

Deep Imitation Learning for Complex Manipulation Tasks from Virtual Reality Teleoperation

Tianhao Zhang*, Zoe McCarthy* , Owen Jow , Dennis Lee , Xi Chen, Ken Goldberg , Pieter Abbeel

Presenter: Muhammad Muaz

October 11, 2022

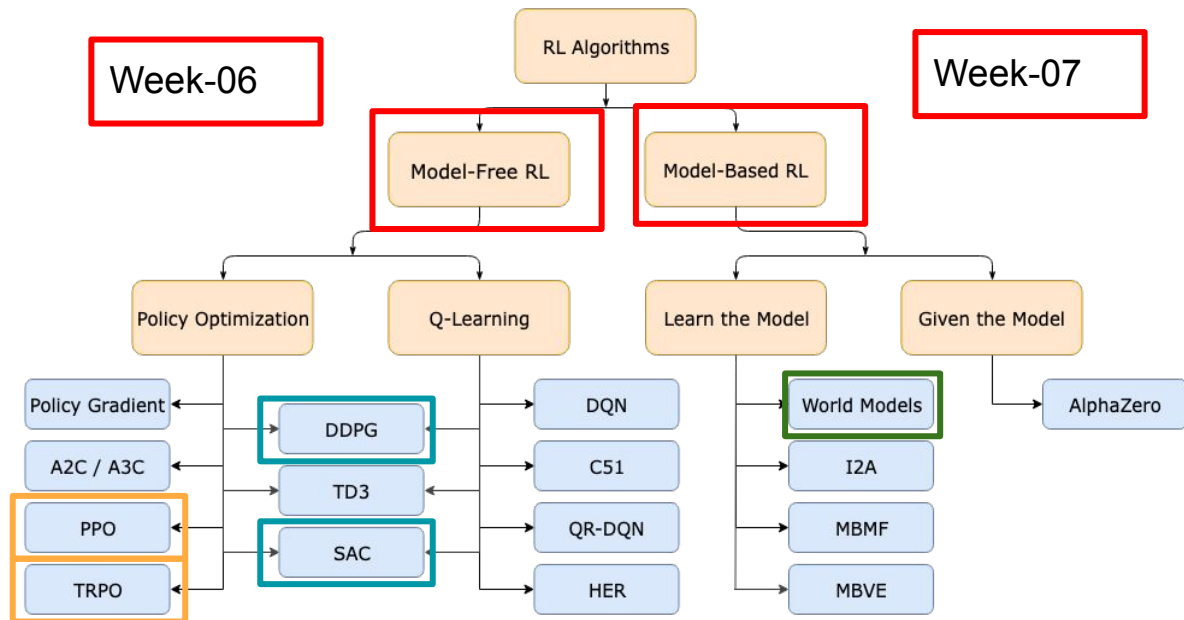
* = *equal contribution*

Talk's Overview

- Recap & Overview of Imitation Learning (Motivation)
- Problem under consideration (Behavior Cloning)
- Related Works
- Proposed Approach
- Experimental Analysis
- Discussion, Limitations, Future Work

What we have seen so far in RL!

- Model-free reinforcement learning
 - TRPO
 - PPO
 - DDPG
 - SAC
- Model-based reinforcement learning
 - Dreamer (World Models)
- Batch (Offline) RL
 - BCQ
 - CQL



[Source: Open AI Spinning Up]

What's next !

- Imitation Learning (IL)

(In a nutshell)

Provided: Expert Demonstrations or Demonstrator

Goal: Learn a policy that mimics the behaviour of demonstrator

- But why do we need IL ?

- Safety Concerns
- Learning by Interaction is very costly in time in real world
- Learn societal norms
- Easier to provide demonstrations than formulate reward function

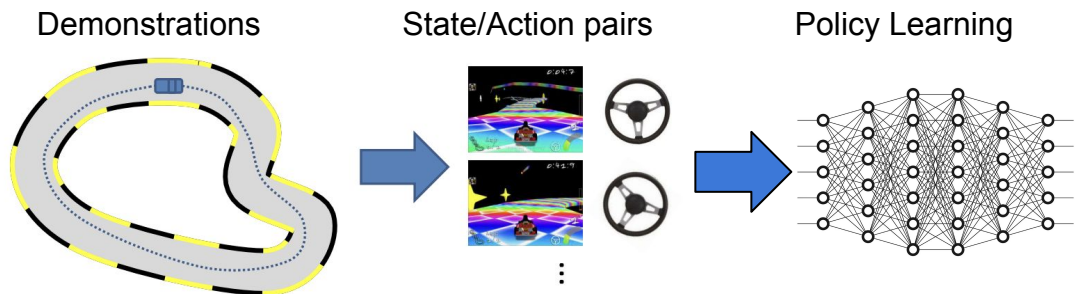
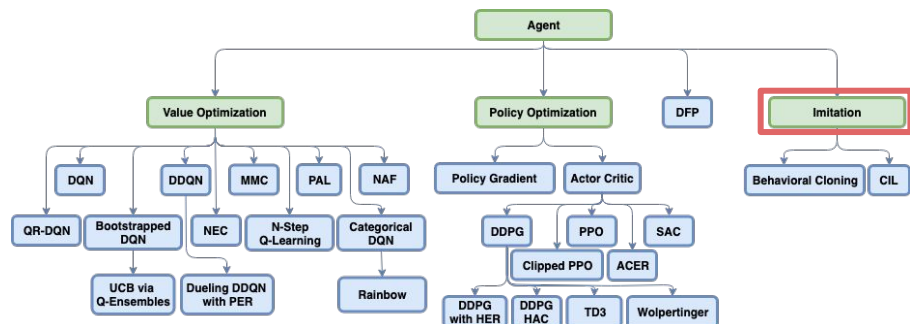


Image sources:

<https://intellabs.github.io/coach/components/agents/index.html>

https://www.cs.cmu.edu/~sross1/publications/ross_phdthesis.pdf

Imitation Learning (Ctd.)

Autonomous Driving



ALVINN [Dean Pomerleau et al., 1989-1999]

Helicopter Acrobatics



An Application of Reinforcement Learning to Aerobatic Helicopter Flight
Pieter Abbeel, Adam Coates, Morgan Quigley, Andrew Y. Ng, NIPS 2006

Problem Formulation & Overview

More **formally**:

Given:

- State space \mathcal{S}_t , Action space \mathcal{U}_t
- Transition model $P(\mathbf{s}'|\mathbf{s},\mathbf{u})$
- No Reward function R
- Set of one or more teacher's demonstrations $(\mathbf{o}_0, \mathbf{u}_0, \mathbf{o}_1, \mathbf{u}_1, \dots)$
(actions drawn from teacher policy π^*)

Goal:

- ❖ **Behavioral Cloning**
 - Can we directly learn expert's policy using supervised learning?
- ❖ **Inverse Optimal Control / RL**
 - Can we recover R and then recover π^* ?


Problem Formulation & Overview (Ctd.)

More **formally**:

Given:

- State space \mathcal{S}_t , Action space \mathcal{U}_t
- Transition model $P(\mathbf{s}'|\mathbf{s},\mathbf{u})$
- No Reward function R
- Set of one or more teacher's demonstrations $(\mathbf{o}_0, \mathbf{u}_0, \mathbf{o}_1, \mathbf{u}_1, \dots)$
(actions drawn from teacher policy π^*)

Goal:

- ❖ **Behavioral Cloning**  **Topic of Today's paper**
 - Can we directly learn expert's policy using supervised learning?
- ❖ Inverse Optimal Control / RL
 - Can we recover R and then recover π^* ?

Behavioral Cloning

Given:

- State space \mathcal{S}_t , Action space \mathcal{U}_t
- Transition model $P(\mathbf{s}'|\mathbf{s},\mathbf{u})$
- No Reward function R
- Set of one or more teacher's demonstrations $\xi = \{(\mathbf{o}_0, \mathbf{u}_0), (\mathbf{o}_1, \mathbf{u}_1), \dots\}$
(actions drawn from teacher policy π^*)

Formulate problem as a standard supervised learning problem to learn $\pi_\theta(\mathbf{u}_t|\mathbf{o}_t)$

- Fix learning algorithm class (e.g., DNN, SVM, Decision Trees etc.)
- Estimate a policy from demonstration set ξ

Optimization Objective

$$\hat{\pi}^* = \arg \min_{\pi} \sum_{\xi \in \Xi} \sum_{\mathbf{x} \in \xi} L(\pi(\mathbf{x}), \pi^*(\mathbf{x}))$$

Set of expert demonstrations

Minimize the loss between learned policy and expert's policy (loss can be KL divergence, p-norms, etc.)

Related Works

❖ Behavioral Cloning:

- Some prior works used low-dimensional representations of environment states [1,2,3,4]
Issue: Hard to extract environment state which makes learning *policies directly from raw pixels desirable* and it has been successful for tasks such as driving [5,6,7], drones [8]
- Kinesthetic teaching (human operator guides robot by force on its body for demonstrations) [9,10]
Issues: - unintuitive, visual artifacts, low quality
- Previous Teleoperation systems for robotic manipulation (da Vinci Surgical System)[11]
Pros: High quality demonstrations without any visual obstructions
Issues: Expensive, Specialized

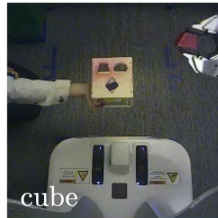


Image sources: <https://tinyurl.com/3ycj87zp>
<https://www.roboticstoday.com/devices/da-vinci-s-hd-surgical-system>

Related Works (Ctd.)

- ❖ Demonstrations from Trajectory Optimization Approaches [12,13,14,15]
Issues: Time-consuming even for experts

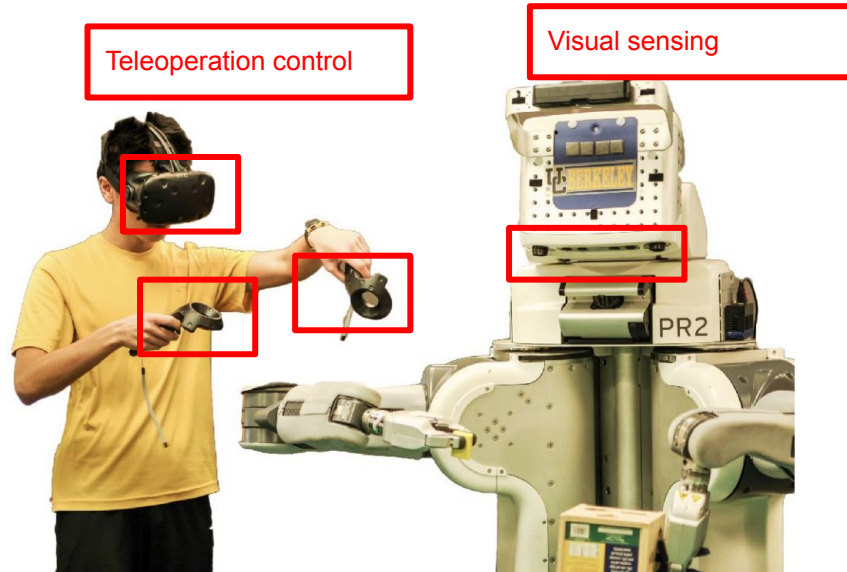
- ❖ Reinforcement Learning for skill acquisition:
Pros: Recent success in learning policies from pixels to actions [16,17,18]
Issues: > Impractical amount of exploration needed for real robots
(Example: Atari results would have taken **40 days** of real-time experience)
> Hard to specify reward functions in practice.



Proposed Approach

Hardware:

- ❖ Use Vive VR System, a consumer-grade VR device for teleoperation
Pros: Very cost-effective \$600
- ❖ Provides headset + 2 hand-controllers (each having 6DoF; provides sub-millimeter pose tracking at 90 Hz in room-scale tracking area)
- ❖ Visual Sensing: Primesense Carmine 1.09
Low cost 3D camera mounted on robot's head and captures RGB-D images at 30 Hz



Visual Reality Teleoperation in action

Proposed Approach

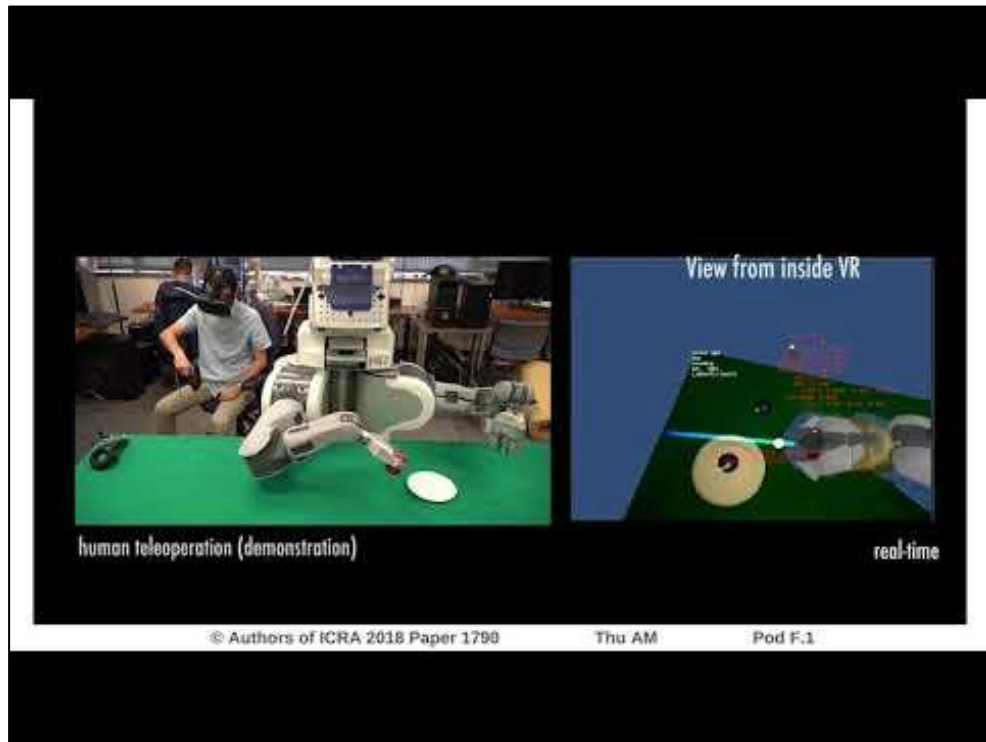
Visual Interface:

- ❖ Use RGB-D images to render coloured point cloud, processed to remove gaps between points as physical objects in virtual environment.
- ❖ Overlay 3D visualizations on the point cloud to assist teleoperation process.

Control Interface:

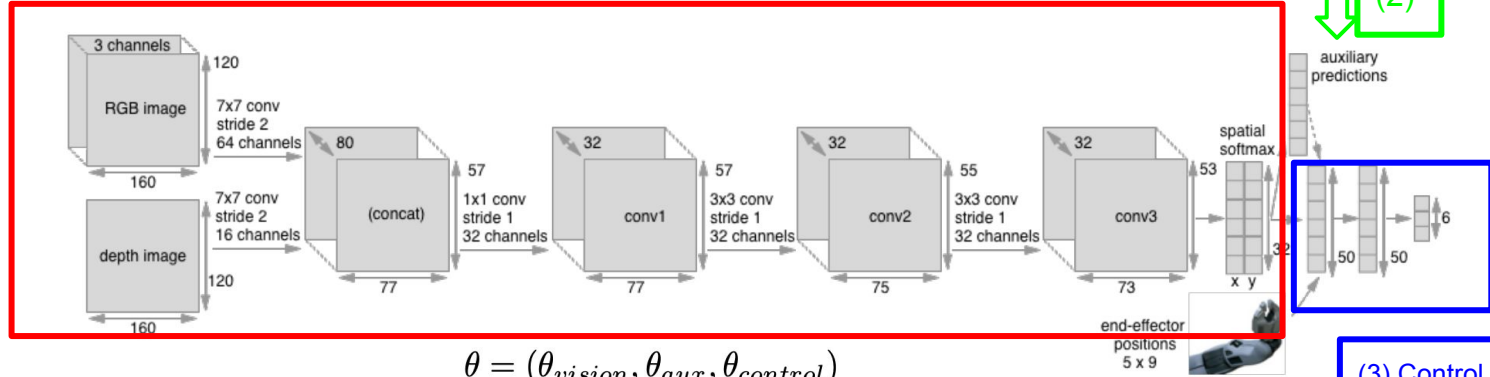
- ❖ Use Vive hand-controller's pose as robot's gripper target pose.
- ❖ Natural control mechanism to provide demonstrations.
- ❖ Intuitive way to apply force control (if robot's gripper is hindered during contact, then $| \text{target pose} - \text{gripper pose} | \propto \text{force exerted by gripper}$)

Benefit: Allow human operator to dynamically vary force



Network Architecture

(1) Vision network:
extract spatial features



(2)

(3) Control output network

❖ **Inputs:**

➤ $o_t = (I_t, D_t, p_{t-4:t})$

I_t : Current RGB image of dimensions (160 x 120 x 3)

D_t : Current Depth image (dim: 160 x 120)

$p_{(t-4:5)}$: three points on end-effector of right arm to capture pose for last 5 steps (# of dim : 45)

❖ **Outputs:**

➤ Angular velocity (# of dim: 3)

➤ Linear Velocity (# of dim: 3)

➤ Gripper open/close state $\epsilon \{0, 1\}$ (for tasks involving grasping)

Loss Functions

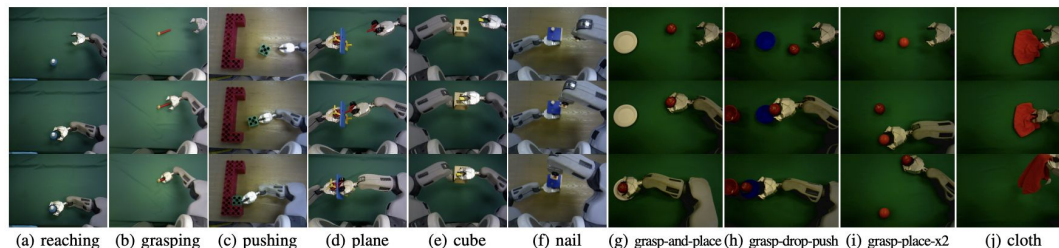
- ❖ Behavioral cloning loss : $\mathcal{L}_{l2} = \|\pi_\theta(o_t) - u_t\|_2^2, \quad \mathcal{L}_{l1} = \|\pi_\theta(o_t) - u_t\|_1$
- ❖ Directional alignment loss : $\mathcal{L}_c = \arccos\left(\frac{u_t^T \pi_\theta(o_t)}{\|u_t\| \|\pi_\theta(o_t)\|}\right)$
- ❖ Sigmoid cross-entropy loss :
(for gripper open/close prediction) $\mathcal{L}_g = g_t \log(\sigma(\hat{g}_t)) - (1 - g_t) \log(1 - \sigma(\hat{g}_t))$
- ❖ Auxiliary loss : $\mathcal{L}_{aux}^{(a)} = \|\text{NN}(f_t; \theta_{aux}^{(a)}) - s_t^{(a)}\|_2^2$
 - Extra source of supervision; making learning goal-oriented; auxiliary labels can be inferred from demonstrations.
 - Examples of auxiliary tasks:
 - Predict current gripper pose
 - Predict final gripper pose
 - Predict current object position (for pushing, grasp-and-place, grasp-drop-push tasks)
- ❖ Overall loss function:

$$\mathcal{L}(\theta) = \lambda_{l2} \mathcal{L}_{l2} + \lambda_{l1} \mathcal{L}_{l1} + \lambda_c \mathcal{L}_c + \lambda_g \mathcal{L}_g + \lambda_{aux} \sum_a \mathcal{L}_{aux}^{(a)}$$

Network Training: Use Stochastic Gradient Descent to train policy using random batches sampled from demonstrations

Experiments

- ❖ Evaluations done on suite of 10 challenging **manipulation** tasks
 - i. Reach a bottle, grasp a tool, push a toy block, attach wheels to a toy plane etc., to name a few.
- ❖ Evaluation criteria/goals:
 - i. Can we use our system to train, with little tuning, successful deep visuomotor policies for a range of challenging manipulation tasks?
 - ii. What is the sample complexity for learning an example manipulation task using our system?
 - iii. Does our auxiliary prediction loss improve data efficiency for learning real-world robotic manipulation?



Experiments (Ctd.)

- ❖ Can we use our system to train, with little tuning, successful deep visuomotor policies for a range of challenging manipulation tasks?

Success rates of learned policies averaged across all initial states during test time

| task | reaching | grasping | pushing | plane | cube | nail | grasp-and-place | grasp-drop-push | grasp-place-x2 | cloth |
|-----------------------|----------|----------|---------|-------|-------|-------|-----------------|-----------------|----------------|-------|
| test | 91.6% | 97.2% | 98.9% | 87.5% | 85.7% | 87.5% | 96.0% | 83.3% | 80% | 97.4% |
| demo time (min) | 13.7 | 11.1 | 16.9 | 25.0 | 12.7 | 13.6 | 12.3 | 14.5 | 11.6 | 10.1 |
| avg length (at 10 Hz) | 41 | 37 | 58 | 47 | 37 | 38 | 68 | 87 | 116 | 60 |
| #demo | 200 | 180 | 175 | 319 | 206 | 215 | 109 | 100 | 60 | 100 |

Statistics of training data

- ❖ The results suggest that a simple imitation learning can train successful control policies for range of real-world manipulation tasks while achieving sample efficiency and good performance
- ❖ Each policy is trained with same hyperparameter settings and NN architecture and uses <30 min of human demonstrations as training data.

Experiments (Ctd.)

- ◆ Moreover, paper mentions the learned policies were able to complete sequence of maneuvers in long running tasks which shows that policies learned how to transition from one skill to another.

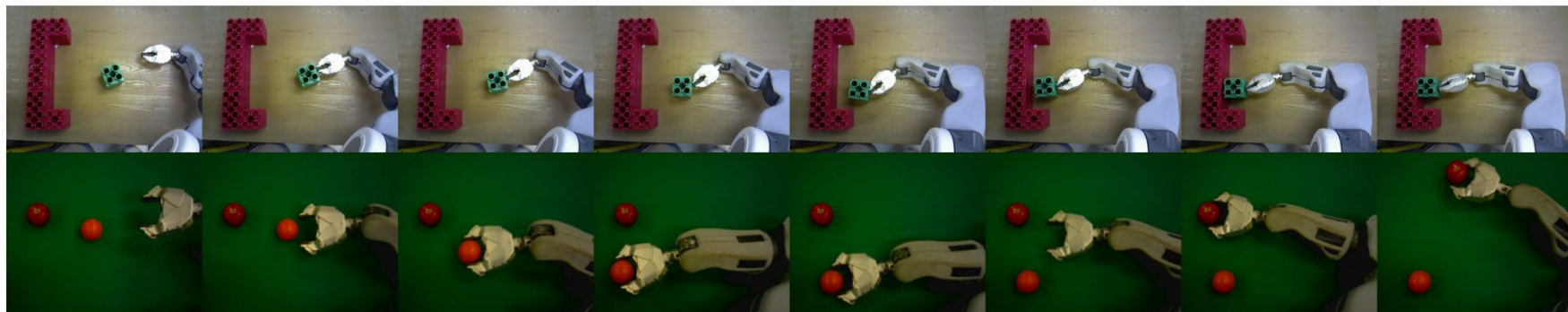


Fig. 6: Example successful trials of the learned policies during evaluation (top: pushing; bottom: grasp-place x2)

Experiments (Ctd.)

- ❖ What is the sample complexity for learning an example manipulation task using our system?

TABLE II: Success rates of policies trained using different numbers of demonstrations for the nail task

| task: nail | | |
|--------------------------|--------------------------------------|---------------|
| number of demonstrations | demonstration time (estimated) (min) | success rates |
| 193 | 12.2 | 88.9% |
| 115 | 7.3 | 77.8% |
| 67 | 4.2 | 50% |

- ❖ Only ~5 minutes of human demonstrations was needed to achieve 50% success for the nail task

Experiments (Ctd.)

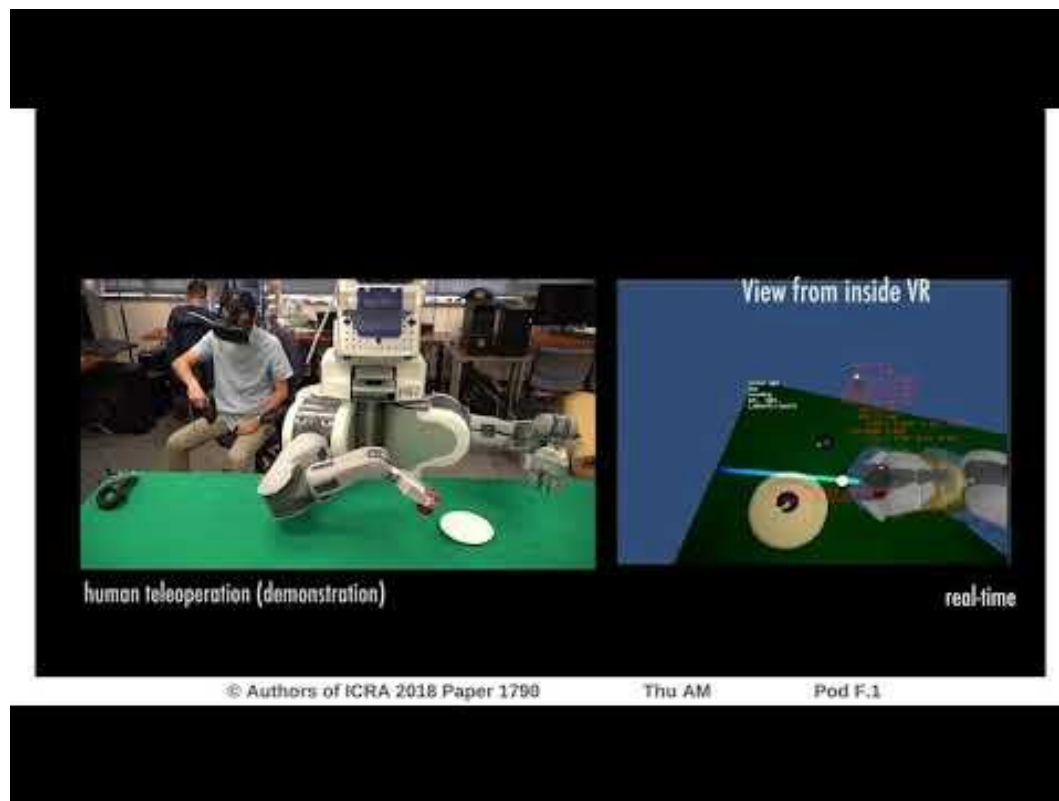
- ❖ Does our auxiliary prediction loss improve data efficiency for learning real-world robotic manipulation?

TABLE III: Comparison of policy when trained with and without auxiliary prediction loss on the grasp-and-place task.

| task: grasp-and-place | | |
|--------------------------|----------------------|-------------------------|
| number of demonstrations | success rates (with) | success rates (without) |
| 109 | 96% | 80% |
| 55 | 53% | 26% |
| 11 | 28% | 20% |

- ❖ Observation: Auxiliary losses empirically improves data efficiency

Learned Policy Visualizations



Critique / Limitations / Open Issues

- ❖ Expert demonstrations **will not** be sampled uniformly across the entire state space.
- ❖ **Distributional mismatch** between training and testing policies due to **covariance shift** - which can be chaotic !

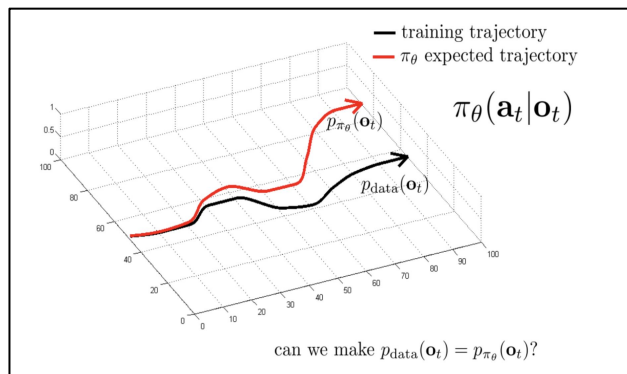
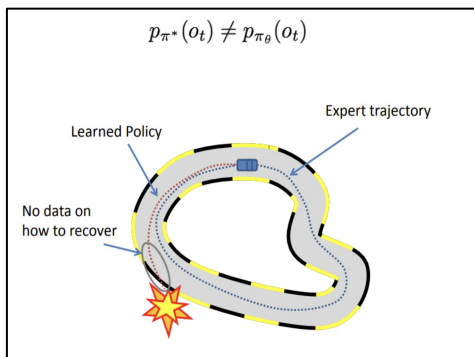
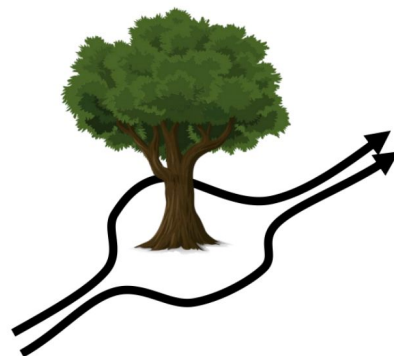


Image sources: <https://tinyurl.com/5xeda32m>

Critique / Limitations / Open Issues (Ctd.)

- ❖ What happens in case expert exhibits **multi-modal behaviour**?



- ❖ For some tasks, humans are not good at providing actions which can lead to **suboptimal policy** learning

Future Directions

- ❖ Can we learn good policies from demonstrations provided in videos?
 - Petr'ik, V., Tapaswi, M., Laptev, I., & Sivic, J. (2020). Learning Object Manipulation Skills via Approximate State Estimation from Real Videos. *CoRL*.
- ❖ Can we train a robot while only using low-level controllers that resembles robot?
 - Kim, H., Ohmura, Y., Nagakubo, A., & Kuniyoshi, Y. (2022). Training Robots without Robots: Deep Imitation Learning for Master-to-Robot Policy Transfer. *ArXiv, abs/2202.09574*.
- ❖ Can we use imitation learning to learn vision-based manipulation policies on new novel tasks?
 - Jang, E., Irpan, A., Khansari, M., Kappler, D., Ebert, F., Lynch, C., Levine, S., & Finn, C. (2021). BC-Z: Zero-Shot Task Generalization with Robotic Imitation Learning. *CoRL*.

Extended Readings

- ❖ Learning from a single demonstration
 - Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017). One-Shot Visual Imitation Learning via Meta-Learning. *CoRL*.
- ❖ Liu, Y., Gupta, A., Abbeel, P., & Levine, S. (2018, May). Imitation from observation: Learning to imitate behaviors from raw video via context translation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)* (pp. 1118-1125). IEEE.
- ❖ Finn, C., Yu, T., Zhang, T., Abbeel, P., & Levine, S. (2017, October). One-shot visual imitation learning via meta-learning. In *Conference on robot learning* (pp. 357-368). PMLR.
- ❖ Suomalainen, M., Karayiannidis, Y., & Kyrki, V. (2022). A Survey of Robot Manipulation in Contact. *Robotics Auton. Syst.*, 156, 104224.
- ❖ Torabi, F., Warnell, G., & Stone, P. (2018). Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*.

Summary

- ❖ Can we use consumer-grade VR device to teleoperate robots for complex manipulation tasks?
- ❖ Control interfaces exist for driving cars/drones, but what about manipulation tasks in robots?
Kinesthetic teaching introduces visual artifacts in vision based tasks?
- ❖ - Developed cost-effective, consumer-grade teleoperation control system
 - Single deep neural network architecture able to perform well on a suite of 10 complex manipulation tasks
 - Auxiliary loss besides behavioral cloning loss introduces self-supervision which provides sample efficiency
- ❖ The key takeaways from this paper was that it is easy to use commercial-grade VR devices to collect high-quality robot manipulation demonstrations suitable for visuomotor learning. Moreover, imitation learning can be surprisingly effective in learning deep policies that map pixel values to action only using small amount of data.

Questions?

References Used

1. A. Billard, Y. Epars, S. Calinon, S. Schaal, and G. Cheng, “Discovering optimal imitation strategies,” *Robotics and autonomous systems*, vol. 47, no. 2, pp. 69–77, 2004.
2. S. Schaal, J. Peters, J. Nakanishi, and A. Ijspeert, “Learning movement primitives,” *Robotics Research*, pp. 561–572, 2005.
3. P. Pastor, H. Hoffmann, T. Asfour, and S. Schaal, “Learning and generalization of motor skills by learning from demonstration,” in *Robotics and Automation, 2009. ICRA’09. IEEE International Conference on*. IEEE, 2009, pp. 763–768.
4. N. Ratliff, J. A. Bagnell, and S. S. Srinivasa, “Imitation learning for locomotion and manipulation,” in *Humanoid Robots, 2007 7th IEEE-RAS International Conference on*. IEEE, 2007, pp. 392–397.
5. D. A. Pomerleau, “Alvinn: An autonomous land vehicle in a neural network,” in *Advances in Neural Information Processing Systems*, 1989, pp. 305–313.
6. M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al., “End to end learning for self-driving cars,” *arXiv preprint arXiv:1604.07316*, 2016.
7. A. Giusti, J. Guzzi, D. C. Ciresan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro, et al., “A machine learning approach to visual perception of forest trails for mobile robots,” *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2016.
8. S. Ross, N. Melik-Barkhudarov, K. S. Shankar, A. Wendel, D. Dey, J. A. Bagnell, and M. Hebert, “Learning monocular reactive uav control in cluttered natural environments,” in *Robotics and Automation (ICRA), 2013 IEEE International Conference on*. IEEE, 2013, pp. 1765–1772.
9. B. Akgun, M. Cakmak, K. Jiang, and A. L. Thomaz, “Keyframe-based learning from demonstration,” *International Journal of Social Robotics*, vol. 4, no. 4, pp. 343–355, 2012.
10. A. Dragan, K. C. Lee, and S. Srinivasa, “Teleoperation with intelligent and customizable interfaces,” *Journal of Human-Robot Interaction*, vol. 1, no. 3, 2013
11. M. Talamini, K. Campbell, and C. Stanfield, “Robotic gastrointestinal surgery: early experience and system description,” *Journal of laparoendoscopic & advanced surgical techniques*, vol. 12, no. 4, pp. 225–232, 2002.

References Used

12. A. E. Bryson, Applied optimal control: optimization, estimation and control. CRC Press, 1975.
13. J. T. Betts, Practical methods for optimal control and estimation using nonlinear programming. SIAM, 2010
14. M. Posa, C. Cantu, and R. Tedrake, “A direct method for trajectory optimization of rigid bodies through contact,” The International Journal of Robotics Research, vol. 33, no. 1, pp. 69–81, 2014.
15. S. Levine and P. Abbeel, “Learning neural network policies with guided policy search under unknown dynamics,” in Advances in Neural Information Processing Systems, 2014, pp. 1071–1079
16. V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.
17. J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, “Trust region policy optimization,” in Proceedings of the 32nd International Conference on Machine Learning (ICML-15), 2015, pp. 1889–1897
18. S. Levine, C. Finn, T. Darrell, and P. Abbeel, “End-to-end training of deep visuomotor policies,” Journal of Machine Learning Research, vol. 17, no. 39, pp. 1–40, 2016.