SHOWMe: Benchmarking Object-agnostic Hand-Object 3D Reconstruction

Anilkumar Swamy[1,2] Vincent Leroy[1] Philippe Weinzaepfel[1] Fabien Baradel[1] Salma Galaaoui[1]
Romain Brégier[1] Matthieu Armando[1] Jean-Sebastien Franco[2] Grégory Rogez[1]

[1]NAVER LABS Europe [2]Inria centre at the University Grenoble Alpes

Figure 1: **The SHOWMe dataset** comprises 96 videos with their associated high-quality textured meshes of a hand holding an object. For two different samples, we show on the left side, row by row, real RGB crops from the dataset, an overlay of the corresponding ground truth textured mesh, and a rendering of the texture-less mesh with Phong shading. On the right, we show the 3D reconstruction of the hand-object obtained from the RGB stream only, using one of the evaluated baselines.

Abstract

Recent hand-object interaction datasets show limited real object variability and rely on fitting the MANO parametric model to obtain groundtruth hand shapes. To go beyond these limitations and spur further research, we introduce the SHOWMe dataset which consists of 96 videos, annotated with real and detailed hand-object 3D textured meshes. Following recent work, we consider a rigid hand-object scenario, in which the pose of the hand with respect to the object remains constant during the whole video sequence. This assumption allows us to register sub-millimeter-precise groundtruth 3D scans to the image sequences in SHOWMe. Although simpler, this hypothesis makes sense in terms of applications where the required accuracy and level of detail is important e.g.*, object hand-over in human-robot collaboration, object scanning, or manipulation and contact point analysis. Importantly, the rigidity of the hand-object systems allows to tackle video-based 3D reconstruction of unknown hand-held objects using a 2-stage pipeline consisting of a rigid registration step followed by a multi-view reconstruction (MVR) part. We carefully evaluate a set of non-trivial baselines for these two stages and show that it is possible to achieve promising object-agnostic 3D hand-object reconstructions employing an SfM toolbox or a hand pose estimator to recover the rigid transforms, and off-the-shelf MVR algorithms. However, these methods remain sensitive to the initial camera pose estimates which might be imprecise due to lack of textures on the objects or heavy occlusions of the hands, leaving room for improvements in the reconstruction. Code and dataset are available at https://europe.naverlabs.com/research/showme/.*

1. Introduction

Understanding interactions between hands and objects from RGB images is a key component towards better understanding human actions and interactions. Such understanding could benefit many applications, from virtual and augmented reality to human-robot interaction and autonomous robotic manipulation via learning by demonstration. For instance, in a scenario where a human is handing over an object to a robot equipped with RGB sensors, we expect the

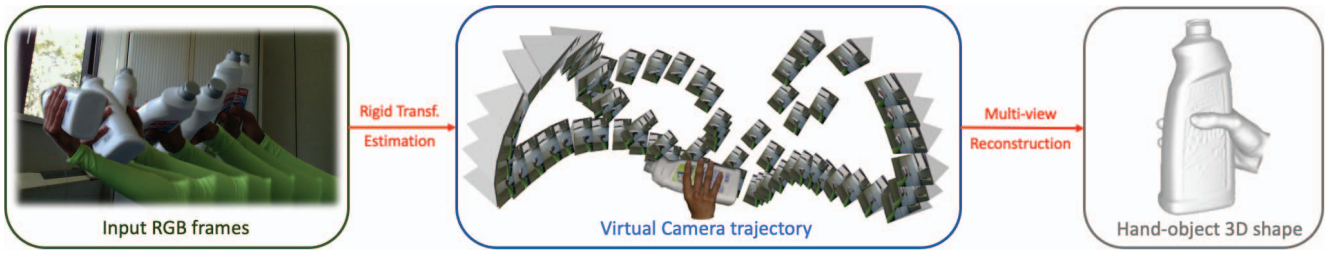

Figure 2: **Hand-Object 2-stage 3D reconstruction pipeline.** Given an RGB video of a hand holding an object (left), the rigid transformation between frames is first estimated. This allows to see the problem as if a set of multiple virtual cameras observe a fixed hand-object system (middle). Multi-view reconstruction can then be employed to estimate an accurate hand-object 3D shape (right). We benchmark several baselines for both stages using the presented dataset.

robot to grasp the object without hurting the person in any way. Such action is likely to require a fine-grained perception of both the object and the hand holding it, and being able to accurately model the hand-object (HO) system in 3D from RGB data would be very useful in such context.

This problem of joint HO 3D reconstruction has been addressed in a large body of recent works [26, 24, 25, 6, 21, 11, 13, 54, 13, 48] that estimate HO 3D shape from single RGB images. These methods often rely on a deformable kinematic model of the human hand, MANO [42], which contains useful priors, but also limits the potential reconstruction accuracy [15] for unseen hand shapes. A second important limitation of most HO reconstruction approaches is that the exact 3D model of the object is often assumed to be known apriori, and they tend to struggle to generalize to objects that fall outside of the training distribution. While single-image HO reconstruction without priors over the objects remains very challenging, exploiting multiple observations of the scene can significantly simplify the task.

One way to obtain more observations is to consider a synchronized multi-camera setup, increasing the complexity of deployment in practice. Another way is to focus on the temporal aspect of the RGB streams as in [27] who recently showed that multiple observations of the scene can be exploited to simplify object-agnostic hand-object 3D reconstruction. However, their method remains limited to close-up fingertip grasps of small objects and cannot be used for natural hand-object interactions. Interestingly, seldom previous work focused on aggregating temporal information of a RGB video for HO reconstruction [27], unless the strong assumption of a known object was made [25, 24].

Following [6, 27], we simplify the problem as an intermediary step towards dynamic temporal integration by assuming that the camera is static and the hand is holding an unknown object rigidly. In this setup, an RGB video can be viewed as multiple observations of the same HO system, which allows to formulate the HO modeling problem in a Multi-View Reconstruction (MVR) setting: the RGB appearance of a HO instance that undergoes a rigid transformations is observed. In order to solve this problem, two unknowns have to be addressed: 1. the rigid transformation and 2. how to aggregate RGB observations. It is worth noting that these points can be addressed either separately or jointly. With the exception of [27] who operates in a rather constrained scenario, no method was specifically designed to solve the challenges raised by this task but, more importantly, there is a need for an evaluation protocol and a specifically designed dataset.

Therefore, we propose a novel dataset consisting of 96 videos of a hand holding an object rigidly and showing this object to the camera. We captured a total of 87K frames depicting 42 unique objects with evenly distributed grasp configurations, handled by 15 subjects reflecting a diversity of gender, color, and hand shape. Importantly, our dataset contains high-precision ground-truth (GT) HO 3D shapes, that we captured using a sub-millimeter precision scanner before capturing each video sequence. The resulting textured 3D meshes are then registered to each frame of the corresponding sequences, in order to provide highly detailed ground truth annotations. In practice, we proceed in two steps: 1) we register the GT HO mesh to the depth map of each frame in the sequence. 2) We refine the registration using a differentiable rendering pipeline to obtain very accurate alignments of the 3D mesh with the RGB frames as shown in Fig. 1. We call our dataset SHOWMe, standing for Single-camera Hand-Object videos With accurate textured 3D Meshes.

Using SHOWMe, we benchmark the 2-stage pipeline consisting of a rigid registration followed by a HO 3D reconstruction from multiple observations, see Fig. 2. In the same spirit as [34] with body shapes, we first estimate the rigid transformations between frames using the output of a hand keypoints detector as in [27]. We compare this approach to a standard structure-from-motion (SfM) approach, namely COLMAP [44]. We find that hand-based estimation of the rigid transformation is more robust for textureless objects but suffers in case of heavy occlusions. Given the rigid registration, the HO reconstruction can be performed using multi-view reconstruction methods. We compare a silhouette based reconstruction method, leveraging hand-object segmentation [32] to more recent approaches based on differentiable rendering method [47] and

dataset	real images	marker-less	# number of img	seq	sbj	obj	image resol.	grasp variability	object scan	hand-obj texture	hand scan	hand annotation
ObMan[26]	×	✓	154k	-	20	3K	256 × 256	+++	✓	×	×	MANO
GRAB[46]	×	×	-	1,335	10	51	-	+++	✓	×	×	MANO
FPHA[20]	✓	×	105k	1,175	6	4	1920×1080	+	✓	×	×	keypoints
ContactPose[6]	✓	×	2,991k	2,303	50	25	960×540	++	✓	×	×	MANO
ARCTIC[17]	✓	×	1,200k	242	9	10	2800 × 2000	+++	✓	×	×	MANO
YCB-Affordance[16]	✓	✓	133k	-	1	21	640 × 480	+++	✓	×	×	MANO
GUN-71[41]	✓	✓	12k	1,680	8	1988	640×480	+++	×	×	×	grasp Id
FreiHand[58]	✓	✓	37k	-	32	27	224×224	++	×	×	×	MANO
Dexter+Object[45]	✓	✓	3k	6	2	2	640×480	+	×	×	×	fingertips
EgoDexter[35]	✓	✓	3k	4	4	-	640×480	+	×	×	×	fingertips
HO3D[21]	✓	✓	78k	27	10	10	640 × 480	+++	✓	×	×	MANO
DexYCB[12]	✓	✓	582k	1,000	10	20	640 × 480	++	✓	×	×	MANO
H2O[31]	✓	✓	571k	-	4	8	1280 × 720	++	✓	×	×	MANO
OakInk[53]	✓	✓	230k	-	12	100	848 × 480	+++	×	×	×	MANO
HOD[27]	✓	✓	126k	70	1	35	2160 × 3840	+	✓(only 14)	×	×	NO Annotations
SHOWMe (Ours)	✓	✓	**87k**	**96**	**15**	**42**	**1280×720**	+++	✓	✓	✓	MANO

Table 1: **Comparison of our dataset with existing hand-object interaction datasets**

neural implicit surfaces [27]. All three obtain extremely accurate results given ground-truth registration. Yet, when considering estimated registrations, results of the best baseline are satisfactory on approximately three quarters of the sequences, and fail on the others. This confirms that HO 3D reconstruction from an RGB video is a difficult task, and we hope our dataset will foster further research on this topic.

In summary, our contribution is twofold. First, we propose a novel hand-object interaction dataset, SHOWMe, and the pipeline we designed to annotate RGB-D videos using high-precision hand-object 3D scans. SHOWMe is the first dataset providing such level of accuracy for the ground-truth hand-object 3D shapes. Second, we evaluate a set of baselines for the MVR-based pipeline for detailed and object-agnostic HO 3D reconstruction in RGB videos.

After discussing related work and existing datasets in Sec. 2, we introduce the SHOWMe dataset and its capturing setup in Sec. 3. We finally present the 2-stage pipeline in Sec. 4 before evaluating several baselines in Sec. 5.

2. Related Work

Our two contributions being a new HO dataset and a benchmark of object-agnostic HO reconstruction baselines, we discuss below the most relevant datasets and methods.

Hand-Object Datasets. Earlier research on hand-object interaction [41, 4, 8, 9, 18] have proposed datasets for grasp classification or action recognition. Despite the importance of the recognition tasks, these datasets were seldom considered for HO reconstruction research due to the unavailability of GT 3D annotations, such as 3D joints or 3D shapes.

Obtaining images with ground-truth 3D information is a tedious problem in general, even for non-hand-related research. The small size of the hands in images make them difficult to annotate manually [45]. The problem is exacerbated when considering a hand interacting with objects. Past work has therefore proposed to consider synthetic data [40, 35, 14, 26, 16], motion capture with markers [6], magnetic sensors [20] or multi-view set-ups [58, 21, 6, 12, 31, 53]. Synthetic data is usually obtained by rendering a parametric model of the hand interacting with objects. Even if realism is sufficient when considering a depth sensor [40], the domain gap between synthetic and real RGB images is often too large to be a valid option on its own. On the other hand, invasive motion capture methods based on magnetic sensors and markers make the hand appearance unrealistic and introduces an undesired bias.

Most of the recent datasets obtained through multi-view set-ups [58, 21, 6, 12, 31, 53] use the multi-view data to fit the MANO parametric model [42] that is then considered as GT hand shape. Although it contains useful priors, MANO cannot represent very detailed hand shapes [15]. In our case, we scan the hand using a high-precision scanner, obtaining a GT shape with sub-millimeter accuracy.

Recent multiview video datasets such as [21, 12, 22] are impressive in terms of scale, markerless nature, and realism in motion but they lack object variability (10 objects for [21] and 20 for [12], both object sets from the YCB dataset [10]). Motions are also limited to the same patterns like lifting the objects from the table and placing it back or handing them over to another person. OakInk [53] provides a much larger variety of objects but with limited motions. The SHOWMe dataset contains more than 40 objects with complex movements showing all sides of the object.

Closer to the proposed SHOWMe dataset are ContactPose [6] and HOD [27] which also consider a static HO configuration during the manipulation. While HOD provides unregistered 3D scans for a subset of the manipulated objects, ContactPose provides groundtruth 3D shapes and poses for both the hand and the object. This dataset is however limited to objects artifically made textureless, that are equipped with intrusive fiducial markers for motion capture purposes. The hand shape is also obtained after fitting the MANO model. Besides, we found that some frames

Figure 3: **Rendering of the textured mesh for few hand-object configurations of our SHOWMe dataset**.

are missing in some sequences leading to discontinuities in HO motion during manipulation and preventing the use of a video-based approach. Our SHOWMe dataset offers more variety in terms of object appearance and grasp types (see Fig. 3) and, importantly, it is the first dataset that provides real ground-truth 3D shape for both the hand and the object. We provide a comparison of SHOWMe to the most relevant and widely used hand-object interaction datasets in Table 1.

Hand-Object Reconstruction from a single RGB image or from a monocular video is an extremely difficult task due to hand-object mutual occlusions, complex hand-object motion and variability in object shapes. That is why earlier work [50, 49, 55, 3, 37] considered RGB-D or multi-view inputs . Recent works on joint HO reconstruction from monocular RGB images have achieved impressive results. These works can be generally categorized into parametric hand model-based methods [26, 42, 11, 33, 30, 39] that assume a known object template (or category [30, 36, 23]) and implicit representation-based methods [29], or a combination of both [54, 13]. While [54] assumes known 3D templates and obtain both hand and object poses from parametric models - using Signed Distance Functions (SDFs) to help reconstruct shape details for both hand and object, [13] only uses a parametric model for the hand prior and reconstruct generic hand-held object without knowing their 3D templates. However, the object reconstruction performance is rather poor as it remains unclear how to learn the implicit representations to reconstruct a large variety of object shapes with a single model as observed in [29]. To achieve reasonable HO results in a fully object-agnostic manner, [27] leverages multiple observations of a HO rigid configuration along a video sequence. The camera motion is recovered using a hand tracker and an implicit neural representation-based method is then employed to reconstruct the SDF and color fields of the hand and object. Similarly to this method, we consider a 2-stage pipeline consisting of a rigid registration followed by MVR and benchmark several baselines for each of these 2 stages.

Other methods have considered hand-object monocular RGB video as input. [24] performs joint HO reconstruction by leveraging photometric consistency over time while in [25], an optimization approach is used. [33] leverages spatial-temporal consistency to select pseudo-labels for self-training. These methods have the biggest caveat of requiring the object template mesh at inference time, which makes the hand-object reconstruction problem a HO 6DOF pose estimation task. We focus on bench-marking object-agnostic methods that can reconstruct any HO shapes.

3. The SHOWMe dataset

In this section, we detail the collection procedure in Sec. 3.1 and the data annotation in Sec. 3.2 (see Fig. 4 for an overview) while Sec. 3.3 details how GT scans are further annotated with hand-object information.

3.1. Dataset collection

We instruct the subject to grasp one object according to different use cases: either a *power-grasp*, *i.e.*, holding the object strongly with all fingers, a *use-grasp*, *i.e.*, holding the object as if the object was going to be used or a *handover-grasp*, *i.e.*, holding the object as if the intent was to give it to someone else. We then record a video with an RGB-D monocular camera of the subject showing every part of the hand-object grasp. In order to ease hand-object segmentation from the arm, which is not the focus of our dataset, the subject is wearing a distinctive sleeve and no other human parts are visible in the video. Once the video is captured, we ask the subject to maintain the same grasp and capture the shape of the HO configuration using a sub-millimeter precision scanner. Fig. 3 shows several captured textured meshes, highlighting the diversity of objects and grasps.

Hardware details. We acquire the videos using a single Intel RealSense L515 RGB-D camera [28], and we capture the GT HO shapes with a Artec Eva 3D scanner [2]. The camera is calibrated in a pre-processing step and is used to capture both depth and RGB streams at a rate up to 30fps and 1280 x 720 resolution. We process the RGB and depth streams to perform pixel alignment and temporal synchronization. We use the software provided by the supplier for obtaining an accurate shape from the scans.

Dataset statistics. We collect 96 sequences from 15 differ-

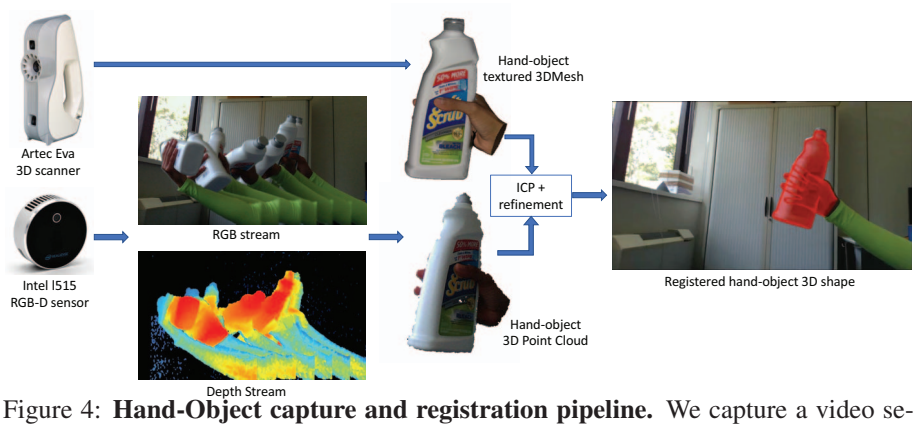

Figure 4: **Hand-Object capture and registration pipeline.** We capture a video sequence of a hand holding rigidly an object and moving in front of an RGB-D camera, and we automatically segment the hand-object system in the images. We reconstruct a precise textured mesh of the hand-object in the exact same pose using an off-the-shelf 3D scanner and register this mesh to each frame to provide ground-truth annotations.

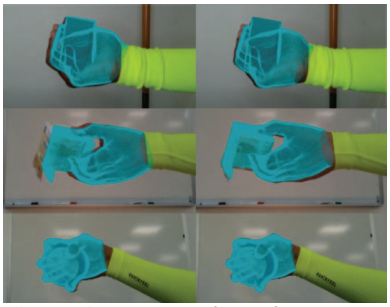

Figure 5: **Mesh registration procedure for data annotation.** Left: the pose of the ground truth HO mesh (light blue) is initialized through ICP registration with the segmented depth map. Right: it is then refined with differentiable rendering and temporal smoothness priors.

ent subjects holding 42 different objects from everyday life, with various sizes and shapes. The subjects reflect diversity in gender, color, and hand shape. The different grasp types (power-grasp, use-grasp, handover-grasp) are evenly represented. Each video sequence lasts an average of 48 seconds. This represents a total of 87,540 frames.

3.2. Ground-truth HO 3D shape annotation

We now detail how we obtain HO segmentation in the RGB images and GT rigid transformation, *i.e*. the alignment between each frame and the 3D mesh obtained from the scanner, allowing its reprojection onto the image.

Segmentation. We first segment the foreground, *e.g*. HO pixels by thresholding the depth values from the input RGB-D stream. This process segments out the wrist and the object, but also the arm which we want to ignore, since it is out of the scope of this work, and it violates the rigidity assumption. We then segment the arm part by thresholding RGB pixels values based on the color of the sleeve. Finally, we combine these two masks to obtain the HO masks which can be applied on both the RGB frames as well as the depth values, that we express as back-projected 3D point clouds.

Rigid transformation from scanned mesh to each frame. For each video, we align all the frames to the scanned GT mesh. The first step of this alignment consists in performing a robust rigid Iterative Closest Point (ICP) [56] between the GT mesh and the aforementioned masked depth point clouds. We manually 3D align to initialize the first frame of each sequence and then automatically align the remaining frames using the previous result as initialization for the next one, to obtain initially aligned poses $\{R_i|t_i\} \in SE(3)$, denoting rotations and translations respectively. We found that such an alignment is already quite satisfactory but some outliers remain, due to sensor noise or invalid local minima of the ICP. We thus refine these aligned poses via a differentiable rendering pipeline that we detail in the following.

For each sequence, let $I_i, i \in \{1..N\}$ denote the N input frames of resolution $H \times W$, S_i be the ground-truth segmentations at the same resolution and $\mathcal{M}$ the GT mesh. This mesh is associated with appearance information acquired from the sensor such that we can render it onto the image planes in a differentiable manner. Our objective is to refine the camera poses $\{R'_i|t'_i\} = \{R_i \text{orth}(R_i^{corr})|t_i + t_i^{corr}\}$ such that the projection of the colored mesh $\mathcal{P}(\mathcal{M}, \{R'_i|t'_i\})$ matches the RGB observations for each frame. We express the pose corrections as offsets over the ICP results. And we parametrize the rotation corrections R_i^{corr} as 2×3 matrices, that we orthonormalize with the Gram-Schmidt process orth() to be rotation matrices, following [7].

More formally, we minimize a masked Mean Square Error (MSE) between rendered image $\hat{I}_i$ and observations:

$$\mathcal{L}_{RGB} = \sum_{i}^{N} \sum_{p}^{H \times W} S_i(p).||\hat{I}_i(p) - I_i(p)||^2. \quad (1)$$

This loss alone does not properly converge for sequences where the RGB information is ambiguous. Thus, we add two regularization terms following two assumptions. We assume the consecutive rotations and translations to be smooth, thus we add a smoothing term $\mathcal{L}_{Smooth} = \mathcal{L}_t + \mathcal{L}_R$ as a combination of two functions that minimize the discrete Laplace operator of transformations, one for rotations $\mathcal{L}_R$ in degrees and one for translations $\mathcal{L}_t$ in centimeters:

$$\mathcal{L}_t = \sum_{i=1}^{N-1} \frac{||2t'_i - \text{sg}(t'_{i-1} + t'_{i+1})||}{2N}, \quad (2)$$

$$\mathcal{L}_R = \sum_{i=1}^{N-1} \frac{\angle\big(\text{sg}(R'_{i-1}), R'_i\big) + \angle\big(R'_i, \text{sg}(R'_{i+1})\big)}{2N}, \quad (3)$$

where sg is the stop-gradient operator and $\angle$ returns the angle between two rotations. sg is needed to prevent collapsing to unique R and T values in our auto-differentiating framework.

These smoothing terms forbid camera transformations that violate the motion smoothness assumption. To incentivize the pose corrections to be small, we add a weight decay regularization term formulated as follows:

$$\mathcal{L}_{wd} = \sum_i \|R_i^{corr} - I\|^2 + \|T_i^{corr}\|^2 \quad (4)$$

where I denotes the identity rotation. Finally, the final loss we optimize is expressed as:

$$\mathcal{L} = \mathcal{L}_{RGB} + \lambda_{Smooth}\mathcal{L}_{Smooth} + \lambda_{wd}\mathcal{L}_{wd} \quad (5)$$

We did not include a loss for the depth information as it would have been computationally demanding. We considered that the ICP-alignment already provided signal from the depth, that is included in the current formulation in $\mathcal{L}_{wd}$. We model the GT geometry in the form of a sparse voxel grid structure in the differentiable rendering framework of [19], each voxel close to the GT mesh having a high opacity. Each non-zero voxel is equipped with appearance information initialized from $\mathcal{M}$. As the appearance of $\mathcal{M}$ was obtained using a scanner, it does not correspond exactly to the RGB observations, so we need to compensate for the appearance to account for sensor-dependent information. We thus optimize for both the camera poses offsets and the appearance of the GT mesh. Please refer to the supplementary material for optimization details.

The effects of this camera refinement procedure are shown in Fig. 5. Thin structures can hardly be correctly aligned via ICP as only very few pixels provide depth information on those regions. In contrast, the RGB based refinement along with the smoothing components help annotate more accurate poses. After manual verification, we managed to improve the annotated poses for 47 out of the 96 sequences both quantitatively in terms of $\mathcal{L}_{RGB}$ and qualitatively. The remaining 49 sequences were already very accurate and the optimization did not help in this case.

3.3. Parametric Model Annotations

For each sequence, we also provide semantic information regarding the depicted grasp. For that purpose, we captured textured 3D scans of the objects alone that we register together with the MANO hand model [42] to the HO meshes. This provides pose and shape annotations regarding both the hand and the object independently, as shown in Fig. 6. This additional information could prove useful for other tasks such as detailed grasp analysis, HRI-related tasks or even hand-object pose estimation although out of the scope of this paper. Importantly, the GT MANO kinematic poses will allow us to benchmark hand pose estimation methods employed to estimate the rigid transformation.

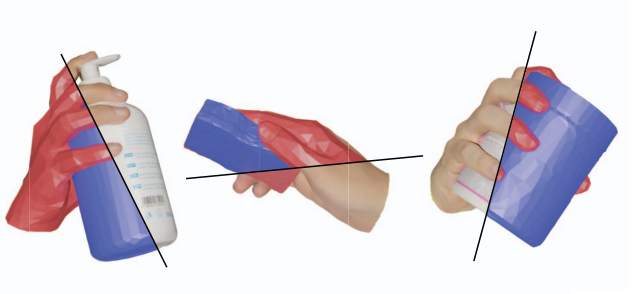

Figure 6: **Hand and Object 3D model annotation.** Partial overlay of the MANO hand model (in red) and a decimated object mesh (in blue) registered to the textured hand-object 3D scan for different sequences of SHOWMe.

Our registration process is semi-automatic and consists of three steps. First, we manually estimate the pose of the object by roughly aligning its mesh to the HO mesh. Second, we estimate MANO hand pose and shape parameters that minimize the squared distance error between 3D keypoints we manually annotated on the HO mesh and corresponding MANO vertices. We use L-BFGS optimization and the differentiable MANO layer of [26]. Third, we refine MANO parameters and object pose to obtain a precise registration, by minimizing:

$$\frac{1}{|\mathcal{HO}|}\sum_{x\in\mathcal{HO}} \min\left(d(x,\mathcal{O}), d(x,\mathcal{H})\right), \quad (6)$$

consisting in the mean *distance* of each point x on the mesh $\mathcal{HO}$ to the closest point on the hand mesh or the object mesh (denoted respectively $\mathcal{H}$ and $\mathcal{O}$). We define the *distance* between a point x of 3D normal n_x and a mesh $\mathcal{M}$ as:

$$d(x,\mathcal{M}) \triangleq \|x-p\|^2 + \lambda\|n_x - n_p\|^2, \quad (7)$$

where p is the point on $\mathcal{M}$ closest to x, and where n_p denotes its 3D normal. We choose $\lambda = 1mm^2$ in practice, and sample uniformly $|\mathcal{HO}| = 30k$ points on the HO mesh to evaluate Eq. (6). We obtain a sub-millimetre residual error after optimization. We provide qualitative visualizations of this registration in Fig. 6 and in the supplementary video.

4. Two-stage reconstruction pipeline

To reconstruct the HO from an RGB video, We use a 2-stage pipeline in Fig. 2: estimating the rigid transformations of the HO in the sequence (Sec. 4.1) and MVR (Sec. 4.2).

4.1. Rigid transformation estimation

We evaluate two methods for estimating the rigid transformation of the HO between frames, either using standard generic SfM toolbox, or using the hand pose as a proxy.

Rigid transformation from a SfM toolbox. We run COLMAP [44] – SfM software recognized for its robustness and efficiency – to estimate the pose of the camera

	Method	Hand pose			Rigid transformation		
		MPJPE ↓	PA-MPJPE ↓	PCK ↑	Rot error ↓	Trans error ↓	Det. rate (%) ↑
image-based	Minimal Hand [57]	85.4	38.1	10.9	-	-	-
	Frankmocap [43]	39.3	14.9	38.3	-	-	-
	HandOccNet [38]	37.4	14.7	45.7	-	-	-
	DOPE [52]	**26.9**	**12.4**	**64.6**	21.0	0.17	99.0
video-based	DOPE [52] + fixed hand pose	**26.2**	12.4	**69.4**	21.5	0.16	99.0
	DOPE [52] + median filtering	**26.2**	12.4	**69.4**	21.3	0.15	**100**
	DOPE [52] + PoseBERT [5]	27.3	**12.3**	58.4	20.6	0.15	**100**
	COLMAP [44]	-	-	-	**14.6**	**0.06**	78.2

Table 2: **MANO Evaluations: Hand pose estimation and associated rigid transformation estimation.** The MPJPE and PA-MPJPE are reported in mm. We use a threshold of 30mm for the PCK. The 'Rot. error' is the geodesic distance expressed in degree with the ground-truth rigid transformation. The 'Trans error' is the MSE.

with respect to the HO system across video frames. We ignore background keypoints using the silhouettes information. **Rigid transformation from hand pose estimation.** In our particular setup, we can also measure the rigid transformation by estimating the HO pose. As in [27], we assume the object to be unknown and we focus on the hand keypoints. We first run an off-the-shelf 2D-3D hand pose estimator, and estimate the rigid transformation between frames by computing the relative transformation of the hand 3D keypoints. As these are centered around the wrist, while 2D keypoints are estimated in the pixel space, we first run a PnP algorithm to obtain 3D keypoints in the scene. Then, we estimate the rigid transformation, *i.e*. camera poses, between frames via Procrustes alignment.

4.2. Reconstruction from multiple observations

Reconstruction from robust visual hulls (VH). First, we consider the silhouette-based formulation from [32] as a baseline for reconstruction, using GT silhouettes. Following their notation, we set $\alpha = N/8$ and $\beta = N/4$.
Reconstruction with fast differentiable rendering (FDR). We also benchmark the recent method from [47]. They propose a coarse-to-fine differentiable rendering method, targeted at multiview surface capture problems.
Reconstruction with neural implicit surfaces. We finally consider the more advanced method proposed in HHOR [27] that combines NeuS [51], a NeRF representation where the density radiance field is replaced with a Signed Distance Field (SDF), with semantic-guided ray sampling (to focus more on the object) and a camera refinement stage. This step simultaneously optimizes SDF and camera poses to compensate for imprecise estimations.

5. Experimental results

We now evaluate the 2 stages of the pipeline, namely rigid registration (Section 5.1) and MVR (Section 5.2).

5.1. Rigid transformation estimation evaluation

We report results for estimating the rigid transformations either from hand poses or from COLMAP in Tab 2. As the performance for the hand-based method is likely correlated with hand pose accuracy, we also evaluate hand 3D pose estimation for 4 different image-based methods: (i) Minimal Hand [57] an easy to use real-time system, (ii) FrankMocap [43], used in IHOI [54] and HHOR [27], (iii) the recent HandOccNet [38] and (iv) the hand module of DOPE [52] which proved to perform well under hand-object interactions [1]. We found that DOPE outperforms the other methods by a large margin and selected it as hand pose estimator.

We also investigate three methods to further smooth the per-frame DOPE predictions: (i) Exploiting the rigid motion assumption, by computing a median pose resulting from an aggregation of all hand poses across the sequence. (ii) By applying a median filter on pose sequences, with a sliding window of 5 frames. (iii) Using PoseBERT [5] a transformer module for smoothing 3D pose sequences. We found simple baselines (i) and (ii) to perform better.

We found that better hand pose estimations tend to lead to better rigid transformations but COLMAP performs the best. However, it yields a lower detection rate compared to its hand pose counterpart (which always provides an estimation), requires accurate segmentation and recovers the camera poses up to an unknown scale factor. Hand-based poses naturally embed a rough scale information and the resulting reconstructions have a similar scale to that of GT meshes, which is an interesting property.

5.2. Hand-object 3D Reconstruction evaluation

In Tab. 3, we report accuracy (acc), completeness (comp), as well as Fscore for the different reconstruction methods after Procrustes in rotation, translation and scale to the GT scans. First, we evaluate the performance of IHOI [54], a recent single-image template-free HO reconstruction method. We use the annotated MANO joints for alignment, which is thus near-perfect. This explains the overall good results despite severe artefacts in the reconstructions (see Supp. Mat.). On the other hand, these results show that a strong hand prior helps for this challenging task. The reconstruction rate reported in the table is expressed frame-wise for this method.

Rigid Transform	Recon. Method	Rec. rate (%) ↑	Acc.† (cm) ↓	Comp.† (cm) ↓	Acc. ratio @5mm (%) ↑	Comp. ratio @5mm (%) ↑	Fscore @5mm (%) ↑
GT	IHOI [54]	87.3	0.79	1.34	41.7	37.8	39.3
GT	VH [32]	93.7	0.42	0.65	67.3	61.6	63.6
GT	FDR [47]	95.8	0.35	0.49	75.8	72.0	73.5
GT	HHOR [27]	**98.9**	**0.34**	**0.31**	**81.0**	**83.7**	**82.2**
DOPE [52]	FDR [47]	**92.7**	1.02	3.18	31.7	15.7	20.0
COLMAP [44]	FDR [47]	76.0	**0.64**	0.79	39.3	36.2	37.6
COLMAP [44]	HHOR [27]	72.9	0.65	**0.73**	**53.7**	**55.2**	**54.2**

Table 3: **Hand-object reconstruction evaluation** using either ground-truth rigid transformations or estimated ones. † means that the metrics are obtained by computing on the reconstructed mesh only, the failing ones are not taken into account, making direct comparison between different methods unfair. DOPE refers to the variant 'DOPE + fixed hand pose' from Tab 2.

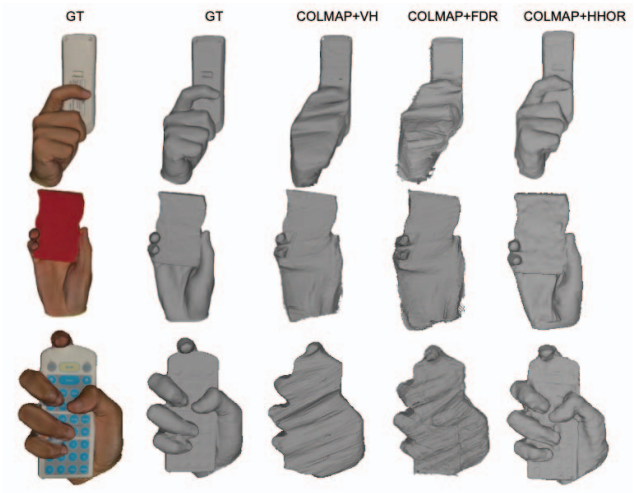

Figure 7: **Qualitative reconstruction results.**

Using GT rigid transforms, all 3 reconstruction methods lead to an excellent result (Fscore above 60% at $5mm$). The recent HHOR method performs better for all metrics. We then evaluate the FDR reconstruction when using estimated rigid transforms, with either hand keypoints or SfM. The performance drops, *e.g.* from a Fscore @5mm from 73.5% to 37.6% using COLMAP, and to 20% using DOPE. Next, we evaluate HHOR and observed a 16.6% boost compared to FDR (vs a 9% boost only when using GT rigid transforms). The camera pose refinement corrects noisy camera poses from COLMAP at the expense of a much heavier computational cost (1 GPU.day per sequence for HHOR *vs.* less than a minute for FDR). We show qualitative results in Fig. 7 and in Supp. Mat. VH cannot reconstruct concavities, *e.g.*, between fingers, while FDR is slightly better. We can appreciate that the shapes reconstructed by HHOR are highly-detailed. Note that HHOR [27] reported very poor results with COLMAP, justifying the use of FrankMocap to estimate the rigid transforms. However, FrankMocap performs very poorly on our dataset of varied HO interactions. We posit that the unrealistic close-up fingertip grasps in their HOD dataset allowed accurate hand pose estimates, and it is not the case at all in our setup.

method	object size	Acc. ratio @5mm (%) ↑	Comp. ratio @5mm (%) ↑	Fscore @5mm (%) ↑
COLMAP+FDR	small	31.78	28.95	30.23
	larger	**50.05**	**46.23**	**47.93**
DOPE+FDR	small	**35.38**	**18.58**	**23.43**
	larger	29.44	13.96	17.85

Table 4: **HO reconstruction evaluation *vs*. object size.**

Detailed analysis. Upon careful analysis, we found that COLMAP failed or performed poorly on objects of small size compared to larger-size objects. To corroborate this, we categorize the objects in our dataset to small and larger (*i.e.*, large and medium) objects and compute reconstruction errors on these two sets of objects. Table 4 shows the reconstruction metrics. We observe that COLMAP leads to better results on larger objects while DOPE is better for small objects. For small objects, there may not be sufficient features detected for the matching step which is critical for camera pose estimation by COLMAP. On the other hand, small objects lead to less hand occlusions and better hand joint estimates, which in turn results in robust rigid-transformation estimation. This strongly emphasizes that a robust hand key points estimator is key for accurate rigid-transformation estimation in the case of small objects with little visual support to perform a standard pose estimation.

6. Conclusion

We introduced the SHOWMe dataset to tackle the problem of detailed 3D reconstruction of a hand holding rigidly an unknown object from a monocular video. We then benchmarked several video-based baselines that follow a common two-stage pipeline consisting of a rigid registration step followed by a multi-view reconstruction. Even if high-quality HO 3D reconstructions are obtained in some cases, their quality highly depends on the initial rigid transformation estimates which can be difficult to obtain in case of texture-less objects or heavy occlusions of the hands. There is still room for improvement regarding the reconstruction quality too and we hope SHOWMe will help foster further research in this direction.

References

[1] Anil Armagan, Guillermo Garcia-Hernando, Seungryul Baek, Shreyas Hampali, Mahdi Rad, Zhaohui Zhang, Shipeng Xie, Mingxiu Chen, Boshen Zhang, Fu Xiong, Yang Xiao, Zhiguo Cao, Junsong Yuan, Pengfei Ren, Weiting Huang, Haifeng Sun, Marek Hrúz, Jakub Kanis, Zdenek Krnoul, Qingfu Wan, Shile Li, Linlin Yang, Dongheui Lee, Angela Yao, Weiguo Zhou, Sijia Mei, Yunhui Liu, Adrian Spurr, Umar Iqbal, Pavlo Molchanov, Philippe Weinzaepfel, Romain Brégier, Grégory Rogez, Vincent Lepetit, and Tae-Kyun Kim. Measuring generalisation to unseen viewpoints, articulations, shapes and objects for 3d hand pose estimation under hand-object interaction. In *ECCV*, 2020. 7

[2] Artec3d. Artec3D structure light 3d scanner. 4

[3] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 640–653. Springer, 2012. 4

[4] Sven Bambach, Stefan Lee, David J. Crandall, and Chen Yu. Lending A Hand: Detecting Hands and Recognizing Activities in Complex Egocentric Interactions. In *ICCV*, 2015. 3

[5] Fabien Baradel, Romain Brégier, Thibault Groueix, Philippe Weinzaepfel, Yannis Kalantidis, and Grégory Rogez. Posebert: A generic transformer module for temporal 3d human modeling. *IEEE Trans. PAMI*, 2022. 7

[6] Samarth Brahmbhatt, Chengcheng Tang, Christopher D Twigg, Charles C Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *ECCV*, 2020. 2, 3

[7] Romain Brégier. Deep regression on manifolds: a 3D rotation case study. In *3DV*, 2021. 5

[8] Ian M Bullock, Thomas Feix, and Aaron M Dollar. The Yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *IJRR*, 2015. 3

[9] Minjie Cai, Kris M Kitani, and Yoichi Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015. 3

[10] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *IEEE Robotics and Automation Magazine*, 2015. 3

[11] Zhe Cao, Ilija Radosavovic, Angjoo Kanazawa, and Jitendra Malik. Reconstructing hand-object interactions in the wild. In *ICCV*, 2021. 2, 4

[12] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, et al. Dexycb: A benchmark for capturing hand grasping of objects. In *CVPR*, 2021. 3

[13] Zerui Chen, Yana Hasson, Cordelia Schmid, and Ivan Laptev. Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In *ECCV*, 2022. 2, 4

[14] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, 2017. 3

[15] Enric Corona, Tomas Hodan, Minh Vo, Francesc Moreno-Noguer, Chris Sweeney, Richard Newcombe, and Lingni Ma. Lisa: Learning implicit shape and appearance of hands. In *CVPR*, 2022. 2, 3

[16] Enric Corona, Albert Pumarola, Guillem Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *CVPR*, 2020. 3

[17] Zicong Fan, Omid Taheri, Dimitrios Tzionas, Muhammed Kocabas, Manuel Kaufmann, Michael J Black, and Otmar Hilliges. Articulated objects in free-form hand interaction. *arXiv preprint arXiv:2204.13662*, 2022. 3

[18] Alireza Fathi, Xiaofeng Ren, and James M Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011. 3

[19] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 6

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

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

[22] Shreyas Hampali, Sayan Deb Sarkar, and Vincent Lepetit. Ho-3d_v3: Improving the accuracy of hand-object annotations of the ho-3d dataset. *arXiv preprint arXiv:2107.00887*, 2021. 3

[23] Shreyas Hampali, Sayan Deb Sarkar, Mahdi Rad, and Vincent Lepetit. Handsformer: Keypoint transformer for monocular 3d pose estimation ofhands and object in interaction. *arXiv preprint arXiv:2104.14639*, 2, 2021. 4

[24] Yana Hasson, Bugra Tekin, Federica Bogo, Ivan Laptev, Marc Pollefeys, and Cordelia Schmid. Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In *CVPR*, 2020. 2, 4

[25] Yana Hasson, Gül Varol, Cordelia Schmid, and Ivan Laptev. Towards unconstrained joint hand-object reconstruction from rgb videos. In *3DV*, 2021. 2, 4

[26] Yana Hasson, Gul Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2, 3, 4, 6

[27] Di Huang, Xiaopeng Ji, Xingyi He, Jiaming Sun, Tong He, Qing Shuai, Wanli Ouyang, and Xiaowei Zhou. Reconstructing hand-held objects from monocular video. In *SIGGRAPH Asia*, 2022. 2, 3, 4, 7, 8

[28] IntelRealsense. Intel realsense l515 rgbd camera. 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 *3DV*, 2020. 4

[30] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. In *IROS*, 2019. 4

[31] Taein Kwon, Bugra Tekin, Jan Stühmer, Federica Bogo, and Marc Pollefeys. H20: Two hands manipulating objects for first person interaction recognition. In *ICCV*, 2021. 3

[32] Vincent Leroy, Jean-Sébastien Franco, and Edmond Boyer. Volume sweeping: Learning photoconsistency for multi-view shape reconstruction. *IJCV*, 2021. 2, 7, 8

[33] Shaowei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. In *CVPR*, 2021. 4

[34] Wei-Chiu Ma, Anqi Joyce Yang, Shenlong Wang, Raquel Urtasun, and Antonio Torralba. Virtual correspondence: Humans as a cue for extreme-view geometry. In *CVPR*, 2022. 2

[35] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In *ICCV*, 2017. 3

[36] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Generalized feedback loop for joint hand-object pose estimation. *IEEE transactions on pattern analysis and machine intelligence*, 42(8):1898–1912, 2019. 4

[37] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 4

[38] JoonKyu Park, Yeonguk Oh, Gyeongsik Moon, Hongsuk Choi, and Kyoung Mu Lee. Handoccnet: Occlusion-robust 3d hand mesh estimation network. In *CVPR*, 2022. 7

[39] Neng Qian, Jiayi Wang, Franziska Mueller, Florian Bernard, Vladislav Golyanik, and Christian Theobalt. HTML: A Parametric Hand Texture Model for 3D Hand Reconstruction and Personalization. In *ECCV*. 2020. 4

[40] Grégory Rogez, James Steven Supancic III, and Deva Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015. 3

[41] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 3

[42] Javier Romero, Dimitrios Tzionas, and Michael J Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM ToG*, 2017. 2, 3, 4, 6

[43] Yu Rong, Takaaki Shiratori, and Hanbyul Joo. Frankmocap: A monocular 3d whole-body pose estimation system via regression and integration. In *ICCV Workshops*, 2021. 7

[44] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *CVPR*, 2016. 2, 6, 7, 8

[45] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016. 3

[46] Omid Taheri, Nima Ghorbani, Michael J Black, and Dimitrios Tzionas. Grab: A dataset of whole-body human grasping of objects. In *ECCV*, 2020. 3

[47] Briac Toussaint, Maxime Genisson, and Jean-Sébastien Franco. Fast gradient descent for surface capture via differentiable rendering. In *3DV*, 2022. 2, 7, 8

[48] Tze Ho Elden Tse, Kwang In Kim, Ales Leonardis, and Hyung Jin Chang. Collaborative learning for hand and object reconstruction with attention-guided graph convolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1664–1674, 2022. 2

[49] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing Hands in Action using Discriminative Salient Points and Physics Simulation. *IJCV*, 2016. 4

[50] Dimitrios Tzionas, Abhilash Srikantha, Pablo Aponte, and Juergen Gall. Capturing hand motion with an rgb-d sensor, fusing a generative model with salient points. In *GCPR*, 2014. 4

[51] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. 2021. 7

[52] Philippe Weinzaepfel, Romain Brégier, Hadrien Combaluzier, Vincent Leroy, and Grégory Rogez. DOPE: Distillation of part experts for whole-body 3D pose estimation in the wild. In *ECCV*, 2020. 7, 8

[53] Lixin Yang, Kailin Li, Xinyu Zhan, Fei Wu, Anran Xu, Liu Liu, and Cewu Lu. Oakink: A large-scale knowledge repository for understanding hand-object interaction. In *CVPR*, 2022. 3

[54] Yufei Ye, Abhinav Gupta, and Shubham Tulsiani. What's in your hands? 3d reconstruction of generic objects in hands. In *CVPR*, 2022. 2, 4, 7, 8

[55] Hao Zhang, Yuxiao Zhou, Yifei Tian, Jun-Hai Yong, and Feng Xu. Single depth view based real-time reconstruction of hand-object interactions. *TOG*, 2021. 4

[56] Juyong Zhang, Yuxin Yao, and Bailin Deng. Fast and robust iterative closest point. *IEEE Trans. PAMI*, 2021. 5

[57] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. Monocular real-time hand shape and motion capture using multi-modal data. In *CVPR*, 2020. 7

[58] 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*, 2019. 3