GanHand: Predicting Human Grasp Affordances in Multi-Object Scenes

Enric Corona[1*], Albert Pumarola[1], Guillem Alenyà[1], Francesc Moreno-Noguer[1], Grégory Rogez[2]

[1]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
[2]NAVER LABS Europe

Figure 1: **GanHand** predicts hand shape and pose for grasping multiple objects given a single RGB image. The figure shows sample results on the YCB-Affordance dataset we propose, the largest dataset of human grasp affordances in real scenes.

Abstract

The rise of deep learning has brought remarkable progress in estimating hand geometry from images where the hands are part of the scene. This paper focuses on a new problem not explored so far, consisting in predicting how a human would grasp one or several objects, given a single RGB image of these objects. This is a problem with enormous potential in e.g. *augmented reality, robotics or prosthetic design. In order to predict feasible grasps, we need to understand the semantic content of the image, its geometric structure and all potential interactions with a hand physical model. To this end, we introduce a generative model that jointly reasons in all these levels and 1) regresses the 3D shape and pose of the objects in the scene; 2) estimates the grasp types; and 3) refines the 51-DoF of a 3D hand model that minimize a graspability loss. To train this model we build the YCB-Affordance dataset, that contains more than 133k images of 21 objects in the YCB-Video dataset [69]. We have annotated these images with more than 28M plausible 3D human grasps according to a 33-class taxonomy. A thorough evaluation in synthetic and real images shows that our model can robustly predict realistic grasps, even in cluttered scenes with multiple objects in close contact.*

1. Introduction

The problem of estimating 3D hand pose from monocular images has made major advances over the past few years [71, 61, 46, 10, 31, 50, 55]. Current approaches can estimate not only the 3D pose of the hand, but also its shape [21], even when manipulating an object [28].

In this paper, we move beyond these works and tackle a new problem which has not been explored so far: *given a single RGB image of a scene with an arbitrary number of objects, we aim to predict human grasp affordances*, *i.e.* predict multiple plausible solutions of how a human would grasp each one of the observed objects. We believe that such a technology would have a great impact in several fields, including virtual and augmented reality, human-robot interaction, robot imitation learning and would also open new avenues in areas like prosthetic design to *e.g.* transfer human hand-like motions to electronic gloves [33].

Predicting human grasps, however, is a very challenging problem as it requires modeling the physical interactions and contacts between a high-dimensional hand model (*e.g.* the shape of MANO model in [58] is ruled by 51-DoF) and a potentially noisy 3D representation of the objects estimated from the input RGB image. Note that this is a significantly more complex problem than that of generating robotic grasps, as robot end-effectors have much less DoF than the human hand. For instance, very recently, Mousavian *et al.* [45] introduced GraspNet to predict 6-DoF for object manipulation. Furthermore, the common practice in robotics is to use RGB-D cameras which, despite simplifying the process of modeling the geometry of the objects, do not have the versatility of standard RGB cameras.

In order to predict feasible human grasps, we introduce GanHand, a multi-task GAN architecture that given solely one input image: 1) estimates the 3D shape/pose of the objects; 2) predicts the best grasp type according to a taxon-

*Work done while visiting NAVER LABS Europe.

omy with 33 classes [18]; 3) refines the hand configuration given by the grasping class, through an optimization of the 51 parameters of the MANO model. This process involves maximizing the number of contact points between the object and the hand shape model while minimizing the interpenetration. Interestingly, our generative model is stochastic, allowing to predict several grasps per object.

Another key contribution of this paper is the YCB-Affordance dataset that we created to train our network. This dataset is based on the 58 household objects of the YCB dataset [11], whose 3D models we have *manually* annotated with a total of 367 plausible human grasps according again to the taxonomy in [18]. The grasps of 21 objects are then transferred to 92 video sequences, depicting scenes with one or several still objects captured by a moving camera. Only feasible grasps where the hand does not collide with other elements of the scene are selected. The total number of annotated frames is 133,936, with more than 28M of realistic grasps, being the *largest dataset of human grasp affordances in real scenes* built so far.

An extensive evaluation on synthetic and real data demonstrates the robustness of GanHand to predict realistic human grasps. We first evaluate our system on the ObMan dataset [28], made of single and synthetic objects from ShapeNet [12] annotated automatically with GraspIt [44]. Despite its realism, the variability of grasp types in this dataset is somewhat reduced, making it suitable for a proof-of-concept of our approach. We finally evaluate GanHand on our challenging YCB-Affordance dataset, and show that it is able to predict realistic human grasps, even in cluttered scenes as those shown in Fig. 1.

2. Related Work

Our work lies in between several areas of both computer vision and robotics. An exhaustive literature review is beyond the scope of this paper, so for practical purposes we have just focused on the most related works.

3D hand pose and shape estimation from single images. Most literature on 3D hand analysis is focused on estimating hand pose, represented by a skeleton with up to 21 joints. This problem has been studied for years, either taking as input RGB-D [63, 13, 47, 70, 26, 32] or RGB images [71, 49, 46, 61, 31, 2, 5, 21]. The community is recently shifting to estimating hand 3D shape from RGB inputs [21, 2]. Ge *et al*. [21] use graph-CNNs to infer the 3D coordinates of a 1280-vertices hand mesh. In [2], the hand is represented by the MANO model [58], that encodes the 3D shape using 51-DoF. While effective, these methods are focused on hands which do not interact with objects.

Modeling hand-object interactions. Hand-object interactions have been analyzed from different perspectives. A broad line of work aims to estimate the pose of a hand manipulating an object, either from RGB-D inputs [26, 25, 62, 65, 66], video sequences [57, 4, 48, 68] or single images [34, 35]. The most relevant work in this line is Hasson *et al*. [28] which, from a single RGB image, jointly reconstruct the shape of hands and manipulated objects.

Hand-object interactions have also been analyzed from a classification angle in which the hand pose is to be classified according to a particular taxonomy [56, 9, 8, 30, 60]. For instance, Rogez *et al*. [56] consider a taxonomy with 73 grasp types [41], used to infer hand-to-object forces and contacts. Pham *et al*. [51] also estimate contact forces using a classification scheme.

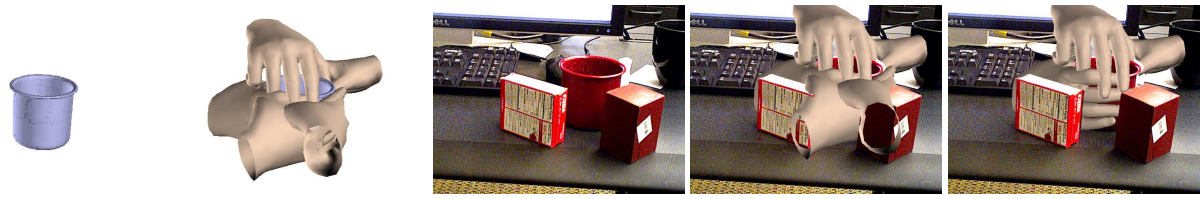
Very recent works aim to build models to understand all actors that appear in hand-object manipulation, namely the 3D hand and object poses, object and action classes and grasp types. Cai *et al*. [8] use a graphical model for this purpose and Tekin *et al*. [64] a multi-task deep architecture.

We will borrow insights from several of these approaches. For instance, we will split our hand prediction problem into a classification and a regression task. Classification will be conducted based on the 33-class grasp taxonomy proposed in [18]. Further, during the regression of the hand parameters we will consider the differentiable hand model used in [28]. Yet, the key difference with our approach and all methods discussed in this section is that in our case hands are not visible in the input images and all reasoning is done from an image of the objects alone.

Affordance prediction. Predicting affordances, defined as *opportunities of interaction in the scene*, is an active research area in the cross-domains of robotics and computer vision. The concept of affordances applies mostly to objects and it is typically posed as a semantic segmentation task, in which the pixels of an input image are classified based on their affordance label. Recent works, address this problem using deep learning [36, 53, 17, 6, 59].

Our work is more related to those approaches that learn human affordances in 3D indoor environments [24, 67, 39, 16]. These works, besides understanding the functionality of the visible elements in an image, predict valid - but coarse - 3D human poses and actions within it. We will go beyond these methods by considering much finer pose configurations of the human hand (33 grasp types) and predicting not only pose-based affordances, but also shape affordances.

Grasp prediction is one of the most important problems in robot manipulation. It involves predicting where to move the gripper (typically controlled by 6 DoF) in order to pickup an object. Some works use Deep convolutional architectures, by directly operating on visual measurements [37, 38, 42, 52, 14, 35]. Recent methods leverage data from 3D object reconstruction [54, 22, 43]. GraspNet [45] formulates the problem using a variational auto-encoder which, given an input point cloud generates several grasp hypotheses, later refined by a second network. This



(a) (b) (c) (d) (e)

Figure 2: **Grasp affordance annotations.** From the CAD model (a), we manually annotate a set of realistic grasps (b). Given a multi-object scene containing the object (c), we transfer all the grasps to this scene (d). Then, we select only valid grasps for which the hand does not collide with other objects, obtaining annotated images (e). Note that we show only three grasps for the considered object to ease visualization but, in practice, we find thousands of them for each annotated scene.

strategy is shown to significantly improve the final grasps quality. A similar scheme is used by GraspIt [44], that generates a large number of grasps and scores them using the metric proposed in [19]. The generation process of GraspIt, however, is based on heuristics that tend to produce large percentages of not feasible grasps. As we will discuss later, this is one of the limitations of the synthetic grasp dataset introduced in [28], as it relies on GraspIt.

As in [45], we adopt a stochastic generative approach that, given a single image, produces several grasp hypotheses. Note, however, that the human hand model we use has 51 DoF (in contrast to the 6 DoF of robotic grippers), which makes the problem considerably more challenging.

Datasets of Human Grasps. Building a large and realistic dataset that captures the shape of the human hand while manipulating objects is fundamental for the future research of the field. Unfortunately, the human hand is very complex, and recording its 3D shape is a major challenge. Previous work has focused on providing alternative information such as grasp taxonomies [60, 56, 7, 30]. Other works provide the position of the hand joints, obtained using manual annotations [3], data gloves [40] or wired sensors [20]. Two recent works [28, 27] annotate images with 3D hand shapes. [28] uses GraspIt to generate synthetic data, although, as discussed above, is penalized by the quality and realism of the grasps. [27] fits the MANO parametric model onto RGB images. This process, however, is not automated and laborious, and comes at the expense of the the variability of the generated grasps. [6] has released dataset of hand-object contact maps obtained with a thermal camera.

Our YCB-Affordance dataset advances state-of-the art in that it is annotated on real images and contains realistic human grasps. These grasps were manually defined, but we automatically transferred them to more than 133K images.

3. YCB-Affordance Dataset

To train our network for grasp affordance prediction in multi-object scenes, we needed natural images showing multiple objects annotated with valid human grasps. We could not find any prior work on this topic and no existing suitable dataset. We thus collected the first large-scale dataset that includes hand pose and shape for natural and realistic grasping in multi-object scenes. To do so, we augmented the YCB-Video Dataset [69] with realistic human grasps. The YCB dataset contains more than 133K frames from videos of 92 cluttered scenes with highly occluded objects whose 6D pose was annotated in camera coordinates. Our dataset, called YCB-Affordance, features grasps for all objects from the YCB Object set [11] for which a CAD model was available. These include 58 diverse household objects of particular interest for grasping and manipulation tasks, such as tools, cutlery, food or more basic shape structures. Each CAD model was first annotated with realistic grasps as explained in Sec. 3.1. Then, the resulting grasps were transferred to the YCB scenes and images as detailed in Sec. 3.2, *yielding more than 28 million grasps for 133K images*. The overall annotation process is depicted in Fig. 2.

3.1. Grasp annotations on 3D models

Realistic grasps were manually annotated to cover all possible ways to naturally pick up or manipulate the objects. We used the visual interface of the GraspIt simulator [44] to manually adapt the hand palm position and rotation, and each of the finger joint angles. We exploited its integration with the SMPL model [58] to directly retrieve the low-dimensional MANO representation and obtain posed and registered hand shape meshes. On average, we annotated the 3D models with 6 distinct grasps for symmetric objects such as cans or bottles, and up to 12 different grasps for more complex objects such as tools or cutlery. In total, we manually annotated 367 different fine-grained grasps that we also assigned to a grasp type within the 33-grasp taxonomy of [18]. This taxonomy was defined considering the position of the contact fingers, the level of power/precision tradeoff in the grasp and the position of the thumb.

We then annotated rotational symmetries in all the objects from the YCB Object set considering each main axis. A rotational symmetry is represented by its order, which indicates the number of times an object can be rotated on a particular axis and results in an equivalent shape [15]. We took advantage of objects' symmetry by simply rotating the hand around the axes, automatically extending the number

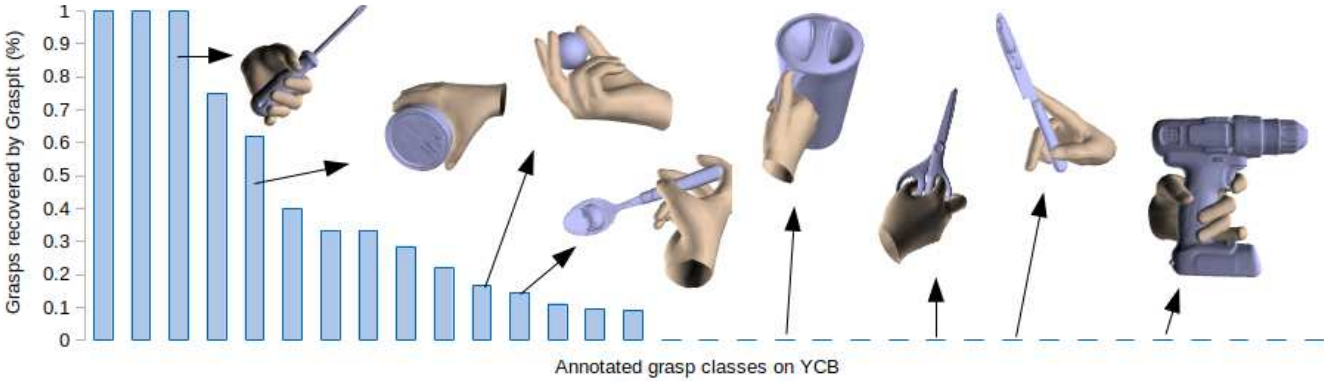


Figure 3: **Percentage of grasps found through simulation compared to our manual annotations.** When provided with the CAD models of the objects, the GraspIt simulator only recovered a portion of the natural grasps that we annotated, legitimating our choice to manually annotate the grasp for more realism. GraspIt fully recovers only three power grasp types (left) while grasps that require abducted thumbs or accurate grasps (right) are often not found at all through simulation.

of grasps *e.g*. repeating grasps along the revolution axis.

Note that simple brute-force generation of grasps using GraspIt simulator only leads to a reduced set of grasps which maximize the analytical grasp score [19] but are not necessarily correct or natural, *e.g*. holding a knife by the blade or grasping a cup with 2 fingers. On the contrary, our YCB-Affordance dataset includes only realistic grasps, including hand shapes that GraspIt would never find such as those shown in Fig. 3, *e.g*. scissors grasp.

3.2. Grasp transfer to YCB scenes

The scenes in the YCB-Video Dataset [69] contain between 3 and 9 objects in close contact. Often, the placement of the objects makes them not easily accessible for grasping without touching other objects. Our goal was to annotate the scenes with valid and feasible grasps only, *i.e*. grasps for which the hand does not collide with other objects. To do so, we exploited the 6D pose annotations of the CAD models in camera coordinates available for the different objects. For a more complete 3D representation of the scene, we manually annotated the position of the table plane . In practice, this was manually done in the first frame of each video and propagated through the remaining frames using the motion of the camera in consecutive frames.

We then transferred all the grasps annotated on the 3D CAD models to the real scenarios, using ground-truth 6D object poses and selecting only valid grasps for which the hand 3D mesh does not intersect with the objects 3D CAD models or the table plane. In most cases, several possible grasps remain valid for each object. However, the YCB-Video dataset does contain a few challenging scenes where an object is placed in a way that other objects occlude it too much for it to be grasped without any collision. In such cases, the object is considered as not reachable and left without grasp annotation. The final dataset contains 133,936 frames with more than 28M realistic grasp annotations, a suitable size to train deep networks.

4. Problem Formulation

Our goal is to predict how a human would naturally grasp one or several objects, given a single RGB image of these objects. This implies producing valid hand configurations showing several contact points with the target object but no intersection with other elements of the scene. Formally, given an image I, we train a model $\mathcal{M}$ that provides a hand pose P and shape V, and grasp type C for every object of interest in I:

$$\mathcal{M} : I \Longrightarrow \{C, V, P\},$$

where shape V is the set of vertices of the hand mesh and C is a coarse hand representation, within the 33-grasp taxonomy of [18]. Hand pose and shape parameters will be represented by the 51-DoF of the MANO model [58]. In the following we will jointly represent them by $H = \{P, V\}$.

5. Method

This section describes GanHand, our multi-task architecture that given solely one input RGB image: 1) estimates the 6D pose for known objects (or 3D pose + shape for unknown objects) in the image; 2) predicts the best grasp types for each object; 3) refines the coarse hand configurations, given by the predicted grasp classes, to gracefully adjust the fingertips to the object shape. Our three-step architecture is depicted in Fig. 4.

5.1. 3D scene understanding

To predict accurate grasps, we need to understand the geometry of the 3D scene. We consider two situations:

1) For *multi-object scenes*, we assume the observed objects are known and integrate the state-of-the-art object pose estimation method [29] to estimate their 6D pose, which we denote as T_{object}. During training, one object is randomly selected at a time, its 3D shape is projected onto the image

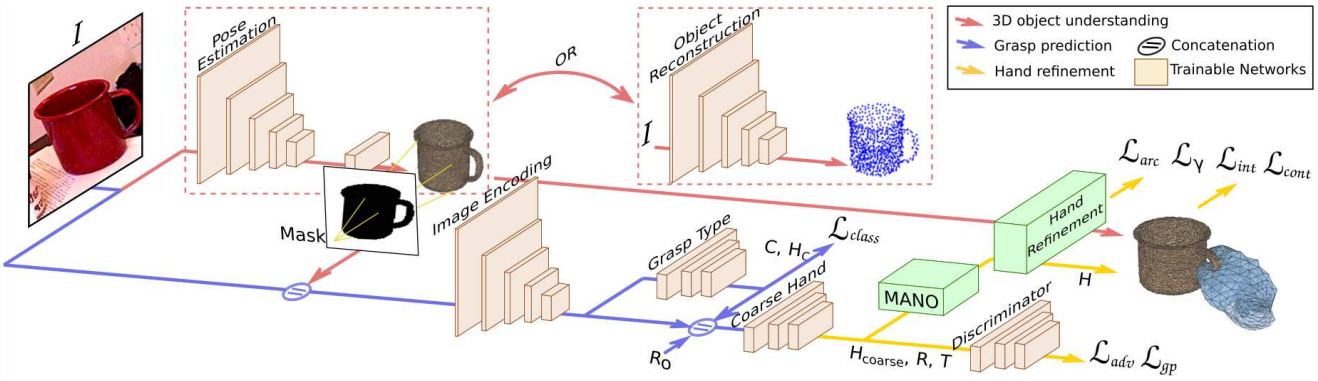


Figure 4: **Architecture of GanHand.** GanHand takes a single RGB image of one or several objects and predicts how a human would grasp these objects naturally. Our architecture consists of three stages. First, the objects' shapes and locations are estimated in the scene using an object 6D pose estimator or a reconstruction network (red). The predicted shape is then projected onto the image plane to obtain a segmentation mask that is concatenated with the input image and fed to the second sub-network for grasp prediction (blue). Finally, we refine the hand parameters and obtain hand final shapes and poses using the parametric model MANO [58] (yellow). The model is trained using adversarial, interpenetration, classification and optimization losses, indicated in bold.

plane to obtain a segmentation mask that is then concatenated with the input image and fed to the grasp prediction network. The mask indicates which object has to be focused on while the original RGB image gives contextual information about the entire scene for a more realistic grasp. At test time, we run a forward pass for each detected object, obtaining a grasp prediction for each one of them.

2) For the simpler case of *single-object scenes*, we used the object reconstruction method AtlasNet [22] retrained on the synthetic ObMan dataset [28] to validate our approach. This reconstruction method does not require to know the object beforehand but is not reliable in case of multiple objects. Once the 3D shape and pose is known, we compute its segmentation mask and proceed as before.

5.2. Predicting grasp type and coarse hand

Inspired by other approaches that tackle grasp recognition as a classification task [56], we propose a coarse-to-fine approach where grasp prediction is first addressed as a classification problem followed by a refinement stage. We predict the grasp class C that best suits the target object from a 33-grasp taxonomy [18]. To do so, we extract a representation of the input image using a pretrained and fine-tuned ResNet-50 , followed by a classification network with a cross entropy loss $\mathcal{L}_{class}$.

The predicted grasp C is associated to a representative hand configuration H_C, centered on itself, that needs to be aligned in the camera coordinate system. For this purpose we represent the absolute translation of the hand w.r.t the camera as $T = T_{object} + \Delta T$. Similarly we represent the absolute hand rotation as $R = R_o + \Delta R$, where at training, R_o is the rotation from a ground truth grasp with added noise. We then build a Fully Connected Network fed with $\{H_C, T_{object}, R_o\}$ that predicts $\{\Delta H, \Delta T, \Delta R\}$, to compute the absolute rigid pose of the hand, and its configuration $H_{coarse} = H_C + \Delta H$, which is still a coarse estimate. We observed that using this strategy of predicting the increment for each of the parameters significantly speeds up convergence during training and improves results.

At test time, we uniformly sample rotation candidates R_o and run the forward pass for multiple proposals, keeping the top-scoring hand for each object.

5.3. Hand refinement for grasping

To improve the fingers location with respect to the object surface and consequently, the quality of the predicted grasp, we propose a new differentiable and parameter-free layer that allows to maximise grasp metric during training. This layer takes as input a MANO representation of H_{coarse} and the 3D model of the object to be grasped, and returns a refined hand pose where the positions of the fingers are optimized to gracefully fit the object 3D surface. For each finger, we consider 3 rotations, one for each articulation. Following the kinematic chain, from the knuckle to the last joint, we bend/flex the finger within its physical limits, until it contacts the object. Formally, this is achieved by minimizing the distance D between the object vertices $\{O_k\}$ and the closest arc obtained when rotating an angle θ the finger's vertices about the joint axes:

$$D_\theta \leftarrow \min_i(\min_k(||A_i^\theta, O_k||_2) \ , \qquad (1)$$

where A_i^θ is the arc obtained when rotating θ degrees the i-th vertex of the finger. Given this equation to compute the arc, we can then estimate the angle γ_j' the finger needs to rotate around the first joint to collide with the object:

$$\gamma_j' \leftarrow \arg\min_\theta D_\theta + \delta, \quad \forall \theta \ s.t. \ D_\theta < t_d \ , \qquad (2)$$

where δ (angle) is a hyperparameter that controls the interpenetration of the hand into the object and hence the grasp stability (we analyze its effect in Sec. 7.1). Additionally, t_d is an upper boundary threshold to consider when there is object-finger contact. In practice, we use $t_d = 2$mm.

From these two equations we can define the following loss functions we will use to train our architecture:

$$\mathcal{L}_{arc} = \frac{1}{|J|} \sum_{j \in J} D_\theta^j \qquad \mathcal{L}_\gamma \leftarrow \sum_j^J ||\gamma'_j - \gamma_j||_2 \ , \quad (3)$$

where $|J| = 5$ is the number of fingers. $\mathcal{L}_{arc}$ aims to minimize the hand-object distances when rotating the first joint of each finger, and $\mathcal{L}_\gamma$ directly operates on the estimated angles and compares them with the ground truth ones γ_j.

This process can be sequentially performed for all three joints (knuckle, proximal and distal) of every finger, following the hand kinematic chain. In the results section we will provide results when optimizing 1, 2 or 3 joints per finger.

5.4. Training the model

So far, we have only defined loss functions that progressively guide the fingers towards the target object. We next define complementary loss functions that aim to generate human-like grasps and prevent interpenetration between the hand and the scene.

First, following [28], we build a contact prior on the hand vertices V_{cont} that are more likely to be in contact with the target object O^t and we minimize the distance between these vertices and the 3D object:

$$\mathcal{L}_{cont} = \frac{1}{|V_{cont}|} \sum_{v \in V_{cont}} \min_k ||v, O_k^t||_2 \ . \quad (4)$$

V_{cont} are computed as the vertices close to the object in at least 8% of the ground truth samples. They are mostly concentrated on the fingertips and the palm of the hand.

A very important loss (also considered in [28]) consists in penalizing the interpenetration between the hand and the object. If we denote by V_i the set of hand vertices that are inside an object, we minimize their distance to their closest object surface point:

$$\mathcal{L}_{int} = \frac{1}{|V_i|} \sum_j^{|O|} \sum_{v \in V_i} \min_k ||v, O_k^j||_2 \ , \quad (5)$$

where $|O|$ is the total number of objects found in the image.

We also penalize hand configurations that are below the table plane, by calculating the signed distance from each hand vertex to the table plane, and favoring this distance to be positive. Formally, if we represent the table plane by a point p_p and a normal v_p pointing upwards, this loss is:

$$\mathcal{L}_p = \sum_v^V \min(0, |(v - p_p) \cdot v_p|) \ . \quad (6)$$

To further enforce our network to generate anthropomorphic hands and realistic grasps we introduce a discriminator D trained using a Wasserstein loss [1]. Formally, let G be the trainable model defined so far, and let H^*, R^*, T^* be the ground truth training samples, and $\widetilde{H}, \widetilde{R}, \widetilde{T}$ interpolations between correct samples and predictions. Then, the adversarial loss is defined as:

$$\mathcal{L}_{adv} = -\mathbb{E}_{H,R,T \sim p(H,R,T)} \left[D(G(I)) \right] + \mathbb{E}_{H,R,T \sim p(H,R,T)} \left[D(H^*, R^*, T^*) \right] . \quad (7)$$

Additionally, to guarantee the satisfaction of the Lipschitz constraint in the W-GAN, we introduce a gradient penalty loss $\mathcal{L}_{gp}$ as proposed in [23].

Finally the total loss $\mathcal{L}$ to be minimized is a linear combination (see weights in Sec. 6) of all previous loss functions: $\mathcal{L}_{class}$, $\mathcal{L}_{arc}$, $\mathcal{L}_\gamma$, $\mathcal{L}_{cont}$, $\mathcal{L}_{int}$, $\mathcal{L}_p$, $\mathcal{L}_{adv}$ and $\mathcal{L}_{gp}$.

6. Implementation Details

We use a pre-trained ResNet-50 as image encoder. The discriminator and hand pose refiner are 4-layer fully connected networks with relu nonlinearities and Xavier initialization. Input images are resized to 256x256. We perform a hyperparameter grid search to maximize [19] and finally train all models using LR=0.0001, BS=32, loss weights $\lambda_{class} = 1$, $\lambda_{arc} = 0.01$, $\lambda_{cont} = 100$, $\lambda_{int} = 4000$, $\lambda_p = 20$, $\lambda_{adv} = 1$ and $\lambda_{gp} = 10$ using Adam optimizer. The Generator is trained once every 5 forward passes to improve the relative quality of the Discriminator.

The model is trained for 5 epochs, and with linear LR decay for 25 epochs more. Training models for single object (ObMan) or multi-object (YCB-Affordance) scenes takes approximately 6 and 8 days respectively on a V100 GPU. More implementation details (*e.g.*, baselines) are provided in Supplementary Material. We plan to release our code.

7. Experiments

In this section, we first evaluate the contribution of our optimization layer when included in a state-of-the-art method for hand shape estimation. Then we validate our grasp prediction method on the single-object synthetic ObMan dataset [28] and fully evaluate it in multi-object scenes with our challenging YCB-Affordance dataset.

Baseline. We consider a baseline made of a pre-trained ResNet-50 model that directly predicts the MANO representation of the hand, rotation and translation. This baseline substitutes the blue sub-network in Fig. 4, still using those layers for '3D scene understanding' and 'hand refinement'. It basically lacks the grasp taxonomy prediction.

Evaluation metrics. There exists many metrics to measure grasp quality, and we consider on several of them. The *analytical grasp metric* from [19] is one of the most common

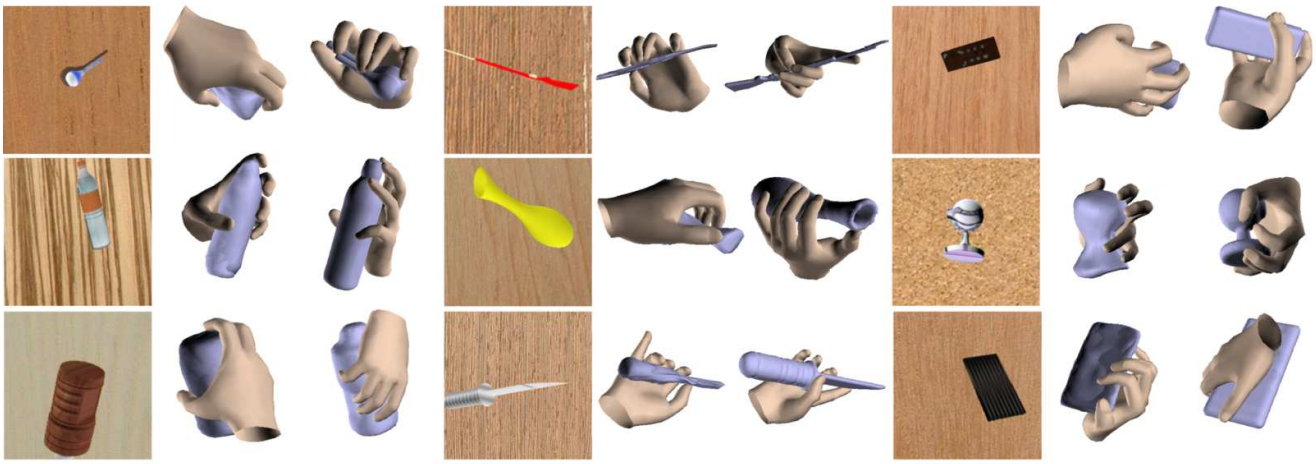


Figure 5: **Sample results on the ObMan dataset** [28]. For each object, we show the input image (left), the predicted grasp when estimating the object 3D shape (middle) and when using the ground-truth object shape (right).

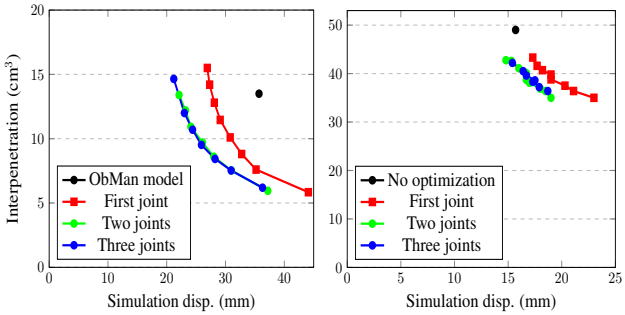


Figure 6: **Impact of the optimization layer.** Trade-off between interpenetration and the simulation displacement (the lower the better), varying δ from Eq. 2. We show the contribution of our layer in the hand reconstruction pipeline from [28] (left) and GanHand for grasp prediction (right).

Model	Baseline				GanHand				GraspIt*
Finger joints optimized	-	1	2	3	-	1	2	3	-
Grasp score [19] ↑	.19	.36	.37	.43	.40	**.60**	.56	.56	.30
# Hand-Obj Contacts ↑	2.6	4.0	4.4	**4.6**	3.0	3.9	4.4	4.4	4.4
Interpenetration ↓	42	27	29	29	48	33	34	34	**10**
Time (sec) ↓	.2	.3	.3	.4	.2	.3	.3	.4	300

Table 1: **Grasp prediction on ObMan** [28]. ↑: the higher the better, ↓: the lower the better. We sample three grasps for both GanHand and baseline, and select the one with highest grasp score, providing a good trade-off between grasp accuracy and running time. We evaluate both methods using our optimization for 1, 2, or 3 joints. Note that we run GraspIt on ground-truth object shapes.

ways to score a grasp. It basically computes an approximation of the minimum force to be applied to break the grasp stability. The *average number of contact fingers* can also be used to measure the quality of a grasp since numerous contact points between hand and object favor a strong grasp. Following [28] we also compute *hand-object interpenetration volume* (cm^3). We voxelize both object and hand, and compute the volume shared by both 3D models, using a voxel size of $0.5cm^3$. More details can be found in [28]. Also from [28] we consider the *simulation displacement* of the object mesh when it is subjected to gravity in simulation. Finally, in multi-object scenes such as YCB-Affordance, we can compute the *percentage of graspable objects* for which a valid grasp, *i.e.* with at least two contact points and no interpenetration, has been predicted.

7.1. Contribution of optimization layer

We first evaluate the contribution of our optimization layer when included in the Hand-Object reconstruction of [28]. We use the released code and trained model, and add our optimization for the first, first+second and all three joints of each finger. We depict the drop in simulation displacement and interpenetration metrics in Fig. 6-left, where we can see that the proposed layer provides a significant improvement in the reconstruction results of [28] by reducing both metrics by more than 30%. Results also improve in our grasp prediction pipeline as shown in Fig. 6-right.

7.2. Validation on synthetic data

The ObMan dataset [28] contains around 150k synthetic hand+object pairs with successful grasps produced using GraspIt for 27k different objects. Around 70k grasps were simulated for each object, keeping only the grasps with highest score. We use the images showing the objects alone and added basic background textures. For training, we used a simplified version of the method which does not consider intersections with other elements of the scene (plane and objects). Quantitative and qualitative results obtained on this dataset are shown in Table 1 and Fig. 5, respectively.

Our optimization layer provides a significant boost on all analysed metrics independently of the employed architecture (baseline or ours). It increases the number of fin-

Figure 7: **Sample results on the YCB-Affordance.** Failure cases in bottom row: The absolute pose of the can and clamps is not accurate and overlapping grasps are produced (left). The cup is detected as a brick, predicting a wrong grasp (right).

gers touching the object while correcting part of the interpenetrations. There is no significant difference in the number of Hand-Obj contacts but grasps predicted by GanHand have consistently better grasp scores. The GraspIt simulator achieves lower interpenetration by handling collisions, at the expense of a lower grasp score but, in practice, it cannot be easily deployed on real systems because of its computation cost and the need for an aligned object shape.

7.3. Evaluation on YCB-Affordance

Finally, we train both models on 80 videos from YCB-Affordance (∼130k frames). Test is evaluated on a different subset of 12 videos (2949 frames), of the same objects seen at train, but different scenes and poses. Numerical results in real multi-object scenes are reported in Table 2. In this case, we sample up to 20 predictions and select the one with least interpenetration with all predicted objects. We found that sampling 20 rotation candidates provided a good compromise between inference time and discretisation of the rotation space, with an average rotation error of .15 (mean L2 error of neighboring quaternions) while allowing inference in 0.3 seconds per object. Both methods leverage the grasp variety of YCB-Affordance dataset predicting a good diversity of grasps. However, GanHand achieves a higher % of graspable objects and a higher accuracy in predicted grasp types compared to the baseline (see Fig. 7). We deem this is a benefit of our grasp prediction module. The plane interpenetration is considerably low for both methods (3 mm), indicating both models learnt to adequately place the hands above the tables. This can be appreciated again in Fig. 7.

Model	Baseline				GanHand			
Finger joints optimized	-	1	2	3	-	1	2	3
% graspable objs ↑	4	21	33	31	**21**	**58**	**57**	**55**
Acc. grasp type % ↑	49	62	57	56	**78**	76	70	76
Grasp score [19] ↑	.37	.45	.44	.45	.36	**.47**	.46	.42
# Hand-Obj Contacts ↑	3.7	3.7	3.7	3.7	3.7	3.7	3.8	**3.9**
Obj. Interp. (cm^3) ↓	38	30	30	30	**26**	27	28	**26**
Plane interp. (cm) ↓	.1	.1	.1	**.1**	.3	.3	.2	.3

Table 2: **Results on YCB-Affordance.** GanHand outperforms the baseline in all metrics, except from plane interpenetration which is negligible for both methods.

8. Conclusion

We have introduced the problem of human grasp prediction in RGB images and proposed GanHand, a generative model that 1) estimates the 3D pose of the objects in the scene; 2) predicts grasp types; and 3) refines a 3D hand mesh model. To train GanHand, we built the YCB-Affordance, the first large-scale dataset of images annotated with plausible human grasps. We have validated our approach in both synthetic and real images showing that our model can robustly predict realistic human grasps. In further work, we could take into account the intended activity and the state of the object to select a more appropriate grasp.

Acknowledgements: This work has been partially funded by the Spanish government with the project HuMoUR TIN2017-90086-R, the ERA-Net Chistera project IPALM PCI2019-103386 and María de Maeztu Seal of Excellence MDM-2016-0656.

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017. 6

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *CVPR*, 2019. 2

[3] Ravi Balasubramanian, Ling Xu, Peter D Brook, Joshua R Smith, and Yoky Matsuoka. Physical human interactive guidance: Identifying grasping principles from human-planned grasps. *Trans. on Robotics*, 28(4):899–910, 2012. 3

[4] Luca Ballan, Aparna Taneja, Jürgen Gall, Luc Van Gool, and Marc Pollefeys. Motion capture of hands in action using discriminative salient points. In *ECCV*, 2012. 2

[5] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 3d hand shape and pose from images in the wild. In *CVPR*, 2019. 2

[6] Samarth Brahmbhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging. In *CVPR*, 2019. 2, 3

[7] Ian M Bullock, Thomas Feix, and Aaron M Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *IJRR*, 34(3):251–255, 2015. 3

[8] Minjie Cai, Kris Kitani, and Yoichi Sato. Understanding hand-object manipulation by modeling the contextual relationship between actions, grasp types and object attributes. *arXiv preprint arXiv:1807.08254*, 2018. 2

[9] Minjie Cai, Kris M Kitani, and Yoichi Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015. 2

[10] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *ECCV*, 2018. 1

[11] Berk Calli, Arjun Singh, Aaron Walsman, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. The ycb object and model set: Towards common benchmarks for manipulation research. In *ICAR*, 2015. 2, 3

[12] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 2

[13] Chiho Choi, Sang Ho Yoon, Chin-Ning Chen, and Karthik Ramani. Robust hand pose estimation during the interaction with an unknown object. In *ICCV*, 2017. 2

[14] Enric Corona, Guillem Alenyà, Antonio Gabas, and Carme Torras. Active garment recognition and target grasping point detection using deep learning. *Pattern Recognition*, 74:629–641, 2018. 2

[15] Enric Corona, Kaustav Kundu, and Sanja Fidler. Pose estimation for objects with rotational symmetry. In *IROS*, 2018. 3

[16] Enric Corona, Albert Pumarola, Guillem Alenyà, and Francesc Moreno-Noguer. Context-aware human motion prediction. In *CVPR*, 2020. 2

[17] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *ICRA*, 2018. 2

[18] Thomas Feix, Javier Romero, Heinz-Bodo Schmiedmayer, Aaron M Dollar, and Danica Kragic. The grasp taxonomy of human grasp types. *Trans. on Human-Machine Systems*, 46(1):66–77, 2015. 2, 3, 4, 5

[19] Carlo Ferrari and John F Canny. Planning optimal grasps. In *ICRA*, 1992. 3, 4, 6, 7, 8

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

[21] Liuhao Ge, Zhou Ren, Yuncheng Li, Zehao Xue, Yingying Wang, Jianfei Cai, and Junsong Yuan. 3d hand shape and pose estimation from a single rgb image. In *CVPR*, 2019. 1, 2

[22] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan Russell, and Mathieu Aubry. AtlasNet: A Papier-Mâché Approach to Learning 3D Surface Generation. In *CVPR*, 2018. 2, 5

[23] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. In *NIPS*, 2017. 6

[24] Abhinav Gupta, Scott Satkin, Alexei A Efros, and Martial Hebert. From 3d scene geometry to human workspace. In *CVPR 2011*, 2011. 2

[25] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *CVPR*, 2010. 2

[26] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *ICCV*, 2009. 2

[27] Shreyas Hampali, Markus Oberweger, Mahdi Rad, and Vincent Lepetit. Ho-3d: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. *arXiv preprint arXiv:1907.01481*, 2019. 3

[28] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 1, 2, 3, 5, 6, 7

[29] Yinlin Hu, Joachim Hugonot, Pascal Fua, and Mathieu Salzmann. Segmentation-driven 6d object pose estimation. In *CVPR*, 2019. 4

[30] De-An Huang, Minghuang Ma, Wei-Chiu Ma, and Kris M Kitani. How do we use our hands? discovering a diverse set of common grasps. In *CVPR*, 2015. 2, 3

[31] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *ECCV*, 2018. 1, 2

[32] Cem Keskin, Furkan Kıraç, Yunus Emre Kara, and Lale Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV*, 2012. 2

[33] Min Ku Kim, Ramviyas Nattanmai Parasuraman, Liu Wang, Yeonsoo Park, Bongjoong Kim, Seung Jun Lee, Nanshu Lu,

Byung-Cheol Min, and Chi Hwan Lee. Soft-packaged sensory glove system for human-like natural interaction and control of prosthetic hands. *NPG Asia Materials*, 11(1):1–12, 2019. 1

[34] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning to estimate pose and shape of hand-held objects from rgb images. In *IROS*, 2019. 2

[35] Mia Kokic, Danica Kragic, and Jeannette Bohg. Learning task-oriented grasping from human activity datasets. *IEEE Robotics and Automation Letters (RA-L)*, 2020. 2

[36] Manfred Lau, Kapil Dev, Weiqi Shi, Julie Dorsey, and Holly Rushmeier. Tactile mesh saliency. *ACM TOG*, 35(4):52, 2016. 2

[37] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *IJRR*, 34(4-5):705–724, 2015. 2

[38] Sergey Levine, Peter Pastor, Alex Krizhevsky, Julian Ibarz, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *IJRR*, 37(4-5):421–436, 2018. 2

[39] Xueting Li, Sifei Liu, Kihwan Kim, Xiaolong Wang, Ming-Hsuan Yang, and Jan Kautz. Putting humans in a scene: Learning affordance in 3d indoor environments. In *CVPR*, 2019. 2

[40] Yun Lin and Yu Sun. Grasp planning based on strategy extracted from demonstration. In *International Conference on Intelligent Robots and Systems*. IEEE, 2014. 3

[41] Jia Liu, Fangxiaoyu Feng, Yuzuko C Nakamura, and Nancy S Pollard. A taxonomy of everyday grasps in action. In *Inter. Conf. on Humanoid Robots*, 2014. 2

[42] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. 2017. 2

[43] Priyanka Mandikal and Venkatesh Babu Radhakrishnan. Dense 3d point cloud reconstruction using a deep pyramid network. In *WACV*, 2019. 2

[44] Andrew T Miller and Peter K Allen. Graspit! a versatile simulator for robotic grasping. 2004. 2, 3

[45] Arsalan Mousavian, Clemens Eppner, and Dieter Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *ICCV*, 2019. 1, 2, 3

[46] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated hands for real-time 3d hand tracking from monocular rgb. In *CVPR*, 2018. 1, 2

[47] Markus Oberweger and Vincent Lepetit. Deepprior++: Improving fast and accurate 3d hand pose estimation. In *ICCV Workshops*, 2017. 2

[48] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 2

[49] Paschalis Panteleris and Antonis Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. In *ICCV Workshops*, 2017. 2

[50] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *WACV*, 2018. 1

[51] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *PAMI*, 40(12):2883–2896, 2017. 2

[52] Lerrel Pinto and Abhinav Gupta. Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours. In *ICRA*, 2016. 2

[53] Lorenzo Porzi, Samuel Rota Bulo, Adrian Penate-Sanchez, Elisa Ricci, and Francesc Moreno-Noguer. Learning depth-aware deep representations for robotic perception. *IEEE Robotics and Automation Letters*, 2(2):468–475, 2016. 2

[54] Albert Pumarola, Stefan Popov, Francesc Moreno-Noguer, and Vittorio Ferrari. C-Flow: Conditional Generative Flow Models for Images and 3D Point Clouds. In *CVPR*, 2020. 2

[55] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain transfer for 3d pose estimation from color images without manual annotations. In *ACCV*. Springer, 2018. 1

[56] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV*, 2015. 2, 3, 5

[57] Javier Romero, Hedvig Kjellström, and Danica Kragic. Hands in action: real-time 3d reconstruction of hands in interaction with objects. In *ICRA*, 2010. 2

[58] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *SIGGRAPH*, 36(6), Nov. 2017. 1, 2, 3, 4, 5

[59] Anirban Roy and Sinisa Todorovic. A multi-scale cnn for affordance segmentation in rgb images. In *ECCV*, 2016. 2

[60] Artur Saudabayev, Zhanibek Rysbek, Raykhan Khassenova, and Huseyin Atakan Varol. Human grasping database for activities of daily living with depth, color and kinematic data streams. *Scientific data*, 5:180101, 2018. 2, 3

[61] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *CVPR*, 2018. 1, 2

[62] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *ECCV*, 2016. 2

[63] James S. Supancic, Grégory Rogez, Yi Yang, Jamie Shotton, and Deva Ramanan. Depth-based hand pose estimation: Methods, data, and challenges. *Int. J. Comput. Vis.*, 126(11):1180–1198, 2018. 2

[64] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In *CVPR*, 2019. 2

[65] Aggeliki Tsoli and Antonis A Argyros. Joint 3d tracking of a deformable object in interaction with a hand. In *ECCV*, 2018. 2

[66] Dimitrios Tzionas and Juergen Gall. 3d object reconstruction from hand-object interactions. In *ICCV*, 2015. 2

[67] Xiaolong Wang, Rohit Girdhar, and Abhinav Gupta. Binge watching: Scaling affordance learning from sitcoms. In *CVPR*, 2017. 2

[68] Yangang Wang, Jianyuan Min, Jianjie Zhang, Yebin Liu, Feng Xu, Qionghai Dai, and Jinxiang Chai. Video-based hand manipulation capture through composite motion control. *ACM TOG*, 32(4):43, 2013. 2

[69] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. In *RSS*, 2018. 1, 3, 4

[70] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *CVPR*, 2018. 2

[71] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *ICCV*, 2017. 1, 2