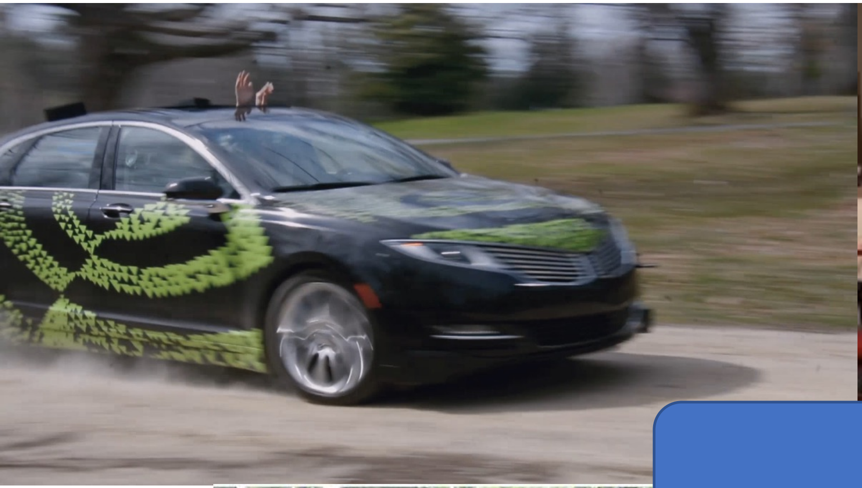
Imitation Learning from Observation

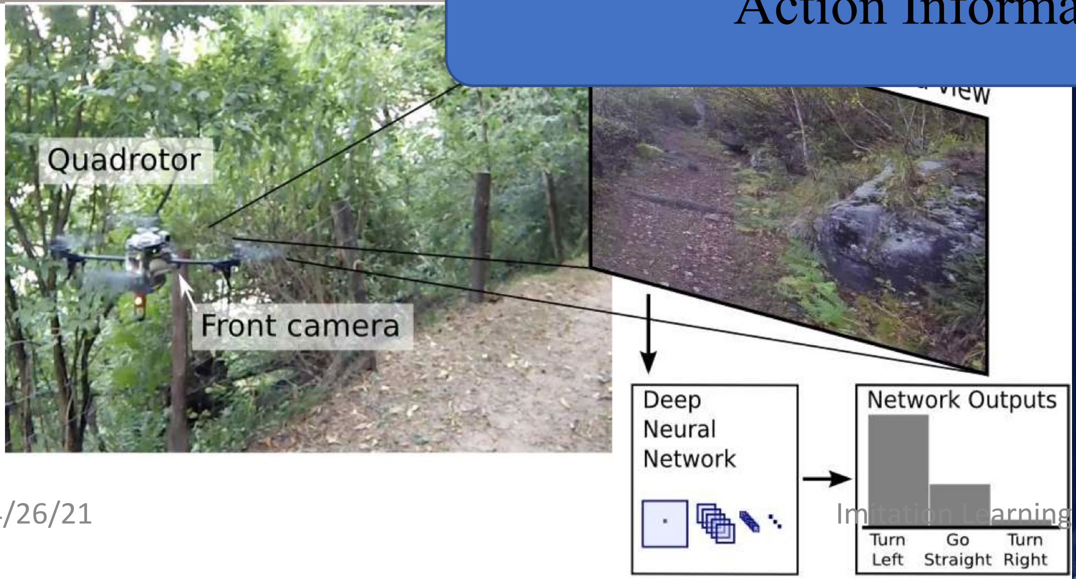
PhD Defense

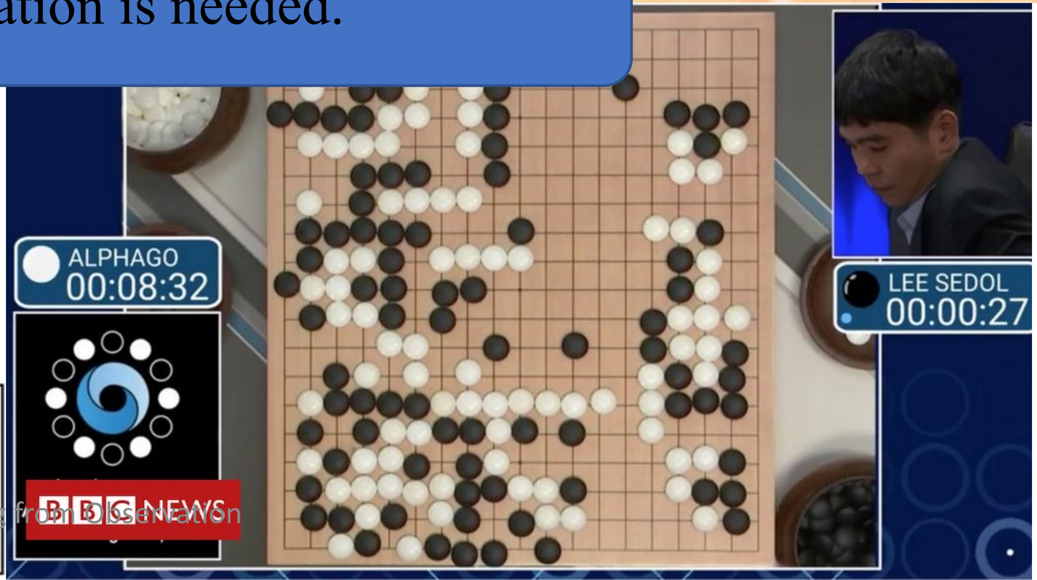
Faraz Torabi
Supervisor: Peter Stone
University of Texas at Austin
faraztrb@cs.utexas.edu

Success of Imitation Learning



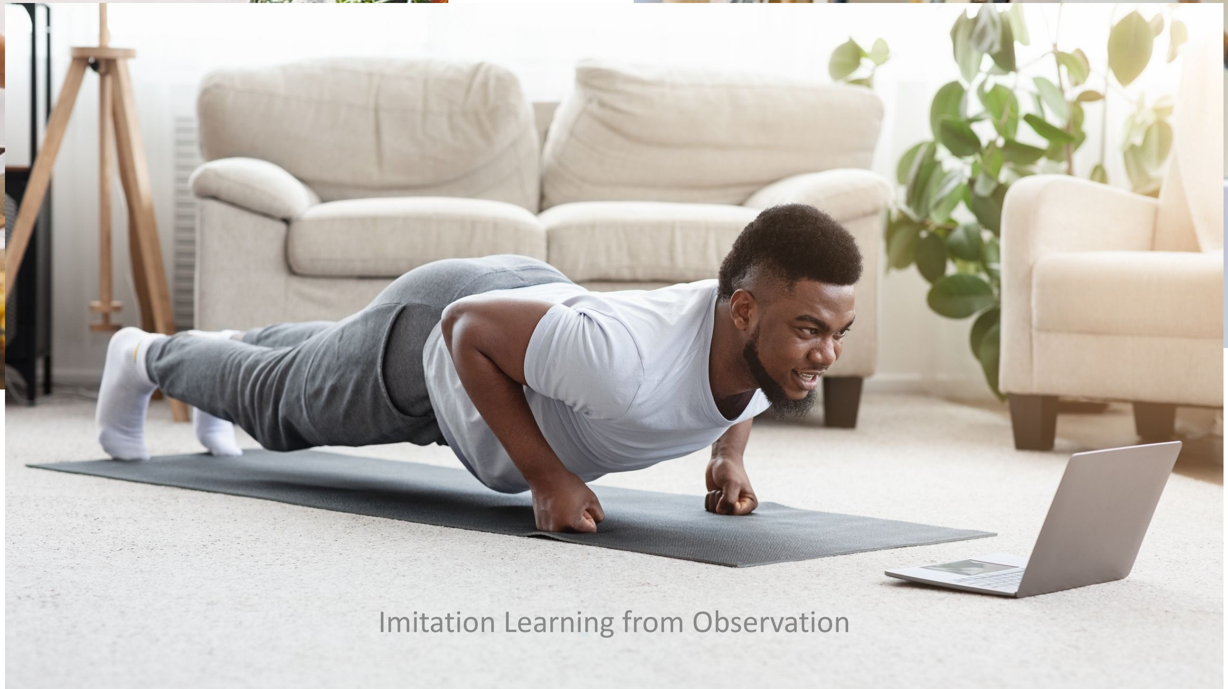
Action Information is needed.





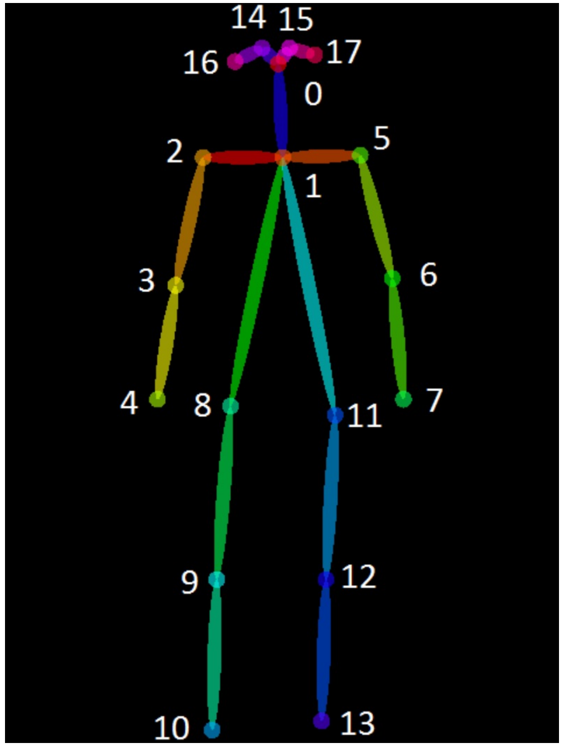
Humans Mostly Learn by Observation

Example: Watching YouTube



Imitation Learning from Observation

Perception



Control

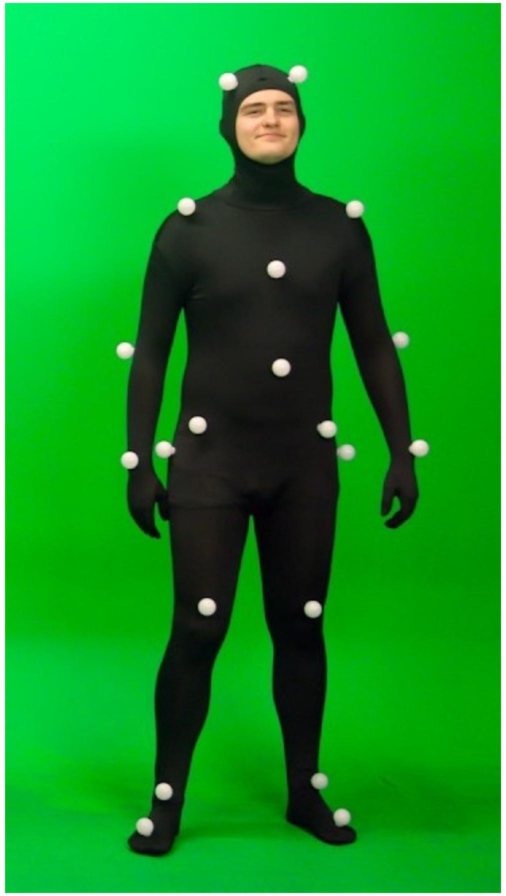
Visual observations

States

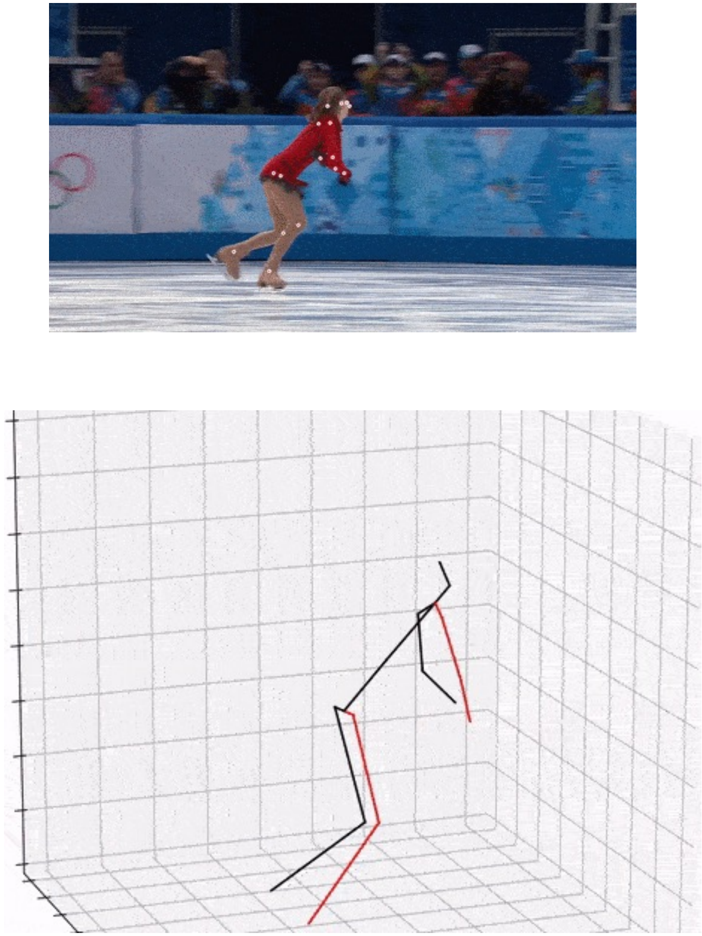
Policy

Perception Module

Sensors

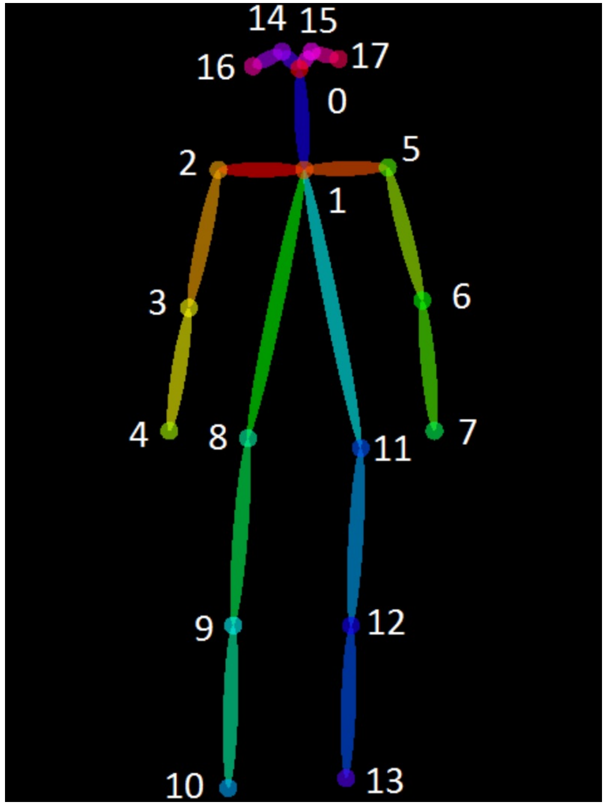


Motion Capture



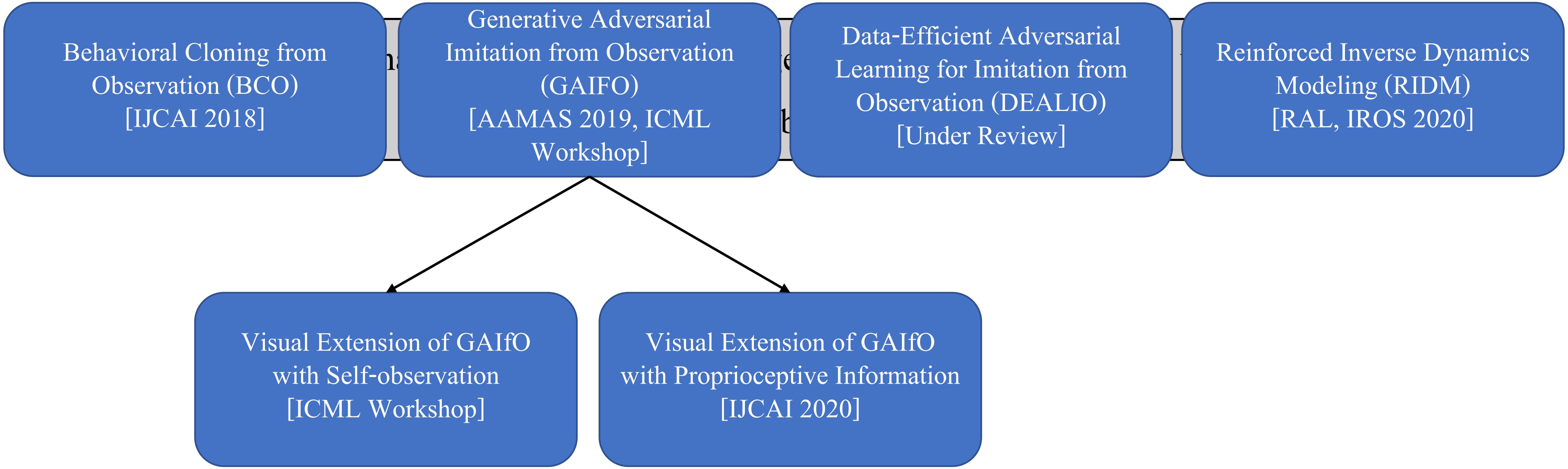
Pose Estimation

Focus of My Research is on the …



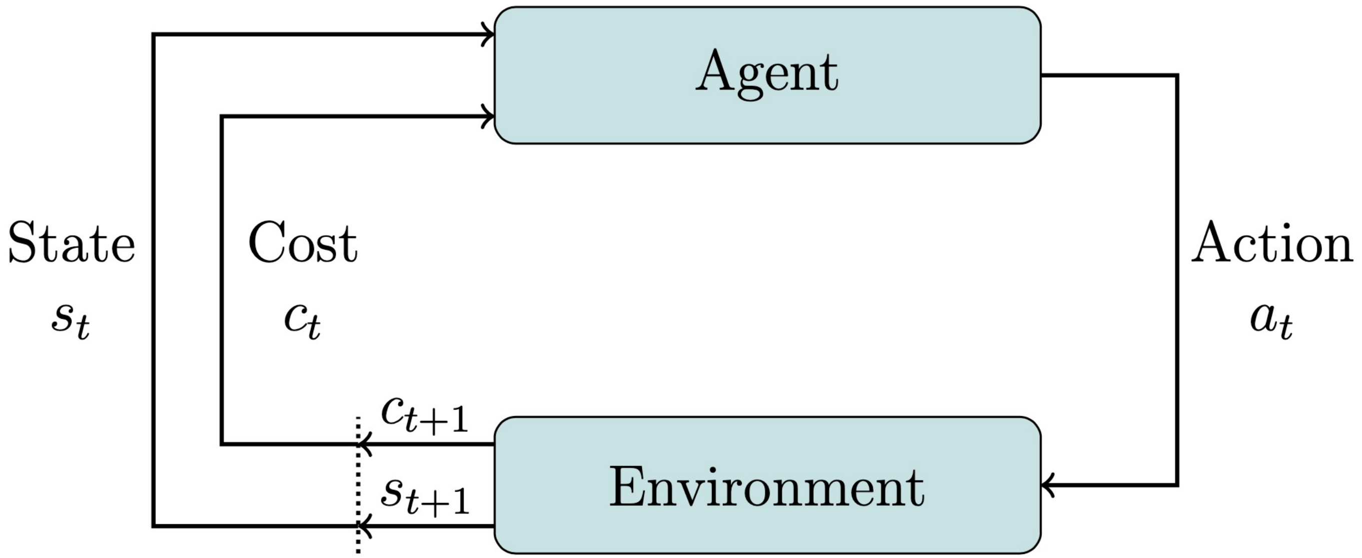
Control

Research Question



Reinforcement Learning

- Goal:
  - Learn how to make decisions by minimizing the cumulative cost feedback.
- $M = < S, A, P, c >$
  - *S: Set of states*
  - *A: Set of actions*
  - P: Transition function
  - c: Cost function
- Learn a policy $\pi: S \rightarrow A$



Reinforcement Learning

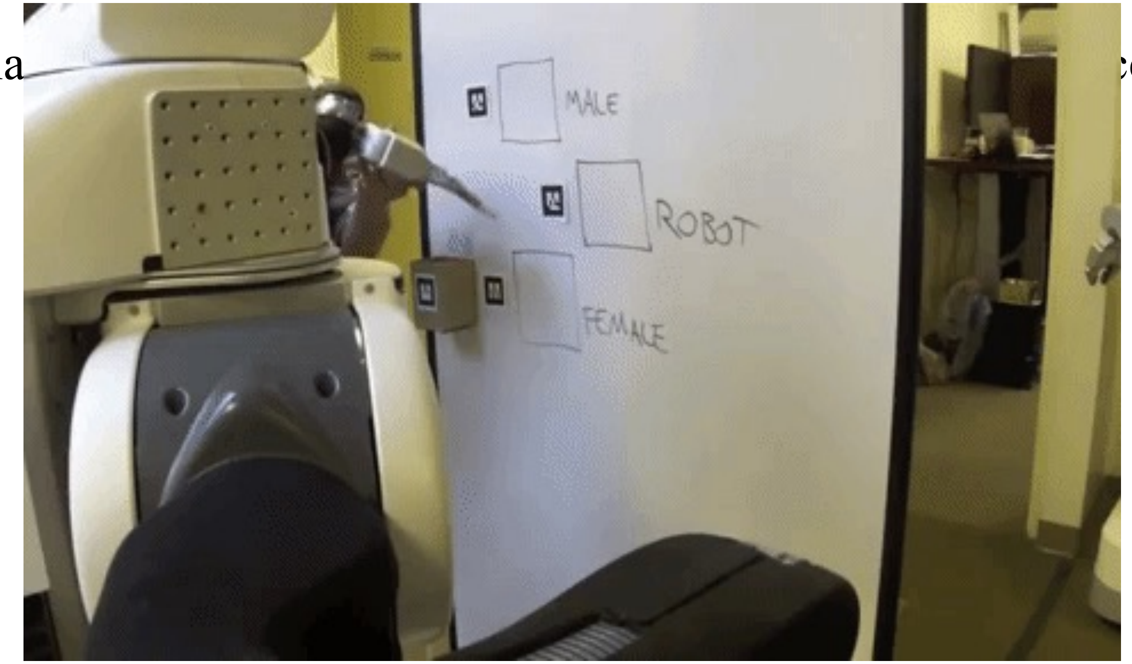
- Algorithms:
  - Model-based:
    - Known model (planning): LQR, MCTS, etc.
    - Unknown model: PILQR, PLATO, etc.
  - Model-free:
    - Policy-based: Reinforce, TRPO, PPO, etc.
    - Value-based: SARSA, Q-learning, etc.

Imitation Learning

- Goal:
  - Learn how to make decisions by trying to imitate another agent.
- $M\backslash$c:
  - Provided: $\tau_i^e = \{(s_0^e, a_0^e), (s_0^e, a_0^e), \dots, (s_N^e, a_N^e)\}_i$
  - Learn: $\pi: s_t \rightarrow a_t$
- Algorithms:
  - Behavioral Cloning (BC)
  - Inverse Reinforcement Learning (IRL)
  - Adversarial Imitation Learning (AIL)

Imitation Learning

- Observations of other agent (demonstrations) consist of state-action pairs.
- Limitation:
  - Precludes using a la ... ces are not given.



Scott Niekum et al. “Learning and generalization of complex tasks from unstructured demonstrations”. In: Intelligent Robots and Systems (IROS), 2012

Imitation Learning from Observation

Control Module

- Goal:
  - Learn how to perform a task ~~by visually observing~~ from state-only demonstrations of an expert.
- $M\backslash$c:
  - Provided: $\cancel{\tau_i^e = \{o_0^e, o_1^e, \ldots, o_N^e\}_i}$ $\tau_i^e = \{s_0^e, s_1^e, \ldots, s_N^e\}_i$
  - Learn: $\pi : s_t \rightarrow a_t$

In what ways can autonomous agents learn to imitate experts using state-only observations?

Behavioral Cloning from Observation (BCO)
[IJCAI 2018]

Generative Adversarial Imitation from Observation (GAIFO)
[AAMAS 2019, ICML Workshop]

Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO)
[Under Review]

Reinforced Inverse Dynamics Modeling (RIDM)
[RAL, IROS 2020]

Visual Extension of GAIfO with Self-observation
[ICML Workshop]

Visual Extension of GAIfO with Proprioceptive Information
[IJCAI 2020]

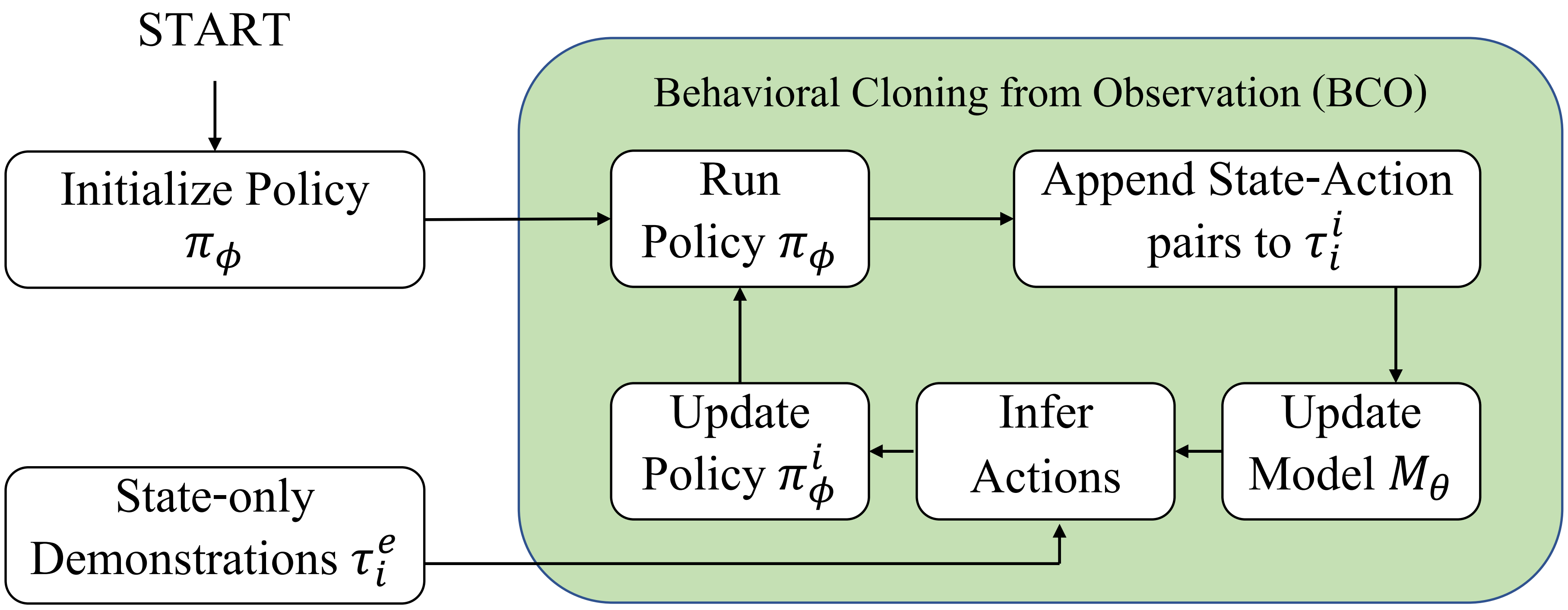
Behavioral Cloning from Observation

- Goal:
  - Propose a ***Model-based Algorithm*** for Imitation from Observation.
- Imitation Learning (IL): $\tau_i^e = \{(s_0^e, a_0^e), (s_0^e, a_0^e), \dots, (s_N^e, a_N^e)\}_i$
- Imitation from Observation (IfO): $\tau_i^e = \{(s_0^e, ?), (s_1^e, ?), \dots, (s_N^e, ?)\}_i$

A Model-based Approach

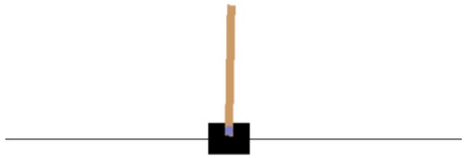
Learn an inverse model of the environment → Infer the missing actions → Perform a conventional IL method

Behavioral Cloning from Observation

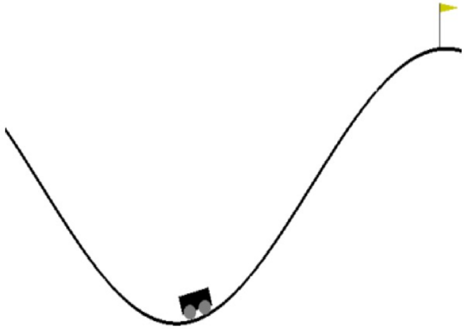


Experiments

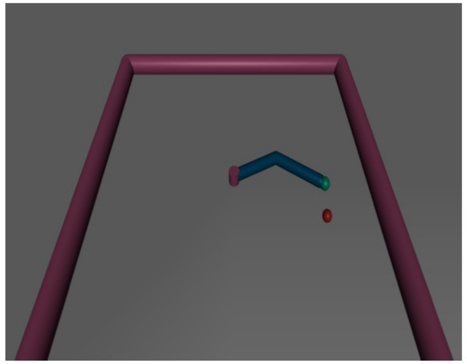
- Tasks:



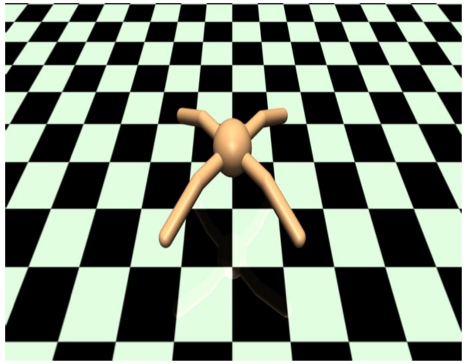
(a) CartPole



(b) MountainCar



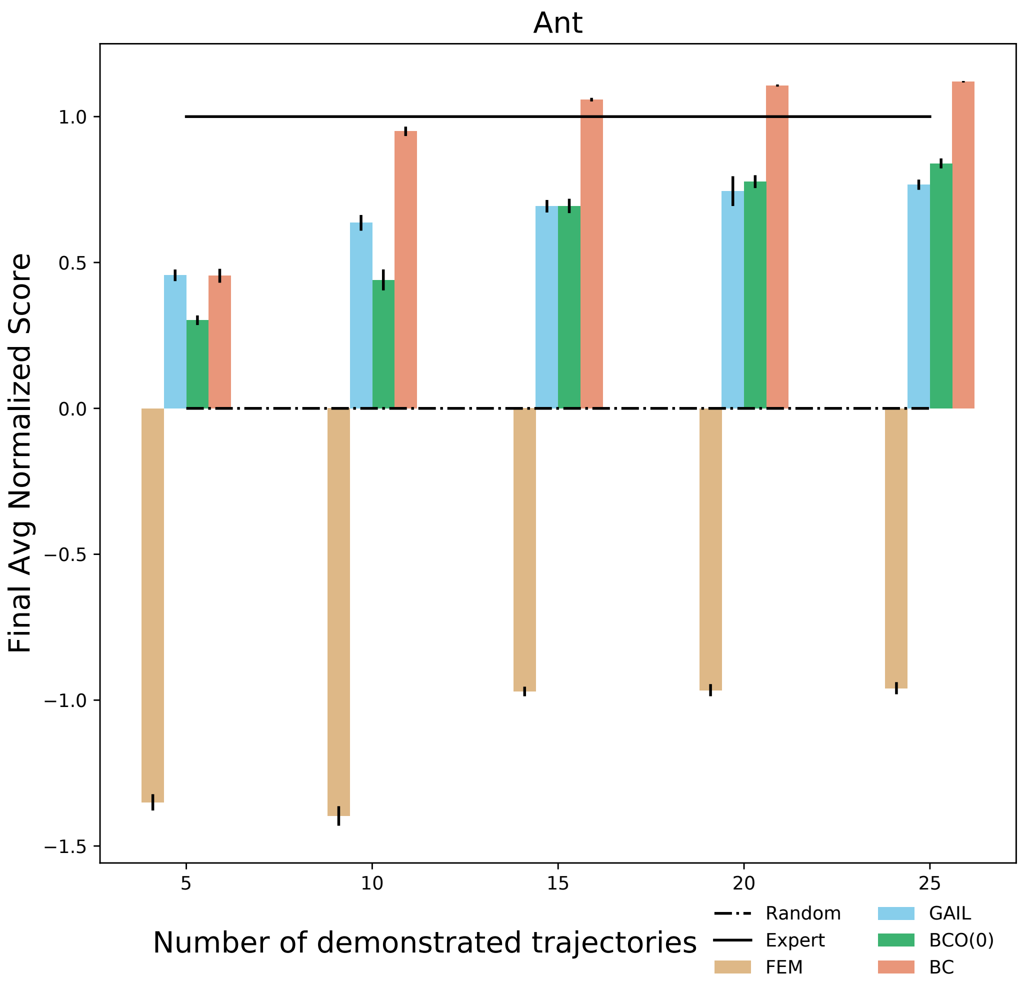
(c) Reacher



(d) Ant

Experiments

- Ant



In what ways can autonomous agents learn to imitate experts using state-only observations?

Behavioral Cloning from Observation (BCO)
[IJCAI 2018]

Generative Adversarial Imitation from Observation (GAIFO)
[AAMAS 2019, ICML Workshop]

Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO)
[Under Review]

Reinforced Inverse Dynamics Modeling (RIDM)
[RAL, IROS 2020]

Visual Extension of GAIfO with Self-observation
[ICML Workshop]

Visual Extension of GAIfO with Proprioceptive Information
[IJCAI 2020]

Generative Adversarial Imitation from Observation

- Goal:
  - Propose a ***Model-free Algorithm*** for Imitation from Observation.
- IfO problem:

$$RL \circ IRLfO_{\psi}(\pi^e) = \underset{\pi \in \prod}{\operatorname{argmin}}\ \underset{c \in R^{S \times S}}{\operatorname{argmax}} -\psi(c) + (\min_{\pi \in \prod} E_{\pi}[c(s,s')]) - E_{\pi^e}[c(s,s')]$$

- Which is a composition of:

$$\mathrm{IRLFO}_{\psi}(\pi^e) = \underset{c \in R^{S \times S}}{\operatorname{argmax}} -\psi(c) + (\min_{\pi \in \prod} E_{\pi}[c(s,s')]) - E_{\pi^e}[c(s,s')]$$

$$RL(\tilde{c}) = \underset{\pi \in \prod}{\operatorname{argmin}}\ E_{\pi}[\tilde{c}(s,s')]$$

- It is equivalent to solving:

$$\underset{\pi \in \prod}{\operatorname{argmin}}\ \underset{D \in (0,1)^{S \times S}}{\operatorname{argmax}}\ E_{\pi}[\log(D(s,s'))] + E_{\pi^e}[\log(1 - D(s,s'))]$$

Difficult to Solve

- $c(s,s')$: Cost as a function of state transition
- π^e: Expert policy
- $\prod$: Set of all possible policies
- $\psi(c)$: Regularizer

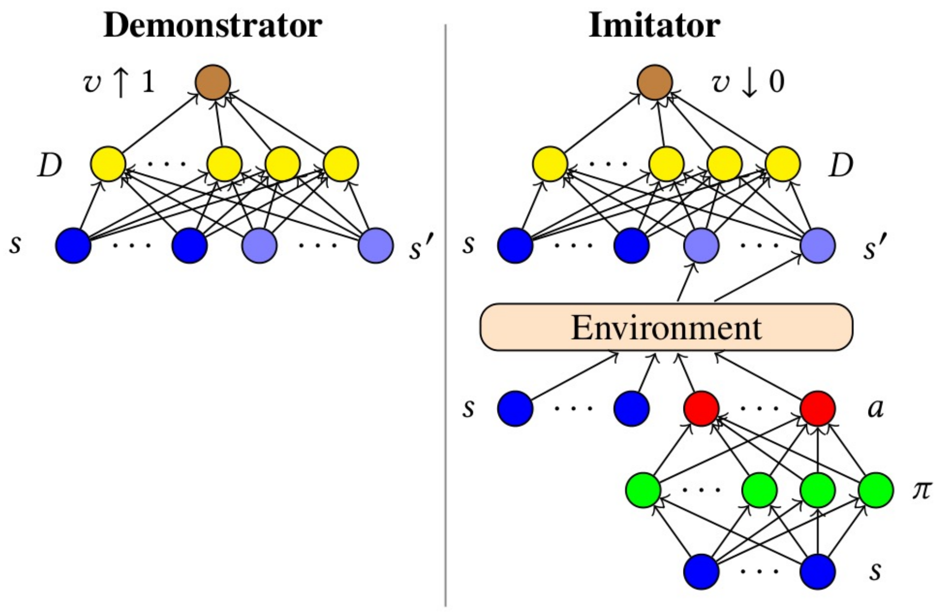
Generative Adversarial Imitation from Observation

- Algorithm:
  - Initialize π_ϕ and D_θ
  - While π_ϕ improves do:
    - Execute π_ϕ and store state transitions $\tau_i^i = \{(s^i)\}_i$
    - Update D_θ using loss:

      $$-(E_\pi[\log(D(s,s'))] + E_{\pi^e}[\log(1 - D(s,s'))])$$

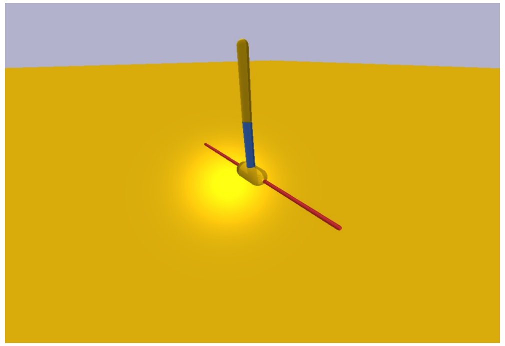
    - Update π_ϕ by performing TRPO updates with cost function:

      $$E_\pi[\log(D(s,s'))]$$

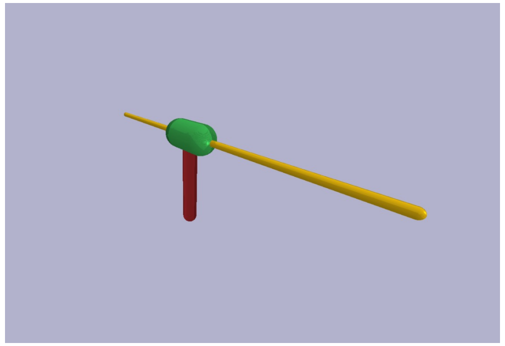


Experiments

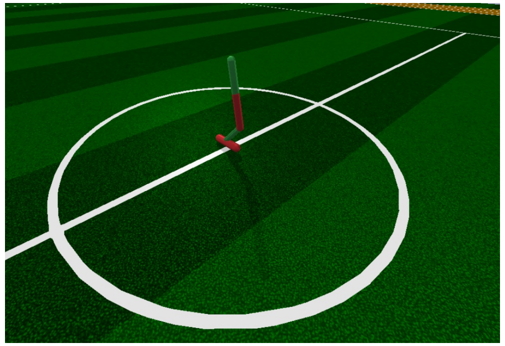
- Tasks:



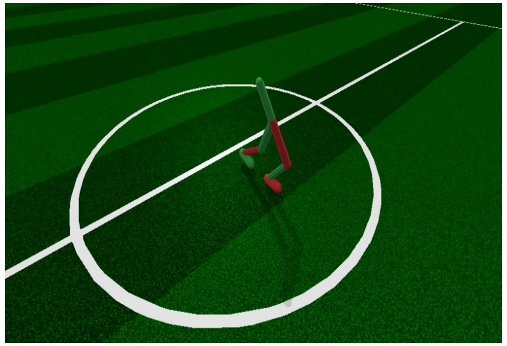
InvertedDoublePendulum



InvertedPendulumSwingup



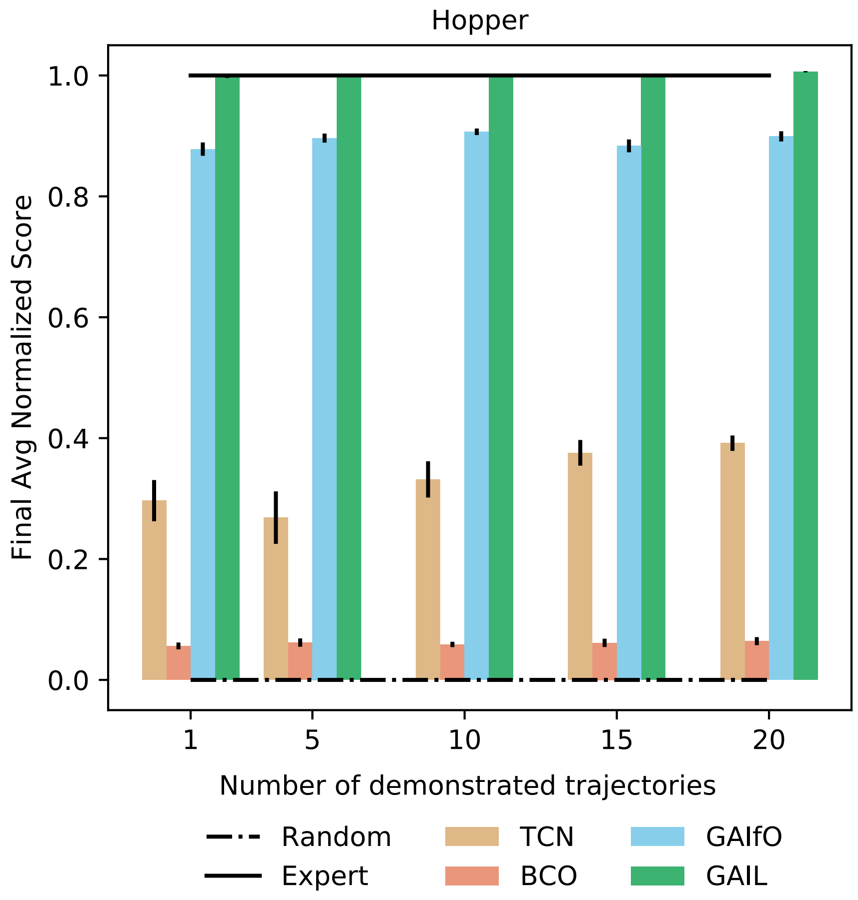
Hopper



Walker2D

Experiments

- Hopper:



Theoretical Contribution

- IfO problem:

$$RL \circ IRLfO_{\psi}(\pi^e) = \underset{\pi \in \prod}{\operatorname{argmin}} \underset{c \in R^{S \times S}}{\operatorname{argmax}} -\psi(c) + (\min_{\pi \in \prod} E_{\pi}[c(s, s')]) - E_{\pi^e}[c(s, s')]$$

- Equivalent to:

Difficult to Solve

$$\underset{\pi \in \prod}{\operatorname{argmin}} \underset{D \in (0,1)^{S \times S}}{\operatorname{argmax}} E_{\pi}[\log(D(s, s'))] + E_{\pi^e}[\log(1 - D(s, s'))]$$

How are they equivalent?

Proposition 5.4.1. $RL \circ IRLfO_{\psi}(\pi^e)$ and $\underset{\pi \in \prod}{\operatorname{argmin}}\, \psi^*(\rho_{\pi}^{s} - \rho_{\pi^e}^{s})$ induce policies that have the same state transition occupancy measure, $\rho_{\tilde{\pi}}^{s}$.

Where $\rho_{\pi}^{s}(s_i, s_j) = \sum_a P(s_j | s_i, a)\pi(a|s_i) \sum_{t=0}^{\infty} \gamma^t P(s_t = s_i | \pi)$

Proposition 5.4.1. $RL \circ IRLfO_{\psi}(\pi^e)$ and $\underset{\pi\in\prod}{\operatorname{argmin}}\, \psi^*(\rho_\pi^s - \rho_{\pi^e}^s)$ induce policies that have the same state transition occupancy measure, $\rho_{\tilde{\pi}}^s$.

$$RL \circ IRLfO_{\psi}(\pi^e) = \underset{\pi\in\prod}{\operatorname{argmin}}\ \underset{c\in R^{S\times S}}{\operatorname{argmax}} -\psi(c) + (\min_{\pi\in\prod} E_\pi[c(s,s')]) - E_{\pi^e}[c(s,s')]$$

$$\overline{RL} \circ \overline{IRLfO}_{\psi}(\pi^e) = \underset{\rho_\pi^s\in P^s}{\operatorname{argmin}}\ \underset{c\in R^{S\times S}}{\operatorname{argmax}} -\psi(c) + (\min_{\rho_\pi^s\in P^s} \sum_{s,s'} \rho_\pi^s(s,s')\, c(s,s')) - \sum_{s,s'} \rho_{\pi^e}^s(s,s') c(s,s')$$

Proof:

$RL \circ IRLfO_{\psi}(\pi^e)$ — Same $\rho_{\tilde{\pi}}^s$ (Proposition 5.4.1) — $\underset{\pi\in\prod}{\operatorname{argmin}}\, \psi^*(\rho_\pi^s - \rho_{\pi^e}^s)$

Lemma 5.4.1 | Same $\rho_{\tilde{\pi}}^s$

Same $\rho_{\tilde{\pi}}^s$ | Lemma 5.4.3

$\overline{RL} \circ \overline{IRLfO}_{\psi}(\pi^e)$ — Same $\rho_{\tilde{\pi}}^s$ (Lemma 5.4.2) — $\underset{\rho_\pi^s\in P^s}{\operatorname{argmin}}\, \psi^*(\rho_\pi^s - \rho_{\pi^e}^s)$

Theoretical Contribution

- New IfO problem:

$$\operatorname*{argmin}_{\pi \in \prod} \psi^*(\rho_\pi^s - \rho_{\pi^e}^s)$$

- Specifying ψ:

$$\psi(c) = \begin{cases} E_{\pi^e}\left[g\big(c(s,s')\big)\right] & if\ c < 0 \\ +\infty & otherwise \end{cases} \quad \text{where} \quad g(x) = \begin{cases} -x - \log(1 - e^x) & if\ x < 0 \\ +\infty & otherwise \end{cases}$$

- The optimization problem becomes [proposition A.1.1]:

$$\operatorname*{argmin}_{\pi \in \prod} \operatorname*{argmax}_{D \in (0,1)^{S \times S}} E_\pi\left[\log\big(D(s,s')\big)\right] + E_{\pi^e}\left[\log\big(1 - D(s,s')\big)\right]$$

Similar to the Generative Adversarial Loss

In what ways can autonomous agents learn to imitate experts using state-only observations?

Behavioral Cloning from Observation (BCO)
[IJCAI 2018]

Generative Adversarial Imitation from Observation (GAIFO)
[AAMAS 2019, ICML Workshop]

Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO)
[Under Review]

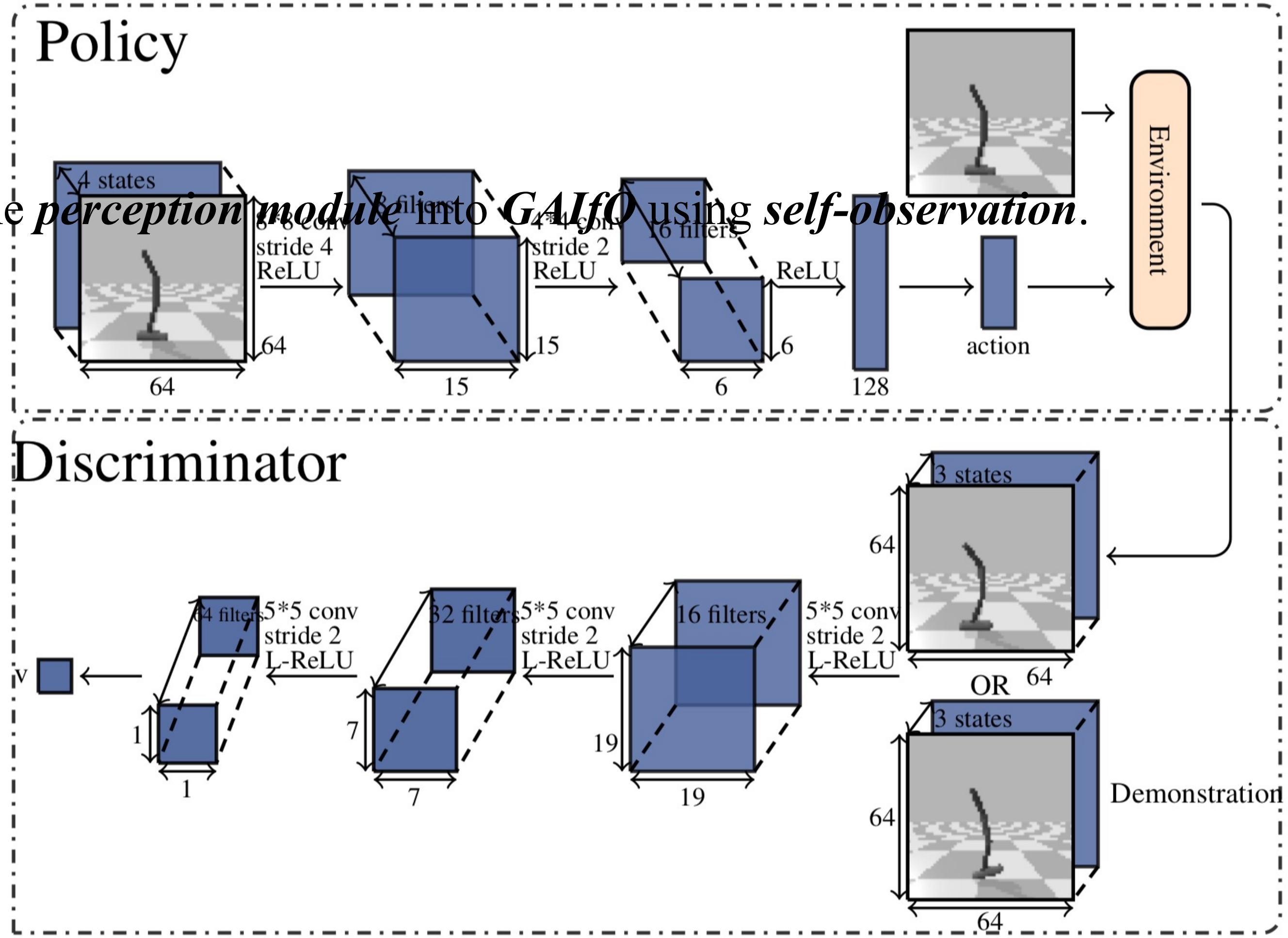
Reinforced Inverse Dynamics Modeling (RIDM)
[RAL, IROS 2020]

Visual Extension of GAIfO with Self-observation
[ICML Workshop]

Visual Extension of GAIfO with Proprioceptive Information
[IJCAI 2020]

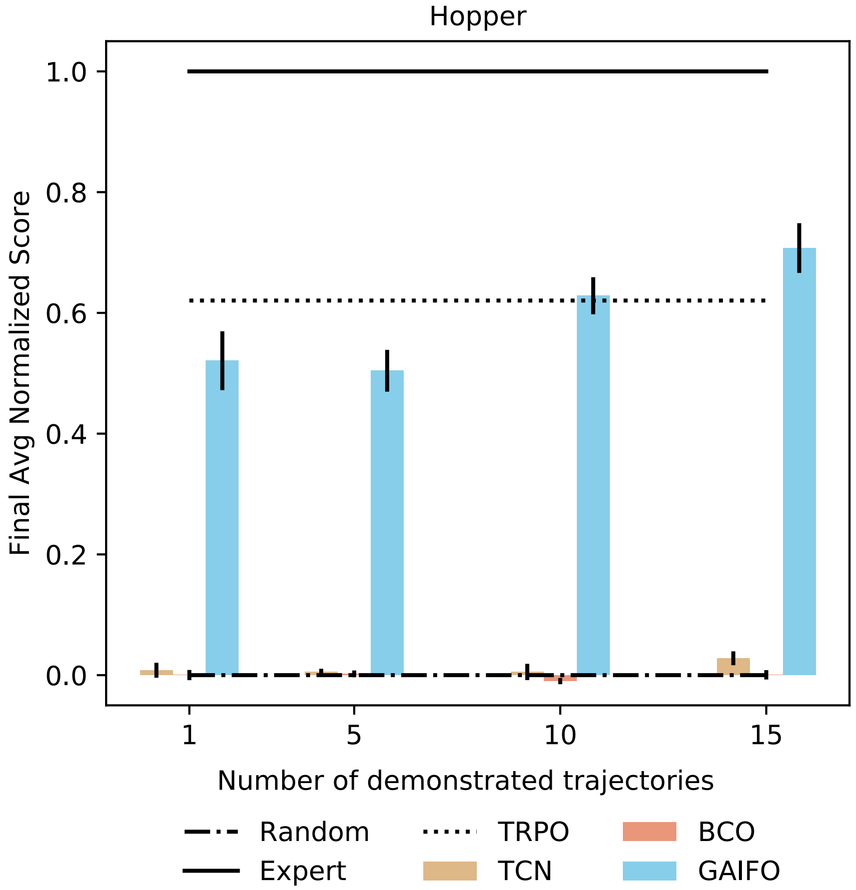
GAIfO with Self-observation

- Goal:
  - Incorporate the ***perception module*** into ***GAIfO*** using ***self-observation***.

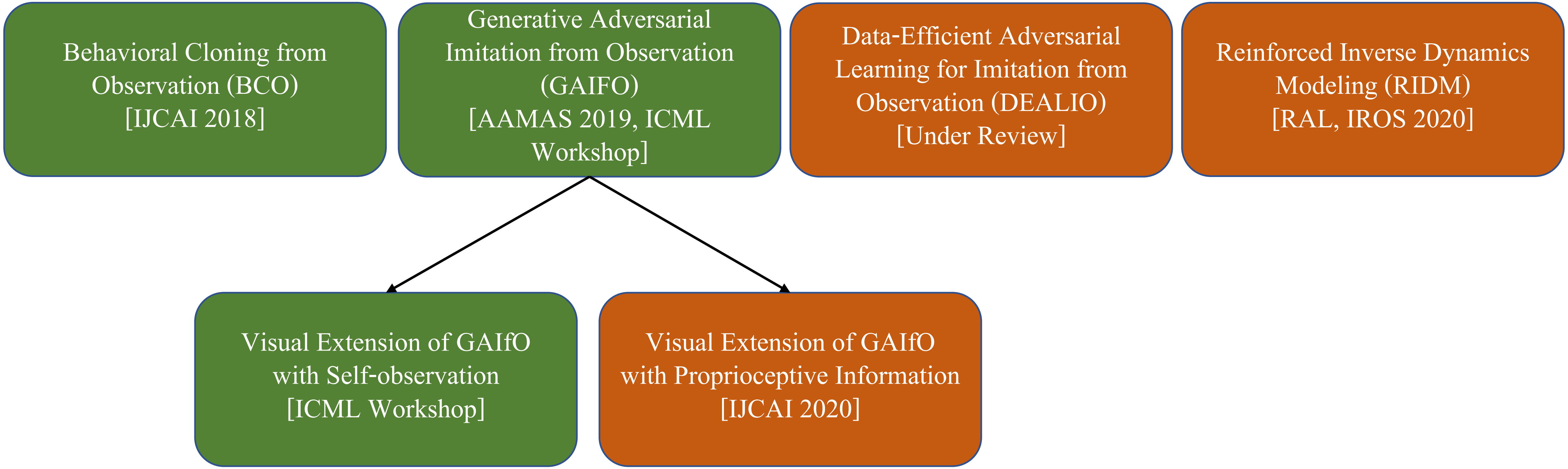


Experiments

- Hopper:



In what ways can autonomous agents learn to imitate experts using state-only observations?



GAIfO with Proprioceptive Information

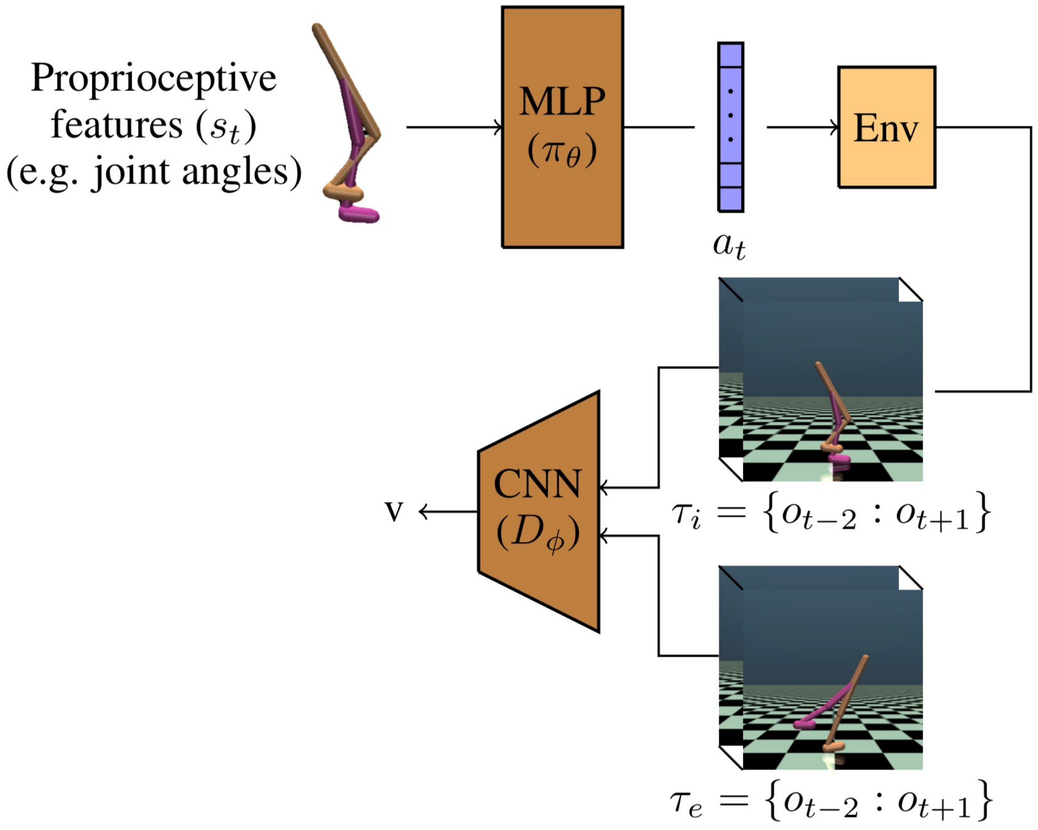
- Goal:
  - To improve the ***performance*** and ***sample-complexity*** of ***GAIfO with self-observation***.

- Hypothesis:

Leveraging ***proprioceptive information*** will help with both issues

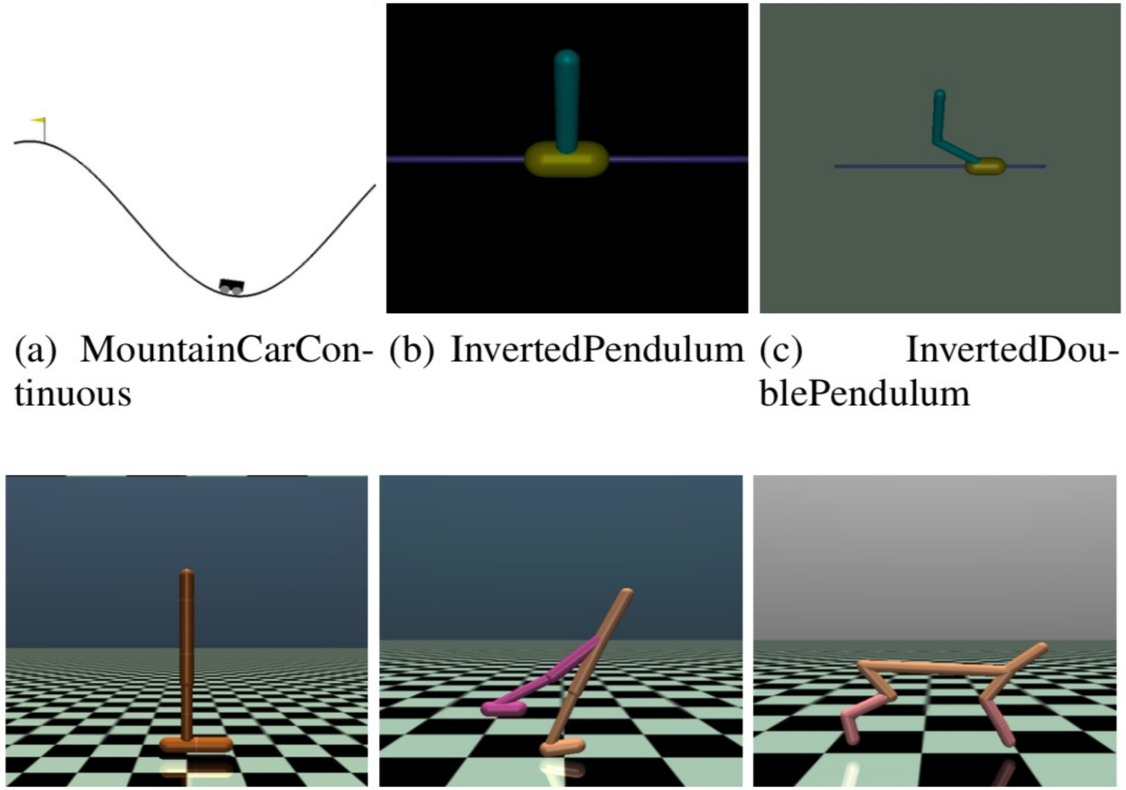
GAIfO with Proprioceptive Information

- Propose an algorithm that uses both proprioceptive and visual information in order to:
  - Improver performance
  - Improve sample complexity



Experiments

- Tasks:
  - OpenAI Gym Environments
- Visual Demonstrations:
  - 64*64 grayscale frames

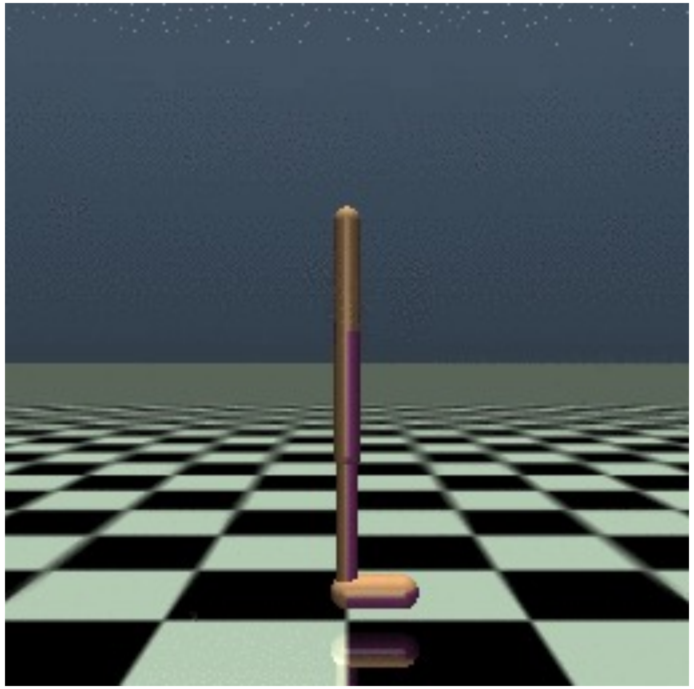


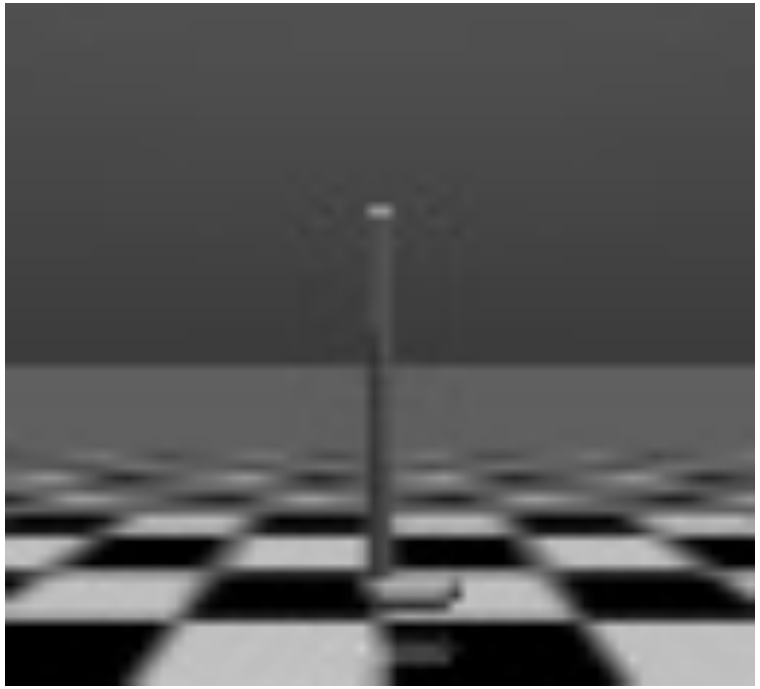
(a) MountainCarContinuous (b) InvertedPendulum (c) InvertedDoublePendulum

(d) Hopper (e) Walker2d (f) HalfCheetah

Experiments

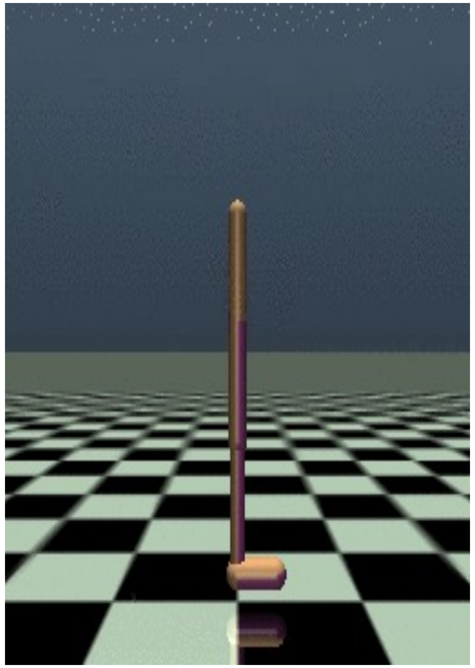
- Walker2D Expert:

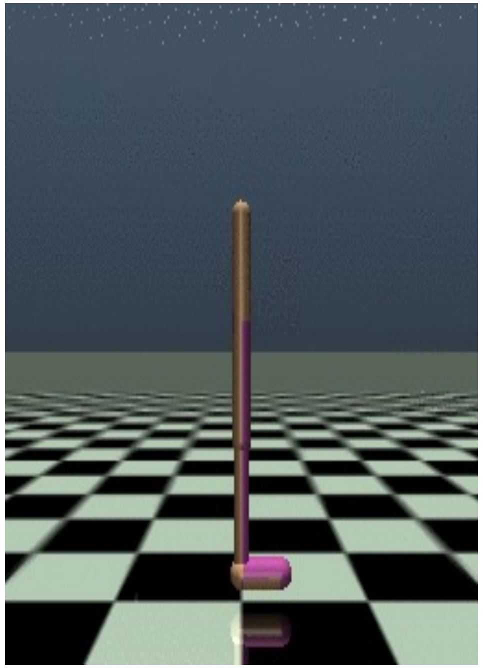


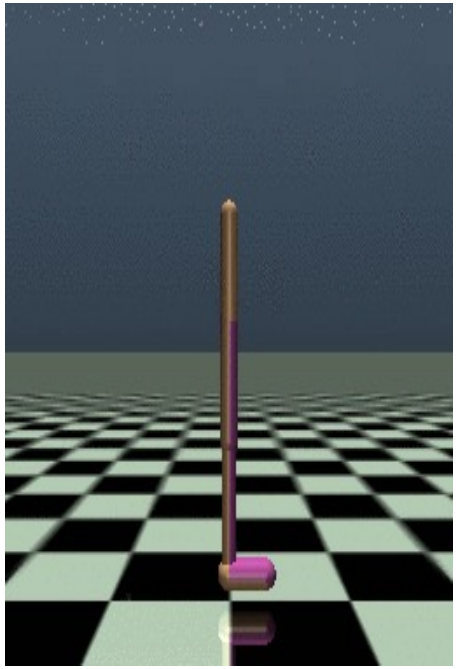


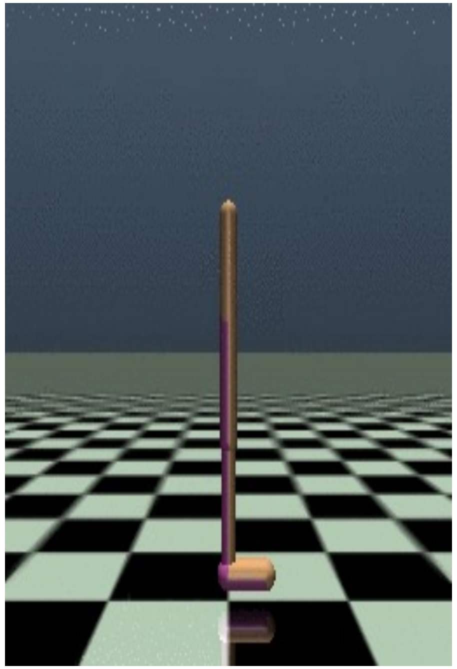
Experiments

- Walker2D


Expert

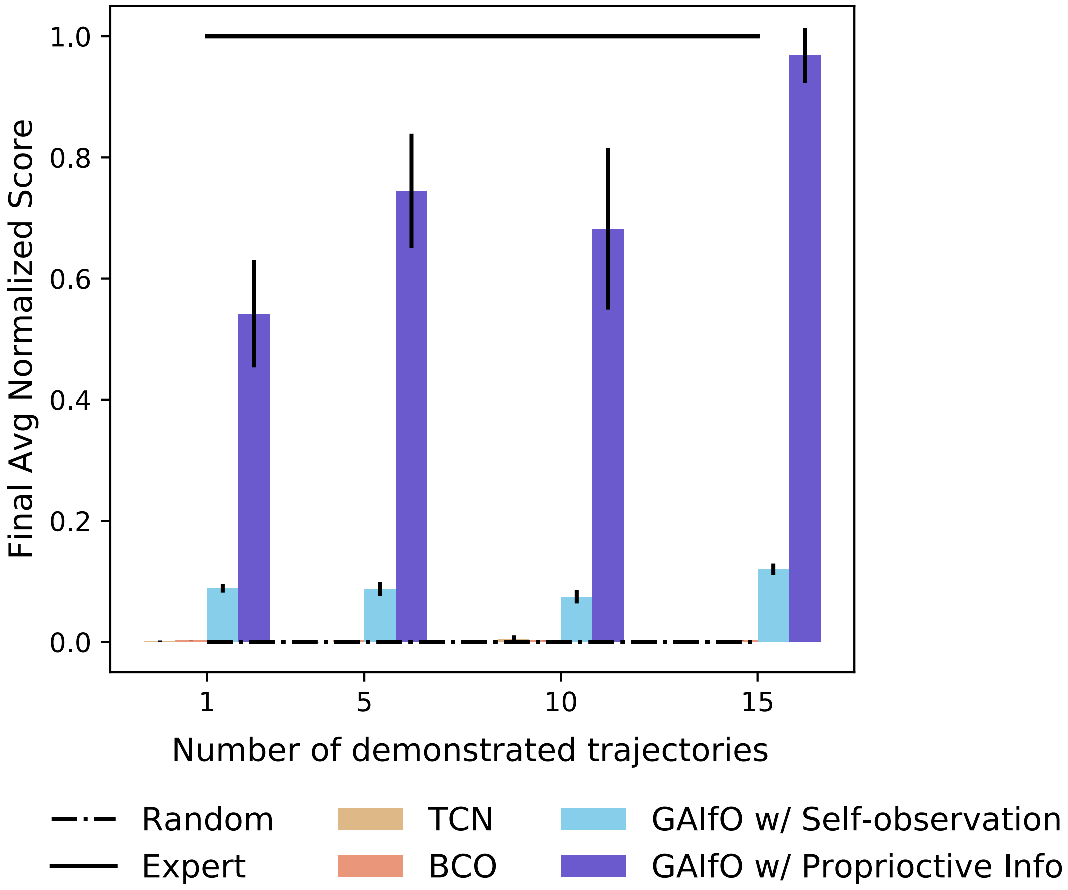

Iteration 0


Iteration 100


Iteration 1942

Experiments

- Walker2D



In what ways can autonomous agents learn to imitate experts using state-only observations?

- Behavioral Cloning from Observation (BCO) [IJCAI 2018]
- Generative Adversarial Imitation from Observation (GAIFO) [AAMAS 2019, ICML Workshop]
  - Visual Extension of GAIfO with Self-observation [ICML Workshop]
  - Visual Extension of GAIfO with Proprioceptive Information [IJCAI 2020]
- Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO) [Under Review]
- Reinforced Inverse Dynamics Modeling (RIDM) [RAL, IROS 2020]

Motivation

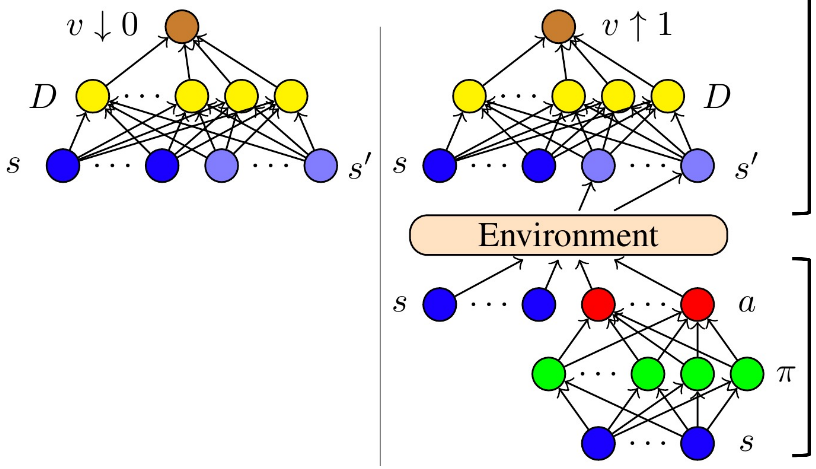
- Goal:
  - Improve ***sample complexity*** of ***GAIfO*** to enable ***application on physical robots***.
- Integrating
  - Sample-efficient RL updates from PILQR [1] with
  - High-performing GAIfO algorithm for IfO.

[1] Chebotar, Yevgen, et al. "Combining model-based and model-free updates for trajectory-centric reinforcement learning." *International conference on machine learning*. PMLR, 2017.

PILQR

- Combines:
  - iterative Linear Quadratic Regulator (iLQR)
  - And, Path Integral Policy Improvement (PI^2)
- iLQR constraints:
  - Linear dynamics $s_{t+1} = F_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + f_t$
  - Quadratic cost function $c(s_t, a_t) = \frac{1}{2} \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T C_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T c_t$
- PILQR:
  - Twice differentiable-cost function
  - iLQR on quadratic approximation of the cost
  - PI^2 policy update on the residual cost
  - Returns a Gaussian controller $p(a|s)$

~~GAIfO~~ DEALIO



Demonstrator

Imitator

Some constraints on the cost

~~Used as cost function~~

PILQR

~~PPO~~ used to update

Twice-differentiable

A function of both states and actions

- We consider:

$$c(s_t, a_t) = \underbrace{\frac{1}{2}\begin{bmatrix} s_t \\ a_t \end{bmatrix}^T C_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T c_t + cc_t}_{\text{Quadratic approximation}}$$

DEALIO

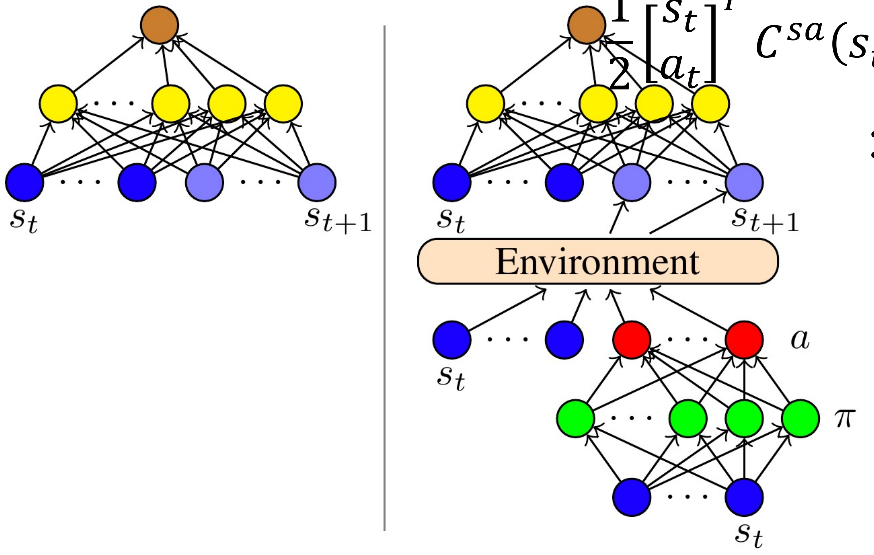
$$c(s_t, a_t) = \frac{1}{2}\begin{bmatrix} s_t \\ a_t \end{bmatrix}^T C_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T c_t + cc_t$$

$v \downarrow 0$ $\quad$ $v \uparrow 1$

$$D_\theta(s_t, s_{t+1}) = \frac{1}{2}\begin{bmatrix} s_t \\ s_{t+1} \end{bmatrix}^T C^{ss}(s_t, s_{t+1}) \begin{bmatrix} s_t \\ s_{t+1} \end{bmatrix} + \begin{bmatrix} s_t \\ s_{t+1} \end{bmatrix}^T c^{ss}(s_t, s_{t+1}) + cc^{ss}(s_t, s_{t+1})$$

$C^{ss}(s_t, s_{t+1}), c^{ss}(s_t, s_{t+1}), cc^{ss}(s_t, s_{t+1})$

$v \downarrow 0$ $\quad$ $v \uparrow 1$



- Having the dynamics model

$$s_{t+1} = F_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + f_t$$

- Substitute s_{t+1} in $D_\theta(s_t, s_{t+1})$ to find

$$\frac{1}{2}\begin{bmatrix} s_t \\ a_t \end{bmatrix}^T C^{sa}(s_t, s_{t+1}) \begin{bmatrix} s_t \\ a_t \end{bmatrix} + \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T c^{sa}(s_t, s_{t+1}) + cc^{ss}(s_t, s_{t+1})$$

$$\vdots$$

$$C_t = C^{sa}(\overline{s_t}, \overline{s_{t+1}})$$

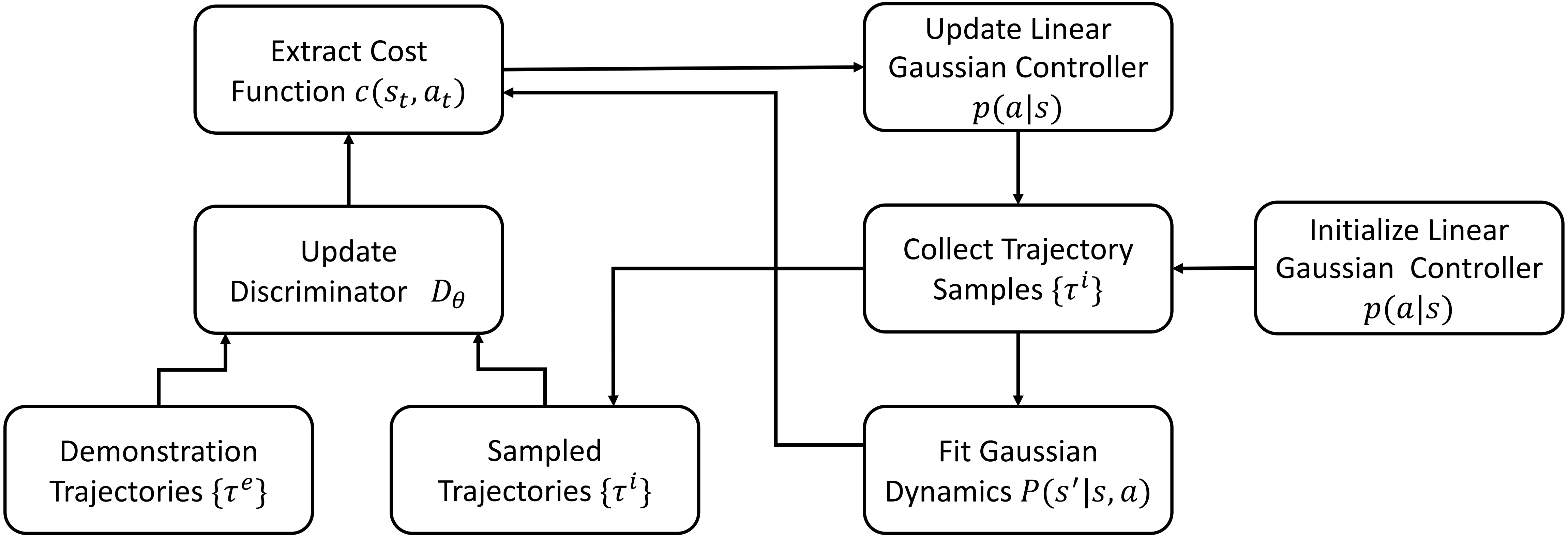
$$c_t = c^{sa}(\overline{s_t}, \overline{s_{t+1}})$$

$$cc_t = cc^{ss}(\overline{s_t}, \overline{s_{t+1}})$$

DEALIO

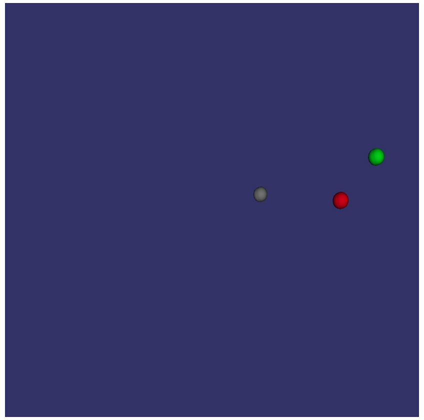
$$D_\theta(s_t, s_{t+1}) = \frac{1}{2}\Big[\dots \qquad c^{ss}(s_t, s_{t+1})$$

$$c(s_t, a_t) = \frac{1}{2}\begin{bmatrix} s_t \\ a_t \end{bmatrix}^T C_t \begin{bmatrix} s_t \\ a_t \end{bmatrix} + \begin{bmatrix} s_t \\ a_t \end{bmatrix}^T c_t + cc_t$$

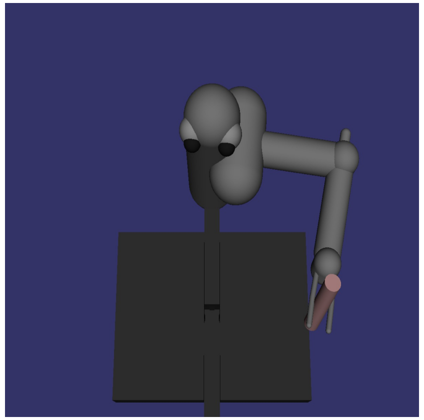


Experiments

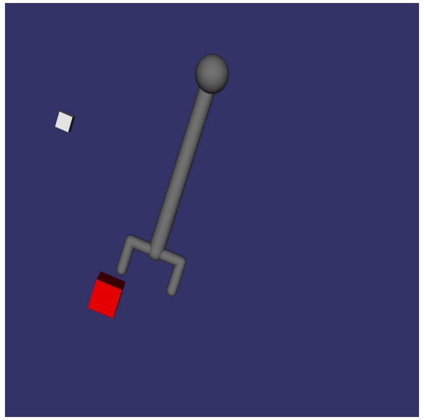
- MuJoCo Simulation Domains:



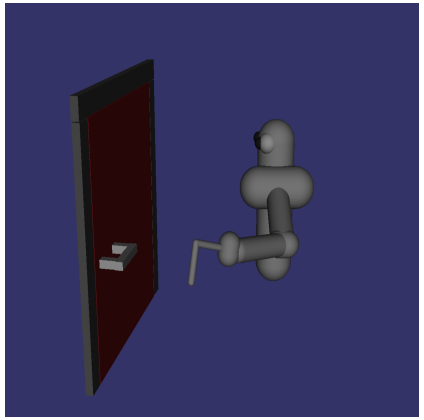
Disc



PegInsertion



GripperPusher

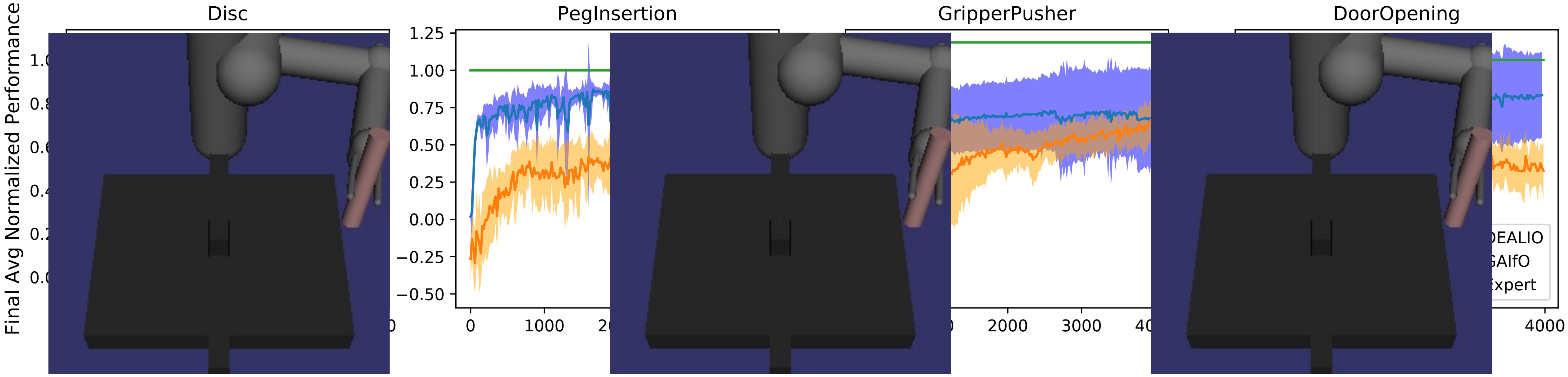


DoorOpening

- Hypothesis:
  - DEALIO is able to learn tasks efficiently compared to GAIfO
  - DEALIO is able to perform better compared to GAIfO

Experiments

- Our experiments on MuJoCo show DEALIO is faster in learning and has higher performance compared to GAIfO.
- PegInsertion:



Demonstration GAIfO DEALIO

In what ways can autonomous agents learn to imitate experts using state-only observations?

Behavioral Cloning from Observation (BCO)
[IJCAI 2018]

Generative Adversarial Imitation from Observation (GAIFO)
[AAMAS 2019, ICML Workshop]

Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO)
[Under Review]

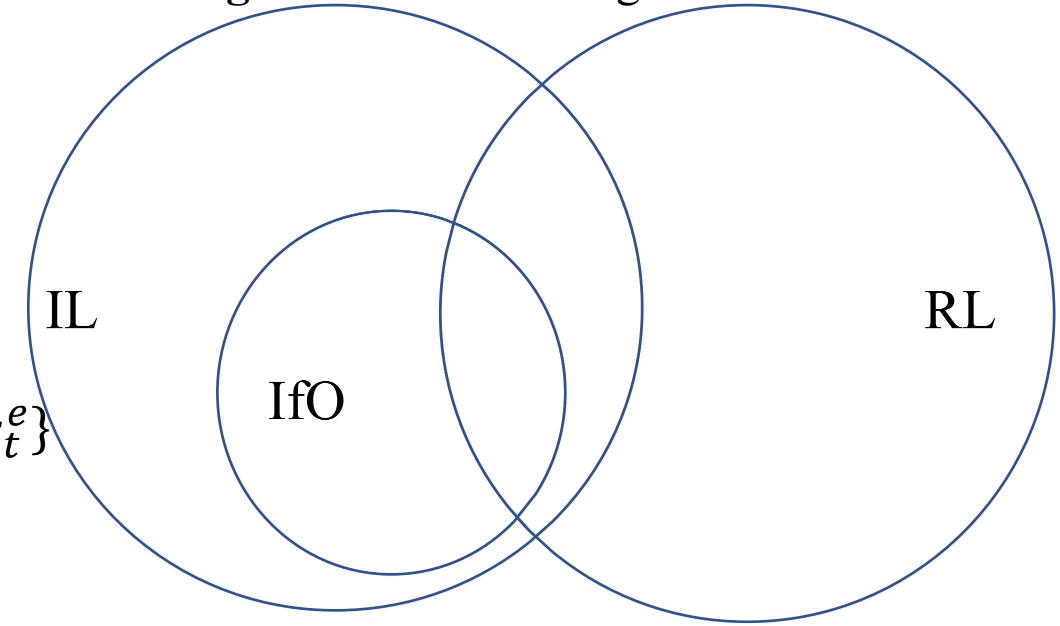
Reinforced Inverse Dynamics Modeling (RIDM)
[RAL, IROS 2020]

Visual Extension of GAIfO with Self-observation
[ICML Workshop]

Visual Extension of GAIfO with Proprioceptive Information
[IJCAI 2020]

Motivation

- Goal:
  - Combine "**imitation from observation**" and "**reinforcement learning**" to enable learning when:
    - The demonstrator is sub-optimal
    - Not many demonstration trajectories are available



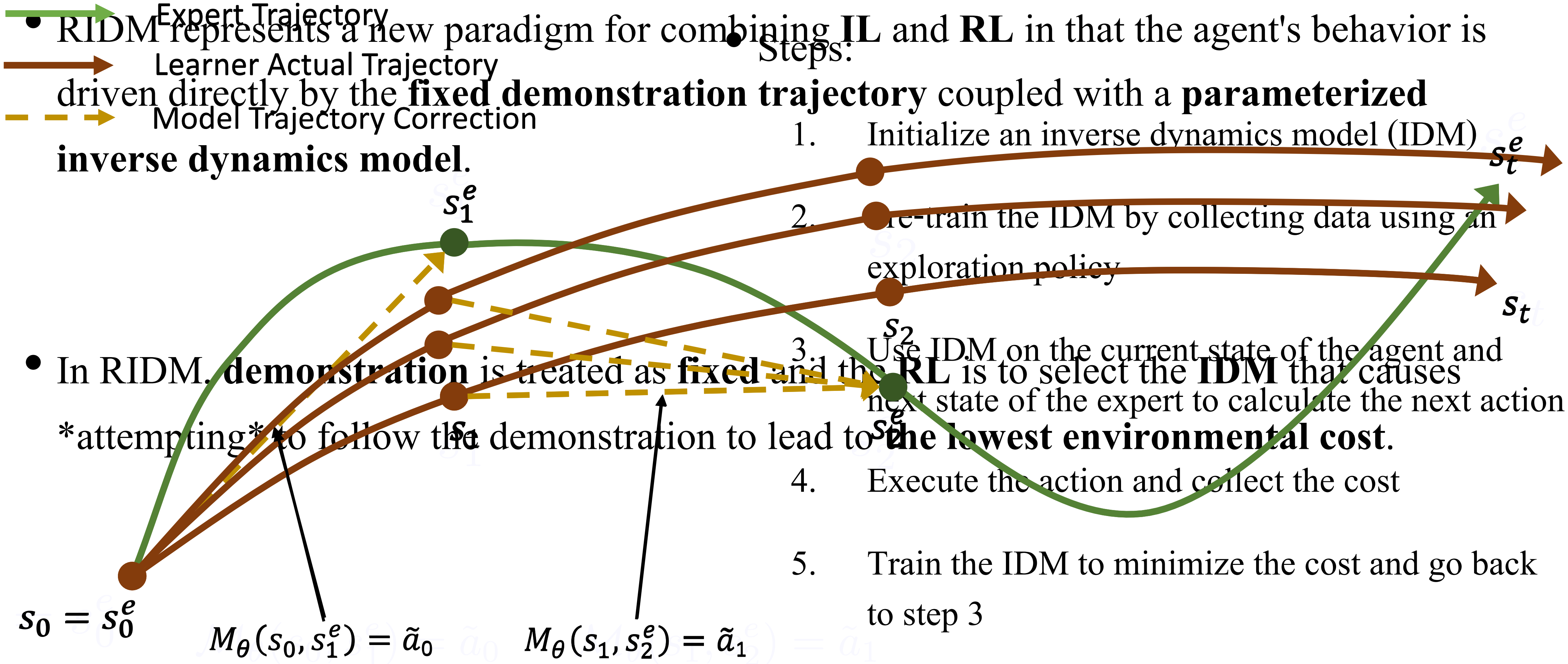
- **Given:**
  - A single state-only (sub-optimal) demonstration: $D^e = \{s_t^e\}$
  - A cost function: c_{env}
- **Learn:**
  - A policy to perform the task

RIDM: Reinforced Inverse Dynamics Modeling

- RIDM represents a new paradigm for combining **IL** and **RL** in that the agent's behavior is driven directly by the **fixed demonstration trajectory** coupled with a **parameterized inverse dynamics model**.

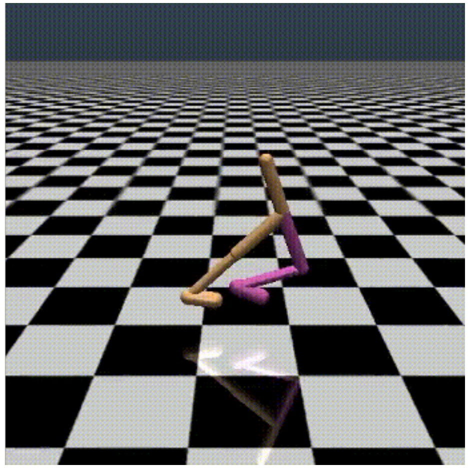
- In RIDM, **demonstration** is treated as **fixed** and the **RL** is to select the **IDM** that causes *attempting* to follow the demonstration to lead to **the lowest environmental cost**.

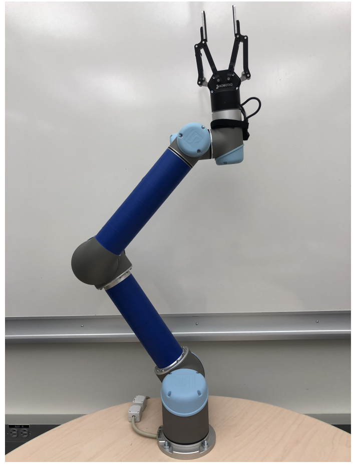
- Steps:
  1. Initialize an inverse dynamics model (IDM)
  2. Pre-train the IDM by collecting data using an exploration policy
  3. Use IDM on the current state of the agent and next state of the expert to calculate the next action
  4. Execute the action and collect the cost
  5. Train the IDM to minimize the cost and go back to step 3



Experiments

- Robot control domains:
  - MuJoCo Simulator
  - SimSpark Simulator
  - UR5 Arm Robot





- Hypothesis:
  - RIDM is able to learn tasks efficiently with comparable performance compared to the demonstrator
  - If the demonstrator is sub-optimal, RIDM is potentially able to outperform the demonstrator

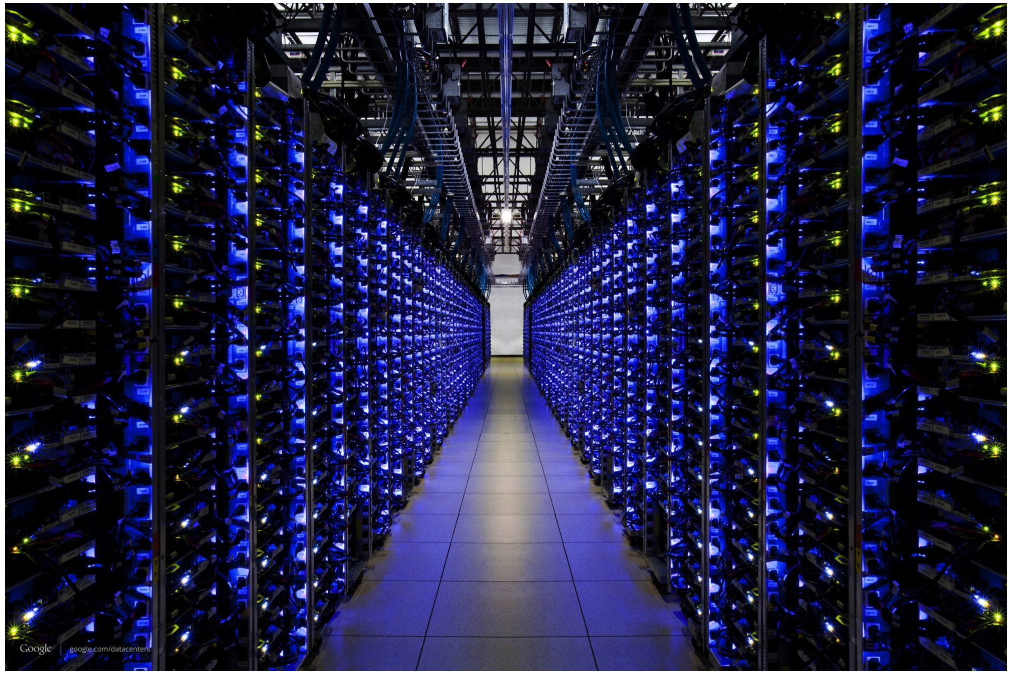
Experiments

- Computationally challenging:
  - Hopper: 4.5 days
  - Nao’s Fast Walk: 2.5 days

Parallelized Over 50 Machines

↓

2-3 hours each experiment



Experiments: SimSpark Robot Soccer

- Used in 3D Simulation RoboCup
- Developing skills such as walk and kick is challenging
- Tasks:
  - Fast Walk
  - Long Kick
- Demonstrators:
  - FUT-K
  - FC Portugal

- Demonstrators are sub-optimal with respect to the designed cost.

Experiments: SimSpark Robot Soccer

- Our experiments on SimSpark 3D simulator show learned behavior outperforms the suboptimal experts.
- Fast Walk (FUT-K): / Long Kick (FUT-K):

Learned Behavior:

Demo:

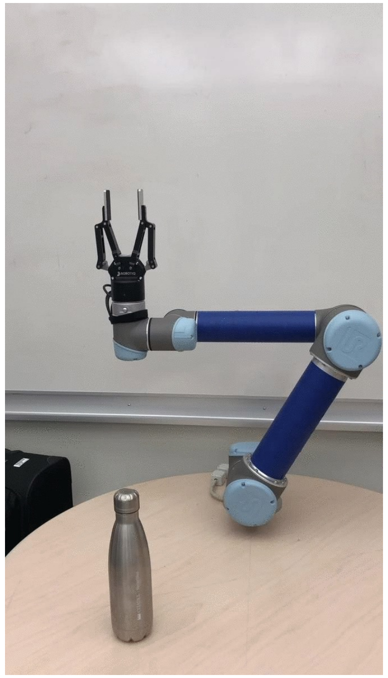
- Fast Walk (FC Portugal): / Long Kick (FC Portugal):

Learned Behavior:

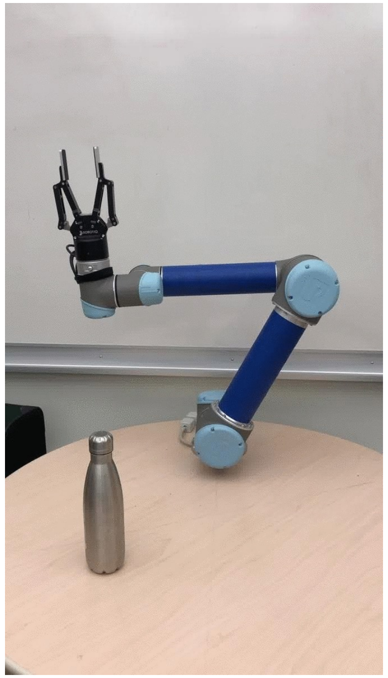
Demo:

Experiments: UR5 Arm Robot

- Our experiments on UR5 robots show learned behavior outperforms robot's default PID performance.
- Pushing Task (10x):



Demonstration

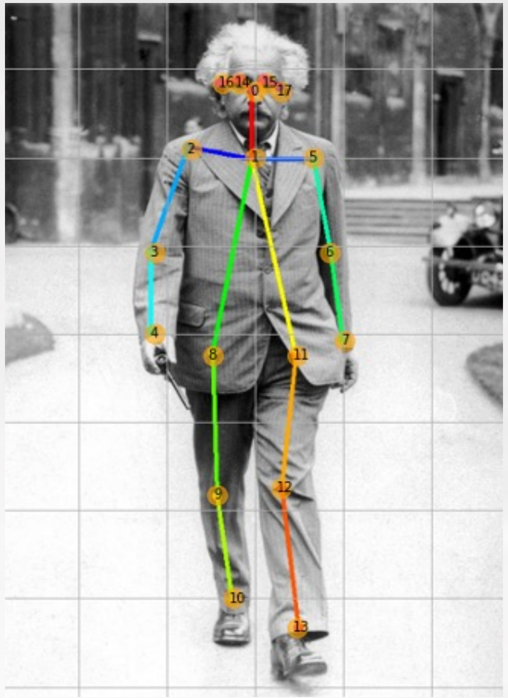


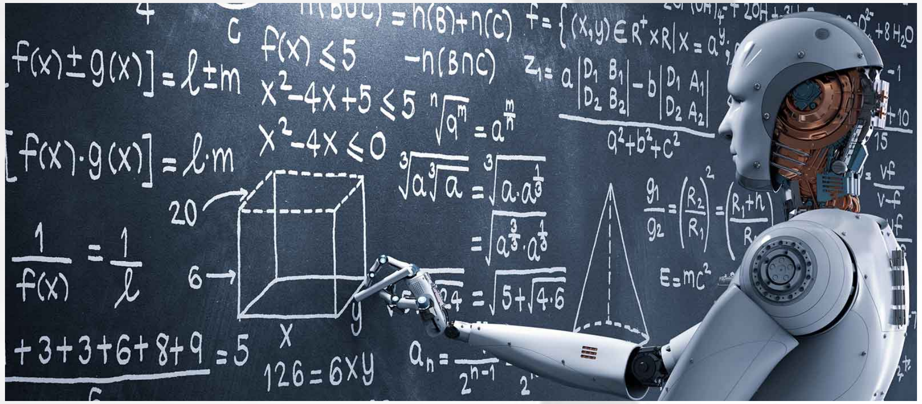
Default UR5 PID

Learned Behavior

Future Work

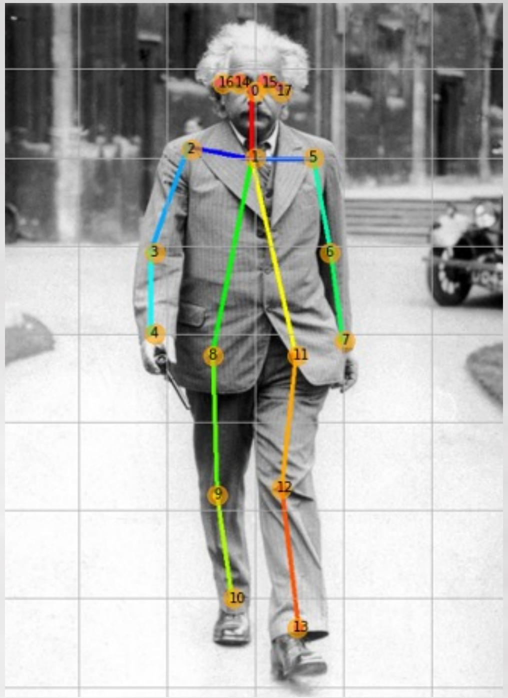
- Perception
- Application to Physical Robots
- Fully-intelligent Agents





Future Work

- Perception Challenges



Integration of Perception and Control

Embodiment Mismatch

CycleGAN [Zhu et al. 2017]

Pix2pix [Isola et al. 2017]

Dual GAN [Yi et al. 2017]

Disco GAN [Kim et al. 2017]

Viewpoint Mismatch

Pose-estimation [Cao et al. 2017, Wang et al. 2019]

Keypoint Detection

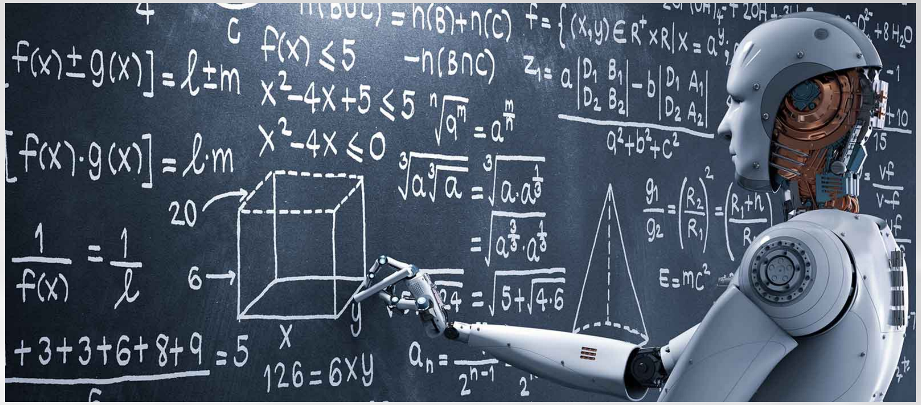
Future Work

- Application to Physical Robots

Sample-efficiency

Safe

Future Work



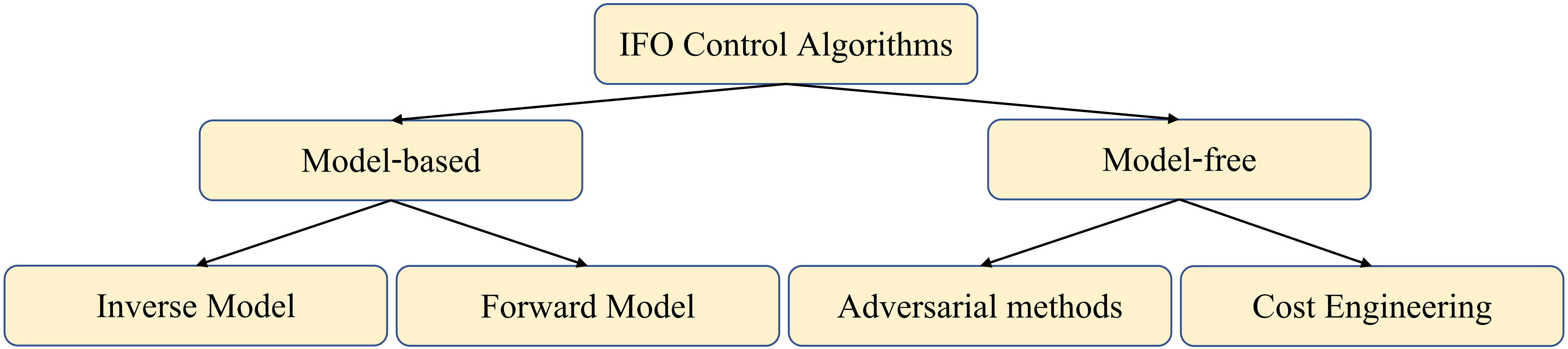
- Fully-intelligent Agents

Reinforcement learning

Imitation Learning

Imitation from Observation

Related Work



Self-supervised [Nair et al. 2017]
GSP [Pathak et al. 2018]
BCO* [Robertson et al. 2020]
A hybrid method [Guo et al. 2019]

ILPO [Edwards et al. 2019]
Future Image Similarity [Wu et al. 2020]

Third Person [Stadie et al. 2018]
OptionGAN [Henderson et al. 2018]
FAIL [Sun et al. 2019]
IDDM [Yang et al. 2019]

TCN [Sermanet et al. 2018]
IFO [Liu et al. 2020]
TDC [Aytar et al. 2018]
Invariant Feature Space [Gupta et al. 2017]

Summary

- Area A:
  - Equivalency of solving the model-free IfO problem and solving the GANs like optimization problem.
- Area B:
  - Implementation of the introduced algorithms.
  - Training the models on hundreds of machines.
  - Extensive hyperparameter search for each algorithm.
- Area C:
  - Modeling the human ability of imitation from observation.
  - Application of the developed algorithms to simulated and physical robots.

Acknowledgements

Peter Stone

Garrett Warnell

Patrick MacAlpine

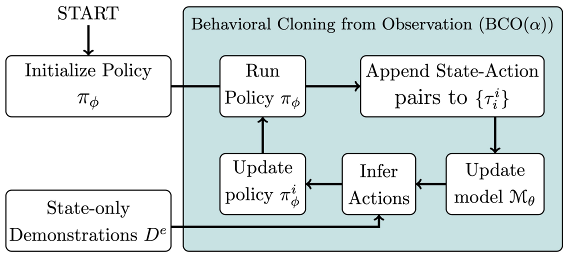
Josiah Hanna

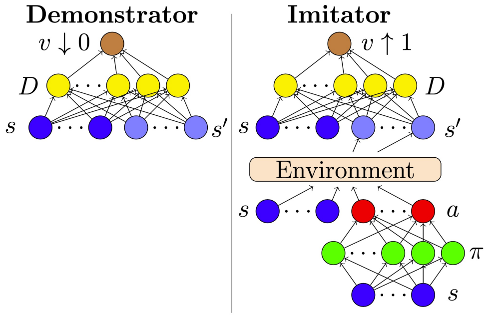
Brahma Pavse

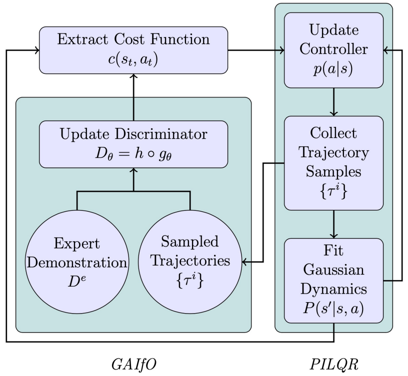
Ruohan Zhang

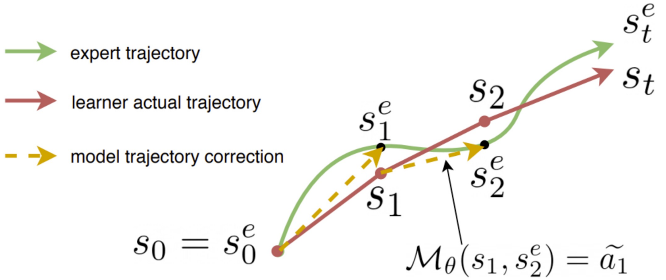
Sean Geiger

In what ways can autonomous agents learn to imitate experts using state-only observations?







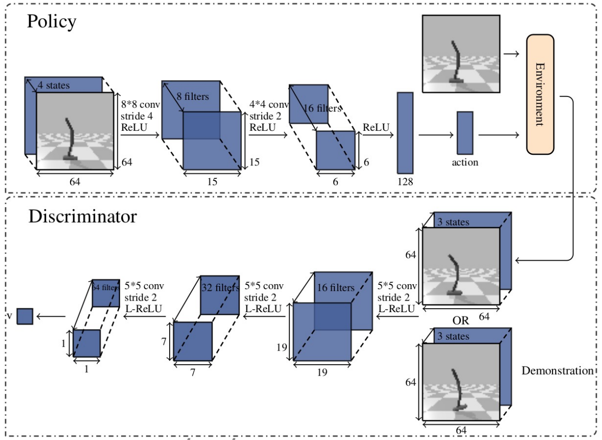


Behavioral Cloning from Observation (BCO)
[IJCAI 2018]

Generative Adversarial Imitation from Observation (GAIFO)
[AAMAS 2019, ICML Workshop]

Data-Efficient Adversarial Learning for Imitation from Observation (DEALIO)
[Under Review]

Reinforced Inverse Dynamics Modeling (RIDM)
[RAL, IROS 2020]



Visual Extension of GAIfO with Self-observation
[ICML Workshop]

Visual Extension of GAIfO with Proprioceptive Information
[IJCAI 2020]