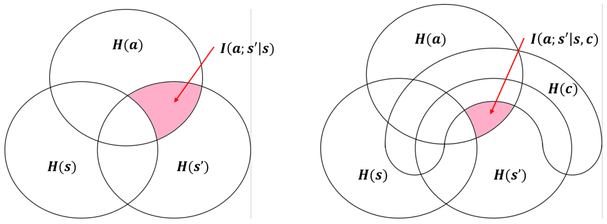
Learning Multi-Task Transferable Reward via Variational Inverse Reinforcement Learning

Se-Wook Yoo [1] Seung-Woo Seo [2]

***Abstract*— We consider a problem of learning the reward and policy from expert examples under multi-task environment with unknown dynamics. In this paper, We extend an empowerment-based regularization technique into multi-task situations on the purpose of restoring nearly-optimal multi-task reward function based on the framework of generative adversarial network. Many robotic tasks are composed of many temporally sustained and correlated sub-tasks under the highly complex environment. It is the key factor to discover situational intentions separated from changing dynamics and discover proper actions deliberating on temporal abstractions in order for effectively solving the problems. For understanding robust intention under unknown dynamics, we define the situational empowerment as the maximum of mutual information that represent how an action that conditioned on certain state and sub-task affects the future. Our proposed method derive the variational lower bound of the situational mutual information to optimize that. we simultaneously learn transferable multi-task reward function and policy by adding the induced term on the objective function. By doing that, the empowerment-based disentangled multi-task reward function helps to learn robust policy. We evaluate our approaches on various high-dimensional complex control tasks. we demonstrate the strong points of multi-task learning and multi-task transfer learning on the side of both randomness and robustness of changing task dynamics. Finally, We demonstrate significantly better performance and data-efficiency than existing imitation learning methods on the various benchmarks.**

I. INTRODUCTION

RECENTLY, in the field of robotics, the combination of Reinforcement Learning (RL) [1] has shown excellent performance on complex sequential decision-making and high-dimensional control tasks. [2]. When RL algorithm applies to various robotic applications such as dexterous manipulation or and autonomous driving, designing general reward function has a lots of challenges. To reduce the burden of designing reward, a paradigm of Learning from Demonstration(LfD) [3] has been developed. Many variants of Generative Adversarial Imitation Learning (GAIL) [4] algorithm successfully have been resolved the simple primitive tasks so far. Nevertheless, there are a variety of hardships to get the breakthrough on the complex multi-task environment. Most of realistic applications consist of multiple tasks that are necessary to set sub-tasks and select the appropriate actions according to predefined hierarchical structure. It is not only difficult to design general hierarchical relationship but also wasteful to label sub-tasks for each with predefined rules. Although a variant of GAIL [5] tried to find hierarchical policy depend on causal information flow in an unsupervised way, the learned policy did not adapt environment where it faces with similar but different cases not seen in the expert demonstration. This is because the discriminator simply compares trajectories sampled from expert demonstrations and generated trajectories. To break the limitation, Variational Inverse Reinforcement Learning (VIRL) based on empowerment-based regularization [6] was proposed. It could restore a robust reward function that depends on state and action pair to successfully resolve ambulation tasks. In spite of adjusting policy to transformed dynamics, it just handles a single task. we focus on the multi-task transfer learning problem and suggest the solution by extending previous works and unraveling the connection among them.



(a) Mutual Information (b) Situational Mutual Information

Fig. 1. **Left:** Information diagram for expowerment-based regularization technique used in EAIRL [6]. The variables are current state s, action a and next state s'. They use mutual information $I(a; s'|s)$ (pink region on (a)) as internal reward **Right:** We expand the mutual information (a) with $I(a; s'|s, c)$ (b) by introducing a new variable c meaning a sub-task.

To infer the sub-task in an unsupervised way without labeling cost, we adopt the architecture of Directed-Info GAIL (DIGAIL) developed from Options framework [8] by using the supervision that acquires from unsegmented demonstrations. However, DIGAIL only dealt with simple a few tasks under the informed dynamics. To handle more complex applications. we consider the relation of current state, action, sub-task and next state as shown information diagram in Fig. 1. The existing empowerment quantify how much a current action conditioned on current state affects the future state as the maximum of mutual information $I(a; s'|s)$. Specifically, The empowerment, the mutual information based theoretic measure, encourages the policy to explore more meaningful future states because it gives reward signals to be robust under unknown dynamics. Inspired by

*This work was supported by Seoul Nation University

[1] Se-Wook Yoo is Researcher with the Department of Electrical and Computer Engineering, Seoul National University, College of Engineering, Seoul, Republic of Korea tpdnr1360@snu.ac.kr

[2] Seung-Woo Seo is with Faculty of Electrical and Computer Engineering, Seoul National University, College of Engineering, Seoul, Republic of Korea sseo@snu.ac.kr

EAIRL, in order to learn reward function that disentangled with task transition dynamics, we redefine the expowerment as the maximum of mutual information $I(a;s'|s,c)$ called situational empowerment by introducing a latent variable c that indicates the current sub-task. After that, we learn the situational empowerment via maximization of the variational information through the tractable optimization explained in section III. Therefore, the learned empowerment prevents the policy from over-fitting expert demonstrations when the relation of task transition is changed. we validate the robustness of changing task dynamics in section IV.

As mentioned above, the causal confusion of task and behavior on a certain state hinders the multi-task reward function being learned. The proposed method to solve the above problems presents the following three main contributions. The first is to successfully learn a potential expression similar to human's temporal abstraction in an unsupervised way to increase the interpretability of deep learning. Secondly, by extending the existing empowerment-based regularization technique to a complex multiple task, it successfully restores the robust reward function conditional on a task separated from the dynamic environment. Finally, We show state-of-the-art performance and sampling complexity that exceed previous baselines in various experiment environment settings, which are composed of simple a few tasks, complex multiple tasks and transferable multiple tasks.

II. Preliminaries

A. Notation

We analyze the problem under the Markov Decision Process (MDP) expressed as tuple of $(\mathcal{S},\mathcal{A},\mathcal{P},\gamma,\mathcal{R})$ where $\mathcal{S}$ denotes the state-space, $\mathcal{A}$ means action-space, $\mathcal{P}:\mathcal{S}\times\mathcal{A}\times\mathcal{S}\rightarrow[0,1]$ represents state transition probability distribution, $\gamma\in(0,1)$ is discount factor and $\mathcal{R}:\mathcal{S}\times\mathcal{A}\rightarrow\mathbb{R}$ corresponds to reward function. Let τ and τ_E be a set of trajectories $(s_1,a_1,\cdots,s_T,a_T)$, generated by stochastic policy $\pi:\mathcal{S}\times\mathcal{A}\rightarrow[0,1]$ and expert policy π_E repectively, where T denotes the terminal time. Each such demonstration has a corresponding sequence of latent variables $c=\{c_1,\cdots,c_{T-1}\}$, meaning the sub-task at any given time step. Let $\Omega(a|s,s'):\mathcal{S}\times\mathcal{S}\times\mathcal{A}\rightarrow[0,1]$ be the inverse model. Finally, let $\Phi(s)$ be an empowerment-based potential function that quantifies a utility of a given state $s\in\mathcal{S}$ to regularize policy under MaxEnt-IRL framework. In our proposed work, we expand the above models into multi-task settings by adding a latent variable c.

B. Imitation Learning

IL [9] aims at directly learning policy π that can mimic expert behaviors π_E from expert trajectories τ_E. For mitigating compounding errors, GAIL [4] realizes the imitation learning problem into an adversarial learning framework. Agent's policy π serves as generator while the discriminator D represents a local reward function that differentiates between expert samples from π_E and generated samples from π. The objective is given as Eq.1.

$$\min_{\pi}\max_{D}\,[\log D(s,a)]+\mathbb{E}_{\pi_E}[1-\log D(s,a)]-\lambda H(\pi) \quad (1)$$

To distinguish different type of behaviors in expert demonstration τ_E, InfoGAIL [17] introduces a latent variable c into the existing policy π. High mutual information $I(c;\tau)$ incentives π to use c based on an information-theoretic regularization, a concept was first derived in InfoGAN [18]. They induce a variational lowerbound $L_q(\pi,Q)$ of the mutual information $I(c;\tau)$ and then add it to to the loss function in GAIL where $H(c)$ is entropy of the posterior Q. To reduce the dependency of trajectory from entire to current time, DIGAIL [5] modifies the $L_q(\pi,Q)$ by replacing $I(c;\tau)$ with the directed or causal information flow $I(\tau\rightarrow c)$. This gives the following lower bound like Eq. 2 where $\tau_{1:t}=(s_1,\cdots,a_{t-1},s_t)$.

$$\sum_t \mathbb{E}_{p(c_{1:t}),\pi(a_{t-1}|s_{t-1},c_{1:t-1})}[\log Q(c_t|c_{1:t-1},\tau_{1:t})]+H(c)$$
$$=L_q(\pi,Q)\;\le\;I(\tau\rightarrow c) \quad (2)$$

The prior distribution $p(c_{1:t})$ is pre-trained from Variational Auto-Encoder (VAE) [19] on expert trajectories and the $H(c)$ term does not considered in practice. The modified objective is denoted as Eq. 3.

$$\min_{\pi,Q}\max_{D}\mathbb{E}_{\pi}[\log D(s,a)]+\mathbb{E}_{\pi_E}[1-\log D(s,a)]$$
$$-\lambda_1 L_q(\pi,Q)-\lambda_2 H(\pi) \quad (3)$$

Unlike the previous work, we aim to recover a portable or transferable reward function that depends on the latent variable$c$. In section III, we will explain how to incorporate the above concept into multi-task IRL framework.

C. Variational Information Maximization

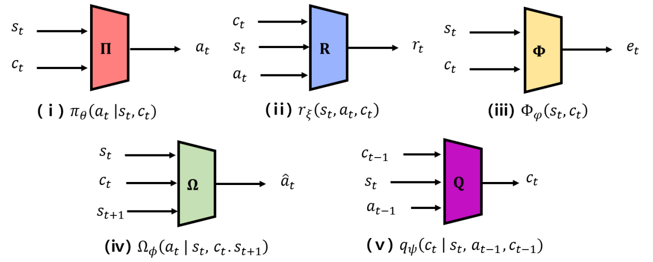
AIRL [15], a variant of GAIL, models the expert trajectory distribution $p(\tau_E)$ with Energy Based Model (EBM) $e^{\sum_{t=0}^{T}r(s_t,a_t)}$ where the energy function corresponds to reward function, connecting with sampling-based MaxEnt-IRL framework [14]. It could restore disentangled reward function from dynamics by adding shaping term like Eq. 4, representing a optimal discriminator with $\frac{e^{f(s,a)}}{e^{f(s,a)}+\pi(a|s)}$.

$$f(s,a)=r(s,a)+\gamma\Phi(s')-\Phi(s) \quad (4)$$

Nevertheless, AIRL could only recover the state dependent reward function because the irregular features of the action hinder the function learning. To solve the above limitation, EAIRL uses the empowered reward $\Phi(s)$ that maximize mutual information $\max I(a;s'|s)$ as the internal reward. This gives the following lower bound denoted as Eq. 5.

$$H(w(a|s))+\mathbb{E}_{\pi}[\log\Omega(a|s,s')]=L_I(w,\Omega)\le I(a;s'|s) \quad (5)$$

The lowerbound $L_I(w,\Omega)$ is optimized by Expectation Maximization (EM) algorithm over the distribution of the



(i) $\pi_\theta(a_t\,|s_t,c_t)$ (ii) $r_\xi(s_t,a_t,c_t)$ (iii) $\Phi_\varphi(s_t,c_t)$

(iv) $\Omega_\phi(a_t\,|\,s_t,\,c_t.\,s_{t+1})$ (v) $q_\psi(c_t\,|\,s_t,\,a_{t-1},c_{t-1})$

Fig. 2. The structures of five networks. We train five models denoted by a policy π, a reward r, an empowerment-based potential Φ, an inverse model Ω and a posterior q with each parameter denoted by $\theta, \xi, \varphi, \phi$ and ψ. The above figure shows the input and output information of each network.

action and the inverse model. For the inverse model Ω parameterized by ϕ, it minimizes the Mean Squared Error (MSE) between actions a conducted by the policy π and predicted action $\hat{a}$ from the inverse model Ω based on a supervised maximum log-likelihood problem, which is formulated as Eq. 6.

$$l_q(s,a,s') = |\Omega_\phi(\hat{a}|s,s') - a|^2 \tag{6}$$

$w(a)$ is the solution of Lagrange dual problem over $\frac{\partial \hat{I}^w}{\partial w} = 0 \;\; s.t \sum_a w(a|s) = 1$. The analytic solution is $w^*(a|s) = \frac{e^{u(s,a)}}{Z(s)}$, where $u(s,a) = \beta \mathbb{E}_P[\log \Omega(a|s,s')]$ and $Z(s) = \sum_a u(s,a)$, where the normalization term $\log Z(s)$ is equivalent to empowerment-based potential function $\Phi(s)$ parameterized by φ. They minimize the MSE between approximated $w_{\theta,\varphi}(a|s) \approx \log \pi_\theta(a|s) + \Phi_\varphi(s)$ and $\log \Omega(a|s,s')$ as the following Eq. 7.

$$l_I(s,a,s') = |\log \Omega(a|s,s') - \{\log \pi_\theta(a|s) + \Phi_\varphi(s)\}|^2 \tag{7}$$

To deal with multi-task problem that resilient to task transition, our proposed method reformulate the empowerment-based regularization technique by adding the latent variable c into the above equations in section III-A.

D. Hindsight Inference

To overcome limitation of MaxEnt-RL [12] or MaxEnt-IRL [13] approaches that focus on single primitive tasks, Hindsight Inference for Policy Improvement (HIPI) [16] extend the policy $\pi(a|s)$ and reward $r(s,a)$ into task-conditioned policy $\pi(a|s,c)$ and reward $r(s,a,c)$ introducing a latent variable c. They reveal that the process of relabeling sub-tasks with a relabeling distribution $Q(c|s,a)$ correspond to multi-task objective like Eq. 8 that both MaxEnt-RL and MaxEnt-IRL maximize at the same time where $q(\tau)$ is a distribution over previously-observed trajectories.

$$\begin{aligned} &- D_{KL}(q(\tau,c|s,a)\;||\;p(\tau,c,s,a)) \\ =&\mathbb{E}_{\substack{c\sim q(c|\tau)) \\ \tau\sim q(\tau)}}\Bigg[\sum_t r(s,a,c) - \log \pi(a|s,c) \\ &- D_{KL}(Q(c|s,a)\;||\;p(c)) - \log Z(c)\Bigg] \end{aligned} \tag{8}$$

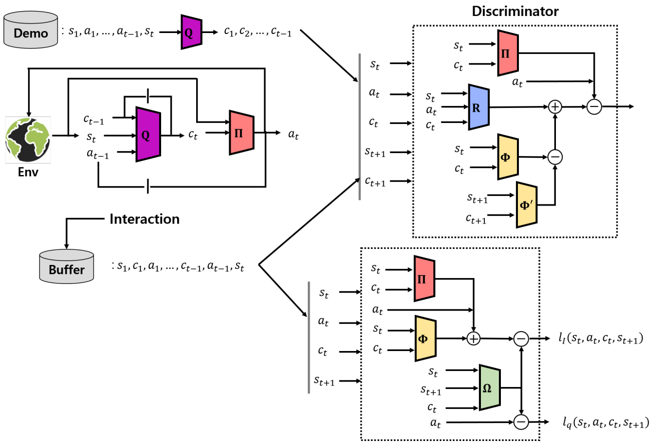

Demo
: $s_1, a_1, \ldots, a_{t-1}, s_t$
Q
$c_1, c_2, \ldots, c_{t-1}$
Discriminator
Env
c_{t-1}
s_t
a_{t-1}
Q
c_t
Π
a_t
Interaction
Buffer
: $s_1, c_1, a_1, \ldots, c_{t-1}, a_{t-1}, s_t$
s_t
a_t
c_t
s_{t+1}
c_{t+1}
Π
R
Φ
Φ′
$l_I(s_t, a_t, c_t, s_{t+1})$
$l_q(s_t, a_t, c_t, s_{t+1})$
Ω

Fig. 3. The overall adversarial architecture of our proposed method. The discriminator models reward function explicitly with the potential function which explains shaping-term. Fake or true sequence is generated as the policy utilizes the pseudo-labels provided by posterior. After that, we use the discriminant signal to train hierarchical policy.

Adding the first and second term means the expanded soft Q-function with latent variable c. In this work, they found the optimal relabeling distribution becomes the exponential family of combination of the expanded soft Q-function and the partition function that normalizes rewards of different scales. In section III-B, we discover our proposed method coincides in the above Eq. 8.

III. Proposed Method

We focus on solving multi-task learning problem without predefined task-specific knowledge and multi-task transfer learning problem where there are situations that have not been seen in training. We assume that expert trajectories $\{\tau^1, \ldots, \tau^N\} \in \tau_E$ are not labeled, sequence length for each episodes is variable and there exists sub-task c_t corresponding to each state-action pair (s_t, a_t). Our proposed method comprises five networks modeled as artificial neural networks with each parameter as shown in the Fig. 2. (i) a policy model $\pi_\theta(a|s,c)$ outputs a distribution over actions given the current state and the current sub-task. (ii) a reward $r_\xi(s,a,c)$ is a function of the state, action and sub-task. (iii) a potential function $\Phi_\varphi(\cdot)$ determines the reward-shaping function $F = \gamma\Phi_\varphi(s',c') - \Phi_\varphi(s,c)$ and also regularize the policy update. (iv) an inverse model $\Omega_\phi(s,c,s')$ outputs a distribution over actions that brings about state transition. (v) a posterior $q_\psi(c_t|s_t,a_{t-1},c_{t-1})$ outputs a distribution over a sub-task given the sub-tasks discovered up to current and the trajectory followed up to current. (See Appendix B for the details of each model structure and training details.) All these models except for the posterior are trained simultaneously based on the objective functions described in Algo. 1. The following sections explains how to recover optimal hierarchical policies and generalizable multi-task reward functions concurrently. Moreover, we reveal that our approach is identical to the objective of the Hindsight Inference approach [16], which solves the MaxEnt-RL and MaxEnt-IRL problems simultaneously in a multi-task setting.

Algorithm 1 Situational Empowerment-based Adversarial Inverse Reinforcement Learning

```
Obtain expert demonstrations $\tau_E$ by running expert policy $\pi_E$ and posterior $Q_\psi$
Pretrain posterior $Q_\psi$ using VAE with the gradient: $-\mathbb{E}_{\tau_E}[\nabla_\theta \log \pi_\theta(a|s,c)] + \nabla_\psi D_{KL}[Q_\psi(c'|c,s',a)\,||\,\mathcal{N}(0,I)]$
Initialize parameters of policy $\pi_\theta$, inverse model $\Omega_\phi$ empowerment $\Phi_\varphi$, reward $r_\xi$ functions
Synchronize the parameters of target empowerment $\Phi_{\varphi'}$ with $\varphi$
for $i \leftarrow 0$ to $N$ do
    Collect trajectories $\tau$ by executing $\pi_\theta$ and $Q_\psi$
    Update $\phi_i$ to $\phi_{i+1}$ with gradient $\mathbb{E}_\tau[\nabla_{\phi_i} l_q(s,a,c,s')]$
    Update $\varphi_i$ to $\varphi_{i+1}$ with gradient $\mathbb{E}_\tau[\nabla_{\varphi_i} l_I(s,a,c,s')]$
    Update $\xi_i$ to $\xi_{i+1}$ with the gradient : $\mathbb{E}_\tau[\nabla_{\xi_i} \log D_{\xi_i,\varphi_{i+1}}(s,a,c,s')] + \mathbb{E}_{\tau_E}[\nabla_{\xi_i}(1 - \log D_{\xi_i,\varphi_{i+1}}(s,a,c,s'))]$
    Update $\theta_i$ to $\theta_{i+1}$ using PPO update rule with the gradient:
        $$\mathbb{E}_\tau[\nabla_{\theta_i} \log \pi_{\theta_i}(a|s,c)\hat{r}_{\xi_{i+1}}(s,a,c,s',c')] - \lambda_I \mathbb{E}_\tau[\nabla_{\theta_i} l_I(s,a,c,s')]$$
    After every n epoch synchronize $\varphi'$ with $\varphi$
end for
```

A. Empowerment-based Regularization for Multi-Tasking

The empowerment-based regularization technique [6] is newly expanded by introducing the intrinsic task variable used in [5] to solve the complex multitask problem with unknown dynamics. To give robustness to environmental changes, we propose a method that recover a portable or transferable multi-task reward function included in discriminator through the empowerment-based regularization technique shown as Fig. 3. Considering the relationship of information diagram in Fig. 1, we define situational empowerment-based potential function $\Phi(s,c) = \underset{w,\Omega}{max}\ I^{w,\Omega}(a;s'|s,c)$ as a new internal reward. The variational lowerbound derived in Eq. 5 is expanded into the following inequality Eq. 9 and the detailed derivation is given in Appendix A.

$$\begin{aligned} I(a;s'|s,c) &= H(a|s,c) - H(a|s,c,s') \\ &\geq H[w(a|s,c)] - H[\Omega(a|s,c,s')] \\ &= \mathbb{E}_\pi[-\log w(a|s,c)] + \mathbb{E}_{Q,w,P}[\log \Omega(a|s,c,s')] \\ &= L_I(w,\Omega) \end{aligned} \tag{9}$$

We optimize the above lowerbound $\hat{I}_{w,\Omega}$ over the distribution w and hierarchical inverse model Ω through the Expectation-maximization (EM) algorithm where w is the action distribution conditioned on state and sub-task pair. For the hierarchical inverse model, we calculate the following objective Eg. 10 based on maximum log-likelihood problem.

$$l_q(s,a,c,s') = |\Omega_\phi(\hat{a}|s,c,s') - a|^2 \tag{10}$$

For $w(a|s,c)$ that means the normalized hierarchical policy, we find optimal solution $w^*(a|s,c) = e^{(\lambda-1)+\beta\mathbb{E}_{Q,P}[\log\Omega(a|s,c,s')]}$ by replacing constrained form $\frac{\partial \hat{I}^{w,\Omega}}{\partial w} = 0\ s.t \sum_a w(a|s,c) = 1$ with unconstrained Lagrange dual problem. We set the potential function $\Phi(s,c)$ as $\log Z(s,c)$ because $e^{(1-\lambda)}$ means partition function $Z(s,c) = \sum_a e^{\beta\mathbb{E}_{Q,P}[\log\Omega(a|s,c,s')]}$. We minimize MSE between approximated $w_{\theta,\varphi}(a|s,c) \approx \log\pi_\theta(a|s,c) + \Phi_\varphi(s,c)$ and $\log\Omega(a|s,c,s')$ as the follwing Eq. 11.

$$\begin{aligned} l_I(s,a,c,s') =&| \log\Omega(a|s,c,s') \\ &- \{\log\pi_\theta(a|s,c) + \Phi_\varphi(s,c)\}|^2 \end{aligned} \tag{11}$$

Finally, the derived objective Eq. 6 and Eq. 7 are extended into Eq. 10 and Eq. 11 by introducing latent code c to deal with the multi-task setting.

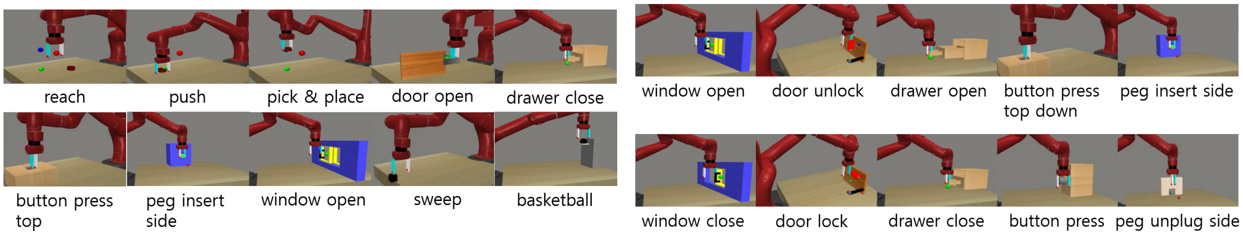
B. Connection with Hindsight Relabeling Framework

We design a reward function represents intention contained in expert distribution with energy function based on sampling-based MaxEnt-IRL [7]. To extent the previous work [15], [6] into the multi-task setting, we present a method to restore the disentangled multi-task reward $f_{\xi,\varphi}(s,a,c,s',c') = r_\xi(s,a,c) + \gamma\Phi_{\varphi'}(s',c') - \Phi_\varphi(s,c)$ with the shaping term added by introducing a latent variable c to generalize the reward function for changes of task transition dynamics. For the stability, we fix the parameter of target situational potential function $\Phi_{\varphi'}$ and update it at several interval. When the discriminator presented in Eq. 3 is substituted for the situational discriminator $D_{\xi,\varphi}(s,a,c,s',c') = \frac{e^{f_{\xi,\phi}(s,a,c,s',c')}}{e^{f_{\xi,\phi}(s,a,c,s',c')}+\pi(a|s,c)}$ and add empowerment-based regularization term Eq. 11, we could get the following objective Eq. 12 for the reward and Eq. 13 for the policy.

$$\begin{aligned} \max_\xi\ & \mathbb{E}_\tau[\log D_{\xi,\phi}(s,a,c,s',c')] \\ & + \mathbb{E}_{\tau_E}[\log(1 - D_{\xi,\phi}(s,a,c,s',c'))] \end{aligned} \tag{12}$$

$$\begin{aligned} \min_\theta\ & \mathbb{E}_{Q,\pi}[-f_{\xi,\varphi}(s,a,c,s',c') + \log\pi_\theta(a|s,c)] \\ & + \lambda_1 L_q(\pi,Q) + \lambda_2 l_I(\pi,\Phi) \end{aligned} \tag{13}$$

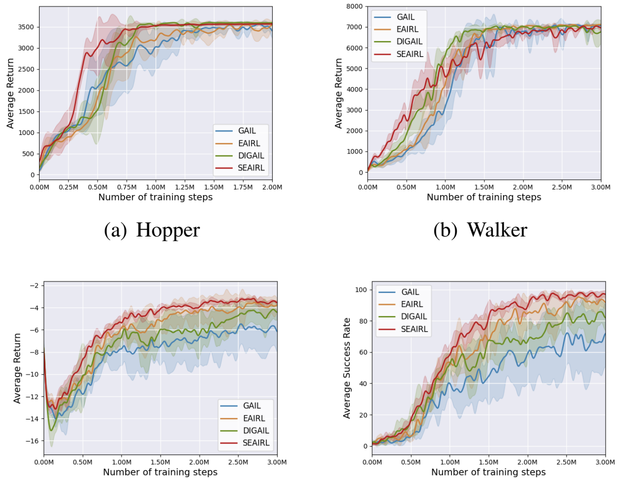
To apply policy gradient, we rewrite the above objective Eq. 13 using the alternative reward $\hat{r}_\xi(s_t,a_t,c_t,s_{t+1},c_{t+1}) = f_{\xi,\varphi}(s_t,a_t,c_t,s_{t+1},c_{t+1}) + \lambda_q \log Q_\psi(c_t|s_t,a_{t-1},c_{t-1})$, then the alternative objective is shown as Eq. 14,



(a) MT10

(b) ML5

Fig. 4. Two environment settings for multi-task (a) and multi-task transfer (b) imitation learning. **Left**: MT10 contains 10 random distinct scenarios where each scenario has 50 random initial object and goal positions. **Right** : The customized environment called ML5 (b) also has equivalent setting with (a) for each scenario. When collecting expert demonstration and pre-training phase, it randomly selects one of the above five scenarios. When transfer learning phase, it chooses one of the below five similar but different scenarios.



(a) Hopper

(b) Walker

(c) Fetch

(d) Fetch

Fig. 5. Learning curves for Hopper-v3 (a), Walker-v2 (b) and FetchPickAndPlace-v1 (c), (d) environments. The darker-colored lines and shaded areas denote the average return or success rate and standard deviations, respectively, computed over 5 random seed.

$$\max_{\theta} \mathbb{E}_{\tau}[\log \pi_{\theta}(a_t|s_t, c_t)\hat{r}_{\xi}(s_t, a_t, c_t, s_{t+1}, c_{t+1})] + \lambda_h H(\pi_{\theta}) - \lambda_I l_I(\pi, \Phi) \quad (14)$$

Therefore, our proposed method is a special case of previous work Eq. 8 because maximizing $\log Q_{\psi}(c_t|s_t, a_{t-1}, c_{t-1})$ is equivalent to minimize $D_{KL}(Q(c, |s, a)||p(c))$.

TABLE I

COMPARISON OF AVERAGE RETURNS WITH STANDARD DEVIATION

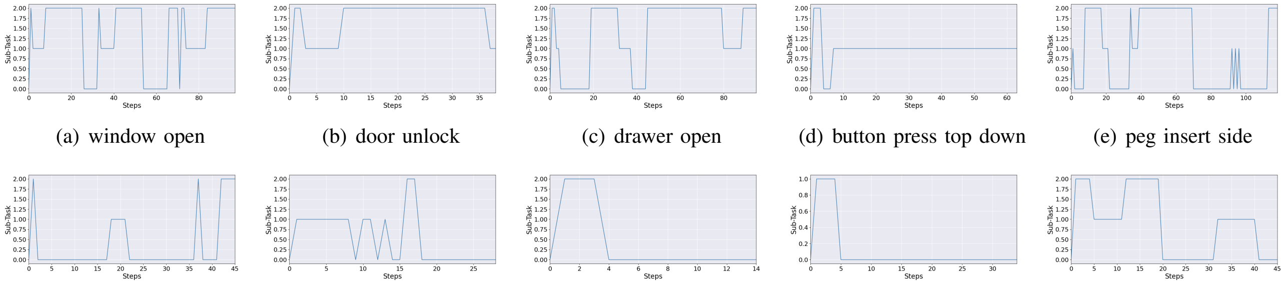
METHOD	HOPPER	WALKER	FETCH
	AVG. RETURN ± STD		
GAIL	3604.94 ± 18.85	7128.18 ± 710.64	-7.65 ± 5.15
DIGAIL	**3632.48** ± 9.37	7262.67 ± 138.09	-6.13 ± 4.99
EAIRL	3615.72 ± 7.54	**7339.17** ± **41.24**	-3.67 ± 2.07
SEAIRL	3630.86 ± **5.69**	7212.81 ± 49.92	**-3.31** ± **1.78**

IV. EXPERIMENTS

We present results on both multi-task learning and multi-task transfer learning environment by utilizing two physics engine (i.e. MuJoCo / Robotics / Meta-World) interfaced within OpenAI Gym [20]. We validate outstanding performance of our proposed method (Situated-EAIRL) comparing with existing baselines (i.e. GAIL / DIGAIL / EAIRL). In the multi-task settings, we show that the more complex the relation of the task transition becomes, the better our proposed model performs. Furthermore, We demonstrate robustness through successful transfer learning in environments with work shifts not seen when acquiring expert trajectories. (The detailed environmental description is given in Appendix B.4)

A. Multi-Task Learning Performance

To validate our proposed approach could acquire the various distinct skills in multi-task settings, we experiment with two types of continuous state-action control tasks. The first type is locomotive tasks moving forward in the same direction, which involves Hopper and Walker environments. The second type is manipulation tasks which is comprised of FetchPickandPlace and MT10 as shown in Fig. 4(a). Since the problem becomes more challenging as the number of scenarios and state-action space increases, we construct the above four environments. Both Hopper, Walker, FetchPickandPlace environments are a single scenario composed of a few sub-tasks. Specifically, for Hopper, the macro action such as jumping, mid air and landing phases are switched periodically. In case of Walker, the agent lift and set down each foot in turn at a regular pace. On the FetchPickandPlace environment, the agent controls fetch's end effector to grasp and lift a block up and reach a goal point where the initial position of the block and goal point are randomly selected. Fig. 5 shows the learning curve of each single scenario environment and Table. I represents the final average return with standard deviation over 100 episodes for each method. The curves describes our method has better overall data-effiency. When comparing between environment without randomness such as Walker and hopper and Fetch environment with randomness, the proposed model called Situational-EAIRL (SEAIRL) shows better performance in an environment with randomness. Meanwhile, for the case that the dimension



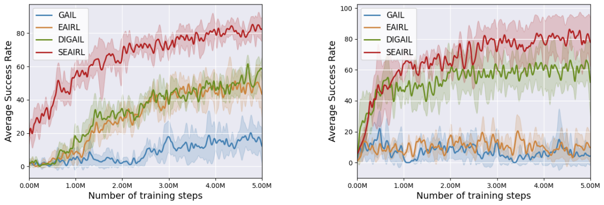
(a) window open (b) door unlock (c) drawer open (d) button press top down (e) peg insert side

(f) window close (g) door lock (h) drawer close (i) button press (j) peg unplug side

Fig. 6. The visualization of sub-task transition to adapt new environment. The sub-task is a three dimensional one-hot vector represented by categorical variable by applying the Gumbel-softmax trick [22] to the output of the posterior.

increases without randomness, the proposed model showed similar final performance with the baselines.

For comparison in environments with more complex task transitions, we setup environments that composed of randomly selected from ten scenarios by employing Meta-World simulator where each scenario has 50 randomized initial object and final goal positions. we expect that an reward function trained with our method provides task specific learning signals to learn composable sub-task policy. After that, the agent could achieve successfully given tasks in various scenarios by combining basic action primitives such as grasping, lifting, reaching, pushing and placing. Compared with baselines in the MT10 environment with various scenarios, Fig. 7(a) describes a significant improvement in performance and convergence speed. The final average success rate over 100 episodes of each method is shown in Table. II. Our proposed method shows 1.18 and 1.78 times higher performance than the existing methods in Fetch and MT10 environment respectively, which means that our proposed approach is better as the randomness and task complexity increase. Fig. 8 shows the average success rate for each scenario. GAIL tends to over-fit only to one scenario. EAIRL has a limitation in targeting multiple tasks although reward signals that quickly adapt to changing circumstances help policy learning. In the case of DIGAIL, as mentioned in [16], it could be difficult for the learned policy to understand the hierarchical structure unless the reward signal is explicitly expressed for each task. Our method compensates for the shortcomings of the two methods by normalizing the reward size for each task and providing a separate reward signal to the hierarchical policy for task change. By doing so, we successfully solve many scenarios where the success rate is zero in the previous methods. As the above experimental results, we validate a strong point of multi-task learning.



(a) MT10 (b) ML5

Fig. 7. Learning curves for MT10 (a) and ML5 (b) environments. The darker-colored lines and shaded areas denote the average success rate and standard deviations, respectively, computed over 5 random seed.

TABLE II
Comparison of Average Success Rate

Method	Fetch	MT10	ML5
	Avg. Success Rate		
GAIL	0.46	0.14	0.20
DIGAIL	0.73	0.46	0.72
EAIRL	0.82	0.43	0.30
SEAIRL	**0.97**	**0.82**	**0.98**

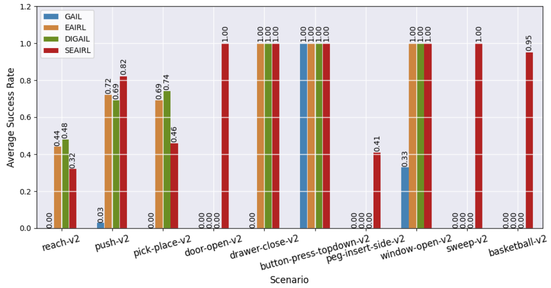


Fig. 8. Average success rate for MT10 environment. The average rate of each scenario is computed over 100 episodes.

B. Multi-Task Transfer Learning Performance

With the objective of confirming the ability to quickly adapt to new tasks in multi-task settings, we build ML5 environment where training scenarios are similar but different from the environment in which expert demonstrations have been obtained as shown in Fig. 4(b). we expect that learned reward disentangled from sub-task could be composed to give learning signals for guiding policy into a new types of behavior. In turn, the policy chains learned macro actions together to make a different desired policy as visualized in Fig. 6. The swapping pattern of the latent variable used in the environment in the first row is newly recombined according to the changed environment in the second row. We plot the

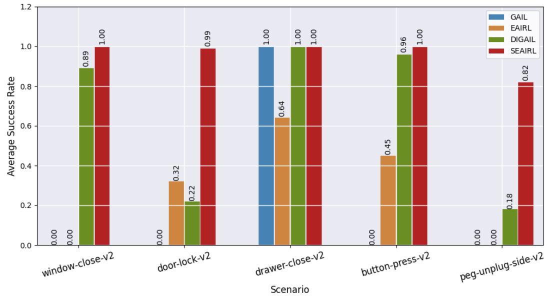


Fig. 9. Average success rate for ML5 environment. The average rate of each scenario is computed over 100 episodes.

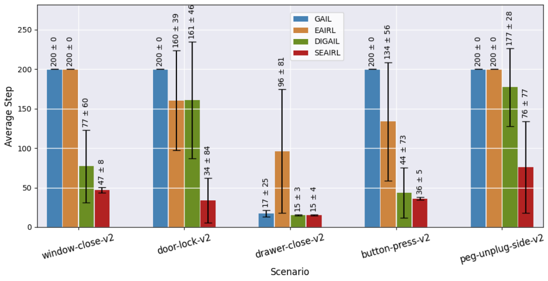


Fig. 10. Average s episode steps for ML5 environment. The average steps of each scenario is computed over 100 episodes. The maximum episode step is set by 200 steps. Black solid line for each bar represents the standard deviations.

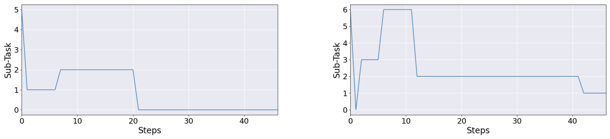
change in values for the higher-dimensional latent variables in Fig. 11 to see if even larger latent sizes identify basic action primitives similar to those performed at latent size of 3. The plot means that in spite of larger dimensionality our approach is able to reuse appropriate context inferred previously.

Our approach outperforms both GAIL and EAIRL which does not focus on multi-tasking as well as DIGIAL which does not model the multitask reward function explicitly in terms of data efficiency and performance, as described in Fig. 7(b). Fig. 9 represents the success rate through transfer learning for a new environment for each scenario. GAIL shows a tendency to over-fit only to one scenario, similar to the MT10 environment. For the case of eairl, it has increased generality to several scenarios, but overall performance is low. It is difficult for DIGAIL to adapt to a new task because of frequent task changeovers unnecessarily like Fig. 12, so that the goal could not be achieved or the episode elapse step could be long rather than proposed method as shown in Fig. 10. According to Table. II, The proposed method results in the state-of-the-art by increasing the success rate for episode 100 by 1.36 times compared to the existing methods in the ML5 environment. Finally, we verify robustness over changing task dynamics in the multiple scenarios.

V. CONCLUSIONS

Our work propose an approach to adversarial transferable reward and adaptable policy learning from unstructured expert demonstrations. The key idea is to learn the situational empowerment through variational information maximization in parallel to learning the reward and policy. We also shows the theoretical connections with the hindsight inference literature as used in multi-task learning. The proposed regularization makes the process of restoring the normalized reward function for each task, which in turn prevents the policy from over-fitting into the local behavior for each task. Outperforming previous imitation learning methods, we show that our policy could handle highly diverse task sets as well as adapt quickly to dynamically or structurally different environments by using the learned portable reward. Our experimental results demonstrate that our learned reward and policy understand a hierarchical task structure in an unsupervised manner without predefined knowledge and lead to meaningful generalization across many tasks and unknown dynamics.

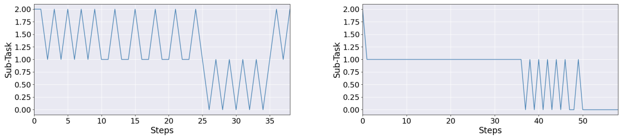
In our future work, we will extend to our work to learn posterior model concurrently based on graphical embedding to naturally deal with long-term or image-based multi-tasks. Another exciting direction would be to investigate how to build an algorithm that learns from sub-optimal demonstrations that contains both optimal and non-optimal behaviors.



(a) Latent Size 6

(b) Latent Size 9

Fig. 11. Visualization of high-dimensional latent variables in a window close environment.



(a) button press

(b) window close

Fig. 12. The visualization of sub-task transition of DIGAIL algorithm for button press (a) and window close (b) environments. sub-task variable is 3 dimensional one-hot vector which is same as the proposed method.

APPENDIX

A. Variational Information Lowerbound

By defining MI as a difference in conditional entropies $I^{w,\Omega}(a;s';|s,c) = H(a|s,c) - H(a|s,c,s')$ as mentioned in section III-A, the variational lower bound representation of MI is derived as follow.

$$
\begin{aligned}
&-H(a_t|s_t,c_t,s_{t+1})\\
&=\mathbb{E}_{\substack{Q(c_t|s_t,a_{t-1},c_{t-1})\\ w(a_t|s_t,c_t)\\ P(s_{t+1}|a_t,s_t)}}[\log p(a_t|s_t,c_t,s_{t+1})]\\
&=\mathbb{E}_{Q,w,P}[\log\frac{p(a_t|s_t,c_t,s_{t+1})\Omega(a_t|s_t,c_t,s_{t+1})}{\Omega(a_t|s_t,c_t,s_{t+1})}]\\
&=\mathbb{E}_{Q,w,P}[\Omega(a_t|s_t,c_t,s_{t+1})+\log\frac{p(a_t|s_t,c_t,s_{t+1})}{\Omega(a_t|s_t,c_t,s_{t+1})}]\\
&=\mathbb{E}_{Q,w,P}[\Omega(a_t|s_t,c_t,s_{t+1})]+D_{KL}(p\,||\,\Omega)\\
&\geq\mathbb{E}_{Q,w,P}[\Omega(a_t|s_t,c_t,s_{t+1})]
\end{aligned}
$$

B. Implementation Details

1) Network Architectures: We briefly describe the network structure of the models included in our method. The reward has two layers of 128 units with Leaky-ReLU activations. The policy and inverse model has two layers of 256 units with ReLU activations. For the continuous control, the policy and inverse model estimate the Gaussian parameters. The posterior has one GRU layers with 256 hidden size to capture the long-term dependencies between sub-tasks. We use Gumbel-softmax trick [22] for posterior to represent the latent code with categorical variable. Finally, the empowerment has two layers of 64 units with ReLU activations.

2) Hyperparameters: Table. III describes the hyperparameters used for our experiments.

TABLE III
HYPERPARAMETER

HYPERPARAMETERS	VALUE
PPO Clipping Factor	0.2
Discount Factor (γ)	0.99
GAE Parameter	0.95
Value Coefficient	0.5
Entropy Coefficient (λ_h)	0.01
Posterior Coefficient (λ_q)	0.01
Empowerment Coefficient (λ_I)	0.001
Target Update Epoch	5
Mini-Batch Size	512
Rollout Horizon	2048
Learning Rate (Policy, Posterior)	0.0003
Learning Rate (others)	0.0001

3) Training Detail: All of the experiments were performed on a PC with a 3.20 GHz Intel Core i7-8700K Processor, and one GTX 1080-Ti GPU. Our method consists of five models in Fig. 2. The overall training procedure of our method is described in Algo. 1 The posterior and policy pretrained with VAE [19] and then we only use the fixed parameter of the posterior. The policy and discriminator are trained with adversarial learning [15]. The policy is optimized with PPO [23] and all networks except for the policy are update with Adam [24] optimizer.

In order to gather demonstrations for both Hopper and Walker and FetchPickAndPlace, we trained expert agents with RL algorithms such as ACKTR [25] and HER [26] in dense reward settings. In MT10 and ML5, a policy designed for each scenario provided by the Meta-World simulator was utilized. we use the number of expert trajectory over {8, 8, 512, 1024, 1024} for each environment with sub-sampling frequency 4.

4) Environment Description: Table. IV provides further details about the environments we used for our experiments.

TABLE IV
BENCHMARK DESCRIPTION

ENVIRONMENT	STATE	ACTION	MAXIMUM STEP
HOPPER	11	3	1000
WALKER2D	17	6	1000
FETCHPICKANDPLACE	28	4	50
MT10	39	4	200
ML5	39	4	200

ACKNOWLEDGMENT

We sincerely appreciate Sang-Hyun Lee, jin-Woo Choi, Chan Kim for their helpful discussion. This work was supported by Seoul Nation University, College of Engineering, Seoul, Republic of Korea.

REFERENCES

[1] Richard S Sutton, Andrew G Barto, et al. Introduction to reinforcement learning, volume 135. MIT press Cambridge, 1998.

[2] Mnih, Volodymyr, et al. ”Playing atari with deep reinforcement learning.” arXiv preprint arXiv:1312.5602 (2013).

[3] Argall, Brenna D., et al. ”A survey of robot learning from demonstration.” Robotics and autonomous systems 57.5 (2009): 469-483.

[4] Ho, Jonathan, and Stefano Ermon. ”Generative adversarial imitation learning.” Advances in neural information processing systems. 2016.

[5] Sharma, Arjun, et al. ”Directed-info GAIL: Learning hierarchical policies from unsegmented demonstrations using directed information.” arXiv preprint arXiv:1810.01266 (2018).

[6] Qureshi, Ahmed H., Byron Boots, and Michael C. Yip. ”Adversarial imitation via variational inverse reinforcement learning.” arXiv preprint arXiv:1809.06404 (2018).

[7] Finn, Chelsea, Sergey Levine, and Pieter Abbeel. ”Guided cost learning: Deep inverse optimal control via policy optimization.” International conference on machine learning. PMLR, 2016.

[8] Sutton, Richard S., Doina Precup, and Satinder Singh. ”Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning.” Artificial intelligence 112.1-2 (1999): 181-211.

[9] Dean A Pomerleau. Alvinn: An autonomous land vehicle in a neural network. In Advances in neural information processing systems, pp. 305–313, 1989.

[10] Abbeel, Pieter, and Andrew Y. Ng. ”Apprenticeship learning via inverse reinforcement learning.” Proceedings of the twenty-first international conference on Machine learning. 2004.

[11] Dean A Pomerleau. Efficient training of artificial neural networks for autonomous navigation. Neural Computation, 3(1):88–97, 1991.

[12] Haarnoja, Tuomas, et al. ”Reinforcement learning with deep energy-based policies.” International Conference on Machine Learning. PMLR, 2017.

[13] Ziebart, Brian D., et al. ”Maximum entropy inverse reinforcement learning.” Aaai. Vol. 8. 2008.

[14] Finn, Chelsea, et al. ”A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models.” arXiv preprint arXiv:1611.03852 (2016).

[15] Fu, Justin, Katie Luo, and Sergey Levine. ”Learning robust rewards with adversarial inverse reinforcement learning.” arXiv preprint arXiv:1710.11248 (2017).

[16] Eysenbach, Benjamin, et al. ”Rewriting History with Inverse RL: Hindsight Inference for Policy Improvement.” arXiv preprint arXiv:2002.11089 (2020).

[17] Li, Yunzhu, Jiaming Song, and Stefano Ermon. ”Infogail: Interpretable imitation learning from visual demonstrations.” Advances in Neural Information Processing Systems. 2017.

[18] Chen, Xi, et al. ”Infogan: Interpretable representation learning by information maximizing generative adversarial nets.” Proceedings of the 30th International Conference on Neural Information Processing Systems. 2016.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114, 2013.

[20] Brockman, Greg, et al. ”Openai gym.” arXiv preprint arXiv:1606.01540 (2016).

[21] Yu, Tianhe, et al. ”Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning.” Conference on Robot Learning. PMLR, 2020.

[22] Jang, Eric, Shixiang Gu, and Ben Poole. ”Categorical reparameterization with gumbel-softmax.” arXiv preprint arXiv:1611.01144 (2016).

[23] Schulman, John, et al. ”Proximal policy optimization algorithms.” arXiv preprint arXiv:1707.06347 (2017).

[24] Kingma, Diederik P., and Jimmy Ba. ”Adam: A method for stochastic optimization.” arXiv preprint arXiv:1412.6980 (2014).

[25] Wu, Yuhuai, et al. ”Scalable trust-region method for deep reinforcement learning using kronecker-factored approximation.” Advances in neural information processing systems 30 (2017): 5279-5288.

[26] Andrychowicz, Marcin, et al. ”Hindsight experience replay.” arXiv preprint arXiv:1707.01495 (2017).