AffordPose: A Large-scale Dataset of Hand-Object Interactions with Affordance-driven Hand Pose

Juntao Jian[1] Xiuping Liu[1] Manyi Li[2⋆] Ruizhen Hu[3] Jian Liu[4⋆]
[1]Dalian University of Technology [2]Shandong University
[3]Shenzhen University [4]Tsinghua University

Abstract

How human interact with objects depends on the functional roles of the target objects, which introduces the problem of affordance-aware hand-object interaction. It requires a large number of human demonstrations for the learning and understanding of plausible and appropriate hand-object interactions. In this work, we present AffordPose, a large-scale dataset of hand-object interactions with affordance-driven hand pose. We first annotate the specific part-level affordance labels for each object, e.g. twist, pull, handle-grasp, etc, instead of the general intents such as use or handover, to indicate the purpose and guide the localization of the hand-object interactions. The fine-grained hand-object interactions reveal the influence of hand-centered affordances on the detailed arrangement of the hand poses, yet also exhibit a certain degree of diversity. We collect a total of 26.7K hand-object interactions, each including the 3D object shape, the part-level affordance label, and the manually adjusted hand poses. The comprehensive data analysis shows the common characteristics and diversity of hand-object interactions per affordance via the parameter statistics and contacting computation. We also conduct experiments on the tasks of hand-object affordance understanding and affordance-oriented hand-object interaction generation, to validate the effectiveness of our dataset in learning the fine-grained hand-object interactions. Project page: https://github.com/GentlesJan/AffordPose

1. Introduction

One of the long-standing goals of robotics is to imitate all kinds of human-centered interactions, especially hand-object interactions, ranging from general grasping to functional interactions such as unscrewing a cap or even tool usage [22, 15]. Performing appropriate hand-object interaction is a complicated decision-making process. The agents need to understand the functional role of the object, select the contacting location, and perform the specific hand pose to complete the task [7, 1, 39].

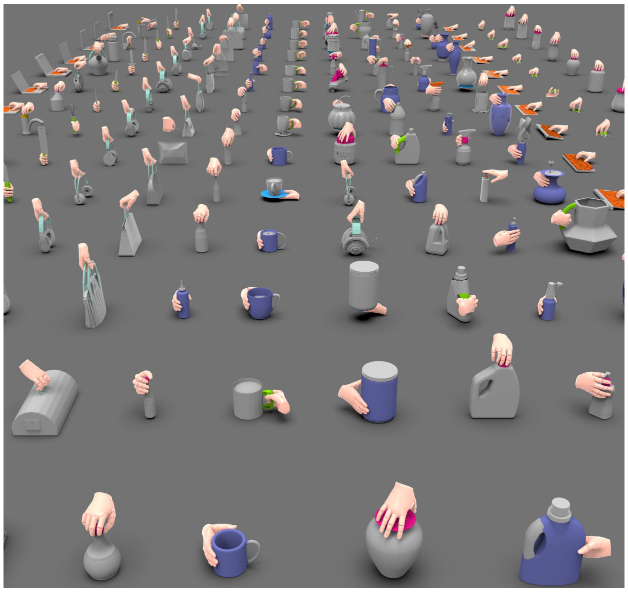

Figure 1: **A Gallery of AffordPose.** AffordPose is the first large-scale dataset for fine-grained hand-object interactions driven by the specific part-level affordance labeling, which reveals the high correlation between the object affordance and the detailed arrangement of hand poses.

There has been a trend to develop deep learning solutions to predict diverse hand-object interactions. The researchers build hand-object interaction datasets, such as HO-3D [13], DexYCB [5], Obman [16], and train different networks [26, 44, 43, 27] to predict the hand poses for the given objects. However, these works only consider the general grasping task and focus on the stability of the generated hand poses, but overlook the semantic meaning of the hand-object interactions. Recently, many related works collect additional annotations, e.g. contact maps [2, 3, 24], grasp type labels [8], and intent labels [40, 51], to learn how human use different objects with appropriate hand-object

⋆ Corresponding Authors: manyili@sdu.edu.cn, jianliu2006@gmail.com interactions, not only to hold the objects tightly but also for the convenient usage being consistent with human habits.

Object usage involves the key characteristics of the interactions, including what interaction to perform, where to interact with the object, and how to perform the interaction to achieve the purpose. As summarized in Table 1, if exists, the object usage, i.e. intents, are often of two types in the relevant datasets. One is human objectives, e.g. use, hand-out, receive as in OakInk [51], which describes the human targets regardless of the object categories and attributes. Although these human objectives illustrate the purpose of hand-object interactions well, these selected human-centered objectives are general goals and take the objects as just geometric shapes rather than functional objects. The other is object-centric affordances, e.g. pour juice in FPHAB dataset [11], which specifies the detailed type of interactions, but with a limited generalization ability.

In this paper, we take a step further to study the hand-object interactions driven by part-level affordances, which provide fine-grained localizations and are generalizable among object categories. We first collect the specific part-level affordances on the objects, i.e. the hand-centered labels such as twist, pull, handle-grasp, and the corresponding parts, instead of the general labels such as use or handover, then manually adapt the hand poses to complete the interaction tasks corresponding to these affordances. As shown in our dataset, the part-level affordances correspond to some common characteristics of the hand-object interactions even with different object categories, yet allows an extent of hand pose diversity, thus improving the understanding and prediction of hand-object interactions.

Our AffordPose is a large-scale dataset of fine-grained hand-object interactions with affordance-driven hand poses. The dataset collects 26.7K manually annotated interactions, each including the 3D object shape, the part-level affordance label, and the parameters of the detailed hand configuration. Moreover, our dataset supports the interactions for different affordances on the same object, exhibiting the distinctiveness of the hand poses w.r.t. the corresponding part-level affordance.

Our contributions are listed as follows:

- We present the AffordPose dataset, a large-scale dataset of fine-grained hand-object interactions with affordance-driven hand pose.
- We provide comprehensive data analysis to understand how affordance affects the detailed arrangement of hand poses to complete the appropriate interaction.
- We conduct experiments on two tasks, i.e. hand-object affordance understanding and affordance-oriented hand-object interaction generation, to validate the effectiveness of our dataset in learning the fine-grained hand-object interactions.

Table 1: Comparison of AffordPose dataset with the existing hand-object interaction datasets.

dataset	mod.	syn/real	#obj	#hand pose	intent
HO3D	RGBD	real	10	68	-
DexYCB	RGBD	real	20	1k	-
YCBAfford	RGB	syn	68	367	-
Obman	RGBD	syn	2.7k	21k	-
FPHAB	RGBD	real	26	273	object affordance
ContactPose	RGBD	real	25	2.3k	human objective
GRAB	Mesh	real	51	1.3k	human objective
OakInk-image	RGBD	real	100	1k	human objective
OakInk-shape	Mesh		1.7k	49k	
Ours	Mesh	syn	641	26k	part affordance

2. Related Works

Hand-Object Interaction Datasets. It's a crucial problem to produce accurate and plausible hand poses to perform hand-object interactions. The researchers have developed different hand pose acquisition, reconstruction, and simulation methods to build large-scale datasets for tasks ranging from hand pose estimation [10, 28], and grasp synthesis [26, 29], to hand-object interaction generation [4, 8].

Some works focus on hand pose estimation from hand-object interaction observations, e.g. RGB, RGB-D, or video sequential inputs. Therefore, the essential datasets should contain a large amount of accurate hand-object interactions. For example, Bighand2.2 [53] collects million-scale hand poses by building a tracking system. Obman [16] utilizes a grasp optimizer to synthesize the hand poses while ensuring the stability of the grasping. HO-3D [13] and DexYCB [5] build the motion capture systems to collect the sequential frames with one or more RGB-D cameras and solves for the 3D hand and object poses to build their datasets. On the other hand, some related works, e.g. DexteriousGrasping [54] and DexGraspNet [45], synthesize for the robotic dexterous hand poses, to complete the grasping task. These constructed datasets provide a wide range of large-scale hand-object interactions for the learning of hand pose estimation from the input observations.

However, as pointed out in ContactGrasp [2], the hand-object interactions are not only stable but also functional. In other words, the hand-object interaction datasets should involve more human annotations, rather than being built automatically, to reveal how human use different objects. Some related works, e.g. ContactGrasp [2], ContactPose [3] and Contact2Grasp [24], require the annotators to specify the contact maps of each object and then optimize the hand poses via simulators. The contact map constrains the functional goal of the interactions, while the simulator optimization ensures the physical feasibility of the hand poses. Alternatively, YCB-Affordance [8] requires the annotators to manually specify the hand position, hand pose, and grasp

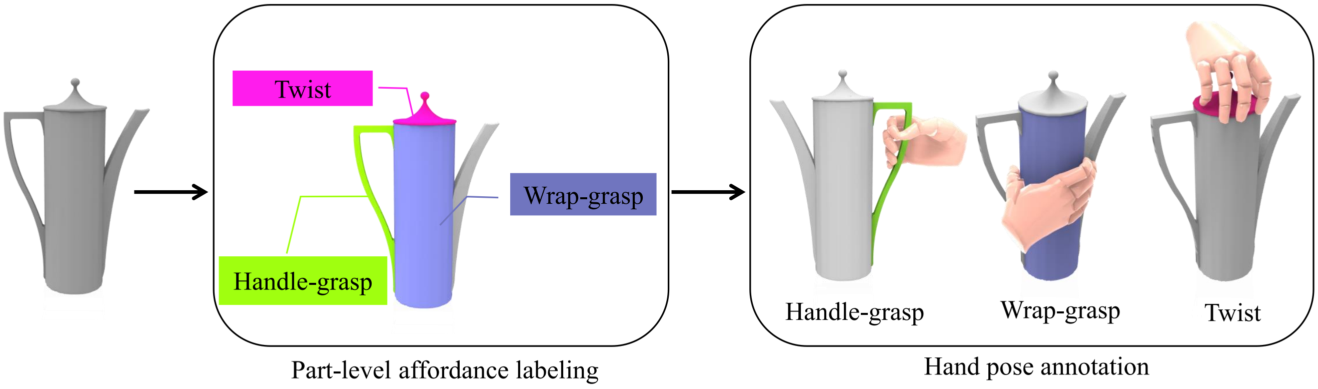

Figure 2: Dataset construction process for AffordPose. We first annotate the part-level affordance labeling, and then use it as guidance for the volunteers to manually adjust the hand pose annotations.

type of each object, and then transfer the grasps to the YCB scenes [49]. These datasets collect natural and realistic hand poses for different objects, which are necessary for the learning of hand-object interaction generation.

Some recent works investigate hand-object interactions with different intents. GRAB [40] captures the whole-body grasps for different interactions, e.g. eating a banana, drinking from a bowl, etc, which are classified into 4 different intents, i.e. use, pass, lift, and off-hand pass. Similarly, OakInk [51] collects affordance-aware and intent-oriented hand-object interactions. That is, the captured hand-object interactions are performed based on the semantic meaning of objects and the specified intents, including use, hold, lift-up, hand-out and receive. H2O [52] provides a particular benchmark for the human-human object handover analysis.

In our work, we build the dataset named AffordPose, which contains large-scale hand-object interactions with affordance-driven hand poses. Our collected data, termed as affordance-driven hand-object interactions, are performed with the guidance of part-level affordance labels such as twist, pull, handle-grasp, etc. It is different from the grasp type labels in the YCB-Affordance dataset [8] which only indicates different joint arrangements of the hands, or the intents in OakInk dataset [51] which only indicates the general task purpose regardless of the object categories. Although people may consider the object affordances while performing the interactions, i.e. affordance-aware, the corresponding affordance for each interaction is ambiguous and not explicitly specified. By contrast, our dataset contains fine-grained hand-object interactions equipped with the corresponding part-level affordance labeling, which reveals the influence of hand-centered affordances on the detailed arrangement of the hand poses and allows the affordance-oriented hand-object interaction generation.

Object Affordance Datasets. The affordances of objects are related to their functionality, i.e. what it offers the agents or environments [12]. Hassanin [14] summarizes the common affordances of man-made objects, including the inherent properties (e.g. pour, contain, display, etc) and hand-centered affordances (e.g. twist, pull, press, etc.). Therefore, many object affordance datasets make their efforts in annotating the affordance labels [18, 37, 25, 38, 46], aiming at the affordance learning tasks such as the affordance categorization [21, 42], detection [33, 39, 34] and segmentation [36, 7].

3D Affordance Net [9] is the first 3D object affordance dataset that provides the dense affordance labels on the 3D point clouds. The per-point affordance distribution highlights the regions where the interactions occur. PartAfford [50] focuses on part-level affordance discovery to build the link between the semantic parts and the affordance labels, such as openable, rollable, etc. These works stimulate a deeper understanding of the objects and their affordance, building the basis for various affordance learning tasks, e.g. task-oriented grasp detection [6], contact map generation [23].

In the spirit of the high correlation between affordance labeling and various hand-object interaction, we further connect the part-level affordance with the detailed configuration of hand-object interactions. In other words, affordance should also reflect how the interactions, especially the hand-object interactions, are performed, which is validated by our data analysis and experiments. Our dataset provides more fine-grained information for the learning of affordance-oriented hand-object interaction.

3. AffordPose Dataset

3.1. Dataset Construction

We recruit volunteers to independently complete the two stages of the data collection, i.e. part-level affordance annotation and hand pose annotation, as illustrated in Figure 2. The collected part-level affordance labeling acts as the guidance for the volunteers to complete the following hand pose

Table 2: Statistics of the 3D AffordPose Dataset. **#Object** denotes the number of objects, **#Afford** denotes the number of affordance and **#Hand** presents the number of hand-object interaction annotations.

Statistics	**All**	**Bag**	**Bottle**	**Dispenser**	**Earphone**	**Faucet**	**Handle bottle**	**Jar**	**Keyboard**	**Knife**	**Laptop**	**Mug**	**Pot**	**Scissors**
#Object	641	53	52	34	50	55	32	45	53	57	50	55	48	57
#Afford	8	2	2	3	1	2	5	3	1	1	1	3	4	1
#Hand	26712	1624	2884	2772	1400	1540	2408	2716	1484	1596	1400	3052	2240	1596

annotation, to make sure that our collected hand-object interactions match the specific affordance.

Data Preparation. We collect 641 objects from 13 categories in PartNet [31] and PartNet-Mobility [48] to annotate our hand-object interactions. The objects of each category are scaled into their normal size w.r.t. human hands in order to perform the realistic interactions. To create the dataset, we organized a panel discussion with experts and selected 8 hand-centered affordance labels, i.e. the affordance involving hand-object interactions such as press and twist. For the details of object and affordance selection, please refer to our supplementary material.

Affordance annotation. We manually annotate part-level functional areas of objects to acquire 3D object affordance annotation. In practice, we present the pre-segmented objects to the volunteers and require them to assign the affordance labels to the object parts. The finest level of the hierarchical semantic segmentation from PartNet [31] is shown to perform the annotation. For each part of an object, we ask 5 volunteers to discuss and finally make a consensus on the most related affordance label. Note that we only focus on parts with functionality. The volunteers are allowed to mark the other object parts without grasping or manipulating functionalities as ”no affordance”.

Hand-Object Interaction annotation. In this stage, we present the 3D objects with each part colored based on the affordance label in the visual interface of GraspIt [30] simulator. The affordance labeling illustrates what and where the hand-object interactions should be. The volunteers need to manually adjust the position and rotation of hand palm, and each of the finger joint angles, to complete the appropriate hand-object interaction of each affordance. During the annotation, the object position and orientation are fixed while one can interactively change the viewpoint and adjust the pose of the hand model. We invited 14 volunteers to annotate an average of 42 interactions for each object model, each with at least 28 hand interactions.

The hand model we use is the widely adopted articulated model named MANO [35]. Following the practice in [16], the hand model is composed of 16 rigid parts, including 3 phalanges of each finger and 1 hand palm. We use the standard hand model with fixed shape parameter β, while allowing the modification of the pose parameters. The pose parameters include the palm pose (extrinsic parameters) $p = \{\mathbf{t}, \mathbf{q}\}$ and the joint configuration (intrinsic parameters) $\theta \in \mathbb{R}^{16}$, where **t** and **q** represents the transformation and rotation of the entire hand, θ is the rotation angle of each joint around its pre-defined axis. During the annotation process, similar to YCBAfford dataset [8], we run the GraspIt! simulator with force analysis to avoid physically implausible hand poses and penetration.

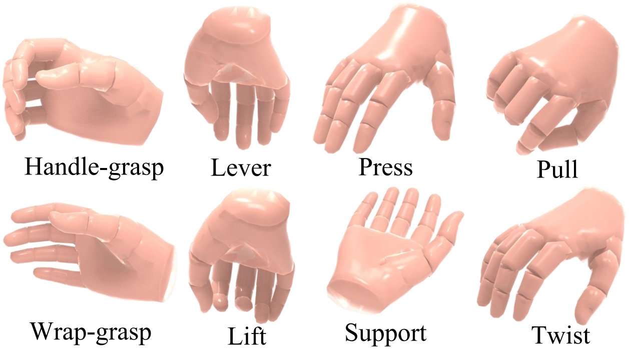

Figure 3: The representative hand poses for each affordance show their distinctive characteristics.

3.2. Dataset Statistics

Our AffordPose dataset records the complete hand-object interaction information, including the 3D object shape, the part-level affordance labels to localize the interaction, and the hand pose parameters to complete the interaction task. As listed in Table 2, the dataset collects a total of about 26.7K affordance-driven hand-object interactions, involving 641 3D objects from 13 different categories and 8 types of affordance, i.e. handle-grasp, press, lift, pull, twist, warp-grasp, support, lever. Each object has 1 to 5 different affordances and each affordance may appear in several object categories, exhibiting rich variation in the collected hand-object interactions. More statistics are listed in our supplementary material.

4. Data Analysis

The hand-object interactions for each affordance exhibit distinctive characteristics and a degree of diversity. We provide the following qualitative and quantitative data analysis

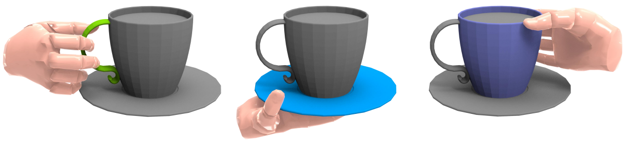

Figure 4: Different affordances, i.e. handle-grasp, support, wrap-grasp, cause significantly different hand poses to interact with the same object.

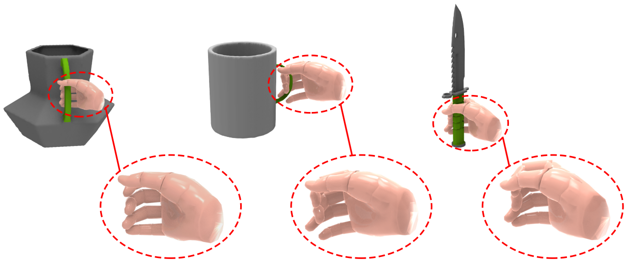

Figure 5: The hand-object interactions for the handle-grasp affordance on different object categories, i.e. pot, mug, knife, from left to right. The hand poses share similar joint configurations for the same affordance.

to understand how affordance affects the detailed arrangements of hand-object interactions.

First, the hand poses vary significantly across their different affordances. We show the per-affordance representative hands in Figure 3. The representative hand is defined as the nearest hand model of the mean intrinsic parameters θ for a specific affordance. Note that we ignore the hand rotation and position (i.e. extrinsic pose parameters) in the computation of the representative hands due to the large variation of the contacting parts from different object categories. Therefore, it only demonstrates the representative joint configurations corresponding to the same affordance label. The representative hands in Figure 3 show the distinctive characteristics for different affordances, e.g. the pinching hand for the pull affordance. The distinctivenss of affordance-driven hand poses is also reflected in Figure 4, where we sample the hand-object interactions on the same mug, but for different affordances such as handle-grasp, support, wrap-grasp. The three hands are completely different in the joint configurations, hand rotations and positions, as well as the contacting parts, w.r.t the corresponding affordances.

Second, one affordance often leads to similar joint configurations for the interactions with several different object categories. For example, we show the handle-grasp interactions with a mug, knife, and pot respectively in Figure 5. Although the rotation and position of the entire hand (i.e. extrinsic pose parameters) varies when contacting with different parts, the joint configurations of the hands look similar and share some common distinctive characteristics for the handle-grasp affordance.

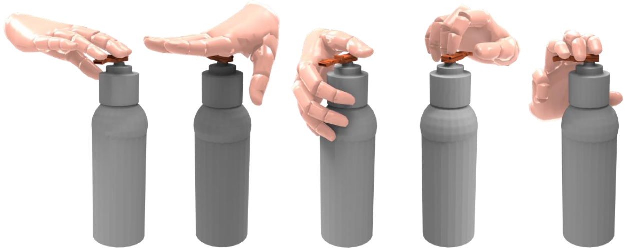

Figure 6: The hand poses on the same object exhibit some diversities even for the same affordance, e.g. press, due to personal habit factors.

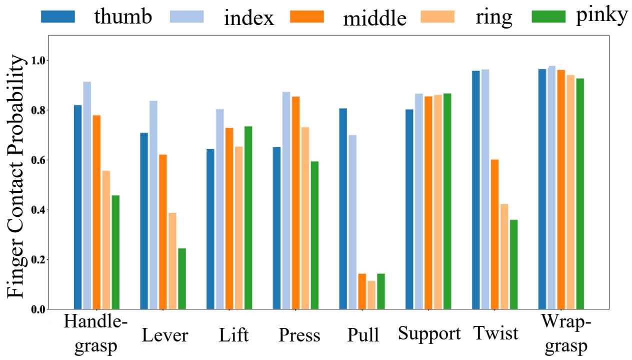

Figure 7: The probabilities of different fingers contacting the objects for each affordance.

Third, the hand-object interactions of the same affordance also exhibit a degree of diversity, which is mainly caused by different personal habits rather than different object shapes. For example, people may press the lid of the same dispenser object with different contact points on their hands, as shown in Figure 6, resulting in diverse hand poses. However, we can still tell the difference between these hand poses and those for other affordances.

One may be interested in how affordance affects the detailed configuration of the hands, i.e. the arrangement of each finger or joint. We compute the contacting frequency of the fingers for each affordance label (in Figure 7) to understand the importance of the fingers when performing different actions. It shows that the thumb, index finger, and middle finger play a significant role in hand-object interaction for most of the affordances. On the other hand, we further quantify the standard deviation of each hand joint angle, to understand the detailed diversity of each affordance. We specially select the two affordance labels, support and pull, to analyze the per-joint variation in Figure 8. We can observe that the DOFs of the root joints (except for the root of the thumb) are often limited for either of the two affordances, i.e. the variations are small for joints I1, M1, R1, P1, but relatively large for I2, M2, R2, P2. Moreover, it's interesting to see how the two quantitative analysis support and complement each other. Taking the pull affordance as

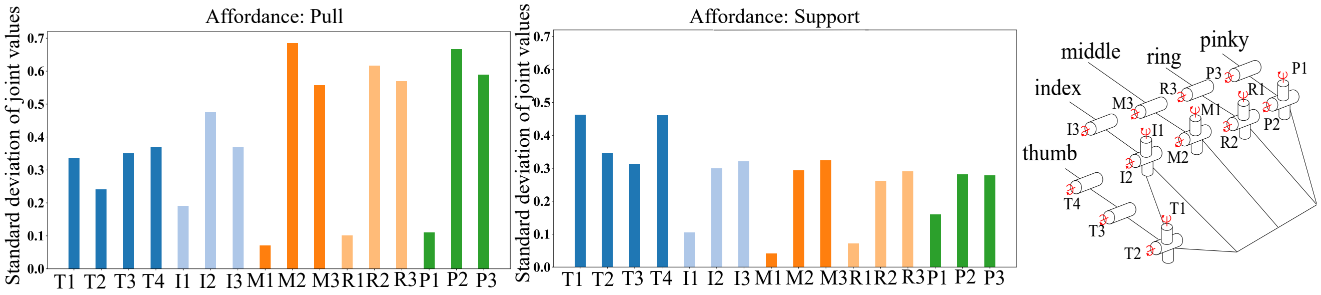

Figure 8: Left: Standard deviations of joint values for the pull and support affordances. Right: The illustration of 16 DoFs (corresponding to the joint values on the left) to represent the kinematic hand model.

an example, Figure 7 shows that the thumb and index finger often contact the object, with the other fingers curling inward, while Figure 8 indicates similar information with lower deviation on the thumb and index fingers and higher deviation on the rest fingers.

5. Experiments

Our AffordPose dataset contains the 3D objects and the hand poses as demonstrations of hand-object interactions, and the affordance labeling to indicate the fine-grained manipulation purpose. Therefore, it enables two related tasks, i.e. the hand-object affordance understanding and the affordance-oriented hand-object interaction generation. The former aims to understand whether one can infer the corresponding affordance from a rough hand demonstration and a target object to guide the interaction, while the latter is to suggest the corresponding hand-object interaction implementation for a specific affordance-related purpose. We also conducted an RGB-based hand-object interaction understanding and mesh recovery experiments.

5.1. Hand-object Affordance Understanding

Due to the high correlation between the affordances and hand-object interactions, inferring the affordance labeling from a rough hand demonstration and a target object, i.e. hand-object affordance understanding, plays an important role to guide or suggest the users what to perform and where to interact in the following interactions, which is useful in many scenarios such as AR applications.

We conduct multiple experiments with different input and output settings. The input hand demonstrations are represented as either the intrinsic hand pose parameters (i.e. the joint configurations of the hands) or all the pose parameters including the hand rotation and position as well. The network outputs the object-level affordance label (i.e. affordance classification experiments) or the per-point affordance labels (i.e. part localization experiments). The per-point affordance label is defined as the part index value if the point belongs to the corresponding part to be interacted with, and 0 otherwise.

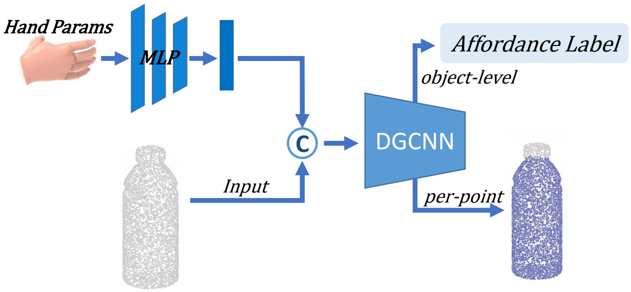

Figure 9: The hand-object affordance understanding network. It takes a hand pose and a target object as input and predicts the object-level or per-point affordance labels with different branches of DGCNN to guide the interactions.

All the experiments are implemented using the network architecture of DGCNN [47], which is developed to process the point clouds with its Edge-Conv modules. Specifically, as shown in Figure 9, we encode the hand pose parameters as a 10D vector with a 4-layer MLP network, and concatenate it with the coordinates of each point to form a high-dimensional point cloud $P' \in R^{N \times 13}$, where N is the number of points. We adopt the two corresponding branches of DGCNN [47], i.e. the classification branch and the segmentation branch, to output the object-level affordance and the per-point affordance labels respectively. The network is trained on 8-1-1 train-val-test split of the dataset.

The quantitative evaluations are reported in Table 3. We compute the affordance accuracy and IoU to measure the performance of the classification (top two rows) and the parts localization (bottom two rows) respectively. The IoU reflects whether the predicted per-point affordance label matches with the region of the corresponding part for the given hand pose. Overall, all the experiments achieve quite high performance, which aligns with our conclusion in the data analysis that there's a high correlation between affordance and hand poses. In addition, the mean affordance labeling from all the parameters consistently outperforms that from the intrinsic parameters only. This implies that although different affordances correspond to certain types

Table 3: The quantitative evaluations of hand-object affordance understanding, including the classification experiments (top two rows) and the part localization experiments (bottom two rows). We report the experiments with two kinds of hand pose settings as input: the intrinsic pose parameters and all the pose parameters.

Methods	Inputs	Handle-grasp	Lever	Lift	Press	Pull	Support	Twist	Wrap-grasp	Mean
Classification	Intrinsic	92.79	99.88	98.43	98.60	99.10	91.58	90.13	93.72	94.40
(Accuracy)%	All	99.50	100	99.16	95.00	91.00	99.65	98.35	98.73	98.39
Localization	Intrinsic	95.62	96.44	97.94	94.89	77.78	88.96	90.80	97.78	95.36
(IoU)%	All	95.05	96.99	97.90	94.42	77.78	96.59	95.17	98.91	96.29

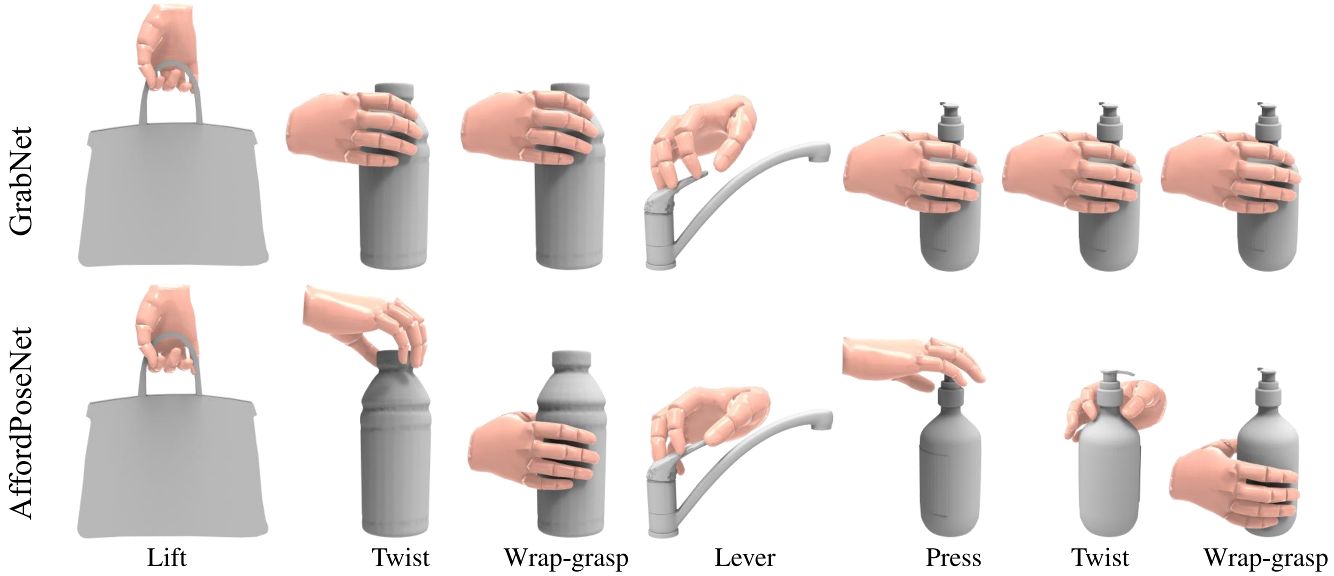

Figure 10: Qualitative results of hand-object interaction generation experiments. Top: GrabNet baseline with only the object as input. Bottom: AffordPoseNet with both the object and affordance condition as input. For the objects which have several different affordances, GrabNet fails in generating diverse hand-object interactions, while AffordPoseNet is able to generate the appropriate hand poses corresponding to the given affordance.

of hand joint configurations, it is also affected by the rough rotation and position of the input hand demonstration.

The per-affordance accuracy of these experiments listed in Table 3 offers more detailed evaluations. For example, the pull affordance obtains the worst performance in all the experiments except for the affordance classification from intrinsic parameters. This is because the main characteristics of these hand poses are the joints of the thumbs and index fingers, included in the intrinsic parameters. However, the contacting part of the pull affordance, i.e. zippers of bags, is relatively small, which affects the performance of the part localizing experiments. Additionally, both the classification and part localization performances of the support and twist affordances are small when taking only the intrinsic parameters as input, but are largely improved when all the parameters are used. The reason might be that the two affordances usually share similar joint configurations, which are represented by the intrinsic parameters, but different hand rotations and positions, which are recorded in the extrinsic parameters of the hand poses.

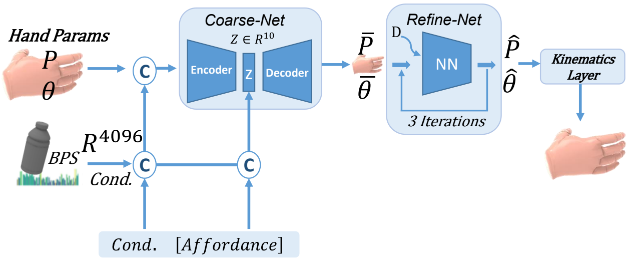

Figure 11: The network architecture of AffordPoseNet for afforfance-oriented hand-object interaction generation.

5.2. Affordance-oriented Interaction Generation

The affordance-oriented hand-object interaction generation aims to predict a possible hand pose, including the intrinsic (joint configurations) and extrinsic (hand rotation and position) parameters, from the input object and a given affordance label. Following the related work [51], we compare the generation ability of the original GrabNet base-

Table 4: The quantitative evaluations of hand-object interaction generation experiments.

Metrics	GrabNet	AffordPoseNet								
		Handle-grasp	Lever	Lift	Press	Pull	Support	Twist	Wrap-grasp	Mean
Penet.Depth(cm) ↓	0.87	1.01	0.60	1.02	0.09	0.94	1.85	0.97	0.94	0.89
Solid.Intsec.Vol(cm^3) ↓	3.20	5.32	2.44	3.37	0.97	2.82	22.93	1.83	6.04	4.57
Contact Ratio(%) ↑	96.06	100	100	92.50	92.86	75	100	96.88	97.26	96.06
Affordance accuracy(%) ↑	-	80	72.73	92.50	95.24	0	87.50	53.13	98.63	83.51

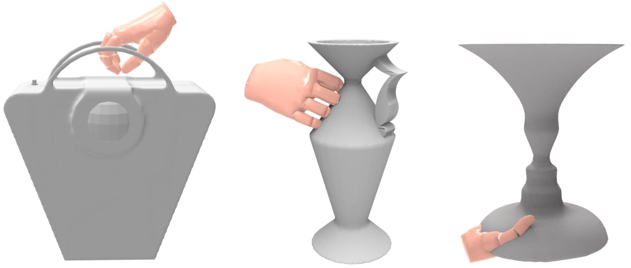

Figure 12: Challenging cases produced by AffordPoseNet for the affordance-oriented interaction generation task.

line [40] and our variant named AffordPoseNet. The former is trained to predict the hand pose from the object only, while the latter takes both the object and the specified affordance condition as input. The network architecture of AffordPoseNet is shown in Figure 11. We encode the affordance label as a one-hot vector and concatenate it with the object feature to generate the appropriate hand pose. Except for the 8-1-1 train-val-test split of our dataset, we add 211 objects with part-level affordance but no hand pose annotations to expand the test set.

Figure 10 shows the qualitative evaluations of the two experiments. As expected, in the first row, although GrabNet [41] is trained with various hand poses for different affordances, it can only produce roughly reasonable but similar hands for each object. Taking the bottle object (the 2nd and 3rd results in the top row) as an example, the predicted hands are similar even if we sample different random vectors from the Gaussian distribution to generate the results. Actually, when one object has several affordances, the predicted hand from GrabNet [41] often contacts the object in the middle of the related parts, with the hand pose corresponding to the most frequent affordance, i.e. wrap grasp for the bottle cases in Figure 10. By contrast, the AffordPoseNet is able to predict the distinctive hand poses for the specified affordance, justifying the effectiveness of the affordance labeling in guiding hand-object interactions.

Some challenging cases produced by AffordPoseNet are shown in Figure 12. We can see that the generated hand poses still successfully reflect the distinctive characteristics of the corresponding affordances, in terms of the hand joint configurations and hand rotations. But the contact regions are inappropriate, resulting in unrealistic hand-object interactions. For example, for the pull affordance, the hand is usually used to interact with the zippers of bags. But it's hard for the network to predict the correct hand pose to accurately contact at this part. To solve this issue, a physical simulator could be used to post-process the generated hand, optimizing it to contact the affordance-related part.

We also adopt the commonly used quantitative metrics [20, 19, 16] to measure the quality of the generated hand-object interactions and whether the results match the affordance conditions. The penetration depth (Penet.Depth) and solid intersection volume (Solid.Intsec.Vol) metrics reflect how much the hand models penetration with the object shape. The contact ratio is defined as the probability of the generated hands contacting the object surface. On the other hand, the affordance accuracy metric computes whether the contacting part of the predicted hand is consistent with the ground-truth part of the input affordance condition. The contacting part is defined as the part with the most contact points ($dis_{o2h} \leq \alpha = 0.004m$). In particular, if there's no contact point under this threshold, we gradually increase α by $step = 0.001m$ until the contact points appear or α reaches 0.01. If there's no contact point when $\alpha = 0.01$, we directly mark the result as a wrongly predicted hand. Note again that we test the performance on a wider range of test set, which has the part-level affordances but without the annotated hand poses, since the above evaluation metrics don't need the ground-truth hand poses for the computation.

Table 4 reports the quantitative performances on the generation of hand-object interactions. Based on the top three rows, the two experiments generate the hand poses with similar quality, although AffordPoseNet has slightly worse performance w.r.t. the solid intersection volume. On the other hand, most of the results of AffordPoseNet match with the input affordance condition, while the GrabNet baseline often generates similar hand poses when the input object has multiple affordances. However, the affordance accuracies of AffordPoseNet are particularly low for the pull and twist affordances. This is because the corresponding parts, e.g. zippers of bags, lids of bottles, are relatively small in contrast with the nearby parts. When the generated hands are not accurately contacting with the appropriate region, the affordance accuracy metric often considers them as wrong

Table 5: Quanlitative results of hand-object interaction classification.

	Handle-grasp	Lever	Lift	Press	Pull	Support	Twist	Wrap-grasp	Mean
Precision	97.95%	99.00%	98.97%	98.15%	96.67%	92.38%	94.53%	97.77%	97.31%
Recall	97.63%	98.30%	99.35%	96.09%	92.06%	96.33%	96.87%	97.10%	97.29%

Table 6: Quanlitative results of hand mesh recovery.

		Handle-grasp	Lever	Lift	Press	Pull	Support	Twist	Wrap-grasp	Mean
Mesh	MPVPE	12.2	18.44	45.36	11.82	24.3	26.36	14.02	9.58	16.4
Recovery	MPJRE	0.2516	0.2455	0.2278	0.1478	0.3796	0.1962	0.2043	0.1279	0.1892

predictions, although the joint configurations still show the correct distinctive characteristics as shown in Figure 12.

5.3. Image-based Applications

Our AffordPose dataset supports RGB-based applications as the other existing datasets do. We render the RGB images of the hand-object interactions from our dataset. Specifically, the objects are centered at the origin of the coordinate frames, and we randomly sample 3 viewpoints around each hand-object interaction instance to render the images for the RGB-based applications.

Hand-Object Interaction Classification. Taking an RGB image of the hand-object interaction as input, we train a network to predict the interaction type, i.e. the affordance label. We adopt ResNet-18 [17] as the network architecture for this classification task. Table 5 reports the classification precision and recall to evaluate the performance. The network consistently performs well on all the hand-object interaction types. We further found a high correlation between classification performance and object functionality. The fewer affordances the object category has, the better classification it obtains. For example, the categories earphone, keyboard, knife, and laptop each have only one affordance type and gain the highest interaction classification results. For more detailed evaluation statistics, please refer to our supplementary material.

Hand Mesh Recovery. We also tested the hand mesh recovery task from the input RGB image, taking the part-level affordance as the input condition. To adapt our hand pose data format, the network architecture is modified from I2L-MeshNet [32]. As shown in Table 6, we are able to reconstruct the hand meshes from images, with MPVPE (mean per vertex position error) equal to $16.4mm$ and MPJRE (mean per joint radian error) equal to $0.1892(AUC)$ on average. Some interaction types, i.e. pull, lift, and support, have relatively worse hand pose reconstruction performance than others. This is probably due to the hand pose diversity of these interaction types, making it hard to predict the accurate hand pose with occlusion in the input image.

6. Conclusions

We present AffordPose, a large-scale dataset of hand-object interactions with affordance-driven hand poses. Our data analysis reveals how affordance affects the detailed configurations of the hand poses to complete the interactions. The additional affordance labeling helps to form the fine-grained hand-object interactions: the hand poses corresponding to the same affordance exhibit some distinctive characteristics as well as a certain degree of diversity. The effectiveness of our dataset is observed in the related tasks, hand-object affordance understanding and affordance-oriented interaction generation, as well as the image-based applications.

To further expand our dataset in the future, we believe that including more types of hand-object interactions with affordance labeling will stimulate a wider range of applications. For example, a series of affordance-driven dynamic interactions will demonstrate how human perform complicated tasks, such as washing face, pouring water from a bottle to a cup, etc. In addition, we are also interested in hand-hand cooperation, e.g. bi-manual manipulations, human-robot cooperation, etc, to investigate how different hands are assigned with appropriate affordances for the cooperation. In order to construct datasets with these diverse and complicated hand-object interactions, it is necessary to develop efficient hand pose annotation methods, e.g. semi-automatric algorithms, to make large-scale and high-quality data for the learning tasks.

Acknowledgements. We thank the anonymous reviewers for their valuable comments. This work was supported by NSFC programs (61976040), Emerging Interdisciplinary Cultivation Project of Jiangxi Academy of Sciences (2022YXXJC0101), Shenzhen Collaborative Innovation Program (CJGJZD2021048092601003), GD Natural Science Foundation (2021B1515020085), and Shenzhen Science and Technology Program (RCYX20210609103121030). Manyi Li is also supported by the Excellent Young Scientists Fund Program (Overseas) of Shandong Province (Grant No.2023HWYQ-034).

References

[1] Paola Ard'on, Éric Pairet, Katrin Solveig Lohan, Subramanian Ramamoorthy, and Ronald P. A. Petrick. Affordances in robotic tasks - a survey. *ArXiv preprint arXiv:2004.07400*, 2020. 1

[2] Samarth Brahmbhatt, Ankur Handa, James Hays, and Dieter Fox. Contactgrasp: Functional multi-finger grasp synthesis from contact. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2386–2393, 2019. 1, 2

[3] Samarth Brahmbhatt, Chengcheng Tang, Christopher D. Twigg, Charles C. Kemp, and James Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, 2020. 1, 2

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

[5] Yu-Wei Chao, Wei Yang, Yu Xiang, Pavlo Molchanov, Ankur Handa, Jonathan Tremblay, Yashraj S. Narang, Karl Van Wyk, Umar Iqbal, Stan Birchfield, Jan Kautz, and Dieter Fox. Dexycb: A benchmark for capturing hand grasping of objects. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9040–9049, 2021. 1, 2

[6] Wenkai Chen, Hongzhuo Liang, Zhaopeng Chen, Fuchun Sun, and Jianwei Zhang. Learning 6-dof task-oriented grasp detection via implicit estimation and visual affordance. *ArXiv preprint arXiv:2210.08537*, 2022. 3

[7] Fu-Jen Chu, Ruinian Xu, and Patricio A. Vela. Learning affordance segmentation for real-world robotic manipulation via synthetic images. *IEEE Robotics and Automation Letters*, 4:1140–1147, 2019. 1, 3

[8] Enric Corona, Albert Pumarola, G. Alenyà, Francesc Moreno-Noguer, and Grégory Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5030–5040, 2020. 1, 2, 3, 4

[9] Sheng Deng, Xun Xu, Chaozheng Wu, Ke Chen, and Kui Jia. 3d affordancenet: A benchmark for visual object affordance understanding. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1778–1787, 2021. 3

[10] Bardia Doosti, Shujon Naha, Majid Mirbagheri, and David J. Crandall. Hope-net: A graph-based model for hand-object pose estimation. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6607–6616, 2020. 2

[11] 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. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 2

[12] James J Gibson. *The ecological approach to visual perception: classic edition*. Psychology press, 2014. 3

[13] Shreyas Hampali, Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Honnotate: A method for 3d annotation of hand and object poses. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3193–3203, 2020. 1, 2

[14] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding. *ACM Computing Surveys (CSUR)*, 54:1 – 35, 2021. 3

[15] Mohammed Hassanin, Salman Khan, and Murat Tahtali. Visual affordance and function understanding: A survey. *ACM Comput. Surv.*, 54(3), 2021. 1

[16] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11799–11808, 2019. 1, 2, 4, 8

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 9

[18] Tucker Hermans, James M Rehg, and Aaron Bobick. Affordance prediction via learned object attributes. In *IEEE international conference on robotics and automation (ICRA): Workshop on semantic perception, mapping, and exploration*, pages 181–184, 2011. 3

[19] Hanwen Jiang, Shaowei Liu, Jiashun Wang, and Xiaolong Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the International Conference on Computer Vision*, 2021. 8

[20] Korrawe Karunratanakul, Jinlong Yang, Yan Zhang, Michael J. Black, Krikamol Muandet, and Siyu Tang. Grasping field: Learning implicit representations for human grasps. *2020 International Conference on 3D Vision (3DV)*, pages 333–344, 2020. 8

[21] Zeyad Osama Khalifa and Syed Afaq Ali Shah. Towards visual affordance learning: A benchmark for affordance segmentation and recognition. *ArXiv preprint arXiv:2203.14092*, 2022. 3

[22] Mike Land, Neil Mennie, and Jennifer M. Rusted. The roles of vision and eye movements in the control of activities of daily living. *Perception*, 28:1311 – 1328, 1999. 1

[23] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Jiming Chen, and Qi Ye. Learning object affordance with contact and grasp generation. *ArXiv preprint arXiv:2210.09245*, 2022. 3

[24] Haoming Li, Xinzhuo Lin, Yang Zhou, Xiang Li, Yuchi Huo, Jiming Chen, and Qi Ye. Contact2grasp: 3d grasp synthesis via hand-object contact constraint. *ArXiv preprint arXiv:2210.09245*, 2022. 1, 2

[25] Wei Liang, Yibiao Zhao, Yixin Zhu, and Song-Chun Zhu. What is where: Inferring containment relations from videos. In *IJCAI*, 2016. 3

[26] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Generating grasp poses for a high-dof gripper using neural networks. *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1518–1525, 2019. 1, 2

[27] Min Liu, Zherong Pan, Kai Xu, Kanishka Ganguly, and Dinesh Manocha. Deep differentiable grasp planner for high-dof grippers. *ArXiv preprint arXiv:2002.01530*, 2020. 1

[28] Shao-Wei Liu, Hanwen Jiang, Jiarui Xu, Sifei Liu, and Xiaolong Wang. Semi-supervised 3d hand-object poses estimation with interactions in time. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14682–14692, 2021. 2

[29] Priyanka Mandikal and Kristen Grauman. Learning dexterous grasping with object-centric visual affordances. *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176, 2021. 2

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

[31] Kaichun Mo, Shilin Zhu, Angel X. Chang, L. Yi, Subarna Tripathi, Leonidas J. Guibas, and Hao Su. Partnet: A large-scale benchmark for fine-grained and hierarchical part-level 3d object understanding. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 909–918, 2019. 4

[32] Gyeongsik Moon and Kyoung Mu Lee. I2l-meshnet: Image-to-lixel prediction network for accurate 3d human pose and mesh estimation from a single rgb image. In *European Conference on Computer Vision (ECCV)*, 2020. 9

[33] Austin Myers, Ching Lik Teo, Cornelia Fermüller, and Yiannis Aloimonos. Affordance detection of tool parts from geometric features. *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1374–1381, 2015. 3

[34] Anh Nguyen, D. Kanoulas, Darwin Gordon Caldwell, and Nikolaos G. Tsagarakis. Object-based affordances detection with convolutional neural networks and dense conditional random fields. *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5908–5915, 2017. 3

[35] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics (TOG)*, 36:1 – 17, 2017. 4

[36] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *European Conference on Computer Vision*, 2016. 3

[37] Johann Sawatzky, Abhilash Srikantha, and Juergen Gall. Weakly supervised affordance detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5197–5206, 2017. 3

[38] Tianmin Shu, Michael S. Ryoo, and Song-Chun Zhu. Learning social affordance for human-robot interaction. *ArXiv preprint arXiv:1604.03692*, 2016. 3

[39] Hyun Oh Song, Mario Fritz, Daniel Goehring, and Trevor Darrell. Learning to detect visual grasp affordance. *IEEE Transactions on Automation Science and Engineering*, 13:798–809, 2016. 1, 3

[40] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 3, 8

[41] Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas. GRAB: A dataset of whole-body human grasping of objects. In *European Conference on Computer Vision (ECCV)*, 2020. 8

[42] Spyridon Thermos, Georgios Th. Papadopoulos, Petros Daras, and Gerasimos Potamianos. Deep affordance-grounded sensorimotor object recognition. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 49–57, 2017. 3

[43] Dylan Turpin, Liquang Wang, Eric Heiden, Yun-Chun Chen, Miles Macklin, Stavros Tsogkas, Sven J. Dickinson, and Animesh Garg. Grasp'd: Differentiable contact-rich grasp synthesis for multi-fingered hands. *ArXiv preprint arXiv:2208.12250*, 2022. 1

[44] Jacob Varley, Jonathan Weisz, Jared Weiss, and Peter K. Allen. Generating multi-fingered robotic grasps via deep learning. *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4415–4420, 2015. 1

[45] Ruicheng Wang, Jialiang Zhang, Jiayi Chen, Yinzhen Xu, Puhao Li, Tengyu Liu, and He Wang. Dexgraspnet: A large-scale robotic dexterous grasp dataset for general objects based on simulation. *ArXiv preprint arXiv:2210.02697*, 2022. 2

[46] X. Wang, Rohit Girdhar, and Abhinav Kumar Gupta. Binge watching: Scaling affordance learning from sitcoms. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3366–3375, 2017. 3

[47] Yue Wang, Yongbin Sun, Ziwei Liu, Sanjay E. Sarma, Michael M. Bronstein, and Justin M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Trans. Graph.*, 38(5), 2019. 6

[48] Fanbo Xiang, Yuzhe Qin, Kaichun Mo, Yikuan Xia, Hao Zhu, Fangchen Liu, Minghua Liu, Hanxiao Jiang, Yifu Yuan, He Wang, Li Yi, Angel X. Chang, Leonidas J. Guibas, and Hao Su. SAPIEN: A simulated part-based interactive environment. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4

[49] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. 2018. 3

[50] Chao Xu, Yixin Chen, He Wang, Song-Chun Zhu, Yixin Zhu, and Siyuan Huang. Partafford: Part-level affordance discovery from 3d objects. *ArXiv preprint arXiv:22022.13519*, 2022. 3

[51] 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. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20921–20930, 2022. 1, 2, 3, 7

[52] Ruolin Ye, Wenqiang Xu, Zhendong Xue, Tutian Tang, Yanfeng Wang, and Cewu Lu. H2o: A benchmark for visual human-human object handover analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15762–15771, 2021. 3

[53] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset

and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017. 2

[54] Tianqiang Zhu, Rina Wu, Xiangbo Lin, and Yi Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15721–15731, 2021. 2