TactoFind: A Tactile Only System for Object Retrieval

Sameer Pai[*,1,2], Tao Chen[*,1,2], Megha Tippur[*,2], Edward Adelson[2], Abhishek Gupta[†,1,2,3], Pulkit Agrawal[†,1,2]

[1]Improbable AI Lab   [2]Massachusetts Institute of Technology [3]University of Washington

[*]Authors contributed equally. [†]Equal Advising.

Abstract— We study the problem of object retrieval in scenarios where visual sensing is absent, object shapes are unknown beforehand and objects can move freely, like grabbing objects out of a drawer. Successful solutions require localizing free objects, identifying specific object instances, and then grasping the identified objects, only using touch feedback. Unlike vision, where cameras can observe the entire scene, touch sensors are local and only observe parts of the scene that are in contact with the manipulator. Moreover, information gathering via touch sensors necessitates applying forces on the touched surface which may disturb the scene itself. Reasoning with touch, therefore, requires *careful* exploration and integration of information over time – a challenge we tackle. We present a system capable of using sparse tactile feedback from fingertip touch sensors on a dexterous hand to localize, identify and grasp novel objects without any visual feedback. Videos are available at **https://sites.google.com/view/tactofind**.

I. INTRODUCTION

Consider the setup in Fig 1 wherein multiple objects are placed in a box with unknown positions and orientations. A multi-fingered robot is tasked to fetch a particular object using only tactile observations from sensors placed on fingertips, without access to visual observations. Such problems are commonly encountered in daily life — retrieving a desired object from the backpack, inside of a drawer, or from a tall cabinet where the topmost shelves are not visually observable. In these situations, the sense of touch can compensate for the lack of visual observations. To the best of our knowledge, the capability of a robotic system successfully retrieving a desired object based only on tactile observations has not been demonstrated in scenes with multiple freely moving and unknown objects.

To understand what makes vision-free object retrieval challenging, first consider a scenario where the robot can visually observe the scene. In such a case, it is (typically) possible to identify the location and identity of all objects from just a single image. Furthermore, the same image communicates enough information about object geometry to plan a grasp. In contrast, tactile observations made by the fingertips are local and sparse. Fingertips only make infrequent contact with a local area of the object. Therefore the robot first needs to explore to localize objects. Then it needs to plan a sequence of touches on each object to gather enough information to identify and grasp it. Unlike visual data, where a single observation is usually sufficient, tactile-based object retrieval requires careful planning of a long sequence of actions for information gathering and mechanisms for temporal integration of this information.
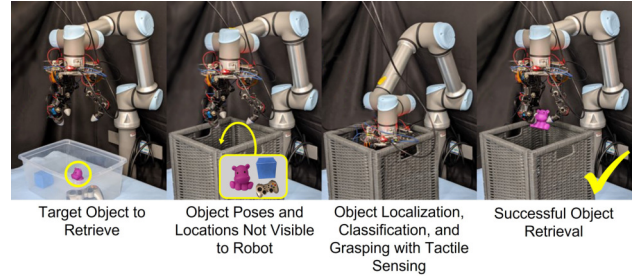

Fig. 1: System setup showing object retrieval from an occluded bin using only tactile sensing in the real world. The agent has to localize and identify the object from fingertip tactile sensors alone

Usually, one would expect that as the robot interacts more with the environment, it gathers more "information". However, this is not always true: generating touch readings requires applying force on an object that might move it. Such motion creates difficulty both in localization and identification. During localization, a moving object might contact another object previously encountered by the robot and potentially invalidate existing state estimates such as the object's location and pose. Similarly, during classification, if the object moves while being touched, the shape estimation is noisier which can make object identification difficult.

To summarize, the two key problems that make object retrieval in the absence of vision challenging are: **Firstly**, there is significant partial observability resulting from local and sparse touch observations. **Secondly**, obtaining touch readings requires force application on objects which may, in turn, lead to a change in object pose. While prior work has studied object classification and localization using tactile observations [1], [2], this has primarily been limited to single object scenes with a few classes of stationary objects.

We present a system that overcomes these challenges and is capable of exploring and retrieving novel objects only from local fingertip tactile observations in the presence of several movable objects, albeit in a simplified setup where individual objects are separated. Our solution employs a localization scheme that minimizes object motion while clustering to find object positions. Once objects are localized, the robot performs directed exploration to infer object shape by making careful touches around the determined object location. The shape information gathered from the sequence of contacts is represented as a point cloud. We leverage self-supervised contrastive learning [3], [4] to embed tactile point clouds into a feature space used for object identification. To retrieve localized and identified objects, a simple grasping system is deployed. We present results on a 3-fingered robotic hand-

arm system in both simulation and the real world. Our system achieves 76% success in simulation and 60% in the real world at retrieving novel objects without visual feedback.

II. RELATED WORK

In recent years, much progress has been made in the development and usage of tactile sensors to improve a robot's ability to complete dexterous manipulation tasks [5]. In particular, sensors utilizing a variety of transduction methods, such as resistance [6] , capacitance [7], piezo-electrical [8], magnetic [9], and optical [10], [11], [12], have been shown to help robots more intelligently interact with the physical world [12], [13], [14], [15]. In our work, we leverage tactile sensing to perform object retrieval completely in the absence of visual input.

A. Object Classification with Tactile Sensors

While there has been much work in object classification using a combination of image and tactile information [16], [17], [18], our problem requires us to rely only on tactile data for classification. [19], [20], and [11] classified different fabrics, textures, and basic geometries (respectively) using only camera-based tactile sensors. Others, such as [1], [21], [22], used force-based tactile sensors to accomplish the classification task but often assumed the object was static. [23] aims to do tactile-based classification by leveraging multi-modal information across vision and touch. A class of methods also aim to make full shape inference from tactile feedback [24], [25], [26], [27], [28] but typically assume the object is static and there is only a single object in the scene. In contrast, in this work, we also operate in the tactile-only regime, but we interact with multiple non-static objects when collecting the surface touches used in classification and just look to make object identification, not ideal shape reconstruction. Moreover, we are able to operate in a regime where we do not require any explicit object models or knowledge of object dynamics at test time, making the method easy to apply broadly to a variety of novel objects.

B. Object Localization with Tactile Sensing

Localization and manipulation of objects using only tactile sensing have been explored in prior work. While some works [29], [30] focused more on the dynamics and kinematics of the robot to perform localization, Li et al. [31] used a GelSight sensor to localize the sensed portion of single small objects like USB cables relative to the gripper. Bauza et al. [2] adopt a similar approach, combining heightmap information obtained from a tactile sensor with the robot's kinematics and the iterated closest point method to localize and even identify small objects. Other works [32] approach the localization problem by also building a visual map, but do so assuming objects are fixed. [33] perform contact-based localization to perform accurate manipulation when the object is in hand. In contrast to these works, we are less focused on fine-grained, precise, and in-hand localization than localization strategies that can deal with multiple, much

larger movable objects while also performing identification and grasping. We compromise significantly on accuracy to obtain a general strategy that can approximately localize objects without requiring known object shapes and dynamics by aiming to keep the scene static.

[34] goes beyond static objects and uses a SLAM approach to do planar object reconstruction and localization of non-static objects in the plane in single object scenes, leveraging a Gaussian process implicit surface method and particle filtering. Methods like [35], [36] also attempted to use only tactile sensing to retrieve a non-static object in a single object scene using particle filtering methods. Additionally, [37] uses a pre-touch sensor to localize and then senses object properties like stiffness and sliding for object discrimination. In contrast, in our work, we interact with multiple movable objects with a dexterous 3-fingered manipulator and perform the whole pipeline of localization, identification *and* grasping for object retrieval only using fingertip tactile sensors.

III. SYSTEM DESCRIPTION

Fig 1 illustrates our system consisting of a three-fingered robotic hand mounted on a UR5e robotic arm (controlled by [38]), equipped with fingertip GelSight 360 [39] for touch sensing. **Robotic Hand**: as shown in Fig 1, we use the three-fingered D'Claw robotic hand system [40]. Each finger has three Dynamixel servo motors connected in series for a total of nine degrees of freedom. The fingers are position controlled. **Tactile Sensor**: we use omnidirectional GelSight sensors [39] on the tip of each finger of the D'Claw hand. These sensors contain fisheye camera lens surrounded by a soft elastomer gel that deforms when the sensor makes contact. The deformations are recorded as a RGB image by the camera. This GelSight sensor has a significantly larger area of contact (approximately 15 cm^2) with objects than most sensors since it wraps completely around the finger. This is important to actually help identify objects with fewer interactions. To obtain binary contact readings from the sensor's raw data (RGB images), we train a force recognition model (as described in [41]) in the real world to map images of contact from the GelSight sensor to binary labels of whether contact occurred or not.

IV. PROBLEM SETUP

We first task the agent by allowing it to interact with a given *target object* at a known location and pose, as shown in Fig 2. The target object is then placed in a planar bin area along with other *distractor objects*, each in a random unknown pose. The agent's goal is to leverage only tactile sensing to retrieve the target object. All objects are rigid and movable, and the agent has no prior knowledge of distractor object shapes. Furthermore, we assume no objects are stacked on each other for tractability. The robot observes fingertip positions $\{p_i \in \mathbb{R}^3, i = 1, 2, 3\}$ estimated through joint encoders, as well as a reading of binary contact $\{c_i \in \{0, 1\}, i = 1, 2, 3\}$ for each finger.
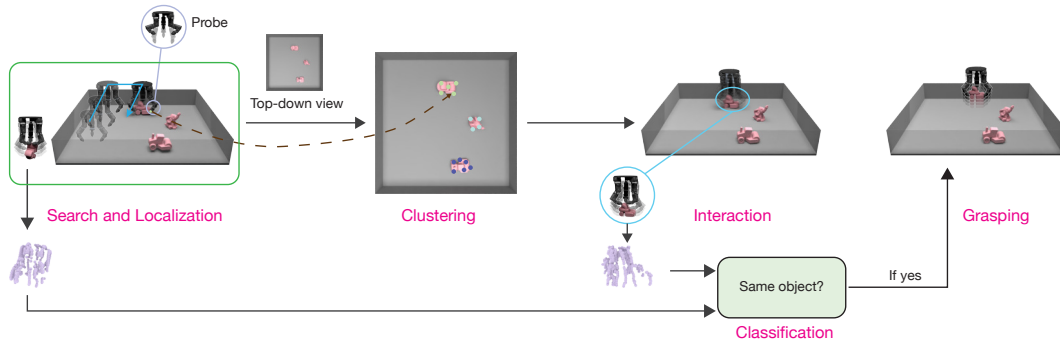
Fig. 2: Depiction of the full pipeline for tactile-only object retrieval. The agent first localizes several objects in the scene by vertically probing in a discrete grid around the environment, while applying minimal force. The object is then interacted with by radially tapping it gently for identification. The collected data is then used to identify if the object matches the object it is tasked to retrieve and then grasping is performed to actually retrieve the object.

V. TACTOFIND: IDENTIFYING AND RETRIEVING OBJECTS IN THE ABSENCE OF VISUAL FEEDBACK

The problem of blind identification and retrieval of novel and movable objects is an instance of *partially observable Markov decision process*, which is known to be intractable in general. To make this problem more approachable, instead of solving end-to-end, we can take a step-wise approach illustrated in Fig 2 and make assumptions detailed below. The robot first performs object localization, then instance identification, followed by grasping. In the *localization phase*, the location of all objects is estimated agnostic of their identity. During *instance identification*, an exploration algorithm is deployed to interact with each object. The contact data resulting from the interaction with each object is matched against the contact data of the target object for identification. Lastly, the position of the identified object is used to *grasp* and retrieve it. In each step, both interaction and inference algorithms are designed to minimize object motion and to integrate sparse tactile information for dealing with the challenges mentioned in Section I. We train the identification model in simulation and transfer it to the real world, whereas the localization and interaction strategies are directly implemented in the real world.

A. Object Localization with Tactile Sensing

For the task of interest, an object localization The main challenge in object localization is that object motion while localizing the N-th object may disrupt the estimates for any of the $N-1$ objects localized previously. Minimizing object motion is further complicated when object masses, shapes, and friction properties are unknown. Assuming that no two objects are vertically stacked and that objects do not touch each other, our approach to object localization consists of two steps: first, the workspace is discretized into a square grid (H) with fixed side length $\delta = 5cm$ (roughly half the average width of the target objects) and every location on the grid is probed to estimate if it is occupied by the object. The result is a binary *object occupancy map* (i.e., $H[i,j] \leftarrow \{0,1\}$). Next, H is clustered to obtain approximate locations for each object.

To obtain the occupancy map, $H[i,j]$ we move the gripper to the center of the grid location (i,j) and at a pre-specified

height above the bottom of the bin/table. The gripper is commanded to move down along the normal to the plane (Fig 2). To minimize object motion, as soon as any contact is encountered or if the tip of the gripper reaches the surface of the table, the arm's motion is stopped. If contact was encountered, the grid position is deemed to be occupied (i.e., $H[i,j] = 1$). The procedure is repeated for all grid locations. The occupancy map is clustered using K-means algorithm [42], with a known value of K corresponding to the number of objects in the bin. This procedure yields a list of approximate object center positions $\left[o_c^1, o_c^2, \ldots, o_c^N\right]$ (Fig 5), which can then be utilized for fine-grained object identification.

B. Dynamic Object Identification

The localization process above provides an occupancy map and object locations, which is insufficient to estimate the $3D$ shape of objects required to perform identification. Prior works have found that random One naive strategy to get more interaction data is to randomly move the fingers, and once in a while, the fingers will touch the object. However, such a naive strategy makes it hard to identify objects. The reason is that the object might move a lot due to random finger motions, causing noise in the data. When the object does not move, the fingertip locations when fingers make contact with the object represent the points on the object's surface. However, if the object moves a lot, the touch points are effects of both the object's geometry as well as the object's motion, which makes it more challenging to identify what object it is. Therefore, it is beneficial to have an interaction policy that explores the object's surface but also tries to minimize the object's motion.

After collecting such data, the next step is identifying whether the object is the target object we are searching for. A straightforward way is to train a classifier that outputs the object ID and see if it matches the target object ID. However, such a classifier cannot generalize to novel objects. Therefore, instead, we use contrastive learning to learn an embedding space in which we can determine whether the object is the same as the target object based on the distance between their embedding vectors.

a) Collecting Interaction Data: Instead of having the hand doing unstructured exploration around the object, which

obtains very few useful interaction data and can cause the object to have a big motion, we devised a radial sliding strategy. It tries to move the fingers to follow the contours of the objects in the vertical direction. We hypothesize that such slices of object contours along the vertical axis would be sufficient for identifying an object (as described in the next section) since they give a notion of object shape. However, in our experiments, we found that having the fingertips move along the object contours when object geometry is complex tends to cause the object to move. Instead, we command the fingers to *tap* along the object contours (*radial tapping strategy*). Tapping significantly reduces the object motion and allows us to identify free objects much more effectively than sliding. This is because the number of points of contact is reduced and the force inward can be controlled more carefully. As shown in Fig 3, we first reset the fingers to their initial positions, and then close the fingers (move the fingertips inwards to touch the objects), when touch happens on any of the fingertips ($c_j^t = 1$), we record the fingertip position p_j^t and stop moving the finger. Since each touch only generates one data point (the fingertip position p_j^t), we get up to three points at each time we close the fingers (potentially one from each finger). We move the fingers upward by a small distance and repeat such a process until the fingers touch the top of the object. By doing so, we get a sequence of contact positions $\mathcal{P}_o^x = \{p_0^0, p_1^0, p_2^0, p_0^1, p_1^1, p_2^1, \ldots, p_0^N, p_1^N, p_2^N\}$ for a series of N taps for object o at an unknown pose x, where each contact point p is a point in $\mathbb{R}^3$.

Some things to note — Firstly, the actual contact forces do not need to be used since we are just detecting binary contact on the surface. Secondly, the object itself may move since the fingers are not *guaranteed* to perfectly cage the object while the contact is being made. To account for potential object motion, we perform object relocalization by estimating the direction of motion from the contact points at each time step. Specifically, at every point in time, we maintain a current estimate $\hat{c}$ for the current object center. Each radial tap gives us a sequence of contact points $\{p_i\}$, which we average to get a single point p. We then perform an exponential smoothing update $\hat{c}_{new} = \gamma \hat{c}_{old} + (1-\gamma)p$ to get a smoothed new estimate, and move the hand in the xy-plane to $\hat{c}_{new}$. This allows us to correct for local object motions that may occur when interacting with the object without knowing the object model beforehand. Note that our goal is not to accurately estimate the center of the object. A rough estimate is sufficient for the hand to track its location. As long as the three fingers can surround the object, we can collect meaningful interaction data.

b) Representation Learning for Object Identification: The contact information we get from the interaction policy is a sequence of contact positions $\mathcal{P}_o^x$, which can be viewed as a very sparse point cloud. Our goal is to identify whether the obtained point cloud $\mathcal{P}_o^{x_1}$ and the point cloud obtained by interacting with the target object $\mathcal{P}_{o_{tgt}}^{x_2}$ belong to the same object. Note that even if they are the same object, the object can be in different poses. Therefore, we need a shape representation that is agnostic to the object orientation

but still allows distinguishing between different objects. We can learn such representation using contrastive learning [43], [4], [44], [45] — represent touches on the same object at different poses similarly and touches across different objects differently. One potential representation learning objective is the InfoNCE loss (as also discussed in [4]). Assuming we have some parametric encoder f with parameters θ, this loss can be expressed as:

$$\mathcal{L}_{\text{NCE}}(\theta) = \mathbb{E}_{(x,x^+)\sim \mathcal{D}_o, \{x_k\}\sim \mathcal{D}} \left[\log \frac{\exp f_\theta(x) \cdot f_\theta(x^+)}{\sum \exp f_\theta(x) \cdot f_\theta(x_k)} \right]$$

where the positive distribution $\mathcal{D}_o$ consists of touch sequences sampled from different poses of the same object o and the negative distribution $\mathcal{D}$ is touch sequences sampled from arbitrary poses of different objects. We treat the touch data (sequences of contact positions $\mathcal{P}_o^x$) as a sequential time serie, and parameterize the encoder f_θ with a transformer architecture.
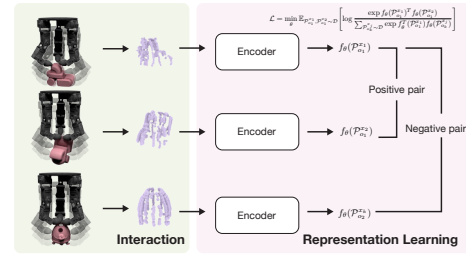

Fig. 3: Depiction of the contrastive representation learning architecture for object identification. The sequence of contact points goes through a transformer based encoder to provide embeddings that are trained with the InfoNCE loss. Affinity in this representation space can then be used for object identification

Once the representation is learned, the object of interest can be identified by comparing the cosine similarity of touch sequences in the representation space $z = f_\theta(\mathcal{P}_o^x)$ of the touches made blindly across various objects in the bin with the touches made on the target object outside of the bin. This identified object can be used for object retrieval using a grasping controller.

While more complex schemes could be developed for grasping [46], [47], we found that the hand lends itself very naturally to grasping with a simple caging policy. Concretely, we simply close the fingers until the fingertips touch the object with some amount of force. The touch sensors on the fingertips are used to tell whether the fingertips touch the object with a sufficient amount of contact area. After all three fingers touch the object with a sufficient amount of contact, we lift the arm to grasp the object. We depict this policy in Fig 2, and deploy it in an object-agnostic manner to grasp objects which have been localized and identified.

VI. EXPERIMENTAL EVALUATION

In this section, we aim to answer the following questions: (1) How effective is the overall pipeline at identifying and grasping particular object instances from tactile sensing alone? (2) Can a dexterous hand with tactile sensing localize

movable objects without visual input? (3) Is the data obtained by interacting with objects using the radial tapping strategy described in Section V-B effective for classification? (4) Is the radial tapping interaction strategy for object identification effective for movable objects? (5) Is representation learning using the scheme in Section V-B effective for identifying novel objects?

We used the 150 object meshes collected in [48]. The meshes are from various datasets including Google Scanned Objects [49] and ShapeNet [50]. We built the simulation environment in PyBullet. The 3-D printed versions of the objects were used for testing in the real world. For evaluation, we create test scenes by randomly choosing K objects from this set and placing them in random poses in a bin.

A. Baselines and Evaluation Metrics

While no entire system exists that completes the entire task described in this work, we compare it with several prior works that perform localization, identification, or grasping. For localization, we compare with [35] which uses particle filters for localization. To understand the importance of our particular interaction strategy, sensing modality and other design choices in terms of being able to collect information without moving the object significantly during object identification, we compared with two baselines: 1) radial tapping with noisy contact detection rather than accurate binary contact direction (noisy contact) to test the importance of accurately binarizing contact, 2) radial tapping without hand relocalization in cases where the object slips as discussed in Section V-B (no relocalization) to show the importance of the relocalization strategy. To understand the impact of specific model architecture, we compared with two baselines: 1) an LSTM [1] rather than a transformer with the InfoNCE objective, 2) using Triplet loss [51], [52] instead of InfoNCE loss function, to understand the importance of the particular choice of transformer and infoNCE objectives. **Evaluation Metrics**: For localization, we measure the error (in cm) between the identified object center and the ground truth object center of mass (clustering error). Additionally, we measure displacement from the original object positions to determine how disruptive exploration is (perturbation error). We evaluated this for scenes with 3 objects. For identification, we measure the success percentage in identifying the correct object out of 5 unique object instances that are randomly placed in the bin, as well as the number of successful taps made with the object.

B. Simulation Results

We report results on localizing objects in different scenes in simulation with various different objects in Table I. We see that our proposed clustering scheme gets within 3.6 cm on average across various objects while ensuring minimal displacement of only 1.7 cm during exploration when evaluated on test objects approximately 10 to 15 cm in diameter. In comparison, the particle filtering scheme used in [35] performs competitively with ours on scenes with single objects, achieving an average of $4.5cm$ of error.
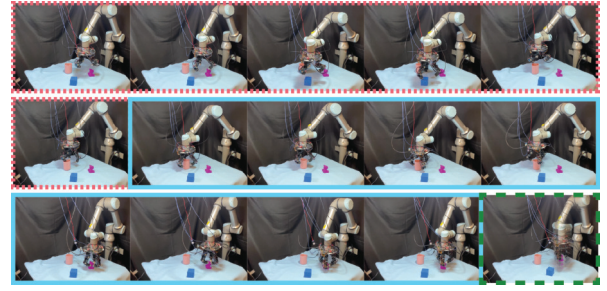

Fig. 4: Film strip depicting phases of real world localization and identification with our pipeline. We see that objects move minimally, while shapes are successfully identified and grasped. Red box shows the localization, blue box shows the interaction, and the green box shows the grasping.

However, when increasing the number of objects to three, the particle filtering struggles and the error increases to $6.4cm$ on average. This is likely due to the fact that increasing the number of objects also increases the dimensionality of the particles, meaning exponentially more particles are needed to achieve more accuracy.

TABLE I: Results on object localization in simulation. For both strategies, we measure the success rate (defined as when each predicted center is within $\delta = 7.5cm$ of the object center), the average prediction error, and the average displacement of the objects during the localization. Our clustering strategy scales to multiple objects much better than particle filtering, likely because multiple objects introduce a higher dimensional search space.

Method	Success Rate	Center Error	Object Displacement
Ours (1 object)	$99.2\% \pm 0.2\%$	$3.2 \pm 0.1cm$	$1.8 \pm 0.1cm$
Particle Filter (1 object) [35]	$94.6\% \pm 0.6\%$	$4.5 \pm 0.3cm$	$1.5 \pm 0.1cm$
Ours (3 objects)	$91.3\% \pm 1.7\%$	$3.6 \pm 0.2cm$	$1.7 \pm 0.1cm$
Particle Filter (3 objects) [35]	$52.8\% \pm 1.4\%$	$6.4 \pm 0.1cm$	$2.4 \pm 0.1cm$

Next, we show in Table II, that our proposed identification technique is able to select the correct object out of five novel objects with 69.8% accuracy.

Interaction: To understand why the interaction strategy is better than alternatives, we study the impact of removing individual elements of the interaction strategy. Table II shows the results of these ablations. First, we add random noise into the sensors, compromising the system's ability to quickly detect contact. This results in identification success dropping from 69.8% to 54.4%, likely due to greatly increased object motion. We also investigate the effects of removing the relocalization policy described in Section V-B. Without the ability to adapt to object motion, identification accuracy drops to 58.5%, a drop of over 10%. Therefore, we can conclude that both our relocalization policy, as well as accurate contact detection, are necessary to achieve our system's accuracy.

Identification: From Table III, we see that the transformer with the InfoNCE objective achieves a higher success rate than the alternatives. In particular, training with InfoNCE loss improves performance over using Triplet loss

Finally, in simulation, we evaluate the combined pipeline of object localization and identification (leaving grasping for real-world evaluation in Section VI-C) in terms of the percentage of trials where the successful object is identified and localized. We find in Table IV that our pipeline is able to accomplish a success rate of 76.8% in simulation.

Method	Accuracy (Validation Set)	Accuracy (Training Set)	Successful Taps
Ours	**69.8% ± 1.2%**	88.4% ± 0.9%	**228.4 ± 0.7**
No Relocalization	58.5% ± 1.4%	61.8% ± 1.3%	157.1 ± 0.8
Noisy Sensors	54.4% ± 1.4%	59.3% ± 1.4%	183.8 ± 0.7

TABLE II: Effectiveness of interaction strategies on object identification in simulation. We measure for each policy the overall accuracy when selecting out of five (previously unseen) objects on both the training and validation object sets, as well as the average number of successful taps the policy gets from the object. These experiments show that both our object position estimation as well as accurate contact detection are essential for successful object identification.

Method	Accuracy (5 objects)	Accuracy (3 objects)
Transformer + InfoNCE loss (Ours)	69.8% ± 1.2%	80.3% ± 1.0%
LSTM + InfoNCE loss	65.7% ± 1.3%	77.0% ± 1.2%
Transformer + Triplet loss	52.3% ± 1.4%	64.8% ± 1.3%

TABLE III: Effects of architecture and objective on object identification in simulation. Using a more expressive architecture like a transformer helps with classification, while the InfoNCE loss outperforms using a triplet loss.
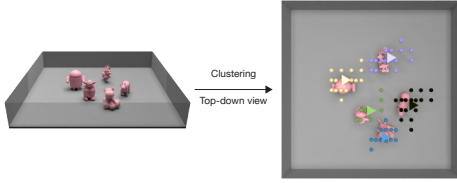

Fig. 5: Visualization of planar occupancy and resulting clusters (the dots) during localization compared with true object centers (the triangles). While showing non-zero error, the relative object centers are well predicted enough for subsequent object identification.

C. Real World Results

Success Rate	Overall	Localization	Identification
Sim	76.8% ± 2.3%	91.3% ± 1.7%	80.3% ± 1.0%
Real	59.4% ± 11.3%	81.0% ± 9.0%	75.7% ± 9.8%

TABLE IV: Results on full pipeline object retrieval in simulation and the real world. We find that there are expected drops in performance from simulation to the real system, but the identification is still able to significantly outperform random chance

The pipeline transfers relatively well to the real world. Due to the lack of ground truth object positions running on the real system, we qualitatively evaluate pipeline stage success. Specifically, we deem localization to be successful if the robot is able to get its three fingers around each object, and grasping to be successful if the robot is able to lift the chosen object. When evaluated on three objects at a time, we find that localization has a success rate of 78%. The object identification model trained in simulation has a success rate of 73% in the real world. And similarly, grasping is able to accomplish a success rate of 71%. When this entire system is executed sequentially, it accomplishes a success rate of 59.4%, as shown in Fig 4. While this leaves room for improvement, in each part of the pipeline, it is significantly better than chance and much better than a random policy.

D. Ablations

To understand how much moving objects affect classification performance, we perform an ablation study in Table V repeating the comparisons in Section VI-B with a fixed stationary object. We find that classification is significantly easier, showing the difficulty of scenarios with moving objects.

Method	Accuracy (5 objects)	No. of Successful Taps
Ours (Moving Objects)	69.8% ± 1.2%	228.4 ± 0.7
Ours (Static Objects)	86.2% ± 0.9%	235.4 ± 1.2

TABLE V: Effectiveness of interaction strategies on object identification in simulation for *static* objects. We run our object interaction strategy on both moving and static objects and find that static objects are easier to classify than moving ones, showing how challenging our problem setting is

Friction Coefficient	$\mu = 0.5$	$\mu = 0.25$	$\mu = 0.1$
Localization Accuracy	91.3% ± 1.7%	82.6% ± 1.8%	17.3% ± 1.7%
Identification	69.8% ± 1.2%	65.3% ± 1.3%	53.8% ± 1.4%

TABLE VI: Effect of the coefficient of friction (μ) on localization and classification performance in simulation. As friction reduce, localization becomes much more difficult and identification accuracy also decreases due to increased object movement.

To understand just how much these methods help with moving objects, we also ran an ablation study where we performed comparisons in Table VI as we change object friction and mass. The localization performance degrades as we use small friction and mass values, especially low friction. This is expected as on a slippery surface, an object moves more easily and suffers from large motion even if the robot hand applies a small force on it. However, our identification pipeline suffers a much smaller performance drop, indicating the capability of our method in handling non-static objects.

VII. DISCUSSION

The pipeline proposed in this work is only a starting point for tactile-only object localization and retrieval. While we have designed a strategy using tapping to minimize object movement, an interesting future research direction would be to explore how to localize, identify and grasp objects that can move substantially.

VIII. AUTHOR CONTRIBUTIONS AND ACKNOWLEDGEMENTS

REFERENCES

[1] Wolfgang Bottcher, Pedro Machado, Nikesh Lama, and Thomas M McGinnity. Object recognition for robotics from tactile time series data utilising different neural network architectures. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[2] Maria Bauza, Oleguer Canal, and Alberto Rodriguez. Tactile mapping and localization from high-resolution tactile imprints. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3811–3817. IEEE, 2019.

[3] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, pages 539–546. IEEE Computer Society, 2005.

[4] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018.

[5] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands. *Robotics and Autonomous Systems*, 74:195–220, 2015.

[6] Subramanian Sundaram, Petr Kellnhofer, Yunzhu Li, Jun-Yan Zhu, Antonio Torralba, and Wojciech Matusik. Learning the signatures of the human grasp using a scalable tactile glove. *Nature*, 569(7758):698–702, 2019.

[7] Clementine M Boutry, Marc Negre, Mikael Jorda, Orestis Vardoulis, Alex Chortos, Oussama Khatib, and Zhenan Bao. A hierarchically patterned, bioinspired e-skin able to detect the direction of applied pressure for robotics. *Science Robotics*, 3(24):eaau6914, 2018.

[8] Mark R Cutkosky, Robert D Howe, and William R Provancher. Force and tactile sensors. *Springer Handbook of Robotics*, 100:455–476, 2008.

[9] Raunaq Bhirangi, Tess Hellebrekers, Carmel Majidi, and Abhinav Gupta. Reskin: versatile, replaceable, lasting tactile skins. *arXiv preprint arXiv:2111.00071*, 2021.

[10] Benjamin Ward-Cherrier, Nicholas Pestell, Luke Cramphorn, Benjamin Winstone, Maria Elena Giannaccini, Jonathan Rossiter, and Nathan F Lepora. The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft robotics*, 5(2):216–227, 2018.

[11] Naveen Kuppuswamy, Alex Alspach, Avinash Uttamchandani, Sam Creasey, Takuya Ikeda, and Russ Tedrake. Soft-bubble grippers for robust and perceptive manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9917–9924. IEEE, 2020.

[12] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[13] Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40(12-14):1385–1401, 2021.

[14] Chen Wang, Shaoxiong Wang, Branden Romero, Filipe Veiga, and Edward Adelson. Swingbot: Learning physical features from in-hand tactile exploration for dynamic swing-up manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5633–5640. IEEE, 2020.

[15] Siyuan Dong, Devesh K Jha, Diego Romeres, Sangwoon Kim, Daniel Nikovski, and Alberto Rodriguez. Tactile-rl for insertion: Generalization to objects of unknown geometry. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6437–6443. IEEE, 2021.

[16] Arkadeep Narayan Chaudhury, Timothy Man, Wenzhen Yuan, and Christopher G Atkeson. Using collocated vision and tactile sensors for visual servoing and localization. *IEEE Robotics and Automation Letters*, 7(2):3427–3434, 2022.

[17] Jingwei Yang, Huaping Liu, Fuchun Sun, and Meng Gao. Object recognition using tactile and image information. In *2015 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, pages 1746–1751. IEEE, 2015.

[18] Tadeo Corradi, Peter Hall, and Pejman Iravani. Object recognition combining vision and touch. *Robotics and biomimetics*, 4(1):1–10, 2017.

[19] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4842–4849. IEEE, 2018.

[20] Rui Li and Edward H Adelson. Sensing and recognizing surface textures using a gelsight sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1241–1247, 2013.

[21] Zachary A. Pezzementi, Caitlin Reyda, and Gregory D. Hager. Object mapping, recognition, and localization from tactile geometry. In *IEEE International Conference on Robotics and Automation, ICRA 2011, Shanghai, China, 9-13 May 2011*, pages 5942–5948. IEEE, 2011.

[22] Somchai Pohtongkam and Jakkree Srinonchat. Tactile object recognition for humanoid robots using new designed piezoresistive tactile sensor and dcnn. *Sensors*, 21(18):6024, 2021.

[23] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. *CoRR*, abs/1903.03591, 2019.

[24] Nawid Jamali, Carlo Ciliberto, Lorenzo Rosasco, and Lorenzo Natale. Active perception: Building objects' models using tactile exploration. In *16th IEEE-RAS International Conference on Humanoid Robots, Humanoids 2016, Cancun, Mexico, November 15-17, 2016*, pages 179–185. IEEE, 2016.

[25] Uriel Martinez-Hernandez, Giorgio Metta, Tony J. Dodd, Tony J. Prescott, Lorenzo Natale, and Nathan F. Lepora. Active contour following to explore object shape with robot touch. In *2013 World Haptics Conference, WHC 2013, Daejeon, Korea (South), April 14-17, 2013*, pages 341–346. IEEE, 2013.

[26] Zhengkun Yi, Roberto Calandra, Filipe Veiga, Herke van Hoof, Tucker Hermans, Yilei Zhang, and Jan Peters. Active tactile object exploration with gaussian processes. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2016, Daejeon, South Korea, October 9-14, 2016*, pages 4925–4930. IEEE, 2016.

[27] Danny Drieß, Peter Englert, and Marc Toussaint. Active learning with query paths for tactile object shape exploration. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2017, Vancouver, BC, Canada, September 24-28, 2017*, pages 65–72. IEEE, 2017.

[28] Jingxi Xu, Shuran Song, and Matei Ciocarlie. Tandem: Learning joint exploration and decision making with tactile sensors. *arXiv preprint arXiv:2203.00798*, 2022.

[29] Yiannis Karayiannidis, Christian Smith, Francisco E. Vina, and Danica Kragic. Online contact point estimation for uncalibrated tool use. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2488–2494, 2014.

[30] Nejc Likar and Leon Žlajpah. External joint torque-based estimation of contact information. *International Journal of Advanced Robotic Systems*, 11(7):107, 2014.

[31] 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 *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3988–3993. IEEE, 2014.

[32] 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.

[33] Michael C. Koval, Nancy S. Pollard, and Siddhartha S. Srinivasa. Pose estimation for planar contact manipulation with manifold particle filters. *Int. J. Robotics Res.*, 34(7):922–945, 2015.

[34] Ghani Kissoum and Veronique Perdereau. Simultaneous tactile localization and reconstruction of an object during robotic manipulation. In *2021 20th International Conference on Advanced Robotics (ICAR)*, pages 948–954. IEEE, 2021.

[35] Adithyavairavan Murali, Yin Li, Dhiraj Gandhi, and Abhinav Gupta. Learning to grasp without seeing. In Jing Xiao, Torsten Kröger, and Oussama Khatib, editors, *Proceedings of the 2018 International Symposium on Experimental Robotics, ISER 2018, Buenos Aires, Argentina, November 5-8, 2018*, volume 11 of *Springer Proceedings in Advanced Robotics*, pages 375–386. Springer, 2018.

[36] Anna Petrovskaya and Oussama Khatib. Global localization of objects via touch. *IEEE Trans. Robotics*, 27(3):569–585, 2011.

[37] Mohsen Kaboli, Di Feng, Kunpeng Yao, Pablo Lanillos, and Gordon Cheng. A tactile-based framework for active object learning and discrimination using multimodal robotic skin. *IEEE Robotics Autom. Lett.*, 2(4):2143–2150, 2017.

[38] Tao Chen, Anthony Simeonov, and Pulkit Agrawal. AIRobot. https://github.com/Improbable-AI/airobot, 2019.

[39] Megha H. Tippur. Design and manufacturing methods for a curved all-around camera-based tactile sensor. page 69, 2020.

[40] Michael Ahn, Henry Zhu, Kristian Hartikainen, Hugo Ponte, Abhishek Gupta, Sergey Levine, and Vikash Kumar. Robel: Robotics benchmarks for learning with low-cost robots. In *Conference on robot learning*, pages 1300–1313. PMLR, 2020.

[41] Huanbo Sun, Katherine J Kuchenbecker, and Georg Martius. A

soft thumb-sized vision-based sensor with accurate all-round force perception. *Nature Machine Intelligence*, 4(2):135–145, 2022.

[42] S. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982.

[43] R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742, 2006.

[44] Qiang Zhang, Yunzhu Li, Yiyue Luo, Wan Shou, Michael Foshey, Junchi Yan, Joshua B. Tenenbaum, Wojciech Matusik, and Antonio Torralba. Dynamic modeling of hand-object interactions via tactile sensing. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2021, Prague, Czech Republic, September 27 - Oct. 1, 2021*, pages 2874–2881. IEEE, 2021.

[45] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. *CoRR*, abs/2002.05709, 2020.

[46] Tao Chen, Jie Xu, and Pulkit Agrawal. A system for general in-hand object re-orientation. In *Conference on Robot Learning*, pages 297–307. PMLR, 2022.

[47] Haonan Duan, Peng Wang, Yayu Huang, Guangyun Xu, Wei Wei, and Xiaofei Shen. Robotics dexterous grasping: The methods based on point cloud and deep learning. *Frontiers Neurorobotics*, 15:658280, 2021.

[48] Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand dexterous manipulation from depth. *arXiv preprint arXiv:2211.11744*, 2022.

[49] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. *arXiv preprint arXiv:2204.11918*, 2022.

[50] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015.

[52] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.