CPF: Learning a Contact Potential Field to Model the Hand-Object Interaction

Lixin Yang[1,2] Xinyu Zhan[1] Kailin Li[1] Wenqiang Xu[1,2] Jiefeng Li[1] Cewu Lu[1,2,†]
[1]Shanghai Jiao Tong University, China [2]Shanghai Qi Zhi Institute
{siriusyang, kelvin34501, kailinli, vinjohn, ljf_likit, lucewu}@sjtu.edu.cn

Abstract

Modeling the hand-object (HO) interaction not only requires estimation of the HO pose, but also pays attention to the contact due to their interaction. Significant progress has been made in estimating hand and object separately with deep learning methods, simultaneous HO pose estimation and contact modeling has not yet been fully explored. In this paper, we present an explicit contact representation namely Contact Potential Field (CPF), and a learning-fitting hybrid framework namely MIHO to Modeling the Interaction of Hand and Object. In CPF, we treat each contacting HO vertex pair as a spring-mass system. Hence the whole system forms a potential field with minimal elastic energy at the grasp position. Extensive experiments on the two commonly used benchmarks have demonstrated that our method can achieve state-of-the-art in several reconstruction metrics, and allow us to produce more physically plausible HO pose even when the ground-truth exhibits severe interpenetration or disjointedness. Our code is available at https://github.com/lixiny/CPF.

1. Introduction

It is essential to model hand-object interaction from a single image for understanding the human activities, in which simulating a physically plausible grasp is also crucial for VR/AR, teleoperation, and grasping applications. Given an image as input, the problem aims to not only estimate proper hand-object pose but also to recover a natural grasp configuration. While estimating hand [36, 31, 58, 4, 19, 56] or object [20, 23, 14, 54, 55] alone has made a considerable success over the past decades, simultaneously estimating hand-object pose [24, 51, 23, 28, 12] with interaction has only emerged in the past few years.

Previous works on joint hand-object estimation usually treat the contact as a result of the correct pose estimation [23, 29, 44]. Apparently, if the hand and object can be perfectly recovered, the contact between them will also be satisfied. Yet, such perfection cannot be achieved in practice. Since contact can provide rich cues to guide accurate pose and natural grasp, more attention has been recently drawn to the contact modeling [5, 7] and contact representation [26, 6]. And several contact datasets [5, 7, 50] have been released to the community. However, a solution of properly integrating contact modeling into the current hand-object pose estimation pipeline has remained an open research question. The existing methods either exploit distance-based attraction and repulsion [24, 26] to mitigate disjointedness and interpenetration, or refine the predicted pose in virtue of physics simulators [28, 29, 17]. While the both solutions are considered to be irrelevant to contact *semantics*, which we will explain later, the latter solutions also lack flexibility on hand pose and shape.

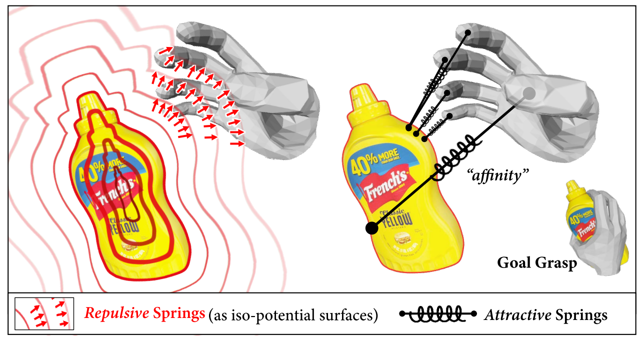

Figure 1. **Illustration of the proposed Contact Potential Field.** The contacts between hand and object vertices are modeled as the attractive (right) and repulsive (left) springs that connect paired vertex on them.

To model the contact, we propose an explicit representation named **C**ontact **P**otential **F**ield (**CPF**, §4). It is built upon the idea that the contact between a hand and an object mesh under grasp configuration is multi-point contact, which involves multiple hand-object vertex pair affinities. These affinities are regarded as the contact *semantics*, which depict the pairing of the hand-object vertices that come into contact with each other during the interaction. When noisy predicted hand and object are disjointed from each other, we shall apply an attraction to pull these vertex pairs close; While the hand and object are intersected, we shall have a

[†]Cewu Lu is the corresponding author. He is the member of Qing Yuan Research Institute and MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, China.

repulsion to push them away. Contacts of those affinitive vertex pairs are the result of equilibrium between the attraction and repulsion. In this paper, we treat each contacting HO vertex pair as a spring-mass system. First, the two end-points of spring is a counterpart of the two HO vertices in affinity. Second, the spring's elastic property is another counterpart of the intensity of the vertex pair affinity. In this way, we can model the HO interaction with a potential field, as we call it CPF, which is determined by minimal elastic energy at the grasp position. Therefore, estimating the HO pose under contact is equivalent to minimizing the elastic energy inside CPF. Representing contact as CPF has two main advantages. First, compared with contact heuristic with proximity metrics [1, 52] or distance field [26, 6], CPF is able to assign per-vertex contact *semantics* (contact points on different hand part) to object mesh. Second, by minimizing the elastic energy, CPF can uniformly avoid interpenetration and control the disjointedness. Based on CPF, we also propose a novel learning-fitting hybrid framework namely for **M**odeling the **I**nteraction of **H**and and **O**bject, as we call it **MIHO** (§5).

Another problem with the existing methods is the representation of the hand model. Most researches adopted a skinning model, MANO [47], to represent hand. MANO is considered to be flexible and deformable with its pose and shape parameters. However, fitting on these high DoFs parameters is prone to anatomical abnormality. Researches in the robotics community adopted a dexterous hand [29, 17] in the off-the-shelf grasping software [35], which can almost guarantee a valid pose. But the rigidity of those rod-like hand is less suitable for applications in CV/CG. To make the best of both worlds, we propose a novel anatomically constrained hand model namely A-MANO (§3). It inherits the formulation of the skinning model and constrains the hand joints' rotation within a proposed *twist-splay-bend* frame (Fig. 2).

For evaluation, we report our scores on FHB [18] and HO3D [22, 21] dataset in terms of reconstruction and physical quality metrics. Note that, the ground truth of FHB is noisy and suffers from severe interpenetration [26]. Since our method can avoid the penetration in the first place, our results are more visually and physically plausible. Therefore, we argue that, in this dataset, a higher reconstruction score does not necessarily benchmark the performance of the method. While on HO3D, we achieve state-of-the-art performance on both reconstruction and physical metrics. The contributions of this paper are as follows.

- We highlight contact in the hand-object interaction modeling task by proposing an explicit representation named CPF.
- We introduce A-MANO, a novel anatomical-constrained hand model that helps to mitigate pose's abnormality during optimization.
- We present a novel framework, MIHO, for modeling hand-object interaction. It can achieve state-of-the-art performance on several benchmarks.

2. Related Work

3D Hand Reconstruction. Most of the existing 3D hand reconstruction methods [4, 58, 2] adopted a parametric skinning hand, *e.g.* MANO [47] as a template. To drive MANO, it is crucial to obtain joint rotation along hand kinematic tree. Boukhayma *et al.* [4] firstly proposed to regress the PCA components of the rotations. Later, directly regressing the full rotations from 3D positions [58, 56] has shown better performance. However, those high DoF regression is prone to pose abnormality. Thus, Spurr *et al.* [49] exploited biomechanical constraints over hand joints in training scheme. Different from [49], we apply rotation constraints over the axes and angles in the proposed *twist-splay-bend* coordinate frame.

Hand-object Pose Estimation. In a wide range of topics in modeling hand-object interaction, the most commonly referred one is HO pose estimation [24, 23, 12, 16, 51]. In this regard, the earlier methods focused on either hand [43, 45, 52] or object [53] pose alone, or estimated hand in grasping pose with knowing object shape prior [15, 8, 9, 10]. Jointly estimating hand and object pose was firstly presented by Romero *et al.* [46] via searching for nearest neighbors in a large database. Recently, learning-based frameworks have emerged in this area. Hasson *et al.* [24, 23] proposed two learning frameworks to recover hand-object meshes, one by synthesizing HO data under manipulation [24] and the other by exploiting photometric consistency over video sequence [23]. Doosti *et al.* [12] employed the graph neural networks [16] to lift the 2D HO keypoints into 3D space. Tekin *et al.* [51] adopted 3D YOLO [41] to predict HO pose in one stage. Korrawe *et al.* [26] recovered HO model in a form of Signed Distance Function [39].

Contact Heuristic. Exploiting contact heuristic in hand-object interaction can be traced back to several decades before [42, 13, 34]. Early works utilized some shape-specified contact physics (*e.g.* cones and blocks [42]) or predefined grasp [34] as prior. Studies on capturing [30] or imitating [3] HO interaction also leveraged contact to satisfy the reality. Later, the studies on grasping synthesis [57, 17, 29] and tracking [38, 32] turned to physical simulators to circumvent model's intersection. Multi-point contact formulation was proposed in [27, 25, 1], which we found useful when applying physical constraints, *e.g.* [27, 25] used contact points to resolve penetration. For unified attraction and repulsion, most works employed heuristic such as proximity metric [24, 1, 52], signed distance function [26, 6], predefined contact pattern [44, 6], or turned to simulator [28, 29] for simplicity. Recently, Antotsiou *et al.* [1] refined

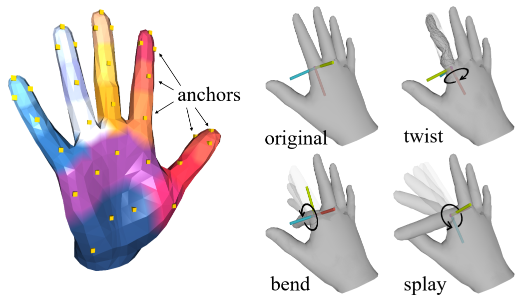

Figure 2. **Illustration of the proposed A-MANO.** Left: the subdivision of hand regions and anchors attached to it. Right: the proposed *twist-splay-bend* frame.

the grasp by attracting fingers to its nearest point on object surface w.r.t distance-based energy. Hasson *et al.* [24] applied well-designed interaction losses which are also based on proximity metric. Although our method differs from all of the previous methods in terms of contact heuristic, we consider that both [1] and [24] are still strong baselines. Thus we will compare our contact heuristic with theirs.

3. Anatomically Constrained A-MANO

The proposed A-MANO inherits from a parametric skinning hand model, MANO [47], which drives an articulated hand mesh with pose parameters $\boldsymbol{\theta}$ and shape parameters $\boldsymbol{\beta}$. $\boldsymbol{\theta} \in \mathbb{R}^{15\times3}$ is 15 joint rotations along the hand kinematic tree. And $\boldsymbol{\beta} \in \mathbb{R}^{10}$ represents the PCA components of hand shape. The main differences of A-MANO from MANO are: 1) the restriction on the joints' rotation axes and angles within the *twist-splay-bend* frame; 2) the *anchor* representation in the subdivision of hand region.

The *Twist-splay-bend* Frame. Fitting on 15 joint rotations of MANO requires high DoFs regression which may cause abnormal hand posture as shown in Fig. 7. Since the human hand can be modeled in a kinematic tree, and the majority of the joints only have one DoF about the *bend* axis, we can impose constraints over the rotation about the unwanted axes. Therefore the proposed *twist-splay-bend* Cartesian coordinate frame can be assigned to each joint along the kinematic tree. The frame' s x, y, z axes are coaxial to the 3 revolute directions: *twist*, *splay*, and *bend* direction on the basis of hand anatomy (Fig. 2 right). Then we can impose axial constraints in the *twist* and *splay* axes, and impose angular constraints w.r.t the *bend* angle. Details of the *twist-splay-bend* frame are elaborated in *Supp* A.1.

Anchors. Since the hand mesh of different subjects are almost identical in the subdivision of hand region (*e.g.* phalanges), we can interpolate several representative points (later we call it *anchors*) on hand mesh to largely reduce the number of HO vertex pairs. Instead of attaching springs from object mesh to all the affinitive vertices on hand mesh, we only attach them on the several hand subregion centers, as we call it *anchors* (Fig. 2 left). According to the statistics [24, 7] on the contact frequency of different hand parts, we first divide the full hand palm into 17 subregions: 3 for each phalange of 5 fingers, 1 for metacarpals, and another for carpals. Then, we interpolate up to 4 anchors for each subregion. We ignore all the vertices on the back side of the hand. Details of subregion division and anchors interpolation are described in *Supp* A.2, A.3.

4. Contact Potential Field

Contact as Spring-Mass System. A single contact is modeled as a spring-mass system which consists of a spring and two mass points on each side (hand and object). When the spring is at its rest position, it does not store energy, whilst it is stretched or compressed, according to Hooke's Law[1], it will store the elastic potential energy with the form: $\frac{1}{2}k|\boldsymbol{\Delta l}|^2$, where k is the spring elasticity, and $|\boldsymbol{\Delta l}|$ is a certain "*distance*" metric w.r.t. the spring's rest position.

In CPF, we define two types of spring: ***attractive*** spring and ***repulsive*** spring. The goal of *attractive* spring is to pull the hand vertex $\boldsymbol{v}^h$ toward the object vertex $\boldsymbol{v}^o$ based on a given HO vertex pair affinity. And the goal of *repulsive* spring is to push the $\boldsymbol{v}^h$ away from $\boldsymbol{v}^o$ along the $\boldsymbol{v}^o$'s normal if the $\boldsymbol{v}^h$ is in the vicinity of $\boldsymbol{v}^o$. Apart from these definitions, we should also point out that the *attractive* spring is bound with a certain pair of HO vertex affinity, while the *repulsive* spring only takes effect in the neighborhood of HO vertex pairs at some point.

- Attractive Spring. We define the rest length of *attractive* spring as 0 in which the hand vertex and object vertex are in perfect contact, and the *distance* metric $|\boldsymbol{\Delta l}|$ as Euclidean distance. Given a HO affinity that includes a vertex pair: $\boldsymbol{v}_i^h$ and $\boldsymbol{v}_j^o$, the $|\boldsymbol{\Delta l}_{ij}^{\mathrm{atr}}|$ is equal to $\left\|\boldsymbol{v}_i^h - \boldsymbol{v}_j^o\right\|_2$. The potential energy of the current *attractive* spring is given by:

$$E_{ij}^{\mathrm{atr}} = \frac{1}{2}k_{ij}^{\mathrm{atr}} * \left\|\boldsymbol{\Delta l}_{ij}^{\mathrm{atr}}\right\|_2^2 \tag{1}$$

- Repulsive Spring. We hope that the repulsion energy is high when $\boldsymbol{v}_i^h$ is penetrating or in the vicinity of $\boldsymbol{v}_j^o$, but gradually decays as the $\boldsymbol{v}_i^h$ moves away from the object, and finally becomes negligible at certain distance. Given a proximate HO vertex pair: $\boldsymbol{v}_i^h$ and $\boldsymbol{v}_j^o$, We define a repulsive spring to model this behavior. Supposing that the *repulsive* spring has the rest position at $+\infty$ away along the object normal $\mathbf{n}_j^o$. We adopt a heuristic *distance* metric $|\boldsymbol{\Delta l}| = e^{-|\boldsymbol{\Delta l}_{ij}^{\mathrm{rpl}}|} - e^{-\infty} = e^{-|\boldsymbol{\Delta l}_{ij}^{\mathrm{rpl}}|}$, where $|\boldsymbol{\Delta l}_{ij}^{\mathrm{rpl}}| = (\boldsymbol{v}_i^h - \boldsymbol{v}_j^o) \cdot \mathbf{n}_j^o$ is the projection of the $(\boldsymbol{v}_i^h - \boldsymbol{v}_j^o)$ on the object normal $\mathbf{n}_j^o$. Thus, the potential energy of the current *repulsive* spring is

$$E_{ij}^{\mathrm{rpl}} = \frac{1}{2}k_{ij}^{\mathrm{rpl}} * \left(e^{-|\boldsymbol{\Delta l}_{ij}^{\mathrm{rpl}}|}\right)^2 \tag{2}$$

[1] https://en.wikipedia.org/wiki/Hookes_law

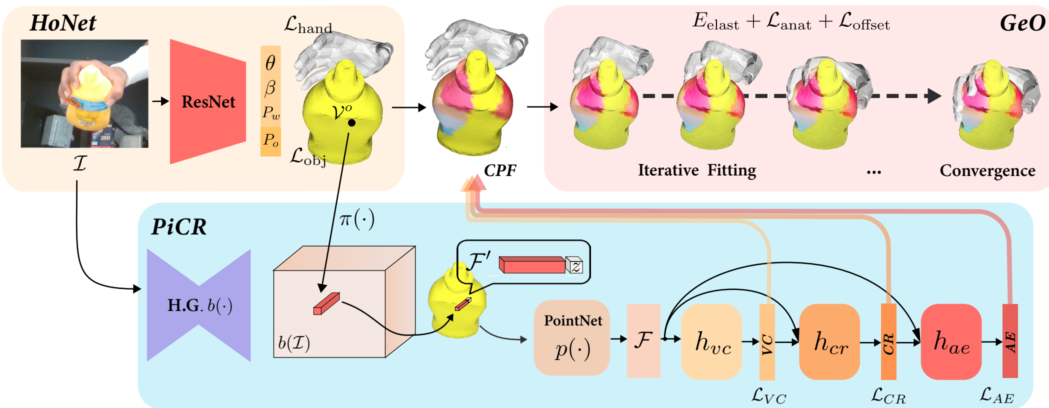

Figure 3. **The architecture of the hybrid model MIHO.** The MIHO consists of three submodules: the first HONet estimates coarse poses of HO meshes, the second PiCR learns to recover the CPF and the last GeO retrieves the refined poses based on the CPF.

In literature, adopting repulsive effect along surface normal can be found in [6, 22]. [22] (Eq. 10) also discussed that $e^{-(\cdot)}$ is an efficient heuristic concerning sub-sampled set of vertices.

Grasping inside Contact Potential Field. By collecting all the *attractive* and *repulsive* springs, to form a natural grasp is equivalent to minimize the elastic energy:

$$E_{\text{elast}} = \sum_i \sum_j (E_{ij}^{\text{atr}} + E_{ij}^{\text{rpl}}) \tag{3}$$

As discussed in §3, the hand vertices can be simplified to subregion $anchors$, which will largely relax the difficulty of learning and fitting inside the CPF. Thus, for *attractive* spring, we replace the $\boldsymbol{\Delta l}_{ij}$ in Eq.1 to $\boldsymbol{\Delta l}'_{ij} = \boldsymbol{a}_i - \boldsymbol{v}_j^o$, where $\boldsymbol{a}_i$ is the closest anchor to $\boldsymbol{v}_i^h$. Besides, we would like to have the repulsion force be only applied to those HO affinity pairs that are of vertices in vicinity. Thus we set zero the repulsion energy when the vertex distance $\|\boldsymbol{v}_j^o - \boldsymbol{v}_i^h\|_2$ is greater than a threshold $t_{\text{rpl}} = 20\ mm$.

Annotation of the *Attractive* Springs (k^{atr}). While the attraction energy is bound with certain HO affinities, the repulsion energy is rather ambient and affinity-agnostic. To integrate the CPF into learning framework, we only consider the k_{ij}^{atr} as the prediction of neural network. To enable this, network shall have the abilities of 1) pairing the hand anchors and object vertices into HO affinity pair, *e.g.* $(\boldsymbol{a}_i, \boldsymbol{v}_j^o)$; and 2) regressing the intensity of those affinity pairs, *e.g.* k_{ij}^{atr}. These require annotation of the *attractive* springs k^{atr}.

Given the ground-truth (*gt.*) HO pose and their mesh model, we automatically annotate each k_{ij}^{atr} based on a heuristic of the $(\boldsymbol{a}_i, \boldsymbol{v}_j^o)$ pair distance. Since each $\boldsymbol{a}_i$ may be included in several affinity pairs, we hope the attraction energy stored in each spring at *gt.* HO pose is well balanced. Thus we assign the *gt.* $\hat{k}_{ij}^{\text{atr}}$ a value that is inverse-proportional to the *gt.* $|\boldsymbol{\Delta}\hat{\boldsymbol{l}}_{ij}^{\text{atr}}|$. In order to train the network, we also bound the magnitude of $\hat{k}_{ij}^{\text{atr}}$ by 0 and 1. Here we only provide a glimpse of the annotation heuristic of $\hat{k}_{ij}^{\text{atr}}$:

$$\hat{k}_{ij}^{\text{atr}} = 0.5 * \cos\left(\frac{\pi}{s} * |\boldsymbol{\Delta}\hat{\boldsymbol{l}}_{ij}^{\text{atr}}|\right) + 0.5 \tag{4}$$

Empirically, we set the scale factor $s = 20\ mm$ and reject those HO affinities with *gt.* $|\boldsymbol{\Delta}\hat{\boldsymbol{l}}_{ij}^{\text{atr}}| \geq 20\ mm$. As for the elasticity of *repulsive* spring, we empirically set all k_{ij}^{rpl} to 1×10^{-3}. Detailed analysis of the *gt.* $\hat{k}^{\text{atr}}$ and the attraction-repulsion equilibrium are provided in *Supp* B.1, B.2.

5. Hybrid Framework – MIHO

With respect to the proposed CPF (§4), our approach MIHO models the hand-object interaction in three stages, namely HoNet (§5.1), PiCR (§5.2), and GeO (§5.3).

As shown in Fig. 3, firstly, given an RGB image $\mathcal{I}$, HoNet predicts a coarse pose of hand mesh $\mathcal{V}^h = \{\boldsymbol{v}_i^h \in \mathbb{R}^3 \mid i \leq N_h\}$ and object mesh $\mathcal{V}^o = \{\boldsymbol{v}_j^o \in \mathbb{R}^3 \mid j \leq N_o\}$, where N_h and N_o are the number of the vertex of hand and object respectively. Then, PiCR learns to construct the CPF and collect the elastic energy E_{elast} in it. Finally, GeO minimizes E_{elast} in CPF to yield the refined HO meshes ${}^*\mathcal{V}^o$, ${}^*\mathcal{V}^h$.

5.1. Hand-object Pose Estimation Network, HoNet

The HoNet first predicts coarse poses of HO meshes by the baseline model *MeshRegNet* as in [23]. The outcomes from the baseline comprise in total 37 coefficients: object 6D pose $\boldsymbol{P}_o \in \mathfrak{se}(3)$ ($\mathbb{R}^6$), hand wrist 6D pose $\boldsymbol{P}_w \in \mathfrak{se}(3)$, PCA components of MANO pose $\boldsymbol{\theta}_{\text{pca}} \in \mathbb{R}^{15}$ and shape $\boldsymbol{\beta} \in \mathbb{R}^{10}$. With these coefficients, HoNet could place the HO meshes into camera space. Details of the baseline can be referred to [23].

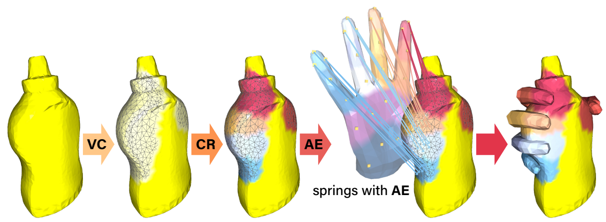

Figure 4. Illustration of assigning *Vertex Contact*, *Contact Region* and *Anchor Elasticity* onto object surface.

5.2. Pixel-wise Contact Recovery Module, PiCR

With the coarse meshes of hand and object in HoNet, PiCR learns to recover the CPF by firstly paring the hand anchors and object vertices into HO affinity pairs and then regressing the spring elasticities that describe the affinities. To achieve this, PiCR yields three cascaded outcomes: 1) *Vertex Contact* (*VC*) decides which vertices on object are in contact with hand; 2) *Contact Region* (*CR*) decides the subregion that is most likely to contact with those vertices in *VC*; 3) *Anchor Elasticity* (*AE*) represents the elasticities of the *attractive* springs. With *VC*, *CR*, and *AE*, we can then recover the CPF as illustrated in Fig. 4.

Vertex Contact. PiCR' s first outcome $VC \in \mathbb{R}^{N_o}$ stands for the contact probability of object vertices. More specifically, $VC[j]$ is a probability that implies the j-th object vertex $\boldsymbol{v}_j^o$ is in contact with hand. The loss function of *VC* is defined as a binary focal loss [33]:

$$\mathcal{L}_{VC} = -\sum_{j}^{N_o} \mathbb{1}_j^{img} * \alpha_j (1 - f_j)^\gamma \log(f_j) \tag{5}$$

where $f_j = p_j$ if the *gt.* $\hat{\boldsymbol{v}}_j^o$ belongs to any HO affinity, otherwise $f_j = (1 - p_j)$, and the p_j is the predicted probability at $VC[j]$. $\mathbb{1}_j^{img}$ denotes whether the vertex $\boldsymbol{v}_j^o$ is projected inside the image. α_j is inverse class frequency and γ is empirically set to 2.

Contact Region. PiCR's second outcome $CR \in \mathbb{R}^{N_o \times 17}$ stands for the subregion probabilities of object vertices. More specifically, for the j-th query, $CR[j]$ contains 17 probabilities that indicates $\boldsymbol{v}_j^o$'s affinity toward 17 hand subregions. The loss function $\mathcal{L}_{CR}$ is defined as a multi-class focal loss.

$$\mathcal{L}_{CR} = -\sum_{j}^{N_o} \mathbb{1}_j^{VC} * \mathbb{1}_j^{img} * (1 - m_j)^\gamma \log(m_j) \tag{6}$$

where the $m_j = \sum(p_j * t_j)$ in which $p_j = CR[j] \in \mathbb{R}^{17}$ is the predicted per-subregion probabilities through *softmax*, and $t_j \in \mathbb{R}^{17}$ is the *gt.* subregion affinity of $\hat{\boldsymbol{v}}_j^o$ as a one-hot vector. $\mathbb{1}_j^{VC}$ denotes that the *gt.* *VC* of $\hat{\boldsymbol{v}}_j^o$ is positive.

Anchor Elasticity. PiCR's third outcome $AE \in \mathbb{R}^{N_o}$ stands for the predicted elasticity of *attractive* springs k^{atr}. More specifically, $AE[j]$ is the elasticity k_{ij}^{atr} of an *attractive* spring that connects $\boldsymbol{v}_j^o$ to its affinitive anchor $\boldsymbol{a}_i$ in the predicted subregion: $argmax(CR[j])$. The loss function $\mathcal{L}_{AE}$ is defined as a binary cross-entropy (BCE):

Algorithm 1: Procedure of recovering CPF

Input: $\mathcal{V}^o, \mathcal{V}^h, VC, CR, AE$
Output: E_{elast}: elastic energy

```
recovery anchors: A ← linear_interpolation(V^h);
foreach j ∈ {j | j ≤ N_o, VC[j] > t_vc} do
    recover subregion id: r ← argmax(CR[j]);
    foreach a_i ∈ A_r (anchors in subregion r) do
        recover elasticity: k_ij^atr ← AE[j];
        E_elast +← 1/2 * k_ij^atr ||a_i − v_j^o||_2^2;
    foreach i ∈ {i | i ≤ N_h, ||v_i^h − v_j^o||_2^2 ≤ t_rpl} do
        E_elast +← 1/2 * k_ij^rpl |exp(−(v_i^h − v_j^o) · n_j^o)|^2;
```

$$\mathcal{L}_{AE} = \sum_{j}^{N_o} \mathbb{1}_j^{VC} * \mathbb{1}_j^{img} * BCE(\, k_{ij}^{\text{atr}},\, \hat{k}_{ij}^{\text{atr}}) \tag{7}$$

where the $\hat{k}_{ij}^{\text{atr}}$ is the *gt.* elasticity described in §4.

With the predicted *VC*, *CR* and *AE*, as well as the coarse meshes $\mathcal{V}^o, \mathcal{V}^h$ in HoNet, PiCR finally recovers the CPF and collects the elastic energy E_{elast} as described in Algm.1. We empirically set the probability threshold of *VC*: $t_{\text{vc}} = 0.8$ and the distance threshold: $t_{\text{rpl}} = 20\ mm$.

PiCR's Framework. The proposed PiCR consists of a backbone b that extracts features from image, an encoder p that converts image features to object vertex features, and 3 heads h_{vc}, h_{cr} and h_{ae} which sequentially convert those features into *VC*, *CR*, and *AE*. As illustrated in Fig. 3, the process of feature extraction in PiCR can be expressed as:

$$\mathcal{F}' = \Big[f\big(\pi(\mathcal{V}^o), b(\mathcal{I})\big), z(\mathcal{V}^o) \Big]; \quad \mathcal{F} = p(\mathcal{F}') \tag{8}$$

where $b(\cdot)$ is the hourglass networks [37], $\pi(\cdot)$ is the perspective camera projection, and $f(\cdot)$ stands for aligning $\mathcal{V}^o$'s 2D projection $\pi(\mathcal{V}^o)$ with the image features $b(\mathcal{I})$ through bilinear sampling. Inspired from Eq.(1) in [48], we also append the object's root-relative z value $z(\mathcal{V}^o)$ at the end of $f(\cdot)$ to form the pixel-wise features $\mathcal{F}'$. Next, a PointNet [40] encoder $p(\cdot)$ is adopted to convert $\mathcal{F}'$ to its point-wise features $\mathcal{F}$.

The process of three PiCR's heads can be expressed as:

$$VC = h_{\text{vc}}(\mathcal{F}); \; CR = h_{\text{cr}}(VC, \mathcal{F}); \; AE = h_{\text{ae}}(CR, \mathcal{F}) \tag{9}$$

where all the heads are presented as multi-layer perceptrons. We provide implementation details in *Supp* D.1.

5.3. Grasping Energy Optimizer, GeO

The fitting part: Grasping Energy Optimizer (GeO) aims to refine the HO pose w.r.t. the recovered CPF. For the object part, we adjust its 6D pose $\boldsymbol{P}_o \in \mathfrak{se}(3)$. For the hand part, we jointly adjust the A-MANO' s 15 joint rotations $\{\boldsymbol{R}_j \in \mathfrak{so}(3) \mid j \leq 15\}$ and a wrist pose $\boldsymbol{P}_w \in \mathfrak{se}(3)$.

In order to mitigate the abnormal hand posture during optimization, we also define an anatomical cost function $\mathcal{L}_{\text{anat}}$ that penalizes the unwanted axial components and angular values of the 15 rotations in the proposed *twist-splay-bend* coordinate frame. First, for the joints along hand kinematic tree, we penalize the component of rotation axis $\mathbf{a}^{rot}$ on *twist* direction: $\mathbf{n}^{twist}$, since any component that causes the finger twisting along its pointing direction is prohibited. Second, for the joints that do not belongs to 5 knuckles, we also penalize the component of $\mathbf{a}^{rot}$ on *splay* direction: $\mathbf{n}^{splay}$. Last, we penalize the rotation angle ϕ^{bend} that revolves about the *bend* axis if it is greater than $\pi/2$. The total anatomical cost can be written as:

$$\begin{aligned}\mathcal{L}_{\text{anat}} = &\sum_{j\in\text{all}} \mathbf{a}_j^{rot} \cdot \mathbf{n}_j^{twist} + \sum_{j\notin\text{knuck}} \mathbf{a}_j^{rot} \cdot \mathbf{n}_j^{splay} \\ &+ \sum_{j\in\text{all}} \max\left((\phi_j^{bend} - \frac{\pi}{2}), 0\right)\end{aligned} \tag{10}$$

We also penalize the offset of the refined hand-object vertices ${}^*\mathcal{V}^o$, ${}^*\mathcal{V}^h$ from their initial estimation $\mathcal{V}^h$, $\mathcal{V}^o$ in form of *l2* distance: $\mathcal{L}_{\text{offset}}$. We implement GeO in PyTorch with Adam solver. The whole optimization process can be expressed as:

$${}^*\mathcal{V}^o, {}^*\mathcal{V}^h \longleftarrow \underset{\boldsymbol{P}_o, \boldsymbol{P}_w, \boldsymbol{R}_j}{argmin}\left(E_{\text{elast}} + \mathcal{L}_{\text{anat}} + \mathcal{L}_{\text{offset}}\right) \tag{11}$$

6. Experiments and Results

6.1. Datasets

We would like to train and evaluate MIHO w.r.t. the real-world dataset that involves human hand interacting with textured object. In the community, there exist mainly four datasets that contain images and ground-truth 3D HO annotation, namely ObMan [24], FHB [18] and HO3D [21, 22] and ContactPose [7]. However, only FHB and HO3D satisfy our requirements in this study.

First-person Hand Action Benchmark, FHB. FHB is a first-person RGBD video dataset of hand in manipulation with objects. The ground-truth of hand poses was captured via magnetic sensors. In our experiments, we use a subset of FHB that contains 4 objects with a scanned model and pose annotation. We adopt the *action* split following the protocol given by [23, 51], and filter out the samples with a minimum HO distance greater than 5 mm, which yields us 7223 samples for training and 7373 for testing.

HO3D. HO3D is another dataset that contains precise hand-object pose during the interaction. Due to historical reasons, there is two versions of HO3D, namely v1 [21] and v2 [22]. In our experiments, we mainly compare our methods with the baseline [23] on HO3Dv1, but also conduct several comparisons with the recently released pretrained model of [23] on HO3Dv2. Similar to FHB, we filter out samples with distance threshold $5mm$. It's also worth mentioning that, since our method requires a known object model, as well as a stable grasping configuration, nearly 5448 samples in HO3Dv2 test set are not suitable for our methods to report. Therefore, we manually select 6076 samples in HO3dv2 test set to compare MIHO with [23]. We call this split by HO3Dv2$^-$. Besides, training HO3Dv1 in previous methods [21, 23] requires an extra synthetic dataset that is not publicly available. Thus we manually augment the HO3Dv1 training set (referred as HO3Dv1$^+$) and reproduce the results (referred as [23]$^+$) comparable with those in [23]. Details of HO3Dv2$^-$ selection and the augmentation procedures are provided in *Supp* C.1, C.2.

6.2. Metrics

Modeling the HO interaction requires not only a proper pose of both hand and object but also a natural grasp configuration. Here, we report 5 metrics in total that cover both reconstruction and grasp quality. Note that, since considering either of those metrics alone may yield misleading comparison, we consider them **together** for evaluation.

MPVPE. We compute the mean per vertex position error for both hand and object in camera space to assess the quality of pose estimation.

Penetration Depth (PD). To measure how deep that the hand is penetrating the object's surface, we calculate the penetration depth that is the maximum distance of all the penetrated hand vertices to their closest object surface.

Solid Intersection Volume (SIV). To measure how much space intersection that occurs during estimation, we voxelize the object mesh into 80^3 voxels, and calculate the sum of the voxel volume inside the hand surface.

Disjointedness Distance (DD). We also encourage stable HO contact, which can be depicted as attracting fingertips onto the object surface. Therefore, we define the disjointedness metrics as the average distance of hand vertices in 5 fingertips region to their closet object surface.

Simulation Displacement (SD). We further evaluate the grasp stability in a modern physics simulator [11]. We measure the average displacement of object's center over a fixed time period by holding the hand steadily and applying gravity to the object [24].

6.3. Comparison with State-of-the-Arts

For the FHB dataset, we compare our methods with the previous SOTA [23, 24] of hand-object reconstruction. For

Datasets	FHB					HO3Dv1$^+$				HO3Dv2$^-$	
Method	Ours†	Ours‡	*gt.*	[23]	ObMan*	Ours†	Ours‡	*gt.*	[23]$^+$	Ours‡	[23]
Hand MPVPE $(mm)\downarrow$	21.16	19.54	0	**17.51**	18.42	24.56	**23.99**	0	24.80	-	-
Object MPVPE $(mm)\downarrow$	**21.06**	21.57	0	**21.06**	21.17	**18.10**	19.15	0	**18.10**	**73.28** $^\diamondsuit$	75.77 $^\diamondsuit$
Penetra. depth $(mm)\downarrow$	**16.13**	16.92	19.55	20.63	19.76	11.87	**11.42**	7.55	18.57	**16.47**	20.02
Solid intersec. vol. $(cm^3)\downarrow$	12.56	**11.76**	20.41	21.10	16.16	3.63	**3.46**	3.57	9.62	**7.44**	9.25
Disjoint. distance $(mm)\downarrow$	24.54	**22.41**	37.28	37.40	27.95	**11.71**	11.83	14.53	18.62	**37.04**	41.41
Displacement $(mm)\downarrow$	58.79	**58.02**	63.40	65.48	59.41	28.16	27.66	12.37	**25.68**	**39.33**	41.03

Table 1. **Quantitative results and detailed comparison with the previous state-of-the-art [23, 24] on the FHB and HO3D datasets.** "*gt.*" denotes the ground-truth. "†" denotes ours *hand-alone* optimization setting and "‡" denotes the jointly *hand-object* setting. "$*$" denotes the reproduced ObMan [24]. "$\diamondsuit$" denotes the *wrist-relative* object vertex error. "-" indicates the results that are not available.

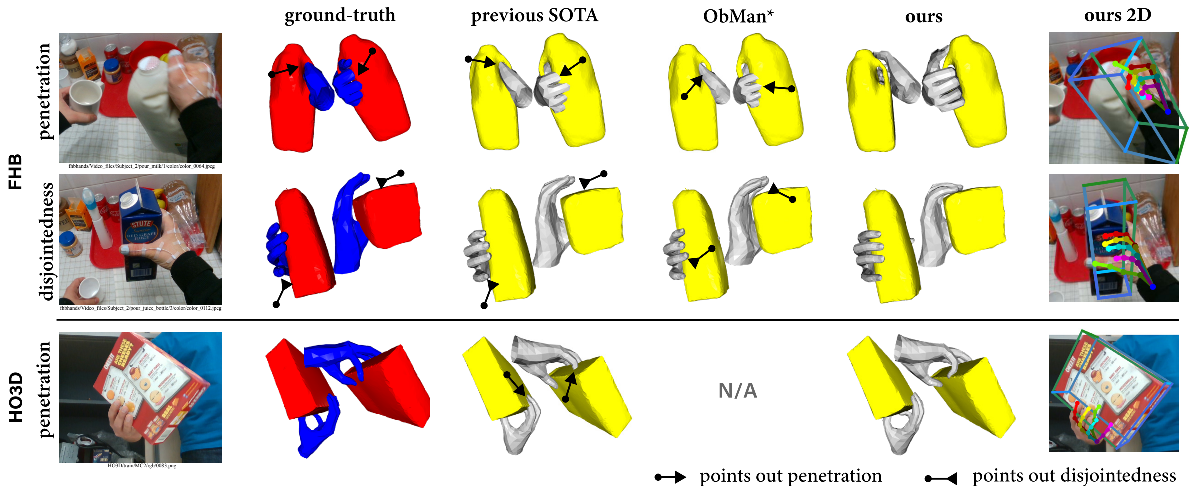

Figure 5. **Qualitative Comparison with ground-truth and previous arts on the FHB and HO3D datasets.**

[23], we select the results under the setting of full data supervision. Since [23] didn't exploit any repulsion and attraction loss during training, direct comparisons on intersection and disjointedness may not be convincing enough. While the contact losses were considered in another work named ObMan [24], it only represented the genus 0 object mesh as a deformable icosphere, which is also not directly comparable with ours (known object model). To ensure rational comparisons, we migrate the *repulsion loss* and *attraction loss* from ObMan to the *MeshRegNet* in [23], and reproduce the results on par with it. We call this adaptation: ObMan*. For the HO3Dv1 dataset, we compare our results with the reproduced [23]$^+$.

We report our results under two experimental settings: 1) ***hand-alone*** that fixes the object at the initial prediction in HoNet, and only optimizes the hand pose in GeO; 2) ***hand-object*** that jointly optimizes the hand and object poses in GeO. In Tab.1 we show our comparisons with the previous SOTA in all 5 metrics. For FHB dataset, as analyzed in [7], its ground-truth suffers from frequent interpenetration. We find that lower vertex error does not necessarily benchmark a higher reconstruction quality. Indeed, as shown in Tab.1 (col. 4, 5), either ground-truth or [23] reveals substantial solid intersection volume, penetration depth and disjointedness. We find that MIHO outperforms [23] by a margin of 3.71 mm in penetration depth, 9.34 cm^3 in solid intersection volume, and 14.99 mm in disjointedness distance, while only suffers from minor performance cost in hand MPVPE of 2.03 mm and object MPVPE of 0.51 mm. In the mean time, our simulation displacement also demonstrates the stability of our predicted grasp. These are consistent with our expectation that the CPF can by nature repulse intersection away and attract disjointedness to touch. As for HO3Dv1 testing set, our method also outperformed the previous SOTA over the most metrics. In terms fo simulation displacement, we found [23]$^+$ slightly outperforms us by 1.98 mm. Based on our inspection in the Bullet [11] simulator, their stability are mainly attributed to the forces resulting from the intersection that balance each other. Visual comparisons are shown in Fig. 5. As for HO3Dv2, since we only test MIHO on the subset: HO3Dv2$^-$, our results are not directly suitable for submitting to its online evaluation server. Thus, we only report the object 3D vertex errors on HO3Dv2$^-$ based on the given annotation. We firstly align the predicted object vertex to the predicted hand wrist joint, then compute the *wrist-relative* object vertex error with those in ground-truth. Detailed comparisons are in Tab. 1 (col. 11, 12).

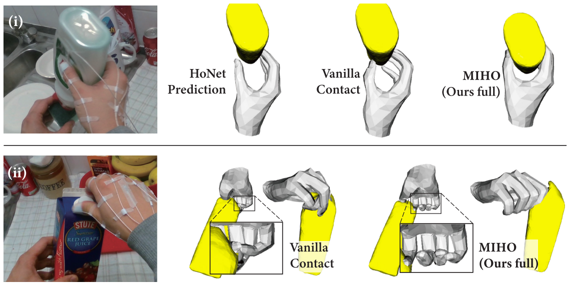

Figure 6. Comparisons of MIHO with simple contact heuristic.

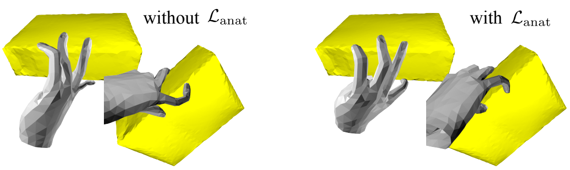

Figure 7. Example to illustrates the efficacy of our proposed A-MANO with anatomical constraints ($\mathcal{L}_{\text{anat}}$).

6.4. Ablation Study

In this experiment, we further evaluate the effectiveness of the proposed CPF and A-MANO. In the main text, we include three of the most representative studies. The ablation studies are mainly conducted on the FHB test set with *action* split. For more studies on 1) impact from the magnitude of k^{rpl}; 2) A-MANO with PCA pose; 3) unwanted twist correction; please visit *Supp* D.2.

Comparison with simple Distance-based Contact Heuristics. To show the superiority of the CPF over the distance-based contact heuristics, we compare the fitting stage of MIHO with two simple yet strong baselines: (a) *Vanilla Contact* that removes the E_{elast} term in Eq. 11 and purely attracts the anchors on fingertips to its nearest object vertex (similar to [1]) in a given threshold which we set as 20 mm; (b) *ObMan Contact* that replaces E_{elast} in Eq. 11 by the well-designed interaction losses in ObMan [24]. All the three experiments start from the same HO pose predicted by HoNet (§5.1). We show in Tab. 2 that by exploiting CPF, MIHO can surpass the simple baselines on most of the metrics. Note that, since both (a) and (b) directly optimize the disjointedness term, their results show better resistance on it. The last column in Tab. 2 shows that our methods can save average time per iteration by 46% compared with *ObMan Contact*. We also conduct two qualitative comparisons in Fig. 6. The first shows that CPF can learn the contact *semantics* to guide the optimization that better matches visual cues, whereas the *Vanilla Contact* fails to form a valid grasp. The second shows that CPF can maintain subtle interaction, as no attraction will be applied on those non-affinitive vertex pairs (see ring and pinky fingers when unscrewing the juice cap).

Effectiveness of *Repulsive* Springs. To measure the efficacy of *repulsive* springs in CPF, we remove all the repulsion energy E^{rpl} induced by them, leaving the attraction as the unique type of energy applied on hand and object. As we expected, the result in Tab. 3 witnesses the accumulation of PD and SIV. To note, even without the *repulsive* springs, we still witness a remarkable improvement of PD and SIV over the FHB ground-truth. This is attributed to the repulsive behavior of the *attractive* springs: when hand is inside the object surface, the energy stored in the *attractive* springs will act as repulsion that pushes out the hand.

Settings	Scores					$t_{\text{iter}}(ms)$
	HE ↓	OE ↓	PD ↓	SIV ↓	DD ↓	
MIHO (ours full)	19.54	21.57	16.92	11.76	22.41	55.77
(a) Vanilla Contact	24.01	24.29	18.36	15.64	16.32	45.40
(b) ObMan Contact	22.15	22.54	15.13	16.20	11.97	103.41

Table 2. Ablation study on the different contact heuristics. HE, OE stands for 3D hand and object vertex error. PD, SIV and DD are the abbreviation of metrics in §6.2

Settings	PD↓	SIV↓	DD↓
with E^{rpl} (ours full)	16.92	11.76	22.41
without E^{rpl}	17.79	13.76	20.27
gt. FHB	19.55	20.41	37.28

Table 3. Ablation study on the *repulsive* springs.

Effectiveness of the Anatomical Constraints. We further highlight the efficacy of adopting the anatomical constraints. We conduct a contrastive experiment whose only difference is the absence of $\mathcal{L}_{\text{anat}}$. Both experiments start from a zero (flat) hand and minimize the E_{elast} based on the same predicted CPF. We show in Fig. 7 that the anatomical constraints are able to effectively prevent abnormality during the optimization.

7. Conclusion

In this work, we propose a novel contact representation named CPF and a learning-fitting hybrid framework MIHO to help modeling hand and object interaction. Comprehensive evaluations show that our methods, while being able to recover precise hand-object pose, can also effectively 1) avoid interpenetration and control disjointedness, and 2) prevent abnormality in hand pose. We hope CPF can serve as an effective contact representation for future works on hand-object interaction. Later, we also plan to develop for an object-agnostic representation of CPF, for the interaction in general cases.

Acknowledgements This work is supported in part by the National Key R&D Program of China, No. 2017YFA0700800, National Natural Science Foundation of China under Grants 61772332 and Shanghai Qi Zhi Institute, SHEITC (018-RGZN-02046).

References

[1] Dafni Antotsiou, Guillermo Garcia-Hernando, and Tae-Kyun Kim. Task-oriented hand motion retargeting for dexterous manipulation imitation. In *ECCV Workshops*, 2018. 2, 3, 8

[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] Christoph W Borst and Arun P Indugula. Realistic virtual grasping. In *VR*, 2005. 2

[4] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 1, 2

[5] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, 2019. 1

[6] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. In *IROS*, 2019. 1, 2, 4

[7] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. ContactPose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020. 1, 3, 6, 7

[8] Minjie Cai, Kris M Kitani, and Yoichi Sato. Understanding hand-object manipulation with grasp types and object attributes. In *RSS*, 2016. 2

[9] Minjie Cai, Kris M Kitani, and Yoichi Sato. An ego-vision system for hand grasp analysis. *IEEE Transactions on Human-Machine Systems*, 2017. 2

[10] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, 2017. 2

[11] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation in robotics, games and machine learning, 2017. 6, 7

[12] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, 2020. 1, 2

[13] George ElKoura and Karan Singh. Handrix: animating the human hand. In *SIGGRAPH*, 2003. 2

[14] Haoqiang Fan, Hao Su, and Leonidas J Guibas. A point set generation network for 3d object reconstruction from a single image. In *CVPR*, 2017. 1

[15] Thomas Feix, Ian M Bullock, and Aaron M Dollar. Analysis of human grasping behavior: Object characteristics and grasp type. *IEEE transactions on haptics*, 2014. 2

[16] Hongyang Gao and Shuiwang Ji. Graph u-nets. *ICLR*, 2019. 2

[17] Guillermo Garcia-Hernando, Edward Johns, and Tae-Kyun Kim. Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. *arXiv preprint arXiv:2008.03285*, 2020. 1, 2

[18] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *CVPR*, 2018. 2, 6

[19] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 1

[20] 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. 1

[21] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. *arXiv preprint arXiv:1907.01481*, 2019. 2, 6

[22] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, 2020. 2, 4, 6

[23] 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, 6, 7

[24] 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, 3, 6, 7, 8

[25] Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *VR*, 2018. 2

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

[27] Jun-Sik Kim and Jung-Min Park. Physics-based hand interaction with virtual objects. In *ICRA*, 2015. 2

[28] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. *arXiv preprint arXiv:1903.03340*, 2019. 1, 2

[29] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *RAL*, 2020. 1, 2

[30] Paul G Kry and Dinesh K Pai. Interaction capture and synthesis. *TOG*, 2006. 2

[31] Dominik Kulon, Riza Alp Guler, Iasonas Kokkinos, Michael M Bronstein, and Stefanos Zafeiriou. Weakly-supervised mesh-convolutional hand reconstruction in the wild. In *CVPR*, 2020. 1

[32] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *CVPR*, 2013. 2

[33] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. Focal loss for dense object detection. In *ICCV*, 2017. 5

[34] C Karen Liu. Dextrous manipulation from a grasping pose. In *SIGGRAPH*, 2009. 2

[35] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 2004. 2

[36] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder

framework for high-fidelity hand mesh modeling. In *ECCV*, 2020. 1
[37] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *ECCV*, 2016. 5
[38] 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
[39] 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
[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *CVPR*, 2017. 5
[41] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2
[42] Hans Rijpkema and Michael Girard. Computer animation of knowledge-based human grasping. *SIGGRAPH*, 1991. 2
[43] Grégory Rogez, Maryam Khademi, JS Supančič III, Jose Maria Martinez Montiel, and Deva Ramanan. 3d hand pose detection in egocentric rgb-d images. In *ECCV*, 2014. 2
[44] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 1, 2
[45] Javier Romero, Hedvig Kjellström, and Danica Kragic. Monocular real-time 3d articulated hand pose estimation. In *IEEE-RAS HUMANOID*, 2009. 2
[46] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 2
[47] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *TOG*, 2017. 2, 3
[48] 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. 5
[49] Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Otmar Hilliges, and Jan Kautz. Weakly supervised 3d hand pose estimation via biomechanical constraints. In *ECCV*, 2020. 2
[50] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 1
[51] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 1, 2, 6
[52] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *IJCV*, 2016. 2
[53] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In *ICCV*, 2015. 2
[54] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *ECCV*, 2018. 1
[55] Jiajun Wu, Yifan Wang, Tianfan Xue, Xingyuan Sun, Bill Freeman, and Josh Tenenbaum. Marrnet: 3d shape reconstruction via 2.5 d sketches. In *NeurIPS*, 2017. 1
[56] Lixin Yang, Jiasen Li, Wenqiang Xu, Yiqun Diao, and Cewu Lu. Bihand: Recovering hand mesh with multi-stage bisected hourglass networks. In *BMVC*, 2020. 1, 2
[57] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *TOG*, 2012. 2
[58] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 1, 2