Dynamic Reconstruction of Hand-Object Interaction with Distributed Force-aware Contact Representation

Zhenjun Yu[1*], Wenqiang Xu[1*], Pengfei Xie[1], Yutong Li[1],
Brian W. Anthony[2], Zhuorui Zhang[2], Cewu Lu[1§]
[1]Shanghai Jiao Tong University, [2]Massachusetts Institute of Technology
[1]{jeffson-yu, vinjohn, pf.xie, davidliyutong, lucewu}@sjtu.edu.cn
[2]{banthony, zhuoruiz}@sjtu.edu.cn

Abstract

We present ViTaM-D, a novel visual-tactile framework for reconstructing dynamic hand-object interaction with distributed tactile sensing to enhance contact modeling. Existing methods, relying solely on visual inputs, often fail to capture occluded interactions and object deformation. To address this, we introduce DF-Field, a distributed force-aware contact representation leveraging kinetic and potential energy in hand-object interactions. ViTaM-D first reconstructs interactions using a visual network with contact constraint, then refines contact details through force-aware optimization, improving object deformation modeling. To evaluate deformable object reconstruction, we introduce the HOT dataset, featuring 600 hand-object interaction sequences in a high-precision simulation environment. Experiments on DexYCB and HOT datasets show that ViTaM-D outperforms state-of-the-art methods in reconstruction accuracy for both rigid and deformable objects. DF-Field also proves more effective in refining hand poses and enhancing contact modeling than previous refinement methods. The code, models, and datasets are available at https://sites.google.com/view/vitam-d/.

1. Introduction

Humans manipulate objects using both visual and tactile feedback intuitively. Modeling and reconstructing this hand-object interaction can empower many downstream tasks, including VR/AR, robotic imitation learning, and human behavior understanding. Existing methods for dynamic hand-object interaction reconstruction [4, 12, 29, 33, 36] primarily rely on vision to recover the global geometry and estimate hand-object poses. However, these methods struggle to capture the occluded contact details, particularly the object deformation at the contact interface.

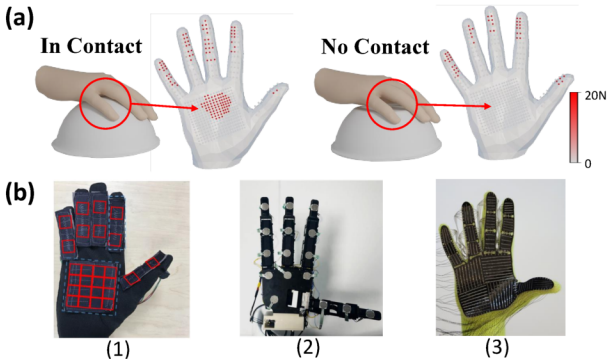

Figure 1. (a): The relationship between tactile information and contact geometry. Grasping different bowls with the same hand poses shows that distributed tactile arrays can capture occlusion contacts and object states. (b): Different types of distributed tactile sensors proposed in previous works (1). [18], (2). [38], (3). [26].

Recent advances in tactile sensing [18, 23, 26, 39] inspire the development of visual-tactile approaches for hand-object reconstruction tasks [25, 30, 31] to leverage fine-grained perceptual signals from contact regions. (Fig. 1.a). Among various tactile sensor types, distributed tactile sensors (Fig. 1.b) [18, 26, 38], in form factors of wearable gloves or attachable sensor arrays, show great promise in human manipulation data collection comparing to other counterparts [16, 23, 27, 39]. A previous work, ViTaM [15] proposed a novel high-density tactile glove system that is capable of generating high-accuracy force data. Meanwhile, ViTaM utilized a transformer module to directly fuse tactile sensory data with visual perception for object reconstruction, which ignores the unbalanced information between forces and point cloud features. Therefore, how to effectively integrate distributed tactile sensors and state-of-the-art visual perception tools for refined hand–object reconstruction remains underexplored.

Since distributed tactile sensors can conform closely to

* indicates equal contributions. §Cewu Lu is the corresponding author.

the hand while maintaining the hand's geometry and natural dexterity during manipulation, we can use an off-the-shelf visual approach to reconstruct the hand-object state and integrate the tactile information to refine the results. In light of this, we propose a novel **D**istributed **F**orce-aware contact representation, **DF-Field**, and a **Vi**sual-**Ta**ctile **M**anipulation reconstruction framework with **D**istributed tactile sensing, **ViTaM-D**. DF-Field models the contact by considering both kinetic and potential energy in hand-object interaction. This approach enables accurate modeling of object deformation. Using DF-Field, ViTaM-D initially reconstructs the hand-object interaction with visual observations and contact state constraints by the proposed visual dynamic tracking network, and then refines the contact details with DF-Field via a force-aware hand-pose optimization process. The proposed framework leverages advances in visual-only hand-object reconstruction methods [4, 12, 29, 36] while seamlessly integrating the tactile information into an existing motion capture or estimation system to enhance the fidelity of contact modeling and overall reconstruction quality.

To build and train such a visual-tactile framework, large-scale datasets that capture diverse hand-object interactions with precise tactile readings are required, especially for deformable objects. However, existing datasets for hand-object interactions [3, 9, 35] only cover rigid or articulated object manipulation and generally lack accurate tactile data. Therefore, alongside using the public dataset (*e.g*. DexYCB [3]) to benchmark on *rigid* objects with common baselines [4, 6], we also introduced a new dataset, **HOT dataset**, to fully benchmark our method on *deformable* object reconstruction. This dataset is built with ZeMa [8], a high-precision physics-based simulation environment that supports penetration-free frictional contact modeling with finite element method (FEM) to model the object deformation. The HOT dataset contains 600 sequences of hand-object interaction, with 30 deformable objects from 5 different categories and 8 camera views for each sequence.

To evaluate the method, we compare our proposed ViTaM-D with previous state-of-the-art methods gSDF [6], HOTrack [4] on DexYCB, and ViTaM[15] on our proposed HOT dataset. Results show that our method significantly improves the reconstruction performance compared to previous works. We successfully demonstrated the superiority of our method in tracking both rigid and deformable objects. Besides, the effectiveness of the force-aware optimization has been validated in refining hand poses to eliminate penetrations and suboptimal contact states.

Our contributions are summarized as follows:

(1) A visual-tactile learning framework, ViTaM-D. It contains a visual dynamic tracking network for reconstructing hand-object interactions, and a force-aware optimization process, which integrates the tactile information into reconstruction refinement based on a novel distributed force-aware contact representation, DF-Field.

(2) A new dataset, HOT. It contains 600 RGB-D manipulation sequences on 30 deformable objects from 5 categories with penetration-free hand-object poses and accurate tactile information.

2. Related Work

Hand-object reconstruction has been richly studied because it tries to recover the full details of hand-object interaction, showing potential applicability for downstream tasks.

Research in this direction starts with static reconstruction. Earlier works have predominantly relied on RGB inputs to estimate hand pose and object geometry. Hasson et al. [11] presented a method that learns hand-object interaction using synthetic RGB images with the grasp poses planned by GraspIt! [20]. Doosti et al. [7] proposed a graph-based network for hand-object pose estimation. Cao et al. [2] introduced a method that operates in-the-wild, estimating the hand pose first and optimizing the solution using a contact representation. The optimization involves push and pull terms to handle the contact, which is purely empirical. Similarly, CPF [33, 36] employs a contact representation using a spring-mass system, adopting an empirical approach to refine the hand-object interaction. ArtiBoost [34] further enhanced the performance of static hand-object reconstruction through data augmentation techniques. AlignSDF [5] leveraged RGB inputs to estimate hand pose and combined point cloud data for object geometry, using SDF-based decoders for both hand and object reconstruction.

Later, hand-object reconstruction in the dynamic setting draws increasing attention since manipulation is naturally dynamic. Approaches have advanced by incorporating temporal information and using more complex representations. Tekin et al. [28] adopted RNN for temporal feature fusion, utilizing egocentric RGB video to capture dynamic hand-object interaction and pose estimation. Hasson et al. [13] improved dynamic reconstruction by enforcing photometric consistency constraints. Ye et al. [37] introduced a diffusion model that guides dynamic hand-object reconstruction via score distillation sampling and differentiable rendering techniques. Recent advances by Fan et al. [10] focused on refining pose estimation using SDF-based representations and volumetric rendering, improving the overall dynamic reconstruction accuracy. Furthermore, gSDF [6] employed transformer architectures and SDF representations to model complex hand-object interactions dynamically.

These works focus more on visual-only inputs. However, due to occlusion between the hand and object during the interaction, visual perception usually lacks information near the contact areas, and such information cannot necessarily be mitigated by cross-frame feature fusion. Therefore, tactile perception comes to supplement the near-contact in-

formation. Zhang et al. [41] utilized a tactile glove [26] to track object movements. However, it focuses more on the dynamic object trajectory rather than contact geometry. Works such as [25, 30] employed optical tactile sensors, trained the model by synthetic data, and applied to rigid object geometry reconstruction. Later, VTacO [31] extended this line of research by using optical tactile sensors to capture object geometry, including deformation, providing a more comprehensive representation of hand-object interactions. Jiang et al. [15] proposed a visual-tactile recording and tracking system ViTaM, which contains a wearable tactile glove along with a visual-tactile reconstruction algorithm. However, it simply used neural networks to fuse tactile and visual data and did not consider the information mismatch between them. Unlike these works, the proposed ViTaM-D inherits the merits from both worlds: it leverages the advanced techniques of visual reconstruction and incorporates the distributed tactile readings to enhance the local contact details.

3. Force-aware Contact Representation

During manipulation, forces are reciprocal, so the force exerted by the hand affects the state of both the hand and the object mutually. To fully capture dynamic contact behaviors, the representation must encode both contact locations and forces. We introduce a distributed force-aware contact representation, **DF-Field** (Sec. 3.1), and apply it using distributed tactile sensors (Sec. 3.2).

3.1. DF-Field Representation

Without loss of generality, we take the object-centric perspective to describe a hand-object interaction process, where the object is assumed to be fixed in the origin point, and the hand moves around the object. In this way, by ignoring the gravitational potential energy, the rest of the contact dynamics is the kinetic energy and deformation potential energy of the hand and object, which are described by a **Relative Potential Energy**. Besides, to prevent penetration between the hand and object contact surface, we assume a virtual **Barrier Energy**. With an established point pair i, j of object vertex $\mathcal{V}_i^o$ and hand vertex $\mathcal{V}_j^h$, and the Euclidean distance of the point pair $l_{ij} = \left\| \mathcal{V}_i^o - \mathcal{V}_j^h \right\|_2$, we define the two energy terms as follows.

Relative Potential Energy. A relative potential energy that describes both the object deformation and the hand movements:

$$E_{ij} = \kappa l_{ij}^2, \tag{1}$$

where κ is a parameter representing the hand interacting with the object vertices. If $\kappa > 0$, the hand and object are in contact. Thus, the distance l_{ij} and the relative energy will be close to 0, indicating that the contact's relative potential energy is satisfied.

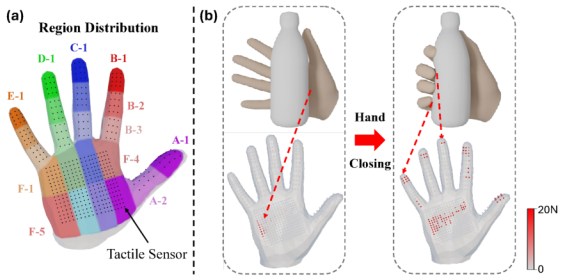

Figure 2. (a): The 22 regions and a typical distributed tactile sensor layout [26]. (b): An example of data collection in the simulation environment, generating depth images and tactile sensor readings.

Barrier Energy. Given a certain threshold distance $\hat{l}$, the barrier energy is:

$$B_{ij} = \begin{cases} -e^{-\kappa}(l_{ij} - \hat{l})^2 \log\left(\frac{l_{ij}}{\hat{l}}\right), & 0 < l_{ij} < \hat{l} \\ 0 & l_{ij} \geq \hat{l} \end{cases} \tag{2}$$

Adopted from [17], the barrier energy aims to push away the point pair when κ is low, thus avoiding penetration issues between the hand and object. The function is defined in this way not only to ensure repulsion becomes smaller when the distance is larger but also to remain smooth for optimization.

Overall Energy. A proper contact is met when both energy terms approach 0:

$$E = \sum_i \sum_j (E_{ij} + B_{ij}). \tag{3}$$

To note, though κ is strongly correlated to the tactile readings, DF-Field can also be tactile-independent, as long as we empirically set the exerted force instead of physics-based. In this way, DF-Field can work with visual-only methods. The details are discussed in Sec. 6.5.

3.2. DF-Field with Distributed Tactile Sensors

As aforementioned, the distributed tactile sensors generally conform to the hand, and different tactile sensors may have different layout configurations. Thus, to adapt to different tactile sensors, we take the hand-centric perspective by first dividing the hand into 22 regions (as shown in the left of Fig. 2.a, 2 areas for the thumb, 3 for other fingers, and 8 for the palm), and assign the tactile sensors to the corresponding region. In each region, we define the center as the hand keypoint, resulting in $\mathcal{K}^h \in \mathbb{R}^{22\times 3}$.

For optimization, we connect only the hand keypoints to object vertices, reducing computational load while enhancing regional interaction information. The force in each region is calculated by averaging the tactile readings $\mathcal{M}^j$. By the definition of the energy, we can obtain that κ is the difference of the exerted force over the distance l_{ij}. In practice, we approximate it by dividing the force by the distance:

$$\kappa_{ij} \sim \frac{\overline{\mathcal{M}^j}}{l_{ij}}. \tag{4}$$

We discuss the difference between using keypoints or all-hand vertices for optimization in the supplementary file.

4. ViTaM-D Method

4.1. Overview

We address the problem of 4D tracking and dynamic reconstruction of hand-object interactions from a visual-tactile perspective. Based on DF-Field, we propose a novel method named **ViTaM-D**, which consists of two stages: visual dynamic tracking and force-aware optimization. We assume that the visual input consists of a live stream of 3D point clouds, denoted as $\{\mathcal{P}_t \in \mathbb{R}^{N\times 3}\}_{t=1}^{n}$, representing hand-object interactions derived from single-view depth images, where N is the number of input point, n is the number of frames. The tactile input $\{\mathcal{M}_t\}_{t=1}^{n}$ contains distributed tactile sensor readings for each frame.

In the first stage, ViTaM-D uses visual input to establish flow features and dynamically reconstruct the hand and object. Contact information is used to constrain the reconstruction to capture global hand-object geometry. In the second stage, tactile information is incorporated to improve contact accuracy and reduce penetration issues based on our proposed DF-Field. The overall pipeline is given in Fig. 3.

4.2. Visual dynamic tracking of hand and object

In the first stage, to introduce correspondence feature from the previous frame, the visual dynamic tracking net predicts the point cloud flow from the last to the current frame using a *flow prediction module*, and extracts a fused visual feature containing both correspondence feature and static information. This feature is forwarded into an *object decoder* and a *hand decoder* to reconstruct the object and hand geometry, respectively. To improve global geometry accuracy, we introduce a novel *contact constraint* that reasons about object geometry based on hand reconstruction. Objects are modeled using signed distance fields (SDF), while the hand is represented using the MANO model [24].

Flow Prediction Module. At frame t, we first extract the per-point features $F_t, F_{t-1} \in \mathbb{R}^{N\times d}$ from $\mathcal{P}_t$ and $\mathcal{P}_{t-1}$ using a backbone network $I(\cdot)$. In practice, we adopt a simple PointNet++ [22] for feature extraction, with 3 layers of set abstraction and 3 layers of feature propagation. Then, we design a flow prediction network $\mathcal{F}_f(\cdot)$ to predict the point cloud flow $f_{t-1\to t} \in \mathbb{R}^{N\times 3}$ from frame $t-1$ to t, which contains the correspondence information between the two frames and represents the hand movement and object deformation:

$$f_{t-1\to t} = \mathcal{F}_f(F_t, F_{t-1}, \mathcal{P}_t, \mathcal{P}_{t-1}). \tag{5}$$

The flow prediction module consists of several MLPs and convolutional networks. After obtaining the point cloud flow $f_{t-1\to t}$, we use another PointNet++ $I'(\cdot)$ to extract the visual correspondence feature F_t^f from $\mathcal{P}_{t-1}$ with $f_{t-1\to t}$ added. F_t^f and F_t are then forwarded to a transformer fusion module to obtain the final visual feature F_v, corresponding to the current frame's point cloud $\mathcal{P}_t$. Specifically, the transformer fusion module first uses self-attention to encode the point cloud features. After applying positional embedding between the feature and its point cloud, a cross-attention module fuses the two features and outputs F_v. This fusion strategy considers both 3D static information of the current frame and the corresponding feature extracted from the flow. We will discuss the detailed architecture of the flow prediction module, the importance of it, and flow prediction results in the supplementary material.

Object Decoder. The *object decoder* processes the final visual feature F_v in two steps: *feature scattering & sampling* and *SDF decoding*, to predict SDF values and reconstruct the object mesh using the Marching Cubes algorithm [19]. Following ConvOccNet [21], we scatter features into a volume of resolution D, process them with a 3D-UNet, and predict signed distances using a 5-layer MLP.

Hand Decoder. For hand reconstruction, we use the parametric MANO model [24], with $\beta \in \mathbb{R}^{10}$ for shape and $\theta \in \mathbb{R}^{51}$ for pose. Based on F_v and the previous frame's point cloud $\mathcal{P}_{t-1}$, we predict hand joint positions $\mathcal{J}_h \in \mathbb{R}^{21\times 3}$ using a voting mechanism inspired by PVN3D [14]. This involves predicting a translation offset $O_t \in \mathbb{R}^{21\times 3}$ and a probability matrix $T_{t-1} \in \mathbb{R}^{N\times 21}$ to compute:

$$\mathcal{J}_h = O_t + \mathcal{P}_{t-1} \times T_{t-1}. \tag{6}$$

Joint poses θ are then estimated via inverse kinematics [43], and the hand mesh is reconstructed using the MANO layer.

Contact Constraint. In order to obtain a relatively reasonable contact state of the reconstructed hand and object, we use a contact constraint to better optimize the prediction of the SDF of every sample position. Assuming we have the contact information $c_x \in \{0, 1\}$ of a sample position $x \in \mathcal{X}$, then if $c_x = 1$, it indicates that at this position, the object is in contact with the hand, meaning its signed distance value s_x should be 0, and otherwise not equal to 0. Therefore, the contact constraint should be as follows:

$$\mathcal{L}_C = \sum_{x\in\mathcal{X}} s_x \cdot \mathbb{1}_{c_x=1}. \tag{7}$$

If the tactile readings $\mathcal{M}$ can be obtained by haptic sensors, we can locate regional tactile arrays $\mathcal{M}^j$ through the hand reconstruction result, and the contact state c_x of sample points whose distances are less than l_c from those active tactile sensors will be set to 1, otherwise 0. In the absence of access to tactile signals, we can also use a simple network to predict contact states, which consists of a PointNet-based classifier with the input of sample positions and the final visual feature F_v. We will discuss the different effects of tactile readings

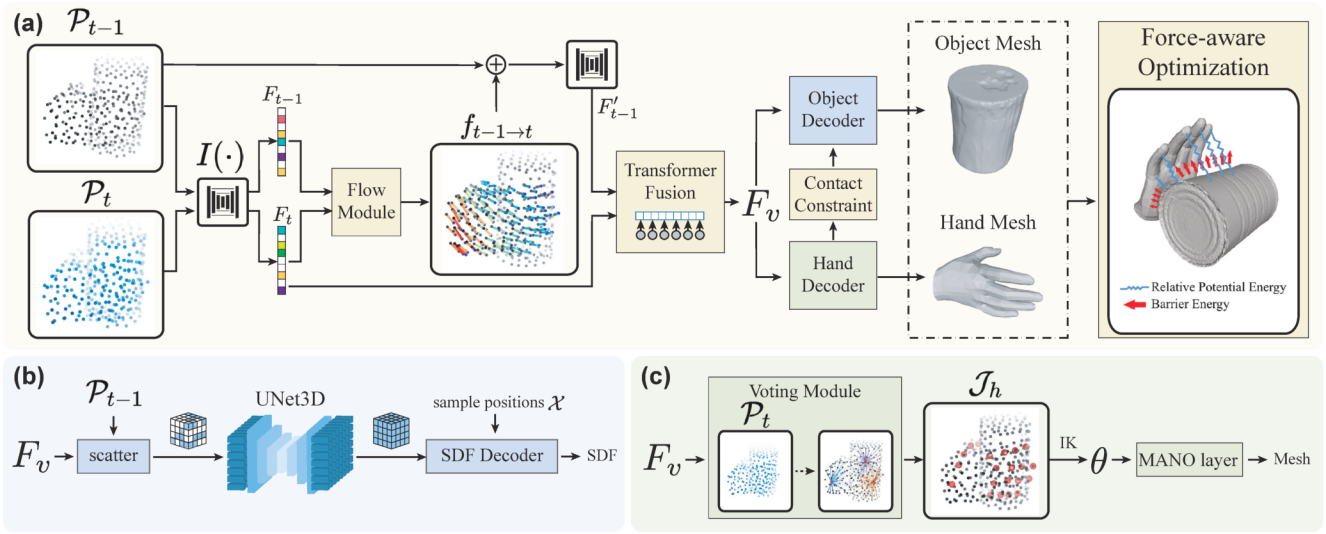

Figure 3. **ViTaM-D** pipeline. (a): The overview of our pipeline, including the Flow Prediction module for visual feature extraction and flow estimation, Hand and Object Decoders, the Contact Constraint, and Force-aware hand-pose Optimization. (b): The Object Decoder to reconstruct the object mesh. (c): The Hand Decoder for estimating hand parameters and reconstructing based on the MANO model.

and network prediction for acquiring interaction states in Sec. 6.5.

Loss Function. The network is trained end-to-end with four loss terms:

$$\mathcal{L} = \lambda_f \mathcal{L}_{flow} + \lambda_S \mathcal{L}_{SDF} + \lambda_H \mathcal{L}_{Hand} + \lambda_C \mathcal{L}_C. \quad (8)$$

Flow loss $\mathcal{L}_{flow}$ uses Chamfer distance between forward- and backward-shifted point clouds. SDF loss $\mathcal{L}_{SDF} = |s - s^*|$ compares predicted and ground truth SDF values. Hand joint loss $\mathcal{L}_{Hand} = \left\|\mathcal{J}_h - \mathcal{J}_h^*\right\|_2^2$ penalizes deviations from ground truth keypoints. The contact constraint $\mathcal{L}_C$ ensures accurate SDF predictions based on contact states.

4.3. Force-aware Hand-pose Optimization

The visual dynamic tracking net can give decent outputs of hand-object reconstruction. However, even though the contact constraint module has taken contact states into account for object reconstruction, the hand pose still needs refinement, for further improving the contact details. Here, we introduce the force-aware optimization for hand pose.

Given the contact energy defined in Sec. 3, we optimize the predicted pose θ from the hand decoder, with respect to the reconstructed object mesh, to obtain a better hand mesh and contact state. Based on Eq. 3, we use the ball-query method to find the corresponding object vertices for each hand region with a radius R, and the point pair will be set up between them and the hand keypoints $\mathcal{J}_h$.

Besides, For a given joint j, we ensure it remains in reasonable poses by penalizing rotations $\mathcal{R}_j$ that are near twisted directions $\mathcal{R}_t$ or if any angle exceeds $\pi/2$, using the L_2 loss. Additionally, we constrain the optimized hand pose ${}^*\theta$ to stay close to the original prediction:

$$\begin{aligned} \mathcal{L}_r &= \left\|\mathcal{R}_j \cdot \mathcal{R}_t\right\|_2^2 + \left\|\max(|\mathcal{R}_j| - \frac{\pi}{2}, 0)\right\|_2^2, \\ \mathcal{L}_o &= \left\|{}^*\theta - \theta\right\|_2^2. \end{aligned} \quad (9)$$

Finally, the optimization target can be demonstrated as:

$${}^*\theta = \arg\min_{\theta}(E + \mathcal{L}_r + \mathcal{L}_o). \quad (10)$$

We use the gradient descent method and the Adam solver for optimization. By minimizing energy and loss, we can obtain a better interaction state between the hand and object, avoid severe penetration problems, and forbid the hand from abnormal poses.

5. Hand-Object Tactile Dataset, HOT Dataset

Obtaining ground truth for object deformation in real-world scenarios is challenging, and prior works [3, 35] on hand-object interaction primarily focus on rigid or articulated objects. To benchmark our method's performance on deformable objects, we introduce the **H**and-**O**bject-**T**actile (**HOT**) dataset, featuring depth images, force sensor readings, and SDF ground truths generated using the ZeMa simulator [8]. ZeMa employs FEM for intersection-free, high-accuracy contact modeling and deformation.

5.1. Data Acquisition

An example of data collection in the simulation environment is shown in Fig. 2.b. To ensure stable grasp with multiple contacts, we use DiPGrasp [32], a fast grasp planner. After obtaining a feasible grasp pose, we position the hand near the

Metrics	IoU(%)↑	CD(mm)↓	MPJPE(mm)↓	PD(mm)↓	CIoU(%)↑
DexYCB					
gSDF [6] (RGB)	86.8	13.4	14.4	**8.9**	31.3
HOTrack [4]	88.2	10.2	25.7	12.3	28.5
Zhang *et al*. [40]	88.1	10.4	15.3	10.6	30.5
Ours (w/o Force Opt.)	**90.1**	**9.6**	**13.2**	9.9	**35.4**
HOT					
Zhang *et al*. [40]	77.9	16.8	15.6	11.9	26.8
ViTaM [15]	80.5	11.5	15.1	10.6	28.5
Ours (w/o Force Opt.)	**81.0**	**10.9**	13.6	10.7	29.8
Ours (w. CPF [33])	*	*	11.9	7.8	37.5
Ours (w. TOCH [42])	*	*	13.5	9.3	34.1
Ours (w. Force Opt.)	*	*	**11.3**	**7.3**	**40.3**

Table 1. Quantitative results on DexYCB and HOT datasets for previous SOTA and our method. We only compare our results on rigid objects with the baselines. * indicates that adding hand-object refinement methods doesn't impact the metrics on reconstructed objects. ↑ /↓ indicates higher scores/lower scores are better. Force Opt. is short for force-aware optimization.

target, add slight rotational perturbations, and set the hand fully open. It then moves toward the target and performs the grasp, generating a sequence of depth images, point clouds, and tactile arrays $\{\mathcal{P}_t, \mathcal{M}_t\}_{t=1}^{n}$ per frame.

5.2. Dataset Statistics

The **HOT** dataset includes objects from the YCB repository [1], with 5-10 objects from the **Bottle** and **Box** categories. We also add 5 **Sponges**, 5 **Plasticines**, and 5 **Stuffed Toys** to evaluate large deformations. Object properties are defined using density ρ, Young's modulus E, and Poisson's ratio ν. For **Bottle** and **Box**, $\rho = 10^3 kg/m^3$, $E = 1.271$GPa, $\nu = 0.28$; for **Sponges**, **Plasticines**, and **Stuffed Toys**, $\rho = 30 kg/m^3$, $E = 0.1$MPa, $\nu = 0.38$. **Plasticine** has a yield stress $s_y = 200$pa.

Each object has 20 trajectories (15 training, 2 testing, 3 validation), totaling 600 sequences of 50-100 frames. Each sequence includes point clouds from 8 camera views.

6. Experiments

6.1. Implementation Details

We use Farthest Point Sampling to downsample the input point cloud size to $N = 1024$ and sample 1×10^6 positions in space, including 2×10^5 surface points. During training, we subsample $M = 2048$ for SDF prediction. The volume scattering feature resolution $D = 64$. The distance threshold in contact constraint $l_c = 3$mm. Loss weights are $\lambda_f = 0.01$, $\lambda_S = 0.5$, $\lambda_H = 1$, and $\lambda_C = 0.05$.

The visual dynamic tracking net is trained on the entire dataset with a batch size of 6, learning rate $1e-4$, and Adam optimizer for 100 epochs. Fine-tuning per object category uses a batch size of 4, learning rate $5e-5$, and 50-100 epochs. Training takes 15 hours on an Nvidia A40 GPU.

For force-aware optimization, the barrier function threshold $\hat{l} = 2$mm and the ball-query radius $R = 5$mm. The energy is minimized for 100 iterations per frame with a learning rate of $2e-3$, taking 3.5 ± 0.5 seconds per frame.

6.2. Datasets

We use **DexYCB** for rigid object testing and the **HOT** dataset for deformable object interactions. DexYCB validates our method against baselines, while HOT evaluates deformation tracking and force-aware contact modeling.

6.3. Metrics

Intersection over Union (IoU) measures mesh overlap between predictions and ground truth.
Chamfer Distance (CD) evaluates object vertex reconstruction accuracy.
Mean Per Joint Position Error (MPJPE) assesses hand joint tracking precision.
Penetration Depth (PD) quantifies hand-object interaction plausibility by measuring maximum hand penetration into the object.
Hand-Object Contact Mask IoU (CIoU) validates contact recovery by comparing contact masks (defined as distances < 3mm) between optimized results and ground truth.

6.4. Results

For hand-object tracking evaluation, we compare our method with state-of-the-art approaches: gSDF [6], HOTrack [4], Zhang *et al*. [40], ViTaM [15], CPF [33], and TOCH [42]. We include gSDF which uses RGB images only for metrics references, and [40] uses depths for SDF-based object tracking. HOTrack predicts object poses from segmented point clouds, and ViTaM, similar to our method, takes unsegmented point clouds and tactile arrays as input but uses

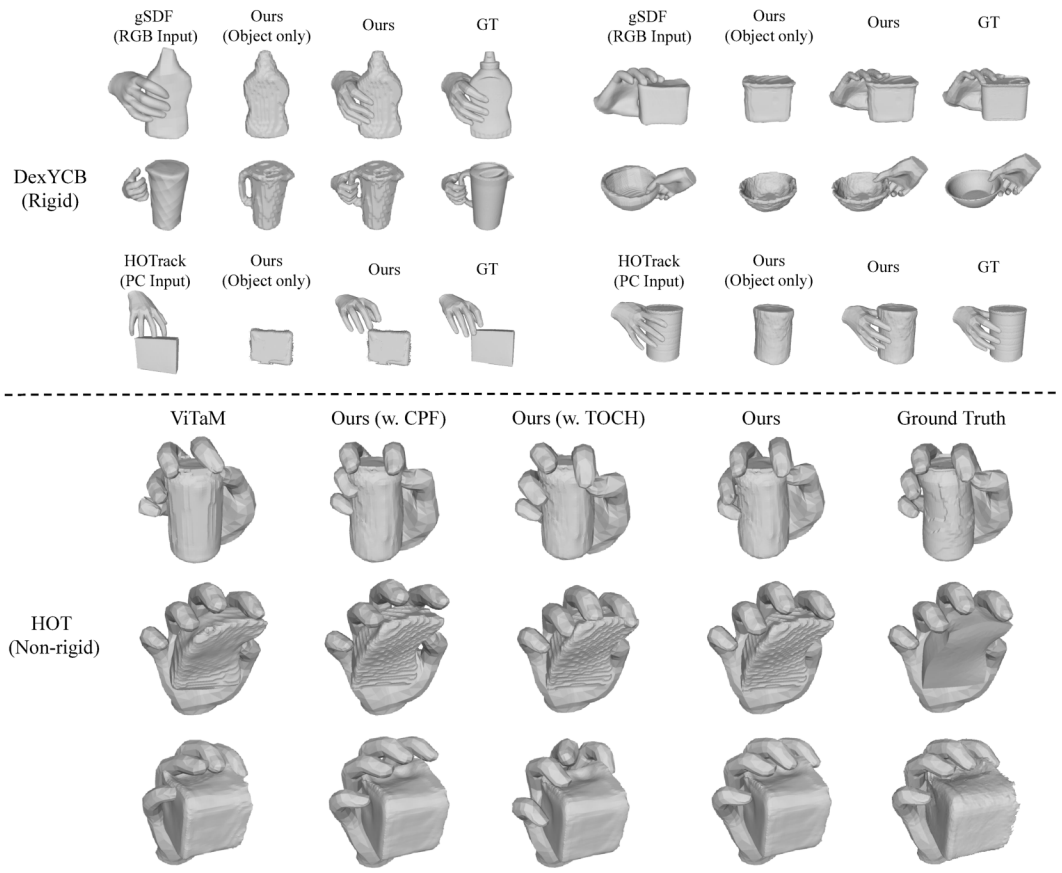

Figure 4. Qualitative results on both DexYCB and HOT datasets. The upper shows our better performances on rigid objects compared to the baselines, and the lower demonstrates the effectiveness of model for dynamic tracking of deformable objects and hands.

WNF representation, making it a strong baseline. We compare gSDF and HOTrack on DexYCB for rigid objects without tactile input, and ViTaM on the HOT dataset due to its tactile requirement. CPF and TOCH are used to benchmark our force-based optimization. We will discuss how to use force-aware optimization empirically with a fixed force setting in Sec. 6.5.

Quantitative results in Tab. 1 show our method outperforms baselines on DexYCB in IoU and Chamfer distance, demonstrating superior object tracking and reconstruction. Our higher MPJPE score indicates better hand-tracking performance. On the HOT dataset, our visual dynamic tracking net matches ViTaM's performance, while force-based optimization significantly improves MPJPE, PD, and CIoU, confirming its effectiveness. This is largely due to our method's ability to obtain accurate contact information through tactile arrays, unlike previous methods that rely on network-based contact modeling.

Figure 4 shows the qualitative results of our method. Our tracking and reconstruction outperform state-of-the-art methods in the DexYCB dataset, and our force-aware optimization process shows excellent performance on deformable objects in the HOT dataset in comparison with previous hand-object optimization baselines. Our method accurately tracks hand movements and object deformations, thanks to the fusion of flow features and original point cloud features using a transformer, which incorporates both the 3D information of the current frame and correspondence features from the previous frame. Figure 5 visualizes the force-aware hand-pose optimization. With force-aware contact representation DF-Field, the optimization can solve most of the penetration problems and refine the contact map based on collected tactile information.

6.5. Ablation Study

We conduct the ablation study on (1) methods for obtaining contact states for the contact constraint (Sec. 4.2) and (2) definitions of force representation (heuristic vs. tactile-aware). Additional discussions on point pair establishment, flow prediction, and tactile array integration in the visual dynamic tracking net are in the supplementary material.

Contact State & Effect of Contact Constraint. We evaluate the impact of the contact constraint by testing the visual dynamic tracking net without it. Additionally, we

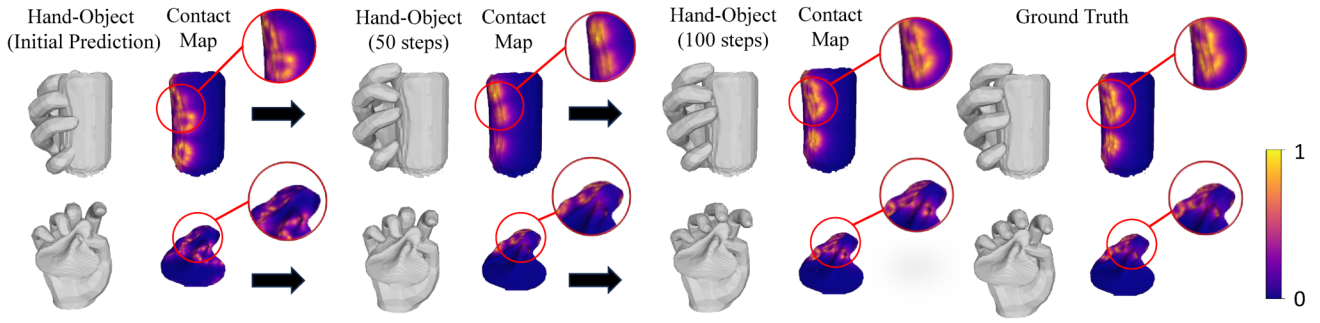

Figure 5. Qualitative results with force-aware optimization. As iteration steps rise, the penetration problem decreases, and the contact map is closer to the ground truth, making the contact more reasonable.

Metrics	IoU↑	CD↓	CIoU↑	MPJPE↓
w/o CC	75.9	12.8	25.3	12.0
CC. (GT)	**81.2**	11.2	29.2	**11.9**
CC. (Tactile)	81.0	**10.9**	**29.8**	**11.9**
CC. (PointNet)	78.3	12.1	27.6	12.1

Table 2. Quantitative results for different contact information acquirement. CC. indicates Contact Constraint, along with approaches for acquiring contact information.

Metrics	MPJPE↓	PD↓	CIoU↑
DexYCB			
w/o Force Opt.	13.2	9.9	35.4
$\mathcal{M}$ fixed Force Opt.	**12.3**	**8.5**	**39.7**
HOT			
w/o Force Opt.	13.6	10.7	29.8
$\mathcal{M}$ fixed Force Opt.	12.9	8.5	36.8
Force Opt.	**11.3**	**7.3**	**40.3**

Table 3. Quantitative results for different force representations in force-aware hand-pose optimization.

compare ground truth contact states with predictions from a PointNet-based network when there is a lack of tactile hardware support. Results in Tab. 2 show that the contact constraint significantly improves object reconstruction and hand-object interaction prediction. The PointNet-based network effectively predicts interaction states from visual inputs, while tactile readings match ground truth accuracy, greatly enhancing object geometry reconstruction.

Different Force Representations. We explore the effect of force representations on contact modeling. For DexYCB (no tactile data), we set $\mathcal{M}_j = 0.5$ for regions near objects ($l = 3mm$) and compare results with the visual dynamic tracking net. For HOT, we test the optimization with $\mathcal{M}_j = 0.5$ to evaluate the impact of removing tactile information.

Results in Tab. 3 demonstrate that even with fixed $\mathcal{M}$, DexYCB performance improves, validating the effective design of DF-Field. On HOT, tactile information significantly enhances penetration resolution and contact recovery, as distributed tactile arrays provide accurate interaction forces, outperforming empirical force settings.

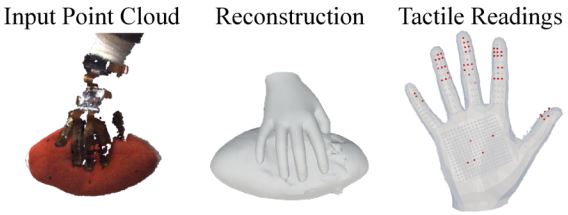

Figure 6. Real-world reconstruction example of an unseen stuffed toy. The overall shape and deformation of the toy have been well reconstructed.

6.6. Real-world Experiments

In order to validate the effectiveness of our method, we conduct real-world experiments with our network trained on simulation data. We bought a ViTaM glove [15] from its authors to capture distributed tactile sensory data and a Photoneo MotionCam M+ high-precision depth camera to record the point cloud series. We can see an example of a reconstruction result on an unseen stuffed toy in Fig. 6. We will demonstrate our real-world setup and more series of tracking results in supplementary material.

7. Conclusion and Future Works

In this work, we introduce ViTaM-D, a dynamic hand-object interaction reconstruction framework that integrates distributed tactile sensing with visual perception. Featuring DF-Field and force-aware optimization, our approach effectively captures fine contact details and object deformation. We also present the HOT Dataset for benchmarking deformable object manipulation. Evaluations show that ViTaM-D outperforms state-of-the-art methods on both rigid and deformable objects. Future works include applying ViTaM-D to data collection for robotic imitation learning, dexterous manipulation, and human-robot interaction.

Acknowledgement

This work was supported by the Shanghai Commitee of Science and Technology, China (Grant No.24511103200) by the National Key Research and Development Project of China (No.2022ZD0160102), Shanghai Artificial Intelligence Laboratory, XPLORER PRIZE grants. This work has also been supported by the Shanghai Municipal Education Commission (No. 2024AIYB010) and the Fundamental Research Funds for the Central Universities (YG2025LC03).

References

[1] 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. 6

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

[3] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9044–9053, 2021. 2, 5

[4] Jiayi Chen, Mi Yan, Jiazhao Zhang, Yinzhen Xu, Xiaolong Li, Yijia Weng, Li Yi, Shuran Song, and He Wang. Tracking and reconstructing hand object interactions from point cloud sequences in the wild. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 304–312, 2023. 1, 2, 6

[5] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *European Conference on Computer Vision*, pages 231–248. Springer, 2022. 2

[6] Zerui Chen, Shizhe Chen, Cordelia Schmid, and Ivan Laptev. gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12890–12900, 2023. 2, 6

[7] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J Crandall. Hope-net: A graph-based model for hand-object pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6608–6617, 2020. 2

[8] Wenxin Du, Siqiong Yao, Xinlei Wang, Yuhang Xu, Wenqiang Xu, and Cewu Lu. Intersection-free robot manipulation with soft-rigid coupled incremental potential contact. *IEEE Robotics and Automation Letters*, 2024. 2, 5

[9] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Arctic: A dataset for dexterous bimanual hand-object manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12943–12954, 2023. 2

[10] Zicong Fan, Maria Parelli, Maria Eleni Kadoglou, Xu Chen, Muhammed Kocabas, Michael J Black, and Otmar Hilliges. Hold: Category-agnostic 3d reconstruction of interacting hands and objects from video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 494–504, 2024. 2

[11] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 2

[12] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11807–11816, 2019. 1, 2

[13] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 571–580, 2020. 2

[14] Yisheng He, Wei Sun, Haibin Huang, Jianran Liu, Haoqiang Fan, and Jian Sun. Pvn3d: A deep point-wise 3d keypoints voting network for 6dof pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11632–11641, 2020. 4

[15] Chunpeng Jiang, Wenqiang Xu, Yutong Li, Zhenjun Yu, Longchun Wang, Xiaotong Hu, Zhengyi Xie, Qingkun Liu, Bin Yang, Xiaolin Wang, et al. Capturing forceful interaction with deformable objects using a deep learning-powered stretchable tactile array. *Nature Communications*, 15(1): 9513, 2024. 1, 2, 3, 6, 8

[16] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020. 1

[17] Minchen Li, Zachary Ferguson, Teseo Schneider, Timothy Langlois, Denis Zorin, Daniele Panozzo, Chenfanfu Jiang, and Danny M. Kaufman. Incremental potential contact: Intersection- and inversion-free large deformation dynamics. *ACM Trans. Graph. (SIGGRAPH)*, 39(4), 2020. 3

[18] Hangxin Liu, Xu Xie, Matt Millar, Mark Edmonds, Feng Gao, Yixin Zhu, Veronica J Santos, Brandon Rothrock, and Song-Chun Zhu. A glove-based system for studying hand-object manipulation via joint pose and force sensing. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6617–6624. IEEE, 2017. 1

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

[20] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 2

[21] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy

networks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 523–540. Springer, 2020. 4

[22] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Advances in neural information processing systems*, 30, 2017. 4

[23] Jieji Ren, Jiang Zou, and Guoying Gu. Mc-Tac: Modular camera-based tactile sensor for robot gripper. In *The 16th International Conference on Intelligent Robotics and Applications (ICIRA)*, 2023. 1

[24] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 2017. 4

[25] Edward Smith, Roberto Calandra, Adriana Romero, Georgia Gkioxari, David Meger, Jitendra Malik, and Michal Drozdzal. 3d shape reconstruction from vision and touch. *Advances in Neural Information Processing Systems*, 33:14193–14206, 2020. 1, 3

[26] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019. 1, 3

[27] Ian H Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim 3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10781–10787. IEEE, 2022. 1

[28] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4511–4520, 2019. 2

[29] Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 484–500, 2018. 1, 2

[30] Shaoxiong Wang, Jiajun Wu, Xingyuan Sun, Wenzhen Yuan, William T Freeman, Joshua B Tenenbaum, and Edward H Adelson. 3d shape perception from monocular vision, touch, and shape priors. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1606–1613. IEEE, 2018. 1, 3

[31] Wenqiang Xu, Zhenjun Yu, Han Xue, Ruolin Ye, Siqiong Yao, and Cewu Lu. Visual-tactile sensing for in-hand object reconstruction. In *CVPR IEEE Conference on Computer Vision and Pattern Recognition*, pages 8803–8812, 2023. 1, 3

[32] Wenqiang Xu, Jieyi Zhang, Tutian Tang, Zhenjun Yu, Yutong Li, and Cewu Lu. Dipgrasp: Parallel local searching for efficient differentiable grasp planning. *IEEE Robotics and Automation Letters*, 2024. 5

[33] 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 IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021. 1, 2, 6

[34] Lixin Yang, Kailin Li, Xinyu Zhan, Jun Lv, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Artiboost: Boosting articulated 3d hand-object pose estimation via online exploration and synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2750–2760, 2022. 2

[35] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20953–20962, 2022. 2, 5

[36] Lixin Yang, Xinyu Zhan, Kailin Li, Wenqiang Xu, Jiefeng Li, and Cewu Lu. Learning a contact potential field for modeling the hand-object interaction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1, 2

[37] Yufei Ye, Poorvi Hebbar, Abhinav Gupta, and Shubham Tulsiani. Diffusion-guided reconstruction of everyday hand-object interaction clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19717–19728, 2023. 2

[38] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023. 1

[39] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017. 1

[40] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *ACM Transactions on Graphics (TOG)*, 40(3):1–12, 2021. 6

[41] Qiang Zhang, Yunzhu Li, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B Tenenbaum, Wojciech Matusik, and Antonio Torralba. Dynamic modeling of hand-object interactions via tactile sensing. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2874–2881. IEEE, 2021. 3

[42] Keyang Zhou, Bharat Lal Bhatnagar, Jan Eric Lenssen, and Gerard Pons-Moll. Toch: Spatio-temporal object-to-hand correspondence for motion refinement. In *European Conference on Computer Vision*, pages 1–19. Springer, 2022. 6

[43] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5346–5355, 2020. 4