

Estimating Global Object Pose from Tactile Images

by

Antonia Bronars

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Master of Science in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2022

© Massachusetts Institute of Technology 2022. All rights reserved.

Author
Department of Mechanical Engineering
Aug 12, 2022

Certified by
Alberto Rodriguez
Associate Professor
Thesis Supervisor

Accepted by
Nicolas Hadjiconstantinou
Department Graduate Officer

Estimating Global Object Pose from Tactile Images

by

Antonia Bronars

Submitted to the Department of Mechanical Engineering
on Aug 12, 2022, in partial fulfillment of the
requirements for the degree of
Master of Science in Mechanical Engineering

Abstract

This work evaluates Tac2Pose, an object-specific approach to tactile pose estimation for known objects. Given the object geometry, we learn a perception model in simulation that estimates a probability distribution over possible object poses given a tactile observation. To do so, we simulate the contact shapes that a dense set of object poses would produce on the sensor. Then, given a new contact shape obtained from the sensor, we match it against the pre-computed set using an object-specific embedding learned using contrastive learning. We obtain contact shapes from the sensor with an object-agnostic calibration step that maps RGB tactile images to binary contact shapes. This mapping, which can be reused across object and sensor instances, is the only step trained with real sensor data. Tac2Pose produces pose distributions and can incorporate additional pose constraints coming from other perception systems, multiple contacts, or priors.

We provide quantitative results for 20 objects. Tac2Pose provides high accuracy pose estimations from distinctive tactile observations while regressing meaningful pose distributions to account for those contact shapes that could result from different object poses. We test Tac2Pose in multi-contact scenarios where two tactile sensors are simultaneously in contact with the object, as during a grasp with a parallel jaw gripper. We further show that when the output pose distribution is filtered with a prior on the object pose, Tac2Pose is often able to improve significantly on the prior. This suggests synergistic use of Tac2Pose with additional sensing modalities (e.g. vision) even in cases where the tactile observation from a grasp is not sufficiently discriminative. Given a coarse estimate, even ambiguous contacts can be used to determine an object’s pose precisely.

We also test Tac2Pose on object models reconstructed from a 3D scanner, to evaluate the robustness to uncertainty in the object model. We show that even in the presence of model uncertainty, Tac2Pose is able to achieve fine accuracy comparable to when the object model is the manufacturer’s CAD model. Finally, we demonstrate the advantages of Tac2Pose compared with three baseline methods for tactile pose estimation: directly regressing the object pose with a neural network, matching an observed contact to a set of possible contacts using a standard classification neural network, and direct pixel comparison of an observed contact with a set of possible contacts.

Thesis Supervisor: Alberto Rodriguez
Title: Associate Professor

Acknowledgments

First and foremost, I want to thank my advisor, Professor Alberto Rodriguez, for his thoughtful guidance, advice, and support over the last two years. When I reflect on our time working together, his intuition and ideas about our research stand out equally to me as his kindness and generosity in supporting me as a person, as well as a researcher. I am truly looking forward to where our collaboration takes us next!

Second, I want to thank the members of the MCube Lab. Starting at a new university in a new city in the middle of the Covid-19 pandemic was not optimal, but my labmates always made me feel welcome and supported, despite not meeting in person for many months. In particular, I want to thank Maria Bauza. The research in this thesis builds upon her work, and great idea to explicitly estimate distributions over object pose using tactile images. So much of what I know about working with robots, probabilistic perception, and how to be an effective researcher I credit to Maria's mentorship.

Third, I want to thank my family for their endless, unconditional support. Their unshakable belief in me is one of the things I am most grateful for. Mom, I learned the thrill and joy of discovery and new ideas from you. Thank you for being a lifelong champion of my curiosity, for always picking up the phone, and for drying my tears with your unique blend of comfort and rationality time and time again. Dad, your unsolicited texts of encouragement and love mean more than you know! Thank you for always making me feel like I'm enough, and for keeping my inbox stocked with Frenchie pictures. Matt, you will always be my little brother, but you have grown into a person I trust, respect, and admire wholeheartedly. Your vote of confidence is one of the ones I value most (and I hope you know that it's reciprocated).

Finally, I want to thank the friends I've met since moving to Cambridge, who have truly become my family. To my roommates Kristen, Ruby, and Rachel - living with you three makes every day an adventure and I wouldn't want it other way. You are some of the most funny, loving, bright, motivated, and interesting people I've had the pleasure of knowing, much less calling my best friends.

How did I get so lucky?!

Contents

1	Introduction	15
1.1	Related Work	17
1.1.1	Tactile Pose Estimation	17
1.1.2	Visual Pose Estimation	18
1.1.3	Tactile Tracking	19
2	Single-Shot Pose Estimation from Tactile Images	21
2.1	Methods	21
2.1.1	Contact Shape Prediction	22
2.1.2	Global Tactile Pose Estimation	23
2.1.3	Real Data Collection	25
2.2	Results	27
2.2.1	Single Contact	28
2.2.2	Parallel Jaw Contacts	31
2.2.3	Parallel Jaw Contacts with a Pose Prior	34
2.2.4	Comparing Grasp Approach Directions	36
2.2.5	Comparing Individual Grasps	36
2.2.6	Comparison with Reconstructed Object Geometry	39
2.3	Baselines	41
2.3.1	Baseline Comparison with Simulated Data	43
2.3.2	Baseline Comparison with Real Data	48

3	Conclusions	53
3.1	Discussion	53
3.2	Ongoing Work	56
3.2.1	Discrete Smoother	57
3.2.2	Continuous Smoother	58

List of Figures

- 2-1 **Tactile pose estimation with Tac2Pose.** (Bottom row) In simulation, we render geometric contact shapes of the object from a dense set of possible contacts between object and tactile sensor. (Top row) The real sensor generates a tactile image from which we estimate its geometric contact shape. We then match it against the simulated set of contact shapes to find the distribution of contact poses that are more likely to have generated it. For efficiency and robustness, we do the contact shape matching in an embedding learned for that particular object. 22
- 2-2 **Object grids.** The four dimensions of the grid with respect to the object are visualized (left): grasp approach direction, which is defined as the direction of the axis of the grasp (blue arrow), x translation (red arrow), y translation (green error), and angle in the plane of the grasp (yellow arrow). Samples of grid elements on the object long grease are shown as black dots, where each black dot represents the center location of the gripper during a grasp. Tac2Pose assigns likelihoods to each gripper location (right), given an observed contact (top). The most likely gripper locations for an observed contact are colored green, while the least likely are colored blue. The ground truth gripper location is shown as a black dot (right, center). 24
- 2-3 **Similarity function.** The similarity function learns to encode contact shapes into a low dimensional space and predicts, given a new contact shape, the likelihood of being the closest match of each contact shape in the pre-computed set. 25

2-4 **Normalized pose errors**, i.e., pose errors with respect to the average random error, for the object *grease*. The closest distance in the grid for this object is 0.03, similar to the first case. The median pose error when using parallel jaw contact information (0.10) corresponds to an error like the second case, and the average random error (1) would match with the last case. Finally, the example with 0.84 normalized error depicts a non-unique contact shape, i.e., the two object poses result in very similar contact shapes at the center that are not possible to distinguish without additional information. 26

2-5 **Error distributions and sample contacts**. Error distributions for snapping (top row), long grease (second row), hydraulic (third row), and cotter (bottom row). Each violin plot contains three distributions: **single contact**, **parallel jaw contact**, and **parallel jaw distribution filtered with a 10mm prior**. The median closest grid error is visualized as a **red** line. To the left and right of each violin plot, we show the localization errors for a sample grasp on each object (true object pose shown as grey mesh). The best match when using a single contact is shown in **green**, parallel jaw contacts in **blue**, and prior in **purple**. 30

2-6 **Comparing grasp approach directions**. Error distributions for the first (left) and second (right) grasp approach directions on **round clip** using parallel jaw contacts. The **red** line represents the median closest grid error. The first grasp approach direction has more non-unique contacts (sample contacts visualized at left of the plot), and therefore has higher median error. Contacts in the second grasp approach direction, on the other hand, are much more unique (sample contacts visualized at right of the plot). . . . 37

2-7 **Uniqueness of individual grasps on long pencil**. Output pose distributions for three grasps on long pencil, using real (2-7a) and simulated (2-7b) parallel jaw contacts. Each dot represents a possible grasp center location, and its color the likelihood that a given grasp generated the tactile observation. The true grasp location is shown as a black dot. 38

2-8 **Baseline results on the object hanger.** Normalized pose error for the object hanger using tactile matching (Tac2Pose) and three baseline methods. We compare the performance of each method on simulated contacts from different grid sizes in (2-8a), and on real vs. simulated versions of the contacts in the hanger dataset in (2-8b). 45

List of Tables

2.1	Median error of the most likely pose. The median error of the most likely pose is reported in mm, and as a normalized error (in parenthesis).	29
2.2	Reconstructed geometry versus manufacturer’s CAD. Pose error and normalized pose error (in parenthesis) for Tac2Pose on manufacturer’s CAD models versus scanned object models. We compare results with parallel jaw contacts + 10mm pose prior.	40
2.3	Baseline performance on simulated data. Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods for a range of grid sizes, using 100 randomly simulated contacts per grid. Single Contact and Parallel Jaw are abbreviated as SC and PJ respectively. Mini One Face and Bigger Mini One Face grids are abbreviated as MOF and BMOF, respectively.	44
2.4	Baseline results on real data. Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods, on <i>real</i> datasets. Columns labelled SC are evaluated with a single contact, while columns labelled PJ use parallel jaw information.	49
2.5	Baseline results on simulated versions of real datasets. Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods, on <i>simulated</i> versions of the real contacts used in Table 2.4. Single Contact and Parallel Jaw are abbreviated as SC and PJ, respectively.	50

Chapter 1

Introduction

Robotics history sends a clear lesson: accurate and reliable perception is an enabler of progress in robotics. From depth cameras to convolutional neural networks, we have seen how advances in perception foster the development of new techniques and applications. For instance, the invention of high-resolution LIDAR fueled self-driving cars, and the generalization capacity of deep neural networks has dominated progress in perception and grasp planning in warehouse automation [45, 29, 36]. A primary goal of this research is to demonstrate the role that image-based tactile sensing has to play in that history. In robotic manipulation applications, occlusions challenge accurate pose estimation and object dynamics are dominated by contact interactions. Image-based tactile sensing provides a local view of the object geometry at contact; therefore, these challenges both fall in the set of problems that image-based tactile sensing is well-poised to address.

In this thesis, we evaluate Tac2Pose, a method for estimating the pose of a touched object from a single tactile image. Given a 3D model of the object, Tac2Pose learns an object-specific perception model in simulation, tailored at estimating the pose of the object from a tactile image without requiring any previous interaction. The proposed approach is motivated by scenarios where the main requirement is estimation accuracy and where object models will be available beforehand. Many industrial scenarios fit this category. Many previous solutions to tactile pose estimation require prior exploration of the object [26, 2]. Acquiring this tactile experience can be expensive, and in many cases, unrealistic. In this work, instead, we learn the perception model directly from the object mesh model, and

demonstrate that the model learned in simulation directly transfers to the real world. We attribute this to the object-specific nature of the learned model, the high resolution nature of the tactile sensors, and the contrastive-based perception framework. Tac2Pose uses an intermediate representation of the tactile image, a binary contact mask, which we are able to generate with high fidelity both in simulation and using the real sensor. Furthermore, because Tac2Pose matches observed contact shapes to a discrete set of simulated contacts, the match doesn't need to be perfect - just better than the other possible options - for the localization to be accurate.

Also key to the approach is that, by simulating a dense set of tactile imprints, the algorithm can reason over pose distributions, not only the best estimate. The learned embedding allows us to efficiently compute the likelihood of each contact shape in the simulated dense set to match with the predicted contact shape from the tactile sensor. Predicting distributions is key given that tactile sensing provides local observations, which often do not discriminate pose globally.

We evaluate Tac2Pose's accuracy with respect to object geometry, uncertainty (from reconstructed object geometry and tactile sensor noise), and other methods for tactile localization. In particular, we show:

1. Quantitative results for 20 objects on three ablations of Tac2Pose using real data: single contact, parallel jaw grasping, and parallel jaw grasping with a pose prior. We achieve high localization accuracy when making contact with distinct object features. Complete results for each object can be seen in Table 2.1.
2. Comparisons for results on 5 objects using reconstructed object shapes instead of manufacturer's CAD mode. The localization error on reconstructed models increases between 0.2 to 1.5mm compared to manufacturer's CAD models. Complete results can be seen in Table 2.2.
3. Better performance of Tac2Pose when compared with three baseline methods. Results on real contact shapes for 5 objects can be seen in Table 2.4.

1.1 Related Work

1.1.1 Tactile Pose Estimation

Low-resolution tactile sensors. Most initial works in tactile localization were focused on low-resolution tactile sensors [35, 9, 30, 6, 4, 34, 19, 7]. Some works explore how to combine multiple tactile readings and reason in the space of contact manifolds [21, 22]. However, these are based on binary contact/no-contact signals, and require many tactile readings to narrow pose estimates.

Given the challenges from the locality of tactile sensing, recent works have gravitated towards two different approaches. Combining tactile and vision to obtain better global estimates of the object pose or using higher-resolution tactile sensors that can better discriminate different contacts. Among the solutions that combine vision and tactile, most rely on tactile sensors as binary contact detectors whose main purpose is to refine the predictions from vision [5, 1, 16, 12, 44].

High-resolution tactile sensors. Other works, more in line with Tac2Pose, have focused on using high-resolution tactile sensors as the main sensing source for object localization. Initial works in this direction used image-based tactile sensors to recover the contact shape of an object and then use it to filter the object pose [32, 28]. However, these approaches only provide results on planar objects and require previous tactile exploration. There has also been some recent work on highly deformable tactile sensors for object localization [23]. These sensors are large enough to fully cover the touched objects, which eases localization.

In this work, we use the image-based tactile sensor GelSlim 3.0 [40]. The sensing capabilities of high-resolution sensors of this kind have already proven useful in multiple robotic applications, including assessing grasp quality [15], improving 3D shape perception [42] or directly learning from tactile images how to do contour following [25] or tactile servoing [41]. For the task of tactile object localization, Li et al. [26] proposed to extract local contact shapes from objects to build a map of the object and then use it to localize new contacts. The approach is meant to deal with small parts with discriminative features. Later Izatt et al. [18] proposed to compute pointclouds from the sensor and use them to

complement a vision-based tracker. Their tracker is fused with vision to deal with the uncertainty that arises from the locality of tactile sensing. Bauza et al. [2] proposed to extract local contact shapes from the sensors and match them to the tactile map of the objects to do object pose estimation. This approach requires the estimation of a tactile map for each object by extensively exploring them with the sensor.

1.1.2 Visual Pose Estimation

Tac2Pose’s approach to tactile pose estimation is related to methods recently explored in the computer vision community where they render realistic images of objects and learn how to estimate the orientation of an object given a new image of it [39, 20]. [27] designs a network to predict a relative pose between an observed image, and a rendered image of the object in a known pose. The method is used to iteratively improve on a coarse initial estimate of object pose. [24] leverages a modified version of [27] to estimate the pose of multiple known objects from a sequence of images from multiple, unknown, viewpoints. The approach consists of predicting object poses relative to the camera within single views, then robustly matching objects and poses between views, before performing a scene-level refinement of both object and camera pose. Although this method does not reason explicitly over pose distributions for a single-view estimate of pose, it is able to handle occlusions and inaccurate single estimates by combining information from multiple viewpoints.

Other methods address occlusion and inaccurate estimates by reasoning over pose distributions directly. [11] estimates the 6D pose (3D translation, and a distribution over a discretization of 3D rotations) of an object in a Rao-Blackwellized particle filtering framework. The distributions generated using this method function as scores on a codebook of possible poses, rather than as well-defined probability distributions. Tac2Pose, in contrast, regresses probability distributions which can be combined with distributions coming from additional contacts, sensing modalities, or priors.

1.1.3 Tactile Tracking

The tracking problem can be defined as solving for the sequence of states that is most likely given a sequence of measurements. Tracking approaches can be coarsely broken down into filtering and smoothing approaches. Filtering approaches make estimates only of the current state, while smoothing approaches make estimates of the full history of states up to the current state. In the most general problem formulation, filtering approaches are more computationally efficient, while smoothing approaches are more robust to noisy or uninformative observations.

For a subset of problems in which the measurement distribution can be represented by a Gaussian noise model, it is paradoxically more efficient to solve for the full history of states, rather than just the most current state. This class of problems can be solved incrementally and efficiently using sparse linear algebra techniques [10], and benefit from the robustness of a smoothing solution without compromising the tractability of the problem. Because tactile sensors provide local, and often ambiguous, measurements of the object state, the measurement distribution cannot be well represented by a Gaussian noise model. Tactile sensor measurement distributions are often represented non-parametrically, and thus smoothing approaches become computationally infeasible.

Filtering approaches. Pezzementi et al. [31] estimates a planar object state from a series of binary contact measurements using two types of non-parametric filters - the histogram and the particle filter. They find that the histogram filter outperformed the particle filter, which suffered from premature convergence as a result of capturing only an approximation of the state distribution. Chhatpar and Branicky [8] use a particle filter to resolve the uncertainty of the lock position in a lock and key problem. They consider scenarios where the lock position uncertainty far exceeds the lock clearance, and focus on resolving the uncertainty using probe readings only (no visual input). The particle filtering strategy was effective at reducing the lock position uncertainty, such that a complaint strategy alone was capable of achieving lock and key assembly. Saund et al. [34] also uses a particle filter to update a non-Gaussian belief of the object pose. They leverage the particle filter structure to anticipate the expected information gain of each probing action, and develop an active

localization framework based on taking the best information gathering action to reduce pose uncertainty.

Smoothing approaches. Smoothing approaches to the tactile tracking problem often require a prior on the object pose, such that the measurement distribution from the tactile sensor can be represented accurately by a Gaussian noise model. Sodhi et al. [37] estimated object pose during planar pushing from a stream of tactile imprints through a factor graph-based estimation framework. Their tactile observation model is trained to predict the relative pose of the object between a pair of non-sequential tactile images. This approach is designed to track the drift of an object from an initial well-known pose. Similarly, Sodhi et al. [38] estimated 3D object pose over a contact sequence through a factor graph-based estimation framework. The tactile observation model maps from tactile images to surface normals using an image-to-image translation network, and uses ICP to determine the relative pose between 3D contact geometry in a contact sequence. This approach is also designed to track the drift of object from an initial well-known pose, where the object itself is arbitrary and unknown. In contrast, our approach assumes that the object geometry is known, but has no prior information about the object pose.

Chapter 2

Single-Shot Pose Estimation from Tactile Images

2.1 Methods

Our approach to object pose estimation (Tac2Pose) is based on tactile sensing and known object models, as illustrated in Fig. 2-1. In an object-specific embedding, we match a dense set of simulated *contact shapes* against the estimated contact shape from a real tactile observation. A contact shape is a binary mask over contact regions on the tactile sensor. This results in a probability distribution over contact poses that can be later refined using other pose constraints. For example, we can combine information from the two contact shapes and the gripper opening obtained by grasping an object with a parallel jaw gripper.

We predict real contact shapes directly from the raw RGB tactile images that the tactile sensor outputs (Section 2.1.1). As seen in the top left quadrant of Figure 2-1, we predict a binary mask over the region of contact (a.k.a. the contact shape) from an RGB tactile image with high fidelity. The next steps of Tac2Pose exploit the object model to estimate the contact pose, and are learned in simulation without using any real tactile observations. First, we render the contact shape for a given object pose, using the object model. Examples of the correspondence between object pose and contact shape obtained through geometric contact rendering [3] can be seen in the bottom half of Figure 2-1. Next, we use geometric contact rendering to generate a dense set of object poses and their respective contact shapes,

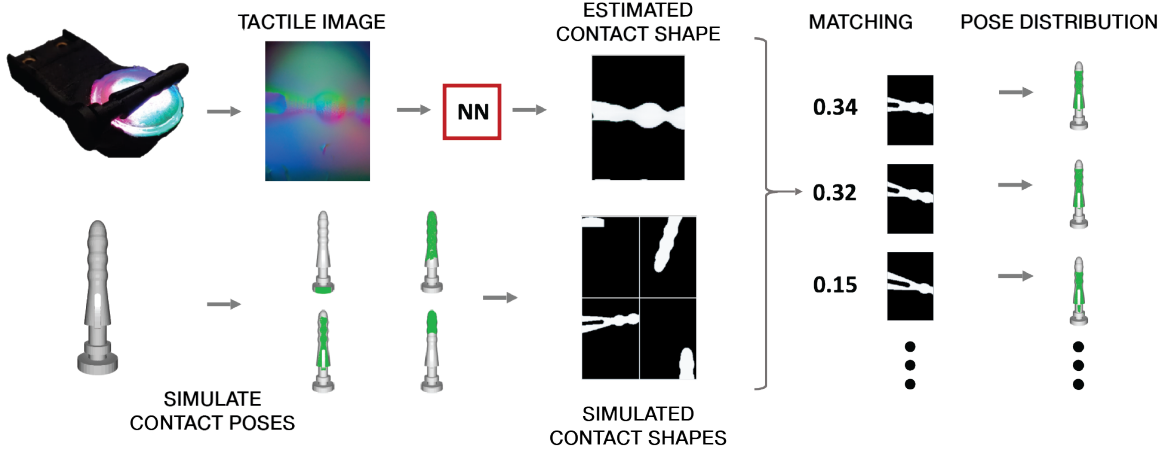


Figure 2-1: **Tactile pose estimation with Tac2Pose.** (Bottom row) In simulation, we render geometric contact shapes of the object from a dense set of possible contacts between object and tactile sensor. (Top row) The real sensor generates a tactile image from which we estimate its geometric contact shape. We then match it against the simulated set of contact shapes to find the distribution of contact poses that are more likely to have generated it. For efficiency and robustness, we do the contact shape matching in an embedding learned for that particular object.

and use contrastive learning to learn an embedding to match contact shapes depending on the closeness of their object poses (Section 2.1.2). As a result, given the estimated contact shape from a real tactile observation, we can match it against this pre-computed dense set to obtain a probability distribution over contact poses.

2.1.1 Contact Shape Prediction

Given a tactile observation, our goal is to extract the contact shape that produces it. To that aim, we train an image translation network (pix2pix) based on Isola et al. [17] to map RGB tactile observations to contact shapes. We train pix2pix using pairs of real RGB tactile images and their corresponding contact shapes. The training data is collected autonomously in a controlled 4-axis stage that generates planned touches on known 3D-printed shapes, following the approach proposed in Bauza et al. [2]. The dataset we use to train pix2pix contains 10,000 RGB tactile image/contact shape pairs from 32 distinct contact geometries. The dataset does not contain examples of contact geometries from any of the objects we use to evaluate tactile localization. Note that the map between RGB

tactile images and contact shapes is independent of the object, and therefore we only need to gather labelled data once. Empirically, we find that a model trained on images from a single GelSlim sensor generalizes well across multiple instances of GelSlim sensors.

2.1.2 Global Tactile Pose Estimation

Once we know how to compute contact shapes both in simulation and from real tactile imprints, we reduce the problem of object pose estimation to finding what contact poses are more likely to produce a given contact shape. We solve this problem by first discretizing the space of possible contact poses as a parametrized grid, and then learning a similarity function that compares contact shapes.

Object-dependent grids. Using the 3D model of an object, we discretize the space of object poses in a multidimensional grid. Building a grid in the space of poses is a well-studied problem [43, 33] that makes finding nearest neighbors trivial. It also allows each point on the grid to be seen as the representative of a volumetric part of the space which helps to reason over distributions. We prune the grid by only keeping object poses that result in contact, and then pair each of them with their respective contact shape. Since we only consider poses that result in contact, this reduces the space of 6D poses to a 5D manifold. Using a grid, a discrete structured set of poses, allows us to easily account for object symmetries which can significantly reduce the grid size.

Our grids cover regions of the object which correspond to feasible grasp locations from the set of stable resting poses of the object on a surface. In particular, we include faces that correspond to feasible *grasp approach directions*, where the grasp approach direction specifies the axis of the grasp relative to the object. For some objects, we also include additional grasp approach directions that have interesting tactile features. For each grasp approach direction, We compute a dense set of contacts with 2.5mm translational resolution, and 6 degrees of rotational resolution around the axis of the grasp. As a result, the grid spans four coordinates (grasp approach direction, x , y , θ) and a contact pose will be no more than 1.25mm from an element of the grid, and often closer. A visualization of the grid dimensions can be seen in Figure 2-2. For the twenty objects we evaluate, the number

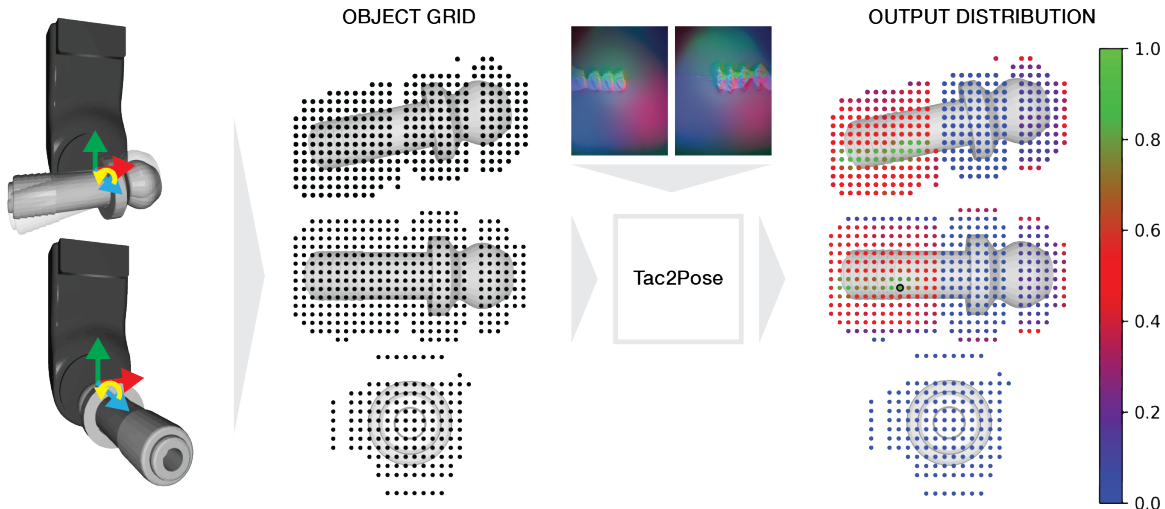


Figure 2-2: **Object grids.** The four dimensions of the grid with respect to the object are visualized (left): grasp approach direction, which is defined as the direction of the axis of the grasp (blue arrow), x translation (red arrow), y translation (green error), and angle in the plane of the grasp (yellow arrow). Samples of grid elements on the object long grease are shown as black dots, where each black dot represents the center location of the gripper during a grasp. Tac2Pose assigns likelihoods to each gripper location (right), given an observed contact (top). The most likely gripper locations for an observed contact are colored green, while the least likely are colored blue. The ground truth gripper location is shown as a black dot (right, center).

of elements in each grid varies between 3.8K and 181.3K, depending on the object size, shape, and number of grasp approach directions.

Similarity metric for contact shapes. Given a new contact shape, we compare it to all pre-computed contact shapes in the grid to find what poses are more likely to produce it. To that aim, we modify Momentum Contrast (MoCo) [14], a widely-used algorithm in contrastive learning, to encode contact shapes into a low dimensional embedding based on the distance between contact poses.

MoCo is able to learn unsupervised embeddings by building "a dynamic dictionary with a queue and a moving-averaged encoder". Instead, Tac2Pose is supervised and the elements in the queue are fixed and assigned to each of the poses in the object's grid. Given a new contact shape, our model learns to predict the likelihood that each pose in the grid has produced the given shape. This likelihood is computed by taking the softmax over the distances between the embedding of the new contact shape and the embeddings saved in the queue, which correspond to the embeddings of each contact shape in the grid. Compared

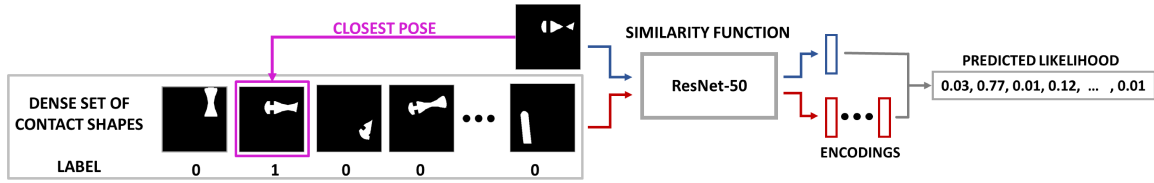


Figure 2-3: **Similarity function.** The similarity function learns to encode contact shapes into a low dimensional space and predicts, given a new contact shape, the likelihood of being the closest match of each contact shape in the pre-computed set.

to the original MoCo, Tac2Pose is supervised because during training, given a new contact shape, we can compute which element in the queue is closest to it.

To implement the encoder in our model, we use a ResNet-50 [13] cropped before the average-pooling layer to preserve spatial information, making it a fully-convolutional architecture. The loss function is the categorical cross-entropy loss which allows us to ensure that the output of the softmax is a well-defined probability distribution. The training data comes from selecting a random contact pose and finding its closest element in the dense grid. Then, we use as desired probabilities a vector of all zeros except for the closest element which gets assigned to probability one (see Figure 2-3). Once we have created a dense grid and trained a similarity encoder for an object, given a new contact shape at test time, we can estimate which poses from the grid are more likely to generate it. To run Tac2Pose in real-time, we first encode the given contact shape and then compare it to all pre-computed encodings from the grid, which requires a single matrix-vector multiplication. Finally, we perform a softmax over the resulting vector of distances to obtain a probability distribution over the contact poses in the grid. We take the best match to be the contact pose with the highest probability, after (if applicable) incorporating additional pose constraints.

2.1.3 Real Data Collection

While Tac2Pose is trained purely with simulated data, it can provide accurate pose estimation when evaluated on real tactile data. To that aim, we design a system that collects tactile observations on accurately-controlled poses. Below we describe the tactile sensor, the robot platform, and the objects used to perform the experiments.

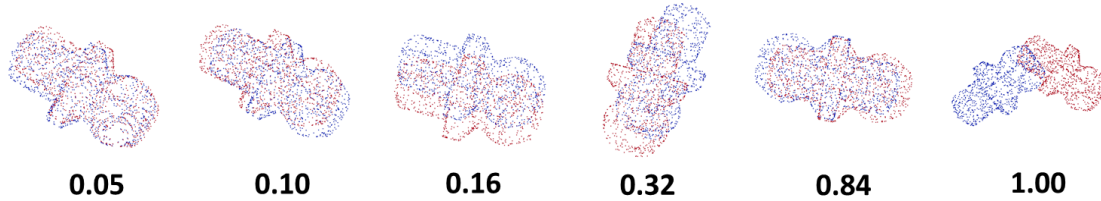


Figure 2-4: **Normalized pose errors**, i.e., pose errors with respect to the average random error, for the object *grease*. The closest distance in the grid for this object is 0.03, similar to the first case. The median pose error when using parallel jaw contact information (0.10) corresponds to an error like the second case, and the average random error (1) would match with the last case. Finally, the example with 0.84 normalized error depicts a non-unique contact shape, i.e., the two object poses result in very similar contact shapes at the center that are not possible to distinguish without additional information.

Tactile sensor. We consider the tactile sensor GelSlim 3.0 [40] which provides high-resolution tactile readings in the form of RGB images. The sensor consists of a membrane that deforms when contacted and a camera that records the deformation. The sensor publishes tactile observations through ROS as 470x470 compressed images at a frequency of 90Hz (see top left portion of Figure 2-1).

Robot platform. We collect labelled datasets of tactile observations of grasps on 20 objects mounted in known positions and orientations in the world. Each dataset contains pairs of RGB tactile images, and the corresponding ground-truth object pose relative to the gripper. The robotic system we use to collect the labelled datasets consists of a dual arm ABB Yumi with two WSG-32 grippers and GelSlim 3.0 tactile sensing fingers.

We collect labelled observations of feasible grasps on each object. We compute the contacts in each dataset by considering the *stable poses* of each object. Stable poses are the poses that an object is most likely to fall in when dropped onto a table. We then determine a set of feasible grasp approach directions for each stable pose. We define the grasp approach direction as the axis of the grasp during a parallel jaw grasp (Figure 2-2). For some objects, we also include additional grasp approach directions that have interesting tactile features. For each grasp approach direction, we collect a set of equispaced real observations, that correspond to a given grasp approach direction at various x,y locations relative to the object. In total, these observations correspond to the set of contacts that we are likely to encounter when picking up each object from a table.

Objects. We test Tac2Pose on 20 objects, derived from CAD models of objects available on McMaster. During the selection of objects, we aimed at covering object features that differently impact tactile localization. The features we considered include:

1. *Contact non-uniqueness*: Contact non-uniqueness is the degree to which a contact is ambiguous, i.e., the contact in one pose resembles the contacts from other object poses (that are not equivalent under any symmetry). Non-unique contacts are more difficult to localize. An example of a non-unique contact on the object grease is shown in Figure 2-4 (second from right).
2. *Symmetry*: Object symmetries are sets of transformations under which the object pose is indistinguishable. Symmetry is a desirable property because it reduces the size of the grid.
3. *Object size*: Larger objects are generally more challenging to localize with tactile sensing alone, since a single touch corresponds to a more local view of the object.
4. *Contact size*: Large regions of contact with the sensor are generally more difficult to localize than smaller ones, because most tactile sensors are less likely to produce crisp imprints for large contacts.

2.2 Results

We evaluate the accuracy of Tac2Pose at estimating object poses from real tactile images. We note that Tac2Pose, even for large grids (100K elements), can easily run at 50Hz allowing real-time estimation of pose distributions. For each object, we evaluate Tac2Pose on 90 to 400 (varies depending on object size and number of grasp approach directions) pairs of real tactile images and object poses per object using the approach described in 2.1.3. Given the ground truth object pose and our best estimate (the object pose with highest likelihood), we measure the resulting *pose error* (or distance between the two poses) by sampling a pointcloud of 10K points from the object 3D model and averaging the distance between these points when the object is at either of the two poses. This distance is sometimes called ADD (average 3D distance) but, for simplicity, we just refer to it as the pose error. To compare errors across shapes and object sizes, we also compute the

normalized pose error which divides the original pose error by the average error obtained from predicting a random contact pose. Figure 2-4 shows examples of different normalized pose errors. For each object, we evaluate the pose error for three ablations of Tac2Pose :

1. *Single Contact*: Estimate object pose from a single tactile image.
2. *Parallel Jaw*: Estimate object pose from a pair of tactile images, collected during a parallel jaw grasp on the object. The estimate also factors in the opening of the gripper during the grasp.
3. *Parallel Jaw + Prior*: Filter the distribution from the parallel jaw estimate to remove any poses that are more than a given pose distance from ground truth. This approximates cases in which an object pose is known within a margin of error, and is relevant when tactile localization is used to refine a coarse estimate of object pose from another sensing modality, like vision. We evaluate prior distances of 10mm and 5mm.





















The median error for each ablation for each object is shown in Table 2.1. We also aggregate the pose errors corresponding to the best estimate for each grasp, then visualize the distribution of errors as a violin plot in Figure 2-5, for selected objects. The medians reported in Table 2.1 correspond to the medians of the distributions visualized in the violin plots.

For each of the selected objects, we show three different error distributions, corresponding to the best estimates when using: a single contact (**green** distribution), parallel jaw contacts (**blue** distribution), and parallel jaw contacts plus a 10mm pose prior (**purple** distribution). To facilitate comparison between the objects, we plot normalized pose errors. Note that a normalized pose error of 1 corresponds to the expected error from selecting a pose at random from the grid. The **red** line measures the median pose error between the ground truth pose and its closest pose in the object’s grid of simulated contacts (a.k.a. the *closest error*). This sets a lower bound on the median performance for any given method.

2.2.1 Single Contact

We first analyze the performance of Tac2Pose with a single contact. Of the 20 objects we evaluate, 9 have median localization error with a single contact that is at least twice

Table 2.1: **Median error of the most likely pose.** The median error of the most likely pose is reported in mm, and as a normalized error (in parenthesis).

		Tactile Only		Pose Prior	
		Single Contact mm (norm)	Parallel Jaw mm (norm)	10mm Prior mm (norm)	5mm Prior mm (norm)
Snap Ring		1.5 (0.10)	1.4 (0.10)	1.4 (0.17)	1.4 (0.38)
Grease		1.3 (0.12)	1.2 (0.10)	1.2 (0.15)	0.9 (0.26)
Slotted Shim		4.0 (0.15)	3.0 (0.12)	2.4 (0.30)	2.3 (0.57)
Round Clip		3.4 (0.17)	11.5 (0.58)	2.0 (0.24)	1.9 (0.50)
Hanger		6.6 (0.19)	2.6 (0.07)	2.4 (0.30)	2.4 (0.57)
Pin		6.5 (0.20)	5.6 (0.17)	4.7 (0.66)	3.6 (1.02)
Big Head		7.8 (0.20)	6.1 (0.16)	4.9 (0.72)	3.9 (1.19)
Holder		5.8 (0.26)	2.2 (0.10)	1.8 (0.23)	1.8 (0.45)
Round Couple		13.6 (0.53)	10.9 (0.43)	6.0 (0.75)	3.5 (0.91)
Hydraulic		14.0 (0.67)	4.9 (0.23)	2.5 (0.31)	2.0 (0.49)
Long Grease		26.6 (0.76)	3.3 (0.09)	2.3 (0.33)	2.3 (0.61)
Stud		33.7 (0.85)	13.4 (0.34)	4.8 (0.57)	3.4 (0.82)
Cotter		19.0 (0.49)	19.6 (0.51)	2.9 (0.38)	2.7 (0.70)
Cable Clip		10.2 (0.59)	11.7 (0.67)	2.5 (0.30)	1.9 (0.45)
Hook		24.5 (0.77)	27.2 (0.85)	3.0 (0.38)	2.4 (0.63)
Couple		20.7 (0.79)	19.9 (0.76)	3.5 (0.42)	2.6 (0.66)
Hose		39.0 (0.62)	41.6 (0.66)	7.8 (1.00)	4.2 (1.03)
Pencil		38.1 (0.69)	41.6 (0.76)	5.0 (0.62)	3.9 (1.05)
Round Hose		37.6 (0.70)	37.3 (0.69)	5.5 (0.71)	4.2 (1.05)
Long Pencil		77.5 (0.96)	78.5 (0.97)	6.9 (0.85)	4.3 (1.01)

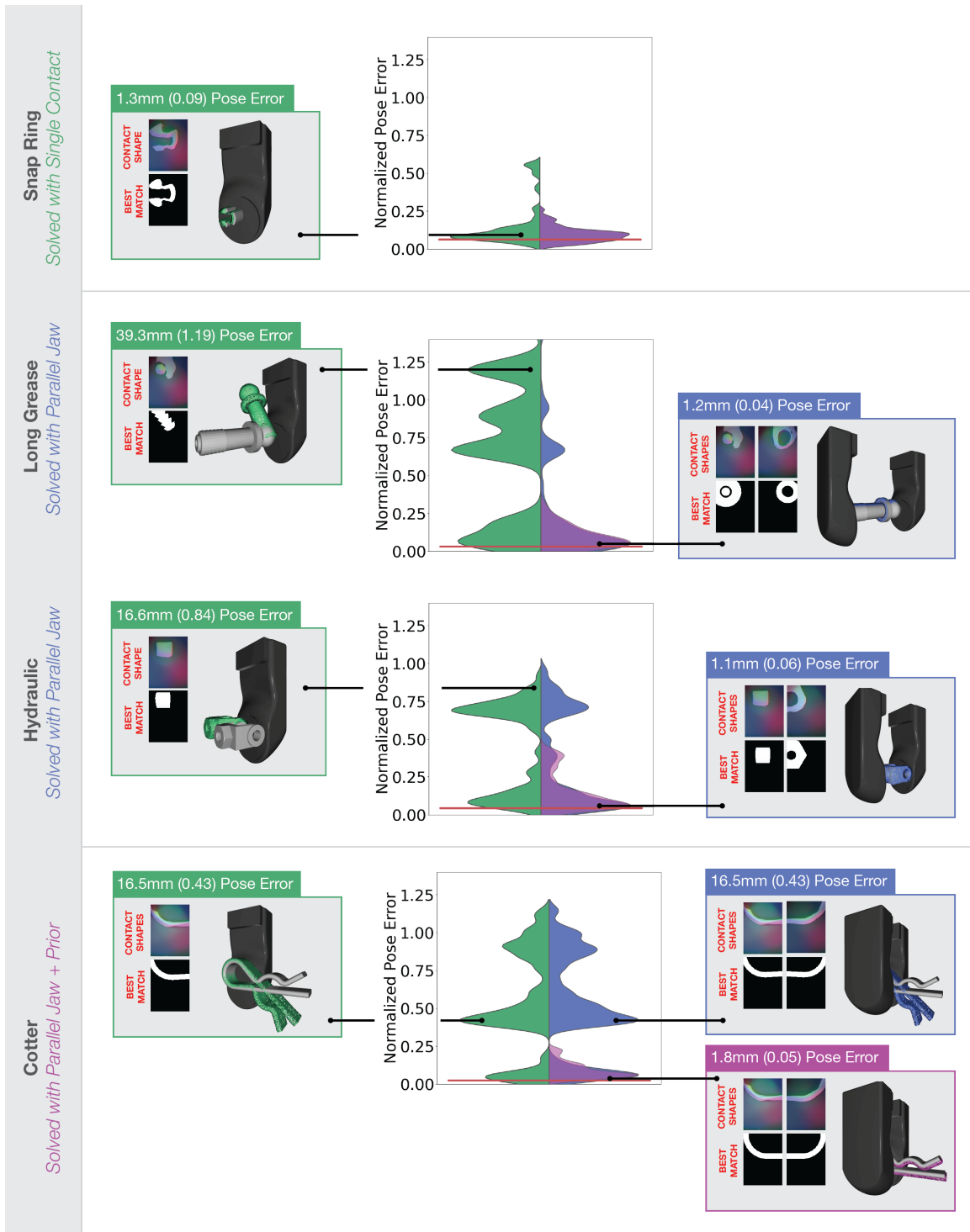


Figure 2-5: **Error distributions and sample contacts.** Error distributions for snap ring (top row), long grease (second row), hydraulic (third row), and cotter (bottom row). Each violin plot contains three distributions: **single contact**, **parallel jaw contact**, and **parallel jaw distribution filtered with a 10mm prior**. The median closest grid error is visualized as a **red** line. To the left and right of each violin plot, we show the localization errors for a sample grasp on each object (true object pose shown as grey mesh). The best match when using a single contact is shown in **green**, parallel jaw contacts in **blue**, and prior in **purple**.

as good as random. These 9 objects have single contact normalized errors less than 0.5 in Table 2.1.

The objects that tend to perform best with a single contact are small objects with unique tactile features. For these objects, a single tactile imprint is often discriminative enough to fully determine the object pose. Consider the error distribution for `snap ring`, visualized in top row of Figure 2-5 as an example. Even with a single contact (**green** distribution), the error distribution has a clear primary mode which is centered near the closest error from the grid (**red** line). When incorporating additional constraints from a parallel jaw grasp, the distribution (**blue**) tightens around the closest error from the grid (**red** line), and secondary, higher error modes disappear. The median normalized error is approximately 0.10 for both the single contact and parallel jaw contact cases (Table 2.1). In other words, choosing the most likely pose using Tac2Pose results in, on average, approximately ten times less error than selecting a pose at random from the grid.

A sample contact that is localized within 1.3mm (0.09 normalized error) of the true pose using a single contact is visualized at left of the violin plot in Figure 2-5. This provides a sense of the scale of the errors.

2.2.2 Parallel Jaw Contacts

We next consider objects that perform significantly better with parallel jaw contacts. Of the 20 objects we evaluate, 11 have a median localization error with parallel jaw contacts that is more than twice as good as random. Of these 11 objects, 5 perform much better with parallel jaw contacts than with a single contact. The 5 objects satisfying both criteria are those in Table 2.1 where the parallel jaw normalized error is less than 0.5 and the single contact normalized error is more than double the parallel jaw contact normalized error. Two such objects are `long grease` and `hydraulic`.

The objects that perform significantly better with parallel jaw contacts compared with a single contact tend to be larger. Contacts with the object may appear as large, featureless patches from the local perspective of the tactile sensor. In general, large, flat contacts are more challenging to reconstruct from tactile images because they do not deform the gel

membrane of the tactile sensor as much. Including the constraints of a second contact and the gripper opening provides higher robustness to noisy contact reconstructions. Furthermore, objects that vary significantly in width depending on where they are grasped benefit from parallel jaw contacts. In these cases, the additional constraint of the gripper opening during grasp can disambiguate between otherwise non-unique contacts. Recall that a contact is non-unique if its look similar to contacts where the object is in a different pose. In these cases, the contact is not a unique indicator of the object pose.

As an example, the error distribution for long grease is visualized as a violin plot in the second row of Figure 2-5. Considering parallel jaw information shifts the mode of the distribution much closer to the closest grid match (red line), and significantly reduces the prevalence of higher error modes. The medians of the error distributions for the single contact and parallel cases are significantly different; with a single contact, the median pose error is 26.6mm (0.76 normalized error) while with parallel jaw contacts, the median pose error is 3.3mm (0.10 normalized error). The random error, in comparison, is 35.2mm. With a single contact, Tac2Pose only performs about 1.3 times better than random. When including parallel jaw contacts, however, Tac2Pose is more 10 times better than random. The second row of Figure 2-5 (left of the violin plot) shows a sample contact on long grease, and the corresponding best match when using only a single contact (green pointcloud). The best match using a single contact is 39.3mm (1.19 normalized error) away from the true pose. The same contact, and the corresponding best match when using parallel jaw contacts (blue pointcloud) is visualized to the right of the violin plot. The best match using parallel jaw contacts is only 1.2mm (0.04 normalized error) away from the true pose. In this case, the additional information from the second contact and gripper opening is enough to resolve the ambiguity and substantially improve the localization.

Parallel jaw contacts can also be an important tool for disambiguating ambiguity inherent to the object geometry. As an example, consider a grasp on the object hydraulic, visualized at left of the violin plot in the third row of Figure 2-5. The true object pose is visualized as a grey mesh, whereas the best match using a single contact is visualized as a green pointcloud. The localization error with a single contact, in this case, is 16.6mm (0.84 normalized error). When considering only contacts from one of the fingers, the con-

tact shapes corresponding to the true object pose and the best match are indistinguishable. Considering parallel jaw contacts, on the other hand, resolves the ambiguity.

The same contact, and the corresponding best match using parallel jaw contacts, is visualized at the right of the violin plot in the third row of the same figure. The localization (blue pointcloud) using parallel jaw contacts is much better because the contact on the second finger provides critical information about the object's pose. The localization improves from 16.6mm (0.84 normalized) error with a single contact, to 1.1mm (0.06 normalized) error with parallel jaw contacts.

All grasps on `hydraulic` in contact with this non-unique feature are subject to the same problem when considering information from only a single contact, and therefore the non-uniqueness impacts the overall localization quality; the median localization error for `hydraulic` with a single contact is 14mm (0.67 normalized error), while with parallel jaw contact is 4.9mm (0.23 normalized error). The violin plot in the third row of Figure 2-5 shows that the distribution of errors when using a single contact (green) is bimodal, with a portion of the second, high error, mode corresponding to flipped localizations, like the one shown at left of the violin plot. When considering the distribution of best match errors using parallel jaw contacts (blue distribution), we see that the density of the high error mode corresponding to the flipped localization decreases, and more probability mass shifts into the low error mode centered around the closest error.

In the case of `hydraulic`, it is informative to break out the results by *grasp approach direction*. The primary reason that the localization error with parallel jaw contacts is, on average, nearly 3 times better than with a single contact is that including parallel jaw contacts resolves the aforementioned non-uniqueness, which only impacts one of the two grasp approach directions. For the grasp approach direction that contains the non-unique feature, including parallel jaw contacts reduces the localization error by nearly 8 times. For the other grasp approach direction, including parallel jaw contacts reduces the localization error by 1.6 times.

2.2.3 Parallel Jaw Contacts with a Pose Prior

Finally, we consider the performance of Tac2Pose with parallel jaw contacts plus a prior on the object pose. Pose priors, in practice, can be obtained using additional sensing modalities (e.g. vision), kinematics, or previous estimates of the object’s pose. Because tactile sensing is inherently local, an object can have non-unique contacts which won’t be possible to fully disambiguate even with parallel jaw contacts. In such cases, a prior on the object’s pose can help resolve these ambiguities. To test the effect of a prior on Tac2Pose, we take the prediction distribution over possible object poses, and filter any pose that is more than a given distance from ground truth. Note that this implementation of a pose prior requires knowledge of the ground truth object pose, but it is a useful proxy for cases in which the object pose is known roughly, but not accurately. We evaluate localization accuracy for prior distances of 10mm and 5mm. Results for both prior distances for each of the 20 objects are listed in Table 2.1.

The objects that benefit most from incorporating a prior on the object pose are those with *discrete non-uniqueness*. An object with discrete non-uniqueness has features that are unique (i.e. create discriminate tactile imprints) relative to most other possible contacts on the object, with a discrete number of exceptions. This type of discrete non-uniqueness can be an inherent feature of the object, or an artifact of noisy and incomplete contact shapes. Because there is a discrete number of possible object poses that are likely to produce such a contact shape, we expect the distribution over object poses to have a discrete number of modes where the majority of probability mass is concentrated. In these cases, incorporating a pose prior can truncate the distribution in such a way that only one of the modes is left. In other words, because there is a discrete number of possible poses that are likely to produce a given contact, if we eliminate the higher error options using a pose prior, we are likely to get a near perfect match. Therefore, for some objects, even a coarse prior on the object pose results in precise localization when combined with tactile information.

We evaluate which objects benefit most from incorporating a pose prior by comparing the pose error of the best match after filtering the pose distribution, with the expected error from selecting a pose at random from the filtered distribution. The normalized error we

report in Table 2.1 for the 10mm and 5mm prior ablations uses the expected random error from the filtered distribution. For the remainder of this section, we consider just the case of a 10mm pose prior. For 12 of the 20 objects we evaluate, using Tac2Pose on top of a 10mm pose prior results in performance more than twice as good as selecting a pose at random from the filtered distribution. For 5 of the 12 objects, incorporating a pose prior plays an important role in driving down the localization error. With just parallel jaw contact information, the median localization error for these 5 objects is not significantly better than selecting a pose at random from the grid (where here we take *significantly better* to be more than twice as good as random). However, after filtering the distribution with a 10mm prior on the object pose, Tac2Pose selects the best pose with more than twice as much accuracy as selecting a pose at random from the filtered distribution. The objects that benefit most from incorporation of a pose prior are those in Table 2.1 with parallel jaw normalized error greater than 0.5, but 10mm prior normalized errors less than 0.5. An example of an object for which the incorporation of a 10mm prior leads to significantly better localization is `cotter`.

By examining the multi-modal error distributions with a single contact, or parallel jaw contact (violin plot in the bottom section of Figure 2-5), we can notice that the object `cotter` is likely to benefit from using pose prior. Its error distributions have a discrete number of modes (three) where the majority of probability mass is concentrated. Note that the symmetry of `cotter` is such that parallel jaw contacts do not provide much new information relative to a single contact, so the distributions of error for a single contact and parallel jaw contacts are very similar. The median error is 19.0mm (0.49 normalized error) when using a single contact, and 19.6mm (0.51 normalized error) when using parallel jaw contacts. Incorporation of a 10mm prior on the object pose eliminates the two higher error modes, and shifts the median of the distribution toward the closest error. The median error becomes 2.9mm (0.38 normalized error), which is nearly seven times lower than using parallel jaw contact information alone. Recall that after incorporating a pose prior, the median error is normalized by the expected error from selecting a pose at random from the filtered distribution rather than from all contact poses.

We show an example of a grasp that benefits from a pose prior at left of the violin plots

in the bottom row of Figure 2-5. With single or parallel jaw contacts, the localization error is 16.5mm (0.43 normalized). The contacts when cotten is in the given pose (grey mesh) are informative, but not completely unique relative to other possible contacts on the object. Incorporation of a coarse 10mm prior on the object pose, though, resolves the ambiguity, and the localization improves from 16.5mm (0.43 normalized) to 1.8mm (0.23 normalized) error for this grasp. The best match after filtering the parallel jaw distribution is visualized as a purple pointcloud on the right side of the violin plot in the bottom row of Figure 2-5.

2.2.4 Comparing Grasp Approach Directions

We evaluate multiple grasp approach directions for 13 of the 20 objects. For these objects, we compare the median localization error using parallel jaw contacts of each grasp approach directions. In practice, knowing which sets of grasps lead to lower localization errors could be leveraged in a grasp planning framework to select more informative grasps. There are 6 objects for which one direction has half as much error or less as the other(s): long grease, grease, hanger, hydraulic, round clip and round couple. As example, we compare the two grasp approach directions for round clip in Figure 2-6. The first grasp approach direction, visualized in the left half of the figure, has non-unique contacts. As a result, the median normalized error is 15.2mm (0.85 normalized) when using parallel jaw contacts. The second grasp approach direction, visualized in the right half of the figure, has much more unique contacts; the median normalized error is 2.2mm (0.11 normalized), and the primary mode of the error distribution is concentrated near the closest grid error. The second grasp approach direction consists of grasps that, on average, lead to about eight times lower localization errors.

2.2.5 Comparing Individual Grasps

Some objects are difficult to localize even in the presence of a pose prior. For these objects, Tac2Pose does not reduce the median error much beyond the prior on the object pose. These objects are characterized by large, continuous, regions of non-unique contacts, and their error distributions tend to be broader, rather than having a discrete number of

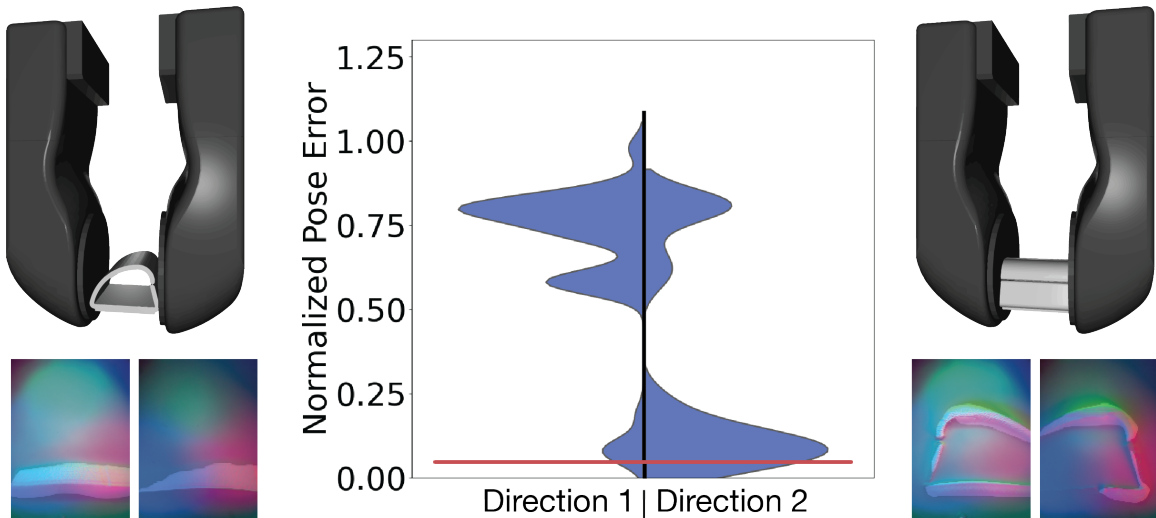
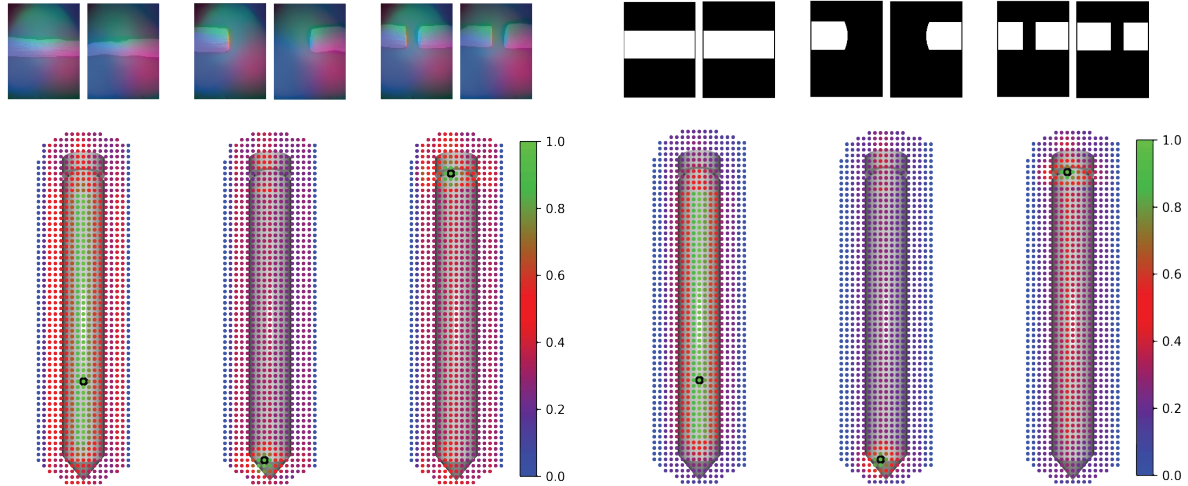


Figure 2-6: **Comparing grasp approach directions.** Error distributions for the first (left) and second (right) grasp approach directions on `round_clip` using parallel jaw contacts. The red line represents the median closest grid error. The first grasp approach direction has more non-unique contacts (sample contacts visualized at left of the plot), and therefore has higher median error. Contacts in the second grasp approach direction, on the other hand, are much more unique (sample contacts visualized at right of the plot).

modes.

`long_pencil` (Figure 2-7a), for example, suffers from having large regions of non-unique contacts. Many of the contact poses along the length of the object result in the same information, and thus on average the match is chosen essentially at random from a large set of possible contacts. Note that this is not a limitation of Tac2Pose for tactile localization, but rather results from the object geometry and the local nature of tactile feedback; most contacts on `long_pencil` are not sufficient to uniquely determine the object pose.

Even for objects with a large amount of non-uniqueness on average, there can be regions of the object which create distinctive and unique tactile imprints. Contacts near either end of `long_pencil`, for example, are easier to localize accurately. Grasping on these regions leads to better localization. Depending on the downstream application, it might be beneficial to plan for grasps on an object that are expected to produce more unique tactile imprints, and thus lower localization errors. The existence of good grasps is something we can detect and exploit to avoid large regions of non-uniqueness, and achieve low localization errors. The distribution over contact poses for selected contacts on unique and



(a) Real RGB tactile images for three grasps on long pencil (top). Distributions over possible contacts using real grasp (bottom).

(b) Simulated contact shapes for three grasps on long pencil (top). Distributions over possible contacts using simulated grasp (bottom).

Figure 2-7: **Uniqueness of individual grasps on long pencil.** Output pose distributions for three grasps on long pencil, using real (2-7a) and simulated (2-7b) parallel jaw contacts. Each dot represents a possible grasp center location, and its color the likelihood that a given grasp generated the tactile observation. The true grasp location is shown as a black dot.

non-unique regions of long pencil is shown in Figure 2-7a. The top row of the figure shows three possible grasps on the object. The bottom row shows the output of Tac2Pose - a distribution over possible contact poses. Each dot overlaid on long pencil represents a possible grasp location, and the color represents the likelihood that a given grasp generated the input tactile observation. The green dots are grasps with highest likelihood, while the blue dots are grasps with lowest likelihood. The black dot represents the true grasp location. Note that for ease of visualization, we show only points that represent center locations of the grasp when long pencil is oriented horizontally. This is a small subset (935) of the total number of contacts (>65k) Tac2Pose reasons over. For contacts near the middle of the object (left of Figure 2-7a, for example), many grasps along the length of the object have high probability (green dots). For instance, the left contact in Figure 2-7a on its own is not enough to uniquely determine the object pose. In fact, the pose with the highest likelihood results in 77.9mm (0.99 normalized) error for this case. This aligns with our intuitive understanding that the best match is chosen essentially at random from a large set of likely contacts (green dots) along the middle of the object. Cases like this highlight the impor-

tance of outputting meaningful distributions over possible contact poses; the true contact pose has high likelihood, so Tac2Pose could be used in combination with information from other sensing modalities, kinematics, or previous tactile estimates to converge on a unique estimate of the true pose.

For contacts near the tips of the pencil (center and right of Figure 2-7a, for example), only grasps immediately surrounding the correct tip have high probability. The cases illustrated result in 6mm (0.09 normalized) for the center contact, and 7mm (0.09 normalized) for the right.

In Figure 2-7b, we show distributions over contact poses for simulated versions of the same three contacts on `long_pencil`. The output distributions when using simulated contacts are qualitatively similar as when using real contacts. Simulated contact shapes can therefore be used to detect regions of non-uniqueness before ever encountering the object. This feature could be used in a grasp planning framework to avoid regions of non-unique contacts and promote accurate, unique pose estimation from a single grasp.

2.2.6 Comparison with Reconstructed Object Geometry








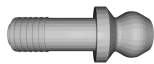



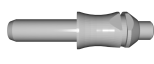
Tac2Pose relies on apriori knowledge of the object geometry, in the form of a 3D CAD model. Previously, we evaluate Tac2Pose on manufacturer’s CAD models. Here, we consider the case where the object model is reconstructed using a 3D scanner. In particular, we use a SOL 3D Scanner by Scan Dimension, to generate the scanned object models.

We simulate a grid of possible object contacts and contact poses using the scanned object model. At test time, we compare contacts observed on the real object (same real datasets used previously) against this imperfect set of simulated contacts.

We evaluate the accuracy of Tac2Pose when using scanned object models for 5 of the 20 objects, `long_grease`, `snap_ring`, `big_head`, `cotter`, and `hanger`. We qualitatively compare the scanned and manufacturer’s CAD models for each of the 5 objects in Table 2.2. The left column shows the result of the 3D scan, and the right column shows the true object model.

When evaluating on the scanned object models, we relax the parameter that controls

Table 2.2: **Reconstructed geometry versus manufacturer’s CAD.** Pose error and normalized pose error (in parenthesis) for Tac2Pose on manufacturer’s CAD models versus scanned object models. We compare results with parallel jaw contacts + 10mm pose prior.

	Scanned Model		Manufacturer’s Model	
		mm (norm)		mm (norm)
Snap Ring		1.6 (0.19)		1.4 (0.17)
Hanger		3.0 (0.39)		2.4 (0.30)
Long Grease		2.5 (0.33)		2.3 (0.33)
Cotter		4.0 (0.51)		2.9 (0.38)
Big Head		6.4 (0.88)		4.9 (0.72)

our confidence in the measurement of the gripper opening to reflect the fact that we expect the scanned object models to provide noisy information about the object geometry. The intuition for this decision is the following: when the object model is known, the gripper opening is a reliable signal of object pose. Therefore, we only want to assign high likelihood to those contact poses that match the observed gripper opening closely. When the object model is reconstructed with a scanner, we cannot be as confident that the gripper opening we measure corresponds closely to the opening of the true nearest contact pose in the grid. This is because the gripper opening computed using the scanned model will be noisy. Therefore, we may want to assign high likelihood to a contact pose, even if the observed gripper opening does not match the opening of the contact pose as closely.

Table 2.2 compares the localization accuracy on scanned models versus manufacturer’s CAD models for the 5 objects, when using parallel jaw contact information with a 10mm pose prior. Although globally the geometry of the object looks similar between the scanned and manufacturer’s CAD models, there are clear differences at the local level. This ablation of Tac2Pose allows us to evaluate whether we are still able to achieve significant pose

refinement, even when the local geometries are noisy.

We find that the error is comparable to that when using a manufacturer’s CAD model for all 5 objects. The normalized error is between 1.01 times (*long grease*) and 1.33 times (*cotter*) higher when using a scanned model compared with the manufacturer’s CAD model. *Cotter* is most impacted by using a scanned object model because the contact non-uniqueness becomes less discrete. When the object model is known exactly, many of the contacts look unique with a discrete number of exceptions. Therefore, even a coarse prior on the object pose leads to precise localization. When the object model is noisy, the discrete nature of the non-uniqueness is impacted for a fraction of the contacts, and the median localization error increases.

For *snap ring*, *hanger*, *long grease*, and *cotter* Tac2Pose is able to improve on the 10mm prior by a significant amount. For the first three objects, including parallel jaw information on top of the 10mm prior results in localization errors twice as low as selecting a pose at random from the filtered distribution (in Table 2.2, the normalized pose errors for the scanned objects are less than 0.5). For *cotter*, the localization error is nearly twice as low as selecting a pose at random from the filtered distribution (in Table 2.2, the normalized pose error is 0.51). For these four objects, Tac2Pose is able to refine the object pose within a small amount of error, even when local geometry is noisy. For *big head*, the amount of error is comparable (1.22 times higher) to that when using a manufacturer’s CAD model. Because *big head* has a symmetry-breaking feature around its principle axis, many contacts are continuously non-unique, and therefore the localization error does not improve much when incorporating a coarse prior, even when the object model is known.

2.3 Baselines

Next, we compare Tac2Pose with three baseline methods for tactile pose estimation:

1. *Pixel*: We perform direct pixel comparison between the observed contact mask, and each of the contact shapes in the grid.
 - (a) *Single Contact*: We select the grid shape that has the most pixels in common with the observed contact mask as the best match. We take the object pose to

be that corresponding to the best match in the grid.

- (b) *Parallel Jaw*: We take into account both tactile images and the gripper opening during a parallel jaw grasp. To do so, we compare the pair of observed contact masks, and the observed gripper opening, with triplets of contact masks and gripper openings from the grid. We sum the pixel error from each of the contacts, and the normalized error between the observed opening and the grid opening, to score each of the parallel jaw triplets in the grid. We take the object pose to be the contact pose corresponding to the triplet with the lowest score.
- 2. *Classification*: We formulate the task of pose prediction as a standard classification problem between the elements of the grid. We train a convolutional neural network (CNN) based on ResNet-50 to predict a distribution over the grid elements from a given contact mask. Recall that in Tac2Pose, the neural network is an encoder that maps contact shapes to vectors and we obtain a distribution over object poses by comparing the encoding from an observed contact shape to all the encodings from the contact shapes in the grid. In contrast, the neural network in the `classification` baseline learns to predict the distributions over object poses directly. We train this baseline by using the same data generation as in Tac2Pose. The inputs are simulated contacts that we obtained from slightly perturbing a pose from the grid and rendering its corresponding contact. For each of these new contacts, we compute its closest grid element and use as training label a vector with length equal to the number of elements as the grid, where all its values are zero except the entry corresponding to the index of the closest element, which has value one. This vector represents a probability distribution that measures the likelihood of each element of the grid to be the closest to the new contact. The loss function is the cross-entropy loss between the predicted likelihood and the classification label. At test time, from an observed tactile image in the form of a contact mask we obtain a distribution over grid elements.
- 3. *Pose*: We train a CNN based on ResNet-50 to regress a nine-element representation of the object pose based on an observed tactile image. The training data consists of contact shapes as inputs, and the contact poses that generated such contacts as labels. Each label is a pose represented by three translational elements, and six rotational

elements that can be mapped into a rotation matrix. In comparison with quaternions, our 6D representation of rotations is continuous (meaning that similar orientations are close together in 6D representation), and therefore better suited to regression. We construct the last column of the rotation matrix from the 6D representation by applying Gram-Schmidt as a post-process on the first two columns, and taking the third column as the cross product of the first two [46]. The loss function is the mean squared error between the output and true nine-element pose. This method for pose prediction does not rely on matching to elements on the grid, but rather predicts the pose of the object directly using supervised regression.

Both Tac2Pose and `pixel` are compatible with using parallel jaw information, while `classification` and `pose` only work with single contacts.

We evaluate the performance of Tac2Pose compared to baselines for 5 of the 20 objects: long grease, snap ring, big head, cotter, and hanger. We select these 5 objects to cover a range of difficulty in terms of grid size and degrees of symmetry and non-uniqueness.

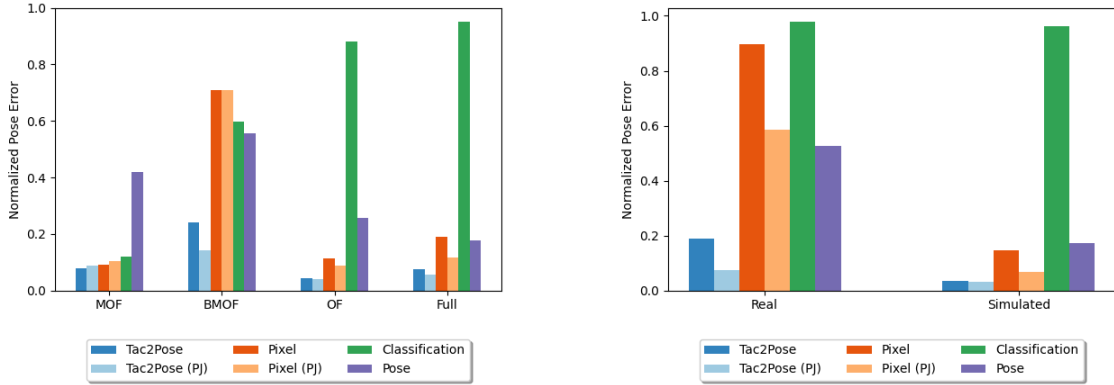
2.3.1 Baseline Comparison with Simulated Data

We first evaluate each method on sets of 100 simulated contacts per object. We sample 100 contact poses from the object’s grid, add noise to each pose, and render the contacts corresponding to the new poses. For each of the 5 baseline objects, we create three smaller, simpler, grids in addition to the original. The reduced grids serve to test each method’s ability to scale to different problem sizes and types of complexity. Recall that the original grids are determined by 4 spatial coordinates (grasp approach direction, x , y , θ) as illustrated in Figure 2-2. They have angular resolution of 6 degrees, and x , y translational resolution of 2.5mm. The size of the original grids for the baseline objects ranges from 8.1k to 91.5k poses. The reduced grids are modified as follows:

1. *Mini One Face*: We include one grasp approach direction, and one angle. We consider translations in x , y with 5mm resolution. This grid is much smaller than the original (29 to 124 poses, depending on the object), and is effectively 2D (x , y).

Table 2.3: **Baseline performance on simulated data.** Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods for a range of grid sizes, using 100 randomly simulated contacts per grid. Single Contact and Parallel Jaw are abbreviated as SC and PJ respectively. Mini One Face and Bigger Mini One Face grids are abbreviated as MOF and BMOF, respectively.

	Tac2Pose		Pixel		Classification	Pose
	SC mm (norm)	PJ mm (norm)	SC mm (norm)	PJ mm (norm)	SC mm (norm)	SC mm (norm)
Long Grease						
MOF	2.3 (0.09)	2.3 (0.09)	2.7 (0.11)	2.4 (0.09)	2.1 (0.08)	16.1 (0.57)
BMOF	18.1 (0.51)	11.5 (0.31)	27.1 (0.76)	16.6 (0.45)	34.8 (0.99)	25.2 (0.71)
One Face	1.5 (0.04)	1.3 (0.04)	12.0 (0.35)	4.8 (0.14)	32.7 (0.96)	7.5 (0.22)
Full	1.3 (0.04)	1.3 (0.04)	5.0 (0.14)	2.2 (0.06)	33.8 (0.92)	21.6 (0.59)
Snap Ring						
MOF	2.0 (0.16)	2.0 (0.16)	2.1 (0.14)	2.1 (0.15)	1.2 (0.11)	8.5 (0.76)
BMOF	3.1 (0.20)	3.1 (0.22)	6.9 (0.44)	6.9 (0.45)	2.6 (0.16)	10.1 (0.63)
Full	1.0 (0.06)	1.0 (0.07)	1.2 (0.08)	1.2 (0.08)	7.2 (0.49)	4.9 (0.33)
Big Head						
MOF	2.5 (0.09)	2.3 (0.09)	3.5 (0.13)	3.2 (0.12)	13.4 (0.50)	21.5 (0.81)
BMOF	37.5 (0.83)	14.6 (0.34)	37.9 (0.89)	31.1 (0.77)	43.4 (1.04)	30.5 (0.73)
One Face	1.7 (0.05)	1.6 (0.05)	4.8 (0.15)	4.7 (0.14)	31.4 (0.97)	15.6 (0.48)
Full	1.6 (0.04)	1.6 (0.04)	9.9 (0.23)	9.9 (0.23)	39.6 (0.97)	13.8 (0.34)
Cotter						
MOF	2.5 (0.08)	2.5 (0.10)	15.5 (0.52)	17.7 (0.60)	2.1 (0.07)	19.6 (0.71)
BMOF	8.3 (0.21)	8.8 (0.21)	34.0 (0.87)	34.0 (0.87)	15.2 (0.39)	22.2 (0.56)
One Face	1.3 (0.03)	1.4 (0.03)	5.8 (0.15)	6.2 (0.16)	35.2 (0.90)	8.5 (0.22)
Full	13.8 (0.34)	8.9 (0.22)	17.1 (0.44)	18.2 (0.42)	34.9 (0.88)	18.2 (0.46)
Hanger						
MOF	2.3 (0.08)	2.2 (0.09)	2.5 (0.09)	2.6 (0.10)	2.8 (0.12)	9.8 (0.42)
BMOF	9.1 (0.24)	5.1 (0.14)	25.2 (0.71)	25.1 (0.71)	21.1 (0.60)	19.6 (0.56)
One Face	1.5 (0.04)	1.4 (0.04)	4.0 (0.11)	3.3 (0.09)	30.9 (0.88)	9.0 (0.26)
Full	2.6 (0.08)	1.9 (0.06)	6.6 (0.19)	4.2 (0.12)	34.2 (0.95)	6.4 (0.18)



(a) Results for 100 randomly sampled *simulated* contacts across four hanger grid sizes. Mini One Face (MOF) is the smallest and simplest grid, while Full is the largest and most complex.

(b) Results on the *hanger* dataset with over 100 contacts, using the full-sized grid for all methods. We evaluate all methods on the same *real* (left) and *simulated* (right) contact shapes.

Figure 2-8: **Baseline results on the object hanger.** Normalized pose error for the object hanger using tactile matching (Tac2Pose) and three baseline methods. We compare the performance of each method on simulated contacts from different grid sizes in (2-8a), and on real vs. simulated versions of the contacts in the hanger dataset in (2-8b).

2. *Bigger Mini One Face*: We include only one grasp approach direction, and 10 angles (angular resolution of 36 degrees). We consider translations in x, y with 5mm resolution. This grid is smaller than the original, and ranges from 331 to 1137 poses, depending on the object. It is effectively 3D (x, y, θ).
3. *One Face*: We include only one grasp approach direction, with full angular (6 degrees) and translational (2.5mm) resolution. This grid ranges in size from 4.8k to 26.8k poses, depending on the object. It is effectively 3D (x, y, θ). For snap ring, which only has one grasp approach direction even in the full sized grid, this grid is equivalent to the original and is therefore omitted.

Comparing results between the reduced grids can provide insight into each method’s sensitivity to problem size, resolution, and complexity. For example, comparing *mini one face* and *one face* can indicate the sensitivity of each method to grid size and resolution. Comparing *one face* and the original grid (called *full* from here on out), which have the same resolution but cover different faces, provides insight into how each method handles complexity in the form of 3D rotations. *Mini one face* and *bigger mini one face* have the same translational resolution, but *bigger mini one face* varies the angle in

the plane of the grasp. Comparing results on these two grids provides insight into how incorporating rotations in the plane of the grasp impacts the accuracy of each method.

We first conduct experiments using only simulated data. The full set of results on the 5 baseline objects are listed in Table 2.3. We consider the results for hanger, visualized in Figure 2-8a, for simulated contacts from `mini one face`, `bigger mini one face`, `one face`, and `full`. The trends we note for this object are representative of the trends for the remaining objects.

The performance of `classification` (green bar) is inversely correlated with grid size; its performance is comparable to Tac2Pose only for `mini one face` (MOF in Figure 2-8a), which has 124 contacts. The normalized error is 0.12, compared with 0.08 for the single contact case of Tac2Pose. This indicates that although `classification` could be effective for very small grid sizes, the grids quickly become too large. Even with `bigger mini one face`, which contains 794 transformations, `classification` has a normalized error of 0.6, making it more than six times worse than Tac2Pose (single contact), which has 0.09 normalized error.

The `classification` baseline struggles with larger grid sizes because it needs to provide a classification over all elements in the grid, which often contains thousands of elements (the `full` grid for hanger, for example, has more than 90k contact poses). In comparison with the `classification` baseline, Tac2Pose scales better to large grid sizes because it learns to generate an embedding space based on the distance between contact poses. The encoder in Tac2Pose learns to push distant contacts away in embedding space, and therefore when comparing a new encoding to all the encodings in the grid, this direct comparison scales better than `classification` when the set of contacts in the grid is large. Instead, in `classification` the learned NN needs to encapsulate all the information relevant about all poses solely in the NN weights. Another way to think about the difference is to consider Tac2Pose as describing each contact in the grid with 1000 parameters (the encoder), whereas with `classification`, each contact in the grid is described by a single parameter (0 or 1).

The `pose` baseline (purple bar), on the other hand, is most comparable to Tac2Pose for the larger grids with high resolution. For `one face` (OF in Figure 2-8a), which covers one

object face with the same translational and rotational resolution as the full grid, the normalized pose error using pose is 0.26, compared with 0.04 for the single contact version of Tac2Pose. For the full grid, the normalized error for pose and the single contact version of Tac2Pose is 0.18 and 0.08, respectively. In the absence of contact shape noise (simulated contacts), Tac2Pose is about 2.3 times better on the full grid. Recall that the pose baseline does not match observed contacts to contact poses on the grid, but instead regresses the corresponding contact pose directly. In pose, the learned NN needs to encapsulate all information relevant about all poses solely in the NN weights, as it doesn't have access to embeddings it can compare against.

The pixel baseline, for both single contact and parallel jaw contact cases, performs most comparably to Tac2Pose for the mini one face grid. The mini one face grid is small and simple enough (one angle, one face) that all methods (excluding pose) perform comparably. For one face and full, Tac2Pose outperforms pixel by about 2-3 times for both single contact and parallel jaw cases, which is still relatively comparable. This implies that for pixel is most successful for high resolution grids. The performance of the pixel baseline method degrades most when the grid is low resolution, particularly when the grid contains object rotations. Bigger mini one face (shown as BMOF in the figure) has 0.71 median normalized error, compared with 0.24 for Tac2Pose (single contact). With parallel jaw contacts, the pixel baseline has 0.71 median normalized error, compared with 0.14 for Tac2Pose. Tac2Pose outperforms pixel by nearly 3 times in the single contact case, and more than 5 times in the parallel jaw case. Furthermore, pixel (both single contact and parallel jaw contact cases) is the least effective method when the grid is low resolution but complex.

When the grid resolution is low, the nearest grid match will be farther (in pixel distance) from an observed contact than when the grid resolution is high. This is particularly harmful for pixel when rotations are introduced, because a rotated version of a remote contact may, by chance, have more pixel overlap with the observed contact than the true closest match. This is much less likely to occur for high resolution grids, in which there exists a very close match on the grid to any observed contact. When the grid is simple (does not contain planar or 3D rotations) pixel can be successful even when the grid resolution is

low (as is the case for `mini one face`) because the best match on the grid is still likely to have the smallest pixel distance. If the grid is sufficiently complex, it must also be high resolution for `pixel` to be relatively successful (as is the case for `one face` and `full`). It is worth noting that a drawback of `pixel`, particularly with high resolution grids, is the execution time. In order to choose the best match, an observed contact is compared with contact shapes corresponding to every contact pose in the grid. Execution time is doubled when considering parallel jaw information, because two contacts are compared for each contact pose. `Pixel` is therefore the slowest method we evaluated by a significant margin. In practice, matching a contact to a high resolution grid in real-time using `pixel` is likely infeasible. Instead, `Tac2Pose` only compares low-dimensional embeddings resulting in at least an order of magnitude speed-up.

Finally, `Tac2Pose` performs consistently for both single and parallel jaw simulated contacts for all grid sizes for `hanger`. `Tac2Pose` also outperforms all baselines for all grid sizes. The parallel jaw case of `Tac2Pose` slightly outperforms the single contact case for all `hanger` grid sizes, except `mini one face`, in which the single contact case outperforms parallel jaw by 0.08 versus 0.09 normalized error. `Tac2Pose` (both single contact and parallel jaw) also performs slightly better for high resolution grids (`one face` and `full`).

2.3.2 Baseline Comparison with Real Data

We next compare `Tac2Pose` against the three baselines, using the real datasets evaluated previously and full sized grids. The full set of results are listed in Table 2.4. We also evaluate simulated versions of the contacts in the real datasets, to assess each method’s sensitivity to noise in the contact shapes. The results on simulated versions of the contacts in the real datasets are listed in Table 2.5. We discuss in detail the results for `hanger`, visualized in Figure 2-8b. The trends for this object are representative of the trends we see overall.

For real data, and full-sized grids, the performance `classification` (`green bar`), and `pixel` with a single contact (`dark orange bar`) are similar to selecting a contact pose randomly from the grid. The median normalized error is 0.98 for `classification`, and 0.90 for `pixel` with a single contact. `Pixel` with parallel jaw contacts and pose perform better,

Table 2.4: **Baseline results on real data.** Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods, on *real* datasets. Columns labelled SC are evaluated with a single contact, while columns labelled PJ use parallel jaw information.

	Tac2Pose		Pixel		Classification	Pose
	SC mm (norm)	PJ mm (norm)	SC mm (norm)	PJ mm (norm)	SC mm (norm)	SC mm (norm)
Long Grease	26.6 (0.76)	3.3 (0.09)	32.8 (0.93)	6.0 (0.17)	33.3 (0.95)	25.3 (0.72)
Snap Ring	1.5 (0.10)	1.4 (0.10)	5.6 (0.39)	2.2 (0.15)	6.0 (0.42)	5.9 (0.41)
Big Head	7.8 (0.20)	6.1 (0.16)	27.6 (0.70)	11.7 (0.30)	35.0 (0.89)	33.8 (0.86)
Cotter	19.0 (0.49)	19.6 (0.51)	31.5 (0.81)	36.7 (0.95)	35.8 (0.93)	38.1 (0.99)
Hanger	6.6 (0.19)	2.6 (0.07)	31.3 (0.90)	20.5 (0.59)	34.2 (0.98)	18.3 (0.53)

and have errors almost two times lower than selecting a pose at random from the grid; the median normalized error is 0.59 for `pixel` with parallel jaw contacts, and 0.53 for `pose`. Tac2Pose (dark blue bar for single contact, light blue bar for parallel jaw contacts) significantly outperforms all baselines. The median normalized error for the single contact case is 0.19, and for the parallel jaw contact case is 0.07. Tac2Pose outperforms the next best baseline (`pose`) by nearly three times for the single contact case, and nearly eight times for the parallel jaw contact case.

On simulated versions of the same contacts, `pose` and `pixel` (single contact and parallel jaw contact cases) both perform better than they do with real contacts. The median normalized error for `pose` shrinks by over three times, dropping to 0.17 with simulated data. Similarly, for `pixel`, the normalized error in simulation shrinks six times for single contact, and over eight times for parallel jaw contacts, dropping to 0.15 and 0.07, respectively. This indicates the sensitivity of both `pose` and `pixel` to noise in the contact shapes. Tac2Pose, in comparison, is more robust to contact shape noise, particularly in the parallel jaw contact case. With simulated contacts, the median normalized error for the parallel jaw case of Tac2Pose shrinks about two times, dropping to 0.03. Furthermore, despite the better performance of `pose` and `pixel` on simulated contacts compared with real contacts, Tac2Pose yields the best performance on simulated contacts as well. With simulated con-

Table 2.5: **Baseline results on simulated versions of real datasets.** Pose error and normalized pose error (in parenthesis) for Tac2Pose versus baseline methods, on *simulated* versions of the real contacts used in Table 2.4. Single Contact and Parallel Jaw are abbreviated as SC and PJ, respectively.

	Tac2Pose		Pixel		Classification	Pose
	SC mm (norm)	PJ mm (norm)	SC mm (norm)	PJ mm (norm)	SC mm (norm)	SC mm (norm)
Long Grease	1.1 (0.03)	1.1 (0.03)	1.8 (0.05)	1.7 (0.05)	33.0 (0.94)	17.0 (0.48)
Snap Ring	1.0 (0.07)	1.0 (0.07)	1.2 (0.08)	1.2 (0.08)	6.0 (0.41)	4.5 (0.31)
Big Head	3.9 (0.10)	3.3 (0.08)	8.7 (0.22)	9.6 (0.24)	34.0 (0.87)	13.8 (0.35)
Cotter	1.3 (0.03)	1.3 (0.03)	17.9 (0.46)	17.7 (0.46)	35.0 (0.91)	13.2 (0.34)
Hanger	1.2 (0.03)	1.2 (0.03)	5.1 (0.15)	2.4 (0.07)	33.6 (0.96)	6.0 (0.17)

tacts, Tac2Pose is more than twice as good as the next best baseline (pixel with parallel jaw contacts).

The primary reason for pixel’s sensitivity to noise in the contact shapes is that it only evaluates the exact 2D location of the contact on the sensor. Tac2Pose, in comparison, is trained to match contact shapes resulting from slight perturbations of grid poses to the closest match on the grid. This makes Tac2Pose more robust to slight discrepancies between observed and grid contact shapes.

The sensitivity of pose to contact shape noise has to do, instead, with the method’s ability to generalize to out of distribution data in the form of real contact shapes. Both pose and Tac2Pose are trained on simulated versions of contact shapes, but Tac2Pose generalizes to real contact shapes much better. Tac2Pose computes the likelihood of an observed contact matching to a discrete set of simulated shapes. Therefore, the match does not need to be exact, just better than other possible options, for the localization to be accurate. The structure imposed by the grid therefore improves the ability of Tac2Pose to generalize to real, noisy contact shapes. Because pose regresses object poses directly, without the structure of the grid, its performance is more brittle and degrades when using real contacts.

The normalized error of classification, even with simulated contacts, is 0.96, re-

maintaining similar to selecting a pose randomly from the grid. The size of the grid is too large to be handled with `classification` (as addressed in Section 2.3.1), so it is not possible to comment on the impact of noisy contacts.

For the 5 objects we evaluate (`long grease`, `snap ring`, `big head`, `hanger`, and `cotter`), `Tac2Pose` performs significantly better than random (here, we define significantly better as more than twice as good, which can be identified as a median normalized error less than 0.5 in Table 2.4) for 4/5 objects in both the single contact and parallel jaw contact cases. In the parallel jaw contact case, the fifth object, `cotter`, has median normalized error of 0.51, which is just below twice as good as random. In comparison, `pose`, `classification`, and `pixel` with a single contact are only significantly better than random for 1/5 objects. `Pixel` with parallel jaw contacts is significantly better than random for 3/5 objects. Furthermore, with a single contact, `Tac2Pose` outperforms all baselines for 4/5 objects. `Tac2Pose` is between 1.5 and 4 times better than the next best baseline for all objects except for `long grease`, which performs marginally worse than `pose` in the single contact case (0.72 median normalized error for `pose`, compared with 0.76 for `Tac2Pose`). With parallel jaw contacts, `Tac2Pose` outperforms `pixel`, which is the only other method compatible with parallel jaw contacts, with by between 1.5 and 9 times, depending on the object.

A final observation is that for some objects, parallel jaw contact versions of `pixel` and `Tac2Pose` are more robust to contact shape noise than single contact versions. We consider `long grease` (results listed in Table 2.4) as an example. `Tac2Pose` with a single real contact has 0.76 median normalized error, compared with 0.09 for the parallel jaw contact case. This means that the localization performance improves around eight times when using parallel jaw contacts. Similarly, `pixel` with a single contact has 0.93 median normalized error, compared with 0.17 median normalized error with parallel jaw contacts; the localization performance is about 5.5 times better in the parallel jaw contact case. We also consider the performance of `Tac2Pose`, and `pixel`, on simulated contacts to evaluate the impact of contact shape noise on the discrepancy between single and parallel jaw contact cases. With simulated contacts, the localization performance is the same for `Tac2Pose` for both a single contact and parallel jaw contacts; the median normalized error is 0.03 in

both cases. The outcome is similar for `pixel`; the median normalized error is 0.05 in both the single contact and parallel jaw contact cases. The consistent localization performance between methods and contact configurations when using simulated contacts indicates that contact shape noise is the differentiating factor between single contact and parallel jaw contact cases when using real contacts. This means that robustness to noise is a key advantage of using parallel jaw contacts over a single contact, for both methods. Both `Tac2Pose` and `pixel` have similar performance in the parallel jaw case with real contacts as they do with idealized (simulated) contacts; it is 3 (`Tac2Pose`) to 3.5 (`pixel`) times easier to localize simulated contacts than real contacts. The discrepancy between simulated and real results is much more pronounced in the single contact case of both methods. This implies that the sim-to-real gap can be bridged, in part, by the inclusion of parallel jaw information.

We conclude this section with some remarks about the challenge of estimating an object's pose from a single noisy contact. `Tac2Pose` outperforms all baselines on 4/5 objects with a single contact, and all objects with parallel jaw contacts. Ultimately, though, even `Tac2Pose` struggles to perform significantly better than selecting a pose at random from the grid with a single contact for 1/5 objects (where significantly better than random is defined as more than twice as good). Estimating the pose of an object from a single, noisy, often non-unique contact is challenging. Therefore, perhaps the most important feature of `Tac2Pose` is that it outputs meaningful pose distributions over possible object poses (Figure 2-7a). This creates a natural framework for incorporating constraints from additional contacts, measurements of the robot state (such as the gripper opening), information from additional sensing modalities, or even previous tactile estimates of the object pose.

Chapter 3

Conclusions

We conclude this thesis with some remarks on Tac2Pose , and directions for future work to improve the single shot estimation step in Section 3.1. We also explore in more detail a direction of ongoing work, that leverages Tac2Pose in a tracking framework to estimate a sequence of poses through a stream of tactile images in Section 3.2. Estimating poses through a trajectory has the benefit of overcoming the non-uniqueness inherent in an individual image. Furthermore, such a perception framework is necessary to supervise in-hand manipulation sequences in which the object moves relative to the manipulator. Supervising this type of in-hand manipulation sequence is a key direction of future work.

3.1 Discussion

This thesis evaluates an approach to tactile pose estimation for objects with known geometry. Tac2Pose relies on learning an embedding completely in simulation that facilitates comparing real and simulated contact shapes. We can reconstruct contact shapes with high fidelity, using images from the real sensor and in simulation. We compare the embeddings of an observed contact shape with those of a precomputed dense set of simulated contact shapes to obtain the distribution over a dense set of possible object poses. The approach therefore allows to reason over pose distributions and to handle additional pose constraints.

We evaluate Tac2Pose on real grasp datasets for 20 objects, and report the accuracy for three ablations of Tac2Pose, corresponding to increasing amounts of information from a

given grasp. First, we evaluate the accuracy using a single tactile image which corresponds to the case where only one contact is available to the algorithm. Second, we consider the case where two tactile images (corresponding to a parallel jaw grasp on the object) and the gripper opening are available. Third, we evaluate the accuracy after filtering the distribution obtained using parallel jaw information with a coarse prior on the object pose, approximating the case where information from an additional sensing modality (e.g. vision) is available.

We find that the amount of information needed to localize an arbitrary grasp on an object accurately is highly dependent on the object geometry. Of the 20 objects we evaluate, 9 can be localized accurately with a single contact. For these objects, more than half of arbitrary grasps on the object can be localized accurately. The objects that can be localized accurately with a single contact tend to be smaller, and have significant regions of unique contacts. When including parallel jaw information, the number of objects that can be localized accurately using the best match from Tac2Pose increases to 12. The objects that can be localized accurately after including parallel jaw information tend to have pseudosymmetrical features which can be disambiguated with a second contact, or vary significantly in width across the length of the object, which makes the gripper opening a discriminative source of information. After filtering the parallel jaw pose distribution with a coarse 10mm prior on the object pose, the number of objects that can be localized accurately increases to 16. The objects that can be localized accurately after including a coarse prior on the object pose tend to be larger, and have *discrete nonuniqueness* (features that are unique, with a discrete number of exceptions on remote regions of the object).

There are, however, four objects which cannot be localized accurately on average even after filtering the parallel jaw distribution with a 10mm prior and selecting the resulting most likely object pose. These objects are large, and have significant, continuous regions of non-unique contacts. The majority of contacts on such objects do not provide enough information to uniquely determine their pose.

Even objects which can be, on average, localized accurately from an arbitrary grasp have regions of non-unique contacts that do not provide enough information to uniquely determine the object pose. It is therefore important to evaluate the localization accuracy of

specific grasps on objects, as well as arbitrary grasps. To this end, we compare the localization accuracy separated out by grasp approach direction, and the localization accuracy of individual grasps on selected objects. We find that many objects have one direction which is significantly easier to localize (half as much error or less) than others. By examining individual grasps on the object `long_pencil`, we find that even objects with large, continuous regions of non-unique contacts have regions that are more unique and easier to localize.

Both of these breakdowns (sets broken out by grasp approach direction, or individual grasps) could ultimately be leveraged in a grasp planning framework, in which grasps that are easier to localize are specifically targeted. In practice, grasps within a given grasp approach direction could be targeted with a very coarse prior on the object pose, while a specific individual grasp could be targeted with somewhat finer information about the object pose. Because Tac2Pose is trained entirely in simulation, it is possible to evaluate which grasps lead to lower localization error in advance of encountering the object. For instance, the output pose distributions we obtain using simulated and real contact shapes on `long_pencil` are qualitatively similar (Figure 2-7a and 2-7b). Simulated contacts can therefore be used to identify unique and non-unique candidate grasps. This feature, too, could be important to leverage in a manipulation planning framework.

Our approach also assumes we have access to accurate geometric models of objects. We demonstrate that Tac2Pose can be effectively combined with object models reconstructed from a 3D scanner. A key feature of Tac2Pose is the ability to refine coarse estimates of object pose with high resolution tactile information. We find that moderate shape noise does not significantly compromise that ability for any of the 5 objects we evaluate.

Future work could learn embeddings that are even more robust to shape uncertainty by corrupting the contact shapes with noise during training. This would not inherently improve the localization accuracy when considering only information from a single grasp, but could yield distributions that are more representative of our true confidence in the estimate. These distributions could be combined effectively with other information (e.g. pose priors, additional tactile measurements, or dynamics) to converge on a unique estimate of the object pose.

We demonstrate the advantages of Tac2Pose compared with baseline methods for estimating object pose from tactile images. Compared with a standard classification approach, direct pose regression, and direct pixel comparison, Tac2Pose scales better to larger, and more complex problems (e.g. regressing 6D object pose rather than simply translation or rotation in plane) and is more robust to contact shape noise.

In summary, we demonstrate the effectiveness of Tac2Pose at pose estimation for 20 objects with information gathered from a single, real grasp on the object. We consider known and reconstructed object shapes, and show that in both cases Tac2Pose outputs meaningful distributions over object pose that can be maximized directly, or combined with additional information to converge on a unique estimate of object pose. While Tac2Pose is trained entirely in simulation, it is robust to real and noisy contact shapes arising from both sensor noise and object shape noise. This robustness stands in contrast baseline methods, which are more sensitive to contact shape noise. This suggests that matching estimated contact shapes to a dense precomputed set opens the door to moving many computations into simulation, without loss of robustness, and improving how robots learn to perceive and manipulate their environment.

3.2 Ongoing Work

Ongoing work tackles the problem of tracking global object pose through trajectories of tactile images, to overcome the non-uniqueness inherent in most individual tactile images. This thesis explores techniques for injecting additional information to recover from non-unique contacts, such as parallel jaw contact information (two contacts on opposite sides of a parallel jaw gripper, and the gripper opening) and a coarse prior on the object pose (in practice, such a prior could come from other sensing modalities, like vision). Ongoing work, described here, takes the opposite approach. Rather than assuming we have more information about a given instance of contact, it takes a stream of single tactile images over time.

We propose a hybrid smoother with three primary goals: first, it should accurately estimate the global object pose through a unique tactile trajectory. Second, it should be fast

enough to ultimately inform real-time control of an in-hand manipulation sequence, which is the intended use-case for such a tracking algorithm. Finally, it should be able to quantify the uniqueness of a tactile trajectory, in order to determine when a tactile trajectory becomes unique. The limit of a trajectory is a single image, so it is crucial to be able to specify when a trajectory contains enough information that the estimate becomes unambiguous. This information can, at a minimum, be used to inform at what point the estimate from the hybrid smoother can be trusted. In the future, such information could also be used in a decision-making policy that determines in which direction to move the sensor to reduce estimate uncertainty most quickly. The hybrid smoothing solution consists of two levels of inference that operate simultaneously: a discrete smoother, and a continuous smoother. At a high-level, the discrete portion guides the continuous portion with coarse, but globally correct estimates of the object pose. The continuous portion then refines those estimates to arbitrary accuracy.

3.2.1 Discrete Smoother

The discrete smoother is able to solve for coarse (both spatially and temporally), but globally correct, estimates of the object pose by reasoning over full measurement distributions in a discretization of pose space. The pose discretization is defined by the object grid. The object grid is the set of possible contacts on the object. It contains pairs of object poses relative to the sensor, and contact images that would result if the object were in that pose.

We estimate the history of object states over a window of time (called a smoothing window), given two types of constraints. First, Tac2Pose assigns likelihood each of the poses in the grid given an observed tactile image. Second, we constrain the type of transitions that can occur. Because we know that the object is sliding continuously, only transitions to near neighbors in the grid are feasible, and equally likely. If we had infinite time and computational resources, the problem would be more or less solved by the discrete smoother. However, a key challenge of the discrete approach is managing complexity. Restricting the spatial and temporal resolution of the discrete smoother allows us to get globally correct, but low resolution in both space and time, updates of the object pose. Combining the

discrete smoother with a continuous smoother allows us to refine the spatial and temporal resolution of our estimates, such that they can be used to supervise an in-hand manipulation sequence.

3.2.2 Continuous Smoother

A continuous smoother is introduced to track the object pose more closely. We guide the continuous smoother with the discrete estimates, and get a combined framework where each piece is strong where the other is weak.

The continuous portion looks very much like a standard smoothing problem, with unimodal gaussian factors. In contrast to the discrete framework, this formulation allows for fast, arbitrarily accurate updates to the object pose, meaning that it can regress any pose in $SO3$, and is not restricted to lie within the grid. Without guidance from the discrete portion, however, it would suffer from nonuniqueness. In parallel to the discrete smoother, the continuous smoother has two types of factors. First, we center a gaussian noise model at a pose guided by the discrete smoother as the measurement model. Next, we center a gaussian noise model at the odometry measurement of the robot as the dynamics model. Layering the continuous smoother on top of the discrete smoother allows for for more continuous, accurate tracking of the object pose along a tactile trajectory.

Bibliography

- [1] Peter K. Allen, Andrew T. Miller, Paul Y. Oh, and Brian S. Leibowitz. Integration of vision, force and tactile sensing for grasping. *Int. J. Intelligent Machines*, 4:129–149, 1999.
- [2] Maria Bauza, Oleguer Canal, and Alberto Rodriguez. Tactile mapping and localization from high-resolution tactile imprints. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2019.
- [3] Maria Bauza, Eric Valls, Bryan Lim, Theo Sechopoulos, and Alberto Rodriguez. Tactile object pose estimation from the first touch with geometric contact rendering. *CoRR*, abs/2012.05205, 2020. URL <https://arxiv.org/abs/2012.05205>.
- [4] J. Bimbo, S. Luo, K. Althoefer, and H. Liu. In-hand object pose estimation using covariance-based tactile to geometry matching. *IEEE Robotics and Automation Letters*, 2016.
- [5] Joao Bimbo, Petar Kormushev, Kaspar Althoefer, and Hongbin Liu. Global estimation of an object’s pose using tactile sensing. *Advanced Robotics*, 29(5):363–374, 2015.
- [6] Maxime Chalon, Jens Reinecke, and Martin Pfanne. Online in-hand object localization. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2977–2984. IEEE, 2013.
- [7] Yevgen Chebotar, Oliver Kroemer, and Jan Peters. Learning robot tactile sensing for object manipulation. In *International Conference on Intelligent Robots and Systems*. IEEE, 2014.
- [8] S.R. Chhatpar and M.S. Branicky. Localization for robotic assemblies using probing and particle filtering. In *Proceedings, 2005 IEEE/ASME International Conference on Advanced Intelligent Mechatronics.*, pages 1379–1384, 2005. doi: 10.1109/AIM.2005.1511203.
- [9] Craig Corcoran. Tracking object pose and shape during robot manipulation based on tactile information. In *International Conference on Robotics and Automation (ICRA)*, 2010.
- [10] Frank Dellaert and Michael Kaess. *Factor Graphs for Robot Perception*. Now Publishers Inc., August 2017.

- [11] Xinke Deng, Arsalan Mousavian, Yu Xiang, Fei Xia, Timothy Bretl, and Dieter Fox. Poserbpf: A rao-blackwellized particle filter for 6d object pose tracking. *CoRR*, abs/1905.09304, 2019. URL <http://arxiv.org/abs/1905.09304>.
- [12] Pietro Falco, Shuang Lu, Andrea Cirillo, Ciro Natale, Salvatore Pirozzi, and Dongheui Lee. Cross-modal visuo-tactile object recognition using robotic active exploration. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 5273–5280. IEEE, 2017.
- [13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. *arXiv preprint arXiv:1911.05722*, 2019.
- [15] Francois Robert Hogan, Maria Bauzá, Oleguer Canal, Elliott Donlon, and Alberto Rodriguez. Tactile regrasp: Grasp adjustments via simulated tactile transformations. *CoRR*, abs/1803.01940, 2018.
- [16] Jarmo Ilonen, Jeannette Bohg, and Ville Kyrki. Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing. *The International Journal of Robotics Research*, 2014.
- [17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.
- [18] Gregory Izatt, Geronimo Mirano, Edward Adelson, and Russ Tedrake. Tracking objects with point clouds from vision and touch. In *International Conference on Robotics and Automation*, 2017.
- [19] Shervin Javdani, Matthew Klingensmith, J Andrew Bagnell, Nancy S Pollard, and Siddhartha S Srinivasa. Efficient touch based localization through submodularity. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2013.
- [20] MyungHwan Jeon and Ayoung Kim. Prima6d: Rotational primitive reconstruction for enhanced and robust 6d pose estimation. *IEEE Robotics and Automation Letters*, PP:1–1, 06 2020.
- [21] M. C. Koval, M. Klingensmith, S. S. Srinivasa, N. S. Pollard, and M. Kaess. The manifold particle filter for state estimation on high-dimensional implicit manifolds. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4673–4680, May 2017.
- [22] Michael C Koval, Nancy S Pollard, and Siddhartha S Srinivasa. Pose estimation for planar contact manipulation with manifold particle filters. *The International Journal of Robotics Research*, 2015.

- [23] Naveen Kuppaswamy, Alejandro Castro, Calder Phillips-Grafflin, Alex Alspach, and Russ Tedrake. Fast model-based contact patch and pose estimation for highly deformable dense-geometry tactile sensors. *IEEE Robotics and Automation Letters*, PP: 1–1, 12 2019. doi: 10.1109/LRA.2019.2961050.
- [24] Yann Labbé, Justin Carpentier, Mathieu Aubry, and Josef Sivic. Cosypose: Consistent multi-view multi-object 6d pose estimation. *CoRR*, abs/2008.08465, 2020. URL <https://arxiv.org/abs/2008.08465>.
- [25] Nathan F Lepora, Alex Church, Conrad De Kerckhove, Raia Hadsell, and John Lloyd. From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor. *IEEE Robotics and Automation Letters*, 4(2):2101–2107, 2019.
- [26] Rui Li, Robert Platt, Wenzhen Yuan, Andreas ten Pas, Nathan Roscup, Mandayam A. Srinivasan, and Edward Adelson. Localization and manipulation of small parts using gelsight tactile sensing. In *Intelligent Robots and Systems (IROS), 2014 IEEE/RSJ International Conference on.*, 2014.
- [27] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. Deepim: Deep iterative matching for 6d pose estimation. *CoRR*, abs/1804.00175, 2018. URL <http://arxiv.org/abs/1804.00175>.
- [28] Shan Luo, Wenxuan Mou, Kaspar Althoefer, and Hongbin Liu. Localizing the object contact through matching tactile features with visual map. *CoRR*, abs/1708.04441, 2017.
- [29] Anton Milan, Trung Pham, K Vijay, Douglas Morrison, Adam W Tow, L Liu, J Erskine, R Grinover, A Gurman, T Hunn, N Kelly-Boxall, D Lee, M McTaggart, G Rallos, A Razjigaev, T Rowntree, T Shen, R Smith, S Wade-McCue, Z Zhuang, C Lehnert, G Lin, I Reid, P Corke, and J Leitner. Semantic segmentation from limited training data. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.
- [30] Anna Petrovskaya and Oussama Khatib. Global localization of objects via touch. *IEEE Transactions on Robotics*, 27(3):569–585, 2011.
- [31] Zachary Pezzementi, Caitlin Reyda, and Gregory D Hager. Object mapping, recognition, and localization from tactile geometry. In *International Conference on Robotics and Automation (ICRA)*, 2011.
- [32] Robert Platt, Frank Permenter, and Joseph Pfeiffer. Using bayesian filtering to localize flexible materials during manipulation. *IEEE Transactions on Robotics*, 27(3):586–598, 2011.
- [33] D Roşca, A Morawiec, and M De Graef. A new method of constructing a grid in the space of 3d rotations and its applications to texture analysis. *Modelling and Simulation in Materials Science and Engineering*, 22(7):075013, 2014.

- [34] Brad Saund, Shiyuan Chen, and Reid Simmons. Touch based localization of parts for high precision manufacturing. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.
- [35] Monika A Schaeffer and Allison M Okamura. Methods for intelligent localization and mapping during haptic exploration. In *International Conference on Systems, Man and Cybernetics*. IEEE, 2003.
- [36] Max Schwarz, Christian Lenz, Germán Martín García, Seongyong Koo, Arul Selvam Periyasamy, Michael Schreiber, and Sven Behnke. Fast object learning and dual-arm coordination for cluttered stowing, picking, and packing. In *International Conference on Robotics and Automation (ICRA)*. IEEE, 2018.
- [37] Paloma Sodhi, Michael Kaess, Mustafa Mukadam, and Stuart Anderson. Learning tactile models for factor graph-based state estimation. *CoRR*, abs/2012.03768, 2020. URL <https://arxiv.org/abs/2012.03768>.
- [38] Paloma Sodhi, Michael Kaess, Mustafa Mukadam, and Stuart Anderson. Patchgraph: In-hand tactile tracking with learned surface normals. *CoRR*, abs/2111.07524, 2021. URL <https://arxiv.org/abs/2111.07524>.
- [39] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.
- [40] Ian Taylor, Siyuan Dong, and Alberto Rodriguez. Gelslim3.0: High-resolution measurement of shape, force and slip in a compact tactile-sensing finger. *CoRR*, abs/2103.12269, 2021. URL <https://arxiv.org/abs/2103.12269>.
- [41] Stephen Tian, Frederik Ebert, Dinesh Jayaraman, Mayur Mudigonda, Chelsea Finn, Roberto Calandra, and Sergey Levine. Manipulation by feel: Touch-based control with deep predictive models. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 818–824. IEEE, 2019.
- [42] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson. 3D Shape Perception from Monocular Vision, Touch, and Shape Priors. *ArXiv e-prints*, August 2018.
- [43] Anna Yershova, Swati Jain, Steven M Lavalley, and Julie C Mitchell. Generating uniform incremental grids on so (3) using the hopf fibration. *The International journal of robotics research*, 2010.
- [44] Kuan-Ting Yu and Alberto Rodriguez. Realtime state estimation with tactile and visual sensing. application to planar manipulation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7778–7785. IEEE, 2018.

- [45] Andy Zeng, Shuran Song, Kuan-Ting Yu, Elliott Donlon, Francois Robert Hogan, Maria Bauza, Daolin Ma, Orion Taylor, Melody Liu, Eudald Romo, Nima Fazeli, Ferran Alet, Nikhil Chavan Dafle, Rachel Holladay, Isabella Morona, Prem Qu Nair, Druck Green, Ian J. Taylor, Weber Liu, Thomas A. Funkhouser, and Alberto Rodriguez. Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2018.
- [46] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. *CoRR*, abs/1812.07035, 2018. URL <http://arxiv.org/abs/1812.07035>.