Towards Broad Generalization in Robotic Manipulation: Data, Algorithms, and Supervision

Suraj Nair
Stanford University

From general object grasping [13] to in-hand manipulation [20], learning has enabled a number of exciting robotic manipulation capabilities in recent years. Despite this, the quintessential home robot that can enter a previously unseen home environment and complete a wide range of tasks like humans can is far from a reality. While there are many problems to solve in accomplishing this goal, one of the central bottlenecks lies in learning control policies that can *generalize to new tasks, objects, and environments*. For example, a robot cooking in a home cannot afford to re-learn from scratch for each new dish, nor is it feasible to hard-code state features for every new kitchen a robot might encounter.

One potential route to accomplishing this generalization is to train the robot on a wide distribution of data that contains many tasks, objects, and environments. Indeed, this recipe of large, diverse datasets combined with scalable offline learning algorithms (e.g. self-supervised or cheaply supervised learning) has been the key behind recent successes in NLP [7, 2] and vision [6, 21]. However, directly extending this recipe to robotics is nontrivial, as we neither have sufficiently large and diverse datasets of robot interaction, nor is it obvious what types of learning algorithms or sources of supervision can enable us to scalably learn useful skills from these datasets.

The goal of my research lies in tackling these challenges, and *replicating the recipe of large-scale data and learning in the context of robotic manipulation*. Towards this goal my research has focused on three key questions. **First**, how do we scalably collect large and diverse datasets of robots interacting in the physical world? **Second,** how do we design self-supervised reinforcement learning algorithms that can consume such broad data, which may come from non-experts and lack reward labels, and from it learn to reach unseen goals. **Third**, how might we unlock the broad sources of data that exist on the web, like videos of humans and natural language to enable more effective learning in our robots?

Scaling Robotic Manipulation Datasets. The first challenge lies in acquiring large amounts of useful robotic interaction data, a particularly difficult problem in robotic manipulation. One approach to tackling this is having humans explicitly collect meaningful interaction via teleoperation [14, 15, 12], however this can be difficult to do at scale due to the immense burden it places on the human operator.

As an alternative to human collected data, in our first paper on the topic we collected a large, 15 million frame dataset of robot interaction through *autonomous data collection* [5]. This dataset spanned multiple universities and robots, and

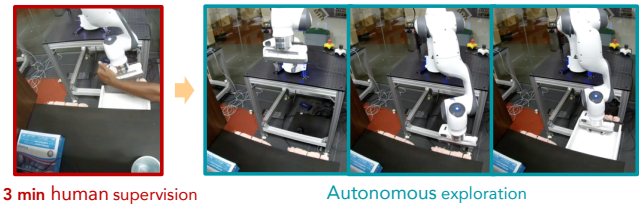

Fig. 1. **Weakly-Supervised Autonomous Data Collection.** Using just a few minutes of human supervision, our method is able to guide robotic exploration towards more meaningful interactions..

in our experiments we found that pre-training on it enabled significantly faster adaptation to a completely unseen robot and scene. However, a limitation of this work is that due to being fully unsupervised, the data contained primarily object picking/pushing, but lacked more interesting behaviors.

Motivated by this, in our follow up work [4] we studied how to balance this tradeoff between (a) the scalability of autonomous collection with (b) a prior over meaningful interaction from human supervision. Our key insight is that by leveraging some *weak human supervision*, we can allow the agent to focus on semantically relevant parts of the state space, greatly accelerating the collection of useful data while still keeping the data collection process fully autonomous. Specifically, a human can communicate a prior over relevant states by simply capturing a handful of images of "interesting" states ahead of time, which a learning-based agent can then use to guide their exploration. Using this approach, our proposed algorithm interacts more than twice as often with relevant objects than prior state-of-the-art unsupervised exploration methods, and as a result collects higher-quality data enabling better downstream, task performance. With less than 3 minutes of human supervision, it can collect high quality data autonomously for multiple days on hard exploration problems from pixels on a Franka Emika Panda robot (See Figure 1).

Self-Supervised Reinforcement Learning Algorithms. Given a large and diverse dataset of robotic interaction, the second challenge lies in effectively learning behavior from such data, such that the learned agent can generalize to a large number of possibly unseen tasks. Towards this goal, a large portion of my prior work has studied *self-supervised offline reinforcement learning*, where given a non-expert dataset of interaction, the agent aims to learn a visuomotor policy $\pi(s, g)$ that can reach goals g, specified by images. Notably, such self-supervised methods do not need reward labels and can consume non-expert data, making them ideal candidates to learn from broad, pre-collected datasets. One such algorithm

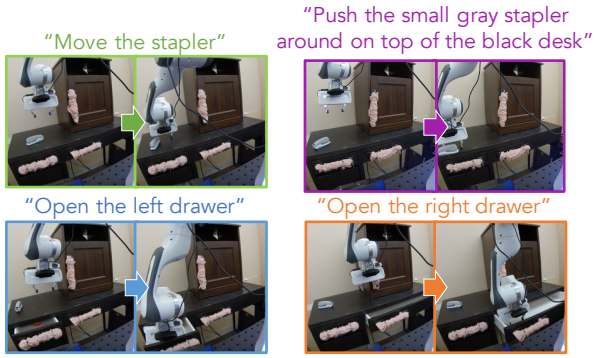

Fig. 2. In [18] we learn language-conditioned visuomotor policies using sub-optimal offline data and crowd-sourced annotation, enabling a real robot to complete natural language specified tasks.

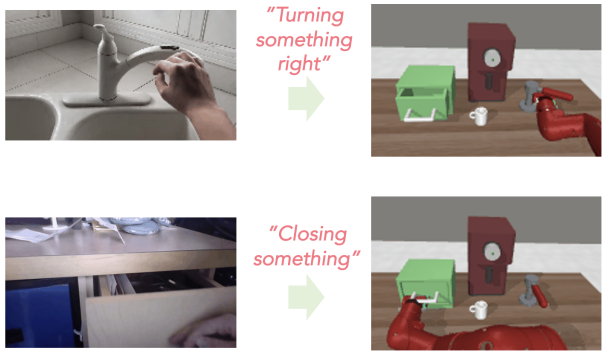

Fig. 3. In [3] we learn a video-conditioned reward function on diverse videos that (1) enables specifying tasks with a human video and (2) generalizes to unseen environments by training on diverse videos.

is visual foresight [9, 8], which takes a model-based planning approach to this problem, learning a visual dynamics model, and performs model-predictive control (MPC) to plan to reach goals. Towards improving such algorithms, my prior work has included a number of techniques for learning better visual dynamics models, including conditioning them on goals to better model goal relevant quantities [17], and designing better architectures and training procedures for video prediction models [23, 1]. Moreover, several of my prior works have studied how to extend the visual foresight framework to handle more challenging, long-horizon tasks. For example in [16], we studied how we could learn a generative model over images in a self-supervised way, and use it to perform collocation based planning in visual space, solving long-horizon tasks and generating visually interpretable subgoals. In [22] we extended this line of work by learning *functional distances* using self-supervised goal-conditioned Q-learning, and using the learned Q function as a cost function for model-predictive control. Finally, in our most recent follow up work [24], we learn individual skills using an approach similar to [22], then combine these skills with a symbolic planner to complete long-horizon planning problems with over 250 steps.

Supervising Robot Learning with Language and Video from the Web. While this recipe of autonomous data collection + self-supervised offline reinforcement learning can scale well with size and diversity of data, it has some important limitations. Critically, without any human supervision, it is forced to model *everything* about a scene (including task-irrelevant features), ultimately limiting performance. Capturing task relevance requires some supervision, and a focus of my recent work is on how we can supervise task relevance *in a scalable way*, by leveraging data that can easily be collected through the web. In my recent work [18], we found that simply by using crowd-sourcing and having annotators describe in natural language what was happening in a video of the robot, we could learn reward functions for learning language-conditioned control (See Figure 2). Moreover, this work showed that with crowd-sourced language we can achieve significantly improved performance over fully self-supervised approaches, with limited impact on scalability.

While the prior work I have described is able to consume large amounts of robot data and generalize to new goals, it is still fundamentally limited by the amount of robot data we can collect. Alternatively, there exists vast amounts of pre-collected video data of humans interacting in semantically interesting ways in their environments [10, 11]. This data is diverse, spanning people and environments across the globe, and tasks ranging from assembling objects to doing the dishes. While the embodiment in these videos is different than that of our robots, if we could leverage this data in robot learning it could greatly boost generalization to unseen tasks, environments, and objects. We study exactly this in our recent work [3], where we learn reward functions on diverse human video data [10], and measure how well the learned reward functions generalize to held out tasks and environments. We observed that by training with diverse human videos, the learned reward performed over 20% more effectively on unseen environments and tasks, suggesting that diverse human videos can be a promising path towards broader generalization (See Figure 3). In my most recent work [19], we pushed on this direction further, pre-training a visual representation on human video data, and showing that this representation enabled more data efficient learning of downstream robotic manipulation tasks.

Future Work. In my ongoing and future research I aim to scale up along all three directions. We still lack large and diverse robot datasets, and in an ongoing project we aim to bring robots into real homes, collecting a robot dataset that is truly reflective of the environments humans operate in. Furthermore, we continue to push on scaling up our offline learning algorithms, specifically towards the goal of a single large model that can consume all of the robot data we have collected, simulated robot data, videos of humans, and natural language, and produce an agent which can perform tasks effectively while also generalizing broadly.

Lastly, a number of other important problems remain towards the goal of generalist robots. First, even with generalizable policy trained on large amounts of data, such a model needs to (1) be safe in the new environment and avoid catastrophic failures, and (2) learn quickly and adapt from any mistakes. Safety during deployment, and the ability to quickly adapt a policy given a few corrections are an important component of my future work, and I believe this direction will be critical to making any robot trained on large scale data deployable in the real world.

References

[1] Mohammad Babaeizadeh, Mohammad Taghi Saffar, Suraj Nair, Sergey Levine, Chelsea Finn, and D. Erhan. Fitvid: Overfitting in pixel-level video prediction. *ArXiv*, abs/2106.13195, 2021.

[2] Tom B. Brown et al. Language models are few-shot learners. *arXiv:2005.14165*, 2020.

[3] Annie S. Chen, Suraj Nair, and Chelsea Finn. Learning generalizable robotic reward functions from "in-the-wild" human videos. *ArXiv*, abs/2103.16817, 2021.

[4] Annie S. Chen, HyunJi Nam, Suraj Nair, and Chelsea Finn. Batch exploration with examples for scalable robotic reinforcement learning. *IEEE Robotics and Automation Letters*, 2021.

[5] Sudeep Dasari, Frederik Ebert, Stephen Tian, Suraj Nair, Bernadette Bucher, Karl Schmeckpeper, Siddharth Singh, Sergey Levine, and Chelsea Finn. Robonet: Large-scale multi-robot learning. In *Conference on Robot Learning*, 2019.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[8] Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv:1812.00568*, 2018.

[9] Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2017.

[10] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5842–5850, 2017.

[11] Kristen Grauman et al. Ego4D: Around the World in 3,000 Hours of Egocentric Video, 2021.

[12] Eric Jang, Alex Irpan, Mohi Khansari, Daniel Kappler, Frederik Ebert, Corey Lynch, Sergey Levine, and Chelsea Finn. Bc-z: Zero-shot task generalization with robotic imitation learning. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 991–1002. PMLR, 08–11 Nov 2022.

[13] Dmitry Kalashnikov, Alex Irpan, Peter Pastor, Julian Ibarz, Alexander Herzog, Eric Jang, Deirdre Quillen, Ethan Holly, Mrinal Kalakrishnan, Vincent Vanhoucke, et al. Scalable deep reinforcement learning for vision-based robotic manipulation. In *Conference on Robot Learning*, pages 651–673. PMLR, 2018.

[14] Ajay Mandlekar, Yuke Zhu, Animesh Garg, Jonathan Booher, Max Spero, Albert Tung, Julian Gao, John Emmons, Anchit Gupta, Emre Orbay, Silvio Savarese, and Li Fei-Fei. Roboturk: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, 2018.

[15] Ajay Mandlekar, Jonathan Booher, Max Spero, Albert Tung, Anchit Gupta, Yuke Zhu, Animesh Garg, Silvio Savarese, and Li Fei-Fei. Scaling robot supervision to hundreds of hours with roboturk: Robotic manipulation dataset through human reasoning and dexterity. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1048–1055. IEEE, 2019.

[16] Suraj Nair and Chelsea Finn. Hierarchical foresight: Self-supervised learning of long-horizon tasks via visual subgoal generation. *ArXiv*, abs/1909.05829, 2020.

[17] Suraj Nair, Silvio Savarese, and Chelsea Finn. Goal-aware prediction: Learning to model what matters. *ArXiv*, abs/2007.07170, 2020.

[18] Suraj Nair, Eric Mitchell, Kevin Chen, Brian Ichter, Silvio Savarese, and Chelsea Finn. Learning language-conditioned robot behavior from offline data and crowd-sourced annotation. In *CoRL*, 2021.

[19] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. *arXiv preprint arXiv:2203.12601*, 2022.

[20] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafal Jozefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation, 2019.

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[22] Stephen Tian, Suraj Nair, Frederik Ebert, Sudeep Dasari, Benjamin Eysenbach, Chelsea Finn, and Sergey Levine. Model-based visual planning with self-supervised functional distances. In *International Conference on Learning Representations*, 2021.

[23] Bo-Han Wu, Suraj Nair, Roberto Martín-Martín, Li Fei-Fei, and Chelsea Finn. Greedy hierarchical variational autoencoders for large-scale video prediction. *ArXiv*, abs/2103.04174, 2021.

[24] Bohan Wu, Suraj Nair, Li Fei-Fei, and Chelsea Finn. Example-driven model-based reinforcement learning for solving long-horizon visuomotor tasks. *ArXiv*, abs/2109.10312, 2021.