Zafeiriou. Single image 3D hand reconstruction with mesh convolutions. In *BMVC*, 2019. 2

[31] Vincent Lepetit. Recent advances in 3D object and hand pose estimation. *arXiv preprint arXiv:2006.05927*, 2020. 2

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

[33] William E Lorensen and Harvey E Cline. Marching Cubes: A high resolution 3D surface construction algorithm. *TOG*, 1987. 3

[34] Jun Lv, Wenqiang Xu, Lixin Yang, Sucheng Qian, Chongzhao Mao, and Cewu Lu. HandTailor: Towards high-precision monocular 3D hand recovery. In *BMVC*, 2021. 2

[35] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3D human pose estimation in the wild using improved CNN supervision. In *3DV*, 2017. 6

[36] Hao Meng, Sheng Jin, Wentao Liu, Chen Qian, Mengxiang Lin, Wanli Ouyang, and Ping Luo. 3D interacting hand pose estimation by hand de-occlusion and removal. In *ECCV*, 2022. 2

[37] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy Networks: Learning 3D reconstruction in function space. In *CVPR*, 2019. 2

[38] 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 *CVPR*, 2018. 1, 2, 4

[39] 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 *CVPR*, 2018. 2

[40] 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. *TOG*, 2019. 2

[41] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full DOF tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 2

[42] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2, 6

[43] Georgios Pavlakos, Xiaowei Zhou, and Kostas Daniilidis. Ordinal depth supervision for 3D human pose estimation. In *CVPR*, 2018. 5

[44] Georgios Pavlakos, Xiaowei Zhou, Konstantinos G Derpanis, and Kostas Daniilidis. Coarse-to-fine volumetric prediction for single-image 3D human pose. In *CVPR*, 2017. 4

[45] James M Rehg and Takeo Kanade. Visual tracking of high DOF articulated structures: an application to human hand tracking. In *ECCV*, 1994. 2

[46] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied Hands: Modeling and capturing hands and bodies together. *TOG*, 2017. 1, 2

[47] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. PiFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV*, 2019. 4

[48] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3D hand pose estimation from monocular RGB via contrastive learning. In *ICCV*, 2021. 2

[49] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from RGB-D input. In *ECCV*, 2016. 2

[50] Xiao Sun, Bin Xiao, Fangyin Wei, Shuang Liang, and Yichen Wei. Integral human pose regression. In *ECCV*, 2018. 2, 4

[51] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3D articulated hand posture. In *CVPR*, 2014. 1, 2

[52] Maxim Tatarchenko, Stephan R Richter, René Ranftl, Zhuwen Li, Vladlen Koltun, and Thomas Brox. What do single-view 3D reconstruction networks learn? In *CVPR*, 2019. 6

[53] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In *CVPR*, 2019. 1, 2

[54] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *CVPR*, 2022. 1, 2

[55] Aggeliki Tsoli and Antonis A Argyros. Joint 3D tracking of a deformable object in interaction with a hand. In *ECCV*, 2018. 2

[56] Dimitrios Tzionas and Juergen Gall. 3D object reconstruction from hand-object interactions. In *ICCV*, 2015. 2

[57] 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. *TOG*, 2020. 2

[58] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. Video-based hand manipulation capture through composite motion control. *TOG*, 2013. 2

[59] Fu Xiong, Boshen Zhang, Yang Xiao, Zhiguo Cao, Taidong Yu, Joey Tianyi Zhou, and Junsong Yuan. A2J: Anchor-to-joint regression network for 3D articulated pose estimation from a single depth image. In *ICCV*, 2019. 2

[60] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. ArtiBoost: Boosting articulated 3D hand-object pose estimation via online exploration and synthesis. In *CVPR*, 2022. 2, 6

[61] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. OakInk: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022. 1, 2

[62] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. CPF: Learning a contact potential field to model the hand-object interaction. In *ICCV*, 2021. 1, 2, 6

[63] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3D reconstruction of generic objects in hands. In *CVPR*, 2022. 2, 3, 4, 6, 8

[64] Frank Yu, Mathieu Salzmann, Pascal Fua, and Helge Rhodin. PCLs: Geometry-aware neural reconstruction of 3D pose with perspective crop layers. In *CVPR*, 2021. 6

[65] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, Junsong

Yuan, Xinghao Chen, Guijin Wang, Fan Yang, Kai Akiyama, Yang Wu, Qingfu Wan, Meysam Madadi, Sergio Escalera, Shile Li, Dongheui Lee, Iason Oikonomidis, Antonis Argyros, and Tae-Kyun Kim. Depth-based 3D hand pose estimation: From current achievements to future goals. In *CVPR*, June 2018. 2

[66] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *CVPR*, 2019. 3, 4

[67] Christian Zimmermann and Thomas Brox. Learning to estimate 3D hand pose from single RGB images. In *ICCV*, 2017. 1, 2