2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Off-policy Imitation Learning from Visual Inputs

Zhihao Cheng, Li Shen, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Recently, various successful applications utilizing expert states in imitation learning (IL) have been witnessed. However, IL from visual inputs (ILfVI), which has a greater promise to be widely applied by using online visual resources, suffers from low data-efficiency and poor performance resulted from on-policy learning and high-dimensional visual inputs. We propose OPIfVI (Off-Policy Imitation from Visual Inputs), which is composed of an off-policy learning manner, data augmentation, and encoder techniques, to tackle the mentioned challenges, respectively. More specifically, to improve data-efficiency, OPIfVI conducts IL in an off-policy manner, with which sampled data used multiple times. In addition, we enhance the stability of OPIfVI with spectral normalization to mitigate the side effect of off-policy training. The core factor, contributing to the poor performance of ILfVI, that we think is agents could not extract meaningful features from visual inputs. Hence, OPIfVI employs data augmentation from computer vision to help train encoders to better extract features from visual inputs. Besides, a specific structure of gradient backpropagation for the encoder is designed to stabilize the encoder training. At last, we demonstrate that OPIfVI can achieve expert-level performance and outperform existing baselines via extensive experiments using DeepMind Control Suite.**

I. Introduction

Imitation learning (IL) empowers agents to learn from expert data instead of designing an explicit reward function [1] and has achieved remarkable successes in graphics [2], online games [3], and robotic manipulation [4]. The expert data in IL can be divided into two categories [5], [6], demonstrations and observations, among which demonstrations contain states and actions of experts' experiences, whereas observations only consist of states. In real world applications, the state is the proprioceptive state of an expert, which could be hard to access and record. By contrast, intelligent creatures grasp knowledge or skills by observing how peer fellows accomplish tasks without knowing their proprioceptive states [7]. In other words, intelligent creatures generally learn with visual inputs rather than state inputs. This learning scheme is more practical, but has been less studied in the IL community.

The thought to enable agents to learn like intelligent creatures leads to a spectrum of IL, imitation learning from visual inputs (or images, or pixels) (ILfVI), which is also referred to as visual imitation learning (VIL) [8], [9]. Here, we denote it as ILfVI to emphasize that visual inputs can be further classified. Corresponding to traditional state demonstrations and observations, visual inputs fall into visual demonstrations and visual observations, where the former contains images and actions while the latter merely includes images. In contrast to dramatic successes in IL from state inputs [1], [10], [11], there is only a little research [10], [12] on ILfVI. Furthermore, the performance and sample efficiency of ILfVI methods is still far from satisfactory to be employed in reality.

Compared to state inputs, the major difference is that images in ILfVI are partially-observed high-dimensional inputs. This difference introduces several challenges: 1) visual inputs are partially observed from states, which converts underlying dynamics from Markov Decision Process (MDP) [13] to Partially Observable MDP (POMDP) [14]; 2) high-dimensional inputs contain a large portion of redundant information [15], distracting agents from extracting useful information for decision-making; 3) the adversarial training paradigm in IL suffers from training instability with high-dimensional inputs [16]. In addition, the process of learning to extract meaningful features in ILfVI further aggravates low data-efficiency for an on-policy training manner, which can only use sampled examples once. Despite these challenges, it is worthy of improving ILfVI owing to its promising applications.

Considering the above challenges, we propose OPIfVI (Off-policy Imitation from Visual Inputs) to improve both data-efficiency and performance. To be more data-efficient, we build OPIfVI in an off-policy manner, where sampled data are stored in a replay buffer such that data can be utilized multiple times. Spectral normalization is adopted to enhance the stability of this off-policy training scheme. We borrow the idea of data augmentation from computer vision to help encoders extract meaningful features from images. Then, extracted features are forwarded to agents to make decisions. Data augmentation enlarges sampled data and can benefit data-efficiency to some extent. Furthermore, we design a specific structure for the gradient backpropagation to train and stabilize encoders. In this structure, the actor and critic in the generator share an encoder, while the discriminator maintains an independent encoder. The encoder in the generator is updated with only gradients backpropagated from the critic, whereas the other encoder is trained with the discriminator loss. Combing these aspects, we propose OPIfVI, which can efficiently and effectively learn from visual inputs. We evaluate OPIfVI with visual demonstrations and visual observations on various DeepMind Control tasks [17], showing that OPIfVI significantly surpasses the other baselines in terms of both data-efficiency and performance.

Zhihao Cheng is partially supported by ARC project FL-170100117.

[1] Z. Cheng is with the University of Sydney, Australia zche3121@uni.sydney.edu.au

[2] L. Shen is with the JD Explore Academy, China mathshenli@gmail.com

[3] D. Tao is with the JD Explore Academy, China, and the University of Sydney, Australia dacheng.tao@gmail.com

II. Related Work

a) IL from State Inputs: IL allows reproducing policies that can imitate expert behaviors with expert data. IL algorithms can be split into different classes from various aspects. For example, based on the learning mechanism, IL can be divided into behavioral cloning (BC) [18] and inverse reinforcement learning (IRL) [19]. BC takes IL as pure supervised learning, while IRL first reconstructs a reward function with expert data and then conducts ordinary RL. For more works, please refer to [1], [10], [11], [20], [21]. From the perspective of expert data, IL can be categorized into learning from demonstration (LfD) and learning from observation (LfO) [5], [6]. Demonstrations contain both states and actions of experts, whereas observations only consist of expert states. LfO algorithms are often extended from the LfD version, such as BCO [22] and GAIfO [10]. Most IRL algorithms employ the on-policy learning to maintain accurate estimations of occupancy measures [1], [10], resulting in low data-efficiency.

b) IL from Visual Inputs: ILfVI is attracting more attention owing to its broad application prospects. However, there is only a little research on ILfVI. InfoGAIL [12] is proposed to deal with visual demonstrations sampled from diverse experts by learning latent variables from expert data. To cope with visual demonstrations, Young *et al.* [8] enhance BC with data augmentation and develop VBC. Due to the compounding error, VBC could perform poorly in complex environments. Experiments of GAIfO with visual observations show that GAIfO only achieves about half of the expert performance in no-trivial environments [10]. Two concurrent works [9], [23] give further insights into ILfVI. Rafailov *et al.* [9] solve ILfVI from a model-based perspective, whose algorithm V-MAIL first learns a world model and then updates the discriminator with on-policy samples from the learned model. Model-based methods could be computation-expensive owing to learning of a model. By contrast, [23] adopts a model-free learning scheme, who employs the encoder in DrQ-v2 [24] to help extract features and achieve expert-level performance.

c) Data Augmentation: In computer vision (CV), data augmentation is one of the basic techniques, which has long been studied [25]–[27]. However, data augmentation is rarely adopted in RL or IL [28] for state inputs because the state s in MDP is unique, and any modification will lead to a state that represents different information. For visual inputs, recently, data augmentation has been employed in RL and significantly improves the performance [28]–[31]. For example, [30] illustrates that general data augmentation methods enable agents to achieve excellent performance in RL from visual inputs. In ILfVI, data augmentation also demonstrates remarkable performance gains [8], [9].

Our work bears some resemblance to the two concurrent works [9], [23]. Compared to [9], we solve ILfVI from a model-free perspective, which merely uses off-policy samples instead of a learned model and even works for visual observations. In contrast to [23], we dynamically calculate rewards with the latest discriminator and employ spectral normalization to stabilize the learning process, surpassing the former in terms of both data-efficiency and performance.

III. Preliminaries

a) Markov Decision Process (MDP [13]): We consider an MDP described by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma)$, with the state space $\mathcal{S}$, the action space $\mathcal{A}$, the transition distribution $\mathcal{T} = \mathcal{T}(s'|s,a)$, the reward function $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, and the discount factor γ. We denote a stochastic policy for the agent as $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to [0,1]$, where $s \in \mathcal{S}$ and $a \in \mathcal{A}$. A trajectory $\tau = \{s_t, a_t\}_{t=0}^{\infty}$ can be obtained via interactions between policy $\pi(a|s)$ and the environment, where t is the current timestep, the initial state s_0 is sampled from the probability distribution $s_0 \sim \rho_0(s_0)$, $a_t \sim \pi(a_t|s_t)$, and $s_{t+1} \sim \mathcal{T}(s_{t+1}|s_t, a_t)$. Then, the expected discounted reward is $J(\pi) = \mathbb{E}_{\tau\sim\pi}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)\right]$. RL algorithms are supposed to find the optimal policy $\pi^*(a|s)$, which can achieve the maximum episode cumulative reward $J^*(\pi)$.

When using visual inputs, the MDP turns to POMDP [28]. Compared to MDP, POMDP can be formulated with a 7-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{T}, r, \gamma, \Omega, O)$, where the two additional elements Ω and O is the space of observations and the probability that $O(o_{t+1}|s_t, a_t)$, respectively. Agents can merely receive partially observed information o instead of state s in POMDP. A routine solution to deal with the partial observability is to stack several adjacent visual inputs together and then regard it as a state $s_t \approx \boldsymbol{o_t} = \{o_t, o_{t-1}, \cdots\}$ [28], [30], [32], which is employed in the paper.

b) Generative Adversarial Imitation Learning: GAIL [1] adopts the framework of GAN training [16] and can be formalized as a minimax problem:

$$\min_{\theta} \max_{\omega} \mathbb{E}_{\pi_\theta}[\log D_\omega(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D_\omega(s,a))] - \beta\mathcal{H}(\pi_\theta), \tag{1}$$

where $D_\omega(s,a)$ is a discriminator that measures the similarity of agent's state-action pairs to expert ones, ω denotes the discriminator's parameter, and $\mathcal{H}(\pi) = \mathbb{E}_\pi[-\log \pi(a|s)]$ is the entropy of policy π with weight $\beta \geq 0$.

c) Generative Adversarial Imitation from Observation: GAIfO [10] is extended from GAIL to learn from expert data with the absence of actions. The only difference lying between the two approaches is that GAIfO uses state-state (s, s') pairs, whereas GAIL utilizes state-action (s, a) pairs. GAIfO is formalized as follows

$$\min_{\theta} \max_{\omega} \mathbb{E}_{\pi_\theta}[\log D_\omega(s,s')] + \mathbb{E}_{\pi_E}[\log(1 - D_\omega(s,s'))].$$

When using single states instead of state-state pairs, GAIfO degrades to GAIfO-s as in [33].

IV. Off-policy Imitation from Visual Inputs

Here, we describe our off-policy IL algorithm for learning from visual inputs, OPIfVI, which is presented in Fig. 1. We begin by formalizing the off-policy ILfVI problem and introducing the challenges responsible for low data-efficiency and poor performance of current IL algorithms. To alleviate these challenges, OPIfVI adopts three major modifications

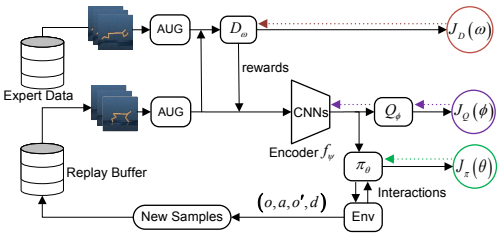


Fig. 1: The framework of OPIfVI. The main components in OPIfVI are a replay buffer, an expert data set, a data augmentation block (called AUG), an encoder f_ψ, a policy π_θ, the Q-value function Q_ϕ, and a discriminator D_ω. The solid lines with arrows denote directions of information flow, while the colored dotted lines represent gradients backpropagated from loss functions.

compared to previous works: 1) an off-policy IL paradigm with enhanced stability (Section IV-A); 2) data augmentation for better feature extraction (Section IV-B); 3) a specifically designed gradients backpropagation scheme for the training of encoders (Section IV-C). By virtue of the three modifications, our algorithm–OPIfVI is able to primely achieve expert-level performance in ILfVI with high data-efficiency, surpassing the other baselines.

a) Problem Formulation: Two major differences are lying between our ILfVI setting and the previous ones [1], [10]. The first is that agents only receive high-dimensional partially-observed images instead of low-dimensional fully-observed states. Although we can stack several consecutive images into $\boldsymbol{o}$ and roughly regard $\boldsymbol{o}$ as a state, the high-dimensional inputs still make it challenging to imitate. We consider the problem of learning from visual demonstrations and observations, *i.e.*, the expert data are $\tau_E = \{(\boldsymbol{o}, a)\}_0^N$ or $\tau_E = \{(\boldsymbol{o}, \boldsymbol{o'})\}_0^N$, respectively. Furthermore, we also study a degraded setting in visual observations, where only single observations rather than neighboring observation pairs are provided such that $\tau_E = \{(\boldsymbol{o})\}_0^N$. To unify the three kinds of expert data, we define a symbol $\boldsymbol{x}$ such that $\boldsymbol{x} = (\boldsymbol{o}, a)$, $(\boldsymbol{o}, \boldsymbol{o'})$, or $(\boldsymbol{o})$. The remaining difference is that we use off-policy samples to conduct IL. With off-policy samples, the source distribution of samples for updating parameters would substantially change, increasing training instability. Our ILfVI problem is formalized as follows:

$$\min_\theta \max_\omega \ \mathbb{E}_{\mu^{\mathcal{R}}(\boldsymbol{x})}[\log D_\omega(\boldsymbol{x})] + \mathbb{E}_{\pi_E}[\log(1 - D_\omega(\boldsymbol{x}))],$$

where $\mu^{\mathcal{R}}(\boldsymbol{x})$ is the distribution of $\boldsymbol{x}$ in a replay buffer $\mathcal{R}$. Samples in the replay buffer are recorded from historical policies, which is considered off-policy.

b) Algorithm Overview: The framework of OPIfVI is presented in Fig. 1. Similar to SAC [34], there are four Q-value functions in OPIfVI, including two alternate Q-value functions (Q_{ϕ_1} and Q_{ϕ_2}) and two target ones ($Q_{\bar{\phi}_1}$ and $Q_{\bar{\phi}_2}$). For simplicity, we denote them with a unified symbol Q_ϕ in Fig. 1. The policy π_θ and Q-value functions share an identical encoder f_ψ that is constructed with convolutional neural networks (CNNs) and is used to extract features from images. With interactions between the policy π_θ and the environment, we collect samples and store them into a replay buffer $\mathcal{R}$. For training, we first randomly sample data from the replay buffer and augment sampled images with random crop. Then the augmented data are used to calculate rewards with the current discriminator D_ω. Subsequently, we input the data into the encoder f_ψ followed by the Q-value function Q_ϕ as well as the policy π_θ. According to the losses in Algorithm 1, we backpropagate gradients to update the parameters ϕ, θ, ψ. Note that the encoder is only trained with gradients from Q-value functions. The discriminator D_ω maintains a separate encoder with the same structure to the one of the policy and Q-value functions. The discriminator uses samples from the replay buffer and expert data set to update its parameter ω with the loss in Algorithm 1. Images inputted to the discriminator are also augmented, and spectral normalization is adopted to enhance the stability of discriminator training.

A. Off-policy Learning

To improve data-efficiency, OPIfVI adopts an off-policy training manner via an off-policy generator SAC [34]. SAC uses a replay buffer $\mathcal{R}$ to store historical samples that are collected by previous policies and randomly samples data from this buffer to train its policy and Q-value networks [34]. Such a replay buffer enables samples to be utilized multiple times for policy improvement, thus making learning more data-efficient. However, this off-policy learning scheme poses a threat to the training stability of OPIfVI.

Compared to on-policy adversarial IL algorithms [10], [11], OPIfVI updates the discriminator with data from a replay buffer to improve its ability on discriminating samples, resulting in an off-policy training mode for the discriminator. It is difficult to estimate the characteristics of the current generator and discriminator from off-policy samples as accurately as on-policy samples. As a result, the off-policy update of the adversarial training structure would be less stable [35]. What's worse, this off-policy regime is likely to over-fit to training data, leading to severe training instability or even failures of imitation as shown in [9], [36]. In OPIfVI, to enhance the training stability against the drawbacks of off-policy learning, we employ spectral normalization [37], [38] to force the discriminator to be local Lipschitz-continuous. Local Lipschitz continuity of the learned reward function is necessary to achieve excellent performance of off-policy adversarial IL algorithms [39].

With spectral normalization, the performance of OPIfVI as well as its stability are significantly improved. In particular, different from [23], OPIfVI does not save rewards into the replay buffer $\mathcal{R}$ and re-calculates rewards with the newest discriminator D_ω for the policy improvement. This dynamical reward calculation provides more accurate rewards for policy update and can learn much faster than [23].

B. Data Augmentation

Unlike state inputs, where every element of a state s stands for practical physical meanings and is irreplaceable, images contain plenty of redundant information. In ILfVI, agents struggle to select actions based on these high-dimensional inputs and need first to extract meaningful features from pixels. For example, in robot locomotion tasks, we suppose agents

Algorithm 1 Off-policy Imitation from Visual Inputs (OPIfVI)

Inputs: Expert trajectories τ_E.
Hyperparameters: Total iteration number M, replay buffer size H, initial number of samples B, image augmentation AUG, minibatch size N, learning rate η, polyak averaging ρ, discount γ, target minimum entropy $\bar{\mathcal{H}}$, and temperature α.
Parameters: Denote the encoder as f_ψ, policy as π_θ, Q-values and target Q-values as Q_{ϕ_i} and $Q_{\bar{\phi}_i}$ $(i \in \{1,2\})$, discriminator D_ω. The parameters of each block are denoted by its subscript.
Initialize Replay Buffer $\mathcal{R}$ ▷ Randomly sample B transitions $(\boldsymbol{o}, a, \boldsymbol{o}', d)$
for $i = 1$ **to** M **do**
  $\{(\boldsymbol{o}, a, \boldsymbol{o}', d)\}_{k=1}^N \sim \mathcal{R}$ ▷ Sample N transitions from replay buffer $\mathcal{R}$
  $\{(\boldsymbol{o}, a, r, \boldsymbol{o}', d)\}_{k=1}^N$ with $r = -\log(D_\omega(\boldsymbol{o}, a))$ ▷ Compute rewards with discriminator D_ω
  $\{(\boldsymbol{o}_e, a_e)\}_{k=1}^N \sim \tau_E$, $x = \text{AUG}(x), x \in \{\boldsymbol{o}, \boldsymbol{o}', \boldsymbol{o}_e\}$ ▷ Sample expert data and data augmentation
  $z = f_\psi(\boldsymbol{o}), z' = f_\psi(\boldsymbol{o}')$ ▷ Extract features with the encoder
  $y(r, z', d) = r + \gamma(1-d)\left(\min_{i=1,2} Q_{\bar{\phi}_i}(z', a') - \alpha \log \pi_\theta(a'|z')\right),\ a' \sim \pi_\theta(\cdot|z')$
  $\nabla_{\phi_i, \psi} \frac{1}{N}(Q_{\phi_i}(z, a) - y(r, z', d))^2$ ▷ Update encoder f_ψ and Q-value Q_ϕ
  $Q_{\bar{\phi}_i} \leftarrow \rho Q_{\bar{\phi}_i} + (1-\rho) Q_{\phi_i},\ i \in \{1,2\}$ ▷ Update target Q-value functions
  $\nabla_\theta \frac{1}{N}\left(\min_{i=1,2} Q_{\phi_i}(z, \bar{a}) - \alpha \log \pi_\theta(\bar{a}|z)\right),\ \nabla_\alpha \frac{1}{N} \alpha(-\log \pi_\theta(\bar{a}|s) - \bar{\mathcal{H}}),\ \bar{a} \sim \pi_\theta(\cdot|z)$ ▷ Update policy π_θ and α
  $\nabla_\omega \frac{1}{N} \mathbb{E}_{(\boldsymbol{o}, a)}[\log D_\omega(\boldsymbol{o}, a)] + \mathbb{E}_{(\boldsymbol{o}_e, a_e)}[\log(1 - D_\omega(\boldsymbol{o}_e, a_e))]$ ▷ Update discriminator D_ω
  $\mathcal{R} \leftarrow \mathcal{R} \cup (\boldsymbol{o}, a, \boldsymbol{o}', d)$ ▷ Sample a new transition to replay buffer $\mathcal{R}$
end for

can accurately estimate the angles and angle velocities of robot joints, which is defined as the state of a robot [40], from images $\boldsymbol{o_t}$ for decision-making. However, it is challenging to learn what is essential for making decisions from several consecutive images. To help extract meaningful features from images, we employ data augmentation in OPIfVI.

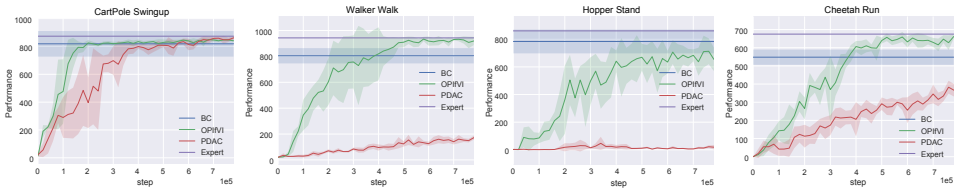
Data augmentation is used to enlarge expert data and agent data, and helps suppress overfitting as well as enhance robustness [41]. In a way, the data-efficiency in OPIfVI is also improved because we can obtain more samples by augmenting $\boldsymbol{o_t}$. Due to the property of images, modifying a series of pixels in an image would not distort the core information [41]. Hence, data augmentation helps OPIfVI to learn invariant features from images. In OPIfVI, we employ random crop to augment visual inputs, which is considered to be simple and can dramatically improve the performance [28]. Other data augmentation methods such as grayscale and color-jitter [30] could also be helpful. Data augmentation is vital for the successful imitation of OPIfVI, which is empirically studied in Section V.

C. Encoder Training Structure

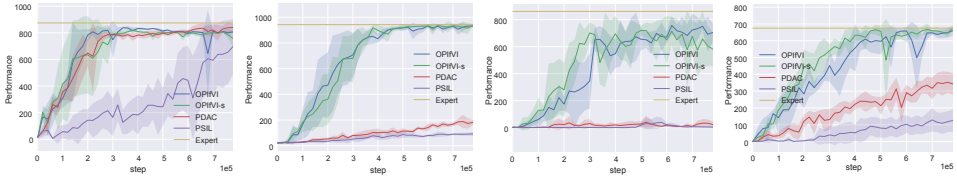
As discussed in Subsection IV-B, images contain a large portion of redundant information that is useless for agents to make decisions. It is significant for ILfVI to extract useful features from images and prevent distracting agents from selecting proper actions by the useless information. Therefore, OPIfVI employs encoders to extract features from images. An encoder perceives an image in RGB form and outputs low-dimensional features [42]. In our framework, three networks (the policy π_θ, Q-value function Q_ϕ, and discriminator D_ω) need to use an encoder to extract features. As a result, how to regulate the encoders of those networks and properly train encoders to improve their ability on extracting features is challenging.

First, we share the encoder between the actor π_θ and the critic Q_ϕ. We denote the encoder with f_ψ, which can output a latent feature from augmented adjacent images, $z_t = f_\psi(\text{AUG}(\boldsymbol{o_t}))$. After encoding, latent features are input to two separate MLPs (multilayer perceptrons) to build the policy and Q-value functions, leading to a policy $\pi(a|\boldsymbol{o_t}) = \pi_\theta(a|z_t)$ and Q-value $Q(\boldsymbol{o_t}, a) = Q_\phi(z_t, a)$. The encoder's parameters are only updated with gradients from Q-value with losses in Algorithm 1. This separate update law is inspired by SAC-AE [29], which shows that training the encoder with only gradients from Q-value network performs better and more stable than training with gradients from both the actor and critic. Second, the discriminator D_ω also needs an encoder block. The problem becomes how to design the encoder for the discriminator network.

Generally, we have three choices: 1) maintain a separate encoder for the discriminator; 2) share the encoder of Q-value functions with the discriminator and co-train the encoder with gradients from both the discriminator and Q-value functions; 3) share the encoder with the discriminator but do not backpropagate gradients from the discriminator to train the encoder. Resembling previous work GAIfO [22], we choose to hold an independent encoder for the discriminator rather than sharing the encoder of Q-value functions. We choose this structure due to the following reasons. The crucial role of a discriminator is to discriminate whether an image is sampled from the agent or the expert. Many successful GAN methods, whose discriminators possess a separate encoder, have been developed [43], [44]. Besides, our discriminator is spectrally normalized, which enforces the Lipshitz continuity of networks. The Lipshitz continuity could impair the representational capacity of neural networks, making it difficult to share parameters of encoders



(a) Visual demonstrations



(b) Visual observations

Fig. 2: Performance of OPIfVI compared to the baselines on DeepMind Control tasks. Performance is measured with episode cumulative rewards, which is averaged across 5 random seeds, and the x-axis is the number of interactions with environments.

between the discriminator and the generator. Furthermore, the discriminator D_ω is trained with losses such as BCE Loss [45]. This kind of loss could distract the encoder from extracting meaningful features for decision-making.

In OPIfVI, we share an encoder between the policy and the Q-value network, but maintain a separate encoder for the discriminator. The encoders help extract features from images for downstream applications, such as selecting actions or discriminating samples. The shared encoder between π_θ and Q_ϕ is trained with mere gradients from Q-value losses, while the encoder of the discriminator is updated with the discriminator loss. This design plays an important role in stabilizing the training of the imitator and achieves better performance compared to previous structures.

V. Experimental Results

We conduct experiments with DeepMind Control Suite [17] to demonstrate the performance of OPIfVI against other baselines. We aim to answer the following questions:

1) Can OPIfVI successfully reproduce expert policies from visual inputs and outperform other baselines regarding both data-efficiency and performance?
2) Does every modification that we adopt help improve the performance or data-efficiency of OPIfVI, and what role does it play. In particular, we investigate the effects of spectral normalization, data augmentation, and the encoder training structure.

a) Setups and Baselines: We choose four typical environments in DeepMind Control Suite [17], *i.e.*, CartPole Swingup, Walker Walk, Hopper Stand, and Cheetah Run, which are shown in Fig. 4. The action repeat for these environments is set to 4. We stack three consecutive frames together to construct $\boldsymbol{o}_t$, whose dimension is $84 \times 84 \times 9$ (channel last). These configurations are coherent with [28]. First, we use DrQ [28] to train experts in these environments and then sample data using trained experts. Visual observations $\tau_E = \{(\boldsymbol{o}, \boldsymbol{o'})\}_0^N$ can be obtained by removing actions in visual demonstrations $\tau_E = \{(\boldsymbol{o}, a)\}_0^N$. For every environment, we sample 5,000 state-action pairs or state-state pairs. The performance of expert data is 873.8 $\pm$ 1.5 (CartPole Swingup), 943.1 $\pm$ 22.4 (Walker Walk), 860.7 $\pm$ 48.6 (Hopper Stand), and 675.0$\pm$ 30.9 (Cheetah Run). Then, we can conduct IL experiments with acquired visual inputs.

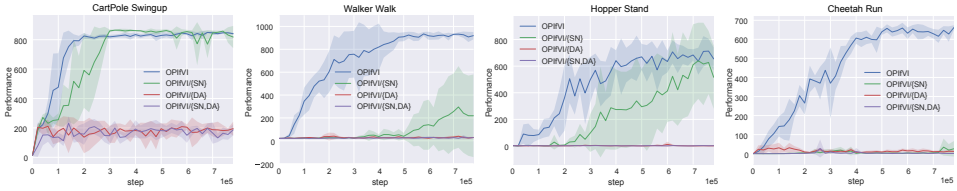
We compare OPIfVI against two spectra of IL algorithms, *i.e.*, visual demonstrations and visual observations. To achieve fair evaluations, we use the identical data augmentation technique as in [28] and neural network architectures for inchoate algorithms.

1) Baselines for visual demonstrations. We select two baselines for visual demonstrations, *i.e.*, VBC [8] and P-DAC [23]. P-DAC in the concurrent work is employed to serve as the baseline because it demonstrates state-of-the-art performance.
2) Baselines for visual observations. We choose P-SIL and P-DAC in [23] as counterparts for visual observations. P-SIL is trained with expert data $\tau_E = \{(\boldsymbol{o})\}_0^N$, which slightly differs from the LfO setting. Hence, we conduct experiments to investigate the performance of OPIfVI with both $\tau_E = \{(\boldsymbol{o}, \boldsymbol{o'})\}_0^N$ and $\tau_E = \{(\boldsymbol{o})\}_0^N$ and denote them as OPIfVI and OPIfVI-s, respectively.

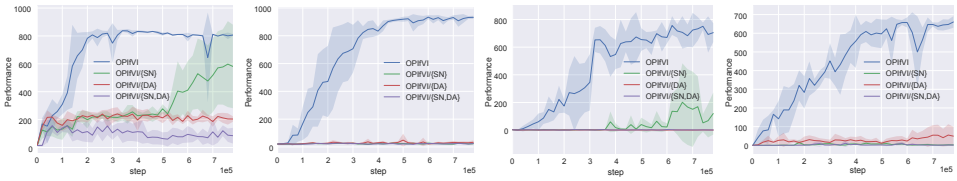
Our algorithm OPIfVI is implemented based on DrQ [28] and OpenAI Baselines [46]. The counterparts that we compare to, P-SIL and P-DAC, are the official implementation in [23]. The hyperparameters for our experiments are inspired by [28], [46] and presented in Table I.

TABLE I: Hyperparameters in experiments.

Hyperparameters	Value
Common parameters	
Activation	ReLU
Batch size	128
Optimizer	Adam
Encoder feature dim	50
MLP network size	(1024,1024)
Actor/Critic update frequency	2/1
Discriminator update frequency	1
SAC parameters	
Discount	0.99
Learning rate	1e-3
Initial temperature α	0.1
Ployak	0.01
Discriminator parameters	
Learning rate	1e-4



(a) Visual demonstrations



(b) Visual observations

Fig. 3: Ablation study of spectral normalization (SN) and data augmentation (DA) in OPIfVI. We use OPIfVI to represent the integrated framework, OPIfVI/x to denote that the framework without modification x. x could be SN, DA, or the union of these two modifications.

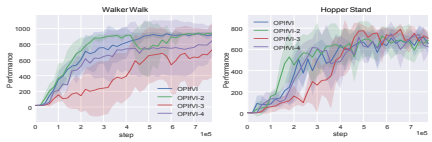
(a) (b) (c) (d)

Fig. 4: Screenshots of DeepMind Control Suite tasks. (a) CartPole Swingup. (b) Walker Walk. (c) Hopper Stand. (d) Cheetah Run.

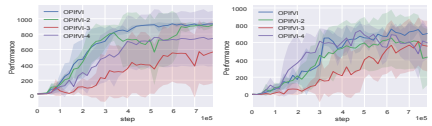
b) Results: We conduct experiments of OPIfVI with both visual demonstrations and visual observations. Besides, baselines corresponding to the two spectra are implemented and evaluated. The learning curves are shown in Figure 2. It is clear from the learning curves that: (1) OPIfVI is able to primely replicate expert behaviors and achieve expert-level performance no matter visual demonstrations and observations are provided, while the other baselines fail to perform as similar as experts with same training steps; (2) OPIfVI noticeably outperforms the baselines regarding both final performance and data-efficiency. For example, at 700 k steps in Walker Walk, OPIfVI achieves about **6.6**× and **5.5**× higher scores compared to PDAC in visual demonstrations and visual observations, respectively.

c) Ablation Studies: First, we study the impacts of spectral normalization and data augmentation in OPIfVI. Concretely, we compare the performance of OPIfVI against its versions without spectral normalization and/or data augmentation, whose results are visualized in Fig. 3. From Fig. 3, we can see that OPIfVI even could not reproduce a satisfactory policy to mimic experts in most environments without either of them. For example, in CartPole Swingup, OPIfVI/DA only achieves about a quarter of the OPIfVI's performance with visual demonstrations. In more complex environments, only OPIfVI is able to achieve expert-level performance, which means that both spectral normalization and data augmentation play an important role in OPIfVI.

Second, we conduct additional experiments to validate the encoder training structure. We test four cases: 1) the discriminator maintains a separate encoder and trains it with discriminator losses from scratch (OPIfVI, the structure we adopt); 2) similar to 1) despite that the encoder is trained with both losses from the actor and critic (OPIfVI-2); 3) the discriminator owns an encoder that is shared from the Q-value network and this encoder is trained with only the loss of Q-value functions (OPIfVI-3); 4) the discriminator, actor, and critic possess an independent encoder, respectively, and separately trains their encoders (OPIfVI-4). From the experimental results in Fig. 5, we can see that OPIfVI demonstrates excellent performance across different environments and tasks. On the contrary, the other encoder structures could be unstable and perform poorly, especially on Walker Walk and Hopper Stand tasks with visual observations.



(a) Visual demonstrations



(b) Visual observations

Fig. 5: Ablation study of encoder structure in OPIfVI.

VI. Conclusion

In this paper, we present an IL algorithm, OPIfVI, which can efficiently and effectively learn from visual inputs. OPIfVI works in an off-policy manner with stability enhanced with spectral normalization, improving learning efficiency. In addition, to deal with visual inputs, we adopt data augmentation and design a specific architecture to train the encoder. These two techniques help agents to better identify meaningful features in visual inputs, thus empowering agents to take correct actions. Compared to previous baselines, OPIfVI outperforms them regarding both data-efficiency and final performance. For future work, we would extend OPIfVI to real robot applications and more complex tasks.

REFERENCES

[1] J. Ho and S. Ermon, "Generative adversarial imitation learning," in *Advances in neural information processing systems*, 2016, pp. 4565–4573.

[2] Y. Yuan and K. Kitani, "3d ego-pose estimation via imitation learning," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 735–750.

[3] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[4] B. Fang, S. Jia, D. Guo, M. Xu, S. Wen, and F. Sun, "Survey of imitation learning for robotic manipulation," *International Journal of Intelligent Robotics and Applications*, vol. 3, no. 4, pp. 362–369, 2019.

[5] W. Goo and S. Niekum, "One-shot learning of multi-step tasks from observation via activity localization in auxiliary video," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7755–7761.

[6] N. Wake, R. Arakawa, I. Yanokura, T. Kiyokawa, K. Sasabuchi, J. Takamatsu, and K. Ikeuchi, "Learning-from-observation framework: One-shot robot teaching for grasp-manipulation-release household operations," *arXiv preprint arXiv:2008.01513*, 2020.

[7] R. Douglas Greer, J. Dudek-Singer, and G. Gautreaux, "Observational learning," *International journal of psychology*, vol. 41, no. 6, pp. 486–499, 2006.

[8] S. Young, D. Gandhi, S. Tulsiani, A. Gupta, P. Abbeel, and L. Pinto, "Visual imitation made easy," *arXiv preprint arXiv:2008.04899*, 2020.

[9] R. Rafailov, T. Yu, A. Rajeswaran, and C. Finn, "Visual adversarial imitation learning using variational models," *arXiv preprint arXiv:2107.08829*, 2021.

[10] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," *arXiv preprint arXiv:1807.06158*, 2018.

[11] X. Zhang, Y. Li, Z. Zhang, and Z.-L. Zhang, "f-gail: Learning f-divergence for generative adversarial imitation learning," *arXiv preprint arXiv:2010.01207*, 2020.

[12] Y. Li, J. Song, and S. Ermon, "Infogail: Interpretable imitation learning from visual demonstrations," in *Advances in Neural Information Processing Systems*, 2017, pp. 3812–3822.

[13] R. Bellman, "A markovian decision process," *Journal of mathematics and mechanics*, pp. 679–684, 1957.

[14] L. P. Kaelbling, M. L. Littman, and A. R. Cassandra, "Planning and acting in partially observable stochastic domains," *Artificial intelligence*, vol. 101, no. 1-2, pp. 99–134, 1998.

[15] P. Ramachandran, N. Parmar, A. Vaswani, I. Bello, A. Levskaya, and J. Shlens, "Stand-alone self-attention in vision models," *arXiv preprint arXiv:1906.05909*, 2019.

[16] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

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

[18] M. Bain and C. Sammut, "A framework for behavioural cloning," *In Machine Intelligence*, vol. 15, pp. 103–129, 1995.

[19] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proceedings of international conference on Machine learning*, 2004.

[20] J. Fu, K. Luo, and S. Levine, "Learning robust rewards with adversarial inverse reinforcement learning," *arXiv preprint arXiv:1710.11248*, 2017.

[21] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal wasserstein imitation learning," in *International Conference on Learning Representations*, 2020.

[22] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," 2018.

[23] S. Cohen, B. Amos, M. P. Deisenroth, M. Henaff, E. Vinitsky, and D. Yarats, "Imitation learning from pixel observations for continuous control," 2022. [Online]. Available: https://openreview.net/forum?id=JLbXkHkLCG6

[24] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, "Mastering visual continuous control: Improved data-augmented reinforcement learning," *arXiv preprint arXiv:2107.09645*, 2021.

[25] A. Fawzi, H. Samulowitz, D. Turaga, and P. Frossard, "Adaptive data augmentation for image classification," in *2016 IEEE international conference on image processing (ICIP)*. Ieee, 2016, pp. 3688–3692.

[26] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 113–123.

[27] N.-T. Tran, V.-H. Tran, N.-B. Nguyen, T.-K. Nguyen, and N.-M. Cheung, "On data augmentation for gan training," *IEEE Transactions on Image Processing*, vol. 30, pp. 1882–1897, 2021.

[28] D. Yarats, I. Kostrikov, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," in *International Conference on Learning Representations*, 2020.

[29] D. Yarats, A. Zhang, I. Kostrikov, B. Amos, J. Pineau, and R. Fergus, "Improving sample efficiency in model-free reinforcement learning from images," *arXiv preprint arXiv:1910.01741*, 2019.

[30] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[31] Y. Seo, L. Chen, J. Shin, H. Lee, P. Abbeel, and K. Lee, "State entropy maximization with random encoders for efficient exploration," *arXiv preprint arXiv:2102.09430*, 2021.

[32] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[33] C. Yang, X. Ma, W. Huang, F. Sun, H. Liu, J. Huang, and C. Gan, "Imitation learning from observations by minimizing inverse dynamics disagreement," in *Advances in Neural Information Processing Systems*, 2019, pp. 239–249.

[34] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th international conference on Machine learning*, 2018, p. 1.

[35] I. Kostrikov, O. Nachum, and J. Tompson, "Imitation learning via off-policy distribution matching," in *International Conference on Learning Representations*, 2019.

[36] H. Hoshino, K. Ota, A. Kanezaki, and R. Yokota, "Opirl: Sample efficient off-policy inverse reinforcement learning via distribution matching," *arXiv preprint arXiv:2109.04307*, 2021.

[37] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[38] Z. Cheng, L. Liu, A. Liu, H. Sun, M. Fang, and D. Tao, "On the guaranteed almost equivalence between imitation learning from observation and demonstration," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2021.

[39] L. Blondé, P. Strasser, and A. Kalousis, "Lipschitzness is all you need to tame off-policy generative adversarial imitation learning," *arXiv preprint arXiv:2006.16785*, 2020.

[40] X. B. Peng, P. Abbeel, S. Levine, and M. van de Panne, "Deepmimic: Example-guided deep reinforcement learning of physics-based character skills," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–14, 2018.

[41] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of Big Data*, vol. 6, no. 1, pp. 1–48, 2019.

[42] S. Wang, Z. Ding, and Y. Fu, "Discerning feature supported encoder for image representation," *IEEE Transactions on Image Processing*, vol. 28, no. 8, pp. 3728–3738, 2019.

[43] I. Skorokhodov, S. Ignatyev, and M. Elhoseiny, "Adversarial generation of continuous images," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 753–10 764.

[44] T. Karras, M. Aittala, S. Laine, E. Härkönen, J. Hellsten, J. Lehtinen, and T. Aila, "Alias-free generative adversarial networks," in *Proc. NeurIPS*, 2021.

[45] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, and J. Peters, "An algorithmic perspective on imitation learning," *arXiv preprint arXiv:1811.06711*, 2018.

[46] P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman, S. Sidor, Y. Wu, and P. Zhokhov, "Openai baselines," https://github.com/openai/baselines, 2017.