RIDE: REWARDING IMPACT-DRIVEN EXPLORATION FOR PROCEDURALLY-GENERATED ENVIRONMENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Exploration in sparse reward environments remains one of the key challenges of model-free reinforcement learning (RL). Instead of solely relying on extrinsic rewards provided by the environment, many state-of-the-art methods use intrinsic rewards to encourage the agent to explore the environment. However, we show that existing methods fall short in procedurally-generated environments where an agent is unlikely to ever visit the same state more than once. We propose a novel type of intrinsic exploration bonus which rewards the agent for actions that change the agent's learned state representation. We evaluate our method on multiple challenging procedurally-generated tasks in MiniGrid, as well as on tasks used in prior curiosity-driven exploration work. Our experiments demonstrate that our approach is more sample efficient than existing exploration methods, particularly for procedurally-generated MiniGrid environments. Furthermore, we analyze the learned behavior as well as the intrinsic reward received by our agent. In contrast to previous approaches, our intrinsic reward does not diminish during the course of training and it rewards the agent substantially more for interacting with objects that it can control.

1 INTRODUCTION

Deep reinforcement learning (RL) is one of the most popular frameworks for developing agents that can solve a wide range of complex tasks (Mnih et al., 2016; Silver et al., 2016; 2017). RL agents learn to act in new environments through trial and error, in an attempt to maximize their cumulative reward. However, many environments of interest, particularly those closer to real-world problems, do not provide a steady stream of rewards for agents to learn from. In such settings, agents require many episodes to come across any reward, often rendering standard RL methods inapplicable.

Inspired by human learning, the use of intrinsic motivation has been proposed to encourage agents to learn about their environments even when extrinsic feedback is rarely provided (Schmidhuber, 1991b; 2010; Oudeyer et al., 2007; Oudeyer & Kaplan, 2009). This type of exploration bonus emboldens the agent to visit new states (Bellemare et al., 2016; Burda et al., 2019b; Ecoffet et al., 2019) or to improve its knowledge and forward prediction of the world dynamics (Pathak et al., 2017; Burda et al., 2019a), and can be highly effective for learning in hard exploration games such as Montezuma's Revenge (Mnih et al., 2016). However, most hard exploration environments used in previous work have either a limited state space or an easy way to measure similarity between states (Ecoffet et al., 2019) and generally use the same "*singleton*" environment for training and evaluation (Mnih et al., 2016; Burda et al., 2019a). Deep RL agents trained in this way are prone to overfitting to a specific environment and often struggle to generalize to even slightly different environments (Rajeswaran et al., 2017; Zhang et al., 2018a;b). As a first step towards addressing this problem, a number of procedurally-generated environments have been recently released, for example DeepMind Lab (Beattie et al., 2016), Sokoban (Racanière et al., 2017), Malmö (Johnson et al., 2016), CraftAssist (Jernite et al., 2019), Sonic (Nichol et al., 2018), CoinRun (Cobbe et al., 2019), Obstacle Tower (Juliani et al., 2019), or Capture the Flag (Jaderberg et al., 2019).

In this paper, we investigate exploration in procedurally-generated sparse-reward environments. Throughout the paper, we will refer to the general problem that needs to be solved as a *task* (*e.g.* find a goal inside a maze) and to the particular instantiation of this task as *environment* (*e.g.* maze layout, colors, textures, locations of the objects, environment dynamics etc.). The environment can

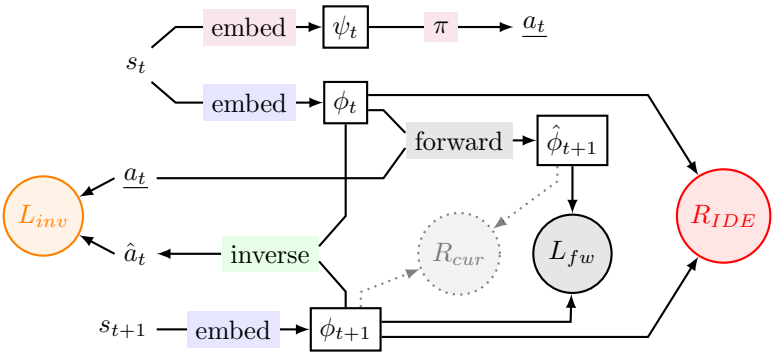

Figure 1: RIDE rewards the agent for actions that have an impact on the state representation (R_{IDE}), which is learned using both a forward (L_{fw}) and an inverse dynamics (L_{inv}) model.

be *singleton* or *procedurally-generated*. Singleton environments are those in which the agent has to solve the same task in the same environment in every episode, *i.e.*., the environment does not change between episodes. A popular example of a hard exploration environment that falls into that category is Montezuma's Revenge. In procedurally-generated environments, the agent needs to solve the same task, but in every episode the environment is initialized differently, making it unlikely to ever visit the same state twice. Thus, agents in such environments have to learn policies that generalize well across a very large state space. We demonstrate that current exploration methods fall short in such environments as they (i) make strong assumptions about the environment (deterministic or resettable to previous states) (Ecoffet et al., 2019; Aytar et al., 2018), (ii) make strong assumptions about the state space (small number of different states or easy to determine if two states are similar) (Ecoffet et al., 2019; Burda et al., 2019b; Bellemare et al., 2016; Ostrovski et al., 2017; Machado et al., 2018a), or (iii) provide intrinsic rewards that can diminish quickly during training (Pathak et al., 2017; Burda et al., 2019a)..

To overcome these limitations, we propose **Rewarding Impact-Driven Exploration** (RIDE), a novel intrinsic reward for exploration in RL that encourges the agent to take actions which result in impactful changes to its representation of the environment state (see Figure 1 for an illustration). We compare against state-of-the-art intrinsic reward methods on singleton environments with visual inputs, as well as on hard-exploration tasks in procedurally-generated grid-world environments. Our experiments show that RIDE outperforms state-of-the-art exploration methods particularly in procedurally-generated environments. Furthermore, we present a qualitative analysis demonstrating that RIDE, in contrast to prior work, does not suffer from diminishing intrinsic rewards during training and encourages the agent substantially more to interact with objects that it can control (relative to other state-action pairs).

2 Related Work

The problem of exploration in reinforcement learning has been extensively studied. Exploration methods encourage RL agents to visit novel states in various ways, for example by rewarding surprise (Schmidhuber, 1991b;a; 2010; 2006; Achiam & Sastry, 2017), information gain (Little & Sommer, 2013; Still & Precup, 2012; Houthooft et al., 2016), curiosity (Pathak et al., 2017; Burda et al., 2019b), empowerment (Klyubin et al., 2005; Rezende & Mohamed, 2015; Gregor et al., 2017), diversity (Eysenbach et al., 2019), feature control (Jaderberg et al., 2017; Dilokthanakul et al., 2019), or decision states (Goyal et al., 2019; Modhe et al., 2019). Another class of exploration methods apply the Thompson sampling heurisitc (Osband et al., 2016; Ostrovski et al., 2017; O'Donoghue et al., 2018; Tang et al., 2017). Osband et al. (2016) use a family of randomized Q-functions trained on bootstrapped data to select actions, while Fortunato et al. (2018) add noise in parameter space to encourage exploration. Here, we focus on intrinsic motivation methods, which are widely-used and have proven effective for various hard-exploration tasks (Mnih et al., 2016; Pathak et al., 2017; Bellemare et al., 2016; Burda et al., 2019b).

Intrinsic motivation can be useful in guiding the exploration of RL agents, particularly in environments where the extrinsic feedback is sparse or missing altogether (Oudeyer et al., 2007; 2008;

Oudeyer & Kaplan, 2009; Schmidhuber, 1991b; 2010). The most popular and effective kinds of intrinsic motivation can be split into two broad classes: count-based methods that encourage the agent to visit new states and curiosity-based methods that encourage the agent to learn about the environment dynamics.

Count-Based Exploration. Strehl & Littman (2008) proposed the use of state visitation counts as an exploration bonus in tabular settings. More recently, such methods were extended to high-dimensional state spaces (Bellemare et al., 2016; Ostrovski et al., 2017; Martin et al., 2017; Tang et al., 2017; Machado et al., 2018a). Bellemare et al. (2016) use a Context-Tree Switching (CTS) density model to derive a state pseudo-count, while Ostrovski et al. (2017) use PixelCNN as a state density estimator. Burda et al. (2019b) employ the prediction error of a random network as exploration bonus with the aim of rewarding novel states more than previously seen ones. However, one can expect count-based exploration methods to be less effective in procedurally-generated environments with sparse reward. In these settings, the agent is likely to characterize two states as being different even when they only differ by features that are irrelevant for the task (*e.g.* the texture of the walls). If the agent considers most states to be "novel", the feedback signal will not be distinctive or varied enough to guide the agent.

Curiosity-Driven Exploration. Curiosity-based bonuses encourage the agent to explore the environment to learn about its dynamics. Curiosity can be formulated as the error or uncertainty in predicting the consequences of the agent's actions (Stadie et al., 2015; Pathak et al., 2017; Burda et al., 2019b). For example, Pathak et al. (2017) learn a latent representation of the state and design an intrinsic reward based on the error of predicting the next state in the learned latent space. While we use a similar mechanism for learning state embeddings, our exploration bonus is very different and builds upon the difference between the latent representations of two consecutive states. As we will see in the following sections, one problem with their approach is that the intrinsic reward can vanish during training, leaving the agent with no incentive to further explore the environment and reducing its feedback to extrinsic reward only.

Generalization in Deep RL. Most of the existing exploration methods that have achieved impressive results on difficult tasks (Ecoffet et al., 2019; Pathak et al., 2017; Burda et al., 2019b; Bellemare et al., 2016; Choi et al., 2019; Aytar et al., 2018), have been trained and tested on the same environment and thus do not generalize to new instances. Several recent papers (Rajeswaran et al., 2017; Zhang et al., 2018a;b; Machado et al., 2018b; Foley et al., 2018) demonstrate that deep RL is susceptible to severe overfitting. As a result, a number of benchmarks have been recently released for testing generalization in RL (Beattie et al., 2016; Cobbe et al., 2019; Packer et al., 2018; Justesen et al., 2018; Leike et al., 2017; Nichol et al., 2018; Juliani et al., 2019). Here, we make another step towards developing exploration methods that can generalize to unseen scenarios by evaluating them on procedurally-generated environments. We opted for MiniGrid (Chevalier-Boisvert et al., 2018) because it is fast to run, provides a standard set of tasks with varied difficulty levels, focuses on single-agent, and does not use visual inputs, which allows us to better isolate the exploration problem.

More closely related to our work are the papers of Marino et al. (2019) and Zhang et al. (2019). Marino et al. (2019) use a reward that encourages changing the values of the non-proprioceptive features for training low-level policies on locomotion tasks. Their work assumes that the agent has access to a decomposition of the observation state into internal and external parts, an assumption which may not hold in many cases and may not be trivial to obtain even if it exists. Zhang et al. (2019) use the difference between the successor features of consecutive states as intrinsic reward. In this framework, a state is characterized through the features of all its successor states. While both of these papers use fixed (*i.e.* not learned) state representations to define the intrinsic reward, we use forward and inverse dynamics models to learn a state representation constrained to only capture elements in the environment that can be influenced by the agent. Lesort et al. (2018) emphasize the benefits of using a learned state representation for control as opposed to a fixed one, which may not contain information relevant for acting in the environment. In the case of Zhang et al. (2019), constructing a temporally extended state representation for aiding exploration is not trivial. Such a feature space may add extra noise to the intrinsic reward due to the uncertainty in the states that might be visited in the future. This is particularly problematic when the agent often encounters states that have never been visited before (as it is the case in procedurally-generated environments) or the environment is highly stochastic.

3 BACKGROUND: CURIOSITY-DRIVEN EXPLORATION

We use the standard formalism of a single agent Markov Decision Process (MDP) defined by a set of states $\mathcal{S}$, a set of actions $\mathcal{A}$, and a transition function $\mathcal{T} : \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{P}(\mathcal{S})$ which gives the probability distribution of the next state given a current state and action. The agent chooses actions by sampling from a stochastic policy $\pi : \mathcal{S} \rightarrow \mathcal{P}(\mathcal{A})$, and receives reward $r : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ at every time step. The agent's goal is to learn a policy which maximizes its discounted expected return $R_t = \mathbb{E}\left[\sum_{k=0}^{T} \gamma^k r_{t+k+1}\right]$ where r_t is the sum of the intrinsic and extrinsic reward received by the agent at time t, $\gamma \in [0, 1]$ is the discount factor, and the expectation is taken with respect to both the policy and the environment. Here, we consider the case of episodic RL in which the agent maximizes the reward received within a finite time horizon.

In this paper we consider that, along with the extrinsic reward r_t^e, the agent also receives some intrinsic reward r_t^i, which can be computed for any (s_{t-1}, a_{t-1}, s_t) tuple. Consequently, the agent tries to maximize the weighted sum of the intrinsic and extrinsic reward: $r_t = r_t^e + \omega_{ir} r_t^i$ where ω_{ir} is a hyper-parameter to weight the importance of both rewards.

We built upon the work of Pathak et al. (2017) who note that some parts of the observation may have no influence on the agent's state. Thus, Pathak et al. propose learning a state representation that disregards those parts of the observation and instead only models (i) the elements that the agent can control, as well as (ii) those that can affect the agent, even if the agent cannot have an effect on them. Concretely, Pathak et al. learn a state representations $\phi(s) = f_{emb}(s;\ \theta_{emb})$ of a state s using an inverse and a forward dynamics model (see Figure 1). The forward dynamics model is a neural network parametrized by θ_{fw} that takes as inputs $\phi(s_t)$ and a_t, predicts the next state representation: $\hat{\phi}(s_{t+1}) = f_{fw}(\phi_t,\ a_t;\ \theta_{fw})$, and is trained to minimize $L_{fw}(\theta_{fw}, \theta_{emb}) = \|\hat{\phi}(s_{t+1}) - \phi(s_{t+1})\|_2^2$. The inverse dynamics model is also a neural network parametrized by θ_{inv} that takes as inputs $\phi(s_t)$ and $\phi(s_{t+1})$, predicts the agent's action: $\hat{a}_t = f_{inv}(\phi_t, \phi_{t+1};\ \theta_{inv})$, and is trained to minimize $L_{inv}(\theta_{inv}, \theta_{emb}) = CrossEntropy(\hat{a}_t, a_t)$ when the action space is discrete. Pathak et al.'s curiosity-based intrinsic reward is then defined as the Euclidean distance between the actual embedding of the next state $\phi(s_{t+1})$ and the one predicted by the forward model $\hat{\phi}(s_{t+1})$.

4 IMPACT-DRIVEN EXPLORATION

Our main contribution is a novel intrinsic reward based on the change in the state representation produced by the agent's action. The proposed method encourages the agent to try out actions that have a significant impact on the environment and we demonstrate that it can promote effective exploration strategies when the feedback from the environment is sparse.

We train a forward and an inverse dynamics model to learn a latent state representation $\phi(s)$ as proposed by Pathak et al. (2017). However, instead of using the Euclidean distance between the predicted next state representation and the actual next state representation as intrinsic reward (R_{cur} in Figure 1), we define impact-driven reward as the change between consecutive state representations (R_{IDE} in Figure 1). Compared to curiosity-driven exploration, impact-driven exploration rewards the agent for very different state-actions, leading to distinct agent behaviors which we analyze in Section 6.1.1.

Stanton & Clune (2018) categorize exploration into: *across-training* and *intra-life* and argue they are complementary. Popular methods such as count-based exploration (Bellemare et al., 2016) encourage agents to visit new states in relation to all prior training episodes (*i.e.* across-training novelty), but they do not consider whether an agent visits new states within some episode (*i.e.* intra-life novelty). As we will see, RIDE combines both types of exploration.

Formally, RIDE is computed as the L_2-norm $\|\phi(s_{t+1}) - \phi(s_t)\|_2$ of the difference in the learned state representation between consecutive states. However, to ensure that the agent does not go back and forth between a sequence of states (with a large difference in their embeddings) in order to gain intrinsic reward, we discount RIDE by episodic state visitation counts. Concretely, we divide the impact-driven reward by $\sqrt{N_{ep}(s_{t+1})}$, where $N_{ep}(s_{t+1})$ is the number of times that state has been visited during the current episode, which is initialized to 1 in the beginning of the episode. In high-dimensional regimes, one can use episodic pseudo-counts instead (Bellemare et al., 2016; Ostrovski

et al., 2017). Thus, the overall intrinsic reward provided by RIDE is calculated as:

$$R_{IDE}(s_t, a_t) \equiv r_t^i(s_t, a_t) = \frac{\|\phi(s_{t+1}) - \phi(s_t)\|_2}{\sqrt{N_{ep}(s_{t+1})}}$$

where $\phi(s_{t+1})$ and $\phi(s_t)$ are the learned representations of consecutive states, resulted from the agent transitioning to state s_{t+1} after taking action a_t in state s_t. The state is projected into a latent space using a neural network with parameters θ_{emb}.

The overall optimization problem that is solved for training the agent is

$$\min_{\theta_\pi, \theta_{inv}, \theta_{fw}, \theta_{emb}} [\omega_\pi L_{RL}(\theta_\pi) + \omega_{fw} L_{fw}(\theta_{fw}, \theta_{emb}) + \omega_{inv} L_{inv}(\theta_{inv}, \theta_{emb})]$$

where θ_π are the parameters of the policy and value network ($a_t \sim \pi(s_t; \theta_\pi)$), and ω_π, ω_{inv} and ω_{fw} are scalars that weigh the relative importance of the reinforcement learning (RL) loss to that of the inverse and forward dynamics losses which are used for learning the intrinsic reward signal. Note that we never update the parameters of the inverse (θ_{inv}), forward (θ_{fw}), or embedding networks (θ_{emb}) using the signal from the intrinsic or extrinsic reward (*i.e.* the RL loss); we only use these learned state embeddings for constructing the exploration bonus and never as part of the agent's policy (Figure 1 highlights that the policy learns its own internal representation of the state ψ_t, which is only used for control and never for computing the intrinsic reward). Otherwise, the agent can artificially maximize its intrinsic reward by constructing state representations with large distances among themselves, without grounding them in observations of the environment.

Note that there is no incentive for the learned state representations to encode features of the environment that cannot be influenced by the agent's actions. Thus, our agent will not receive rewards for reaching states that are inherently unpredictable, making exploration robust with respect to distractor objects or other inconsequential sources of variation in the environment. As we will later show, RIDE is robust to the well-known noisy-TV problem in which an agent rewarded for errors in the prediction of its forward model (such as the one proposed in Pathak et al. (2017)) gets attracted to local sources of entropy in the environment. Furthermore, the difference of consecutive state representations is unlikely to go to zero during learning as they are representations of actual states visited by the agent and constrained by the forward and inverse model. This is in contrast to Pathak et al. (2017) and Burda et al. (2019b) where the intrinsic reward goes to zero as soon as the forward model becomes sufficiently accurate or the agent's policy only explores well known parts of the state space.

5 EXPERIMENTS

We evaluate RIDE on procedurally-generated environments from MiniGrid, as well as two existing singleton environments used in prior work, and compare it against both standard RL and three commonly used intrinsic reward methods for exploration. For all our experiments, we show the mean and standard deviation of the average return across 5 different seeds for each model. The average return is computed as the rolling mean over the past 100 episodes.

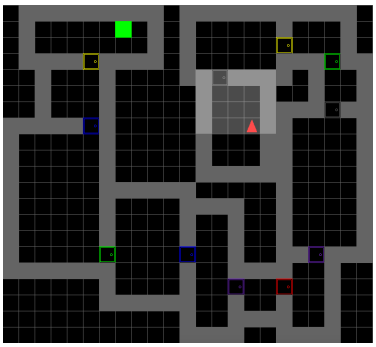

Figure 2: Rendering of a procedurally-generated environment from MiniGrid's MultiRoomN12S10 task.

5.1 ENVIRONMENTS

The first set of environments are procedurally-generated gridworlds in MiniGrid (Chevalier-Boisvert et al., 2018). We consider three types of hard exploration tasks: *MultiRoomNXSY*, *KeyCorridorS3R3*, and *ObstructedMaze2Dlh*.

In MiniGrid, the world is a partially observable grid of size NxN. Each tile in the grid contains at most one of the following objects: wall, door, key, ball, box and goal. The agent can take one of seven actions: turn left or right, move forward, pick up or drop an object, toggle or done. More details about the MiniGrid environment and tasks can be found in A.3.

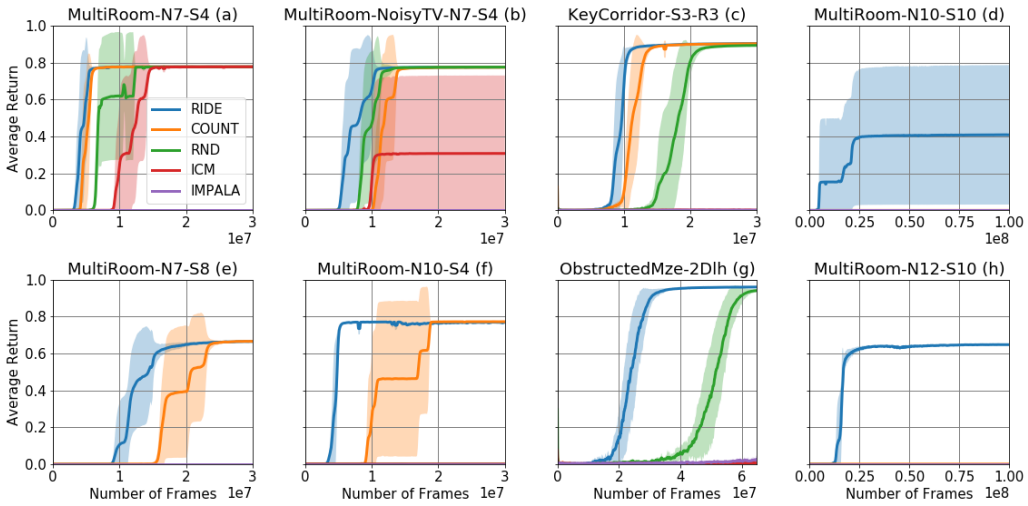

Figure 3: Performance of RIDE, Count, RND, ICM and IMPALA on a variety of hard exploration problems in MiniGrid. Note RIDE is the only one that can solve the hardest tasks.

For the sole purpose to compare in a fair way to the curiosity-driven exploration work by Pathak et al. (2017), we ran a one-off experiment on their Mario (singleton) environment. We train our model with and without extrinsic reward on the first level of the game.

The last (singleton) environment we evaluate on is *VizDoom* (Kempka et al., 2016). Details about the environment can be found in A.4.

5.2 Baselines

For all our experiments, we use IMPALA (Espeholt et al., 2018) as the base RL algorithm, and RMSProp (Tieleman & Hinton, 2012) for optimization. All models use the same basic RL algorithm and network architecture for the policy and value functions (see Appendix A.2 and Appendix A.1 for details regarding the hyperparameters and network architectures), differing only in how intrinsic rewards are defined. In our experiments we compare with the following baselines: **Count**: Count-Based Exploration by Bellemare et al. (2016) which uses state visitation counts to give higher rewards for new or rarely seen states. **RND**: Random Network Distillation Exploration by Burda et al. (2019b) which uses the prediction error of a random network as exploration bonus with the aim of rewarding novel states more than previously encountered ones. **ICM**: Intrinsic Curiosity Module by Pathak et al. (2017) (see Section 3). **IMPALA**: Standard RL approach by Espeholt et al. (2018) that uses only extrinsic reward and encourages random exploration by adding the entropy of the policy to the objective function.

6 Results and Discussion

We present results of RIDE in comparison to popular exploration methods, as well as an analysis of the learned policies and properties of the intrinsic reward generated by different methods.

6.1 MiniGrid

Figure 3 summarizes our results on various hard MiniGrid tasks. Note that the standard RL approach IMPALA (purple) is not able to learn in any of the environments since extrinsic reward is too sparse. Furthermore, our results reveal that RIDE is more sample efficient compared to all the other exploration methods across all MiniGrid tasks considered here. While other exploration bonuses seem effective on easier tasks and are able to learn optimal policies where IMPALA fails, the gap between our approach and the others is increasing with the difficulty of the task. Furthermore, RIDE manages to solve some very challenging tasks on which the other methods fail to get any reward even after training on over 100M frames (Figure 3).

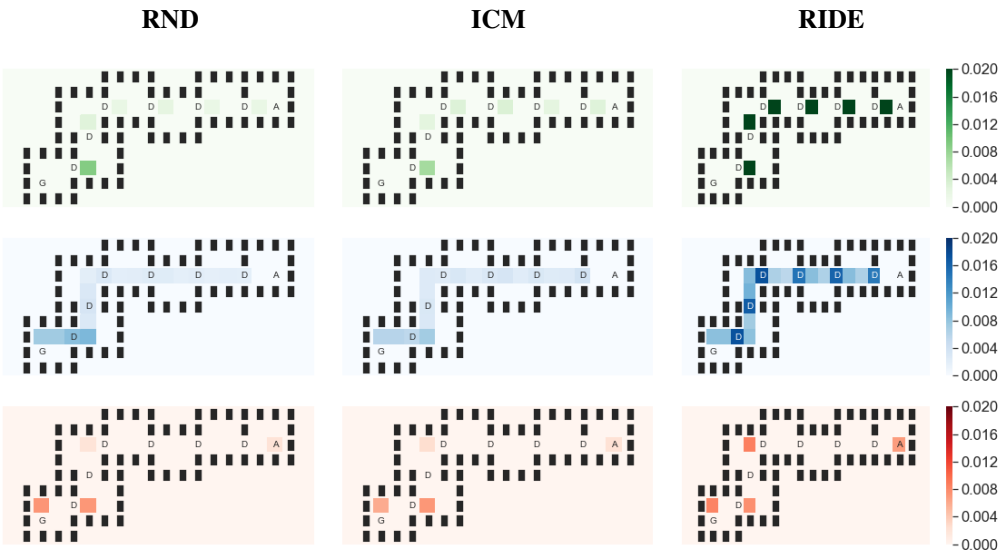

Figure 4: Intrinsic reward heatmaps for RND, ICM, and RIDE (from left to right) for opening doors (green), moving forward (blue), or turning left or right (red) on a random environment from the MultiRoomN7S4 task. A is the agent's starting position, G is the goal position and D are doors that have to be opened on the way.

	Open Door		Turn Left / Right		Move Forward	
Model	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
RIDE	0.0490	0.0019	0.0071	0.0034	0.0181	0.0116
RND	0.0032	0.0018	0.0031	0.0028	0.0026	0.0017
ICM	0.0055	0.0003	0.0052	0.0003	0.0056	0.0003

Table 1: Mean intrinsic reward per action over 100 episodes on a random maze in MultiRoomN7S4.

In addition to existing MiniGrid tasks, we also tested the ability of the models to deal with stochasticity in the environment by adding a "noisy TV" in the MiniGridN7S4 task, resulting in the new MiniGirdN7S4NoisyTV task (left-center plot in the top row of Figure 3). The noisy TV is implemented as a ball that changes its color to a randomly picked one whenever the agent takes a particular action. As expected, the performance of ICM drops as the agent becomes attracted to the ball as it gets rewarded for not being able to predict the next color. The Count model also needs more time to train, likely caused by the increasing number of rare and novel states (due to the changing color of the ball).

We include results for ablations to our model in Appendix A.5, highlighting the importance of combining impact-driven exploration with episodic state visitation discounting.

6.1.1 Analysis of the Intrinsic Reward

To better understand the effectiveness of different exploration methods, we investigate the intrinsic reward an agent receives for certain trajectories in the environment.

Figure 4 shows a heatmap of the intrinsic reward received by RND, ICM, and RIDE on a sampled environment after having been trained on procedurally-generated environments from the MultiRoomN7S4 task. While all three methods can solve this task, the intrinsic rewards received are different. Specifically, the RIDE agent is rewarded in a much more structured manner for opening doors, entering new rooms and turning at decision points. Table 1 provides quantitative numbers for this phenomenon. We record the intrinsic rewards received for each type of action, averaged over 100 episodes. We found that the RIDE is putting more emphasis on actions interacting with the door than for moving forward or turning left or right, while the other methods reward actions more uniformly.

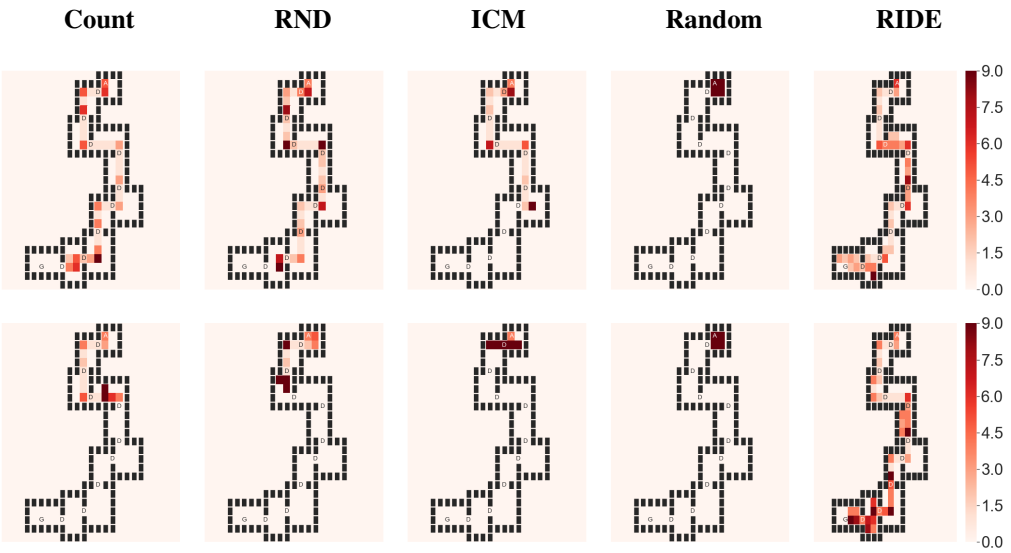

Figure 6: State visitation heatmaps for Count, RND, ICM, Random, and RIDE models (from left to right) trained for 50m frames without any extrinsic reward on a singleton maze (top row) and on procedurally-generated mazes (bottom row) in MultiRoomN10S6.

Figure 12 and Table 3 in A.6.2 show a similar pattern in the intrinsic rewards for agents trained on the MultiRoomN12S10 task, while Figure 13 and Table 4 in A.6.3 contain the equivalent analysis for agents trained on ObstructedMaze2Dlh. As emphasized there, RIDE is rewarded more for interactions with objects relative to actions for moving around in the maze, a characteristic which is not as prevalent in the other models.

Figure 5 shows the mean intrinsic reward of all models while training on the MultiRoomN12S10 task. While the ICM, RND, and Count intrinsic reward converges to very low values quite early in the training process, the RIDE bonus keeps changing and has a higher value even after training on 100m frames. Hence, RIDE constantly encourages the agent to take actions that change the local environment. In contrast, Count, RND, and Curiosity may not consider certain states to be "novel" or "surprising" after longer periods of training as they have seen similar states in the past or learned to almost perfectly predict the next state in a subset of the environment states. Consequently, their intrinsic rewards diminish during training, so the agent will have a difficult time distinguishing between actions that lead to novel or surprising states and those that do not and it may get trapped in some parts of the state space (see Figure 12).

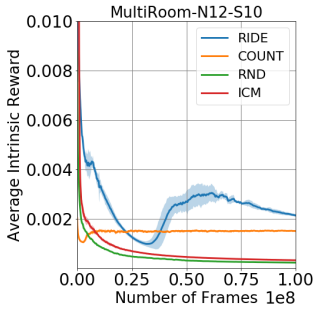

Figure 5: Mean intrinsic reward for models trained on MultiRoomN12S10.

6.1.2 Singleton versus Procedurally-Generated Environments

It is important to understand and quantify how much harder it is to train existing deep RL exploration methods on tasks in procedurally-generated environments compared to a singleton environment.

To investigate this dependency, we trained the models on a singleton environment of the the *ObstructedMaze2Dlh* task so that at the beginning of every episode, the agent is spawned in exactly the same maze with all objects located in the same positions. In this setting, we see that Count, RND, and IMPALA are also able to solve the task (see Figure 7 and compare with the center-right plot in the bottom row of Figure 3 for procedurally-generated environments of the same task). As expected, this emphasizes that training an agent in procedurally-generated environments creates significant challenges over training on a singleton environment for the same task. Moreover, it highlights the importance of training on a variety of environments to avoid overfitting to the idiosyncrasies of a particular environment.

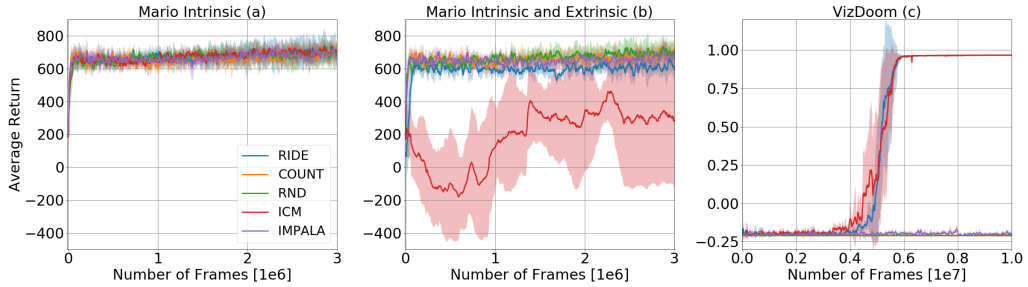

Figure 8: Performance on Mario with intrinsic reward only (a), with intrinsic and extrinsic reward (b), and VizDoom (c). Note that IMPALA is trained with extrinsic reward only in all cases.

6.1.3 No Extrinsic Reward

To analyze the way different methods explore environments without depending on the chance of running into extrinsic reward (which can dramatically change the agent's policy), we analyze agents that are trained without any extrinsic reward on both singleton and procedurally-generated environments.

The top row of Figure 6 shows state visitation heatmaps for all the models in a singleton environment on the MultiRoomN10S6 task, after training all of them for 50m frames with intrinsic reward only. The agents are allowed to take 200 steps in every episode. The figure indicates that all models have effective exploration strategies when trained on a singleton maze, the 10th, 9th and 6th rooms are reached by RIDE, Count/RND, and ICM, respectively. The Random policy fully explores the first room but does not get to the second room within the time limit.

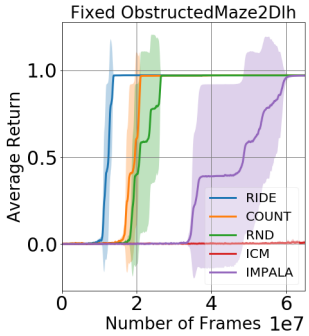

Figure 7: Training on a singleton instance of ObstructedMaze2Dlh.

When trained on procedurally-generated mazes, existing models are exploring much less efficiently as can be seen in the bottom row of Figure 6. Here, Count, RND, and ICM only make it to the 4th, 3rd and 2nd rooms respectively within an episode, while RIDE is able to explore all rooms. This further supports that RIDE learns a state representation that allows generalization across different mazes and is not as distracted by less important details that change from one procedurally-generated environment to another.

6.2 Mario and VizDoom

Our results on the first level of the Mario environment in order to compare to Pathak et al. (2017) (see Figure 8a and b) suggest that this environment may not be as challenging as previously believed, given that all the methods evaluated here, including vanilla IMPALA, can learn similarly good policies after training on only 1m frames even without any intrinsic reward (left figure). Note that we are able to reproduce the results mentioned in the original ICM paper (Pathak et al., 2017). However, when training with both intrinsic and extrinsic reward (center figure), the curiosity-based exploration bonus (ICM) hurts learning, converging later and to a lower value than the other methods evaluated here.

For VizDoom (see Figure 8c) we observe that RIDE performs as well as ICM, while all the other baselines fail to learn an effective policy given the same amount of training. Note that our ICM implementation can reproduce the results in the original paper on this task, achieving a 100% success rate after training on approximately 60m frames (Pathak et al., 2017).

7 CONCLUSION AND FUTURE WORK

In this work, we propose rewarding impact-driven exploration (RIDE), an intrinsic reward bonus that encourages the agent to explore actions that substantially change the state of the environment, as measured in a learned latent space. RIDE has a number of desirable properties: it attracts the agent to states where it can affect the environment, it provides a signal to the agent even after training for a long time, and it is conceptually simple as well as compatible with other intrinsic or extrinsic rewards and any deep RL algorithm.

Our approach is particularly effective in procedurally-generated sparse-reward environments where it significantly outperforms IMPALA (Espeholt et al., 2018), as well as some of the most popular exploration methods: Count (Bellemare et al., 2016), RND (Burda et al., 2019b), ICM (Pathak et al., 2017). Furthermore, RIDE explores procedurally-generated environments more efficiently than the other exploration methods.

However, there are still many ways to improve upon RIDE. For example, one can make use of symbolic information to measure or characterize the agent's impact, consider longer-term effects of the agent's actions, or promote diversity among the kinds of changes the agent makes to the environment. Another interesting avenue for future research is to develop algorithms that can distinguish between desirable and undesirable types of impact the agent can have in the environment, thus constraining the agent to act safely and avoid distractions (*i.e.* actions that lead to large changes in the environment but that are not useful for a given task). Such different kinds of impact might correspond to distinctive skills or low-level policies that a hierarchical controller can use to learn more complex policies or better exploration strategies.

REFERENCES

Joshua Achiam and Shankar Sastry. Surprise-based intrinsic motivation for deep reinforcement learning. *CoRR*, abs/1703.01732, 2017. URL http://arxiv.org/abs/1703.01732.

Yusuf Aytar, Tobias Pfaff, David Budden, Thomas Paine, Ziyu Wang, and Nando de Freitas. Playing hard exploration games by watching youtube. In *Advances in Neural Information Processing Systems*, pp. 2930–2941, 2018.

Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. *CoRR*, abs/1612.03801, 2016. URL http://arxiv.org/abs/1612.03801.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pp. 1471–1479, 2016.

Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

Yuri Burda, Harrison Edwards, Deepak Pathak, Amos J. Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019a. URL https://openreview.net/forum?id=rJNwDjAqYX.

Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019b. URL https://openreview.net/forum?id=H1lJJnR5Ym.

Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic gridworld environment for openai gym. https://github.com/maximecb/gym-minigrid, 2018.

Jongwook Choi, Yijie Guo, Marcin Moczulski, Junhyuk Oh, Neal Wu, Mohammad Norouzi, and Honglak Lee. Contingency-aware exploration in reinforcement learning. In *7th International*

Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, 2019. URL https://openreview.net/forum?id=HyxGB2AcY7.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.07289.

Karl Cobbe, Oleg Klimov, Christopher Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, pp. 1282–1289, 2019. URL http://proceedings.mlr.press/v97/cobbe19a.html.

Nat Dilokthanakul, Christos Kaplanis, Nick Pawlowski, and Murray Shanahan. Feature control as intrinsic motivation for hierarchical reinforcement learning. *IEEE transactions on neural networks and learning systems*, 2019.

Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. Go-explore: a new approach for hard-exploration problems. *arXiv preprint arXiv:1901.10995*, 2019.

Lasse Espeholt, Hubert Soyer, Rémi Munos, Karen Simonyan, Volodymyr Mnih, Tom Ward, Yotam Doron, Vlad Firoiu, Tim Harley, Iain Dunning, Shane Legg, and Koray Kavukcuoglu. IMPALA: scalable distributed deep-rl with importance weighted actor-learner architectures. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 1406–1415, 2018. URL http://proceedings.mlr.press/v80/espeholt18a.html.

Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=SJx63jRqFm.

John Foley, Emma Tosch, Kaleigh Clary, and David Jensen. Toybox: Better atari environments for testing reinforcement learning agents. *CoRR*, abs/1812.02850, 2018. URL http://arxiv.org/abs/1812.02850.

Meire Fortunato, Mohammad Gheshlaghi Azar, Bilal Piot, Jacob Menick, Matteo Hessel, Ian Osband, Alex Graves, Volodymyr Mnih, Rémi Munos, Demis Hassabis, Olivier Pietquin, Charles Blundell, and Shane Legg. Noisy networks for exploration. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018. URL https://openreview.net/forum?id=rywHCPkAW.

Anirudh Goyal, Riashat Islam, Daniel Strouse, Zafarali Ahmed, Hugo Larochelle, Matthew Botvinick, Yoshua Bengio, and Sergey Levine. Infobot: Transfer and exploration via the information bottleneck. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=rJg8yhAqKm.

Karol Gregor, Danilo Jimenez Rezende, and Daan Wierstra. Variational intrinsic control. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*, 2017. URL https://openreview.net/forum?id=Skc-Fo4Yg.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pp. 1109–1117, 2016.

Max Jaderberg, Volodymyr Mnih, Wojciech Marian Czarnecki, Tom Schaul, Joel Z. Leibo, David Silver, and Koray Kavukcuoglu. Reinforcement learning with unsupervised auxiliary tasks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. URL https://openreview.net/forum?id=SJ6yPD5xg.

Max Jaderberg, Wojciech M Czarnecki, Iain Dunning, Luke Marris, Guy Lever, Antonio Garcia Castaneda, Charles Beattie, Neil C Rabinowitz, Ari S Morcos, Avraham Ruderman, et al. Human-level performance in 3d multiplayer games with population-based reinforcement learning. *Science*, 364(6443):859–865, 2019.

Yacine Jernite, Kavya Srinet, Jonathan Gray, and Arthur Szlam. Craftassist instruction parsing: Semantic parsing for a minecraft assistant. *CoRR*, abs/1905.01978, 2019. URL http://arxiv.org/abs/1905.01978.

Matthew Johnson, Katja Hofmann, Tim Hutton, and David Bignell. The malmo platform for artificial intelligence experimentation. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, pp. 4246–4247, 2016. URL http://www.ijcai.org/Abstract/16/643.

Arthur Juliani, Ahmed Khalifa, Vincent-Pierre Berges, Jonathan Harper, Ervin Teng, Hunter Henry, Adam Crespi, Julian Togelius, and Danny Lange. Obstacle tower: A generalization challenge in vision, control, and planning. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pp. 2684–2691, 2019. doi: 10.24963/ijcai.2019/373. URL https://doi.org/10.24963/ijcai.2019/373.

Niels Justesen, Ruben Rodriguez Torrado, Philip Bontrager, Ahmed Khalifa, Julian Togelius, and Sebastian Risi. Illuminating generalization in deep reinforcement learning through procedural level generation. *arXiv preprint arXiv:1806.10729*, 2018.

Michał Kempka, Marek Wydmuch, Grzegorz Runc, Jakub Toczek, and Wojciech Jaśkowski. Vizdoom: A doom-based ai research platform for visual reinforcement learning. In *2016 IEEE Conference on Computational Intelligence and Games (CIG)*, pp. 1–8. IEEE, 2016.

Alexander S Klyubin, Daniel Polani, and Chrystopher L Nehaniv. All else being equal be empowered. In *European Conference on Artificial Life*, pp. 744–753. Springer, 2005.

Jan Leike, Miljan Martic, Victoria Krakovna, Pedro A Ortega, Tom Everitt, Andrew Lefrancq, Laurent Orseau, and Shane Legg. Ai safety gridworlds. *arXiv preprint arXiv:1711.09883*, 2017.

Timothée Lesort, Natalia Díaz Rodríguez, Jean-François Goudou, and David Filliat. State representation learning for control: An overview. *Neural Networks*, 108:379–392, 2018. doi: 10.1016/j.neunet.2018.07.006. URL https://doi.org/10.1016/j.neunet.2018.07.006.

Daniel Ying-Jeh Little and Friedrich Tobias Sommer. Learning and exploration in action-perception loops. *Frontiers in neural circuits*, 7:37, 2013.

Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. *arXiv preprint arXiv:1807.11622*, 2018a.

Marlos C Machado, Marc G Bellemare, Erik Talvitie, Joel Veness, Matthew Hausknecht, and Michael Bowling. Revisiting the arcade learning environment: Evaluation protocols and open problems for general agents. *Journal of Artificial Intelligence Research*, 61:523–562, 2018b.

Kenneth Marino, Abhinav Gupta, Rob Fergus, and Arthur Szlam. Hierarchical RL using an ensemble of proprioceptive periodic policies. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019. URL https://openreview.net/forum?id=SJz1x20cFQ.

Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pp. 2471–2478, 2017. doi: 10.24963/ijcai.2017/344. URL https://doi.org/10.24963/ijcai.2017/344.

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 *International conference on machine learning*, pp. 1928–1937, 2016.

Nirbhay Modhe, Prithvijit Chattopadhyay, Mohit Sharma, Abhishek Das, Devi Parikh, Dhruv Batra, and Ramakrishna Vedantam. Unsupervised discovery of decision states for transfer in reinforcement learning. *CoRR*, abs/1907.10580, 2019. URL http://arxiv.org/abs/1907.10580.

Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, and John Schulman. Gotta learn fast: A new benchmark for generalization in rl. *arXiv preprint arXiv:1804.03720*, 2018.

Brendan O'Donoghue, Ian Osband, Rémi Munos, and Volodymyr Mnih. The uncertainty bellman equation and exploration. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pp. 3836–3845, 2018. URL http://proceedings.mlr.press/v80/o-donoghue18a.html.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped dqn. In *Advances in neural information processing systems*, pp. 4026–4034, 2016.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pp. 2721–2730. JMLR. org, 2017.

Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 2009.

Pierre-Yves Oudeyer, Frdric Kaplan, and Verena V Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE transactions on evolutionary computation*, 11(2):265–286, 2007.

Pierre-Yves Oudeyer, Frederic Kaplan, et al. How can we define intrinsic motivation. In *Proc. of the 8th Conf. on Epigenetic Robotics*, volume 5, pp. 29–31, 2008.

Charles Packer, Katelyn Gao, Jernej Kos, Philipp Krähenbühl, Vladlen Koltun, and Dawn Song. Assessing generalization in deep reinforcement learning. *arXiv preprint arXiv:1810.12282*, 2018.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 16–17, 2017.

Sébastien Racanière, Theophane Weber, David P. Reichert, Lars Buesing, Arthur Guez, Danilo Jimenez Rezende, Adrià Puigdomènech Badia, Oriol Vinyals, Nicolas Heess, Yujia Li, Razvan Pascanu, Peter W. Battaglia, Demis Hassabis, David Silver, and Daan Wierstra. Imagination-augmented agents for deep reinforcement learning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 5690–5701, 2017. URL http://papers.nips.cc/paper/7152-imagination-augmented-agents-for-deep-reinforcement-learning.

Aravind Rajeswaran, Kendall Lowrey, Emanuel V Todorov, and Sham M Kakade. Towards generalization and simplicity in continuous control. In *Advances in Neural Information Processing Systems*, pp. 6550–6561, 2017.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, pp. 1530–1538, 2015. URL http://proceedings.mlr.press/v37/rezende15.html.

Jürgen Schmidhuber. Curious model-building control systems. In *Proc. international joint conference on neural networks*, pp. 1458–1463, 1991a.

Jürgen Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *Proc. of the international conference on simulation of adaptive behavior: From animals to animats*, pp. 222–227, 1991b.

Jürgen Schmidhuber. Developmental robotics, optimal artificial curiosity, creativity, music, and the fine arts. *Connection Science*, 18(2):173–187, 2006.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

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*, 550(7676):354, 2017.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Christopher Stanton and Jeff Clune. Deep curiosity search: Intra-life exploration can improve performance on challenging deep reinforcement learning problems. *arXiv preprint arXiv:1806.00553*, 2018.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in neural information processing systems*, pp. 2753–2762, 2017.

T Tieleman and G Hinton. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *Tech. Rep., Technical report*, pp. 31, 2012.

Amy Zhang, Nicolas Ballas, and Joelle Pineau. A dissection of overfitting and generalization in continuous reinforcement learning. *arXiv preprint arXiv:1806.07937*, 2018a.

Chiyuan Zhang, Oriol Vinyals, Remi Munos, and Samy Bengio. A study on overfitting in deep reinforcement learning. *arXiv preprint arXiv:1804.06893*, 2018b.

Jingwei Zhang, Niklas Wetzel, Nicolai Dorka, Joschka Boedecker, and Wolfram Burgard. Scheduled intrinsic drive: A hierarchical take on intrinsically motivated exploration. *arXiv preprint arXiv:1903.07400*, 2019.

A APPENDIX

A.1 NETWORK ARCHITECTURES

All our models use the same network architecture for the policy and value networks. The input is passed through a sequence of three (for MiniGrid) or four (environments used by Pathak et al. (2017)) convolutional layers with 32 filters each, kernel size of 3x3, stride of 2 and padding of 1. An exponential linear unit (ELU; (Clevert et al. (2016))) is used after each convolution layer. The output of the last convolution layer is fed into a LSTM with 256 units. Two separate fully connected layers are used to predict the value function and the action from the LSTM feature representation.

For the singleton environments used in prior work, the agents are trained using visual inputs that are pre-processed similarly to Mnih et al. (2016). The RGB images are converted into gray-scale and re-sized to 42 42. The input given to both the policy and the state representation networks consists of the current frame concatenated with the previous three frames. In order to reduce overfitting, during training, we use action repeat of four. At inference time, we sample the policy without any action repeats.

A.2 HYPERPARAMETERS

We ran grid searches over the learning rate $\in [0.0001, 0.0005, 0.001]$, batch size $\in [8, 32]$ and unroll length $\in [20, 40, 100, 200]$. The best values for all models can be found in Table 2.

Parameter	**Value**
Learning Rate	0.0001
Batch Size	32
Unroll Length	100
Discount	0.99
RMSProp Momentum	0.0
RMSProp ϵ	0.01
Clip Gradient Norm ϵ	40.0

Table 2: Hyperparameters common to all experiments.

The learning rate is linearly annealed to 0 in all experiments.

We also ran grid searches over the intrinsic reward coefficient $\in [1.0, 0.5, 0.1, 0.05, 0.01, 0.005, 0.001]$ for all the models and found that the best values were 0.1 for RIDE, ICM and RND, and 0.05 for Count. We also ran grid searches over the entropy coefficient $\in [0.01, 0.005, 0.001, 0.0005, 0.0001, 0.00005]$ for all the models and found that the best values were 0.0005 for RIDE and 0.0001 for Count, RND, and ICM. In all experiments presented here, we use the best values found for each model.

A.3 MINIGRID ENVIRONMENT

In MiniGrid, the world is a partially observable grid of size NxN. Each tile in the grid contains exactly zero or one object. The possible object types are wall, door, key, ball, box and goal.

Each object in MiniGrid has an associated discrete color, which can be one of red, green, blue, purple, yellow and grey. By default, walls are always grey and goal squares are always green. Rewards are sparse for all MiniGrid environments.

There are seven actions in MiniGrid: turn left, turn right, move forward, pick up an object, drop an object, toggle and done. The agent can use the turn left and turn right action to rotate and face one of 4 possible directions (north, south, east, west). The move forward action makes the agent move from its current tile onto the tile in the direction it is currently facing, provided there is nothing on that tile, or that the tile contains an open door. The agent can open doors if they are right in front of it by using the toggle action.

Observations in MiniGrid are partial and egocentric. By default, the agent sees a square of 7x7 tiles in the direction it is facing. These include the tile the agent is standing on. The agent cannot see

through walls or closed doors. The observations are provided as a tensor of shape 7x7x3. However, note that these are not RGB images. Each tile is encoded using 3 integer values: one describing the type of object contained in the cell, one describing its color, and a flag indicating whether doors are open or closed. This compact encoding was chosen for space efficiency and to enable faster training. For all tasks, , the agent gets an egocentric view of its surroundings, consisting of 33 pixels. A neural network parameterized as a CNN is used to process the visual observation.

The MultiRoomNXSY environment consists of X rooms, with size at most Y , connected in random orientations. The agent is placed in the first room and must navigate to a green goal square in the most distant room from the agent. The agent receives an egocentric view of its surrounding, consisting of 33 pixels. The task increases in difficulty with X and Y. Episodes finish with a positive reward when the agent reaches the green goal square. Otherwise, episodes are terminated with zero reward after a maximum of 20xN steps.

In the KeyCorridorS3R3 environment, the agent has to pick up an object which is behind a locked door. The key is hidden in another room, and the agent has to explore the environment to find it. Episodes finish with a positive reward when the agent picks up the ball behind the locked door or after a maximum of 270 steps.

In the ObstructedMaze2Dlh environment, the agent has to pick up a box which is placed in a corner of a 3x3 maze. The doors are locked, the keys are hidden in boxes and doors are obstructed by balls. Episodes finish with a positive reward when the agent picks up the ball behind the locked door or after a maximum of 576 steps.

In the DynamicObstacles environment, the agent has to navigate to a fixed goal location while avoiding moving obstacles. In our experiments, the agent is randomly initialized to a location in the grid. If the agent collides with an obstacles, it receives a penalty of -1 and the episode ends.

A.4 VizDoom Environment

We consider the Doom 3-D navigation task where the action space of the agent consists of four discrete actions: move forward, move left, move right and no-action. Our testing setup in all the experiments is the *DoomMyWayHome-v0* environments which is available as part of OpenAI Gym (Brockman et al., 2016). Episodes are terminated either when the agent finds the vest or if the agent exceeds a maximum of 2100 time steps. The map consists of 9 rooms connected by corridors and the agent is tasked to reach some fixed goal location from its spawning location. The agent is always spawned in Room-13 which is 270 steps away from the goal under an optimal policy. A long sequence of action sis required to reach the goals from these rooms, making this setting a hard-exploration problem. The agent is only provided a sparse terminal reward of +1 if it finds the vest and 0 otherwise. While this environment has sparse reward, it is not procedurally generated, so the agent finds itself in exactly the same environment in each episode and does not need to generalize to different environment instantiations. This environment is identical to the "sparse" setting used in Pathak et al. (2017).

A.5 Ablations

In this section, we aim to better understand the effect of using episodic discounting as part of the intrinsic reward, as well as that of using entropy regularization as part of the IMPALA loss.

Figure 9 compares the performance of our model on different MiniGrid tasks with that of three ablations. The first one only uses episodic state counts as exploration bonus without multiplying it by the impact-driven intrinsic reward (*OnlyEpisodicCounts*), the second one only uses the impact-driven exploration bonus without multiplying it by the episodic state count term (*NoEpisodicCounts*), while the third one is the *NoEpisodicCounts* model without the entropy regularization term in the IMPALA loss (*NoEntropyNoEpisodicCounts*).

OnlyEpisodicCounts does not solve any of the tasks. *NoEntropyNoEpisodicCounts* either converges to a suboptimal policy or completely fails. In contrast, *NoEpisodicCounts* can solve the easier tasks but it requires more interactions than RIDE and fails to learn on the hardest domain. During training, *NoEpisodicCounts* can get stuck cycling between two states (with a large distance in the embedding states) but due to entropy regularization, it can sometimes escape such local optima

(unlike *NoEntropyNoEpisodicCounts*) if it finds extrinsic reward. However, when the reward is too sparse, *NoEpisodicCounts* is insufficient while RIDE still succeeds, indicating the effectiveness of augmenting the impact-driven intrinsic reward with the episodic count term.

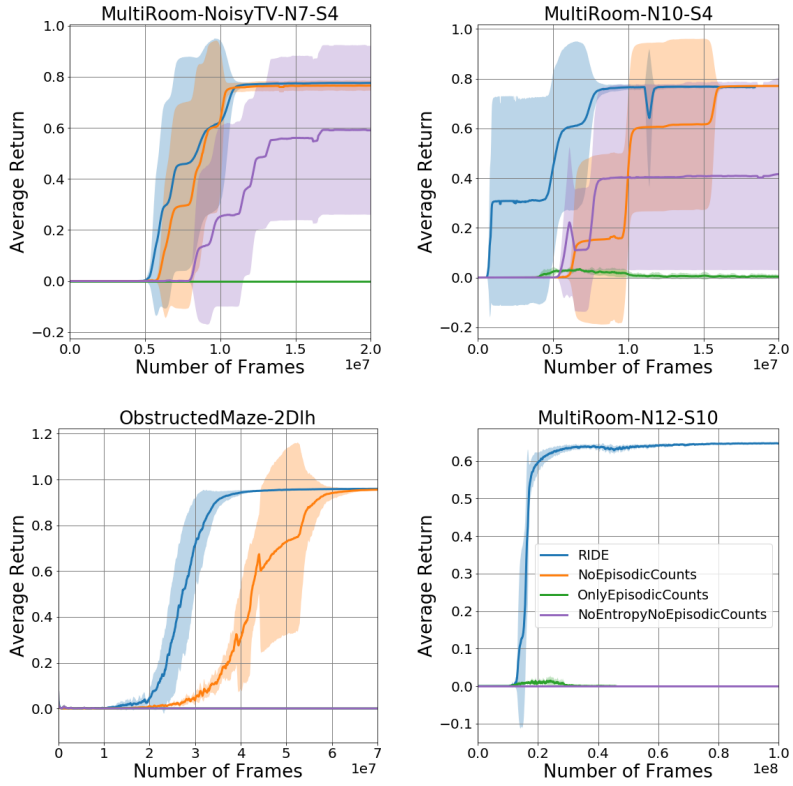

Figure 9: Comparison between the performance of RIDE and three ablations: *OnlyEpisodicCounts*, *NoEpisodicCounts*, and *NoEntropyNoEpisodicCounts*.

Figure 10 shows the average number of states visited during an episode of MultiRoomN12S10, measured at different training stages for our full RIDE model and the *NoEpisodicCounts* ablation. While the *NoEpisodicCounts* ablation always visits a low number of different states each episode ($\lesssim 10$), RIDE visits an increasing number of states throughout training (converging to ~ 100 for an optimal policy). Hence, it can be inferred that *NoEpisodicCounts* revisits some of the states. This claim can be further verified by visualizing the agents' behaviors. After training, *NoEpisodicCounts* goes back and forth between two states, while RIDE visits each state once on its path to the goal. Consistent with our intuition, discounting the intrinsic reward by the episodic state-count term does help to avoid this failure mode.

A.6 Analysis

A.6.1 State Visitation in MultiRoomN12S10

In this section, we analyze the behavior learned by the agents. Figure 11 shows the state visitation heatmaps for all models trained on 100m frames of MultiRoomN12S10, which has a very sparse reward. Note that while our model has already reached the goal in the farthest room of the maze, Count has explored about half of the maze, while RND and ICM are still in the first two rooms.

A.6.2 Intrinsic Reward in MultiRoomN12S20

Figure 12 shows a heatmap of the intrinsic reward received by RIDE, RND, and ICM agents trained on the procedurally-generated MultiRoomN12S10 environment. Table 3 shows the corresponding intrinsic rewards received for each type of action, averaged over 100 episodes, for the trained models. This environment is very challenging since the chance of randomly stumbling upon extrinsic reward

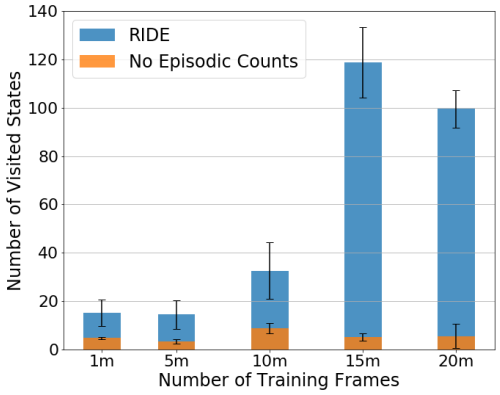

Figure 10: Average number of states visited during an episode of MultiRoomN12S10, measured at different training stages for our full RIDE model (blue) and the *NoEpisodicCounts* ablation (orange).

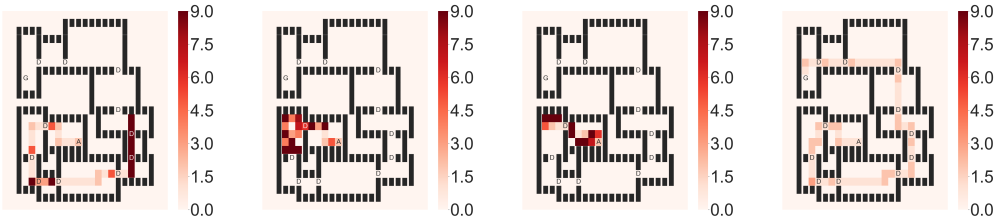

Figure 11: State visitation heatmaps for Count, RND, ICM, and RIDE (from left to right) trained for 100m frames on MultiRoomN12S10.

Figure 12: Intrinsic reward heatmaps for RND, ICM, and RIDE (from left to right) on MultiRoomN12S10.

	Open Door		Turn Left / Right		Move Forward	
Model	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
RIDE	0.0116	0.0011	0.0042	0.0020	0.0032	0.0016
RND	0.0041	0.0016	0.0035	0.0013	0.0034	0.0012
ICM	0.0082	0.0003	0.0074	0.0005	0.0086	0.0002

Table 3: Mean intrinsic reward per action computed over 100 episodes on a random map from MultiRoomN12S10.

is extremely low. Thus, we see that while the intrinsic reward provided by RIDE is still effective at exploring the maze and finding extrinsic reward, the exploration bonuses used by RND and ICM are less useful, leading to agents that do not go beyond the second room, even after training on 100m frames.

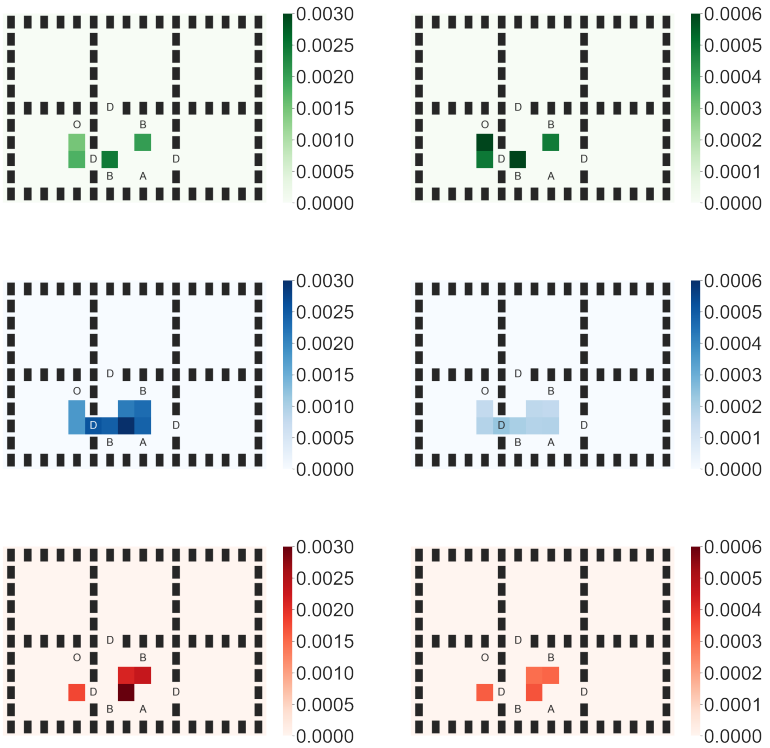

Figure 13: Intrinsic reward heatmaps for RND, and RIDE (from left to right) for interacting with objects (i.e. open doors, pick up / drop keys or balls) (green), moving forward (blue), or turning left or right (red) on a random map from ObstructedMaze2Dlh. A is the agent's starting position, K are the keys hidden inside boxes (that need to be opened in order to see their colors), D are colored doors that can only be opened by keys with the same color, and B is the ball that the agent needs to pick up in order to win the game. After passing through the door the agent also needs to drop the key in order to be able to pick up the ball since it can only hold one object at a time.

	Open Door		**Pick Ball**		**Pick Key**		**Drop Key**		**Other**	
Model	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**	**Mean**	**Std**
RIDE	0.0005	0.0002	0.0004	0.0001	0.0004	0.00001	0.0004	0.00007	0.0003	0.00001
RND	0.0034	0.0015	0.0027	0.0006	0.0026	0.0060	0.0030	0.0010	0.0025	0.0006

Table 4: Mean intrinsic reward per action computed over 100 episodes on a random map from ObstructedMaze2Dlh.

A.6.3 Intrinsic Reward in ObstructedMaze2Dlh

In order to understand how various interactions with objects are rewarded by the different exploration methods, we also looked at the intrinsic reward in the ObstructedMaze2Dlh environment which contains multiple objects . However, the rooms are connected by locked doors and the keys for unlocking the doors are hidden inside boxes. The agent does not know in which room the ball is located and it needs the color of the key to match that of the door in order to open it. Moreover, the agent cannot hold more than one object so it needs to drop one in order to pick up another.

Figure 13 and Table 4 indicate that RIDE rewards the agent significantly for interacting with various objects (e.g. opening the box, picking up the key, opening the door, dropping the key, picking up the ball) relative to other actions such as moving forward or turning left and right. In contrast, RND again rewards all actions much more uniformly and often times, within an episode, it rewards the interactions with objects less than the ones for moving around inside the maze.

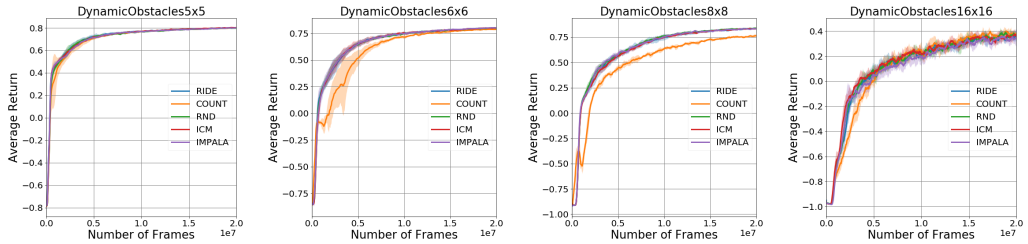

Figure 14: Performance on DynamicObstacles with varying degrees of difficulty.

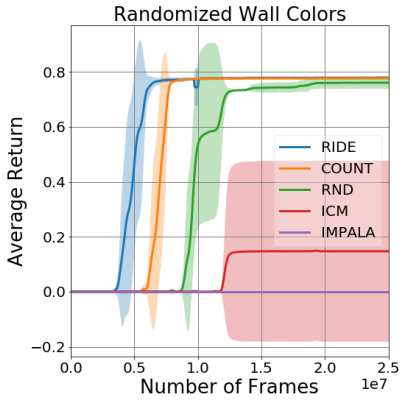

Figure 15: Performance on a version of the MiniGridRoomN7S4 in which the colors of the walls and goals are randomly picked from a set of 4 colors at the beginning of each episode.

A.7 Dynamic Obstacles Environment

One potential limitation of RIDE is that it may be drawn to take actions that significantly change the environment, even when those actions are undesirable. In order to test the limits of RIDE, we ran experiments on the DynamicObstacles environment in MiniGrid. As seen in Figure 14, RIDE learns to solve the task of avoiding the moving obstacles in the environment, even if chasing them provides large intrinsic rewards. Hence, RIDE is still able to learn effectively in certain scenarios in which high-impact actions are detrimental to solving the task.

A.8 Generalization to Unseen Colors

We also ran experiments on a version of MultiRoomN7S4 in which the colors of the walls and the goal change at each episode. The models are trained on a set of 4 colors and tested on a held-out set of 2 colors. As seen in Figure 15 and Table 5, RIDE and Count learn to solve this task and can generalize to unseen colors at test time without any extra fine-tuning. RND and ICM perform slightly worse on the test environments, and only one out of five seeds of ICM converges to the optimal policy on the train environments. The best seed for each model was used to evaluate on the test set.

A.9 Other Practical Insights

While developing this work, we also experimented with a few other variations of RIDE that did not work. First, we tried to use observations instead of learned state embeddings for computing the RIDE reward, but this was not able to solve any of the tasks. Using a common state representation for both the policy and the embeddings also fell short.

	Test Return	
Model	**Mean**	**Std**
RIDE	0.77	0.02
Count	0.77	0.02
RND	0.76	0.11
ICM	0.73	0.03
IMPALA	0.00	0.00

Table 5: Average return over 100 episodes on a version of MultiRoomN7S4 in which the colors of the walls and goals change with each episode. The models were trained until convergence on a set of 4 colors and tested on a held-out set of 2 colors.