ContactPose: A Dataset of Grasps with Object Contact and Hand Pose

Samarth Brahmbhatt[1], Chengcheng Tang[3], Christopher D. Twigg[3], Charles C. Kemp[1], and James Hays[1,2]

[1] Institute for Robotics and Intelligent Machines, Georgia Tech, Atlanta GA, USA
{samarth.robo,hays}@gatech.edu, charlie.kemp@bme.gatech.edu
[2] Argo AI
[3] Facebook Reality Labs
{chengcheng.tang,chris.twigg}@oculus.com

Abstract. Grasping is natural for humans. However, it involves complex hand configurations and soft tissue deformation that can result in complicated regions of contact between the hand and the object. Understanding and modeling this contact can potentially improve hand models, AR/VR experiences, and robotic grasping. Yet, we currently lack datasets of hand-object contact paired with other data modalities, which is crucial for developing and evaluating contact modeling techniques. We introduce ContactPose, the first dataset of hand-object contact paired with hand pose, object pose, and RGB-D images. ContactPose has 2265 unique grasps of 25 household objects grasped with functional intents by 50 participants. Analysis of ContactPose data reveals interesting relationships between hand pose and contact. We use this data to rigorously evaluate various data representations, heuristics from the literature, and learning methods for contact modeling.

Keywords: contact modeling, hand-object contact, functional grasping

1 Introduction

A person's daily experience includes numerous and varied hand-object interactions. Understanding and reconstructing hand-object interaction has received growing attention from the computer vision, computer graphics, and robotics communities. Most research has focused on hand pose estimation [13, 42, 45, 47], realistic hand and body reconstruction [18, 19, 49, 53], and robotic grasp prediction for anthropomorphic hands [4, 28]. In this paper, we address the underexplored problem of *hand-object contact modeling*, including predicting which points on the object are in contact with the hand based on other information about the grasp, such as the 3D pose of the hand and images of the grasp. Accurate contact models have numerous applications in computer interfaces, understanding social interaction, object manipulation, and safety. For example, a hand contact model could interpret computer commands from physical interactions with a 3D printed replica object, or estimate if pathogens from a

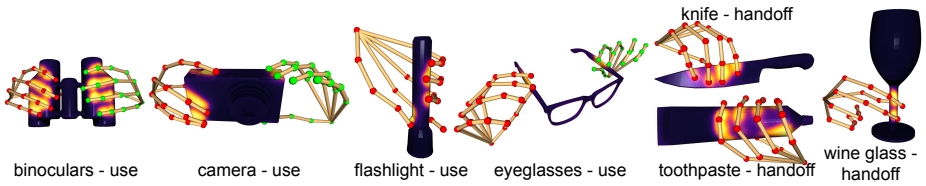


Fig. 1: Examples from ContactPose, a dataset capturing grasps of household objects. ContactPose includes high-resolution contact maps (object meshes texture-mapped with hand-object contact), 3D joint locations, and multi-view RGB-D videos of grasps (not shown here). Left hand joints are **green**, right hand joints are **red**.

contaminated surface were transmitted through contact. More broadly, accurate contact modeling can improve estimation of grasp dynamics [10, 29, 32, 37], which can lead to better VR simulations of grasping scenarios and grasping with soft robotic hands [8, 22].

Lack of ground-truth data has likely played a role in the under-exploration of this problem. Typically, the contacting surfaces of a grasp are occluded from direct observation with visible light imaging. Approaches that instrument the arm or hand with sensorized gloves [44, 50] can subtly influence natural grasping behavior, and do not measure contact on the object surface. Approaches that intersect hand models with object models to infer contact require careful selection of proximity thresholds or specific contact points on the hand [19, 49]. In addition, they cannot account for the effects of soft hand tissue deformation, since existing state-of-the-art hand models [40] are rigid.

Brahmbhatt *et al.* [3] recently introduced thermal cameras as sensors for capturing detailed ground-truth contact. Their method observes the heat transferred from the (warm) hand to the object through a thermal camera after the grasp. We adopt their method in ContactPose because it avoids the pitfalls mentioned above and allows for evaluation of contact modelling approaches with ground-truth data. However, it also imposes some constraints. 1) The objects have a plain visual texture since they are 3D printed to ensure consistent thermal properties. This does not affect contact modeling methods that rely on 3D shape and not texture, like 3D hand pose-based methods and many practical applications like simulations for VR and robotic grasping. It does limit the generalization ability of RGB image-based methods, which can potentially be mitigated by use of depth images and synthetic textures. 2) The grasps are static, because in-hand manipulation results in multiple overlapping thermal hand-prints that depend on timing and other factors. Contact modeling for static grasps is still an unsolved problem, and forms the basis for future work on dynamic grasps. The contact modeling methods we present here could be applied to dynamic scenarios on a frame-by-frame basis.

In addition, we develop a data collection protocol that captures multi-view RGB-D videos of the grasp, and an algorithm for 3D reconstruction of hand joints (§ 3.1). To summarize, we make the following contributions:

- **Data**: We introduce ContactPose, a dataset that captures 50 participants each grasping 25 objects with two different functional intents. In addition to high-quality contact maps for each grasp, it includes over 1.5 M RGB-D images from 3 viewpoints, with object pose and 3D hand joints annotated in each frame. We will make this datset available for public use to encourage research in contact modelling, and in contact-aware hand- and object-pose estimation broadly.
- **Analysis**: We dissect this data in various ways to explore the interesting relationship between contact and hand pose. This reveals some surprising patterns, and confirms some common intuitions.
- **Algorithms**: We explore various representations of object shape, hand pose, contact, and network architectures for learning-based contact modeling. Importantly, we rigorously evaluate these methods (and heuristic methods from the literature) against ground-truth unique to ContactPose.

2 Related Work

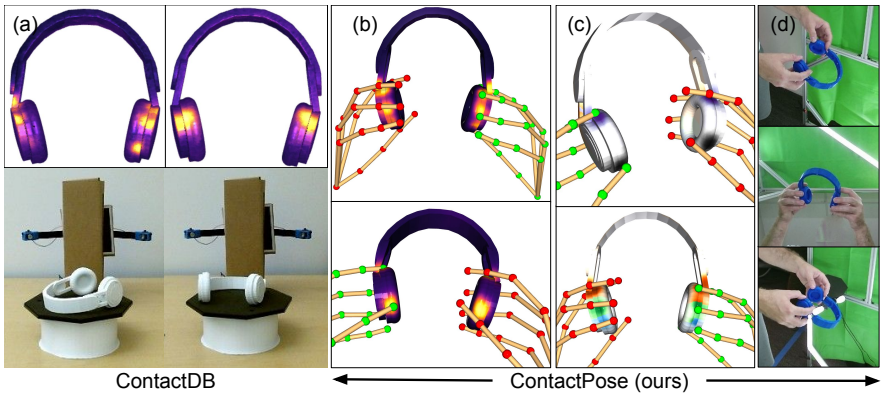


Fig. 2: Comparison to ContactDB [3]. It includes contact maps and turntable RGB-D images (a), which are often not enough to fully interpret the grasp e.g. it is not clear which fingers generated the contact. In contrast, ContactPose includes 3D joint locations (b), which allows association of contacted areas to hand parts (c), and multi-view RGB-D grasp images (d). These data enable a more comprehensive interpretation of the grasp.

Capturing and modeling contact: Previous works have instrumented hands and/or objects to capture contact. Sundaram *et al.* [44] used a tactile glove

Feature	FPHA [13]	HO-3D [17]	FreiHand [55]	STAG [44]	ContactDB [3]	Ours
3D joints	✓	✓	✓	×	×	✓
Object pose	✓	✓	×	×	✓	✓
Grasp RGB images	✓	✓	✓	✓	×	✓
Grasp Depth images	✓	✓	×	×	×	✓
Natural hand appearance	×	✓	✓	×	×	✓
Natural object appearance	×	✓	✓	✓	×	×
Naturally situated	✓	×	×	×	×	×
Multi-view images	×	×	✓	×	×	✓
Functional intent	✓	×	×	×	✓	✓
Hand-object contact	×	×	×	✓	✓	✓
# Participants	6	8	32	1	50	50
# Objects	4	8	35	26	50	25

Table 1: Comparison with existing hand-object interaction datasets. ContactPose (ours) stands out for its size, and paired hand-object contact, hand pose and object pose.

to capture hand contact during grasping. Brahmbhatt *et al.*[3] used a thermal camera after the grasp to observe the heat residue left by the warm hand on the object surface. However, these methods did not capture hand pose or grasp images, which are necessary for developing applicable contact models (Figure 2). Pham *et al.*[35] tracked hands and simple objects in videos, and trained models to predict contact forces at fingertips that explain the motion, which are evaluated against embedded force transducer data from sparse object points, in [36]. In contrast, we focus on detailed contact modeling for complex objects and grasps, evaluated against contact maps over the entire object surface.

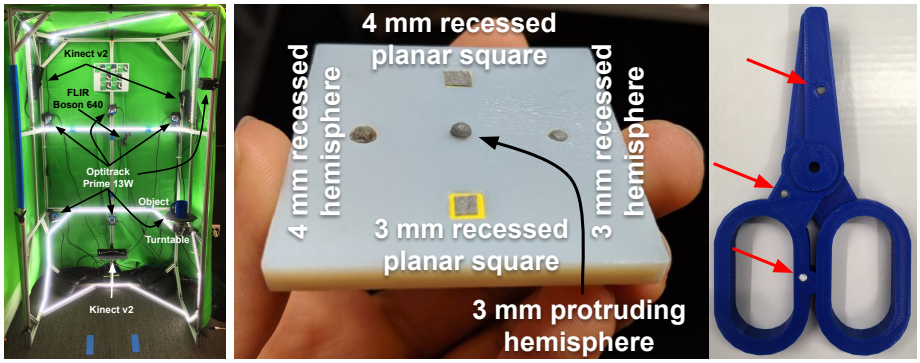
Contact heuristics: Some heuristic methods have been proposed to detect hand-object contact, often aimed at improving hand pose estimation. Hamer *et al.*[16] performed joint hand tracking and object reconstruction, and inferred contact only at fingertips using proximity threshold. In simulation [51] and robotic grasping [30, 32], contact is often determined similarly, or through collision detection [26, 46]. Ballan *et al.*[2] defined a cone circumscribing object mesh triangles, and penalized penetrating hand points (and vice versa). This formulation has also been used to penalize self-penetration and environment collision [34, 49]. While such methods were evaluated only through proxy tasks (*e.g.* hand pose estimation), ContactPose enables evaluation against ground-truth contact (§ 6).

Grasp Datasets: Focusing on datasets involving hand-object interaction, hand pose has been captured in 3D with magnetic trackers [13], optimization [17], multi-view boot-strapping [42], semi-automated human-in-the-loop [55], manually [43], synthetically [19], or as instances of a taxonomy [5, 9, 39] along with RGB-D images depicting the grasps. However, none have contact annotations (see Table 1), and suffer additional drawbacks like lack of object information [42, 55] and simplistic objects [13, 43] and interactions [19, 43], which make them unsuitable for our task. In contrast, ContactPose has ground-truth contact, and real RGB-D images of complex (including bi-manual) functional grasps for complex objects. The plain object texture is a drawback of ContactPose. Tradeoffs for this in the context of contact modelling are discussed in § 1.

3 The ContactPose Dataset

In ContactPose, hand-object contact is represented as a contact map on the object mesh surface, and observed through a thermal camera. Hand pose is represented as 3D hand(s) joint locations in the object frame, and observed through multi-view RGB-D video clips. The cameras are calibrated and object pose is known, so that the 3D joints can be projected into images (examples shown in supplementary material). Importantly, we avoid instrumenting the hands with data gloves, magnetic trackers or other sensors. This has the dual advantage of not interfering with natural grasping behavior and allowing us to use the thermal camera-based contact capture method from [3]. We develop a computational approach (Section 3.2) that optimizes for the 3D joint locations by leveraging accurate object tracking and aggregating over multi-view and temporal information. Our data collection protocol, described below, facilitates this approach.

3.1 Data Capture Protocol and Equipment



(a) (b)

Fig. 3: (a) Our setup consists of 7 Optitrack Prime 13W tracking cameras, 3 Kinect v2 RGB-D cameras, a FLIR Boson 640 thermal camera, 3D printed objects, and a turntable. (b) **Left**: Different object tracking marker configurations we investigate. **Right**: 3D printed object with recessed 3 mm hemispherical markers (highlighted by **red** arrows) offer a good compromise between unobtrusiveness and tracking performance.

We invite able-bodied participants to our laboratory and collect data through the following IRB-approved protocol. Objects are placed at random locations on a table in orientation normally encountered in practice. Participants are instructed to grasp an object with one of two functional intents (either using the object, or handing it off). Next, they stand in the data collection area (Figure 3a) and move the object for 10-15 s in the cubical space. They are instructed to hold

their hand joints steady, but are free to arbitrarily rotate the wrist and elbow, and to grasp objects with both hands or their dominant hand. This motion is recorded by 3 Kinect v2 RGB-D cameras (used for hand pose) and an Optitrack motion capture (mocap) system (used for object pose). Next, they hand the object to a researcher, who places it on a turntable, handling it with gloved hands. The object is recorded with the mocap system, Kinect v2, and a FLIR Boson 640 thermal camera as the turntable rotates a circle. Thermal images are texture-mapped to the object mesh to construct the contact maps.

Object Selection and Fabrication: We capture grasps on a subset of 25 objects from [3] that are applicable for both 'use' and 'hand-off' functional grasping (see supplementary material for a list). The objects are 3D printed in blue for good contrast with hands and the green background of our capture area. 3D printing the objects ensures consistent thermal properties and ensures geometric consistency between real world objects in capture sessions and the 3D models in our dataset.

Mocap recovers the object pose using retro-reflective markers, whose the placement on the object requires some care. Attaching a large 'marker tree' would block interactions with a significant area of the surface. Placing hemispherical markers on the surface is more promising, but a sufficient number (8+) are needed to ensure visibility during hand occlusion and the resulting 'bumps' can be uncomfortable to touch, which might influence natural grasping behavior. We investigate a few alternative marker configurations (Figure 3b). Flat pieces of tape were more comfortable but only tracked well when the marker was directly facing the camera. A good compromise is to use 3 mm hemispherical markers but to recess them into the surface by adding small cut-outs during 3D printing. These are visible from a wide range of angles but do not significantly affect the user's grip. Fixing the marker locations also allows for simple calibration between the Optitrack rigid body and the object's frame.

3.2 Grasp Capture without Hand Markers

Each grasp is observed through N time frames, each containing RGB-D images from C cameras. We want to estimate the 3D joint locations in every frame. Assuming that the hand pose relative to the object is fixed, and given the 6-DOF object pose for each frame, we aggregate the noisy per-frame 2D joint detections into a single set of high-quality 3D joint locations, which can be transformed by the frame's object pose.

For each RGB frame, we use Detectron [20] to locate the wrist, and run the OpenPose hand keypoint detector [42] on a 200×200 crop around the wrist. This produces 2D joint detections $\{\mathbf{x}^{(i)}\}_{i=1}^{N}$ and confidence values $\{\mathbf{w}^{(i)}\}_{i=1}^{N}$, following the 21-joint format from [42]. One option is to lift these 2D joint locations to 3D using the depth image [47], but that biases the location toward the camera and the hand surface (our goal is to estimate joint locations internal to the hand). Furthermore, the joint detections at any given frame are unreliable. Instead, we use our hand-object rigidity assumption to estimate the 3D joint locations $^{\mathbf{o}}\mathbf{X}$

in the object frame that are consistent with all NC images. This is done by minimizing the average re-projection error:

$$\min_{^{\mathbf{o}}\mathbf{X}} \sum_{i=1}^{N} \sum_{c=1}^{C} \mathcal{D}\left(\mathbf{x}_c^{(i)}, \pi\left({}^{\mathbf{o}}\mathbf{X}; K_c, {}^{c}T_w \; {}^{w}T_o^{(i)}\right); \mathbf{w}_c^{(i)}\right) \tag{1}$$

where $\mathcal{D}$ is a distance function, and $\pi(\cdot)$ is the camera projection function using camera intrinsics K_c and object pose w.r.t. camera at frame i, ${}^{c}T_o^{(i)} = {}^{c}T_w \; {}^{w}T_o^{(i)}$. Our approach requires the object pose w.r.t. world at each frame ${}^{w}T_o^{(i)}$ i.e. object tracking. This is done using an Optitrack motion capture system tracking markers embedded in the object surface.

In practice, the 2D joint detections are noisy and object tracking fails in some frames. We mitigate this by using the robust Huber function [23] over Mahalanobis distance ($\mathbf{w}^{(i)}$ acting as variance) as $\mathcal{D}$, and wrapping Eq. 1 in a RANSAC [12] loop. A second pass targets frames that fail the RANSAC inlier test due to inaccurate object pose. Their object pose is estimated through the correspondence between their 2D detections and the RANSAC-fit 3D joint locations, and they are included in the inlier set if they pass the inlier test (re-projection error less than a threshold). It is straightforward to extend the optimization described above to bi-manual grasps. We get a low re-projection error of 3-5 pixels w.r.t. (inherently noisy) 2D joint detections over > 100 3-view frames for each grasp, indicating that participants indeed followed the static grasp instruction. We manually curated the dataset, including clicking 2D joint locations to aid the 3D reconstruction in some cases, and discarding some obviously noisy data.

Hand Mesh Models: In addition to capturing grasps, hand shape information is collected through palm contact maps on a flat plate, and multi-view RGB-D videos of the participant performing 7 known hand gestures (shown in the supplementary material). Along with accurate 3D joints, this data enables fitting of the MANO hand mesh model [40] to each grasp (Figure 4). We use these hand meshes for some of the analysis and learning experiments discussed below.

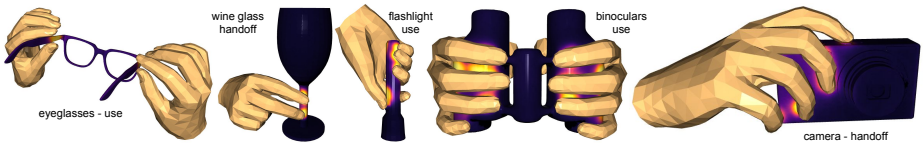
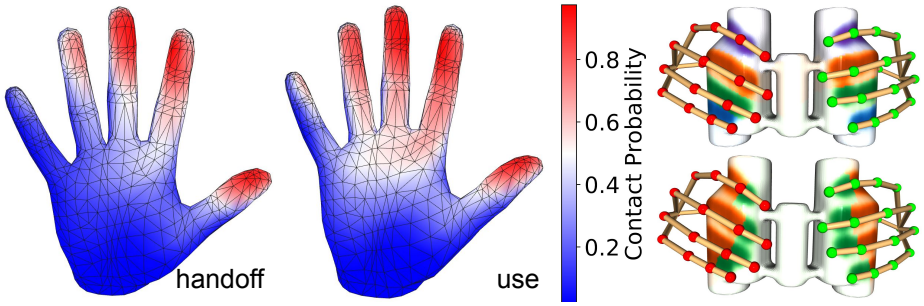


Fig. 4: MANO hand meshes [40] fit to ContactPose data. Both hand pose and shape parameters are optimized to minimize the distance of MANO joints from ContactPose 3D joint annotations.

4 Data Analysis

All contact maps are normalized to $[0, 1]$ following the sigmoid fitting procedure from [3].

4.1 Association of Contact to Hand Parts



(a) (b)

Fig. 5: (a) Hand contact probabilities estimated from the entire dataset. (b) Association of contacted binoculars points with fingers (top) and sets of phalanges at the same level of wrist proximity (bottom), **indicated by different colors**.

It has been observed that certain fingers and parts of fingers (e.g. fingertips) are more frequently in contact with the object than other parts [5, 6]. ContactPose allows us to quantify this. This can potentially inform grasp synthesis, anthropomorphic robotic hand design, and tactile sensor (e.g. BioTac [1]) placement in robotic hands.

For each grasp, we threshold the contact map at 0.4 and associate each contacted object point with its nearest hand point from the fitted MANO hand mesh. A hand point is considered to be in contact if one or more contacted object points are associated with it. A coarser analysis at the phalange level is possible by modeling phalanges as line segments connecting joints. In this case, the distance from an object point to a phalange is the distance to the closest point on the line segment.

Figure 5a shows the contact probabilities averaged over 'use' and 'handoff' grasps. Not surprisingly, the thumb, index, and middle finger are the most contacted fingers, and tips are the most contacted phalanges. Even though fingertips receive much attention in grasping literature, the contact probability for all three phalanges of the index finger is *higher* than the contact probability of the pinky fingertip. Proximal phalanges and palm also have significant contact probabilities. This is consistent with observations made by Brahmbhatt et al [3]. Interestingly, contact is more concentrated at the thumb and index finger for

'hand-off' than 'use'. 'Use' grasps have an average contact area of 35.87 cm^2 compared to 30.58 cm^2 for 'hand-off'. This analysis is similar to that in Fig. 3 of Hasson *et al.*[19], but supported by ground-truth contact rather than synthetic grasps.

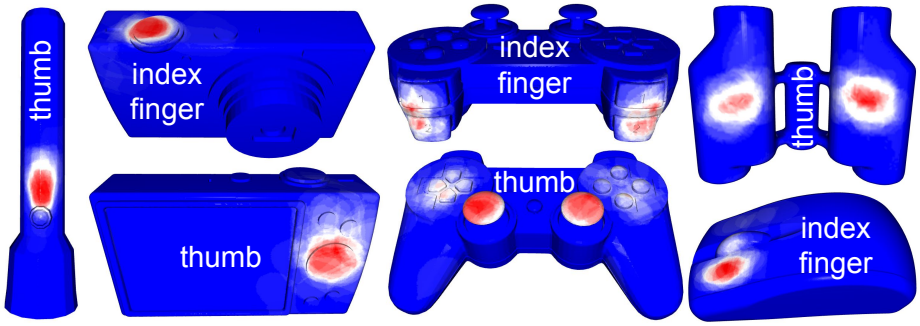


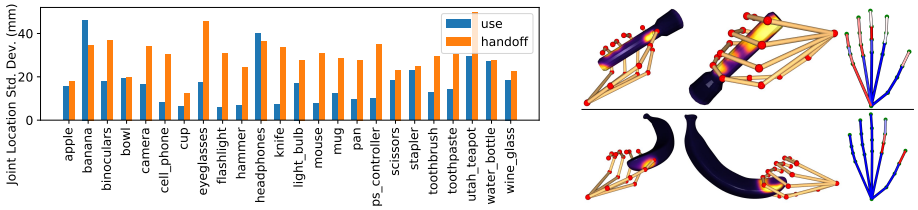
Fig. 6: Automatic 'active area' discovery: Contact probability for various hand parts on the object surface.

4.2 Automatic Active Area Discovery

Brahmbhatt et al [3] define active areas as regions on the object highly likely to be contacted. While their analysis was limited to manually selecting active areas and measuring their probability of being contacted by any part of the hand, ContactPose allows us to 'discover' active areas automatically and for specific hand parts. We use the object point-phalange association from § 4.1 (*e.g.* Fig. 5b) to estimate the probability of each object point being contacted by a given hand part (e.g. index finger tip), which can be thresholded to segment the active areas. Figure 6 shows this probability for the index fingertip and thumb, for 'use' grasps of some objects. This could potentially inform locations for placing contact sensors (real [36] or virtual for VR) on objects.

4.3 Grasp Diversity

We further quantify the effect of intent on grasping behavior by measuring the standard deviation of 3D joint locations over the dataset. The mean of all 21 joint standard deviations is shown in Figure 7a. It shows that 'hand-off' grasps are more diverse than 'use' grasps in terms of hand pose. We accounted for symmetrical objects (e.g. wine glass) by aligning the 6 palm joints (wrist + 5 knuckles) of all hand poses for that object to a single set of palm joints, where the only degree of freedom for alignment is rotation around the symmetry axis. Hand size is normalized by scaling all joint location such that the distance from wrist to middle knuckle is constant.



(a) (b)

Fig. 7: (a) Per-object standard deviation in 3D joint locations, for 'use' and 'hand-off'. 'Hand-off' grasps consistently exhibit more diversity than 'use' grasps. (b) A pair of grasps with similar hand pose but different contact characteristics. Hand contact feature color-coding is similar to Figure 5a.

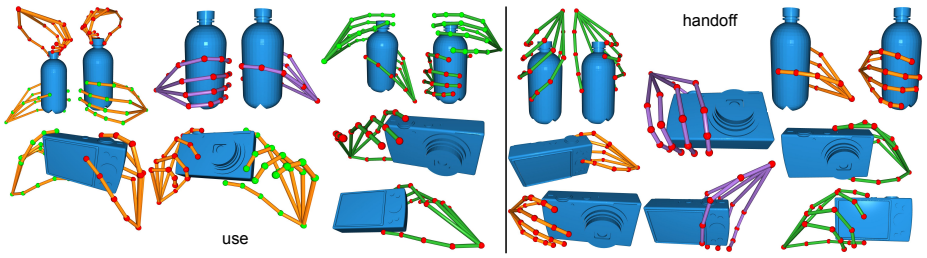


Fig. 8: Examples from hand pose clusters for 'use' and 'hand-off' grasps. Grasps from different clusters are shown **with different colors** (some grasps are bimanual). Left hand joints are **green**, right hand joints are **red**.

Organizing the grasps by clustering these aligned 3D joints (using L2 distance and HDBSCAN [7]) reveals the diversity of grasps captured in ContactPose (Figure 8). 'Hand-off' grasps exhibit a more continuous variation than 'use' grasps, which are tied more closely to the function of the object.

Figure 7b shows pair of grasps found by minimizing hand pose distance and maximizing hand contact distance over the entire dataset. We use the phalange-level contact association described in § 4.1. Summing the areas of all object mesh triangles incident to all vertices associated with a phalange creates a 20-dimensional vector. We use L2 distance over this vector as contact distance. Figure 7b shows that grasps with similar hand pose can contact different parts of the object and/or hand, inducing different forces and manipulation possibilities [13] and emphasizing that hand pose alone provides an inadequate representation of grasping.

5 Contact Modeling Experiments

This section describes our experiments on *contact modeling* given the hand pose (3D joint locations) or RGB grasp image(s), assuming known object geometry and pose. Our experiments focus on finding good data representations and learning algorithms, and evaluating techniques against ground-truth.

Object Shape Representation: We represent the object geometry through either a pointcloud densely sampled from the object body (with 1K to 30K points based on the object size), or a 64^3 voxel occupancy grid. Features encoding the input hand pose are associated with individual points or voxels. The entire pointcloud or voxel grid is then processed to predict contact values for points or surface voxels.

Hand Pose Representation: Features relating object shape to hand pose are computed for each point or voxel. These features have varying levels of richness of hand shape encoding. To simulate occlusion and noisy pose perception for the first 4 features, we sample a random camera pose and drop (set to 0) all features associated with the farthest 15% of the joints from the camera.

- `simple-joints`: We start by simply using the 21 3D joint locations w.r.t. the object coordinate system as 63-dimensional features for every point. For bi-manual grasps, the hand with the closest joint to a point is used to provide features for that point.
- `relative-joints`: Since contact at an object surface point depends on the *relative* position of the finger, we next calculate relative vectors from an object point to every joint of the hand closest to it. Contact also depends on the surface geometry: a finger is more likely to contact an object point if the vector to it is parallel to the surface normal at that point. Hence we use unit-norm surface normals and the relative joint vectors to form $63+3 = 66$-dimensional features for every point.
- `skeleton`: To better capture hand joint connectivity, we compute relative vectors from an object point to the nearest point on phalanges, modeled as line segments. 40-dimensional features for each object point are constructed

by concatenating the lengths of 20 such vectors (one for each phalange), and their dot product with the surface normal at that object point.

- `mesh`: These features leverage the rich hand geometry present in the MANO hand model. A relative vector is constructed from the object point to its closest hand mesh point. 23-dimensional features are constructed from the length of this vector, its dot product with the surface normal, and distances to 21 hand joints.
- Grasp Image(s): To investigate if CNNs can extract relevant information directly from images, we extract dense 40-dimensional features from 256×256 crops of RGB grasp images using a CNN encoder-decoder inspired by U-Net [41] (see supplementary material for architecture). These images come from the same time instant. We investigate both 3-view and 1-view settings, with feature extractor being shared across views for the former. Features are transferred to corresponding 3D object points using the known object pose and camera intrinsics, averaging the features if multiple images observe the same 3D point (Figure 10a). Points not visible from any image have all features set to 0. Image backgrounds are segmented by depth thresholding at the 20th percentile, and the foreground pixels are composited onto a random COCO [27] image. This investigation is complementary to recent work on image-based estimation of object geometry [15, 54], object pose [14, 48], and hand pose [17, 42, 45, 53, 55].

Contact Representation: We observed in early experiments that contact maps supervised with a mean squared error loss were blurred and saturated. We conjecture that this is due contact value occurrence imbalance and discontinuous contact boundaries for smooth input features. Hence, we discretize the $[0, 1]$ normalized values into 10 equal bins and treat contact prediction as a classification problem, inspired by the image colorization approach from Zhang et al [52]. We use the weighted cross entropy loss, where the weight for each bin is proportional to a linear combination of the inverse occurrence frequency of that bin and a uniform distribution (Eq. 4 from [52] with $\lambda = 0.4$). Following [52], we derive a point estimate for contact in $[0, 1]$ from classification outputs using the annealed mean ($T = 0.1$).

Learning Algorithms: Given the hand pose features associated with points or voxels, the entire pointcloud or voxel grid is processed by a neural network to predict the contact map. We use the PointNet++ [38] architecture implemented in pytorch-geometric [11, 33] (modified to reduce the number of learnable parameters) for pointclouds, and the VoxNet [31]-inspired 3D CNN architecture from [3] for voxel grids (see the supplementary material for architectures). For voxel grids, a binary feature indicating voxel occupancy is appended to hand pose features. Since hand pose features are related to surface quantities, they are set to 0 for voxels inside the object. Because the features are rich and provide fairly direct evidence of contact, we include a simple learner baseline of a multi-layer perceptron (MLP) with 90 hidden nodes, parametric ReLU [21] and batchnorm [24].

Contact Modeling Heuristics: We also investigate the effectiveness of heuristic techniques, given detailed hand geometry through the MANO hand mesh. Specifically, we use the conic distance field Ψ from [2, 49] as a proxy for contact intensity. To account for imperfections in hand modelling (due to rigidity of the MANO mesh) and fitting, we compute Ψ not only for collisions, but also when the hand and object meshes are closer than 1 cm. Finally, we calibrate Ψ to our ground truth contact through least-squares linear regression on 4700 randomly sampled contact points. Both these steps improve the technique's performance.

6 Results

Learner	Features	mug	pan	wine-glass	Average	Rank
None	Heuristic [2, 49]	78.47	83.06	81.79	81.11	3
VoxNet [3, 31]	skeleton	73.97	82.12	76.30	77.46	
MLP	simple-joints	74.69	79.89	73.68	76.09	
	relative-joints	73.20	79.70	75.91	76.27	
	skeleton	76.75	80.72	80.81	79.43	5
	mesh	81.29	**85.83**	82.52	**83.21**	1
PointNet++	simple-joints	74.64	73.32	67.92	71.96	
	relative-joints	73.89	74.91	74.88	74.56	
	skeleton	78.84	79.06	82.24	80.05	4
	mesh	**82.92**	83.13	**83.33**	83.13	2
Image enc-dec,	images (3-view)	76.28	81.56	80.14	79.33	
PointNet++	images (1-view)	71.28	78.71	72.64	74.21	

Table 2: Contact prediction re-balanced AuC (%) (higher is better) for various combinations of features and learning methods.

In this section, we evaluate various combinations of features and learning algorithms described in § 5. The metric for quantitative evaluation is the area under the curve formed by calculating accuracy at increasing contact difference thresholds. Following [52], this value is re-balanced to account for varying occurrence frequencies of values in the 10 contact bins. Following [3], we hold out all grasps for 3 objects (mug, pan and wine glass) for evaluation, and train our models on the rest.

Table 2 shows the re-balanced AuC values averaged over joint drop probability $\in [0, 0.3]$ and 3 runs for all the non-image-based variants. We observe that features capturing richer hand shape information perform better (*e.g.* `simple-joints` vs. `skeleton` and `mesh`). Learning-based techniques with `mesh` features that operate on pointclouds are able to outperform heuristics, even though the latter has access to the full high-resolution object mesh, while the former makes predictions on a pointcloud. Learning also enables `skeleton` features, which

have access to only the 3D joint locations, to perform competitively against mesh-based heuristics and features. While image-based techniques are not yet as accurate as the hand pose-based ones, a significant boost is achieved with multi-view inputs.

Figure 9 shows qualitative results for contact prediction from hand pose (predictions are transferred from the pointcloud to high-resolution meshes for better visualization). The `skeleton`-PointNet++ combination is able to predict plausible contact patterns for dropped-out parts of the hand, and capture some of the nuances of palm contact. The `mesh`-PointNet++ combination captures more nuances, especially at the thumb and bottom of the palm. In contrast, `relative-joints` features-based predictions are diffused, lack finer details, and have high contact probability in the gaps between fingers, possibly due to lack of access to information about joint connectivity and hand shape.

Figure 10b shows qualitative results for contact prediction from RGB images. These predictions have less high-frequency details compared to hand pose based predictions. They also suffer from depth ambiguity – the proximal part of the index appears to be in contact from the mug images, but is actually not. This can potentially be mitigated by use of depth images.

7 Conclusion and Future Work

We introduced ContactPose, the first dataset of paired hand-object contact, hand pose, object pose, and RGB-D images for functional grasping. Data analysis revealed some surprising patterns, like higher concentration of hand contact at the first three fingers for 'hand-off' vs. 'use' grasps. We also showed how learning-based techniques for geometry-based contact modeling can capture nuanced details missed by heuristic methods.

An interesting direction for future work could be using this contact ground-truth to develop more realistic, deformable hand mesh models. State-of-the-art models (*e.g.* [25, 40]) are rigid, while the human hand is covered with soft tissue. As the Future Work section of [40] notes, they are trained with meshes from which objects are manually removed, and do not explicitly reason about hand-object contact. The high-quality contact ground truth (along with RGB-D data) can help in the development, and more importantly, evaluation of hand mesh deformation algorithms.

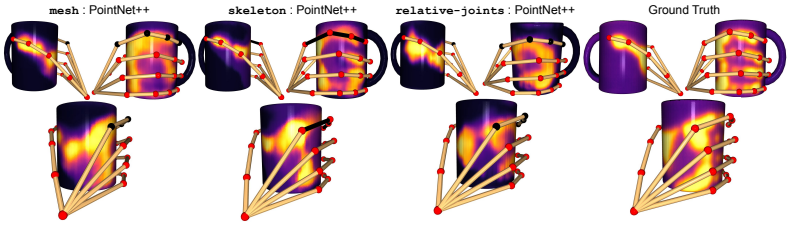


Fig. 9: Contact prediction from hand pose. All input features related to black line segments and joints were dropped (set to 0). Notice how the `mesh`- and `skeleton`-PointNet++ predictors is able to capture nuances of palm contact, thumb and finger shapes.

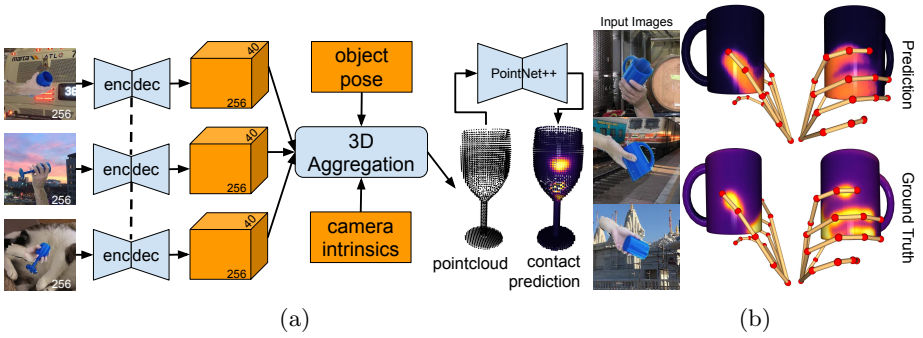


Fig. 10: (a) Image-based contact prediction architecture. (b) Contact prediction from RGB images, using networks trained with 3 views. Hand poses shown only for reference.

Bibliography

[1] BioTac. https://www.syntouchinc.com/robotics/, accessed: 2020-03-05 8

[2] Ballan, L., Taneja, A., Gall, J., Van Gool, L., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: European Conference on Computer Vision. pp. 640–653. Springer (2012) 4, 13

[3] Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019) 2, 3, 4, 5, 6, 8, 9, 12, 13

[4] Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2019) 1

[5] Bullock, I.M., Feix, T., Dollar, A.M.: The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. The International Journal of Robotics Research **34**(3), 251–255 (2015) 4, 8

[6] Bullock, I.M., Zheng, J.Z., De La Rosa, S., Guertler, C., Dollar, A.M.: Grasp frequency and usage in daily household and machine shop tasks. IEEE transactions on haptics **6**(3), 296–308 (2013) 8

[7] Campello, R.J.G.B., Moulavi, D., Zimek, A., Sander, J.: Hierarchical density estimates for data clustering, visualization, and outlier detection. ACM Trans. Knowl. Discov. Data **10**(1), 5:1–5:51 (Jul 2015). https://doi.org/10.1145/2733381, http://doi.acm.org/10.1145/2733381 11

[8] Deimel, R., Brock, O.: A novel type of compliant and underactuated robotic hand for dexterous grasping. The International Journal of Robotics Research **35**(1-3), 161–185 (2016) 2

[9] Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The grasp taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems **46**(1), 66–77 (2015) 4

[10] Ferrari, C., Canny, J.: Planning optimal grasps. In: Proceedings 1992 IEEE International Conference on Robotics and Automation. pp. 2290–2295. IEEE (1992) 2

[11] Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019) 12

[12] Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM **24**(6), 381–395 (1981) 7

[13] Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (2018) 1, 4, 11

[14] Garon, M., Lalonde, J.F.: Deep 6-dof tracking. IEEE transactions on visualization and computer graphics **23**(11), 2410–2418 (2017) 12

[15] Groueix, T., Fisher, M., Kim, V.G., Russell, B.C., Aubry, M.: A papier-mâché approach to learning 3d surface generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 216–224 (2018) 12

[16] Hamer, H., Gall, J., Weise, T., Van Gool, L.: An object-dependent hand pose prior from sparse training data. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 671–678. IEEE (2010) 4

[17] Hampali, S., Oberweger, M., Rad, M., Lepetit, V.: HO-3D: A multi-user, multi-object dataset for joint 3d hand-object pose estimation. arXiv preprint arXiv:1907.01481 (2019) 4, 12

[18] Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3d human pose ambiguities with 3d scene constraints. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 1

[19] Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11807–11816 (2019) 1, 2, 4, 9

[20] He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (Oct 2017) 6

[21] He, K., Zhang, X., Ren, S., Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: Proceedings of the IEEE international conference on computer vision. pp. 1026–1034 (2015) 12

[22] Homberg, B.S., Katzschmann, R.K., Dogar, M.R., Rus, D.: Haptic identification of objects using a modular soft robotic gripper. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 1698–1705. IEEE (2015) 2

[23] Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics, pp. 492–518. Springer (1992) 7

[24] Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning. pp. 448–456 (2015) 12

[25] Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8320–8329 (2018) 14

[26] Larsen, E., Gottschalk, S., Lin, M.C., Manocha, D.: Fast distance queries with rectangular swept sphere volumes. In: Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). vol. 4, pp. 3719–3726. IEEE (2000) 4

[27] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 12

[28] Lu, Q., Chenna, K., Sundaralingam, B., Hermans, T.: Planning multi-fingered grasps as probabilistic inference in a learned deep network. In: International Symposium on Robotics Research (2017) 1

[29] Mahler, J., Matl, M., Satish, V., Danielczuk, M., DeRose, B., McKinley, S., Goldberg, K.: Learning ambidextrous robot grasping policies. Science Robotics **4**(26), eaau4984 (2019) 2

[30] Mahler, J., Pokorny, F.T., Hou, B., Roderick, M., Laskey, M., Aubry, M., Kohlhoff, K., Kröger, T., Kuffner, J., Goldberg, K.: Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards. In: 2016 IEEE international conference on robotics and automation (ICRA). pp. 1957–1964. IEEE (2016) 4

[31] Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928. IEEE (2015) 12, 13

[32] Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine **11**(4), 110–122 (2004) 2, 4

[33] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017) 12

[34] Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (Jun 2019), http://smpl-x.is.tue.mpg.de 4

[35] Pham, T.H., Kheddar, A., Qammaz, A., Argyros, A.A.: Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2810–2819 (2015) 4

[36] Pham, T.H., Kyriazis, N., Argyros, A.A., Kheddar, A.: Hand-object contact force estimation from markerless visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**, 2883–2896 (2018) 4, 9

[37] Pollard, N.S.: Parallel methods for synthesizing whole-hand grasps from generalized prototypes. Tech. rep., MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB (1994) 2

[38] Qi, C.R., Yi, L., Su, H., Guibas, L.J.: Pointnet++: Deep hierarchical feature learning on point sets in a metric space. In: Advances in neural information processing systems. pp. 5099–5108 (2017) 12

[39] Rogez, G., Supancic, J.S., Ramanan, D.: Understanding everyday hands in action from rgb-d images. In: Proceedings of the IEEE international conference on computer vision. pp. 3889–3897 (2015) 4

[40] Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) **36**(6), 245 (2017) 2, 7, 14

[41] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015) 12

[42] Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017) 1, 4, 6, 12

[43] Sridhar, S., Mueller, F., Zollhöfer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from rgb-d input. In: European Conference on Computer Vision. pp. 294–310. Springer (2016) 4

[44] Sundaram, S., Kellnhofer, P., Li, Y., Zhu, J.Y., Torralba, A., Matusik, W.: Learning the signatures of the human grasp using a scalable tactile glove. Nature **569**(7758), 698 (2019) 2, 3, 4

[45] Tekin, B., Bogo, F., Pollefeys, M.: H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4511–4520 (2019) 1, 12

[46] Teschner, M., Kimmerle, S., Heidelberger, B., Zachmann, G., Raghupathi, L., Fuhrmann, A., Cani, M.P., Faure, F., Magnenat-Thalmann, N., Strasser, W., et al.: Collision detection for deformable objects. In: Computer graphics forum. vol. 24, pp. 61–81. Wiley Online Library (2005) 4

[47] Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics (ToG) **33**(5), 169 (2014) 1, 6

[48] Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. arXiv preprint arXiv:1809.10790 (2018) 12

[49] Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision **118**(2), 172–193 (2016) 1, 2, 4, 13

[50] Wade, J., Bhattacharjee, T., Williams, R.D., Kemp, C.C.: A force and thermal sensing skin for robots in human environments. Robotics and Autonomous Systems **96**, 1–14 (2017) 2

[51] Ye, Y., Liu, C.K.: Synthesis of detailed hand manipulations using contact sampling. ACM Transactions on Graphics (TOG) **31**(4), 41 (2012) 4

[52] Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016) 12, 13

[53] Zhang, X., Li, Q., Mo, H., Zhang, W., Zheng, W.: End-to-end hand mesh recovery from a monocular rgb image. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 1, 12

[54] Zhou, X., Leonardos, S., Hu, X., Daniilidis, K.: 3d shape estimation from 2d landmarks: A convex relaxation approach. In: proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4447–4455 (2015) 12

[55] Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., Brox, T.: Freihand: A dataset for markerless capture of hand pose and shape from

single rgb images. In: The IEEE International Conference on Computer Vision (ICCV) (October 2019) 4, 12