Grounded Semantic Reasoning for Robotic Perception and Manipulation

Weiyu Liu

I. Introduction

For robots to autonomously operate in unstructured human environments, they need to effectively manipulate novel objects while adapting to changes in environments and goals. Semantic knowledge about objects, which represent relations between object categories, locations, properties, and uses, can help robots develop a mental model of objects and serve as a structured prior that relates different tasks [19, 2, 14, 30]. Reasoning about the encoded knowledge further enables robots to act according to what are implicitly represented rather than the subbody of knowledge they have access to. For example, from the observation that a bowl is made of ceramic and the knowledge that the ceramic material is fragile, a robot can perform deductive reasoning to infer that the bowl is fragile and needs to be handled cautiously.

A crucial challenge for applying semantic reasoning to robotic perception and manipulation is to connect symbolic representations (e.g., knowledge graphs and textual facts) with robots' sensorimotor data (e.g., object point clouds, 6-dof poses, spectroscopic reading of surfaces), also known as the symbol grounding problem [16]. These two types of the data have drastically different characteristics, in terms of modality, granularity, and diversity. A useful connection between the two types of data should not only be a direct mapping from invariant features in the sensorimotor data to predefined categorical representations but rather a deeper interaction that reveals meaningful patterns of object properties and uses.

Prior work has demonstrated that symbolic knowledge are useful for tasks involving high-level decision making, including semantic navigation [32], human-robot interaction [17], and task planning [11]. Large knowledge bases, such as KnowRob [41] and RoboBrain [35], provide a unified knowledge representation to store both concept taxonomy and sensorimotor data; however, reasoning is lacking. Recent methods from relational machine learning on knowledge graphs [29] has enabled robots to model soft statistical patterns [10] and perform multi-hop reasoning [21]. These techniques, however, are often applied to high-level commonsense knowledge sources such as ConceptNet [38] and WordNet [27], without connection to sensorimotor knowledge.

Huge stride has been made on the development of structured sensorimotor representation for integrating perception with manipulation [37, 44]. Vision-based object representations, such as affordance segmentation [12] and keypoint [25], enable robots to generalize skills to novel instances of objects in the same category. Understanding spatio-semantic relations (e.g., left, contain) has also shown to be useful for many applications such as language-conditioned object retrieval [36, 28] and moving object to achieve desired pairwise spatial relations [31, 26]. Interactive perception [4] utilizes different sensor modalities and exploratory actions to ground semantic object properties (e.g., thin, rough and compressible) [7, 42], but the inherent relations between object properties are not currently exploited. In general, existing manipulation and perception methods lack generalization to novel object categories, tasks, and more complex real-world environments.

My research aims to *combine high-level conceptual knowledge and low-level skills by building semantic reasoning frameworks that are capable of modeling higher-order semantic relations grounded in robots' sensorimotor data.*

II. Predicting Multimodal Object Properties with N-Ary Relations

Prior work has encoded relations between object properties, especially those that can be grounded in multimodal interactive perception, primarily as binary relations between an object's *class* label and its *semantic properties* (e.g., (*cup*, *is*, *fragile*)) [10, 6, 46, 40, 35]. Such representation fails to integrate closely with robot perception; for example, observing that the cup is wet does not help to infer that the cup is more likely to be located in sink than in cabinet. In our work [22], we propose to use n-ary relations to model higher-order interactions between object properties. However, collecting semantically meaningful n-ary relations is challenging because it requires various object properties to be conditioned on each other. We obtain n-ary observations, each representing a set of identified properties of an object instance within a particular environmental context (e.g., a *small silver metal cup* that is *wet* and *in sink*). We then mine generalizable patterns from n-ary observations with a permutation-invariant transformer neural network trained with an autoencodeing objective. Since the learned model implicitly encodes statistical rules (e.g., *paper* is *light*) that apply to any objects, the model can predict properties of novel objects in different environmental contexts given different amounts of observed information. We validate on a unique dataset we crowdsource that contains 15 multimodal properties types and 200 total properties. Compared to the prior state of the art Markov Logic Network [34], our model obtains a 10% improvement in metric score while reducing training and inference time by 150 times. We also apply our model to a mobile manipulator, demonstrating the ability to retrieve objects based on desired properties of objects (Fig. 1 leftmost) and actively detect object properties.

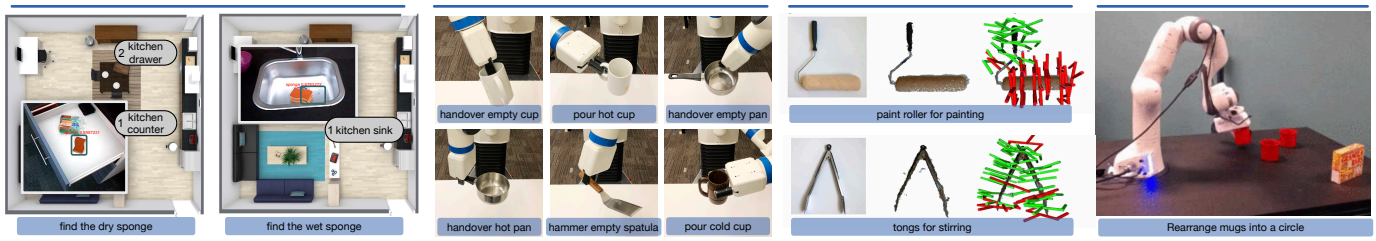

Fig. 1. My work has enriched robots' spatial understanding as well as semantic understanding of object properties and uses.

III. Reasoning for Context-Aware Grasping.

We further apply higher-order reasoning of object properties to the sensorimotor skill of task-oriented grasping from object point clouds. In contrast to classic grasping which has focused on getting hold of objects [3, 18, 24], task-oriented grasping [9] resembles how humans grasp an object [8]; we do so with a clear purpose of preparing the object to be used for an intended manipulation task. For example, when we grasp a cup for drinking from it, we use the handle though other stable grasps exist. In our work [20], we reason about a broader range of context for achieving purposeful grasping, including object affordance, material, state, and task. To discover structure from the high-dimensional, irregular, and multi-modal data associated with the varying context, we leverage semantic features as an abstract intermediate representation, which can be acquired with perception modules such as pixel-wise affordance detection [12] and material classification from spectroscopic data [13]. Compared to existing methods that consider less context or directly learn from low-level sensor data, our method more effectively captures the complex reasoning patterns for selecting suitable grasps and can generalize to a broader range of novel situations with statistic significance. Deployed on a mobile manipulator, the robot can extract and reason about semantic information to execute semantically correct grasps on everyday objects (Fig. 1 middle left).

IV. Generalizing Task-Oriented Grasping with A Knowledge Graph

Besides reasoning about semantic representations of environmental contexts, we also investigate whether a knowledge graph encoding semantic relations between objects and tasks (e.g., (*cup*, *is_a*, *container*) and (*tongs*, *used_for*, *stirring*)) can be used as a structured prior to generalize grasping to novel objects and tasks, and skills beyond an object's prototypical use. To systematically study generalization, we collect a dataset of 250K task-oriented grasps for 56 tasks and 191 objects (Fig. 1 middle right). We introduce a framework based on a Graph Convolutional Network (GCN) that incorporates the knowledge graph into the end-to-end learning of task-oriented grasping from object point clouds. We further leverage word embeddings trained on large-scale linguistic datasets and commonsense knowledge bases to initialize the node embeddings in the GCN to provide additional prior information. Our method shows a significant improvement of 12% and 3.5% on zero-shot generalization to novel tasks and object categories, respectively, compared to baselines which do not incorporate semantic knowledge. The method is deployed to a 7-DOF Sawyer Robot for executing task-oriented stable grasps for novel objects and tasks.

V. Learning Multi-Object Spatial Structure for Language-Conditioned Rearrangement

Continuing the study of semantic and spatial relations, we further examine multi-object spatio-semantic relations for object rearrangement, which has many real-world applications (e.g., setting the table and loading the dishwasher) and has been recognized as a benchmark for embodied AI [1]. We focus on the problem of semantic rearrangement, where a robot must move a set of novel objects to form a spatial structure that satisfies a high-level language instruction (e.g., *put the red mugs in a row*, *build a circle of the wine bottles*, and *set the table*). Compared to rearrangement based on visual goals [33, 15, 43], language provides an intuitive input modality for untrained users [39]; however, it brings with it challenges in inferring the implied object configuration directly from symbolic goal specifications. In our work [23], we introduce a novel framework that uses transformer encoders to jointly reason about both language instructions as well as semantic and geometric features of objects extracted from segmented point clouds. The encoders can directly predict what objects to move and also provide a multi-object context for an autoregressive transformer decoder to predict target 6-DoF poses representing where the objects should go and how they should be oriented. We validate on a procedurally generated dataset for different structures (circles, lines, towers, and table settings) using 335 3D object models from ShapeNet [5]. We show through rigorous experiments that our model enables robots to rearrange novel objects into meaningful structures with multi-object relational constraints inferred from the language command (Fig. 1 rightmost), also more effectively than prior methods modeling pairwise spatial relations.

VI. Conclusion

To address the challenges associated with operation in real-world domains, robots must effectively generalize knowledge, learn, and be transparent in their decision making. My research has demonstrated that semantic knowledge grounded in perception and manipulation provide robots (1) a *unified representation* to identify the meaningful patterns of multi-modal object properties and uses, (2) a *structured prior* to help robots efficiently generalize sensorimotor skills to novel

object categories and tasks, and (3) *intuitive input modalities* to receive instructions from human users.

This work, however, is a first step towards integrating reasoning into the loop of perception and manipulation. First, I plan to develop an active perception method that combine n-ary relational knowledge with sequential decision making to guide interactive perception of object properties. Second, I am excited to extend the work of task-oriented grasping and my previous research on affordance-based keypoint [45] to learn a 3D semantic representation of objects, with the goal of generalize manipulation skills to novel objects and tasks under language guidance.

References

[1] Dhruv Batra, A. Chang, S. Chernova, A. Davison, Jun Deng, V. Koltun, Sergey Levine, J. Malik, Igor Mordatch, R. Mottaghi, M. Savva, and Hao Su. Rearrangement: A challenge for embodied ai. *ArXiv*, abs/2011.01975, 2020.

[2] Michael Beetz, Raja Chatila, Joachim Hertzberg, and Federico Pecora. Ai reasoning methods for robotics. In *Springer Handbook of Robotics*, pages 329–356. Springer, 2016.

[3] Antonio Bicchi and Vijay Kumar. Robotic grasping and contact: A review. In *Proceedings 2000 ICRA. Millennium Conference. IEEE International Conference on Robotics and Automation. Symposia Proceedings (Cat. No. 00CH37065)*, volume 1, pages 348–353. IEEE, 2000.

[4] Jeannette Bohg, Karol Hausman, Bharath Sankaran, Oliver Brock, Danica Kragic, Stefan Schaal, and Gaurav S Sukhatme. Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33(6):1273–1291, 2017.

[5] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[6] Sonia Chernova, Vivian Chu, Angel Daruna, Haley Garrison, Meera Hahn, Priyanka Khante, Weiyu Liu, and Andrea Thomaz. Situated bayesian reasoning framework for robots operating in diverse everyday environments. In *International Symposium on Robotics Research*, 2017.

[7] Vivian Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Trevor Darrell, and Katherine J Kuchenbecker. Robotic learning of haptic adjectives through physical interaction. *Robotics and Autonomous Systems*, 63:279–292, 2015.

[8] F Cini, V Ortenzi, P Corke, and M Controzzi. On the choice of grasp type and location when handing over an object. *Science Robotics*, 4(27), 2019.

[9] Hao Dang and Peter K Allen. Semantic grasping: Planning robotic grasps functionally suitable for an object manipulation task. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1311–1317. IEEE, 2012.

[10] Angel Daruna, Weiyu Liu, Zsolt Kira, and Sonia Chetnova. Robocse: Robot common sense embedding. In *2019 International Conference on Robotics and Automation (ICRA)*, 2019.

[11] Angel Daruna, Lakshmi Velayudhan Nair, Weiyu Liu, and Sonia Chernova. Towards robust one-shot task execution using knowledge graph embeddings. In *International Conference on Robotics and Automation (ICRA)*, 2021.

[12] Thanh-Toan Do, Anh Nguyen, and Ian Reid. Affordancenet: An end-to-end deep learning approach for object affordance detection. In *2018 IEEE international conference on robotics and automation (ICRA)*, pages 5882–5889. IEEE, 2018.

[13] Zackory Erickson, Nathan Luskey, Sonia Chernova, and Charles C Kemp. Classification of household materials via spectroscopy. *IEEE Robotics and Automation Letters*, 4(2):700–707, 2019.

[14] Mustafa Ersen, Erhan Oztop, and Sanem Sariel. Cognition-enabled robot manipulation in human environments: Requirements, recent work, and open problems. *IEEE Robotics & Automation Magazine*, 2017.

[15] Ankit Goyal, Arsalan Mousavian, Chris Paxton, Yu-Wei Chao, Brian Okorn, Jia Deng, and Dieter Fox. Ifor: Iterative flow minimization for robotic object rearrangement. *arXiv preprint arXiv:2202.00732*, 2022.

[16] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335–346, 1990.

[17] Séverin Lemaignan, Raquel Ros, Lorenz Mösenlechner, Rachid Alami, and Michael Beetz. Oro, a knowledge management platform for cognitive architectures in robotics. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 3548–3553. IEEE, 2010.

[18] Ian Lenz, Honglak Lee, and Ashutosh Saxena. Deep learning for detecting robotic grasps. *The International Journal of Robotics Research*, 34(4-5):705–724, 2015.

[19] Weiyu Liu, Angel Daruna, and Sonia Chernova. A survey of semantic reasoning frameworks for robotic systems. *Under review at Robotics and Autonomous Systems*.

[20] Weiyu Liu, Angel Daruna, and Sonia Chernova. Cage: Context-aware grasping engine. In *International Conference on Robotics and Automation (ICRA)*, 2020.

[21] Weiyu Liu, Angel Daruna, Zsolt Kira, and Sonia Chernova. Path ranking with attention to type hierarchies. In *Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[22] Weiyu Liu, Dhruva Bansal, Angel Daruna, and Sonia Chernova. Learning Instance-Level N-Ary Semantic Knowledge At Scale For Robots Operating in Everyday Environments. In *Proceedings of Robotics: Science and Systems*, 2021.

[23] Weiyu Liu, Chris Paxton, Tucker Hermans, and Dieter Fox. Structformer: Learning spatial structure for language-guided semantic rearrangement of novel objects. In *2022 International Conference on Robotics and Automation (ICRA)*, 2022.

[24] Jeffrey Mahler, Jacky Liang, Sherdil Niyaz, Michael Laskey, Richard Doan, Xinyu Liu, Juan Aparicio Ojea, and Ken Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. *Robotics: Science and Systems (RSS)*, 2017.

[25] Lucas Manuelli, Wei Gao, Peter Florence, and Russ Tedrake. kpam: Keypoint affordances for category-level robotic manipulation. In *ISRR*, 2019.

[26] Oier Mees, Alp Emek, Johan Vertens, and Wolfram Burgard. Learning object placements for relational instructions by hallucinating scene representations. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 94–100. IEEE, 2020.

[27] George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

[28] Thao Nguyen, Nakul Gopalan, Roma Patel, Matt Corsaro, Ellie Pavlick, and Stefanie Tellex. Affordance-based robot object retrieval. *Autonomous Robots*, 46(1):83–98, 2022.

[29] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2015.

[30] David Paulius and Yu Sun. A survey of knowledge representation in service robotics. *Robotics and Autonomous Systems*, 118:13–30, 2019.

[31] Chris Paxton, Chris Xie, Tucker Hermans, and Dieter Fox. Predicting stable configurations for semantic placement of novel objects. In *Conference on Robot Learning (CoRL)*, 2021.

[32] Andrzej Pronobis and Patric Jensfelt. Large-scale semantic mapping and reasoning with heterogeneous modalities. In *Robotics and Automation (ICRA), 2012 IEEE International Conference on*, pages 3515–3522. IEEE, 2012.

[33] Ahmed Qureshi, Arsalan Mousavian, Chris Paxton, Michael Yip, and Dieter Fox. Nerp: Neural rearrangement planning for unknown objects. In *Proceedings of Robotics: Science and Systems*, 2021.

[34] Matthew Richardson and Pedro Domingos. Markov logic networks. *Machine learning*, 62(1-2):107–136, 2006.

[35] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K Misra, and Hema S Koppula. Robobrain: Large-scale knowledge engine for robots. *arXiv preprint arXiv:1412.0691*, 2014.

[36] Mohit Shridhar, Dixant Mittal, and David Hsu. Ingress: Interactive visual grounding of referring expressions. *The International Journal of Robotics Research*, 39(2-3):217–232, 2020.

[37] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022.

[38] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence*, 2017.

[39] Stefanie Tellex, Nakul Gopalan, H. Kress-Gazit, and Cynthia Matuszek. Robots that use language. *Annual review of control, robotics, and autonomous systems*, (3), 2020.

[40] Moritz Tenorth and Michael Beetz. Representations for robot knowledge in the knowrob framework. *Artificial Intelligence*, 247:151–169, 2017.

[41] Moritz Tenorth, Georg Bartels, and Michael Beetz. Knowledge-based specification of robot motions. In *ECAI*, pages 873–878, 2014.

[42] Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond Mooney. Learning multi-modal grounded linguistic semantics by playing "I spy". In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2016.

[43] Hongtao Wu, Jikai Ye, Xin Meng, Chris Paxton, and Gregory Chirikjian. Transporters with visual foresight for solving unseen rearrangement tasks. *arXiv preprint arXiv:2202.10765*, 2022.

[44] Danfei Xu, Suraj Nair, Yuke Zhu, Julian Gao, Animesh Garg, Li Fei-Fei, and Silvio Savarese. Neural task programming: Learning to generalize across hierarchical tasks. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1–8. IEEE, 2018.

[45] Ruinian Xu, Fu-Jen Chu, Chao Tang, Weiyu Liu, and Patricio Vela. An affordance keypoint detection network for robot manipulation. *IEEE Robotics and Automation Letters*, pages 1–1, 2021. doi: 10.1109/LRA.2021.3062560.

[46] Yuke Zhu, Alireza Fathi, and Li Fei-Fei. Reasoning about object affordances in a knowledge base representation. In *European conference on computer vision*, pages 408–424. Springer, 2014.