Flow-based Intrinsic Curiosity Module

Hsuan-Kung Yang[*] , **Po-Han Chiang**[*] , **Min-Fong Hong** and **Chun-Yi Lee**

Elsa Lab, Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan

{hellochick, ymmoy999, romulus, cylee}@gapp.nthu.edu.tw

Abstract

In this paper, we focus on a prediction-based novelty estimation strategy upon the deep reinforcement learning (DRL) framework, and present a flow-based intrinsic curiosity module (FICM) to exploit the prediction errors from optical flow estimation as exploration bonuses. We propose the concept of leveraging motion features captured between consecutive observations to evaluate the novelty of observations in an environment. FICM encourages a DRL agent to explore observations with unfamiliar motion features, and requires only two consecutive frames to obtain sufficient information when estimating the novelty. We evaluate our method and compare it with a number of existing methods on multiple benchmark environments, including *Atari* games, *Super Mario Bros.*, and *ViZDoom*. We demonstrate that FICM is favorable to tasks or environments featuring moving objects, which allow FICM to utilize the motion features between consecutive observations. We further ablatively analyze the encoding efficiency of FICM, and discuss its applicable domains comprehensively. See ***here*** for our codes and demo videos.

1 Introduction

A wide range of decision learning domains based on deep reinforcement learning (DRL) usually require an agent to explore environments containing moving objects, which are common in applications such as game playing [Mnih *et al.*, 2015] and robot navigation [Zhang *et al.*, 2016]. The rapidly changing parts or moving patterns between an agent's consecutive observations can be interpreted as an important indicator of information when exploring an environment. A biological analogy is that animals often posses the capability to discover whether or not newcomers have intruded into their familiar territories. It is the motions of unfamiliar feature patterns appearing in an animal's field of view that arouses its instinct of curiosity. Inspired by such a natural instinct, we argue that exploiting such motions features during the exploration phase might potentially add positive impacts to a DRL

---

[*]Equal contribution

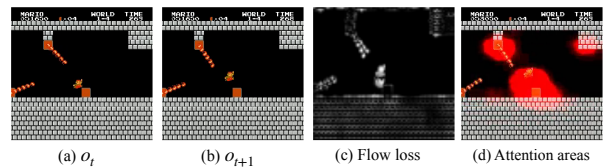(a) o_t (b) o_{t+1} (c) Flow loss   (d) Attention areas

Figure 1: Illustration of the relation between the prediction errors (i.e., flow loss) and the attention areas of an agent. (a) and (b) are two consecutive frames o_t and o_{t+1}. (c) shows the flow loss, while (d) depicts the attention areas [Greydanus *et al.*, 2018] of the agent.

agent, especially for the above-mentioned domains.

Exploration incentives leveraging motion features captured between consecutive observations have not been discussed by the DRL community. Past researchers have attempted various ways to provide an agent with exploration bonuses (also known as "*intrinsic rewards*") to encourage it to explore even when the reward signals from environments are sparse. These bonus rewards are associated with state novelty to incentivize an agent to explore novel states. A number of novelty measurement strategies have been proposed in the past few years, such as the use of information gain [Houthooft *et al.*, 2016], count-based methods utilizing counting tables [Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017], and prediction-based methods exploiting prediction errors of dynamics models [Stadie *et al.*, 2015; Pathak *et al.*, 2017; Burda *et al.*, 2019a; Burda *et al.*, 2019b]. The prediction-based novelty measurement strategies, which are the primary focus and interest of this paper, differ in the targets of prediction. Researchers in [Pathak *et al.*, 2017; Burda *et al.*, 2019a] introduce a forward dynamics model for predicting the feature representation of the next state of an agent, which is known as *next-frame prediction*. Next-frame prediction for complex or rapid-changing motions, however, is rather difficult for forward dynamics models, 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 prediction typically requires complex feature representations [Xue *et al.*, 2016; Lotter *et al.*, 2017]. On the other hand, the researchers in [Burda *et al.*, 2019b] proposed a *self-frame prediction* strategy by employing a random distillation network (RND) to predict the feature representation of the current state from a fixed and randomly initialized target neural network. Nev-
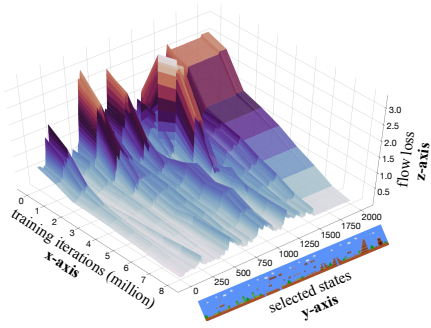
Figure 2: Prediction errors of FICM versus training iterations for selected states.
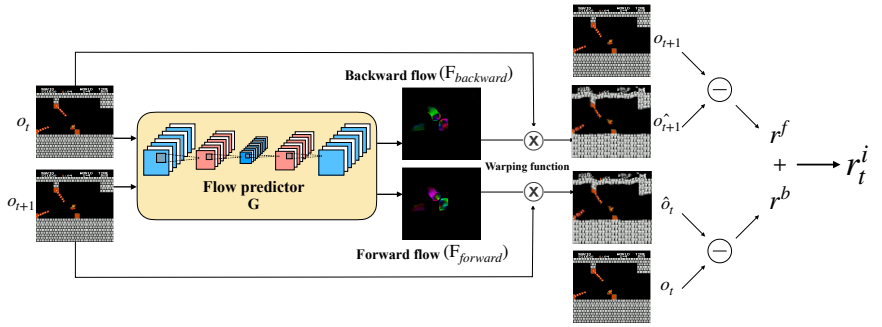

Figure 3: The workflow of the proposed flow-based intrinsic curiosity module (FICM).

ertheless, the attempt of self-frame prediction inherently neglects motion features between consecutive observations.

As a result, in this paper we introduce a **f**low-based **i**ntrinsic **c**uriosity **m**odule, called **FICM**, for evaluating the novelty of observations and generating intrinsic rewards. FICM is motivated by the issues mentioned above, and is designed to focus on the motion features of the objects extracted from two consecutive observations. FICM generates intrinsic rewards based on the prediction errors of optical flow estimation (i.e., the flow loss). 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 of FICM for *Super Mario Bros*. It can be observed that the brighter parts (which correspond to higher flow loss) in Fig. 1 (c) align with the attention areas [Greydanus *et al.*, 2018] in Fig. 1 (d), implying that the agent is intrinsically intrigued by the motion parts between the two observations.

We validate the above properties of FICM in a variety of benchmark environments, including *Atari 2600* [Bellemare *et al.*, 2013], *Super Mario Bros.*, and *ViZDoom* [Wydmuch *et al.*, 2018]. We demonstrate that FICM is preferable to a number of previous prediction-based exploration methods in terms of concentrating on motion features between consecutive observations in several tasks and environments. We further provide a set of ablation analyses for comparing the encoding efficiency of FICM and the existing approaches, and comprehensively discuss their implications. The primary contributions of this paper are thus summarized as follows. First, we propose FICM, which leverages on the prediction errors of optical flow estimation to encourage a DRL agent to explore unfamiliar motion features between consecutive observations in a comprehensive manner. Second, FICM is able to encode high dimensional inputs (e.g., RGB frames), and demands only two consecutive frames to obtain sufficient information when evaluating the novelty of the observations.

2 Background

In this section, we first introduce the previous curiosity-driven exploration methodologies. We then discuss the two representative strategies employed by them: the next-frame prediction and the self-frame prediction strategies. Finally, we briefly review the concepts of optical flow estimation.

2.1 Curiosity-Driven Exploration Methodologies

Curiosity-driven exploration is an exploration strategy adopted by a number of DRL researchers in recent years [Houthooft *et al.*, 2016; Bellemare *et al.*, 2016; Ostrovski *et al.*, 2017; Pathak *et al.*, 2017; Burda *et al.*, 2019a; Burda *et al.*, 2019b] in order to explore environments more efficiently. While conventional random exploration strategies are easily trapped in local minima of state spaces for complex or sparse reward environments, curiosity-based methodologies tend to discover relatively un-visited regions, and therefore are likely to explore more effectively. In addition to the extrinsic rewards provided by the environments, most curiosity-driven exploration strategies introduce intrinsic rewards generated by the agent to encourage itself to explore novel states. In this paper, we primarily focus on the prediction-based curiosity-driven exploration methods, which include two branches: the next-frame and the self-frame prediction strategies.

Next-frame prediction strategy. In [Pathak *et al.*, 2017; Burda *et al.*, 2019a], the authors formulate the exploration bonuses as the prediction errors of a forward dynamics network. Given an observation o_t, a feature representation ϕ_t is generated by an embedding network. The forward dynamics network f is then used to predict the representation of the next observation by ϕ_t and the agent's taken action a_t. The mean-squared error (MSE) $||f(\phi_t, a_t) - \phi_{t+1}||^2$ is minimized by f and serves as the intrinsic reward signal. When the same observations are visited multiple times, their MSE errors drop, indicating that the agent are becoming familiar with them.

Self-frame prediction strategy. In [Burda *et al.*, 2019b], the authors introduce RND to measure state novelty. RND is composed of two networks: a fixed and randomly initialized target network g, and a predictor network $\hat{g}$. For the current observation o_t, the objective of RND is to minimize the MSE $||g(o_t) - \hat{g}(o_t)||^2$, which serves as the intrinsic reward signal. By introducing a deterministic g, RND avoids the problem of stochastic prediction errors occurred in next-frame prediction.
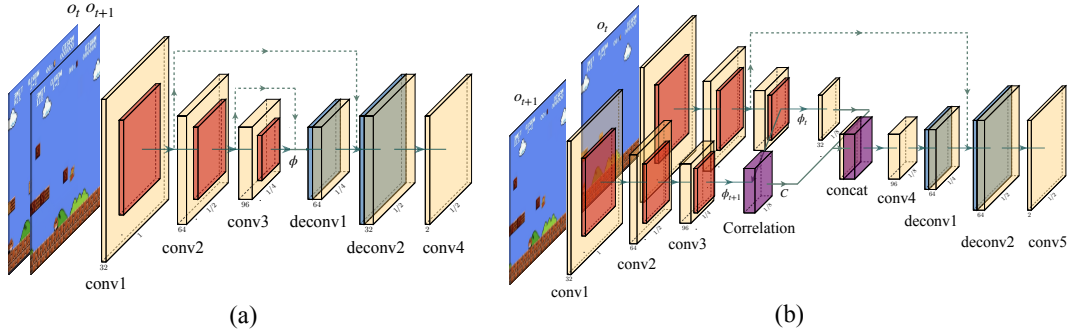
Figure 4: The flow predictor architectures of (a) FICM-S and (b) FICM-C, which based on *FlowNetS* and *FlowNetC* introduced by FlowNet 2.0 [Ilg *et al.*, 2017], respectively.

2.2 Optical Flow Estimation

Optical flow estimation is a technique for evaluating the motion of objects between consecutive images, which typically requires a reference image and a target image. Optical flow is usually represented as a vector field containing displacement vectors assigned to the pixels of the reference image. These vectors indicate the shifts of the corresponding pixels from the target image, and can be exploited as the motion features. In recent years, a number of deep learning approaches have been proposed [Fischer *et al.*, 2015; Ilg *et al.*, 2017] for enhancing the prediction accuracy and resolving large displacement issues of optical flow estimation.

3 Methodology

In this section, we present the motivation and design overview of FICM. We first introduce the essential concepts of FICM. Then, we formulate them into mathematical equations. Finally, we investigate two different implementations of FICM.

3.1 Flow-Based Curiosity-Driven Exploration

Our main motivation and objective is to develop a new exploration bonus method that is able to (1) understand the motion features between two consecutive observations, (2) encode the observations efficiently, and (3) estimate the novelty of different observations. As a result, we propose to embrace optical flow estimation [Ilg *et al.*, 2017; Meister *et al.*, 2018], a technique that has commonly been leveraged by the field of computer vision for interpreting motion features as displacement vectors between consecutive frames, as our novelty measurement strategy. We validate the qualification of FICM to serve as a novelty estimator by an experiment performed on *Super Mario Bros.*, and illustrate the decreasing trend of the flow loss for millions of iterations. Fig. 2 shows that FICM yields higher intrinsic rewards for unfamiliar observations than familiar ones. When the agent encounters unfamiliar observations, FICM motivates the agent to revisit these observations, and gradually learns the motion features of them. The workflow of FICM and the overview of the flow predictor design are illustrated in Figs. 3 and 4, respectively.

3.2 Flow-Based Intrinsic Curiosity Module

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 o_t and o_{t+1} perceived at consecutive timesteps t and $t+1$, FICM takes the 2-tuple (o_t, o_{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:

$$F_{forward} = G(o_t, o_{t+1}, \Theta_f)$$
$$F_{backward} = G(o_{t+1}, o_t, \Theta_f). \tag{1}$$

$F_{forward}$ and $F_{backward}$ are used to generate the observations $\hat{o}_t$ and $\hat{o_{t+1}}$ via a warping function defined in [Fischer *et al.*, 2015; Ilg *et al.*, 2017], where $\hat{o}_t$ and $\hat{o_{t+1}}$ are formulated as:

$$\hat{o}_t(x) = o_{t+1}(x + F_{forward}(x) * \beta)$$
$$\hat{o_{t+1}}(x) = o_t(x + F_{backward}(x) * \beta), \tag{2}$$

where β is a flow scaling factor, x denotes the pixel index, and '$*$' represents element-wised multiplication. We use the bilinear interpolation method to implement the warping function. Please note that in this work, we employ the reverse mapping approach discussed in [Beier and Neely, 1992], in which the warping procedure looks for the corresponding pixel coordinate in the source image for each pixel in the destination image. The forward mapping approach is not adopted in this work because some pixels in the destination image might not be found and correctly mapped from the source image.

With the predicted observations $\hat{o}_t$ and $\hat{o_{t+1}}$, Θ_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 Θ_f is expressed as:

$$\min_{\Theta_f} L_G = \min_{\Theta_f}(L^f + L^b)$$
$$= \min_{\Theta_f}(||o_{t+1} - \hat{o_{t+1}}||^2 + ||o_t - \hat{o}_t||^2), \tag{3}$$

where (L^f, L^b) are derived from the MSE between $(o_{t+1}, \hat{o_{t+1}})$ and $(o_t, \hat{o}_t)$, respectively. In this work, L_G is interpreted by FICM as a measure of novelty, and serves as
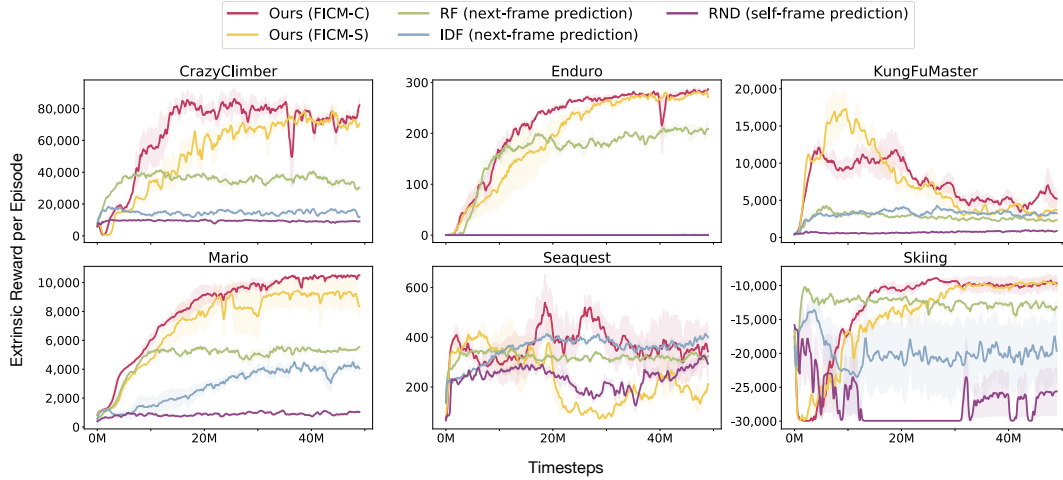
Figure 5: Mean extrinsic rewards (with standard errors) obtained by the agents fed only with intrinsic rewards during the training phase.

the intrinsic reward signal r^i presented to the agent. The expression of r^i can therefore be formulated as follows:

$$
\begin{aligned}
r^i = r^f + r^b &= \frac{\zeta}{2}(L^f + L^b) \\
&= \frac{\zeta}{2}L_G = \frac{\zeta}{2}(||o_{t+1} - \hat{o_{t+1}}||^2 + ||o_t - \hat{o_t}||^2),
\end{aligned}
\tag{4}
$$

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, and FICM only takes two consecutive observations as inputs. The results presented in Section 4 validate the effectiveness of the reward r^i and our FICM.

3.3 Implementations of FICM

In this work, we propose two different implementations of FICM: FICM-S and FICM-C, as depicted in Fig. 4 (a) and Fig. 4 (b), 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 primary difference between the two predictors lies in their feature extraction strategy from the input observations o_t and o_{t+1}. FICM-S encodes the stacked observations $\langle o_t, o_{t+1} \rangle$ together to generate a single feature embedding ϕ, while the other one encodes o_t and o_{t+1} in different paths to generate two separate feature embeddings ϕ_t and ϕ_{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 [Fischer *et al.*, 2015]. The fused feature map is then processed by another convolutional layer at the end to predict the optical flow from o_t to o_{t+1}. Please note that the two input paths of Fig. 4 (b) are share-weighted in order to generate comparable embeddings. The flow predictor of FICM-C additionally employs a correlation layer from [Fischer *et al.*, 2015] to perform multiplicative patch comparisons between ϕ_t and ϕ_{t+1}.

4 Experiments

In this section, we begin with a pure curiosity-driven learning task in the absence of any extrinsic reward signal in Section 4.1, to validate the exploitation capability of motion features. Then, we evaluate the performance of FICM on *Viz-Doom* as well as six established hard exploration *Atari* games, combining intrinsic rewards with sparse extrinsic rewards in Sections 4.2 and 4.3, respectively. In these experiments, we compare FICM against three baselines: next-frame prediction with (a) inverse dynamics features (*IDF*) [Pathak *et al.*, 2017] and (b) random features (*RF*) [Burda *et al.*, 2019a], as well as (c) self-frame prediction with *RND* [Burda *et al.*, 2019b], respectively. To present the best settings in terms of the agents' performance, FICM takes RGB frames as the inputs of the flow predictor while the baselines maintain their default settings described in their original papers. An ablation analysis of the encoding efficiency between (a) RGB and gray-scale and (b) stack and non-stacked frames are further discussed in Section 4.4. Finally, we discuss the impacts of FICM on the agents in Section 4.5. Unless stated otherwise, all of the experiments are conducted with proximal policy optimization (PPO) [Schulman *et al.*, 2017] and all of the presented curves are the average of three runs based on different random seeds.

4.1 Experiments on FICM's Capability of Exploiting Motion Features for Exploration

Environments. We first validate FICM's capability of exploiting motion features to motivate agents for exploration by feeding the agents only with intrinsic rewards. We similarly feed the baseline agents with only the intrinsic rewards specified in their original papers for a fair comparison. In our experiments, we remove the end-of-episode signal, and measure the extrinsic rewards obtained by the agents during the training phase for *Super Mario Bros.* and five *Atari* games, including *CrazyClimber*, *Enduro*, *KungFuMaster*, *Seaquest*, and *Skiing*. These games are characterized by moving objects that require the agents to concentrate on and interact with. More precisely, the motions of the foreground and background components (including the controllable agent and the
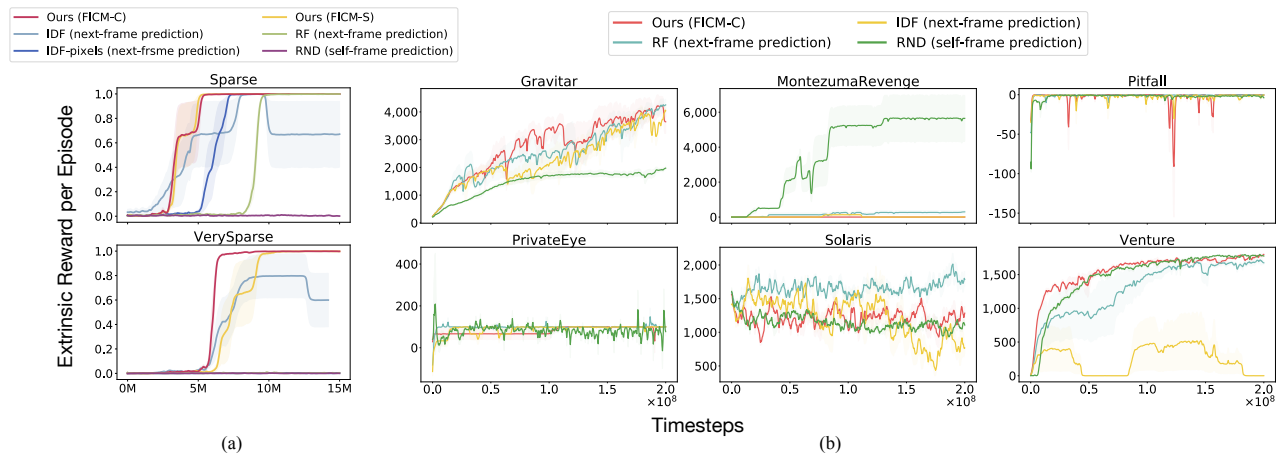
Figure 6: Comparison of the evaluation curves in (a) *ViZDoom* with *sparse* and *very sparse* extrinsic reward setups, and (b) six established hard exploration *Atari* games.

uncontrollable objects) are related to the performance (i.e., obtainable scores) of the agent. Taking *Enduro* for example, the agent not only has to understand the motions of its controllable car, but is also required to perceive and comprehend the motions of the other cars, as their motions are directly related to the final score of the agent. We provide additional discussions on the above properties quantitatively in Section 4.5.

Results and discussions. In Fig. 5, we plot the evaluation curves of the extrinsic rewards obtained by the agents versus timesteps for our methods (denoted as *FICM-S* and *FICM-C*) and the baselines in different environments. At each timesteps, different agents may focus their attention on exploring different subjects. The resultant fluctuations in the curves (e.g., *KungFuMaster* and *Seaquest*) directly reflect the pure exploration processes (and hence, the capability) of the agents, since there is no extrinsic reward provided to the them. Please note that the same fluctuations in the curves also occurred in the experiments conducted in [Burda *et al.*, 2019a]. The trends of the curves reveal that our methods enable the agents to explore states with high extrinsic rewards more often than the baselines in *Super Mario Bros.* (e.g., discovering secret rooms and defeating Bowser), *KungFuMaster*, *CrazyClimber*, *Enduro*, and *Skiing*, while delivering comparable performance to the baselines in *Seaquest*. As these games are characterized by moving objects that require the agents to concentrate on and interact with, they are favorable to *FICM-S* and *FICM-C* that are capable of capturing motion features and perceiving the differences between observations.

4.2 Experiments on Exploration with Sparse Extrinsic Rewards

Environments. We further conduct experiments on the *DoomMyWayHome-v0* environment in *ViZDoom* [Wydmuch *et al.*, 2018], using the same *sparse reward* and *very sparse reward* setups as those conducted in [Pathak *et al.*, 2017]. This environment features first-person perspective observations. Besides the default baselines mentioned in Section 4, we additionally include another next-frame prediction baseline *IDF-pixels* (also known as *ICM-pixels* in [Pathak *et al.*,

2017]) for a fair comparison. In these experiments, the DRL agents are all based on Asynchronous Advantage Actor-Critic (A3C) [Mnih *et al.*, 2016]. As we intend to reproduce the results of [Pathak *et al.*, 2017] and compare with them, we adopt the officially released implementation of [Pathak *et al.*, 2017] as our baselines. For a fair comparison, we modify the feature learning modules of [Pathak *et al.*, 2017] for the other baselines to generate different types of intrinsic rewards. Please note that in this environment, a single value head [Burda *et al.*, 2019b] is applied to *RND*. The summation of the extrinsic and intrinsic rewards are provided to the DRL agents in all of the experiments considered in this section.

Results and discussions. Fig. 6 (a) plots the evaluation curves of the experiments. It is observed that the agents based on FICM are able to converge faster than the baselines and maintain stable performance consistently in both the *sparse* and *very sparse* reward setups. The *RF* and *IDF-pixels* baselines are able to succeed in the *sparse reward* setup, but fail in the *very sparse reward* one. The *IDF* baseline is able to reach the final destination in both cases, however, sometimes suffers from performance drops and is not always guaranteed to stably maintain its performance. *RND* fails to guide the agent to reach the final destination in both the *sparse reward* and *very sparse reward* setups. Since *FICM-C* is stabler than *FICM-S* in our results demonstrated in Sections 4.1 and 4.2, we present only *FICM-C* in the remaining sections.

4.3 Experiments on Hard Exploration Games

Environments. We evaluate FICM on six established hard exploration *Atari* games categorized by [Bellemare *et al.*, 2016], including *Venture*, *Gravitar*, *Montezuma's Revenge*, *Pitfall*, *PrivateEye*, and *Solaris*. These games feature sparse reward environments. In the experiments, the agents trained with the proposed methods and the baselines are fed with both extrinsic and intrinsic rewards. Please note that solving hard exploration games is not the focus of this paper. We provide these comparisons just for benchmark purposes, and validate if FICM is beneficial for sparse-reward games.
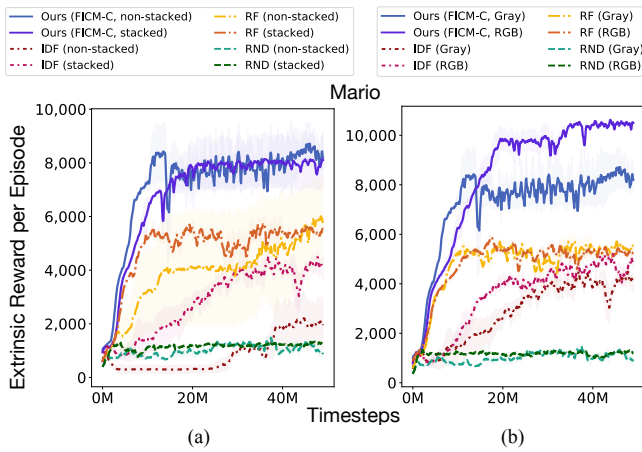
Figure 7: Encoding efficiency analysis: (a) stacked v.s. non-stacked frames and (b) RGB v.s. gray-scale.

(a) Extrinsic   (b) Intrinsic   (c) Both

(d) Evaluation curve

Figure 8: Visualization of the agents' attention areas (purple) and the evaluation curves for three reward setups.

Results and discussions. The evaluation curves are plotted and compared in Fig. 6 (b). It is worth noticing that in [Burda *et al.*, 2019b], *RND* established the state of art performance in *Montezuma's Revenge*, *Gravitar*, and *Venture*. From Fig. 6 (b), it can be observed that FICM outperforms the baselines in *Gravitar*, and is able to converge faster than them in *Venture*. As the moving objects in these two games are related to the obtainable scores, focusing on motion features offers benefits for our agents to explore the environments. In *Solaris*, the scores of our method is comparable to those of *RND* and *IDF*, while *RF* achieves relatively higher performance. On the other hand, all of the methods in Fig. 6 (b) achieve similar results in *Pitfall*, *PrivateEye*, and *Montezuma's Revenge*, except for *RND* in *Montezuma's Revenge*. The rationale is that these games demand reasoning and problem-solving skills, and are thus relatively harder to be solved solely by the observations perceived by the agents.

4.4 Ablation Analysis

In this section, we present a set of ablation analysis to investigate the feature encoding abilities of FICM in two aspects: (a) the number of the stacked frames and (b) the dimensionality of the input frames. The experiments are evaluated on *Super Mario Bros.*, and the results are plotted in Fig. 7, respectively.

Stacked versus non-stacked frames. We consider two different setups in terms of the numbers of stacked frames: (a) two consecutive single frames (denoted as *non-stacked*) and (b) a pair of four stacked frames (denotes as *stacked*). The results show that both *FICM-C (stacked)* and *FICM-C (non-stacked)* maintain consistent performance, indicating that FICM is able to encode the information efficiently even with only two consecutive frames. In contrast, it is observed that the evaluation curves of *RF (non-stacked)* and *IDF (non-stacked)* exhibit high variance (highlighted as the wide shaded areas of the evaluation curves) and achieve lower scores as compared to their *stacked* versions, indicating that they suffer from unstable performance due to the poor encoding efficiency. On the other hand, *RND* shows similar encoding efficiency for the *stacked* version and the *non-stacked* version.
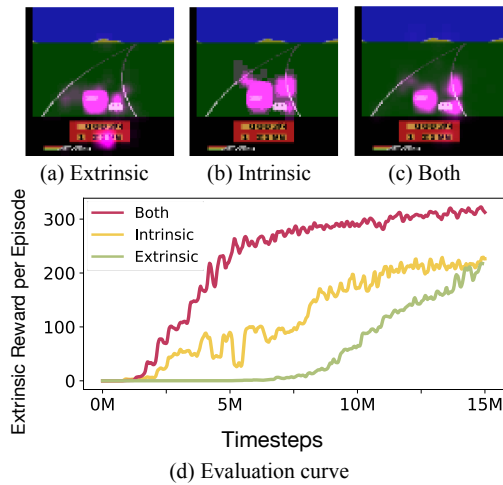
RGB versus gray-scale frames. We hypothesize that RGB frames contain more useful information than gray-scale ones to be utilized by prediction-based approaches. We examine whether FICM and the baselines are able to make use of the extra information to guide the agent for exploration. We present the experiments of utilizing RGB and gray-scale input frames for both FICM and the baselines in Fig. 7 (b). It is observed that *RF* and *RND* perform nearly the same for both settings. In contrast, FICM with RGB frames outperforms the one with gray-scale frames, indicating that FICM can encode the features more efficiently and utilize the information contained in RGB channels. The superior encoding efficiency is mainly due to the capability of the flow predictor in perceiving the motion features with color information, as it is much easier for it to distinguish between still and moving objects. It is observed that *IDF* is also slightly benefited from the color information as compared to the one with gray-scale inputs. On the contrary, it is difficult for *RF* and *RND* to obtain appropriate representations of RGB frames, since they mainly rely on random features and are therefore lack of eligible embedding networks to encode the additional color information.

4.5 Discussions on the Impacts of FICM

We first discuss why FICM is favorable for environments characterizing moving objects, which may include controllable agents and uncontrollable parts, by an illustrative example performed on *Enduro*. Fig. 8 visualizes the evaluation curves obtained with the same settings discussed in Section 4.1 as well as the attention areas [Greydanus *et al.*, 2018] for three different reward setups: (1) extrinsic rewards only, (2) intrinsic rewards only, and (3) both extrinsic and intrinsic rewards. It is observed that in Fig. 8 (a), the agent misfocuses its attention on the scoreboard at the bottom, which is not directly related to the performance of the agent. In contrast, Fig. 8 (b) shows that the agent trained with FICM concentrates more on the moving parts (i.e., the controllable car and opponents). The evaluation curves in Fig. 8 (d) also indicate that the agent trained with intrinsic rewards only is able to
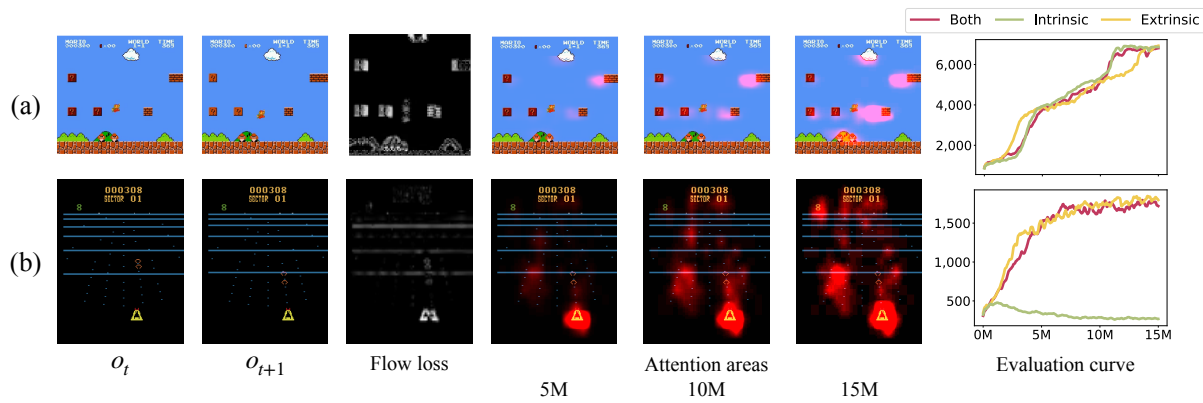
Figure 9: FICM's influence in (a) *Super Mario Bros.* and (b) *BeamRider*. The 1st-3rd columns show o_t, o_{t+1}, and the flow loss of them, where the brighter parts reflect higher flow loss. The 4th-6th columns highlight the agents' attention areas in red at different timesteps. The last one shows the evaluation curves.

visit states with high extrinsic rewards more often than that trained with extrinsic rewards only. Fig. 8 (d) further reveals that when applying both extrinsic and intrinsic rewards together during the train phase, the agent is able to perform significantly better than the other two cases. It is observed that the attention areas depicted in Fig. 8 (c) do not cover the scoreboard part anymore, implying that the intrinsic rewards generated by FICM bring positive impacts on guiding the learning process of the agent toward a beneficial direction.

Despite the benefits that FICM brings, FICM suits better for environments where the motions of the controllable and uncontrollable parts are relevant to the resultant performance. FICM is not favorable in the following scenarios: (1) insufficiency of movements from the controllable and uncontrollable objects (e.g., enemies), and (2) excessive changes of irrelevant components (e.g., backgrounds) that might distract the focus of FICM. For instance, if the objects that the agent is supposed to pay attention to barely move, FICM generates negligible intrinsic rewards due to little discrepancy between two consecutive observations. On the other hand, if some irrelevant parts of the environment move relatively faster than those pertinent to the resultant performance, FICM may be distracted to focus on those irrelevant regions or components.

To prove and validate the hypothesis above, we further illustrate how FICM guides the agent to explore the environments with only intrinsic rewards by visualizing the corresponding flow loss and the attention areas [Greydanus *et al.*, 2018] of the agents during the training phase in Fig. 9. 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 become brighter at those regions at later timesteps. Flow loss may as well distract the agent to focus on irrelevant parts of the agents' observations. Take *BeamRider* for example, the agent is misled to pay attention to the background by the rich flow loss in it, limiting the extent of FICM's benefits. However, please note that the overall performance of the agents fed with both extrinsic and intrinsic rewards illustrated in the evalua-

tion curves of Fig. 9 are not degraded when they are compared with the agents fed with extrinsic rewards only, implying that the intrinsic rewards generated by FICM do not bring negative impacts to the agent's learning process. On the other hand, although the agent is able to master *Super Mario Bros.* with different reward setups, the intrinsic rewards generated by FICM is able to support the agent's learning procedure to converge faster than that with only extrinsic rewards. This is because FICM is able to help the agent to concentrate on the enemies, and not being distracted by the irrelevant objects.

5 Conclusions

In this paper, we proposed FICM for evaluating the novelty of observations. FICM generates intrinsic rewards by leveraging on the prediction errors of optical flow estimation to encourage a DRL agent to explore unfamiliar motion features between consecutive observations in a 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*. We validated that FICM is capable of focusing on the motion features and guiding the agents for exploration. Moreover, we presented a set of ablation analysis to investigate the encoding efficiency of FICM and discussed the impacts of it on the agents. Based on the experimental evidence discussed above, we thus conclude that FICM is a promising method to be employed by the agents, especially in environments featuring moving objects.

Acknowledgments

References

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

[Bellemare *et al.*, 2013] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *J. Artificial Intelligence Research (JAIR)*, 47:253–279, May 2013.

[Bellemare *et al.*, 2016] Marc G. Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Rémi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1471–1479, Dec. 2016.

[Burda *et al.*, 2019a] Yuri Burda, Harrison Edwards, Deepak Pathak, Amos J. Storkey, Trevor Darrell, and Alexei A. Efros. Large-scale study of curiosity-driven learning. In *Proc. Int. Conf. Learning Representation (ICLR)*, May 2019.

[Burda *et al.*, 2019b] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *Proc. Int. Conf. Learning Representations (ICLR)*, May 2019.

[Fischer *et al.*, 2015] Philipp Fischer, Alexey Dosovitskiy, and Eddy Ilg. *et al.* FlowNet: Learning optical flow with convolutional networks. In *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, pages 2758–2766, May 2015.

[Greydanus *et al.*, 2018] Samuel Greydanus, Anurag Koul, Jonathan Dodge, and Alan Fern. Visualizing and understanding atari agents. In *Int. Conf. Machine Learning (ICML)*, pages 1787–1796, Jun. 2018.

[Houthooft *et al.*, 2016] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational information maximizing exploration. In *Proc. Advances in Neural Information Processing Systems (NeurIPS)*, pages 1109–1117, Dec. 2016.

[Ilg *et al.*, 2017] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas 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.

[Lotter *et al.*, 2017] William Lotter, Gabriel Kreiman, and David D. Cox. Deep predictive coding networks for video prediction and unsupervised learning. *arXiv:1605.08104*, Mar. 2017.

[Meister *et al.*, 2018] Simon Meister, Junhwa Hur, and Stefan 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.

[Mnih *et al.*, 2015] Volodymyr Mnih, Koray Kavukcuoglu, and David Silver. *et al*. Human-level control through deep reinforcement learning. *Nature*, vol. 518, no. 7540, pp. 529-533, Feb. 2015.

[Mnih *et al.*, 2016] Volodymyr Mnih, Adrià Puigdomènech Badia, and Mehdi Mirza. *et al.* Asynchronous methods for deep reinforcement learning. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 1928–1937, Jun. 2016.

[Ostrovski *et al.*, 2017] Georg Ostrovski, Marc G. Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 2721–2730, Jun. 2017.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proc. Int. Conf. Machine Learning (ICML)*, pages 2778–2787, May 2017.

[Schulman *et al.*, 2017] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, Jul. 2017.

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

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

[Xue *et al.*, 2016] Tianfan Xue, Jiajun Wu, Katherine L. Bouman, and Bill 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.

[Zhang *et al.*, 2016] Marvin Zhang, Zoe McCarthy, Chelsea Finn, Sergey Levine, and Pieter Abbeel. Learning deep neural network policies with continuous memory states. In *Proc. Int. Conf. Robotics and Automation (ICRA)*, pages 520–527, May 2016.