Stability-driven Contact Reconstruction From Monocular Color Images

Zimeng Zhao Binghui Zuo Wei Xie Yangang Wang*

Southeast University, China

Abstract

Physical contact provides additional constraints for hand-object state reconstruction as well as a basis for further understanding of interaction affordances. Estimating these severely occluded regions from monocular images presents a considerable challenge. Existing methods optimize the hand-object contact driven by distance threshold or prior from contact-labeled datasets. However, due to the number of subjects and objects involved in these indoor datasets being limited, the learned contact patterns could not be generalized easily. Our key idea is to reconstruct the contact pattern directly from monocular images, and then utilize the physical stability criterion in the simulation to optimize it. This criterion is defined by the resultant forces and contact distribution computed by the physics engine. Compared to existing solutions, our framework can be adapted to more personalized hands and diverse object shapes. Furthermore, an interaction dataset with extra physical attributes is created to verify the sim-to-real consistency of our methods. Through comprehensive evaluations, hand-object contact can be reconstructed with both accuracy and stability by the proposed framework.

1. Introduction

Monocular hand-object contact recovery has wide applications, which can enable accurate interactions in metaverse and telepresence robot control. Traditional methods often judge contact regions by the closest distances between surfaces of the hand and object in an optimization strategy [58], where the recovered contact highly depends on the accuracy of hand-object pose estimation. However, this accuracy is hard to be guaranteed by monocular reconstruction. Recent approaches [5, 17, 62] learn the hand-object contact prior from well-labeled datasets [3, 52], yet their performance rely on the diversity of the contact data.

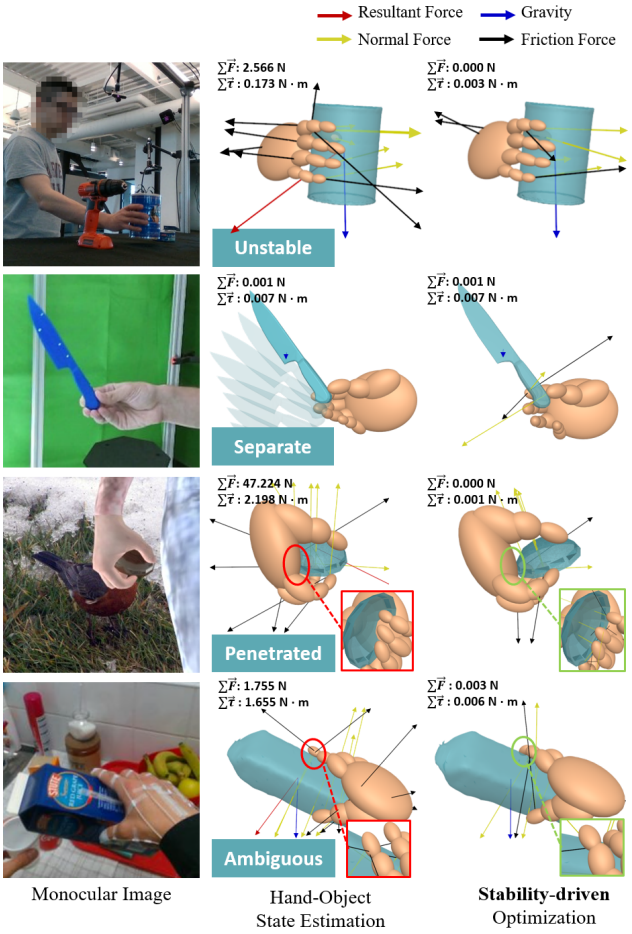


Figure 1. **Stability-driven contact reconstruction**. Each row illustrates the hand-object state represented by multiple ellipsoids. The resultant forces and torques on the objects are calculated by the physics engine [8].

Generally speaking, a credible contact is to ensure the interacting stability between a hand and object in the physical world, either to keep the object stationary or with a required acceleration. To reconstruct this stable contact, our key idea is to **reconstruct the contact pattern driven by the physical criteria (*i.e.*, the balancing of forces and torques) cal-**

*Corresponding author. E-mail: yangangwang@seu.edu.cn. This work was supported in part by the National Natural Science Foundation of China (No. 62076061), the "Young Elite Scientists Sponsorship Program by CAST" (No. YES20200025), and the "Zhishan Young Scholar" Program of Southeast University (No. 2242021R41083).

culated by a physics engine. It is noted that most existing methods [7, 21, 27, 29] utilize the relative object displacements to evaluate the contact stability. However, the criteria cannot be directly used to drive optimization due to the shortcomings in both hand modeling and stability evaluation.

Regarding hand modeling, traditional methods for simulation utilize either a whole mesh [21] or multiple mesh segments without connectivity [34, 55, 58]. Such models are difficult to perform robot control and force analysis due to the lack of a kinematic tree. To overwhelm the limitations, we adopt a structured multi-body for dynamics simulation, whose rigid parts can be automatically adjusted according to the personalized information estimated from an image. Specifically, those hand rigid parts and the object are jointly represented as a series of ellipsoidal primitives, and our front-end network is used to estimate the state parameters for composing these primitives. Compared with MANO [47] parameters, regressing this state not only brings acceleration to the calculation of self and mutual collision during network training but also facilitates the construction of our multi-body in a physics engine.

We argue that the stability could not be fully evaluated by the displacement of the contact object, which is an average effect of the resultant force. Alternatively, some novel stability criteria are proposed with more considerations on physical factors related to contact. Considering the contact constraints are unilateral, we use sampling-based optimization rather than gradient-based methods to make the estimated state meet the above stability requirements. A hand-object contact dataset is further built to analyze the sim-to-real gap in our simulation. In addition to images and meshes of hands and objects, our proposed dataset includes physical properties and stability evaluations for each interaction scene. The stability of a real interaction scene is mainly evaluated by the additional balancing force that needs to be applied to the object when capturing contact by our multi-view system.

Summarily, we make the following contributions.

- A regression-optimization framework for reconstructing hand-object contacts and physical correlation from monocular images guided by stability;
- A hand-object representation and learning strategy based on ellipsoid primitives, which brings convenience to the process of both deep learning inference and physical simulation;
- A hand-object interaction dataset containing physical attributes and stability metrics, which validates the sim-to-real consistency of related methods.

The dataset and codes will be publicly available at https://www.yangangwang.com.

2. Related Work

The reconstruction method discussed in this part mainly takes the monocular color image as input and considers the interaction between one hand and one object.

Hand-Object State Estimation. With the rapid increase of 3D hand datasets [16, 33, 61, 64, 68] and object datasets [23, 33, 61], data-driven methods [2, 15, 25, 26, 30, 37, 43, 54, 60, 63, 66, 67] become popular in the community. However, when the hand interacts with the object, the problem becomes further complicated because of severe occlusions. The representation in pioneer datasets [14] and methods [10, 53] only contained hand skeletons and object bounding boxes. Subsequent work [6, 18] provided more fine-grained hand-object surfaces depicted by MANO parameters [47] and specific object categories [4, 61]. With more synthetic data, Hasson *et al*. [21] explored the scheme to reconstruct the shape and pose of hand-object through a unified network. Other methods [5, 17, 19, 20] placed more emphasis on the hand state and object pose. This work also relies on providing object meshes in the simulation. However, the object pose and hand features are estimated from the input images.

Contact Estimation. There is a trend [40, 48] to understand the interaction pattern directly at the image level. Since the contact area is generally invisible in the image, more methods explore it from 3D states. To enable data-driven methods, many pioneers [3, 11, 44] utilized expensive sensors, ingenious deployment, and manual labor to obtain actual contact information without affecting the hand-object appearance (marker-less). Others [21, 52] used the distance between the hand and the object surface in Mocap data as the criterion to annotate the contact. Benefit from these datasets, recent methods [5, 17, 62] learn the contact area prior in advance and then iteratively optimize the hand-object state according to the prior. In the evaluation stage, some approaches treat the state with more contact coverage ratio as stable [17]. But this may exacerbate unreasonable penetrations rather than improve the contact quality. As [21] pointed out, this can be compensated for by evaluation methods based on the physical simulation [58]. With considering more contact-related physics, we create a stability criterion to effectively optimize the hand-object state without the prior dependence.

Hand Collision Shape. Although hand meshes [47] are convenient for rendering, they require expensive computation for collision detection on each vertex [17, 21, 38, 58]. For articulated objects such as the human body and hand, collisions occur not only with other objects but also between different links of themselves. Several attempts [9, 28, 29, 35] have been made to implicitly represent the surface with the neural occupancy function, but they are ineffective for self-intersection [35]. By contrast, approximation of articulated objects using geometric primitives, *e.g*. capsules [13, 46],

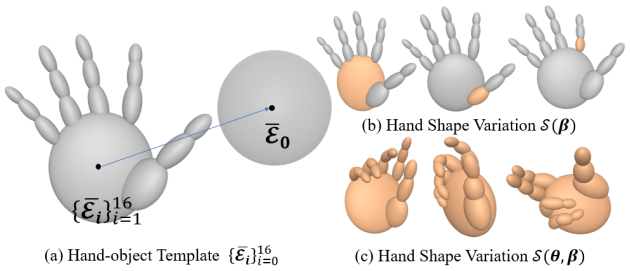


Figure 2. **Hand-object state representation**. The part colored in brown indicates a shape variation w.r.t. the template. (a) Hand-object template consists of 17 ellipsoids; (b) Template hands with shape variations. (c) Personalized hands with pose variations.

spheres [39, 45, 49–51, 59] or mixtures [41, 42] are more intuitive to tackle both kind of intersections. [56, 57] presented a conversion method from implicit spheres to smooth triangular mesh. We propose a more concise scheme to represent the hand-object as a series of ellipsoids. It builds a bridge between network regression and optimization in the simulated environment.

3. Method

We take two steps to reconstruct the state $\mathcal{S}$ of the hand-object and their physical contact $\mathcal{R}$ from monocular color images. First, a network is built to regress the shape and coarse pose of the hand-object represented by ellipsoidal parameters(Sec. 3.1). The above parameters are applied to create dynamics scenes as the initial state of contact optimization(Sec. 3.2). To facilitate the formulation, the tilde superscripts represent the variables regressed from the network, the hat superscripts represent the variables optimized in the simulation, and the star superscripts represent the ground truth.

3.1. Hand-Object State Estimation

Ellipsoid representation. To directly import scene into the physics engine, the states of the personalized hand and object are uniformly represented as a series of ellipsoids rather than MANO [47]. Specifically, a hand is approximated as 16 articulated ellipsoids and an object is approximated as one ellipsoid. Each ellipsoid can be implicitly represented as the zero isosurface of the quadratic form function:

$$\mathcal{E}(\boldsymbol{x}|\boldsymbol{c}, \boldsymbol{r}, \boldsymbol{a}) = (\boldsymbol{x} - \boldsymbol{c})^T A(\boldsymbol{r}, \boldsymbol{a})(\boldsymbol{x} - \boldsymbol{c}) - 1 \quad (1)$$

where $\boldsymbol{c}$ is the ellipsoid center, $\boldsymbol{r}$ is the radii, $\boldsymbol{a}$ is the orientation represented as axis-angle. It should be noted that the decomposition of the symmetric matrix $A(\boldsymbol{r}, \boldsymbol{a}) = R(\boldsymbol{a})^T \mathrm{diag}(\boldsymbol{r})^{-2} R(\boldsymbol{a})$ is not unique, *e.g.* $A((a, b, c)^T, (0, 0, 0)^T)$ vs $A((b, a, c)^T, (0, 0, 0.5\pi)^T)$. Therefore, we adopt a traditional strategy [47] to create a hand template $\{\bar{\mathcal{E}}_i\}_{i=1}^{16}$ shown in Fig. 2. With this template, our hand-object state can be formulated as:

$$\begin{aligned} &\mathcal{S}(\{\mathcal{E}_i\}_{i=0}^{16}) = \mathcal{S}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi};\ \{\bar{\mathcal{E}}_i\}_{i=1}^{16}) \\ &\boldsymbol{\beta} \triangleq \{\delta \boldsymbol{r}_i\}_{i=1}^{16}, \boldsymbol{\theta} \triangleq \{\delta \boldsymbol{a}_i\}_{i=1}^{16}, \boldsymbol{\phi} \triangleq \{\delta \boldsymbol{r}_0, \delta \boldsymbol{a}_0, \delta \boldsymbol{c}_0\} \end{aligned} \quad (2)$$

In this model, each ellipsoid can be scaled by $\delta \boldsymbol{r}_i$ and rotated by $\delta \boldsymbol{a}_i$ w.r.t. its local frame. The center of the palm is used as the coordinate origin and the camera coordinate system is adopted in the network prediction phase. Other ellipsoid centers can be constrained adaptively according to the connection of the ellipsoid to its parent. Because interacting objects usually keep a comparable scale and orientation to the palm, $\{\delta \boldsymbol{r}_0, \delta \boldsymbol{a}_0\}$ as well as the center offset $\delta \boldsymbol{c}_0$ of the object are relative to the $\mathcal{E}_1$.

Mesh conversion. The explicit surface mesh is acquired from implicit primitives in three steps shown in Fig. 4(a 1-3). According to [1, 56], the zero isosurface of the following function corresponds a mesh surface:

$$\mathcal{M}(\boldsymbol{x}) = \min\{\mathcal{E}_i(\boldsymbol{x}|\boldsymbol{c}_i, \boldsymbol{r}_i, \boldsymbol{a}_i)\}_{i=1}^{16} \quad (3)$$

Additional convex hull calculations will make its surface smoother. We use this approach to project the reconstructed hand model into the image to calculate the error. On the other hand, as shown in Fig. 4(b 1-3), diverse LBS hand meshes [32, 38, 47, 58, 64] are first segmented according to the skinning weights. Then oriented bounding box is created for each segment, and the final ellipsoid maintains the same radii and orientation as the box. This approach is used to convert those existing mesh-labeled datasets to the ground truth of $\boldsymbol{\beta}^\star, \boldsymbol{\theta}^\star$ for our training process.

Network architecture. The network is structurally designed as an encoder-decoder. To retain more network attention on the hand-object RoI, pixel-wise features including 2D heatmap, Z-maps of hand joints and object center, hand mask, and object mask are decoded and supervised. The backbone of its encoder is ResNet18 [22] with extra connections to its decoder. Those encoded features are then encoded again and concatenated with the previous features to predict our state parameters $\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\phi}$. In addition, joint regressor $\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is required to regress the joint position between adjacent ellipsoids. It is designed as a two-layer MLP and used to regress joint coordinates $X \in \mathbb{R}^{3\times 21}$ from the explicit mesh vertices of each ellipsoid.

Training process. Because the coordinate of our representation is hand-centered, the datasets with only hand mesh annotation [32, 38, 64, 68] can be used in our training. In the first stage, semi-supervised paradigm is adopted to pre-train our network using those datasets with hand-only or object-only annotations. The overall loss includes:

$$\begin{aligned} L_{\mathrm{S1}} = &\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star\|_2^2 + \|\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^\star\|_2^2 + \|\tilde{X} - X^\star\|_2^2 \\ &+ L_{2D} + L_{in}(\tilde{\mathcal{S}}) \end{aligned} \quad (4)$$

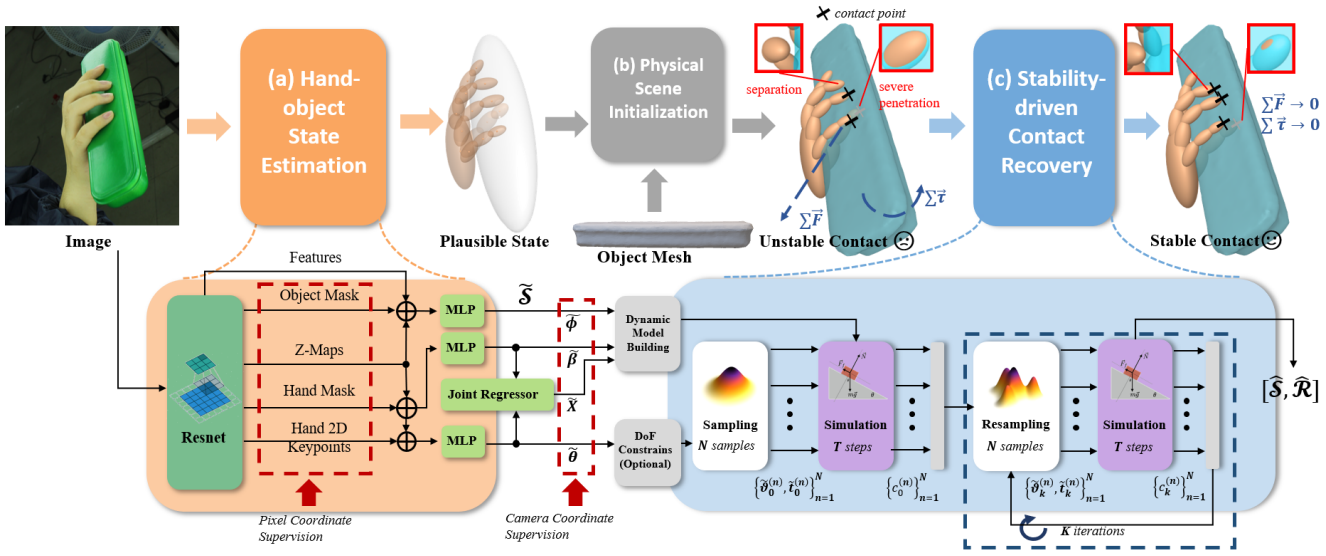


Figure 3. **Stable contact reconstruction pipeline**. (a) Hand-object state represented by implicit ellipsoids is estimated from the input image; (b) Simulated interaction scene is direct constructed from the estimated parameters; (c) The optimization process is driven by the stability cost in simulation to get more reliable states iteratively.

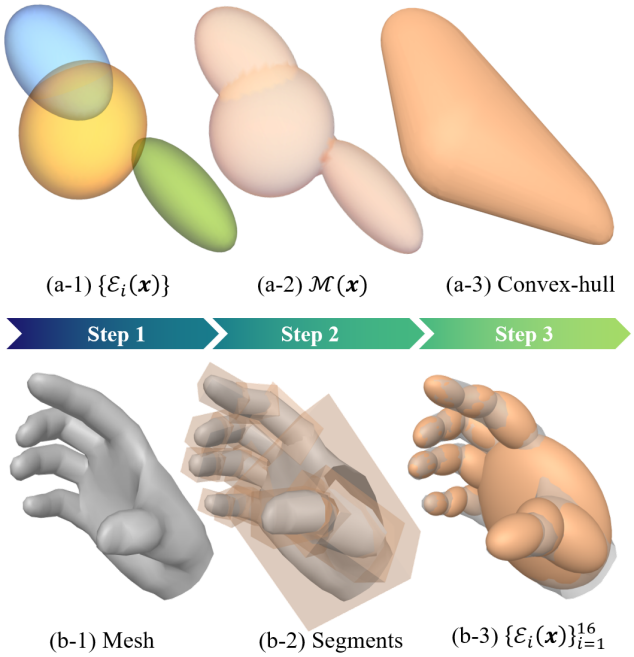


Figure 4. **Conversions between implicit and explicit hand shape**. (a 1-3) Steps from implicit ellipsoids to mesh. To show more details, 3 ellipsoids with great directional variation are used to illustrate; (b 1-3) Steps from explicit mesh to ellipsoids.

The first three items are the hand 3D reconstruction errors. The joint location is estimated with the help of our joint regressor $\tilde{X} = \mathcal{J}(\tilde{\boldsymbol{\beta}}, \tilde{\boldsymbol{\theta}})$. L_{2D} contains the error of all the 2D information regressed in the intermediate steps. Some datasets may not have all annotations, then the corresponding term is also not supervised. The last term is the contact loss designed as point-based [28, 35] to penalize the collision among ellipsoids:

$$L_{in}(\tilde{\mathcal{S}}) = -\sum_{\boldsymbol{x}\in\Omega(\mathcal{E}_i)}\sum_{j\neq i}\mathcal{E}_j(\boldsymbol{x}|\tilde{\mathcal{S}}),\ \text{where}\ \mathcal{E}_j(\cdot) < 0 \quad (5)$$

In practice, 872 vertices uniformly distributed on $\Omega(\mathcal{E})$ are sampled in advance, whose actual coordinates $\boldsymbol{x}$ on $\mathcal{E}_i$ are determined by the ellipsoidal parameters $\tilde{\mathcal{S}}$.

In the second stage, we use the datasets with full annotations [3, 6, 18, 21] to train our network thoroughly:

$$L_{S2} = L_{S1} + \|\tilde{\boldsymbol{\phi}} - \boldsymbol{\phi}^{\star}\|_2^2 + \|\Pi(\tilde{\mathcal{S}}) - \Pi(\mathcal{S}^{\star})\|_2^2 \quad (6)$$

where Π denotes the differentiable projection process to generate the hand and object mask through orthogonal projection. The camera parameters can be obtained by comparing the scale and translation of the hand model with 2D key-points in the image.

$\mathcal{J}(\boldsymbol{\beta}, \boldsymbol{\theta})$ is trained independently. Since it is a mapping from the surface vertices and joints of the hand model during the movement, we obtain a large amount of pairwise training data through the forward dynamic of our hand model in the physics engine.

Implementation details. Our networks are trained on a single NVIDIA GeForce RTX 3090 GPU at a base learning rate of 1e-4, an input image size of 256×256, and a batch size of 64, respectively. We use Adam solver [31] in PyTorch as the optimizer in our training.

3.2. Physical Contact Recovery

Our optimization process is driven by the physical stability evaluated on each sample.

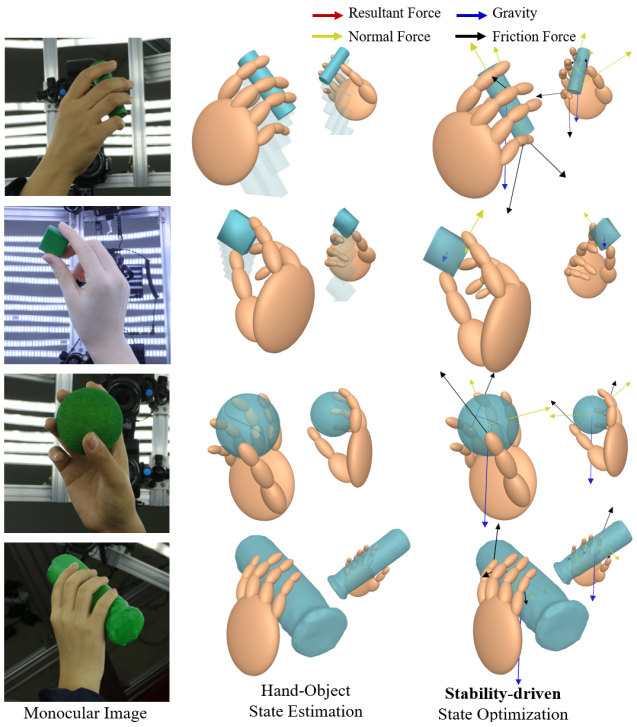


Figure 5. **Qualitative results on our dataset**. For each sample image, the estimated results from the network and optimized results are displayed from two views.

Scene initialization. The estimated state $\tilde{\mathcal{S}}$ are used to initialize the interaction scene in the physics engine [8]. Firstly, the hand template $\{\bar{\mathcal{E}}_i\}_{i=1}^{16}$ with personalized variation $\tilde{\boldsymbol{\beta}}$ are used to construct a dynamic multi-body with 16 ellipsoidal links and fixed root at the origin. The detailed object mesh is loaded to the scene with the position $\tilde{\boldsymbol{p}}$ and orientation $\tilde{\boldsymbol{q}}$ which are determined by $\tilde{\boldsymbol{\phi}}$. Because it is challenging to estimate the linear acceleration $\vec{\boldsymbol{a}}$ and angular acceleration $\vec{\boldsymbol{\alpha}}$ of the object from a single image, they are simply set to zero in the following steps. To facilitate the sampling, the hand pose $\boldsymbol{\theta}$ represented by the axis-angles is converted into $\boldsymbol{\vartheta}$ represented by the Euler angles. Two schemes are adopted for local DoF: retain all 45 or only 20 physically plausible ones [62, 65], *i.e.* $|\boldsymbol{\vartheta}| = 48$ or 23. The hand root is constrained at the origin before optimization, and is allowed to reach a new location $\tilde{\boldsymbol{t}}$ during sampling. As a result, $(\tilde{\boldsymbol{\vartheta}}, \tilde{\boldsymbol{t}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{q}})$ are involved in the next step.

Stability evaluation. The actual physical contact of the given hand-object state is calculated in the impulse-based simulation [36]. Specifically, the collisions are detected among the hand links and object. Based on the Coulomb friction model [12, 24], normal force and lateral friction forces at each contact point are calculated based on the penetration depth. The hand is maintained at a given target pose driven by the PD controller, and the object moves passively due to its own gravity and hand contact forces. Consequently, the contact is evaluated by the stability cost:

$$C = C_{\mathcal{S}}(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}, \hat{\boldsymbol{\vartheta}}, \hat{\boldsymbol{t}}) + C_{\mathcal{R}}(\vec{\boldsymbol{f}}, \vec{\boldsymbol{\tau}}, m) \tag{7}$$

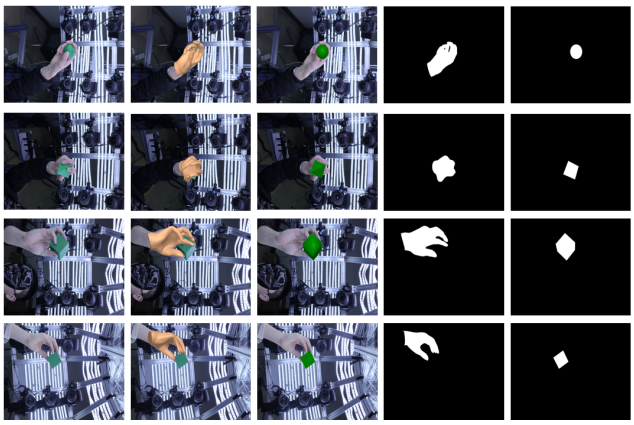
where $C_{\mathcal{S}}$ measures the change in hand-object state before and after simulation, $C_{\mathcal{R}}$ measures the physical relationship including the resultant force $\vec{\boldsymbol{f}}(t)$, torque $\vec{\boldsymbol{\tau}}(t)$ and the number of contact points $m(t)$ collected in $0 < t < T$:

$$\begin{cases} C_{\mathcal{S}} = \|\hat{\boldsymbol{p}} - \tilde{\boldsymbol{p}}\|_2 + L_Q(\hat{\boldsymbol{q}}^{-1}\tilde{\boldsymbol{q}}) + \dfrac{1}{|\boldsymbol{\vartheta}|}\|\hat{\boldsymbol{\vartheta}} - \tilde{\boldsymbol{\vartheta}}\|_1 + \|\hat{\boldsymbol{t}} - \tilde{\boldsymbol{t}}\|_2 \\ C_{\mathcal{R}} = \dfrac{1}{T}\sum_{t=0}^{T} \dfrac{\|\vec{\boldsymbol{f}}(\boldsymbol{t}) - M_o\vec{\boldsymbol{a}}\|_2^2}{\|M_o\vec{\boldsymbol{g}}\|_2^2} + \dfrac{\|\vec{\boldsymbol{\tau}}(t) - I_o\vec{\boldsymbol{\alpha}}\|_2^2}{\|I_o\vec{\boldsymbol{\alpha}}\|_2^2} + e^{-m(t)} \end{cases} \tag{8}$$

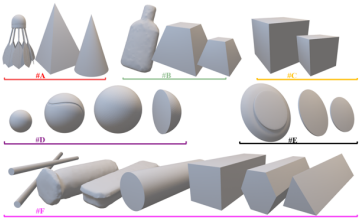
In practice, the change of object direction is measured by the angular difference. The normalization of $\vec{\boldsymbol{f}}(t)$ and $\vec{\boldsymbol{\tau}}(t)$ based on the mass M_o and moment of inertia I_o of the particular object could avoid the impact of the stability cost due to the variation of objects. To prevent the object from flying out of the hand operating area in each simulation step t, the state of the object would be reset if $\|\hat{\boldsymbol{p}}(t) - \tilde{\boldsymbol{p}}\|_2 > 0.1$ or $L_Q(\hat{\boldsymbol{q}}(t)^{-1}\tilde{\boldsymbol{q}}) > 0.3\pi$.

Iterative sampling. Due to the contact constraint being unilateral which may fail to compute the gradient, We use sampling-based optimization driven by the above stability criterion. The distribution $D(\hat{\boldsymbol{\vartheta}})$ is initialized by Gaussian with $\tilde{\boldsymbol{\vartheta}}$ as the center and 0.1π as the variance of each dimension, and the distribution $D(\hat{\boldsymbol{t}})$ is initialized by Gaussian with $\mathbf{0}$ as the center and 0.05 as the variance of each dimension. In each iteration k, the samples $\{\tilde{\boldsymbol{\vartheta}}_k^{(n)}, \tilde{\boldsymbol{t}}_k^{(n)}\}_{n=1}^{N}$ with lower cost are given greater weight. Using these weighted samples, the variance of each dimension is updated before the resampling. In the last round, the lowest cost state, together with the contact point and contact force, is the result of hand-object interaction reconstruction.

Implementation details. In our experiment, the number of sampling iterations is set to $K = 30$, and the number of samples is set to $N = 300$. For each state sample, the interaction process is performed $T = 120$ steps in the physics engine. All the samples in the same iteration are simulated in parallel. The time step follows the default 240Hz setting in the bullet physics [8], *i.e.* each simulation process corresponds to 0.5s in the real physical world. Distance is measured in meters, mass in kilogram, and force in Newton. The gravity direction is considered to be down along the Y-axis in the image coordinates. For objects from other datasets [3, 21, 52, 61], the mass is proportional to its volume, and the density is uniformly set to $500\mathrm{kg/m}^3$. The restitution coefficient of both hand and object is set to 1.0. The friction coefficient between the hand and objects is set to 0.8. For the objects in our dataset, the mass and friction settings follow the actual measurement results contained in our **Sup. Mat** .



(a) Multi-view images with annotations



(b) Objects classified by shape

(c) Objects and measurements

Figure 6. **Interaction dataset with physical attributes**. (a) Multi-view dataset with mesh and mask annotations; (b) 20 objects classified into 6 categories; (c) Real models of our objects.

3.3. Interaction Dataset Preparation

As shown in Fig. 6, we created a dataset containing multi-view color images, hand and object visible masks, physical attributes, and stability degree measured by the magnitude of the extra balancing force. In summary, it contains 1K scenes of 20 subjects interacting with 20 objects captured by 25 cameras. These objects are classified into 6 categories according to their shape, including A) cones, B) prisms, C) cubes, D) spheres, E) disks, and F) columns. For more details about our dataset, please refer to **Sup. Mat** .

4. Experiments

In this section, the evaluating datasets and the criteria are first defined in Sec. 4.1. Our method is compared with the SOTA methods in Sec. 4.2. Detailed ablation studies are also conducted to our key components in Sec. 4.3.

4.1. Datasets and Metrics

Datasets. The existing dataset contains two main types. The first type [6, 14, 18] records real RGB images and the whole hand-object interaction process including approach, contact, and manipulation. This kind of data is used to test our entire pipeline. To reduce the ambiguity in the selection of interacted objects, we follow the method [21, 62] to filter these datasets with the 3D distance between the hand-object not exceeding 5mm as the threshold. The official testing set from HO3D [18] is not used due to the lack of hand mesh ground truth. In the end, the data used for testing contains 7,373 samples in FPHB [14], 69,292 samples in HO3Dv3 [18], and 93,264 samples in DexYCB [6]. Another type [3, 21, 52] focuses on recording the contact pattern of the hand-object. For each sequence in GRAB [52], we extracted the interaction sub-sequences containing contact between the right hand and object with 50 frames as an interval. This dataset is denoted as GRAB_{rh50}. In the end, the data used for testing contains 2,259 samples in ContactPose [18] and 19,008 samples in GRAB_{rh50} [52].

Datasets	ContactPose [3]			GRAB_{rh50}		
Methods	GT.	[17]	Ours‡	GT.	[17]	Ours‡
Max Pene.$(\text{mm})\downarrow$	11.62	12.07	**8.54**	10.33	12.38	**7.54**
Inter.$(\text{cm}^3)\downarrow$	12.24	12.35	**6.13**	14.62	13.97	**7.28**
Disp. $(\text{mm})\downarrow$	4.68	4.35	**1.02**	4.25	4.47	**1.23**
SC. $\downarrow$	1.46	1.03	**0.27**	1.34	1.28	**0.44**

Table 1. **Evaluations for Hand-Object Contact Estimation.** '$\text{Ours}^{\ddagger}$" denotes our method with optimization only.

State Error. Due to the hand mesh reconstructed by our method being different from MANO [47], the mean per-point position error (*MPJPE*) of 21 hand joints is chosen to evaluate the 3D reconstruction error. In 2D, the mean intersection over union (*mIOU*) is adopted to evaluate the re-projection error between the conversed mesh and the ground truth. As for the object, the vertices of the posed object are obtained by aligning the object reference mesh with the estimated ellipsoids. The mean per-vertex position error (*MPVPE*) and the *mIOU* are adopted to evaluate the object error.

Contact Quality. First, *max penetration* (Max Pene.) and *intersection volume* (Inter.) [21] are adopted to evaluate the geometric relationship. Then, *simulation displacement* (Disp.) [21] and our *stability cost* (SC.) defined in Sec. 3.2 are used to evaluate the contact stability in the same simulation settings. For a fair comparison, the ellipsoid hand is converted to the convex-hull mesh when computing these intersection metrics according to Sec. 3.1.

Sim-to-Real Gap. For each scene in our dataset, the correlation between the balancing force and the corresponding cost is used to evaluate the simulation effectiveness.

4.2. Comparisons

State Estimation. In the task of estimating hand-object state from monocular images, our method is compared with the methods using pure regression [19, 21] and the methods with additional optimization [5, 62]. As shown in Tab. 2, the hand-object state estimated from our front-end network has a better performance than the direct regressing method, and our full pipeline achieves the best results across data sets. This demonstrates that our approach outperforms other MANO-based regressions in terms of representation, and

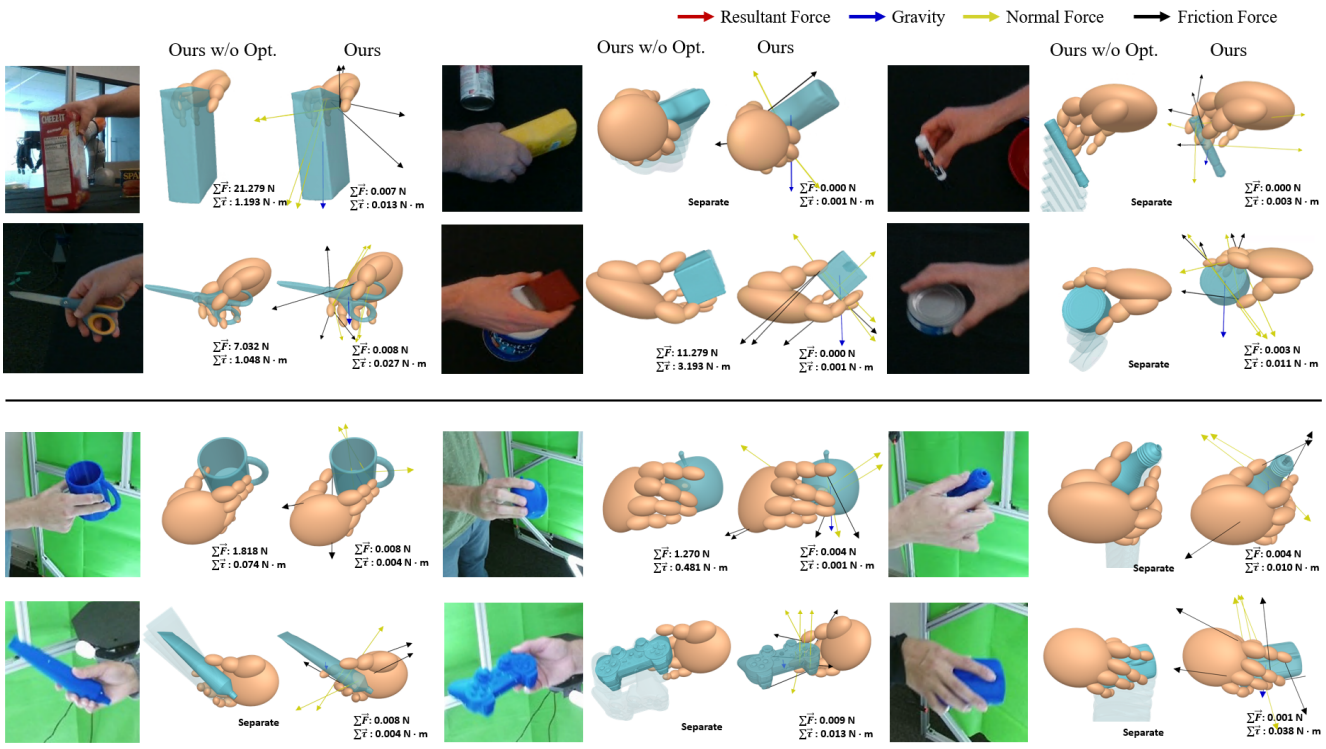


Figure 7. **Qualitative results on DexYCB [6] and ContactPose [3].** For each sample, the optimized state increases the stability of the contact while ensuring the consistency of the initial state estimated by the front-end network.

our recovery module can achieve effective optimization of the contact pattern. In some cases, the position accuracy of hand-object may be slightly influenced by the optimization with their stability increasing. This may be caused by the difference between real and simulated conditions.

Contact Recovery. By taking the hand-object state, our recovery module is compared with [17] under ContactPose and GRAB$_{rh50}$. As shown in Tab. 1, our method increases the stability of the contact while reducing the penetration. This further illustrates that the contact has been optimized more comprehensively with our method.

4.3. Ablation Study

Due to the ContactPose [3] having both images and accurate contact, most of our ablation experiments are based on this dataset. Among them, the results of completely using our entire pipeline are in the last row of Tab. 3.

Training Paradigms. The verifications of two key components in the training process, including semi-supervised pre-training and contact loss, are shown in the first two rows of Tab. 3. The lack of collision loss may worsen the initial state of the hand-object before optimization, which in turn affects the whole optimization process. On the other hand, the network without pre-training is less robust to the diversity of hand posture, the change of perspectives, and the occlusion during hand-object interaction, which may have similar effects on the entire pipeline.

Stability Cost. We compared the importance of each term in our stability cost, as shown in the middle 6 rows of Tab. 3. The lack of each item would weaken the final result, among which the force item has the greatest impact. Further, we replaced the stability cost with the displacement defined on our hand model as the objective of driving optimization, while the result becomes worse as shown in row 8 of Tab. 3. The main reason may be that the object displacement can only reflect the contact stability in fewer simulation steps. Therefore, our criteria could measure contact patterns more generally. The method with only collision detection in the physics engine is also employed, which does not have enough stability either.

Hand Model. As shown in row 9 and row 10 of Tab. 3, the choices of collision shape and local DoFs of our hand model are also explored under the same simulation conditions. Among them, the hand consisting of mesh segments leads to poorer stability. This may be caused by the fact that the mesh collision shape in the physics engine is automatically approximated as a convex hull, which changes the accuracy of collision detection. On the other hand, the hand with more local DoFs has lower accuracy because it increases the difficulty of optimization. To improve the efficiency of sampling and optimization methods, the methods with 20 local DoFs were adopted.

Datasets	FPHB [14]					HO3Dv3 [18]					DexYCB [6]		
Methods	[21]	[19]	[62]	Ours†	Ours	[19]	[62]	[5]	Ours†	Ours	GT.	Ours†	Ours
$mIoU_H(\%)\uparrow$	-	54.54	-	59.34	**62.01**	64.04	-	-	**61.52**	61.43	-	62.64	**63.52**
$\text{MPJPE}_H(\text{mm})\downarrow$	28.80	19.32	-	19.10	**18.56**	14.32	-	9.50	10.96	**9.14**	-	11.32	**11.15**
$mIoU_O(\%)\uparrow$	-	66.10	-	71.34	**72.58**	75.26	-	-	**82.53**	82.47	-	80.66	**81.34**
$\text{MPVPE}_O(\text{mm})\downarrow$	-	21.07	21.57	21.14	**20.96**	20.08	73.28$^\diamond$	-	**19.34**	19.45	-	18.61	**18.84**
Max Pene.$(\text{mm})\downarrow$	15.12	18.08	16.92	15.07	**11.43**	10.29	16.47	-	16.85	**11.36**	10.65	7.32	**6.72**
Inter.$(\text{cm}^3)\downarrow$	10.90	11.05	11.76	10.12	**6.23**	12.26	7.44	-	7.32	**6.19**	14.76	6.94	**6.61**

Table 2. **Evaluations for Hand-object State Estimation.** "Ours†" denotes our method without optimization, "Ours" denotes our full pipeline. The item marked by "-" indicates that the work has not been trained or tested on the relevant dataset. The item marked by "$\diamond$" denotes the wrist-relative object vertex error.

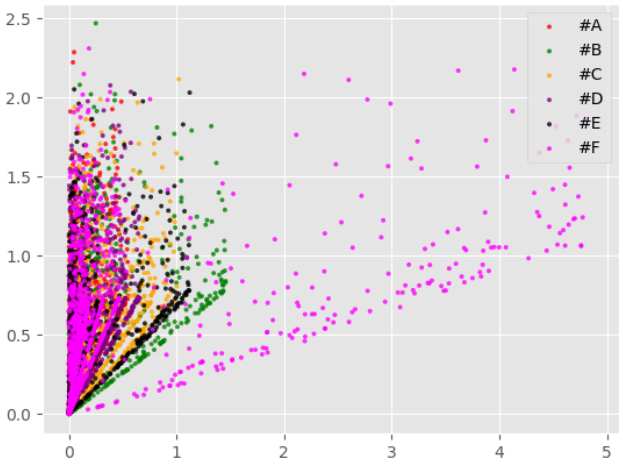


Figure 8. **Correlation between stability cost and real forces**. The horizontal axis is the real-measured force, and the vertical axis is the corresponding stability cost of the same hand-object state reconstructed in the simulation environment. The data corresponding to 6 types of objects are marked with different colors.

Sim-to-Real Correlation. With interaction scenes in our dataset, we quantitatively analyze the relationship between simulation stability cost and actual compensation force. For more details about the balancing force measurements and corresponding physical properties, please refer to our **Sup. Mat** . In the experiment, each hand-object scene reconstructed was used to directly initialize our simulation interaction scene, and then their stability cost was calculated. Each reconstructed hand object scene is used to directly initialize our simulation interaction scene. Their actual stability during capturing is measured by the scale of the balancing force, and the stability in the simulation is measured by the stability cost. The mass and friction coefficient of the object in the simulation is set to be the same as the actual measured values. As shown in Fig. 8, for objects with different shapes, the actual stability and simulation stability have different correlations. Among them, objects in class F (*i.e.* columns) correspond to multiple slopes, which is caused by the great scale variations within the categories.

Method	Inter.$(\text{cm}^3)\downarrow$	Disp. $(\text{mm})\downarrow$	SC.$\downarrow$
w/o L_{in}	7.41	4.65	0.86
w/o pre-trained	7.34	3.77	0.51
w/o C_{S} Opt.	6.32	3.43	4.62
w/o $L_{cnt.}$	6.28	2.39	0.58
w/o $L_{frc.}$	6.23	2.66	0.73
w/o $L_{tau.}$	6.37	2.17	0.64
w/o C_{stab} Opt.	7.32	1.92	1.44
Opt. with Disp.	6.94	3.43	4.62
w/o Ellipsoids	6.36	1.59	0.64
with $\lvert\boldsymbol{\vartheta}\rvert = 48$	6.32	1.47	0.47
Ours	**6.24**	**1.13**	**0.31**

Table 3. **Ablation study on ContactPose.** The components in network training paradigm, optimization function and physical hand model are evaluated.

5. Conclusion

This paper proposes a novel monocular hand-object contact recovery scheme driven by the simulated stability criteria in the physics engine. Through sampling-based optimization, a more stable contact pattern is obtained without data prior dependence. A hand-object ellipsoid representation further promotes the effective implementation of our regression-optimization pipeline. It enables personalized hand shape variations at the same time. The sim-to-real consistency is verified later by our contact scene dataset with real physical properties and stability evaluation.

Limitations and Future Work. Although our method is robust under existing datasets, it may become invalid in a complex scene with severe occlusion or multiple hands/objects. Getting rid of object mesh dependence is also significant for the improvement of our approach. In the future, rewards with our stability cost considerations could more effectively guide reinforcement learning methods to reconstruct hand-object interaction sequences.

References

[1] Jules Bloomenthal and Ken Shoemake. Convolution surfaces. In *Proceedings of the 18th annual conference on Computer graphics and interactive techniques*, pages 251–256, 1991. 3

[2] Samarth Brahmbhatt, Cusuh Ham, Charles C Kemp, and James Hays. Contactdb: Analyzing and predicting grasp contact via thermal imaging. In *CVPR*, pages 8709–8719, 2019. 2

[3] 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*, pages 361–378. Springer, 2020. 1, 2, 4, 5, 6, 7

[4] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *2015 international conference on advanced robotics (ICAR)*, pages 510–517. IEEE, 2015. 2

[5] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, pages 12417–12426, 2021. 1, 2, 6, 8

[6] 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*, pages 9044–9053, 2021. 2, 4, 6, 7, 8

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

[8] Erwin Coumans et al. Bullet physics library. *Open source: bulletphysics. org*, 15(49):5, 2013. 1, 5

[9] Boyang Deng, John P Lewis, Timothy Jeruzalski, Gerard Pons-Moll, Geoffrey Hinton, Mohammad Norouzi, and Andrea Tagliasacchi. Nasa neural articulated shape approximation. In *ECCV*, pages 612–628. Springer, 2020. 2

[10] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *CVPR*, pages 6608–6617, 2020. 2

[11] Kiana Ehsani, Shubham Tulsiani, Saurabh Gupta, Ali Farhadi, and Abhinav Gupta. Use the force, luke! learning to predict physical forces by simulating effects. In *CVPR*, pages 224–233, 2020. 2

[12] Roy Featherstone. *Rigid body dynamics algorithms*. Springer, 2014. 5

[13] Shachar Fleishman, Mark Kliger, Alon Lerner, and Gershom Kutliroff. Icpik: Inverse kinematics based articulated-icp. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–35, 2015. 2

[14] 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*, pages 409–419, 2018. 2, 6, 8

[15] 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*, pages 10833–10842, 2019. 2

[16] Francisco Gomez-Donoso, Sergio Orts-Escolano, and Miguel Cazorla. Large-scale multiview 3d hand pose dataset. *Image and Vision Computing*, 81:25–33, 2019. 2

[17] Patrick Grady, Chengcheng Tang, Christopher D Twigg, Minh Vo, Samarth Brahmbhatt, and Charles C Kemp. Contactopt: Optimizing contact to improve grasps. In *CVPR*, pages 1471–1481, 2021. 1, 2, 6, 7

[18] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. In *CVPR*, pages 3196–3206, 2020. 2, 4, 6, 8

[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*, pages 571–580, 2020. 2, 6, 8

[20] Yana Hasson, Gül Varol, Ivan Laptev, and Cordelia Schmid. Towards unconstrained joint hand-object reconstruction from rgb videos. *arXiv preprint arXiv:2108.07044*, 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*, pages 11807–11816, 2019. 2, 4, 5, 6, 8

[22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 3

[23] Stefan Hinterstoisser, Vincent Lepetit, Slobodan Ilic, Stefan Holzer, Gary Bradski, Kurt Konolige, and Nassir Navab. Model based training, detection and pose estimation of texture-less 3d objects in heavily cluttered scenes. In *ACCV*, pages 548–562. Springer, 2012. 2

[24] Markus Höll, Markus Oberweger, Clemens Arth, and Vincent Lepetit. Efficient physics-based implementation for realistic hand-object interaction in virtual reality. In *2018 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pages 175–182. IEEE, 2018. 5

[25] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *CVPR*, pages 3385–3394, 2019. 2

[26] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, pages 118–134, 2018. 2

[27] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *ICCV*, pages 11107–11116, 2021. 2

[28] Korrawe Karunratanakul, Adrian Spurr, Zicong Fan, Otmar Hilliges, and Siyu Tang. A skeleton-driven neural occupancy representation for articulated hands. In *International Conference on 3D Vision (3DV)*, pages 11–21. IEEE, 2021. 2, 4

[29] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. In *International Conference on 3D Vision (3DV)*, pages 333–344. IEEE, 2020. 2

[30] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *ICCV*, pages 1521–1529, 2017. 2

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

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

[33] Xingyu Liu, Shun Iwase, and Kris M Kitani. Stereobj-1m: Large-scale stereo image dataset for 6d object pose estimation. In *ICCV*, pages 10870–10879, 2021. 2

[34] Khaled Mamou, E Lengyel, and A Peters. Volumetric hierarchical approximate convex decomposition. In *Game Engine Gems 3*, pages 141–158. AK Peters, 2016. 2

[35] Marko Mihajlovic, Yan Zhang, Michael J Black, and Siyu Tang. Leap: Learning articulated occupancy of people. In *CVPR*, pages 10461–10471, 2021. 2, 4

[36] Brian Vincent Mirtich. *Impulse-based dynamic simulation of rigid body systems*. University of California, Berkeley, 1996. 5

[37] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *ECCV*, pages 752–768. Springer, 2020. 2

[38] Gyeongsik Moon, Takaaki Shiratori, and Kyoung Mu Lee. Deephandmesh: A weakly-supervised deep encoder-decoder framework for high-fidelity hand mesh modeling. In *ECCV*, pages 440–455. Springer, 2020. 2, 3

[39] Franziska Mueller, Micah Davis, Florian Bernard, Oleksandr Sotnychenko, Mickeal Verschoor, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Real-time pose and shape reconstruction of two interacting hands with a single depth camera. *ACM TOG*, 38(4):1–13, 2019. 3

[40] S Narasimhaswamy, T Nguyen, and M Hoai. Detecting hands and recognizing physical contact in the wild. *Advances in neural information processing systems*, 2020. 2

[41] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, volume 1, page 3, 2011. 3

[42] 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*, pages 2088–2095. IEEE, 2011. 3

[43] Sida Peng, Yuan Liu, Qixing Huang, Xiaowei Zhou, and Hujun Bao. Pvnet: Pixel-wise voting network for 6dof pose estimation. In *CVPR*, pages 4561–4570, 2019. 2

[44] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE TPAMI*, 40(12):2883–2896, 2017. 2

[45] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *CVPR*, pages 1106–1113, 2014. 3

[46] James M Rehg and Takeo Kanade. Visual tracking of high dof articulated structures: an application to human hand tracking. In *ECCV*, pages 35–46. Springer, 1994. 2

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

[48] Dandan Shan, Jiaqi Geng, Michelle Shu, and David F Fouhey. Understanding human hands in contact at internet scale. In *CVPR*, pages 9869–9878, 2020. 2

[49] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. Fast and robust hand tracking using detection-guided optimization. In *CVPR*, pages 3213–3221, 2015. 3

[50] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *ICCV*, pages 2456–2463, 2013. 3

[51] Srinath Sridhar, Helge Rhodin, Hans-Peter Seidel, Antti Oulasvirta, and Christian Theobalt. Real-time hand tracking using a sum of anisotropic gaussians model. In *International Conference on 3D Vision (3DV)*, volume 1, pages 319–326. IEEE, 2014. 3

[52] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, pages 581–600. Springer, 2020. 1, 2, 5, 6

[53] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, pages 4511–4520, 2019. 2

[54] Bugra Tekin, Sudipta N Sinha, and Pascal Fua. Real-time seamless single shot 6d object pose prediction. In *CVPR*, pages 292–301, 2018. 2

[55] Daniel Thul, L'ubor Ladickỳ, Sohyeon Jeong, and Marc Pollefeys. Approximate convex decomposition and transfer for animated meshes. *ACM TOG*, 37(6):1–10, 2018. 2

[56] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. Sphere-meshes for real-time hand modeling and tracking. *ACM TOG*, 35(6):1–11, 2016. 3

[57] Anastasia Tkach, Andrea Tagliasacchi, Edoardo Remelli, Mark Pauly, and Andrew Fitzgibbon. Online generative model personalization for hand tracking. *ACM TOG*, 36(6):1–11, 2017. 3

[58] 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*, 118(2):172–193, 2016. 1, 2, 3

[59] Chengde Wan, Thomas Probst, Luc Van Gool, and Angela Yao. Self-supervised 3d hand pose estimation through training by fitting. In *CVPR*, pages 10853–10862, 2019. 3

[60] Jiayi Wang, Franziska Mueller, Florian Bernard, Suzanne Sorli, Oleksandr Sotnychenko, Neng Qian, Miguel A Otaduy, Dan Casas, and Christian Theobalt. Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. *ACM TOG*, 39(6):1–16, 2020. 2

[61] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017. 2, 5

[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*, pages 11097–11106, 2021. 1, 2, 5, 6, 8

[63] Xiong Zhang, Qiang Li, Hong Mo, Wenbo Zhang, and Wen Zheng. End-to-end hand mesh recovery from a monocular rgb image. In *ICCV*, pages 2354–2364, 2019. 2

[64] Zhengyi Zhao, Tianyao Wang, Siyu Xia, and Yangang Wang. Hand-3d-studio: A new multi-view system for 3d hand reconstruction. In *ICASSP*, pages 2478–2482. IEEE, 2020. 2, 3

[65] Zimeng Zhao, Xi Zhao, and Yangang Wang. Travelnet: Self-supervised physically plausible hand motion learning from monocular color images. In *ICCV*, pages 11666–11676, 2021. 5

[66] 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*, pages 5346–5355, 2020. 2

[67] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, pages 4903–4911, 2017. 2

[68] Christian Zimmermann, Duygu Ceylan, Jimei Yang, Bryan Russell, Max Argus, and Thomas Brox. Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In *ICCV*, pages 813–822, 2019. 2, 3