# Integrating High-Resolution Tactile Sensing into Grasp Stability Prediction

Lachlan Chumbley<sup>1</sup>, Morris Gu<sup>1</sup>, Rhys Newbury<sup>1</sup>, Jürgen Leitner<sup>2</sup> and Akansel Cosgun<sup>1</sup> <sup>1</sup>Monash University, Australia <sup>2</sup>LYRO Robotics, Australia

{lachlan.chumbley,morris.gu,rhys.newbury,akansel.cosgun}@monash.edu, juxi@lyro.io

Abstract—We investigate how high-resolution tactile sensors can be utilized in combination with vision and depth sensing, to improve grasp stability prediction. Recent advances in simulating high-resolution tactile sensing, in particular the TACTO simulator, enabled us to evaluate how neural networks can be trained with a combination of sensing modalities. With the large amounts of data needed to train large neural networks, robotic simulators provide a fast way to automate the data collection process. We expand on the existing work through an ablation study and an increased set of objects taken from the YCB benchmark set. Our results indicate that while the combination of vision, depth, and tactile sensing provides the best prediction results on known objects, the network fails to generalize to unknown objects. Our work also addresses existing issues with robotic grasping in tactile simulation and how to overcome them.

#### I. INTRODUCTION

Advancements in grasping and manipulation abilities are one of the major factors that will allow robots to move outside the world of manufacturing. Moving from structured environments to dynamic, complex human environments will allow robots to assist in homes, offices, and hospitals. This transition to operate in these human-centered environments will require high levels of dexterity, intelligence and versatility. As such, robots will require human-like abilities in grasping and manipulation.

Current approaches to robotic grasping often utilize a single sensing modality, commonly RGB-D cameras [1]. It remains uncommon to use tactile sensors for grasping, especially in conjunction with another sensing modality such as vision. In contrast, humans heavily rely on the sense of touch when they are manipulating objects [2]. This discrepancy is due to the difficulty in creating robotic end-effectors that are as capable as human hands, owing to a lack of sensors and actuators equivalent in size, precision, and efficiency to human skin and muscles [3].

A major challenge with machine learning in robotics is insufficient amounts of training data due to the limitation of the data collection speed with real hardware [4]. To combat this, simulated environments are often used to collect datasets much larger and faster than previously obtained using a real robot [5]–[7]. However, due to the difficulty of simulating vision-based tactile sensing, it has been largely excluded from these simulated grasping experiments. The recent development of TACTO [8], an open-source simulator for high-resolution vision-based tactile sensors, has bridged this gap and allowed



Fig. 1: Two examples of input to our approach for a successful grasp. Each row of three images, shows the output of the TACTO simulator, for both left (left image) and right gripper fingers (middle image), with the image on the right showing a camera view of the scene. Our network can successfully predict grasp success using this data. The intensity of the pixel roughly represents the depth of displacement for the gel inside the DIGIT sensor, where the darker the pixel, the greater the depth.

for the large-scale simulated grasp sampling required to train a deep neural network. [8] demonstrates the feasibility of the simulator with an example scenario where a network is trained to predict the grasp success of a single, rectangular object given visual tactile readings, however, to our knowledge, extensive experiments with TACTO has not been conducted in the literature.

We pose grasp stability prediction as a supervised learning task and use Convolutional Neural Networks (CNN) for function approximation. The input to the network is a set of images coming from multiple sensors and the output is a binary label indicating whether the grasp would be successful if the object is lifted. In this work, we use the TACTO simulator to train a network for predicting grasp success for a given grasp from tactile, visual and depth sensor data. Our grasping experiments are conducted using a subset of the YCB object set [9], a well-known benchmark object set in robotic grasping research, extending previous work using only a single object [8].

The primary contributions of this paper are:

- 1) Ablation studies with different combinations of sensing modalities including RGB and depth images from a side camera and tactile (DIGIT) sensor images.
- An investigation into validating the TACTO simulation of tactile sensors for robotic grasping.
- 3) A data filtering process for robotic grasping dataset collection, particularly for handling tactile information.

# II. RELATED WORK

#### A. Tactile Sensors

Developing robotic skins is an active research area [10], however, these sensors remain niche, hence are not commonplace. An alternative to robotic skins are high-resolution vision-based tactile sensors [11]–[17], which provide a high spatial resolution by typically outputting an image that encodes the deformation on the contact surface.

Data-driven methods alongside the development of sensors, particularly depth sensors, have allowed deep learning training to be performed with multiple modalities, significantly improving the robotic grasping capabilities [1]. Recently, the development of high-resolution tactile sensors has allowed for tactile information to be incorporated into the training, improving performance in both grasping [18], [19] and manipulation [20]–[22].

Early work in tactile sensors largely focused on measuring force and torque applied to the end-effector or the sensor's pressure distribution over the sensor [23].

There are at least eight main tactile sensor types, including [24]; piezoresistive [25], capacitive [25], piezoelectric [26], quantum tunnel effect [27], barometric measurements based [28], multi-modal [29], structure-borne sound [30] and vision based [11]–[16].

High-resolution vision-based tactile sensors such as Gel-Sight [11], [12], Gelslim [13], [14], FingerVision [15], OmniTact [17] and DIGIT [16] have been applied to robotic manipulation [20]–[22], [31], [32] and slip detection [12], [33] with some success. The vision-based sensors have a high spatial resolution and ability to provide a better generalization to objects of different geometry compared to Force/Torque sensors [22]. These sensors observe the topography of the contact surface, which is often a deformable elastomer material to measure contact forces. They also allow for the use of standard visual sensing convolutional neural network architectures since they typically output standard 2D images, making them significantly easier to incorporate into a multi-modal model [18]. We explore the ability of the DIGIT [16] sensor to improve robotic grasping ability in a simulation environment.

#### B. Grasping with Tactile sensing

Tactile information in robotics research has been used in a variety of tactile-relevant applications, which include: tactile exploration, grasping, in-hand manipulation, locomotion, tool manipulation, human-robot interaction, and non-prehensile manipulation [34]. Although, recent work has shown the strong

abilities of analytical approaches to use tactile information, with tasks such as manipulation [32]. However, these approaches often rely on assumptions of the geometry of the objects, robot, and environment. On the other hand, learningbased grasping methods do not rely on these assumptions and have gained prominence in recent years for these tasks.

Tactile information has been used in various learningbased approaches, most commonly in supervised learning [18], [19], [35]–[37]. Tactile data has also been integrated into reinforcement learning approaches for tasks such as in-hand manipulation [38] and object manipulation [22]. Tactile data has also been used in other learning-based approaches such as unsupervised learning [20], [39], self-supervised learning [31], transfer-learning [40], and active learning [41]. Our work extends previous research [8] by introducing a variety of real-world object models for predicting simulated grasping outcomes.

### III. METHODOLOGY

#### A. Problem Definition

Both simulation [8] and real-world experiments [18], [19] have shown that high-resolution tactile sensing improves the grasp stability estimation for singulated objects compared to more traditional sensing modalities. Our work extends previous simulation studies by using a more complex and standardized benchmark object set designed specifically for robotic grasping.

We assume that a singulated, rigid object is sitting stably on a flat table surface in front of the robot manipulator with a parallel gripper. Further, we assume that a top-down grasp pose is given, either manually or by a grasp detection algorithm, and the robot moves to that grasp pose before closing the gripper. We are interested in solving the problem of estimating whether the grasp would be successful if the object was lifted up without changing the end-effector orientation, given the vision, depth and tactile sensor readings before the lift. The success of a grasp is measured by whether or not the object is above a threshold height after lifting the gripper vertically upwards for a fixed distance and time.

#### B. Object Models

The YCB object set [9] is prominently used as a benchmark in robotic grasping and consists of 77 standard household items such as food, toys, and tools with different shapes, sizes, textures, weight, and rigidity. We use 20 objects from this set for training (shown in Fig. 2) and another 4 objects for validation (shown in Fig. 6). These objects were selected based on their ability to be grasped successfully by the robotic manipulator, which will be detailed in Sec. III-C. The YCB dataset provides scanned 3D models of each object obtained from a scanning rig. During preliminary experiments, we observed that grasping physics with these scanned 3D models is often unstable as the gripper moves through the objects during grasp attempts. To improve robustness post-processed watertight 3D models [42] of the YCB objects are used.



Fig. 2: 20 YCB object models used in training and testing. Watertight mesh equivalents of the object were used to help with the simulation accuracy.

However, no visual textures are available and therefore they were rendered in grey.

Fig. 3 shows the complexity and grasp difficulty of the objects used in our study, as defined by the EGAD! object set [43]. The grasp difficulty is quantified by the 75th percentile grasp of all sampled grasps, originally proposed by Wang et al. [44], Dex-Net [5] is used for grasp sampling, and the Ferrari-Canny metric is used for robustness. Morphological complexity is used to measure the complexity of the shape [45], [46]. Objects used for evaluation are shown with the black border, and the color represents the likelihood of collecting good tactile data on a grasp attempt for an object (red is low at approximately. 3%, green is high at approximately 88%).

#### C. Object Selection

We selected 24 objects among the 77 original YCB objects for our experiments. Object selection occurred in two stages. The first stage involved performing 100 grasp attempts on each watertight object model, scaled by 0.6 to 0.9 in increments of 0.1. This produced a total of 400 grasps per object. We then chose any object with a grasping success greater than 25% for any individual scaling and selected the scaling with the highest success rate. This is done to eliminate any object models that are inherently too difficult to grasp. This reduced the number of objects to 28. The second stage involved an initial stage of data collection, where any objects which did not produce tactile information on both fingers from 150 sampled grasps were removed. This reduced the number of objects to 24.

#### D. Data Collection

We use PyBullet and TACTO [8] for the simulation. We collected 10,000 data points to make the data comparable to the real-world experiment by Calandra et al. [18].

A single object is placed on the table at a uniformly random position in the workspace and yaw rotation at each iteration. To collect a data point, we complete three stages of grasping, detailed below.

- Select grasp pose: GGCNN [47] is used to calculate a top-down grasping pose. GGCNN uses a depth image to calculate the quality of grasp at each pixel. We choose the grasp with the highest quality to maximize the grasp success.
- 2) Move and close gripper: Once the grasp pose is chosen, the robot moves its end-effector to the grasp pose, and the gripper is closed at a constant velocity until at least 2N of force is read on both fingers. Data is only saved if tactile readings on both fingers are detected. This is determined by having at least 100 depth pixels with data greater than 0.0001m. This is because we are interested in both gripper fingers are touching the object, rather than cases where the object has already slipped from the gripper during closing. Furthermore, it would only make sense to try to lift an object in the real-world if both fingers are touching the object.
- 3) Lift object: Once the gripper fingers are closed, determined using force feedback, the robot attempts to lift the object vertically upwards. The robot moves at a fixed speed for a fixed amount of time, if the object moves at least 80% of the vertical distance, the attempt is labeled as successful.

#### E. Dataset Filtering

Simulation of contact is not perfect. For example, the tactile sensors would often not provide information, as fingers were simulated to go through an object. The success rate of obtaining tactile input from both fingers differed per object, causing the dataset to have an uneven distribution of datapoints per object. The data is filtered to reduce the effect of unbalanced data on the training results. We limit the number of grasp attempts per object to 500 and filtered to create a success/failure split that is as even as possible given collected data. For example, if an object had 700 grasp attempts, 200 successful and 500 unsuccessful, only 300 unsuccessful would be used for the dataset together with the 200 successful ones. Overall, this filtering provided a dataset with 66.7% successful grasp labels.



Fig. 3: The objects selected to validate grasp predictions plotted by grasp difficulty against shape complexity. The background colour per object indicates the proportion of grasps that contain tactile information on both sensors. From red (low,  $\sim 3\%$ ) to green (high,  $\sim 88\%$ ).

Four of the selected objects presented less than 500 grasps attempts with tactile information. These objects were reserved for the evaluation set.



Fig. 4: The basic architecture of the networks used to predict grasp success. The inputs to the network were various combinations of sensor modalities.

## F. Training

Convolutional neural networks with the ResNet-18 backbone [48] were trained using a combination of various modalities, the network architecture is illustrated in Fig. 4. A ResNet-18 backbone was chosen to directly compare it to the realworld experiment by Calandra et al. [18].

	Test Accuracy (%)	
Vision + Depth + Touch (Both)	82.7±0.6	
Vision + Touch (Left only)	81.9±1.5	
Vision + Touch (Both)	81.6±0.1	
Vision	81.5±0.8	
Vision + Depth	80.8±0.4	
Depth + Touch (Both)	80.8±1.2	
Depth	80±2.1	
Touch (Both)	78.9±1.2	
Touch (Left only)	76±1.5	

TABLE I: Ablation study for input sensor modalities. All networks are trained with 10,000 data samples.

In addition to the modalities originally presented [8], we added the modalities that included depth (see Table I). In accordance to the original TACTO paper, we trained the networks on 1,000, 2500, 5000 and 10,000 samples in total (split 80%/20% into training and testing). The networks were trained for ten epochs using Binary Cross-Entropy loss, and three-fold cross validation.

#### IV. RESULTS

# A. Grasp Stability Prediction (Known Objects)

1) Input Sensor Modalities: We perform an ablation study for a combination of input modalities, results are shown in Fig. 5 and Tab. I. Our results suggest that using all sensors (vision, depth, and touch) led to the highest accuracy (Fig. 5),



Fig. 5: Test accuracy of trained networks on 1,000, 2,500, 5,000 and 10,000 for various input modalities. The high-lighted area bounds the training results from all five cross validation trials.

	Vision+ Depth+ Touch (Both)	Vision+ Touch (Left only)	Vision	
Flat Screwdriver	68.6±2	67.1±4.1	68.1±7.7	
Large Clamp	57.4±10.8	59.9±4.4	57.4±5	
Spoon	57.6±5.9	56.5±6.3	64.1±4.2	
Phillips Screwdriver	62.8±2.1	66.9±9.3	60.4±8.8	
Softball	87.5±1.7	85.4±6.9	81.2±7.4	
Scissors	53.1±4.4	67.2±6	59.2±2.1	
Tomato Soup Can	90.3±3.7	89.5±6.3	85.1±6.7	
Windex Bottle	81.6±1.6	83.8±2	83.6±4	
Mini Soccer Ball	92.6±2.7	93.6±0.6	92.3±2.7	
Potted Meat Can	97.7±0.8	97.9±0.8	98.8±0.5	
Masterchef Can	93.1±5.1	83.7±2.5	80.8±1.5	
Power Drill	86±3.6	90.8±4.7	90.1±2.1	
Pitcher Base	91.1±0.7	88.3±8.4	87.3±5	
Plate	69.7±1.6	79.3±5.7	74.8±2.6	
Pudding Box	98.2±2.7	98.6±0.9	97.7±1.1	
Nine Hole Peg Test	84±1.8	75.4±5.1	78±1.9	
Mustard Bottle	83.8±5.3	82.9±3.7	80.5±5.4	
Bleach	86.7±4.4	82.5±1.8	84.6±5	
Sugar Box	96.8±1.5	96±1.9	97.1±0.7	
Wood Block	92.3±1.4	90±5.7	95±2.1	
Average	82.7±0.6	81.9±1.5	81.5±0.8	

TABLE II: Grasp stability prediction accuracy of the 20 "known objects". The networks are trained on 10,000 data samples. The table is sorted by the shape complexity of the objects (high to low, top to bottom).

albeit with a narrow margin with larger number of samples, as there was only 6.7 percentage points between the best- and worst-performing models for 10,000 samples. We note that the accuracy didn't necessarily increase with the number of samples for lower sample sizes. However this can likely be explained by the higher variance present for lower sample sizes as the test set size scaled with the training set.

The prediction accuracy for each modality increases with the number of training samples provided. Moreover, this convergence, particularly of vision, appears to happen significantly faster than previous grasping work using the TACTO simulator [8].



TABLE III: Vision + Depth + Touch (Both) model predicted grasp success against the ground truth for four samples on the mustard bottle object.

While the different modalities converge when 10,000 samples are used (Table I), integrating touch with visual and depth sensing provides significant improvement when the data availability is limited (83.5% when trained with 1,000 samples).

Another interesting observation was that the networks that utilized input from the left tactile sensor alone performed just as well as the corresponding networks which used input from both tactile sensors. This result is likely a byproduct of the filtering process during data collection (see Sec. III-E), which ensured tactile readings from both sensors. Another possible factor is that fewer input images were easier to train as there are fewer network weights to learn. This observation suggests that a single high-resolution vision-based tactile sensor could be sufficient when combined with a potentially cheaper tactile sensor that only detects whether the gripper finger has made contact with the object.

2) Performance on individual objects: The grasp stability prediction accuracy of the top-performing networks on the objects used in training are detailed in Table II. The objects are ordered in descending levels of complexity. We notice a large variation among objects in terms the grasp stability accuracy: for the highest performing modality (all sensors combined), the Pudding Box object had 98% accuracy, whereas the Scissors object had only 53% accuracy.

Our results show that, overall, the inclusion of tactile information marginally increases overall network performance on known objects. Additionally, networks that include the touch modality tend to perform better on objects of higher complexity. Furthermore, there tends to be an advantage of using tactile information for thinner objects, such as the two Screwdriver objects and the Scissors object. However, this is not always true, as evident by the vision-only model performing best on the Spoon, Potted Meat Can, Sugar Box and Wood Block objects. We notice a trend that larger box shaped objects tended to have higher accuracy for the visiononly network, but we have no explicit explanation for why this is the case. A further trend was for thin objects (e.g Spoon,

	Rubik's Cube	Cracker Box	Large Marker	Gelatin Box	Average
Vision + Depth + Touch (Both)	49.1±0.6	87.8±4.7	50.4±4.7	47.8±4.0	58.0±3.5
Vision + Touch (Left only)	49.9±1.4	79.8±10.8	47.0±6.0	41.1±7.8	53.9±1.9
Vision + Touch (Both)	48.6±6.4	88.9±7.5	48.9±7.8	51.9±12.1	58.5±3.8
Vision	48.9±0.5	81.2±3.7	42.9±5.4	$32.9 \pm 4.1$	51.1±2.7
Vision + Depth	52.8±2.0	69.6±7.4	39.8±2.5	34.1±3.1	48.5±1
Depth + Touch (Both)	45.3±2.8	93.5±0.9	56.4±2.0	52.9±10.6	61.3±2.8
Depth	44.6±1.2	79.5±13.4	41.3±5.9	42.1±1.0	51.6±4.9
Touch (Both)	44.2±1.7	93.3±1.9	55.8±4.2	53.5±5.5	60.9±3.1
Touch (Left only)	47.7±6.2	86.2±5.5	58.2±5.5	51±11.1	60.3±4.4

TABLE IV: Grasp stability prediction accuracy on unknown objects for networks trained on 10,000 samples. The overall performance is much lower than for known objects.

Screwdrivers, Large Clamp, Scissors) to have relatively low accuracy. This could be because there was greater difficulty in predicting the outcome, resulting in inconsistent results over different modalities.

*3) Qualitative Analysis:* Table III illustrates four examples of successful and unsuccessful predictions of the network for the Vision + Depth + Touch (Both) model.

The top left image illustrates a correctly predicted successful grasp which is likely due to its clear vision of the object and centered grasp.

A common failure of the network is due to partial occlusion of the object caused by the gripper. An example of this can be seen in the top right image resulting in an unsuccessful grasp being incorrectly predicted.

The bottom left image depicts the network incorrectly predicting that the grasp would be unsuccessful. Likely, this example could also be suffering from vision occlusion of the point contacts by the gripper.

The bottom right image of a correctly predicted unsuccessful grasp illustrates the network's potential ability to recognize a falling object and identify the low chance of a successful grasp.

These examples support our previous finding that the vision modality is heavily relied on for grasp prediction accuracy. This observation indicates perhaps more work is needed to rely more on tactile information in poor visual situations.



Cracker Box Large Marker Gelatin Box

Fig. 6: The four YCB object models used as unknown objects in our experiments

# B. Grasp Stability Prediction (Unknown Objects)

Rubik's Cube

The results on unknown objects in Table IV reveal that the best performing networks on the known objects did not align with the highest performing networks on unknown objects. This illustrates an overall issue around generalization and the potential the networks over-fitted to the training data. This is particularly evident by the poor results on the "Gelatin Box" (48%) compared to the very similarly shaped "Pudding Box" (98%). However, the networks that utilized tactile information did generalize better – the three worst-performing networks did not incorporate tactile information.

#### V. DISCUSSION

**Best sensing modalities:** Combining Vision + Depth + Touch (Both) modalities created the most successful predictions for known objects. It also provided high levels with only limited data (1,000 samples instead of 10,000). The same modality showed third-best performance for unknown objects. This result indicates that there is no clear winner in terms of which sensor modality should be utilized for grasp stability prediction. Additionally, this result may be subject to overfitting of the networks to the training objects as suggested in Table IV. This could be a result of a lack of data augmentation to allow the network to generalise.

For known objects, the three worst-performing modalities do not include vision illustrating the criticality of vision for predicting grasp success. There is further evidence on this observation, as we noticed that in many cases where the network was not able to predict grasp success, the object was occluded by the gripper. This problem could be alleviated with a multi-camera or an eye-in-hand setup.

**Comparison to existing literature:** Our best performing modality on known objects, Vision + Depth + Touch (Both), was actually not tested by Calandra et al. [18], who did not explore depth sensing in detail. [18]'s best performing modality was Touch + Vision, however our work suggests that adding depth could further improve their results.

We observe that the performance of different modalities in our work was much closer to each other compared to previous work on real robots [18] and in the TACTO simulator [8]. However, we did observe a similar trend of performance increasing with the number of modalities utilized, confirming this previous work that the addition of sensing modalities can improve grasp prediction performance.

**Problems with the simulator:** We encountered several difficulties during the data collection phase. When attempting to grasp the object, the gripper would often pass through the object. This occurred very frequently when we used the original 3D scans from the YCB dataset. This was significantly reduced after switching to the watertight models, however, these cases still happened, therefore, faulty data needed to be filtered out. Furthermore, we often encountered successful

grasps with no tactile information on one or both DIGIT sensors and no force feedback information. As such, we needed to perform significant filtering of our data to remove such cases. This extensive filtering left approximately 20% of the original data collected, notably increasing the time required for data collection. As well, due to this extensive filtering, there is a possibility that this may bias results as certain grasp types may be more likely to be included.

**Object diversity:** As depicted in Fig. 3, we observe that shapes of higher complexity tended to be more difficult to obtain grasping data with valid tactile information. However, this was not true for grasp difficulty, which did not significantly affect on the amount of data that could be collected. This illustrates a barrier in the simulator of dealing with objects of higher complexity, but not grasp difficulty.

# VI. CONCLUSION AND FUTURE WORK

The movement towards learning-based approaches in robotic grasping has increased the need for efficient largescale data collection and subsequently, the ability to accurately simulate sensors during robotic grasping. In this paper, we investigate the use of high-resolution tactile sensor information in a data-driven approach to estimate the success of robotic grasping.

Building on previous work [18] in a simulation setting, we used a subset of objects from the YCB benchmark. We found evidence that using multiple modalities helped with predicting whether a grasp would be successful. Our attempt at training and performing robotic grasping of benchmark objects in the TACTO simulator revealed several points of difficulty, particularly in data collection. For example, the simulator requires watertight object models and filtering to remove data inconsistencies, especially to ensure tactile information is present.

Our results suggest that a combination of visual, depth, and tactile sensing provides good predictions for known objects, especially in data limited scenarios. We also see, in our findings, the possibility of using a single vision-based tactile sensor with a force/torque sensor could provide the sufficient tactile information needed to predict grasp success on par with using two vision-based tactile sensors. This could be an interesting direction for future work. Additionally, we observed many failure cases occurred in visually occluded settings, and as such future work could pursue the ability to rely on tactile information more heavily in these situations.

#### REFERENCES

- A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017. [Online]. Available: https://doi.org/10.1177/0278364917735594
- [2] R. Johansson and J. Flanagan, "Coding and use of tactile signals from the fingertips in object manipulation tasks," *Nature reviews. Neuroscience*, vol. 10, pp. 345–59, 05 2009.
- [3] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, 2019.
- [4] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, p. 1677–1734, Aug 2020.

- [5] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg, "Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics," 2017.
- [6] A. Depierre, E. Dellandréa, and L. Chen, "Jacquard: A large scale dataset for robotic grasp detection," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 3511–3516.
- [7] H.-S. Fang, C. Wang, M. Gou, and C. Lu, "Graspnet-Ibillion: A large-scale benchmark for general object grasping," in 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11 441–11 450.
- [8] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible and open-source simulator for high-resolution vision-based tactile sensors," 2020.
- [9] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in 2015 International Conference on Advanced Robotics (ICAR), 2015, pp. 510–517.
- [10] D. Hughes, J. Lammie, and N. Correll, "A robotic skin for collision avoidance and affective touch recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1386–1393, 2018.
- [11] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017.
- [12] S. Dong, W. Yuan, and E. H. Adelson, "Improved gelsight tactile sensor for measuring geometry and slip," 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Sep 2017.
- [13] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactilesensing finger," in 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2018, pp. 1927–1934.
- [14] I. Taylor, S. Dong, and A. Rodriguez, "Gelslim3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger," 2021.
- [15] A. Yamaguchi and C. G. Atkeson, "Implementing tactile behaviors using fingervision," in 2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids), 2017, pp. 241–248.
- [16] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "Digit: A novel design for a low-cost compact highresolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.
- [17] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine, "Omnitact: A multi-directional high-resolution touch sensor," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 618–624.
- [18] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine, "The feeling of success: Does touch sensing help predict grasp outcomes?" in *CoRL*, 2017.
- [19] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, p. 3300–3307, Oct 2018.
- [20] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 818–824.
- [21] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 8863– 8869.
- [22] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," in 2021 IEEE International Conference on Robotics and Automation (ICRA), 2021, pp. 6437–6443.
- [23] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors and Actuators A: Physical*, vol. 167, no. 2, pp. 171–187, 2011, solid-State Sensors, Actuators and Microsystems Workshop.
- [24] Z. Kappassov, J.-A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands — review," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.
- [25] R. Kõiva, M. Zenker, C. Schürmann, R. Haschke, and H. J. Ritter, "A highly sensitive 3d-shaped tactile sensor," in 2013 IEEE/ASME

International Conference on Advanced Intelligent Mechatronics, 2013, pp. 1084–1089.

- [26] A. Schmitz, P. Maiolino, M. Maggiali, L. Natale, G. Cannata, and G. Metta, "Methods and technologies for the implementation of largescale robot tactile sensors," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 389–400, 2011.
- [27] M. W. Strohmayr and D. Schneider, "The dlr artificial skin step ii: Scalability as a prerequisite for whole-body covers," in 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems, 2013, pp. 4721–4728.
- [28] J. A. Fishel, V. J. Santos, and G. E. Loeb, "A robust micro-vibration sensor for biomimetic fingertips," in 2008 2nd IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics, 2008, pp. 659–663.
- [29] H. Hasegawa, Y. Mizoguchi, K. Tadakuma, A. Ming, M. Ishikawa, and M. Shimojo, "Development of intelligent robot hand using proximity, contact and slip sensing," in 2010 IEEE International Conference on Robotics and Automation, 2010, pp. 777–784.
- [30] L.-T. Jiang and J. R. Smith, "Seashell effect pretouch sensing for robotic grasping," in 2012 IEEE International Conference on Robotics and Automation, 2012, pp. 2851–2858.
- [31] M. A. Lee, Y. Zhu, K. Srinivasan, P. Shah, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks," in 2019 International Conference on Robotics and Automation (ICRA), 2019, pp. 8943–8950.
- [32] F. R. Hogan, J. Ballester, S. Dong, and A. Rodriguez, "Tactile dexterity: Manipulation primitives with tactile feedback," in 2020 IEEE International Conference on Robotics and Automation (ICRA), 2020, pp. 8863– 8869.
- [33] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a gelsight tactile sensor," in 2015 IEEE International Conference on Robotics and Automation (ICRA), 2015, pp. 304–311.
- [34] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.
- [35] F. Veiga, H. van Hoof, J. Peters, and T. Hermans, "Stabilizing novel objects by learning to predict tactile slip," in 2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2015, pp. 5065–5072.
- [36] F. Veiga, J. Peters, and T. Hermans, "Grip stabilization of novel objects using slip prediction," *IEEE Transactions on Haptics*, vol. 11, no. 4, pp. 531–542, 2018.
- [37] J. Hoelscher, J. Peters, and T. Hermans, "Evaluation of tactile feature extraction for interactive object recognition," in 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015, pp. 310–317.
- [38] H. van Hoof, T. Hermans, G. Neumann, and J. Peters, "Learning robot in-hand manipulation with tactile features," in 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids), 2015, pp. 121–127.
- [39] D. Cockbum, J.-P. Roberge, T.-H.-L. Le, A. Maslyczyk, and V. Duchaine, "Grasp stability assessment through unsupervised feature learning of tactile images," in 2017 IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 2238–2244.
- [40] P. Falco, S. Lu, C. Natale, S. Pirozzi, and D. Lee, "A transfer learning approach to cross-modal object recognition: From visual observation to robotic haptic exploration," *IEEE Transactions on Robotics*, vol. 35, no. 4, p. 987–998, Aug 2019.
- [41] D. Driess, P. Englert, and M. Toussaint, "Active learning with query paths for tactile object shape exploration," in 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2017, pp. 65–72.
- [42] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez, "Ganhand: Predicting human grasp affordances in multi-object scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5031–5041.
- [43] D. Morrison, P. Corke, and J. Leitner, "Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 4368–4375, 2020.
- [44] D. Wang, D. Tseng, P. Li, Y. Jiang, M. Guo, M. Danielczuk, J. Mahler, J. Ichnowski, and K. Goldberg, "Adversarial grasp objects," in 2019

IEEE 15th International Conference on Automation Science and Engineering (CASE), 2019, pp. 241–248.

- [45] D. Page, A. Koschan, S. Sukumar, B. Roui-Abidi, and M. Abidi, "Shape analysis algorithm based on information theory," in *Proceedings 2003 International Conference on Image Processing (Cat. No.03CH37429)*, vol. 1, 2003, pp. I–229.
- [46] J. E. Auerbach and J. C. Bongard, "Environmental influence on the evolution of morphological complexity in machines," *PLOS Computational Biology*, vol. 10, no. 1, pp. 1–17, 01 2014. [Online]. Available: https://doi.org/10.1371/journal.pcbi.1003399
- [47] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *Proc.* of Robotics: Science and Systems (RSS), 2018.
- [48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision* and Pattern Recognition (CVPR), June 2016.