TARS: Tactile Affordance in Robot Synesthesia for Dexterous Manipulation

Qiwei Wu, Haidong Wang, Jiayu Zhou, Xiaogang Xiong, Yunjiang Lou

Abstract— **In the field of dexterous robotic manipulation, integrating visual and tactile modalities to inform manipulation policies presents significant challenges, especially in non-contact 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. Our project page is available at https://github.com/NathanWu7/VisualTactile.**

I. INTRODUCTION

In the daily life of humans, we effortlessly combine hand-eye coordination to perform precise manipulation. For example, when we see an object, we grasp it with our hands. If the object becomes obscured by our hand during the grasp, we rely on tactile perception to sense the object's state. This synergy between visual and tactile feedback grants us dexterous manipulation capabilities, essential for shifting tasks from a coarse-grained to a fine-grained level. For robots, however, naturally integrating visual and tactile modalities to accomplish manipulation 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 non-contact 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.

II. RELATED WORK

1) Visual-Tactile Coordination in Robotic Manipulation: Recently, various vision-based tactile sensors [1]–[3] have

This work was supported partially by Shenzhen Fundamental Research Program under the following research Grants KQTD20190929172545139, GXWD 20231130153844002 and partially supported by Guangdong Research Foundation 2022A1515011521, (Corresponding authors: Xiaogang Xiong and Yunjiang Lou).

Special thanks to Ruixuan Jiao from Southeast University, China, for his contributions to the tactile simulation in our simulation environment setup.

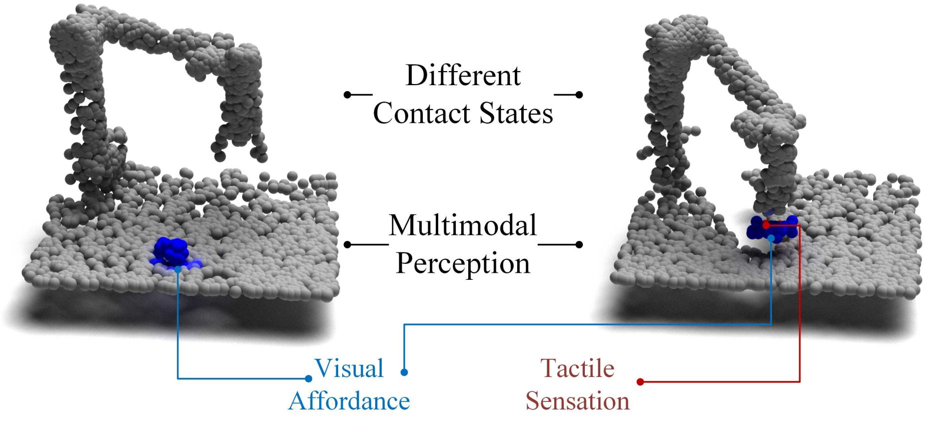


Fig. 1. TARS (Tactile Affordance in Robot Synesthesia) provides sufficient information for manipulation tasks in both contact and non-contact states and under visuo-tactile multimodal conditions. We show different contact states during a grasping task.

been widely applied in robotic manipulation research, significantly enhancing robotic dexterity by estimating object shapes, positions, and contact forces in various tasks [4]–[8]. 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, 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.

III. 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.

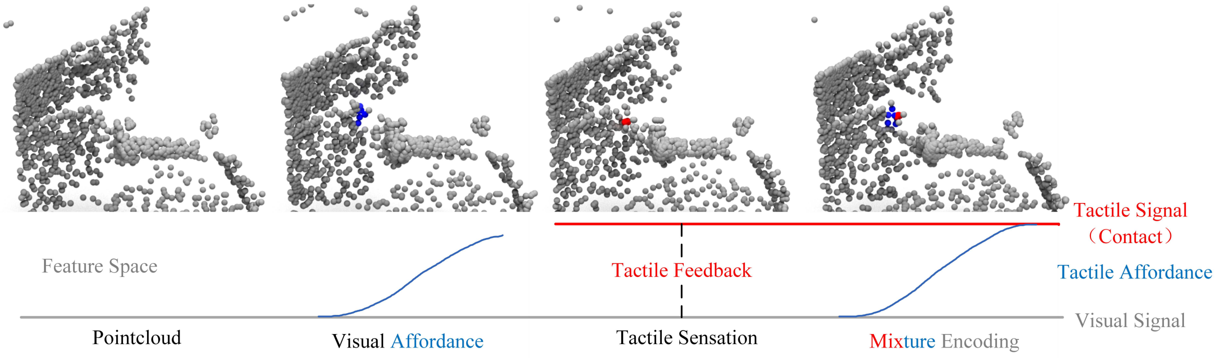


Fig. 2. **The characteristics of TARS.** For point cloud features, visual-tactile synesthesia uses visual and tactile one-hot classification encoding, while visual affordance uses only affordance information. TARS's mixed encoding unifies these features, ensuring that the point cloud has a continuous feature distribution in any state, enabling unified decision-making.

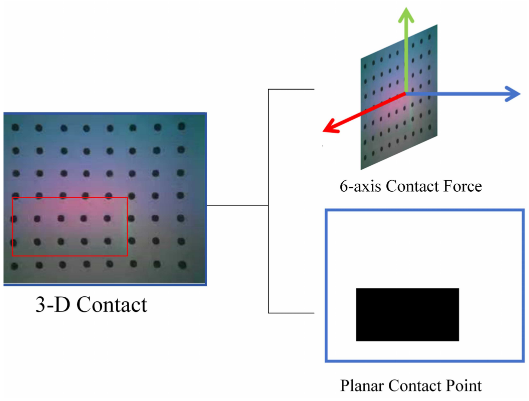


Fig. 3. **Tactile information decoupling.** We decompose the three-dimensional tactile information into planar binary contact points and six-axis contact forces.

A. *Simulation of Tactile Point Cloud*

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

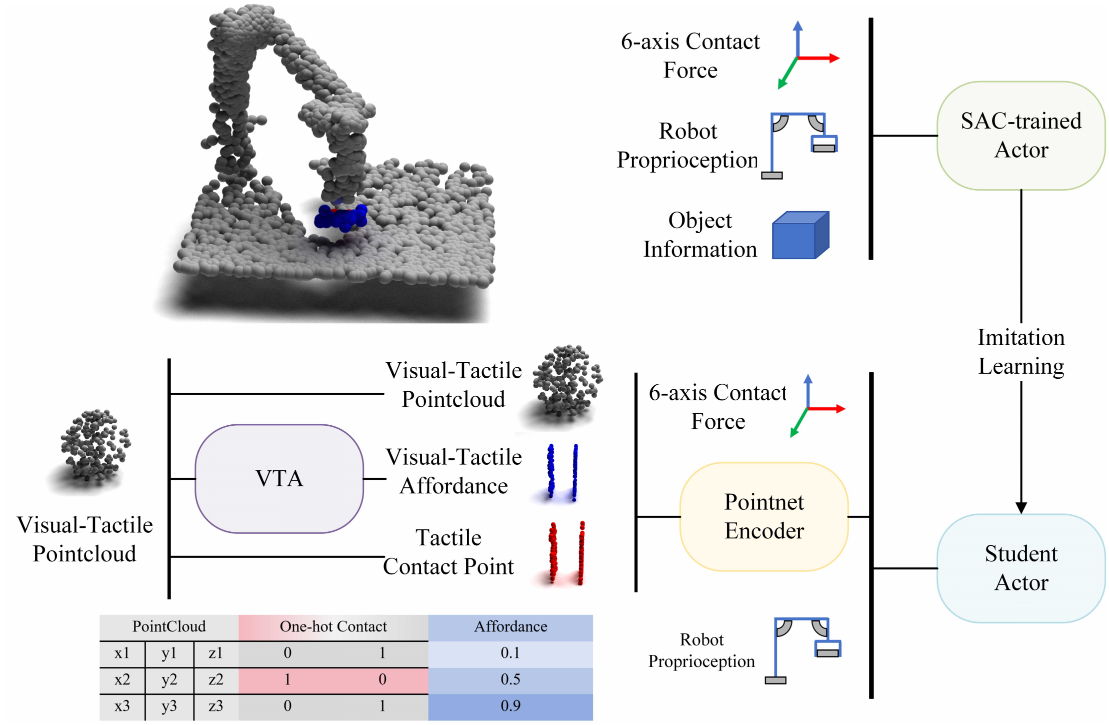
B. *Visual-Tactile Affordance*

We propose a bootstrapping-based approach for end-to-end key feature learning through iterative tactile interactions, eliminating the need for prior CAD model point clouds. In each step of every parallel environment, we set and save a classification label, denoted to *TAL* (Tactile Affordance Label). This label distinguishes between two categories of tactile points: (1) Contact points: Detected by the tactile sensor when in contact with the object. (2) Non-contact points: Identified through a fusion of camera and tactile sensor data when not in direct contact with the object. Subsequently, we employ a unified default feature to fuse visual and tactile points and utilize this integrated representation as input to the Visual-Tactile Affordance (VTA) module.

The VTA model uses the point cloud segmentation structure of PointNet++ [41] as the backbone network. After outputting the multi-dimensional features of the embedding, a few layers of MLP (Multilayer Perceptron) are used to obtain the final one-dimensional feature. VTA's input consists of the coordinates of the points and a default feature value of 1, and its output is a prediction value between 0 and 1, denoted to *TAP* (Tactile Affordance Prediction). We calculate the deviation between *TAP* and *TAL*, using this as the loss function to optimize VTA. We use binary cross-entropy loss as the loss function for the VTA module as follows:

$$\mathscr{L}_{VTA} = -\frac{1}{N}\sum_{i=1}^{N}\left[TAL\log(TAP)+(1-TAL)\log(1-TAP)\right]. \tag{1}$$

Once trained, owing to the VTA module's input comprising a fusion of visual and tactile points, it demonstrates the capability to predict object affordances through the mixed point cloud, both in the presence and absence of tactile points. The specific process is shown in Alg. 1, where the number of contact environments is uncertain, implying that the number of elements of the set K is also uncertain.



PointCloud			One-hot Contact		Affordance
x1	y1	z1	0	1	0.1
x2	y2	z2	1	0	0.5
x3	y3	z3	0	1	0.9

Fig. 4. **Training Pipeline.** Our teacher policy takes robot proprioception, binary contact, and object pose as input. After training the teacher policy via RL, we distill it into a visual-tactile-based student policy. Besides robot proprioception and touch signal, the student policy takes point clouds from depth-camera and tactile sensors. We used a joint encoding composed of visual-tactile synesthetic classification one-hot encoding and affordance encoding as the feature for each point.

Therefore, we established a buffer and updated it in batches of a certain size to ensure the stable convergence of the VTA model.

Algorithm 1 VTA Learning.

```
Require: $E_{1,2,...,n}$: the n parallel environments, $TC_i$: the tactile point cloud for environment $E_i$, $VC_i$: the visual point cloud for environment $E_i$, $\pi$: Trained RL policy, $K$: a set of the index of environments in which contact occurs.
Ensure: $\theta_o$: The initial parameters of VTA
repeat
    $VC_{1,2,...,n}, TC_{1,2,...,n} \leftarrow getPointcloud(E_{1,2,...,n})$;
    $K \leftarrow getIndex(TC_{1,2,...,n})$;  ▷ In which environments contact occurred
    $TAL \leftarrow getTactileAffordanceLabel(TC_K, VC_K)$;  ▷ Label contact points and non-contact points
    $VC'_K \leftarrow SampleVisualPoints(VC_K)$;  ▷ Adjust Input shape
    $TAP \leftarrow predictAffordance(VTA, VC'_K, TC_K)$
    $\theta' \leftarrow update(\theta, TAL, TAP)$;
    $E'_{1,2,...,n} \leftarrow RL(E_{1,2,...,n})$;  ▷ Update environments using RL policy $\pi$
until convergence and return optimal $\theta^*$
```

C. Visual-Tactile Policy

With the privileged information provided by the simulation environment, such as the position and pose of target objects, we can quickly obtain stable reinforcement learning policies through the parallelized simulation environment. However, this privileged information cannot be directly accessed in actual POMDP (Partially Observable Markov Decision Process) robotic systems. Therefore, we use a teacher-student learning approach to distill policies that can be applied in real-world environments from the trained models. The VTP (Visual-Tactile Policy) framework is illustrated in Fig. 4.

We use the affordance trained by VTA and the visual-tactile one-hot classification encoding together as point features, which ensures that our feature space is smooth. The point features have three dimensions: the first dimension is the affordance prediction ranging from 0 to 1, and the second and third dimensions represent the tactile and visual classification information. We will validate the roles of these features in Sec. IV-C.

Through an encoder of PointNet, we can encode the coordinates and feature information of the point cloud into a feature vector. The student policy also employs a MLP as the decision network and we add a fully connected layer to output a Gaussian mixture density. We utilize a Gaussian Mixture Density Model (GMDM) to handle scenarios where multiple paths are planned for the same task by teacher policies. The final strategy samples one feasible path from the mixture density model. The loss function for the VTP module is shown as follows:

$$\mathscr{L}_{vtp} = -\log\left(\sum_{i=0}^{m-1} \alpha_i \phi_i(a^t|x)\right), \tag{2}$$

where $\phi_i(a^t|x)$ is a kernel function in the form of a multivariate Gaussian distribution with parameters $\{\mu_i, \sigma_i\}$, a^t represents the output action of teacher policy that should be taken, and x is the observation. The loss function (2) simultaneously trains the PointNet encoder network and the MLP policy network. The network models a probability density function (PDF) $p(a^s|x)$ as a mixture of m PDFs with the mixing coefficient $\prod = \{\alpha_0, \alpha_1, ..., \alpha_{m-1}\}$. The student action a_s will be generated by sampling from $p(a^s|x)$.

A parallelized teacher-student framework is then established for model training, incorporating our improvements. The VTP uses the policy trained by SAC as the teacher policy, while the DAgger [42] method mixes the decisions of the teacher and student policies. Additionally, a replay buffer was leveraged to utilize the data, continuously supplemented with new data through the parallelized environment of the Isaac Gym.

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

IV. EXPERIMENTS

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

A. Experimental Setup and Tasks Description

We evaluate our proposed method and comparison methods in the Isaac Gym physics simulator. In the simulation environment, we uniformly use the UR5 robotic arm and the Gelsight Mini tactile sensor simulation. We set the number of input points to 8192, including 128 tactile sampling points from two sensors in total, and tested this configuration across the four tasks. Additionally, we performed 4× point cloud downsampling, denoted as *DS*, and directly tested it on some tasks without altering the policy model.

We selected single-stage tasks such as Lift Objects, Pull Drawer, and Open Door, as shown in Fig. 5. We guided the tasks through rewards to use the tactile sensors on the two-finger gripper to complete these tasks. Here is a brief introduction to these tasks:

Lift Objects: There are irregular objects on the table with random initial positions and orientations. The agent needs to locate the object and identify its key parts, then use the tactile sensors to lift the object.

Open Door: In the initial state, the door is closed. The agent needs to use the two tactile sensors on the parallel gripper to grasp the door handle and open the door to a specific angle. This task requires the agent to observe key positions of the door and achieve the task with a specific posture, making it very challenging.

Pull Drawer: A drawer is initially closed, similar to *open door*, the agent needs to use the two tactile sensors on the parallel gripper to open the drawer to a specific distance.

We also selected the multi-stage Pick and Place task for evaluation, as shown in Fig. 5:

Pick and Place: An object with a random position and orientation is on the table. The agent needs to use tactile sensors to pick it up and place it at a target point on a separate, higher table.

All the tasks mentioned above were trained using reinforcement learning with the oracle observation, resulting in high-success-rate teacher policies.

B. Compared Methods

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

2) Variants: We evaluated several variants of our model. For our method, we also tested its robustness under different settings. First, we examined whether the combined visual-tactile perception maintained robustness with point clouds of varying scales. In the Lift task, we tested the applicability of the policy trained on the Lightbulb object directly on other objects without modification. We also compared the impact of three different encoding inputs on the policy: our proposed TARS, the visual-tactile direct concatenation *PN+MLP*, and the *PN+MLP* without the tactile perception component. Additionally, we investigated the performance of policies based on different modalities during the training process of multi-stage pick and place tasks. This was done

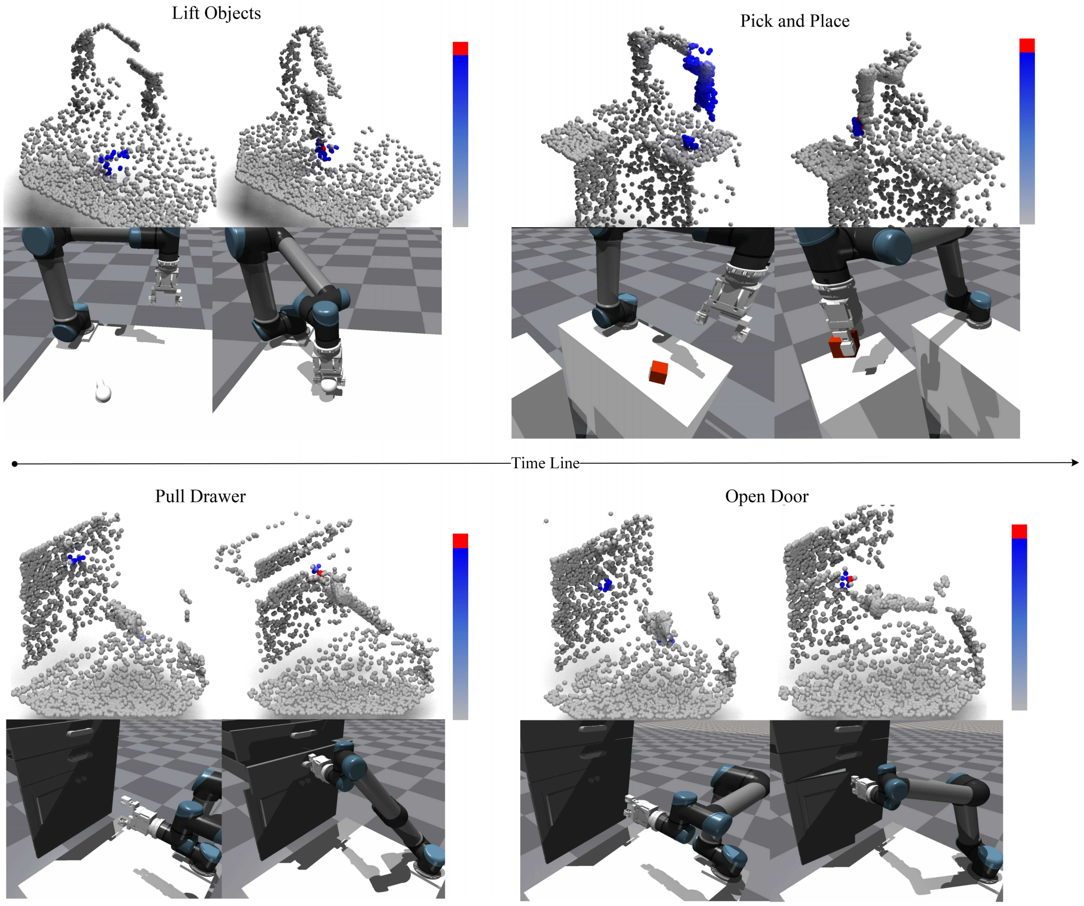


Fig. 5. **Simulations.** For each task, top: the simulation task scenario transitions from a non-contact state to a contact state with the object, bottom: The affordance transitions from a non-contact state to a contact state with the object.

to validate the impact and contribution of our visual-tactile method on policy effectiveness.

C. Simulation Results

1) Comparisons to Baselines: The comparison results, as shown in Tab. I, demonstrate that our method, which combines visuo-tactile classification encoding and visual affordance, achieves the best overall performance after extensive testing. In tasks involving rich contact, the *RS* method based on visuo-tactile classification encoding shows a significant improvement over the *PN+MLP* method. Similarly, in tasks with numerous non-contact states, the *VA* method based on visual affordance also demonstrates substantial improvement compared to the *PN+MLP* method. However, since non-contact scenarios are less frequent, the enhancement provided by *VA* is not as pronounced as that of *RS*. Additionally, our method shows its robustness to point cloud inputs of varying scales, indicating that the VTA module effectively learns the key tactile features of objects, enabling the policy to utilize this information effectively.

2) Variants: In the Lift task, we conducted direct tests without replacing the policy model. We selected six test objects out of twenty that were somewhat similar to the training object. We also used a visual point cloud policy as a baseline and evaluated the impact of removing the simulated tactile point cloud from the policy. The results, shown in Tab. II, indicate that our policy has strong generalization ability and that the local tactile perception significantly enhances the policy's performance. Among the test objects, the *Apple* produced anomalous results, likely due to its larger volume affecting the gripping policy across all three methods.

In addition to evaluating the performance of the policy upon completion of the task training, we also measured the policy's performance at different training steps and recorded the results in Tab. III. The results indicate that during training, policies based solely on visual modality showed limited improvement after reaching a certain success rate. In contrast, visual affordance and tactile information were effective at different stages of training, with visual information aiding in the early stages and tactile information contributing in the later stages. This synergy resulted in our visual-tactile method achieving the best performance.

D. Real-World Transfer

We constructed a digital twin [43] system in both simulation and real-world environments. We selected a robotic system composed of a UR5 robotic arm, a DaHuan PGI model parallel gripper, and a Gelsight Mini tactile sensor,

TABLE I

OUR METHOD DEMONSTRATED GOOD PERFORMANCE ACROSS DIFFERENT TASKS.

Model	Lift	Pull Drawer	Pick and Place	Open Door	Lift(DS)	Pull Drawer(DS)
RL Teacher	0.989	1.0	0.802	0.953	0.989	1.0
TARS(Ours)	**0.783**	**0.954**	**0.426**	**0.248**	**0.697**	**0.556**
RS	0.692	0.936	0.403	0.138	0.494	0.534
VA	0.602	0.778	0.271	0.014	0.321	0.279
PN+MLP	0.685	0.127	0.302	0.003	0.665	0.011

TABLE II

OUR METHOD DEMONSTRATES GOOD GENERALIZATION CAPABILITIES.

Model(Lift Task)	Lightbulb	Banana	Mouse	Apple	Rabbit	Telescope	Elephant
TARS(Ours)	**0.783**	**0.301**	**0.261**	0.411	**0.340**	**0.212**	**0.318**
PN+MLP	0.685	0.272	0.243	0.467	0.335	0.201	0.262
PN+MLP(Camera Only)	0.626	0.261	0.18	**0.471**	0.218	0.177	0.243

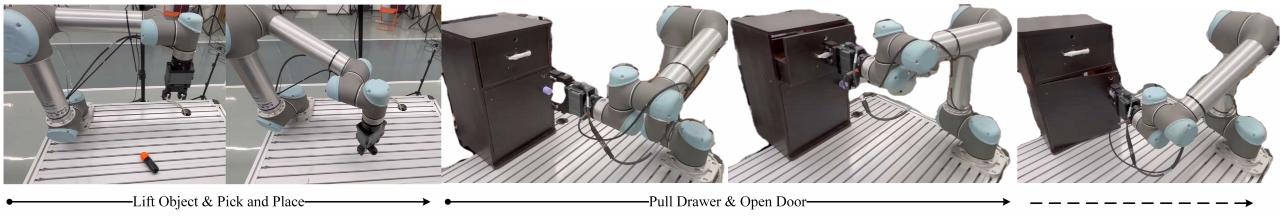


Fig. 6. **Real-world deployment.** From the left, the first image and third image: initial position of the robotic arm. The second image: Lift Object and Pick and Place tasks. The fourth and fifth images: Pull Drawer and Open Door tasks.

TABLE III

THE PERFORMANCE OF DIFFERENT ALGORITHMS IN THE PICK-AND-PLACE TASK AT DIFFERENT TRAINING STEPS.

Epochs	TARS(Ours)	RS	PN+MLP	VA
20k	**0.173**	0.023	0.007	0.093
40k	**0.302**	0.190	0.278	0.175
60k	**0.346**	0.163	0.263	0.269
80k	**0.370**	0.292	0.370	0.267
100k	**0.426**	0.403	0.346	0.271

ensuring consistency between the robot's movements in both simulation and real-world scenarios. To minimize discrepancies between the simulation and real-world environments, we applied linear mappings to the object positions in the real system relative to those in the simulated system.

In our real-world experiments, our inputs were divided into two folds: point clouds from the simulator and proprioceptive feedback from the real-world robot. we used hand-eye calibration to transform both the tactile sensor point cloud and the camera point cloud into the robotic arm coordinate system, consistent with the simulation cases. The action policy was primarily executed based on the policy obtained from the simulator. Additionally, for real-world scenarios, we imposed some simple constraints on the actions to ensure the safety of the sensors, robotic arm, and objects. The experimental setup is illustrated in Fig. 6. Our trained policy model successfully completed tasks in real-world scenarios.

V. CONCLUSION

We proposed the TARS framework, leveraging visual-tactile synesthesia to provide continuous tactile affordance distribution in the absence of tactile signals and enhance contact information when tactile signals are present. This framework simplifies training by enabling the direct acquisition of visual-tactile affordance from the camera without prior object knowledge. Through mixed encoding, it ensures a smooth transition between contact and non-contact states by maintaining a continuous distribution of visual-tactile point cloud features.

Our experiments demonstrate that the TARS framework outperforms baseline methods on visual-tactile point clouds, tested with a robotic arm, parallel gripper, and optical tactile sensor system. The approach shows superior performance across different tasks, and ablation experiments validate the effectiveness of its various modules. Additionally, our policy exhibits better generalization compared to baselines, as evidenced by experiments involving downsampling, removal of local tactile point clouds, and substitution with different objects. We will continue to enhance our real-world policy transfer experiments and expand our simulation framework to support a wider range of sensors and robotic systems.

References

[1] C. Lin, H. Zhang, J. Xu, L. Wu, and H. Xu, "9dtact: A compact vision-based tactile sensor for accurate 3d shape reconstruction and generalizable 6d force estimation," 2023.

[2] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, 2017.

[3] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. 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*, vol. 5, no. 3, p. 3838–3845, Jul. 2020.

[4] J. Varley, D. Watkins-Valls, and P. K. Allen, "Multi-modal geometric learning for grasping and manipulation," *CoRR*, vol. abs/1803.07671, 2018. [Online]. Available: http://arxiv.org/abs/1803.07671

[5] A. N. Chaudhury, T. Man, W. Yuan, and C. G. Atkeson, "Using collocated vision and tactile sensors for visual servoing and localization," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, p. 3427–3434, Apr. 2022.

[6] D. Zhao, F. Sun, Z. Wang, and Q. Zhou, "A novel accurate positioning method for object pose estimation in robotic manipulation based on vision and tactile sensors," *The International Journal of Advanced Manufacturing Technology*, vol. 116, 10 2021.

[7] M. C. Welle, M. Lippi, H. Lu, J. Lundell, A. Gasparri, and D. Kragic, "Enabling robot manipulation of soft and rigid objects with vision-based tactile sensors," 2023.

[8] A. Murali, Y. Li, D. Gandhi, and A. Gupta, "Learning to grasp without seeing," 2018.

[9] W. Zheng, H. Liu, and F. Sun, "Lifelong visual-tactile cross-modal learning for robotic material perception," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 3, pp. 1192–1203, 2021.

[10] F. Yang, C. Ma, J. Zhang, J. Zhu, W. Yuan, and A. Owens, "Touch and go: Learning from human-collected vision and touch," in *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022. [Online]. Available: https://openreview.net/forum?id=ZZ3FeSSPPblo

[11] V. Dave, F. Lygerakis, and E. Rueckert, "Multimodal visual-tactile representation learning through self-supervised contrastive pre-training," 2024.

[12] H. Liu, Y. Yu, F. Sun, and J. Gu, "Visual-tactile fusion for object recognition," *IEEE Transactions on Automation Science and Engineering*, vol. 14, pp. 1–13, 01 2016.

[13] Y. Chen, A. Sipos, M. V. der Merwe, and N. Fazeli, "Visuo-tactile transformers for manipulation," 2022.

[14] T. Wang, C. Yang, F. Kirchner, P. Du, F. Sun, and B. Fang, "Multimodal grasp data set: A novel visual–tactile data set for robotic manipulation," *International Journal of Advanced Robotic Systems*, vol. 16, p. 172988141882157, 01 2019.

[15] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine, "More than a feeling: Learning to grasp and regrasp using vision and touch," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, p. 3300–3307, Oct. 2018.

[16] T. Taunyazov, W. Sng, H. H. See, B. Lim, J. Kuan, A. F. Ansari, B. C. K. Tee, and H. Soh, "Event-driven visual-tactile sensing and learning for robots," 2020.

[17] J. Hansen, F. Hogan, D. Rivkin, D. Meger, M. Jenkin, and G. Dudek, "Visuotactile-rl: Learning multimodal manipulation policies with deep reinforcement learning," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8298–8304.

[18] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang, "Robot synesthesia: In-hand manipulation with visuotactile sensing," 2023.

[19] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," 2022.

[20] W. Liang, F. Fang, C. Acar, W. Q. Toh, Y. Sun, Q. Xu, and Y. Wu, "Visuo-tactile manipulation planning using reinforcement learning with affordance representation," 2022.

[21] A. Khazatsky, A. Nair, D. Jing, and S. Levine, "What can i do here? learning new skills by imagining visual affordances," 2021.

[22] J. Borja-Diaz, O. Mees, G. Kalweit, L. Hermann, J. Boedecker, and W. Burgard, "Affordance learning from play for sample-efficient policy learning," 2022.

[23] A. Simeonov, Y. Du, B. Kim, F. R. Hogan, P. Agrawal, and A. Rodriguez, "Learning to plan with pointcloud affordances for general-purpose dexterous manipulation," in *Conference on Robot Learning*, 2020.

[24] Y. Geng, B. An, H. Geng, Y. Chen, Y. Yang, and H. Dong, "End-to-end affordance learning for robotic manipulation," 2022.

[25] K. Mo, L. Guibas, M. Mukadam, A. Gupta, and S. Tulsiani, "Where2act: From pixels to actions for articulated 3d objects," in *International Conference on Computer Vision (ICCV)*, 2021.

[26] R. Wu, Y. Zhao, K. Mo, Z. Guo, Y. Wang, T. Wu, Q. Fan, X. Chen, L. J. Guibas, and H. Dong, "Vat-mart: Learning visual action trajectory proposals for manipulating 3d articulated objects," *CoRR*, vol. abs/2106.14440, 2021.

[27] Y. Wang, R. Wu, K. Mo, J. Ke, Q. Fan, L. J. Guibas, and H. Dong, "Adaafford: Learning to adapt manipulation affordance for 3d articulated objects via few-shot interactions," *CoRR*, vol. abs/2112.00246, 2021. [Online]. Available: https://arxiv.org/abs/2112.00246

[28] C. Gao, Z. Xue, S. Deng, T. Liang, S. Yang, L. Shao, and H. Xu, "Riemann: Near real-time se(3)-equivariant robot manipulation without point cloud segmentation," 2024.

[29] C. Bao, H. Xu, Y. Qin, and X. Wang, "Dexart: Benchmarking generalizable dexterous manipulation with articulated objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 21 190–21 200.

[30] Z. Ding, Y.-Y. Tsai, W. W. Lee, and B. Huang, "Sim-to-real transfer for robotic manipulation with tactile sensory," 2021.

[31] Y. Du, G. Zhang, and M. Y. Wang, "3d contact point cloud reconstruction from vision-based tactile flow," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 12 177–12 184, 2022.

[32] P. K. Murali, M. Gentner, and M. Kaboli, "Active visuo-tactile point cloud registration for accurate pose estimation of objects in an unknown workspace," 2021.

[33] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," 2018.

[34] H. Li, S. Nam, Z. Lu, C. Yang, E. Psomopoulou, and N. F. Lepora, "Biotactip: A soft biomimetic optical tactile sensor for efficient 3d contact localization and 3d force estimation," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5314–5321, 2024.

[35] C. Zhang, S. Cui, Y. Cai, J. Hu, R. Wang, and S. Wang, "Learning-based six-axis force/torque estimation using gelstereo fingertip visuotactile sensing," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3651–3658.

[36] V. Kakani, X. Cui, M. Ma, and H. Kim, "Vision-based tactile sensor mechanism for the estimation of contact position and force distribution using deep learning," *Sensors*, vol. 21, no. 5, 2021.

[37] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 754–10 761, 2022.

[38] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.

[39] Y. Zhao, K. Qian, B. Duan, and S. Luo, "Fots: A fast optical tactile simulator for sim2real learning of tactile-motor robot manipulation skills," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5647–5654, 2024.

[40] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac gym: High performance gpu-based physics simulation for robot learning," 2021.

[41] C. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," in *Neural Information Processing Systems*, 2017. [Online]. Available: https://api.semanticscholar.org/CorpusID:1745976

[42] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," 2011.

[43] K. Xia, C. Sacco, M. Kirkpatrick, C. Saidy, L. Nguyen, A. Kircaliali, and R. Harik, "A digital twin to train deep reinforcement learning agent for smart manufacturing plants: Environment, interfaces and intelligence," *Journal of Manufacturing Systems*, vol. 58, pp. 210–230, 2021, digital Twin towards Smart Manufacturing and Industry 4.0.