Robust Visual Imitation Learning with Inverse Dynamics Representations

Siyuan Li[1*], Xun Wang[2*], Rongchang Zuo[1], Kewu Sun[2], Lingfei Cui[3], Jishiyu Ding[2], Peng Liu[1], Zhe Ma[2†]

[1]Harbin Institute of Technology
[2]Intelligent Science & Technology Academy Limited of CASIC
[3]Institute of Computer Application Technology, Norinco Group
siyuanli@hit.edu.cn, wxhelloworld@outlook.com

Abstract

Imitation learning (IL) has achieved considerable success in solving complex sequential decision-making problems. However, current IL methods mainly assume that the environment for learning policies is the same as the environment for collecting expert datasets. Therefore, these methods may fail to work when there are slight differences between the learning and expert environments, especially for challenging problems with high-dimensional image observations. However, in real-world scenarios, it is rare to have the chance to collect expert trajectories precisely in the target learning environment. To address this challenge, we propose a novel robust imitation learning approach, where we develop an inverse dynamics state representation learning objective to align the expert environment and the learning environment. With the abstract state representation, we design an effective reward function, which thoroughly measures the similarity between behavior data and expert data not only element-wise, but also from the trajectory level. We conduct extensive experiments to evaluate the proposed approach under various visual perturbations and in diverse visual control tasks. Our approach can achieve a near-expert performance in most environments, and significantly outperforms the state-of-the-art visual IL methods and robust IL methods.

Introduction

Imitation learning (IL) has gained encouraging success in various domains, e.g., games (Scheller, Schraner, and Vogel 2020; Baker et al. 2022), robotics (Hua et al. 2021; Wang and Chang 2021), and autonomous driving (Hawke et al. 2020; Hu et al. 2022). By learning behaviors directly from expert demonstrations, IL provides a way of sparing the burden of designing delicate reward functions (Hussein et al. 2017; Arora and Doshi 2021). However, every coin has two sides. Despite no dependence on the rewards, collecting expert datasets requires much effort. Furthermore, in some scenarios, it is quite difficult to collect expert trajectories exactly in the target learning environment. For example, although the aim is to learn a manipulation policy for a real robot arm, since the hardware is sophisticated and expensive, the expert demonstrations could only be collected in a simulator, which is similar to the target real-world environment, but has differences. The difference between the expert environment and the target learning environment induces an important challenge for current IL approaches, especially for tasks with high-dimensional visual observations.

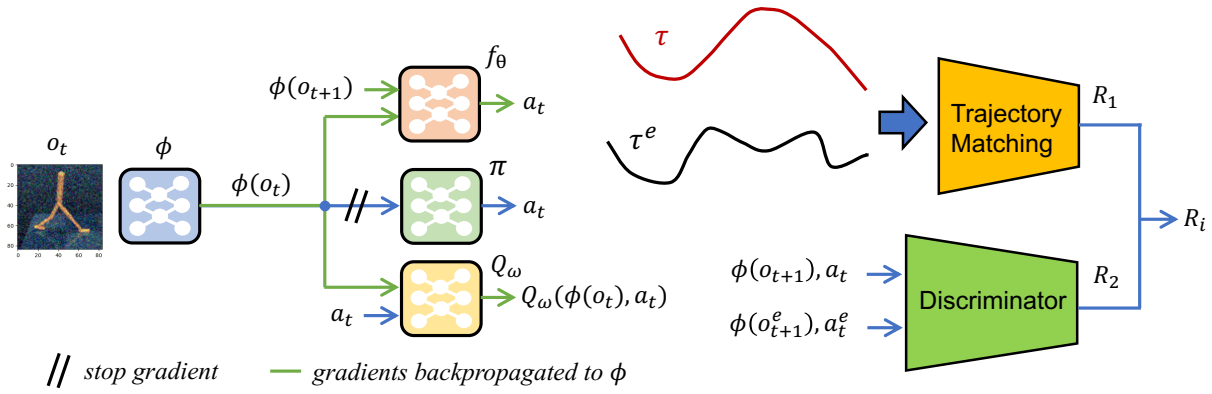
(a) Expert environment

(b) Learning environment

Figure 1: We refer the environment of collecting the expert dataset as *expert environment*, and the environment of learning the target policy as *learning environment*.

Behavior cloning (BC) (Pomerleau 1991; Torabi, Warnell, and Stone 2018a; Shafiullah et al. 2022) is a conventional IL method, which maximizes the likelihood of taking the demonstrated action in a supervised learning manner. Although simple, BC requires a large amount of expert data and suffers from compounding error (Ross and Bagnell 2010; Ross, Gordon, and Bagnell 2011). As there is no online interaction, BC cannot handle the difference between the expert environment and the learning environment. Inverse reinforcement learning (IRL) (Ng and Russell 2000; Ziebart et al. 2008; Han et al. 2022) is a popular learning paradigm for IL problems, which generates rewards by measuring the difference between expert data and behavior data. With the generated reward function, IRL employs an off-the-shelf RL algorithm to learn policies. Note that in challenging tasks with high-dimensional observation space, estimating differences between observations is difficult. Furthermore, when the learning environment and the expert environment are not exactly the same, IRL approaches may misuse the difference between environments to generate uninformative rewards. For example, even if the robot positions in Figure 1(a) and Figure 1(b) are similar (both standing), the background in Figure 1(b) may disturb the reward generation, which leads to a small reward to punish the observation in Figure 1(b). In fact, as the underlying state of this observa-

[*]These authors contributed equally.
[†]Corresponding author.



(a) state representation learning (b) imitative reward generation

Figure 2: Robust visual imitation learning with inverse dynamics representations.

tion is similar to that of the expert observation, this observation should be encouraged. In addition, current IRL approaches measure the difference between expert data and behavior data either from the element-wise view (Ho and Ermon 2016; Kostrikov et al. 2019; Liu et al. 2023) or from the trajectory view (Haldar et al. 2023; Papagiannis and Li 2023; Dadashi et al. 2021), which does not fully utilize the expert dataset to generate effective rewards.

To align the learning environment and the expert environment, we propose a novel imitation learning approach based on inverse dynamics representation learning. The inverse dynamics objective serves to extract the action-related features from the high-dimensional observations, so that we could obtain a common state representation space between the learning and expert environments. Thanks to the inverse dynamics state representation, the proposed approach could deal with challenging IL tasks with visual observations, and is robust to the difference between the learning environment and the expert environment. By measuring the similarity between expert state embeddings and behavior state embeddings, we develop an imitative reward function, which not only considers the element-wise similarity of observation-action pairs, but also takes the trajectory-level similarity into consideration. This thorough reward function improves the previous IRL methods which only generate rewards from a single perspective.

We conduct extensive experiments on a set of visual control tasks in Meta-World domain (Yu et al. 2020) and DeepMind Control Suite (DMC) (Tassa et al. 2018). The experiment results demonstrate that the proposed approach significantly outperforms the state-of-the-art visual IL methods and robust IL methods in terms of learning efficiency and convergent performance. To probe into the reason for the great performance, we further analyze the learned state representation in detail. Moreover, we conduct several ablation studies to validate the effectiveness of the various components in the proposed approach. It is hoped that these results could provide some insights for representation learning and reward design in robust visual imitation learning.

Preliminaries

We formulate the learning problem with a discounted finite-horizon Markov Decision Process (MDP). The MDP is of the form (O, A, P, R, γ), where O is the observation space, A is the action space, $P(o_{t+1}|o_t, a_t)$ is the transition function specifying the probability distribution over the next observation given the current observation and action, $R : O \times A \to \mathbb{R}$ is the reward function, and $\gamma \in [0, 1)$ is the discount factor. In this paper, we focus on the challenging problems where the observations are high-dimensional images. The goal is to learn a policy $\pi : O \to A$ that maximizes the expected cumulative discounted reward: $\max_\pi \mathbb{E}_{P,\pi}[\sum_{t=1}^{T} \gamma^t R(o_t, a_t)]$, where T denotes the horizon length.

In the IL setting, there is no available reward function for an agent to infer. Instead, the agent is provided with a demonstration dataset $\mathcal{T}^e = \{\tau^e_n|^N_{n=1}\} = \{(o^e_t, a^e_t)^T_{t=1}|^N_{n=1}\}$, which includes n trajectories collected by experts. IRL approaches (Ng and Russell 2000; Abbeel and Ng 2004) try to solve the IL problem by generating rewards based on expert trajectories $\mathcal{T}^e$, and then optimize policy π to maximize the cumulative rewards. Our work falls in the IRL paradigm, and uses the actor-critic learning framework to conduct policy optimization. Specifically, we employ the Twin Delayed Deep Deterministic policy gradient algorithm (Fujimoto, Hoof, and Meger 2018) as the base RL optimizer, which alleviates the Q overestimation issue with the twin delayed critic networks.

Approach

This paper considers an IL setting where the learning environment is different from the environment of collecting the expert dataset (expert environment), e.g., the learning environment is perturbed by visual distractors (e.g., the white noise in o_t in Figure 2), but the expert environment is free from noise. The similarity between the learning and expert environments is that they share the same task. Previous IRL methods (Ho and Ermon 2016; Torabi, Warnell, and Stone 2018b; Kostrikov et al. 2019) generate rewards based on

the similarity between expert data and behavior data. However, under the visual perturbations, it is challenging to accurately measure the similarity of the underlying states, which severely hurts imitation policy learning.

To fully exploit the expert trajectories even not in the expert environment, we propose a Robust visual Imitation Learning approach based on Inverse dynamics Representations (RILIR), as depicted in Figure 2. To avoid the negative influence of visual perturbations, we design a state representation module to extract prominent features from image observations. Based on the abstract state representation, we design a reward function that measures the similarity between the learned policy and the expert policy from not only the single-step perspective but also the trajectory perspective. The following of this section first elaborates on the state representation learning module, and then describes the way of using the state embeddings to generate rewards.

State Representation Learning

To facilitate reward generation and imitation policy learning, we aim to learn a representation function ϕ, which extracts state embedding $z_t = \phi(o_t)$ from high-dimensional observation o_t. The features related to actions are important for decision-making, and with the action-related state representation, we could better measure the similarity between the behavior trajectories and the expert trajectories. To learn such state representations, we build an inverse dynamics network $f_\theta(\phi(o_t), \phi(o_{t+1}))$, which takes the subsequent state embeddings as inputs and predicts the action a_t in the transition. In addition to the action-related property, the state representation needs to be beneficial to imitation policy learning as well, since state representation learning and policy learning influence each other in a loop: the training data for the state representation is collected by the learned policy, and the rewards for policy learning are dependent on the learned state representation. Therefore, we augment the inverse dynamics objective with the value function optimization in an end-to-end manner. The overall loss function of representation network ϕ, inverse dynamics network f_θ, and the Q network $Q_\omega(\phi(o_t), a_t)$ is as follows:

$$
\begin{aligned}
L_{\phi,\theta,\omega} = \mathbb{E}_{(o_t,a_t,o_{t+1})\sim\{\tau,\tau^e\}}[(f_\theta(\phi(o_t),\phi(o_{t+1})) - \hat{a}_t)^2] \\
+ \mathbb{E}_{(o_t,a_t,o_{t+1})\sim\tau, r_t=R_i(\tau,\tau^e)}[(r_t + \gamma \min_{i=1,2} \hat{Q}^i(\phi(o_{t+1}), a_{t+1}) \\
- Q^i_\omega(\phi(o_t), a_t))^2|_{i=1,2}],
\end{aligned} \tag{1}
$$

where R_i denotes the imitative reward function, $\hat{Q}$ denotes the target Q network, and a_{t+1} is the action taken by the policy π on observation o_{t+1} with exploration:

$$
a_{t+1} \leftarrow \pi(\phi(o_{t+1})) + \epsilon, \epsilon \sim clip(\mathcal{N}(0,\sigma), -c, c). \tag{2}
$$

The loss in Equation (1) influences the representation network ϕ through stochastic gradient descent, as depicted by the reverse directions of the green lines in Figure 2(a).

To boost the diversity of the training data for state representation, the inverse dynamics objective is optimized with both the expert and behavior trajectories. As we focus on the tasks with continuous action space, the inverse dynamics network is optimized with the mean-squared loss. For tasks with discrete action space, the first term in Equation (1) could be alternated with a cross-entropy loss. Note that the online update of the representation function induces non-stationarity for the rewards, we employ a target representation network for reward generation, where the target representation network is synchronized with the learned representation network for each Δt timesteps.

Imitative Reward Generation

A key component in IRL is generating effective rewards to learn policies resembling the expert policy. The rewards generated by the discriminator in the adversarial IRL frameworks suffer from the non-stationary issue, as the discriminator is updated online with policy learning. Recent IL works (Cohen et al. 2021; Dadashi et al. 2021; Haldar et al. 2023; Papagiannis and Li 2023) propose to generate rewards with trajectory matching, which measures the similarity between the expert trajectories and the behavior trajectories with optimal transport. Since trajectory matching is non-parameteric, the rewards in these methods are stationary. However, as the optimal-transport based trajectory matching methods emphasize the trajectory similarity as a whole, the element-wised similarity, i.e. the similarity of state-action pairs, may be neglected. To solve this problem, we propose a novel reward function that combines the advantages of trajectory matching and state-action pair similarity. The following of this subsection first gives the formulations of trajectory matching rewards and discriminator rewards, and then elaborates on how to integrate them.

Trajectory Matching Rewards Inspired by previous works (Haldar et al. 2023), we compute the closeness between the expert trajectories $\mathcal{T}^e$ and behavior trajectories $\mathcal{T}$ by measuring the optimal transport of probability mass from $\mathcal{T} \to \mathcal{T}^e$. Given a cost function $c : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ defined in the state representation space $\mathcal{Z}$ and an optimal transport objective g, the optimal alignment between an expert trajectory τ^e and a behavior trajectory τ is shown in Equation (3).[1] Note that it is necessary to define the cost function in the state representation space, since the distance in the high-dimensional image observation space cannot measure the similarity between states.

$$
\mu* \in \arg\min_{\mu\in\mathcal{M}} g(\mu, f_\theta(\tau), f_\theta(\tau^e), c), \tag{3}
$$

where $\mathcal{M} = \{\mu \in \mathbb{R}^{T\times T} : \mu\mathbf{1} = \mu^{\mathrm{T}}\mathbf{1} = \frac{1}{T}\mathbf{1}\}$ is the coupling matrix set and the cost function c could be Euclidean or cosine distance.[2] Specifically, we utilize the Wasserstein distance with cosine cost as the optimal transport metric, and then

$$
\begin{aligned}
g(\mu, f_\theta(\tau), f_\theta(\tau^e), c) &= \mathcal{W}^2(f_\theta(\tau), f_\theta(\tau^e)) \\
&= \sum_{t,t'=1}^{T} C_{t,t'}\mu_{t,t'},
\end{aligned} \tag{4}
$$

[1] Here we overwrite the notation f_θ for trajectory embeddings: $f_\theta(\tau) = [f_\theta(o_1), ..., f_\theta(o_T)]$.

[2] We provide an ablation study on the cost function c in Appendix D, and using cosine distance performs slightly better than Euclidean distance.

where the cost matrix $C_{t,t'} = c(f_\theta(o_t), f_\theta(o^e_{t'}))$. By maximizing the rewards in Equation (5), the agent could learn the policy that closely matches the expert trajectories,

$$R_1(o_t) = -\sum_{t'=1}^{T} C_{t,t'}\mu^*_{t,t'}. \tag{5}$$

Solving Equation (3) to compute μ^* is computationally expensive, so we employ the Sinkhorn algorithm (Knight 2008) to obtain an approximate solution. As there are multiple expert trajectories, previous works (Haldar et al. 2023; Cohen et al. 2021) search the expert dataset, and use the nearest trajectory with the current behavior trajectory to compute the rewards in Equation (5). However, the ignored expert trajectories may contain certain state-action pairs, which are closer to the current behavior trajectories but not used to generate rewards, since a trajectory is considered as a whole. To thoroughly utilize the expert dataset, we propose to augment the trajectory-matching rewards with the following discriminator rewards, which estimate the similarity of state-action pairs.

Discriminator Rewards Following the GAIL method (Ho and Ermon 2016), we train a discriminator D that differentiates between the agent's samples and the expert data with the following loss function:

$$\begin{aligned} L_D = &- \mathbb{E}_{(o^e_t, a^e_t)\sim\mathcal{T}^e}[\log D(f_\theta(o^e_t), a^e_t)] \\ &- \mathbb{E}_{(o_t, a_t)\sim\mathcal{T}}[\log(1 - D(f_\theta(o_t), a_t))]. \end{aligned} \tag{6}$$

Similar to trajectory matching, the discriminator D also works in the state representation space. Beyond that, D takes the actions into account, hence the similarity between the expert trajectories and the behavior trajectories could be measured more accurately. By training the discriminator, we derive another reward function:

$$R_2(o_t, a_t) = -\log D(f_\theta(o_t), a_t). \tag{7}$$

To summarize, the trajectory matching reward is derived from a macro view, and the discriminator reward is formulated from a micro view. They are both heavily dependent on the state representation learning in the previous section. By integrating these two kinds of rewards together, we obtain a reward function R_i which could thoroughly describe the similarity between the expert data and the behavior data:

$$R_i(o_t, a_t) = R_1(o_t) + \eta R_2(o_t, a_t). \tag{8}$$

η is a scaling factor balancing these two kinds of imitative rewards. In Appendix A, we provide the pseudocode and the algorithmic details of RILIR.

Related Work

Imitation Learning IL aims to learn policies from demonstrations without access to the environment rewards (Hussein et al. 2017). There are three major paradigms in IL. (1) Behavior cloning (BC) (Pomerleau 1991; Torabi, Warnell, and Stone 2018a; Shafiullah et al. 2022) treats policy learning as a supervised learning problem over state-action pairs. While these methods are appealingly simple, they suffer from compounding errors caused by covariate shift (Ross and Bagnell 2010; Ross, Gordon, and Bagnell 2011). An improved BC method (Brantley, Sun, and Henaff 2020) alleviates the covariate shift problem in specific tasks. (2) Inverse reinforcement learning (IRL) (Ng and Russell 2000; Ziebart et al. 2008; Han et al. 2022) infers rewards from the given demonstrations. Compounding error is not an issue for these methods (Mendez, Shivkumar, and Eaton 2018; Zeng et al. 2022). However, IRL is extremely expensive regarding samples, since after inferring rewards, it still needs to run RL methods in an inner loop to learn the policies. (3) Generative adversarial imitation learning (GAIL) (Ho and Ermon 2016) is an adversarial learning based formulation inspired by maximum entropy IRL (Ziebart et al. 2008) and GANs (Goodfellow et al. 2014). Compared to IRL, this line of research does not need to infer the rewards while regarding the discriminator results as an auxiliary (Dadashi et al. 2021; Papagiannis and Li 2023), which contributes to more efficient learning. Previous works (Baram et al. 2017; Sun et al. 2021) combine model-based learning with GAIL to construct a fully differentiable frame and enable more accurate gradient estimation. Different from these model-based IL methods, the inverse dynamics model in this work is designed for an abstract state representation, and the discriminator reward is inspired by the GAIL paradigm.

Robust Imitation Learning Robustness against the variations between learning and expert environments has recently received much attention. Domain adaptive IL methods (Cetin and Çeliktutan 2021; Kim et al. 2020) seek the consistency between the expert and learning domains with a set of prior data pre-collected in both two domains, and use this consistency for policy learning in a related target task in the learning domain. In contrast to domain-adaptive IL, the proposed approach does not need a prior dataset, and can learn directly in the learning environment instead. Chae et al. (2022) proposed an IL method to deal with the dynamics variance in the learning and expert environments. The proposed approach aims to solve the visual variance problem, which commonly occurs in real-world tasks, e.g., an occluded camera. The SeMAIL method (Wan et al. 2023) tries to solve the visual distractors in IL with model-based learning. However, as the high-dimensional visual observations are hard to reconstruct, the forward model learning in SeMAIL is sample inefficient. In contrast, the proposed approach learns an inverse dynamics model, which has achieved a better performance, as shown in the experiment section.

Visual Representation Learning (Visual RepL) Learning from visual observations is an important problem due to its potential impact on fields like robotics (Hua et al. 2021), autonomous driving (Hawke et al. 2020) and video games (Ye et al. 2023). OpenAI seeks to train general-purpose foundation models for sequential decision-making by utilizing freely available internet-scale unlabeled visual datasets via imitation learning (Baker et al. 2022). However, this paradigm suffers from low sample efficiency and poor generalization ability due to the high-dimensional visual space. Therefore, Visual RepL plays a critical role, which could be divided into the following categories. (1) Representation learning with auxiliary objectives, e.g., learning

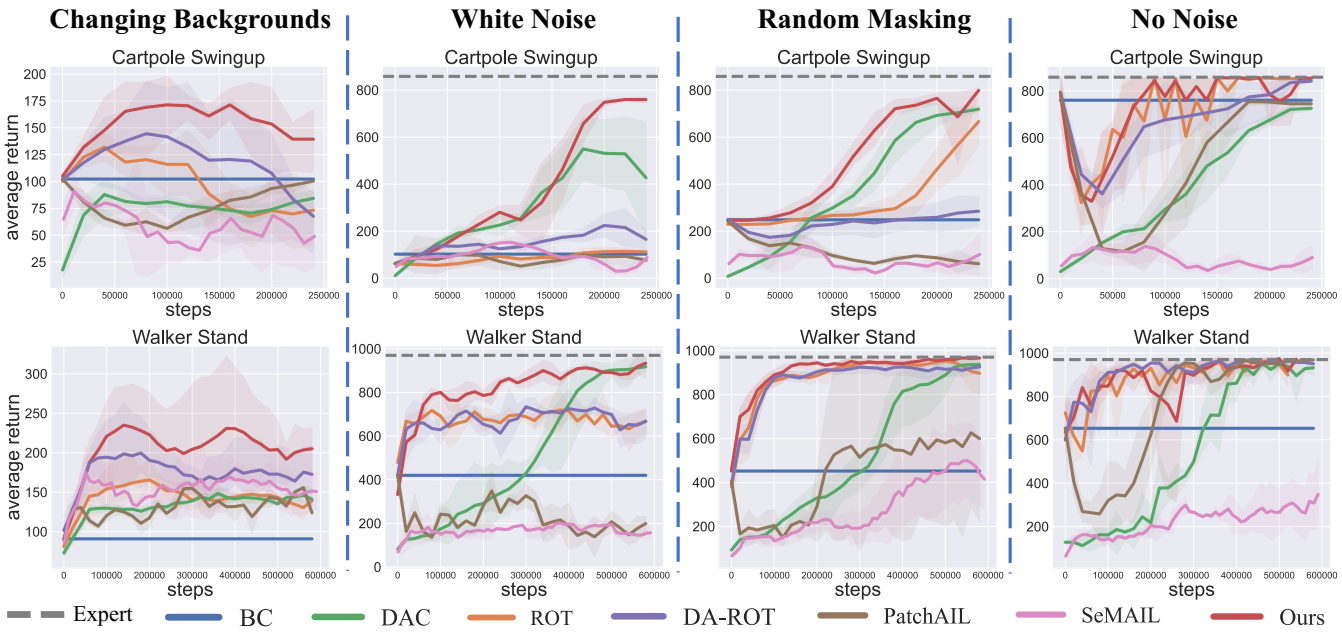


Figure 3: Experiment results under various types of visual shifts. The videos of the learned policies are shown in https://sites.google.com/view/rilir-aaai.

compact state representations with an encoder-decoder architecture (Rafailov et al. 2021), enhancing discriminators with patches (Liu et al. 2023), joint optimization of model, representation and policy (Ghugare et al. 2022). (2) World model based methods (Hu et al. 2022; Wu, Piergiovanni, and Ryoo 2019; Seo et al. 2022). These methods first construct an evolution world by a latent dynamics model that predicts latent states from visual observations, and then the learned latent state is fed to the policy network as an input. Rolling out within the learned world model can help reduce the sample number in the real environment for RL methods (Hafner et al. 2019, 2020). (3) Image augmentation (Yarats, Kostrikov, and Fergus 2021; Laskin, Srinivas, and Abbeel 2020), which is widely studied in the computer vision domain (Hendrycks et al. 2020; Sohn et al. 2020). Recently it has also been utilized to promote representation learning in RL and IL (Wang and Chang 2021; Chen et al. 2022). Our work belongs to the auxiliary-task paradigm, and could be easily combined with the data augmentation methods.

Experiments

We evaluate the proposed approach RILIR on a set of challenging visual control environments with perturbations, aiming at answering the following questions: (1) Can RILIR generally work under various types of visual perturbations? (2) Can RILIR work in different types of tasks, including locomotion and manipulation? (3) How is the state representation learned by RILIR? (4) How important are the various components of the RILIR approach?

Experiment Results Under Various Visual Shifts

In this subsection, we aim to evaluate the ability of RILIR to work under different types of visual shifts, and two tasks from the DeepMind Control (DMC) suite (Tassa et al. 2018) are used for evaluation. RILIR's ability to work in diverse tasks is evaluated in the next subsection. RILIR is compared with state-of-the-art IL methods, including BC, adversarial IL methods, visual IL methods, and robust IL methods. A brief description of the baselines is as follows, and in Appendix C, we provide the hyperparameters for all the baselines and the proposed approach.

- Behavior cloning (BC): Supervised learning method trained with expert demonstrations.
- Discriminator Actor-Critic (DAC) (Kostrikov et al. 2019): An adversarial IL method, which outperforms prior works such as GAIL (Ho and Ermon 2016) and AIRL (Fu, Luo, and Levine 2018).
- Regularized Optimal Transport (ROT) (Haldar et al. 2023): A state-of-the-art visual IL method based on trajectory matching, which not only takes the BC policy as an initialization, but also regularizes the policy updates with the BC objective.
- DA-ROT (Yarats, Kostrikov, and Fergus 2021): As data augmentation has shown its strength in handling visual shifts, we compare RILIR with the DrQ data augmentation version of ROT.
- PatchAIL (Liu et al. 2023): A visual IL method which generates rewards based on patches of images.
- SeMAIL (Wan et al. 2023): A robust IL method which aims to eliminate visual distractors via separated models.

In the experiments, we consider three types of visual shifts, including changing the backgrounds, adding white noise, and random masking.[3] For the experiments of chang-

[3]The experiment results including more aspects of robustness, e.g., changing colors, changing sizes, and adding objects, are shown in Appendix D.

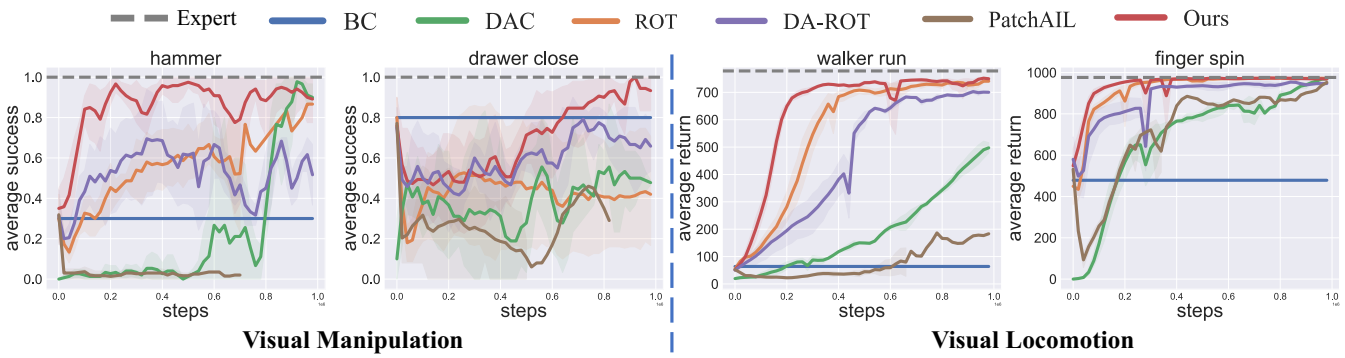


Figure 4: Experiment results in various types of tasks with random masking noises.

ing backgrounds, we use the environments in the easy version of the distracting control suite (Stone et al. 2021). In contrast to the noisy learning environments, the environments for collecting the expert demonstrations are free of noise. We provide the details of the learning environments and the way of collecting expert datasets in Appendix B. Besides, we conduct comparative experiments in "clean" environments without visual distractors as well. Figure 3 demonstrates the results in the CartPole Swingup task and the Walker Stand task under various types of visual shifts. The y axis shows the average return over 10 episodes. Each line is the mean of 3 runs with shaded regions corresponding to a confidence interval of 95%. All the curves have been smoothed equally for visual clarity. The code to reproduce these results is available in the supplementary material.

As shown in the first three columns in Figure 3, the proposed approach significantly outperforms the baseline methods.[4] In the "clean" environments, the performance of all the methods is nearly the same. Similar to ROT, our approach and PatchAIL also utilize the behavior-cloned policy as an initialization for the policy network, so there is a jump start in these learning curves. Comparing the columns in Figure 3, we find that changing backgrounds is severely more challenging than other visual shifts. Even in this challenging environment, the proposed approach can achieve positive learning and perform better than the baselines. In other types of environments, the proposed approach accomplishes a near-expert learning performance using much fewer samples than the baselines.

Note that ROT (Haldar et al. 2023) is a state-of-the-art visual IL method, which has shown better performance than DAC, and we also compare the proposed approach with the DrQ data augmented version of ROT (DA-ROT). Data augmentation improves the learning performance of ROT in some cases, but not all cases, possibly because data augmentation induces a heavier computation burden and cannot cover all types of visual shifts. PatchAIL has a similar performance with DAC in the "clean" environments, but the performance of PatchAIL drops severely in the noisy environments. This phenomenon implies that segmenting images into patches and generating rewards by taking averages over patches can hardly help the agent adapt to visual shifts. SeMAIL is a robust visual IL method with separated forward models. This method achieves a comparable performance with other baselines in the Walker Stand task with changing backgrounds. However, SeMAIL is not as efficient as other baselines in the environments with easier visual perturbations due to the heavy burden of optimizing a forward model. We provide the results of running SeMAIL with more steps in Appendix D.

[4]These experiments have been run with A100 GPUs, and each run takes no more than 1 day.

Experiment Results in Diverse Tasks

To evaluate RILIR's ability to work in diverse tasks, we conduct experiments in two domains: manipulation and locomotion. Two tasks in the Meta-World benchmark (Yu et al. 2020) are used as the manipulation tasks, and two tasks in the DMC suite are used as the locomotion tasks. These tasks are with medium difficulty, and the difficulty of tasks is measured by the dimension of action spaces. The learning results in harder tasks are provided in Appendix D. To evaluate the robustness of these IL methods, the visual observations in the learning environments are randomly masked.

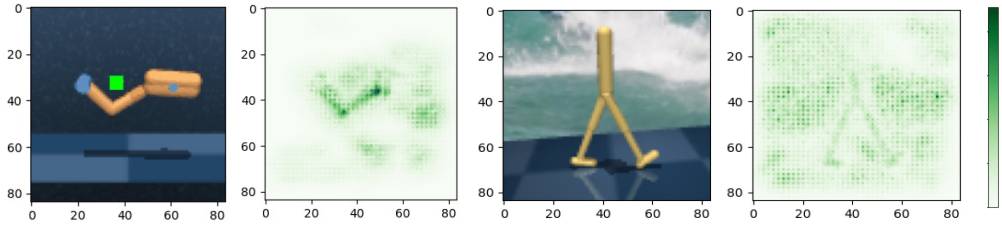
As shown in Figure 4, in both manipulation and locomotion domains, the proposed approach significantly outperforms the baselines regarding learning efficiency and convergent performance. Benefiting from the inverse dynamics representation learning and effective imitative rewards, the return or success gap between the policy learned by the proposed method and the expert policy is less than 5%. Similar to the results in Figure 3, PatchAIL suffers from the random masking noises. In the following subsection, we analyze the state representation learning in our method to probe into the reason why it can achieve such a good performance.

Analysis of the State Representation

In this subsection, we analyze the learned representation in the RILIR approach. Specifically, we visualize the state representations in the form of saliency maps (Simonyan, Vedaldi, and Zisserman 2013) to analyze which regions have been paid more attention by the representation function. The saliency map is calculated as follows,

$$\sum_i \left| \frac{\partial(\phi_i(o_t))}{\partial o_t} \right|, \tag{9}$$

where $\phi_i(o_t)$ denotes the i-th element of the state representation $\phi(o_t)$. A pixel with a larger saliency value has a larger



(a) Finger Spin (b) Saliency Map (c) Walker Stand (d) Saliency Map

Figure 5: (a)(b) Example observation in the Finger Spin task with random masking perturbation, and the corresponding saliency map at 200k steps. (c)(d) Example observation in the Walker Stand task with changing background distractors from the Distracting Control suite (Stone et al. 2021), and the corresponding saliency map at 600k steps. Note that the saliency maps are only used to analyze the learned representation model, which has not been involved in the training process.

influence on the Q network and the policy network, since $\phi(o_t)$ is the input of these two networks.

A darker green color denotes a larger saliency value. In the locomotion tasks, the proprioception features of the robot are substantially important. As shown in Figure 5, our method clearly extracts the position of the *Finger* in the Finger Spin task and ignores the random masking distractors. With a changing background, our approach can still extract the leg positions of the simulated Walker robot, as shown in Figure 5 (c) and (d). The state representations in more tasks are shown in https://sites.google.com/view/rilir-aaai, and we provide a comparison with the representations learned by the baselines in Appendix D.

Ablation Studies

To validate the effectiveness of various components in the proposed approach, we conduct ablation studies in two tasks, Hammer and Drawer Close, with random masking perturbations in the environments. Hammer is relatively easy, and Drawer Close is a difficult task, as it takes nearly 1 million steps for RILIR to converge. In this subsection, we have respectively ablated the representation learning module and the rewards in RILIR. "Ours w/o representation" in Figure 6 denotes the experiments removing the inverse dynamics representation objective, and "Ours w/o discriminator" denotes removing the discriminator rewards, i.e., only using the trajectory matching rewards. More ablation studies in both the noisy environment and the "clean" environment are provided in Appendix D.

Figure 6: Ablation studies of the state representation learning and the reward function in the proposed approach RILIR.

The results in Figure 6 show that both the representation learning module and the discriminator rewards are effective. Specifically, the discriminator rewards mainly affect learning efficiency, and the inverse dynamics representation learning influences convergent performance. Beyond that, in more challenging tasks (Drawer Close), the influence is more obvious. Furthermore, among the three curves, the shaded areas of the proposed approach are the smallest. This implies that the state representation learning objective and the compound rewards help stabilize the learning process.

Conclusion and Limitations

In this work, we have proposed a robust visual imitation learning approach based on inverse dynamics representation learning, dubbed RILIR, which is able to resist the difference in the expert environment and the learning environment, since the representation module extracts the common parts in these two environments. Based on this abstract state representation, we develop a thorough reward function, which considers the similarity of expert data and behavior data from both an element-wise view and a trajectory-level view. Extensive experiment results in a set of challenging visual control tasks demonstrate that the proposed approach has achieved substantially better performance than prior visual IL works and robust IL works. Furthermore, we have conducted ablation studies to validate the effectiveness of the representation learning module and the reward function in the RILIR approach.

However, we recognize a few limitations in this work: (a) The proposed approach may not work well when the observations in the learning environments are significantly different from those in the expert demonstrations, and more advanced representation learning methods need to be investigated in the robust visual IL domain. (b) The inverse dynamics objective may not be the right thing when the dynamics models are not consistent between the learning environment and the expert environment, as this objective relies on the invariant dynamics model. A recent work (Chae et al. 2022) proposed to solve the robust IL problem with various dynamics by imitating multiple experts simultaneously.

Acknowledgements

The authors would like to thank the anonymous reviewers for their valuable comments and helpful suggestions. The work is supported by National Natural Science Foundation of China (62306088), National Natural Science Foundation of China (62103386), and Young Elite Scientists Sponsorship Program by CAST (2022QNRC001).

References

Abbeel, P.; and Ng, A. Y. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, 1.

Arora, S.; and Doshi, P. 2021. A survey of inverse reinforcement learning: Challenges, methods and progress. *Artificial Intelligence*, 297: 103500.

Baker, B.; Akkaya, I.; Zhokov, P.; Huizinga, J.; Tang, J.; Ecoffet, A.; Houghton, B.; Sampedro, R.; and Clune, J. 2022. Video PreTraining (VPT): Learning to Act by Watching Unlabeled Online Videos. In *NeurIPS*.

Baram, N.; Anschel, O.; Caspi, I.; and Mannor, S. 2017. End-to-end differentiable adversarial imitation learning. In *International Conference on Machine Learning*, 390–399. PMLR.

Brantley, K.; Sun, W.; and Henaff, M. 2020. Disagreement-Regularized Imitation Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Cetin, E.; and Çeliktutan, O. 2021. Domain-Robust Visual Imitation Learning with Mutual Information Constraints. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Chae, J.; Han, S.; Jung, W.; Cho, M.; Choi, S.; and Sung, Y. 2022. Robust Imitation Learning against Variations in Environment Dynamics. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvári, C.; Niu, G.; and Sabato, S., eds., *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, 2828–2852. PMLR.

Chen, X.; Toyer, S.; Wild, C.; Emmons, S.; Fischer, I.; Lee, K.; Alex, N.; Wang, S. H.; Luo, P.; Russell, S.; Abbeel, P.; and Shah, R. 2022. An Empirical Investigation of Representation Learning for Imitation. *CoRR*, abs/2205.07886.

Cohen, S.; Amos, B.; Deisenroth, M. P.; Henaff, M.; Vinitsky, E.; and Yarats, D. 2021. Imitation Learning from Pixel Observations for Continuous Control. In *Deep RL Workshop NeurIPS*.

Dadashi, R.; Hussenot, L.; Geist, M.; and Pietquin, O. 2021. Primal Wasserstein Imitation Learning. In *ICLR 2021-Ninth International Conference on Learning Representations*.

Fu, J.; Luo, K.; and Levine, S. 2018. Learning Robust Rewards with Adverserial Inverse Reinforcement Learning. In *International Conference on Learning Representations*.

Fujimoto, S.; Hoof, H.; and Meger, D. 2018. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, 1587–1596. PMLR.

Ghugare, R.; Bharadhwaj, H.; Eysenbach, B.; Levine, S.; and Salakhutdinov, R. 2022. Simplifying Model-based RL: Learning Representations, Latent-space Models, and Policies with One Objective. *CoRR*, abs/2209.08466.

Goodfellow, I. J.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A. C.; and Bengio, Y. 2014. Generative Adversarial Nets. In Ghahramani, Z.; Welling, M.; Cortes, C.; Lawrence, N. D.; and Weinberger, K. Q., eds., *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, 2672–2680.

Hafner, D.; Lillicrap, T. P.; Ba, J.; and Norouzi, M. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hafner, D.; Lillicrap, T. P.; Fischer, I.; Villegas, R.; Ha, D.; Lee, H.; and Davidson, J. 2019. Learning Latent Dynamics for Planning from Pixels. In Chaudhuri, K.; and Salakhutdinov, R., eds., *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, 2555–2565. PMLR.

Haldar, S.; Mathur, V.; Yarats, D.; and Pinto, L. 2023. Watch and match: Supercharging imitation with regularized optimal transport. In *Conference on Robot Learning*, 32–43. PMLR.

Han, D.; Kim, H.; Lee, H.; Ryu, J.; and Zhang, B. 2022. Robust Imitation via Mirror Descent Inverse Reinforcement Learning. In *NeurIPS*.

Hawke, J.; Shen, R.; Gurau, C.; Sharma, S.; Reda, D.; Nikolov, N.; Mazur, P.; Micklethwaite, S.; Griffiths, N.; Shah, A.; and Kendall, A. 2020. Urban Driving with Conditional Imitation Learning. In *2020 IEEE International Conference on Robotics and Automation, ICRA 2020, Paris, France, May 31 - August 31, 2020*, 251–257. IEEE.

Hendrycks, D.; Mu, N.; Cubuk, E. D.; Zoph, B.; Gilmer, J.; and Lakshminarayanan, B. 2020. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Ho, J.; and Ermon, S. 2016. Generative adversarial imitation learning. *Advances in neural information processing systems*, 29.

Hu, A.; Corrado, G.; Griffiths, N.; Murez, Z.; Gurau, C.; Yeo, H.; Kendall, A.; Cipolla, R.; and Shotton, J. 2022. Model-Based Imitation Learning for Urban Driving. In *NeurIPS*.

Hua, J.; Zeng, L.; Li, G.; and Ju, Z. 2021. Learning for a robot: Deep reinforcement learning, imitation learning, transfer learning. *Sensors*, 21(4): 1278.

Hussein, A.; Gaber, M. M.; Elyan, E.; and Jayne, C. 2017. Imitation Learning: A Survey of Learning Methods. *ACM Comput. Surv.*, 50(2): 21:1–21:35.

Kim, K.; Gu, Y.; Song, J.; Zhao, S.; and Ermon, S. 2020. Domain adaptive imitation learning. In *International Conference on Machine Learning*, 5286–5295. PMLR.

Knight, P. A. 2008. The Sinkhorn–Knopp algorithm: convergence and applications. *SIAM Journal on Matrix Analysis and Applications*, 30(1): 261–275.

Kostrikov, I.; Agrawal, K. K.; Dwibedi, D.; Levine, S.; and Tompson, J. 2019. Discriminator-Actor-Critic: Addressing Sample Inefficiency and Reward Bias in Adversarial Imitation Learning. In *International Conference on Learning Representations*.

Laskin, M.; Srinivas, A.; and Abbeel, P. 2020. CURL: Contrastive Unsupervised Representations for Reinforcement Learning. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, 5639–5650. PMLR.

Liu, M.; He, T.; Zhang, W.; Yan, S.; and Xu, Z. 2023. Visual Imitation Learning with Patch Rewards. *CoRR*, abs/2302.00965.

Mendez, J. A.; Shivkumar, S.; and Eaton, E. 2018. Lifelong Inverse Reinforcement Learning. In Bengio, S.; Wallach, H. M.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, 4507–4518.

Ng, A. Y.; and Russell, S. 2000. Algorithms for Inverse Reinforcement Learning. In Langley, P., ed., *Proceedings of the Seventeenth International Conference on Machine Learning (ICML 2000), Stanford University, Stanford, CA, USA, June 29 - July 2, 2000*, 663–670. Morgan Kaufmann.

Papagiannis, G.; and Li, Y. 2023. Imitation learning with sinkhorn distances. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part IV*, 116–131. Springer.

Pomerleau, D. 1991. Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Comput.*, 3(1): 88–97.

Rafailov, R.; Yu, T.; Rajeswaran, A.; and Finn, C. 2021. Visual Adversarial Imitation Learning using Variational Models. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y. N.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, 3016–3028.

Ross, S.; and Bagnell, D. 2010. Efficient Reductions for Imitation Learning. In Teh, Y. W.; and Titterington, D. M., eds., *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15, 2010*, volume 9 of *JMLR Proceedings*, 661–668. JMLR.org.

Ross, S.; Gordon, G. J.; and Bagnell, D. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In Gordon, G. J.; Dunson, D. B.; and Dudík, M., eds., *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2011, Fort Lauderdale, USA, April 11-13, 2011*, volume 15 of *JMLR Proceedings*, 627–635. JMLR.org.

Scheller, C.; Schraner, Y.; and Vogel, M. 2020. Sample efficient reinforcement learning through learning from demonstrations in minecraft. In *NeurIPS 2019 Competition and Demonstration Track*, 67–76. PMLR.

Seo, Y.; Hafner, D.; Liu, H.; Liu, F.; James, S.; Lee, K.; and Abbeel, P. 2022. Masked World Models for Visual Control. In Liu, K.; Kulic, D.; and Ichnowski, J., eds., *Conference on Robot Learning, CoRL 2022, 14-18 December 2022, Auckland, New Zealand*, volume 205 of *Proceedings of Machine Learning Research*, 1332–1344. PMLR.

Shafiullah, N. M.; Cui, Z. J.; Altanzaya, A.; and Pinto, L. 2022. Behavior Transformers: Cloning k modes with one stone. In *NeurIPS*.

Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Sohn, K.; Berthelot, D.; Carlini, N.; Zhang, Z.; Zhang, H.; Raffel, C.; Cubuk, E. D.; Kurakin, A.; and Li, C. 2020. FixMatch: Simplifying Semi-Supervised Learning with Consistency and Confidence. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Stone, A.; Ramirez, O.; Konolige, K.; and Jonschkowski, R. 2021. The Distracting Control Suite–A Challenging Benchmark for Reinforcement Learning from Pixels. *arXiv preprint arXiv:2101.02722*.

Sun, J.; Yu, L.; Dong, P.; Lu, B.; and Zhou, B. 2021. Adversarial inverse reinforcement learning with self-attention dynamics model. *IEEE Robotics and Automation Letters*, 6(2): 1880–1886.

Tassa, Y.; Doron, Y.; Muldal, A.; Erez, T.; Li, Y.; Casas, D. d. L.; Budden, D.; Abdolmaleki, A.; Merel, J.; Lefrancq, A.; et al. 2018. Deepmind control suite. *arXiv preprint arXiv:1801.00690*.

Torabi, F.; Warnell, G.; and Stone, P. 2018a. Behavioral Cloning from Observation. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 4950–4957. ijcai.org.

Torabi, F.; Warnell, G.; and Stone, P. 2018b. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*.

Wan, S.; Wang, Y.; Shao, M.; Chen, R.; and Zhan, D.-C. 2023. SeMAIL: Eliminating Distractors in Visual Imitation via Separated Models. *arXiv preprint arXiv:2306.10695*.

Wang, T.; and Chang, D. E. 2021. Robust Navigation for Racing Drones based on Imitation Learning and Modularization. In *IEEE International Conference on Robotics and Automation, ICRA 2021, Xi'an, China, May 30 - June 5, 2021*, 13724–13730. IEEE.

Wu, A.; Piergiovanni, A. J.; and Ryoo, M. S. 2019. Model-based Behavioral Cloning with Future Image Similarity Learning. In Kaelbling, L. P.; Kragic, D.; and Sugiura, K., eds., *3rd Annual Conference on Robot Learning, CoRL 2019, Osaka, Japan, October 30 - November 1, 2019, Proceedings*, volume 100 of *Proceedings of Machine Learning Research*, 1062–1077. PMLR.

Yarats, D.; Kostrikov, I.; and Fergus, R. 2021. Image Augmentation Is All You Need: Regularizing Deep Reinforcement Learning from Pixels. In *International Conference on Learning Representations*.

Ye, W.; Zhang, Y.; Abbeel, P.; and Gao, Y. 2023. Become a Proficient Player with Limited Data through Watching Pure Videos. In *The Eleventh International Conference on Learning Representations*.

Yu, T.; Quillen, D.; He, Z.; Julian, R.; Hausman, K.; Finn, C.; and Levine, S. 2020. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on robot learning*, 1094–1100. PMLR.

Zeng, S.; Li, C.; Garcia, A.; and Hong, M. 2022. Maximum-Likelihood Inverse Reinforcement Learning with Finite-Time Guarantees. In *NeurIPS*.

Ziebart, B. D.; Maas, A. L.; Bagnell, J. A.; and Dey, A. K. 2008. Maximum Entropy Inverse Reinforcement Learning. In Fox, D.; and Gomes, C. P., eds., *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, AAAI 2008, Chicago, Illinois, USA, July 13-17, 2008*, 1433–1438. AAAI Press.