EXPLORATION VIA FLOW-BASED INTRINSIC REWARDS

Hsuan-Kung Yang*, Po-Han Chiang*, Min-Fong Horng, and Chun-Yi Lee
Elsa Lab, Department of Computer Science
National Tsing Hua University
{hellochick,ymmoy999,romulus,cylee}@gapp.nthu.edu.tw

ABSTRACT

Exploration bonuses derived from the novelty of observations in an environment have become a popular approach to motivate exploration for reinforcement learning (RL) agents in the past few years. Recent methods such as curiosity-driven exploration usually estimate the novelty of new observations by the prediction errors of their system dynamics models. In this paper, we introduce the concept of optical flow estimation from the field of computer vision to the RL domain, and utilize the errors from optical flow estimation to evaluate the novelty of new observations. We introduce a flow-based intrinsic curiosity module (FICM) capable of learning the motion features and understanding the observations in a more comprehensive and efficient fashion. We evaluate our method and compare it with a number of baselines on several benchmark environments, including Atari games, Super Mario Bros, and VizDoom. Our results show that the proposed method is superior to the baseline in certain environments, especially for those featuring sophisticated moving patterns or with high-dimensional observation spaces. We further analyze the hyper-parameters used in the training phase, and discuss our insights into them.

1 INTRODUCTION

Reinforcement learning (RL) algorithms are aimed at developing the policy of an agent to maximize the cumulative rewards collected in an environment, and have gained considerable attention in a wide range of application domains, such as game playing (Mnih et al., 2015; Silver et al., 2016) and robot navigation (Zhang et al., 2016). In spite of their recent successes, however, one of the key constraints of them is the requirement of sufficiently dense reward signals. In environments where the reward signals are spare, it becomes extremely challenging for an agent to explore and learn an useful policy. Although simple heuristics such as ϵ-greedy (Sutton & Barto, 1998; Mnih et al., 2015) or entropy regularization (Mnih et al., 2016) were proposed, they are still far from satisfactory in such environments.

Researchers in recent years have attempted to deal with the challenge by providing an agent with exploration bonuses (also known as "intrinsic rewards") whenever an unfamiliar state or unexpected observation is encountered. These bonus rewards are provided by novelty measurement strategies to encourage the agent to explore those states with intrinsic motivation. A number of such strategies have been proposed in the past few years, such as the use of information gain (Houthooft et al., 2016), counting table (Tang et al., 2017), and prediction errors of system dynamics models (Stadie et al., 2015; Pathak et al., 2017; Burda et al., 2019). Among these approaches, curiosity-driven exploration (Pathak et al., 2017; Burda et al., 2019) has been recognized effective in several tasks which demand extensive exploration for the sparsely distributed reward signals. It introduces a forward dynamics model for predicting the next state feature embedding from the current state embedding and the action taken by the agent. The discrepancy between the predicted embedding and the actual next state embedding serves as the curiosity-based intrinsic reward. Although the use of the forward dynamics model is sufficient for novelty measurement for low-dimensional observations, however, it becomes difficult for it to perform such evaluation for high-dimensional inputs. In has been widely recognized that performing next frame or next embedding prediction typically requires complex feature representations (Kingma & Welling, 2014; Goodfellow et al., 2014; Mirza & Osindero, 2014; Lotter et al., 2017; Xue et al., 2016). This prohibits the forward model from guiding the agent to explore the environment efficiently.

*Equal contribution

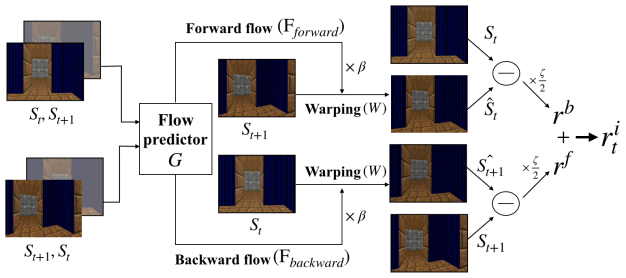


Figure 1: The workflow of the proposed flow-based intrinsic module (FICM).

Instead of directly predicting the exact feature embeddings of next observations as proposed in Pathak et al. (2017); Burda et al. (2019), in this paper we introduce a flow-based intrinsic curiosity module, called FICM, for evaluating the novelty of observations. FICM generates intrinsic rewards based on the prediction errors of optical flow estimation. The optical flow predictor in FICM is trained to extract displacement features of objects between consecutive observations, such that the relatively shallow model is enforced to learn high-level motion features in a more comprehensive and efficient manner. One of the key advantages of FICM is that the feature embeddings extracted by its flow predictor are inherently compact and sufficient, as the training procedure of FICM necessitates efficient encoding of input observations as well as indifference to irrelevant features, enabling the encoded feature embeddings to concentrate on important high-level information. Moreover, the stability of intrinsic rewards is ensured by our training method tailored for FICM discussed in Section 3.3. We validate the performance of FICM in a variety of benchmark environments, including Atari 2600 (Bellemare et al., 2013) and ViZDoom (Wydmuch et al., 2018), and demonstrate that FICM is superior to a number of baseline novelty estimation methods in terms of the learning efficiency and model complexity of the agent in several tasks, especially for those featuring sophisticated moving patterns or with high-dimensional observation spaces. The contributions of this paper are thus summarized as follows:

- We propose a new intrinsic reward module, called FICM, for evaluating the novelty of states based on the prediction errors of optical flow estimation between consecutive observations.
- We employ the mean squared error (MSE) between the warped observation and the ground truth observation to serve directly as the intrinsic reward signal. The straightforward implementation allows FICM to consist of only a single model instead of two (i.e., the feature extractor and the forward dynamics model in Pathak et al. (2017); Burda et al. (2019).
- We eliminate the requirement of the actions of the agent when estimating the novelty of states. FICM demands only two consecutive frames (i.e., observations) as its input, significantly more efficient than that of the forward dynamics model (which requires eight input frames).

The rest of this paper is organized as follows. Section 2 presents the proposed framework. Section 3 demonstrate the experimental results and discusses their implications. Section 4 concludes the paper.

2 Methodology

In this section, we present the design and implementation details of our methodology. We first provide an introduction to the concepts of the proposed flow-based curiosity driven exploration. Then, we formulate these concepts into mathematical equations, and discuss our training objectives. Finally, we explore two different implementations of FICM, and discuss the features and advantages of them.

2.1 Flow-Based Curiosity Driven Exploration

We propose to embrace optical flow estimation (Ilg et al., 2017; Meister et al., 2018), a popular technique commonly used in the field of computer vision for interpreting displacement of objects in consecutive frames, as our novelty measurement scheme. Fig. 1 illustrates the workflow of the proposed FICM. FICM takes two consecutive observations as its input, and predicts a forward flow $F_{forward}$ and a backward flow $F_{backward}$ from the pair of its input observations. The forward flow $F_{forward}$ is the optical flow inferenced from the consecutive observations ordered in time (i.e., t to $t+1$), while the backward flow $F_{backward}$ is the optical flow inferenced from the same observations but in the opposite direction (i.e., $t+1$ to t). The input observations S_t and S_{t+1} are then warped by the flows to generate the predicted observations $\hat{S}_t$ and $\hat{S_{t+1}}$. The losses of these predicted observations then serve as the partial intrinsic reward signals r^b and r^f, respectively. The sum of r^f and r^b forms the final intrinsic reward r^i presented to the RL agent. Based on the framework, FICM

(a) FICM-S (b) FICM-C

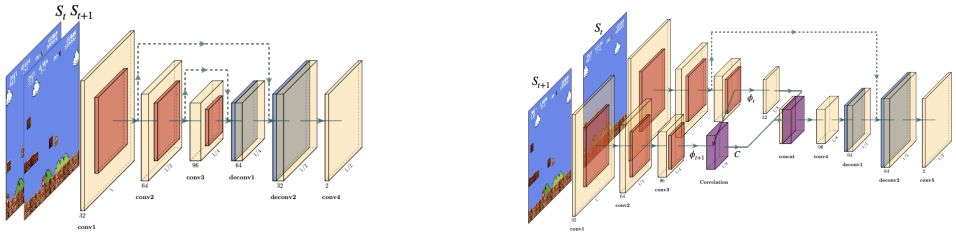


Figure 2: The flow predictor architectures in FICM-S and FICM-C.

yields higher intrinsic rewards when the agent encounters unfamiliar pairs of observations. It then motivates the agent to revisit those observations, and gradually learns the features of them over time.

2.2 Flow-Based Intrinsic Curiosity Module (FICM)

In this section, we formulate the procedure of FICM as formal mathematical equations. The main objective of FICM is to leverage the optical flow between two consecutive observations as the encoded representation of them. Given two raw input observations S_t and S_{t+1} observed at consecutive timesteps t and $t+1$, FICM takes the 2-tuple (S_t, S_{t+1}) as its input, and predicts a forward flow $F_{forward}$ and a backward flow $F_{backward}$ by its flow predictor G parameterized by a set of trainable parameters Θ_f. The two flows $F_{forward}$ and $F_{backward}$ can therefore be expressed as the following:

$$\begin{aligned} \mathrm{F_{forward}} &= G(\mathrm{S_t}, \mathrm{S_{t+1}}, \Theta_f) \\ \mathrm{F_{backward}} &= G(\mathrm{S_{t+1}}, \mathrm{S_t}, \Theta_f). \end{aligned} \tag{1}$$

$F_{forward}$ and $F_{backward}$ are then used to generate the predicted observations $\hat{S}_t$ and $\hat{S_{t+1}}$ via a warping function $W(*)$ defined in Ilg et al. (2017). The predicted $\hat{S}_t$ and $\hat{S_{t+1}}$ are thus expressed as:

$$\begin{aligned} \hat{\mathrm{S}}_\mathrm{t} &= W(\mathrm{S_{t+1}}, \mathrm{F_{forward}}, \beta) \\ \hat{\mathrm{S}_{\mathrm{t+1}}} &= W(\mathrm{S_t}, \mathrm{F_{backward}}, \beta), \end{aligned} \tag{2}$$

where β is the flow scaling factor. $W(*)$ warps $\mathrm{S_{t+1}}$ to $\hat{S}_t$ and $\mathrm{S_t}$ to $\hat{S_{t+1}}$ via $\mathrm{F_{forward}}$ and $\mathrm{F_{backward}}$ respectively using bilinear interpolation and element-wise multiplication with β. The interested reader is referred to Fischer et al. (2015); Ilg et al. (2017) for more details of the warping algorithm. Please note that in this work, $W(*)$ employs inverse mapping instead of forward mapping to avoid the common duplication problem in flow warping (Beier & Neely, 1992). With the predicted observations $\hat{\mathrm{S}}_\mathrm{t}$ and $\hat{\mathrm{S}_{\mathrm{t+1}}}$, Θ_f is iteratively updated to minimize the loss function L_G of the flow predictor G, which consists of a forward loss L^f and a backward loss L^b. The goal of Θ_f is given by:

$$\min_{\Theta_f} L_G = \min_{\Theta_f}(L^f + L^b) = \min_{\Theta_f}(||\mathrm{S_{t+1}} - \hat{\mathrm{S}_{\mathrm{t+1}}}||^2 + ||\mathrm{S_t} - \hat{\mathrm{S}}_\mathrm{t}||^2), \tag{3}$$

where (L^f, L^b) are derived from the mean-squared error (MSE) between $(\mathrm{S_{t+1}}, \hat{S_{t+1}})$ and $(\mathrm{S_t}, \hat{S}_t)$, respectively. In this work, L_G is interpreted by FICM as a measure of novelty, and serves as an intrinsic reward signal r^i presented to the DRL agent. The expression of r^i is therefore formulated as:

$$r^i = r^f + r^b = \frac{\zeta}{2}(L^f + L^b) = \frac{\zeta}{2} L_G = \frac{\zeta}{2}(||\mathrm{S_{t+1}} - \hat{\mathrm{S}_{\mathrm{t+1}}}||^2 + ||\mathrm{S_t} - \hat{\mathrm{S}}_\mathrm{t}||^2), \tag{4}$$

where ζ is the reward scaling factor, and r^f and r^b are the forward and backward intrinsic rewards scaled from L^f and L^b, respectively. Please note that r^i is independent of the action taken by the agent, which distinguishes FICM from the intrinsic curiosity module (ICM) proposed in Pathak et al. (2017). FICM only takes two consecutive input observations for estimating the prediction errors of optical flows, which serve as a more meaningful measure for evaluating and memorizing the novelty of observations in environments with high-dimensional observation spaces and sparse extrinsic reward signals. The results presented in Section 3 validate the effectiveness of r^i and FICM.

2.3 Implementations of FICM

In this work, we propose two different implementations of FICM: FICM-S and FICM-C. These implementations adopt different flow predictor architectures based on *FlowNetS* and *FlowNetC*

introduced by FlowNet 2.0 (Ilg et al., 2017), respectively. We employ different implementations to validate that Eqs. (1)-(4) are generalizable to different architectures, rather than restricted to specific predictor designs. The flow predictor architectures are depicted in Fig. 2, and are described as follows.

FICM-S. The flow predictor in FICM-S consists of several convolutional and deconvolutional layers. The module first stacks two consecutive observations S_t and S_{t+1} together. It then feeds the stacked observations $\langle S_t, S_{t+1} \rangle$ into three convolutional layers, followed by two deconvolutional layers. The encoded features are fused with the feature maps from the shallower parts of the network by adding skips (Fischer et al., 2015). This skip layer fusion architecture allows the flow predictor to preserve both coarse, high-level information and fine, low-level information. At the end, the fused feature map is passed into another convolutional layer with two filters to predict the optical flow from S_t to S_{t+1}.

FICM-C. The flow predictor in FICM-C encodes two consecutive observations S_t and S_{t+1} separately instead of stacking them together. The input observations are passed through three convolutional layers to generate feature embeddings ϕ_t and ϕ_{t+1}. The convolutional layers of the two paths are share-weighted in order to generate comparable representations of ϕ_t and ϕ_{t+1}, as input observations S_t and S_{t+1} usually contain same or similar patterns. The embeddings ϕ_t and ϕ_{t+1} are then fed into a correlation layer introduced by Fischer et al. (2015). The correlation layer performs multiplicative patch comparisons between ϕ_t and ϕ_{t+1} to estimate their correspondences c, given by:

$$c(x_1, x_2) = \sum_{o \in [-k,k] \times [-k,k]} \langle (\phi_t(x_1 + o), \phi_{t+1}(x_2 + o)) \rangle \,, \tag{5}$$

where x_1 and x_2 are the patch centers in ϕ_t and ϕ_{t+1}, respectively. Once c is estimated, it is concatenated with the feature map from the shallower part, and together fed into one convolutional and one deconvolutional layers. The output feature map is then fused with the feature map came from the skip, and forwarded to another deconvolutional layer. Similar to FICM-S, the final feature map traverses through a convolutional layer with two filters to generate the optical flow from S_t to S_{t+1}.

After estimating the optical flow from the flow predictor, the predicted optical flow is used to warp S_t and S_{t+1} forward and backward using Eq. (2), and then derive L_G and r^i based on Eqs. (3) and (4).

3 Experimental Results

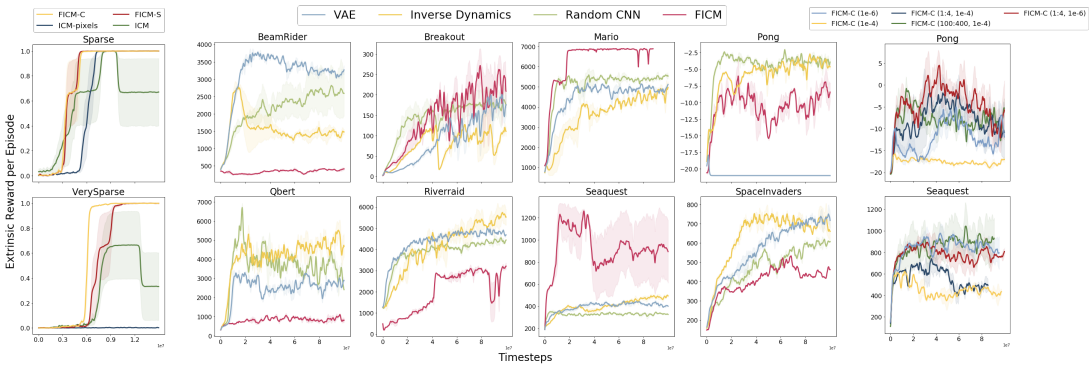
In this section, we present the experimental results on a number of environmental settings. We start by comparing the proposed methodology with the previous approaches on ViZDoom (Wydmuch et al., 2018) with only sparse and very sparse extrinsic rewards. Next, we evaluate the performance of FICM on Atari 2600 games (Bellemare et al., 2013) and Super Mario Bros., without any extrinsic reward signals. Finally, we discuss the techniques for stabilizing the training process of our FICM.

3.1 Exploration with Sparse and Very Sparse Extrinsic Rewards in High-Dimensional Observation Space

We perform experiments on the ViZDoom environment, *DoomMyWayHome-v0*, the same as those conducted in Pathak et al. (2017). In this environment, the agent is required to reach the fixed goal from its spawning location in a 9-rooms map, and only receives an extrinsic reward of +1 if it accomplishes. We adopt two setups, *sparse* and *very sparse* reward settings, to evaluate the exploration ability of an agent. The two settings are different in the distance between the initial spawning location of the agent and the fixed goal. We compare FICM with the baseline approaches presented in Pathak et al. (2017), and plot the results in Fig. 3 (a). In both the *sparse* and *very sparse* settings, our methods and the baselines are able to guide the agent to reach the goal. However, for the *very sparse* reward setting, it is observed that the baselines sometimes suffer from performance drop, and are not always able to to obtain the maximum performance. In contrast, our methods are able to converge faster than them, and maintain stable performance consistently over different initial seeds.

3.2 Exploration without Extrinsic Reward

We further perform experiments on seven different Atai games and *Super Mario Bros*. without using any extrinsic reward or end of episode signal during the training phase, the same as those conducted in Burda et al. (2019). We compare the performance of our method against the three baselines implemented with different forward dynamics models presented in Burda et al. (2019): *VAE*, *Random CNN*, and *Inverse Dynamics*, and plot the learning curves in Fig. 3 (b). It is observed that out method (denoted as *FICM-C*) significantly outperforms the baselines in games *Breakout*, *Seaquest* and *Mario*, while delivering comparable performance to the baselines in game *Pong*. These



(a) (b) (c)

Figure 3: (a) A comparison of the learning curves in the ViZDoom environment with sparse and very sparse extrinsic rewards. (b) A comparison of FICM and the baseline methods on seven selected Atari games and *Super Mario Bros.*, with no extrinsic reward. (c) A comparison of different stabilization techniques for training FICM.

games are characterized by moving objects that require the agent to concentrate on and explore from. As a result, they are favorable to the proposed methodology, as optical flow estimation is capable of learning motion features and understanding the changes in observations in a more comprehensive fashion. On the contrary, the proposed methodology does not deliver satisfactory results and perform unstably in games *BeamRider*, *Qbert*, *River Raid*, and *Space Invaders*. These are primarily due to the lack of movements of primary objects (e.g., enemies), or excessive changes of irrelevant components (e.g., backgrounds) that distract the focus of FICM. FICM generates negligible intrinsic rewards if the objects that the agent should pay attention to barely move (e.g., *Space Invaders*), since there is only little discrepancy between two consecutive frames. On the other hand, if some irrelevant parts of the environments move relatively faster than the agent (e.g., *BeamRider*), FICM may be distracted to focus on incorrect regions or components, leading to unstable performance. The above results suggest that the proposed method is preferable to the baselines in exploring certain environments.

3.3 Analysis of the Stabilization Technique

In this section, we discuss the stability issue of the proposed methodology, and analyze the techniques for dealing with it. It is observed our method suffers from poor and unstable performances in a few Atari game when the same set of hyper-parameters used for training FICM in the VizDoom environment is employed. The results of *Pong* and *Seaquest* are plotted as the yellow curves in Fig. 3 (c). We consider that the above issue is mainly caused by the imbalance of learning speed between the agent and FICM. When FICM learns faster than the agent and transfers its attention (i.e., curiosity) quickly, the intrinsic rewards generated by FICM turn into an easily consumable resource, causing the agent to fall behind and unable to explore and collect sufficient data samples to update its policy. We propose two heuristics to balance the learning speeds of FICM and the agent by adjusting the learning rate and the update period of FICM. In our experiments, we adopt two different choices of the learning rate: $1\mathrm{e}{-4}$ and $1\mathrm{e}{-6}$. In addition, we examine two different update periods of FICM, $1:4$ and $100:400$, where the notation $k1:k2$ indicates that the parameters of FICM are updated for $k1$ iterations and fixed for $k2$ iterations. Fig. 3 (c) presents an ablative analysis of the two proposed heuristics. In Fig. 3 (c), the fixed period $k2$ is assumed to be zero if not specified. It is observed that decreasing the learning rate leads to obvious overall improvement. In addition, the existence of the fixed period $k2$ also enhances the performance. According to our experiments, the combination of learning rate $1\mathrm{e}{-6}$ and update period $1:4$ tends to perform better in these two games. Further investigations into the balancing issues are left as our future research directions.

4 Conclusions

In this paper, we proposed a flow-based intrinsic curiosity module (FICM) for evaluating the novelty of observations in RL exploration. FICM employs optical flow estimation errors as a measure for generating intrinsic rewards, which allow an RL agent to explore environments featuring moving objects or with high-dimensional observation spaces in a more comprehensive and efficient manner. We validated the proposed methodology and compared it against a number of baselines on Atari games, Super Mario Bros., and ViZDoom. According to our experiments, we observed that the proposed FICM is capable of focusing on important objects, and guiding the RL agent to deliver superior performance to the baselines in certain environments. We further provided our insights into the stabilization techniques of FICM, and analyzed the results of different hyper-parameter settings.

REFERENCES

T. Beier and S. Neely. Feature-based image metamorphosis. In *Special Interest Group on Computer Graphics (SIGGRAPH)*, pp. 35–42, Jul. 1992.

M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artificial Intelligence Research (JAIR)*, 47:253–279, May 2013.

Y. Burda, H. Edwards, D. Pathak, A. J. Storkey, T. Darrell, and A. A. Efros. Large-scale study of curiosity-driven learning. In *Proc. Int. Conf. Learning Representation (ICLR)*, May 2019.

P. Fischer, A. Dosovitskiy, and E. IlgA. *et al*. FlowNet: Learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pp. 2758–2766, May 2015.

I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. W.-Farley, S. Ozair, A. C. Courville, and Y. Bengio. Generative adversarial networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2672–2680, Dec. 2014.

R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel. VIME: Variational information maximizing exploration. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 1109–1117, Dec. 2016.

E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, pp. 1647–1655, Dec. 2017.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv:1312.6114*, May 2014.

W. Lotter, G. Kreiman, and D. D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv:1605.08104*, Mar. 2017.

S. Meister, J. Hur, and S. Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Proc. of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 7251–7259, 2018.

M. Mirza and S. Osindero. Conditional generative adversarial nets. *arXiv:1411.1784*, Nov. 2014.

V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. A. Riedmiller, A. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, vol. 518, no. 7540, pp. 529-533, Feb. 2015.

V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. P. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 1928–1937, Jun. 2016.

D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proc. Int. Conf. Machine Learning (ICML)*, pp. 2778–2787, May 2017.

D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. P. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis. Mastering the game of Go with deep neural networks and tree search. *Nature*, vol. 529, no. 7587, pp. 484-489, Jan. 2016.

B. C. Stadie, S. Levine, and P. Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv:1507.00814*, Nov. 2015.

R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

H. Tang, R. Houthooft, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. D. Turck, and P. Abbeel. #Exploration: A study of count-based exploration for deep reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 2750–2759, Dec. 2017.

M. Wydmuch, M. Kempkaand, and W. Jaśkowski. ViZDoom competitions: Playing Doom from pixels. *IEEE Trans. Games*, Oct. 2018.

T. Xue, J. Wu, K. L. Bouman, and B. Freeman. Visual dynamics: Probabilistic future frame synthesis via cross convolutional networks. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pp. 91–99, Dec. 2016.

M. Zhang, Z. McCarthy, C. Finn, S. Levine, and P. Abbeel. Learning deep neural network policies with continuous memory states. In *Proc. Int. Conf. Robotics and Automation (ICRA)*, pp. 520–527, May 2016.