FoX: Formation-Aware Exploration in Multi-Agent Reinforcement Learning

Yonghyeon Jo, Sunwoo Lee, Junghyuk Yeom, Seungyul Han[*]

Artificial Intelligence Graduate School, UNIST, Ulsan, South Korea
{yonghyeonjo, sunwoolee, yumjunghyuk, syhan}@unist.ac.kr

Abstract

Recently, deep multi-agent reinforcement learning (MARL) has gained significant popularity due to its success in various cooperative multi-agent tasks. However, exploration remains a challenging problem in MARL due to the partial observability of the agents and the exploration space that can grow exponentially as the number of agents increases. Firstly, in order to address the scalability issue of the exploration space, we define a formation-based equivalence relation on the exploration space and aim to reduce the search space by exploring only meaningful states in different formations. Then, we propose a novel formation-aware exploration (FoX) framework that encourages partially observable agents to visit the states in diverse formations by guiding them to be well aware of their current formation solely based on their observations. Numerical results show that the proposed FoX framework significantly outperforms the state-of-the-art MARL algorithms on Google Research Football (GRF) and sparse Starcraft II multi-agent challenge (SMAC) tasks.

Introduction

In recent years, multi-agent reinforcement learning (MARL) has successfully solved various real-world application challenges such as traffic control (Chu et al. 2019; Wang et al. 2020d), games (Vinyals et al. 2019), and robotic controls (Kalashnikov et al. 2018). With its increasing popularity, many multi-agent learning algorithms have been introduced (Tan 1993; Lowe et al. 2017; Rashid et al. 2020). The MARL algorithms can be broadly classified into three categories. First, the fully decentralized methods (Tan 1993; Zhang et al. 2018a) train agents independently. Second, the fully centralized methods (Gupta, Egorov, and Kochenderfer 2017) share information among agents for more efficient training. Finally, The CTDE methods (Foerster et al. 2018; Sunehag et al. 2017; Rashid et al. 2020) provide global information during training to alleviate the nonstationarity while maintaining scalability. As a result, MARL algorithms have achieved significant success in various applications.

However, MARL algorithms still face challenges from partial observability (Sukhbaatar, Fergus et al. 2016). Even in the traditional RL, exploration is critical to prevent the
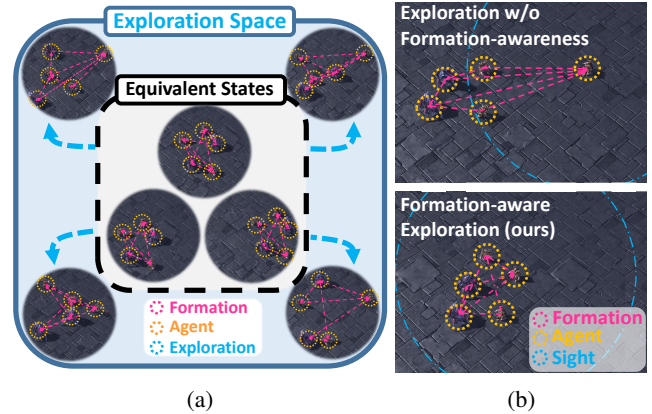
---
[*]Corresponding author

Figure 1: (a) Illustration of formation-based state equivalence. Defining state equivalence under formations can reduce the search space efficiently. (b) Illustration of formation-awareness. Our method encourages agents to be fully aware of the formation.

agent from converging to sub-optimal policies (Ostrovski et al. 2017). To address exploration in traditional RL, exploration methods based on curiosity (Ostrovski et al. 2017; Pathak et al. 2017; Burda et al. 2018) have been proposed. Among the curiosity explorations, count-based exploration has demonstrated simple yet efficient exploration in single-agent RL problems (Burda et al. 2018; Tang et al. 2017). Unfortunately, the single-agent approaches may struggle in the context of MARL as partial observability makes efficient exploration even more challenging. As consideration for agent relationships exponentially increases the search space, (Wang et al. 2019) techniques such as utilizing the influence between agents (Wang et al. 2019) or social influence (Jaques et al. 2019) have been introduced, but despite these efforts, efficient exploration in MARL still remains a challenge (Zheng et al. 2021).

In realistic scenarios, soccer for example, it is often impractical to consider the entire information of the field upon formulating a strategy. Rather, coaches formulate strategies based on a formation, which contains the distance information between the players in soccer games. For a team with high speed, a formation that allows quick transition between offense and defense would be beneficial while fac-

ing an aggressive opponent might be helpful to adopt a defensive formation. Thus, in the soccer game, formation information is critical to winning the game against the opponents. Inspired by the real soccer game example, we define a formation in a multi-agent environment based on the differences between the agents. Based on the formation, we propose a novel formation-aware exploration (FoX) framework to visit diverse formations instead of visiting large search spaces. In FoX, we have two main contributions: 1) By defining the formation-based equivalence relation on the exploration space, we can efficiently visit states in diverse formations as illustrated in Figure 1(a), and 2) To overcome the partial observability of agents, we design an intrinsic reward to encourage each agent to be aware of the formation viewed by the agent as shown in Figure 1(b). In order to show the superiority of our exploration method, we conduct performance comparisons on various challenging multi-agent environments such as StarCraft II Multi-Agent Challenge (SMAC) (Samvelyan et al. 2019), and Google Research Football (GRF) (Kurach et al. 2020)

Related Work

Deep Multi-Agent Reinforcement Learning

With its increasing popularity, numerous MARL algorithms (Tan 1993; Foerster et al. 2016; Sunehag et al. 2017; Yang et al. 2020; Iqbal and Sha 2019; Yu et al. 2022) have been proposed. The fully decentralized methods (Tan 1993; Zhang et al. 2018b) consider agents independently during training. To address partial observability under decentralized settings, the CTDE paradigm provides global information during training. QMIX (Rashid et al. 2020) utilizes a mixing network for individual Q values of each agent to decompose the team's expected return. QTRAN (Son et al. 2019) and QPLEX (Wang et al. 2020a) extend the value decomposition according to the principle of IGM. While the previous methods approach MARL problems with a value-based approach, (Foerster et al. 2018; Lowe et al. 2017; Wang et al. 2020e; Peng et al. 2021) address the problem via a policy-based approach.

Exploration in State Space

In environments where rewards are sparse or the state is very complex, the agent must be provided with inner motivation for efficient exploration. (Osband et al. 2016; Houthooft et al. 2016a; Nair et al. 2018; Jin et al. 2020; Machado, Bellemare, and Bowling 2020) utilizes prediction error as intrinsic rewards for exploration. (Strehl and Littman 2008; Bellemare et al. 2016; Ostrovski et al. 2017; Tang et al. 2017) designs their intrinsic rewards based on visitation count methods. While (Strehl and Littman 2008) proposes using a density model as an approximator for the number of visits to a state, (Tang et al. 2017) hashes similar states to discretize the high-dimensional state space. (Houthooft et al. 2016b; Haarnoja et al. 2018; Han and Sung 2021; Eysenbach et al. 2018)approaches exploration via information-theoretical objectives.

Intrinsic Motivations in MARL

As partial observability constraints efficient exploration in MARL, many studies have been conducted to address the issue (Du et al. 2019; Wang et al. 2022; Gupta et al. 2021; Liu et al. 2021). (Mahajan et al. 2019) proposes maximizing the mutual information of a latent variable and agent trajectories. Using external state information such as the agents' location (Liu et al. 2023) can greatly assist training, but assuming such information may affect generality in the application. (Jaques et al. 2019; Wang et al. 2019) addresses the challenge by exploring based on the influence between the agents, (Wang et al. 2020b,c) assigns roles to the agents for task-appropriate behavior. (Jeon et al. 2022) targets observations with the highest sum of individual and global value as sub-goals. On the other hand, (Zheng et al. 2021) defines curiosity based on prediction errors on individual Q-values, whereas (Li et al. 2021) encourages agent diversity via local trajectory.

Preliminaries

Decentralized POMDP

A fully cooperative multi-agent task can be seen as a decentralized partially observable Markov decision process (Dec-POMDP) (Oliehoek, Amato et al. 2016), represented as a tuple $G = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, Z, O, \mathcal{O}, I, n, \gamma \rangle$. $\mathcal{S}$ is the true state space of the environment or the observation product space of all n agents, where I is a finite set of n agents. $\mathcal{A}$ is the set of actions. At each time step t, each agent $i \in I$ receives d-dimensional observation vector $o_t^i \in \mathcal{O}^i$ from the environment according to the observation function $O^i(s)$. Then, agents select joint action $\mathbf{a}_t = (a_t^0, \cdots, a_t^{n-1})$, and the next state s_{t+1} and the global reward $r_t = R(s_t, \mathbf{a}_t)$ are generated from the environment based on the transition function $P(\cdot|s_t, \mathbf{a}_t)$ and the reward function R. Z is the latent space, and $\gamma \in [0, 1)$ is the discount factor. Each agent conditions its policy $\pi^i(a_t^i|\tau_t^i)$, where $\tau_t^i = (o_0, a_0, \cdots, o_t)$ is a trajectory of i-th agent, as the agents choose their action based on their local observations. Individual policies π_i form the joint policy $\pi = \prod_{i=0}^{n-1} \pi^i$, and the main objective of RL is to maximize the cumulative reward sum $\mathbb{E}_{s_0, \mathbf{a}_0, \cdots}[\sum_{t=0}^{T-1} r_t]$ given from the environment, where T is the episode length.

Centralized Training Decentralized Execution

The CTDE methods provide information of the full information of agents only during training to mitigate the challenges from partial observability while maintaining decentralized execution. Among the CTDE methods are the value decomposition methods where the global value function $Q^{tot}(\boldsymbol{\tau}, \mathbf{a})$ is decomposed into individual values. A popular value decomposition method is QMIX (Rashid et al. 2020), which decomposes the global value function through a mixing network to individual agent utility functions Q^i as follows:

$$\operatorname*{argmax}_{\mathbf{a}} Q^{tot}(\boldsymbol{\tau}, \mathbf{a}) = \prod_{i=0}^{n-1} \operatorname*{argmax}_{a^i} Q^i(\tau^i, a^i), \quad (1)$$

where $\boldsymbol{\tau} = (\tau_0, \cdots, \tau_{n-1})$ is the joint trajectory. Furthermore, the individual-global-max (IGM) condition must be
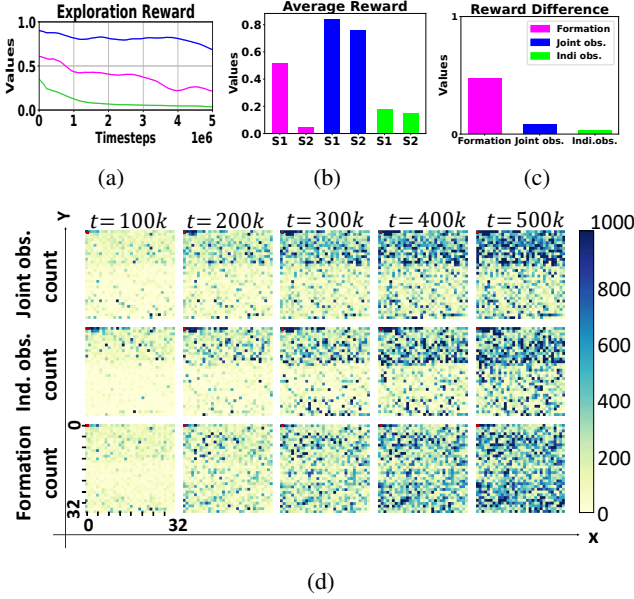
(a)　　　　　(b)　　　　　(c)

(d)

Figure 2: (a) Pure exploration reward based on visitation count graph. (b) Average reward graph on set of rarely visited formation S_1 and frequently visited formation S_2. (c) The reward difference graph for each component. (d) Heatmap of diverse formation-based exploration space. With initial spawn formation at (0,0), a farther point in the heatmap indicates a larger difference in formations.

satisfied to maintain consistency between local and global greedy actions (Son et al. 2019). The consistency between the global and local greedy actions allows greedy local actions to result in optimal global actions.

FoX: Formation-aware Exploration

In this section, we introduce FoX, a novel exploration framework for cooperative multi-agent reinforcement learning.

Motivation of Formation-aware Exploration

In general, the assumption that agents have information about the true global state may compromise generality. Thus, in this paper, we consider a more general scenario that agents explore the observation product space, so we define an exploration space $\mathcal{S}^e$ as the observation product space $\mathcal{S}^e := \prod_{i \in I} \mathcal{O}^i$. Then, agents should explore diverse exploration states defined as joint observations $s_e := (o_0, \cdots, o_{n-1}) \in \mathcal{S}^e$ to experience various interactions between agents well.

However, visiting all exploration states is a challenging problem since $\mathcal{S}^e$ grows exponentially if the number of agents and the dimension of observation space increase, and it suffers from the curse of dimensionality (Bellman 1966). Thus, as motivated in real soccer games, we define the formation $\mathcal{F}$ based on the observation differences between agents, and aim to explore diverse formations instead of the entire exploration space to reduce the search space.

In order to show the efficiency of our formation-based exploration, we conducted pure-exploration experiments in

a modified 2s3z environment of SMAC (Samvelyan et al. 2019). Following the previous works on visitation count methods (Bellemare et al. 2016), in this setup, the agent only gets the pure exploration reward for count-based exploration $r_t \propto \frac{1}{\sqrt{N(s)}}$ to explore rarely visited components s, where $N(s)$ is the count for visitation on state s. Here, we consider three types of components s for comparison: 1) joint observations $(o^0, \cdots, o^{n-1})$, 2) individual observations o^i, and 3) proposed formations $\mathcal{F}$, and details of visitation count are provided in Appendix A. For all three cases, we simultaneously count the visitation of formation $N(\mathcal{F})$ during training, and compute the average r_t of two state sets: a set of states that rarely visit the same formation $S_1 = \{s_t \in \mathcal{S} | 1 \leq N(\mathcal{F}(s_t)) \leq 100\}$ and another set of states that frequently visit the same formation $S_2 = \{s_t \in \mathcal{S} | N(\mathcal{F}(s_t)) > 100\}$ at $t = 500k$. For efficient exploration, it is important to rapidly decrease rewards for frequently visited states S_2, while maintaining high rewards for novel states S_1.

Figure 2 shows the result of pure exploration, where Figure 2(a) shows the average pure exploration reward, Figure 2(b) shows the average reward sets S_1 and S_2, Figure 2(c) shows the reward difference between the two sets S_1 and S_2, and Figure 2(d) illustrates the heatmap of the visitation frequency of diverse formations in the exploration space. In the heatmap, a farther point in the heatmap indicates the larger difference in formations with initial spawn formation at (0,0). From Figure 2(a), we can observe no significant change in reward in the case of joint observations, and rapid decrease in the reward in the case of individual observations. According to Figure 2(b) and Figure 2(c), exploration based on joint observations or individual observations is unable to highlight the difference in rewards between the set of novel states S_1, and the set frequently visited states S_2. On the other hand, in the case of our proposed formation, the reward difference between the two sets is prominently displayed, as shown in Figure 2(c). Such a tendency in the intrinsic rewards demonstrates that the reward decrease of our proposed formation shown in Figure 2(a) is indeed appropriate, and that formation-based exploration enables efficient exploration in MARL.

The efficiency of formation-based exploration is also illustrated in Figure 2(d). According to our heatmap, the joint observation or the individual observation-based exploration only explores a limited area of the heatmap while in the case of the individual observation, it visits a farther area in the heatmap than the joint observation case, but it still cannot visit the diverse area in the heatmap. On the other hand, we can observe that our formation-based exploration allows the agents to successfully visit more diverse areas in the heatmap. Thus, the pure exploration example shows that our method yields better exploration to visit diverse formations in exploration space.

Formation Arrangement

In order to reduce the search space, we aim to define a formation based on the observation differences $o^i - o^j$ between i-th agent and j-th agent, as explained in previous

sections. However, The dimension of the observation difference $o^i - o^j$ is still so large that it makes exploration difficult, we first reduce the dimension of the observation difference $o_i - o_j$ as a 2-dimensional vector $D^{ij} := (||o^i - o^j||_2, c(o^i - o^j)) \in \mathbb{R} \times \mathbb{Z}$, where $|| \cdot ||_2$ indicates a 2-norm operator, and $c : \mathbb{R}^d \to \mathbb{Z}$ is a mapping function that transforms the angle of inputs into integers based on hash coding (Aggarwal and Verma 2015). For hash coding, SimHash (Tang et al. 2017) is used to measure the approximate angle of the observation difference, and detail of hash coding is provided in Appendix A. Then, D^{ij} contains the distance and angle information of $o^i - o^j$ in a much lower dimension.

Now, using the compressed difference D^{ij}, we define the formation $\mathcal{F}_{F^0, \cdots, F^{n-1}}$ of the exploration state $s^e = (o^0, \cdots, o^{n-1}) \in \mathcal{S}^e$ as follows:

$$\mathcal{F}_{F^0, \cdots, F^{n-1}}(s_e) = (\mathcal{F}^0_{F^0}, \cdots, \mathcal{F}^{n-1}_{F^{n-1}}), \qquad (2)$$

where n is the number of agents, $F^i = \{j_0, \cdots, j_{k-1}\} \subset I$ is an arbitrary ordered agent index set, and $\mathcal{F}^i_{F^i}$ is the individual formation viewed by i-th agent, defined as

$$\mathcal{F}^i_{F^i} = (D^{ij_0}, \cdots, D^{ij_{k-1}}). \qquad (3)$$

From the definition, the formation of the exploration state s^e contains all the difference information of (i, j) agent pairs for all $i \in I$, $j \in F^i$. From now on, we denote the formation as $\mathcal{F}$ for better clarity.

Finally, we can define a formation-based binary relation on the exploration state space as

$$\sim_{\mathcal{F}} := \{(s_1, s_2) \in \mathcal{S}^e \times \mathcal{S}^e \mid \mathcal{F}(s_1) = \mathcal{F}(s_2)\}, \qquad (4)$$

and we can easily prove that $\sim_{\mathcal{F}}$ is an equivalence relation on the exploration space $\mathcal{S}^e$ by Lemma 1.

Lemma 1 (Formation-based equivalence relation). *The binary relation $\sim_{\mathcal{F}}$ is an equivalence relation on the exploration space $\mathcal{S}^e$, i.e., two exploration states s_1 and s_2 in the exploration state $\mathcal{S}^e$ are equivalent under $\sim_{\mathcal{F}}$ if $\mathcal{F}(s_1) = \mathcal{F}(s_2)$.*

Proof) Proof of Lemma 1 is provided in Appendix B.

Here, the dimension of formation $\mathcal{F}$ is nk and the dimension of exploration space is nd. k depends on the selection of F^i, but in most cases, the observation dimension d is much larger than k. Thus, we can efficiently reduce the dimension of search space. In addition, based on the equivalence relation $\sim_{\mathcal{F}}$, the agents explore the exploration states in diverse formations without visiting the equivalent states that have the same formation, and it can further reduce the search space and enables better exploration as shown in Figure 2.

Formation-aware Exploration Objective

In this subsection, we propose a formation-aware exploration objective to explore states in diverse formations as

$$\underbrace{\mathcal{H}(\mathcal{F}_t)}_{(a)} + \frac{1}{n(k+1)} \sum_{i=0}^{n-1} \sum_{j \in F^{i+}} \underbrace{\mathcal{I}(\mathcal{F}^i_{F^i, t+1}; z^j_t | \tau^{F^{i+}}_t, z^{F^{i+} \setminus \{j\}}_t)}_{(b)},$$

where $F^{i+} = F^i \cup \{i\}$, a^F indicates the set $\{a^i | i \in F\}$, $\mathcal{H}(X) = -p(X) \log p(X)$ is the information entropy of a
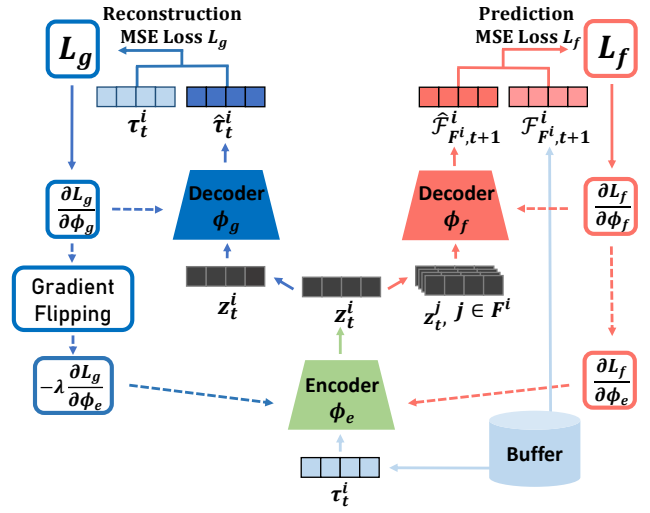

Figure 3: $\mathcal{F}$-Net Architecture

random variable X, and $\mathcal{I}(X; Y) = \mathcal{H}(X) - \mathcal{H}(X|Y)$ is the mutual information between random variables X and Y.

In the exploration objective, (a) is the entropy of formations, so maximizing (a) encourages the agents to visit states in diverse formations as illustrated in Figure 1(a). Here, maximizing (a) can be accomplished by visitation count-based exploration (Ostrovski et al. 2017). We define a formation-based count exploration reward $r^{exp}_t = \frac{1}{\sqrt{N(\mathcal{F}_t)}}$, then agents get higher rewards if they visit rarely visited formations. In order to count the number of formation $N(\mathcal{F}_t)$, we use $round(\cdot)$ method to discretize the formation $\mathcal{F}_t$, and then count the visitation number in the discretized bins. A detailed explanation for counting is provided in Appendix A. However, agents are partially observable in MARL, so visiting diverse formations by maximizing (a) can be a challenging problem if the agents do not know the formation information at all.

For instance, the upper illustration of Figure 1(b) depicts a situation where an agent lacks knowledge of the formation information. In the SMAC (Samvelyan et al. 2019) environment, each agent is assigned a sight range. While an agent can observe other agents within its sight range, it remains uninformed about agents outside its sight range. Consequently, if the other agents are positioned outside its sight range, the agent may not be fully aware of the formation, as the formation includes information about all agents involved. Therefore, we consider an additional formation-aware objective (b), which is the mutual information between the latent variable z^j_t, $\forall j \in F^{i+}$ and the next formation $\mathcal{F}^i_{F^i, t+1}$, where the trajectory τ^j_t, the trajectories $\tau^{F^{i+} \setminus \{j\}}_t$ and the latent variables $z^{F^{i+} \setminus \{j\}}_t$ of other agents are given. Then, maximizing (b) guides the i-th agent to be aware of not only its own next formation $\mathcal{F}^i_{F^i, t+1}$ but also all formations $\mathcal{F}^j_{F^j, t+1}$, $\forall j$ s.t. $i \in F^j$ to which i-th agent belongs as illustrated in Figure 1(b). Thus, we maximize both (a) and (b) to make agents visit exploration states in diverse
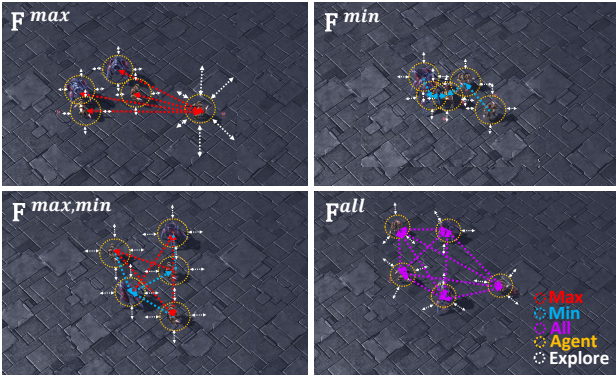
Figure 4: Formations with various agent index set F. While $F^{i,max}$, $F^{i,min}$ focuses on particular agent relationships, $F^{i,all}$ and $F^{i,max,min}$ considers extensive agent relationships.

formations while recognizing their formations well.

In order to maximize (b), we derive an evidence lower bound for the mutual information term based on the variational inference approaches (Kingma and Welling 2013; Li et al. 2021), and then we will maximize the evidence lower bound to increase the mutual information.

Lemma 2 (Evidence lower bound). *Mutual information (b) can be lower-bounded as*

$$
\mathcal{I}(\mathcal{F}^i_{F^i,t+1}; z^j_t | \tau^{F^{i+}}_t, z^{F^{i+}\backslash\{j\}}_t) \geq
$$
$$
\mathbb{E}_{\mathcal{F}^i_{F^i,t+1}, z^j_t \sim q_{\phi_e}} \big[\log q_{\phi_f}(\mathcal{F}^i_{F^i,t+1} | \tau^{F^{i+}}_t, z^{F^{i+}}_t) \quad (5)
$$
$$
- \log p(\mathcal{F}^i_{F^i,t+1} | \tau^{F^{i+}}_t, z^{F^{i+}\backslash\{j\}}_t) \big],
$$

where $q_{\phi_f}(\cdot | \tau^{F^{i+}}_t, z^{F^{i+}}_t)$ *is an arbitrary posterior distribution, and* ϕ_f *and* ϕ_e *parameterize the distributions* q_{ϕ_e} *and* q_{ϕ_f}, *respectively.*

Proof) Proof of Lemma 2 is provided in Appendix B.

In the proof of Lemma 2, z^j_t is averaged over the distribution $q_{\phi_e}(\cdot | \tau^{F^{i+}}_t, z^{F^{i+}\backslash\{j\}}_t)$, but it makes difficult to utilize the latent variable z^j_t in the decentralized execution setup since j-th agent cannot exploit the other agent's trajectory information. Thus, for practical implementation, we consider the approximate distribution $q_{\phi_e}(\cdot | \tau^j_t)$ instead of $q_{\phi_e}(\cdot | \tau^{F^{i+}}_t, z^{F^{i+}\backslash\{j\}}_t)$. Then, we can formulate the encoder-decoder structure such as variational auto-encoder (VAE) (Kingma and Welling 2013), and we define an encoder $E_{\phi_e}(\tau^i_t)$ whose outputs are the mean and standard deviation of the latent variable z^i_t to sample the latent variable and a decoder $D_{\phi_f}(z^{F^{i+}}_t)$ whose output is the prediction of next formation $\hat{\mathcal{F}}^i_{F^i,t+1}$. Here, note that we exploit the agent's trajectory τ^i_t to extract z^i_t from the encoder, but the trajectory τ^i_t may contain the redundant information regardless of the formation prediction, and it can be delivered to the latent variable z^i_t. The irrelevant information can disturb formation-aware exploration as in the case of Pathak et al. (2017). Thus, we aim to consider a gradient flipping

(GF) technique proposed in the field of domain adaptation (Ganin and Lempitsky 2015) to remove the redundant information. In order to apply GF in our setup, we consider an additional trajectory decoder $D_{\phi_g}(z^i_t)$ whose output is the reconstruction of trajectory $\hat{\tau}^i_t$ and we update the encoder parameter ϕ_e to prevent the trajectory reconstruction, while decoder strives to reconstruct the trajectory. Then, we can prevent delivering useless information from the trajectory τ^i_t to the latent variable z^i_t as proposed in Ganin and Lempitsky (2015).

In summary, we can construct $\mathcal{F}$-Net architecture as shown in Figure 3. We define a formation prediction loss $L_f(\phi_e, \phi_f) = \frac{1}{n} \sum_{i=0}^{n-1} MSE(\mathcal{F}^i_{F^i,t+1}, \hat{\mathcal{F}}^i_{F^i,t+1})$, a trajectory reconstruction loss $L_g(\phi_e, \phi_g) = \frac{1}{n} \sum_{i=0}^{n-1} MSE(\tau^i_t, \hat{\tau}^i_t)$, and Kullback-Leibler divergence loss $L_{KL}(\phi_e) = D_{KL}(q_{\phi_e}(\cdot | \tau^i_t) || \mathcal{N}(0, \mathbf{I}))$ as in Kingma and Welling (2013); Ganin and Lempitsky (2015), where $MSE(x, y) = \mathbb{E}[(x - y)^2]$ is the mean square error loss and $\mathcal{N}(0, \mathbf{I})$ is the multi-variate standard normal distribution with identity matrix $\mathbf{I}$. Based on the loss functions, we can update the encoder-decoder parameters ϕ_e, ϕ_f, ϕ_g as follows:

$$
\phi_e \leftarrow (1 - \alpha)\phi_e + \alpha \left(\frac{\partial L_f}{\partial \phi_e} + \frac{\partial L_{KL}}{\partial \phi_e} - \lambda_{GF} \frac{\partial L_g}{\partial \phi_e} \right)
$$
$$
\phi_f \leftarrow (1 - \alpha)\phi_f + \alpha \frac{\partial L_f}{\partial \phi_f}, \ \phi_g \leftarrow (1 - \alpha)\phi_g + \alpha \frac{\partial L_g}{\partial \phi_g},
$$
$$(6)$$

where α is a learning rate, λ_{GF} is a hyperparameter that controls the gradient flipping effect, and we fix $\lambda_{GF} = 0.1$ in our setup. Note that ϕ_e maximizes the reconstruction loss L_g to prevent the trajectory reconstruction.

Finally, in order to maximize the evidence lower bound in (5) to be aware of the formation information, we design the intrinsic reward r^{aware} that approximately represents the evidence lower bound as follows:

$$
r^{aware}_t = \frac{1}{n} \sum_{i=0}^{n-1} \left(\log q_{\phi_f}(\hat{\mathcal{F}}^i_{t+1} | \tau^{F^{i+}}_t, z^{F^{i+}}_t) \right.
$$
$$
\left. - \frac{1}{k+1} \sum_{j \in F^{i+}} \log p(\hat{\mathcal{F}}^i_{t+1} | \tau^{F^{i+}}_t, z^{F^{i+}\backslash\{j\}}_t) \right), \quad (7)
$$

the detailed implementation of r^{aware}_t using formation prediction loss is provided in Appendix C.

Finally, with the extrinsic reward r^{ext}_t given from the environment and intrinsic rewards r^{exp}_t and r^{aware}_t for formation-aware exploration, we can define the total reward $r^{tot}_t = r^{ext}_t + \beta_1 r^{exp}_t + \beta_2 r^{aware}_t$, where β_1 and β_2 are hyperparamers to control the effect of our intrinsic rewards r^{exp}_t and r^{aware}_t, respectively. Then, an overall temporal difference loss to update the total Q-function Q^{tot}_θ in (1) parameterized by θ is defined as

$$
L_{TD}(\theta) = (r^{tot}_t + \gamma \max_{a'} Q^{tot}_{\theta^-}(s_{t+1}, a') - Q^{tot}_\theta(s_t, a_t))^2
$$
$$(8)$$

where θ^- is a parameter for the target network updated by the exponential moving average (EMA). We summarize the proposed FoX framework as Algorithm 1 and Figure 5.
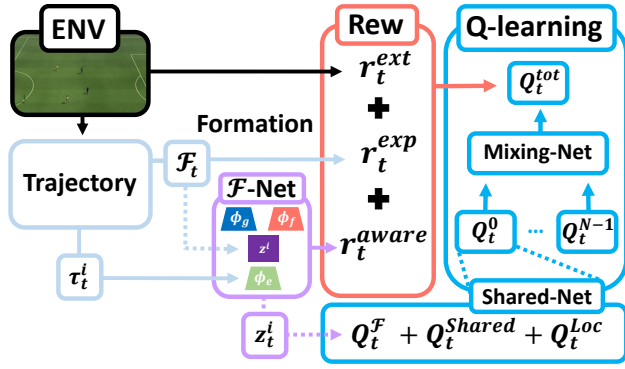
Figure 5: Schema of FoX framework

Algorithm 1: FoX framework

Initialize $\phi_e, \phi_f, \phi_g, \theta, \theta^-$

1: **for** each epoch **do**
2:    **for** each gradient step **do**
3:       Obtain trajectory samples from environment
4:       Choose random batch from buffer $\mathcal{D}$
5:       Calculate formation $\mathcal{F}_t$ from $\mathbf{o}_t$ in (2)
6:       Compute intrinsic rewards r^{exp}, r^{aware}
7:       Compute loss functions L_f, L_g, L_{KL}, L_{TD}
8:       Update $\mathcal{F}$-Net parameters ϕ_e, ϕ_f, ϕ_g based on (6)
9:       Update Q-function parameter θ
10:      Update target parameter θ^- using EMA.
11:      Store samples in buffer $\mathcal{D}$
12:    **end for**
13: **end for**

Selection of Index Set F^i

In arranging formations, various agent index sets F^i can be considered. For example, $F^{i,max} := \{\arg\max_{j \in I \setminus \{i\}} d(i,j)\}$ forms a formation among agents with the biggest difference. In our experiments, we observe the exploration behaviors of various formations based on $F^{i,max}$, $F^{i,min} = \{\arg\min_{j \in I \setminus \{i\}} d(i,j)\}$, $F^{i,max,min} = F^{i,max} \cup F^{i,min}$, and $F^{all} = \{i | i \in I \setminus \{i\}\}$ which is illustrated in Figure 4. We conduct an ablation study for the selection of F^i, and the result shows that $F^{i,max,min}$ shows the best performance.

Formation-based Shared Network

Based on the shared network structure for individual Q-function proposed in Li et al. (2021), we aim to exploit the latent variable z_t^i. Thus, individual Q-function Q^i in (1) can be decomposed as three Q functions as follows:

$$Q^i(\tau_t^i, z_t^i, \cdot) = Q^{Shared}(\tau_t^i, \cdot) + Q^{Loc,i}(\tau_t^i, \cdot) + Q^{\mathcal{F}}(z_t^i, \cdot), \quad (9)$$

where Q^{Shared} is a shared Q-function, $Q^{Loc,i}$ is i-th local Q-function, and $Q^{\mathcal{F}}$ is a Q-function that exploits formation information using z_t^i. As proposed in Li et al. (2021), we also consider $l1$ regularization for $Q^{Loc,i}$ with the regularization coefficient $\lambda_{reg} = 0.1$.

Experiments

In our experiments, we evaluate the performance of the proposed FoX framework in the challenging cooperative multi-agent environments: the sparse StarCraft multi-agent challenge (SMAC) (Samvelyan et al. 2019) and Google Research Football (GRF) (Kurach et al. 2020). The source code of our proposed algorithm is available at https://github.com/hyeon1996/FoX. Detailed settings of the environment can be found in D. In the SMAC and GRF environments, the observation of an agent or enemy force identified for elimination or out of sight is automatically set to zero or remains constant. Thus the difference between observation of such agent remain unchanged, becoming ineffective in formation computation. Therefore, in the SMAC and the GRF environment, our formation can effectively represent the topological changes. For hyperparameters β_1, β_2, we have tested $\beta_1 \in \{0.001, 0.005, 0.01, 0.02, 0.1\}$ and $\beta_2 \in \{0.001, 0.005, 0.01, 0.05\}$ for both environments. We used the best parameters for each environment to demonstrate our experimental results. Detailed hyperparameter setups can be found in Appendix E. We averaged 5 seeds for SMAC tasks and 5 seeds for GRF tasks, and the solid line in the performance graph indicates the average performance, and the shaded part represents the standard deviation.

Performance on Sparse SMAC

In the original scenarios of SMAC, agents are densely rewarded for the damage they have dealt or taken in addition to sparse rewards upon the deaths of an ally or the enemy and defeating the enemy. In the sparse SMAC environment, the agents must learn with only sparse rewards without the dense rewards thus making the task more challenging. Detailed reward settings of the dense and sparse SMAC can be found in Table 1. To leverage the fact that formation-aware exploration is not value-dependent, we conducted experiments in sparse environments. In sparse SMAC, we evaluate the performance of FoX in 6 scenarios: 3m, 8m, 2s3z, 2m_vs_1z, 3s5z, and 1c3s5z. While the 2m_vs_1z scenario has the fewest number of agents with $n = 2$, the 1c3s5z scenario has the highest number of agents with $n = 9$.

The exploration scheme of FoX can be easily applied to other existing MARL algorithms, but as we implement Fox on QMIX (Rashid et al. 2020) during our experiments, we tested the performance of FoX against several QMIX-based algorithms including EMC (Zheng et al. 2021), MASER (Jeon et al. 2022), and ROMA (Wang et al. 2020b). EMC employs exploration based on the prediction error of agent

Event	Dense reward	Sparse reward
Death of an enemy	+10	+10
Death of an ally	-5	-5
Win	+200	+200
Enemy hit-point	-Enemy hit-point	-
Ally hit-point	+Ally hit-point	-
Other elements	+/- other components	-

Table 1: The reward setting for dense and sparse SMAC.
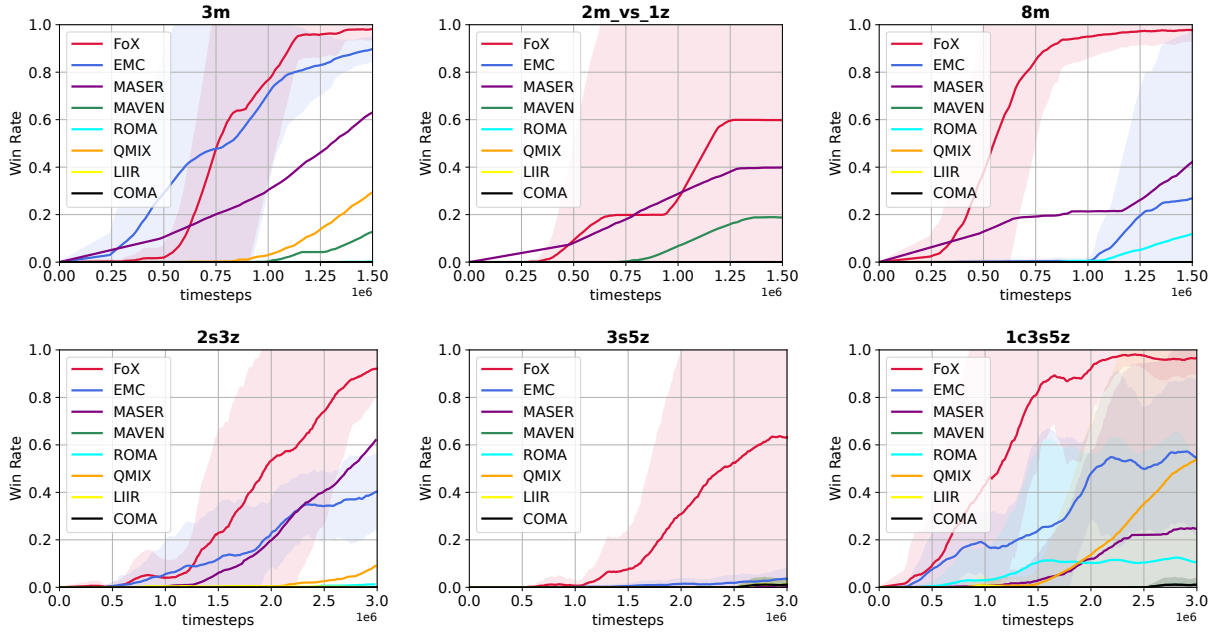
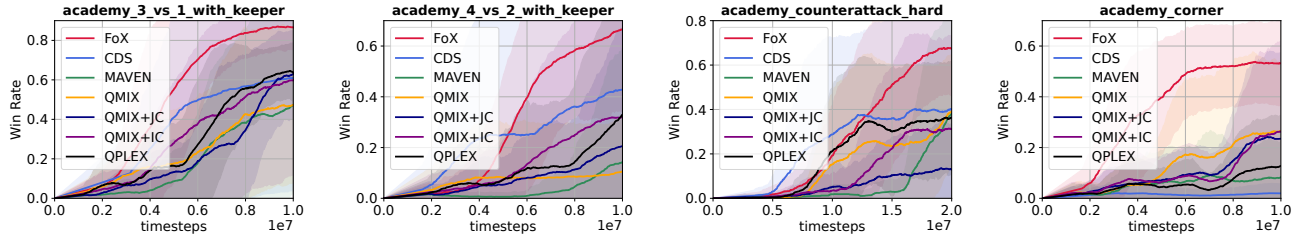Figure 6: Performance results on SMAC(sparse)

Figure 7: Performance results on GRF

influence, MASER assigns subgoals with the highest individual and joint Q value, and ROMA assigns roles based on agent trajectories. FoX was also compared with MAVEN (Mahajan et al. 2019) in the context of exploration via variational inference, LIIR (Du et al. 2019) for intrinsic reward learning, and COMA (Foerster et al. 2018) as a policy-gradient method. All experiments were conducted with the author-provided codes. As both FoX and EMC can be implemented on QMIX or QPLEX (Wang et al. 2020a), we conduct our experiments with QMIX-based FoX and EMC for fair comparison. From the experimental results, we can see that in the relatively easier environment of 3m, EMC shows comparable results to FoX. However, FoX significantly outperforms the other baselines in the other scenarios, where the results in 2m_vs_1z are especially notable as the other baselines struggled to learn the environment.

Performance on GRF

In Google Research Football (Kurach et al. 2020) the agents' observation contains the location of the players, the opponents, and the ball. In our implementation, the episode is ter-

minated and a negative reward is given to the agents when the ball goes to the other half of the court. The reward for the GRF environment is described in Table 2. In GRF, we compare the FoX framework with baselines such as QPLEX (Wang et al. 2020a), CDS (Li et al. 2021), MAVEN (Mahajan et al. 2019), and QMIX (Rashid et al. 2020). QPLEX incorporates the IGM principle into a neural network architecture, utilizing a duplex dueling structure, CDS allows the agents to adaptively decide whether to pursue diversity or share information based on agent trajectories. In addition, we also compare the visitation count method implemented on QMIX to show the efficiency of our formation-based exploration. The variations of QMIX are denoted as QMIX+JC for QMIX with joint observation visitation counts and QMIX+IC for the individual observation visitation counts. In our experiments, we evaluate the performance of FoX on four academy scenarios: $Academy\_3\_vs\_1\_with\_keeper$, $Academy\_4\_vs\_2\_with\_keeper$, $Academy\_counteratta$ ck_hard, and $Academy\_corner$. While the $Academy\_$ $3\_vs\_1\_with\_keeper$ scenario has the least number of agents with $n = 3$, the $Academy\_corner$ scenario has the
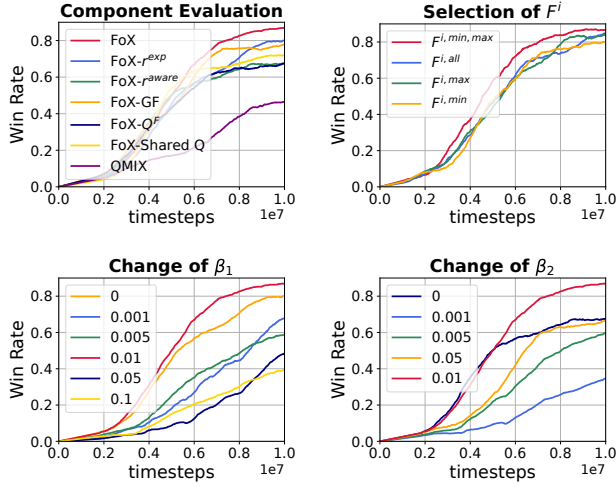
Figure 8: Ablation studies on GRF

most number of agents with $n = 10$, corresponding to all ally players except the goalie. Figure 7 illustrates the performance of FoX and the baseline algorithms on the GRF scenarios. FoX shows outstanding performance in all four scenarios of the experiment, outperforming the other baselines. Especially, FoX outperforms the other visitation count baselines, supporting the idea that formation-based state equivalence efficiently reduces the search space in MARL. In addition, the exploration path of FoX agents in the GRF environment can be found in Appendix F, showing how agents become aware of their formation as the learning step increases.

Ablation Study

In this section, we analyze the contributions of FoX. As $\beta_1 = 0.01$ and $\beta_2 = 0.01$ showed the best performance among all GRF scenarios, we conduct our study in the GRF $Academy\_3\_vs\_1\_with\_keeper$ scenario with a default setting of $\beta_1, \beta_2 = 0.01$ in our ablation studies.

Component Evaluation

To measure the effects of each component in the FoX framework, we remove each component from the FoX algorithm. Component evaluation graph in Figure 8 shows the effect of various components of the intrinsic rewards r^{exp} and r^{aware} for formation-aware exploration, the gradient flipping (GF), Q-function $Q^{\mathcal{F}}$ with formation information, and shared Q learning. In Figure 8, FoX-r^{exp} and FoX-r^{aware} both show a decrease in their performance, so it shows that the intrinsic rewards are crucial components for improving the per-

Event	Checkpoint reward	Score reward
Our team scores	-	+1
Enemy team scores	-	-1
Checkpoint scores	$+0.1 \cdot$ checkpoints	-

Table 2: The reward setting for GRF.

formance in FoX. Furthermore, it is noticeable that excluding $Q^{\mathcal{F}}$ and Shared Q significantly reduces the performance. Lastly, the importance of minimizing irrelevant information from individual trajectories upon latent variable extraction is shown through FoX-GF. As gradient flipping allows z to capture only the formation relevant information from τ, removing gradient flipping shows a decrease in performance.

The Effect of Intrinsic Rewards

As seen in the graph in Figure 8, formation-aware exploration proves its effectiveness in the $Academy\_3\_vs\_1\_with\_keeper$ environment since $\beta_1 = 0.01$ works better than $\beta_1 = 0$, but excessive exploration with $\beta_1 > 0.01$ can slow down the learning for excessive exploration. Similarly, guiding the agents to be formation-aware helps the learning of the environment. According to the change in reward when $\beta_2 > 0.01$, the agents are guided to be aware of the formation. Awareness of the formation formulated by the other agents mitigates the information bottleneck from partial observability, boosting the learning process.

Formation Selection

According to Figure 8, formation-based on $F^{i,max}$ and $F^{i,min}$ may hold enough information to represent the agent relationship since $Academy\_3\_vs\_1\_with\_keeper$ presents a relatively small number of agents $n = 3$. Surprisingly, F^{all} showed a decrease in performance even though it should hold a similar amount of information to formation based on $F^{i,max,min}$. Such behavior emphasizes the importance of formation selection. To see the effect of formation selection we have to consider the environment with a large number of agents such as 2s3z as represented in Figure 4.

Conclusion

Exploration space which can increase exponentially upon the number of agents makes efficient exploration in MARL very challenging. In this paper, we reduce the exploration space while capturing the condensed information of the state through formations, ensuring visitation to diverse formations with awareness. Finally, the efficiency of formation-aware exploration is demonstrated in the StarCraft II Multi-Agent Challenge and Google Research FootBall, where FoX shows state-of-the-art performance.

Acknowledgements

References

Aggarwal, K.; and Verma, H. K. 2015. Hash_RC6—Variable length Hash algorithm using RC6. In *2015 International Conference on Advances in Computer Engineering and Applications*, 450–456. IEEE.

Bellemare, M.; Srinivasan, S.; Ostrovski, G.; Schaul, T.; Saxton, D.; and Munos, R. 2016. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29.

Bellman, R. 1966. Dynamic programming. *Science*, 153(3731): 34–37.

Burda, Y.; Edwards, H.; Pathak, D.; Storkey, A.; Darrell, T.; and Efros, A. A. 2018. Large-scale study of curiosity-driven learning. *arXiv preprint arXiv:1808.04355*.

Chu, T.; Wang, J.; Codecà, L.; and Li, Z. 2019. Multi-agent deep reinforcement learning for large-scale traffic signal control. *IEEE Transactions on Intelligent Transportation Systems*, 21(3): 1086–1095.

Du, Y.; Han, L.; Fang, M.; Liu, J.; Dai, T.; and Tao, D. 2019. Liir: Learning individual intrinsic reward in multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 32.

Eysenbach, B.; Gupta, A.; Ibarz, J.; and Levine, S. 2018. Diversity is all you need: Learning skills without a reward function. *arXiv preprint arXiv:1802.06070*.

Foerster, J.; Assael, I. A.; De Freitas, N.; and Whiteson, S. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems*, 29.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Ganin, Y.; and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, 1180–1189. PMLR.

Gupta, J. K.; Egorov, M.; and Kochenderfer, M. 2017. Cooperative multi-agent control using deep reinforcement learning. In *Autonomous Agents and Multiagent Systems: AAMAS 2017 Workshops, Best Papers, São Paulo, Brazil, May 8-12, 2017, Revised Selected Papers 16*, 66–83. Springer.

Gupta, T.; Mahajan, A.; Peng, B.; Böhmer, W.; and Whiteson, S. 2021. Uneven: Universal value exploration for multi-agent reinforcement learning. In *International Conference on Machine Learning*, 3930–3941. PMLR.

Haarnoja, T.; Zhou, A.; Abbeel, P.; and Levine, S. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, 1861–1870. PMLR.

Han, S.; and Sung, Y. 2021. A max-min entropy framework for reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 25732–25745.

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2016a. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29.

Houthooft, R.; Chen, X.; Duan, Y.; Schulman, J.; De Turck, F.; and Abbeel, P. 2016b. Vime: Variational information maximizing exploration. *Advances in neural information processing systems*, 29.

Iqbal, S.; and Sha, F. 2019. Actor-attention-critic for multi-agent reinforcement learning. In *International conference on machine learning*, 2961–2970. PMLR.

Jaques, N.; Lazaridou, A.; Hughes, E.; Gulcehre, C.; Ortega, P.; Strouse, D.; Leibo, J. Z.; and De Freitas, N. 2019. Social influence as intrinsic motivation for multi-agent deep reinforcement learning. In *International conference on machine learning*, 3040–3049. PMLR.

Jeon, J.; Kim, W.; Jung, W.; and Sung, Y. 2022. Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer. In *International Conference on Machine Learning*, 10041–10052. PMLR.

Jin, C.; Krishnamurthy, A.; Simchowitz, M.; and Yu, T. 2020. Reward-free exploration for reinforcement learning. In *International Conference on Machine Learning*, 4870–4879. PMLR.

Kalashnikov, D.; Irpan, A.; Pastor, P.; Ibarz, J.; Herzog, A.; Jang, E.; Quillen, D.; Holly, E.; Kalakrishnan, M.; Vanhoucke, V.; et al. 2018. Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation. *arXiv preprint arXiv:1806.10293*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kurach, K.; Raichuk, A.; Stańczyk, P.; Zając, M.; Bachem, O.; Espeholt, L.; Riquelme, C.; Vincent, D.; Michalski, M.; Bousquet, O.; et al. 2020. Google research football: A novel reinforcement learning environment. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, 4501–4510.

Li, C.; Wang, T.; Wu, C.; Zhao, Q.; Yang, J.; and Zhang, C. 2021. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34: 3991–4002.

Liu, B.; Pu, Z.; Pan, Y.; Yi, J.; Liang, Y.; and Zhang, D. 2023. Lazy agents: a new perspective on solving sparse reward problem in multi-agent reinforcement learning. In *International Conference on Machine Learning*, 21937–21950. PMLR.

Liu, I.-J.; Jain, U.; Yeh, R. A.; and Schwing, A. 2021. Cooperative exploration for multi-agent deep reinforcement learning. In *International Conference on Machine Learning*, 6826–6836. PMLR.

Lowe, R.; Wu, Y. I.; Tamar, A.; Harb, J.; Pieter Abbeel, O.; and Mordatch, I. 2017. Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Machado, M. C.; Bellemare, M. G.; and Bowling, M. 2020. Count-based exploration with the successor representation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 5125–5133.

Mahajan, A.; Rashid, T.; Samvelyan, M.; and Whiteson, S. 2019. Maven: Multi-agent variational exploration. *Advances in neural information processing systems*, 32.

Nair, A.; McGrew, B.; Andrychowicz, M.; Zaremba, W.; and Abbeel, P. 2018. Overcoming exploration in reinforcement learning with demonstrations. In *2018 IEEE international conference on robotics and automation (ICRA)*, 6292–6299. IEEE.

Oliehoek, F. A.; Amato, C.; et al. 2016. *A concise introduction to decentralized POMDPs*, volume 1. Springer.

Osband, I.; Blundell, C.; Pritzel, A.; and Van Roy, B. 2016. Deep exploration via bootstrapped DQN. *Advances in neural information processing systems*, 29.

Ostrovski, G.; Bellemare, M. G.; Oord, A.; and Munos, R. 2017. Count-based exploration with neural density models. In *International conference on machine learning*, 2721–2730. PMLR.

Pathak, D.; Agrawal, P.; Efros, A. A.; and Darrell, T. 2017. Curiosity-driven exploration by self-supervised prediction. In *International conference on machine learning*, 2778–2787. PMLR.

Peng, B.; Rashid, T.; Schroeder de Witt, C.; Kamienny, P.-A.; Torr, P.; Böhmer, W.; and Whiteson, S. 2021. Facmac: Factored multi-agent centralised policy gradients. *Advances in Neural Information Processing Systems*, 34: 12208–12221.

Rashid, T.; Samvelyan, M.; De Witt, C. S.; Farquhar, G.; Foerster, J.; and Whiteson, S. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *The Journal of Machine Learning Research*, 21(1): 7234–7284.

Samvelyan, M.; Rashid, T.; De Witt, C. S.; Farquhar, G.; Nardelli, N.; Rudner, T. G.; Hung, C.-M.; Torr, P. H.; Foerster, J.; and Whiteson, S. 2019. The starcraft multi-agent challenge. *arXiv preprint arXiv:1902.04043*.

Son, K.; Kim, D.; Kang, W. J.; Hostallero, D. E.; and Yi, Y. 2019. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*, 5887–5896. PMLR.

Strehl, A. L.; and Littman, M. L. 2008. An analysis of model-based interval estimation for Markov decision processes. *Journal of Computer and System Sciences*, 74(8): 1309–1331.

Sukhbaatar, S.; Fergus, R.; et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems*, 29.

Sunehag, P.; Lever, G.; Gruslys, A.; Czarnecki, W. M.; Zambaldi, V.; Jaderberg, M.; Lanctot, M.; Sonnerat, N.; Leibo, J. Z.; Tuyls, K.; et al. 2017. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*.

Tan, M. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, 330–337.

Tang, H.; Houthooft, R.; Foote, D.; Stooke, A.; Xi Chen, O.; Duan, Y.; Schulman, J.; DeTurck, F.; and Abbeel, P. 2017. # exploration: A study of count-based exploration for deep reinforcement learning. *Advances in neural information processing systems*, 30.

Vinyals, O.; Babuschkin, I.; Czarnecki, W. M.; Mathieu, M.; Dudzik, A.; Chung, J.; Choi, D. H.; Powell, R.; Ewalds, T.; Georgiev, P.; et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782): 350–354.

Wang, J.; Ren, Z.; Liu, T.; Yu, Y.; and Zhang, C. 2020a. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*.

Wang, L.; Zhang, Y.; Hu, Y.; Wang, W.; Zhang, C.; Gao, Y.; Hao, J.; Lv, T.; and Fan, C. 2022. Individual reward assisted multi-agent reinforcement learning. In *International Conference on Machine Learning*, 23417–23432. PMLR.

Wang, T.; Dong, H.; Lesser, V.; and Zhang, C. 2020b. Roma: Multi-agent reinforcement learning with emergent roles. *arXiv preprint arXiv:2003.08039*.

Wang, T.; Gupta, T.; Mahajan, A.; Peng, B.; Whiteson, S.; and Zhang, C. 2020c. Rode: Learning roles to decompose multi-agent tasks. *arXiv preprint arXiv:2010.01523*.

Wang, T.; Wang, J.; Wu, Y.; and Zhang, C. 2019. Influence-based multi-agent exploration. *arXiv preprint arXiv:1910.05512*.

Wang, X.; Ke, L.; Qiao, Z.; and Chai, X. 2020d. Large-scale traffic signal control using a novel multiagent reinforcement learning. *IEEE transactions on cybernetics*, 51(1): 174–187.

Wang, Y.; Han, B.; Wang, T.; Dong, H.; and Zhang, C. 2020e. Dop: Off-policy multi-agent decomposed policy gradients. In *International conference on learning representations*.

Yang, Y.; Hao, J.; Liao, B.; Shao, K.; Chen, G.; Liu, W.; and Tang, H. 2020. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv preprint arXiv:2002.03939*.

Yu, C.; Velu, A.; Vinitsky, E.; Gao, J.; Wang, Y.; Bayen, A.; and Wu, Y. 2022. The surprising effectiveness of ppo in cooperative multi-agent games. *Advances in Neural Information Processing Systems*, 35: 24611–24624.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018a. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, 5872–5881. PMLR.

Zhang, K.; Yang, Z.; Liu, H.; Zhang, T.; and Basar, T. 2018b. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, 5872–5881. PMLR.

Zheng, L.; Chen, J.; Wang, J.; He, J.; Hu, Y.; Chen, Y.; Fan, C.; Gao, Y.; and Zhang, C. 2021. Episodic multi-agent reinforcement learning with curiosity-driven exploration. *Advances in Neural Information Processing Systems*, 34: 3757–3769.