

Supplementary Materials

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

Woojin Cho¹, Jihyun Lee¹, Minjae Yi¹, Minje Kim¹, Taeyun Woo¹,
Donghwan Kim¹, Taewook Ha¹, Hyokeun Lee³, Je-Hwan Ryu⁴,
Woontack Woo¹, Tae-Kyun Kim^{1,2}

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

1 Additional Evaluation

Here, we present additional experimental results with HOGraspNet that are not included in the main paper. We provide detailed descriptions of the evaluation setups and the quantitative results across each split type we defined. Additionally, we show qualitative comparisons between HOGraspNet and the publicly available datasets [19,17,4].

1.1 Evaluation Setup

Split Protocol. The S0 split follows the traditional sequence split, where the first trial of each subject and grasp scenario is assigned to the test set. For the S1 split, the 99 subjects are randomly divided into train and test sets at a ratio of roughly 7:3. For the S2 split, the data from the four camera views are divided at a ratio of 3:1. Each view corresponds to a different side of the subjects (back, left, front, and right), and the right-side camera is selected as the test set. In the S3 split, 30 objects are divided into train and test sets at a ratio of 23:7, ensuring that all grasp classes are represented in the train set. The cumulative grasp classes for the objects in both sets are shown in Fig. 1. Lastly, the S4 split separates the total of 28 grasp classes at a ratio of 6:1, with four classes identified as *Intermediate* (please refer to Fig. 11) assigned to the test set.

Baseline Network. We report the hand reconstruction results using HFL-net [12] trained on each split. We train each network for 5 epochs using the Adam optimizer with a learning rate of 1e-4 and a decaying gamma of 0.9 per 2 epochs, while other parameters are set to the default values.

1.2 Evaluation Results

Hand Pose Estimation. We summarize the hand estimation results in Fig. 2. The S0 split exhibits the highest performance, while the S2 split demonstrates the lowest, with S4 showing the second lowest performance. As the S0 partitions the dataset only by sequences, it generally shares common pose spaces between

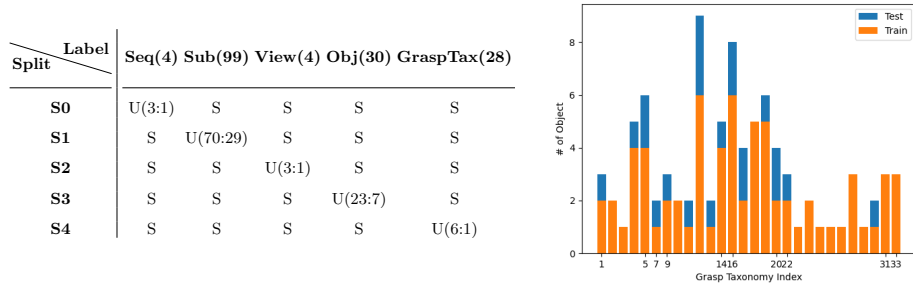


Fig. 1. (left) Detailed experimental protocols (train:test splits). **U** denotes unseen factor, **S** denotes seen factors. We consider sequence, subject, view, object, and grasp taxonomy as split factors. **(right) Distribution of grasp taxonomy in S3.** All grasp classes are present in S3 (unseen object) train set.

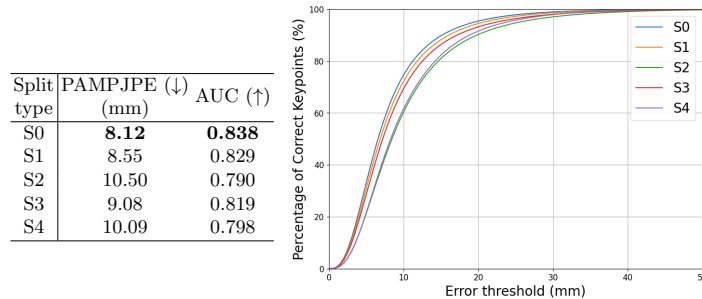


Fig. 2. (left) 3D hand pose estimation results, (right) PCK curve for each split type.

the train and test sets, which results in the lowest pose error. Conversely, the S2 split with viewpoint exhibits the lowest performance. As our dataset contains 4 sparse viewpoints (back, left, front, and right) following the existing literature [10], the RGB frames among different viewpoints may exhibit high disparity, potentially making the network generalization difficult. Thus, we have anticipated that S2 would be the most challenging split due to the unseen camera view in the test set. The second lowest performance is shown by the S3 split categorized by grasp taxonomy. As mentioned earlier, there are distinct variations in hand poses among grasp taxonomies, which results in inconsistencies in the performance of the trained model on the test set. On the other hand, in the S1 split with unseen subjects, shape variations exist between the train and test sets. However, since the same grasp guidelines were provided to all subjects, differences in hand poses among subjects are relatively small, resulting in relatively good joint pose estimation performance. The S3 split for unseen objects, which includes all grasps but not all objects, shows relatively lower performance due to

Table 1. 6D object pose estimation results in ADD-0.1D(\uparrow) per object. We used HFL-Net [12] trained on each split.

	S0	S1	S2	S3	S4	avg		S0	S1	S2	S3	S4	avg
1: cracker_box	51.46	39.55	22.80	-	-	37.94	16: golf_ball	1.29	1.47	0.06	-	-	0.94
2: potted_meat_can	19.19	12.95	8.18	0.07	-	10.10	17: credit_card	2.15	3.31	1.17	-	1.97	2.15
3: banana	14.43	20.12	0.55	-	-	11.70	18: dice	0.0	0.02	0.04	-	0.04	0.03
4: apple	11.36	14.57	1.40	0.24	-	6.89	19: disk_lid	30.61	32.12	1.98	-	-	21.64
5: wine_glass	22.96	15.64	1.76	-	-	13.46	20: smartphone	12.81	16.40	2.75	0.06	9.69	7.93
6: bowl	26.78	19.43	6.34	0.03	22.82	14.07	21: mouse	12.13	12.16	2.42	-	-	8.90
7: mug	25.23	21.36	3.60	-	-	16.73	22: tape	9.61	1.75	7.78	-	5.04	6.05
8: plate	23.76	29.73	1.52	-	25.81	20.21	23: master_chef_can	29.51	28.15	16.87	-	-	24.85
9: spoon	15.64	19.07	12.87	-	16.70	16.08	24: scrub_cleanser_bottle	52.61	40.23	42.35	-	-	45.07
10: knife	12.92	10.97	3.95	5.31	-	8.29	25: large_marker	7.65	5.05	3.19	-	-	5.30
11: small_marker	6.17	6.91	3.53	1.39	-	4.50	26: stapler	26.57	26.97	6.80	-	-	20.12
12: spatula	2.15	20.64	0.94	0.04	-	12.10	27: note	43.62	61.99	0.69	-	41.92	37.06
13: flat_screwdriver	13.99	12.08	0.46	-	-	8.85	28: scissors	20.87	14.66	1.26	-	7.40	11.05
14: hammer	28.24	30.79	1.98	-	-	20.34	29: foldable_phone	8.68	8.00	0.78	-	6.28	5.94
15: baseball	7.24	5.77	2.75	-	-	5.56	30: cardboard_box	6.44	5.14	1.55	-	-	4.38
Avg		S0: 18.86		S1: 18.10		S2: 5.15	S3: 1.02		S4: 13.76				

the characteristics of the HFL baseline model, which estimates hand pose given the object information.

Object Pose Estimation. We show the object pose estimation results in Tab. 1. Recall that S3 is an unseen object split and S4 is an unseen grasp taxonomy split. However, S4 does not consist of all objects in the training set due to the fixed grasps per object, leading to several invalid object poses in both splits (denoted as -). Overall, the baseline achieves the best object pose estimation performance on the S0 split, which has a minimal disparity between the train and test sets. Notably, small objects like dice, golf balls, and credit cards, which are mostly absent in existing datasets, show the lowest performance due to significant occlusion caused by the interaction with the hand. Furthermore, objects with intricate textures, such as cracker boxes and scrub cleanser bottles, generally yield high accuracy due to providing reliable cues for the network model. Unlike the object pose estimation experiment in the main paper (Sec. 4.3), this experiment does not account for the object’s symmetry during the metric calculation, resulting in lower accuracy for symmetric objects like markers, cardboard boxes, and tape.

Qualitative Comparison Between Datasets. Fig. 3 shows the qualitative results of randomly sampled data from each recent dataset. Minor discrepancies between real and rendered meshes are found across examined datasets, with overall annotation quality being supposed to be consistent. Additionally, mesh discrepancies in the wrist area are observed in several samples, likely resulting from inherent limitations in the representation of the MANO [16] hand model.

Refinement results on noisy hand-object contact maps and poses Regarding in-the-wild experiments, recent works train a generative prior on 3D shapes and use it to regularize the plausibility of poses estimated from in-the-wild images, achieving new state-of-the-art accuracy [20,11]. The diversity in the training 3D shape set is the key to building a strong prior. We thus believe that HOGraspNet can play an important role in this direction. Although this



Fig. 3. Qualitative comparison of random samples of different datasets. (a) HOGraspNet (ours). (b) OakInk [19]. (c) SHOWMe [17]. (d) DexYCB [4].

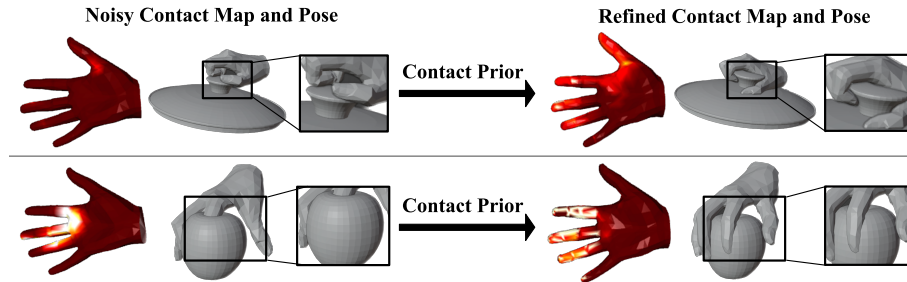


Fig. 4. Refinement results of our prior model trained on HOGraspNet on noisy hand-object contact maps and poses.

was not the original scope of our dataset collection work, we report the preliminary results following the experimental setup in [9], where we trained a prior model of hand-object shapes represented as contact maps and used it to refine the noisy hand-object poses (e.g., estimated from images). Fig. 4 shows that the prior model trained on HOGraspNet can successfully refine the noisy contacts and enforces more plausible pose estimation. We observed that incorporating our per-sample grasp type (which is not provided by most of the existing datasets) further improves the performance of the prior model, increases the hand and object contact F1 scores by 6% and 5%, respectively. We will further investigate this direction as future work.

2 Additional Survey on Interaction Datasets

In this section, we provide more details on the existing hand-object interaction datasets discussed in Sec. 2 in the main paper. We then report the missing grasp types in the existing datasets [10,4,3] to show that ours capture more comprehensive grasps.

More Details on Hand-Object Datasets. **FPHA** [8] captures first-person dynamic hand actions interacting with 3D objects. It consists of over 100K RGB-D frames of 1.1K action samples across 45 categories by manipulating 26 objects. The hand joint is annotated using six magnetic sensors and inverse kinematics. However, the FPHA does not fully provide the 6D object pose, and the presence of magnetic sensors on the hand corrupts the RGB-D images. **YCB-Affordance** [5] consists of 133K annotated frames featuring over 28M synthetic grasps to depict diverse human grasp affordances. It has manual annotations of 367 different hand-object poses, according to the 33-grasp taxonomy [7]. These hand poses are rendered upon the YCB-Video [18] dataset. However, YCB-Affordance captures less diverse grasp configurations since the hand poses are replicated through the rotation symmetry of the object. Also, the synthetic images may lead to less effective network training due to the discrepancy between the synthetic and real image domains. **ContactDB** [1] captures detailed hand-object contact of grasping using a thermal camera. It consists of 3.7K 3D meshes of 50 household objects textured with contact maps and 375K frames of synchronized RGB-D+thermal images. **ContactPose** [2] expands ContactDB by including 3D joint locations and multi-view RGB-D grasp images. It has 2.3K unique grasps of 25 household objects by 50 participants and more than 2.9M RGB-D grasp images. However, ContactPose does not support dynamic grasp, where the thermal camera can only capture one contact map for each interaction sequence. **EPIC-KITCHEN** [6] is a large-scale egocentric hand-object dataset. It captures the kitchen activities of 32 identities of 10 different nationalities. However, it does not provide mesh annotations; thus, it cannot be used to train hand-object dense shape reconstruction networks. **HOI4D** [13] captures 4D egocentric category-level human-object interaction. HOI4D consists of 2.4M RGB-D egocentric video frames over 4K sequences collected by 4 participants interacting with 800 object instances from 16 categories over 610 indoor rooms. HOI4D provides annotations for frame-wise panoptic segmentation, motion segmentation, 3D hand pose, rigid and articulated object pose, and action segmentation. Although diverse objects were obtained, it is limited to the 16 types of object categories, which do not cover full taxonomies (e.g., grasps that can only be acquired through flat or small objects).

Missing Grasp Classes in Existing Datasets. In Fig. 5, we present a list of grasp classes that are absent in the existing datasets (MOW [3], HO3D [10], and DexYCB [4]), which we manually identified. This further demonstrates that our dataset more comprehensively captures hand grasps, covering more grasp types.

















































	Missing grasps of each dataset for 33 grasp taxonomy																		
Ours																			
MOW (ICCV21)																			
HO3D (CVPR20)																			
DexYCB (CVPR21)																			

Fig. 5. Missing grasp types in the existing datasets (MOW [3], HO3D [10], and DexYCB [4]).

3 Details of Annotation Procedure

In this section, we provide more details on our camera calibration (Sec. 3.1), segmentation (Sec. 3.2), and hand-object model fitting procedures (Sec. 3.3).

3.1 Camera Calibration

We use 4 RGB-D cameras to record the frames and 8 IR cameras to acquire object 6D poses. The RGB-D cameras and the IR cameras are temporally synchronized with electric signals, and we manually align the starting frames between them with a blinking LED. The transformations among the RGB-D cameras are calibrated with a checkerboard and a T-shaped wand with optical markers for IR cameras. To align coordinates between the RGB-D and IR cameras, three optical markers are placed on the checkerboard, which is positioned on the table. Although the capture system remains fixed, this calibration process is performed for each sequence.

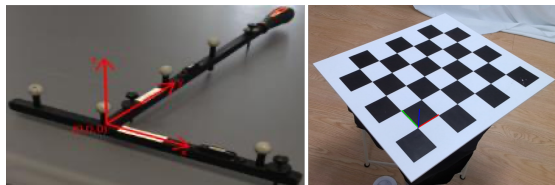


Fig. 6. (left) **T-shaped wand.** IR cameras capture the visible optical markers of a traversing wand. (right) **Checkerboard.** Optical markers are placed at the end of red and green lines.

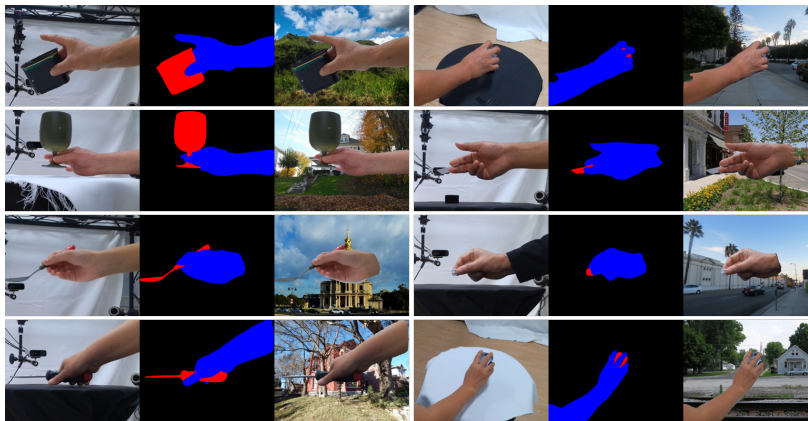


Fig. 7. Examples of our pseudo GT segmentation masks and background augmentation.

3.2 Segmentation

The quality of the pseudo-ground truth in our annotation pipeline is crucially tied to the final annotation quality. The hand and object masks were improved using a fine-tuned segmentation model, with examples in Fig. 7. These results demonstrate plausible masks for each object, achieved through tuning the model with our manually annotated segmentation masks.

3.3 MANO and Object Model Fitting

In this subsection, we provide more details on our MANO [16] and object model fitting process discussed in Sec. 3.5 in the main paper. Recall that our goal is to align MANO hand and object mesh models with RGB-D data captured from multiple cameras based on the initial poses of the hand and the object. Here, the MANO model is parameterized by pose parameter $\theta \in \mathbb{R}^{48}$ and shape parameter $\beta \in \mathbb{R}^{10}$. The object model is represented using the standard 6D pose representation $\phi \in \mathbb{R}^6$, consisting of a 3D rotation matrix and a 3D translation vector.

Multi-view Multi-frame Gradual Hand-Object Model Fitting. To annotate hand and object parameters, we formulate an optimization problem with the objective discussed in the main paper (Eq. 1). To avoid local minima, we gradually fit the partial hand pose parameters across several stages. Through preliminary experiments, we found that optimizing the hand pose parameters starting from the wrist outward yielded promising results. With this insight, we perform the optimization progressively in the subsequent order: (1) global hand orientation, (2) partial hand poses extended from the wrist, and (3) the full hand and object parameters.

Details on Loss Terms. We now provide the details on each loss term in Eq. 1 in the main paper. \mathcal{L}_h^{2D} is 2D hand joint loss, which is defined as the L2 distance between pseudo GT 2D joint j_i and 2D projection Π^c of the models 3D joint \tilde{j}_i^{3D} over camera view c with the visibility v_i :

$$\mathcal{L}_h^{2D} = \lambda_h^{2D} \sum_c \sum_{i=1}^{21} \left\| j_i^c - \Pi^c(\tilde{j}_i^{3D}) \right\|_2 \cdot v_i. \quad (1)$$

\mathcal{L}_o^{3D} is 3D object marker loss, which is also defined as the L2 distance between the 3D marker position m_n^{3D} and the corresponding vertex position of object model v_n :

$$\mathcal{L}_o^{3D} = \lambda_o^{3D} \cdot \sum_c \sum_{n=1}^{3\sim 5} \left\| v_n^c - \Pi^c(m_n^{3D}) \right\|_2. \quad (2)$$

\mathcal{L}_{seg} is segmentation loss computed for hand and object via the L1 distance between predicted mask M and the rendered mask from mesh model \tilde{M} over the camera view c :

$$\mathcal{L}_{seg} = \lambda_{seg} \cdot \sum_c \left\| M^c - \tilde{M}^c \right\|_1. \quad (3)$$

\mathcal{L}_{depth} is depth loss, which is also computed as the L1 distance between captured depth image D and the rendered depth map \tilde{D} but for both hand and object:

$$\mathcal{L}_{depth} = \lambda_{depth} \cdot \sum_c \left\| D^c - \tilde{D}^c \right\|_1. \quad (4)$$

Following [21], we also incorporate the pose and shape regularize \mathcal{L}_{reg} to the MANO model to regularize the hand model and additional L2 norm between current hand model parameter θ_t, β_t and previous parameter $\theta_{t-1}, \beta_{t-1}$ for temporal consistency:

$$\mathcal{L}_{reg} = \lambda_{pose} \cdot \left\| \tilde{\theta} \right\|_2 + \lambda_{shape} \cdot \left\| \tilde{\beta} \right\|_2 + \lambda_{temporal} \cdot (\|\theta_t - \theta_{t-1}\|_2 + \|\beta_t - \beta_{t-1}\|_2). \quad (5)$$

To additionally regularize the fitted hand and object meshes to be physically plausible, we incorporate a 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 [10]. Let $\mathbf{V}_h \in \mathbb{R}^{N_h \times 3}$ and $\mathbf{V}_o \in \mathbb{R}^{N_o \times 3}$ be the vertex matrices of hand and object, respectively. Then, the penetration loss \mathcal{L}_{pen} is defined as:

$$\mathcal{L}_{pen} = \sum_{i,j \in \mathcal{S}(\mathbf{V}_h, \mathbf{V}_o)} \max(-(\mathbf{n}_o^j)^\top \mathbf{V}_h^i - \mathbf{V}_o^j, 0), \quad (6)$$

where $\mathcal{S}(\cdot, \cdot)$ is a function that returns the hand vertex index $i \in \mathbb{N}$ and its nearest object index $j \in \mathbb{N}$ in the Euclidean space, and \mathbf{n}_o^j denotes a normal vector at j -th object vertex. Equation 6 projects a vector joining the nearest vertices of the hand and object onto the normal vector at the object to approximate the amount of penetration in a differentiable manner.

\mathcal{L}_{pen} enforces the penetrated hand and object surfaces to repel each other, however, it is also important to regularize the closely located hand and object surface regions to be in actual contact to model physically plausible grasps. We use contact loss $\mathcal{L}_{contact}$ that aims to minimize the distances between hand and object vertices below a distance threshold τ :

$$\mathcal{L}_{contact} = \sum_{i,j \in \mathcal{S}(\mathbf{V}_h^i, \mathbf{V}_o^j)} \|\mathbf{V}_h^i - \mathbf{V}_o^j\|_2, \quad \text{where } d(\mathbf{V}_h^i, \mathbf{V}_o^j) < \tau. \quad (7)$$

In the above equation, $d(\cdot, \cdot)$ denotes a distance function¹. Note that we set τ as 8mm in our experiments.

Finally, we implement the overall fitting process using PyTorch[14] to minimize:

$$\hat{\theta}, \hat{\beta}, \hat{\phi} = \underset{\tilde{\theta}, \tilde{\beta}, \tilde{\phi}}{\operatorname{argmin}} (\mathcal{L}(\tilde{\theta}, \tilde{\beta}, \tilde{\phi})). \quad (8)$$

4 Taxonomies and Statistics for HOGraspNet

In this section, we report additional statistics of the HOGraspNet and the details of the taxonomies. We organized the subject’s gender with a ratio of 50:49 for males to females. The age distribution is also evenly dispersed, with proportions of 23%, 26%, 25%, and 25% across 20-year intervals from age 0 to 80. Hand sizes, measured from the wrist to the tip of the middle finger, are distributed as shown in Fig. 8. Fig. 9 shows each 3 grasp types for each of the 30 types of objects, resulting in a total of 90 interaction scenarios. We also show the data samples from HOGraspNet for each combination in Fig. 10. Lastly, the hand grasping and object taxonomies in Fig. 4 in the paper are provided on a larger scale in Fig. 11 and Fig. 12.

¹ We use `point_to_mesh_distance` function in Pytorch3D [15].

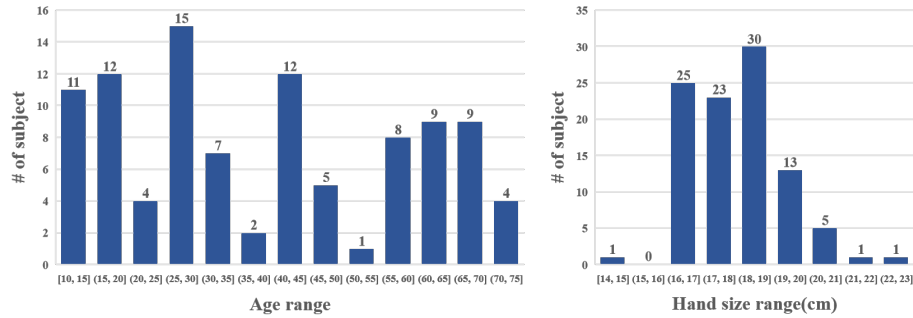


Fig. 8. Additional statistics of HOGraspNet

References

1. Brahmabhatt, S., Ham, C., Kemp, C.C., Hays, J.: Contactdb: Analyzing and predicting grasp contact via thermal imaging. In: CVPR (2019)
2. Brahmabhatt, 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)
3. Cao, Z., Radosavovic, I., Kanazawa, A., Malik, J.: Reconstructing hand-object interactions in the wild. In: ICCV (2021)
4. 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)
5. Corona, E., Pumarola, A., Alenya, G., Moreno-Noguer, F., Rogez, G.: Ganhand: Predicting human grasp affordances in multi-object scenes. In: CVPR (2020)
6. 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)
7. 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)
8. 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)
9. Grady, P., Tang, C., Twigg, C.D., Vo, M., Brahmabhatt, S., Kemp, C.C.: Contactopt: Optimizing contact to improve grasps. In: CVPR (2021)
10. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: CVPR (2020)
11. 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)
12. Lin, Z., Ding, C., Yao, H., Kuang, Z., Huang, S.: Harmonious feature learning for interactive hand-object pose estimation. In: CVPR (2023)
13. Liu, Y., Liu, Y., Jiang, C., Lyu, K., Wan, W., Shen, H., Liang, B., Fu, Z., Wang, H., Yi, L.: Hoi4d: A 4d egocentric dataset for category-level human-object interaction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21013–21022 (2022)

Object Name / Image	Taxonomies per Object			Object Name / Image	Taxonomies per Object				
1: cracker_box		13: Precision Sphere	19: Distal	22: Parallel Extension	16: golfball		14: Tripod	27: Quadpod	33: Inferior Pincer
2: potted_meat_can		1: Large Diameter	12: Precision Disk	19: Distal	17: credit card		16: Lateral	23: Addaction Grp	24: Tip Pinch
3: banana		3: Medium Wrap	4: Adducted Thumb	14: Tripod	18: dice		9: Palmar Pinch	14: Tripod	16: Lateral
4: apple		11: Power Sphere	13: Precision Sphere	14: Tripod	19: disk lid		10: Power Disk	18: Extension Type	28: Sphere 3 Finger
5: wine_glass		7: Prismatic 3 Finger	10: Power Disk	24: Tip Pinch	20: smartphone		12: Precision Disk	16: Lateral	22: Parallel Extension
6: bowl		12: Precision Disk	16: Lateral	30: Palmar	21: mouse		12: Precision Disk	19: Distal	26: Sphere 4 Finger
7: mug		5: Light Tool	12: Precision Disk	31: Ring	22: tape		5: Light Tool	12: Precision Disk	25: Lateral Tripod
8: plate		16: Lateral	18: Extension Type	30: Palmar	23: master_chef_can		1: Large Diameter	12: Precision Disk	31: Ring
9: spoon		5: Light Tool	20: Writing Tripod	29: Stuck	24: scrub_cleanser_bottle		1: Large Diameter	19: Distal	31: Ring
10: knife		5: Light Tool	17: IndexFinger Extension	20: Writing Tripod	25: large_marker		5: Light Tool	9: Palmar Pinch	20: Writing Tripod
11: small marker		5: Light Tool	7: Prismatic 3 Finger	9: Palmar Pinch	26: stapler		4: Adducted Thumb	17: Index Finger Extension	33: Inferior Pincer
12: spatula		4: Adducted Thumb	17: IndexFinger Extension	20: Writing Tripod	27: note		16: Lateral	18: Extension Type	22: Parallel Extension
13: flat_screwdriver		2: Small Diameter	4: Adducted Thumb	14: Tripod	28: scissors		12: Precision Disk	16: Lateral	19: Distal
14: hammer		2: Small Diameter	4: Adducted Thumb	17: IndexFinger Extension	29: foldable phone		12: Precision Disk	16: Lateral	18: Extension Type
15: baseball		11: Power Sphere	28: Sphere 3 Finger	33: Inferior Pincer	30: cardboard box		18: Extension Type	19: Distal	28: Sphere 3 Finger

Fig. 9. Grasp classes for each object.

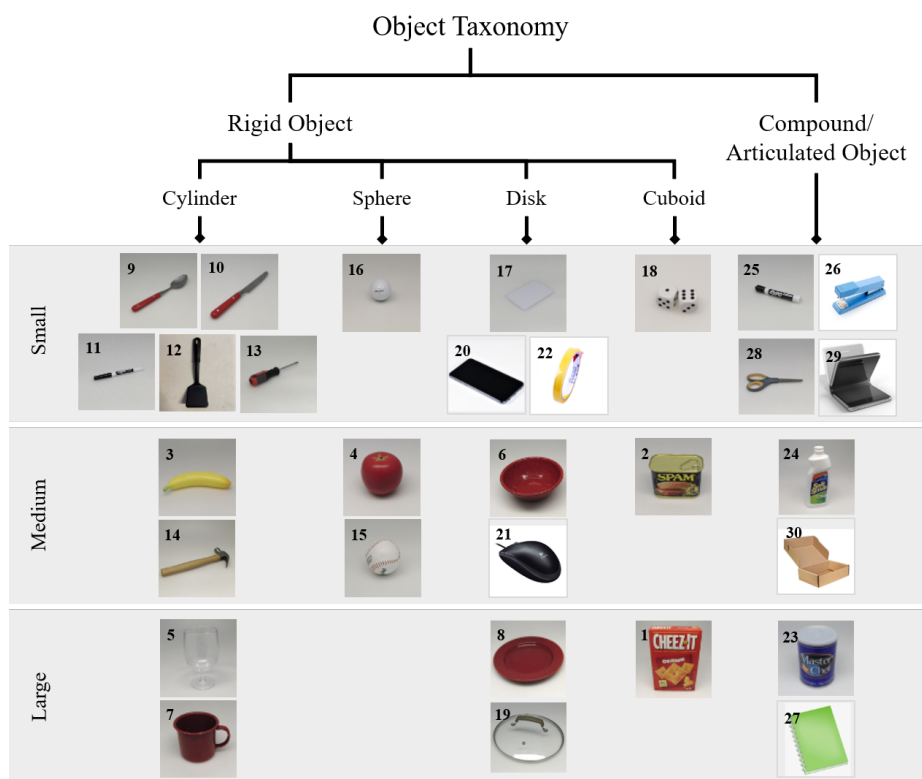


Fig. 12. Object taxonomy.

14. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* **32** (2019)
15. Ravi, N., Reizenstein, J., Novotny, D., Gordon, T., Lo, W.Y., Johnson, J., Gkioxari, G.: Accelerating 3d deep learning with pytorch3d. In: *CoRR*. vol. arXiv:2007.08501 (2020)
16. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. *ACM TOG* (2017)
17. Swamy, A., Leroy, V., Weinzaepfel, P., Baradel, F., Galaoui, S., Brégier, R., Armando, M., Franco, J.S., Rogez, G.: Showme: Benchmarking object-agnostic hand-object 3d reconstruction. In: *ICCV* (2023)
18. Xiang, Y., Schmidt, T., Narayanan, V., Fox, D.: Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes (2018)
19. 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)
20. Ye, Y., Gupta, A., Kitani, K., Tulsiani, S.: G-hop: Generative hand-object prior for interaction reconstruction and grasp synthesis. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 1911–1920 (2024)

21. 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)