Generalizable 6D Object Pose Tracking using Visuo-Haptic Sensing

Hongyu Li[1], Mingxi Jia[1], Tuluhan Akbulut[1], Yu Xiang[2], George Konidaris[1], Srinath Sridhar[1]
[1]Brown University [2]UT Dallas

***Abstract*— Humans naturally integrate vision and haptics for robust object perception during manipulation. The loss of either modality significantly degrades performance. Inspired by this multisensory integration, prior object pose estimation research has attempted to combine visual and haptic/tactile feedback. Although these works demonstrate improvements in controlled environments or synthetic datasets, they often underperform vision-only approaches in real-world settings due to poor generalization across diverse grippers, sensor layouts, or sim-to-real environments. Furthermore, they typically estimate the object pose for each frame independently, resulting in less coherent tracking over sequences in real-world deployments. To address these limitations, we introduce a novel unified haptic representation that effectively handles multiple gripper embodiments. Building on this representation, we introduce a new visuo-haptic transformer-based object pose tracker that seamlessly integrates visual and haptic input. We validate our framework in our dataset and the Feelsight dataset, demonstrating significant performance improvement on challenging sequences. Notably, our method achieves superior generalization and robustness across novel embodiments, objects, and sensor types (both taxel-based and vision-based tactile sensors). In real-world experiments, we demonstrate that our approach outperforms state-of-the-art visual trackers by a large margin. We further show that we can achieve precise manipulation tasks by incorporating our real-time object tracking result into motion plans, underscoring the advantages of visuo-haptic perception. A complete version of our paper can be found at [14].**

I. Introduction

Accurately tracking object poses is a core capability for robotic manipulation, and would enable contact-rich and dexterous manipulations with efficient imitation or reinforcement learning [8, 10, 23]. Recent state-of-the-art object pose estimation methods, such as FoundationPose [24], have significantly advanced visual tracking by leveraging large-scale datasets. However, relying solely on visual information to perceive objects can be challenging, particularly in contact-rich or in-hand manipulation scenarios involving high occlusion and rapid dynamics.

The cognitive science findings show that humans naturally integrate visual and haptic information for robust object perception during manipulation [5, 9, 18]. For instance, Gordon et al. [7] demonstrated that humans use vision to hypothesize object properties and haptics to refine precision grasps. The human "sense of touch" consists of two distinct senses [2, 16]: the *cutaneous sense*, which detects stimulation on the skin surface, and *kinesthesis*, which provides information on static and dynamic body posture. This integration, known as **haptic perception**, allows humans to effectively perceive and manipulate objects [9]. In robotics, analogous capabilities are achieved through tactile sensors (cutaneous sense) and joint sensors (kinesthesis) [18].

Drawing inspiration from these human capabilities, researchers have explored the integration of vision and touch in robotics for decades. As early as 1988, Allen [1] proposed an object recognition system that combined these modalities. More recently, data-driven approaches have emerged to tackle object pose and shape estimation using visuo-tactile information [3, 6, 11–13, 17, 19–22]. Although promising, these methods face two major barriers that hinder their broader applicability: (i) **Cross-embodiment**: Most approaches overfit specific grippers or tactile sensor layouts, limiting their adaptability. (ii) **Domain generalization**: Compared to visual-only baselines, visuo-tactile approaches struggle to generalize, hindered by insufficient data diversity and model scalability. Moreover, they typically process each frame independently, which can result in less coherent object pose tracking over sequences in real-world deployments. As a result, existing methods are challenging to deploy broadly and often require significant customization to specific robotic platforms.

To address these challenges, we propose **V-HOP**: a two-fold solution for generalizable visuo-haptic 6D object pose tracking. First, we introduce a novel unified haptic representation that facilitates cross-embodiment learning. We consider the combination of tactile and kinesthesis in the form of a point cloud, addressing a critical yet often overlooked aspect of visuo-haptic learning. Second, we propose a transformer-based object pose tracker to fuse visual and haptic features. We leverage the robust visual prior captured by the visual foundation model while incorporating haptics. V-HOP accommodates diverse gripper embodiments and various objects and generalizes to novel embodiments and objects.

We build a multi-embodied dataset with eight grippers using the NVIDIA Isaac Sim simulator for training and evaluation. Compared to FoundationPose [24], our approach achieves 5% improvement in the accuracy of object pose estimation in terms of ADD-S [25] in our dataset. These results highlight the benefit of fusing visual and haptic sensing. In the FeelSight dataset [20], we benchmark against NeuralFeels [20], an optimization-based visuo-tactile object pose tracker, achieving a 32% improvement in the ADD-S metric and *ten times faster* run-time speed. Finally, we perform the sim-to-real transfer experiments using Barrett Hands. Our method demonstrates remarkable robustness and significantly outperforms FoundationPose, which could lose object tracks entirely. When integrated into motion plans, our

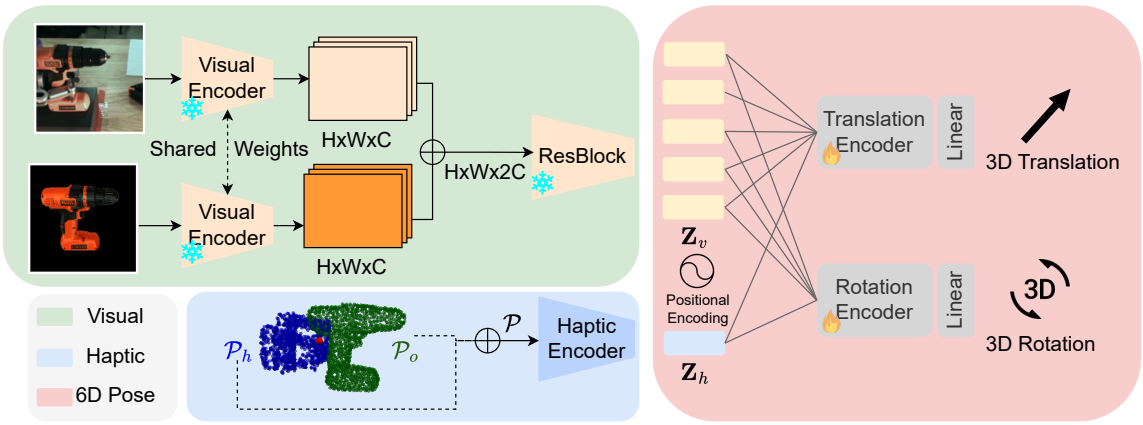

Fig. 1: **Network design of V-HOP.** The visual modality, based on FoundationPose [24], uses a visual encoder to process RGB-D observations (real and rendered) into feature maps, which are concatenated and refined through a ResBlock to produce visual embeddings [4]. The haptic modality encodes a unified hand-object point cloud, derived from 9D hand $\mathcal{P}_h$ and object $\mathcal{P}_o$ point clouds, into a haptic embedding that captures hand-object interactions. The **red dot** in the figure denotes the activated tactile sensor. These visual and haptic embeddings are processed by Transformer encoders to estimate 3D translation and rotation.

approach achieves 40% higher average task success rates. *To the best of our knowledge, V-HOP is the first data-driven visuo-haptic approach to demonstrate robust generalization across both taxel-based tactile sensors (e.g., Barrett Hand) and vision-based tactile sensors (e.g., DIGIT sensors), as well as on novel embodiments and objects.*

In conclusion, our contributions to this paper are two-fold:

1) **Unified haptic representation**: we introduce a novel haptic representation, enabling cross-embodiment learning and addressing the **cross-embodiment challenge** by improving adaptability across diverse embodiments and objects.
2) **Visuo-haptic transformer**: We present a transformer model that integrates visual and haptic data, improving pose tracking consistency and addressing the **domain generalization challenge**.

II. Methodology

We propose V-HOP, a data-driven approach that fuses visual and haptic modalities to achieve accurate 6D object pose tracking. Our goal is to build a *generalizable* visuo-haptic pose tracker that accommodates diverse embodiments and objects. Our choice for the representations follows the spirit of the render-and-compare paradigm [15]. An overview of our network design is at Fig. 1.

III. Experiments

We compare V-HOP against the current state-of-the-art approaches in visual pose tracking (FoundationPose [24], or FP in short) and visuo-tactile pose estimation (ViTa [3]). To ensure a fair comparison, we finetune FoundationPose and train ViTa on our multi-embodied dataset. To verify the generalizability of the novel object and novel gripper, we exclude one object (pudding_box) and one gripper (D'Claw) during training.

Object Name	AUC Metric	ViTa	FP	V-HOP
master_chef_can	ADD	5.61	**64.95**	62.88
	ADD-S	80.51	84.60	**86.38**
sugar_box	ADD	11.09	73.21	**74.75**
	ADD-S	74.34	85.27	**89.35**
tomato_soup_can	ADD	32.08	57.02	**59.13**
	ADD-S	84.19	78.45	**83.30**
mustard_bottle	ADD	7.23	72.65	**74.07**
	ADD-S	73.49	86.05	**88.57**
pudding_box (Unseen)	ADD	N/A	69.87	**70.75**
	ADD-S	N/A	84.63	**88.20**
gelatin_box	ADD	43.20	63.89	**69.75**
	ADD-S	86.66	80.16	**86.87**
potted_meat_can	ADD	34.13	65.62	**68.29**
	ADD-S	86.77	82.67	**87.21**
banana	ADD	23.93	63.87	**69.72**
	ADD-S	71.67	79.99	**85.79**
mug	ADD	35.05	**59.60**	58.42
	ADD-S	86.58	82.16	**84.10**
power_drill	ADD	2.58	67.21	**68.56**
	ADD-S	61.02	80.77	**85.77**
scissors	ADD	23.34	66.23	**70.67**
	ADD-S	65.56	81.27	**85.08**
large_marker	ADD	42.43	61.74	**71.10**
	ADD-S	73.69	75.45	**85.00**
large_clamp	ADD	30.56	71.64	**75.63**
	ADD-S	79.20	86.07	**89.09**
All	ADD ↑	23.93	66.29	**68.90**
	ADD-S ↑	76.87	82.37	**86.62**

TABLE I: **Per-object comparison of AUC metrics for ADD and ADD-S**. The row of novel object is grayed out. Both metrics are the higher, the better. The best results are **bolded**.

In Tab. I, we show the performance for each object. V-HOP consistently outperforms ViTa and FoundationPose (FP) on most objects with respect to ADD and across all objects in terms of ADD-S. On average, our approach delivers an improvement of 4% in ADD and 5% in ADD-S compared to FoundationPose. Notably, V-HOP demonstrates strong performance on unseen objects, highlighting the potential of our model to generalize effectively to novel objects.

References

[1] Peter K. Allen. Integrating Vision and Touch for Object Recognition Tasks. *The International Journal of Robotics Research*, 7(6):15–33, December 1988.

[2] Ravinder S. Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile Sensing—From Humans to Humanoids. *IEEE Transactions on Robotics*, 26(1): 1–20, February 2010.

[3] Snehal Dikhale, Karankumar Patel, Daksh Dhingra, Itoshi Naramura, Akinobu Hayashi, Soshi Iba, and Nawid Jamali. VisuoTactile 6D Pose Estimation of an In-Hand Object Using Vision and Tactile Sensor Data. *IEEE Robotics and Automation Letters*, 7(2): 2148–2155, April 2022.

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021.

[5] Marc O. Ernst and Martin S. Banks. Humans integrate visual and haptic information in a statistically optimal fashion. *Nature*, 415(6870):429–433, January 2002.

[6] Yuan Gao, Shogo Matsuoka, Weiwei Wan, Takuya Kiyokawa, Keisuke Koyama, and Kensuke Harada. In-Hand Pose Estimation Using Hand-Mounted RGB Cameras and Visuotactile Sensors. *IEEE Access*, 11: 17218–17232, 2023.

[7] A. M. Gordon, H. Forssberg, R. S. Johansson, and G. Westling. The integration of haptically acquired size information in the programming of precision grip. *Experimental Brain Research*, 83(3):483–488, February 1991.

[8] Cheng-Chun Hsu, Bowen Wen, Jie Xu, Yashraj Narang, Xiaolong Wang, Yuke Zhu, Joydeep Biswas, and Stan Birchfield. SPOT: SE(3) Pose Trajectory Diffusion for Object-Centric Manipulation, November 2024.

[9] Simon Lacey and K. Sathian. Chapter 7 - Visuo-haptic object perception. In K. Sathian and V. S. Ramachandran, editors, *Multisensory Perception*, pages 157–178. Academic Press, January 2020.

[10] Albert H. Li, Preston Culbertson, Vince Kurtz, and Aaron D. Ames. DROP: Dexterous Reorientation via Online Planning, October 2024.

[11] Hongyu Li, Snehal Dikhale, Soshi Iba, and Nawid Jamali. ViHOPE: Visuotactile In-Hand Object 6D Pose Estimation With Shape Completion. *IEEE Robotics and Automation Letters*, 8(11):6963–6970, November 2023.

[12] Hongyu Li, Snehal Dikhale, Jinda Cui, Soshi Iba, and Nawid Jamali. HyperTaxel: Hyper-Resolution for Taxel-Based Tactile Signals Through Contrastive Learning. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7499–7506, October 2024.

[13] Hongyu Li, James Akl, Srinath Sridhar, Tye Brady, and Taskin Padir. ViTa-Zero: Zero-shot Visuotactile Object 6D Pose Estimation, April 2025.

[14] Hongyu Li, Mingxi Jia, Tuluhan Akbulut, Yu Xiang, George Konidaris, and Srinath Sridhar. V-HOP: Visuo-Haptic 6D Object Pose Tracking, February 2025.

[15] Yi Li, Gu Wang, Xiangyang Ji, Yu Xiang, and Dieter Fox. DeepIM: Deep Iterative Matching for 6D Pose Estimation. pages 683–698, 2018.

[16] Jack M. Loomis and Susan J. Lederman. Tactual perception. In *Handbook of perception and human performance, Vol. 2: Cognitive processes and performance*, pages 1–41. John Wiley & Sons, Oxford, England, 1986.

[17] Shan Luo, Wenxuan Mou, Kaspar Althoefer, and Hongbin Liu. iCLAP: shape recognition by combining proprioception and touch sensing. *Autonomous Robots*, 43(4):993–1004, April 2019.

[18] Nicolás Navarro-Guerrero, Sibel Toprak, Josip Josifovski, and Lorenzo Jamone. Visuo-haptic object perception for robots: an overview. *Autonomous Robots*, 47(4):377–403, April 2023.

[19] Alireza Rezazadeh, Snehal Dikhale, Soshi Iba, and Nawid Jamali. Hierarchical Graph Neural Networks for Proprioceptive 6D Pose Estimation of In-hand Objects. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2884–2890, May 2023.

[20] Sudharshan Suresh, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. NeuralFeels with neural fields: Visuotactile perception for in-hand manipulation. *Science Robotics*, November 2024.

[21] Yuyang Tu, Junnan Jiang, Shuang Li, Norman Hendrich, Miao Li, and Jianwei Zhang. PoseFusion: Robust Object-in-Hand Pose Estimation with SelectLSTM. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 6839–6846, October 2023.

[22] Zhaoliang Wan, Yonggen Ling, Senlin Yi, Lu Qi, Wang Wei Lee, Minglei Lu, Sicheng Yang, Xiao Teng, Peng Lu, Xu Yang, Ming-Hsuan Yang, and Hui Cheng. VinT-6D: A Large-Scale Object-in-hand Dataset from Vision, Touch and Proprioception. In *Proceedings of the 41st International Conference on Machine Learning*, pages 49921–49940. PMLR, July 2024.

[23] Bowen Wen, Wenzhao Lian, Kostas Bekris, and Stefan Schaal. You Only Demonstrate Once: Category-Level Manipulation from Single Visual Demonstration, May 2022.

[24] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. FoundationPose: Unified 6D Pose Estimation and Tracking of Novel Objects. In *CVPR*, 2024.

[25] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. PoseCNN: A Convolutional Neural Network for 6D Object Pose Estimation in Cluttered Scenes. In *RSS*, volume 14, June 2018.