Adversarial Imitation Learning from Video using a State Observer

Haresh Karnan[1], Faraz Torabi[2], Garrett Warnell[3] and Peter Stone[4]

***Abstract*— The imitation learning research community has recently made significant progress towards the goal of enabling artificial agents to imitate behaviors from video demonstrations alone. However, current state-of-the-art approaches developed for this problem exhibit high sample complexity due, in part, to the high-dimensional nature of video observations. Towards addressing this issue, we introduce here a new algorithm called Visual Generative Adversarial Imitation from Observation using a State Observer (VGAIfO-SO). At its core, VGAIfO-SO seeks to address sample inefficiency using a novel, self-supervised *state observer*, which provides estimates of lower-dimensional proprioceptive state representations from high-dimensional images. We show experimentally in several continuous control environments that VGAIfO-SO is more sample efficient than other IfO algorithms at learning from video-only demonstrations and can sometimes even achieve performance close to the Generative Adversarial Imitation from Observation (GAIfO) algorithm that has privileged access to the demonstrator's proprioceptive state information.**

I. Introduction

Imitation Learning (IL) [1], [2] is a framework in which an autonomous agent learns to imitate a demonstration provided by an expert agent, typically in the form of state and control signals. Specifically, in this work, we are interested in the problem of Imitation from Observation (IfO), a sub-problem of IL that does not assume access to the control signals in the expert demonstrations. Several IfO algorithms proposed in the past [3]–[11] have been shown to be successful at imitating expert's state-only demonstrations in several continuous control, robotics domains.

While these prior IfO algorithms have enjoyed some success, they typically assume that the demonstration includes proprioceptive state information (i.e, the most basic, internal state information that is available to the agent such as joint angles and velocities of a legged robot). While some algorithms, such as GAIfO, can learn from video demonstrations instead, the additional sample complexity incurred as a result can be detrimental in domains such as robotics where data collection during online learning is costly. For example, for the common benchmark task of Hopper-v2 on MuJoCo [12], we found that an IfO algorithm using video demonstrations took over three times more timesteps to learn a good policy, compared to the case where a proprioceptive demonstration was available (see Fig. 1). This problem is especially relevant in scenarios where proprioceptive information about the demonstrator might not be available at all. For example, it may be expensive or impossible to fit specialized sensors on the demonstrator agents to record the proprioceptive state information. In some cases, we may not have access to the demonstrator at all—for example, when learning to imitate skills from YouTube videos, or learning using video demonstrations collected from robots with proprietary code that one would like to mimic on the same or different hardware.

To alleviate this sample inefficiency problem, an extension to GAIfO for learning from video-only demonstrations was proposed [4] (referred to as VGAIfO hereafter) that was shown to have better sample efficiency compared to GAIfO. VGAIfO achieves improved sample efficiency when learning to imitate video-only demonstrations by leveraging the freely-available proprioceptive state information from the imitator. Although VGAIfO is better than GAIfO at imitation from video-only demonstration data, there is still a significant gap in performance compared to GAIfO that can also imitate from proprioceptive state-only expert demonstrations. Fig.1 shows this difference in performance between GAIfO (which has privileged access to proprioceptive states of the expert) and VGAIfO (which has access only to visual observations of the expert). The performance gap between these two techniques motivates us here to seek a new method to improve both sample efficiency and performance when learning to imitate from video-only demonstrations.

We hypothesize that improved sample efficiency when learning to imitate from video-only demonstrations can be achieved by ensuring that the imitation learning process itself happens over lower-dimensional quantities like proprioceptive state information, rather than high-dimensional visual observations for both the generator and the discriminator. Therefore, we propose here a novel self-supervised *state observer* function that estimates the proprioceptive states of the agent from high-dimensional observations. Based on this state observer, we propose a novel IfO algorithm called Visual Generative Adversarial Imitation from Observation using a State Observer (VGAIfO-SO). In VGAIfO-SO, the state observer is jointly learned along with the generator

*This work has taken place in the Learning Agents Research Group (LARG) at UT Austin. LARG research is supported in part by NSF (CPS-1739964, IIS-1724157, NRI-1925082), ONR (N00014-18-2243), FLI (RFP2-000), ARO (W911NF-19-2-0333), DARPA, Lockheed Martin, GM, and Bosch. Peter Stone serves as the Executive Director of Sony AI America and receives financial compensation for this work. The terms of this arrangement have been reviewed and approved by the University of Texas at Austin in accordance with its policy on objectivity in research.

[1]Haresh Karnan is with the Department of Mechanical Engineering, The University of Texas at Austin, USA haresh.miriyala@utexas.edu

[2]Faraz Torabi is with the Department of Computer Science, The University of Texas at Austin, USA faraztrb@cs.utexas.edu

[3]Garrett Warnell is with Army Research Laboratory, USA and the Department of Computer Science, The University of Texas at Austin, USA garrett.a.warnell.civ@army.mil

[4]Peter Stone is with the Department of Computer Science, The University of Texas at Austin and Sony AI, USA pstone@cs.utexas.edu

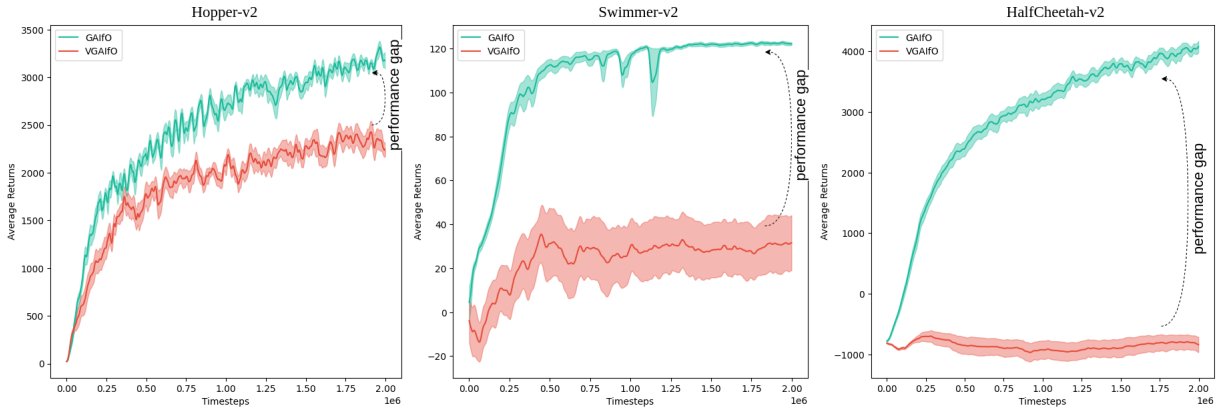

Fig. 1: Demonstrating the need for more sample efficiency and improved performance in imitation learning from video-only demonstrations. Here we show the learning curves of two representative AIL algorithms, GAIfO—that learns to imitate an expert with privileged access to proprioceptive-state-only demonstrations and VGAIfO—that learns to imitate an expert from video-only demonstrations, without access to expert's proprioceptive state information. The x-axis shows number of timesteps of interactions for the imitator agents with the environment and the y-axis shows the task reward function from OpenAI gym [13] used only for evaluation. Evident from the performance gap in their respective learning curves, we notice that GAIfO is more sample efficient than VGAIfO in imitation learning. In this work, we seek to improve both sample efficiency and imitation learning performance by proposing a new algorithm called VGAIfO-SO.

(imitator's policy) and the discriminator within an adversarial IfO framework [3]. We experimentally show that using the low-dimensional proprioceptive state predictions of the state observer as input to the discriminator network leads to improved sample efficiency and performance when learning to imitate from video-only demonstrations. Although the use of AIL for IfO is not a new concept, we show here for the first time that using a state observer to estimate proprioceptive states from video demonstrations significantly improves sample efficiency when learning to imitate from video-only observations.

This paper makes three main contributions: *1)* we show that there exists a significant gap in sample efficiency between the two IfO algorithms GAIfO (which has privileged access to proprioceptive states of the expert) and VGAIfO (which *does not* have access to proprioceptive states of the expert); *2)* we propose a novel algorithm called VGAIfO-SO, which explicitly seeks to perform imitation learning over low-dimensional quantities such as proprioceptive states of the agent, using a novel state observer network; *3)* we show on a suite of MuJoCo benchmark environments [12], [13] that VGAIfO-SO narrows the gap in performance between GAIfO and VGAIfO by improving the sample efficiency, and performs better than other baseline IfO algorithms.

II. Related Work

Learning from Demonstration. Learning from Demonstration (LfD) is a machine learning framework in which, an autonomous agent learns to imitate an expert demonstration of a behavior. The demonstrations usually contains sequences of state-action pairs generated by an expert policy when deployed in the environment. Several algorithms have been proposed to solve the imitation learning problem, and they can be broadly classified into two main categories [4]—Behavior Cloning (BC) [14]–[16] in which the imitative policy is learned directly from demonstrations using supervised learning, and Inverse Reinforcement Learning (IRL) [17]–[19] where a reward function is first learned from the demonstrations, which is then used to learn the policy using Reinforcement Learning (RL) [20]. Recently, a third category of imitation learning algorithms, called Adversarial Imitation Learning (AIL) was proposed [21]–[23], which utilizes an adversarial learning setup similar to a GAN [24] that learns both a discriminator network that classifies experiences from the expert and the imitator, and a generator network (imitator's policy) that imitates the expert demonstrations. Generative Adversarial Imitation Learning (GAIL) [21] was one of the earliest proposed AIL algorithms that was shown to be very successful at learning to imitate expert demonstrations. One shortcoming is that all of these algorithms, including GAIL, assumes access to actions performed by the expert in the demonstration, which may not be available necessarily.

Imitation Learning from Observation-only Demonstrations. Fortunately, advances in the Imitation from Observation (IfO) community have addressed the problem of performing imitation learning when the demonstration data lacks action information. In the IfO problem setting, the demonstrations consists of states-only or observations-only sequences. Behavior Cloning from Observation (BCO) [5] is an IfO algorithm that uses behavior cloning [25] to learn the imitative policy given the expert's state-only demonstrations. However, behavior cloning has been shown to suffer from the well-known compounding-error issue [15], [26] and BCO [5] is no exception [3]. Generative Adversarial Imitation from Observation (GAIfO) [3], on the other hand, overcomes this issue by incorporating reinforcement learning and explo-

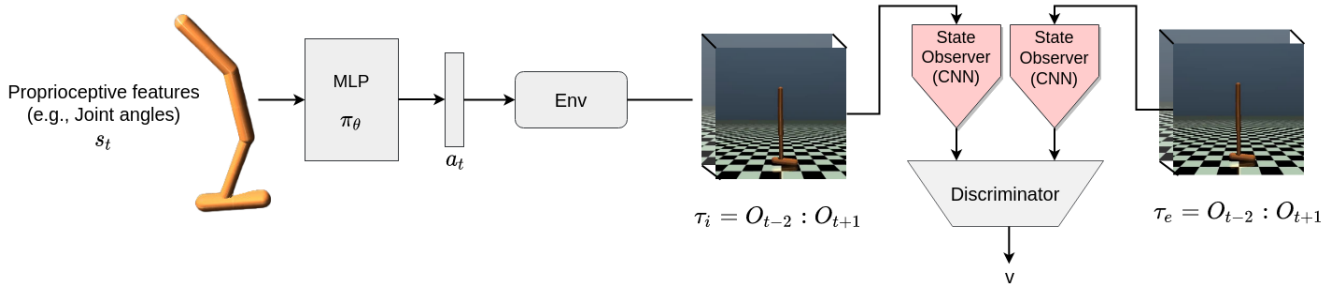

Fig. 2: A diagrammatic representation of the VGAIfO-SO algorithm proposed in this work. Different from existing AIL methods, we propose here to utilize a novel *state observer* module, which serves to simplify discriminator training. More specifically, the state observer learns to map high-dimensional visual observations to low-dimensional proprioceptive states of the agent. While traditional methods like VGAIfO directly utilize the high-dimensional observations in optimizing the discriminator's objective, in this work, we instead use the low-dimensional proprioceptive state predictions of the state observer to optimize the discriminator objective.

ration, which has led to its state-of-the-art IfO performance. An extension of GAIfO [3] called VGAIfO [4] was recently proposed that improves upon the poor sample efficiency and performance of GAIfO when learning to imitate from video-only demonstrations of the expert. VGAIfO leverages the already available proprioceptive states of the imitator as inputs to the imitative policy, which makes it more sample efficient than GAIfO with video-only demonstrations. Time Contrastive Networks (TCN) [27] is another IfO algorithm that was shown to be successful at imitation learning from video-only demonstrations. TCN achieves self-supervised imitation learning by first learning an embedding of the visual observations in the demonstrations using time-contrastive metric learning which is then used to drive a reinforcement learning procedure. Note that the version of TCN we use as a comparison point to VGAIfO-SO is also called single-view TCN. The multi-view TCN approach for IfO deals with addressing the problem of viewpoint and domain mismatch, similar to other related approaches such as TPIL [11], among other methods [9], [10].

Sample efficiency in IfO. Sample efficiency is a desirable property of algorithms that enable robots to acquire skills through imitation. Using off policy learning within the AIL framework has been proposed as a way to improve sample efficiency [22], [28], [29]. Another way to improve sample efficiency is to use model-based reinforcement learning techniques within AIL [7]. We note that using off-policy RL or model-based RL to improve sample efficiency in imitation learning is orthogonal to this work. However, the techniques introduced in this work are not restricted to on-policy IL and can be combined with off-policy and model-based RL algorithms as well to further improve performance. We view this work as a stepping stone towards the greater goal of IfO by focusing on the problem of sample inefficiency when learning to imitate from video-only observations, a relevant problem in robotics domains. While addressing domain, viewpoint, and embodiment mismatches are also important challenges to address in imitation learning, we leave incorporating them into VGAIfO-SO for future work.

Algorithm 1 VGAIfO-SO

Input: Initialize imitator policy π_θ randomly
Initialize discriminator network D_ϕ randomly
Initialize state observer network S_η randomly
Obtain video-only demonstrations $\tau_e = \{\tau_{e_1}, \tau_{e_2}, \ldots\}$, where $\tau_{e_k} = \{O_1, O_2, \ldots\}$;
for $i = 0, 1, 2, \ldots N$ **do**
  Execute π_θ and obtain paired proprioceptive-and-visual state trajectories $\tau_i = \{\tau_{i_1}, \tau_{i_2}, \ldots\}$ where $\tau_{ij} = \{(s_1, O_1), (s_2, O_2), \ldots\}$;

  Train State Observer : Update parameters η of S_η using the regression loss $-\mathbb{E}_{\tau_i}[(S_\eta(O_t) - s_t)^2]$ for N_s epochs and freeze network parameters η;

  Update parameters ϕ of D_ϕ using gradient descent to minimize $-\Big(\mathbb{E}_{\tau_i}[\log(D_\phi(\hat{s}, \hat{s'}))] + \mathbb{E}_{\tau_e}[\log(1 - D_\phi(\hat{s}, \hat{s'}))\Big)$ where $\hat{s} = S_\eta(O)$, $\hat{s'} = S_\eta(O')$;

  Update parameters θ of π_θ using *PPO* updates with reward $-[\log D_\phi(\hat{s}, \hat{s'})]$;
end

III. Method

In this section, we begin by formally introducing the imitation learning problem. We then describe the algorithm introduced in this work— Visual Generative Adversarial Imitation from Observation using a State Observer (VGAIfO-SO).

A. Preliminaries

We formulate sequential decision making as a Markov Decision Process (MDPs) [20]. An MDP is a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{T}, \gamma, \rho_0 \rangle$ where $\mathcal{S}$ and $\mathcal{A}$ are state and action spaces of the agent respectively, and $\mathcal{R}$ is a task reward function. When the agent performs an action $a_t \in \mathcal{A}$ in the environment at a state $s_t \in \mathcal{S}$, it transitions to a new state

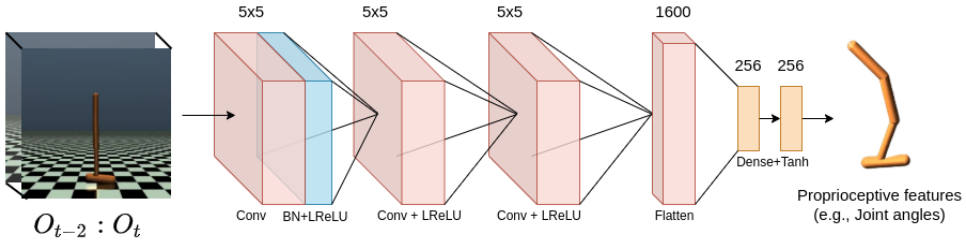

Fig. 3: Network architecture of the state observer function proposed in this work. The state observer takes as input three consecutive visual frames of observations, and produces as output a prediction $\hat{s}_t$ of the proprioceptive features of the agent.

$s_{t+1} \in \mathcal{S}$ according to the transition dynamics $\mathcal{T}(s_{t+1}|s_t, a_t)$ that is unknown to the agent. γ is the discount factor that controls the relative long term utility of the reward and ρ_0 is the initial state distribution of the environment. Since we are concerned with visual observations, we denote the observation space of the agent as $\mathcal{O}$. We are interested in learning a policy $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that the agent can use to select actions that result in behavior similar to what was demonstrated. In the imitation learning setting, agents do not receive a reward $r_t \in \mathcal{R}$ from the environment. Instead, the imitative agents have access to an expert demonstration $\mathcal{D}_E = \{(s_0, a_0), (s_1, a_1) \ldots\}$ consisting of state-action pairs. However, in this work, we focus on the problem of Imitation from Observation (IfO) consisting of video-only demonstrations $\mathcal{D}_E = \{O_0, O_1, \ldots\}$ where $O \in \mathcal{O}$.

The camera observations of the robots considered in this work are floating third person views of the robot performing the task in the environment, rendered using the MuJoCo simulator [12], [13], as shown in Fig. 2.

B. Visual Generative Adversarial Imitation from Observation using a State Observer

We adopt the general adversarial framework for IfO as initially proposed in GAIfO [3], and also followed in VGAIfO [4]. In this IfO framework, a parameterized generator function (imitative policy) π_θ represented by a Multi-Layer Perceptron (MLP) takes as its inputs proprioceptive states of the agent and acts in the environment, so as to produce state transitions that imitate the expert demonstrations. Observations of the imitator's behavior are then used to train the parameterized discriminator function D_ϕ to classify between observation transitions from the imitator and the demonstration sequence. The output of the discriminator network is then used as a reward signal to drive an update to the imitator via RL. Both the discriminator and the generator are updated using the adversarial IfO objective proposed in GAIfO [3]. We hypothesize that the poor sample efficiency of VGAIfO compared to GAIfO, as shown in Fig. 1 is due to the fact that the discriminator objective is optimized over high-dimensional images.

To solve the sample efficiency problem when learning from video-only demonstrations, we propose the novel VGAIfO-SO algorithm. In VGAIfO-SO, a state observer function represented as a convolutional neural network (CNN) (see Fig.3) is learned to circumvent optimizing the discriminator objective directly from high-dimensional visual observations. The parameterized state observer function $\mathcal{S}_\eta$ is trained using self-supervision. More specifically, $\mathcal{S}_\eta$ is trained to minimize the mean-squared error $\mathbb{E}_{\tau_i}[(S_\eta(O_t) - s_t)^2]$ between the predicted and ground truth proprioceptive states of the agent. The training data is obtained from already available experience gathered by the imitator when exploring in the environment. In our implementation, we alleviate partial observability issues by performing frame stacking with three consecutive frames of visual observations as is commonly done in the reinforcement learning community [30], although one can also use other approaches such as LSTM [31] for sequence modelling.

Note that in this work, we neither assume access to the proprioceptive states of the expert nor its actions in the demonstration. We posit that a more natural, and less restrictive form of the IL problem is to imitate from video-only demonstrations, assuming we have access to the proprioceptive state and visual observations of the imitator that we have physical access to. The VGAIfO-SO algorithm is provided in Alg. 1. Note that in Alg. 1, O' and s' are the sequentially next observation and states after O and s respectively. When training the discriminator, the state observer network is frozen and used to infer proprioceptive states from high-dimensional visual observations of both the expert and the imitator. Although one can directly use the known proprioceptive states s_t in place of predicted proprioceptive states $\hat{s}_t$ for the imitator agent, we find empirically that using predicted proprioceptive states performs better. The discriminator and generator networks are updated with the well-known GAN-like loss for IfO [3]. The imitator policy is updated using the PPO [32], [33] algorithm. All three networks are updated iteratively until the imitator successfully imitates the expert.

IV. Experiments

In this section, we perform experiments to evaluate our Visual Generative Adversarial Imitation from Observation using a State Observer (VGAIfO-SO) algorithm. Note that VGAIfO-SO is applicable in problems where the proprioceptive state representation of the agent contains all necessary information to solve the task. In manipulation domain, for example, if the task involves manipulating objects, the proprioceptive state should contain sensed poses of those objects, along with the proprioceptive state information of the

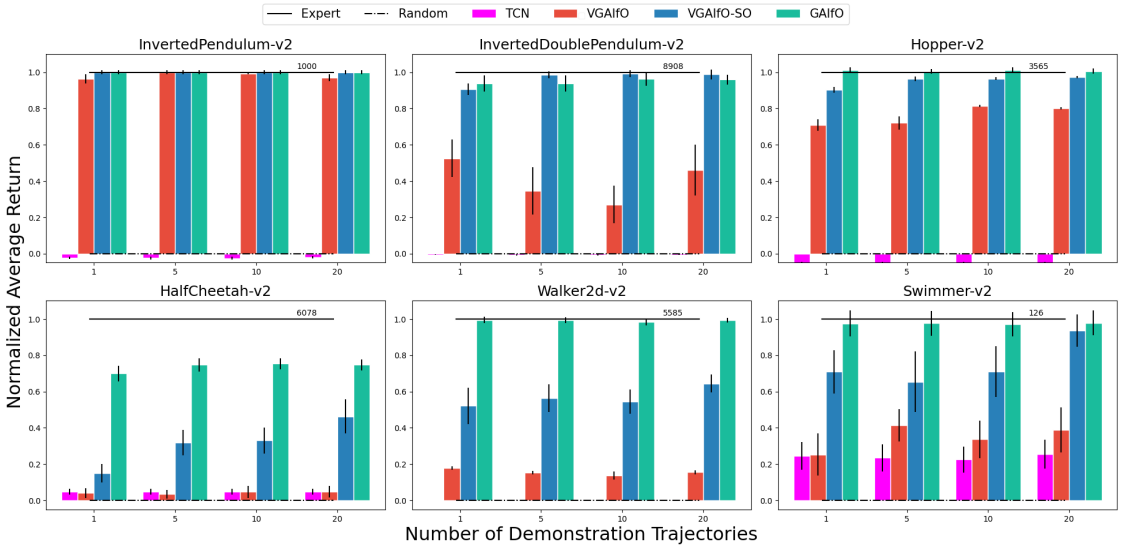

Fig. 4: Performance of different IfO algorithms with varied numbers of expert demonstration trajectories, trained for two million timesteps of interaction with the environment. We see that GAIfO [3] performs the best due to its privileged access to expert's proprioceptive state information. VGAIfO-SO, the algorithm introduced in this work, outperforms VGAIfO [4] and TCN [27] on all six environments and achieves performance similar to GAIfO in *InvertedPendulum*, *InvertedDoublePendulum* and *Hopper*, without access to the expert's true proprioceptive states.

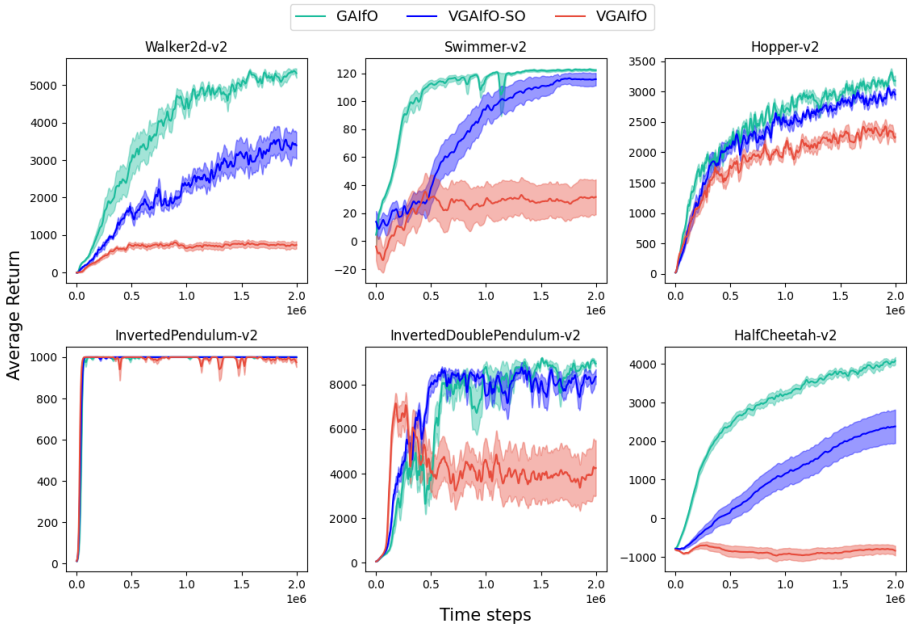

Fig. 5: Learning curves of the three adversarial IfO algorithms GAIfO, VGAIfO and VGAIfO-SO (ours). The learning curves here show that VGAIfO-SO has better sample efficiency than VGAIfO in imitation learning from video-only demonstrations and achieves performance close to the state-of-the-art GAIfO algorithm that has privileged access to expert's proprioceptive states. The x-axis shows timesteps of interactions of the agents with the environment and the y-axis shows the task reward from OpenAI gym [13] (used only for evaluation).

manipulator. However, in this work, we restrict our analysis to the locomotion domain that we are primarily interested in. We hypothesize that, compared to baseline approaches, using the predictions of the state observer in optimizing the discriminator's objective improves the imitation learning performance in terms of both sample efficiency and performance under two million environment interaction timesteps.

We test our hypothesis by evaluating VGAIfO-SO on a suite of continuous control environments in MuJoCo [12], [13], that were also used to evaluate other related IL algorithms [3], [4], [21] in the past. We compare VGAIfO-SO with two representative AIL algorithms GAIfO (with privileged access

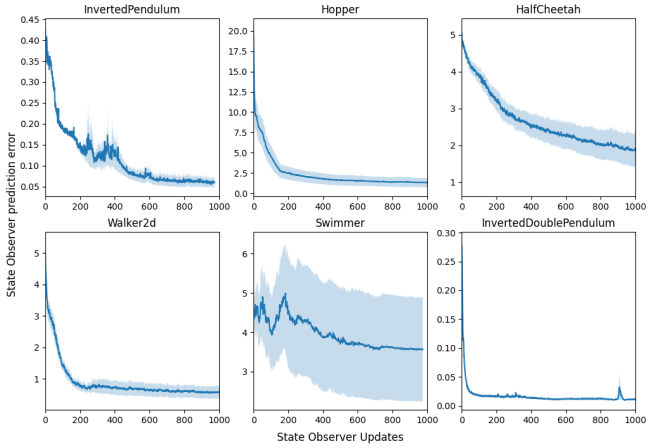

Fig. 6: **State Observer Analysis** Average L^2 norm of proprioceptive state prediction error of the state observer on the expert's demonstration sequence, per update iteration. We see that in all environments, the error has a downward trend which shows that as the learning progresses, the state observer tends towards predicting the true proprioceptive states of the demonstrator (unseen data) as well as the imitator (training data).

to expert's proprioceptive states) and VGAIfO. Note that GAIfO is expected to perform better than VGAIfO-SO since it learns to imitate from low-dimensional proprioceptive state-only demonstrations of the expert. We also compare VGAIfO-SO against single-view TCN [27], an algorithm for self-supervised imitation learning from video-only demonstrations. In our implementation of TCN, instead of using PILQR [34], we use PPO [32], [33] to learn the imitative policy. Note that TCN also uses the proprioceptive states of the imitator as input to the imitative policy, similar to GAIfO, VGAIfO, and VGAIfO-SO, making it a fair baseline for comparison. We additionally perform analysis experiments on VGAIfO-SO to qualitatively evaluate the proprioceptive state predictions from the state observer. In all experiments reported, the results are averaged across ten different random initial seeds.

Figures 4 and 5 depict our experiments comparing VGAIfO-SO with GAIfO, VGAIfO and TCN on the six continuous control environments in MuJoCo [12], [13]. We see that VGAIfO-SO performs significantly better than VGAIfO and TCN, and achieves performance close to GAIfO and the expert policy in *InvertedPendulum*, *InvertedDoublePendulum* and *Hopper*. Fig. 4 shows that TCN does not achieve performance anywhere near the expert. A similar observation was made earlier by Torabi et al. [4], perhaps due to the tasks considered here being cyclical in nature and not well suited to the time-dependent embeddings learned by TCN. The learning curves of the three AIL algorithms is shown in Fig. 5, validating the superior sample efficiency of VGAIfO-SO compared to VGAIfO.

On VGAIfO-SO, we performed a coarse hyperparameter search on number of training epochs on the three networks per adversarial training iteration and found that performing a single epoch per iteration on the networks worked best, except environments *HalfCheetah, Walker2d and Swimmer* that required 20 state observer epochs per iteration. We performed similar hyperparameter searches across other baseline algorithms GAIfO, VGAIfO and TCN and report results for the best set of hyperparameters.

State Observer Analysis. While the state observer is trained exclusively using experience from the imitator, we use it to predict proprioceptive states for both the imitator and the demonstrator. Therefore, it is important to ensure that it behaves as expected on the demonstration observations as well. To determine if the state observer is actually predicting the correct proprioceptive state, we study the prediction errors of the state observer in estimating the expert's proprioceptive states averaged across the demonstration sequence. Fig. 6 shows the average L^2 norm of the prediction error after every state observer update epoch. While the prediction errors are initially high, these errors decrease with more and more training. Therefore, we conclude that the state observer is also learning to predict the true proprioceptive state of the expert agent as well.

V. Discussion

Summary. In this work, we introduced Visual Generative Adversarial Imitation from Observation using a State Observer (VGAIfO-SO), an IfO algorithm that learns to imitate an expert policy from video-only demonstrations. We showed that, by regressing an intermediate state representation such as the proprioceptive states of the agent from visual observations using a *state observer network*, the imitation learning performance can be improved significantly. We compared our approach against similar baselines and showed that VGAIfO-SO learns to imitate the expert with better sample efficiency, and also achieves better overall imitation performance.

Limitations and Future Work. While we have shown that VGAIfO-SO performs as well as GAIfO in some domains, it does not in all experimental domains we studied (e.g., in some relatively harder tasks such as *HalfCheetah* and *Walker2d*, there still exists a gap in performance between GAIfO and VGAIfO-SO). This performance gap can potentially be narrowed further by extending VGAIfO-SO with recent advances in the reinforcement learning community involving image augmentations towards sample efficient learning [35], [36]. Another approach is to jointly train the state observer and the discriminator with their combined losses, end-to-end, to further improve performance. In this work, we assume no viewpoint or domain mismatch between the demonstrator and the imitator since the focus of this work is in improving sample efficiency when learning to imitate from video-only demonstrations. However, viewpoint and domain mismatch are also important challenges to overcome when performing imitation learning from video-only demonstrations in the real-world. VGAIfO-SO can be potentially combined with algorithms such as TPIL [11] to overcome this limitation.

REFERENCES

[1] S. Schaal, "Learning from demonstration," in *Proceedings of the 9th International Conference on Neural Information Processing Systems*, ser. NIPS'96. Cambridge, MA, USA: MIT Press, 1996, p. 1040–1046.

[2] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and Autonomous Systems*, vol. 57, no. 5, pp. 469–483, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0921889008001772

[3] F. Torabi, G. Warnell, and P. Stone, "Generative adversarial imitation from observation," 2019.

[4] T. Faraz, W. Garrett, and S. Peter, "Imitation learning from video by leveraging proprioception," *CoRR*, vol. abs/1905.09335, 2019. [Online]. Available: http://arxiv.org/abs/1905.09335

[5] ——, "Behavioral cloning from observation," *CoRR*, vol. abs/1805.01954, 2018. [Online]. Available: http://arxiv.org/abs/1805.01954

[6] F. Torabi, G. Warnell, and P. Stone, "Recent advances in imitation learning from observation," *CoRR*, vol. abs/1905.13566, 2019. [Online]. Available: http://arxiv.org/abs/1905.13566

[7] T. Faraz, W. Garrett, and S. Peter, "Dealio: Data-efficient adversarial learning for imitation from observation," 2021.

[8] Y. Liu, A. Gupta, P. Abbeel, and S. Levine, "Imitation from observation: Learning to imitate behaviors from raw video via context translation," *CoRR*, vol. abs/1707.03374, 2017. [Online]. Available: http://arxiv.org/abs/1707.03374

[9] A. D. Edwards, H. Sahni, Y. Schroecker, and C. L. I. Jr., "Imitating latent policies from observation," *CoRR*, vol. abs/1805.07914, 2018. [Online]. Available: http://arxiv.org/abs/1805.07914

[10] K. Schmeckpeper, O. Rybkin, K. Daniilidis, S. Levine, and C. Finn, "Reinforcement learning with videos: Combining offline observations with interaction," *CoRR*, vol. abs/2011.06507, 2020. [Online]. Available: https://arxiv.org/abs/2011.06507

[11] B. C. Stadie, P. Abbeel, and I. Sutskever, "Third-person imitation learning," *CoRR*, vol. abs/1703.01703, 2017. [Online]. Available: http://arxiv.org/abs/1703.01703

[12] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2012, pp. 5026–5033.

[13] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," 2016.

[14] M. Bain and C. Sammut, "A framework for behavioural cloning," in *Machine Intelligence 15*, 1995.

[15] S. Ross, G. J. Gordon, and J. A. Bagnell, "A reduction of imitation learning and structured prediction to no-regret online learning," 2011.

[16] S. Daftry, J. A. Bagnell, and M. Hebert, "Learning transferable policies for monocular reactive mav control," 2016.

[17] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video," 2018.

[18] N. Ratliff, D. Bradley, J. A. Bagnell, and J. Chestnutt, "Boosting structured prediction for imitation learning," in *Proceedings of the 19th International Conference on Neural Information Processing Systems*, ser. NIPS'06. Cambridge, MA, USA: MIT Press, 2006, p. 1153–1160.

[19] C. L. Baker, R. Saxe, and J. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, pp. 329–349, 2009.

[20] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[21] J. Ho and S. Ermon, "Generative adversarial imitation learning," *CoRR*, vol. abs/1606.03476, 2016. [Online]. Available: http://arxiv.org/abs/1606.03476

[22] I. Kostrikov, O. Nachum, and J. Tompson, "Imitation learning via off-policy distribution matching," 2019.

[23] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin, "Primal wasserstein imitation learning," *CoRR*, vol. abs/2006.04678, 2020. [Online]. Available: https://arxiv.org/abs/2006.04678

[24] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," 2014.

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

[26] S. Ross and D. Bagnell, "Efficient reductions for imitation learning," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, Y. W. Teh and M. Titterington, Eds., vol. 9. Chia Laguna Resort, Sardinia, Italy: PMLR, 13–15 May 2010, pp. 661–668. [Online]. Available: http://proceedings.mlr.press/v9/ross10a.html

[27] P. Sermanet, C. Lynch, Y. Chebotar, J. Hsu, E. Jang, S. Schaal, and S. Levine, "Time-contrastive networks: Self-supervised learning from video," 2018.

[28] Z. Zhu, K. Lin, B. Dai, and J. Zhou, "Off-policy imitation learning from observations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 402–12 413.

[29] E. Hudson, G. Warnell, and P. Stone, "Rail: A modular framework for reinforcement-learning-based adversarial imitation learning," 2021.

[30] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015.

[31] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, 12 1997.

[32] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017. [Online]. Available: http://arxiv.org/abs/1707.06347

[33] I. Kostrikov, "Pytorch implementations of reinforcement learning

algorithms," https://github.com/ikostrikov/pytorch-a2c-ppo-acktr-gail, 2018.

[34] Y. Chebotar, K. Hausman, M. Zhang, G. S. Sukhatme, S. Schaal, and S. Levine, "Combining model-based and model-free updates for trajectory-centric reinforcement learning," *CoRR*, vol. abs/1703.03078, 2017. [Online]. Available: http://arxiv.org/abs/1703.03078

[35] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas, "Reinforcement learning with augmented data," *CoRR*, vol. abs/2004.14990, 2020. [Online]. Available: https://arxiv.org/abs/2004.14990

[36] I. Kostrikov, D. Yarats, and R. Fergus, "Image augmentation is all you need: Regularizing deep reinforcement learning from pixels," *CoRR*, vol. abs/2004.13649, 2020. [Online]. Available: https://arxiv.org/abs/2004.13649