

000 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 TACTILE AFFORDANCE FOR ROBOT SYNESTHESIA FOR DEXTEROUS MANIPULATION

Anonymous authors

Paper under double-blind review

ABSTRACT

In the field of dexterous robotic manipulation, integrating visual and tactile modalities to inform manipulation policies presents significant challenges, especially in noncontact scenarios where reliance on tactile perception can be inadequate. Visual affordance techniques currently offer effective manipulation-centric semantic priors focused on objects. However, most existing research is limited to using camera sensors and prior object information for affordance prediction. In this study, we introduce a unified framework called Tactile Affordance in Robot Synesthesia (TARS) for dexterous manipulation that employs robotic synesthesia through a unified point cloud representation. This framework harnesses the visuo-tactile affordance of objects, effectively merging comprehensive visual perception from external cameras with tactile feedback from local optical tactile sensors to handle tasks involving both contact and non-contact states. We simulated tactile perception in a virtual environment and trained task-oriented manipulation policies. Subsequently, we tested our approach on four distinct manipulation tasks, conducting extensive experiments to evaluate how different modules within our method optimize the performance of these manipulation policies.

1 INTRODUCTION

Tactile information is needed for robots to perform safe and precise manipulation in emerging domains like healthcare Erickson et al. (2018) and industrial assembly Tang et al. (2023). Recent years have seen a rapid growth in interest in vision-based tactile sensors like Gelsight Yuan et al. (2017a) and Punyo Alspach et al. (2019), which can provide high-resolution contact region observations by sensing the deformation internally via embedded cameras. Vision-based tactile sensors can be used to perceive objects' material properties Yuan et al. (2017b) and perform efficient manipulation She et al. (2021).

In human daily life, we perform fine motor actions through hand-eye coordination in a manner that requires almost no conscious effort. For instance, when our gaze captures an object, we use our hands to grasp it. During the grasping process, if the object is obscured by the hand, humans rely on tactile perception to grasp the object's immediate state. This close collaboration between visual and tactile feedback grants us the ability to manipulate objects dexterously, which is particularly crucial for the transition from coarse to fine task execution. However, for robots, naturally integrating visual and tactile modalities to accomplish operational tasks remains a significant challenge.

In the decision-making sequence of a manipulation task, tactile feedback is not always available. During these intervals, a robot can only rely on visual information to analyze the environment. However, when the robot's end-effector interacts with an object, visual information may be partially obscured, leading to the loss of critical data. The integration of visual and tactile information is particularly crucial for the precise manipulation of small objects. This dual reliance introduces two critical challenges: (i) the manipulation policy must effectively manage transitions between contact and noncontact states, and (ii) the policy must seamlessly integrate information from the inherently different visual and tactile modalities. Most existing research predominantly focuses on visual-tactile coordination in contact-rich scenarios or addresses visual and tactile information separately through distinct modules in tasks with limited contact.

In this paper, we present TARS, a framework designed to uniformly handle both contact and non-contact states while integrating visual and tactile modalities. Drawing on research in visual-tactile

synesthesia and visual affordances, we are the first to apply these concepts to a robotic system using optical tactile sensors and external cameras. We developed a unified point cloud visual-tactile processing module and a multi-state, multi-modal feature processing method trained through visual-tactile affordances. Additionally, we implemented a novel training-deployment framework based on the widely used Teacher-Student reinforcement learning framework for robotic tactile manipulation. Our framework can infer tactile affordances from visual input alone and supplement visual data with tactile information when available. This unified approach enables smooth transitions between contact and non-contact states, integrating visuo-tactile modalities to accomplish various manipulation tasks.

In our study, we used a widely adopted setup comprising of an external camera, a robotic arm, a two-finger parallel gripper, and an optical-based tactile sensor, which is prevalent in both academia and industry settings. We designed four manipulation tasks: Lift, Pick and Place, Pull Drawer, and Open Door. To increase the complexity, we restricted the completion of these tasks to the gripping actions of the two tactile sensors, making them more difficult than tasks without such restrictions. Additionally, unlike some studies that provide prior shape information, we relied solely on data from the external camera to emphasize generalization. In our ablation experiments, we validated the effectiveness of different modules within our framework and assessed its robustness under various physical conditions. Furthermore, we successfully conducted real-world experiments to demonstrate the applicability of our approach.

The rest of this paper is organized as follows: Section II presents related works, while the proposed TARS is detailed in Section III. In Section IV, we compare TARS with existing approaches through different manipulation experiments. Finally, Section V offers concluding remarks and outlines future work.

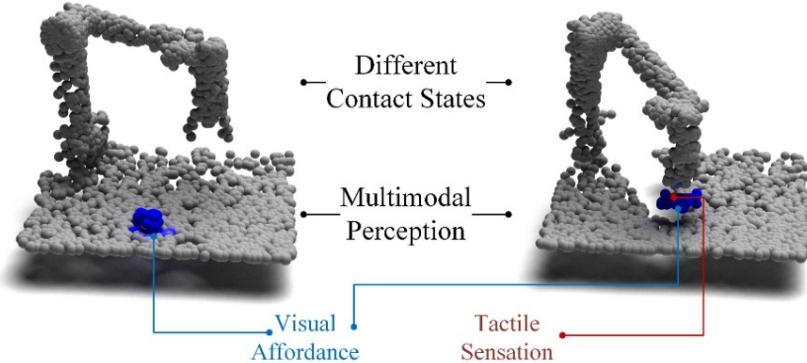
2 RELATED WORK

1) Visual-Tactile Coordination in Robotic Manipulation: Recently, there has been an increasing interest in designing novel tactile sensors Chi et al. (2018). Vision-based tactile sensors such as the Gelsight Yuan et al. (2017a), TacTip Ward-Cherrier et al. (2018) and DIGIT Lambeta et al. (2020) can provide high-resolution contact observations and have been proven useful in manipulation tasks Shah et al. (2021). Conversely, external RGB-D cameras provide critical global information but are prone to interference and occlusion. Some studies [9]–[13] have achieved visual-tactile modality fusion in contact-rich scenarios using sparse representations and self-attention methods. However, the sparse nature of tactile signals in many tasks limits these approaches’ applicability. In low-interaction scenarios, image-based studies [14]–[17] have processed visual and tactile images separately using event cameras or gating mechanisms. These methods heavily rely on camera images when tactile signals are sparse, increasing the Sim2Real challenge. Additionally, research on visual-tactile coordination from a point cloud perspective [18], [19] has primarily focused on dexterous hands and force-tactile sensors, with limited exploration in diverse scenarios. In contrast, TARS builds on point cloud-based visual-tactile coordination methods to achieve a natural integration of visual and tactile modalities in robotic manipulation, as shown in Fig. 1. Our proposed framework incorporates robotic arms, parallel grippers, cameras, and optical tactile sensors. This framework facilitates seamless transitions between different contact states, leveraging both visual and tactile information to enhance robot performance, representing a significant advancement over existing approaches.

2) Visual-Tactile Affordance: Affordance is essential for robotic object manipulation, as it provides actionable information about how an object can be interacted with by robots. Several studies [20]–[23] have highlighted its effectiveness. For example, in point cloud-based robotic manipulation, [24] developed an end-to-end affordance method using reinforcement learning. Other works, such as [25]–[27], collected interaction data to pre-train affordance models before training manipulation policies based on these affordances. These methods, however, often require sampling surface point clouds from 3D CAD (Computer-Aided Design) models to obtain local contact points, relying on prior object information, which can be cumbersome. To address this issue, we conducted contact sampling on objects using both simulated and real optical tactile sensors to obtain precise local information. This approach simplifies the process of acquiring local points, making the affordance acquisition process independent of prior object information, thereby enhancing flexibility and applicability in various scenarios.

3) Point Cloud Based Visual-Tactile Synesthesia: In robotic manipulation research,

108
109
110
111
112
113
114
115
116
117
118
119
120



121 Figure 1: TARS (Tactile Affordance in Robot Synesthesia) provides sufficient information for manip-
122 ulation tasks in both contact and non-contact states and under visuo-tactile multimodal conditions.
123 We show different contact states during a grasping task.

124
125
126
127
128
129
130
131
132
133
134
135
136

some approaches [28], [29] rely solely on visual point clouds to enhance policy robustness, while others [30], [31] use tactile points to improve local perception. Studies on estimating object states [6], [32] often begin with a rough estimate using visual point clouds, refining the estimation with tactile point clouds for greater accuracy. This method encodes visual and tactile point clouds into a coherent 3D space, a concept known as robotic synesthesia [18], which has demonstrated strong capabilities in dexterous manipulation. However, these studies are generally limited to contact-rich states or in-hand manipulations. As shown in Fig. 2, our approach introduces visual-tactile synesthesia encoding based on optical tactile sensors, combined with visual-tactile affordance features to create a unified feature space. This method provides affordance perception through visual-tactile synesthesia in non-contact states and accurate visuo-tactile information in contact states, ensuring smooth transitions between these states. This integration enhances the continuity and effectiveness of manipulation policies across different interaction scenarios.

137
138
139
140
141
142
143
144
145
146

3 OUR APPROACH

We set up a robotic simulation environment in Isaac Gym and implemented tactile simulation using our method. Then, we use the Soft Actor-Critic (SAC) [33], a reinforcement learning algorithm, to train teacher policies for different tasks in the simulation environment with oracle observation. These policies are employed to train the Visual-Tactile Affordance (VTA) and Visual-Tactile Policy (VTP) modules within TARS. Finally, we deploy the trained VTA and VTP modules to realize robotic manipulation tasks.

147

3.1 SIMULATION OF TACTILE POINT CLOUD

148
149
150
151
152
153
154
155
156
157
158
159
160
161

The optical tactile sensor provides tactile information through images. By calibrating and modeling the sensor, we can extract simulated three-dimensional contact information from the two-dimensional tactile image data. In our framework, we decouple this tactile information by decomposing the three-dimensional contact information into a planar contact point and six-axis force information (Fig. 3). The six-dimensional contact force can be obtained in various ways through optical tactile sensors in real environments [34]–[36]. Building on these methods, we use tactile images from the real system as the input and employ a convolutional neural network (CNN) to predict six-dimensional contact forces. These predicted forces are then linearly adjusted to match the contact forces obtained in the simulations. By comparing the tactile images with reference images, we can obtain the planar contact points. In the real system, these points can be mapped to a contact point cloud through the calibration of the robotic arm’s coordinate system with the camera’s coordinate system. For the simulation environment, there are already many environments for tactile simulation [37]–[39], we aim for parallel training of our policy and choose Isaac Gym [40]. In our simulation, we modeled the contact scenario of Gelsight Mini [2] with a depth camera and force sensors to simulate contact states.

162 To represent visual and tactile data using a unified point cloud, we randomly sampled the simulated
 163 tactile depth images to obtain the contact point cloud.
 164

165 **3.2 VISUAL-TACTILE AFFORDANCE**
 166

167 The goal of the membrane model component is to establish a relationship between deformation of the
 168 bubble and their resulting forces. We model the bubble sensor as a homogeneous thin membrane,
 169 similar to Kuppuswamy et al. (2020). Our model considers three types of forces acting on each
 170 element of the bubble: Tension forces from neighboring elements, external forces from contact with
 171 the environment, and pressure force from the air inside. We consider the bubble deformation to be
 172 quasi-static, and solve for static equilibrium:
 173

$$F_{tension} + F_{pressure} + F_{external} = 0. \quad (1)$$

174 We further linearize Equation 1 about a fixed reference configuration of the bubble, defined in
 175 subsection 3.1:
 176

$$\delta F_{tension} + \delta F_{pressure} + F_{external} = 0, \quad (2)$$

177 where $\delta F_{external} = F_{external}$ since the reference configuration has no external forces acting on it.
 178

179 We impose these equilibrium conditions at every mesh vertex and solve for values of these quantities
 180 at those vertices. For this discussion, $\delta F_i \in \mathbb{R}^{M \times 3}$ will denote a matrix of stacked node forces, for
 181 each of the three force quantities.
 182

183 We first lump the change in pressure force to each vertex to compute $\delta F_{pressure}$. We lump the
 184 area to each vertex as $\vec{a} \in \mathbb{R}^M a_i = \frac{1}{3} \sum \text{area}(\Delta) \forall \Delta \in M_i \in \Delta$. Similarly, we compute outward
 185 area-weighted vertex normals $\vec{n}_i \in \mathbb{R}^3$ for each vertex i . $\delta F_{pressure}$ can be computed as:
 186

$$\delta F_{pressure} = \delta p a_1 \vec{n}_1^T a_2 \vec{n}_2^T : a_M \vec{n}_M^T \quad (3)$$

187 where δp is the difference between the observed pressure reading and the reference pressure reading.
 188 $\delta F_{tension}$ is computed by assuming linear elasticity. We characterize the material by its Young's
 189 modulus E and Poisson ratio ν . Following Reissner-Mindlin plate theory, we assume that the
 190 membrane thickness does not change during deformation and that the out of plane stress is zero
 191 (Hughes (2000 - 1987), Section 5.2). Additionally, we assume the bubble's bending stiffness is zero
 192 because the membrane is very thin (0.65mm) compared to its radius of curvature.
 193

194 For a point on the mesh surface, we define the local coordinate system ijk such that i and j are
 195 tangent to the surface and k is perpendicular to the surface. (Refer to Figure 2.) Let σ_{ab} refer to
 196 element (a, b) of the stress tensor σ . ε_{xx} refers to the normal strain in direction x , and γ_{xy} refers
 197 to the shear strain in plane xy . From the frame of reference of such a point, the linear elasticity
 198 equations reduce to:
 199

$$\begin{aligned} \varepsilon_{ii} &= \frac{\sigma_{ii}}{E} - \nu \frac{\sigma_j}{E}, & \varepsilon_{jj} &= \frac{\sigma_{jj}}{E} - \nu \frac{\sigma_i}{E}, \\ \gamma_{ij} &= \frac{2(1+\nu)}{E} \sigma_{ij}, \end{aligned} \quad (4)$$

200 where the assumptions of no thickness change and no bending stiffness give additional constraints
 201

$$\varepsilon_{kk} = \sigma_{kk} = \sigma_{jk} = \sigma_{ik} = 0. \quad (5)$$

202 These equations are applied on a triangular mesh, treating each triangle individually as a flat surface,
 203 following the procedure in Hughes (2000 - 1987) (Section 6.2.13).
 204

205 As part of this simplification, when considering displacements of points in 3D, we project them onto
 206 the triangle plane for the purposes of calculating strain within each individual triangle. Let B be the
 207 partial derivative matrix for a linear triangle element in 2D, such that
 208

$$\varepsilon_{ii} \varepsilon_{jj} \gamma_{ij} = B \vec{u}_{o,2D}^T \vec{u}_{p,2D}^T \vec{u}_{q,2D}^T, \quad (6)$$

209 where $\vec{u}_{v,2D}$ refers to the 2D displacement of vertex v . For our simplification, we set
 210

$$\vec{u}_{v,2D} = \text{project}(\vec{u}_v) = \hat{i} \hat{j}^T \vec{u}_v, \quad (7)$$

216
217
218
219
220
221
222
223
224
225
226
227
228

229
230
231
232
233
234
235
236
237
238
239
240
241
242 
243
244
245
246
247
248
249
250
251
252
253
254
255

256
257 Figure 2: An illustration of the FEM setup. A local coordinate frame is computed for each triangle,
258 and displacements are projected back to the local ij plane to compute strains. [Best viewed in color]
259
260
261
262
263
264
265
266
267
268
269

270 where \vec{u}_v refers to the observed 3D displacement of vertex v :
 271

$$272 \quad \vec{u}_v = \vec{x}_{v,cur} - \vec{x}_{v,ref} \quad (8)$$

273 Refer to Figure 2. The 2D computed node forces are translated back to 3D in an analogous fashion:
 274

$$275 \quad \vec{f}_v = \hat{i}\hat{j}\vec{f}_{v,2D}. \quad (9)$$

277 Following the standard FEM assembly procedure, we get a linear map K between node displacements
 278 and node tension force changes:
 279

$$280 \quad \delta F_{tension} = K \vec{u}_1^T \vec{u}_2^T \dots \vec{u}_{|M|}^T = K(\vec{U}) \quad (10)$$

281 where $\vec{U} \in \mathbb{R}^{3M}$ denotes a vector of stacked node displacements.
 282

283 Substituting Equation 3 and Equation 10, in Equation 2, we obtain an expression for the unknown
 284 contact force $F_{external}$ as a function of the observed displacements \vec{U} and pressure change δp :
 285

$$286 \quad F_{external}(\vec{U}, \delta p) = -K(\vec{U}) + \delta p a_1 \vec{n}_1^T a_2 \vec{n}_2^T \dots a_M \vec{n}_M^T \quad (11)$$

288 To compute contact pressures $P \in \mathbb{R}^{M \times 3}$ at each node, we divide the computed node force by the
 289 area at each vertex:
 290

$$P_{contact} = diag(\vec{a})^{-1} F_{external}. \quad (12)$$

291 We additionally define the continuous pressure distribution $\mathcal{P}_{contact}$ by barycentric interpolation of
 292 $P_{contact}$ within each triangular face.
 293

294 The total contact force \vec{f}_{net} is computed by summing the node forces across the mesh surface,
 295 excluding the boundary:
 296

$$297 \quad \vec{f}_{net} = \sum_{v \in M \setminus \partial M} F_{external,v} \quad (13)$$

298 3.3 VISUAL-TACTILE POLICY

300 With the privileged information provided by the simulation environment, such as the position and pose
 301 of target objects, we can quickly obtain stable reinforcement learning policies through the parallelized
 302 simulation environment. However, this privileged information cannot be directly accessed in actual
 303 POMDP (Partially Observable Markov Decision Process) robotic systems. Therefore, we use a
 304 teacher-student learning approach to distill policies that can be applied in real-world environments
 305 from the trained models. The VTP (Visual-Tactile Policy) framework is illustrated in Fig. 4. We
 306 use the affordance trained by VTA and the visual/tactile one-hot classification encoding together
 307 as point features, which ensures that our feature space is smooth. The point features have three
 308 dimensions: the first dimension is the affordance prediction ranging from 0 to 1, and the second and
 309 third dimensions represent the tactile and visual classification information. We will validate the roles
 310 of these features in Sec. IV-C. Through an encoder of PointNet, we can encode the coordinates and
 311 feature information of the point cloud into a feature vector. The student policy also employs a MLP
 312 as the decision network and we add a fully connected layer to output a Gaussian mixture density. We
 313 utilize a Gaussian Mixture Density Model (GMDM) to handle scenarios where multiple paths are
 314 planned for the same task by teacher policies. The final strategy samples one feasible path from the
 315 mixture density model. The loss function for the VTP module is shown as follows:
 316

317 where $i(a|t|x)$ is a kernel function in the form of a multivariate Gaussian distribution with parameters
 318 μ_i, σ_i , a represents the output action of teacher policy that should be taken, and x is the observation.
 319 The loss function (2) simultaneously trains the PointNet encoder network and the MLP policy network.
 320 The network models a probability density function (PDF) $p(a|s|x)$ as a mixture of m PDFs with the
 321 mixing coefficient $\alpha_i = 0, 1, \dots, m$. The student action a will be generated by sampling from $p(a|s|x)$.
 322 A parallelized teacher-student framework is then established for model training, incorporating our
 323 improvements. The VTP uses the policy trained by SAC as the teacher policy, while the DAgger
 324 [42] method mixes the decisions of the teacher and student policies. Additionally, a replay buffer
 325 was leveraged to utilize the data, continuously supplemented with new data through the parallelized
 326 environment of the Isaac Gym. In summary, our work establishes the synergistic using vision and

324 touch during manipulation processes. The pipeline of our framework is illustrated in Fig. 4, where
 325 our TARS framework comprises two key components: the VTA module, which provides affordance
 326 information, and the VTP module, which makes decisions using mixed encoding. Another significant
 327 aspect is our decoupling of tactile modality information to mitigate the transfer difficulty of the optical
 328 tactile sensor in sim-to-real scenarios. We decompose the tactile information provided by the optical
 329 tactile sensor into contact shape and contact force, and implement this method to achieve tactile
 330 perception in the simulation environment. This tactile decoupling approach enables the deployment
 331 of the VTA and VTP modules on real-world robotic systems.

333 4 EXPERIMENTS

335 The subsequent section evaluates TARS’s performance in comparison to baselines and other variants
 336 in simulations. We focus on three key research questions: (1) How do visualtactile classification
 337 encoding and visual-tactile affordance contribute to policy performance? (2) How does the tactile
 338 point cloud influence grasping decisions? (3) Is our policy robust? These questions will be addressed
 339 in the following experiments.

341 4.1 EXPERIMENTAL SETUP AND TASKS DESCRIPTION

343 We evaluate our proposed method and comparison methods in the Isaac Gym physics simulator. In the
 344 simulation environment, we uniformly use the UR5 robotic arm and the Gelsight Mini tactile sensor
 345 simulation. We set the number of input points to 8192, including 128 tactile sampling points from
 346 two sensors in total, and tested this configuration across the four tasks. Additionally, we performed
 347 4 \times point cloud downsampling, denoted as DS, and directly tested it on some tasks without altering
 348 the policy model. We selected single-stage tasks such as Lift Objects, Pull Drawer, and Open Door,
 349 as shown in Fig. 5. We guided the tasks through rewards to use the tactile sensors on the two-finger
 350 gripper to complete these tasks. Here is a brief introduction to these tasks: Lift Objects: There are
 351 irregular objects on the table with random initial positions and orientations. The agent needs to locate
 352 the object and identify its key parts, then use the tactile sensors to lift the object. Open Door: In
 353 the initial state, the door is closed. The agent needs to use the two tactile sensors on the parallel
 354 gripper to grasp the door handle and open the door to a specific angle. This task requires the agent
 355 to observe key positions of the door and achieve the task with a specific posture, making it very
 356 challenging. Pull Drawer: A drawer is initially closed, similar to open door, the agent needs to use
 357 the two tactile sensors on the parallel gripper to open the drawer to a specific distance. We also
 358 selected the multi-stage Pick and Place task for evaluation, as shown in Fig. 5: Pick and Place: An
 359 object with a random position and orientation is on the table. The agent needs to use tactile sensors
 360 to pick it up and place it at a target point on a separate, higher table. All the tasks mentioned above
 361 were trained using reinforcement learning with the oracle observation, resulting in high-success-rate
 362 teacher policies.

363 4.2 COMPARED METHODS

364 1) **Baselines and Ablations:** We compared our TARS with three main baselines. Our first baseline RS
 365 (Robot Synesthesia) refers to the SOTA (State of the Art) approach in [18], [19], where we use only
 366 the visual and tactile classification one-hot encoding for the features of the visual and tactile point
 367 clouds. In the second baseline VA (Visual Affordance), we did not use classification encoding for the
 368 visual and tactile point clouds; instead, we treat them uniformly as visual encoding and add our VTA
 369 module for prediction, referring to [24], [26]. In the third baseline PN+MLP (PointNet+Multilayer
 370 Perceptron), we retained only the positional features of the visual and tactile point clouds, setting
 371 other features to a uniform value [29]. The results of this approach will be further discussed in
 372 section IV-C. Additionally, following the approach in [24], we considered an end-to-end training
 373 method, where the policy network and the affordance network are trained simultaneously through the
 374 technique of reinforcement learning. However, we were unable to achieve successful convergence, so
 375 these results were not included in the comparisons.

376 2) **Variants:** We evaluated several variants of our model. For our method, we also tested its robustness
 377 under different settings. First, we examined whether the combined visualtactile perception maintained
 robustness with point clouds of varying scales. In the Lift task, we tested the applicability of the policy

378 trained on the Lightbulb object directly on other objects without modification. We also compared the
 379 impact of three different encoding inputs on the policy: our proposed TARS, the visual-tactile direct
 380 concatenation PN+MLP, and the PN+MLP without the tactile perception component. Additionally,
 381 we investigated the performance of policies based on different modalities during the training process
 382 of multi-stage pick and place tasks. This was done to validate the impact and contribution of our
 383 visual-tactile method on policy effectiveness.

384 4.3 SIMULATION RESULTS

385 1) Comparisons to Baselines: The comparison results, as shown in Tab. I, demonstrate that our
 386 method, which combines visuo-tactile classification encoding and visual affordance, achieves the
 387 best overall performance after extensive testing. In tasks involving rich contact, the RS method based
 388 on visuo-tactile classification encoding shows a significant improvement over the PN+MLP method.
 389 Similarly, in tasks with numerous non-contact states, the VA method based on visual affordance also
 390 demonstrates substantial improvement compared to the PN+MLP method. However, since noncontact
 391 scenarios are less frequent, the enhancement provided by VA is not as pronounced as that of RS.
 392 Additionally, our method shows its robustness to point cloud inputs of varying scales, indicating that
 393 the VTA module effectively learns the key tactile features of objects, enabling the policy to utilize this
 394 information effectively. 2) Variants: In the Lift task, we conducted direct tests without replacing the
 395 policy model. We selected six test objects out of twenty that were somewhat similar to the training
 396 object. We also used a visual point cloud policy as a baseline and evaluated the impact of removing
 397 the simulated tactile point cloud from the policy. The results, shown in Tab. II, indicate that our
 398 policy has strong generalization ability and that the local tactile perception significantly enhances
 399 the policy’s performance. Among the test objects, the Apple produced anomalous results, likely due
 400 to its larger volume affecting the gripping policy across all three methods. In addition to evaluating
 401 the performance of the policy upon completion of the task training, we also measured the policy’s
 402 performance at different training steps and recorded the results in Tab. III. The results indicate
 403 that during training, policies based solely on visual modality showed limited improvement after
 404 reaching a certain success rate. In contrast, visual affordance and tactile information were effective at
 405 different stages of training, with visual information aiding in the early stages and tactile information
 406 contributing in the later stages. This synergy resulted in our visual-tactile method achieving the best
 407 performance.

409 5 CONCLUSION

410 We presented a finite element force estimation method for soft-bubble grippers with only three
 411 parameters that can be calibrated with small amounts of data. Our model can run in near real-time and
 412 produce force predictions with accuracy beyond the current state of the art, especially for shear forces.
 413 In future work, we hope to develop a more accurate physical model for the bubble’s deformation: Our
 414 current bubble model uses a simplified model of membrane deformations, ignoring any effect from
 415 changes in the bubble’s curvature or from large displacements changing the orientation of individual
 416 mesh elements. Higher order elements including curvature effects should also improve accuracy. We
 417 also hope to achieve speed improvements by implementation in a compiled language.

420 6 ACKNOWLEDGEMENTS

421 This project has received funding from the National Key R&D Program of China (grant
 422 No.2022YFB4700400), National Natural Science Foundation of China (grant No.62073249), Key
 423 R&D Program of Hubei Province (grant No.2023BBB011).

427 REFERENCES

428 Alexander Alspach, Kunimatsu Hashimoto, Naveen Suresh Kuppuswamy, and Russ Tedrake. Soft-
 429 bubble: A highly compliant dense geometry tactile sensor for robot manipulation. *2019 2nd*
 430 *IEEE International Conference on Soft Robotics (RoboSoft)*, pp. 597–604, 2019. URL <https://api.semanticscholar.org/CorpusID:102336499>.

432 Cheng Chi, Xuguang Sun, Ning Xue, Tong Li, and Chang Liu. Recent progress in technologies
 433 for tactile sensors. *Sensors*, 18(4), 2018. ISSN 1424-8220. doi: 10.3390/s18040948. URL
 434 <https://www.mdpi.com/1424-8220/18/4/948>.

435 Zackory Erickson, Henry M. Clever, Greg Turk, C. Karen Liu, and Charles C. Kemp. Deep haptic
 436 model predictive control for robot-assisted dressing. In *2018 IEEE International Conference on*
 437 *Robotics and Automation (ICRA)*, pp. 4437–4444, 2018. doi: 10.1109/ICRA.2018.8460656.

438 Thomas J. R. Hughes. *The finite element method : linear static and dynamic finite element analysis*.
 439 Dover Publications, Mineola, NY, 2000 - 1987. ISBN 0486411818.

440 Naveen Kuppuswamy, Alejandro Castro, Calder Phillips-Grafflin, Alex Alspach, and Russ Tedrake.
 441 Fast model-based contact patch and pose estimation for highly deformable dense-geometry tactile
 442 sensors. *IEEE Robotics and Automation Letters*, 5(2):1811–1818, 2020. doi: 10.1109/LRA.2019.
 443 2961050.

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

448 Umer Hameed Shah, Rajkumar Muthusamy, Dongming Gan, Yahya Zweiri, and Lakmal Seneviratne.
 449 On the design and development of vision-based tactile sensors. *Journal of Intelligent & Robotic*
 450 *Systems*, 102(82), 2021. ISSN 1573-0409. doi: 10.1007/s10846-021-01431-0. URL <https://doi.org/10.1007/s10846-021-01431-0>.

451 Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable
 452 manipulation with a tactile-reactive gripper. *The International Journal of Robotics Research*, 40
 453 (12-14):1385–1401, 2021. doi: 10.1177/02783649211027233. URL <https://doi.org/10.1177/02783649211027233>.

454 Bingjie Tang, Michael A Lin, Iretiayo A Akinola, Ankur Handa, Gaurav S Sukhatme, Fabio Ramos,
 455 Dieter Fox, and Yashraj S Narang. IndustReal: Transferring Contact-Rich Assembly Tasks from
 456 Simulation to Reality. In *Proceedings of Robotics: Science and Systems*, Daegu, Republic of
 457 Korea, July 2023. doi: 10.15607/RSS.2023.XIX.039.

458 Benjamin Ward-Cherrier, Nicholas Pestell, Luke Cramphorn, Benjamin Winstone, Maria Giannaccini,
 459 Jonathan Rossiter, and Nathan Lepora. The tactip family: Soft optical tactile sensors with 3d-
 460 printed biomimetic morphologies. *Soft Robotics*, 5, 01 2018. doi: 10.1089/soro.2017.0052.

461 Wenzhen Yuan, Siyuan Dong, and Edward H. Adelson. Gelsight: High-resolution robot tactile
 462 sensors for estimating geometry and force. *Sensors*, 17(12), 2017a. ISSN 1424-8220. doi:
 463 10.3390/s17122762. URL <https://www.mdpi.com/1424-8220/17/12/2762>.

464 Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A. Srinivasan, and Edward H. Adelson.
 465 Shape-independent hardness estimation using deep learning and a gelsight tactile sensor. In *2017*
 466 *IEEE International Conference on Robotics and Automation (ICRA)*, pp. 951–958, 2017b. doi:
 467 10.1109/ICRA.2017.7989116.

468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485