ContactDB: Analyzing and Predicting Grasp Contact via Thermal Imaging

Samarth Brahmbhatt[1], Cusuh Ham[1], Charles C. Kemp[1] and James Hays[1,2]
[1]Institute for Robotics and Intelligent Machines, Georgia Tech [2]Argo AI
{samarth.robo,cusuh}@gatech.edu, charlie.kemp@bme.gatech.edu, hays@gatech.edu

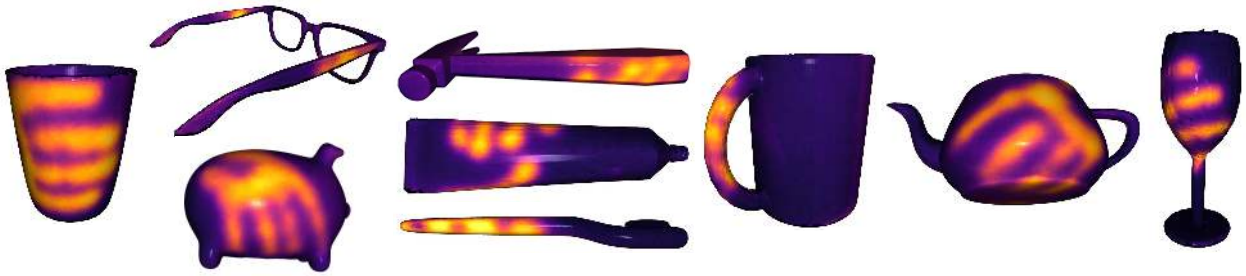


Figure 1: Example contact maps from ContactDB, constructed from multiple 2D thermal images of hand-object contact resulting from human grasps.

Abstract

Grasping and manipulating objects is an important human skill. Since hand-object contact is fundamental to grasping, capturing it can lead to important insights. However, observing contact through external sensors is challenging because of occlusion and the complexity of the human hand. We present ContactDB, a novel dataset of contact maps for household objects that captures the rich hand-object contact that occurs during grasping, enabled by use of a thermal camera. Participants in our study grasped 3D printed objects with a post-grasp functional intent. ContactDB includes 3750 3D meshes of 50 household objects textured with contact maps and 375K frames of synchronized RGB-D+thermal images. To the best of our knowledge, this is the first large-scale dataset that records detailed contact maps for human grasps. Analysis of this data shows the influence of functional intent and object size on grasping, the tendency to touch/avoid 'active areas', and the high frequency of palm and proximal finger contact. Finally, we train state-of-the-art image translation and 3D convolution algorithms to predict diverse contact patterns from object shape. Data, code and models are available at https://contactdb.cc.gatech.edu.

1. Introduction

Humans excel at grasping and then performing tasks with household objects. Human grasps exhibit contact locations, forces and stability that allows post-grasp actions with objects, and are also significantly influenced by the post-grasp intent [8, 2, 45]. For example, people typically grasp a knife by the handle to use it, but grasp it by the blunt side of the blade to hand it off.

A large body of previous work [20, 29, 36, 46, 49, 3, 50, 52, 21, 36, 21, 6, 46] has recorded human grasps, with methods ranging from data gloves that measure joint configuration to manually arranged robotic hands. ContactDB differs significantly from these previous datasets by *focusing primarily on the contact* resulting from the rich interaction between hand and object. Specifically, we represent contact through the texture of 3D object meshes, which we call 'contact maps' (see Figure 1).

There are multiple motivations for recording grasping activity through contact maps. Since it is *object-centric*, it enables detailed analysis of grasping preferences influenced by functional intent, object shape, size and semantic category, and learning object shape features for grasp prediction, and grasp re-targeting to kinematically diverse hand models. Previously employed methods of recording grasping activity do not easily support such analysis, as we discuss in Section 2.

We created ContactDB by recording human participants grasping a set of 3D printed household objects in our laboratory, with two different post-grasp functional intents–using the object and handing it off. See Section 3 for more details on the data collection procedure, size of the dataset and the kinds of data included.

Except for contact edges viewed from select angles, and

contact with transparent objects, contact regions are typically occluded from visual light imaging. Hence, existing studies on the capture and analysis of hand-object contact are extremely limited. Fundamental questions such as the role of the palm in grasping everyday objects are unanswered. We propose a novel procedure to capture contact maps on the object surface at unprecedented detail using an RGB-D + thermal camera calibrated rig.

We make the following contributions in this paper:

- **Dataset**: Present a dataset recording functional human grasping consisting of 3750 meshes textured with contact maps and 375K frames of paired RGBD-thermal data.
- **Analysis**: Demonstrate the influence of object shape, size and functional intent on grasps, and show the importance of non-fingertip contact.
- **Prediction**: Explore data representations and diverse prediction algorithms to predict contact maps from object shape.

2. Related Work

2.1. Datasets of Human Grasps

Since contact between the human hand and an object is fundamental to grasping and manipulation, capturing this contact can potentially lead to important insights about human grasping and manipulation. In practice, however, this has been a challenging goal. The human hand is highly complex with extensive soft tissue and a skeletal structure that is often modeled with 26 degrees of freedom. Hence, previous work has focused on recording grasping activity in other forms like hand joint configuration by manual annotation [49, 3], data gloves [20, 29] or wired magnetic trackers [54, 16] (which can interfere with natural grasping), or model-based hand pose estimation [50]. At a higher level, grasping has been observed through third-person [52, 21, 36] or first-person [21, 6, 46] videos, in which frames are annotated with the category of grasp according to a grasp taxonomy [12, 23]. Tactile sensors are embedded on a glove [4] or in the object [38] to record grasp contact points. Such methods are limited by the resolution of tactile sensors. Puhlmann et al [39] capture hand-table contact during grasping with a touchscreen. Rogez et al [42] manually configure a hand model to match grasps from a taxonomy, and use connected component analysis on hand vertices intersecting with an object model to estimate contact regions on the hand.

Due to hand complexity and lack of understanding of how humans control their hands, approaches like those mentioned above have so far been limited to providing coarse or speculative contact estimates. In contrast, our approach allows us to directly observe where contact between the object and the human hand has taken place with an unprecedented level of fidelity.

2.2. Predicting Grasp Contact

Our work is related to that of Lau et al [26], which crowdsources grasp tactile saliency. Online annotators are instructed to choose a point they would prefer to touch, from a pair sampled from the object surface. This pairwise information is integrated to construct the tactile saliency map. In contrast, ContactDB contact maps are full observations of real human grasps with functional intent (see supplementary material for a qualitative comparison). Akizuki et al [1] use hand pose estimation and model-based object tracking in RGB-D videos to record a set of contact points on the object surface. This is vulnerable to inaccuracies in the hand model and hand pose tracking. Hamer at al [19] record human demonstrations of grasping by registering depth images to get object geometry and object- and hand-pose. Contact is approximated as a single point per fingertip. A large body of work in robotics aims to predict a configuration of the end-effector [32, 9, 28] suitable for grasping. In contrast to ContactDB, these works model contact as a single point per hand digit, ignoring other contact.

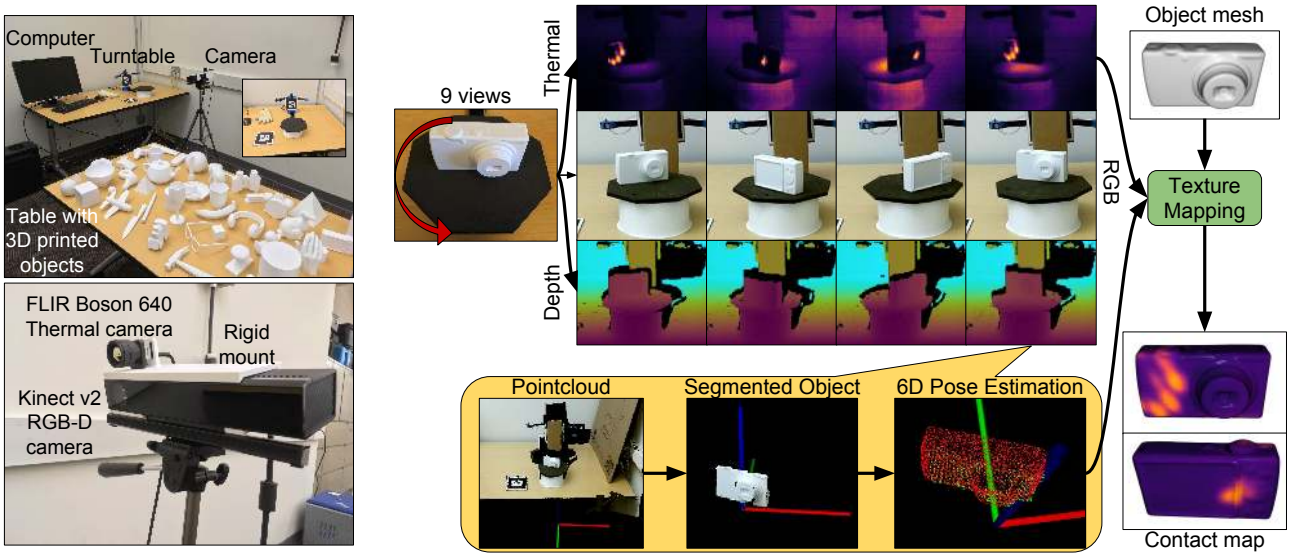
Diverse Predictions: Grasping is a task where multiple predictions can be equally correct. Lee et al [27] and Firman et al [14] have developed theoretical frameworks allowing neural networks to make diverse and meaningful predictions. Recently, Ghazaei et al [17] have used similar techniques to predict diverse grasp configurations for a parallel jaw gripper.

3. The ContactDB Dataset

Here we present the design choices and process in creating the ContactDB, which consists of 50 3D printed household objects being grasped with two functional intents by 50 participants (see Table 1).

Observing Contact Through a Thermal Camera. At the core of our data collection process is the use of a thermal camera to observe the precise locations of contact between human hand and object. Thermal cameras have recently been used to capture humans and their interaction with the environment. For example, Luo et al [31] observe humans interacting with objects for egocentric SLAM, while Larson et al [25] observe human finger interaction with arbitrary surfaces to make them interactive. Both note the phenomenon of thermally observable contact, but do not investigate it rigorously or collect a large-scale dataset.

When a participant grasps an object, heat from the hand transfers onto the object surface. If the object material does not dissipate the heat rapidly, the precise contact areas can be clearly observed in the thermal image after the object is released (see Figure 2b). Intensity at a pixel in the thermal image is a function of the infrared energy emitted by the corresponding world point [51]. Hence, object pixel in-



(a) Data collection area setup

(b) Data processing pipeline, explained in detail in Section 3.3

Figure 2: Data collection and processing for ContactDB. Participants grasp 3D printed objects and put them on the rotating turntable. Thermal images from multiple views are texture-mapped to the object mesh.

	Functional Intent		Total
	Use	Hand-off	
Participants	50	50 (same)	
Objects	27	48 (overlapping)	50
Textured meshes	1350	2400	**3750**
RGBD-Thermal frames	135K	240K	**375K**

Table 1: Size of the ContactDB Dataset

tensity in our thermal images is related to heat of the skin, duration of contact, heat conduction (including diffusion to nearby object locations), and contact pressure. By keeping these factors roughly constant during data collection, we verified empirically that heat conduction from hand-object contact is the dominant factor in the observed thermal measurements. See the supplementary material for more discussion on heat dissipation and accuracy.

3.1. Object Selection and Fabrication

We decided to focus on household objects since an understanding of contact preferences and the ability to predict them are most likely to improve human-robot interaction in household settings. Other standard grasping datasets [7] and competitions [10] have a similar focus. We started with the YCB dataset [7] to choose the 50 objects in our dataset. We excluded similarly-shaped objects (e.g. cereal and cracker boxes) that are unlikely to produce different kinds of grasps, deformable objects (e.g. sponge, plastic chain, nylon rope), very small (e.g. dominoes, washers), and very large objects (e.g. cooking skillet, Windex bottle). We added common ones such as flashlight, eyeglasses, computer mouse, and objects popular in computer graphics (e.g. Stanford bunny and Utah teapot). Since object size has been shown to influence the grasp [11, 8] and we are interested in contact during grasping of abstract shapes, we included 5 primitive objects–cube, cylinder, pyramid, torus and sphere–at 3 different scales (principal axes 12, 8 and 4 cm). See the supplementary material for a full object list.

We chose to 3D print all the objects to ensure uniform heat dissipation properties. Additionally, we empirically found that the PLA material used for 3D printing is excellent for retaining thermal handprints. We used open-source resources to select suitable models for each object, and printed them at 15% infill density using white PLA filament on a Dremel 3D20 printer. 3D printing the objects has additional advantages. Having an accurate 3D model of the object makes 6D pose estimation of the object from recorded pointcloud data easier (see Section 3.3), which we use for texture mapping contact maps to the object mesh. 3D printing the objects also allows participants to focus on the object geometry during grasping.

3.2. Data Collection Protocol

Figure 2a shows our setup. We rigidly mounted a FLIR Boson 640 thermal camera on a Kinect v2 RGB-D sensor. The instrinsics of both the cameras and extrinsics between them are calibrated using ROS [41], so that both RGB and depth images from the Kinect can be accurately registered to the thermal image. We invited 50 participants (mostly 20-25 years of age, able-bodied males and females), and

used the following protocol approved by the Georgia Tech Institutional Review Board.

50 3D printed objects were placed at random locations on a table in orientations commonly encountered in practice. Participants were asked to grasp each object with a post-grasp functional intent. They held the object for 5 seconds to allow heat transfer from the hand to the object, and then hand it to an experimenter. The experimenter wore an insulating glove to prevent heat transfer from their hand, and places the object on a turntable about 1 m away from the cameras. Participants were provided with chemical hand warmers to increase the intensity of thermal handprints. The cameras recorded a continuous stream of RGB, depth and thermal images as the turntable rotated in a 360 degree arc. The turntable paused at 9 equally spaced locations on this arc, where the rotation angle of the turntable was also recorded. In some cases, objects were flipped and scanned a second time to capture any thermal prints that were unseen in the previous rotation.

We used two post-grasp *functional intents*: 'use' and 'hand-off'. Participants were instructed to grasp 48 objects with the intent of handing them off to the experimenter, and to grasp a subset of 27 objects (after the previous thermal handprints had dissipated) with the intent of using them. We used only a subset of 27 objects for 'use', since other objects (e.g. pyramid, Stanford bunny) lack clear use cases. See the supplementary material for specific use instructions. Participants were asked to avoid in-hand manipulation after grasping to avoid smudging the thermal handprints.

3.3. Data Processing

As the turntable rotates with the object on it, the stream of RGB-D and thermal images capture the object from multiple viewpoints. The aim of data processing is to texture-map the thermal images to the object 3D mesh and generate a coherent contact map (examples are shown in Figure 1).

The entire process is shown in Figure 2b. We first extracted the corresponding turntable angle and RGB, depth and thermal images at the 9 locations where the turntable pauses. Next, we converted the depth maps to pointclouds and useed a least-squares estimate of the turntable plane and white color segmentation to segment the object. We used the Iterative Closest Point (ICP) [5] algorithm implemented in PCL [44] to estimate the full 6D pose of the object in the 9 segmented pointclouds. Object origins in the 9 views were used to get a least squares estimate of the 3D circle described by the moving object. This circle was used to interpolate the object poses for views which are unsuitable for the ICP step because of noise in the depth map or important shape elements of the object being hidden in that view, or for rotating symmetric objects around the axis of symmetry.

Finally, the 3D mesh along with the 9 pose estimates and thermal images were input to the colormap optimization algorithm of [55], which is implemented in Open3D [56]. It locally optimizes object poses to minimize the photometric texture projection error and generates a mesh coherently textured with contact maps.

Active Area	handoff	use
Banana tip (either tip)	22.45	63.27
Binoculars (both barrels)	12.50	93.88
Camera shutter button	34.00	69.39
Eyeglasses (both temples)	4.00	64.58
Flashlight button	28.00	62.00
Hammer (head)	38.00	0.00
Mouse (both click buttons)	16.00	84.00
PS controller (both front buttons)	2.00	40.81
PS controller (both analog sticks)	2.00	22.44
Scissors (handle)	38.00	100.00
Scissors (blade)	60.00	0.00
Water-bottle cap	16.00	67.35
Wine glass stem	56.00	30.61

Table 2: Fraction of participants that touched active areas for different functional intents. See Fig. 3 for examples.

4. Analysis of Contact Maps

In this section we present analysis of some aspects of human grasping, using the data in ContactDB. We processed each contact map separately to increase contrast by applying a sigmoid function to the texture-mapped intensity values that maps the minimum to 0.05 and maximum to 0.95.

Effect of Functional Intent. We observed that the functional intent ('use' or 'hand off') significantly influences the contact patterns for many objects. To show qualitative examples, we clustered the contact maps within each object and functional intent category using k-medoids clustering [24] ($k = 3$) on the XYZ values of points which have contact value above 0.4. The distance function between two sets of points was defined as $d(\mathbf{p}_1, \mathbf{p}_2) = \left(\bar{d}(\mathbf{p}_1, \mathbf{p}_2) + \bar{d}(\mathbf{p}_2, \mathbf{p}_1)\right) / \left(|\mathbf{p}_1| + |\mathbf{p}_2|\right)$, where $\bar{d}(\mathbf{p}_1, \mathbf{p}_2) = \sum_{i=1}^{|\mathbf{p}_1|} \min_{j=1}^{|\mathbf{p}_2|} ||\mathbf{p}_1^{(i)} - \mathbf{p}_2^{(j)}||_2$. For symmetric objects, we chose the angle of rotation around the axis of symmetry that minimized $d(\mathbf{p}_1, \mathbf{p}_2)$. Figure 3 shows dominant contact maps (center of the largest cluster) for the two different functional intents.

To quantify the influence of functional intent, we define 'active areas' (highlighted in green in Figure 3) on the surface of some objects and show the fraction of participants that touched that area (evidenced by the map value being greater than 0.4) in Table 2.

Effect of object size. Figure 4 shows the dominant contact maps for objects of the same shape at three different sizes. Small objects exhibit grasps with two or three fingertips, while larger objects are often grasped with more fingers and more than the fingertips in contact with the ob-

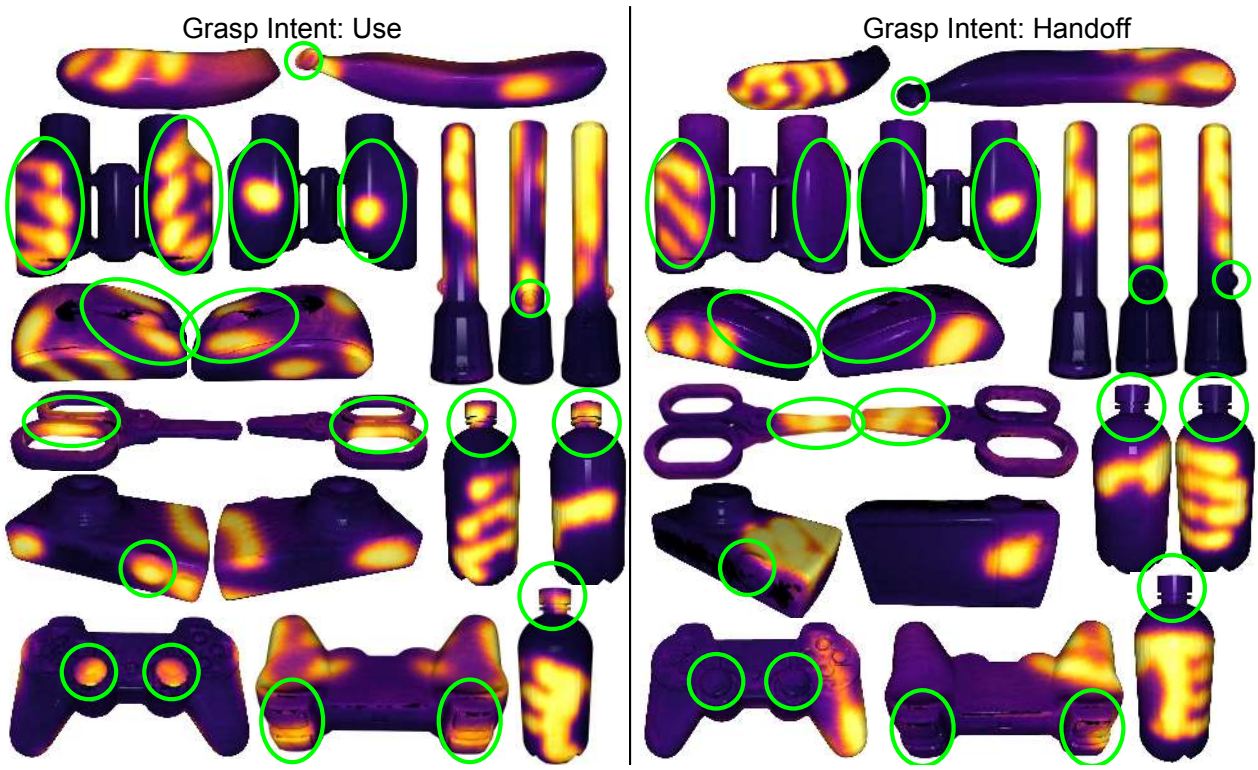


Figure 3: Influence of functional intent on contact: Two views of the dominant grasp (center of the largest cluster after k-medoids clustering across participants). Green circles indicate 'active areas'. This influence is quantified in Table 2.

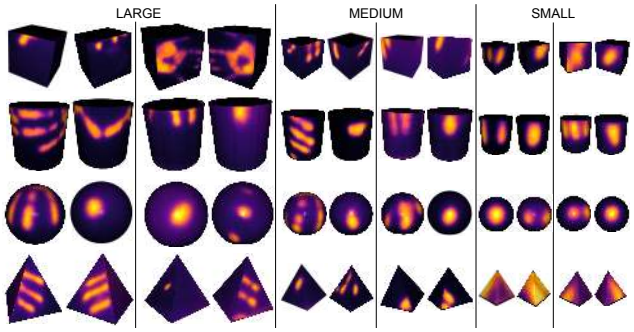


Figure 4: Influence of object size on contact: Two dominant grasps for objects of same shape and varying size.

ject. Grasps for large objects are bi-modal: bimanual using the full hands, or single-handed using fingertips. To quantify this, we manually labelled grasps as bimanual/single-handed, and show their relation to hand size in Fig. 6. The figure shows that people with smaller hands prefer to grasp `large` objects (for 'handoff') with bimanual grasps. No bimanual grasps were observed for the `medium` and `small` object sizes.

How much of the contact is fingertips? Contact is traditionally modelled in robotics [47] and simulation [53] as a single point. However, the contact maps in Figures 1, 3 and 4 show that human grasps have much more than fingertip contact. Single-point contact modeling is inspired by the prevalence of rigid manipulators on robots, but with the recent research interest in *soft robots* [13, 15], we now have access to manipulators that contact the object at other areas on the finger. Data in ContactDB shows the use of non-fingertip contact for highly capable soft manipulators: human hands. For each contact map, we calculated the contact area by integrating the area of all the contacted faces in the mesh. A face is contacted if any of its three vertices have a contact value greater than 0.4. Figures 5(b) and 5(c) show the contact areas for all objects under both functional intents, averaged across participants. Next, we calculated an upper bound on the contact area if only all 5 fingertips were touching the object. This was done by capturing the participants' palm print on a flat plate, where it is easy to manually annotate the fingertip regions (shown in Figure 5(a)). The total surface area of fingertips in the palm print is the desired upper bound. It was doubled for objects for which we observe bimanual grasps. This upper bound was averaged across four participants, and is shown as the red line in Figures 5(b) and 5(c). Note that this is a loose upper bound, since many real-world fingertip-only grasps don't involve all five fingertips, and we mark the entire object category

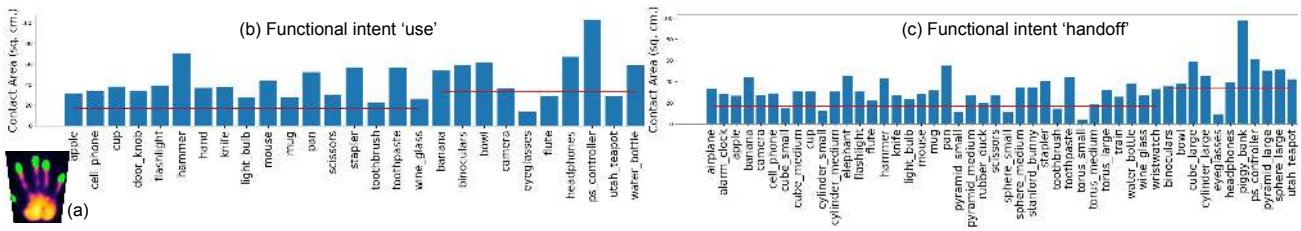


Figure 5: (a): Palm contact on plate, annotated fingertips. (b, c): Contact areas for objects in ContactDB, averaged across participants. The red line indicates a loose upper bound on contact area for a fingertip-only grasp, which is doubled for objects which have bimanual grasps.

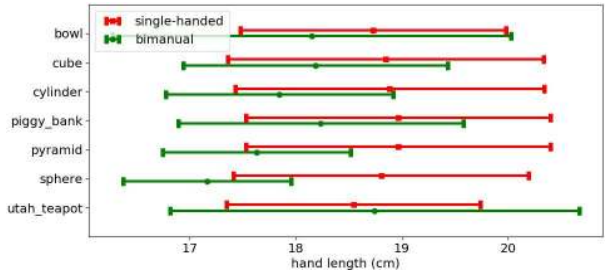


Figure 6: Relationship between hand length (wrist to mid fingertip) and single-handed/bimanual grasps. The intervals show mean and 1 standard deviation. Cube, cylinder, pyramid and sphere are of the `large` size.

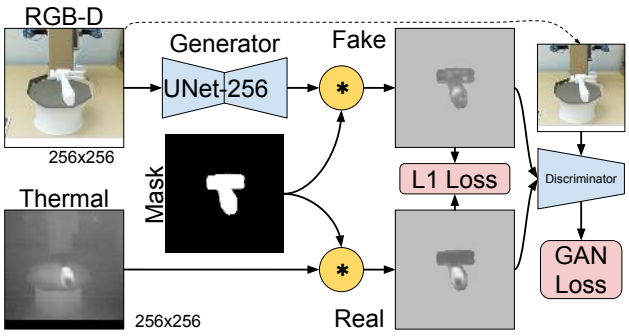


Figure 7: Training procedure for single-view contact map prediction. The discriminator has 5 conv layers followed by batch norm and leaky ReLU.

as bimanual if even one participant performs a bimanual grasp. Total contact area for many objects is significantly higher than the upper bound on fingertip-only contact area, indicating the large role that the soft tissue of the human hand plays in grasping and manipulation. This motivates the inclusion of non-fingertip areas in grasp prediction and modeling algorithms, and presents an opportunity to inform the design of soft robotic manipulators. Interestingly, the average contact area for some objects (e.g. bowl, mug, PS controller, toothbrush) differs across functional intent, due to different kinds of grasps used.

5. Predicting Contact Maps

In this section, we describe experiments to predict contact maps for objects based on their shape. ContactDB is the first large scale dataset that enables training data-intensive deep learning models for this task. Since ContactDB includes diverse contact maps for each object, the mapping from object shape to contact map is one-to-many and makes the task challenging. We explore two representations for object shape: single-view RGB-D, and full 3D. Since the contact patterns are significantly influenced by the functional intent, we train separate models for 'hand-off' and 'use'.

5.1. Single-view Prediction

Object shape is represented by an RGB-D image, and a 2D contact map is predicted for the visible part of the object. A single view might exclude information about important aspects of the object shape, and 'interesting' parts of the contact map might lie in the unseen half of the object. However, this representation has the advantage of being easily applicable to real-world robotics scenarios where mobile manipulators are often required to grasp objects after observing them from a single view. We used generative adversarial network (GAN)-based image-to-image translation [22, 57, 30] for this task, since the optimization procedure of conditional GANs is able to model a one-to-many input-output mapping [35, 18].

Figure 7 shows our training procedure and network architecture, which has roughly 54M and 3M parameters in the generator and discriminator respectively. We modified pix2pix [22] to accept a 4-channel RGB-D input and predict a single-channel contact map. The RGB-D stream from object scanning was registered to the thermal images, and used as input. Thermal images were used as a proxy for the single-view contact map. To focus the generator and discriminator on the object, we cropped a 256×320 patch around the object and masked all images by the object sil-

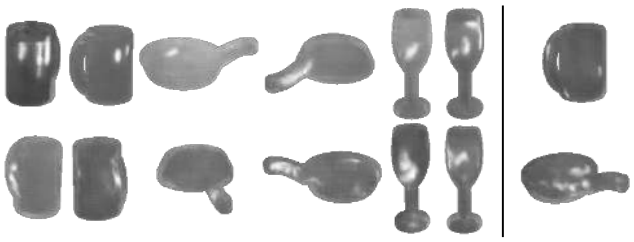


Figure 8: Single-view predictions from the pix2pix model for three *unseen* object classes: mug, pan and wine glass. Top: handoff intent, bottom: use intent. Rightmost column: uninterpretable predictions.

houette. All images from mug, pan, and wineglass were held out and used for testing. Figure 8 shows some predicted contact maps for these unseen objects, selected for looking realistic. Mug predictions for use have finger contact on the handle, whereas contact is observed over the top for handoff. Pan use predictions show grasps at the handle, while handoff predictions additionally show a bimanual grasp of the handle and side. Similarly, the wine glass indicates contact with a side grasp for use and over the opening for handoff.

5.2. 3D Prediction

Full 3D representation gives access to the entire shape of the object, and alleviates the view-consistency problems observed during single-view prediction.

Learning a one-to-many-mapping. Stochastic Multiple Choice Learning [27] (sMCL) trains an ensemble of k predictors to generate k contact maps for each input (see Figure 9a). Each input has multiple equally correct ground truth maps. During training, the loss is backpropagated from each ground truth contact map to the network that makes the prediction closest to it. To encourage all members of the ensemble to be trained equally, as mentioned in [43], we made this association soft by routing the gradient to the closest network with a 0.95 weight and distributed the rest equally among other members of the ensemble, and randomly dropped entire predictions with a 0.1 probability. We trained models with $k = 1$ and $k = 10$.

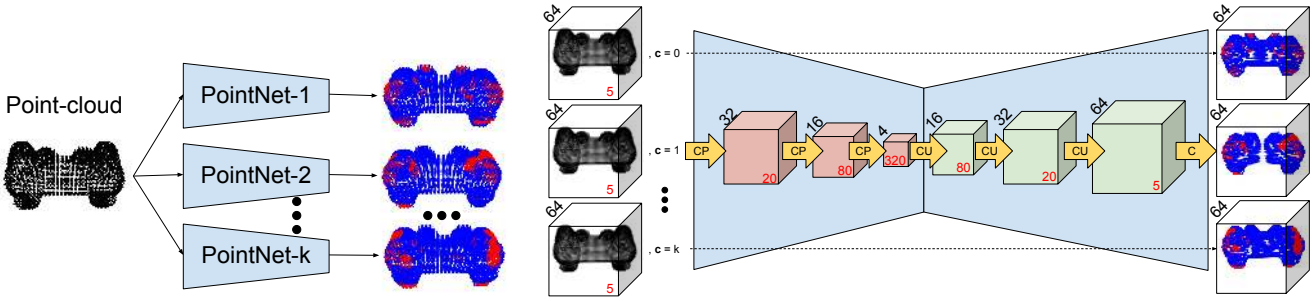
In contrast, DiverseNet [14] generates diverse predictions from a single predictor network by changing the value of a one-hot encoded control variable $\mathbf{c}$ that is concatenated to internal feature maps of the network (See Figure 9b). Each ground truth contact map is associated with the closest prediction and gradients are routed through the appropriate $\mathbf{c}$ value. Diverse predictions can be generated at test time by varying $\mathbf{c}$. Compared to sMCL, DiverseNet requires significantly fewer trainable parameters. We used 10 one-hot encoded $\mathbf{c}$ values in our experiments.

3D representation. We represented the 3D object shape in two forms: pointcloud and voxel occupancy grid. PointNet [40] operates on a pointcloud representation of the object shape, with points randomly sampled from the object surface. We normalized the XYZ position of each point to fit the object in a unit cube. The XYZ position and the normalization scale factor were used as 4-element features for each point. The network was trained by cross entropy loss to predict whether each voxel is in contact. We used a PointNet architecture with a single T-Net and 1.2M parameters.

VoxNet [33] operates on a solid occupancy grid of the object in a 64^3 voxelized space, and predicts whether each voxel is contacted. It uses 3D convolutions to learn shape features. The four features used for PointNet were used in addition to the binary occupancy value to form a 5-element feature vector for each voxel. Cross entropy loss was enforced only on the voxels on the object surface. The network architecture is shown in Figure 9b, and has approximately 1.2M parameters.

Experiments We conducted experiments with both VoxNet and PointNet, using the sMCL and DiverseNet strategies for learning a one-to-many-mapping. For DiverseNet, we concatenated $\mathbf{c}$ to the output of the first and fifth conv layers in VoxNet, and to the input transformed by T-Net and the output of the second-last MLP in PointNet. Voxelization of the meshes was done using the algorithm of [37] implemented in binvox [34]. The PointNet input was generated by randomly sampling 3000 points from the object surface. We thresholded the contact maps at 0.4 after applying the sigmoid described in Section 4, to generate ground truth for classification. We augmented the dataset by randomly rotating the object around the yaw axis. PointNet input was also augmented by randomly choosing an axis and scaling the points along that axis by a random factor in [0.6, 1.4]. Dropout with $p = 0.2$ was applied to VoxNet-DiverseNet input. We found that similar dropout did not improve results for other models. Random sampling of surface points automatically acts like dropout for PointNet models, and sMCL models already incorporate a different dropout strategy as mentioned in Section 5.2. The cross entropy loss for contacted voxels was weighted by a factor of 10, to account for class imbalance. All models were trained with SGD with a learning rate of 0.1, momentum of 0.9 and weight decay of 5e-4. Batch size was 5 for models with $k = 10$, and 25 for models with $k = 1$.

Table 3 shows results on held-out test objects (mug, pan and wine glass). We conclude that the voxel occupancy grid representation is better for this task, and that a model limited to making a single prediction does not capture the complexity in ContactDB. Figures 10a and 10b show some of the 'use' intent predictions for unseen object classes and unseen shapes of training object classes respectively, selected for looking realistic. Mug predictions show horizontal grasps around the body. Predictions for the pan are



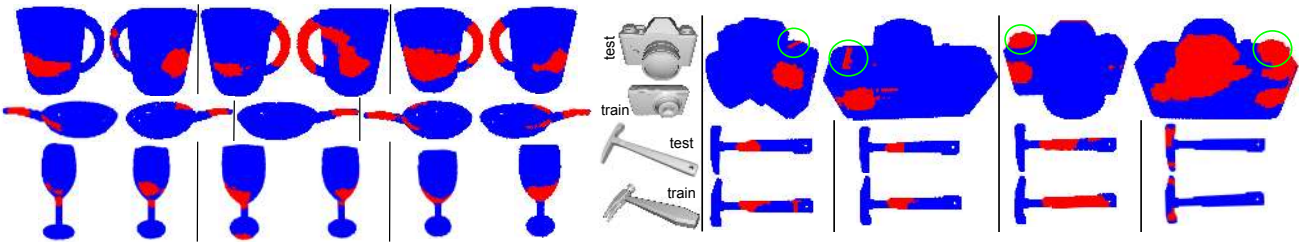
(a) sMCL with a PointNet predictor

(b) DiverseNet with a VoxNet predictor. CP: 3^3 conv with batch norm, ReLU and max pooling, CU: 3^3 conv with batch norm, ReLU and nearest neighbor upsampling. Black numbers: size of voxel grid, red numbers: number of channels.

Figure 9: 3D data representations and training strategies for predicting diverse contact maps. sMCL [27] requires multiple instances of a network, while DiverseNet [14] uses a single instance with an integer valued control variable. PointNet [40] operates on unordered point-clouds, whereas VoxNet [33] uses voxel occupancy grids.

Test object	Handoff						Use					
	sMCL ($k=1$)		sMCL ($k=10$)		DiverseNet ($k=10$)		sMCL ($k=1$)		sMCL ($k=10$)		DiverseNet ($k=10$)	
	VoxNet	PointNet	VoxNet	PointNet	VoxNet	PointNet	VoxNet	PointNet	VoxNet	PointNet	VoxNet	PointNet
pan	76.80	-	**7.13**	20.43	8.48	19.68	17.22	-	8.25	43.57	**5.12**	22.58
wine glass	59.37	-	**11.11**	14.59	28.69	17.28	50.18	-	11.06	14.79	13.98	**10.47**
mug	29.93	-	16.68	27.10	**15.77**	21.60	66.03	-	32.51	31.30	**7.06**	32.41
average	55.37	-	**11.64**	20.71	17.65	19.52	44.48	-	17.27	29.89	**8.72**	21.82

Table 3: Diverse 3D contact map prediction errors (%) for the models presented in Section 5.2. Errors were calculated by matching each ground truth contact map with the closest from k diverse predictions, discarding predictions with no contact. '-' indicates that no contact was predicted.



(a) Contact map predictions for unseen object classes

(b) Contact map predictions for an unseen shape of training object classes

Figure 10: Two views of diverse 3D contact map predictions. (a) *Unseen* object classes: mug, pan, and wine glass, (b) *Unseen shape* of training object classes: camera and hammer. Intent: use, Model: VoxNet-DiverseNet, Red: contact.

concentrated at the handle, with one grasp being bimanual. Wine glass predictions show grasps at the body-stem intersection. Camera predictions show contact at the shutter button and sides, while predictions for the hammer show contact at the handle (and once at the head).

6. Conclusion and Future Work

We presented ContactDB, the first large-scale dataset of contact maps from functional grasping, analyzed the data to reveal interesting aspects of grasping behavior, and explored data representations and training strategies for predicting contact maps from object shape. We hope to spur future work in multiple areas. Contact patterns could inform the design of soft robotic manipulators by aiming to be able to cover object regions touched by humans. Research indicates that in some situations hand pose can be guided by contact points [53, 48]. Using contact maps to recover and/or assist in predicting the hand pose in functional grasping is an exciting problem for future research.

Acknowledgements: We would like to thank Varun Agrawal for lending the 3D printer, Ari Kapusta for initial discussions about thermal cameras, and NVIDIA for a Titan Xp GPU grant.

References

[1] Shuichi Akizuki and Yoshimitsu Aoki. Tactile logging for understanding plausible tool use based on human demonstration. In *British Machine Vision Conference 2018, BMVC 2018, Northumbria University, Newcastle, UK, September 3-6, 2018*, page 334, 2018. 2

[2] Caterina Ansuini, Livia Giosa, Luca Turella, Gianmarco Altoè, and Umberto Castiello. An object for an action, the same object for other actions: effects on hand shaping. *Experimental Brain Research*, 185(1):111–119, 2008. 1

[3] Ravi Balasubramanian, Ling Xu, Peter D Brook, Joshua R Smith, and Yoky Matsuoka. Physical human interactive guidance: Identifying grasping principles from human-planned grasps. *IEEE Transactions on Robotics*, 4(28):899–910, 2012. 1, 2

[4] Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruediger Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *IEEE Transactions on Robotics*, 21(1):47–57, 2005. 2

[5] PJ Besl and Neil D McKay. A method for registration of 3-d shapes. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(2):239–256, 1992. 4

[6] Ian M Bullock, Thomas Feix, and Aaron M Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *The International Journal of Robotics Research*, 34(3):251–255, 2015. 1, 2

[7] Berk Calli, Aaron Walsman, Arjun Singh, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Benchmarking in manipulation research: The ycb object and model set and benchmarking protocols. *arXiv preprint arXiv:1502.03143*, 2015. 3

[8] Umberto Castiello. The neuroscience of grasping. *Nature Reviews Neuroscience*, 6(9):726, 2005. 1, 3

[9] Changhyun Choi, Wilko Schwarting, Joseph DelPreto, and Daniela Rus. Learning object grasping for soft robot hands. *IEEE Robotics and Automation Letters*, 2018. 2

[10] Nikolaus Correll, Kostas E Bekris, Dmitry Berenson, Oliver Brock, Albert Causo, Kris Hauser, Kei Okada, Alberto Rodriguez, Joseph M Romano, and Peter R Wurman. Analysis and observations from the first amazon picking challenge. *IEEE Transactions on Automation Science and Engineering*, 15(1):172–188, 2018. 3

[11] Raymond H Cuijpers, Jeroen BJ Smeets, and Eli Brenner. On the relation between object shape and grasping kinematics. *Journal of Neurophysiology*, 91(6):2598–2606, 2004. 3

[12] Mark R Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE Transactions on robotics and automation*, 5(3):269–279, 1989. 2

[13] Raphael Deimel and Oliver Brock. A novel type of compliant and underactuated robotic hand for dexterous grasping. *The International Journal of Robotics Research*, 35(1-3):161–185, 2016. 5

[14] Michael Firman, Neill DF Campbell, Lourdes Agapito, and Gabriel J Brostow. Diversenet: When one right answer is not enough. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5598–5607, 2018. 2, 7, 8

[15] Kevin C Galloway, Kaitlyn P Becker, Brennan Phillips, Jordan Kirby, Stephen Licht, Dan Tchernov, Robert J Wood, and David F Gruber. Soft robotic grippers for biological sampling on deep reefs. *Soft robotics*, 3(1):23–33, 2016. 5

[16] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 2

[17] Ghazal Ghazaei, Iro Laina, Christian Rupprecht, Federico Tombari, Nassir Navab, and Kianoush Nazarpour. Dealing with ambiguity in robotic grasping via multiple predictions. *arXiv preprint arXiv:1811.00793*, 2018. 2

[18] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014. 6

[19] Henning Hamer, Juergen Gall, Thibaut Weise, and Luc Van Gool. An object-dependent hand pose prior from sparse training data. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 671–678. IEEE, 2010. 2

[20] Guido Heumer, Heni Ben Amor, Matthias Weber, and Bernhard Jung. Grasp recognition with uncalibrated data gloves-a comparison of classification methods. In *Virtual Reality Conference, 2007. VR'07. IEEE*, pages 19–26. IEEE, 2007. 1, 2

[21] De-An Huang, Minghuang Ma, Wei-Chiu Ma, and Kris M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[22] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976. IEEE, 2017. 6

[23] Noriko Kamakura, Michiko Matsuo, Harumi Ishii, Fumiko Mitsuboshi, and Yoriko Miura. Patterns of static prehension in normal hands. *American Journal of Occupational Therapy*, 34(7):437–445, 1980. 2

[24] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids*. North-Holland, 1987. 4

[25] Eric Larson, Gabe Cohn, Sidhant Gupta, Xiaofeng Ren, Beverly Harrison, Dieter Fox, and Shwetak Patel. Heatwave: Thermal imaging for surface user interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2565–2574, New York, NY, USA, 2011. ACM. 2

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

[27] Stefan Lee, Senthil Purushwalkam Shiva Prakash, Michael Cogswell, Viresh Ranjan, David Crandall, and Dhruv Batra.

Stochastic multiple choice learning for training diverse deep ensembles. In *Advances in Neural Information Processing Systems*, pages 2119–2127, 2016. 2, 7, 8

[28] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015. 2

[29] Yun Lin and Yu Sun. Grasp planning based on strategy extracted from demonstration. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 4458–4463. IEEE, 2014. 1, 2

[30] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*, pages 700–708, 2017. 6

[31] Rachel Luo, Ozan Sener, and Silvio Savarese. Scene semantic reconstruction from egocentric rgb-d-thermal videos. In *2017 International Conference on 3D Vision (3DV)*, pages 593–602. IEEE, 2017. 2

[32] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *arXiv preprint arXiv:1703.09312*, 2017. 2

[33] Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 922–928. IEEE, 2015. 7, 8

[34] Patrick Min. binvox. http://www.patrickmin.com/binvox, 2004 - 2017. Accessed: 2018-11-16. 7

[35] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *CoRR*, abs/1411.1784, 2014. 6

[36] Yuzuko C Nakamura, Daniel M Troniak, Alberto Rodriguez, Matthew T Mason, and Nancy S Pollard. The complexities of grasping in the wild. In *Humanoid Robotics (Humanoids), 2017 IEEE-RAS 17th International Conference on*, pages 233–240. IEEE, 2017. 1, 2

[37] Fakir S. Nooruddin and Greg Turk. Simplification and repair of polygonal models using volumetric techniques. *IEEE Transactions on Visualization and Computer Graphics*, 9(2):191–205, 2003. 7

[38] Tu-Hoa Pham, Nikolaos Kyriazis, Antonis A Argyros, and Abderrahmane Kheddar. Hand-object contact force estimation from markerless visual tracking. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2883–2896, 2018. 2

[39] Steffen Puhlmann, Fabian Heinemann, Oliver Brock, and Marianne Maertens. A compact representation of human single-object grasping. In *2016 IEEE International Conference on Intelligent Robots and Systems (IROS)*, page 1954–1959. IEEE, 2016. 2

[40] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 652–660, 2017. 7, 8

[41] Morgan Quigley, Josh Faust, Tully Foote, and Jeremy Leibs. Ros: an open-source robot operating system. 3

[42] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3889–3897. IEEE, 2015. 2

[43] Christian Rupprecht, Iro Laina, Robert DiPietro, Maximilian Baust, Federico Tombari, Nassir Navab, and Gregory D Hager. Learning in an uncertain world: Representing ambiguity through multiple hypotheses. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3591–3600, 2017. 7

[44] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011. 4

[45] Luisa Sartori, Elisa Straulino, and Umberto Castiello. How objects are grasped: the interplay between affordances and end-goals. *PloS one*, 6(9):e25203, 2011. 1

[46] Artur Saudabayev, Zhanibek Rysbek, Raykhan Khassenova, and Huseyin Atakan Varol. Human grasping database for activities of daily living with depth, color and kinematic data streams. *Scientific data*, 5, 2018. 1, 2

[47] Tanner Schmidt, Katharina Hertkorn, Richard Newcombe, Zoltan Marton, Michael Suppa, and Dieter Fox. Depth-based tracking with physical constraints for robot manipulation. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*, pages 119–126. IEEE, 2015. 5

[48] Srinath Sridhar, Franziska Mueller, Michael Zollhöfer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. Real-time joint tracking of a hand manipulating an object from rgb-d input. In *European Conference on Computer Vision*, pages 294–310. Springer, 2016. 8

[49] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014. 1, 2

[50] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014. 1, 2

[51] Michael Vollmer and Klaus-Peter Möllmann. *Infrared thermal imaging: fundamentals, research and applications*. John Wiley & Sons, 2017. 2

[52] Yezhou Yang, Cornelia Fermuller, Yi Li, and Yiannis Aloimonos. Grasp type revisited: A modern perspective on a classical feature for vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 1, 2

[53] Yuting Ye and C Karen Liu. Synthesis of detailed hand manipulations using contact sampling. *ACM Transactions on Graphics (TOG)*, 31(4):41, 2012. 5, 8

[54] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2. 2m benchmark: Hand pose dataset and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4866–4874, 2017. 2

[55] Qian-Yi Zhou and Vladlen Koltun. Color map optimization for 3d reconstruction with consumer depth cameras. *ACM Transactions on Graphics (TOG)*, 33(4):155, 2014. 4

[56] Qian-Yi Zhou, Jaesik Park, and Vladlen Koltun. Open3D: A modern library for 3D data processing. *arXiv:1801.09847*, 2018. 4

[57] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *IEEE International Conference on Computer Vision*, 2017. 6