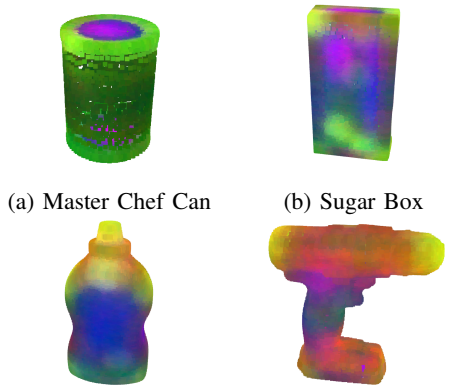
2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

HyperTaxel: Hyper-Resolution for Taxel-Based Tactile Signals Through Contrastive Learning

Hongyu Li[†,‡], Snehal Dikhale[†], Jinda Cui[†], Soshi Iba[†], and Nawid Jamali[†]

***Abstract*— To achieve dexterity comparable to that of humans, robots must intelligently process tactile sensor data. Taxel-based tactile signals often have low spatial-resolution, with non-standardized representations. In this paper, we propose a novel framework, HyperTaxel, for learning a geometrically-informed representation of taxel-based tactile signals to address challenges associated with their spatial resolution. We use this representation and a contrastive learning objective to encode and map sparse low-resolution taxel signals to high-resolution contact surfaces. To address the uncertainty inherent in these signals, we leverage joint probability distributions across multiple simultaneous contacts to improve taxel hyper-resolution. We evaluate our representation by comparing it with two baselines and present results that suggest our representation outperforms the baselines. Furthermore, we present qualitative results that demonstrate the learned representation captures the geometric features of the contact surface, such as flatness, curvature, and edges, and generalizes across different objects and sensor configurations. Moreover, we present results that suggest our representation improves the performance of various downstream tasks, such as surface classification, 6D in-hand pose estimation, and sim-to-real transfer.**



(a) Master Chef Can (b) Sugar Box

(c) Mustard Bottle (d) Power Drill

Fig. 1: A visualization of tactile embeddings for different YCB [19] objects. The embeddings capture the geometric features of the contact surface, such as flatness, curvature, and edges, and are consistent across different objects.

I. Introduction

Tactile sensing is a critical modality for humans to interact with everyday objects [1]. Tactile sensors can be divided into two broad categories [2]: vision-based [3, 4], and taxel-based [5–7]. Recently, vision-based tactile sensors have gained popularity, partly due to their pixel-based representation, which makes them amenable to deep learning approaches [8–10]. However, their size limits full coverage on multi-fingered hands [11, 12]. In contrast, taxel-based sensors remain underexplored because they present many challenges to deep learning approaches, including low spatial resolution and a lack of consensus on how to represent and process taxel-based sensors. However, they continue to remain of interest to the robotic manipulation community due to their unique ability to directly respond to the underlying phenomena measured, thereby offering valuable opportunities for enhancing robotic manipulation [13, 14].

The encoding and processing of taxel signals is still an open research question. Taxel-based sensors present unique challenges before they can be used in downstream tasks. These challenges include 1) the development of effective representations for tactile sensor data, and 2) their inherently low resolution [2, 12, 15], which has been a long-standing barrier hindering tactile dexterous manipulation [15] and perception, such as in-hand 6D pose estimation [16–18]. As suggested by Dahiya et al. [15], one promising line of research is the use of super-resolution algorithms.

In this paper, we present a two-stage solution to address the aforementioned challenges. In the first stage, we propose a method for learning a representation of the tactile signals in an embedding space using contrastive learning. This approach generalizes across various taxel layouts, different objects, and multiple tasks. Our key intuition is that by exploiting the correspondence between the taxel signals and their contact surface, we can learn a geometrically-informative representation. To this end, we propose graphs to represent the tactile signals, with a novel graph construction strategy and convolution kernel for tactile processing.

In the second stage, we map low-resolution taxel signals to a high-resolution three-dimensional surface using a multi-contact strategy to reduce uncertainty in taxel signals, a process we term hyper-resolution. Unlike super-resolution, which focuses on upscaling data within the same domain, hyper-resolution extends beyond mere upscaling within the same domain across different domains and modalities. For example, in image processing, super-resolution produces the same image with a higher resolution. However, hyper-resolution maps low-resolution taxel signals to capture various object properties, such as the three-dimensional surface of the object, surface texture, etc. This distinction allows hyper-resolution to provide informative data beneficial for tasks such as 6D pose estimation.

[†] Honda Research Institute USA. {snehalsubhash_dikhale, jinda_cui, siba, njamali}@honda-ri.com.

[‡] Brown University. This work was done during his internship at Honda Research Institute USA. hongyu@brown.edu.

Our contributions can be summarized as follows: 1) We propose a novel taxel encoder. 2) We introduce a novel representation learning approach for signals from taxel-based tactile sensors. 3) We propose a novel hyper-resolution algorithm for taxel-based tactile sensors, which leverages the proposed taxel encoder.

We present results of qualitative analysis that suggest our tactile representation captures geometric features of the contact surface, such as flatness, curvature, and edges, and generalizes across different objects and sensor configurations. Furthermore, we benchmark our proposed framework against two seminal taxel encoders and present quantitative results that demonstrate the effectiveness of our approach. We also perform a comparative analysis across different representations and confirm the effectiveness of the graph representation. We also assess the quality of our hyper-resolution using 6D object pose estimation and show that using our hyper-resolved data improves performance. In the end, we verify the sim-to-real transferability of our learned tactile representation on the surface classification task on the real robot.

II. Related Work

Representation learning is the process of encoding informative features from raw data to make it suitable for machine learning tasks. Most of the earlier works use supervised learning methods [20]. Recently, there has been a growing interest in self-supervised learning [21–23] and multi-modal learning [24]. While most of the prior studies focus on domains such as vision [21–23] and language [24–26], in this paper, we are interested in whether the same paradigm can be applied in the tactile domain.

Several prior works have investigated tactile representation learning. In the image-based sensor domain, Villalonga et al. [27] leverage the contrastive framework MoCo [22], and Caddeo et al. [28] utilize an autoencoder to learn a representation. However, their transfer from image-based to taxel-based data is non-trivial. Guzey et al. [14] learn a representation for taxel-based sensors using BYOL [21]. However, their primary focus is on dexterous manipulation, and they do not explore various downstream tasks or representation learning approaches. Therefore, the most effective paradigm for taxel-based signals remains a topic for further exploration.

Contact localization has been utilized to achieve tactile super-resolution. In contact localization, the goal is to estimate the contact location from a given tactile observation. Early works use probabilistic approaches to estimate the probability distribution of the contact location [29–32]. Piacenza et al. [12] adopt a data-driven approach for contact localization in 3D space. With the advancements in computer vision, recent works have increasingly explored vision-based tactile sensors [27, 28, 30, 33–35], which lend themselves well to deep learning techniques due to their pixel-based output. However, the fusion of taxel-based sensors with deep learning methodologies remains relatively underexplored.

In the seminal work Lepora et al. [36] propose taxel-based tactile super-resolution using the Bayesian perception method. Recently, there has been a shift from probabilistic-based approaches to learning-based ones. Wu et al. [2] propose a method wherein the taxel-based tactile signal is interpreted as a 2D image, subsequently enhanced using SRGAN [37]. This process results in a higher-resolution representation of the contact surface between the sensor and the object. However, interpreting tactile signals as a 2D image limits their application to 2D arrangements; the 3D arrangement found in curved fingers cannot be accurately represented. Moreover, they do not utilize the geometric information of the object in contact. In this paper, we present a geometrically-informed hyper-resolution algorithm invariant to sensor arrangements.

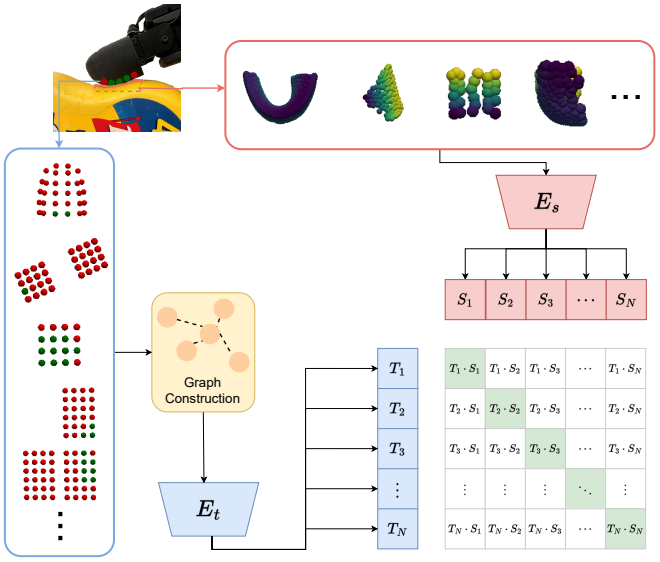


Fig. 2: Overview of our proposed tactile representation learning framework. The tactile signal, left blue box, is represented as a graph and encoded using tactile encoder E_t, and the corresponding contact surface patch, top red box, is encoded using surface encoder E_s.

III. Methodology

Given a sparse taxel-based tactile signal, our goal is to obtain a high-resolution depiction of the contact surface between the sensor and the object of interest. To achieve this, we propose a two-stage solution. The first stage, representation learning, involves using a graph neural network and contrastive learning to learn a geometrically-informed representation of the tactile signals. The second stage, hyper-resolution, uses the learned representation to map low-resolution taxel signals into a high-resolution contact surface using multi-contact localization.

A. Representation Learning

Figure 2 shows an overview of the proposed tactile representation learning framework. In this section, we detail the different components of our framework.

1) Taxel Representaion: Tactile data can be represented as point clouds, images, or graphs. The point cloud representation, however, fails to encode the absence of contact, and the image representation struggles to capture the 3D spatial arrangement of tactile sensors. In our research, we opted for the graph representation because it encodes both the spatial arrangement of the taxel signals and the absence of contact.

We use an undirected spatial graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ to represent the taxel data, where $\mathcal{V}$ and $\mathcal{E}$ are the vertices and edges of the graph, respectively. Each vertex corresponds to a taxel of the tactile sensor, and each edge represents the spatial proximity between two taxels. The vertices have two types of features: the 3D coordinates of each taxel, $\mathcal{X} \in \mathbb{R}^{t\times 3}$, and their corresponding signals, $K \in \mathbb{R}^t$, where t is the number of taxels in the tactile sensors. For taxel signals with three axes, K is considered the Euclidean norm of the signals from all three axes. We combine X and K into a matrix $\mathcal{V} = [\mathcal{X}|K]$ of dimension $\mathbb{R}^{t\times 4}$. To improve sim-to-real transfer, we simplify the taxel signal into a boolean activation state [38, 39].

Edges $\mathcal{E}$ connect the taxel vertices $\mathcal{V}$ in the taxel graph and construct the graph. Since we are considering tactile sensors with arbitrary spatial arrangements, we propose that tactile message passing should be relative to the spatial distance. We use a radius graph to construct $\mathcal{E}$. In this graph, an edge connects two vertices if their distance is within a certain radius. This approach ensures that the interaction between sensors is stronger when they are closer to each other.

2) Tactile Encoder E_t: We process the constructed taxel graph using a graph neural network (GNN). Specifically, the taxel graph passes through three message passing layers [40], a pooling layer, and a non-linear layer output head. The message-passing layer is defined as

$$n_i' = \gamma(n_i, \bigoplus_{j\in\mathcal{N}(i)} \phi(n_i, n_j, e_{j,i})), \tag{1}$$

where n_i is the vertex feature, and $e_{j,i}$ is the edge feature between vertices i and j. $\bigoplus_{j\in\mathcal{N}(i)}$ is a differentiable aggregation function, such as maximum or summation, and γ and ϕ are two differentiable functions such as MLPs. Our observation is that the taxel signals rely on relative features with respect to their neighbors instead of absolute features. For example, a 4×4 taxel pad with evenly high activation signals and evenly low activation signals should represent the same contact surface (flat surface). Therefore, we propose to use the EdgeConv operator [41], which leverages the relative features between vertices n_i and n_j

$$n_i' = \sum_{j\in\mathcal{N}(i)} (\phi(x_i, h_i, x_j - x_i, h_j - h_i)), \tag{2}$$

where h_i and h_j are the hidden features of node i and j.

3) Sensor-Object Contact Surface Representation: To map the low-resolution tactile signals to the high-resolution surface shape, we need to represent the contact surface between the sensor and the object. To this end, we represent the sensor-object contact surface as a cube encapsulating the object's surface that is in contact with the sensor. The cube's height and width are set to the sensor's dimensions, and its depth, δ_p, represents the penetration into the object's surface. In our experiments, we set $\delta_p = 0.8\ cm$, as lower values of δ_p increased the risk of mesh collision during initialization. The intersection between the object $\mathcal{O}$ and this cube is referred to as a contact surface patch. The contact surface patch captures the local geometry of the object at the contact point, and can be used to learn a correspondence between the low-resolution tactile signal and the high-resolution object surface shape. We view the contact surface patch as the hyper-resolution space of that respective tactile sensor.

4) Learning the Tactile Representation: We propose the contrastive learning framework shown in Fig. 2 to learn the tactile representation that leverages the inherent relationship between the contact surface and the tactile signals. For example, when the sensor is pressed against a flat surface, the taxel signals should demonstrate the flatness feature. On the contrary, when the sensor is pressed against a curved surface, the signals should demonstrate the curvature feature and distinguish itself from the flatness feature. Our key insight for learning an effective representation of tactile signals is to exploit this correspondence.

Inspired by the vision-language learning framework CLIP [24], we draw N random pairs of tactile sensor signals (the blue box) and corresponding contact surfaces (the red box) for each data sample. While the CLIP framework is typically used for visual-language tasks, we adapt it for tactile representation learning. This adaptation requires two significant modifications to the original formulation. 1) Due to the lack of an existing dataset for taxel-based tactile sensors, we need to collect the required paired data. 2) Since the original encoders (vision and language) are incompatible with taxel signals, we need to design a neural network model to encode taxel signals.

The tactile graph described in Section III-A.1 is passed to the tactile encoder E_t (Sec. III-A.2) and encoded into tactile embedding $T \in \mathbb{R}^n$, where n represents the embedding size. Second, the contact surface data (red box), represented as a point cloud, is encoded through the surface encoder E_s into surface embedding $S \in \mathbb{R}^n$, which has the same size as the tactile embedding $\mathbb{R}^n$. We choose PointNet [42] as our surface encoder. The embedding size n is empirically set as 128.

The final step involves learning tactile representation. This is achieved by computing the dot product of the tactile embedding and surface embedding $T \cdot S^\intercal$, resulting in a $N \times N$ matrix as depicted in the bottom right corner of Fig. 2. This matrix contains N positive pairs and $N^2 - N$ negative pairs. The dot product operation measures the cosine similarity between the tactile embedding and surface embedding. We optimize both encoders using a symmetric cross-entropy loss [24] such that the $N \times N$ matrix turns into an identity matrix $\mathbb{I}^{N\times N}$. By doing so, we learn a representation that brings matching pairs closer together and pushes non-matching pairs farther apart in the embedding space.

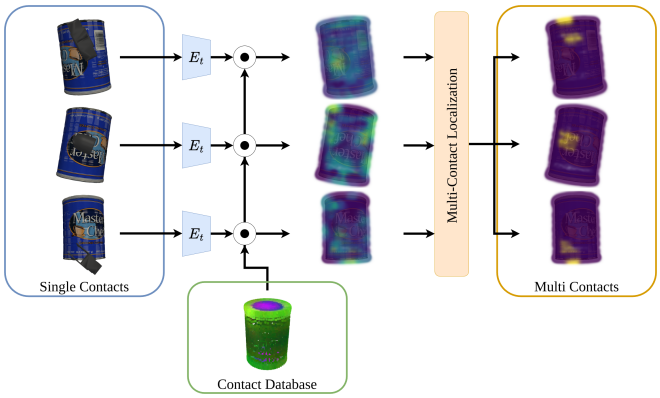


Fig. 3: A visual illustration of multi-contact localization. The first column shows the contact locations. The second and third columns show the likelihood map of contact location using single-, and multi-contact reasoning, respectively.

B. Multi-Contact Localization for Hyper-Resolution

We develop an innovative approach that used multi-contact localization to transform sparse touch data into detailed object surface geometry, thereby achieving hyper-resolution from tactile sensors with limited spatial resolution. Consider a collection of tactile patterns and their corresponding object surface details, analogous to a library of tactile experiences. Given the spatial sparsity inherent in these tactile patterns, confidently associating a tactile pattern with its object surface is a non-trivial task. To address this challenge, we reason over multiple simultaneous contacts with the object to increase confidence in our estimates.

Having a set of objects $\mathcal{O}$, we first collect a contact database $\mathcal{B}_o$ for each object $o \in \mathcal{O}$ offline, which consists of the contact surface patches $\mathcal{S}_{o,c}$, the corresponding contact signals $\mathcal{C}_o$, and the 6D poses of the sensor $\mathcal{P}_{o,c}$ in a common frame of reference F_c. We omit the subscript o in the following paragraph for simplicity and note that these data are object o specific. For each sample b_i in the database $\mathcal{B}_o$, $b_i = (s_{c,i} \in \mathcal{S}_c, c_i \in \mathcal{C}_o, p_{c,i} \in \mathcal{P}_c)$. The 6D pose $p_{c,i} \in \mathbb{R}^7$ is represented as the concatenation of 3D translation $\mathbb{R}^3$ and 3D rotation in quaternion form $\mathbb{R}^4$. To ease online computation, we preprocess the contact surface patches $\mathcal{S}_c$ and encode them into embeddings $\mathcal{S}_l$ using the pretrained surface encoder E_s.

During deployment, we assume there is N_c number of sensors that are in contact with the object o. We denote $\mathcal{J}$ as the set of these sensors and the actual pose of each sensor $j \in \mathcal{J}$ as $\mathcal{P}_{u,j} \in \mathbb{R}^7$. The robot's forward kinematics is used to transform the sensor poses to a common frame of reference. Each sensor j has a sensor reading d_j represented as the taxel graph (Sec. III-A.1). We encode the collection of sensor readings $\mathcal{D} = \{d_i \mid i \in 1, 2, \cdots, N_c\}$ into taxel embeddings $\mathcal{T}$ using tactile encoder E_t such that $\mathcal{T} = \{T_i = E_t(d_i) \mid \forall d_i \in \mathcal{D}\}$. We measure the similarity between $\mathcal{T}$ and the surface embeddings $\mathcal{S}_l$ stored in the database and rank the candidate poses $\mathcal{P}_c$ accordingly. We take top δ_C candidates to reduce the computation in later steps and obtain a distribution of contact locations $\Omega_i \subseteq \mathcal{P}_c$ [28].

Fig. 4: Illustration of data collection using NVIDIA Isaac Sim simulator. The figure shows the curved fingertip sensor (left) and the 4×4 flat pad sensor (right), respectively.

For any two sensors in contact $j_a, j_b \in \mathcal{J}$, we obtain their respective distribution Ω_a, Ω_b for contact location candidates. We filter the pair-wise Euclidean distance between each candidate location $p_{c,a} \in \Omega_a, p_{c,b} \in \Omega_b$ using the actual sensor poses $\mathcal{P}_a, \mathcal{P}_b$

$$\begin{aligned} \Omega_{d,a} &= \{p_{c,a} \mid \|p_{c,a} - p_{c,b}\| - \|\mathcal{P}_a - \mathcal{P}_b\| \leq \delta_n\} \\ \Omega_{d,b} &= \{p_{c,b} \mid \|p_{c,a} - p_{c,b}\| - \|\mathcal{P}_a - \mathcal{P}_b\| \leq \delta_n\}, \end{aligned} \tag{3}$$

resulting in two distance filtered sets $\Omega_{d,a}$ and $\Omega_{d,b}$. δ_n is a threshold to offset noises, such as in calibration and forward kinematics. Repeating this operation on all N_c sensors in contact, we obtain N_c distance filter sets $\Omega_{d,1}, \Omega_{d,2}, \cdots, \Omega_{d,N_c}$.

We desire to find an optimal solution $\Psi \in \Omega_{d,1} \times \Omega_{d,2} \times \cdots \times \Omega_{d,N_c}$ that maximizes the similarities between taxel embeddings $\mathcal{T}$ and the surface embeddings $\mathcal{S}_l$. To achieve this, we build a multipartite graph K with N_c partites. Each candidate left in Ω_d is added as a node, and edges are added if the distance constraint (Eqn. 3) is satisfied. We use Paton's algorithm [43] to find the cycle Ψ with the largest joint probability.

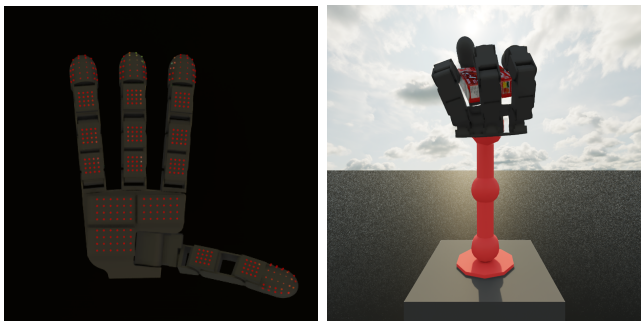
Figure 3 illustrates the process of Hyper-Resolution.

IV. Datasets

We collected two datasets using NVIDIA Isaac Sim for a subset of YCB objects [19]. We used an Allegro Hand equipped with XELA tactile sensors, resembling our real-world setup. The Allegro Hand has eleven 4×4 flat pad sensors, three 4×6 flat pad sensors, and four curved tip sensors (each has 30 taxels). The first dataset, Section IV-A, is a comprehensive database of tactile sensors interacting with the objects. This dataset serves as our tactile experience library and is used to evaluate tactile representation learning. The second dataset, Section IV-B, consists of the Allegro Hand holding an object and executing random trajectories. This dataset is used to evaluate the performance of our methods on a downstream task, namely the in-hand 6D pose estimation task.

A. Contact Database

We first construct a dataset that captures tactile experiences across the entire surface of an object at various points. The tactile sensor is simulated using the `Contact Sensor` provided by Isaac Sim. We sample 2048 points on the object mesh using Poisson disk. Each point corresponds to



(a) (b)

Fig. 5: Allegro Hand equipped with XELA tactile sensors: a) taxel distribution, b) grasping an object

a 3D position $\mathbb{R}^3$ and its respective surface normal $\mathbb{R}^3$. Like previous works [27, 28, 34], we randomly chose a subset of points to collect our tactile experience. For each selected point, we align the tactile sensor's z-axis with the surface normal and conduct eight contact trials. In each trial, the sensor is rotated 45°. We start by positioning the sensor 2.5 cm away from the point and then gradually push it towards the surface along the normal. Once the sensor is in contact with the object, we collect the tactile observations and their corresponding poses. This process is repeated for each type of taxel sensor on the Allegro Hand, which includes $4 \times 4, 4 \times 6$, and curved tips, to compile a comprehensive contact database.

B. In-hand Object Dataset

To evaluate our framework on downstream tasks such as 6D in-hand pose estimation, we also collected a simulated dataset, which consists of the Allegro Hand holding an object and executing random trajectories. For each object, we collect 16,000 samples for the training set and 4,000 samples for the validation set. The dataset is collected using the following procedures:

1) The hand is initialized to face upwards.
2) The object is dropped from a height of $2cm$ above the hand, with its pose randomly initialized.
3) The hand performs the grasping action.
4) If the grasp fails (e.g., the object falls), return to step 1.

V. Experiments

We utilize the AdamW [44] optimizer with a learning rate of 0.001. We pre-train the tactile encoder for 100 epochs using all objects and optimize the pose estimation model for each object for 500 epochs.

We begin by qualitatively assessing the tactile representation learned through our approach (Section V-A). Next, we examine the effectiveness of our method in the hyper-resolution task (Section V-B). We then ablate the chosen graph operators and constructors (Section V-C) and study the impact of multi-contact localization on hyper-resolution (Section V-D). In addition, we test our approach on two downstream tasks: in-hand object pose estimation (Section V-E) and surface classification (Section V-F).

TABLE I: Comparison of Hyper-Resolution Performance using Different Tactile Representations.

Method	Rank ↓	CD ↓
Image-based (CNN)	104.40	0.82
Point cloud-based (PointNet)	150.31	0.85
Graph-based (Ours)	**84.05**	**0.74**

A. Qualitative Analysis of Learned Tactile Representation

To evaluate the quality of our learned tactile representation, we performed a qualitative analysis using visualizations of the tactile embeddings. We used the contact database (Section IV-A) to generate the tactile embeddings for each contact using our learned tactile encoder. We then applied principal component analysis (PCA) to reduce the dimensionality of the embeddings to 3, and used the resulting values as RGB colors for visualization.

The results, as depicted in Fig. 1, reveal that the tactile embeddings effectively capture the geometric features of the contact surface, such as flatness, curvature, and edges. For example, the flat surfaces on the Master Chef Can, Sugar Box, and Mustard Bottle are all represented in shades of purple and blue, while the curved surfaces are depicted in yellow and green. The edges of the Master Chef Can are highlighted in lime green, indicating a stark contrast between the neighboring points. Notably, the tactile embeddings demonstrate consistency across different objects and sensor types, demonstrating the generalization of our representation.

B. Hyper-Resolution Performance Evaluation

In this section, we evaluate the performance of our proposed hyper-resolution algorithm, which maps the low-resolution taxel signals to high-resolution contact surface patches using a contact database. We compare our method with two baselines: image-based approaches (CNN) [2, 14], and point-cloud-based approaches (PointNet) [17, 18].

We use two metrics to measure the accuracy of our hyper-resolution: Chamfer distance (CD) and rank. Chamfer distance computes the average minimum distance between two point sets, and reflects the geometric similarity between the estimated and ground truth contact surfaces. Rank measures the precision of identifying the correct surface based on the similarity between the tactile embeddings and the surface embeddings. The rank of an algorithm is then the average rank it assigns to the ground truth surface across all tactile contact points in the database in Section IV-A. A lower rank means a better performance. Table I shows the results of these experiments. We observe that our method outperforms both of the baselines on both metrics, demonstrating the effectiveness of our hyper-resolution algorithm.

C. Comparison of Graph Operators and Constructors

We ablate the impact of different graph operators and constructors on the quality of the learned tactile representation. We compared our proposed EdgeConv operator with two seminal works: TacGNN [45], GCN [46]. We also examined different graph constructors, such as KNN and radius graphs, with different parameters. Table II shows the results of these

TABLE II: Performance of Different Graph Operators and Constructors on the Hyper-Resolution Task.

Graph Operator	Graph Constructor	Rank ↓	CD ↓
TacGNN	KNN ($n = 1$)	111.03	0.99
TacGNN [45]	KNN ($n = 3$)	87.11	0.79
TacGNN	KNN ($n = 5$)	114.10	1.09
TacGNN	Radius ($r = 0.005$)	201.06	2.47
TacGNN	Radius ($r = 0.01$)	115.71	1.12
TacGNN	Radius ($r = 0.015$)	117.57	1.10
GCN [46]	KNN ($n = 3$)	111.77	0.89
EdgeConv	KNN ($n = 1$)	92.31	0.81
EdgeConv	KNN ($n = 3$)	84.44	0.75
EdgeConv	KNN ($n = 5$)	84.85	0.75
EdgeConv	Radius ($r = 0.005$)	168.37	2.25
EdgeConv (our)	**Radius** ($r = 0.01$)	**84.05**	**0.74**
EdgeConv	Radius ($r = 0.015$)	85.21	0.76

Fig. 6: Effect of Multi-Contact Localization on Hyper-Resolution.

experiments. We observe that our EdgeConv operator combined with the radius graph constructor (r = 0.01) achieves the best performance in terms of rank and CD metrics. This indicates that our operator can effectively capture the relative features between taxels and that the radius graph can better reflect the spatial distance between taxels.

D. Effect of Multi-Contact Localization on Hyper-Resolution

In this section, we evaluate how the number of contacts affects our hyper-resolution algorithm. Taxel-based sensors perceive coarser geometry features, making it challenging to estimate the corresponding surface from a single observation. Previous studies [27, 28] have confirmed performance gains by incorporating multi-contacts on vision-based tactile sensors. In this study, we extend this concept to taxel-based sensors.

Figure 6 shows the results of the quantitative analysis of this experiment. We observe that the CD decreases as the number of contacts increases, indicating that the hyper-resolution quality improves with more contacts. This is because more contacts provide more information and constraints about the object surface, reducing the ambiguity and uncertainty in the hyper-resolution.

We provide a visualization sample in Fig. 3. The second and third column shows the likelihood map of a single contact and multi-contact with the object in each row, respectively. Brighter colors indicate a higher likelihood. We notice the curved surfaces are accurately depicted with brighter colors, suggesting that the algorithm correctly identifies these contacts as originating from a curved surface. The third column shows the refined likelihood map after applying multi-contact reasoning. After multi-contact reasoning, the true contact areas are brightly colored while all other areas are dark, accurately pinpointing the potential origin of the tactile input on the object.

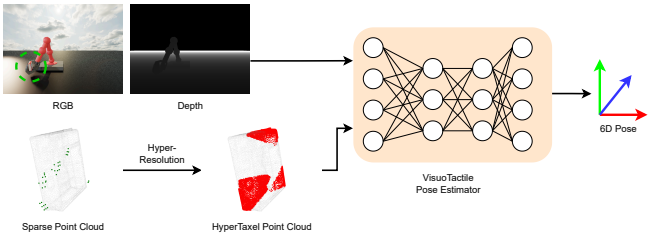


Fig. 7: An illustration of the modified ViTa algorithm with our hyper-resolution module. The tactile data is enhanced by our module to produce a high-resolution representation of the object surface. This representation is then fed into the ViTa algorithm as its tactile input.

TABLE III: Comparative Analysis of the Effect of HyperTaxel on In-Hand 6D Pose Estimation.

Method	Angular Error	Position Error	ADD
DenseFusion	10.52 ± 0.12	0.46 ± 0.00	0.87 ± 0.01
ViTa	8.85 ± 0.10	0.43 ± 0.00	0.77 ± 0.01
ViTa+HyperTaxel	**8.56 ± 0.10**	**0.40 ± 0.00**	**0.74 ± 0.01**

E. Effect of HyperTaxel on In-Hand 6D Pose Estimation

In this section, we evaluate the effectiveness of our approach by integrating it with ViTa [17], an existing visuotactile model for 6D pose estimation. ViTa uses visual and tactile data to represent the object's surface, but low-resolution tactile data can affect its performance. Fig. 7 shows the modified pipeline, which includes our hyper-resolution method to map the sparse low-resolution tactile data to a high-resolution object surface representation. This representation is then fed into the ViTa algorithm without any further changes. Following prior works [16, 17], we evaluate the performance using three metrics: position error (cm), angular error (deg), and ADD (cm). Position error is the L2 norm of the difference between the estimated and ground truth translation vectors, $||t - \hat{t}||_2$. Angular error is the inverse cosine of the inner product of the estimated and ground truth quaternions, $\cos^{-1}(2\langle R, \hat{R}\rangle^2 - 1)$, and ADD measures the pairwise distances between the 3D model points transformed using estimated and ground truth 6D poses, $\frac{1}{m}\sum_{x \in o} ||(Rx + T) - (\hat{R}x + \hat{T})||$, where x is the 3D point, and m is the number of 3D points on the object model o.

We verify the performance of our proposed hyper-resolution algorithm in the synthetic pose estimation data collected in Sec. IV-B. Table III shows the results. We first compare the vision-only baseline DenseFusion [47] with the visuotactile baseline ViTa [17]. By adding tactile information, ViTa has a 1.67 degrees lower angular error and 0.03 cm lower position error. ViTa+HyperTaxel outperforms all of them. An object-wise analysis, depicted in Fig. 8, reveals

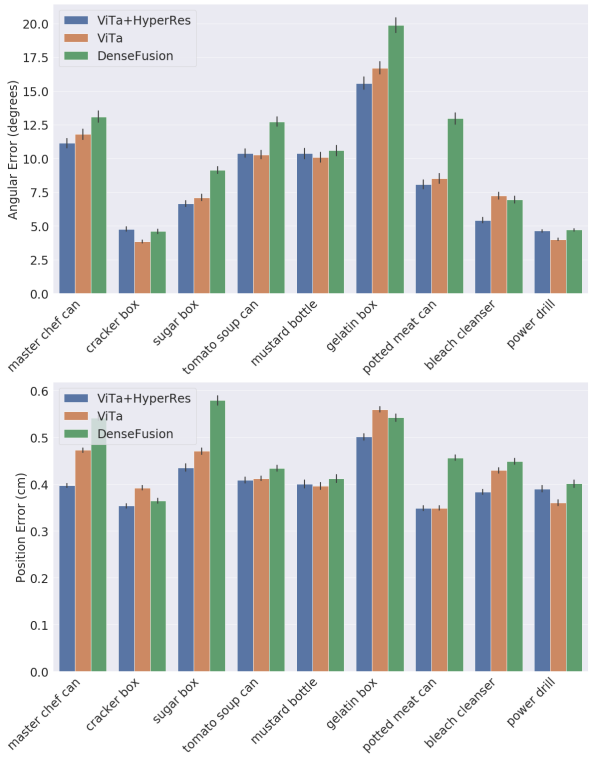


Fig. 8: Pose estimation performance on YCB objects.

Fig. 9: Demonstration of data collection process. We collect a real-world dataset by pressing the tactile sensor-equipped fingertip on flat surfaces (left) and curved surfaces (right).

enhancements for most objects. Notably, our approach faces challenges on power drill objects which reveals one potential limitation. Since our offline collected database relies on random sampling on the object model, our current choice of sample number might not capture the complex geometry of the power drill accurately. In the future, this limitation might be lifted by scaling up the samples on the object.

F. Real Robot Results

We deployed our model, trained on synthetic data on a multi-fingered gripper (Allegro Hand equipped with XELA tactile sensors) affixed to a Sawyer robot. The tactile sensors capture surface contact points on the object. We use real YCB objects to evaluate the performance of our framework in a real-world robot environment.

A good representation should demonstrate ability to distinguish different surface types. We evaluate our representation on the surface classification task to demonstrate its ability to cluster tactile signals based on the geometric features of the contact surfaces, such as flatness and curvature. To conduct this experiment, we use a real-world object that has both flat and curved surfaces: the Master Chef can. As shown in Fig. 9, we press the tactile sensors (both the curved tip and the 4×4 flat pad) on different parts of the can. We collect the tactile signals from multiple contacts on each surface, covering the flat and curved areas as evenly as possible. A video demonstration of this process is available in our supplementary material.

TABLE IV: Surface Classification Performance in Real-Robot Experiments.

Method	RI ↑	ARI ↑	Acc ↑
BYOL	0.546	0.091	83.5
AE	0.502	0.004	83.5
Raw	0.520	0.038	72.8
Ours	**0.586**	**0.171**	**85.4**

We then encode the tactile signals using three representations: BYOL, AE, and ours. We also compare against directly using the raw data (Raw). We then apply the K-means clustering algorithm to classify them into two classes: flat and curved. Table IV shows the results of this experiment, where we measure the performance of our representation using three metrics: random index (RI), adjusted random index (ARI) [48], and accuracy. RI measures the similarity between the estimated clustering and the ground-truth clustering. ARI takes into account the expected value of the RI, which is the random guessing probability. We include the accuracy metric, following the linear classification protocol used to evaluate representation quality in self-supervised learning [21–23] by freezing the learned representation and adding a linear layer to predict the surface class. We observe that our method outperforms the baselines on all metrics, indicating that our representation can effectively cluster the tactile signals based on the geometric features of the contact surfaces.

VI. Conclusions

In this paper, we presented a novel framework, HyperTaxel, for learning a geometrically-informed representation of taxel-based tactile signals to achieve hyper-resolution of contact surfaces between the sensor and the object. We introduced a graph-based representation of tactile signals and a contrastive learning objective to learn a correspondence between the low-resolution taxel signals and the high-resolution contact surfaces. We proposed a multi-contact localization algorithm to reduce the uncertainty and ambiguity in the taxel signals and map them to the object surface geometry. We conducted extensive experiments on synthetic and also presented real-world experiments and showed that our framework outperforms the baselines. We demonstrated that the learned representation can capture the geometric features of the contact surface and generalize across different objects and taxel arrangements. We also showed that the hyper-resolution algorithm can improve the performance of the visuotactile pose estimation model and enable robust sim-to-real transfer.

Our framework opens up new possibilities for leveraging taxel-based tactile sensors for dexterous manipulation and perception. Some future directions for improving the framework include incorporation of temporal information, expanding the contact database, and applying the framework to other modalities.

REFERENCES

[1] R. S. Dahiya *et al.*, "Tactile Sensing—From Humans to Humanoids," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, Feb. 2010.

[2] B. Wu *et al.*, "Tactile Pattern Super Resolution with Taxel-based Sensors," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2022, pp. 3644–3650.

[3] W. Yuan *et al.*, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," *Sensors*, vol. 17, no. 12, p. 2762, Dec. 2017.

[4] E. Donlon *et al.*, "GelSlim: A High-Resolution, Compact, Robust, and Calibrated Tactile-sensing Finger," May 2018, arXiv:1803.00628 [cs].

[5] T. P. Tomo *et al.*, "A New Silicone Structure for uSkin—A Soft, Distributed, Digital 3-Axis Skin Sensor and Its Integration on the Humanoid Robot iCub," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2584–2591, Jul. 2018.

[6] N. Wettels *et al.*, "Biomimetic Tactile Sensor Array," *Advanced Robotics*, vol. 22, no. 8, pp. 829–849, Jan. 2008.

[7] Z. Ding *et al.*, "Sim-to-Real Transfer for Robotic Manipulation with Tactile Sensory," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2021, pp. 6778–6785.

[8] S. Li *et al.*, "Visual–Tactile Fusion for Transparent Object Grasping in Complex Backgrounds," *IEEE Transactions on Robotics*, pp. 1–19, 2023.

[9] Q. K. Luu *et al.*, "Simulation, Learning, and Application of Vision-Based Tactile Sensing at Large Scale," *IEEE Transactions on Robotics*, vol. 39, no. 3, pp. 2003–2019, Jun. 2023.

[10] F. Yang *et al.*, "Touch and Go: Learning from Human-Collected Vision and Touch," Jun. 2022.

[11] S. Li *et al.*, "TaTa: A Universal Jamming Gripper with High-Quality Tactile Perception and Its Application to Underwater Manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*, May 2022, pp. 6151–6157.

[12] P. Piacenza *et al.*, "Data-Driven Super-Resolution on a Tactile Dome," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1434–1441, Jul. 2018.

[13] G. Büscher *et al.*, "Augmenting curved robot surfaces with soft tactile skin," in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2015, pp. 1514–1519.

[14] I. Guzey *et al.*, "Dexterity from Touch: Self-Supervised Pre-Training of Tactile Representations with Robotic Play," in *Proceedings of The 7th Conference on Robot Learning*. PMLR, Dec. 2023, pp. 3142–3166, iSSN: 2640-3498.

[15] R. S. Dahiya *et al.*, "Directions Toward Effective Utilization of Tactile Skin: A Review," *IEEE Sensors Journal*, vol. 13, no. 11, pp. 4121–4138, Nov. 2013.

[16] H. Li *et al.*, "ViHOPE: Visuotactile In-Hand Object 6D Pose Estimation With Shape Completion," *IEEE Robotics and Automation Letters*, vol. 8, no. 11, pp. 6963–6970, Nov. 2023.

[17] S. Dikhale *et al.*, "VisuoTactile 6D Pose Estimation of an In-Hand Object Using Vision and Tactile Sensor Data," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2148–2155, Apr. 2022.

[18] A. Rezazadeh *et al.*, "Hierarchical Graph Neural Networks for Proprioceptive 6D Pose Estimation of In-hand Objects," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 2884–2890.

[19] B. Calli *et al.*, "The YCB object and Model set: Towards common benchmarks for manipulation research," in *2015 International Conference on Advanced Robotics (ICAR)*, Jul. 2015, pp. 510–517.

[20] K. He *et al.*, "Deep Residual Learning for Image Recognition," 2016, pp. 770–778.

[21] J.-B. Grill *et al.*, "Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning," in *Advances in Neural Information Processing Systems*, 2020.

[22] K. He *et al.*, "Momentum Contrast for Unsupervised Visual Representation Learning," 2020, pp. 9729–9738.

[23] ——, "Masked Autoencoders Are Scalable Vision Learners," Dec. 2021, arXiv:2111.06377 [cs].

[24] A. Radford *et al.*, "Learning Transferable Visual Models From Natural Language Supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, Jul. 2021, pp. 8748–8763, iSSN: 2640-3498.

[25] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," Jul. 2020, arXiv:2005.14165 [cs].

[26] F. Yang *et al.*, "Binding Touch to Everything: Learning Unified Multimodal Tactile Representations," Jan. 2024, arXiv:2401.18084 null.

[27] M. B. Villalonga *et al.*, "Tactile Object Pose Estimation from the First Touch with Geometric Contact Rendering," in *Proceedings of the 2020 Conference on Robot Learning*. PMLR, Oct. 2021, pp. 1015–1029, iSSN: 2640-3498.

[28] G. M. Caddeo *et al.*, "Collision-aware In-hand 6D Object Pose Estimation using Multiple Vision-based Tactile Sensors," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, May 2023, pp. 719–725.

[29] C. Corcoran and R. Platt, "A measurement model for tracking hand-object state during dexterous manipulation," in *2010 IEEE International Conference on Robotics and Automation*, May 2010, pp. 4302–4308, iSSN: 1050-4729.

[30] S. Luo *et al.*, "Localizing the Object Contact through Matching Tactile Features with Visual Map," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*, May 2015, pp. 3903–3908, arXiv:1708.04441 [cs].

[31] J. Bimbo *et al.*, "In-Hand Object Pose Estimation Using Covariance-Based Tactile To Geometry Matching," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 570–577, Jan. 2016.

[32] M. C. Koval *et al.*, "The manifold particle filter for state estimation on high-dimensional implicit manifolds," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 4673–4680.

[33] R. Gao *et al.*, "The ObjectFolder Benchmark: Multisensory Learning with Neural and Real Objects," Jun. 2023, arXiv:2306.00956 [cs].

[34] S. Suresh *et al.*, "MidasTouch: Monte-Carlo inference over distributions across sliding touch," in *Proceedings of The 6th Conference on Robot Learning*. PMLR, Mar. 2023, pp. 319–331, iSSN: 2640-3498.

[35] F. Yang *et al.*, "Generating Visual Scenes from Touch," 2023, pp. 22 070–22 080.

[36] N. F. Lepora *et al.*, "Tactile Superresolution and Biomimetic Hyperacuity," *IEEE Transactions on Robotics*, vol. 31, no. 3, pp. 605–618, Jun. 2015.

[37] C. Ledig *et al.*, "Photo-Realistic Single Image Super-Resolution Using a Generative Adversarial Network," 2017, pp. 4681–4690.

[38] Z.-H. Yin *et al.*, "Rotating without Seeing: Towards In-hand Dexterity through Touch," vol. 19, Jul. 2023.

[39] Z. Xue *et al.*, "ArrayBot: Reinforcement Learning for Generalizable Distributed Manipulation through Touch," Jun. 2023, arXiv:2306.16857 [cs].

[40] J. Gilmer *et al.*, "Neural Message Passing for Quantum Chemistry," in *Proceedings of the 34th International Conference on Machine Learning*. PMLR, Jul. 2017, pp. 1263–1272, iSSN: 2640-3498.

[41] Y. Wang *et al.*, "Dynamic Graph CNN for Learning on Point Clouds," Jun. 2019, arXiv:1801.07829 [cs].

[42] C. R. Qi *et al.*, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017.

[43] K. Paton, "An algorithm for finding a fundamental set of cycles of a graph," *Communications of the ACM*, vol. 12, no. 9, pp. 514–518, Sep. 1969.

[44] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," Sep. 2018.

[45] L. Yang *et al.*, "TacGNN: Learning Tactile-Based In-Hand Manipulation With a Blind Robot Using Hierarchical Graph Neural Network," *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3605–3612, Jun. 2023.

[46] A. Garcia-Garcia *et al.*, "TactileGCN: A Graph Convolutional Network for Predicting Grasp Stability with Tactile Sensors," in *2019 International Joint Conference on Neural Networks (IJCNN)*, Jul. 2019, pp. 1–8.

[47] C. Wang *et al.*, "DenseFusion: 6D Object Pose Estimation by Iterative Dense Fusion," 2019, pp. 3343–3352.

[48] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec. 1985.