Robotics: Science and Systems 2023
Daegu, Republic of Korea, July 10-July 14, 2023

Self-Supervised Visuo-Tactile Pretraining to Locate and Follow Garment Features

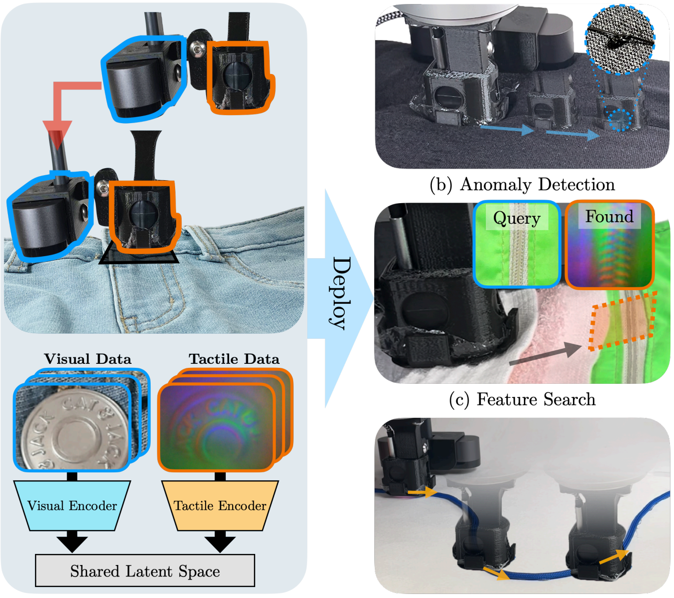
Justin Kerr*[†], Huang Huang*[†], Albert Wilcox[†], Ryan Hoque[†],
Jeffrey Ichnowski[†¶], Roberto Calandra[‡§], and Ken Goldberg[†]
*Equal contribution
[†]The AUTOLab at UC Berkeley
[‡]Learning, Adaptive Systems, and Robotics (LASR) Lab, TU Dresden
[§]The Centre for Tactile Internet with Human-in-the-Loop (CeTI)
[¶]Carnegie Mellon University

***Abstract*—Humans make extensive use of vision and touch as complementary senses, with vision providing global information about the scene and touch measuring local information during manipulation without suffering from occlusions. While prior work demonstrates the efficacy of tactile sensing for precise manipulation of deformables, they typically rely on supervised, human-labeled datasets. We propose Self-Supervised Visuo-Tactile Pretraining (SSVTP), a framework for learning multi-task visuo-tactile representations in a self-supervised manner through cross-modal supervision. We design a mechanism that enables a robot to autonomously collect precisely spatially-aligned visual and tactile image pairs, then train visual and tactile encoders to embed these pairs into a shared latent space using cross-modal contrastive loss. We apply this latent space to downstream perception and control of deformable garments on flat surfaces, and evaluate the flexibility of the learned representations *without fine-tuning* on 5 tasks: feature classification, contact localization, anomaly detection, feature search from a visual query (e.g., garment feature localization under occlusion), and edge following along cloth edges. The pretrained representations achieve a 73-100% success rate on these 5 tasks.**

I. Introduction

Despite significant recent progress in applications of tactile sensing and deep learning to deformable object manipulation [37, 40, 27, 43], many techniques still rely on human-labeled data for training these systems. If robots could leverage self-supervised data instead, scaling learned methods would be far more practical as robots could autonomously collect their own data. How can we achieve self-supervision for training flexible tactile sensing systems? Natural intelligence suggests one answer: multi-sensory fusion of vision and touch, learned through self-supervision, plays a critical role in early human cognitive development of manipulation capabilities [36, 14]. In this paper we explore how cross-modal learning from paired visual and tactile images collected in a self-supervised manner can be used to locate and follow garment features like seams, buttons, zippers, and edges.

Prior work on visuo-tactile representation learning typically relies on significant human supervision for data collection [44, 27, 39, 42] or learns task-specific representations [4, 1, 20, 25, 31]. In this work, we propose a self-supervised data collection pipeline that uses a robot with a custom end effector to autonomously collect visual and tactile images that are *spatially aligned*. This allows the learned representation to capture subtle local details in appearance and texture, making it useful for cross-modal zero-shot queries for downstream tasks. We use the paired data to train a visuo-tactile latent space with a cross-modal contrastive loss, a technique shown to be highly effective for vision and language representations like CLIP [32]. We decouple texture from rotation by training the latent space to be rotation-invariant and training a separate network to predict the relative rotation between a visual image and tactile image. We refer to this method as Self-Supervised Visuo-Tactile Pretraining (SSVTP) because it pretrains representations *once* on self-supervised data, then deploys on multiple tasks.



(b) Anomaly Detection

(c) Feature Search

(a) Visuo-Tactile Pretraining

(d) Edge Following

Fig. 1: **SSVTP Overview.** (a) We design a self-supervised framework described in Section IV-A to collect 4500 spatially aligned visual and tactile images, and use this dataset to learn a shared visuo-tactile latent space $\mathcal{Z}$ as described in Section IV-B. We apply this latent space without fine-tuning for 3 active sliding perception tasks: Anomaly Detection (b), Feature Search (c), and Edge Following (d).

We show that SSVTP can be used *without fine-tuning* for 3 active sliding perception tasks and 2 passive perception tasks on a range of garment seams, buttons, zippers, strings, and edges. An overview of the proposed approach is presented in Figure 1. This paper makes the following contributions:

1) A novel framework and hardware design for self-supervised real-world data collection of spatially aligned visuo-tactile image pairs on deformable surfaces containing garment features.
2) A novel representation learning method decoupling texture and rotation by training a rotation-invariant latent space using contrastive loss and a separate network to predict differences in rotation.
3) Experiments using the learned representations for 3 sliding perception tasks (tactile anomaly detection, vision-guided search, and active tactile servoing) and 2 passive tasks (localization and classification).
4) A publicly available dataset with 4500 samples of real-world, spatially aligned visual and tactile images of deformable surfaces (in contrast to prior large datasets without such alignment, or on rigid objects [44, 47]).

II. Related Work

A. Tactile Sensing for Robotics

Tactile sensing is a longstanding challenge in robotics [30]; in 1984 Goldberg and Bajcsy [11] explored active touch sensing using an elasto-resistive foam on a "finger." Since then, various tactile sensors with different sensing techniques, resolutions, and form factors have been studied such as force-torque sensors [5, 7], haptic devices [30, 3], and soft "skin" that mimics human skin [29]. New sensors widely used in the past few years are the DIGIT [17] and the GelSight [9]. These sensors illuminate a flexible membrane with colored LED light from different directions and observe the back of the membrane with a camera. When the membrane is pressed into surfaces, indentations are illuminated to form an image.

Tactile sensing is of renewed interest for robotic learning and manipulation. Reinforcement learning (RL) policies trained with tactile information can generalize to new object geometries for insertion tasks [10]. Lu et al. [26] develop a tactile based SLAM method to reconstruct object geometries. Object pose estimation also benefits from tactile information [2] and can be applied to manipulation tasks such as bottle opening [15]. Tactile servoing, the use of tactile sensing to control the robot along local features, is another active area of research with both learning-based and classical approaches [35, 21, 23, 22]. Tactile information generation in simulation has been studied [12] and can be used with RL for zero-shot sim-to-real transfer to manipulation tasks [6]. Sunil et al. [37], Suresh et al. [38] incorporate tactile sliding for cloth unfolding and touch localization on known object geometries, respectively. In this work, though we implement sliding tactile primitives on cloth, our method is purely self-supervised in contrast to [37, 35]. Prior work on tactile localization uses contact depth estimates trained in simulation [38, 26], which transfers well to rigid objects but cannot capture the high-frequency texture information required for perceiving fine garment features. In addition, simulating realistic tactile texture is currently prohibitively expensive for complex objects like deformable surfaces [34]. This work learns a visuo-tactile association from real data and applies it to robot tasks on garments without fine-tuning.

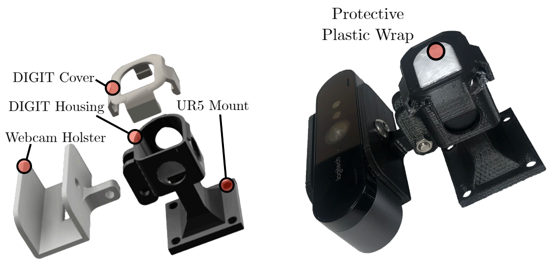
B. Visuo-Tactile Cross-Modal Learning

Learning a representation from tactile and visual inputs has been studied in the context of lifelong learning [46] and synthesizing sensory inputs from their counterparts [39, 24]. Takahashi and Tan [39] use a low-resolution 16×16 grid of pressure sensing "taxels," and train an encoder-decoder structure to predict taxel values based on visual images. Li et al. [24], Lee et al. [19] use video frames to predict outputs of a GelSight tactile sensor given a visual image of a robot touching a surface and vice versa using a conditional Generative Adversarial Network (cGAN) architecture. Recent work uses a cGAN conditioned on the output of a neural radiance field [28] to synthesize tactile images [47]. In contrast, the goal of this paper is not to synthesize tactile images from vision input but instead learn a shared representation from real vision and tactile sensor data for application to robot perception and control.

Yuan et al. [44] learn an embedding space by jointly training CNNs for three modalities: RGB images, depth images, and tactile images; they use the distance in the embedding space to match among modalities and apply their methods to infer tactile properties of fabric from vision. Similarly, Luo et al. [27] propose a fusion method to associate tactile and vision data of fabrics by applying maximum covariance analysis for the latent vectors. Zambelli et al. [45] study several approaches for cross-modal learning of vision and touch from self-supervised data. However, these papers do not explicitly apply their representations to robot control tasks. Additionally, the visual and tactile images we use in this work are spatially aligned while most visual inputs in prior work have a field of view orders of magnitude larger than the tactile images, making it difficult for networks to learn or actively perceive subtleties in local texture important for identifying garment features.

III. Problem Statement

We consider the problem of learning a cross-modal representation between vision and tactile data that can be used for localizing and tracking garment features without fine-tuning. We define *deformable surfaces* as 2.5-dimensional environments of area up to 50cm×50cm and height up to 1cm consisting of layers of textured deformable objects. These deformable layers can include *garment features* with varying textures like edges, corners, seams and embedded rigid objects like buttons, zippers, and rivets. Figure 3 provides an example of a deformable surface environment. Given a robot with a tactile sensor and RGB camera rigidly attached to each other and to a robot end effector and a set of deformable surfaces, the goal is to output a function $f : O_{\text{vis}} \times O_{\text{tac}} \to \mathcal{Z}$ where



(a) CAD Model (b) 3D Printed Hardware

Fig. 2: **Custom Hardware Design:** (a) A CAD model of the mechanical mount designed for data collection with the RGB camera and tactile sensor. (b) The actual sensors and the mount.

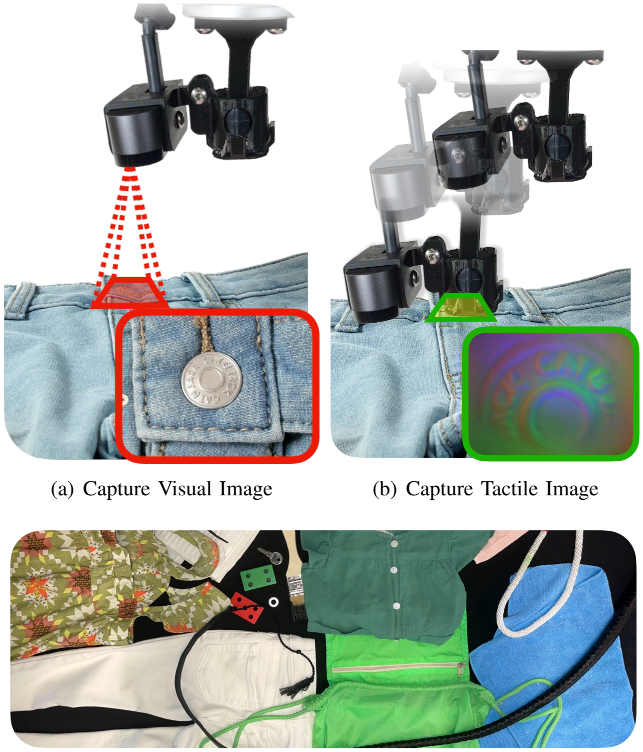
$O_{\text{vis}} = \mathbb{R}^{H\times W}$ and $O_{\text{tac}} = \mathbb{R}^{H\times W\times 3}$ are the input spaces of grayscale visual images and tactile images respectively, and $\mathcal{Z}$ is an output representation space of much lower dimension than inputs.

We make the following assumptions:

1) Objects in deformable surface environments are 1-dimensional or 2-dimensional deformable objects (e.g., fabric or smal rigid features like buttons).
2) Deformable surface environments are arranged on a flat tabletop in stable poses.
3) Surfaces in downstream tasks are within the training distribution (i.e. similar deformable objects).

During evaluation, we keep f constant across tasks and evaluate by measuring the performance of policies using f on five tasks involving locating and follow garment features:

1) *Anomaly Detection* (Active Sliding): While collecting a stream of tactile readings during a sliding motion (see Section IV-D), the task objective is to localize where a tactile anomaly occurs (e.g., a knot in an otherwise straight thread). A trial is successful if the robot stops while touching the anomaly.
2) *Feature Search* (Active Sliding): Given a visual query image, the task objective is to search the workspace with the tactile sensor to find a match. A trial is successful if the robot stops while touching the feature contained in the query image.
3) *Edge Following* (Active Sliding): Given a reference visual image of a uniform directional feature (e.g., cable or towel edge), and a line in the image frame specifying which direction to slide in, the objective is to servo the tactile sensor along the edge without losing contact. We measure the distance followed without losing continuous contact between the sensor and edge.
4) *Feature Classification* (Passive): Given a tactile query image and a set of canonical visual images (one image per class), the task objective is to determine which class the tactile image corresponds to.
5) *Contact Localization* (Passive): Given a tactile query image and a visual observation of the entire 2D workspace, the task objective is to localize in the visual image where the tactile reading was taken.



(a) Capture Visual Image (b) Capture Tactile Image

(c) Example Deformable Surface

Fig. 3: **Self-Supervised Visuo-Tactile Data Collection:** An overview of the data collection pipeline. (a) For each sample the robot first uses an RGB camera to take a closely cropped image of the texture. (b) The robot then adjusts its end effector to take a tactile reading at the same location. (c) We collect data on 10 deformable surface environments, one of which is shown here.

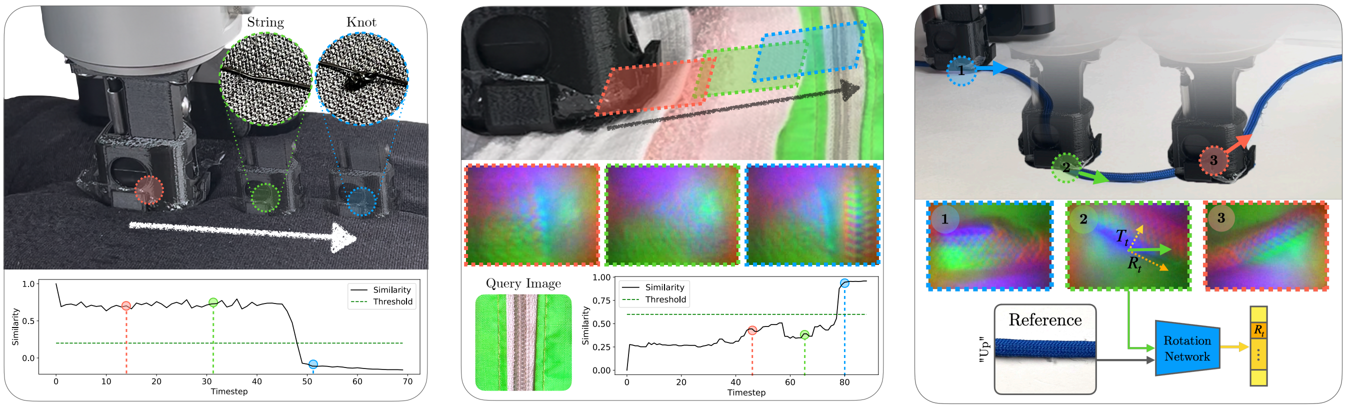
IV. Methods

SSVTP collects self-supervised data using custom hardware (Section IV-A). The collected dataset is then used to train a visuo-tactile latent space (Section IV-B) and a rotation prediction network (Section IV-C). These networks are used in active sliding perception primitives (Section IV-D) and passive perception modules (Section IV-E) for five downstream tasks.

A. Self-Supervised Data Collection

To permit cross-modal representation learning we wish to collect *spatially aligned* pairs of visual and tactile images. To accomplish this, we design a self-supervised data collection pipeline that uses a robot to automatically collect paired images on deformable surfaces defined in Section III. We design a custom end-effector (Figure 2) to hold both sensors parallel to each other with a fixed offset d between the camera and tactile sensor. Unlike many prior works that train on images with a large field of view, we zoom in closely when collecting visual images so that their field of view is similar to the tactile images, allowing the visual images to show detailed local textures. We capture visual images with a 2× wider field of view to permit rotation augmentations without border artifacts in the data.

At the beginning of each round of data collection, we normally set up the workspace by laying out 5-15 different



(a) Anomaly Detection

(b) Feature Search

(c) Edge Following

Fig. 4: **Active Sliding Perception Tasks.** (a) Anomaly Detection: The robot uses the learned tactile encoder to detect a knot while sliding over a black thread on a black surface. (b) Feature Search: Here the system is given a query visual image of a zipper and searches the workspace for a tactile reading that is sufficiently similar to the query (i.e., the texture of a zipper). (c) Edge Following: The robot uses rotation predictions to follow a curved cable. Visual inputs are colorized for clarity, the visual network takes in grayscale.

objects at stable poses on a flat surface to form a deformable surface. The robot then samples planar translations on a predefined grid with a resolution of 2cm and a fixed height offset above the surface. For each sample, the system generates random noise uniformly between -1cm and 1cm in translation and between 0° and 45° in rotation to mitigate aliasing induced by grid sampling. The robot moves the camera above the noised position, takes a visual image using the camera, then translates the constant d offset to align the tactile sensor above the camera observation center. The robot moves down in a stamping motion, stopping when the arm registers a force of 20N. This allows good contact with the sensor without harming the tactile membrane. Thus, the collected tactile and vision images are automatically spatially aligned with a similar field of view. The pipeline takes 11 samples per minute, with minimal human involvement required to periodically reset the surface with new objects. We distribute LED lights close to the workspace to limit self-shadows from the arm. This pipeline is shown in Figure 3.

B. Latent Space Training

We train a vision encoder $f_\theta : O_{\text{vis}} \rightarrow \mathcal{Z}$ and a tactile encoder $g_\phi : O_{\text{tac}} \rightarrow \mathcal{Z}$, where $\mathcal{Z} = \{z \in \mathbb{R}^d : ||z|| = 1\}$ is a shared output representation space with $d \ll H \times W \times 3$. We use the collected data (Section IV-A) and contrastive learning to learn a visuo-tactile association in a latent space. We choose a contrastive loss because of its prior success in embedding different modalities within the same space [32, 42]. We use grayscale camera images as input to the visual encoder to avoid overfitting to colors of materials, and we augment all images with flips and color jitter to make the network less sensitive to input image orientation and lighting. For each visuo-tactile image pair we also apply random rotation uniformly between 0° and 360° to encourage rotation-invariant learning. During training we use InfoNCE loss on the pairwise dot products of the embedding vectors, which intuitively pushes the correct pair of visuo-tactile readings together and all others apart. We constrain embedding vectors to have unit length, as in CLIP [32]. Mathematically, this loss is $L = \text{cross\_entropy}(I_e \times T_e^{\text{T}}, [0...B])$, where T_e and I_e are batched embeddings of the tactile and visual images respectively and B is the batch size. The outputs of the trained encoders are d-dimensional embeddings in the learned latent space. See Appendix VIII-A for more details on parameters and augmentations.

C. Rotation Prediction Network

We train a separate rotation network $h_\psi : O_{\text{vis}} \times O_{\text{tac}} \rightarrow [0, 1]^{N_b}$ on the same dataset to learn the difference in rotation between a vision image and a tactile image. We use color augmentation to prevent potential spurious lighting statistics in the data [8]. We discretize the rotations into N_b buckets uniformly between 0° and 360° and formulate the problem as classification where the input is a concatenated visuo-tactile image pair and the output is a distribution over bucket indices. Classification aids in learning multi-modality in rotations (±180°). See Appendix VIII-B1 for further details.

D. Sliding Perception Primitives

We implement three active sliding perception primitives for use in downstream tasks on garment features (Figure 4) using the trained encoders and rotation network without fine-tuning. The first two focus on *locating* features of interest, and the third on *following* these physically with the sensor.

1) Anomaly Detection: To inspect garments, humans often slide their hands across the surface to detect textures which stand out from the background. We mimic this behavior by sliding the tactile sensor across a deformable surface and using the trained tactile encoder to localize tactile anomalies in materials. The system takes tactile readings while sliding and encodes them with g_ϕ. The system maintains a sliding window $\mathcal{B}_h$ containing past tactile embeddings and continually evaluates the median cosine similarity between the current embedding and the embeddings in the window. If the median

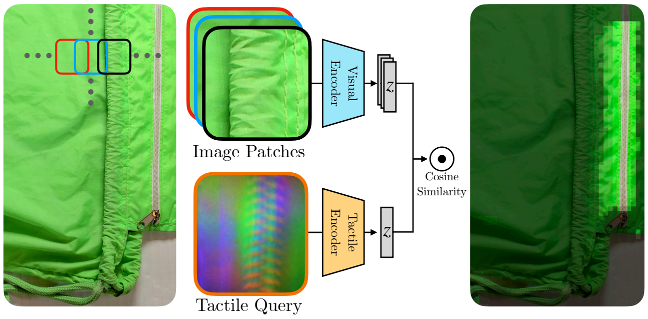


Fig. 5: **Tactile Localization:** An input tactile image of the zipper is compared to discreteized patches from a visual image of the entire scene. In this example, the heatmap shows high probability of a match near the zipper. Note that visual images are colorized for clarity, the visual encoder takes in grayscale.

similarity crosses a predefined threshold m_1, the system outputs that it has detected an anomaly candidate and slows the robot down to take n_1 more readings. If the median similarity remains below m_1, the system outputs an anomaly; otherwise, the system resumes sliding at the original speed.

2) Feature Search: When searching for a specific tactile feature in a garment (e.g., a button in a shirt), anomaly detection is insufficient due to the existence of distractors (e.g. seams). To remedy this, we search for a tactile reading which matches a *visual* query image of the target. This visual image is embedded into the shared latent space. While sliding, the system computes the cosine similarity (dot product) between the current tactile embedding and the query image embedding. If the similarity is above m_2, the policy slows down the robot to takes n_2 more tactile readings. If the median similarity remains greater than m_2, the system outputs that it has found the target; otherwise, it resumes sliding at the original speed.

3) Edge Following: Another useful primitive for tactile manipulation is actively servoing along garment features like edges, cables, and seams. Prior approaches utilize analytic methods [35] or supervised learning with human labels [37, 21]. In this work, the robot servos along material features specified purely from visual images, leveraging the visuo-tactile pretraining to avoid the need for a servoing dataset. Given an input visual image specifying the feature to follow as well as the direction to follow it in, the system uses the rotation network to predict the relative rotation between the query visual image and the current tactile image. This rotation provides a translation in the tactile frame, R_t, to move in which follows the feature. Sometimes the garment feature may drift orthogonally to R_t, so we also estimate a feature-centering offset. To do this, we approximate network confidence as the variance of the predicted rotation distribution. We translate the input image orthogonally to R_t in both directions by half the image width, and choose the offset as the translation which maximizes confidence. Maximizing confidence serves as a centering signal because higher confidence corresponds to views with the feature more visible.

E. Passive Perception Modules

We develop two modules that use the pretrained, frozen encoders for passive tactile perception: a localizer module and a classifier module.

1) Contact Localizer: Given a top-down visual image of the entire workspace (a sort of "visual map") and a tactile image, this module outputs a probability distribution over possible contact locations in the visual image. We divide the input visual image into a grid with uniform patch size and resolution and compute cosine similarity of each patch embedding with the tactile embedding as shown in Figure 5. This produces a distribution over patch locations of the predicted similarity between vision and touch.

2) Feature Classifier: Classification of features in input tactile images (e.g., distinguishing between towel edges and corners) is a common task. Prior work performs tactile classification with supervised learning [37], but human supervision is time-consuming. Instead, we construct a classifier from trained encoders by providing canonical visual images of the classes we consider, which can be rapidly captured. We augment these visual images with the same augmentation as performed during training and embed them into $\mathcal{Z}$. We use the weighted *k*-nearest neighbors algorithm to classify the tactile input (Figure 6), a technique shown to be effective in classification with latent spaces [41] in computer vision.

V. Physical Experiments

We assess the utility of SSVTP through performance on a diverse range of tasks centered around garment feature localization and following. Throughout all experiments, the same pretrained networks are used without fine-tuning, to evaluate the flexibility of the representations learned from visuo-tactile pretraining.

We collect self-supervised data (Section IV-A) with a UR5 robot arm, a DIGIT tactile sensor [17], and a Logitech BRIO webcam. This same setup is used for physical experiments. The data consist of 4500 visuo-tactile pairs over 10 deformable surface environments. It took 7 hours of unsupervised robot time to collect all data samples, which involved about 20 minutes of human time to reset the environment with new objects between sampling.

Both tactile and visual networks use an input resolution of 128 × 128, to which all input images are resized to. We train a texture latent space (Section IV-B) with an embedding dimension of $d = 8$. Visual and tactile encoders use a RegNetY_800MF backbone [13] with an altered final layer because of its performance per MFLOP. The rotation network uses a RegNetY_800MF backbone with an altered first layer to support 2 stacked images, and a final layer to output the rotation bins. We initialize all networks with weights pretrained on ImageNet [33] classification and train all parameters. See the Appendix for more training and parameter details.

A. Active Sliding Tasks

For these tasks, the experiment starts with the robot pushing down at the given start position until a force of 20N is reached,

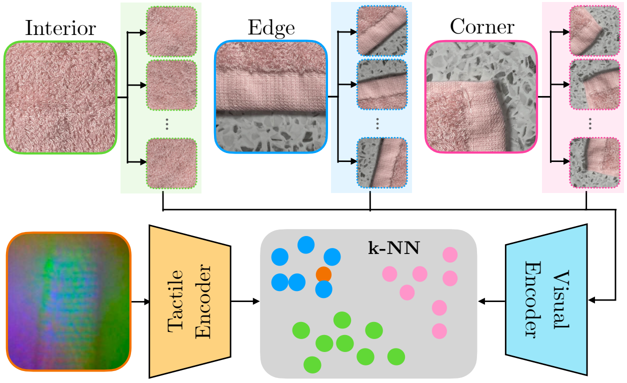


Fig. 6: **Tactile Classification:** Given canonical visual images of categories (images in solid line), we augment these inputs (images in dashed line) and embed them into the latent space, then apply k-NN classification on tactile input images. Images are colorized for clarity.

TABLE I: Anomaly detection success rate.

Algorithm	Knot 3mm	Knot 1mm	Lump
Image diff	**9/10**	0/10	2/10
SSVTP	**9/10**	**8/10**	**5/10**

TABLE II: Feature search success rate.

Algorithm	Zipper	Garment	Button
Anomaly Det.	0/10	1/10	5/10
SSVTP	**10/10**	**10/10**	**10/10**

TABLE III: Edge following distance % completed.

Algorithm	S-Shaped Cable	Dress Seam	Shirt Seam
Ellipse Fit	12.7 ± 24.0	43.1 ± 35.4	18.8 ± 9.6
SSVTP	$\mathbf{78.7 \pm 3.4}$	$\mathbf{92.2 \pm 23.3}$	$\mathbf{92.4 \pm 19.1}$

TABLE IV: Classification results on towel edges, corners, and interiors. We use K-nearest neighbor classification on the latent embeddings of a small set of canonical *visual* images to classify new *tactile* images (Sec. IV-E2). N refers to the number of datapoints, and "primary confusor" is the mode of mis-classifications.

Category	**Accuracy (%)**	**Primary Confusor**	**N**
Edge	75	Corner	36
Corner	93	Edge	30
Interior	74	Edge	43
Overall	80	–	109

to establish good tactile contact. The robot slides at a speed of 1cm/s acquiring tactile readings at a rate of 10hz. We do not use closed-loop control on the force applied, since using the UR5's position control is sufficient to maintain consistent contact on the flat surfaces we test on. All the experiments below are conducted without vision input other than the query or reference visual images. The tactile sensor mount includes a thin layer of protective plastic wrap over the sensor which aids sliding primitives by mitigating friction.

1) Anomaly Detection: For this experiment we set $|\mathcal{B}_h| = 40$, $m_1 = 0.3$, and $n_1 = 20$. The robot is slowed to 20% speed when an anomaly candidate is detected. A trial is considered successful if the robot stops at a tactile reading that contains the anomaly. Each experiment consists of 10 trials with varying start locations. Figure 4(a) shows an example. As a baseline, we apply the same buffer and active confirmation technique but using a threshold on the L2 pixel distance between tactile images over time which detects sharp changes in the appearance of the tactile image. We consider the following 2 settings: **Knot detection in string**: We fix a 1 black surgical thread on top of black cloth and tie a knot anomaly to detect. We conduct experiments on a 1 knot and a 3 knot respectively. **Lump defect detection**: To simulate a manufacturing defect in a piece of soft foam, we embed a 3mm hex nut in foam, 1cm beneath the surface. Results (Table I) show that SSVTP achieves an average 73% success rate on the three settings, while L2 distance achieves an average 37% success rate. The baseline's performance is appreciable for prominent features like a large knot, but it is insufficiently sensitive for more subtle features like the smaller knot. Both techniques achieve lower success rates on lump detection due to the more subtle changes in the tactile readings, though SSVTP is more robust.

2) Feature Search: In these experiments, to test if SSVTP can differentiate targets from multiple distractors, we include false tactile anomalies along the path of the robot to the correct feature. We use $m_2 = 0.6$ and $n_2 = 60$. The robot is slowed to 20% speed when a target candidate is detected. Figure 4(b) shows an example. We use Anomaly Detection as a baseline to test if the distractions are indeed distractors, and run 10 trials for each experiment varying the start location. An experiment is a success if the robot stops with the feature specified by the visual image in the tactile image. We consider the following three settings: **Multi-texture garment**: The robot searches for a ruffled portion of a dress amidst seams and crumples, a subtle difference in tactile texture to detect. **Finding a zipper among multiple edges**: The robot searches for a zipper on a multi-textured surface. We layer two different towels on a bag with a zipper, and the query image is a visual image of the zipper (Figure 4(b)). **Finding a button in a blouse**: The robot slides on a blouse until a button is found, ignoring seams along the way.

The results, summarized in Table II, show that SSVTP achieves a 100% success rate while the baseline achieves 20% average success rate, suggesting that the learned latent space can successfully correlate tactile and vision information, while anomaly detection easily mis-fires on distractor features.

3) Edge Following: For each experiment, we provide a visual image of a target feature and specify the direction to follow in that image. We smooth the direction vector R_t obtained from the rotation network by $\hat{R}_t = \lambda R_t + (1-\lambda)R_{t-1}$, where $\lambda = 0.8$ is chosen empirically. This results in the overall sliding direction of $\hat{R}_t + T_t$ (see Sec IV-D3 for details). We measure the distance successfully traveled without losing contact. As a baseline, we use the contact ellipse estimation from PyTouch [18] and move along the major axis. We average over 10 trials for each test object. We conduct the experiments on a 2mm diameter, 30cm length **S-shaped cable** fixed on the

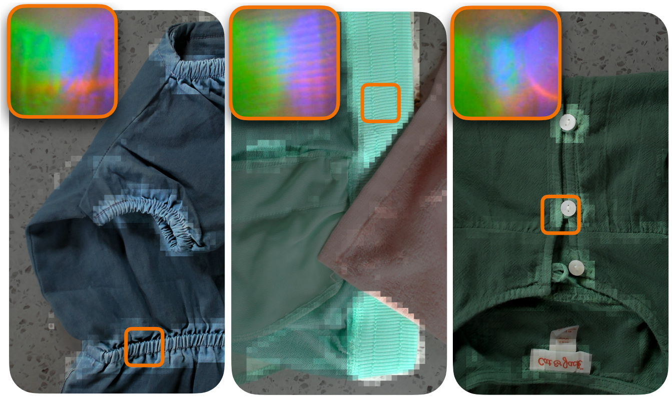


Fig. 7: **Localization Visualizations:** We show 3 illustrative contact localization results; a wrinkled dress seam, an elastic band, and blouse buttons. The heatmaps are composited onto the original image by scaling them between $[0, 1]$, applying gamma correction $\gamma = 1.6$, and multiplying them by the value channel. Orange boxes denote the location of the tactile query.

work surface, a **dress seam** in the middle of a dress with wrinkles on the side and a **shirt seam** near a shirt collar.

The percent distance traveled is in Table III. While the baseline travels up to 43.1% of the total distance, our policy travels at least 78.7% of the total distance. The baseline is sensitive to flat textures that confuse the ellipse fitting, while our rotation network prediction is robust to local textures. The major failure cases of our method are ambiguity in servoing direction (±180°) and errors from the centering offset T_t if the feature drifts too much. This effect is especially visible in the S-shaped cable, where the direction changes drastically during a rollout, whereas the seams only have a gently curve.

B. Passive Perception

1) Contact Localization: Since many locations in the visual image may generate nearly identical tactile readings, such as the interior of a fabric, we provide qualitative visualizations of localization heatmaps in Figure 7. Similar textures as the ground truth touch location are smoothly activated, suggesting the encoder has learned a mapping between visual and tactile appearance rather than overfitting to specific locations.

2) Feature Classification: We manually collect a dataset of 109 tactile images of corners, edges, and interiors of 2 different towels, and use 4 canonical visual images per category to instantiate the K-NN classifier. We use 100 augmentations of the 4 canonical images to smooth decision boundaries and k=50. Results are shown in Table IV. We hypothesize that the edges are occasionally confused as corners because they look locally similar, while the interiors are misclassified as edges due to the similarity in texture between the interior of one towel to the edge texture of the other.

VI. Limitations

This approach has notable limitations. First, the similarity scores obtained via the latent space can be difficult to interpret, making performance on query tasks sensitive to a chosen threshold. Techniques to avoid thresholding could incorporate probabilistic estimation methods like particle filters to account for uncertainty, analogous to [38]. In addition, the deformable surfaces were largely planar. Some human time is still required to reset the workspace between rounds of data collection; fully automating the data collection procedure could improve scalability. In this work the test garments were contained within the self-supervised dataset, so the generalization performance remains to be studied at scale. Another consideration is the size of the cropped visual query images: during usage the input visual images must be within the training distribution or results can become unreliable. In this work the scale was manually chosen, but in the future the system could select this scale based on depth readings from an RGBD camera, or automatically based on the similarity score.

VII. Conclusion

In this paper, we propose Self-Supervised Visuo-Tactile Pretraining: a self-supervised framework for learning visuo-tactile representations. We apply the learned representations to 3 active and 2 passive perception tasks on garment features. Results across this diverse range of tasks suggest the flexibility of the learned representations for diverse tasks including contact localization, tactile feature search, feature classification, anomaly detection, and edge following. While we focus our attention on garment features in this work, SSVTP is not limited to this application and poses a step toward task-agnostic visuo-tactile representation learning for robot control.

Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 2146752. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. This research was performed at the AUTOLAB at UC Berkeley in affiliation with the Berkeley AI Research (BAIR) Lab, and the CITRIS "People and Robots" (CPAR) Initiative. Funded by the German Research Foundation (DFG, Deutsche Forschungsgemeinschaft) as part of Germany's Excellence Strategy – EXC 2050/1 – Project ID 390696704 – Cluster of Excellence "Centre for Tactile Internet with Human-in-the-Loop" (CeTI) of Technische Universität Dresden.

References

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

[2] Maria Bauza, Antonia Bronars, and Alberto Rodriguez. Tac2pose: Tactile object pose estimation from the first touch. *arXiv preprint arXiv:2204.11701*, 2022.

[3] Miguel Altamirano Cabrera, Oleg Sautenkov, Jonathan Tirado, Aleksey Fedoseev, Pavel Kopanev, Hiroyuki Kajimoto, and Dzmitry Tsetserukou. Deepxpalm: Tilt and

position rendering using palm-worn haptic display and cnn-based tactile pattern recognition. In *2022 IEEE Haptics Symposium (HAPTICS)*, pages 1–6, 2022. doi: 10.1109/HAPTICS52432.2022.9765571.

[4] Roberto Calandra, Andrew Owens, Dinesh Jayaraman, Justin Lin, Wenzhen Yuan, Jitendra Malik, Edward H. Adelson, and Sergey Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *IEEE Robotics and Automation Letters*, 3(4):3300–3307, oct 2018. doi: 10.1109/lra.2018.2852779. URL https://doi.org/10.1109%2Flra.2018.2852779.

[5] Max Yiye Cao, Stephen Laws, and Ferdinando Rodriguez y Baena. Six-axis force/torque sensors for robotics applications: A review. *IEEE Sensors Journal*, 21(24):27238–27251, 2021. doi: 10.1109/JSEN.2021.3123638.

[6] Alex Church, John Lloyd, Nathan F Lepora, et al. Tactile sim-to-real policy transfer via real-to-sim image translation. In *Conference on Robot Learning*, pages 1645–1654. PMLR, 2022.

[7] Mark R. Cutkosky, Robert D. Howe, and William R. Provancher. *Force and Tactile Sensors*, pages 455–476. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008. ISBN 978-3-540-30301-5. doi: 10.1007/978-3-540-30301-5_20. URL https://doi.org/10.1007/978-3-540-30301-5_20.

[8] Carl Doersch, Abhinav Kumar Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1422–1430, 2015.

[9] Siyuan Dong, Wenzhen Yuan, and Edward H. Adelson. Improved GelSight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, sep 2017. doi: 10.1109/iros.2017.8202149. URL https://doi.org/10.1109%2Firos.2017.8202149.

[10] 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.

[11] Ken Goldberg and Ruzena Bajcsy. Active touch and robot perception. *Cognition and Brain Theory*, 7(2):199–214, 1984.

[12] Daniel Fernandes Gomes, Paolo Paoletti, and Shan Luo. Generation of gelsight tactile images for sim2real learning. *IEEE Robotics and Automation Letters*, 6(2):4177–4184, 2021.

[13] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[14] R. Held, Y. Ostrovsky, B. de Gelder, T. Gandhi, S. Ganesh, U. Mathur, and P. Sinha. The newly sighted fail to match seen with felt. *Nature Neuroscience*, 14: 551–553, 2011.

[15] Tarik Kelestemur, Robert Platt, and Taskin Padir. Tactile pose estimation and policy learning for unknown object manipulation. *arXiv preprint arXiv:2203.10685*, 2022.

[16] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference for Learning Representations*, abs/1412.6980, 2015.

[17] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, Dinesh Jayaraman, and Roberto Calandra. 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, jul 2020. doi: 10.1109/lra.2020.2977257. URL https://doi.org/10.1109%2Flra.2020.2977257.

[18] Mike Lambeta, Huazhe Xu, Jingwei Xu, Po-Wei Chou, Shaoxiong Wang, Trevor Darrell, and Roberto Calandra. Pytouch: A machine learning library for touch processing. *CoRR*, abs/2105.12791, 2021. URL https://arxiv.org/abs/2105.12791.

[19] Jet-Tsyn Lee, Danushka Bollegala, and Shan Luo. "touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 4276–4282. IEEE, 2019.

[20] Michelle A. Lee, Yuke Zhu, Krishna Parasuram Srinivasan, Parth Shah, Silvio Savarese, Li Fei-Fei, Animesh Garg, and Jeannette Bohg. Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks. *2019 International Conference on Robotics and Automation (ICRA)*, pages 8943–8950, 2019.

[21] Nathan F. Lepora and John Lloyd. Pose-based tactile servoing: Controlled soft touch using deep learning. *IEEE Robotics Automation Magazine*, 28(4):43–55, 2021. doi: 10.1109/MRA.2021.3096141.

[22] Nathan F. Lepora, Kirsty Aquilina, and Luke Cramphorn. Exploratory tactile servoing with active touch. *IEEE Robotics and Automation Letters*, 2(2):1156–1163, 2017. doi: 10.1109/LRA.2017.2662071.

[23] Qiang Li, Carsten Schürmann, Robert Haschke, and Helge J. Ritter. A control framework for tactile servoing. In *Robotics: Science and Systems*, 2013.

[24] Yunzhu Li, Jun-Yan Zhu, Russ Tedrake, and Antonio Torralba. Connecting touch and vision via cross-modal prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10609–10618, 2019.

[25] Justin Lin, Roberto Calandra, and Sergey Levine. Learning to identify object instances by touch: Tactile recognition via multimodal matching. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 3644–3650. IEEE, 2019.

[26] Yujie Lu, Jianren Wang, and Vikash Kumar. Curiosity driven self-supervised tactile exploration of unknown

objects. *arXiv preprint arXiv:2204.00035*, 2022.

[27] Shan Luo, Wenzhen Yuan, Edward Adelson, Anthony G Cohn, and Raul Fuentes. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2722–2727. IEEE, 2018.

[28] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[29] Mostafa A. Mousa, MennaAllah Soliman, Mahmood A. Saleh, and Ahmed G. Radwan. Tactile sensing biohybrid soft e-skin based on bioimpedance using aloe vera pulp tissues. *Scientific Reports*, 11(3054), 2021.

[30] Claudio Pacchierotti, Stephen Sinclair, Massimiliano Solazzi, Antonio Frisoli, Vincent Hayward, and Domenico Prattichizzo. Wearable haptic systems for the fingertip and the hand: Taxonomy, review, and perspectives. *IEEE Transactions on Haptics*, 10(4):580–600, 2017. doi: 10.1109/TOH.2017.2689006.

[31] Leszek Pecyna, Siyuan Dong, and Shan Luo. Visual-tactile multimodality for following deformable linear objects using reinforcement learning. *ArXiv*, abs/2204.00117, 2022.

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[34] Daniel Seita, Aditya Ganapathi, Ryan Hoque, Minho Hwang, Edward Cen, Ajay Kumar Tanwani, Ashwin Balakrishna, Brijen Thananjeyan, Jeffrey Ichnowski, Nawid Jamali, Katsu Yamane, Soshi Iba, John F. Canny, and Ken Goldberg. Deep imitation learning of sequential fabric smoothing from an algorithmic supervisor. *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9651–9658, 2020.

[35] Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward H. Adelson. Cable manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40:1385 – 1401, 2020.

[36] Linda Smith and Michael Gasser. The development of embodied cognition: six lessons from babies. *Artificial Life*, 11:13–29, 2005.

[37] Neha Sunil, Shaoxiong Wang, Yu She, Edward Adelson, and Alberto Rodriguez Garcia. Visuotactile affordances for cloth manipulation with local control. In *Conference on Robot Learning (CoRL)*, 2022.

[38] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. Midastouch: Monte-carlo inference over distributions across sliding touch. In *Conference on Robot Learning (CoRL)*, 2022.

[39] Kuniyuki Takahashi and Jethro Tan. Deep visuo-tactile learning: Estimation of tactile properties from images. In *2019 International Conference on Robotics and Automation (ICRA)*, pages 8951–8957. IEEE, 2019.

[40] Sashank Tirumala, Thomas Weng, Daniel Seita, Oliver Kroemer, Zeynep Temel, and David Held. Learning to singulate layers of cloth using tactile feedback, 2022. URL https://arxiv.org/abs/2207.11196.

[41] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance-level discrimination. *CoRR*, abs/1805.01978, 2018. URL http://arxiv.org/abs/1805.01978.

[42] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. *arXiv preprint arXiv:2211.12498*, 2022.

[43] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward Adelson. Active clothing material perception using tactile sensing and deep learning, 2017. URL https://arxiv.org/abs/1711.00574.

[44] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2017.

[45] Martina Zambelli, Yusuf Aytar, Francesco Visin, Yuxiang Zhou, and Raia Hadsell. Learning rich touch representations through cross-modal self-supervision. In *Conference on Robot Learning (CoRL)*, 2020.

[46] Wendong Zheng, Huaping Liu, and Fuchun Sun. Lifelong visual-tactile cross-modal learning for robotic material perception. *IEEE Transactions on Neural Networks and Learning Systems*, 32(3):1192–1203, 2021. doi: 10.1109/TNNLS.2020.2980892.

[47] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning (CoRL)*, 2022.

VIII. Appendix

A. Encoder training details

We choose the encoder dimension $d = 8$ via grid search over $d \in [8, 16, 32, 64]$, and selected based on qualitative performance on a held-out test set of contact localization and classification. Though it may seem surprising that a lower dimension performs better, this is because given a relatively small dataset (4500 pairs), a lower dimension latent space has a regularizing effect preventing overfitting. We optimize using

Adam [16] with default parameters, and exponential learning rate decay with $\gamma = 0.95$, gradient clipping at 0.5, and a batch size of B $= 512$. This batch size was the largest that could fit in our GPU memory, following literature from the vision-language supporting large batch sizes as key for latent spaces[32].

1) Augmentation: For training the latent space, we apply color augmentations to both tactile and visual images. Though visual images are grayscale, some of these still affect the appearance like contrast and brightness. For tactile images, these we use PyTorch's ColorJitter with parameters `brightness=(0.9,1.1)`, `contrast=(.9,1.1)`, `saturation=0.2`, `hue=0.05`. For visual images, these are `brightness=(0.8,1.1)`, `contrast=(.7,1.3)`, `saturation=0.2` We also apply spatial augmentations *independently* to tactile and visual images. These are rotation augmentation (±360°), translation (2% of image scale), and horizontal flips. These augmentations are applied randomly during dataloading; as such there is no fixed dataset size used, just procedurally generated augmentations of the original 4500 pairs.

B. Rotation network training

We train the rotation network (Section IV-C) with a batch size of B = 32, $N_b = 18$, and Cross Entropy Loss.

1) Augmentation: During rotation network training we use the same color augmentations as in the latent space encoder, and apply rotation augmentations to the visuo-tactile pairs *non-independently*, so that the rotation correlation between pairs is preserved. Then, we randomly select a *relative* rotation to transform the tactile image, thus obtaining a training datapoint with a labeled, random rotation. This rotation is converted to the bin index for the label.