1802.01557 - One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning

One-Shot Imitation from Observing Humans via Domain-Adaptive Meta-Learning

Tianhe Yu*, Chelsea Finn*, Annie Xie, Sudeep Dasari, Tianhao Zhang, Pieter Abbeel, Sergey Levine
University of California, Berkeley
Email: {tianhe.yu,cbfinn,anniexie,sdasari,tianhao.z,pabbeel,svlevine}@berkeley.edu
* denotes equal contribution

- **Yunqiu Xu**
- MAML → MIL → this work
    - Other ref: 机器人模仿人类动作一学就会，还能举一反三了 | 论文
    - 之前搞MIL时有提及future work: 接受人类操作视频作为demonstration, 从而实现one-shot imitation via meta learning
    - 类似sim-to-real, 接受人类操作视频, 也需要考虑substantial domain shift (perspective, environment, embodiment)
    - 前人工作: 手动指定相关性, explicit human pose detection
    - 本工作: 通过meta learning学习prior knowledge, 然后在新任务实现one-shot
    - 实验: 机器人手臂 (place, push, pick-and-place)

---

1. Introduction

- Challenge:
    - 机器人模仿学习与人/动物的学习过程还存在很大不同
        - 机器人需要的demonstration形式为 kinesthetic teaching 或者 teleoperation (遥控)
        - 而人类仅仅需要观看他人即可进行模仿, 且仅仅需要很少的demonstrations
    - 直接从raw visual observations进行学习存在的挑战

        - Systematic domain shift
        - A substantial amount of data
    - Prior work: 手动指定机器和人类动作行为的相关性 → 复杂, 且动作机理有区别

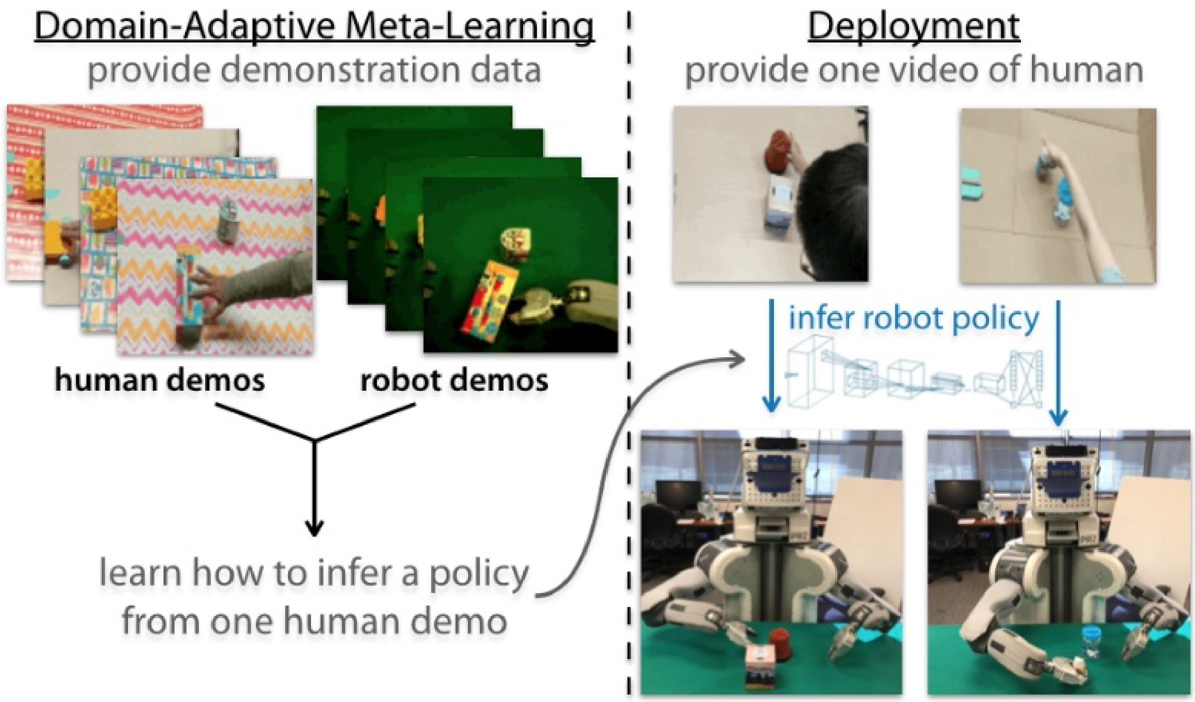

Fig. 1. After meta-learning with human and robot demonstration data, the robot learns to recognize and push a new object from one video of a human.

- Our work: let robot learn how to learn from humans from similar tasks
    - 首先利用meta-training 学习 rich prior knowledge, 其中demonstration包括人类的和遥控的
    - 机器人学习如何利用数据模仿人类动作
    - meta-training结束后, 基于之前学到的prior knowledge, 机器人可以实现one-shot imitation learning, with only one human video demonstration
- Contribution: 应该是MIL的更进一步吧, 之前MIL是MAML和模仿学习的一个结合, 现在扩宽了模仿学习中demonstration的范围, 便于构建数据集

2. Related Work

- Imitation learning

Prior work	Our work
Configuration-space trajectories level : kinesthetic teaching / teleoperation / sensors on the demonstrator	Imitate by watching human demonstrator
Manually resolve correspondence problem	Learn the correspondence implicitly
Explicit hand tracking / precise visual recognition system	Extract human's activity that are the most relevant for choosing actions
Explicitly determine the goal and reward, then optimize via inverseRL	Their work can not handle one-shot learning

- Meta learning:
  - Learning from similar tasks
  - 本工作可以认为是MAML / MIL的延伸 → MAML with domain shift between training and testing demonstration (e.g. learning from human videos)
- Domain shift:
  - Method 1 : domain adaptation
    - Find a representation that is domain invariant
    - Vary visual domains and sim-to-real
  - Method 2 : map datapoints from one domain to another
  - Human imitation problem:
    - Developing invariances: 光影 / 背景的变化
    - **人类和机器人行为间的physical correspondence不是invariant的, 也没法直接进行域间迁移 → 因此我们需要从视频中隐式识别人类行为的目标, 并选取相应动作**

3. Learning from Humans

3.1 Problem Setup

- What prior knowledge should we learn:

- Visual and physical understanding of the world
  - What kinds of outcomes the human want to achieve
  - Which actions can robot choose to get the outcome
- Demonstrations
  - Human $d^h =< o_1, \dots, o_T >$, a sequence of image observations
  - Robot $d^r =< o_1, s_1, a_1, \dots, o_T, s_T, a_T >$, image observations , robot states and robot actions
  - No assumptions about the similarty between human and robot demonstrations $\rightarrow$ can be different appearance of arms, background clutter and cameta viewpoint
- Two phases of our approach
  - Meta-learning phase:
    - Try to learn prior knowledge over tasks using both human and robot demonstration
    - For each training task T_i , we have demonstration datasets $(D^h_{T_i}, D^r_{T_i})$
  - Testing phase:
    - Combine prior knowledge with one human demonstration
    - Try to infer policy parameters ϕ_T to solve the new task

3.2 Domain-Adaptive Meta-Learning

- Compared with MIL, we can not use standard imitation learning loss, since human actions are inaccessible or can not correspond to robot's actions directly
- Adaptation objective L_ψ:
  - Does not need actions
  - Operates only on the policy actions
- Meta training
  - Learn both policy parameters θ and adaptation parameters ψ : **used for choosing actions to match robot demonstrations in D_T^{val}**
  - Imitation learning (behavioral cloning) objective
$$L_{BC}(\phi, d^r) = L_{BC}(\phi, \{o_{1:T}, s_{1:T}, a_{1:T}\}) = \sum_t log\pi_\phi(a_t|o_t, s_t)$$

- Combine L_{BC} with inner GD adaptation $\rightarrow$ meta-training objective

$$min_{\theta,\psi} \sum_{T\sim p(T)} \sum_{d^h\sim D^h_T} \sum_{d^r\sim D^r_T} L_{BC}(\theta - \alpha\nabla_\theta L_\psi(\theta, d^h), d^r)$$

Algorithm 1 Meta-imitation learning from humans

Require: $\{(\mathcal{D}^h_{\mathcal{T}_i}, \mathcal{D}^r_{\mathcal{T}_i})\}$: human and robot demonstration data for a set of tasks $\{\mathcal{T}_i\}$ drawn from $p(\mathcal{T})$
Require: α, β: inner and outer step size hyperparameters
while training **do**
Sample task $\mathcal{T} \sim p(\mathcal{T})$ {or minibatch of tasks}
Sample video of human $\mathbf{d}^h \sim \mathcal{D}^h_{\mathcal{T}}$
Compute policy parameters $\phi_{\mathcal{T}} = \theta - \alpha\nabla_\theta\mathcal{L}_\psi(\theta, \mathbf{d}^h)$
Sample robot demo $\mathbf{d}^r \sim \mathcal{D}^r_{\mathcal{T}}$
Update $(\theta,\psi) \leftarrow (\theta,\psi) - \beta\nabla_{\theta,\psi}\mathcal{L}_{\text{BC}}(\phi_{\mathcal{T}}, \mathbf{d}^r)$
end while
Return θ, ψ

- Meta testing on a new task T
  - Given a human demonstration d^h
  - Use gradient descent starting from θ with learned loss L_ψ to infer new policy parameters

$$\phi_T = \theta - \alpha\nabla_\theta L_\psi(\theta, d^h)$$

Algorithm 2 Learning from human video after meta-learning

Require: meta-learned initial policy parameters θ
Require: learned adaptation objective $\mathcal{L}_\psi$
Require: one video of human demo $\mathbf{d}^h$ for new task $\mathcal{T}$
Compute policy parameters $\phi_{\mathcal{T}} = \theta - \alpha\nabla_\theta\mathcal{L}_\psi(\theta, \mathbf{d}^h)$
return π_ϕ

3.3 Learned Temporal Adaptation Objectives

- Why we need this adaptation objective :
  - 该目标可从人类视频中捕捉有用信息, 如intention或者和任务相关的对象
  - 且可以在不获取真实动作的前提下提供合适的梯度信息 → 这对于传统的BC loss (frame之间彼此独立)太难了
  - 确定demonstrate哪种行为以及哪些对象是需要的, 需要同时检测多个frame以确定人类动作
- 我们需要引入**temporal convolutions** 来表示adaption objective L_ψ (相关文献: 1609.03499 Wavenet: A generative model for raw audio)
  - Use multiple layers of 1D convlutions over time
  - Effective at processing temporal and sequential data
  - 我们的改进: 用类似LSTM的方式使用temporal convolutions

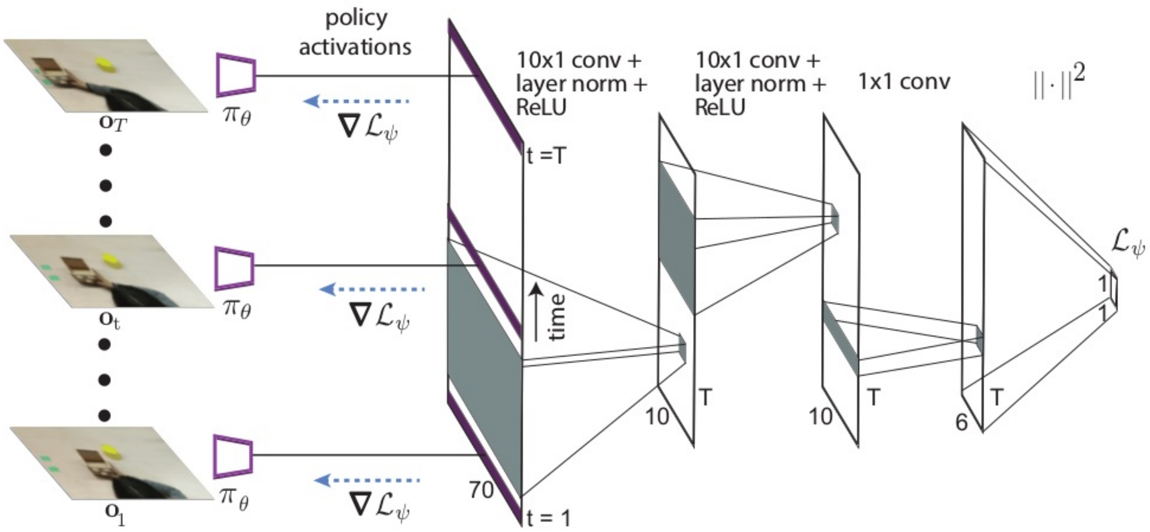

Fig. 2. Visualization of the learned adaptation objective, which uses temporal convolutions to integrate temporal information in the video demonstration.

- MIL引入了双头结构用于one-shot imitation
  - 一个头 pre-update demonstration, 另一个头 post-update policy
  - 双头结构可被解读为在最后一层网络的某种linear loss function, 该函数生效于特定timestep
  - 计算loss和梯度: averaging over all timesteps in the demonstration
  - 在本工作中, single timestep是不够的, **因此我们工作比之前的双头结构更好?**

3.4 Probabilistic Interpretation

- Adaptation的意义: GD on learned loss $L_\psi(\phi, D^{tr}_T)$, rather than likelihood $logp(D^{tr}_T|\phi)$

$$p(\phi|\mathcal{D}^{\text{tr}}_{\mathcal{T}},\theta) \propto p(\phi,\mathcal{D}^{\text{tr}}_{\mathcal{T}}|\theta) \propto \underbrace{p(\phi|\theta)}_{\text{from GD}} \quad \underbrace{\Psi(\phi,\mathcal{D}^{\text{tr}}_{\mathcal{T}})}_{\exp(-\mathcal{L}_\psi(\phi,\mathcal{D}^{\text{tr}}_{\mathcal{T}}))} .$$

- Visual illustration of the graphical model

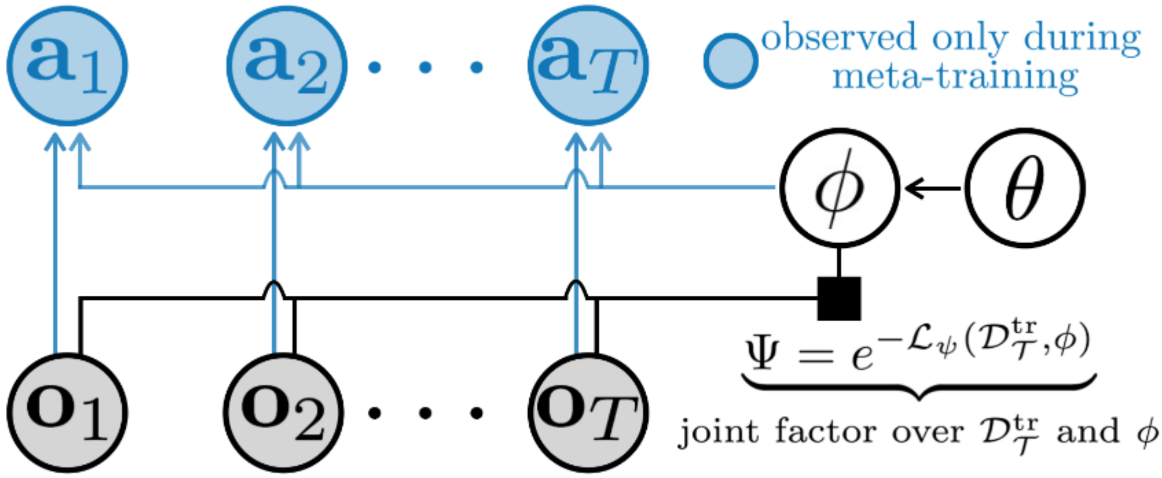

Fig. 3. Graphical model underlying our approach. During meta-training, both the observations $\mathbf{o}_t$ and the actions $\mathbf{a}_t$ are observed, and our method learns θ and Ψ. During meta-testing, only the observations are available, from which our method combines with the learned prior θ and factor Ψ to infer the task-specific policy parameters ϕ.

4. Network Architectures

4.1 Policy π

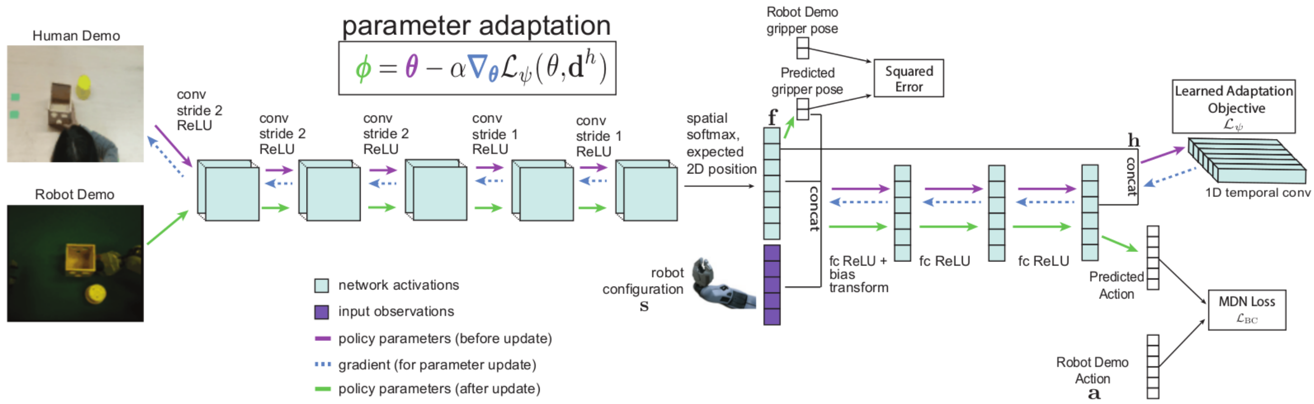

Fig. 4. Illustration of the policy architecture. The policy consists of a sequence of five convolutional (conv) layers, followed by a spatial soft-argmax and fully-connected (fc) layers. The learned adaptation objective $\mathcal{L}_\psi$ is further illustrated in Figure 2. Best viewed in color.

4.2 Learned Adaptation Objective L_ψ

- We need to update both perception and control
- 将预测特征 f 与policy最后一层隐藏层 h 相连构建adaptation objective
- 因此learned loss可以越过控制层被直接应用于卷积层的权重
- 我们用之前构建的temporal adaptation objective更新关于这个任务的策略权重

$$\phi = \theta - \alpha\nabla_\theta L_\psi(\theta, d^h)$$

- Adaptation objective 会被解构成两部分

$$L_\psi = L_{\psi_1}(f_{1:T}) + L_{\psi_2}(h_{1:T})$$

  - L_{ψ_1} 与 L_{ψ_2} 网络结构相同 (Fig 2)

5. Experiments

- Questions:
  - 1. 我们的方法能否有效达到one-shot imitation with human video?
  - 1. 给定新的human demonstrator, 我们的方法能否以不同于机器人的视角泛化于human demonstration
  - 1. 和其他meta-learning方法的性能比较
  - 1. Temporal adaptation objective对于我们工作的重要性
- Baselines:
  - Contextual policy: 输入机器观察值及人类视频的最后一帧(task), 输出预测的动作
  - DA-LSTM policy: RNN, 直接输入人类视频和机器观察值, 输出预测的动作, domain-

adaptive version of Duan's work (One-shot Imitation Learning)
  - DAML, linear loss: our work with linear per-timestep adaptation obective
  - DAML, temporal loss

5.1 PR2 Placing, Pushing, and Pick & Place

Fig. 5. Example placing (left), pushing (middle), and pick-and-place (right) tasks, from the robot's perspective. The top row shows the human demonstrations used in Section VI-A while the bottom shows the robot demonstration.

	placing	pushing	pick and place
DA-LSTM	33.3%	33.3%	5.6%
contextual	36.1%	16.7%	16.7%
DAML, linear loss	76.7%	27.8%	11.1%
DAML, temporal loss (ours)	**93.8%**	**88.9%**	**80.0%**

TABLE I. One-shot success rate of PR2 robot placing, pushing, and pick-and-place, using human demonstrations from the perspective of the robot. Evaluated using held-out objects and a novel human demonstrator.

5.2 Demonstrations with Large Domain Shift

- 不同房间环境, 不同视角

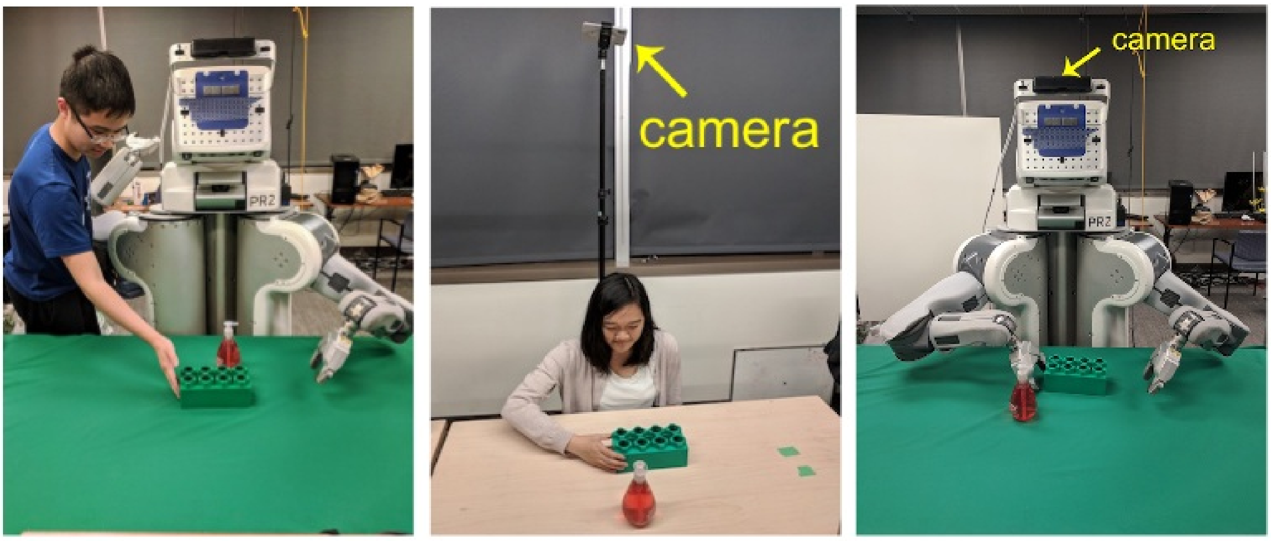

Fig. 6. The PR2 experimental set-up. Left & Middle: human demonstration set-up from Sections VI-A and VI-B respectively. Right: test-time set-up.

Fig. 8. Human and robot demonstrations used for meta-training for the experiments in Section VI-B with large domain shift. We used ten different diverse backgrounds for collecting human demonstrations.

Fig. 9. Frames from the human demos used for evaluation in Section VI-B, illustrating the background scenes. The leftmost background was in the meta-training set (seen bg), whereas the right two backgrounds are novel (novel bg1 and novel bg2). The objects and human demonstrator are novel.

pushing	seen bg	novel bg 1	novel bg 2
DAML, temporal loss (ours)	**81.8%**	**66.7%**	**72.7%**

Failure analysis of DAML	seen bg	novel bg 1	novel bg 2
# successes	27	22	24
# failures from task identification	1	5	4
# failures from control	5	6	5

TABLE II. Top: One-shot success rate of PR2 robot pushing, using videos of human demonstrations in a different scene and camera, with seen and novel backgrounds. Evaluated using held-out objects and a novel human. Bottom: Breakdown of the failure modes of our approach.

5.3 Sawyer Experiments

- 在不同机器人和不同机器demonstration集合类型中的泛化性

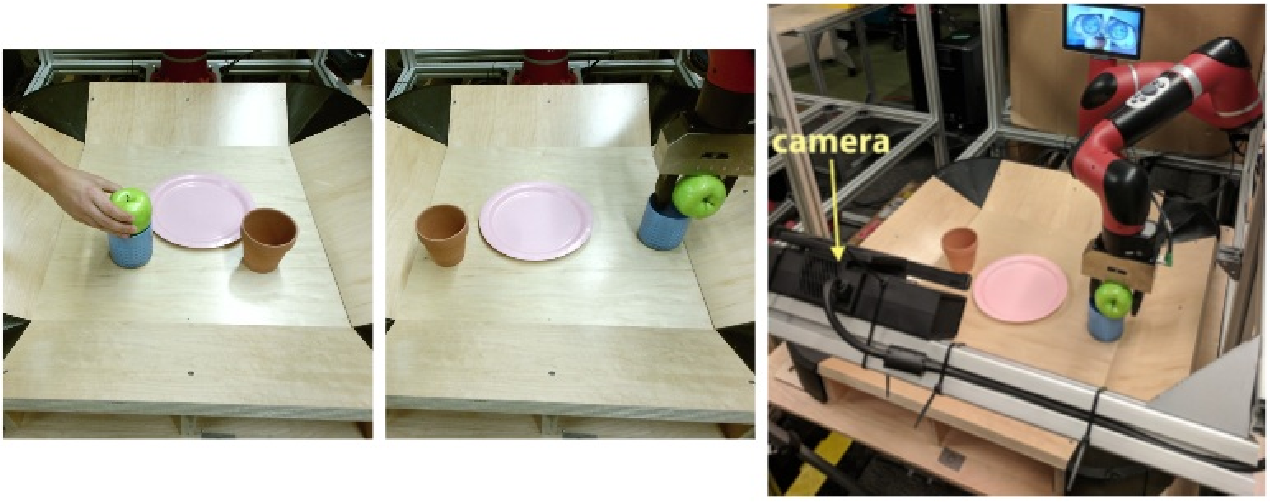

Fig. 10. Sawyer robot set-up. From left the right: a human demo from the robot's perspective, the policy execution from the robot's perspective, and an photo illustrating the experimental set-up.

5.4 Learned Adaptation Objective Ablation

	simulated pushing no domain shift
LSTM [10]	34.23%
contextual	56.98%
MIL, linear loss [15]	66.44%
MIL, temporal loss (ours)	**80.63%**

TABLE III. One-shot success rate of simulated 7-DoF pushing using video demonstrations with no domain shift

6. Summary

- 在MIL基础上进行了延伸, 通过构造adaptation objective, 令机器人可以接受人类视频实现one-shot imitation learning
- 不足:

- Meta-test 时的任务和训练时的任务很相似, 在未来的工作中, 我们希望能够处理 unseen objects and demonstrators
- The amount of demonstration per object is too low than general imitation learning, 在未来的工作中我们会尝试构建更泛化的机器人 (对一个任务可以有更多更宽泛的demonstration) **加噪音可否实现这一功能?**