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

Samarth Brahmbhatt[1][0000−0002−3732−8865], Chengcheng Tang[3], Christopher D. Twigg[3], Charles C. Kemp[1], and James Hays[1,2]

[1] Georgia Tech, Atlanta GA, USA {samarth.robo,hays}@gatech.edu, charlie.kemp@bme.gatech.edu
[2] Argo AI
[3] Facebook Reality Labs {chengcheng.tang,cdtwigg}@fb.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 2306 unique grasps of 25 household objects grasped with 2 functional intents by 50 participants, and more than 2.9 M RGB-D grasp images. 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. Data, code, and trained models are available at https://contactpose.cc.gatech.edu.

**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 [14,46,50,52], realistic hand and body reconstruction [21,22,54,58], and robotic grasp prediction for anthropomorphic hands [4,31]. In this paper, we address the under-explored problem of *hand-object contact modeling i.e.* predicting object contact with the hand, based on other information about the grasp, such as the 3D hand pose and grasp images. 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

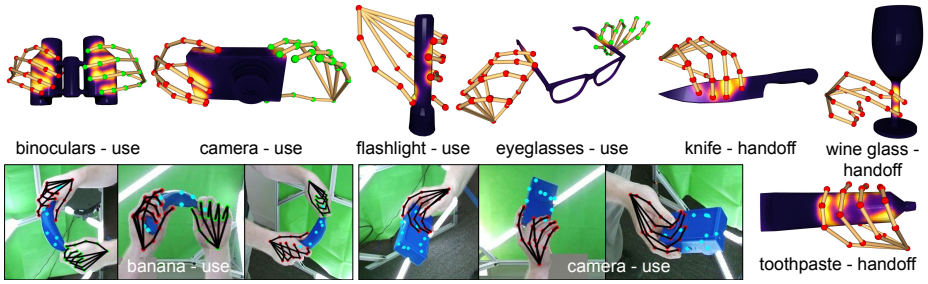



Fig. 1: Examples from ContactPose, a dataset capturing grasps of household objects. ContactPose includes high-resolution contact maps (object meshes textured with contact), 3D joints, and multi-view RGB-D videos of grasps. Left hand joints are **green**, right hand joints are **red**.

pathogens from a contaminated surface were transmitted through contact. More broadly, accurate contact modeling can improve estimation of grasp dynamics [11, 32, 35, 41], which can lead to better VR simulations of grasping scenarios and grasping with soft robotic hands [8, 25].

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 hand with gloves [48, 55] can subtly influence natural grasping behavior, and do not measure contact on the object surface. Approaches that intersect hand models with object models require careful selection of proximity thresholds or specific contact hand points [22, 54]. They also cannot account for soft hand tissue deformation, since existing state-of-the-art hand models [44] 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 because it avoids the pitfalls mentioned above and allows for evaluation of contact modeling approaches with ground-truth data. However, it also imposes some constraints. 1) Objects have a plain visual texture since they are 3D printed to ensure consistent thermal properties. This does not affect 3D hand pose-based contact modeling methods and VR/robotic grasping simulators, since they rely on 3D shape and not texture. It does limit the generalization ability of RGB-based methods, which can potentially be mitigated by using 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 methods we present here could be applied to dynamic scenarios frame-by-frame.

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**: Our dataset (ContactPose) captures 50 participants grasping 25 objects with 2 functional intents. It includes high-quality contact maps for each grasp, over 2.9 M RGB-D images from 3 viewpoints, and object pose and 3D hand joints for each frame. We will make it publicly available to encourage research in hand-object interaction and pose estimation.
- **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

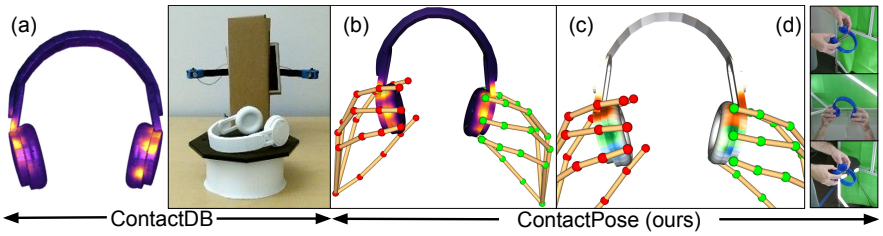



Fig. 2: Comparison to ContactDB [3]. It includes contact maps and turntable RGB-D images (a), which are often not enough to fully interpret the grasp e.g. it is not clear which fingers generated the contact. In contrast, ContactPose includes 3D joint locations (b), which allows association of contacted 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. Bernardin *et al.* [2] and Sundaram *et al.* [48] used a tactile glove 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 datasets lacked either hand pose or grasp images, which are necessary for developing applicable contact models (Figure 2). Pham *et al.* [39, 40] and Ehsani *et al.* [9] tracked hands and objects in videos, and trained models to predict contact forces and locations at fingertips that explain observed object motion. In contrast, we focus on detailed contact modeling for complex objects and grasps, evaluated against contact maps over the entire object surface.

**Contact heuristics**: Heuristic methods to detect hand-object contact are often aimed at improving hand pose estimation. Hamer *et al.* [18] performed joint

| Feature | FPHA [14] | HO-3D [20] | FreiHand [62] | STAG [48] | 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 datasets. ContactPose stands out for its size, and paired hand-object contact, hand pose and object pose.

hand tracking and object reconstruction [19], and inferred contact only at fingertips using proximity threshold. In simulation [56] and robotic grasping [33, 35], contact is often determined similarly, or through collision detection [29, 51]. Ballan *et al.* [1] 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 [38, 54]. While such methods were evaluated only through proxy tasks (*e.g.* hand pose estimation), ContactPose enables evaluation against ground-truth contact (§ 6).

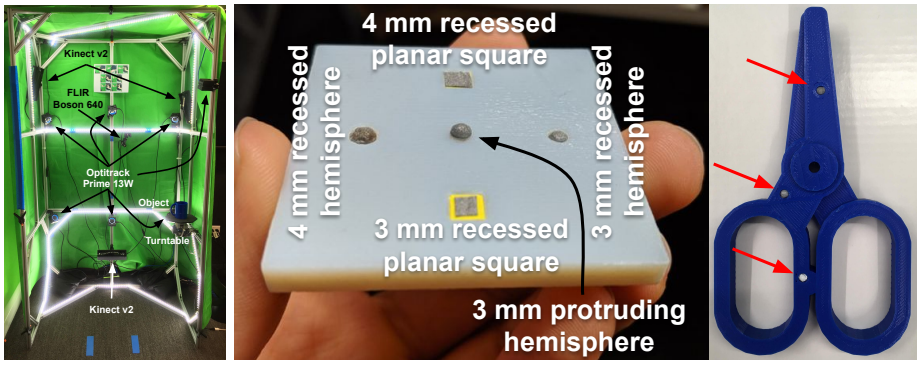
**Grasp Datasets**: Focusing on datasets involving hand-object interaction, hand pose has been captured in 3D with magnetic trackers [14], gloves [2, 16], optimization [20], multi-view boot-strapping [46], semi-automated human-in-the-loop [62], manually [47], synthetically [22], or as instances of a taxonomy [5, 10, 43] along with RGB-D images depicting the grasps. However, none of these have contact annotations (see Table 1), and suffer additional drawbacks like lack of object information [46, 62] and simplistic objects [14, 47] and interactions [22, 47], which make them unsuitable for our task. In contrast, ContactPose has a large amount of 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 modeling 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.

**Contact Capture**: Thermal images are texture-mapped to the object mesh using Open3D [59, 60]. As shown in [3] and the supp. mat., the resulting mesh textures (called contact maps) accurately capture hand-object contact.
**Object Selection and Fabrication**: We capture grasps on a subset of 25 objects from [3] that are applicable for both 'use' and 'hand-off' grasping (see supp. mat. 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 $N$ frames of RGB-D images from $C$ cameras. We assume that the hand is fixed relative to the object, and the 6-DOF object pose for each frame is given. So instead of estimating 3D joints separately in each frame, we can aggregate the noisy per-frame 2D joint detections into a single set of high-quality 3D joints, which can be transformed by the frame's object pose.

For each RGB frame, we use Detectron [23] to locate the wrist, and run the OpenPose hand keypoint detector [46] 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 [46]. One option is to lift these 2D joint locations to 3D using the depth image [52], 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 [26] over Mahalanobis distance ($\mathbf{w}^{(i)}$ acting as variance) as $\mathcal{D}$, and wrapping Eq. 1 in a RANSAC [13] 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 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 3D joints, this data can potentially enable fitting of the MANO hand mesh model [44] to each grasp [36]. In this paper, we use meshes fit to 3D joints (Figure 4, see supp. mat. for details) for some of the analysis and learning experiments discussed below.

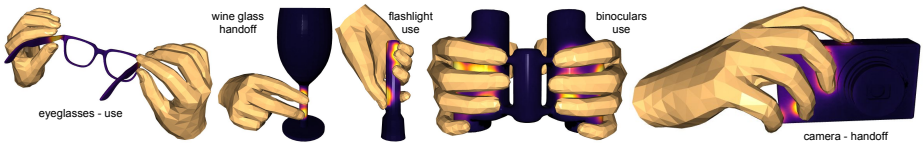


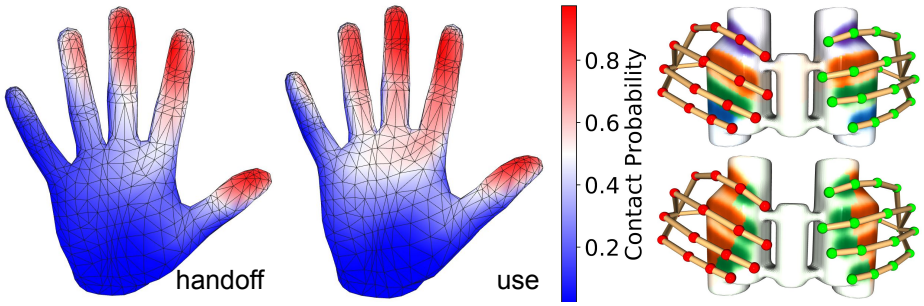

Fig. 4: MANO hand meshes [44] fit to ContactPose data. Both hand pose and shape parameters are optimized to minimize the distance of MANO joints from ContactPose 3D joint annotations.

## 4 Data Analysis

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

**Association of Contact to Hand Parts**: It has been observed that certain fingers and parts (e.g. fingertips) are contacted more frequently than others [5,6]. ContactPose allows us to quantify this. This can potentially inform anthropomorphic robotic hand design and tactile sensor (e.g. BioTac [49]) 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 contacted 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.

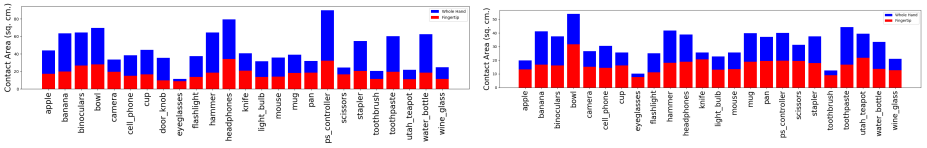



(a) (b)

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**.

Figure 5a shows the contact probabilities averaged over 'use' and 'hand-off' 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 that 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.* [22], but supported by ground-truth contact rather than synthetic grasps.

Comparison of the average fingertip vs. whole-hand contact areas (Figure 6) shows that non-fingertip areas play a significant role in grasp contact, confirming the approximate analysis in [3].



(a) 'use' grasps (b) 'handoff' grasps

Fig. 6: Comparing average fingertip (**red**) vs. whole-hand (**blue**) contact areas.

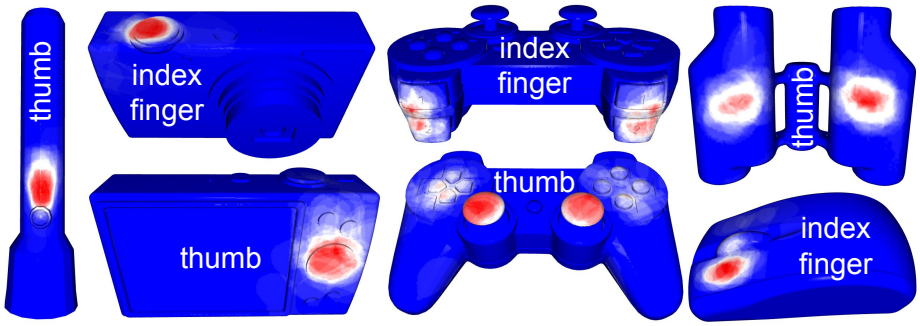
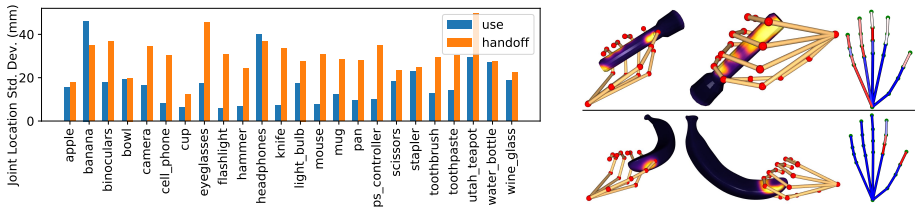



Fig. 7: Automatic 'active area' discovery: Contact probability for various hand parts on the object surface.

**Automatic Active Area Discovery**: Brahmbhatt et al [3] define active areas as regions on the object highly likely to be contacted. While they manually selected active areas and measured 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 described above (*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 7 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 [40] or virtual for VR) on objects.




(a) (b)

Fig. 8: (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.

**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 8a. It shows that 'hand-off' grasps are more diverse than 'use' grasps in terms of hand pose. We

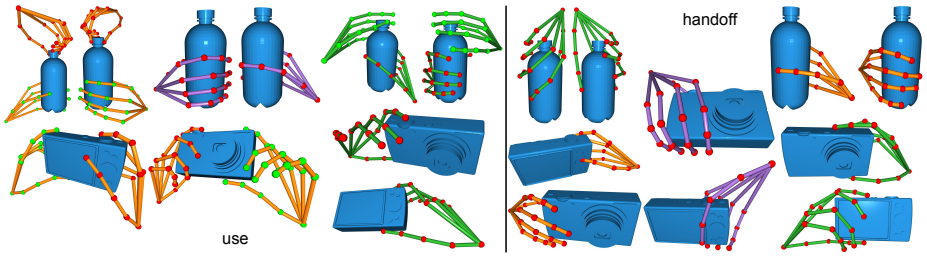



Fig. 9: Examples from hand pose clusters for 'use' and 'hand-off' grasps. Grasps from different clusters are shown **with different colors** (some grasps are bimanual). Left hand joints are **green**, right hand joints are **red**.

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.

Organizing the grasps by clustering these aligned 3D joints (using L2 distance and HDBSCAN [7]) reveals the diversity of grasps captured in ContactPose (Figure 9). 'Hand-off' grasps exhibit a more continuous variation than 'use' grasps, which are tied more closely to the function of the object. The average intra-cluster distance for 'use' grasps is 32.5% less than that for 'handoff' grasps.

Figure 8b shows pair of grasps found by minimizing hand pose distance and maximizing hand contact distance. We use the phalange-level contact association described above. 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. It shows that grasps with similar hand pose can contact different parts of the object and/or hand, inducing different forces and manipulation possibilities [14] 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 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. By providing high-quality contact output from readily available input modalities, such models can enable better hand-object dynamics simulation in AR/VR and soft robotic grasping.

**Object Shape Representation**: We represent the object shape through either a pointcloud densely sampled from the surface (1K-30K points based on size), or a $64^3$ voxel occupancy grid. Features encoding the input hand pose are associated with individual points (voxels). The entire pointcloud (voxel grid) is then processed to predict contact values for points (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, points use the hand with the closest joint.
- `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 MANO hand model geometry. 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 $256\times256$ crops of RGB grasp images using a CNN encoder-decoder inspired by U-Net [45] (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 11a). 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 [30] image. This investigation is complementary to recent work on image-based estimation of object geometry [17,61], object pose [15,53], and hand pose [20,46,50,58,62].

**Contact Representation**: We observed in early experiments that the mean squared error loss resulted in blurred and saturated contact predictions. This might be due to 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 Zhang et al [57]. 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 [57] with $\lambda = 0.4$). Following [57], 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++ [42] architecture implemented in pytorch-geometric [12, 37] (modified to reduce the number of learnable parameters) for pointclouds, and the VoxNet [34]-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. Following [3], hand pose features 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 [24] and batchnorm [27].

**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 $\Psi$ from [1, 54] as a proxy for contact intensity. To account for imperfections in hand modelling (due to rigidity of the MANO mesh) and fitting, we compute $\Psi$ not only for collisions, but also when the hand and object meshes are closer than 1 cm. Finally, we calibrate $\Psi$ 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 | Participant Split | | Object Split | |
|---|---|---|---|---|---|
| | | **AuC (%)** | **Rank** | **AuC (%)** | **Rank** |
| None | Heuristic [1, 54] | 78.31 | 5 | 81.11 | 4 |
| VoxNet [3, 34] | `skeleton` | 77.94 | | 79.99 | |
| MLP | `simple-joints` | 75.11 | | 77.83 | |
| | `relative-joints` | 75.39 | | 78.83 | |
| | `skeleton` | 80.78 | 3 | 80.07 | |
| | `mesh` | 79.89 | 4 | **84.74** | 1 |
| PointNet++ | `simple-joints` | 71.61 | | 73.67 | |
| | `relative-joints` | 74.51 | | 77.10 | |
| | `skeleton` | 81.15 | 2 | 81.49 | 3 |
| | `mesh` | **81.29** | 1 | 84.18 | 2 |
| Image enc-dec, PointNet++ | images (1-view) | 72.89 | | 77.09 | |
| | images (3-view) | 78.06 | | 80.80 | 5 |

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 [57], this value is re-balanced to account for varying occurrence frequencies of values in the 10 contact bins. We create two data splits: the *object split* holds out mug, pan and wine glass following [3], and the *participant split* holds out participants 5, 15, 25, 35, and 45. The held out data is used for evaluation, and models are trained on the rest.

Table 2 shows the re-balanced AuC values averaged over held out data for the two splits. 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 10 shows contact prediction results from hand pose for mug, an unseen object. 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 11b shows contact prediction results from RGB images for mug, an unseen object. 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 finger 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.

Using this contact ground-truth to develop more realistic, deformable hand mesh models could be an interesting research direction. State-of-the-art models (*e.g.* [28, 44]) are rigid, while the human hand is covered with soft tissue. As the Future Work section of [44] notes, they are trained with meshes from

which objects are manually removed, and do not explicitly reason about hand-object contact. ContactPose data can potentially help in the development and evaluation of hand mesh deformation algorithms.

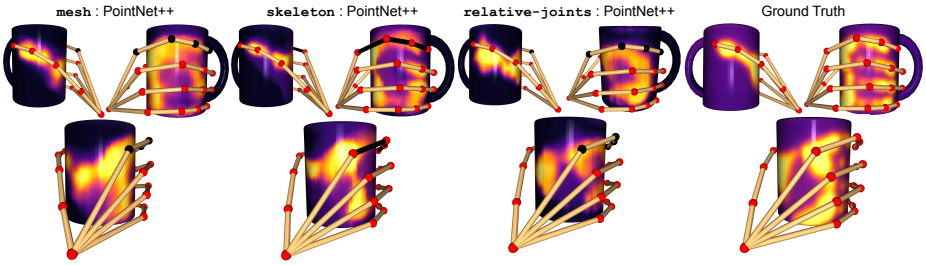



Fig. 10: Contact prediction for mug (an unseen object) from hand pose. All input features related to black line segments and joints were dropped (set to 0). Notice how the `mesh`- and `skeleton`-PointNet++ predictors is able to capture nuances of palm contact, thumb and finger shapes.

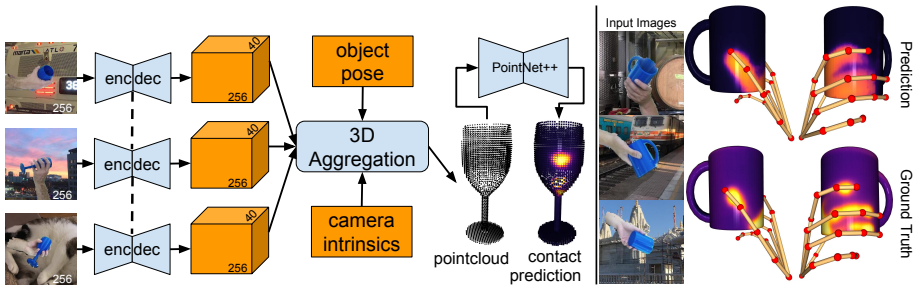



Fig. 11: (a) Image-based contact prediction architecture. (b) Contact prediction for mug (an unseen object) from RGB images, using networks trained with 3 views. Hand poses shown only for reference.

**Acknowledgements**: We are thankful to the anonymous reviewers for helping improve this paper. We would also like to thank Elise Campbell, Braden Copple, David Dimond, Vivian Lo, Jeremy Schichtel, Steve Olsen, Lingling Tao, Sue Tunstall, Robert Wang, Ed Wei, and Yuting Ye for discussions and logistics help.

## References


1. 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, 12
2. Bernardin, K., Ogawara, K., Ikeuchi, K., Dillmann, R.: A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. IEEE Transactions on Robotics **21**(1), 47–57 (2005) 3, 4
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, 7, 8, 9, 12, 13
4. Brahmbhatt, S., Handa, A., Hays, J., Fox, D.: ContactGrasp: Functional Multi-finger Grasp Synthesis from Contact. In: 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, 7
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) 7
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 10
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. Ehsani, K., Tulsiani, S., Gupta, S., Farhadi, A., Gupta, A.: Use the force, luke! learning to predict physical forces by simulating effects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 3
10. 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
11. Ferrari, C., Canny, J.: Planning optimal grasps. In: Proceedings IEEE International Conference on Robotics and Automation. pp. 2290–2295. IEEE (1992) 2
12. Fey, M., Lenssen, J.E.: Fast graph representation learning with PyTorch Geometric. In: ICLR Workshop on Representation Learning on Graphs and Manifolds (2019) 12
13. 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
14. 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, 10
15. Garon, M., Lalonde, J.F.: Deep 6-dof tracking. IEEE transactions on visualization and computer graphics **23**(11), 2410–2418 (2017) 11
16. Glauser, O., Wu, S., Panozzo, D., Hilliges, O., Sorkine-Hornung, O.: Interactive hand pose estimation using a stretch-sensing soft glove. ACM Transactions on Graphics (TOG) **38**(4), 1–15 (2019) 4

17. 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) 11
18. Hamer, H., Gall, J., Weise, T., Van Gool, L.: An object-dependent hand pose prior from sparse training data. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 671–678. IEEE (2010) 3
19. Hamer, H., Schindler, K., Koller-Meier, E., Van Gool, L.: Tracking a hand manipulating an object. In: 2009 IEEE 12th International Conference on Computer Vision. pp. 1475–1482. IEEE 4
20. Hampali, S., Rad, M., Oberweger, M., Lepetit, V.: Honnotate: A method for 3d annotation of hand and object poses. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020) 4, 11
21. 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
22. 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, 8
23. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: IEEE International Conference on Computer Vision (ICCV). pp. 2980–2988 (Oct 2017) 6
24. 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
25. 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
26. Huber, P.J.: Robust estimation of a location parameter. In: Breakthroughs in statistics, pp. 492–518. Springer (1992) 7
27. 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
28. 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) 13
29. Larsen, E., Gottschalk, S., Lin, M.C., Manocha, D.: Fast distance queries with rectangular swept sphere volumes. In: IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065). vol. 4, pp. 3719–3726. IEEE (2000) 4
30. 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) 11
31. 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
32. 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

33. 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: IEEE international conference on robotics and automation (ICRA). pp. 1957–1964. IEEE (2016) 4
34. Maturana, D., Scherer, S.: Voxnet: A 3d convolutional neural network for real-time object recognition. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 922–928. IEEE (2015) 12
35. Miller, A.T., Allen, P.K.: Graspit! a versatile simulator for robotic grasping. IEEE Robotics & Automation Magazine **11**(4), 110–122 (2004) 2, 4
36. Moon, G., Yong Chang, J., Mu Lee, K.: V2V-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In: Proceedings of the IEEE conference on computer vision and pattern Recognition. pp. 5079–5088 (2018) 7
37. 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
38. 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
39. 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) 3
40. 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) 3, 9
41. 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
42. 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
43. 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
44. 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, 13
45. 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) 11
46. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR (2017) 1, 4, 6, 11
47. 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
48. 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

49. SynTouch LLC: BioTac. https://www.syntouchinc.com/robotics/, accessed: 2020-03-05 7
50. 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, 11
51. 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
52. 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
53. Tremblay, J., To, T., Sundaralingam, B., Xiang, Y., Fox, D., Birchfield, S.: Deep object pose estimation for semantic robotic grasping of household objects. In: Conference on Robot Learning (CoRL) (2018), https://arxiv.org/abs/1809.10790 11
54. 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, 12
55. 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
56. Ye, Y., Liu, C.K.: Synthesis of detailed hand manipulations using contact sampling. ACM Transactions on Graphics (TOG) **31**(4), 41 (2012) 4
57. Zhang, R., Isola, P., Efros, A.A.: Colorful image colorization. In: European conference on computer vision. pp. 649–666. Springer (2016) 11, 12, 13
58. 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, 11
59. Zhou, Q.Y., Koltun, V.: Color map optimization for 3d reconstruction with consumer depth cameras. ACM Transactions on Graphics (TOG) **33**(4), 1–10 (2014) 6
60. Zhou, Q.Y., Park, J., Koltun, V.: Open3D: A modern library for 3D data processing. arXiv:1801.09847 (2018) 6
61. 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) 11
62. 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, 11