Exploration via Flow-Based Intrinsic Rewards

Hsuan-Kung Yang
Department of Computer Science
National Tsing Hua University
hellochick@gapp.nthu.edu.tw

Po-Han Chiang
Department of Computer Science
National Tsing Hua University
ymmy999@gapp.nthu.edu.tw

Min-Fong Hong
Department of Computer Science
National Tsing Hua University
romulus@gapp.nthu.edu.tw

Chun-Yi Lee
Department of Computer Science
National Tsing Hua University
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 baselines 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 [1,2] and robot navigation [3]. In spite of their recent successes, however, one of the key constraints of them is the requirement of dense reward signals. In environments where the reward signals are sparse, it becomes extremely challenging for an agent to explore and learn a useful policy. Although simple heuristics such as ε-greedy [4] or entropy regularization [5] 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 an 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 [6], counting table [7], and prediction errors of system dynamics models [8,10]. Among these approaches, curiosity-driven exploration [9] 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

*Equal contribution

Preprint. Under review.
next state feature embedding of an agent from its current state feature 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.

Next frame prediction or next state feature embedding prediction for complex or rapid-changing observations, however, is rather difficult for the forward dynamics model, especially when the prediction is made solely based on the current state and the taken action. It has been widely recognized that performing next frame or next embedding prediction typically requires complex feature representations [11–15]. In addition, rapid changes or moving patterns in two consecutive observations essentially serve as important signals to motivate an agent for exploration. As a result, 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. Observations with low prediction errors (i.e., low intrinsic rewards) indicate that the agent has seen the observations plenty of times. On the contrary, observations are considered seldom visited when the errors from the predicted flow are non-negligible. The latter case then prompts FICM to generate high intrinsic rewards to encourage the agent for exploration. Fig. 1 provides an illustrative example for FICM. Fig. 1 (a) and 1 (b) are two consecutive frames from Super Mario Bros., Fig. 1 (c) visualizes the prediction errors (i.e., flow loss) of FICM where the brighter parts represent areas with higher prediction errors, and Fig. 1 (d) depicts the attention areas [16] of the agent. It can be observed that the bright parts in Fig. 1 (c) align with the attention areas in Fig. 1 (d), implying the significance of motion features in intrinsically motivated exploration.

We validate the performance of FICM in a variety of benchmark environments, including ViZDoom [17], Super Mario Bros., and Atari 2600 [18]. We demonstrate that FICM is preferable to the baselines in terms of game mastering and learning efficiency of the agent in several tasks and environments, especially for those featuring sophisticated moving patterns. Additionally, we analyze the encoding efficiency and different implementations of FICM in a number of experiments quantitatively and qualitatively. We further provide a comprehensive set of ablation analysis for FICM. The contributions of this paper are thus summarized as follows:

- We propose a new flow-based intrinsic curiosity module, called FICM, which leverages on existing methods of optical flow estimation methods from the field of computer vision (CV) to evaluate the novelty of observations for complex or rapid-changing observations.
- The proposed FICM encourages an RL agent to learn the motion features and understand the observations from an environment in a more comprehensive manner.
- The proposed FICM is able to encode high dimensional inputs (e.g. RGB frames) and utilize the information more effectively and efficiently than the baseline approaches.
- Moreover, the proposed FICM requires only two consecutive frames to obtain sufficient information when estimating the novelty of observations, while the forward dynamics model in [9] requires eight frames and an additional action input.

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. For more details of the background materials, implementation details, source codes, and our demo video, please refer to our supplementary materials.
2 Methodology

In this section, we present the design and implementation overview 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 [19, 20], 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. 2 illustrates the workflow of the proposed FICM. FICM takes two consecutive observations as its input, and predicts a forward flow $F_{\text{forward}}$ and a backward flow $F_{\text{backward}}$ from the pair of its input observations. The forward flow $F_{\text{forward}}$ is the optical flow inferenced from the consecutive observations ordered in time (i.e., $t$ to $t + 1$), while the backward flow $F_{\text{backward}}$ 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 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 $F_{\text{forward}}$ and a backward flow $F_{\text{backward}}$ by its flow predictor $G$ parameterized by a set of trainable parameters $\Theta_f$. The two flows $F_{\text{forward}}$ and $F_{\text{backward}}$ can therefore be expressed as the following:

$$
F_{\text{forward}} = G(S_t, S_{t+1}, \Theta_f) \\
F_{\text{backward}} = G(S_{t+1}, S_t, \Theta_f).
$$

(1)

$F_{\text{forward}}$ and $F_{\text{backward}}$ are then used to generate the predicted observations $\hat{S}_t$ and $\hat{S}_{t+1}$ via a warping function $W(\cdot)$ defined in [19][21]. The predicted $\hat{S}_t$ and $S_{t+1}$ are thus expressed as:

$$
\hat{S}_t = W(S_{t+1}, F_{\text{forward}}, \beta) \\
S_{t+1} = W(S_t, F_{\text{backward}}, \beta),
$$

(2)
where $\beta$ is the flow scaling factor. $W(\cdot)$ warps $S_{t+1}$ to $\hat{S}_t$ and $S_t$ to $\hat{S}_{t+1}$ via $F_{\text{forward}}$ and $F_{\text{backward}}$ respectively using bilinear interpolation and element-wise multiplication with $\beta$. The interested reader is referred to [19, 21] for more details of the warping algorithm. Please note that in this work, $W(\cdot)$ employs inverse mapping instead of forward mapping to avoid the common duplication problem in flow warping [22].

With the predicted observations $\hat{S}_t$ and $\hat{S}_{t+1}$, $\Theta_f$ is iteratively updated to minimize the flow loss function $L_G$ of the flow predictor $G$, which consists of a forward flow loss $L^f$ and a backward flow loss $L^b$. The goal of $\Theta_f$ is given by:

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

where $(L^f, L^b)$ are derived from the mean-squared error (MSE) between $(S_{t+1}, \hat{S}_{t+1})$ and $(S_t, \hat{S}_t)$, 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:

$$
\begin{align*}
    r^i &= r^f + r^b \\
    &= \zeta (L^f + L^b) \\
    &= \frac{\zeta}{2} L_G = \frac{\zeta}{2} (||S_{t+1} - \hat{S}_{t+1}||^2 + ||S_t - \hat{S}_t||^2),
\end{align*}
$$

where $\zeta$ 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 $r^b$ is independent of the action taken by the agent, which distinguishes FICM from the intrinsic curiosity module (ICM) proposed in [9]. 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 two implementations adopt different flow predictor architectures based on FlowNetS and FlowNetC introduced by FlowNet 2.0 [19], respectively. We employ different implementations to validate that Eqs. (1)-(4) are generalizable to different architectures, rather than restricted to any specific predictor designs. The flow predictor architectures of FICM-S and FICM-C are depicted in Fig. 3 (a) and Fig. 3 (b), respectively. They both consist of a number of convolutional and de-convolutional layers for predicting the optical flow between the two input observations. The primary difference between the two predictors lies in their feature extraction strategy from the input observations $S_t$ and $S_{t+1}$. The flow predictor of FICM-S encodes the stacked observations $(S_t, S_{t+1})$ together to generate a single feature embedding $\phi$, while the other one encodes $S_t$ and $S_{t+1}$ in different paths to generate two separate feature embeddings $\phi_t$ and $\phi_{t+1}$. In order to preserve both coarse, high-level information and fine, low-level information for enhancing the flow prediction accuracy, the feature embeddings in the two predictor designs are later fused with the feature maps from their shallower parts of the
networks by skips [21]. The fused feature map is then processed by another convolutional layer at the end with two filters to predict the optical flow from \( S_t \) to \( S_{t+1} \). Please note that the two input paths of Fig.3(b) are share-weighted in order to generate comparable embeddings. The predictor of FICM-C additionally employs a correlation layer [21] to perform multiplicative patch comparisons between \( \phi_t \) and \( \phi_{t+1} \). The detailed configurations of the predictors are provided in our supplementary materials.

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 [17] with only sparse and very sparse extrinsic rewards. Next, we evaluate the performance of FICM on a number of Atari 2600 games [18] and Super Mario Bros., without any extrinsic reward signals. Then, we conduct experiments to examine the encoding efficiency of FICM. Finally, we present the ablation set of analysis of FICM from three different aspects. We further provide the training details and the setting of hyper-parameters in the supplementary materials.

3.1 Experiments on Exploration with Sparse and Very Sparse Extrinsic Rewards

Environments. The environment we evaluate on is the ViZDoom [17] game. ViZDoom is a popular Doom-based platform for AI research especially reinforcement learning to test agents’ ability to process raw visual information. We conduct experiments on the gaming environment, DoomMyWayHome-v0, the same as those conducted in [9]. In this environment, the agent is required to reach the fixed goal from certain spawning location in a nine-rooms map, and only receives an extrinsic reward of ‘+1’ if it accomplishes the task. 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 goal. The farther the goal is from the spawning location, the harder the map is for the agent to explore.

Baseline approaches. We compare FICM with two baselines, ‘ICM + A3C’ (denoted as ICM) and ‘ICM-pixels + A3C’ (denoted as ICM-pixels) which are both proposed in [9]. These baselines combines intrinsic curiosity module with training algorithms A3C [23]. ICM-pixels is close to ICM in architecture except without the inverse dynamics model, and it computes curiosity reward only dependent on forward model loss in next frame prediction.

Results. We analyze and compare the evaluation curves of our proposed methods (denoted as FICM-C and FICM-S) with those of the baseline methods. The experimental results on the ViZDoom environment are depicted in Fig. 4 (a). In the sparse reward setting, our methods and the baselines are able to guide the agent to reach the goal. However, in the very sparse reward setting, it is observed that ICM sometimes suffers from performance drop and is not always able to obtain the maximum performance, while ICM-pixels even fails to reach the goal in this setting. In contrast, our methods are able to converge faster than the baselines and maintain stable performance consistently in both sparse and very sparse reward settings over different initial seeds.

3.2 Experiments on Exploration with no Extrinsic Reward

Environments. We further perform experiments on Super Mario Bros. and seven different Atari games, including BeamRider, Breakout, Pong, Qbert, RiverRaid, SeaQuest, and SpaceInvaders. During the training phase, the agents are not provided with any extrinsic reward or end-of-episode signal, the same as the experiments conducted in [10]. The setting is different from that of Section 3.1.

Baseline approaches. We compare the performance of our method against the three baselines implemented in [10]: VAE, Random CNN, and Inverse Dynamics. These three baselines are different in their feature encoding methods while sharing the same design of the forward dynamics models.

Results. We plot the evaluation curves of our method (denoted as FICM-C) and the baselines in different environments in Fig. 4(b) during the training phase. The curves illustrate the mean reward (with standard error) of the agents trained purely by intrinsic rewards. It is observed that our method significantly outperforms the baselines in games Super Mario Bros., Breakout, and SeaQuest, while delivering comparable performance to the baselines in games Pong, RiverRaid, and SpaceInvaders.
Figure 4: Comparison of the evaluation curves in (a) ViZDoom with sparse and very sparse extrinsic rewards, and (b) seven selected Atari games and Super Mario Bros. with no extrinsic reward.

Figure 5: Visualization of flow loss and attention areas at different timesteps in (a) Seaquest & (b) BeamRider.

Figure 6: (a) RGB versus gray-scale frames, and (b) stacked versus non-stacked frames.

These games are characterized by moving objects that require the agents to concentrate on and interact with. As a result, they are favorable to the proposed exploration method, as the optical flow estimator is capable of capturing motion features and perceiving the changes in observations in a more comprehensive fashion. On the contrary, our method does not deliver satisfactory results and perform poorly in games BeamRider and Qbert. This is primarily due to an insufficiency of movements of the major objects (e.g., enemies), or excessive changes of irrelevant components (e.g., backgrounds) that distract the focus of FICM. If the objects that the agent should pay attention to barely move, FICM generates negligible intrinsic rewards 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 crucial objects (e.g., BeamRider), FICM may be distracted to focus on incorrect regions or components, leading to unsatisfactory performance. The above results suggest that the proposed method is preferable to the baselines in exploring certain environments.

Since there is no extrinsic reward provided to the agent, which is different from the experiments conducted in Section 3.1. The fluctuations and sudden drops in the evaluation curves in Fig. 4(b) (e.g., SeaQuest) only reflect the exploration processes of the agents. At different timesteps, the agent may focus its attention on exploring different states, leading to fluctuations in the evaluation curve. High scores in the curves indicate that the agents are able to explore hard-to-visit states (which usually associated with high extrinsic rewards) at certain timesteps. We further illustrate the best extrinsic return curves in our supplementary materials.

We further illustrate how FICM guides the agent to explore the environments by visualizing the flow loss and the attention areas [15] of the agents during training phase in Fig. 5. Figs. 5(a) and 5(b) present visualizations of SeaQuest and BeamRider, respectively. The first two columns depict two
consecutive frames $S_t$ and $S_{t+1}$. The third column shows the flow loss of the two frames, where the brighter parts reflect higher flow loss. The remaining three columns highlight the attention areas of the agents in red at 5M, 10M, and 15M timesteps. It is observed that the flow loss is highly correlated to the attention areas of the agents at different timesteps for both games. The evidence reveals that FICM is indeed able to guide the agents to concentrate on the regions with high flow loss during the training phase, as the attention areas grow wider and becomes brighter at those regions in later timesteps. However, flow loss may as well distract the agent to focus on irrelevant parts of the environments. Take BeamRider for example, the agent is misled to concentrate on the background by the rich flow loss in it, resulting in its poor performance in Fig. 4 (b). On the other hand, the agent is able to master SeaQuest as it concentrates on the enemies and is not distracted by irrelevant objects. The visualization validates that FICM is able to guide an agent to focus on moving objects and encourage exploration in an environment.

3.3 Analysis of the Feature Encoding Ability

In this section, we investigate the feature encoding abilities of our method and the Random CNN baseline introduced in Section 3.2. We compare their efficiency of extracting information from the input frames in terms of (1) the dimensionality of the frames and (2) the number of stacked frames.

**RGB versus gray-scale frames.** We hypothesize that RGB frames contain more useful information than gray-scale ones to be utilized by the curiosity modules. In order to validate this assumption, in the first experiment, we eliminate the procedure of RGB to gray-scale conversion on the input frames when generating the intrinsic rewards. For a fair comparison, the RL agents still take gray-scale frames as their inputs. We examine whether FICM and the baseline are able to make use of the extra information to guide the agent for exploration. The experiments are evaluated on Super Mario Bros., and the results are plotted in Fig. 6 (a). We present the results with and without RGB to gray-scale conversion as two different settings for both FICM and the baseline. According to the figure, the baseline performs nearly the same for both settings. In contrast, FICM using RGB frames outperforms the one using only gray-scale frames, indicating that FICM can encode the features more efficiently and utilize the information contained in RGB channels. The superior performance is mainly due to the fact that the optical flow estimator is more capable of perceiving the motions of objects with color information and predict optical flow more accurately. The accurately predicted flow allows FICM to warp the observation more precisely. Moreover, the RGB information does not impair the ability of the warping function, as the displacement vectors \[21\] are assigned to each dimension of the current observation. On the contrary, the baseline performs next embedding prediction solely based on the encoded state and the taken action. Fig. 6 (a) reveals its inefficiency in encoding RGB information.

**Stacked versus non-stacked frames.** We conduct another experiment to further investigate the encoding efficiency of FICM and the baseline. In addition to the existing setups of FICM with two consecutive input frames (denoted as FICM-C (non-stacked)) and the baseline with a pair of stacked frames (four frames in a stack, denoted as Random CNN (stacked)), we introduce two extra settings of our method using a pair of stacked frames (denoted as FICM-C (stacked)) and the baseline using two consecutive frames (denoted as Random CNN (non-stacked)). The comparison results are plotted in Fig. 6 (b). The results show that both FICM-C (stacked) and FICM-C (non-stacked) outperform the baseline counterparts. Furthermore, it is observed that the evaluation curve of Random CNN (non-stacked) exhibits high variance highlighted as the wide shaded area of the evaluation curve, indicating that Random CNN (non-stacked) suffers from unstable performance due to its poor encoding efficiency. This observation again proves that FICM is able to encode the input frames more efficiently than the baseline. Moreover, since there is no significant difference in the evaluation curves of FICM-C (stacked) and FICM-C (non-stacked), we suggest that FICM only requires non-stacked frames, instead of stacked frames, to generate intrinsic rewards in our proposed method.

3.4 Ablation Analysis

In this section, we present a set of ablation analysis to validate the proposed FICM in terms of three different aspects: flow predictor architectures, learning rates, and flow loss.

**FICM-S versus FICM-C.** We employ different implementations to validate the design of FICM. We first compare the evaluation curves of our method using FICM-S and FICM-C, and plot the results
Figure 7: (a) Comparison of Flow-S and Flow-C, (b) learning rate analysis in Super Mario Bros. and SeaQuest, and (c) flow loss versus training iteration for selected states.

in Fig. 7(a). The setting is the same as that discussed in Section 3.2. It is observed that the two evaluation curves exhibit similar trends. We observe that there is a performance gap between the two curves in SeaQuest. This discrepancy is primarily caused by the difference in learning speeds of the two FICM architectures. This difference then leads to different learning speeds of the agents as well as the exploration efficiency. This is further discussed in the next paragraph. We thus conclude that FICM is generalizable to different architectures, rather than restricted to any specific architectures.

Learning rate analysis. We notice that when the same set of hyper-parameters used for training FICM in the VizDoom environment is employed in a few Atari games and Super Mario Bros., our method suffers from poor and unstable performance. We consider that the above issue is mainly caused by the imbalance of learning speeds 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 a heuristic to balance the learning speeds of the agent and FICM by adjusting the learning rate of FICM. In our experiments, we adopt two different choices of the learning rate: $1e^{-4}$ and $1e^{-6}$. Fig. 7(b) presents an analysis of the proposed heuristic. It is observed that decreasing the learning rate leads to an obvious overall improvement. According to our experiments, a learning rate of $1e^{-6}$ tends to perform better in these games, and is therefore adopted for FICM in all of our experiments in Sections 3.2, 3.3, and 3.4. Please note that the learning rate of the agent remains $1e^{-4}$ in these sections.

Flow loss versus training iteration. In curiosity-driven exploration, intrinsic rewards motivating the agent to explore novel states are supposed to decrease by time and eventually diminish to low values near zero. As the agent visits more states through the training phase, it becomes more familiar with the environment. This leads to a decrease in the generated intrinsic rewards. We plot the magnitude of flow loss for some selected states from Super Mario Bros. versus training iterations in Fig. 7(c). In the figure, the x-axis represents the training iterations, the y-axis denotes the selected states, and the z-axis corresponds to the magnitude of the flow loss. It is observed that the flow loss starts with high magnitudes at the beginning of the training phase, and gradually declines to low values near zero for each selected state after millions of iterations. It is this property that qualifies flow loss as a promising candidate to serve as an intrinsic reward signal. Please refer to our supplementary materials for another illustrative example of this property.

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 (i.e., RGB inputs) in a more comprehensive and efficient manner. We demonstrated 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 additionally analyzed the encoding efficiency of FICM, and provided our insights from different aspects. Moreover, we presented a comprehensive set of ablation analysis, and validated the effectiveness of the proposed
FICM in terms of its architectural implementations, flow losses, and learning rates. We conclude that FICM is a promising scheme for encouraging RL exploration, especially for complex and rapid-changing observations.

References

[1] 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.

[2] 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.

[3] 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)*, pages 520–527, May 2016.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. *arXiv:1312.5602*, Dec. 2013.

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

[6] 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)*, pages 1109–1117, Dec. 2016.

[7] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1471–1479, Dec. 2016.

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

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

[10] 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.

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

[12] 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)*, pages 2672–2680, Dec. 2014.

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

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

[15] 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)*, pages 91–99, Dec. 2016.

[16] S. Greydanus, A. Koul, J. Dodge, and A. Fern. Visualizing and understanding atari agents. In *Int. Conf. Machine Learning (ICML)*, pages 1787–1796, Jun. 2018.
[17] M. Wydmuch, M. Kempkaand, and W. Jaśkowski. ViZDoom competitions: Playing Doom from pixels. IEEE Trans. Games, Oct. 2018.

[18] 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.

[19] 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), pages 1647–1655, Dec. 2017.

[20] 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), pages 7251–7259, 2018.

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

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

[23] 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), pages 1928–1937, Jun. 2016.

References

[1] 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.

[2] 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.

[3] 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), pages 520–527, May 2016.

[4] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller. Playing atari with deep reinforcement learning. arXiv:1312.5602, Dec. 2013.

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

[6] 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), pages 1109–1117, Dec. 2016.

[7] M. G. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos. Unifying count-based exploration and intrinsic motivation. In Advances in Neural Information Processing Systems (NeurIPS), pages 1471–1479, Dec. 2016.

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

[9] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In Proc. Int. Conf. Machine Learning (ICML), pages 2778–2787, May 2017.
[10] 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.

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

[12] 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), pages 2672–2680, Dec. 2014.

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

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

[15] 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), pages 91–99, Dec. 2016.

[16] S. Greydanus, A. Koul, J. Dodge, and A. Fern. Visualizing and understanding atari agents. In Int. Conf. Machine Learning (ICML), pages 1787–1796, Jun. 2018.

[17] M. Wydmuch, M. Kempka, and W. Jaśkowski. ViZDoom competitions: Playing Doom from pixels. IEEE Trans. Games, Oct. 2018.

[18] 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.

[19] 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), pages 1647–1655, Dec. 2017.

[20] 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), pages 7251–7259, 2018.

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

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

[23] 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), pages 1928–1937, Jun. 2016.