Self-Supervised Exploration via Temporal Inconsistency in Reinforcement Learning

Zijian Gao, Kele Xu, Member, IEEE, Yuanzhao Zhai, Bo Ding, Dawei Feng, Xinjun Mao, and Huaimin Wang

Abstract—In sparse extrinsic reward settings, reinforcement learning remains a challenge despite increasing interest in this field. Existing approaches suggest that intrinsic rewards can alleviate issues caused by reward sparsity. However, many studies overlook the critical role of temporal information, essential for human curiosity. This article introduces a novel intrinsic reward mechanism inspired by human learning processes, where curiosity is evaluated by comparing current observations with historical knowledge. Our method involves training a self-supervised prediction model, periodically saving snapshots of the model parameters, and employing the nuclear norm to assess the temporal inconsistency between predictions from different snapshots as intrinsic rewards. Additionally, we propose a variational weighting mechanism to adaptively assign weights to the snapshots, enhancing the model’s robustness and performance. Experimental results across various benchmark environments demonstrate the efficacy of our approach, which outperforms other state-of-the-art methods without incurring additional training costs and exhibits higher noise tolerance. Our findings indicate that leveraging temporal information in intrinsic rewards can significantly improve exploration performance, motivating future research to develop more robust and accurate reward systems for reinforcement learning.

Impact Statement—Reinforcement learning plays a vital role in a wide range of research domains. However, existing algorithms are insufficient to deal with sparse-reward scenarios and perform poorly. To solve this problem, inspired by human learning, intrinsic rewards are proposed to mimic curiosity and improve exploration performance. Previous attempts have a significant limitation: they ignore important temporal information and are not robust enough to demonstrate competent performance. Instead, this article proposes a new temporal information-based approach to accurately calculate the intrinsic rewards by measuring the inconsistency between the prediction snapshots. Moreover, the nuclear norm and variational weighting mechanism are designed to improve the performance and tolerance of noise. Preliminary results show that our method significantly improves performance and noise tolerance over existing competitive methods. Our method will motivate future research to design more robust and accurate intrinsic rewards for efficient exploration.

I. INTRODUCTION

OVER the past few years, deep reinforcement learning (DRL) has made significant progress on several challenging sequential decision-making tasks, such as Atari games [1], board games [2], and video games [3]. Moreover, DRL is also shown to be an effective tool in a wide range of practical scenarios such as traffic control [4], [5], [6], [7]. Despite these advancements, the success of DRL is often contingent upon the availability of well-defined, dense external rewards. Regrettably, crafting such rewards for each specific task is frequently impractical [8]. This limitation becomes a bottleneck, impeding the performance of numerous DRL algorithms that rely on these external incentives.

In stark contrast to artificial agents, human learners possess an innate drive to explore and learn from their environment, even in the absence of explicit rewards [9], [10], [11]. This intrinsic motivation, particularly the concept of curiosity, has been a subject of growing interest in the DRL community. It echoes the human tendency to pursue self-motivated goals, despite the scarcity of external rewards. Drawing inspiration from this natural curiosity, researchers have proposed intrinsic rewards to enrich DRL agents’ capacity for self-supervised learning in environments with sparse extrinsic feedback [8], [12], [13]. Most intrinsic rewards can be divided into two broad categories: count-based methods and prediction-based methods. Count-based methods are designed to encourage agents to venture into less-explored areas, yet they often falter when scaled to larger, more complex domains [8], [14], [15], [16], [17], [18]. On the other hand, prediction-based methods leverage a predictive model to anticipate the subsequent state given a state-action pair, using the divergence between prediction and actual outcome as a measure of intrinsic reward [9], [13], [19], [20], [21].

While the potential of intrinsic rewards in DRL is evident, existing methods have not fully realized their promise. A significant oversight in many curiosity-driven approaches is the neglect of temporal dynamics. Humans, when confronted with a problem, instinctively draw upon their accumulated
knowledge to devise a solution. The realization of a gap in one’s understanding sparks curiosity [22], underscoring the value of historical context in learning. Curiosity in humans is not solely a function of the present state but is deeply rooted in our evolving understanding over time. Consider yourself a newborn baby in your parents’ arms in a park. The erratic movement of the leaves is appealing at first. Still, your attention soon shifts to other, more exciting things because you accumulate historical information over a long observation period and conclude that the movement of the leaves is meaningless and uninteresting. Thus, without considering the historical generations of curiosity, it may be difficult to evaluate it accurately as noise may disrupt it.

The challenge, however, lies in the accurate assessment of curiosity, which can be obscured by the inherent noise in learning processes. The stochastic nature of parameter initialization, exploration strategies, replay mechanisms, and environment dynamics introduces variability and uncertainty into the learning process [23]. This variability can undermine the effectiveness of curiosity-driven methods, leading to unreliable predictions and rewards. Traditional measures of prediction error and count error, while indicative of uncertainty, fall short in the face of such stochasticity.

To circumvent these issues, our work advocates for the use of historical knowledge to mitigate high-variance estimates and offers a more nuanced understanding of curiosity. We introduce the temporal Inconsistency-based intrinsic reward (TIR), a novel approach that harnesses the temporal dynamics within a predictive model to enhance self-supervised exploration (as depicted in Fig. 1). By training a model on state–action pairs and maintaining a series of snapshots of these model parameters over time, we can identify regions of the state space that exhibit higher temporal inconsistency, suggesting greater potential for intrinsic reward. In a matrix game scenario (illustrated in Fig. 2), we demonstrate that areas infrequently visited by the agent exhibit higher levels of temporal inconsistency compared to those that are more frequently explored. The nuclear norm serves as a robust measure of this inconsistency, providing a quantifiable metric for intrinsic rewards in DRL. Additionally, we propose a novel variational weighting mechanism that dynamically assigns importance to different snapshots, thereby enhancing the overall performance and robustness of the model.

In brief, our main contributions include.

1) We introduce the TIR, a method that employs a weighted nuclear norm to gauge the divergence between the predictions of various model snapshots, effectively utilizing these measures as intrinsic rewards in DRL.

2) We devise a variational weighting mechanism for the snapshots, aimed at refining performance and bolstering the model’s resilience to noise. This mechanism is not only straightforward to implement but also remarkably potent.

3) Through extensive benchmarks, we validate that TIR surpasses existing methodologies across diverse settings, including the DeepMind Control Suite (DMC) [24] and Atari games [1]. TIR’s superiority is particularly pronounced in Atari games, and it achieves state-of-the-art results in 8 out of 12 tasks within the DMC, outperforming both intrinsic reward-based approaches and other pretraining strategies.

The remainder of this article is organized as follows: In Section II, we review the related work about the intrinsic reward and low-rank regularization (LRR). In Section III, we introduce a novel intrinsic reward method named TIR and the mechanism that enables the agent to explore the environment by itself more efficiently. Further, in Section IV, we show the performance of and analyze the main components in Section IV-D, which illustrates that TIR can achieve promising performance and robustness compared to strong baselines. The main conclusion are covered in Section V.

II. Related Work

We will discuss the related work of intrinsic rewards and LRR in this section. We also talk about the differences and connections between our method and previous related works.
A. Intrinsic Reward-Based Exploration

In DRL, many previous exploratory attempts employ prediction error [19], prediction uncertainty [13], or variants of uncertainty as self-driving motivations in the learning phase [25], [26], [27]. Most intrinsic reward-based methods can be interpreted as exploring areas of high uncertainty, and curiosity is measured by estimating the deep learning model’s uncertainty (confidence). Some attempts use count-based [28], [29] or pseudocount-based exploration [15], which struggle with the scalability to higher-dimensional state spaces [14], [15], [16]. Motivated by counting, random network distillation (RND) [8] proposes to measure the gap between the predictor and the target as an intrinsic reward, which is easy to implement and can be more efficient. Meanwhile, there are some works to self-generate goals and guide the agent to improve the exploration and sample efficiency [30], [31], [32]. They also prove the effectiveness of intrinsic rewards.

Recently, curiosity via self-supervised prediction [9], [13], [19], [21] is one prominent way to self-supervised exploration. The prediction-based intrinsic reward is defined as the error between the predicted state and the ground truth of the next state. However, accurate prediction can be difficult to obtain in practical settings. For this reason, disagreement [13] proposes to use the variance of predictions by ensembling multiple models’ predictions, instead of using a single model. Deep ensemble [33], [34] technique is widely used to estimate the uncertainty in DRL, which proposes to train an ensemble of deep models rather than a single model and achieves superior results but is also computationally expensive.

However, in prediction model-based methods, training multiple deep networks can be extremely time-consuming and may introduce more stochasticity due to different initial parameters, which may cause inaccurate intrinsic rewards. More importantly, prediction-based methods mainly evaluate the intrinsic rewards by only using the current prediction. Our method also employs a self-supervised exploration paradigm but relies on the temporal inconsistency between multiple snapshots. The set of snapshots is essentially the deep ensemble on the temporal dimension and can avoid the variance introduced by stochastic multiple parameters without additional training costs. With multiple generations of the predictive model, the intrinsic reward can be more accurate and better encourage agents to investigate more novel states using the temporal information [12], [35]. It has been proved that the temporal information can be dedicated to improve the efficiency in the agent learning process [36], [37], [38], [39] and has the potential to be an efficient way to improve the overall performance in a wide range of scenarios. Our work makes this point clearer and prove that the temporal information hidden in the checkpoint can be used to access intrinsic rewards more accurately.

B. LRR

This subsection discusses the relationship between the proposed method and the LRR, as we utilize the weighted nuclear norm to measure the temporal inconsistency between multiple snapshots. Low-rank matrices can be found in multiple fields, such as computer vision and recommendation systems [40], [41]. Hu et al. [42] provide a comprehensive survey of LRR, focusing on the impact of using LRR as a loss function via nuclear norm. Xiong et al. [43] apply the nuclear norm to avoid prediction diversity collapse and improve the generalization ability of the classification model. The nuclear norm (NN) is a popular convex surrogate function of the rank function widely used for LRR. Essentially, compared to the $f^2$ norm, which is widely used in previous attempts to measure the distance, the nuclear norm not only encourages diversity but also distinguishes irrelevance. In contrast to previous attempts, our work aims to maximize the weighted nuclear norm of the prediction matrix introduced by multiple snapshots of the prediction model.

III. METHODOLOGY

In this section, we first introduce the details of temporal inconsistency-based curiosity, then describe how to assign the weights to different snapshots via the proposed variational weighting mechanism.

We first train a prediction model to predict the next state $s_{t+1}$ with given state-action pairs $\{s_t, a_t\}$

$$\hat{s}_{t+1} = f(s_t, a_t; \theta_p)$$

(1)

where $\hat{s}_{t+1}$ is the prediction and the network parameter $\theta_p$ is optimized by minimizing the loss function $L_p$

$$L_p = \frac{1}{2} \| \hat{s}_{t+1} - s_{t+1} \|^2_2.$$  

(2)

A. Temporal Inconsistency-Based Curiosity

During the training phase, we save a snapshot of the model parameters along the optimization path, for every $j$ epochs. Therefore, as shown in Fig. 3, we can get $n$ models in temporal order, which can be denoted as $\{f_1, f_2, \ldots, f_n\}$. The sliding window strategy is used, and we replace the older model using the new one. Given a state-action pair $[s_t, a_t]$, we can obtain a
set of predicted values $\delta_{t+1}$ for the next state using $n$ snapshots of the predictive model, then we use the prediction matrix $P$ containing the predicted values to evaluate intrinsic rewards.

For a subset of the state space that the agent has been thoroughly explored, most of the snapshots can show similar prediction ability after sufficient sampling and training, so higher consistency can be obtained across multiple snapshots. For underexplored parts, the snapshots show high prediction errors and uncertainties because of insufficient training and different parameters, resulting in higher inconsistency over time along the temporal order. Obviously, the stronger the temporal inconsistency between predicted values are, the less exploration of state-action pairs can be, which has been validated in the toy example (Fig. 2). However, how to assess the temporal inconsistency of predicted values is also an under-explored topic.

### B. Measuring Inconsistency Using Weighted Nuclear Norm

To measure the distance between multiple vectors, different norms can be used, such as the $\ell_1$, $\ell_2$, and $\ell_F$. We utilize the nuclear norm to measure the inconsistency between different snapshots, due to the superior performance and high noise tolerance. Moreover, we propose a novel weighting mechanism with the goal of assigning different weights to the snapshots.

1) **Nuclear Norm-Based Intrinsic Reward:** For disagreement [13], the inconsistency of prediction models can be measured by the variance of the predictions. Unlike the disagreement method, which trains multiple models in parallel, we only train one prediction model and get $n$ snapshots of the model for free without incurring additional training costs. Concretely, the prediction matrix $P$ consists of $n$ predictions from the snapshots, and each prediction is a vector of $m$-dimension. Thus, $P$ is with the size of $n \times m$. As each row of the matrix represents the probability distribution of the next state’s prediction, higher inconsistency means greater differences between the rows of the matrix $P$. Hence, a naive way to measure the prediction inconsistency is the matrix rank($P$), which can represent the linear irrelevance between the rows. Accordingly, we can generate our intrinsic reward by maximizing rank($P$), with the goal of encouraging the agent to explore the states with higher temporal inconsistency.

However, directly maximizing the rank($P$) is an NP-hard, nonconvex problem as the value of matrix rank is discrete. Thus, rank($P$) cannot be employed directly for the DRL purpose. Moreover, due to the contamination of the noise, such as the environment stochasticity, the prediction values can be inaccurate, which can incur additional noise. A flurry of studies suggest that the matrix rank can be replaced by the nuclear norm, which is the convex envelope of rank($P$). Specifically, the nuclear norm can be formulated as

$$\|P\|_* = \text{tr} \left( \sqrt{P^T P} \right) = \sum_{i=1}^{D} \sigma_i$$  \hspace{1cm} (3)

where $\text{tr}(.)$ refers to the trace of the $P$, and $D = \min(n, m)$, $\sigma_i$ denotes the $i$th largest singular value of $P$. Mathematically, the nuclear norm and Frobenius norm ($F$ norm) are the boundaries of each other, and the $F$-norm can be formulated as follows:

$$\|P\|_F = \sqrt{\sum_{i=1}^{n} \sum_{j=1}^{n} |P_{i,j}|^2}.$$  \hspace{1cm} (4)

According to [44], [45], the mutual boundary relationships between $\|P\|_*$ and $\|P\|_F$ could be formulated as follows:

$$\frac{1}{\sqrt{D}} \|P\|_* \leq \|P\|_F \leq \sqrt{D} \cdot \|P\|_*.$$  \hspace{1cm} (5)

Furthermore, Cui et al. [46] proved that $\|P\|_F$ is strictly opposite to Shannon entropy in monotony, and maximizing $\|P\|_F$ is equal to minimizing entropy. So, the nuclear norm can assess not only the diversity but also distinguish irrelevance, which can be deployed to measure the inconsistency of the prediction matrix $P$.

It is well-known that the scale of reward value is also a key factor for DRL. Accordingly, to adapt the intrinsic reward to a suitable range, we introduce the parameter $\lambda$ to scale $\|P\|_*$, to the same scale of the $\|P\|_F$. Thus, we can define our intrinsic reward as

$$j^\text{int} = \lambda \|P\|_*$$  \hspace{1cm} (6)

and the parameter $\lambda$ is set as 0.001 for all experiments.

2) **Variational Weighting Mechanism:** In our practical implementations, we find that the early snapshots of the model may be underfitting and produce inaccurate predictions. The inaccurate predictions will result in large singular values of the matrix, which is a perturbation for generating accurate intrinsic rewards. However, the standard nuclear norm treats each singular value by the averaging weight, which leads to suboptimal performance. As a result, we propose an adaptive weight of the singular values to avoid the excessive influence of under-fitting snapshots. Here, a smaller weight will be assigned to the larger singular value. The weighted nuclear norm can be formulated as follows:

$$\|P\|_{w,*} = \sum_{i=1}^{D} w_i \sigma_i$$  \hspace{1cm} (7)

where $w_i \geq 0$ is a nonnegative weight for $\sigma_i$.

Nevertheless, in practice, it can be difficult to manually design the weight $w_i$ for each singular value in every specific scenario. Here, we propose an adaptive weighted nuclear norm according to singular values themselves.

$$\|P\|_{k,*} = \sum_{i=1}^{D} \sigma_i^{1/k}$$  \hspace{1cm} (8)

where $k = \{2, 3, \ldots\}$ and $\sigma_i^{(1-k)/k}$ is treated as the weight for $\sigma_i$. It’s obvious that in this formula the bigger $\sigma_i$ is, the smaller the weight is, which is in line with our principle. Besides that, the greater $k$, the more severe the inhibition on the large singular value. The adaptive weighted nuclear norm can avoid overestimation of intrinsic rewards due to insufficient prediction ability. However, as training continues, the predictive ability of the snapshots can be improved, which can provide more accurate
Algorithm 1: Temporal Inconsistency-based Intrinsic Rewards

**Initialization:** policy network \( \pi_{\xi}(a|s) \), prediction model \( f_p \), the number of snapshots \( N \), the training step counter \( c \), the maximum episode step \( T \), snapshot models \( \{f_1, f_2, \ldots, f_n\} \), coefficient of intrinsic reward \( \alpha \), coefficient of extrinsic reward \( \beta \), coefficient \( k_{ini} \), coefficient \( k \), and interval \( j \).

1: \( c = 0 \)
2: while \( c < 5c^j \) do
3: for \( t = 1, \ldots, T \) do
4: \( c = c + 1 \)
5: Receive observation \( s_t \) from environment
6: \( a_t \leftarrow \pi_{\xi}(a|s) \) based on policy network \( \pi_{\xi} \)
7: Take action \( a_t \), receive observation \( s_{t+1} \) and extrinsic reward \( r_{ext}^{ext} \) from environment
8: if the episode is finished then
9: Store episodic data and end for
10: end if
11: \( s_t \leftarrow s_{t+1} \)
12: end for
13: Sample batch data as \( \{(s_i, a_i, r_{ext}^{ext}, s_{i+1})\}_{i=1}^N \) from reply buffer
14: for each \( i = 1, \ldots, N \) do
15: Predict the next state \( \hat{s}_{i+1} = f_1(s_i, a_i), \hat{s}_{j+1} = f_j(s_j, a_j) \)
16: Form matrix \( P \) by concatenating each state vector \( \hat{s}_{i+1}, \hat{s}_{j+1}, \ldots, \hat{s}_{i+1} \)
17: Calculate intrinsic reward \( r_i^{int} = \lambda \|P\|_{k_x} = \lambda \sum_{i=1}^D \sigma_i^{1/k} \)
18: Calculate total reward \( r_i^{total} = \alpha r_i^{int} + \beta r_i^{ext} \)
19: end for
20: Update \( f_p \) with sampled data by minimizing loss with Equation 2
21: Every \( j \) epochs, replace the earliest snapshot model with \( f_p \)
22: \( k = \text{max}(1, k_{ini} - j) \)
23: Update \( \xi \) with sampled data by maximizing \( r^{total} \) using RL algorithm
24: end while

predictions. As a result, the inhibition on large singular values should be reduced, and the parameter \( k \) should be variable, gradually decreasing to 1 as the experiment progresses.

To conclude, the whole intrinsic reward can be formulated as the following equation, leveraging the variational weighting nuclear norm:

\[
p^{int} = \lambda \|P\|_{k_x} = \lambda \sum_{i=1}^D \sigma_i^{1/k}. \tag{9}
\]

Thus, we can get the optimization goal of the agent

\[
\max \xi \pi_{\pi}(S_i; \xi) \sum_{i=1}^N \gamma^j (\alpha r_i^{int} + \beta r_i^{ext}) \tag{10}
\]

where \( \gamma \) is the discount factor and \( \xi \) represents parameters of policy \( \pi \), \( \alpha \), and \( \beta \) are the coefficient of intrinsic reward and extrinsic reward, respectively. Fig. 3 and Algorithm 1 present the whole framework and pseudocode of TIR.

IV. EXPERIMENTAL RESULTS

A. Experimental Settings

To conduct a fair quantitative comparison with competing approaches, we evaluate the proposed method using two widely used benchmarks, including the DMC Suite [24] and the Atari games [1]. We perform the evaluation using two settings: pre-training accompanying fine-tuning and traditional RL settings. For the Atari environment, we follow the settings of disagreement [13] and settings of unsupervised reinforcement learning benchmark (URLB) [35], which is a pretraining benchmark for DMC. We utilize the proximal policy optimization (PPO) algorithm [47] to train our agent within the Atari Suite and achieve this by leveraging distributed agents to gather experience from the environment. For our experiments within the Atari Suite, we execute them over 30 000 rollouts, with each rollout consisting of 128 steps per environment. This process occurs across 128 parallel environments, and we set the maximum number of episode steps to 4500. In the context of the DMC, and to ensure equitable comparisons, we adhere to the benchmark implementation established by the URLB [35] and apply the deep deterministic policy gradient (DDPG) algorithm [48]. The learning rate is set to 0.0001 in all experiments and all comparison experiments use the same settings, and we run experiments on the device with an NVIDIA RTX 3090 GPU. Our algorithm costs about an average of 8.2 h per training in Atari games and about an average of 2.2 h per training in DMC. In all experiments, the number of snapshots is set to 5, and the hyperparameters \( j \) and \( k \) are set to 4 and 2, respectively.

We run the experiments across three random seeds and employ 50M running steps—equivalent to 200M frames in Atari. The hyperparameters \( \gamma \), \( \alpha \), and \( \beta \) are set to 0.99, 1, and 0.

B. TIR Surpasses Previous Curiosity-Based Methods

For the Atari environment, we evaluate agents’ performance by only using intrinsic rewards. In other words, they are only motivated by intrinsic rewards for self-supervised exploration. Table I lists the aggregate metrics and scores of four competitive intrinsic rewards (intrinsic curiosity module (ICM), disagreement, RND, and our proposed TIR), on the subset of Atari games. Human and random scores are adopted from [49]. Following the evaluation settings of previous works [12], [50], [51], we normalize the episode reward as human-normalized scores (HNS), which can be formulated as

\[
\text{HNS} = \frac{\text{agent score} - \text{random score}}{\text{human score} - \text{random score}}. \tag{11}
\]

#SOTA means the game numbers exceed other methods, and the mean HNS is the average of HNS across all games. TIR displays an overwhelming superiority over ICM, disagreement, and RND with its highest mean HNS and #SOTA. Furthermore, the mean HNS score of TIR in a wide range of Atari games represents a 114% relative improvement over other baselines, and our intrinsic reward-based method even outperforms the human level in some scenarios without any feedback from the external environment. Among all 21 games, TIR outperforms ICM in 17 games, disagreement in 19 games, and RND in 17 games. These results underscore the robustness and efficacy of TIR in a majority of the evaluated scenarios. The disagreement method, which operates by ensembling multiple models to capture uncertainty, serves as an important point of comparison. In contrast, TIR leverages a temporal ensemble of models, capitalizing on historical information to derive a more
Table I

| Game      | Random | Human | ICM | Disagreement | RND | TIR |
|-----------|--------|-------|-----|---------------|-----|-----|
| Alien     | 227.8  | 7127.7| 374.2| 316.6         | 206.1| 532.7|
| Amidar    | 5.8    | 1719.5| 99.4 | 91.6          | 308.9| 41.3 |
| Assault   | 222.4  | 742.0 | 59.4 | 366.3         | 644.5| 7580.0|
| Asterix   | 210.0  | 8503.3| 2367.6| 540.0         | 848.5| 738.0|
| Bank Heist| 14.2   | 753.1 | 96.8 | 113.7         | 16.6 | 425.4|
| BattleZone| 2360.0 | 37187.5| 2794.5| 3663.3         | 6445.0| 7580.0|
| Breakout  | 1.7    | 30.5  | 264.2| 246.6         | 35.5 | 269.2|
| ChopperCommand | 811.0 | 7387.8| 122.4| 371.0         | 320.0| 825.0|
| Demon Attack | 107805.0 | 35829.4| 30.9 | 25.3          | 36.1 | 44.1 |
| Freeway   | 0.0    | 29.6  | 0    | 0.2           | 3.1  | 0.3  |
| Gopher    | 257.6  | 2412.5| 2763.8| 2456.6         | 1054.8| 5921.6|
| Hero      | 1027.0 | 30826.4| 2559.2| 2749.7         | 2130.4| 2889.0|
| Jamesbond | 29.0   | 302.8 | 494.9| 308.8         | 365.5| 4775.0|
| Kangaroo  | 52.0   | 3035.0| 557.0| 514.0         | 412.0| 1682.8|
| Kung Fu Master | 258.5 | 22736.3| 4226.9| 2179.5         | 11350| 10844.1|
| Ms Pacman | 307.3  | 6951.6| 412.7| 291.0         | 607.2| 997.7 |
| Pong      | -20.7  | 14.6  | -9.1 | -7.9          | -11.7| -4.2 |
| Private Eye | 24.9  | 69571.3| -500 | 36.0          | -997.5| 0.0  |
| Qbert     | 1639.0 | 13455.0| 531.3| 616.0         | 1349.0| 646.2|
| Seaquest  | 68.4   | 42054.7| 471.9| 303.6         | 357.8| 646.2|
| Up N Down | 533.4  | 11693.2| 15815.1| 8189.8        | 8736.1| 18314.2|
| Mean HNS  | 0.0    | 1.0   | 0.767 | 0.627 | 0.284 | 1.647 |
| #SOTA     | N/A    | N/A   | 3     | 1           | 2    | 15   |

Note: The bold font indicates the best scores.

Fig. 4. Performance comparison on subsets of Atari games by only using the intrinsic reward. Five competitive methods are compared, including disagreement, ICM, RND, and our proposed TIR.

nuanced and accurate intrinsic reward signal. The experimental outcomes corroborate our hypothesis, demonstrating that the incorporation of historical knowledge allows TIR to provide a more reliable measure of curiosity. This is evidenced by the consistent outperformance of TIR across the majority of games, suggesting that the temporal dynamics and historical context are critical components in the formulation of intrinsic rewards.

Fig. 4 provides the learning curves of TIR with four baselines on eight randomly selected Atari games. Of the eight games, our method outperforms the disagreement in eight, the ICM in six, and the RND in eight. TIR has clearly established benefits in terms of performance and learning speed in most games, as seen in the figure. Particularly on the Jamesbond and MsPacman, the convergent episode reward of TIR is more than double that of the other approaches. In brief, our method outperforms other intrinsic reward-based methods in Atari Suite, demonstrating TIR’s superiority in providing more accurate intrinsic rewards, which is critical in self-supervised exploration. Meanwhile, through the figure, we can compare the confidence intervals of different approaches. This comparative visualization is expected to yield insights into the relative stability of each method, as indicated by the overlap and width of their respective confidence intervals. Preliminary observations suggest that the variance in performance of our method is comparable to, if not better than, the benchmark techniques. The compact nature of TIR’s confidence intervals implies a high degree of reliability in its performance, as it indicates that the mean performance is a stable and consistent estimator of the method’s true merit. The analysis of confidence intervals also substantiate the robustness of our method.
reward-based methods (ICM, disagreement, RND, APT) [12] and other pretraining strategies (ProtoRL [52], SMM [53], DIAYN [54], and APS [55]). The quantitative results in Table II clearly show that TIR achieves state-of-the-art results in 8 of 12 tasks, and the average performances in all domains surpass all the other methods. Especially, in the hardest domain, our approach improves average performance by 42.0%, which demonstrates TIR’s great potential to improve model performance and robustness in the pretraining paradigm. Furthermore, an analysis of the overall quantitative standard deviations reveals that our TIR method exhibits significantly lower variability in its outcomes compared to alternative approaches. This reduced variance underscores the superior stability and reliability of TIR, thereby highlighting its robustness as an intrinsic reward mechanism in reinforcement learning scenarios. The consistency of TIR’s performance across various metrics serves as compelling evidence of its ability to maintain efficacy under diverse conditions, further validating its superiority.

To conclude, large-scale experimental results successfully confirm our analysis that the prediction error may be noisy and inaccurate. Disagreement [13] seeks to address the issue by introducing multiple models to access model uncertainty and generate the lower variance estimate. However, due to different initial parameters and stochasticity, it is extremely difficult to compare the final scores and standard deviations of TIR, with the intrinsic reward-based methods (ICM, disagreement, RND, APT [12]).

C. TIR Surpasses Previous Pretraining Strategies

The pretraining accompanying fine-tuning paradigm is widely used in the deep learning field, and it is also the key technique to improve sample efficiency for DRL [35]. DMC environment encompasses a variety of difficult domains and tasks of different complexity, and it is built on robot simulations, which makes modeling the robot dynamics inherently tough. We conduct extensive experiments on the DMC to demonstrate the superiority of the TIR based on the pretraining paradigm. We evaluate all baselines and TIR from the easiest to the most difficult domains (Walker, Quadruped, and Jaco), which each have four tasks. During the pretraining phase, an agent is trained for 2 million steps with only intrinsic reward and only 100k steps with extrinsic reward during the fine-tuning stage.

Fig. 5 plots the learning curves (fine-tuning phase) of the aforementioned four methods. As shown in the figure, TIR not only provides a faster convergence speed than other intrinsic reward-based methods but the convergence result also significantly outperforms baselines. We suppose the improvements in the convergence speed rest from historical information provided by multiple snapshots. Additionally, we compare the final scores and standard deviations of TIR, with the intrinsic reward-based methods (ICM, disagreement, RND, APT [12]) and other pretraining methods (ProtoRL [52], SMM [53], DIAYN [54], and APS [55]). The quantitative results in Table II clearly show that TIR achieves state-of-the-art results in 8 of 12 tasks, and the average performances in all domains surpass all the other methods. Especially, in the hardest domain, our approach improves average performance by 42.0%, which demonstrates TIR’s great potential to improve model performance and robustness in the pretraining paradigm. Furthermore, an analysis of the overall quantitative standard deviations reveals that our TIR method exhibits significantly lower variability in its outcomes compared to alternative approaches. This reduced variance underscores the superior stability and reliability of TIR, thereby highlighting its robustness as an intrinsic reward mechanism in reinforcement learning scenarios. The consistency of TIR’s performance across various metrics serves as compelling evidence of its ability to maintain efficacy under diverse conditions, further validating its superiority.

To conclude, large-scale experimental results successfully confirm our analysis that the prediction error may be noisy and inaccurate. Disagreement [13] seeks to address the issue by introducing multiple models to access model uncertainty and generate the lower variance estimate. However, due to different initial parameters and stochasticity, it is extremely difficult to compare the final scores and standard deviations of TIR, with the intrinsic reward-based methods (ICM, disagreement, RND, APT [12]).

C. TIR Surpasses Previous Pretraining Strategies

The pretraining accompanying fine-tuning paradigm is widely used in the deep learning field, and it is also the key technique to improve sample efficiency for DRL [35]. DMC environment encompasses a variety of difficult domains and tasks of different complexity, and it is built on robot simulations, which makes modeling the robot dynamics inherently tough. We conduct extensive experiments on the DMC to demonstrate the superiority of the TIR based on the pretraining paradigm. We evaluate all baselines and TIR from the easiest to the most difficult domains (Walker, Quadruped, and Jaco), which each have four tasks. During the pretraining phase, an agent is trained for 2 million steps with only intrinsic reward and only 100k steps with extrinsic reward during the fine-tuning stage.

Fig. 5 plots the learning curves (fine-tuning phase) of the aforementioned four methods. As shown in the figure, TIR not only provides a faster convergence speed than other intrinsic reward-based methods but the convergence result also significantly outperforms baselines. We suppose the improvements in the convergence speed rest from historical information provided by multiple snapshots. Additionally, we compare the final scores and standard deviations of TIR, with the intrinsic reward-based methods (ICM, disagreement, RND, APT [12]) and other pretraining strategies (ProtoRL [52], SMM [53], DIAYN [54], and APS [55]). The quantitative results in Table II clearly show that TIR achieves state-of-the-art results in 8 of 12 tasks, and the average performances in all domains surpass all the other methods. Especially, in the hardest domain, our approach improves average performance by 42.0%, which demonstrates TIR’s great potential to improve model performance and robustness in the pretraining paradigm. Furthermore, an analysis of the overall quantitative standard deviations reveals that our TIR method exhibits significantly lower variability in its outcomes compared to alternative approaches. This reduced variance underscores the superior stability and reliability of TIR, thereby highlighting its robustness as an intrinsic reward mechanism in reinforcement learning scenarios. The consistency of TIR’s performance across various metrics serves as compelling evidence of its ability to maintain efficacy under diverse conditions, further validating its superiority.

To conclude, large-scale experimental results successfully confirm our analysis that the prediction error may be noisy and inaccurate. Disagreement [13] seeks to address the issue by introducing multiple models to access model uncertainty and generate the lower variance estimate. However, due to different initial parameters and stochasticity, it is extremely
difficult for models to achieve an agreement, resulting in a worse performance than ICM. Instead, we reuse previous snapshots as historical knowledge to avoid additional stochasticity (such as initial parameters), which is also employed by averaged-DQN [56] to estimate the Q value. As a result, our curiosity formulation outperforms the others, and TIR shows significant performance improvement without any additional training costs. There is also convincing evidence showing that when pretrained with the prediction model, the pretrained policy for 500k steps outperforms the pretrained policy for 2M steps in downstream tasks, indicating that temporal information should be considered [57].

D. Further Analysis

In this subsection, we mainly investigate the robustness of TIR with different settings.

1) Tolerance of Ensemble Size: The number of snapshots, which is the size of the ensemble, is a critical TIR setting. To investigate the sensitivity to ensemble size, we compare TIR performance with a different number of snapshots of ensemble size $n$. The performance comparison is shown in Fig. 6, and we can observe that TIR with different settings all show promising performance, demonstrating robustness. Especially, when the number of snapshots is set to five, the method exhibits the most pronounced reduction in variability, as evidenced by the significantly smaller error bars in our performance metrics. This observation suggests that an optimal number of snapshots can substantially enhance the consistency and reliability of the TIR method. To balance the performance and computation costs, in this article, we set the number of snapshots as five.

2) Tolerance of Noise: As previously stated, predicting the future is extremely difficult, particularly in the presence of noise. We introduce 0.25 standard deviation Gaussian noise to the feature vector to verify the robustness of baselines and TIR [25]. Fig. 7 plots the performance of agents perturbed with and without noise across two games. We can find that the performance of Disagreement degrades apparently in both two games and the performance of ICM also degrades. On the contrary, TIR with noise still shows great performance compared to baselines, which proves the robustness of TIR over Disagreement and ICM.

3) Sensitivity Analysis for Hyperparameters: In TIR, there is a tradeoff between updating the snapshots too often and updating the snapshots too slowly, and hyperparameter $j$ denotes the temporal gap between different snapshots. To analyze the sensitivity, we evaluate TIR with different settings hyperparameter $j$. As illustrated in Table III, all values of the hyperparameter $j$ between 4 and 7 provide a significant advantage over both baselines and human scores, evidently proving the stability and robustness of TIR. Additionally, we can observe the wild variation in performance between the different values of $j$ in the Jamesbond scenario. It’s because when $j$ is too big, the snapshot parameters may be very different, and the snapshot ensemble can be regarded as a variant of the ensemble in Disagreement, which limits the performance of TIR. Through meticulous investigation, we have ascertained that setting the hyperparameter $j$ to either 4 or 5 yields optimal results in terms of overall performance and stability.

The hyperparameter $k$ can adjust the weight of singular values, and the bigger $k$ is, the smaller the weight is. We set
scores in all different settings of TIR demonstrate promising performance. By comparing the nuclear norm and fix weighted nuclear norm settings, we can validate our hypothesis that inaccurate predictions result in large singular values of the prediction matrix and inaccurate rewards, and we need to avoid the influence of underfitting snapshots with the adaptive weighted nuclear norm. For this reason, in Jamesbond and MsPacman, the scores of TIR exhibit significant advantages over both baselines and other settings, indicating the importance of temporal information and the effectiveness of the variational weighting mechanism.

V. CONCLUSION

To address the challenge of sparse rewards, we propose a simple yet effective intrinsic reward for DRL. Specifically, we train a self-supervised prediction model and maintain a set of snapshots of the prediction model parameters for free. We further utilize the nuclear norm to evaluate the temporal inconsistency between the predictions of different snapshots, which can be further deployed as the intrinsic reward. Moreover, the variational weighting mechanism is proposed to assign weights to different snapshots adaptively. Our solution does not incur additional training costs while maintaining higher noise tolerance. TIR shows a promising advantage in two widely used benchmarks compared to other competing intrinsic reward-based methods. In the future, we will further investigate improving the prediction performance with more trajectory information instead of using only the information from one transition. Overall, our research presents a valuable framework that enhances exploration and learning across various environments, with broad applications for improving agent performance in numerous real-world scenarios (such as navigation systems). The core idea and implementations of utilizing temporal information can motivate more relevant works in DRL. Looking ahead, we intend to investigate how to augment the efficiency of our approach through an array of ensemble techniques and how to utilize temporal information inherent in state sequences.

REFERENCES

[1] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, “The arcade learning environment: An evaluation platform for general agents,” J. Artif. Intell. Res., vol. 47, pp. 253–279, Jun. 2013.
[2] G. Tesauro, “Temporal difference learning and TD-Gammon,” Commun. ACM, vol. 38, no. 3, pp. 58–68, 1995.
[3] V. Mnih et al., “Human-level control through deep reinforcement learning,” Nature, vol. 518, no. 7540, pp. 529–533, 2015.
[4] D. Li, F. Zhu, J. Wu, Y. D. Wong, and T. Chen, “Managing mixed traffic at signalized intersections: An adaptive signal control and CA V coordination system based on deep reinforcement learning,” Expert Syst. Appl., vol. 238, Mar. 2024, Art. no. 121959.
[5] L. de Azevedo Takara, A. A. P. Santos, V. C. Mariani, and L. dos Santos Coelho, “Deep reinforcement learning applied to a sparse-reward trading environment with intraday data,” Expert Syst. Appl., vol. 238, Mar. 2024, Art. no. 121897.
[6] G. Papadopoulos et al., “Deep reinforcement learning in service of air traffic controllers to resolve tactical conflicts,” Expert Syst. Appl., vol. 236, Feb. 2024, Art. no. 121234.
[7] H. Zhao, C. Dong, J. Cao, and Q. Chen, “A survey on deep reinforcement learning approaches for traffic signal control,” Eng. Appl. Artif. Intell., vol. 133, Jul. 2024, Art. no. 108100.
[8] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, “Exploration by random network distribution,” in Proc. Int. Conf. Learn. Representations, 2019.

[9] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, “Large-scale study of curiosity-driven learning,” in Proc. Int. Conf. Learn. Representations, 2018.

[10] R. M. Ryan and E. L. Deci, “Intrinsic and extrinsic motivations: Classic definitions and new directions,” Contemporary Educational Psychol., vol. 25, no. 1, pp. 54–67, 2000.

[11] L. Smith and M. Gasser, “The development of embodied cognition: Six lessons from babies,” Artif. Life, vol. 11, nos. 1–2, pp. 13–29, 2005.

[12] H. Liu and P. Abbeel, “Behavior from the void: Unsupervised active pre-training,” in Proc. Adv. Neural Inf. Process. Syst., vol. 34, 2021, pp. 18459–18473.

[13] D. Pathak, D. Gandhi, and A. Gupta, “Self-supervised exploration via disagreement,” in Proc. Int. Conf. Mach. Learn., PMLR, 2019, pp. 5062–5071.

[14] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, “Unifying count-based exploration and intrinsic motivation,” Adv. Neural Inf. Process. Syst., vol. 29, pp. 1471–1479, Dec. 2016.

[15] J. Martin, S. S. Narayanan, T. Everitt, and M. Hutter, “Count-based exploration in feature space for reinforcement learning,” in Proc. 26th Int. Joint Conf. Artif. Intell., 2017, pp. 2471–2478.

[16] G. Ostrovski, M. G. Bellemare, A. Oror, and R. Munos, “Count-based exploration with neural density models,” in Proc. Int. Conf. Mach. Learn., PMLR, 2017, pp. 2721–2730.

[17] M. Lopes, T. Lang, M. Toussaint, and P.-Y. Oudeyer, “Exploration in model-based reinforcement learning by empirically estimating learning progress,” in Proc. Adv. Neural Inf. Process. Syst., vol. 25, 2012, pp. 206–214.

[18] P. Poupaut, N. Vlassis, J. Hoey, and K. Regan, “An analytic solution to discrete Bayesian reinforcement learning,” in Proc. 23rd Int. Conf. Mach. Learn., 2006, pp. 697–704.

[19] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction,” in Proc. Int. Conf. Mach. Learn., PMLR, 2017, pp. 2778–2787.

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

[21] K. Kim, M. Sano, J. De Freitas, N. Haber, and D. Yamins, “Active world model learning with progress curiosity,” in Proc. Int. Conf. Mach. Learn., PMLR, 2020, pp. 5306–5315.

[22] J. I. Rotgans and H. G. Schmidt, “The role of interest in learning: Knowledge acquisition at the intersection of situational and individual interest,” in The Science of Interest. Cham, Switzerland: Springer, 2017, pp. 69–93.

[23] L. Liang et al., “Reducing variance in temporal-difference value estimation via ensemble of deep networks,” in Proc. Int. Conf. Mach. Learn., PMLR, 2022, pp. 13285–13301.

[24] S. Tunyasuvunakool et al., “DM control: Software and tasks for continuous control,” Softw. Impacts, vol. 6, Nov. 2020, Art. no. 100022.

[25] V. Dean, S. Tulsiani, and A. Gupta, “See, hear, explore: Curiosity via audio-visual association,” Adv. Neural Inf. Process. Syst., vol. 33, 2020, pp. 5306–5315.

[26] R. Houthooft, X. Chen, Y. Duan, J. Schulman, D. F. Turck, and F. Mannella, V. G. Santucci, E. Somogyi, L. Jacquey, K. J. O’Regan, and G. Baldassarre, “Know your body through intrinsic goals,” Frontiers Neurorobot., vol. 12, 2018.

[27] A. Péret, S. Forestier, O. Sigaud, and P.-Y. Oudeyer, “Unsupervised learning of goal spaces for intrinsically motivated goal exploration,” 2018, arXiv:1803.00781.

[28] T. G. Dietterich, “Ensemble methods in machine learning,” in Proc. Int. Workshop Multiple Classifier Syst., Berlin, Heidelberg: Springer, 2000, pp. 1–15.

[29] J. Wu and H. Li, “Deep ensemble reinforcement learning with multiple deep deterministic policy gradient algorithm,” Math. Problems Eng., vol. 2020, 2020.

[30] M. Laskin et al., “URLB: Unsupervised reinforcement learning benchmark,” in Proc. Deep RL Workshop NeurIPS, 2021.

[31] P.-Y. Oudeyer, F. Kaplan, and V. Hafner, “Intrinsic motivation systems for autonomous mental development,” IEEE Trans. Evol. Comput., vol. 11, no. 2, pp. 265–286, Feb. 2007.

[32] M. B. Hafez, C. Weber, M. Kerzel, and S. Wermter, “Deep intrinsically motivated continuous actor-critic for efficient robotic visuomotor skill learning,” 2018, arXiv:1810.11388.

[33] M. Luciw, V. Graziano, M. Ring, and J. Schmidhuber, “Artificial curiosity with planning for autonomous perceptual and cognitive development,” in Proc. IEEE Int. Conf. Develop. Learn. (ICDL), vol. 2, 2011, pp. 1–4.

[34] M. Burhan Hafez and S. Wermter, “Behavior self-organization supports task inference for continual robot learning,” in Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS), 2021, pp. 6739–6746.

[35] J.-B. Li, S.-Y. Lin, Y.-H. Hsu, and Y.-C. Huang, “An empirical study of alternating least squares collaborative filtering recommendation for movielens on Apache Glodon and spark,” Int. J. Grid Util Comput., vol. 11, no. 5, pp. 674–682, 2020.

[36] M. Udell and A. Townsend, “Why are big data matrices approximately low rank?” SIAM J. Math. Data Sci., vol. 1, no. 1, pp. 144–160, 2019.

[37] Z. Hu, F. Nie, R. Wang, and X. Li, “Low rank regularization: A review,” Neural Netw., vol. 136, pp. 218–232, Apr. 2021.

[38] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and I. R. Rehg, “Counterfactual rewards via disagreement,” in Proc. Int. Conf. Mach. Learn., 2021, pp. 6739–6746.

[39] M. Schwarzer, A. Anand, R. Goel, R. D. Hjelm, A. Courville, and R. Salakhutdinov, “Efficient exploration via state marginal matching,” in Proc. 25th Conf. Neural Inf. Process. Syst., vol. 1, no. 1, pp. 144–160, 2018.

[40] M. Hessel et al., “Rainbow: Combining improvements in deep reinforcement learning,” in Proc. 32nd AAAI Conf. Artif. Intell., 2018, pp. 3215–3222.

[41] D. Yarats, I. Kostrikov, and R. Fergus, “Image augmentation is all you need: Regularizing deep reinforcement learning from pixels,” in Proc. Int. Conf. Learn. Representations, 2020.

[42] M. Schwarzer, A. Anand, R. Goel, R. D. Hjelm, A. Courville, and R. Bachman, “Data-efficient reinforcement learning with self-predictive representations,” in Proc. Int. Conf. Learn. Representations, 2020.

[43] D. Yarats, R. Fergus, A. Lazaric, and L. Pinto, “Reinforcement learning with prototypical representations,” in Proc. Int. Conf. Mach. Learn., PMLR, 2021, pp. 11 923–11 931.

[44] L. Lee, B. Eysenbach, E. Parisotto, E. Xing, S. Levine, and R. Salakhutdinov, “Efficient exploration via state marginal matching,” 2019, arXiv:1906.05274.

[45] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, “Diversity is all you need: Learning skills without a reward function,” 2018, arXiv:1802.06970.

[46] H. Liu and P. Abbeel, “APS: Active pretraining with successor features,” in Proc. Int. Conf. Mach. Learn., PMLR, 2021, pp. 6736–6747.

[47] O. Anchel, N. Baram, and N. Shimkin, “Averaged-DQN: Variance reduction and stabilization for deep reinforcement learning,” in Proc. Int. Conf. Mach. Learn., PMLR, 2017, pp. 176–185.

[48] Y. Yuan et al., “EUCLID: Towards efficient unsupervised reinforcement learning with multi-choice dynamics model,” 2022, arXiv:2210.00498.