PreND: Enhancing Intrinsic Motivation in Reinforcement Learning through Pre-trained Network Distillation

Mohammadamin Davoodabadi **Negin Hashemi Dijujin** **Mahdieh Soleymani Baghshah**

Department of Computer Engineering
Sharif University of Technology
{mohammadamin.davoodabadi, n.hashemi94, soleymani}@sharif.edu

Abstract

Intrinsic motivation, inspired by the psychology of developmental learning in infants, stimulates exploration in agents without relying solely on sparse external rewards. Existing methods in reinforcement learning like Random Network Distillation (RND) face significant limitations, including (1) relying on raw visual inputs, leading to a lack of meaningful representations, (2) the inability to build a robust latent space, (3) poor target network initialization and (4) rapid degradation of intrinsic rewards. In this paper, we introduce ***Pre**-trained **N**etwork **D**istillation* (**PreND**), a novel approach to enhance intrinsic motivation in reinforcement learning (RL) by improving upon the widely used prediction-based method, RND. PreND addresses these challenges by incorporating pre-trained representation models into both the target and predictor networks, resulting in more meaningful and stable intrinsic rewards, while enhancing the representation learned by the model. We also tried simple but effective variants of the predictor network optimization by controlling the learning rate. Through experiments on the Atari domain, we demonstrate that PreND significantly outperforms RND, offering a more robust intrinsic motivation signal that leads to better exploration, improving overall performance and sample efficiency. This research highlights the importance of target and predictor networks representation in prediction-based intrinsic motivation, setting a new direction for improving RL agents' learning efficiency in sparse reward environments.

1 Introduction

Reinforcement Learning (RL) has been acquired to solve many complex problems such as robot navigation [24], stratospheric balloons navigation [3], playing Go [35], and instruction-tuning large language models [28, 6]; however, by definition, it relies on reward signals from the environment which could be very cumbersome and labor-intensive to specify properly in the real-world tasks [12]. Many tasks especially in goal-oriented settings [9] involve sparse rewards which provide weak learning signals for the agent and result in an ambiguous understanding of the environment.

One of the methods proposed in the literature to overcome this challenge is *intrinsic motivation* which is mainly inspired by the psychology and developmental learning of skills in babies [1, 11]. Intrinsic motivation approaches involve different methods to improve exploration, learn diverse skills and enhance overall performance in complex environments using internally generated signals besides the external motivations [1]. To this end, various approaches have been proposed which could be categorized into two main groups: 1) *knowledge acquisition* and 2) *skill learning*. Knowledge acquisition methods involve exploration-oriented methods based on prediction error [5], state novelty such as count-based methods [36, 29], and information gain [18], or other methods based

Accepted at the Intrinsically Motivated Open-ended Learning workshop at NeurIPS 2024.

on empowerment [8] and state representation learning [26]. On the other hand, the skill learning category includes skill abstraction [15, 34] and curriculum learning methods [33]. Among all these categories, prediction-based methods and count-based methods have been mostly adopted throughout the literature due to their simplicity and promising performance [19]. However, count-based methods are not scalable in continuous and large state spaces while prediction-based methods handle these scenarios more efficiently [1].

One of the most important methods in prediction-based category is *Random Network Distillation (RND)* [5] which counts as the basis of many other studies [37, 17, 19, 2]. This method derives an intrinsic reward based on the prediction error between a random fixed function of states/observations and a learnable predictor network with the same input. The predictor network tries to learn the output of the fixed target network using a mean squared error objective; hence rewards the agent for visiting unfamiliar states that has seen less during training. Although this approach reduces uncertainty about the environment and seems suitable for exploring large state spaces, it is sensitive to target network initialization, produces rewards with low variance, and suffers from catastrophic forgetting throughout training [30, 19]. Recent studies [30, 19] have emphasized the shortcomings of RND. Specifically in [30], they propose self-supervised regularization of the target network to enhance representations of the target network. Target network representation also plays a crucial role as long-term memory keys in overcoming catastrophic forgetting, and also in specializing ensembles of curiosity modules over observation subspaces [19].

In this paper, we emphasize the importance of target and predictor network representations in producing meaningful intrinsic rewards and improving the overall performance of the agent. Our analyses reveal the shortcomings of basic RND in generating meaningful and varied intrinsic rewards. We suggest that by changing design choices in random network distillation, network target initialization problems and low reward variance could be resolved. We explore using pre-trained models on the Atari domain from the recent Atari-PB benchmark [20] as part of the target network initialization, and separate training speed for the overall policy and the predictor network to prevent its early overfitting to target network and low reward variance. Our experiments in Atari environment suggest effectiveness of these techniques.

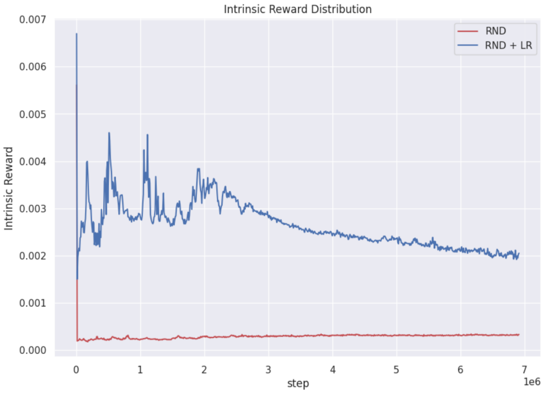


Figure 1: Effect of predictor's learning rate on intrinsic trend and oscillation of intrinsic reward

2 Method

In this section, we present the methodology behind our proposed approach, ***Pre**-trained **N**etwork **D**istillation* (**PreND**). As discussed, RND relies on the difference between the outputs of a target network and a predictor network to generate intrinsic motivation. However, it has the following problems:

1. A random target network might not represent the semantics of the input observations/states in producing the intrinsic reward. This might happen due to poor latent space modeling in a random network. Similar observations might scatter through the latent space while different observations could be mapped to closer latent points. This can confuse the predictor and produce false surprise signals. We calculated the correlation between intrinsic rewards

difference and state embedding distance for RND. The resulting number (≈ 0.39 on the scale of $[-1, 1]$) indicates a weak correlation [14], leaving room for improvement (See Appendix Figure 3 for visualizations). Indeed, several parts of the state space might be far apart (high embedding distance) yet yield similar intrinsic rewards, breaking the correlation. This issue underscores the objective of using richer representations, which can better distinguish states in terms of similarity, addressing the limitations of RND's less structured latent space.

2. The intrinsic reward drops quickly after very few initial iterations, leading to poor reward variance (Figure 1). Similar rewards make it more difficult for the agent to discriminate between states while the whole point of prediction-based intrinsic reward is to help the agent visit under-explored states by giving it relatively considerable rewards. Previous research has noted that the lack of variance in the predictor's output leads to a diminished motivational signal [30], primarily due to the predictor network's very fast adaptation to the frozen target model. This rapid adaptation occurs much faster than the policy can learn, causing the intrinsic motivation to diminish too quickly, preventing the agent from gathering sufficient data to effectively train its policy. While Figure 1 shows the batch-average intrinsic reward rather than direct variance, it serves as a proxy to understand the overall trend in reward variance across states. A consistently low batch-average reward likely indicates low variance in intrinsic rewards, indicating that the predictor network quickly converges to mimicking the target network, thereby reinforcing our conclusion.

To this end, we propose PreND, consisting of the following techniques to improve the abovementioned weaknesses:

Pre-trained Feature Extractor: Using pre-trained models for the target network, we can improve the fixed random target network problem. These models can provide a meaningful latent representation space that captures the underlying structure of observations and induces relatively suitable prediction errors among observations/states. This approach could also enhance the robustness of networks against noisy observations and distractor objects by relying on meaningful learned representations.

More specifically, in our experiments on the Atari environment, we propose to use the domain-specific pre-trained backbone model from the Atari-PB benchmark to extract high-level representations. We then employed a randomly initialized network, similar to the neck model architecture from Atari-PB (a transformer-based random fixed network), as the target network, while a similar learnable model was selected for the predictor. This ensures that the predictor and target are complex enough to understand meaningful features of the input, and the predictor is capable of modeling the output of the target network (See Appendix C for more information).

Our target network maintains the crucial property of keeping distant frames far apart in latent space while bringing consecutive frames closer. Unlike RND, where the predictor also had to learn the spatial and temporal features, PreND ensures that the predictor focuses solely on predicting the reward, resulting in a more robust motivational signal.

Slower Predictor Optimization: To alleviate the fast degradation of intrinsic reward, we suggest lowering the optimization speed in the predictor network. Normally, RND uses the same optimizer for both predictor and value/policy heads. Since fitting RL components are known to be more sample-inefficient compared to supervised tasks, an equal optimization speed increases the chance of collapse in intrinsic reward. By changing the pace via learning rate, similar to previous studies [22], we hope to improve intrinsic reward patterns during training.

To summarize, our experiments demonstrate that PreND outperforms RND by providing a richer representation of the input throughout the training, leading to better correlation between changes in input and changes in reward. This alignment is critical for prediction-based intrinsic motivation methods, and PreND offers an effective solution for addressing the limitations of RND.

3 Experiments

We evaluated our approach on two Atari environments: `Boxing`, and `Riverraid` (See Appendix Figure 4). These environments were chosen because they share several characteristics that are particularly relevant to our study. Each game contains highly irrelevant features, such as detailed backgrounds, which our pre-trained representation model should ideally ignore. Moreover, the objects in these games are relatively large, which provides a more favorable setting for RND to compete fairly. In contrast, RND's performance typically suffers in environments with small moving objects, like the ball in Pong [19], as its random target network struggles to capture the subtle movements.

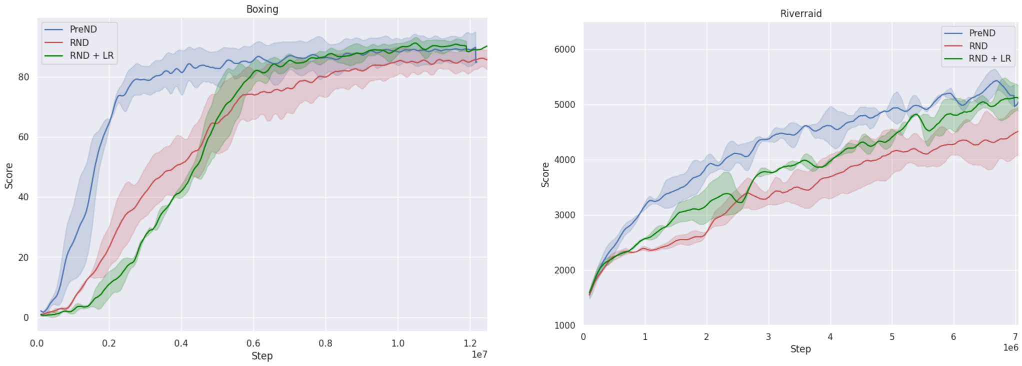


Figure 2: Game score comparison on `Boxing` and `Riverraid`. PreND works better in the low data regimes, showing its sample efficiency. It is also the best model among the settings at the end of the training steps.

For each environment, we ran our experiments with two different seeds and eight parallel episodes, and compared the following three models:

- **RND**[5]: This was our main competitor, offering a prediction-based intrinsic reward to drive exploration. RND utilizes CNNs as the predictor and target networks.
- **RND + LR**: A variant of RND where we adjusted the learning rate to prevent the predictor network from overcoming the target network. We consider a separate optimization process for fitting the predictor network and multiply its learning rate by 0.01. In traditional RND, the learning rate of both the target and predictor networks is the same, which can lead to the predictor network quickly overpowering the target network and learning it too fast. To address this issue, reducing the learning rate of the predictor network relative to the target network was our initial step in modifying the traditional RND.
- **PreND**: Our proposed method, leverages a pretrained representation model for intrinsic reward generation. We use the pre-trained ResNet-50 backbone and SiamMAE neck from Arati-PB [20] as the target and predictor network structure. Backbone (which is pre-trained on several Atari games) retains its pre-trained weights and acts as a feature encoder for observations. We randomly initialize the neck for both the target and predictor networks to prevent game-specific bias of the pre-trained weights. The intrinsic reward in calculated over the neck's output with a size of 512, as in RND.

During our experimentation, we also explored several modifications and variants aimed at further improving RND. These included reducing the capacity of the predictor network by removing additional layers and incorporating techniques such as spectral normalization (commonly used in GAN variants [25]) to the predictor hoping to make its learning process slower and improve the intrinsic reward variance. However, none of these modifications led to improvements. The main results have been shown in Figure 2.

4 Conclusion

In this paper, we introduced Pre-trained Network Distillation, or PreND for short, a novel approach to improving intrinsic motivation in reinforcement learning by leveraging pre-trained representation models. Through experiments in Atari games, we demonstrated that PreND performs better than both Random Network Distillation (RND) and a variant of RND with a modified learning rate. While the modified RND showed improvements over the baseline by addressing some of the RND's inherent issues, PreND's use of richer input representations and pre-trained models seems even more effective. The stable and informative reward signal generated by PreND allowed for better exploration and learning efficiency, confirming that target and predictor network representations play a critical role in producing meaningful intrinsic rewards.

Our results suggest the probable benefits of PreND in the Atari domain, but future work could extend it to more complex environments like DMLab or robotics and explore its use with model-based RL algorithms in more extensive trials. Additionally, experimenting with lighter pre-trained models, such as ResNet-18 instead of ResNet-50, could offer more computational efficiency while maintaining effectiveness.

Our approach employs a domain-specific pretrained model to extract high-level representations, and we acknowledge that this introduces a domain-specific prior. While this may seem to give an advantage over RND, the primary aim here is to address the intrinsic limitations of RND's random latent space and demonstrate the potential of better-structured representations. For future work, task-agnostic or domain-agnostic pretraining methods (e.g., self-supervised learning) could make the approach more generalizable.

References

[1] Arthur Aubret, Laetitia Matignon, and Salima Hassas. A survey on intrinsic motivation in reinforcement learning. *arXiv preprint arXiv:1908.06976*, 2019.

[2] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, et al. Never give up: Learning directed exploration strategies. *arXiv preprint arXiv:2002.06038*, 2020.

[3] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[4] Jonathan Binas, Sherjil Ozair, and Yoshua Bengio. The journey is the reward: Unsupervised learning of influential trajectories. *arXiv preprint arXiv:1905.09334*, 2019.

[5] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. Exploration by random network distillation. *arXiv preprint arXiv:1810.12894*, 2018.

[6] Stephen Casper, Xander Davies, Claudia Shi, Thomas Krendl Gilbert, Jérémy Scheurer, Javier Rando, Rachel Freedman, Tomasz Korbak, David Lindner, Pedro Freire, et al. Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*, 2023.

[7] Nuttapong Chentanez, Andrew Barto, and Satinder Singh. Intrinsically motivated reinforcement learning. *Advances in neural information processing systems*, 17, 2004.

[8] Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational empowerment as representation learning for goal-based reinforcement learning. *arXiv preprint arXiv:2106.01404*, 2021.

[9] Cédric Colas, Tristan Karch, Olivier Sigaud, and Pierre-Yves Oudeyer. Autotelic agents with intrinsically motivated goal-conditioned reinforcement learning: a short survey, 2022.

[10] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[11] Chris Doyle, Sarah Shader, Michelle Lau, Megumi Sano, Daniel Yamins, and Nick Haber. Intrinsically motivated social play in virtual infants. In *Intrinsically-Motivated and Open-Ended Learning Workshop@ NeurIPS2023*, 2023.

[12] Gabriel Dulac-Arnold, Nir Levine, Daniel J Mankowitz, Jerry Li, Cosmin Paduraru, Sven Gowal, and Todd Hester. Challenges of real-world reinforcement learning: definitions, benchmarks and analysis. *Machine Learning*, 110(9):2419–2468, 2021.

[13] Adrien Ecoffet, Joost Huizinga, Joel Lehman, Kenneth O Stanley, and Jeff Clune. First return, then explore. *Nature*, 590(7847):580–586, 2021.

[14] James D Evans. *Straightforward statistics for the behavioral sciences.* Thomson Brooks/Cole Publishing Co, 1996.

[15] Benjamin Eysenbach, Abhishek Gupta, Julian Ibarz, and Sergey Levine. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*, 2018.

[16] Carlos Florensa, David Held, Markus Wulfmeier, Michael Zhang, and Pieter Abbeel. Reverse curriculum generation for reinforcement learning. In *Conference on robot learning*, pages 482–495. PMLR, 2017.

[17] Zhaohan Guo, Shantanu Thakoor, Miruna Pîslar, Bernardo Avila Pires, Florent Altché, Corentin Tallec, Alaa Saade, Daniele Calandriello, Jean-Bastien Grill, Yunhao Tang, et al. Byol-explore: Exploration by bootstrapped prediction. *Advances in neural information processing systems*, 35:31855–31870, 2022.

[18] Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29, 2016.

[19] Jaedong Hwang, Zhang-Wei Hong, Eric Chen, Akhilan Boopathy, Pulkit Agrawal, and Ila Fiete. Neuro-inspired fragmentation and recall to overcome catastrophic forgetting in curiosity. *arXiv preprint arXiv:2310.17537*, 2023.

[20] Donghu Kim, Hojoon Lee, Kyungmin Lee, Dongyoon Hwang, and Jaegul Choo. Investigating pre-training objectives for generalization in vision-based reinforcement learning. *arXiv preprint arXiv:2406.06037*, 2024.

[21] Marlos C Machado, Marc G Bellemare, and Michael Bowling. Count-based exploration with the successor representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5125–5133, 2020.

[22] Kanika Madan, Nan Rosemary Ke, Anirudh Goyal, Bernhard Schölkopf, and Yoshua Bengio. Fast and slow learning of recurrent independent mechanisms. *arXiv preprint arXiv:2105.08710*, 2021.

[23] Jarryd Martin, Suraj Narayanan Sasikumar, Tom Everitt, and Marcus Hutter. Count-based exploration in feature space for reinforcement learning. *arXiv preprint arXiv:1706.08090*, 2017.

[24] Takahiro Miki, Joonho Lee, Jemin Hwangbo, Lorenz Wellhausen, Vladlen Koltun, and Marco Hutter. Learning robust perceptive locomotion for quadrupedal robots in the wild. *Science robotics*, 7(62):eabk2822, 2022.

[25] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks, 2018.

[26] Ofir Nachum, Shixiang Gu, Honglak Lee, and Sergey Levine. Near-optimal representation learning for hierarchical reinforcement learning. *arXiv preprint arXiv:1810.01257*, 2018.

[27] Georg Ostrovski, Marc G Bellemare, Aäron Oord, and Rémi Munos. Count-based exploration with neural density models. In *International conference on machine learning*, pages 2721–2730. PMLR, 2017.

[28] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[29] Deepak Pathak, Dhiraj Gandhi, and Abhinav Gupta. Self-supervised exploration via disagreement. In *International conference on machine learning*, pages 5062–5071. PMLR, 2019.

[30] Matej Pecháč, Michal Chovanec, and Igor Farkaš. Self-supervised network distillation: An effective approach to exploration in sparse reward environments. *Neurocomputing*, page 128033, 2024.

[31] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[32] Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Nikolai Gerasimenko, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. Your transformer is secretly linear, 2024.

[33] Martin Riedmiller, Roland Hafner, Thomas Lampe, Michael Neunert, Jonas Degrave, Tom Wiele, Vlad Mnih, Nicolas Heess, and Jost Tobias Springenberg. Learning by playing solving sparse reward tasks from scratch. In *International conference on machine learning*, pages 4344–4353. PMLR, 2018.

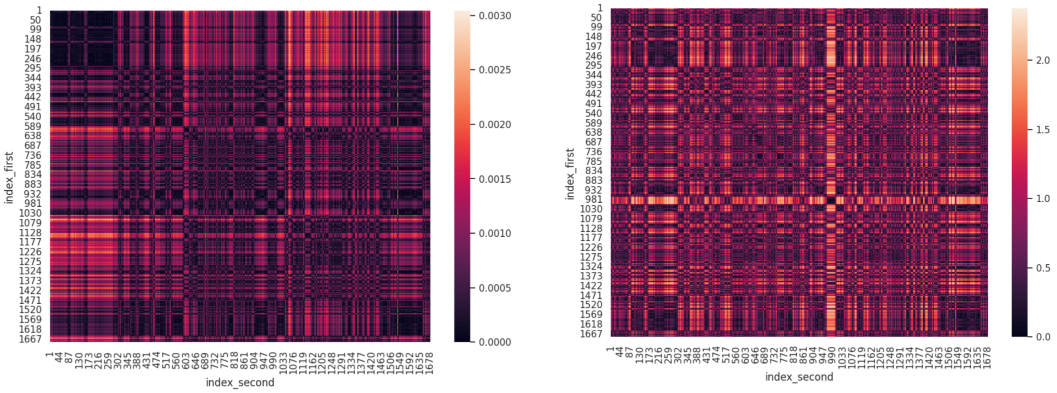
[34] Archit Sharma, Shixiang Gu, Sergey Levine, Vikash Kumar, and Karol Hausman. Dynamics-aware unsupervised discovery of skills. *arXiv preprint arXiv:1907.01657*, 2019.

[35] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

[36] Haoran Tang, Rein Houthooft, Davis Foote, Adam Stooke, OpenAI Xi Chen, Yan Duan, John Schulman, Filip DeTurck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30, 2017.

[37] Huilin Yin, Shengkai Su, Yinjia Lin, Pengju Zhen, Karin Festl, and Daniel Watzenig. Random network distillation based deep reinforcement learning for agv path planning. *arXiv preprint arXiv:2404.12594*, 2024.

Appendix



(a) Pairwise intrinsic reward difference heatmap

(b) Pairwise observation distance heatmap

Figure 3: Heatmaps of pairwise training variables; The correlation between these two is ≈ 0.39 on the scale of $[-1, 1]$. X and Y axes are both the number of training iterations. In both a) and b), the brighter cells indicate higher values (More distance). While prediction-based methods like RND aim to generalize state-novelty by associating similar states with lower prediction errors, we observe that there is no strong relation between the input states RND processes and the intrinsic rewards it produces. This suggests limitations in how RND models and utilizes state similarities in its latent space.

A Related Work

Intrinsic motivation aims to improve exploration at every timestep. Exploration in learning agents helps them acquire various skills required to accomplish diverse tasks [7]. Intrinsic motivation in reinforcement learning is inspired by the psychology and developmental learning of skills in babies and helps tackle several RL challenges [1] such as sparse rewards [5], representation learning [4], skill learning [15, 34], and curriculum learning [16]. There are different approaches to improving exploration using intrinsic motivation which mainly include count-based [36, 13] and prediction-based auxiliary rewards [5, 19, 17, 2]. Count-based approaches are unsuitable for large or continuous state spaces. Some studies try using pseudo-counts and neural density modules to alleviate the scalability problem in count-based methods [27, 21, 23]. On the other hand, prediction-based approaches which are mainly represented by RND [5] could suffer from inappropriate target network initialization, high surprise variance, and early degradation of surprise overtime [30]. In [30], the authors propose improving target network weights using self-supervised regularization techniques. Another approach might rely on using pre-trained representations in the target network. Such representations are abundantly available for various real-world tasks employing large pre-trained vision models [31, 10]. Also in more narrow environments such as Atari, domain-specific pre-trained models [20] offer a chance to examine this hypothesis in the context of prediction-based models such as RND. In this study, we incorporate ResNet-based backbone networks in prediction and target networks to improve the quality of surprise.

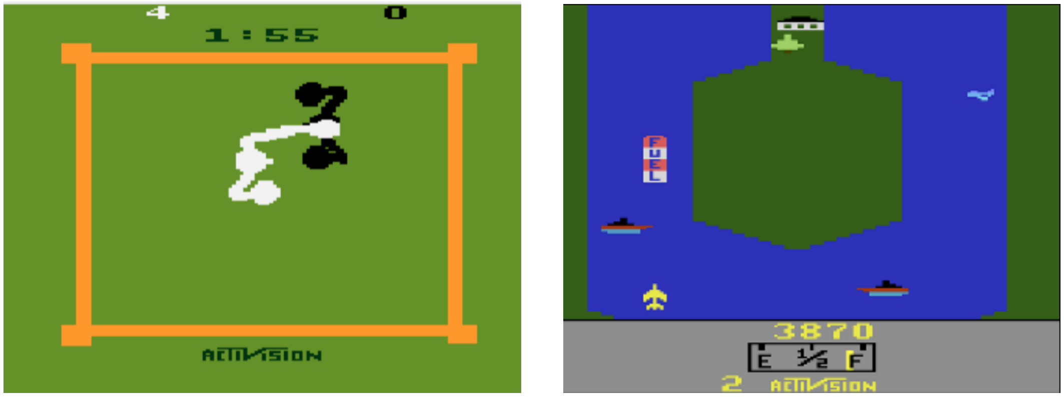
B Experiments Details

We have used the codebase from [19] to run RND and implemented our method on top of it using the codebase of [20]. We ran our reported experiments on two Nvidia GeForce RTX 3090 GPUs and each run took nearly 24 hours.

C Model Architecture

Backbone: The backbone network is a spatial feature extractor, independent of the game, based on ResNet-50 with group normalization. It processes a (4, 84, 84) input and outputs a (2048, 6, 6) feature map, encoding essential features as the input of the target and predictor network of PreND.

Neck: The neck transforms the feature map from the backbone into a 512-dimensional vector by applying spatial pooling, instance normalization, and a 2-layer MLP. It includes game-specific spatial embedding. In the SiamMAE, the output from the neck is further processed by the transformer decoder, which uses cross-attention between the current and masked future frames to reconstruct the image. As [32] said, the transformers maintain the linear property; which is important to ensure that the model keeps the similar frames close to each other in the latent space and the dissimilar ones far from each other.



(a) Boxing

(b) River Raid

Figure 4: Atari games