SAEIR: Sequentially Accumulated Entropy Intrinsic Reward for Cooperative Multi-Agent Reinforcement Learning with Sparse Reward

Xin He, **Hongwei Ge***, **Yaqing Hou** and **Jincheng Yu**
School of Computer Science and Technology, Dalian University of Technology
hx_dlut@mail.dlut.edu.cn, hwge@dlut.edu.cn

Abstract

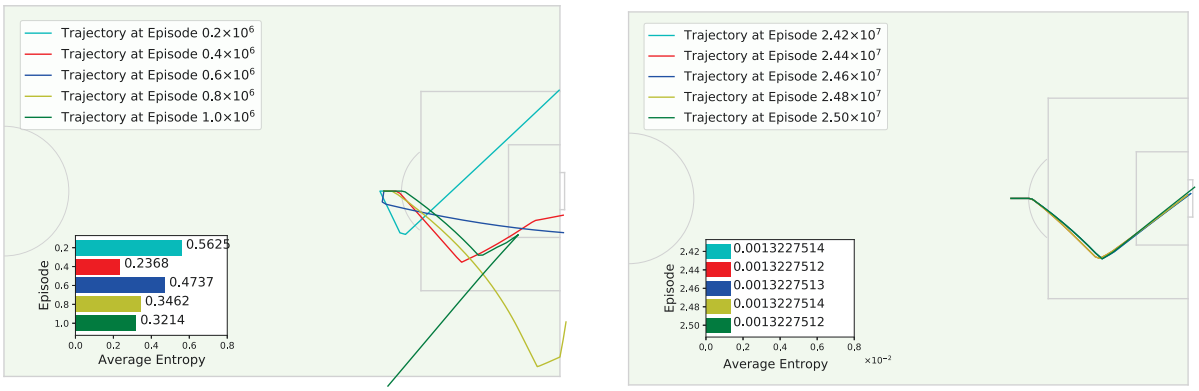
Multi-agent reinforcement learning (MARL) performs well for solving complex cooperative tasks when the scenarios have well-defined dense rewards. However, there are usually sparse reward settings in many real-world multi-agent systems, which makes it difficult for MARL algorithms to successfully learn an effective strategy. To tackle this problem, we propose a novel sequentially accumulated entropy intrinsic reward named SAEIR, which utilizes the entropy of multi-agent system as a bonus to accelerate learning. Specifically, the multi-scale hypergraph critic is proposed to obtain high-order system state representation, which also enhances the ability to effectively evaluate the action produced by the actor. Based on the comprehensive and compact system state representation, the orderliness of multi-agent systems can be measured to determine the highly valuable states for adding entropy-based intrinsic rewards which leads to a highly efficient learning process. Empirical results demonstrate that our proposed method achieves state-of-the-art performance in several complex cooperative multi-agent environments with sparse reward settings.

1 Introduction

Deep multi-agent reinforcement learning (MARL) is an active research field with its successful application to multi-agent cooperative tasks such as multi-player video games [Zhang and Yu, 2023], robot swarms control [Gu *et al.*, 2023], smart grid control [Jin and Ma, 2019], autonomous vehicle coordination [Guo *et al.*, 2023], and resource management [Li *et al.*, 2023]. Despite the progress in value-based and policy-based MARL algorithms [Sunehag *et al.*, 2018; Rashid *et al.*, 2018; Son *et al.*, 2019; Wang *et al.*, 2021; Yu *et al.*, 2022; Lowe *et al.*, 2017], their efficiency relies heavily on expert-designed dense reward feedback. However, there are only sparse reward settings in many real-world multi-agent scenarios [Pathak *et al.*, 2017], where non-zero rewards are provided just when certain conditions are satisfied rather than at every time step. Without the guidance of dense reward feedback, the agents require many episodes to come across any reward and have difficulty in learning effective cooperative strategies, which restricts the application of multi-agent algorithms. Therefore, solving cooperative tasks with sparse rewards becomes a critical challenge for multi-agent reinforcement learning.

A common approach to addressing this challenge in current MARL algorithms is to apply intrinsic motivation techniques during the training procedure, which has been shown to be effective in some cooperative tasks. Due to classical curiosity-based methods [Pathak *et al.*, 2017; Savinov *et al.*, 2019; Yang *et al.*, 2020] performing well in single-agent scenarios, some works extend the intrinsic curiosity reward to MARL algorithms [Iqbal and Sha, 2019b; Zheng *et al.*, 2021; Li and Gajane, 2023; Yu *et al.*, 2023]. There are also many works that utilize the curiosity bonus as part of intrinsic reward and introduce novel prior knowledge to improve exploration, such as interaction influencing [Wang *et al.*, 2020] and counterfactual reasoning [Yang *et al.*, 2021]. Besides using the concept of curiosity to design the intrinsic reward, current works apply effective formulations to generate the intrinsic reward, such as the theory of mind [Ma *et al.*, 2022], subgoal assignment [Jeon *et al.*, 2022; Xie *et al.*, 2023], and causal inference [Liu *et al.*, 2023]. These methods combine intrinsic reward with the extrinsic reward provided by the environment to encourage exploration. However, the above methods ignore the great potential for entropy, the natural property in multi-agent systems, as an intrinsic reward.

Entropy originates from information theory, which is used to characterize the disorder of a system. During the training procedure, the multi-agent system transforms from stochastic processes (disorder) to stationary processes (order), where the strategies learned by agents transform from random to deterministic and the system entropy gradually decreases. As shown in Fig. 1, in the case of solving a multi-agent football game, the multi-agent system states using soccer trajectory as visualization demonstrate this phenomenon. For the visualization of different strategies, we observe that the soccer trajectories implemented by early strategies are not uniform, which means the strategies learned in the early stages are stochastic. On the contrary, the soccer trajectories implemented by later strategies are similar, which means the strategies learned in the later stages are stationary. At the same time, we observe that early strategies do not achieve goals, while later strategies all win the match. Based on this ob-



(a) Early Strategies (Stochastic Process)

(b) Later Strategies (Stationary Processes)

Figure 1: Visualization of the multi-agent system state under different stage strategies using soccer trajectory.

servation, we focus on the highly valuable states that result in a decrease in the entropy of multi-agent systems. Intrinsic rewards should be supplied to these states, substantially accelerating learning when the original environmental rewards are sparse.

In this paper, we propose a novel Sequentially Accumulated Entropy Intrinsic Reward (SAEIR) to address the challenge of sparse reward encountered within cooperative multi-agent reinforcement learning. To precisely characterize the state of the multi-agent system, the high-order state representations are extracted from the environment via the multi-scale hypergraph critic. Based on these comprehensive and compact system state representations, the change of multi-agent system entropy in a trajectory under current policy can be measured. Then, the accumulated system entropy is calculated as an intrinsic reward supplied to such highly valuable states that can decrease the system entropy. We summarize our contributions as follows:

- We propose an effective intrinsic reward, using sequentially accumulated system entropy as a dense bonus to accelerate the learning of successful strategies in multi-agent systems with sparse reward settings.
- A multi-scale hypergraph critic is proposed to extract accurate system state representations, which are used to measure the change of multi-agent system entropy between sequential states. It also enhances the ability to effectively evaluate the action produced by the actor.
- Empirical results in cooperative multi-agent environments with sparse rewards demonstrate our method outperforms the state-of-the-art methods and promotes agents to generate better cooperative policies.

2 Related Work

2.1 Intrinsic Motivation for Multi-Agent Reinforcement Learning with Sparse Rewards

Benefiting from the significant performance of intrinsic motivation techniques in the sparse-reward SARL domain, most MARL methods addressing the challenge of sparse rewards have been adapted from single-agent intrinsic incentives. Multi [Iqbal and Sha, 2019b] designs various types of count-based intrinsic rewards by considering coordination among agents and utilizes a hierarchical policy to dynamically select exploration modalities trained on diverse intrinsic rewards. EDTI and EITI [Wang *et al.*, 2020] consider the agents' interactions during their training process to coordinate their exploration and use information- and decision-theoretic influence to obtain intrinsic rewards respectively. CIExplore [Yang *et al.*, 2021] combines a joint curiosity-based reward and an influence reward motivated by counterfactual reasoning as an intrinsic reward, avoiding training instability and enabling agents to generate cooperative behavior. Elign [Ma *et al.*, 2022] designs an intrinsic reward inspired by the self-organization principle in Zoology [Couzin, 2007]. It is a simple and effective way to improve the learning ability of agents in the alternative setup of decentralized training or sparse rewards. LIGS [Mguni *et al.*, 2022] employs an adaptive learner to construct intrinsic rewards online for performing coordinated joint behavior and achieving an efficient learning process. MASER [Jeon *et al.*, 2022] generates subgoals from the experience replay buffer and designs individual intrinsic rewards for each agent using actionable representations, helping agents maximize the joint action value while achieving their subgoals. IPERS [Xie *et al.*, 2023] utilizes prioritized experience replay along with subgoals to enhance the training convergence rate in environments with sparse rewards. Subgoals are employed to provide beneficial intrinsic rewards. SAME [Xu *et al.*, 2023b] gets a bonus from a special structural prior on the reward function and combines it with a count-based bonus, which not only encourages agents to explore sub-state paces with higher uncertainty but also preserves novel states in the full-state space level. USM [Yu *et al.*, 2023] is an intrinsic reward that focuses on the target-related attributes of the underexplored subspaces rather than the whole state space to accelerate learning. LJIR [Chen *et al.*, 2023] introduces a general framework to construct a compound intrinsic reward online by agents' state and joint actions, which helps agents find the best joint actions via the combination of state novelty

and joint-action novelty. The work [Xu *et al.*, 2023a] improves traditional count-based methods by introducing constrained joint policy diversity, making the traditional count-based methods work well with neural networks and achieve significant performance. PDP [Sun *et al.*, 2023] actively amplifies the diversity among agents' policies during training and designs an intrinsic reward based on the diversity. LAIES [Liu *et al.*, 2023] defines the concept of lazy agents in MARL and proposes two intrinsic rewards based on intrinsic motivation for individual diligence and collaborative diligence derived from these definitions. I-Go-Explore [Li and Gajane, 2023] integrates an intrinsic curiosity module with the Go-Explore framework [Ecoffet *et al.*, 2021] to address the limitations of the intrinsic curiosity module. However, despite SAME employing the entropy of the state distribution to find the sub-state space with higher uncertainty, none of the above methods explicitly considers mapping the state entropy as an intrinsic reward, which is a natural property in multi-agent systems and a suitable intrinsic motivation that does not require domain knowledge. We aim to exploit the state entropy to facilitate learning coordinated policies in complex cooperative multi-agent environments with sparse rewards.

2.2 Enhanced Critic for Multi-Agent Reinforcement Learning with Sparse Rewards

The reliable and accurate critic can improve the performance of MARL algorithms [Foerster *et al.*, 2018; Iqbal and Sha, 2019a; Liu *et al.*, 2022]. Therefore, some works focus on utilizing the enhanced critic to address the challenge of sparse reward. IRAT [Wang *et al.*, 2022] introduces a framework that assists the team policy learning from critic with sparse reward by the individual policy learned using critic with dense reward. SN-MAPPO [Mehta *et al.*, 2023] utilizes the spectral normalization technique to regularize the plain critic. It enables agents to quickly and stably learn from the setup of sparse rewards. Our method develops an enhanced critic via the multi-scale hypergraph network, which not only enhances the ability to effectively evaluate the action produced by the actor but also extracts the accurate high-order system state representations.

3 Preliminaries

3.1 Problem Formulation

The multi-agent cooperation problem can be formulated as a decentralized partially observable Markov decision process (Dec-POMDP) [Oliehoek and Amato, 2016]. Generally, it is defined as a tuple $< \mathcal{N}, \mathcal{S}, \mathcal{A}, \mathcal{O}, \mathcal{T}, \mathcal{R}, \gamma >$, where $\mathcal{N}$ is the set of n agents indexed by $1, 2, \ldots, n$; $\mathcal{S}$ is the finite set of environment states; $\mathcal{A} = A_1 \times A_2 \times \ldots \times A_n$ is the set of joint actions and A_i represents the set of actions available for agent i ($i \in \mathcal{N}$); $\mathcal{O} = O_1 \times O_2 \times \ldots \times O_n$ is the set of joint observations and O_i denotes all possible observation of agent i on $\mathcal{S}$; $\mathcal{T} : \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ formulates the transition probability function, and at each time step, the agents select actions $a \in \mathcal{A}$ at state $s \in \mathcal{S}$, then reach the next new state $s' \sim \mathcal{T}(s, a)$; $\mathcal{R} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^n$ is the reward function. The action policy $\pi_i : O_i \times A_i \to [0, 1]$ represents the probability of agent i selecting an action $a_i \in A_i$ based on its partial observation $o_i \in O_i$. Given an initial state s, all the agents work cooperatively to maximize the total system's discounted cumulative reward: $J(\tau) = \sum_{t=0}^{\infty} \gamma^t r^t$, where r^t is the reward obtained by agents at time step t according to the police π_i, $\gamma \in [0, 1]$ is a discount factor.

3.2 Multi-agent Reinforcement Learning Algorithms

The Actor-Critic method is widely used in the multi-agent deep reinforcement learning framework [Lowe *et al.*, 2017; Foerster *et al.*, 2018; Yu *et al.*, 2022]. The most AC methods commonly use advantage function A^π to optimize policy π:

$$A^\pi(s_t, \mathbf{a}_t) := r_t + \gamma V^\pi(s_{t+1}) - V^\pi(s_t) \tag{1}$$

and the policy objective function is:

$$\mathcal{J} = \sum_{i=1}^{N} E_{s_t, a_t \sim \pi}[\nabla \log \pi(a_t^i | o_t^i) A(s_t, \mathbf{a}_t)] \tag{2}$$

3.3 Multi-Agent Proximal Policy Optimization

IPPO [de Witt *et al.*, 2020] trains an independent PPO-learned [Schulman *et al.*, 2017] strategy with a parameter-sharing technique [Gupta *et al.*, 2017] for each agent in the multi-agent system. MAPPO [Yu *et al.*, 2022] extends the independent critics of IPPO to a centralized value function accepting global information, which shows the significant performance in various cooperative multi-agent tasks. It optimizes the policy network θ by maximizing the following objective:

$$\mathcal{J}(\theta) = \sum_{i=1}^{N} \mathbb{E}[\min\left(\eta_t^i(\theta)\hat{A}_t^i, clip\left(\eta_t^i(\theta), 1 \pm \epsilon\right)\hat{A}_t^i\right)] \tag{3}$$

where $\eta_t^i(\theta) = \frac{\pi_\theta(a_t^i|\tau_t^i)}{\pi_{\theta_{old}}(a_t^i|\tau_t^i)}$ denotes the probability ratio. The function $clip(\cdot)$ removes $\eta_t^i(\theta^i)$ outside of the interval $[1 - \epsilon, 1 + \epsilon]$ parameterized by ϵ, which approximates the KL-divergence constraint. $\hat{A}_i^t$ is a generalized advantage estimator (GAE) [Schulman *et al.*, 2016]

$$\hat{A}_i^t = \sum_{l=0}^{T} (\gamma\lambda)^l A_i^{t+l} \tag{4}$$

where T is the episode time horizon.

3.4 Hypergraph Neural Network

Hypergraph neural network [Feng *et al.*, 2019] is a generalization of general graph neural network, in which a hyperedge can link any number of vertices. Hence, the hypergraph can capture high-order representations through group-wise embeddings instead of pair-wise ones used in the general graph, which has proven to be effective in a wide range of applications. Mathematically, a hypergraph is defined as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of vertices in the hypergraph (N represents the number of vertices), and $\mathcal{E} = \{e_i := (v_1^{(i)}, v_2^{(i)}, \ldots, v_k^{(i)}) | i = 1, 2, \ldots M\}$ denotes the set of hyperedges (M represents the number of hyperedges, k represents the number of vertices in a hypergraph and k can be different in different hyperedges).

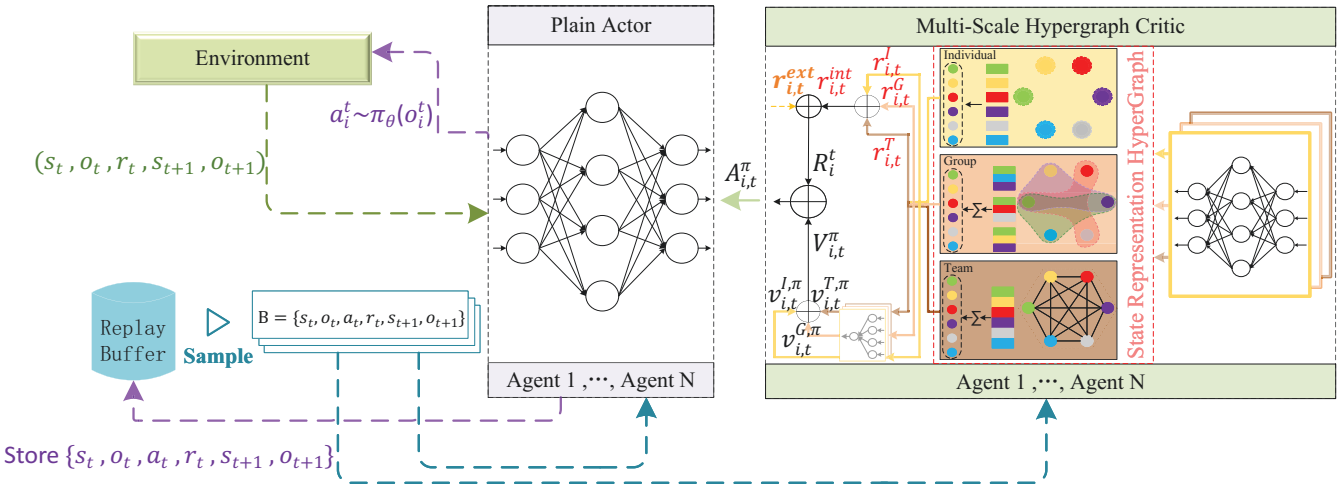


Figure 2: Overview of the proposed approach.

4 Method

This section describes our proposed intrinsic reward function SAEIR for effectively addressing the challenge in cooperative MARL with sparse reward settings. We follow the Centralized Training and Decentralized Execution (CTDE) framework, where agents are trained with access to global information but make decisions based on their own local information in execution. We implement the SAEIR on the representative Actor-Critic (AC) style CTDE algorithm, which has a plain actor and a multi-scale hypergraph critic shown in Fig. 2. The agent executes action sampled from the policy learned by the plain actor (a three-layer Multi-Layer Perceptron with ReLU non-linearities) and interacts with the environment. The multi-scale hypergraph critic provides the state value for the plain actor to update. It also offers a comprehensive and compact system state representation to calculate the intrinsic reward. The sequentially accumulated entropy intrinsic reward is calculated based on these system state representations and supplied to the highly valuable states, leading to a highly efficient learning process. We then introduce the details of the multi-scale hypergraph critic and the calculation of the SAEIR intrinsic reward.

4.1 Multi-Scale Hypergraph Critic

To extract the high-order system state representations and provide the accurate state value for the plain actor, we develop the multi-scale hypergraph critic (MSHG Critic) that comprises the state feature encoder, the state representation hypergraph network, and the state value decoder.

State Feature Encoder: The encoder is composed of a three-layer Multi-Layer Perceptron (MLP) and a Gated Recurrent Unit (GRU). For each agent i at time step t, the system state s_i^t is encoded using the MLP to create the encoded state feature $\bar{e}_{i,t}$. Then, the GRU takes the encoded state feature $\bar{e}_{i,t}$ and the hidden state $h_{i,t-1}$ as inputs to generate the next hidden state $h_{i,t}$ as shown in Eq. 5

$$\begin{cases} \bar{e}_{i,t}^K = & \text{MLP}(s_i^t; W_{\text{MLP}}^K) \\ h_{i,t}^K = & \text{GRU}(\bar{e}_{i,t}^K, h_{i,t-1}^K; W_{\text{GRU}}^K) \end{cases} \tag{5}$$

where $K \in \{Individual(I), Group(G), Team(T)\}$. $\bar{e}_{i,t}^K$ and $h_{i,t}^K$ respectively denote the encoded state feature and the hidden state used for extracting the state representation at the scale of individual, group, and team. W_{MLP}^K and W_{GRU}^K are the learnable weights.

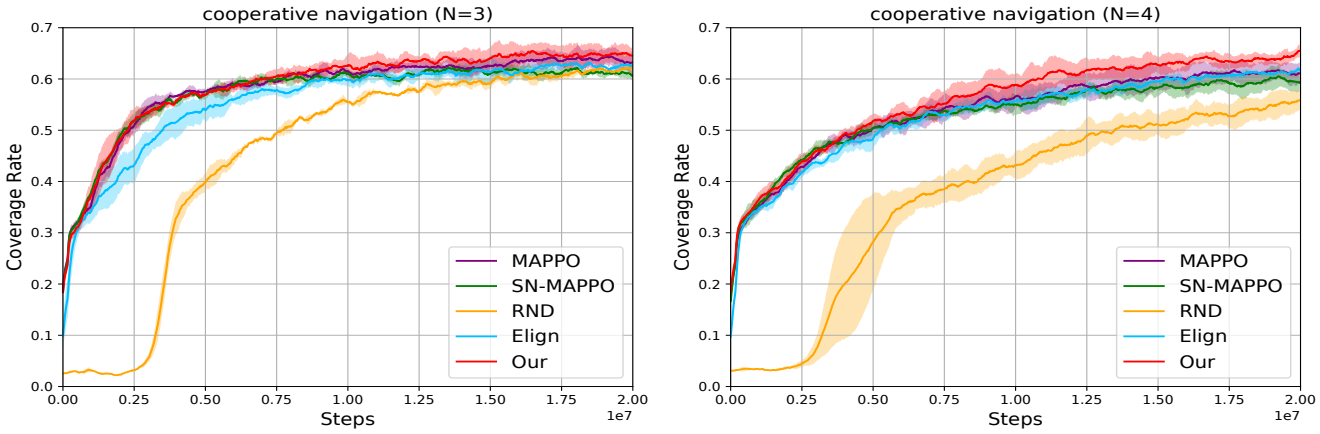
State Representation Hypergraph Network: The comprehensive and compact state representation is extracted under multiple scales via the hypergraph network. Mathematically, let $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denote a vertex set containing N agents and $\mathcal{E}^I = \{e_1^I, e_2^I, \ldots, e_N^I\}$, $\mathcal{E}^G = \{e_1^G, e_2^G, \ldots, e_{M_s}^G\}$ and $\mathcal{E}^T = \{e_1^T\}$ denote three hyperedge sets with the scale of individual, group, and team respectively, where $1 < M_s < N$ and it is decided by a hyperparameter N_G that denotes the number of agents in a group. Therefore, the multiple scale hypergraph networks are formalized as $\mathcal{G}^I = \{\mathcal{V}, \mathcal{E}^I\}$, $\mathcal{G}^G = \{\mathcal{V}, \mathcal{E}^G\}$ and $\mathcal{G}^T = \{\mathcal{V}, \mathcal{E}^T\}$. The topology of each $\mathcal{G}^K$ can be constructed as an incidence matrix $\mathrm{H}^K \in \mathbb{R}^{|\mathcal{V}|\times|\mathcal{E}^K|}$, where $\mathrm{H}_{i,m}^K = 1$ if the ith vertex is included in the mth hyperedge, otherwise $\mathrm{H}_{i,m}^K = 0$. Concretely, H^I equals to the identity matrix $\mathbb{I}^{N\times N}$ and H^T equals to the vector with all 1 elements $\mathbf{1}^{N\times 1}$. For H^G, we first compute an affinity matrix $\mathrm{A}^G \in \mathbb{R}^{N\times N}$ to measure the correlation between the ith agent and the jth agent:

$$\mathrm{A}_{i,j}^G = h_i^{G^\top} h_j^G / (||h_i^G||_2 ||h_j^G||_2) \tag{6}$$

Then, the hyperedges in $\mathcal{E}^G$ are formed by searching the high-density submatrices in the affinity matrix A^G via a greedy algorithm approximation, where agents in the submatrices have high correlation with each other and form a group:

$$\begin{aligned} e_m^G &= \underset{\Omega\subseteq\mathcal{V}}{\text{argmax}} \, ||\mathrm{A}_{\Omega,\Omega}^G||_1, \\ s.t. &|\Omega| = N_G; v_i \in \Omega, i = 1, 2, \ldots, N \end{aligned} \tag{7}$$

where $||\cdot||_1$ denotes the matrix entrywise L_1-norm. $H^G \in \mathbb{R}^{N\times M_s}$ is composed of a stack of $e_m^G (m = 1, 2, \ldots, M_s)$. The obtained hyperedge contains the most correlated agents



(a) Cooperative Navigation (N=3)

(b) Cooperative Navigation (N=4)

Figure 3: The training curves of coverage rate for our approach and baseline algorithms in Cooperative Navigation environment.

and links them together to consider the state representation from a group perspective. According to the hyperedge sets $\mathcal{E}^K$, the state representation $\bar{s}_i^K$ is obtained through:

$$\begin{aligned} \mathbf{e}_m^K &= w_m^K \sum_{v_i \in e_m^K} a_{m,i}^K \mathcal{F}_{e_m^K}^K(\mathbf{v}_i^K) \\ \mathbf{v}_i^K &\leftarrow \mathcal{F}_{v_i}^K([\mathbf{v}_i^K, \sum_{e_m^K \in \mathcal{E}_{v_i}^K} \mathbf{e}_m^K]) \\ \bar{s}_i^K &= \mathcal{F}_{\bar{s}}^K(\mathbf{v}_i^K) \end{aligned} \tag{8}$$

where $\mathbf{e}_m^K$ denotes the state feature embedding from vertex to hyperedge for the mth hyperedge in each K scale, $\mathbf{v}_i^K$ is the state representation for the ith agent in each K scale and initial $\mathbf{v}_i^K$ equals to the hidden state $h_{i,t}^K$, w_m^K is the learnable weight for weighting the state feature embedding of each hyperedge, $a_{m,i}^K$ denotes the attention coefficient of each vertex v_i in e_m^K, $\mathcal{F}_{e_m^K}^K(\cdot)$ is a three-layer MLP, $\mathcal{F}_{\bar{s}}^K(\cdot)$ and $\mathcal{F}_{v_i}^K(\cdot)$ are a single-layer network respectively, $[\cdot,\cdot]$ denotes the concatenation of all state representations about one vertex contained in the associated hyperedges.

State Value Decoder: The decoder uses a fully connection (FC) layer to generate the state value $v_{i,t}^{K,\pi}$ for agent i at time step t under current policy π, which takes each $\bar{s}_{i,t}^K$ as the input and is shown in Eq. 9:

$$v_{i,t}^{K,\pi} = \text{FC}(\bar{s}_{i,t}^K; W_{FC}^K) \tag{9}$$

The final state value is

$$V_{i,t}^{\pi} = \sum_K w_{i,t}^K v_{i,t}^{K,\pi} \tag{10}$$

where $w_{i,t}^K$ are the weight coefficient.

4.2 Sequentially Accumulated Entropy Intrinsic Reward

To determine the highly valuable state for adding a bonus, we utilize the entropy-based technique to characterize the disorder of a multi-agent system. According to the comprehensive and compact system state representation obtained from the MSHG critic, the multi-agent state is measured by the weighted sum of state representation with attention coefficients:

$$\hat{s}_{i,t}^K = \mathbf{1}^\mathsf{T} \cdot a_{i,t}^K \odot \bar{s}_{i,t}^K \tag{11}$$

where $\odot$ denotes the Hadamard product operator. In practice, we set a^K to be W_{FC}^K to maintain consistency between the estimation of the system state and the state value provided by the critic. To identify the highly valuable states that lead to a decrease in the disorder of the multi-agent system, the sequentially accumulated entropy of the system state in a trajectory is calculated as follows:

$$\begin{aligned} p_{i,t}^K[\hat{s}_{i,t}^K(T)] &= \frac{\exp(\hat{s}_{i,t}^K)}{\sum_{t'=1}^T \exp(\hat{s}_{i,t'}^K)} \\ \hat{e}_{i,t}^K(T) &= -p_{i,t}^K[\hat{s}_{i,t}^K(T)] \log p_{i,t}^K[\hat{s}_{i,t}^K(T)] \end{aligned} \tag{12}$$

where $T \equiv t$ denotes the time length of the trajectory until time step t. According to Eq. 12, the highly valuable states that decrease the disorder of the system are searched to be added intrinsic rewards, and the sequentially accumulated entropy intrinsic reward is designed as:

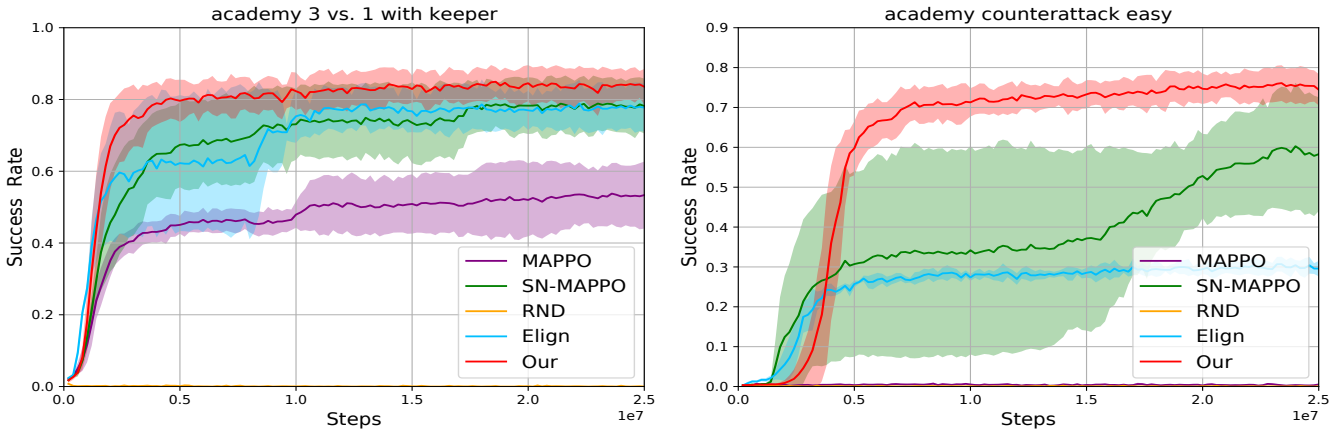
$$r_{i,t}^{int} = \sum\nolimits_K w_{i,t}^K \hat{e}_{i,t}^K \mathbb{I}[\hat{e}_{i,t}^K(T), \hat{e}_{i,t}^K(T+1)] \tag{13}$$

where $w_{i,t}^K$ are the weights for three scales of intrinsic rewards and $\sum_K w_{i,t}^K = 1$, $\mathbb{I}$ is an indicator function:

$$\mathbb{I}(\cdot,\cdot) = \begin{cases} 1, & \hat{e}_{i,t}^K(T) > \hat{e}_{i,t}^K(T+1); \\ 0, & otherwise. \end{cases} \tag{14}$$

which represents the state s_t decreasing the accumulated system entropy $\hat{e}_t$ within the sequentially time length $T+1$ of trajectory is identified as highly valuable state, and the bonus should be supplied to this state. While the intrinsic reward is obtained according to Eq. 13, the augmented reward for agent i at time step t is:

$$R_i^t = r_{i,t}^{ext} + \beta\, r_{i,t}^{int} \tag{15}$$



(a) Academy 3 vs. 1 with Keeper

(b) Academy Counterattack Easy

Figure 4: The training curves of success rate for our approach and baseline algorithms in Google Football Research environment.

Methods	Cooperative Navigation (N=3)		Cooperative Navigation (N=4)		3 vs.1 with Keeper		Counterattack Easy	
	Coverage Rate	Episode Rewards	Coverage Rate	Episode Rewards	Success Rate	Episode Rewards	Success Rate	Episode Rewards
MAPPO	64.43 ± 1.56	3.38 ± 0.06	64.26 ± 3.14	4.59 ± 0.21	53.77 ± 8.53	9.86 ± 2.50	0.73 ± 0.17	-0.03 ± 0.10
SN-MAPPO	62.23 ± 0.53	3.34 ± 0.04	62.37 ± 2.37	4.57 ± 0.09	78.87 ± 7.36	14.46 ± 2.02	60.27 ± 15.93	9.23 ± 1.47
RND	62.33 ± 1.17	3.28 ± 0.09	58.40 ± 1.52	4.27 ± 0.10	0.75 ± 0.14	0.00 ± 0.05	0.22 ± 0.11	-0.21 ± 0.20
Elign	63.16 ± 0.42	3.36 ± 0.02	64.45 ± 1.88	4.69 ± 0.13	78.87 ± 6.62	14.61 ± 2.67	30.70 ± 1.87	6.37 ± 0.46
Our	$\mathbf{65.48 \pm 1.59}$	$\mathbf{3.46 \pm 0.07}$	$\mathbf{66.80 \pm 1.54}$	$\mathbf{4.79 \pm 0.02}$	$\mathbf{84.95 \pm 4.10}$	$\mathbf{15.06 \pm 1.50}$	$\mathbf{76.12 \pm 4.48}$	$\mathbf{11.58 \pm 1.19}$

Table 1: The exact results in four sparse reward scenarios for our method and compared baselines.

where $r_{i,t}^{ext}$ is the external reward given by the environment, $r_{i,t}^{int}$ is the intrinsic reward calculated by Eq. 13, and β is a positive scalar that weights the intrinsic reward.

During the training procedure, the disorder of the system state gradually decreases and intrinsic rewards finally converge to a constant close to 0. Therefore, the advantage estimate $\bar{A}^{\pi}(s_t, \mathbf{a}_t)$ with intrinsic rewards can be proved:

$$\begin{aligned}\lim_{t\to\infty} \bar{A}^{\pi}(s_t,\mathbf{a}_t) &= \lim_{t\to\infty}[R^t + \gamma V^{\pi}(s_{t+1}) - V^{\pi}] \\ &= \lim_{t\to\infty}[r_t^{ext} + \beta r_t^{int} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}] \\ &= \lim_{t\to\infty}[r_t^{ext} + \gamma V^{\pi}(s_{t+1}) - V^{\pi}] + \beta \lim_{t\to\infty} r_t^{int} \\ &= \lim_{t\to\infty} A^{\pi}(s_t, \mathbf{a}_t) \propto r_t^{ext}\end{aligned} \tag{16}$$

This means that after a certain amount of training, the policy is only influenced by external rewards, ensuring policy invariance.

4.3 Training Process

In this section, we show the overall training process of the SAEIR applied to the agent in Algorithm 1. The plain actor network is trained to minimize the Eq. 17 and the multi-scale hypergraph critic network is trained to minimize the Eq. 18:

$$\begin{aligned}\mathcal{L}^{\pi}(\theta) = &\frac{1}{N}\sum_{i}^{N} \min[r_i^{\theta}\hat{A}_i^{\pi}, \text{clip}(r_i^{\theta}, 1\pm\epsilon)\hat{A}_i^{\pi}] \\ &+ \sigma\frac{1}{N}\sum_{i=1}^{N} S[\pi_{\theta}(o_i)]\end{aligned} \tag{17}$$

$$\begin{aligned}\mathcal{L}^{C}(\phi) = &\frac{1}{N}\sum_{i=1}^{N} max[(V_{\theta}(s_i) - R_i^t(\gamma))^2, \\ &[\text{clip}(V_{\theta}(s_i), V_{\theta_{old}}(s_i)\pm\epsilon) - R_i^t(\gamma)]^2\end{aligned} \tag{18}$$

where $r_i^{\theta} = \frac{\pi_{\theta}(a_i|o_i)}{\pi_{\theta_{old}}(a_i|o_i)}$, S is the policy entropy, σ is the entropy coefficient hyperparameter, and $R_i^t(\gamma)$ is the discounted reward-to-go of the augmented reward with factor γ.

5 Experiments and Results

In this section, We evaluate SAEIR in two complex cooperative multi-agent environments: Cooperative Navigation (CN, [Lowe *et al.*, 2017]) and Google Football Research (GRF, [Kurach *et al.*, 2020]). In all environments, scenarios with sparse reward settings are considered. We compare the performance of SAEIR against a classical Actor-Critic MARL algorithm MAPPO [Yu *et al.*, 2022], an Enhanced-Critic MARL algorithm SN-MAPPO [Mehta *et al.*, 2023], and two intrinsic reward MARL algorithms RND [Burda *et al.*, 2019] (extended to MARL version), Elign [Ma *et al.*, 2022]. All experiments run with five random seeds to maintain reliability and consistency.

5.1 Experimental Settings

For the CN environment, it requires N agents cooperating to reach L landmarks while avoiding collision. The locations of agents and landmarks are generated randomly. Each agent gets a positive reward related to the number of occupying landmarks and a negative reward when collisions oc-

Algorithm 1 Sequentially Accumulated Entropy Intrinsic Reward Applied to Agents (SAEIRA)

Initialize the parameters θ for policy π, the parameters set ϕ^K for multi-scale critic V^K, $K \in \{individual, group, team\}$, using Orthogonal initialization.
Set learning rate α.

while $step \leq step_{max}$ **do**
  set data buffer $D = \{\}$
  for $b = 1$ **to** $batch\_size$ **do**
    $\tau = [\,]$ empty list
    for $t = 1$ **to** T **do**
      for all agents i **do**
        $p_i^t, h_i^{t-1} = \pi(o_i^t, h_i^t; \theta)$
        $a_i^t \sim p_i^t$
        $v_{i,t}^K, h_{i,t-1}^K, \hat{s}_{i,t}^K = V^K(s_t; \phi^K)$ // Eq. 5 - Eq. 9
        $V_{i,t}^\pi = \sum_K w_{i,t}^K v_{i,t}^K$ // Eq. 10
      end for
      Execute actions $\mathbf{a}^t$, observe $r_t^{ext}, s_{t+1}, \mathbf{o}_{t+1}$
      Calculate intrinsic reward $r_{i,t}^{int}$ using Eq. 13
      $\tau += [s_t, \mathbf{o}_t, \mathbf{h}^t, \mathbf{h}_t^K, \mathbf{a}_t, r_t^{ext}, \mathbf{r}_t^{int}, s_{t+1}, \mathbf{o}_{t+1}]$
    end for
    Calculate advantages estimate $\hat{A}^\pi$ for plain actor by GAE on τ, using PopArt
    Calculate reward-to-go $\mathbf{R}_t(\gamma)$ of augmented reward by GAE on τ with normalizing PopArt
    $D = D \cup \tau$
  end for
  for $p = 1$ **to** $epoch_{train}$ **do**
    Sample data d from D
    Adam update θ on $\mathcal{L}^\pi(\theta)$ with data d // Eq. 17
    Adam update ϕ on $\mathcal{L}^C(\phi)$ with data d // Eq. 18
  end for
  for $e = 1$ **to** $epoch_{eval}$ **do**
    Evaluate policy π_i of each agent i
  end for
end while

cur. The evaluation metric is the coverage rate per episode during the training process. In the GRF environment, each scenario requires agents to learn collaborative skills to score, such as dribbling, passing, and moving. Each agent only obtains a $+1$ reward for scoring a goal and a -1 reward when conceding one to the opposing team, which corresponds to the SCORING reward function provided by the environment. The evaluation metric is the winning rate per episode during the training process. More details are within supplementary.

5.2 Results in Cooperative Navigation

Fig. 3 shows the training curves of the coverage rate for agents reaching goal landmarks. According to Fig. 3, we can find that the performance of RND deteriorates as the number of agents increases. Although SN-MAPPO and Elign learn coherent policies, their performance slightly lags behind MAPPO. We think this is mainly because the outputs of critic in SN-MAPPO are too smooth and the dynamics models learned in Elign are suboptimal. SAEIR achieves better performance than the other baselines. With the increase in the number of agents, SAEIR shows increasingly pronounced performance improvements compared to the other methods.

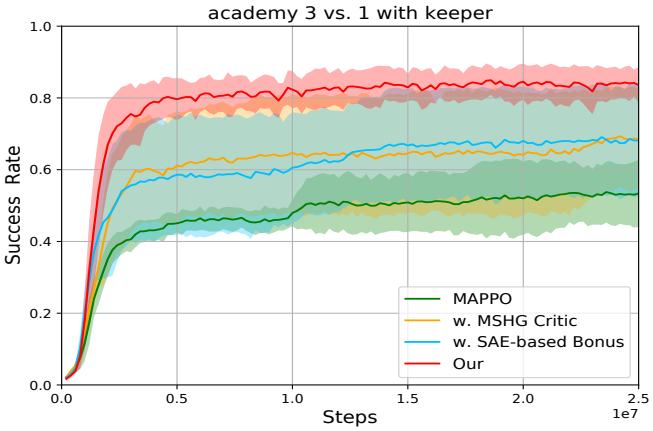


Figure 5: The training curves of success rate for ablation methods.

5.3 Results in Google Football Research

Fig. 4 shows the training curves of the success rate for agents completing a football game. As shown in Fig. 4, RND does not learn a winning strategy in two sparse reward scenarios. Although MAPPO solves the task in Academy 3 vs.1 with keeper scenario, it failed in another scenario. For SN-MAPPO and Elign, they can learn winning strategies in two sparse reward scenarios but the performances are unstable. In both sparse reward scenarios, we can see that SAEIR achieves the highest success rate among the baselines. Table 1 shows the exact results in all sparse reward scenarios.

5.4 Ablation Study

We present the results of the ablation studies. To confirm the performance individually using the multi-scale hypergraph critic and the sequentially accumulated entropy-based bonus, we choose the academy 3 vs.1 with keeper scenario to make the report. Fig. 5 shows the training curves of the success rate. Although the MSHG critic and the SAE-based bonus learn comparable winning strategies, the performance improvements do not achieve expectations. Benefiting from the advantages of both two components, our method achieves the highest performance and is quite stable with a small variance.

6 Conclusion and Future Work

In this paper, we propose a novel intrinsic reward called SAEIR to successfully learn an effective strategy in sparse reward environments. The proposed method is based on the assumption that the highly valuable states making a decrease in the entropy of multi-agent systems should be added a bonus. SAEIR guides multiple agents to find better states that decrease the disorder of system and facilitates the learning of successful strategies. The experimental results demonstrate the effectiveness of SAEIR compared to state-of-the-art methods. One limitation of the current work is that it has a cold start issue sometimes. In the future, we will make efforts to address this limitation and explore the possibility of combining our method with the subspace or subgoal technique.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (61976034,U1808206), the Dalian Science and Technology Innovation Fund (2022JJ12GX013), the Natural Science Foundation of Liaoning Province (2022YGJC20), and the Fundamental Research Funds for the Central Universities (DUT21YG106).

References

[Burda *et al.*, 2019] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. In *Proc. 7th Int. Conf. Learn. Represent.*, 2019.

[Chen *et al.*, 2023] Zihan Chen, Biao Luo, Tianmeng Hu, and Xiaodong Xu. LJIR: learning joint-action intrinsic reward in cooperative multi-agent reinforcement learning. *Neural Networks*, 167:450–459, 2023.

[Couzin, 2007] Iain Couzin. Collective minds. *Nature*, 445(7129):715–715, 2007.

[de Witt *et al.*, 2020] Christian Schröder de Witt, Tarun Gupta, Denys Makoviichuk, Viktor Makoviychuk, Philip H. S. Torr, Mingfei Sun, and Shimon Whiteson. Is independent learning all you need in the starcraft multi-agent challenge? 2020.

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

[Feng *et al.*, 2019] Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. Hypergraph neural networks. In *Proc. 33th AAAI Conf. Artif. Intell.*, pages 3558–3565, 2019.

[Foerster *et al.*, 2018] Jakob N. Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proc. 32th AAAI Conf. Artif. Intell.*, pages 2974–2982, 2018.

[Gu *et al.*, 2023] Shangding Gu, Jakub Grudzien Kuba, Yuanpei Chen, Yali Du, Long Yang, Alois C. Knoll, and Yaodong Yang. Safe multi-agent reinforcement learning for multi-robot control. *Artif. Intell.*, 319:103905, 2023.

[Guo *et al.*, 2023] Jiaying Guo, Long Cheng, and Shen Wang. Cotv: Cooperative control for traffic light signals and connected autonomous vehicles using deep reinforcement learning. *IEEE Trans. Intell. Transp. Syst.*, 2023.

[Gupta *et al.*, 2017] Jayesh K. Gupta, Maxim Egorov, and Mykel J. Kochenderfer. Cooperative multi-agent control using deep reinforcement learning. In *Proc. 16th Int. Conf. Auto. Agents MultiAgent Syst.*, volume 10642, pages 66–83, 2017.

[Iqbal and Sha, 2019a] Shariq Iqbal and Fei Sha. Actor-attention-critic for multi-agent reinforcement learning. In *Proc. 36th Int. Conf. Mach. Learn.*, volume 97, pages 2961–2970, 2019.

[Iqbal and Sha, 2019b] Shariq Iqbal and Fei Sha. Coordinated exploration via intrinsic rewards for multi-agent reinforcement learning. 2019.

[Jeon *et al.*, 2022] Jeewon Jeon, Woojun Kim, Whiyoung Jung, and Youngchul Sung. MASER: multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *Proc. 39th Int. Conf. Mach. Learn.*, volume 162, pages 10041–10052, 2022.

[Jin and Ma, 2019] Junchen Jin and Xiaoliang Ma. A multi-objective agent-based control approach with application in intelligent traffic signal system. *IEEE Trans. Intell. Transp. Syst.*, 20(10):3900–3912, 2019.

[Kurach *et al.*, 2020] Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michal Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. In *Proc. 34th AAAI Conf. Artif. Intell.*, pages 4501–4510, 2020.

[Li and Gajane, 2023] Jiong Li and Pratik Gajane. Curiosity-driven exploration in sparse-reward multi-agent reinforcement learning. 2023.

[Li *et al.*, 2023] Sichen Li, Weihao Hu, Di Cao, Zhe Chen, Qi Huang, Frede Blaabjerg, and Kaiji Liao. Physics-model-free heat-electricity energy management of multiple microgrids based on surrogate model-enabled multi-agent deep reinforcement learning. *Applied Energy*, 346:121359, 2023.

[Liu *et al.*, 2022] Haonan Liu, Liansheng Zhuang, Yihong Huang, and Cheng Zhao. VAAC: v-value attention actor-critic for cooperative multi-agent reinforcement learning. In *Proc. 29th Int. Conf. Neural Inf. Process.*, volume 13623, pages 562–573, 2022.

[Liu *et al.*, 2023] Boyin Liu, Zhiqiang Pu, Yi Pan, Jianqiang Yi, Yanyan Liang, and Du Zhang. Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *Proc. 40th Int. Conf. Mach. Learn.*, volume 202, pages 21937–21950, 2023.

[Lowe *et al.*, 2017] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Proc. Int. Conf. Neural Inf. Process. Syst. 30*, pages 6379–6390, 2017.

[Ma *et al.*, 2022] Zixian Ma, Rose Wang, Fei-Fei Li, Michael S. Bernstein, and Ranjay Krishna. ELIGN: expectation alignment as a multi-agent intrinsic reward. In *Proc. Int. Conf. Neural Inf. Process. Syst. 35*, 2022.

[Mehta *et al.*, 2023] Kinal Mehta, Anuj Mahajan, and Pawan Kumar. Effects of spectral normalization in multi-agent reinforcement learning. In *Proc. IEEE Int. Joint Conf. Neural Netw. 2023*, pages 1–8. IEEE, 2023.

[Mguni *et al.*, 2022] David Henry Mguni, Taher Jafferjee, Jianhong Wang, Nicolas Perez Nieves, Oliver Slumbers, Feifei Tong, Yang Li, Jiangcheng Zhu, Yaodong Yang, and

Jun Wang. LIGS: learnable intrinsic-reward generation selection for multi-agent learning. In *Proc. 10th Int. Conf. Learn. Represent.*, 2022.

[Oliehoek and Amato, 2016] Frans A. Oliehoek and Christopher Amato. *A Concise Introduction to Decentralized POMDPs*. Springer Briefs in Intelligent Systems. Springer, 2016.

[Pathak *et al.*, 2017] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *Proc. 34th Int. Conf. Mach. Learn.*, volume 70, pages 2778–2787, 2017.

[Rashid *et al.*, 2018] Tabish Rashid, Mikayel Samvelyan, Christian Schröder de Witt, Gregory Farquhar, Jakob N. Foerster, and Shimon Whiteson. QMIX: monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proc. 35th Int. Conf. Mach. Learn.*, volume 80, pages 4292–4301, 2018.

[Savinov *et al.*, 2019] Nikolay Savinov, Anton Raichuk, Damien Vincent, Raphaël Marinier, Marc Pollefeys, Timothy P. Lillicrap, and Sylvain Gelly. Episodic curiosity through reachability. In *Proc. 7th Int. Conf. Learn. Reprsent.*, 2019.

[Schulman *et al.*, 2016] John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation. In *Proc. 4th Int. Conf. Learn. Represent.*, 2016.

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

[Son *et al.*, 2019] Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Hostallero, and Yung Yi. QTRAN: learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *Proc. 36th Int. Conf. Mach. Learn.*, volume 97, pages 5887–5896, 2019.

[Sun *et al.*, 2023] Shaoqi Sun, Yuanzhao Zhai, Kele Xu, Dawei Feng, and Bo Ding. Progressive diversifying policy for multi-agent reinforcement learning. In *Proc. IEEE Int. Conf. Acoust., Speech and Signal Process.*, pages 1–5, 2023.

[Sunehag *et al.*, 2018] Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinícius Flores Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z. Leibo, Karl Tuyls, and Thore Graepel. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *Proc. 17th Int. Conf. Auto. Agents MultiAgent Syst.*, pages 2085–2087, 2018.

[Wang *et al.*, 2020] Tonghan Wang, Jianhao Wang, Yi Wu, and Chongjie Zhang. Influence-based multi-agent exploration. In *Proc. 8th Int. Conf. Learn. Reprsent.*, 2020.

[Wang *et al.*, 2021] Yihan Wang, Beining Han, Tonghan Wang, Heng Dong, and Chongjie Zhang. DOP: off-policy multi-agent decomposed policy gradients. In *Proc. 9th Int. Conf. Learn. Represent.*, 2021.

[Wang *et al.*, 2022] Li Wang, Yupeng Zhang, Yujing Hu, Weixun Wang, Chongjie Zhang, Yang Gao, Jianye Hao, Tangjie Lv, and Changjie Fan. Individual reward assisted multi-agent reinforcement learning. In *Proc. 39th Int. Conf. Mach. Learn.*, volume 162, pages 23417–23432, 2022.

[Xie *et al.*, 2023] Zaipeng Xie, Yufeng Zhang, Chentai Qiao, and Sitong Shen. IPERS: individual prioritized experience replay with subgoals for sparse reward multi-agent reinforcement learning. In *Proc. 26th Eur. Conf. Artif. Intell.*, volume 372, pages 2760–2767, 2023.

[Xu *et al.*, 2023a] Pei Xu, Junge Zhang, and Kaiqi Huang. Exploration via joint policy diversity for sparse-reward multi-agent tasks. In *Proc. 32th Int. Joint Conf. Artif. Intell.*, pages 326–334, 2023.

[Xu *et al.*, 2023b] Pei Xu, Junge Zhang, Qiyue Yin, Chao Yu, Yaodong Yang, and Kaiqi Huang. Subspace-aware exploration for sparse-reward multi-agent tasks. In *Proc. 37th AAAI Conf. Artif. Intell.*, pages 11717–11725, 2023.

[Yang *et al.*, 2020] Hsuan-Kung Yang, Po-Han Chiang, Min-Fong Hong, and Chun-Yi Lee. Flow-based intrinsic curiosity module. In *Proc. 29th Int. Joint Conf. Artif. Intell.*, pages 2065–2072, 2020.

[Yang *et al.*, 2021] Huanhuan Yang, Dianxi Shi, Chenran Zhao, Guojun Xie, and Shaowu Yang. Ciexplore: Curiosity and influence-based exploration in multi-agent cooperative scenarios with sparse rewards. In *Proc. 30th ACM Int. Conf. Inf. Knowl. Manage.*, pages 2321–2330, 2021.

[Yu *et al.*, 2022] Chao Yu, Akash Velu, Eugene Vinitsky, Jiaxuan Gao, Yu Wang, Alexandre M. Bayen, and Yi Wu. The surprising effectiveness of PPO in cooperative multi-agent games. In *Proc. Int. Conf. Neural Inf. Process. Syst. 35*, 2022.

[Yu *et al.*, 2023] Yang Yu, Qiyue Yin, Junge Zhang, Hao Chen, and Kaiqi Huang. Underexplored subspace mining for sparse-reward cooperative multi-agent reinforcement learning. In *Proc. IEEE Int. Joint Conf. Neural Netw. 2023*, pages 1–8, 2023.

[Zhang and Yu, 2023] Yucong Zhang and Chao Yu. EXPODE: exploiting policy discrepancy for efficient exploration in multi-agent reinforcement learning. In *Proc. 22th Int. Conf. Auto. Agents MultiAgent Syst.*, pages 58–66, 2023.

[Zheng *et al.*, 2021] Lulu Zheng, Jiarui Chen, Jianhao Wang, Jiamin He, Yujing Hu, Yingfeng Chen, Changjie Fan, Yang Gao, and Chongjie Zhang. Episodic multi-agent reinforcement learning with curiosity-driven exploration. In *Proc. Int. Conf. Neural Inf. Process. Syst. 34*, pages 3757–3769, 2021.