Boulder

Reinforcement Learning; Real Environments

YouTube Playlist

Maziar Raissi

Assistant Professor

Department of Applied Mathematics

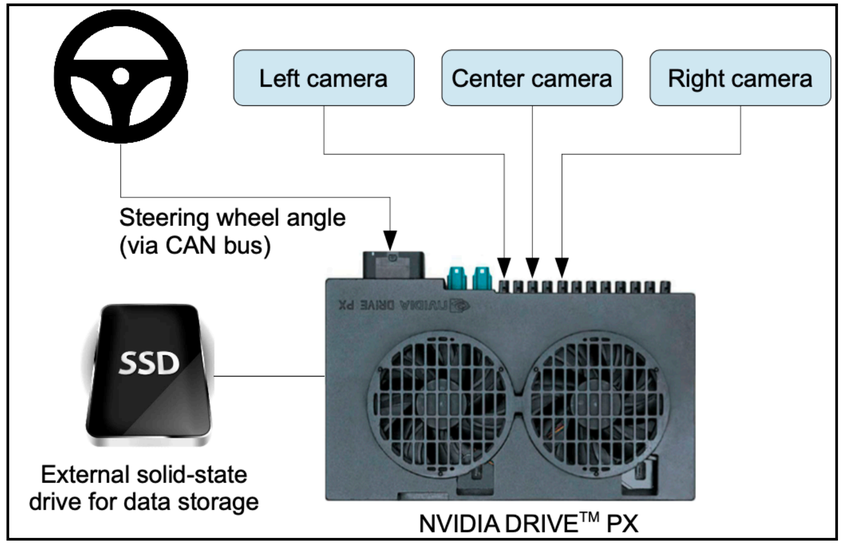
University of Colorado Boulder

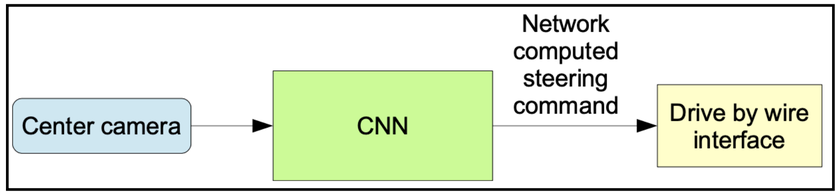
maziar.raissi@colorado.edu

End to End Learning for Self-Driving Cars

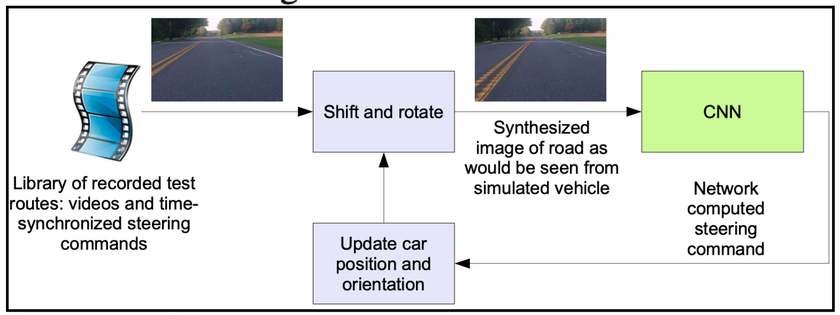
YouTube Video

High-level view of the data collection system.





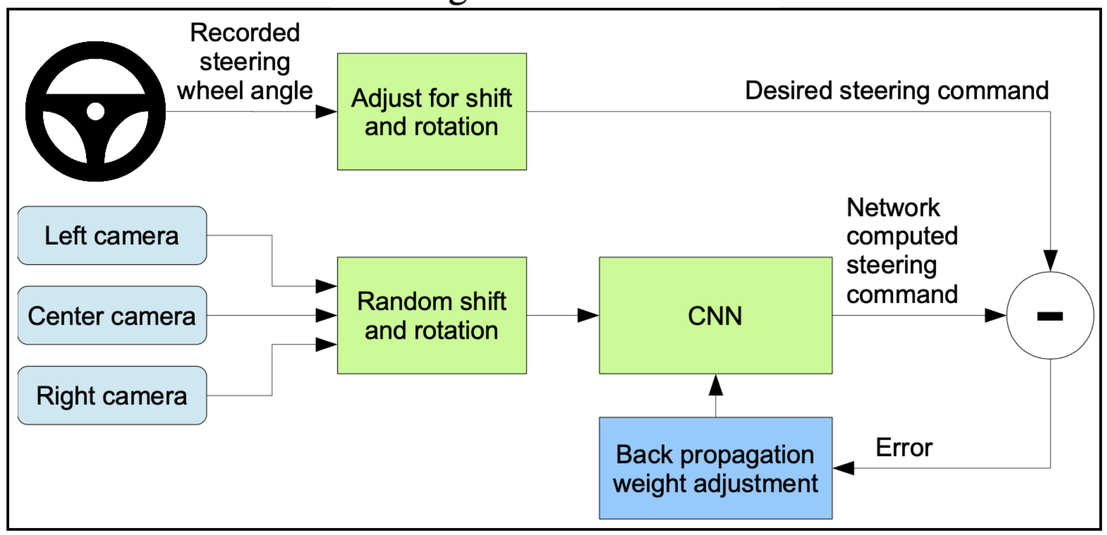
Block-diagram of the drive simulator.

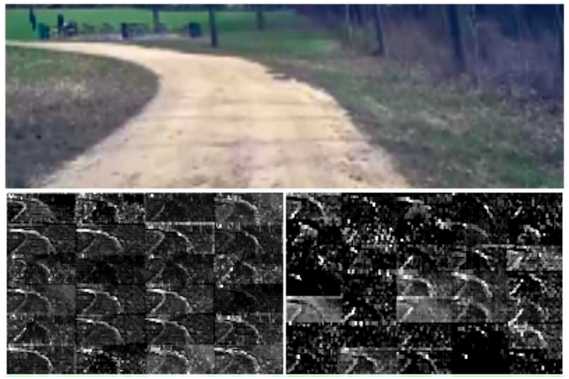


Evaluation Metric

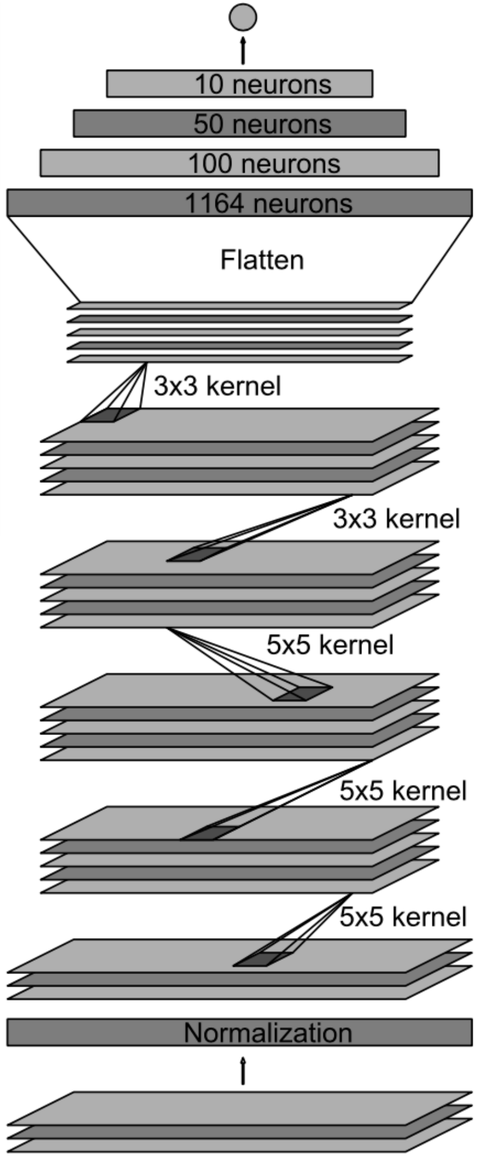
$$\text{autonomy} = (1 - \frac{(\text{number of interventions}) \cdot 6 \text{ seconds}}{\text{elapsed time [seconds]}}) \cdot 100$$

Training the neural network.





Activations of the first & second layer feature maps.



Output: vehicle control

Fully-connected layer
Fully-connected layer
Fully-connected layer

Convolutional feature map 64@1x18

Convolutional feature map 64@3x20

Convolutional feature map 48@5x22

Convolutional feature map 36@14x47

Convolutional feature map 24@31x98

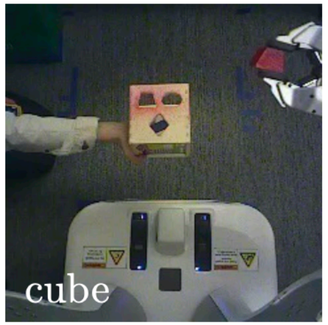
Normalized input planes 3@66x200

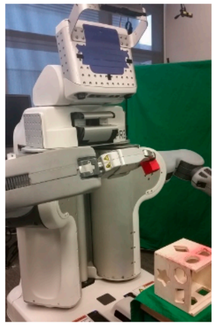
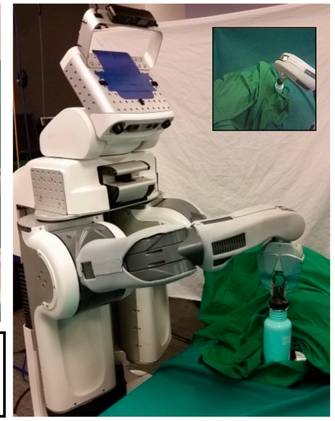
Input planes 3@66x200

Bojarski, Mariusz, et al. "End to end learning for self-driving cars." *arXiv preprint arXiv:1604.07316* (2016).

End-To-End Training Of Deep Visuomotor Policies

YouTube Video



http://sites.google.com/site/visuomotorpolicy

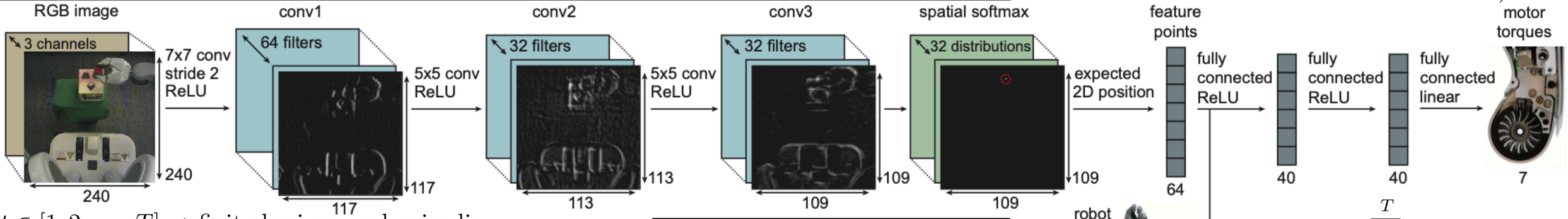
$\pi_\theta(u_t|o_t) \rightarrow$ policy

$u_t \rightarrow$ actions (motor torque commands)

$o_t \rightarrow$ observations (an image from robot's onboard camera)

$x_t \rightarrow$ states (joint angles of the robot, positions of objects in the world, and their time derivatives)

Inserting a block into a shape sorting cube, screwing a cap onto a bottle, fitting the claw of a toy hammer under a nail with various grasps, and placing a coat hanger on a rack with a PR2 robot



$t \in [1, 2, \ldots, T] \rightarrow$ finite horizon and episodic

$p(x_{t+1}|x_t, u_t) \rightarrow$ state dynamics (not assumed to be known)

$$f_{cx} = \sum_{ij} s_{cij} x_{ij} \text{ and } f_{cy} = \sum_{ij} s_{cij} y_{ij}$$

$p(o_t|x_t) \rightarrow$ observations are a stochastic consequence of the states (not assumed to be known)

$\pi_\theta(u_t|x_t) = \int \pi_\theta(u_t|o_t) p(o_t|x_t) do_t \rightarrow$ distribution over actions under the policy conditioned on the state

$\tau = \{x_1, u_1, x_2, u_2, \ldots, x_T, u_T\} \rightarrow$ trajectory

$$p(\mathbf{u}_t|\mathbf{x}_t) = \mathcal{N}(\mathbf{K}_t \mathbf{x}_t + \mathbf{k}_t, \mathbf{C}_t)$$

$\pi_\theta(\tau) = p(x_1) \prod_{t=1}^{T} \pi_\theta(u_t|x_t) p(x_{t+1}|x_t, u_t) \rightarrow$ distribution over trajectories

$\ell(x_t, u_t) \rightarrow$ cost function

$$p(\mathbf{x}_{t+1}|\mathbf{x}_t, \mathbf{u}_t) = \mathcal{N}(f_{\mathbf{x}t}\mathbf{x}_t + f_{\mathbf{u}t}\mathbf{u}_t + f_{ct}, \mathbf{F}_t)$$

$$\min_\theta \mathbb{E}_{\pi_\theta(\tau)}[\underbrace{\sum_{t=1}^{T} \ell(x_t, u_t)}_{=:\ell(\tau)}]$$

Guided Policy Search

$$\min_{\theta, p} \quad \mathbb{E}_{p(\tau)}[\ell(\tau)]$$

$$\text{s.t.} \quad p(u_t|x_t) = \pi_\theta(u_t|x_t)$$

$p(\tau) \rightarrow$ guiding distribution

Alternating Direction Method of Multipliers

Levine, Sergey, et al. "End-to-end training of deep visuomotor policies." *The Journal of Machine Learning Research* 17.1 (2016): 1334-1373.

Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection

Boulder

YouTube Video

https://youtu.be/cXaic_k80uM

Visual servoing: vision-based robot control

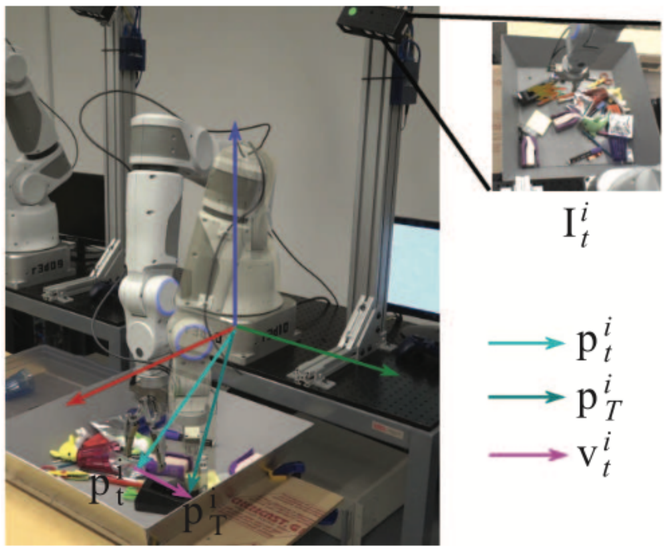
$g \rightarrow$ grasp success prediction network

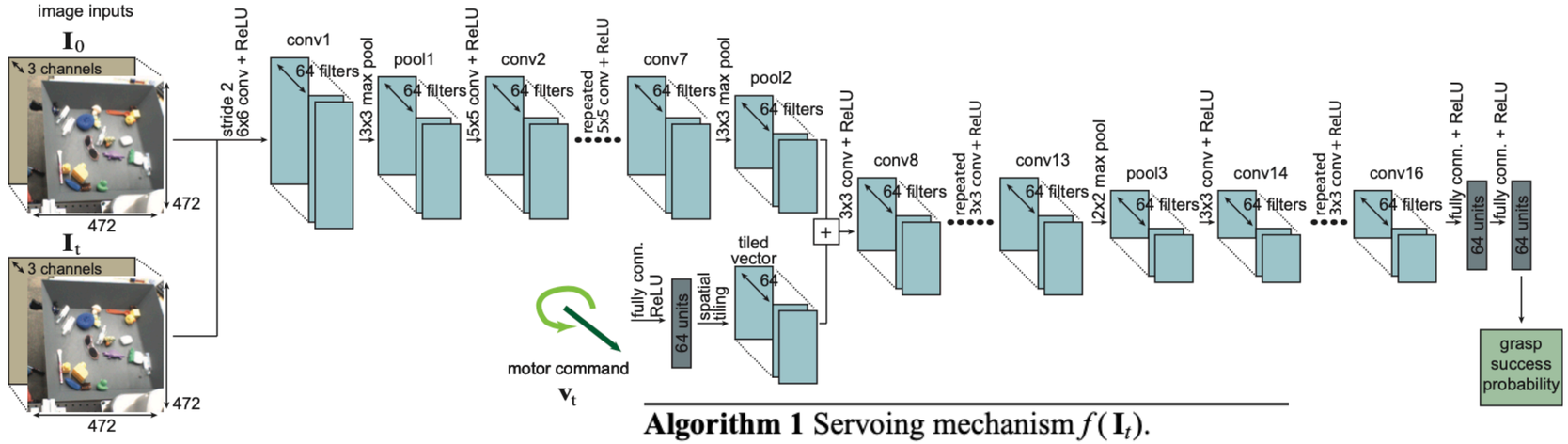
$I_t \rightarrow$ visual input

$v_t \rightarrow$ task-space motion command

$g(I_t, v_t) \rightarrow$ predicted probability that executing the command v_t will produce a successful grasp

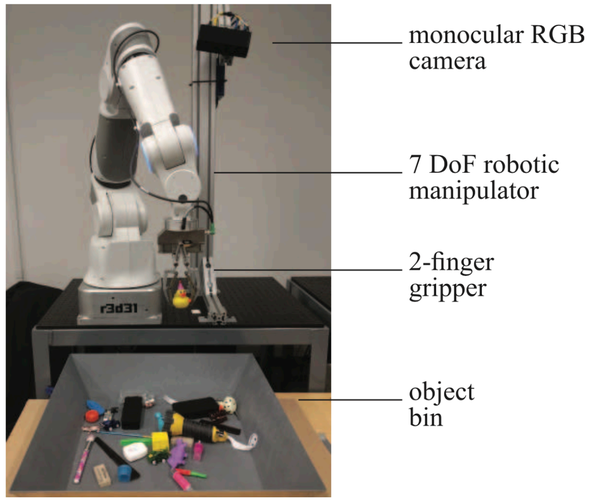
$f(I_t) \rightarrow$ servoing mechanism





$(I_t^i, \underbrace{P_T^i - p_t^i}_{v_t^i}, \ell_i) \rightarrow T$ training samples

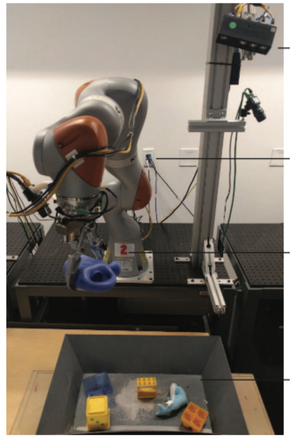
label



Algorithm 1 Servoing mechanism $f(\mathbf{I}_t)$.

1: Given current image $\mathbf{I}_t$ and network g.
2: Infer $\mathbf{v}_t^\star$ using g and CEM.
3: Evaluate $p = g(\mathbf{I}_t, \emptyset) / g(\mathbf{I}_t, \mathbf{v}_t^\star)$.
4: **if** $p > 0.9$ **then**
5: Output $\emptyset$, close gripper.
6: **else if** $p \leq 0.5$ **then**
7: Modify $\mathbf{v}_t^\star$ to raise gripper height and execute $\mathbf{v}_t^\star$.
8: **else**
9: Execute $\mathbf{v}_t^\star$.
10: **end if**

No Motion



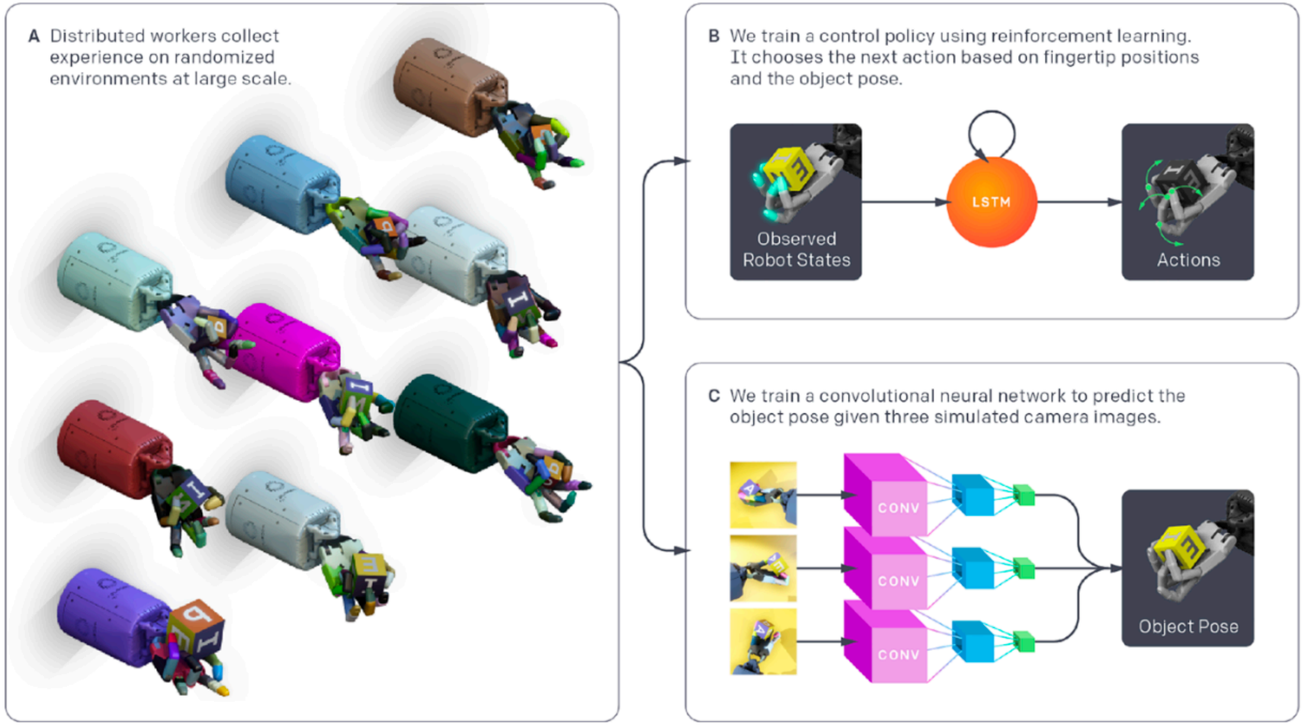
Cross-Entropy Method (CEM)

A simple derivative-free optimization algorithm

1) Sample N=64 values from $\mathcal{N}(0, I)$
2) Choose the best M = 6 and fit a Gaussian to these M samples
3) Sample N=64 value from this Gaussian
4) Repeat steps 2 & 3 three times
5) Return the best v_t^*

Levine, Sergey, et al. "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection." *The International Journal of Robotics Research* 37.4-5 (2018): 421-436.

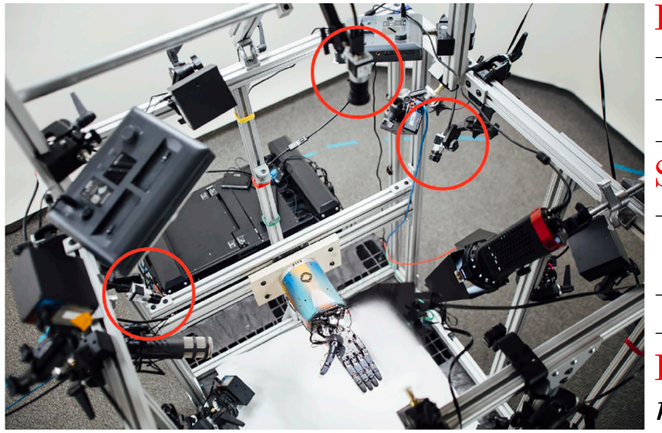
Learning Dexterous In-Hand Manipulation



Andrychowicz, OpenAI: Marcin, et al. "Learning dexterous in-hand manipulation." *The International Journal of Robotics Research* 39.1 (2020): 3-20.

Uncut and real-time video footage of a trial in which the robot hand successfully performed 50 consecutive rotations: https://youtu.be/DKe8FumoD4E

Generalized Advantage Estimator (GAE) & Proximal Policy Optimization (PPO)



Hardware
- Shadow Dexterous Hand
- PhaseSpace tracking
- RGB cameras

Simulation
- MuJoCo physics engine
  http://mujoco.org/book/
- Unity
- OpenAI Gym

Rewards

$r_t = d_t - d_{t+1}$

d_t and d_{t+1} are the rotation angles between the desired and current object orientations before and after the transition, respectively.

Reward of 5 whenever a goal is achieved and a reward of -20 (penalty) whenever the object is dropped.

Reality Gap

We face a dilemma: we cannot train on the physical robot because deep RL algorithms require millions of samples; conversely, training only in simulation results in policies that do no transfer well due to the gap between the simulated and real environments.

Domain Randomization

Distribution over many simulations: randomizing most aspects of the simulated environment (e.g., Observation noise, Physics randomizations, Unmodeled effects, and Visual appearance randomizations)

Physical task	Mean	Median	Individual trials (sorted)									
Block (vision)	15.2±14.3	11.5	46	28	26	15	13	10	8	3	2	1

Boulder

Questions?

YouTube Playlist