DISCRIMINATIVE PARTICLE FILTER REINFORCEMENT LEARNING FOR COMPLEX PARTIAL OBSERVATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

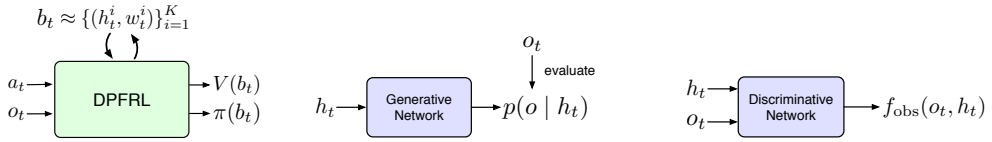
Deep reinforcement learning has succeeded in sophisticated games such as Atari, Go, etc. Real-world decision making, however, often requires reasoning with partial information extracted from complex visual observations. This paper presents *Discriminative Particle Filter Reinforcement Learning* (DPFRL), a new reinforcement learning framework for partial and complex observations. DPFRL encodes a differentiable particle filter with learned transition and observation models in a neural network, which allows for reasoning with partial observations over multiple time steps. While a standard particle filter relies on a generative observation model, DPFRL learns a discriminatively parameterized model that is training directly for decision making. We show that the discriminative parameterization results in significantly improved performance, especially for tasks with complex visual observations, because it circumvents the difficulty of modeling observations explicitly. In most cases DPFRL outperforms state-of-the-art POMDP RL models in Flickering Atari Games, an existing POMDP RL benchmark, and in *Natural* Flickering Atari Games, a new, more challenging POMDP RL benchmark that we introduce. We further show that DPFRL performs well for visual navigation with real-world data.

1 INTRODUCTION

Deep Reinforcement Learning (DRL) has attracted significant interest with applications ranging from game playing (Mnih et al., 2013; Silver et al., 2017) to robot control and visual navigation (Levine et al., 2016; Kahn et al., 2018; Savva et al., 2019). However, more natural or real-world environments pose significant challenges for current DRL methods (Arulkumaran et al., 2017), in part because they require (1) reasoning in a partially observable environment (2) reasoning with complex observations, e.g. visually rich images. For example, a robot navigating in a new environment has to (1) localize and plan a path having only partial information of the environment (2) extract the traversable space from image pixels, where the relevant geometric features are tightly coupled with irrelevant visual features, such as wall textures and lighting.

Decision making under partial observability can be formulated as a *partially observable Markov decision process* (POMDP). Solving POMDPs requires tracking the posterior distribution of the states, called the *belief*. Most POMDP RL methods track the belief, represented as a vector, using a recurrent neural network (RNN) (Hausknecht & Stone, 2015; Zhu et al., 2018). RNNs are model-free generic function approximators, and without appropriate structural priors they may need large amounts of data to learn tracking a complex belief.

Model-based DRL methods aim to reduce the sample complexity by learning and environment model simultaneously with the policy. In particular, to deal with partial observability, Igl et al. (2018) recently proposed DVRL that learns a generative observation model incorporated into the policy through a Bayes filter. Because the Bayes filter tracks the belief explicitly, DVRL performs much better than generic RNNs under partial observability. However, a Bayes filter normally assumes a generative observation model, that defines the probability $p(o \mid h_t)$ of receiving an observation $o = o_t$ given the history h_t of past observations and actions (Fig. 1b). Learning this model can be very challenging. When o_t is an image, $p(o \mid h_t)$ is a distribution over all possible images, e.g., parameterized by independent pixel-wise Gaussians with learned mean and variance. This means, e.g., to navigate in a previously unseen environment, we need to learn the distribution of all possible environments with their visual appearance, lighting condition, etc. — possibly a much harder task than learning to extract features relevant to navigation, e.g. the traversable space.



(a) DPFRL Overview (b) Generative Observation Model (c) Discriminative Observation Model

Figure 1: (a) DPFRL tracks a learned latent belief with a differentiable particle filter and learns a policy conditioned on the particle belief. (b) Generative models learn distribution $p(o \mid h_t)$ over the entire observation space o and evaluate an observation sample o_t to get a likelihood estimate. (c) The discriminatively parameterized observation model of DPFRL predicts observation likelihood values directly, using a neural network $f_{\text{obs}}(o_t, h_t)$, where o_t and h_t are inputs.

We introduce the *Discriminative Particle Filter Reinforcement Learning* (DPFRL), a POMDP RL method that learns to explicitly track a latent belief, while circumventing the difficulty of generative observation modelling, and learns to make decisions based on features of the latent belief (Fig. 1a). DPFRL approximates the belief by a set of weighted learnable latent particles $\{(h_t^i, w_t^i)\}_{i=1}^K$, and it tracks the particle belief by a non-parametric Bayes filter algorithm, a *particle filter*, encoded as a differentiable computational graph in the neural network architecture. Transition and observation models for the particle filter are neural networks learned jointly end-to-end, optimized for the overall policy. Importantly, we use a *discriminatively parameterized* observation model, $f_{\text{obs}}(o_t, h_t)$, a neural network that takes in o_t and h_t and outputs a single value, a direct estimate of the log-likelihood as shown in Fig. 1c. The discriminative parameterization avoids having to explicitly model complex observations with a generative distribution. The intuition is similar to that of, e.g., energy-based models (LeCun et al., 2006) and contrastive predictive coding (Oord et al., 2018), but here the learning signal comes directly from the RL objective, backpropagating through the differentiable particle filter, thus $f_{\text{obs}}(o_t, h_t)$ only needs to model the observation features relevant to decision making. In addition, to summarize the particle belief, we introduce novel learnable features based on *Moment-Generating Functions* (MGFs) (Bulmer, 1979). MGF features are computationally efficient and permutation invariant, and they can be directly optimized to provide useful higher-order moment information for learning the policy. MGF features could be also used as learned features of any empirical distribution in application beyond RL.

We evaluate DPFRL on a range of POMDP RL domains: a continuous control task from Igl et al. (2018), Flickering Atari Games (Hausknecht & Stone, 2015), Natural Flickering Atari Games, a new domain with more complex observations that we introduce, and the Habitat visual navigation domain using real-world data (Savva et al., 2019). DPFRL outperforms state-of-the-art POMDP RL methods in most cases. Results show that the particle filter structure is effective for handling partial observations, and the discriminative parameterization allows for complex observations.

We summarize our contributions as follows. 1) We introduce a differentiable particle filter based method with a discriminatively parameterized observation model for RL with partial and complex observations. 2) We introduce effective MGF features for empirical distributions, such as particles of a particle filter. 3) We introduce a new RL benchmark, Natural Flickering Atari Games, that introduces both partial observability and complex visual observations to the popular Atari domain. We will open source the code to enable future work.

2 Related Work

Real-world decision-making problems are often formulated as POMDPs given the partial observations. POMDPs are notoriously hard to solve; in the worst case, they are computationally intractable (Papadimitriou & Tsitsiklis, 1987). Approximate POMDP solvers have made dramatic progress on solving large-scale POMDPs (Kurniawati et al., 2008). Particle filters have been widely adopted as belief tracker for POMDP solvers (Silver & Veness, 2010; Somani et al., 2013) with the flexibility to model complex and multi-modal distributions, compared to Gaussian and Kalman filters. However, a predefined model and state representations are required for these methods (see e.g. Bai et al. (2015)).

Given the advance in generative neural network models, various neural models (Chung et al., 2015; Maddison et al., 2017; Le et al., 2018; Naesseth et al., 2018) have been proposed for belief tracking. DVRL (Igl et al., 2018) uses Variational Sequential Monte-Carlo method (Naesseth et al., 2018), which is similar to the particle filters that we use, for belief tracking in reinforcement learning.

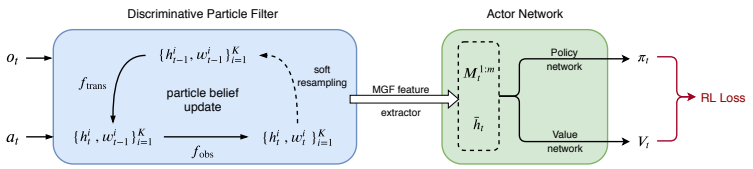


Figure 2: **DPFRL Network.** In DPFRL, latent particles $\{(h^i_t, w^i_t)\}^K_{i=1}$ are maintained by a differentiable discriminative particle filter algorithm, which includes transition function f_{trans}, discriminative observation function f_{obs}, and a differentiable soft-resampling. The policy and value function is conditioned on the belief, which is summarized by the mean particle $\bar{h}_t$, and m moment generating function features, $M^{1:m}_t$.

This gives better belief tracking capabilities, but as we demonstrate in our experiments, generative modeling is not robust in complex observation spaces with high-dimensional irrelevant features.

Learning a robust latent representation and avoiding reconstructing observations are of great interest for RL (Oord et al., 2018; Hung et al., 2018; Gregor et al., 2019; Gelada et al., 2019). Discriminative RNNs have also been widely used for belief approximation in partially observable domains (Bakker, 2002; Wierstra et al., 2007; Foerster et al., 2016). The latent representation is directly optimized for the policy $p(a|h_t)$ that skips observation modeling. In particular, Hausknecht & Stone (2015); Zhu et al. (2018) tackle the partially observable Flickering Atari Games by extending DQN (Mnih et al., 2013) with an LSTM memory. Our experiments demonstrate that the additional structure provided by using a particle filter to track beliefs can give improved performance in reinforcement learning.

Embedding algorithms into neural networks to allow end-to-end discriminative training have gained attention recently. For belief tracking, the idea has been used in the differentiable histogram filter (Jonschkowski & Brock, 2016), Kalman filter (Haarnoja et al., 2016) and particle filter (Karkus et al., 2018). Karkus et al. (2017) combined a learnable histogram filter with the Value Iteration Network (Tamar et al., 2016) and introduced the learnable POMDP planner, QMDP-net. However, these methods require a predefined state representation and are limited to relatively small state spaces. Ma et al. (2019) integrated the particle filter algorithm with standard RNNs, e.g., LSTM, and introduced a discriminative PF-RNNs for sequence prediction. We build on the work in (Ma et al., 2019) demonstrating its advantages for reinforcement learning with complex partial observations and introducing MGF features to the method for improved decision making from particle beliefs. The discriminative reinforcement framework and algorithms proposed in this paper can be applied to all discriminative particle filters, including most of the methods mentioned here.

3 Discriminative Particle Filter Reinforcement Learning

We introduce DPFRL, a framework for reinforcement learning under partial and complex observations. The DPFRL architecture is shown in Fig. 2. It has two main components, a discriminative particle filter that tracks a latent belief b_t, and an actor network that learns a policy $p(a \mid b_t)$ conditioned on the belief b_t.

3.1 Discriminative Particle Filter for Tracking the Latent Belief

State Representation. In DPFRL, we use a fully differentiable particle filter algorithm to maintain a belief state b_t. More specifically, we approximate the belief state with a set of weighted latent particles $b_t \approx \{(h^i_t, w^i_t)\}^K_{i=1}$, where $\{h^i_t\}^K_{i=1}$ are K latent states learned by policy-oriented training, and $\{w^i_t\}^K_{i=1}$ represents the corresponding weights. Each latent state h^i_t stands for a hypothesis in the belief; the set of latent particles approximates the statistics of the belief.

Belief Update. We update the latent particles according to particle filter algorithm

$$h^i_t \sim f_{\text{trans}}(h^i_{t-1}, a_t, u_t(o_t)) = p(h \mid h^i_{t-1}, a_t, u_t(o_t)) \tag{1}$$

$$w^i_t = \eta f_{\text{obs}}(o_t, h^i_t) w^i_{t-1}, \quad \eta = 1/\sum_{i=1}^{K} w^i_t \tag{2}$$

$$\{(h'^i_t, w'^i_t)\}^K_{i=1} = \text{Soft-Resampling}(\{(h^i_t, w^i_t)\}^K_{i=1}) \tag{3}$$

Transition update. Eqn. 1 implements the transition update step. We sample the next state h_t^i from $f_{\text{trans}}(h_{t-1}^i, a_t, u_t(o_t)) = p(h \mid h_{t-1}^i, a_t, u_t(o_t))$, where a_t is the agent action and $u_t(o_t)$ parameterized function that can learn the environment dynamics. This formulation assumes a fully controlled system where a learned latent state that captures the dynamics of both the agent and the environment. Similar formulation has been used for sequence prediction (Ma et al., 2019). In our experiments, f_{obs} is implemented by a gated function following PF-GRU (Ma et al., 2019), where features are extracted from o_t with a neural network designed according to the observation space of the tasks. Further details are in the Appendix.

Measurement update. Eqn. 2 implements the measurement update. We directly replace the observation likelihood function $p(o \mid h_t^i)$ with a discriminative function $f_{\text{obs}}(h_t^i, o_t)$. This formulation avoids modeling the whole observation space, like generative observation models. It is optimized directly for the policy $p(a \mid b_t)$, skipping modeling $p(o \mid h_t^i)$, and learns to extract only the relevant features for decision making. In our experiment, f_{obs} is a fully connected layer that takes the h_t^i and the encoded observation o_t as the inputs, and output a single value that estimates the observation likelihood. Note that more complex network architectures could be considered to further improve the capability of the f_{obs} and we leave it for future study.

Differentiable Particle Resampling. To avoid particle degeneracy, i.e., most of the particles having near-zero weight, we adopt the differentiable soft-resampling strategy (Karkus et al., 2018; Ma et al., 2019). Instead of sampling from $p_t(i) = w_t^i$, we sample particles $\{h_t'^i\}_{i=1}^K$ from a softened proposal distribution $q(i) = \alpha w_t^i + (1-\alpha)1/K$, where α is an trade-off parameter. The new weights are derived using importance sampling: $w'^i_t = \frac{w_t^i}{\alpha w_t^i + (1-\alpha)1/K}$. We can have the final particle belief distribution as $\{(h_t'^i, w_t'^i)\}_{i=1}^K = \text{Soft-Resampling}(\{(h_t^i, w_t^i)\}_{i=1}^K)$. As a result, f_{obs} can be optimized with global belief information and model shared useful features across multiple time steps.

3.2 Belief-Conditional Actor Network

Conditioning a policy directly onto a particle belief distribution is non-trivial. To feed it to the networks, we need to summarize it into a single vector.

We introduce a novel feature extraction method for empirical distributions based on *Moment-Generating Functions* (MGFs). The MGF of an n-dimensional random variable $\mathbf{X}$ is given by $M_{\mathbf{X}}(\mathbf{v}) = \mathbb{E}[e^{\mathbf{v}^{\mathrm{T}}\mathbf{X}}], \mathbf{v} \in \mathbb{R}^n$. In statistics, MGF is an alternative specification of its probability distribution (Bulmer, 1979). Since particle belief distribution b_t is an empirical distribution, the moment generating function of b_t can be denoted as $M_{b_t}(\mathbf{v}) = \sum_{i=1}^{K} w_t^i e^{\mathbf{v}^{\mathrm{T}} h_t^i}$.

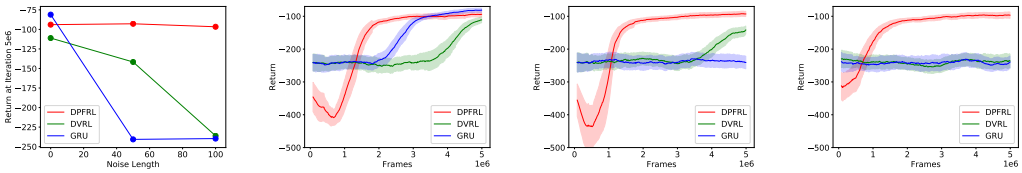
In DPFRL, we treat $\mathbf{v}$ as learnable parameters, and define m MGF features, determined by $\mathbf{v}^{1:m}$. The j-th MGF feature is given by $M_{b_t}^j(\mathbf{v}^j)$. For a clean notation, we use M_t^j in place of $M_{b_t}^j(\mathbf{v}^j)$. As a result, we summarize the belief distribution with $b_t \approx \left[\bar{h}_t, M_t^{1:m}\right]$, where $\bar{h}_t = \sum_{i=1}^{K} w_t^i h_t^i$ is the mean particle. The mean particle $\bar{h}_t$, as the first-order moment, and m additional MGF features, give a summary of the belief distribution characteristics. The number of MGF features, m, controls how much additional information we extract from the belief and we will empirically study the influence of MGF features in ablation studies.

Compared to Ma et al. (2019) that uses the mean as the belief estimate, MGF features provide additional features from the empirical distribution. Compared to DVRL (Igl et al., 2018) that treats the Monte-Carlo samples as a sequence and merge them by an RNN, MGF features are permutation-invariant, computationally efficient and easy to optimize, especially when the particle set is large.

Given $b_t \approx \left[\bar{h}_t, M_t^{1:m}\right]$, we compute the policy $p(a \mid b_t)$ with a policy network $\pi(b_t)$. In actor-critic setups, an additional value network $V(b_t)$ is introduced to assist learning. In our experiment, we evaluated on the A2C algorithm (Mnih et al., 2016).

4 Experiments

We evaluate DPFRL in a range of POMDP RL domains with increasing belief tracking and observation modeling complexity. We first use benchmark domains from the literature, Mountain Hike and 10



Effect of varying l | $l = 0$ | $l = 50$ | $l = 100$

Figure 4: Results for Mountain Hike where useful observations are concatenated with a noise vector of length l.

different Flickering Atari Games. We then introduce a new, more challenging domain, Natural Flickering Atari Games, that use a random video stream as the background. Finally we apply DPFRL to a challenging visual navigation domain with RGB-D observations rendered from real-world data.

We compare DPFRL with a GRU network, a state-of-the-art POMDP RL method, DVRL, and ablations of the DPFRL architecture. As a brief conclusion, we show that: 1) DPFRL significantly outperforms GRU in most cases because of its explicit structure for belief tracking; 2) DPFRL outperforms the state-of-the-art DVRL in most cases even with simple observations, and its benefit increases dramatically with more complex observations because of DPFRL's discriminative observation model; 3) MGF features are more effective for summarizing the latent particle belief than alternatives from the literature.

4.1 Experimental Setup

We plot the accumulated rewards and all reported results are averages over 3 different random seeds. The curves are smoothed over time and averaged over parallel environment executions.

To be comparable with the GRU and DVRL baselines we train DPFRL with the same A2C algorithm, and use a similar network architecture and hyperparameters as the original DVRL implementation. DPFRL and DVRL differ in the particle belief update structure, but they use the same latent particle size $\dim(h)$ and the same number of particles K as in the DVRL paper ($\dim(h) = 128$ and $K = 30$ for Mountain Hike, $\dim(h) = 256$ and $K = 15$ for Atari games and visual navigation). The effect of the number of particles is discussed in Sect. 4.5. We train all models for the same number of iterations using the RMSProp (Tieleman & Hinton, 2012) optimizer. Learning rates and gradient clipping values are chosen based on a search in the BeamRider Atari game independently for each model. Further details on the network structures and training setup can be found in the Appendix.

We have not performed additional search for the network architecture and other hyper-parameters, nor tried other advanced RL algorithm, such as PPO (Schulman et al., 2017), which may all improve our results.

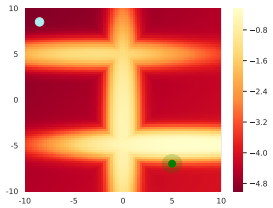
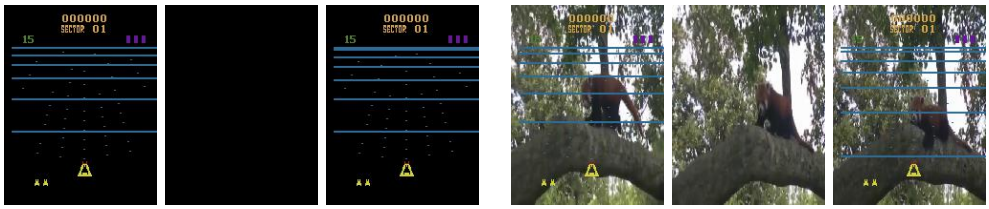


Figure 3: **Mountain Hike Task**. An agent navigates on the map from the start position (white dot) to the goal (green dot with the shaded area as the threshold). Partial observation is disturbed by a Gaussian noise and appended with a long noise vector of length l. The reward $r(x, y)$ for position (x, y) is given by the heat map.

4.2 Mountain Hike.

Mountain Hike has been introduced by Igl et al. (2018) to demonstrate the benefit of belief tracking for POMDP RL. It is a continuous control problem where an agent navigates on a fixed 20×20 map under partial observability due to observation noise. In the original task observations correspond to the agent's location with additive noise. To illustrate the effect of observation complexity in natural environments, we concatenate the original observation vector with a random noise vector. The complexity of the optimal policy remains unchanged, but the relevant information is now coupled with irrelevant observation features. More specifically, the state space and action space in Mountain Hike are defined as $\mathcal{S} = \mathcal{A} = \mathbb{R}^2$, where $s_t = [x_t, y_t]$ and $a_t = [\delta x_t, \delta y_t]$. Transitions of the agent are stochastic with an additive Gaussian noise: $s_{t+1} = s_t + a_t + \epsilon_a$, where $\epsilon_a \sim \mathcal{N}(0, 0.25)$. The observation space is $\mathcal{O} = \mathbb{R}^{2+l}$, where l is a predefined constant and $l = 0$ corresponds to the original setting. Observations are $o_t = [o_t^s, o_t^n]$, where $o_t^s = s_t + \epsilon_s, \epsilon_s \sim \mathcal{N}(0, 1)$, and



Flickering Atari Games Natural Flickering Atari Games

Figure 5: **Partially Observable Atari Games.** In Flickering Atari Games frames are randomly dropped and replaced with a blank frame. In Natural Flickering Atari Games the background is replaced with a random video stream and the Atari components of the image are randomly dropped.

Table 1: Results for Flickering Atari Games and Natural Flickering Atari Games

	Flickering Atari Games			Natural Flickering Atari Games		
	DPFRL	DVRL(Igl et al., 2018)	GRU(Igl et al., 2018)	DPFRL	DVRL	GRU
Pong	15.40±0.76	**18.17±2.67**	6.33±3.03	**15.65±1.99**	-19.78±0.06	2.62±0.93
ChopperCommand	**8,086±159.1**	6,602±449	5,150± 488.1	**1,566±67.03**	1,068±128.9	1,418±5.08
MsPacman	**3,028±545.3**	2,221±199	2,312±358	**2,106±123.9**	1,358±155.4	1,833±45.45
Centipede	**4,849±291.4**	4,240±116	4,395±224	**4,164±23.0**	3,154±335.9	3,679±116.4
BeamRider	**3,940±107.4**	1,663±183	1,801±65	**682.9±37.42**	437.3±46.31	525.6±25.25
Frostbite	293.5±5.06	297.0±7.85	254.0±0.45	**260.2±4.60**	252.4±6.48	254.3±9.20
Bowling	**33.89±0.34**	29.53 ±0.23	30.0±0.18	**29.45±0.13**	24.80±0.31	27.13±0.41
IceHockey	**-4.06±0.02**	-4.88±0.17	-7.10±0.60	-6.08±0.18	-8.79±0.12	-5.30±0.66
DDunk	-11.25±1.25	**-5.95±1.25**	-15.88±0.34	-15.36±0.96	-17.62±0.16	-14.31±0.37
Asteroids	**1,948±202.6**	1,539±73	1,545±51	1,489±15.76	1,406±132.3	1,675±571.5

$o_t^n \in \mathbb{R}^l$ is sampled from a uniform distribution $\mathcal{U}(-10, 10)$. The reward for each step is given by $r_t = r(x_t, y_t) - 0.01||a_t||$ where $r(x_t, y_t)$ is shown in Fig. 3. Episodes end after 75 steps.

We train models for different settings of the noise vector length l, from $l = 0$ to $l = 100$. Results are shown in Fig. 4. Detailed results are in the Appendix. We observe that DPFRL learns faster than the DVRL and GRU in all cases, including the original setting $l = 0$. Importantly, as the noise vector length increases, the performance of DVRL and GRU degrades, while DPFRL is unaffected. This demonstrates the ability of DPFRL to track a latent belief without having to explicitly model complex observations.

4.3 Atari Games with Partial Observability.

Atari games are one of the most popular benchmark domains for RL methods (Mnih et al., 2013). Their partially observable variants, Flickering Atari Games, have been used to benchmark POMDP RL methods (Hausknecht & Stone, 2015; Zhu et al., 2018; Igl et al., 2018). Here image observations are single frames randomly replaced by a blank frame with a probability of 0.5. In Flickering Atari Games, the model is required to effectively aggregate the history information over the long observation sequence while simultaneously making reasonable decisions with high sample efficiency. Another variant, Natural Atari Games (Zhang et al., 2018), replaces the simple black background of the frames of an Atari game with a randomly sampled video stream. This modification brings the Atari domain closer to the visually rich real-world, where relevant information is encoded in complex observations. As shown by Zhang et al. (2018), this poses a significant challenge to existing RL methods.

We propose a new RL domain, *Natural Flickering Atari Games*, that introduces both challenges of real-world environments to the Atari domain: partial observability simulated by flickering frames, and complex observations simulated by random background videos. We sample the background video from the ILSVRC dataset (Russakovsky et al., 2015). Examples for the BeamRider game are shown in Fig. 5. Details are in the Appendix. We will publish our code to enable future research.

We evaluate DPFRL for both Flickering Atari Games and Natural Flickering Atari Games. We use the same set of games as Igl et al. (2018). To ensure a fair comparison, we take the GRU and DVRL results from the paper for Flickering Atari Games, use the same training iterations as in Igl et al. (2018), and we use the official DVRL open source code to train for Natural Flickering Atari Games. Results are summarized in Table 1. We highlight the best performance in bold where the difference is statistically significant ($p = 0.05$). Detailed training curves are in the Appendix.

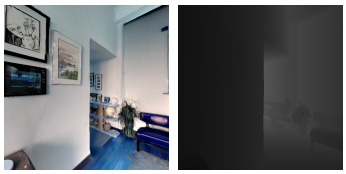


Figure 6: RGB-D Habitat Observations

Table 2: Visual Navigation Results

	SPL	Success Rate	Reward
DPFRL	**0.79**	**0.88**	**12.82±5.82**
DVRL	0.09	0.11	5.22±2.24
GRU	0.63	0.74	10.14±2.82
PPO(Savva et al., 2019)	0.70	0.80	—

We observe that DPFRL significantly outperforms GRU in almost all games, which indicates the importance of explicit belief tracking, and shows that DPFRL can learn a useful latent belief representation. Despite the simpler observations, DPFRL significantly outperforms DVRL and achieves state-of-the art results on 5 out of 10 standard Flickering Atari Games (ChopperCommand, MsPacman, BeamRider, Bowling, Asteroids), and it performs comparably in 3 other games (Centipede, Frostbite, IceHockey). The strength of DFPRL shows even more clearly in the Natural Flickering Atari Games, where it significantly outperforms DVRL on 7 out of 10 games, and performs similarly in the rest. In some games, e.g. in Pong, DPFRL performs similarly with and without videos in the background (15.65 vs. 15.40), while the DVRL performance degrades substantially (-19.78 vs. 18.17). These results show that while the architecture of DPFRL and DVRL are similar, the policy-oriented discriminative observation model of DPFRL is much more effective for handling complex observations, and the MGF features provide a more powerful summary of the particle belief for decision making.

4.4 Visual Navigation

We further evaluate DPFRL on a challenging domain, visual navigation in the Habitat Environment (Savva et al., 2019), using the real-world Gibson dataset (Xia et al., 2018). In this domain a robot needs to navigate to goals in previously unseen environments. In each time step it receives a first-person RGB-D camera image, and its distance and relative orientation to the goal. The main challenge lies in the partial and complex observations: first-person view images only provide partial information about the unknown environment; and the relevant information for navigation, traversability, is encoded in rich RGB-D observations along with many irrelevant features, e.g., the texture of the wall. We use the full Gibson dataset (572 full buildings with 1447 floors, covering a total area of 211,000 m^2) with the given training and validation splits.

We train models with the same architecture as for the Atari games, except for the observation function that accounts for the different observation format. We evaluate models in unseen environments from the validation split and compute the same set of metrics as in the paper, SPL and Success Rate, as well as average rewards. Results are shown in Table 2. Further details and learning curves are in the Appendix.

DPFRL significantly outperforms both DVRL and GRU in this challenging domain. DVRL performs especially poorly, demonstrating the difficulty of learning a generative observation model in realistic, visually rich domains. DPFRL also outperforms the PPO baseline from Savva et al. (2019).

We note that submissions to the recently organized Habitat Challenge 2019 (Savva et al., 2019), such as (Chaplot et al., 2019), have demonstrated better performance than the PPO baseline (while our results are not directly comparable because of the closed test set of the competition). However, these approaches rely on highly specialized structures, such as 2D mapping and 2D path planning, while we use the same generic network as for Atari games. Future work may further improve our results by adding task-specific structure to DPFRL or training with PPO instead of A2C.

4.5 Ablation Study

We conduct an extensive ablation study on the Natural Flickering Atari Games to understand the influence of each DPFRL component. The results are presented in Table 3.

Discriminative parameterization is more effective than generative parameterization. DPFRL-generative replaces the discriminative observation function of DPFRL with a generative observation function, where grayscale image observations are modeled by pixel-wise Gaussian distributions with learned mean and variance. Unlike DVRL, DPFRL-generative only differs from DPFRL in the

Table 3: Ablation Study Results for the Natural Flickering Atari Games

Envs	DPFRL	DPFRL-generative	DPFRL-P1	DPFRL-mean	DPFRL-GRUmerge
Pong	**15.65±1.99**	-20.21±0.02	-18.60±0.08	-5.53±14.35	13.14±4.01
ChopperCommand	**1,566±67.03**	1,027± 12.94	1,287±255.0	1,091±109.9	1,530±29.31
MsPacman	2,106±123.9	2,130±182.3	**2,233±0.47**	1,878±63.86	1,930±48.54
Centipede	**4,164±23.0**	3,194±339.4	3,557±398.1	3,599±439.8	4,093±76.4
BeamRider	**682.9±37.42**	498.1±8.38	570.7±172.3	645.5±227.4	603.8±40.25
Frostbite	**260.2±4.60**	255.9±2.78	178.2±81.5	178.4±81.70	252.1±0.48
Bowling	**29.45±0.13**	24.68±0.13	25.94±0.55	26.0±0.81	**29.50±0.33**
IceHockey	-6.08±0.18	-7.88±0.30	-6.02±1.03	-6.25±1.96	-5.85±0.30
DDunk	-15.36±0.96	-15.59±0.06	-13.28±0.96	-14.42±0.18	-14.39±0.24
Asteroids	1,406±132.3	1,415±5.33	**1,618±64.45**	1,433±40.73	1,397±11.44

parameterization of the observation function, the rest of the architecture and training loss remains the same. In most cases, the performance for DPFRL-generative degrades significantly compared to DPFRL. These results are aligned with our earlier observations and indicate that the discriminative parameterization is indeed important to extract the relevant information from complex observations without having to learn a more complex generative model.

More particles perform better. DPFRL with 1 particle performs poorly on most of the tasks (DPFRLL-P1). This indicates that a single latent state is insufficient to represent a complex latent distribution that is required for the task, and that more particles can be expected to improve performance.

MGF features are useful. We compare DPFRL using MGF features with DPFRL-mean that only uses the mean particle, and with DPFRL-GRUmerge that uses a separate RNN to summarize the belief similar to DVRL. Results show that DPFRL-mean does not work as well as the standard DPFRL, especially for tasks that may need complex belief tracking, e.g., Pong. This can be attributed to the more rich belief statistics provided by MGF features, and that they do not constrain the learned belief representation to be always meaningful when averaged. Comparing to DPFRL-GRUmerge shows that MGF features generally perform better. While an RNN may learn to extract the useful features from the latent belief, optimizing the RNN parameters is harder, because they are not permutation invariant to the set of particles and they result in a long backpropagation chain.

5 Conclusion

We have introduced DPFRL, a principled framework for POMDP RL in natural environments. DPFRL combines the strength of Bayesian filtering and policy-oriented discriminative modeling: it performs explicit belief tracking with discriminative learnable particle filters optimized directly for the RL policy. DPFRL achieved state-of-the-art results on POMDP RL benchmarks from prior work, Mountain Hike and a number of Flickering Atari Games, and it significantly outperformed alternative methods in a new, more challenging domain, Natural Flickering Atari Games, as well as for visual navigation using real-world data. We have also proposed a novel MGF feature extraction method to extract statistics from an empirical distribution. MGF feature extraction could be applied beyond RL, e.g., for general sequence prediction.

DPFRL does not perform well in some particular cases, e.g., in the game DoubleDunk. While our discriminatively parameterized observation function is less susceptible to observation noise, unlike a generative model, it does not allow for additional learning signals that improve sample efficiency, e.g., through a reconstruction loss. A possible future direction would be to combine both generative and discriminative modelling, for which the latent particle filter in DPFRL provides a promising, principled framework.

REFERENCES

Kai Arulkumaran, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. A brief survey of deep reinforcement learning. *arXiv preprint arXiv:1708.05866*, 2017.

Haoyu Bai, Shaojun Cai, Nan Ye, David Hsu, and Wee Sun Lee. Intention-aware online pomdp planning for autonomous driving in a crowd. In *2015 ieee international conference on robotics and automation (icra)*, pp. 454–460. IEEE, 2015.

Bram Bakker. Reinforcement learning with long short-term memory. In *Advances in neural information processing systems*, 2002.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym, 2016.

Michael George Bulmer. *Principles of statistics*. Courier Corporation, 1979.

Devendra Singh Chaplot, Saurabh Gupta, Abhinav Gupta, and Ruslan Salakhutdinov. Modular visual navigation using active neural mapping. *Winner CVPR 2019 Habitat Navigation Challenge*, 2019.

Junyoung Chung, Kyle Kastner, Laurent Dinh, Kratarth Goel, Aaron C Courville, and Yoshua Bengio. A recurrent latent variable model for sequential data. In *Advances in Neural Information Processing Systems*, pp. 2980–2988, 2015.

Jakob N Foerster, Yannis M Assael, Nando de Freitas, and Shimon Whiteson. Learning to communicate to solve riddles with deep distributed recurrent q-networks. *arXiv preprint arXiv:1602.02672*, 2016.

Carles Gelada, Saurabh Kumar, Jacob Buckman, Ofir Nachum, and Marc G Bellemare. Deepmdp: Learning continuous latent space models for representation learning. *arXiv preprint arXiv:1906.02736*, 2019.

Karol Gregor, Danilo Jimenez Rezende, Frederic Besse, Yan Wu, Hamza Merzic, and Aaron van den Oord. Shaping belief states with generative environment models for rl. *arXiv preprint arXiv:1906.09237*, 2019.

Tuomas Haarnoja, Anurag Ajay, Sergey Levine, and Pieter Abbeel. Backprop KF: Learning discriminative deterministic state estimators. In *Advances in Neural Information Processing Systems*, pp. 4376–4384, 2016.

Matthew Hausknecht and Peter Stone. Deep recurrent q-learning for partially observable mdps. *CoRR, abs/1507.06527*, 7(1), 2015.

Chia-Chun Hung, Timothy Lillicrap, Josh Abramson, Yan Wu, Mehdi Mirza, Federico Carnevale, Arun Ahuja, and Greg Wayne. Optimizing agent behavior over long time scales by transporting value. *arXiv preprint arXiv:1810.06721*, 2018.

Maximilian Igl, Luisa Zintgraf, Tuan Anh Le, Frank Wood, and Shimon Whiteson. Deep variational reinforcement learning for POMDPs. In *Proceedings of the International Conference on Machine Learning*, pp. 2117–2126, 2018.

Rico Jonschkowski and Oliver Brock. End-to-end learnable histogram filters. In *NeurIPS Workshop on Deep Learning for Action and Interaction*, 2016.

Gregory Kahn, Adam Villaflor, Bosen Ding, Pieter Abbeel, and Sergey Levine. Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 1–8. IEEE, 2018.

Peter Karkus, David Hsu, and Wee Sun Lee. QMDP-net: Deep learning for planning under partial observability. In *Advances in Neural Information Processing Systems*, pp. 4694–4704, 2017.

Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter networks with application to visual localization. In *Proceedings of the Conference on Robot Learning*, pp. 169–178, 2018.

Hanna Kurniawati, David Hsu, and Wee Sun Lee. Sarsop: Efficient point-based pomdp planning by approximating optimally reachable belief spaces. In *Robotics: Science and systems*, 2008.

Tuan Anh Le, Maximilian Igl, Tom Rainforth, Tom Jin, and Frank Wood. Auto-encoding sequential monte carlo. 2018.

Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. A tutorial on energy-based learning. *Predicting structured data*, 1(0), 2006.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

Xiao Ma, Peter Karkus, David Hsu, and Wee Sun Lee. Particle filter recurrent neural networks. *arXiv preprint arXiv:1905.12885*, 2019.

Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In *Advances in Neural Information Processing Systems*, pp. 6573–6583, 2017.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proceedings of the International Conference on Machine Learning*, pp. 1928–1937, 2016.

Christian Naesseth, Scott Linderman, Rajesh Ranganath, and David Blei. Variational sequential monte carlo. In *International Conference on Artificial Intelligence and Statistics*, pp. 968–977, 2018.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Christos H Papadimitriou and John N Tsitsiklis. The complexity of markov decision processes. *Mathematics of operations research*, 1987.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NeurIPS Autodiff Workshop*, 2017.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115 (3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied AI research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

David Silver and Joel Veness. Monte-carlo planning in large pomdps. In *Advances in neural information processing systems*, pp. 2164–2172, 2010.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 2017.

Adhiraj Somani, Nan Ye, David Hsu, and Wee Sun Lee. DESPOT: Online POMDP planning with regularization. In *Advances in neural information processing systems*, pp. 1772–1780, 2013.

Aviv Tamar, Yi Wu, Garrett Thomas, Sergey Levine, and Pieter Abbeel. Value iteration networks. In *Advances in Neural Information Processing Systems*, 2016.

Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *International Conference on Artificial Neural Networks*. Springer, 2007.

Fei Xia, Amir R. Zamir, Zhi-Yang He, Alexander Sax, Jitendra Malik, and Silvio Savarese. Gibson env: real-world perception for embodied agents. In *Computer Vision and Pattern Recognition (CVPR), 2018 IEEE Conference on*. IEEE, 2018.

Amy Zhang, Yuxin Wu, and Joelle Pineau. Natural environment benchmarks for reinforcement learning. *arXiv preprint arXiv:1811.06032*, 2018.

Pengfei Zhu, Xin Li, Pascal Poupart, and Guanghui Miao. On improving deep reinforcement learning for pomdps. *arXiv preprint arXiv:1804.06309*, 2018.

A EXPERIMENT DETAILS

A.1 IMPLEMENTATION DETAILS

Observation Encoders: For the observation encoders, we used the same structure with DVRL (Igl et al., 2018) for a fair comparison. For Mountain Hike, we use two fully connected layers with batch normalization and ReLU activation as the encoder. The dimension for both layers is 64. For the rest of the domains, we first down-sample the image size to 84×84, then we process images with 3 2D-convolution layers with channel number (32, 64, 32), kernel sizes (8, 4, 3) and stride (4, 2, 1), without padding. The compass and goal information are a vector of length 2; they are appended after the image encoding as the input.

Observation Decoders: Both DVRL and PFGRU-generative need observation decoders. For the Mountain Hike, we use the same structure as the encoder with a reversed order. The transposed 2D-convolutional network of the decoder has the reversed structure. The decoder is processed by an additional fully connected layer which outputs the required dimension (1568 for Atari and Habitat Navigation, both of which have 84 × 84 observations).

Stochastic Transition Function: We directly use the transition function in PF-GRU (Ma et al., 2019) for $p(h \mid h^i_{t-1}, a_t, o_t)$, which is a stochastic function with GRU gated structure. Action a_t is first encoded by a fully connected layer with batch normalization and ReLU activation. The encoding dimension for Mountain Hike is 64 and 128 for all the rest tasks. The mean and variance of the normal distribution are learned again by two additional fully connected layers; for the variance, we use Softplus as the activation function.

Observation Function: f_{obs} is implemented by a single fully connected layer without activation. In DVRL, the observation function is parameterized over the full observation space o and $p(o \mid h^i_{t-1}, a^i_t)$ is assumed as a multivariate independent Bernoulli distribution whose parameters are again determined by a neural network (Igl et al., 2018). For numerical stability, all the probabilities are stored and computed in the log space and the particle weights are always normalized after each weight update.

Soft-resampling: The soft-resampling hyperparameter α is set to be 0.9 for Mountain Hike and 0.5 for the rest of domains. Note that the soft-resampling is used only for DPFRL, not including DVRL. DVRL averages the particle weights to $1/K$ after each resampling step, which makes the resampling step cannot be trained by the RL.

Belief Summary: The GRU used in DVRL and DPFRL-GRUmerge is a single layer GRU with input dimension equals the dimension of the latent vector plus 1, which is the corresponding particle weight. The dimension of this GRU is exactly the dimension of the latent vector. For the MGF features, we use fully connected layers with feature dimension as the number of MGF features. The activation function used is the exponential function. We could potentially explore the other activation functions to test the generalized-MGF features, e.g., ReLU.

Model Learning: For RL, we use a A2C algorithm with 16 parallel environments for both Mountain Hike and Atari games; for Habitat Navigation, we only use 6 parallel environments due to the GPU memory constraints. The loss function for DPFRL and GRU-based policy is just the standard A2C loss, $L^{\text{A2C}}_t = \mathcal{L}^A_t + \lambda^V \mathcal{L}^V_t + \lambda^H \mathcal{L}^H_t$, where $\mathcal{L}^A_t$ is the policy loss, $\mathcal{L}^V_t$ is the value loss, $\mathcal{L}^H_t$ is the entropy loss for encouraging exploration, and λ^V and λ^H are two hyperparameters. For all experiments, we use $\lambda_V = 0.5$ and $\lambda^H = 0.01$. For DVRL, an additional encoding loss L^E_t is used to train the sequential VAE, which gives a loss function $L^{\text{DVRL}}_t = L^{\text{A2C}}_t + \lambda^E \mathcal{L}^E_t$. We follow the default setting provided by Igl et al. (2018) and set $\lambda^E = 0.1$. The rest of the hyperparameters, including learning rate, gradient clipping value and α in soft-resampling are tuned according to the BeamRider and directly applied to all domains due to the highly expensive experiment setups. The learning rate for all the networks are searched among the following values: $(3 \times 10^{-5}, 5 \times 10^{-5}, 1 \times 10^{-4}, 2 \times 10^{-4}, 3 \times 10^{-4})$; the gradient clipping value are searched among $\{0.5, 1.0\}$; the soft-resampling α is searched among $\{0.5, 0.9\}$. The best performing learning rates were 1^{-4} for DPFRL and GRU, and 2^{-4} for DVRL; the gradient clipping value for all models was 0.5; the soft-resampling α is set to be 0.9 for Mountain Hike and 0.5 for Atari games.

A.2 EXPERIMENT SETUPS

Natural Flickering Atari games We follow the setting of the prior works (Zhu et al., 2018; Igl et al., 2018): 1) 50% of the frames are randomly dropped 2) a frameskip of 4 is used 3) there is a 0.25 chance of repeating an action twice. In our experiments, we sample background videos from the ILSVRC dataset (Russakovsky et al., 2015). Only the videos with the length longer than 500 frames are sampled to make sure the video length is long enough to introduce variability. For each new episode, we first sample a new video from the dataset, and a random starting pointer is sampled in this video. Once the video finishes, the pointer is reset to the first frame (not the starting pointer we sampled) and continues from there.

Experiment platform: We implement all the models using PyTorch (Paszke et al., 2017) with CUDA 9.2 and CuDNN 7.1.2. Flickering Atari environments are modified based on OpenAI Gym (Brockman et al., 2016) and we directly use Habitat APIs for visual navigation. For Mountain Hike and Atari games, we run our experiments on servers with 4 NVidia GTX1080Ti GPUs on each server. For Habitat visual navigation, we run the experiment on servers with 4 NVidia RTX2080Ti GPUs on each server.

B PF-GRU NETWORK ARCHITECTURE

We implement DPFRL with gated transition and observation functions for particle filtering similar to PF-GRU (Ma et al., 2019).

In standard GRU, the memory update is implemented by a gated function:

$$h_t = (1 - z_t) \circ \tanh(n_t) + z_t \circ h_{t-1}, \qquad n_t = W_n[r_t \circ h_{t-1}, x_t] + b_n \tag{4}$$

where W_n and b_n are the corresponding weights and biases, and z_t r_t are the learned gates.

PF-GRU introduces stochastic cell update by assuming the update to the memory, n_t^i, follows a parameterized Gaussian distribution

$$n_t^i = W_n[r_t^i \circ h_{t-1}^i, x_t] + b_n + \epsilon_t^i, \quad \epsilon_t^i \sim \mathcal{N}(0, \Sigma_t^i), \quad \Sigma_t^i = W_\Sigma[h_{t-1}^i, x_t] + b_\Sigma \tag{5}$$

With $x_t = [f_{\text{enc}}^o(o_t), f_{\text{enc}}^a(a_t)]$, we implement the transition function $h_{t+1}^i \sim f_{\text{trans}}(h_t^i, o_t, a_t)$, where f_{enc}^o is the encoding network for observation and f_{enc}^a is the encoding network for the actions.

For the observation function, we directly use a fully connected layer $f_{\text{obs}}(h_t^i, o_t) = W_o \left[h_t^i, o_t\right] + b_o$, where W_o and b_o are the corresponding weights and biases.

C ADDITIONAL RESULTS

C.1 VISUAL NAVIGATION

We present the reward curve for habitat visual navigation task below. DPFRL outperforms both GRU-based policy and DVRL given the same training time. DVRL struggles on training the observation model and fails during the first half of the training time. GRU based policy learns fast; given only the model-free belief tracker, it struggles to achieve higher reward after certain point.

We only provide the reward curve here as SPL and success rate are only evaluated after the training is finished.

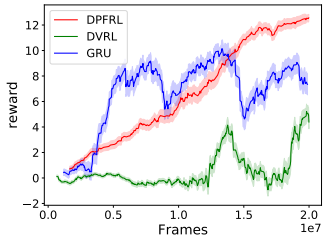


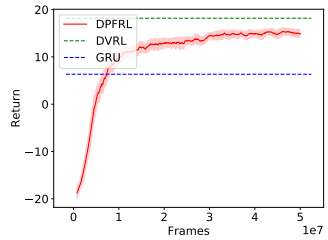
Figure 7: Habitat Visual Navigation Reward

C.2 FLICKEIRNG ATARI GAMES PLOTS

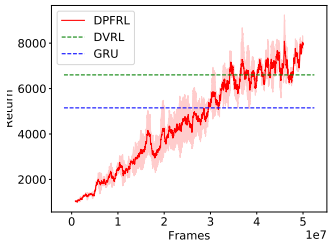
We provide the accumulated reward curves for Atari experiments in this section.

C.2.1 STANDARD FLICKERING ATARI GAMES

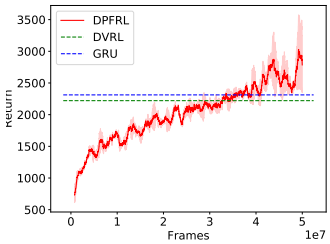
For the standard Flickering Atari Games, no validation environment is provided. We directly provide the training curves below. Results for DVRL and GRU are directly taken from Igl et al. (2018).



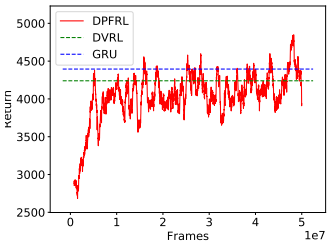
(a) Flickering Pong



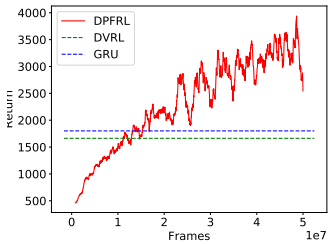
(b) Flickering ChopperCommand



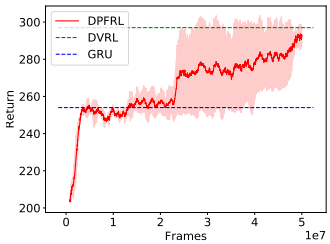
(c) Flickering MsPacman



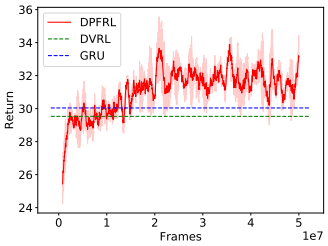
(d) Flickering Centipede



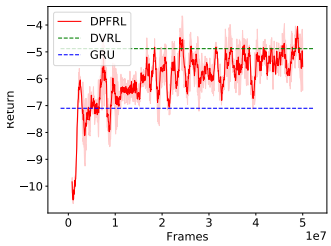
(e) Flickering BeamRider



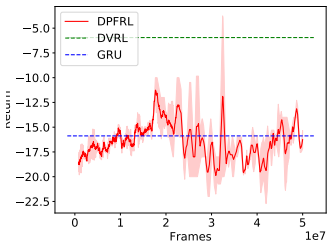
(f) Flickering Frostbite



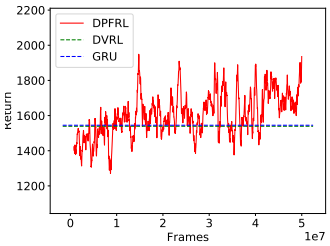
(g) Flickering Bowling



(h) Flickering IceHockey



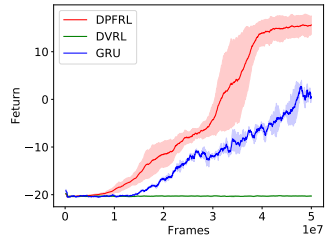
(i) Flickering DoubleDunk



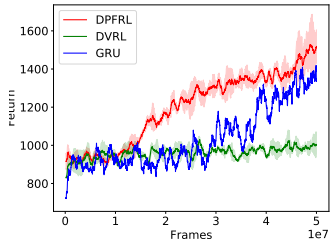
(j) Flickering Asteroids

C.2.2 NATURAL FLICKERING ATARI GAMES

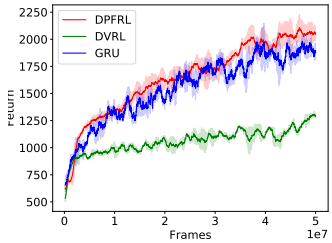
For Natural Flickering Atari Games, we have a separate validation set that contains videos different from the training set. The validation environment step once after every 100 training iterations.



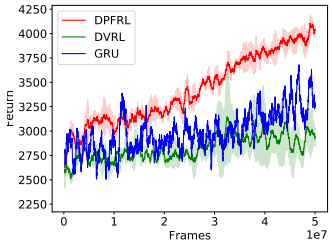
(a) Natural Flickeirng Pong



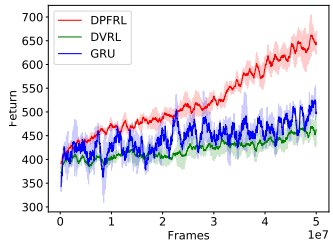
(b) Natural Flickeirng ChopperCommand



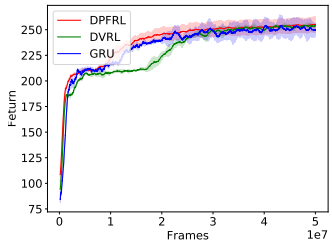
(c) Natural Flickeirng MsPacman



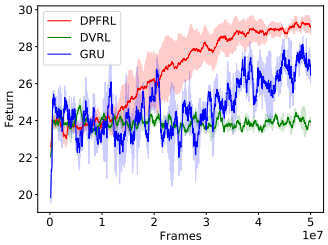
(d) Natural Flickeirng Centipede



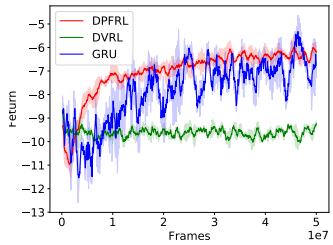
(e) Natural Flickeirng BeamRider



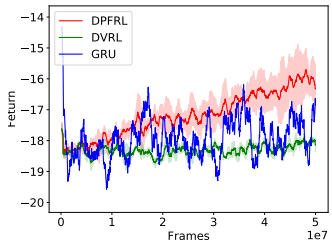
(f) Natural Flickeirng Frostbite



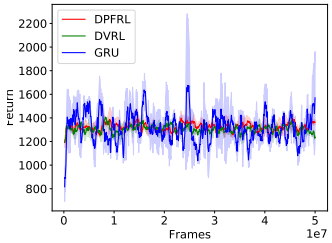
(g) Natural Flickeirng Bowling



(h) Natural Flickeirng IceHockey



(i) Natural Flickeirng DoubleDunk



(j) Natural Flickeirng Asteroids