Dense Hand-Object(HO) GraspNet with Full Grasping Taxonomy and Dynamics

Woojin Cho[1], Jihyun Lee[1], Minjae Yi[1], Minje Kim[1], Taeyun Woo[1], Donghwan Kim[1], Taewook Ha[1], Hyokeun Lee[3], Je-Hwan Ryu[4], Woontack Woo[1], Tae-Kyun Kim[1,2]

[1]KAIST,[2]Imperial College London,[3]Kwangwoon University,[4]Surromind

Abstract. Existing datasets for 3D hand-object interaction are limited either in the data cardinality, data variations in interaction scenarios, or the quality of annotations. In this work, we present a comprehensive new training dataset for hand-object interaction called HOGraspNet. It is the only real dataset that captures full grasp taxonomies, providing grasp annotation and wide intraclass variations. Using grasp taxonomies as atomic actions, their space and time combinatorial can represent complex hand activities around objects. We select 22 rigid objects from the YCB dataset and 8 other compound objects using shape and size taxonomies, ensuring coverage of all hand grasp configurations. The dataset includes diverse hand shapes from 99 participants aged 10 to 74, continuous video frames, and a 1.5M RGB-Depth of sparse frames with annotations. It offers labels for 3D hand and object meshes, 3D keypoints, contact maps, and *grasp labels*. Accurate hand and object 3D meshes are obtained by fitting the hand parametric model (MANO) and the hand implicit function (HALO) to multi-view RGBD frames, with the MoCap system only for objects. Note that HALO fitting does not require any parameter tuning, enabling scalability to the dataset's size with comparable accuracy to MANO. We evaluate HOGraspNet on relevant tasks: grasp classification and 3D hand pose estimation. The result shows performance variations based on grasp type and object class, indicating the potential importance of the interaction space captured by our dataset. The provided data aims at learning universal shape priors or foundation models for 3D hand-object interaction. Our dataset and code are available at https://hograspnet2024.github.io/.

Keywords: hand-object interaction · grasp taxonomy · 3D shape and pose estimation · new benchmark

1 Introduction

The importance of modeling and inferring 3D hand-object interactions is growing. While earlier works focused on single object instances [25,65,69,70,75,77,82,87,88,90,92,95–97], recent efforts have been made on multiple 3D objects and their complex interactions [4,8,16,19,24,28,33,39,40,55,58,59,72,74,81,86,94]. The human hand is the most dexterous and important testbed, and its research is extendable to human bodies or faces in similar articulated and deformable categories. We observe a few new benchmarks on hand-object interaction each latest year. However, existing datasets are limited either in the cardinality of

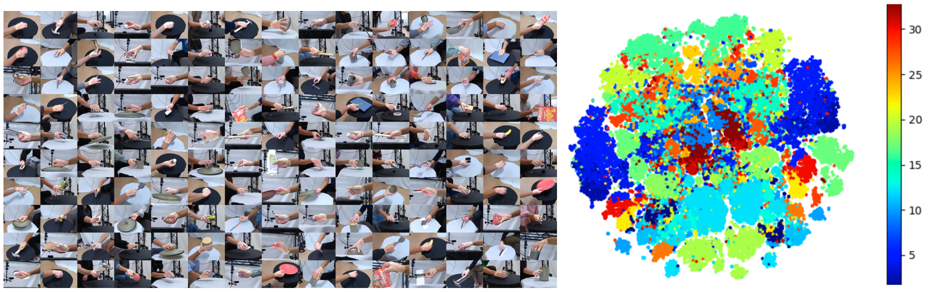

Fig. 1: (left) Diverse samples in HOGraspNet (best viewed with zoom-in). HOGraspNet captures all hand-object grasp taxonomies with high-quality 3D annotations. **(right) Grasp Taxonomy t-SNE.** It covers well the grasp taxonomy space with intra-class variations.

data, the amount of data variations in hands/objects, or the quality of annotations. See Tab. 1, where HO3D [28] and DexYCB [8] include only 15 and 14 (out of 33) grasping taxonomies respectively. YCB Affordance [16] is only an existing benchmark that represents all grasp taxonomies and provides grasp labels, but synthetic; ARCTIC places visible markers on hands in RGB images, and OakInk (the closest to ours) does not provide grasp labels with fewer subjects but more objects. More comparison with OakInk is shown in the t-SNE plot Fig. 5.

We introduce HOGraspNet, an extensive multi-view RGBD training dataset for hand, object, and their interaction with grasp annotations. Based on the existing hand grasping taxonomy [20], our design redefines 28 of the 33 grasps by merging geometrically similar or uncommon poses. Our dataset is the only real dataset covering all grasp taxonomies, including grasp labels and a wide range of intraclass variations. We exploit 22 rigid objects from the YCB dataset [5] with 8 other compound/articulated objects. As an example shown in Fig. 4, 3 distinct hand grasps are performed for each object, totaling 90 interaction scenarios. Note that 30 objects are chosen enough to cover all grasp taxonomies, while textures and shapes beyond grasp areas can be synthetically augmented with 3D models. The dataset comprises a diverse range of hand identities from 99 participants aged 10 to 74. Overall, the dataset contains 1.5M RGB-Depth frames from 4 viewpoints, with annotations for 3D meshes, 3D keypoints, contact maps, and *grasp labels*. We adopt the hand parametric model (MANO [68]) and the hand implicit function (HALO [41]) individually to annotate the hand mesh. Presenting the novel annotation pipeline using the hand implicit function that requires simpler settings (i.e., less hyper-parameters) than MANO with accuracy and continuous shape representation. Considering the small objects in HOGraspNet, we utilize optical markers for MoCap only to obtain object 6D pose. We report experimental results of grasp classification, and SOTA hand-object 3D pose estimation methods. The new dataset demonstrates its comprehensiveness and potential.

Further possibilities from the presented dataset are to 1) synthetically augment the data by changing backgrounds or object textures and shapes beyond grasp areas, 2) provide environments for learning a grasping agent in physi-

Table 1: Comparison of hand-object interaction datasets. Interaction info is the key criteria utilized in terms of hand-object interaction.

Dataset	Type	#image	#views	#obj	#subj	#Grasps in [20]	Real	Video	Marker-less hand	Dynamic interaction	Hand-obj contactmap	Grasp variation	Grasp annotation	Interaction info.
Obman(CVPR19) [34]	RGBD	154k	1	3k	20		✗	✗	✓	✗	✗	✗	✗	
YCB-Affordance(CVPR20) [16]	RGB	133k	1	58	-	100%	✗	✓	✓	✗	✗	✓	✓	Grasp
FreiHAND(ICCV18) [97]	RGB	37k	8	2	32		✓	✗	✓	✗	✗	✗	✗	
MOW(ICCV21) [6]	RGB	500	1	~500	-	82%	✓	✗	✓	✗	✗	✗	✗	
DexYCB(CVPR21) [8]	RGBD	582k	8	20	10	42%	✓	✓	✓	✗	✗	✗	✗	
FPHA(CVPR18) [24]	RGBD	105k	1	4	6		✓	✓	✗	✓	✗	✗	✗	Action
HO3D(CVPR20) [28]	RGBD	78k	1	10	10	45%	✓	✓	✓	✓	✗	✗	✗	
SHOWMe(ICCVW23) [73]	RGBD	87k	1	42	15	61%	✓	✓	✓	✓	✗	✗	✗	
ContactPose(ECCV20) [4]	RGBD	2.9M	3	25	50		✓	✓	✓	✓	✓	✗	✗	Intent
H2O(ICCV21) [42]	RGBD	571k	5	8	4		✓	✓	✓	✓	✓	✗	✗	
ARCTIC(CVPR23) [19]	RGBD	2.1M	9	10	9		✓	✓	✗	✓	✓	✓	✗	Intent
OakInk(CVPR22) [84]	RGBD	230k	4	100	12		✓	✓	✓	✓	✓	✓	✗	
Ours	RGBD	1.5M	4	30	99	85%	✓	✓	✓	✓	✓	✓	✓	Grasp

cal simulators, 3) extend to non-grasping actions, e.g., pushing, throwing, or deformed processes of non-rigid objects by hand. We hope the new dataset serves as a basis for understanding and modeling diverse inference models of hand-object interactions and that learned knowledge applies to human-object or human-human interactions.

2 Survey on Interaction Datasets

This section provides a comprehensive overview of existing datasets on the interaction of 3D shapes, i.e., single hand, hand-object, hand-hand, human-object, and human-human. We also briefly discuss the existing literature on hand-object reconstruction, which is used for benchmarking our dataset (in Section 4). Further survey results are available in the supplementary materials.

Single-Hand Datasets. Earlier research efforts to build a hand dataset have focused on capturing single hands from RGB [25,69,92,95,97], depth [65,75,77,82,88], or RGBD [70,87,90,96], stimulating various learning-based methods for hand reconstruction [7,23,47,60,63]. These datasets can be categorized via three characteristics: (1) whether the captured hand frames are synthetic [69,96] or real [25,65,69,70,75,77,82,87,88,90,92,95,97], (2) whether the hand is annotated as sparse keypoints [25,69,75,77,82,88,92,92,96,97] or mesh [87,95], and (3) whether the annotation is obtained via marker-based [25,65,75,88] or marker-less system [69,70,77,82,87,90,92,95–97]. More recently, various hand datasets aim to capture hands in interaction with an object [3,4,8,16,17,19,24,28,33,59,73,74] or another hand [48,57,58,79,98]. Since the goal of our work is to collect a dataset that comprehensively captures hand-object interactions, we focus on discussing the existing hand-object datasets in the following.

Hand-Object Datasets. Recently, various hand-object datasets [3,4,8,16,17,19,24,28,33,59,73,74] have been proposed. Regarding **(1) annotation method**, most of the earlier datasets collect synthetic RGB and/or depth images rendered from a parametric hand model (MANO [68]) and template object models [16,33,59], or collect real images with markers [4,19,74] or magnetic sensors [24] to obtain hand annotations. However, these samples lack realism due to the rendering of synthetic models or the presence of visible sensors. Thus, many recent datasets use a markerless system to fit the MANO model to RGB-D images captured in a multi-view setup while using a minimal number of markers to obtain object poses [8,28,73]. Our work also follows such marker-less capture system to provide MANO-based hand annotations while additionally fitting an implicit function-based hand model (HALO [41]) to provide supplementary hand shape information. Regarding the **(2) characteristics of captured data**, existing datasets are limited either in data cardinality, the number of object categories or hand identities, or interaction taxonomies (please refer to Tab. 1). For example, HO3D [28] and DexYCB [8] (which are the most widely used hand-object datasets) only consider 10 object categories and capture 10 and 20 hand identities, respectively. While ObMan [33] and SHOWMe [73] capture more diverse object categories, they are limited in the number of hand identities (20 and 15, respectively) and the data cardinality (154K and 87K, respectively). ARCTIC [19] and OakInk [84] are recently proposed datasets that capture dexterous interactions between hands and objects, containing a range of motion variations. However, they do not cover the diverse grasp poses for each object, as they instruct participants to assume poses based on their intent to interact with the object. Our work aims to collect a training dataset that is more comprehensive in terms of interaction scenarios based on grasps, object categories, hand identities, and data cardinality. We also note that most of the existing datasets do not provide a grasping type of each sample, which can further provide a useful prior for the captured hand-object interaction [16,26,50]. Our work also carefully identifies a taxonomy of 33 grasping types and provides grasping type annotation for each sample. For comparison with other datasets, we conducted a thorough survey to ascertain the number of grasp classes present among the 33 grasp taxonomies in Feix et al. [20] and reported in Tab. 1. The detailed results are provided in the supplementary material.

Two-Hand Datasets. Similar to hand-object datasets, various interacting two-hand datasets have been proposed. HIC [79] and RGB2Hands [80] are some of the earliest two-hand interaction datasets, but their data cardinality and interaction diversity are small compared to more recent datasets. InterHand2.6M [58] is the most widely-used large-scale dataset, which captures interacting hands in a multi-view markerless motion capture setup. More recently, Re:InterHand [57] is proposed to capture more diverse two-hand interactions in terms of image appearances and interaction poses via the use of environment maps and hand relighting. TwoHand500K [98] is another recently proposed dataset consisting of (1) real data captured using a marker-based system and (2) synthetic data obtained via combining poses randomly sampled from single-hand datasets. These datasets have inspired various methods for interacting two-hand reconstruction [44,46,89] and generation [43,48].

Interacting Human Datasets. In addition, research attention to interacting hands and several datasets contain interacting humans. These can be categorized as human-object interaction, human-human interaction, and human-scene interaction. **(1) Human-object interaction**: As demonstrated in Tab. 2, several datasets have been released to depict various aspects of human-object

Table 2: Comparison of Human-object interaction datasets.

Dataset	Type	#image	#views	#obj	#subj	#Kinects	Label	Contact annotation	Whole body interact.	Marker-less hand	Natural scene	Scene interact.	Dynamic hand	Articulated object
EgoBody(ECCV22) [91]	RGBD	220K	3∼5	15	36	3∼5	SMPL-X	✗	✓	✓	✓	✓	✗	✗
GRAB(ECCV20) [74]	Mesh	1.6M	-	51	10	-	SMPL-X	✓	✗	✗	✗	✗	✓	✗
ARCTIC(CVPR23) [19]	RGB	2.1M	9	10	9	-	SMPL-X	✓	✗	✗	✗	✗	✓	✓
CHAIRS(ICCV23) [38]	RGBD	1.7M	4	81	46	-	SMPL-X	✓	✓	✗	✗	✗	✗	✓
BEHAVE(CVPR22) [2]	RGBD	15K	5	20	8	-	SMPL	✓	✓	✗	✓	✗	✗	✗
InterCap(IJCV24) [37]	RGBD	67K	6	10	10	6	SMPL-X	✓	✓	✓	✓	✗	✗	✗
PROX(ICCV19) [30]	RGBD	100K	3	12	20	1	SMPL-X	✓	✓	✓	✓	✓	✗	✗
RICH(CVPR22) [36]	RGBD	540K	6∼8	5	22	1	SMPL-X	✓	✓	✓	✓	✓	✗	✗

interactions. GRAB [74] and ARCTIC [19] propose datasets with 3d human mesh which focus on hand-object interaction. These three datasets [2, 37, 38] extend the interacting region to the whole body. Leveraging SMPL-X as mesh templates, [2, 19, 37] contain contact supervision, and [19, 38] focus on articulated objects. **(2) Human-human interaction**: PanopticStudio [39], MuCo-3DHP [55] and MuPoTS-3D [55] propose datasets with 3d human sparse keypoints. More recently, there has been increased interest on reconstructing 3d human mesh [21,40,45,61,64,72,81,86,94]. The majority of them employ parametric models like SMPL [52] or SMPL-X [62], except some datasets [61, 64, 81, 86, 94] that provides textured scans, which is beneficial to represent geometric details. **(3) Human-scene interaction**: Certain datasets [30,36,91] broaden the scope of the interaction to include scenes, utilizing SMPL-X as a mesh template. PROX [30] and RICH [36] provide contact supervision between humans and scenes. RICH [36], in particular, extends its scope to outdoor scenes, and EgoBody [91] provides motion text labels.

Hand-Object Reconstruction. Reconstructing hand and object in interaction has been actively explored. Most of the recent works can be categorized into optimization or learning-based methods. Optimization-based methods [6, 27, 32, 78, 85] typically fit MANO [68] hand and template object models based on contact or other physical constraints (e.g., attraction and repulsion [85], friction [35]). Learning-based methods [10–14, 18, 29, 31, 33, 51, 76] directly regress hand and object poses via a neural network, while focusing on exploring an effective architecture for feature learning [10, 18, 29, 31, 33, 51, 76] and/or shape representation [11, 12]. In our work, we benchmark the RGB-based reconstruction task on our HOGraspNet dataset using HFL-Net [49], which is the most recent state-of-the-art method.

3 HOGraspNet

3.1 Dataset Overview

The dataset includes continuous video images and 1,489,112 annotated RGB-D frames, covering 28 hand grasp classes. We redefined the grasp classes by merging visually similar configurations (see supplementary) using 30 objects. The frames were captured at 4 distinct viewpoints and performed by 99 participants aged 10 to 74. Along with diverse hand shapes, a good scope of intra-class variations within each grasp class has been collected, which is important as a training dataset. Each RGB-D frame is annotated with the 3D hand pose for 21 joints

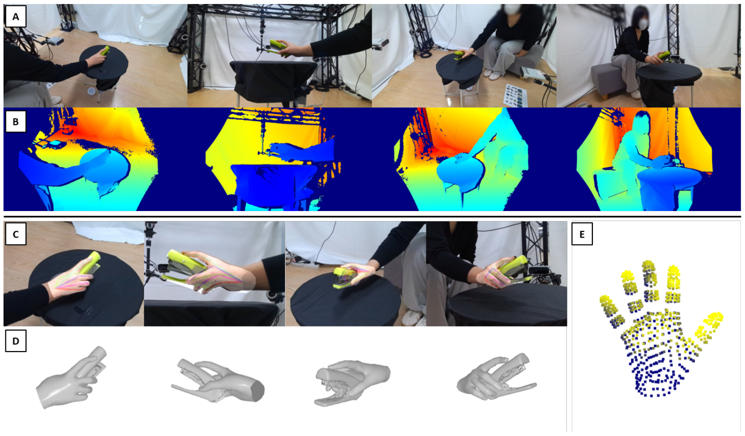

Fig. 2: Structure of HOGraspNet. It captures diverse hand-object grasping at 4 different viewpoints. Example RGB images (A) and depth images (B) are shown, while the fitted hand and object meshes are visualized in (C) and (D). (E) shows the contact map.

and mesh, corresponding grasp class, 6D object pose, and contact map between the hand and object. Hand mesh models are obtained fitting MANO [68] and/or HALO [41], while all object mesh models (3d shapes and textures) are pre-scanned and provided. In Fig. 2, we present examples of the data types included in the dataset.

3.2 Object Categories and Grasp Taxonomy

While interacting with various objects, we often take specific grasp poses based on the object's shape and intention. To capture a wide range of hand pose space, especially to cover all grasping taxonomies, we identified 22 types of objects from the YCB dataset [5] and 8 other daily objects. These objects are selected considering factors, such as the primitive shapes of objects (cylinder, sphere, disk, cube), especially grasp areas, object sizes, and additional articulated/compound objects (see Fig. 3 (right)). Referring to previous studies on hand-object grasp taxonomy [1, 15, 20, 50, 71], we captured the three most common grasp classes for each object. Example per-object taxonomies are shown in Fig. 4. Also, some grasp classes out of 33 are seemingly redundant, visually hard to distinguish, and geometrically close; we redefined them to 28 grasp classes, with their indices following those presented in [20]. Compared to the existing benchmarks (MOW [6] and OakInk [84]), we have a relatively smaller set of objects. However, we span more grasp space with large intra-class variations, thanks to diverse hand shapes and the number of frames. We consider synthetic object augmentation in textures and shapes beyond grasp areas and background augmentation as future work.

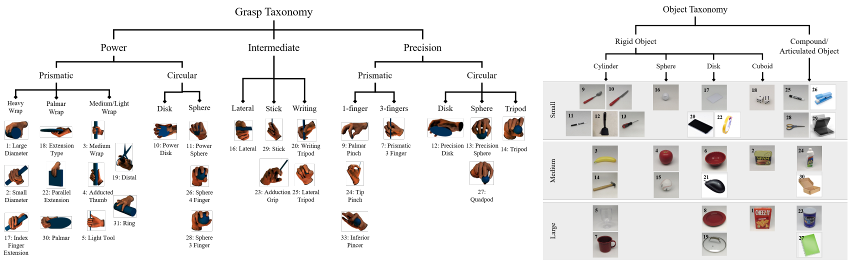

Fig. 3: (left) 33 hand grasping taxonomies, (right) 30 objects used in the dataset. The object types are cylinder, sphere, disk, cuboid, or compound/articulated. They are further dividend to small/medium/large sizes, purporting to cover all grasp taxonomies.

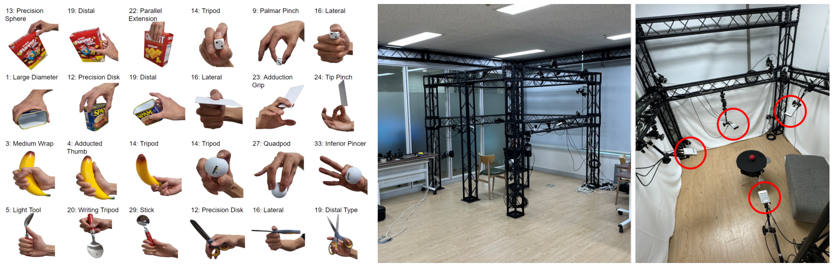

Fig. 4: (left) Per-object taxonomy examples (right) System setup. The full list is shown in the supplementary.

3.3 Hardware Setup and Data Collection

Sensors. Fig. 4 shows the recording studio setup, where 4 temporally synchronized RGB-D cameras (Azure Kinect) are positioned around the designated space. The one in the backside is roughly at users' eye locations, imitating a fixed egocentric view. The cameras capture RGB and depth at 1920x1080 resolution and 30 FPS. For object poses, 8 IR cameras with a frame rate of 120 FPS were set, and 3 to 5 optical markers(3mm) were attached to each object. Notably, no markers were used on hands to maintain their realistic appearance, thereby minimizing the potential degradation of image features in networks trained using our dataset due to RGB image contamination (cf. ARCTIC [19]). However, markers on objects can still limit hand poses, so we minimized this impact by placing markers on regions least likely to be grasped (e.g., the blade of scissors). Note that all objects were symmetrical enough to place markers while avoiding contact areas. Temporal synchronization between the RGB-D and IR cameras was obtained by manually aligning the starting frames during each recording session through a start blink of the LED.

Data acquisition. We conducted data capture involving 99 participants with diverse hand sizes, shapes, and textures. Detailed instructions regarding grasp classes for each object were provided, and participants were requested to grasp each object with their right hand according to the specified grasp while freely

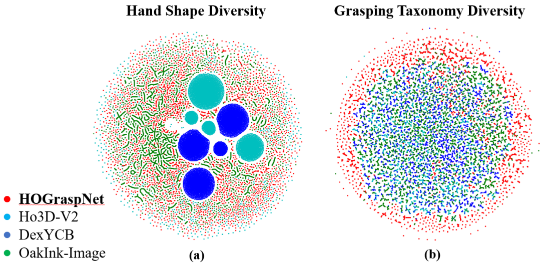

Fig. 5: t-SNE [54] visualization of (left) MANO [68] shape parameter distributions and (right) grasp feature distributions.

performing pose variations such as translation and rotation. Each participant completed the procedure 2 to 4 times, with each trial recorded for 20 seconds to adequately capture actions ranging from reaching for the object to freely manipulating it in the air and eventually placing it back down. This way, diverse intra-class variations were captured.

3.4 Data Distributions

To further demonstrate that our dataset captures more comprehensive hand grasps, we visualize our data distribution in comparison to HO3D [28], DexYCB [8], and OakInk [84]. In Fig. 5(a), we show t-SNE [54] visualizations of the four datasets in the MANO [68] shape parameter space. Our dataset captures more hand shape diversity than the others, as a larger number of hand identities were included (as shown in Tab. 1). Fig. 5(b) shows the t-SNE visualizations in the grasp feature space. For grasp feature extraction, we train a hand auto-encoder with mesh reconstruction loss and the auxiliary contact reconstruction and grasp classification losses (see Sec. 4.2 for more details) to obtain features that capture hand pose and grasp configurations. Ours is shown to be significantly more diverse than the other datasets in this feature space as well, thanks to our data acquisition process associated with carefully determined grasping types. We hope that the comprehensiveness of our dataset can serve as an effective prior for the downstream tasks related to hand-object interaction.

3.5 MANO and Object Annotation

For MANO [68] hand and object annotation, we use an automatic annotation and verification pipeline inspired by prior studies [8,28]. In the following subsections, we discuss each step of our pipeline, while more details can be found in the supplementary.

Data Preprocessing. We downsample the captured RGB-D frames from 30FPS to 10FPS to filter temporally redundant samples. We also prepare the hand and object segmentation masks using the DeepLabv3 [9] model, which is fine-tuned using our data with a few manually annotated segmentation masks for each object.

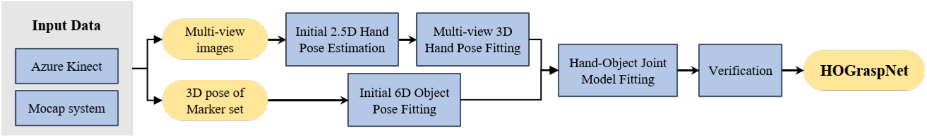
Input Data
Azure Kinect
Mocap system
Multi-view images
3D pose of Marker set
Initial 2.5D Hand Pose Estimation
Multi-view 3D Hand Pose Fitting
Initial 6D Object Pose Fitting
Hand-Object Joint Model Fitting
Verification
HOGraspNet

Fig. 6: MANO [68] and object annotation pipeline (Section 3.5).

Initial Hand Keypoint Estimation. To prepare the initial hand keypoints used for MANO fitting, we use MediaPipe [53] hand pose estimator, which is known to have high generalization ability. We estimate 2.5D hand keypoints from each multi-view frame and lift them to 3D keypoints via triangulation. However, such keypoint estimates may be noisy for viewpoints with high hand-object occlusion. To overcome this, we introduce a novel bootstrapping procedure to achieve a better 3D keypoint lifting quality. Given the 2.5D keypoint estimates from our four viewpoints $\{vp_i\}_{i=0,1,2,3}$, we obtain the lifted 3D keypoints $\{\hat{J}_i\}_{i=0,1,2,3}$, where $\hat{J}_i \in \mathbb{R}^{21\times 3}$ denotes 3D keypoints lifted using three viewpoints while *excluding* vp_i. We assume that if the MPJPE between (1) $\hat{J}_i$ projected onto vp_i and (2) the original 2.5D keypoint estimates from vp_i is above a threshold τ, then the original estimates from vp_i is an outlier. In this way, we filter out noisy 2.5D keypoints during 3D lifting procedure to obtain more robust 3D keypoints per frame. The valid hand poses for each viewpoint and the visibility v_i for each joint i (computed using depth maps) are also stored to serve as pseudo-ground truth (GT) data in the following steps.

Initial Object Pose Estimation. To obtain the initial 6D poses of an object, we attach optical sensors to the predefined surface locations of each object. Using multiple IR cameras, the 3D positions of each optical sensor are collected through in-built software. Then, the object's 3D rotation and translation are computed via Least-Squares Fitting to the marker positions.

Multi-view Multi-frame Gradual Hand-object Model Fitting. In this stage, our goal is to fit MANO [68] hand and object template models to multi-view RGB-D frames and the initial hand and object poses. To this end, we formulate an optimization-based fitting scheme similar to previous works [8, 97]. To avoid local minima, we further propose to gradually fit the MANO parameters, such that our optimization consists of three stages: (1) fitting global hand transformation, (2) fitting partial hand poses extended from the wrist, and (3) fitting the full hand and object pose (see the supplementary for details).

Our overall loss function for the MANO pose $\theta \in \mathbb{R}^{48}$ and shape $\beta \in \mathbb{R}^{10}$ parameters and the object 6D pose $\phi \in \mathbb{R}^{6}$ can be written as:

$$\mathcal{L} = \lambda_h^{2D}\mathcal{L}_h^{2D} + \lambda_o^{3D}\mathcal{L}_o^{3D} + \lambda_{seg}\mathcal{L}_{seg} + \lambda_{depth}\mathcal{L}_{depth} + \lambda_{reg}\mathcal{L}_{reg} + \lambda_{phy}\mathcal{L}_{phy}. \quad (1)$$

$\mathcal{L}_h^{2D}$ measures the L2 distance between the pseudo GT 2D joints and the 2D projection of the MANO 3D joints weighted by the visibility v_i. $\mathcal{L}_o^{3D}$ computes the L2 distance between the 3D marker positions and the corresponding vertex position of an object model. $\mathcal{L}_{seg}$ and $\mathcal{L}_{depth}$ measures the L1 distance between the GT and the rendered segmentation masks and depth maps, respectively.

Following [97], we also incorporate a regularization term $\mathcal{L}_{reg} = ||\tilde{\theta}||_2 + ||\tilde{\beta}||_2 + ||\theta_t - \theta_{t-1}||_2 + ||\beta_t - \beta_{t-1}||_2$, which (1) penalizes MANO pose and shape parameters that deviate too much from the mean zero vectors and (2) encourages the previous and current hand parameters to be close for temporal consistency. To additionally regularize the fitted hand and object meshes to be physically plausible, we incorporate another regularization term $\mathcal{L}_{phy}$, which is designed as a weighted sum of penetration loss and contact loss: $\mathcal{L}_{phy} = \lambda_{pen}\mathcal{L}_{pen} + \lambda_{contact}\mathcal{L}_{contact}$. For penetration loss λ_{pen}, we use a vertex normal projection-based technique used in [28]. For contact loss $\mathcal{L}_{contact}$, we minimize the distances between hand and object vertices below a distance threshold τ to encourage physical contact. In Equation 1, $\{\lambda_i\}_{i=h,\,o,\,seg,\,depth,\,reg,\,phy}$ is a set of scalar values to control the weighting between the loss terms. Please see the supplementary for more details about the annotation procedure.

Post-verification. We conduct both automatic and manual verification steps to further filter out the noisy annotations. We compute the Intersection over Union (IoU) between the pseudo-GT and the rendered segmentation masks, filtering out annotations with an IoU below 0.6 in any view. Subsequently, we perform manual verification through crowdsourcing using LabelOn(https://www.labelon.kr/). Each crowdsourcer identifies misprocessed data that significantly deviates from the hand and object meshes or results from operational errors.

3.6 HALO Annotation

HALO [41] fitting. For hand shape annotation, we additionally provide the hand implicit surface based on HALO [41], which is a neural implicit representation that parameterizes an articulated occupancy field [56] with 3D hand keypoints. Thus, a straightforward approach to fit HALO to our collected data would be to use the 3D hand keypoints lifted from the multi-view 2D keypoint estimates (as described in Section 3.5) as an input to the HALO model. However, we observe that it leads to a less plausible implicit hand surface since the keypoints are not guaranteed to form a valid kinematic structure of the hand. Thus, we postprocess the lifted keypoints to the nearest keypoints on the hand space learned by MANO via the inverse kinematics algorithm in [11]. Our HALO fitting results are shown in Fig. 7.

Comparisons with MANO [68] Fitting. As HALO [41] is an occupancy function that takes hand keypoints as input, it does not require many hyperparameters (except for an occupancy threshold [56]) for model fitting, while MANO fitting typically requires numerous loss weighting terms (i.e., λ_* in Equation 1) for defining the optimization objective. Thus, HALO can be more convenient and scalable for annotating a large-scale dataset. Also, HALO can produce hand shapes in a resolution higher than MANO due to its resolution-independent nature (see Fig. 8(a)). However, HALO is shown to capture less hand shape variation than MANO (see the red circle in Fig. 8(b)) due to its keypoint-based parameterization for hand shape. As shown in Fig. 8(c)), the IoU distribution of HALO is marginally worse than that of MANO, but it still achieves comparable mean IoU results (MANO: 0.739, HALO: 0.719) despite its simple annotation procedure.

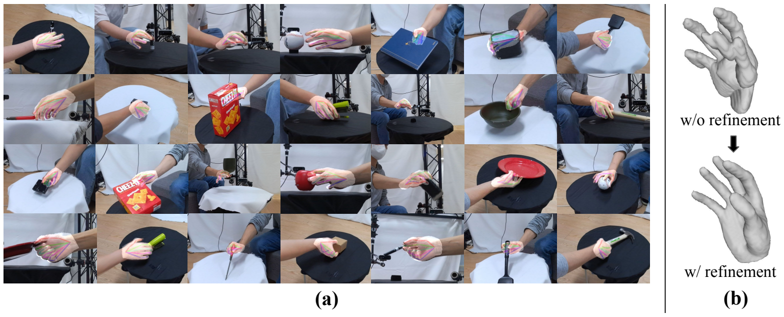

Fig. 7: HALO [41] fitting results. (a) Annotated HALO hand examples. **(b)** Comparisons between the HALO shapes with and without applying inverse kinematics-based keypoint refinement [11].

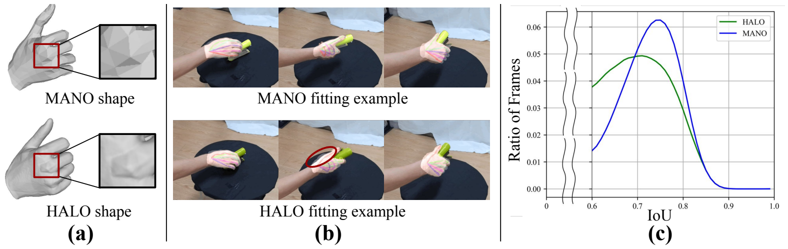

Fig. 8: Comparisons between HALO [41] and MANO [68] fitting results. (a) Hand shapes. **(b)** Fitting examples. **(c)** IoU distributions after post-verification stage.

4 Experimental Results

In this section, we first report the split protocols of HOGraspNet (Section 4.1). We then present our experimental results on grasp classification (Section 4.2) and hand-object pose estimation (Section 4.3) using our dataset.

4.1 Split Protocols

For the evaluation setup, we generated five distinct train/test splits based on key components within our dataset:

- **S0 (default).** This split encompasses all subjects, views, objects, and grasp classes. The dataset is split by sequences, with the first sequence of each subject selected as the test set and the remaining sequences used for training.
- **S1 (unseen subjects).** The dataset is split by subjects, following a 7:3 train/test ratio.
- **S2 (unseen views).** The dataset is split by camera views, following a 3:1 train/test ratio.
- **S3 (unseen objects).** The dataset is split by objects. 7 objects that collectively represent all 28 grasps are selected as the test set, while the other 23 objects are used for training.
- **S4 (unseen taxonomy).** The dataset is divided by the grasping taxonomy. All the *intermediate* grasp types in Fig. 3 are selected as the test set, while others are used for training.

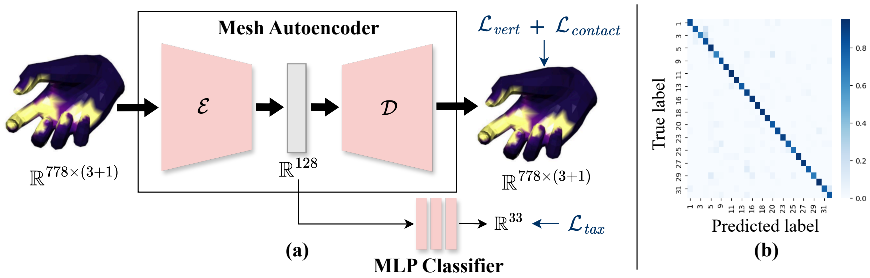

Fig. 9: (left) The network architecture for grasp classification, (right) confusion matrix.

Note that we will release the exact split configurations through code. Also, refer to the supplementary for the benchmarking hand-object reconstruction results for each split.

4.2 Grasp Classification

We evaluate grasp classification performance on HOGraspNet using our S0 (Section 4.1) benchmark setup. For the classification network, we modify the existing convolutional mesh autoencoder (CoMA [67]) to take as input a hand mesh with per-vertex contact value as an additional vertex feature. The bottleneck feature of the autoencoder is fed to an MLP-based classifier to predict a grasp type. To train our model, we use L1 loss ($\mathcal{L}_{vert}$) that learns vertex reconstruction, and two cross-entropy losses that learn grasp taxonomy classification ($\mathcal{L}_{tax}$) and contact classification [27] ($\mathcal{L}_{contact}$), where the range of contact value [0,1] is split into 10 bins. Our overall network architecture is shown in Fig. 9. Note that we utilize auxiliary reconstruction losses for the classification task to obtain richer grasp features, which are also utilized for t-SNE visualization in Sec. 3.4. Our model achieves 0.95 in f1 score for contact map reconstruction and 0.88 in accuracy for taxonomy classification. This experimental validation demonstrates that our grasping taxonomy can be delineated using hand meshes with contact maps without considering an object as input, indicating that our grasp annotation is generalized well across the samples.

4.3 Hand-Object Pose Estimation

In this section, we present the benchmarking results on hand-object pose estimation on HOGraspNet using S0 split. We use HFL-Net [49] as a baseline, as it is the current state-of-the-art network on hand-object reconstruction. HFL-Net jointly estimates a hand mesh and 6D pose of an object from the input image via attention modules (please refer to [49] and Sec. 2 for more details).

In Fig. 10, we visualize the hand pose estimation results in PA-MPJPE for each grasp classes (left). The baseline achieves high accuracy across all grasp types with mean PA-MPJPE value of 5.67mm, which is comparable to the state-of-the-art hand pose estimation results on other widely-used hand-object

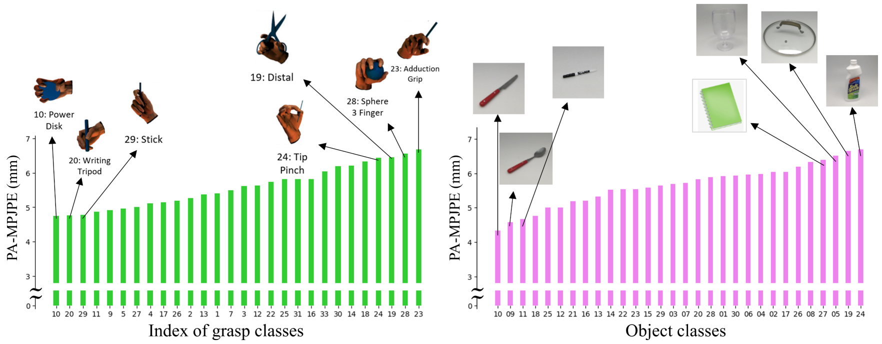

Fig. 10: Hand pose estimation results in PA-MPJPE (mm) (left) per grasp class and (right) object class.

Table 3: 6D object pose estimation results in ADD-0.1D per object class using HOGraspNet.

	ADD-0.1D(↑)		ADD-0.1D(↑)
1: cracker_box	88.79	16: golf_ball	46.27
2: potted_meat_can	59.47	17: credit_card	34.06
3: banana	58.21	18: dice	2.44
4: apple	75.74	19: disk_lid	99.29
5: wine_glass	97.13	20: smartphone	52.22
6: bowl	94.64	21: mouse	41.40
7: mug	72.04	22: tape	62.34
8: plate	99.42	23: master_chef_can	88.75
9: spoon	50.60	24: scrub_cleanser_bottle	89.80
10: knife	38.17	25: large_marker	34.59
11: small_marker	28.29	26: stapler	61.40
12: spatula	67.16	27: note	88.70
13: flat_screwdriver	63.52	28: scissors	54.34
14: hammer	82.99	29: foldable_phone	25.02
15: baseball	73.72	30: cardboard_box	77.80
Avg	63.61		

datasets [22, 66, 83, 93]. This further verifies that the quality of our hand annotation is decent, which allows for effective learning of the downstream hand pose-related task. We found that three of the top four grasp classes with the highest errors were not included in the DexYCB [8] and HO3D [28] datasets, respectively. This implies that our dataset's newly introduced real grasp poses might be challenging for the hand pose estimation model trained on existing datasets. The supplementary materials provide details of the missing grasp classes per dataset. In Fig. 10 (right), we also shows the hand pose metric per object classes. As expected, larger objects with more occlusion showed higher errors. Tab. 3 shows the object pose estimation results in ADD-0.1D per object class. We again achieve reasonable results that are comparable to the state-of-the-art object pose estimation performance on the other datasets [8, 49], except for *Dice* class. As the *Dice* has small visible regions due to hand-object occlusions, increasing the ill-posedness of the pose estimation task. Additional results using other split protocols (S1-S4) can be found in the supplementary material.

Table 4: Cross-benchmark results on hand pose estimation using HFL-Net [49].

Train Set	Test Set	MPJPE (mm)	PA-MPJPE (mm)
HO3D [28]	DexYCB [8]	57.31	10.31
HOGraspNet	DexYCB [8]	**42.65**	**9.36**

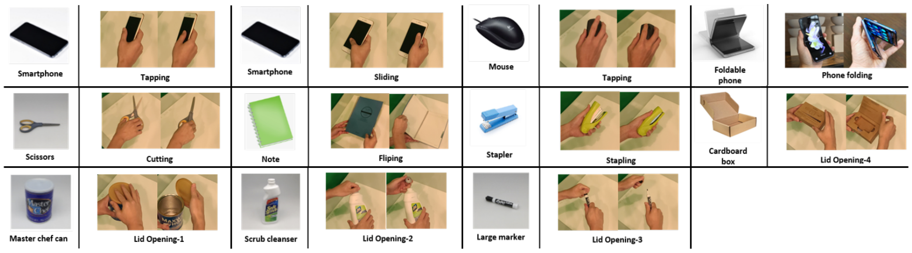

Fig. 11: Non-grasping action sequences in our dataset.

4.4 Cross-benchmark results on Hand Pose Estimation

We additionally report the cross-validation results on hand pose estimation, following the experimental setup used in [84]. Since HFL-Net [49] is an object-aware network, we conducted experiments on samples with object classes that mutually exist in all the datasets to perform fair comparisons. In Tab. 4, the network trained on HOGraspNet achieves better estimation accuracy than the network trained on HO3D [28], indicating the comprehensiveness of HOGraspNet.

5 Conclusions

We have proposed a real RGB-D dataset, HOGraspNet, featuring comprehensive grasp labels. We have also presented the experimental results on grasp classification, and hand-object pose estimation. Our dataset captures diverse hand-object interactions involving 30 objects, 99 participants, and 90 interaction scenarios. It includes MANO [68] and HALO [41] 3D hand meshes, 3D keypoints, object meshes, contact maps, and grasp annotations for every sequence. The benchmark notably improves accuracy across datasets by a broader range of interaction scenarios compared to the existing datasets.

Limitations and Future Work. We aimed to incorporate various compound and articulated objects to capture dynamic actions. However, we currently treat them as rigid objects. Nevertheless, interaction actions like tapping, folding, and opening have already been recorded, as illustrated in Fig. 11. We plan to update the dataset with the object articulation annotations in the future. Furthermore, the dataset can be improved by including non-grasping actions such as pushing, throwing, squeezing, or deforming non-rigid objects like plastic bottles and sponges, which will be addressed as our future work.

Acknowledgement This work was in part sponsored by NST grant (CRC 21011, MSIT), IITP grant (No.2019-0-01270 and RS-2023-00228996, MSIT).

References

1. Arapi, V., Della Santina, C., Averta, G., Bicchi, A., Bianchi, M.: Understanding human manipulation with the environment: a novel taxonomy for video labelling. IEEE Robotics and Automation Letters **6**(4), 6537–6544 (2021) 6
2. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: CVPR (2022) 5
3. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: CVPR (2019) 3
4. Brahmbhatt, S., Tang, C., Twigg, C.D., Kemp, C.C., Hays, J.: Contactpose: A dataset of grasps with object contact and hand pose. In: ECCV (2020) 1, 3
5. Calli, B., Singh, A., Walsman, A., Srinivasa, S., Abbeel, P., Dollar, A.M.: The ycb object and model set: Towards common benchmarks for manipulation research. In: ICAR (2015) 2, 6
6. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021) 3, 5, 6
7. Caramalau, R., Bhattarai, B., Kim, T.K.: Active learning for bayesian 3d hand pose estimation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3419–3428 (2021) 3
8. Chao, Y.W., Yang, W., Xiang, Y., Molchanov, P., Handa, A., Tremblay, J., Narang, Y.S., Van Wyk, K., Iqbal, U., Birchfield, S., et al.: Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR (2021) 1, 2, 3, 4, 8, 9, 13, 14
9. Chen, L.C., Papandreou, G., Schroff, F., Adam, H.: Rethinking atrous convolution for semantic image segmentation. arXiv preprint arXiv:1706.05587 (2017) 8
10. Chen, Y., Tu, Z., Kang, D., Chen, R., Bao, L., Zhang, Z., Yuan, J.: Joint hand-object 3d reconstruction from a single image with cross-branch feature fusion. TIP (2021) 5
11. Chen, Z., Chen, S., Schmid, C., Laptev, I.: gsdf: Geometry-driven signed distance functions for 3d hand-object reconstruction. In: CVPR (2023) 5, 10, 11
12. Chen, Z., Hasson, Y., Schmid, C., Laptev, I.: Alignsdf: Pose-aligned signed distance fields for hand-object reconstruction. In: ECCV (2022) 5
13. Cho, W., Park, G., Woo, W.: Tracking an object-grabbing hand using occluded depth reconstruction. In: ISMAR-Adjunct (2018) 5
14. Cho, W., Park, G., Woo, W.: Bare-hand depth inpainting for 3d tracking of hand interacting with object. In: ISMAR (2020) 5
15. Cini, F., Ortenzi, V., Corke, P., Controzzi, M.: On the choice of grasp type and location when handing over an object. Science Robotics **4**(27), eaau9757 (2019) 6
16. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: CVPR (2020) 1, 2, 3, 4
17. Damen, D., Doughty, H., Farinella, G.M., Furnari, A., Kazakos, E., Ma, J., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al.: Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV (2022) 3
18. Doosti, B., Naha, S., Mirbagheri, M., Crandall, D.J.: Hope-net: A graph-based model for hand-object pose estimation. In: CVPR (2020) 5
19. Fan, Z., Taheri, O., Tzionas, D., Kocabas, M., Kaufmann, M., Black, M.J., Hilliges, O.: ARCTIC: A dataset for dexterous bimanual hand-object manipulation. In: CVPR (2023) 1, 3, 4, 5, 7

20. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on human-machine systems **46**(1), 66–77 (2015) 2, 3, 4, 6
21. Fieraru, M., Zanfir, M., Oneata, E., Popa, A.I., Olaru, V., Sminchisescu, C.: Three-dimensional reconstruction of human interactions. In: CVPR (2020) 5
22. Fu, Q., Liu, X., Xu, R., Niebles, J.C., Kitani, K.M.: Deformer: Dynamic fusion transformer for robust hand pose estimation. arXiv preprint arXiv:2303.04991 (2023) 13
23. Garcia-Hernando, G., Johns, E., Kim, T.K.: Physics-based dexterous manipulations with estimated hand poses and residual reinforcement learning. In: IROS (2020) 3
24. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: CVPR (2018) 1, 3
25. Gomez-Donoso, F., Orts-Escolano, S., Cazorla, M.: Large-scale multiview 3d hand pose dataset. IVC (2019) 1, 3
26. Goyal, M., Modi, S., Goyal, R., Gupta, S.: Human hands as probes for interactive object understanding. In: CVPR (2022) 4
27. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmbhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: CVPR (2021) 5, 12
28. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020) 1, 2, 3, 4, 8, 10, 13, 14
29. Hampali, S., Sarkar, S.D., Rad, M., Lepetit, V.: Keypoint transformer: Solving joint identification in challenging hands and object interactions for accurate 3d pose estimation. In: CVPR (2022) 5
30. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constraints. In: ICCV (2019) 5
31. Hasson, Y., Tekin, B., Bogo, F., Laptev, I., Pollefeys, M., Schmid, C.: Leveraging photometric consistency over time for sparsely supervised hand-object reconstruction. In: CVPR (2020) 5
32. Hasson, Y., Varol, G., Schmid, C., Laptev, I.: Towards unconstrained joint hand-object reconstruction from rgb videos. In: 3DV (2021) 5
33. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11807–11816 (2019) 1, 3, 4, 5
34. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: CVPR (2019) 3
35. Hu, H., Yi, X., Zhang, H., Yong, J.H., Xu, F.: Physical interaction: Reconstructing hand-object interactions with physics. In: SIGGRAPH Asia (2022) 5
36. Huang, C.H.P., Yi, H., Höschle, M., Safroshkin, M., Alexiadis, T., Polikovsky, S., Scharstein, D., Black, M.J.: Capturing and inferring dense full-body human-scene contact. In: CVPR (2022) 5
37. Huang, Y., Taheri, O., Black, M.J., Tzionas, D.: InterCap: Joint markerless 3D tracking of humans and objects in interaction from multi-view RGB-D images. IJCV (2024) 5
38. Jiang, N., Liu, T., Cao, Z., Cui, J., Zhang, Z., Wang, H., Zhu, Y., Huang, S.: Full-body articulated human-object interaction. In: ICCV (2023) 5

39. Joo, H., Liu, H., Tan, L., Gui, L., Nabbe, B., Matthews, I., Kanade, T., Nobuhara, S., Sheikh, Y.: Panoptic studio: A massively multiview system for social motion capture. In: ICCV (2015) 1, 5
40. Joo, H., Neverova, N., Vedaldi, A.: Exemplar fine-tuning for 3d human pose fitting towards in-the-wild 3d human pose estimation. In: 3DV (2020) 1, 5
41. Karunratanakul, K., Spurr, A., Fan, Z., Hilliges, O., Tang, S.: A skeleton-driven neural occupancy representation for articulated hands. In: 3DV (2021) 2, 4, 6, 10, 11, 14
42. Kwon, T., Tekin, B., Stühmer, J., Bogo, F., Pollefeys, M.: H2o: Two hands manipulating objects for first person interaction recognition. In: ICCV (2021) 3
43. Lee, J., Saito, S., Nam, G., Sung, M., Kim, T.K.: Interhandgen: Two-hand interaction generation via cascaded reverse diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 527–537 (2024) 4
44. Lee, J., Sung, M., Choi, H., Kim, T.K.: Im2hands: Learning attentive implicit representation of interacting two-hand shapes. In: CVPR (2023) 4
45. Leroy, V., Weinzaepfel, P., Brégier, R., Combaluzier, H., Rogez, G.: Smply benchmarking 3d human pose estimation in the wild. In: 3DV (2020) 5
46. Li, M., An, L., Zhang, H., Wu, L., Chen, F., Yu, T., Liu, Y.: Interacting attention graph for single image two-hand reconstruction. In: CVPR (2022) 4
47. Lin, K., Wang, L., Liu, Z.: Mesh graphormer. In: ICCV (2021) 3
48. Lin, P., Xu, S., Yang, H., Liu, Y., Chen, X., Wang, J., Yu, J., Xu, L.: Handdiffuse: Generative controllers for two-hand interactions via diffusion models. In: CoRR. vol. abs/2312.04867 (2023) 3, 4
49. Lin, Z., Ding, C., Yao, H., Kuang, Z., Huang, S.: Harmonious feature learning for interactive hand-object pose estimation. In: CVPR (2023) 5, 12, 13, 14
50. Liu, J., Feng, F., Nakamura, Y.C., Pollard, N.S.: A taxonomy of everyday grasps in action. In: 2014 IEEE-RAS International Conference on Humanoid Robots. pp. 573–580. IEEE (2014) 4, 6
51. Liu, S., Jiang, H., Xu, J., Liu, S., Wang, X.: Semi-supervised 3d hand-object poses estimation with interactions in time. In: CVPR (2021) 5
52. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: SMPL: A skinned multi-person linear model. ACM TOG (2015) 5
53. Lugaresi, C., Tang, J., Nash, H., McClanahan, C., Uboweja, E., Hays, M., Zhang, F., Chang, C.L., Yong, M.G., Lee, J., et al.: Mediapipe: A framework for building perception pipelines. arXiv preprint arXiv:1906.08172 (2019) 9
54. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. J. Mach. Learn. Res. (2008) 8
55. Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Sridhar, S., Pons-Moll, G., Theobalt, C.: Single-shot multi-person 3d pose estimation from monocular rgb. In: 3DV (2018) 1, 5
56. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: Learning 3d reconstruction in function space. In: CVPR (2019) 10
57. Moon, G., Saito, S., Xu, W., Joshi, R., Buffalini, J., Bellan, H., Rosen, N., Richardson, J., Mize, M., De Bree, P., et al.: A dataset of relighted 3d interacting hands. NeurIPS (2024) 3, 4
58. Moon, G., Yu, S.I., Wen, H., Shiratori, T., Lee, K.M.: Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: ECCV (2020) 1, 3, 4
59. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: ICCV (2017) 1, 3

60. Park, G., Kim, T.K., Woo, W.: 3d hand pose estimation with a single infrared camera via domain transfer learning. In: ISMAR (2020) 3
61. Patel, P., Huang, C.H.P., Tesch, J., Hoffmann, D.T., Tripathi, S., Black, M.J.: AGORA: Avatars in geography optimized for regression analysis. In: CVPR (2021) 5
62. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: CVPR (2019) 5
63. Pavlakos, G., Shan, D., Radosavovic, I., Kanazawa, A., Fouhey, D., Malik, J.: Reconstructing hands in 3d with transformers. In: CVPR (2024) 3
64. Pumarola, A., Sanchez, J., Choi, G., Sanfeliu, A., Moreno-Noguer, F.: 3DPeople: Modeling the Geometry of Dressed Humans. In: ICCV (2019) 5
65. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: CVPR (2014) 1, 3
66. Qu, W., Cui, Z., Zhang, Y., Meng, C., Ma, C., Deng, X., Wang, H.: Novel-view synthesis and pose estimation for hand-object interaction from sparse views. In: ICCV (2023) 13
67. Ranjan, A., Bolkart, T., Sanyal, S., Black, M.J.: Generating 3d faces using convolutional mesh autoencoders. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018) 12
68. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM TOG (2017) 2, 3, 5, 6, 8, 9, 10, 11, 14
69. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017) 1, 3
70. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using rgb and depth data. In: ICCV (2013) 1, 3
71. Stival, F., Michieletto, S., Cognolato, M., Pagello, E., Müller, H., Atzori, M.: A quantitative taxonomy of human hand grasps. Journal of neuroengineering and rehabilitation **16**, 1–17 (2019) 6
72. Sun, Y., Liu, W., Bao, Q., Fu, Y., Mei, T., Black, M.J.: Putting People in their Place: Monocular Regression of 3D People in Depth. In: CVPR (2022) 1, 5
73. Swamy, A., Leroy, V., Weinzaepfel, P., Baradel, F., Galaaoui, S., Brégier, R., Armando, M., Franco, J.S., Rogez, G.: Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In: ICCV (2023) 3, 4
74. Taheri, O., Ghorbani, N., Black, M.J., Tzionas, D.: Grab: A dataset of whole-body human grasping of objects. In: ECCV 2020 (2020) 1, 3, 5
75. Tang, D., Jin Chang, H., Tejani, A., Kim, T.K.: Latent regression forest: Structured estimation of 3d articulated hand posture. In: CVPR (2014) 1, 3
76. Tekin, B., Bogo, F., Pollefeys, M.: H+o: Unified egocentric recognition of 3d hand-object poses and interactions. In: CVPR (2019) 5
77. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM TOG (2014) 1, 3
78. Tse, T.H.E., Zhang, Z., Kim, K.I., Leonardis, A., Zheng, F., Chang, H.J.: S2 contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In: ECCV (2022) 5
79. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. IJCV (2016) 3, 4
80. Wang, J., Mueller, F., Bernard, F., Sorli, S., Sotnychenko, O., Qian, N., Otaduy, M.A., Casas, D., Theobalt, C.: Rgb2hands: real-time tracking of 3d hand interactions from monocular rgb video. ACM TOG (2020) 4

81. Wen, G., Xiaoyu, B., Xavier, A.P., Francesc, M.N.: Multi-person extreme motion prediction. CVPR (2022) 1, 5
82. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: ICCV (2013) 1, 3
83. Xu, H., Wang, T., Tang, X., Fu, C.W.: H2onet: Hand-occlusion-and-orientation-aware network for real-time 3d hand mesh reconstruction. In: CVPR (2023) 13
84. Yang, L., Li, K., Zhan, X., Wu, F., Xu, A., Liu, L., Lu, C.: Oakink: A large-scale knowledge repository for understanding hand-object interaction. In: CVPR (2022) 3, 4, 6, 8, 14
85. Yang, L., Zhan, X., Li, K., Xu, W., Li, J., Lu, C.: Cpf: Learning a contact potential field to model the hand-object interaction. In: ICCV (2021) 5
86. Yin, Y., Guo, C., Kaufmann, M., Zarate, J., Song, J., Hilliges, O.: Hi4d: 4d instance segmentation of close human interaction. In: CVPR (2023) 5
87. Yu, Z., Yang, L., Chen, S., Yao, A.: Local and global point cloud reconstruction for 3d hand pose estimation. In: BMVC (2021) 1, 3
88. Yuan, S., Ye, Q., Stenger, B., Jain, S., Kim, T.K.: Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. In: CVPR (2017) 1, 3
89. Zhang, B., Wang, Y., Deng, X., Zhang, Y., Tan, P., Ma, C., Wang, H.: Interacting two-hand 3d pose and shape reconstruction from single color image. In: ICCV (2021) 4
90. Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., Yang, Q.: 3d hand pose tracking and estimation using stereo matching. In: ICIP (2017) 1, 3
91. Zhang, S., Ma, Q., Zhang, Y., Qian, Z., Kwon, T., Pollefeys, M., Bogo, F., Tang, S.: Egobody: Human body shape and motion of interacting people from head-mounted devices. In: ECCV (2022) 5
92. Zhang, X., Huang, H., Tan, J., Xu, H., Yang, C., Peng, G., Wang, L., Liu, J.: Hand image understanding via deep multi-task learning. In: ICCV (2021) 1, 3
93. Zheng, X., Wen, C., Xue, Z., Ren, P., Wang, J.: Hamuco: Hand pose estimation via multiview collaborative self-supervised learning. In: ICCV (2023) 13
94. Zheng, Y., Shao, R., Zhang, Y., Yu, T., Zheng, Z., Dai, Q., Liu, Y.: Deepmulticap: Performance capture of multiple characters using sparse multiview cameras. In: ICCV (2021) 1, 5
95. Zimmermann, C., Argus, M., Brox, T.: Contrastive representation learning for hand shape estimation. In: GCPR (2021) 1, 3
96. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single rgb images. In: ICCV (2017) 1, 3
97. Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from single rgb images. In: ICCV (2019) 1, 3, 9, 10
98. Zuo, B., Zhao, Z., Sun, W., Xie, W., Xue, Z., Wang, Y.: Reconstructing interacting hands with interaction prior from monocular images. In: ICCV (2023) 3, 4