Master's Thesis

FoX: Formation-based exploration with formation-awareness for partially observable agents in multi-agent reinforcement learning

Yonghyeon Jo

Graduate School of Artificial Intelligence

Ulsan National Institute of Science and Technology

2024

FoX: Formation-based exploration with formation-awareness for partially observable agents in multi-agent reinforcement learning

Yonghyeon Jo

Graduate School of Artificial Intelligence

Ulsan National Institute of Science and Technology

FoX: Formation-based exploration with formation-awareness for partially observable agents in multi-agent reinforcement learning

A thesis/dissertation submitted to
Ulsan National Institute of Science and Technology
in partial fulfillment of the
requirements for the degree of
Master of Science

Yonghyeon Jo

12.05.2023 of submission

Approved by

Advisor

Seungyul Han

FoX: Formation-based exploration with formation-awareness for partially observable agents in multi-agent reinforcement learning

Yonghyeon Jo

This certifies that the thesis/dissertation of Yonghyeon Jo is approved.

12.05.2023 of submission

Signature

Advisor: Seungyul Han

Signature

Jeong hwan Jeon

Signature

Gi-Soo Kim

Abstract

Recently, the surge in interest surrounding deep multi-agent reinforcement learning (MARL) can be attributed to its remarkable success in tackling various cooperative multi-agent tasks. Despite these advancements, the challenge of effective exploration persists in MARL, primarily owing to the inherent partial observability of agents and the exponential growth of the exploration space with an increasing number of agents. To address the formidable scalability issue associated with exploration, we introduce a formation-based equivalence relation on the exploration space. This novel approach aims to streamline the exploration process by narrowing down the search space to states that bear meaningful distinctions in different formations.

Building upon this foundational concept, we present the Formation-aware Exploration (FoX) framework, a pioneering approach designed to guide partially observable agents toward states within diverse formations. FoX achieves this by fostering an acute awareness of the agents' current formation based solely on their individual observations. By encouraging agents to navigate and explore the exploration space through the lens of formations, FoX seeks to provide a more nuanced and targeted exploration strategy.

In our numerical evaluations, the results unequivocally demonstrate the superior performance of the proposed FoX framework when compared to state-of-the-art MARL algorithms. These assessments were conducted on challenging tasks in both the Google Research Football (GRF) environment and the sparse StarCraft II multi-agent challenge (SMAC), showcasing the efficacy and versatility of FoX across diverse scenarios..

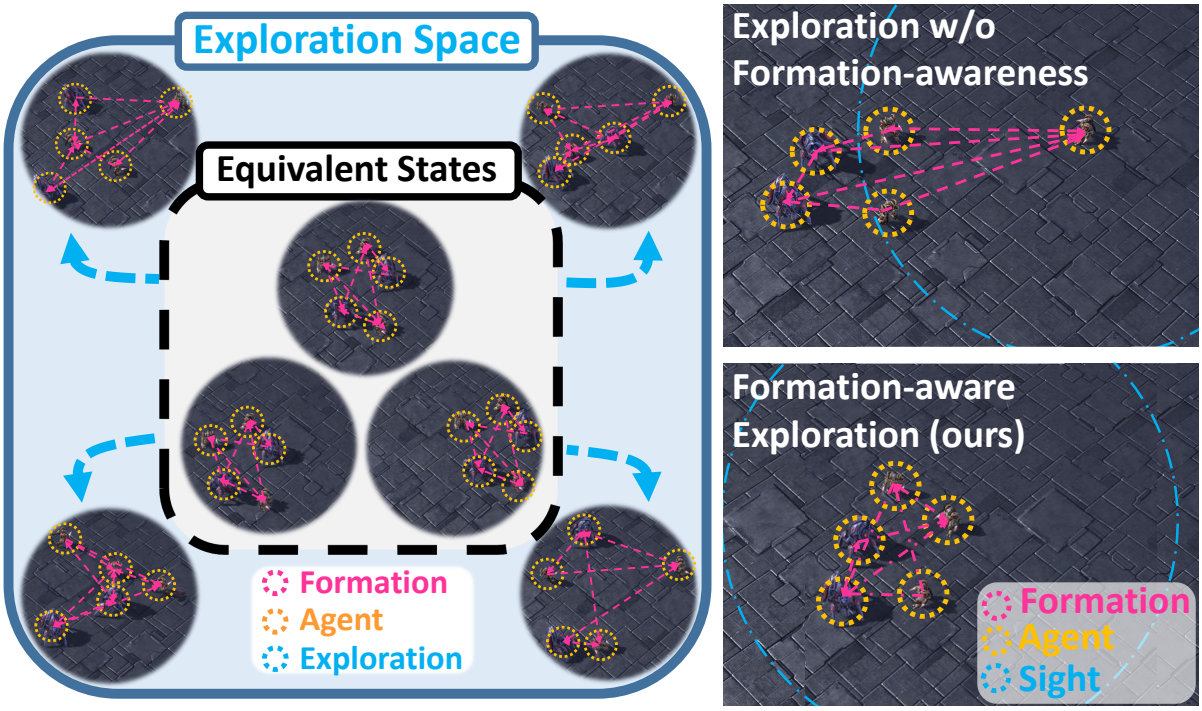
Contents

List of Figures

I Introduction

In recent years, the field of multi-agent reinforcement learning (MARL) has emerged as a powerful paradigm, demonstrating remarkable success in addressing complex real-world challenges across diverse domains. Applications such as traffic control [1, 2], games [3], and robotic controls [4] have witnessed substantial advancements owing to the application of MARL techniques. The growing popularity of MARL is underscored by the continual introduction of novel algorithms designed to enhance the collective decision-making and learning capabilities of autonomous agents [5–7].

MARL algorithms exhibit a rich diversity, broadly falling into three distinctive categories based on their approaches to learning and execution. First, fully decentralized methods [5, 8] pursue a strategy where agents are trained independently, allowing them to make decisions autonomously without explicit communication or coordination. Second, fully centralized methods [9] leverage shared information among agents to enhance training efficiency, fostering collaborative decision-making. Finally, centralized training with decentralized execution (CTDE) methods [7, 10, 11] represent a hybrid approach that provides global information during the training phase. This approach aims to address issues related to nonstationarity while preserving scalability during execution. The incorporation of global information allows agents to adapt to changing environments more effectively.



(a) (b)

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. Agents cannot be fully aware of formations shaped outside their sight range, so exploration without formation awareness may lead to an inaccurate perception of the current formation. By encouraging the agents to be aware of the current formation, FoX mitigates the challenges from partial observability.

Nevertheless, despite the impressive strides made by MARL algorithms, they encounter formidable challenges stemming from the inherent complexities of partial observability [12]. The issue of partial observability poses a substantial hurdle, particularly in dynamic environments where agents possess limited knowledge about the overall state of the system. Even in the realm of traditional reinforcement learning (RL), effective exploration is deemed crucial to prevent agents from converging prematurely to sub-optimal policies [13]. In addressing the exploration-exploitation trade-off in traditional RL, innovative methods grounded in the concept of curiosity have emerged [13–15].

Curiosity-based exploration methods, particularly those employing count-based strategies, have proven to be both simple and efficient in the context of single-agent RL problems [15, 16]. However, the transition from single-agent scenarios to MARL introduces additional layers of complexity, magnifying the challenges associated with exploration in the presence of partial observability. The exponential growth in the search space due to considerations for agent relationships necessitates novel approaches. Techniques such as leveraging the influence between agents [17] or incorporating social influence dynamics [18] have been proposed to enhance exploration capabilities in MARL. Despite these endeavors, the efficient exploration of multi-agent systems remains an ongoing challenge [19]. The intricate interplay of agents within dynamic and partially observable environments calls for continued exploration into innovative methodologies to unlock the full potential of MARL algorithms.

In practical scenarios, such as soccer matches, considering the entire information of the playing field when formulating strategies becomes impractical. Coaches often rely on strategic formations, which encapsulate the distance information and influence between players. In soccer games, for instance, the formation adopted is instrumental in achieving victory against opponents. Drawing inspiration from the intricate strategies of real soccer games, we extend the concept of formations to the realm of multi-agent environments. In this context, we define a formation based on the distinct differences between agents, recognizing the pivotal role of these formations in shaping effective strategies.

Motivated by the soccer game analogy, we introduce a novel framework known as Formation-aware exploration (FoX). The FoX framework aims to revolutionize exploration strategies by prioritizing the discovery of diverse formations rather than navigating expansive search spaces. Our contributions to the FoX framework are twofold: 1) Formation-based Equivalence Relation: We define a formation-based equivalence relation on the exploration space, enabling efficient visits to states in diverse formations as illustrated in Figure 1(a). 2) In order 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). To demonstrate the effectiveness and superiority of our proposed exploration method, we conduct comprehensive performance comparisons in various challenging multi-agent environments, including the StarCraft II Multi-Agent Challenge (SMAC) [20] and Google Research Football (GRF) [21]. Through these evaluations, we aim to showcase the potential of the FoX framework in enhancing exploration strategies in dynamic and complex multi-agent scenarios.

II Related Works

2.1 Deep Multi-Agent Reinforcement Learning

The surge in popularity of multi-agent reinforcement learning (MARL) has prompted the development of a myriad of algorithms [5, 11, 22–25]. Fully decentralized methods [5, 26] treat individual agents as independent entities during training, while the centralized training with decentralized execution (CTDE) paradigm integrates global information for more robust learning in the face of partial observability. QMIX [7] adopts a mixing network to decompose team-wide expected returns into individual Q values. Extensions of this approach, such as QTRAN [27] and QPLEX [28], further refine value decomposition principles, aligning with the Interactive Graphical Model (IGM) framework. In contrast, policy-based methods [6, 10, 29, 30] approach MARL problems through policy optimization, diversifying the landscape of techniques for addressing multi-agent learning challenges.

2.2 Exploration in State Space

In environments characterized by sparse rewards or intricate state spaces, ensuring efficient exploration becomes paramount. Strategies such as utilizing prediction errors as intrinsic rewards [31–35] have proven effective. Another avenue involves designing intrinsic rewards based on visitation count methods [13, 16, 36, 37]. While [36] suggests using a density model to approximate the number of visits to a state, [16] takes a different approach by hashing similar states to discretize the high-dimensional state space. Exploration strategies have also been pursued through information-theoretical objectives [38–41], further broadening the repertoire of techniques available for navigating complex state spaces.

2.3 Intrinsic Motivations in MARL

The challenge of partial observability poses a significant hurdle to efficient exploration in MARL, prompting a surge in studies to address this issue [42–45]. Leveraging external state information, such as the agents' locations [46], can significantly enhance training, but such assumptions may impact generality across applications. Approaches like [17, 18] tackle this challenge by exploring based on the influence between agents, while [47, 48] assigns roles to agents for task-appropriate behavior. Sub-goal targeting is explored by [49], aiming for observations with the highest sum of individual and global values. On a different note, [50] introduces a latent space based on a hierarchical level to propose a latent variable shared among the agents so that the agents may explore in an extended fashion. [19] defines curiosity based on prediction errors on individual Q-values, while [51] encourages agent diversity through local trajectory considerations. These intrinsic motivation strategies contribute to the robustness and adaptability of MARL algorithms in various scenarios.

III Preliminaries

3.1 Reinforcement Learning

Reinforcement Learning (RL) serves as a foundational paradigm in machine learning, offering a principled framework for modeling and addressing decision-making problems. Fundamentally, RL deals with the interaction between an autonomous agent and its dynamic environment, where the agent seeks to optimize a cumulative reward signal through a sequence of decisions. This interaction is encapsulated within the formal structure of a Markov Decision Process (MDP), a mathematical framework that rigorously captures the essential elements of sequential decision-making.

Formally, an MDP is defined by the tuple (S, A, P, R), where S represents the true state space of the environment, A denotes the set of permissible actions available to the agent, P characterizes the state transition probability function, and R specifies the immediate reward function. The state space S comprises all possible situations the environment can assume, while the action space A enumerates the feasible decisions that the agent can undertake. The transition probability function P delineates the likelihood of transitioning between states based on chosen actions, and the reward function R quantifies the immediate rewards associated with state-action transitions. These components collectively define the dynamics of the decision-making process within the MDP.

A cornerstone concept within RL is the Bellman Equation. The Bellman Equation is a fundamental recurrence relation that characterizes the value of a state in terms of the expected immediate reward and the value of the next state under an optimal policy. Mathematically, for a state s and action a in the Markov Decision Process (MDP), the Bellman Equation is expressed as:

$$V^*(s) = \max_{a \in A} \left\{ R(s,a) + \gamma \sum_{s' \in S} P(s'|s,a) \cdot V^*(s') \right\} \tag{1}$$

where $V^*(s)$ represents the optimal value of state s, $R(s,a)$ denotes the immediate reward obtained when taking action a in state s, γ is the discount factor that models the agent's preference for immediate rewards over delayed rewards, $P(s'|s,a)$ is the transition probability from state s to state s' given action a, The summation is taken over all possible successor states s', The maximization is performed over all possible actions a in state s.

Q-Learning is a model-free RL algorithm that focuses on estimating the optimal action-value function, denoted as $Q^*(s,a)$, representing the expected cumulative reward of taking action a in the state s and following the optimal policy thereafter. The Q-Learning update rule, based on the temporal difference (TD) learning principle, is given by:

$$Q(s,a) \leftarrow (1-\alpha) \cdot Q(s,a) + \alpha \cdot \left[R(s,a) + \gamma \cdot \max_{a'} Q(s',a') \right] \tag{2}$$

Here: $Q(s,a)$ is the current estimate of the action-value function for state s and action a, α is the learning rate that determines the weight given to new information, $R(s,a)$ is the immediate reward obtained when taking action a in state s, γ is the discount factor, s' is the next state after taking action a.

3.2 Decentralized POMDP

A fully cooperative multi-agent task can be seen as a decentralized partially observable Markov decision process (Dec-POMDP) [52], represented as a tuple $G = \langle \mathscr{S}, \mathscr{A}, \mathscr{P}, r, Z, O, \mathscr{O}, I, n, \gamma \rangle$. $\mathscr{S}$ 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. $\mathscr{A}$ is the set of actions. At each time step t, each agent $i \in I$ receives d-dimensional observation vector $o_t^i \in \mathscr{O}^i$ 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.

3.3 Centralized Training Decentralized Execution

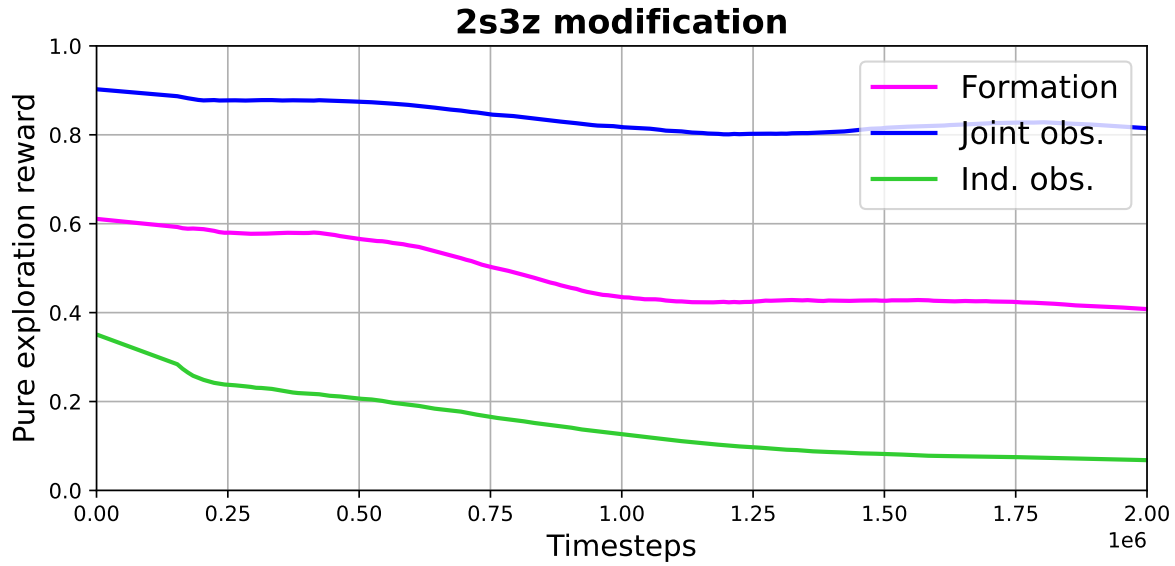
The CTDE methods provide information of the full information of agents only during training to mitigate the challenges from partial observability while maintaining decentralized execution. Among the CTDE methods are the value decomposition methods where the global value function $Q^{tot}(\boldsymbol{\tau}, \mathbf{a})$ is decomposed into individual values. A popular value decomposition method is QMIX [7], which decomposes the global value function through a mixing network to individual agent utility functions Q^i as follows:

$$\underset{\mathbf{a}}{\operatorname{argmax}}\, Q^{tot}(\boldsymbol{\tau}, \mathbf{a}) = \prod_{i=0}^{n-1} \underset{a^i}{\operatorname{argmax}}\, Q^i(\tau^i, a^i), \tag{3}$$

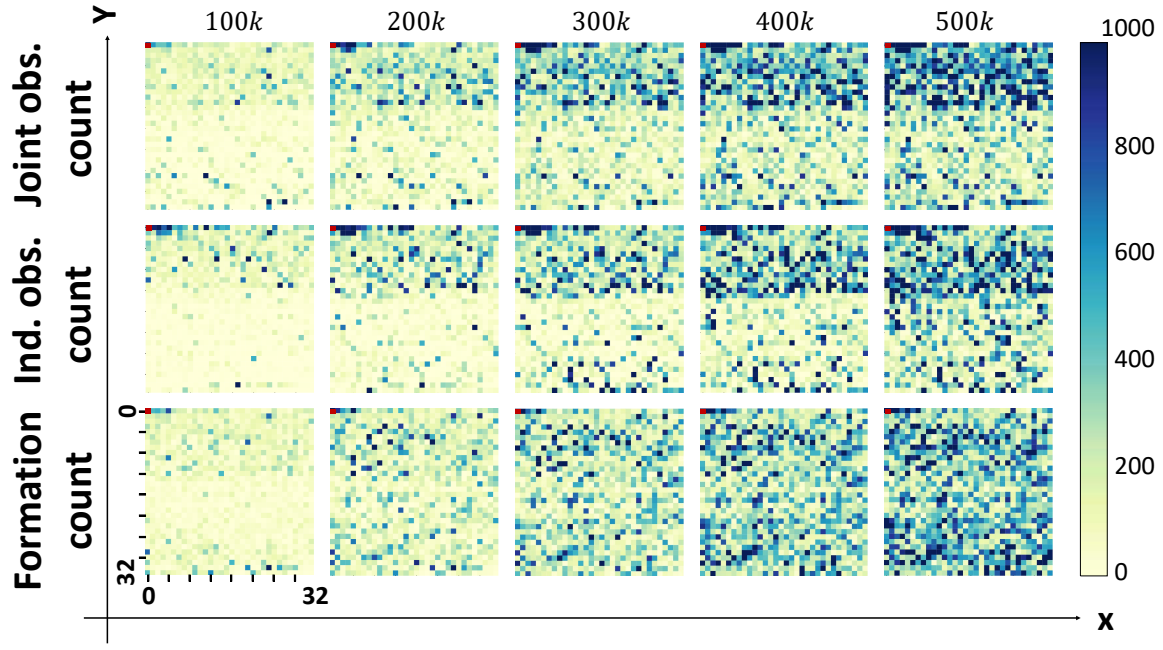
where $\boldsymbol{\tau} = (\tau_0, \cdots, \tau_{n-1})$ is the joint trajectory. Furthermore, the individual-global-max (IGM) condition must be satisfied to maintain consistency between local and global greedy actions [27]. The consistency between the global and local greedy actions allows greedy local actions to result in optimal global actions.

IV FoX: Formation-aware Exploration

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



(a)



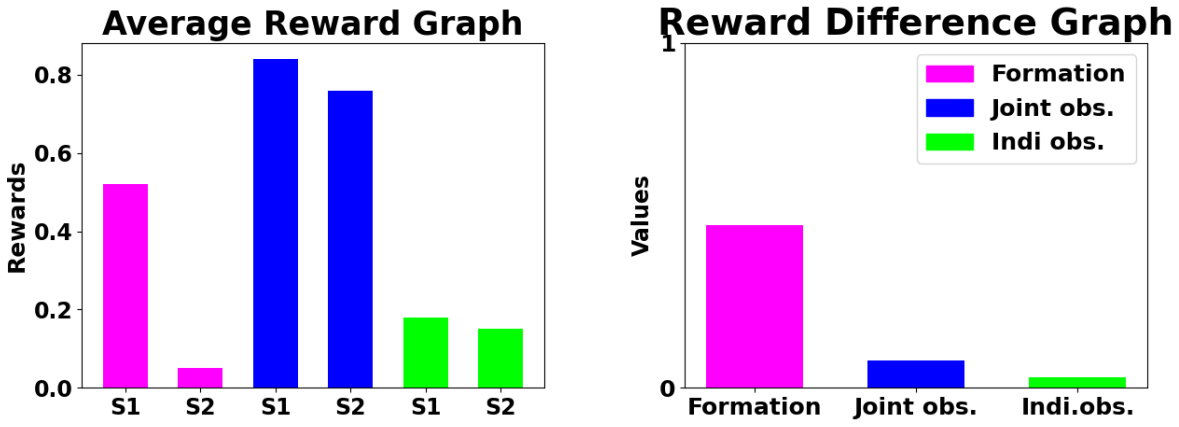
(b)

Figure 2: (a) Pure exploration reward based on visitation count graph. (b) 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.

4.1 Motivation of Formation-aware Exploration

In the broader context of multi-agent reinforcement learning (MARL), it is common to assume that agents possess information about the true global state of the environment during centralized training. However, such an approach may compromise the generality. To overcome this constraint, we adopt a more inclusive perspective by considering a scenario in which agents explore the observation product space. In this context, the exploration space, denoted as $\mathscr{S}^e$, is introduced as the observation product space, defined as $\mathscr{S}^e := \prod_{i \in I} \mathscr{O}^i$. Here, each element of $\mathscr{S}^e$ encapsulates joint observations from all agents, forming a comprehensive representation of the exploration space. Consequently, agents are encouraged to explore diverse exploration states, characterized as joint observations $s_e := (o_0, \cdots, o_{n-1}) \in \mathscr{S}^e$. This broader exploration perspective aims to facilitate agents in experiencing a rich array of interactions and dependencies between themselves, enhancing their ability to adapt to various dynamic and complex multi-agent scenarios.

However, navigating all possible exploration states in the observation product space ($\mathscr{S}^e$) presents a formidable challenge, especially as the dimensionality of the observation space and the number of agents increases. The exponential growth of $\mathscr{S}^e$ exacerbates this challenge, a phenomenon commonly referred to as the curse of dimensionality [53]. To address this issue, inspired by the strategic considerations in real soccer games, this paper introduces the concept of a formation, denoted as $\mathscr{F}$. Formations are defined based on the differences in observations between agents, providing a structured and meaningful representation of the exploration space. The primary objective is to explore a diverse array of formations rather than exhaustively sampling the entire $\mathscr{S}^e$, thereby mitigating the complexity associated with the expansive search space. By strategically focusing exploration efforts on distinct formations, the proposed approach aims to enhance the efficiency of learning in multi-agent systems, drawing inspiration from real-world scenarios where strategic formations play a pivotal role in achieving success.



(a) (b)

Figure 3: (a) Average reward graph of exploration method based on visitation counts of the three components joint observations, individual observations, and formations. (b) Graph of average reward difference between two sets S_1 and S_2 for each component.

To demonstrate the importance of formation-based exploration, we conducted pure-exploration experiments within the modified 2s3z environment of StarCraft II Multi-Agent Challenge (SMAC) [20]. In this experimental setup, agents are exclusively rewarded for count-based exploration, given by $r_t \propto \frac{1}{\sqrt{N(s)}}$, emphasizing the exploration of rarely visited components s. Here, we consider three distinct types of components s for comparative analysis: 1) joint observations $(o^0, \cdots, o^{n-1})$, 2) individual observations o^i, and 3) proposed formations $\mathscr{F}$. The outcomes of pure exploration are visualized in Figure 2, with Figure 2(a) displaying the average pure exploration rewards and Figure 2(b) illustrating the heatmap of visitation frequency for diverse formations in the exploration space.

For joint observations, the average reward does not exhibit a significant decrease, implying that most joint observations are perceived as novel. However, this limited decrease results in exploration being confined to a specific area, as depicted in Figure 2(b). In the heatmap illustrated in Figure 2(b), points farther from the origin indicate larger differences in formations. In the case of individual observations, the visitation reward diminishes rapidly, extending exploration to areas farther from the origin than the joint observation case. Despite this, individual observations still fall short of thoroughly exploring the diverse regions as depicted in the heatmap. Conversely, in the case of proposed formations, agents successfully navigate and visit a more extensive range of areas within the heatmap, with the pure exploration reward appropriately decreasing, as shown in Figure 2(a). This pure exploration experiment illustrates that the proposed formation-based exploration method facilitates superior exploration, enabling agents to visit a more diverse array of formations in the exploration space.

Additionally, we provide a comprehensive understanding of the proposed exploration strategy according to Figure 3, which demonstrates the intrinsic reward tendencies (r_t). Figure 3(a) depicts the total average of r_t (corresponding to Figure 2(a)), while Figure 3(b) focuses on the average r_t of two distinct state sets: $S_1 = \{s_t \in \mathscr{S} | 1 \leq N(\mathscr{F}(s_t)) \leq 100\}$, consisting of states that rarely visit the same formation, and $S_2 = \{s_t \in \mathscr{S} | N(\mathscr{F}(s_t)) > 100\}$, where $N(\mathscr{F}(s_t)$ is the number of visits to the formation at state s_t, to encompass the states that are frequently visiting the same formation at $t = 500k$.

In Figure 3(a), the reward difference between the two sets S_1 and S_2 is evident. For the cases of joint observations and individual observations, the average rewards exhibit either a slow decay or a rapid decline, resulting in relatively small reward differences between S_1 and S_2 as illustrated in Figure 3(a) and Figure 3(b). Conversely, in the case of our proposed formation-based exploration, the reward appropriately decays for S_2 only. Consequently, the average intrinsic reward of S_1 is substantially higher than that of S_2, leading to a notable reward difference compared to the other two cases. This observation implies that the agent achieves significantly higher rewards when exploring rarely visited formations (S_1), highlighting the efficacy of the proposed exploration strategy in promoting more effective exploration.

4.2 Formation Arrangement

To address the challenge posed by the high dimensionality of observation differences $o^i - o^j$, we propose a strategy to define a formation based on these differences, as elaborated in previous sections. However, the dimension of the observation difference $o^i - o^j$ remains prohibitively large, posing difficulties for

effective exploration. Consequently, we initiate a dimensionality reduction process, transforming the observation difference $o_i - o_j$ into a 2-dimensional vector $D^{ij} := (||o^i - o^j||_2,\ c(o^i - o^j)) \in \mathbb{R} \times \mathbb{Z}$.

Here, $||\cdot||_2$ denotes a 2-norm operator, and $c : \mathbb{R}^d \to \mathbb{Z}$ is a mapping function that converts the angle of inputs into integers based on hash coding [54]. Specifically, we employ SimHash [16] for hash coding. SimHash utilizes a random vector v and computes the hash code h by applying the sign function $\text{sign}(\cdot)$ defined as $\text{sign}(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases}$. The generated hash code $h = \text{sign}(v \cdot \frac{o^i - o^j}{||o^i - o^j||_2})$ is then converted into an integer number $c(o_i - o_j) = h_{\text{int}} \in \{0, \cdots, 2^m - 1\}$, discretizing the angle between the observation difference $o^i - o^j$ into 2^m categories. We opt for $m = 9$ in our experiments, as this choice provides a sufficiently expressive representation for the angle of the observation difference, offering 512 categories.

In the context of visitation count, we introduce the round function, denoted as $\text{round}(\cdot)$ to refine the discretization process more accurately than SimHash. The $\text{round}(\cdot)$ function, implemented in Python, rounds the input value to the l-th decimal place using the following formulation:

$$\text{round}(x) = \begin{cases} \text{floor}(x \cdot 10^l) \cdot 10^{-l}, & \text{if } x \cdot 10^l - \text{floor}(x \cdot 10^l) < 0.5 \\ (\text{floor}(x \cdot 10^l) + 1) \cdot 10^{-l}, & \text{otherwise} \end{cases}.$$

For a balanced treatment of the input data, an overly coarse discretization that would occur with excessively small values of l must be prevented. On the contrary, using a very high value of l may distinguish all inputs differently. Therefore, the selection of an appropriate l value is crucial in the context of visitation count methods [55]. In our experimental setup, we find that $l = 1$ or 3 yields effective results, particularly in scenarios involving SMAC and GRF environments.

Following the $\text{round}(\cdot)$ discretization, we proceed to calculate the visitation count $N(s)$ based on the output of this function. The variable s can represent various components, such as joint observations $(o^0, \cdots, o^{n-1})$, individual observations o^j, and our proposed formation $\mathscr{F}$. Subsequently, leveraging the visitation count information, we formulate the pure exploration reward $r^{exp} \propto \frac{1}{\sqrt{N(s)}}$. This formulation incorporates the inverse square root of the visitation count $N(s)$ for a particular state s, effectively incentivizing the exploration of rarely visited components. This comprehensive strategy aligns with the broader exploration objectives within the multi-agent reinforcement learning framework, promoting diverse and informative agent interactions across the environment.

In light of the reduced dimensionality achieved through the transformation of D^{ij}, encapsulating both the distance and angle information of $o^i - o^j$, we proceed to define the formation $\mathscr{F}_{F^0,\cdots,F^{n-1}}$ for an exploration state $s^e = (o^0,\ \cdots,\ o^{n-1}) \in \mathscr{S}^e$. This formation is characterized by the composition:

$$\mathscr{F}_{F^0,\cdots,F^{n-1}}(s_e) = (\mathscr{F}^0_{F^0}, \cdots, \mathscr{F}^{n-1}_{F^{n-1}}), \tag{4}$$

Here, n denotes the number of agents, $F^i = \{j_0, \cdots, j_{k-1}\} \subset I$ represents an arbitrary ordered agent index set, and $\mathscr{F}^i_{F^i}$ signifies the individual formation as perceived by the i-th agent, defined as:

$$\mathscr{F}^i_{F^i} = (D^{ij_0}, \cdots, D^{ij_{k-1}}). \tag{5}$$

Through this definition, the formation of the exploration state s^e comprehensively encompasses all the difference information pertaining to the pairs of agents (i, j), where $i \in I$ and $j \in F^i$. Going forward, we adopt the notation $\mathscr{F}$ to refer to the formation, enhancing clarity in our discussions. This strategic reduction of dimensionality not only facilitates a more efficient representation of the exploration state space but also enables a streamlined exploration process by capturing the essential agent interaction dynamics in a compact and informative manner.

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

$$\sim_{\mathscr{F}} := \{(s_1, s_2) \in \mathscr{S}^e \times \mathscr{S}^e \mid \mathscr{F}(s_1) = \mathscr{F}(s_2)\}, \tag{6}$$

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

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

Proof 1 *To prove Lemma 1, we have to show that $\sim_{\mathscr{F}}$ satisfies 3 properties of equivalence relation: $s_1 \sim_{\mathscr{F}} s_1$ (reflexity) from the definition, if $s_1 \sim_{\mathscr{F}} s_2$ then $s_1 \sim_{\mathscr{F}} s_2$ (symmetry), and if $s_1 \sim_{\mathscr{F}} s_2$ and $s_2 \sim_{\mathscr{F}} s_3$, then $s_1 \sim_{\mathscr{F}} s_3$ (transitivity), $\forall s_1,\ s_2,\ s_3 \in \mathscr{S}^e$.*

1) Reflexity: For all $s_1 \in \mathscr{S}^e$, $\mathscr{F}(s_1) = \mathscr{F}(s_1)$, showing $s_1 \sim_{\mathscr{F}} s_1$

2) Symmetry: For all $s_1,\ s_2 \in \mathscr{S}^e$ such that $s_1 \sim_{\mathscr{F}} s_2$, $\mathscr{F}(s_1) = \mathscr{F}(s_2)$ implying $\mathscr{F}(s_2) = \mathscr{F}(s_1)$. Therefore, $s_2 \sim_{\mathscr{F}} s_1$

3) Transitivity: Suppose $\forall s_1,\ s_2,\ s_3 \in \mathscr{S}^e$ such that $s_1 \sim_{\mathscr{F}} s_2$ and $s_2 \sim_{\mathscr{F}} s_3$. As $\mathscr{F}(s_1) = \mathscr{F}(s_2)$ and $\mathscr{F}(s_2) = \mathscr{F}(s_3)$ from the definition of $\sim_{\mathscr{F}}$, we have $\mathscr{F}(s_1) = \mathscr{F}(s_3)$, indicating $s_1 \sim_{\mathscr{F}} s_3$.
Thus, $\sim_{\mathscr{F}}$ is an equivalence relation on $\mathscr{S}^e$.

It is essential to highlight the dimensional characteristics of the formation, denoted as $\mathscr{F}$, and its implications for the exploration space. The formation's dimensionality is expressed as nk, where n signifies the number of agents and k is contingent on the chosen agent index set F^i. In practice, the selection of F^i allows for adaptability, catering to specific scenarios and agent interactions. Notably, the dimensionality of the formation, nk, is often considerably smaller than that of the exploration space, which is nd, with d representing the observation dimension. This distinction in dimensionality is pivotal for computational efficiency, as the observation dimension is frequently larger than the chosen formation dimension. Consequently, this strategic reduction in dimensionality aids in streamlining the search space, enhancing the exploration process by focusing on essential agent interaction dynamics.

Furthermore, the concept of equivalence relation $\sim_{\mathscr{F}}$ plays a crucial role in the exploration strategy. By adhering to this relation, agents navigate through exploration states in diverse formations, avoiding visits to equivalent states that share identical formations. This avoidance of redundant visits not only aligns with the equivalence relation but also contributes to a substantial reduction in the effective search space. The practical manifestation of this reduction is evident in Figure 2, showcasing the enhanced exploration facilitated by the formation-based equivalence relation. This insightful visualization underscores the efficiency and efficacy of the proposed exploration scheme, where agents systematically

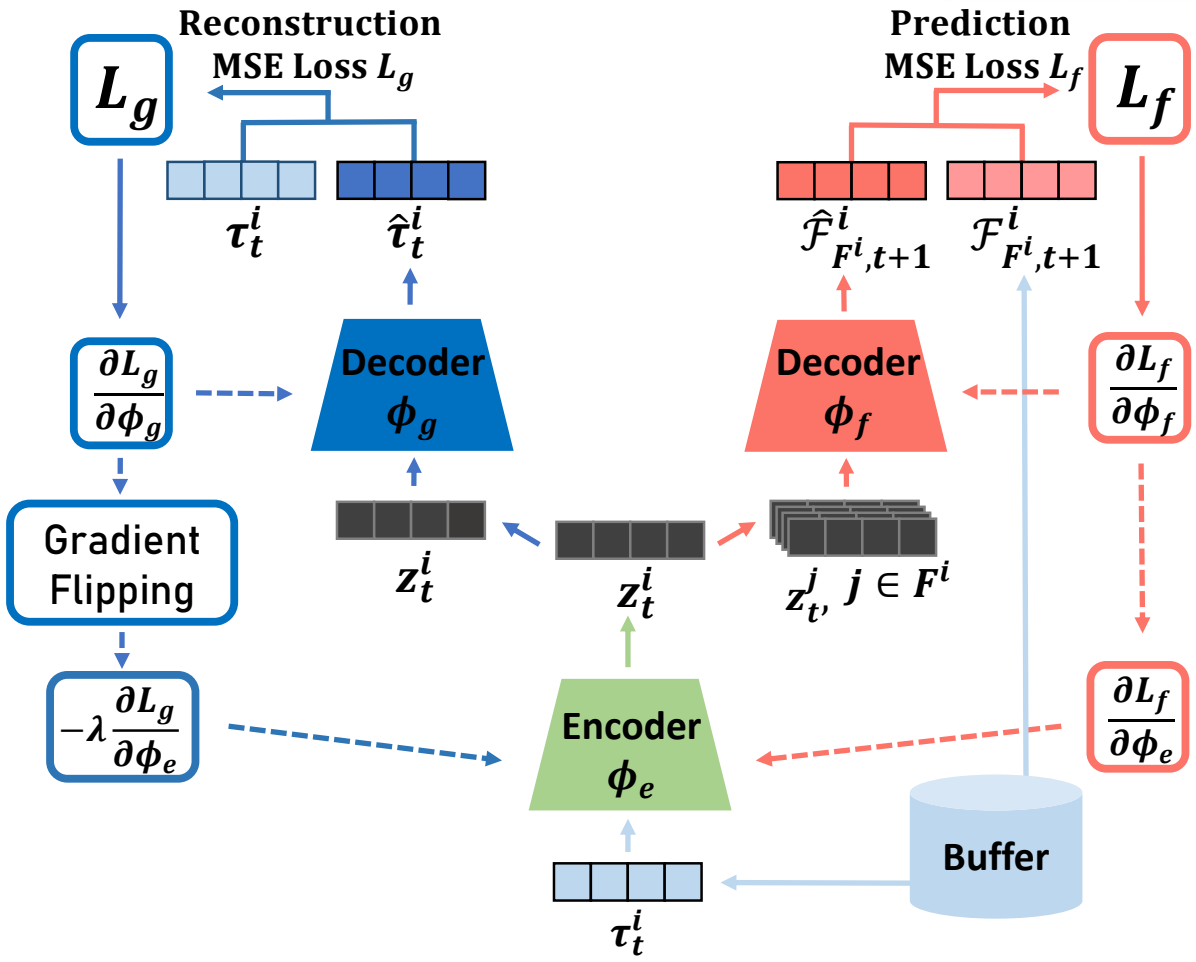


Figure 4: $\mathscr{F}$-Net Architecture

explore diverse formations, leading to a more comprehensive understanding of the exploration space and, consequently, improved exploration outcomes.

4.3 Formation-aware Exploration Objective

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

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

where $F^{i+} = F^i \cup \{i\}$, a^F indicates the set $\{a^i | i \in F\}$, $\mathscr{H}(X) = -p(X)\log p(X)$ is the information entropy of a random variable X, and $\mathscr{I}(X;Y) = \mathscr{H}(X) - \mathscr{H}(X|Y)$ is the mutual information between random variables X and Y.

In pursuit of the exploration objective, component (a) involves maximizing the entropy of formations, a strategy that encourages agents to explore states in a myriad of formations, as vividly depicted in Figure 1(a). The overarching goal is to foster a rich and diverse exploration experience. Achieving the maximization of (a) entails leveraging visitation count-based exploration methodologies [13]. Specifically, we introduce a formation-based count exploration reward, denoted as $r_t^{exp} = \frac{1}{\sqrt{N(\mathscr{F}_t)}}$. This

reward mechanism offers higher rewards to agents who venture into states characterized by rarely visited formations, thereby promoting a balanced and thorough exploration of the exploration space.

To effectively quantify the number of occurrences of a formation, $N(\mathscr{F}_t)$, we employ the $round(\cdot)$ method. This method serves the purpose of discretizing the formation $\mathscr{F}_t$, subsequently facilitating the counting of visitations within discretized bins. The discretization process enhances the precision of visitation count calculations, allowing for a nuanced understanding of the exploration landscape. This meticulous approach to counting formations aligns with the broader objective of maximizing the entropy of formations, fostering an exploration strategy that not only explores diverse formations but also captures the richness and variety within each formation, contributing to a comprehensive exploration endeavor.

Navigating the challenge of partially observable environments in MARL, achieving the exploration goal of maximizing the entropy of formations (a) becomes intricate when agents lack complete information about the formations. In light of this, we introduce an additional formation-aware objective (b) to address this challenge. Objective (b) centers around enhancing the mutual information between the latent variable z_t^j for all $j \in F^{i+}$ and the subsequent formation $\mathscr{F}^i_{F^i,t+1}$. In this context, given the trajectory τ_t^j, the trajectories $\tau_t^{F^{i+}} \setminus \{j\}$, and the latent variables $z_t^{F^{i+}\setminus\{j\}}$ of other agents, maximizing (b) guides the i-th agent not only to be cognizant of its own next formation $\mathscr{F}^i_{F^i,t+1}$ but also to understand the formations $\mathscr{F}^j_{F^j,t+1}$ for all agents j such that $i \in F^j$. This strategic approach, illustrated in Figure 1(b), empowers agents to maximize both (a) and (b), encouraging them to explore diverse formations while gaining a nuanced understanding of their own and others' formations. By jointly optimizing these objectives, agents contribute to a more comprehensive and informed exploration strategy, aligning with the overarching goal of navigating diverse formations in partially observable environments.

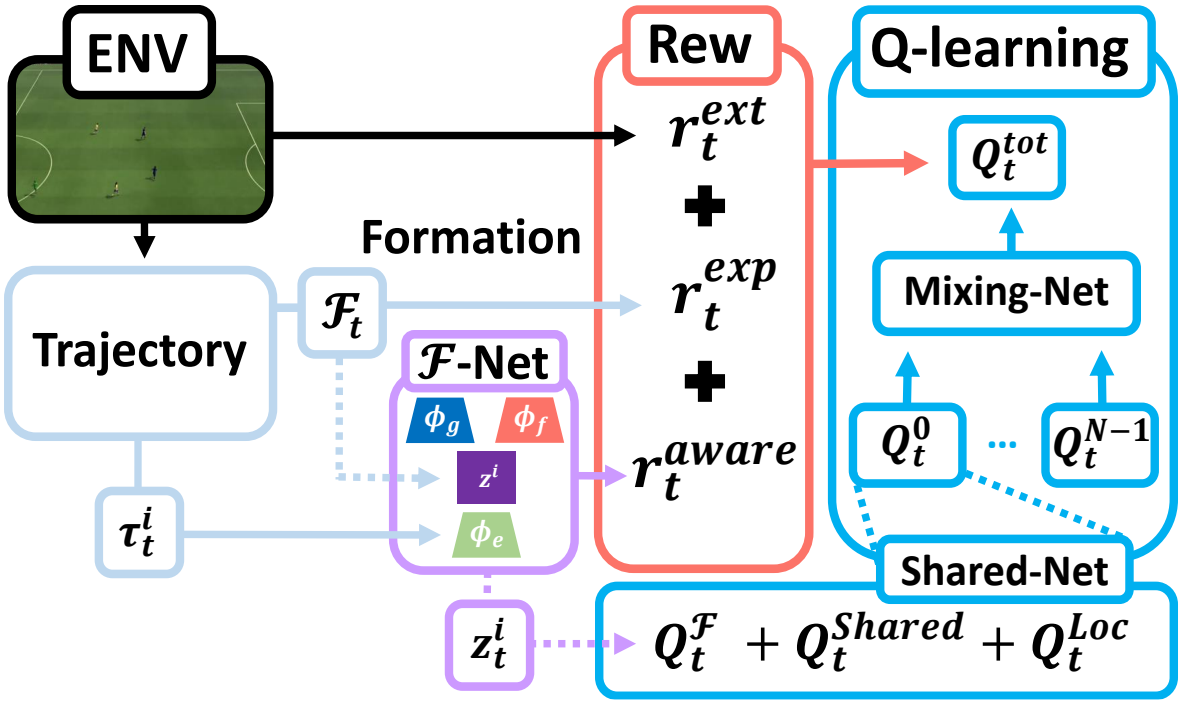


Figure 5: Schema of FoX framework

In order to maximize (b), we derive an evidence lower bound for the mutual information term based on the variational inference approaches [51, 56], 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*

$$\begin{aligned}&\mathscr{I}(\mathscr{F}^i_{F^i,t+1};z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\geq\\&\mathbb{E}_{\mathscr{F}^i_{F^i,t+1},z^j_t\sim q_{\phi_e}}\left[\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\right],\end{aligned}\tag{7}$$

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 2

$$\begin{aligned}&\mathscr{I}(\mathscr{F}^i_{F^i,t+1};z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\\&=\sum_{\mathscr{F}^i_{F^i,t+1}}\sum_{z^j_t}p(\mathscr{F}^i_{F^i,t+1},z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\left\{\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\right\}\\&=\sum_{\mathscr{F}^i_{F^i,t+1}}\sum_{z^j_t}p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)q_{\phi_e}(z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\{\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\\&\qquad-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\}\\&=\sum_{\mathscr{F}^i_{F^i,t+1}}\sum_{z^j_t}p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)q_{\phi_e}(z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\{\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\\&\qquad-\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)+\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\}\\&=\sum_{z^j_t}q_{\phi_e}(z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\cdot\underbrace{\sum_{\mathscr{F}^i_{F^i,t+1}}p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\log\frac{p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)}{q(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)}}_{D_{KL}(p(\cdot|\tau^{F^{i+}}_t,z^{F^{i+}}_t)||q(\cdot|\tau^{F^{i+}}_t,z^{F^{i+}}_t))\geq 0}\\&+\sum_{\mathscr{F}^i_{F^i,t+1}}\sum_{z^j_t}p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)q_{\phi_e}(z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\{\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\\&\qquad-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\Big\}\\&\geq\sum_{\mathscr{F}^i_{F^i,t+1}}\sum_{z^j_t}p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)q_{\phi_e}(z^j_t|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\Big\{\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\\&\qquad-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)\Big\}\\&=\mathbb{E}_{\mathscr{F}^i_{F^i,t+1},z^j_t\sim q_{\phi_e}}\left[\log q_{\phi_f}(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}}_t)-\log p(\mathscr{F}^i_{F^i,t+1}|\tau^{F^{i+}}_t,z^{F^{i+}\setminus\{j\}}_t)\right],\end{aligned}\tag{8}$$

which concludes the proof.

Algorithm 1 FoX framework

Initialize ϕ_e, ϕ_f, ϕ_g, θ, θ^-

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

In the proof of Lemma 2, z_t^j is averaged over the distribution $q_{\phi_e}(\cdot|\tau_t^{F^{i+}}, z_t^{F^{i+}\setminus\{j\}})$, but it makes difficult to utilize the latent variable z_t^j in the decentralized execution setup since j-th agent cannot exploit the other agent's trajectory information. Thus, For practical implementation, we opt for the approximate distribution $q_{\phi_e}(\cdot|\tau_t^j)$ over $q_{\phi_e}(\cdot|\tau_t^{F^{i+}}, z_t^{F^{i+}\setminus\{j\}})$ leading us to formulate an encoder-decoder structure resembling a variational auto-encoder (VAE) [56]. The encoder, denoted as $E_{\phi_e}(\tau_t^i)$, produces the mean and standard deviation of the latent variable z_t^i for sampling purposes. The decoder, represented by $D_{\phi_f}(z_t^{F^i+})$, generates the prediction for the next formation $\hat{\mathscr{F}}^i_{F^i,t+1}$. However, the trajectory τ_t^i may contain information unrelated to the formation prediction, which could be transmitted to the latent variable z_t^i [14]. To mitigate this, we introduce gradient flipping (GF) technique inspired by domain adaptation [57]. To implement GF, we incorporate an additional trajectory decoder $D_{\phi_g}(z_t^i)$ that reconstructs the trajectory $\hat{\tau}_t^i$. We update the encoder parameter ϕ_e to impede trajectory reconstruction, while the decoder strives to reconstruct the trajectory. This approach ensures that irrelevant information from the trajectory τ_t^i is avoided in delivering to the latent variable z_t^i, aligning with the principles outlined in [57].

In summary, we can construct $\mathscr{F}$-Net architecture as shown in Figure 4. We define a formation prediction loss $L_f(\phi_e,\phi_f) = \frac{1}{n}\sum_{i=0}^{n-1} MSE(\mathscr{F}^i_{F^i,t+1},\ \hat{\mathscr{F}}^i_{F^i,t+1})$, a trajectory reconstruction loss $L_g(\phi_e,\phi_g) = \frac{1}{n}\sum_{i=0}^{n-1} MSE(\tau_t^i,\ \hat{\tau}_t^i)$, and Kullback-Leibler divergence loss $L_{KL}(\phi_e) = D_{KL}(q_{\phi_e}(\cdot|\tau_t^i)||\mathscr{N}(0,\mathbf{I}))$ as in [56, 57], where $MSE(x,y) = \mathbb{E}[(x-y)^2]$ is the mean square error loss and $\mathscr{N}(0,\mathbf{I})$ is the multi-variate standard normal distribution with identity matrix $\mathbf{I}$. Based on the loss functions, we can update the encoder-decoder parameters $\phi_e,\ \phi_f,\ \phi_g$ as follows:

$$\begin{aligned}\phi_e &\leftarrow (1-\alpha)\phi_e + \alpha\left(\frac{\partial L_f}{\partial \phi_e} + \frac{\partial L_{KL}}{\partial \phi_e} - \lambda_{GF}\frac{\partial L_g}{\partial \phi_e}\right)\\ \phi_f &\leftarrow (1-\alpha)\phi_f + \alpha\frac{\partial L_f}{\partial \phi_f},\ \phi_g \leftarrow (1-\alpha)\phi_g + \alpha\frac{\partial L_g}{\partial \phi_g},\end{aligned} \tag{9}$$

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

$$r_t^{aware} = \frac{1}{n}\sum_{i=0}^{n-1}\Bigg(\log q_{\phi_f}(\hat{\mathscr{F}}^i_{t+1}|\tau_t^{F^{i+}}, z_t^{F^{i+}}) - \frac{1}{k+1}\sum_{j\in F^{i+}}\log p(\hat{\mathscr{F}}^i_{t+1}|\tau_t^{F^{i+}}, z_t^{F^{i+}\setminus\{j\}})\Bigg), \tag{10}$$

Since q_{ϕ_f} is an arbitrary posterior distribution in Lemma 2, for implementation of r^{aware}, we set the posterior distribution $\log q_{\phi_f}$ in the evidence lower bound term as the prediction error of next formation, i.e., $\log q_{\phi_f}(\hat{\mathscr{F}}^i_{t+1}|\tau_t^{F^{i+}}, z_t^{F^{i+}}) \approx -MSE(\mathscr{F}^i_{F^i,t+1}, \hat{\mathscr{F}}^i_{F^i,t+1})$. Also, we approximate the latter term $\log p$ using the prediction error as $\log p(\hat{\mathscr{F}}^i_{t+1}|\tau_t^{F^{i+}}, z_t^{F^{i+}\setminus\{j\}}) \approx \log \mathbb{E}_{z_j\sim q_e}[q_{\phi_f}(\hat{\mathscr{F}}^i_{t+1}|\tau_t^{F^{i+}}, z_t^{F^{i+}\setminus\{j\}})] \approx \log \mathbb{E}_{z_j\sim q_e}[\exp(-MSE(\mathscr{F}^i_{F^i,t+1}, \hat{\mathscr{F}}^i_{F^i,t+1}))]$. However, in this paper, we just use more simple approximation $\log p \approx \mathbb{E}_{z^j\sim\mathscr{N}(0,\mathbf{I})}[-MSE(\mathscr{F}^i_{F^i,t+1}, \hat{\mathscr{F}}^i_{F^i,t+1})]$ for practical implementation under the assumption that q_{ϕ_e} is close enough to $\mathscr{N}(0,\mathbf{I})$ by reducing L_{KL}.

With the extrinsic reward r_t^{ext} given from the environment and intrinsic rewards r_t^{exp} and r_t^{aware} for formation-aware exploration, we can define the total reward $r_t^{tot} = r_t^{ext} + \beta_1 r_t^{exp} + \beta_2 r_t^{aware}$, where β_1 and β_2 are hyperparamers to control the effect of our intrinsic rewards r_t^{exp} and r_t^{aware}, respectively. Then, an overall temporal difference loss to update the total Q-function Q_θ^{tot} in (3) parameterized by θ is defined as

$$L_{TD}(\theta) = (r_t^{tot} + \gamma \max_{a'} Q_{\theta^-}^{tot}(s_{t+1}, a') - Q_\theta^{tot}(s_t, a_t))^2 \tag{11}$$

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.

4.4 Selection of Index Set F^i

In the context of arranging formations, the choice of agent index sets denoted as F^i, introduces variability in the formation configurations. For instance, one could consider $F^{i,max} := \{\text{argmax}_{j\in I\setminus\{i\}} d(i,j)\}$, which assembles a formation among agents with the most substantial differences from the focal agent i. Similarly, $F^{i,min} = \{\text{argmin}_{j\in I\setminus\{i\}} d(i,j)\}$ forms a configuration involving agents with the smallest differences. Exploring the spectrum, $F^{i,max,min} = F^{i,max} \cup F^{i,min}$ encompasses both extremes, and $F^{all} = \{i | i \in I\setminus\{i\}\}$ includes all agents except the focal one. The impact of these different formations on exploration behaviors is examined in our experiments, as illustrated in Figure 6. To delve deeper into the influence of F^i selection, we conduct an ablation study. Surprisingly, the results reveal that the combined set $F^{i,max,min}$ outperforms the others, showcasing its efficacy in promoting exploration.

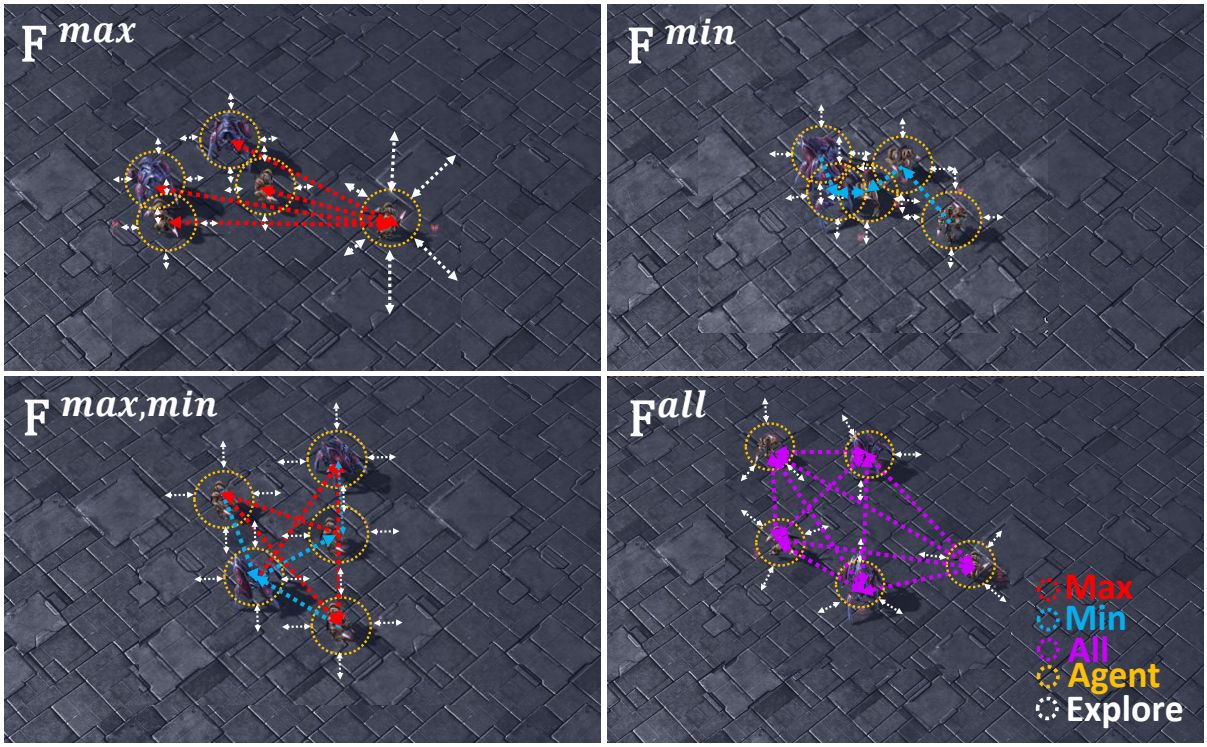


Figure 6: Formations with various agent index set F. While $F^{i,max}$, $F^{i,min}$ focuses on particular agent relationships, $F^{i,all}$ and $F^{i,max,min}$ considers extensive agent relationships.

4.5 Formation-based Shared Network

Based on the shared network structure for individual Q-function proposed in [51], we aim to exploit the latent variable z_t^i. Thus, individual Q-function Q^i in (3) can be decomposed as three Q functions as follows:

$$Q^i(\tau_t^i, z_t^i, \cdot) = Q^{Shared}(\tau_t^i, \cdot) + Q^{Loc,i}(\tau_t^i, \cdot) + Q^{\mathcal{F}}(z_t^i, \cdot), \tag{12}$$

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 [51], we also consider $l1$ regularization for $Q^{Loc,i}$ with the regularization coefficient $\lambda_{reg} = 0.1$.

V Experiments

Our implementations are based on the open-source code provided by [20] and are developed using the PyTorch framework. Leveraging the computational power of an NVIDIA RTX 3090 GPU and an Intel Xeon Gold 6348 CPU, our experiments were conducted on an Ubuntu 20.04 system. The training process for the SMAC scenarios requires less than a single day to complete. On the other hand, training on the more complex scenarios from the GRF environment concludes in less than 36 hours for every 5,000,000 time steps which is 20∼30% more training time than QMIX. Even considering this, we believe that FoX has sufficient benefits compared to QMIX These specific details regarding the hardware and time requirements underscore the feasibility and scalability of our approach across different multi-agent reinforcement learning scenarios.

5.1 StarCraft Multi-Agent Challenge

The StarCraft Multi-Agent Challenge (SMAC) provides a dynamic and challenging environment where two teams, each comprised of StarCraft units, clash in combat with the overarching goal of maximizing their win rates. This unique and complex scenario serves as a prominent benchmark for assessing the efficacy of multi-agent systems, particularly in the context of cooperation among units. The adversarial nature of the environment is accentuated by the fact that the enemy team is composed of units controlled by the built-in game AI, employing predefined non-learned heuristics. On the other side, the army team consists of individual units, each treated as a distinct agent, and undergoes training using state-of-the-art reinforcement learning algorithms. This distinctive setup not only underscores the intricacies of multi-agent interactions but also emphasizes the need for intelligent decision-making strategies among team members to navigate the challenges presented by the opposing forces.

The units are equipped with local observations that offer insights into their immediate surroundings. These local observations form a vector encompassing crucial details such as distances, relative x and y coordinates, health and shield status, as well as the unit types of both allied and enemy units within the unit's visual range. Complementing these local observations is a centralized global state designed for facilitating centralized learning. This global state includes comprehensive information about all units

Event	Dense reward	Sparse reward
Death of single enemy	+10	+10
Death of single ally	-5	-5
Win	+200	+200
Enemy hit-point	-Enemy hit-point	-
Ally hit-point	+Ally hit-point	-
Other elements	+/- elements such as 'shield'	-

Table 1: The reward setting for dense and sparse SMAC environment.

Figure 7: The Starcraft Multi-agent Challenge environment.

on the map, featuring coordinates, local observations, energy levels of Medivacs, cooldowns of allied units, and the most recent actions taken by all agents. The action space available to agents operates in a discrete manner, allowing actions such as move[direction], attack[enemy_id], stop, and no-op.

The reward system for SMAC can be categorized into two main settings: dense rewards and sparse rewards. Dense rewards encompass rewards for specific in-game achievements, including hit-point damage dealt, enemy units eliminated, and overall success in battle. On the other hand, sparse rewards provide a stark contrast, offering a reward of +1 for winning an episode and -1 for losing. Additionally, there exists a nuanced reward setting known as 'partially sparse rewards,' wherein rewards are distributed for actions such as eliminating an opponent's unit, losing one of our units, or emerging victorious in a battle. It is noteworthy that our primary focus lies on utilizing this 'partially sparse rewards' setting as it provides a more intricate and nuanced evaluation of agent performance. This detailed reward structure not only influences the learning dynamics of the agents but also adds complexity to the training process, requiring strategic decision-making to optimize for both dense and sparse reward scenarios.

One noteworthy aspect of SMAC Sparse is the intricate nature of rewards. When agents receive positive rewards beyond what the environment provides, there's a tendency for agents to prioritize prolonging the episode rather than focusing on defeating the enemy units, potentially falling into local optima. To mitigate this issue, we employ a normalization strategy for our intrinsic rewards, namely r^{exp} and r^{aware}, ensuring they are less than or equal to zero. This normalization proves crucial, as it penalizes agents for unnecessarily extending the episode, thus steering them away from local optima. This strategic use of intrinsic rewards aligns with similar approaches found in prior works, notably in settings with comparable sparse reward structures, as evidenced by studies like [49]. By addressing the intricacies of sparse rewards in this manner, we enhance the learning dynamics of our agents, promoting more effective and strategic decision-making throughout the training process.

Event	Checkpoint reward	Score reward
Our team scores	-	+1
Opposing team scores	-	-1
Score from checkpoint	+0.1 · checkpoints	-

Table 2: The reward setting for GRF environment.

5.2 Google Research Football

Google Research Football provides a comprehensive simulation environment capturing the fundamental aspects of football, including goals, fouls, corners, penalty kicks, and offsides. In its original form, the opposing team is comprised of rule-based AI bots from GameplayFootball, allowing for adjustable parameters related to the bot's reaction time and decision-making speed. The state of the game is intricately defined by various factors such as the ball's position, possession, coordinates of all players, the active player, and the overall game state.

Observations within this environment are presented in three distinct representations to offer flexibility and cater to different learning paradigms. The pixel representation presents a 1280x720 RGB image that encompasses critical information like the scoreboard and a small map, allowing for a comprehensive view of all player positions on the field. The super mini-map representation, on the other hand, is a 72x96 matrix that encodes the positions of the home team, away team, ball, and the active player for the current time step, providing a condensed yet informative snapshot of the game.

In our specific implementation, we opt for the floats representation method. This method delivers a 115-dimensional encoded vector encapsulating essential information such as players' coordinates, ball possession and direction, and the active player's game mode. By leveraging the floats method, we can precisely capture and process the intricate dynamics of the football simulation, enabling our agents to make informed decisions based on a rich and detailed set of observations. This strategic choice aligns with our goal of fostering effective and nuanced learning within the Google Research Football environment.

The action space within the Google Research Football environment is characterized by its diversity, encompassing 8-directional standard move actions, various kicking techniques, sprinting, slide tackling, and dribbling. This intricate set of actions allows our agents to engage in a wide array of maneuvers, providing them with the versatility needed to navigate the dynamic and complex game of football. The reward system in this environment is designed with granularity, incorporating two distinct functions: scoring rewards and checkpoint rewards. Scoring rewards are straightforward, offering a +1 reward when our team scores a goal and a -1 penalty if the opposing team scores.

Checkpoint rewards add an additional layer of nuance to the reward system by dividing the area around the opposing goal into 10 checkpoints based on Euclidean distance. Agents are then rewarded with +0.1 for successfully scoring from a specific checkpoint, with the reward gradually accumulating to a maximum of +1. Our experimental setup focuses on utilizing the scoring rewards mechanism to

Figure 8: The Google Research Football environment.

guide the learning process effectively.

To facilitate and expedite the learning iterations, Google Research Football provides a football academy featuring 11 simpler scenarios. Each scenario emulates specific situations that commonly occur in actual soccer matches, allowing agents to hone their skills in targeted environments. Episodes in these scenarios can conclude when the ball is lost or our team scores a goal before reaching a predefined limit. In our specific experiments, we employ the *Academy_3_vs_1_with_keeper*, *Academy_counterattack _hard*, and *Academy_corner* environments from the football academy. Notably, the choice of observation scale can significantly impact the consistency of the reward scale when using prediction error as a measure for exploration bonus [58]. Consequently, in the Google Research Football environment, we strategically normalize r^{aware} to scale progressively with the prediction error, ensuring a balanced and effective exploration strategy within the complex and dynamic scenarios presented by the game.

5.3 Common Hyperparameter Setup

The FoX algorithm is implemented on top of QMIX [7], leveraging its default hyperparameters as suggested by the original paper. QMIX undergoes optimization using the RMSprop optimizer, with a learning rate set at 5×10^{-4}. The architecture of each agent network is designed with a fully connected layer followed by a GRU layer featuring a hidden state dimensionality of 64. The exploration strategy incorporates ε-greedy exploration, initializing ε at 1.0 and annealing it to 0.05 within the first 50,000 timesteps. The mixing network, a critical component of QMIX, is composed of 32 dimensions, and we update the target network every 200 episodes using the most recent 5000 episodes from the buffer.

Furthermore, each algorithm underwent rigorous testing with 4 random seeds for SMAC and 5 random seeds for GRF, ensuring a comprehensive evaluation across diverse scenarios. Table 3 provides an insightful overview of the common parameters employed consistently in our experiments, consolidating the key elements that contribute to the effectiveness of our approach in multi-agent reinforcement learning settings.

Parameters	FoX	CDS	EMC	MASER	QMIX	ROMA	MAVEN	LIIR	COMA
Optimizer	RMSProp	RMSProp	RMSProp	RMSProp	RMSProp	RMSProp	RMSProp	RMSProp	RMSProp
Agent Runner	Episode	Episode	Episode	Episode	Episode	Parallel	Episode	Parallel	Parallel
ε anneal step	50000	50000	50000	50000	50000	50	50000	50000	100000
Replay buffer size	5000	5000	5000	5000	5000	5000	5000	32	8
Target update interval	200	200	200	200	200	200	200	200	200
Mini-batch size	32	32	32	32	32	32	32	32	8
Mixing network dimension	32	32	32	32	32	32	32	-	-
Discount factor γ	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Learning rate	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005	0.0005
λ_{reg}	0.1	0.1	-	-	-	-	-	-	-

Table 3: Values for hyperparameters used in the experiment.

5.4 Implementation Details of FoX

$\mathscr{F}$-Net consists of an encoder and two decoders, each comprising two fully-connected layers culminating in the output layer. The encoder and decoder structures are strategically crafted to enhance the learning capabilities of the model. The hidden state of the network is set at 128 dimensions, providing the model with sufficient capacity to capture intricate patterns and relationships within the data. For optimization, we employed the Adam optimizer, utilizing a learning rate of 1×10^{-3} to strike a balance between convergence speed and stability during training. Each training iteration involved the use of 25 samples, contributing to the model's exposure to diverse scenarios and enhancing its adaptability.

The agent network within $\mathscr{F}$-Net is a pivotal component, and it is thoughtfully crafted to serve distinct purposes. Comprising individual Q-functions denoted as $Q^{Loc,i}$, a shared Q-function denoted as Q^{Shared}, and a formation-based Q-function denoted as $Q^{\mathscr{F}}$, this network inherits the structural characteristics of the agent networks employed in QMIX. Specifically, it features a fully connected layer followed by a GRU layer, with the hidden state dimensionality set at 64. This configuration ensures that the agent network possesses the necessary expressive power to effectively model the complex dynamics inherent in multi-agent environments, striking a balance between model complexity and computational efficiency. The shared structural elements across QMIX and $\mathscr{F}$-Net provide a basis for comparative analysis and facilitate a deeper understanding of the nuanced contributions introduced by the formation-based components in $\mathscr{F}$-Net.

FoX incorporates a set of hyperparameters crucial for shaping its exploration and awareness mechanisms. These include β_1 and β_2, which govern the influence of the exploration reward r^{exp} and the awareness reward r^{aware}, respectively. In the hyperparameter tuning process, we systematically explored the space of β_1 and β_2 values, conducting a search over the sets $\beta_1 \in \{0.001, 0.005, 0.01, 0.02, 0.1\}$ and $\beta_2 \in \{0.001, 0.005, 0.01, 0.05\}$. Another key hyperparameter is l, dictating the binning capacity of the $round()$ function. We performed a search over $l \in \{0, 1, 3\}$ to determine the most effective binning strategy for the visitation count.

Moreover, the parameter m, representing the length of the hash code, plays a pivotal role in shaping the dimensionality of the hash codes generated by SimHash. In our comprehensive experiments, we

Environment	β_1	β_2	l	m	λ_{GF}
Starcraft Multi-agent Challenge (Sparse)					
2m_vs_1z (Sparse)	0.1	0.01	3	9	0.1
3m (Sparse)	0.02	0.02	1	9	0.1
2s3z (Sparse)	0.01	0.01	3	9	0.1
8m (Sparse)	0.02	0.02	1	9	0.1
Google Research Football					
Academy_3_vs_1_with_keeper	0.01	0.01	3	9	0.1
Academy_counterattack_hard	0.01	0.01	3	9	0.1
Academy_corner	0.01	0.01	3	9	0.1

Table 4: Best hyperparameter setup.

set $m = 9$ for the hash code length, with the exception of the modified 2s3z, to represent the angle of observation differences in 512 distinct categories. Notably, the hyperparameter λ_{GF}, controlling the strength of gradient flipping, was set to $\lambda_{GF} = 0.1$ across all experiments, highlighting its consistent role in preventing the delivery of redundant information. The selection of hyperparameters contributes to the versatility and effectiveness of FoX in a range of multi-agent scenarios, as showcased in the experimental results detailed in the paper. The specific parameter values utilized in the reported experiments are meticulously documented in Table 4 for clarity and reproducibility.

VI Results

6.1 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 from the change of health points, making the task more challenging. 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 four scenarios: 3m, 8m, 2s3z, 2m_vs_1z. The exploration scheme of FoX can be easily applied to other existing MARL algorithms, but as we implement FoX on QMIX during our experiments, we tested the performance of FoX against several QMIX-based algorithms, including EMC [19], MASER [49], and ROMA [47].

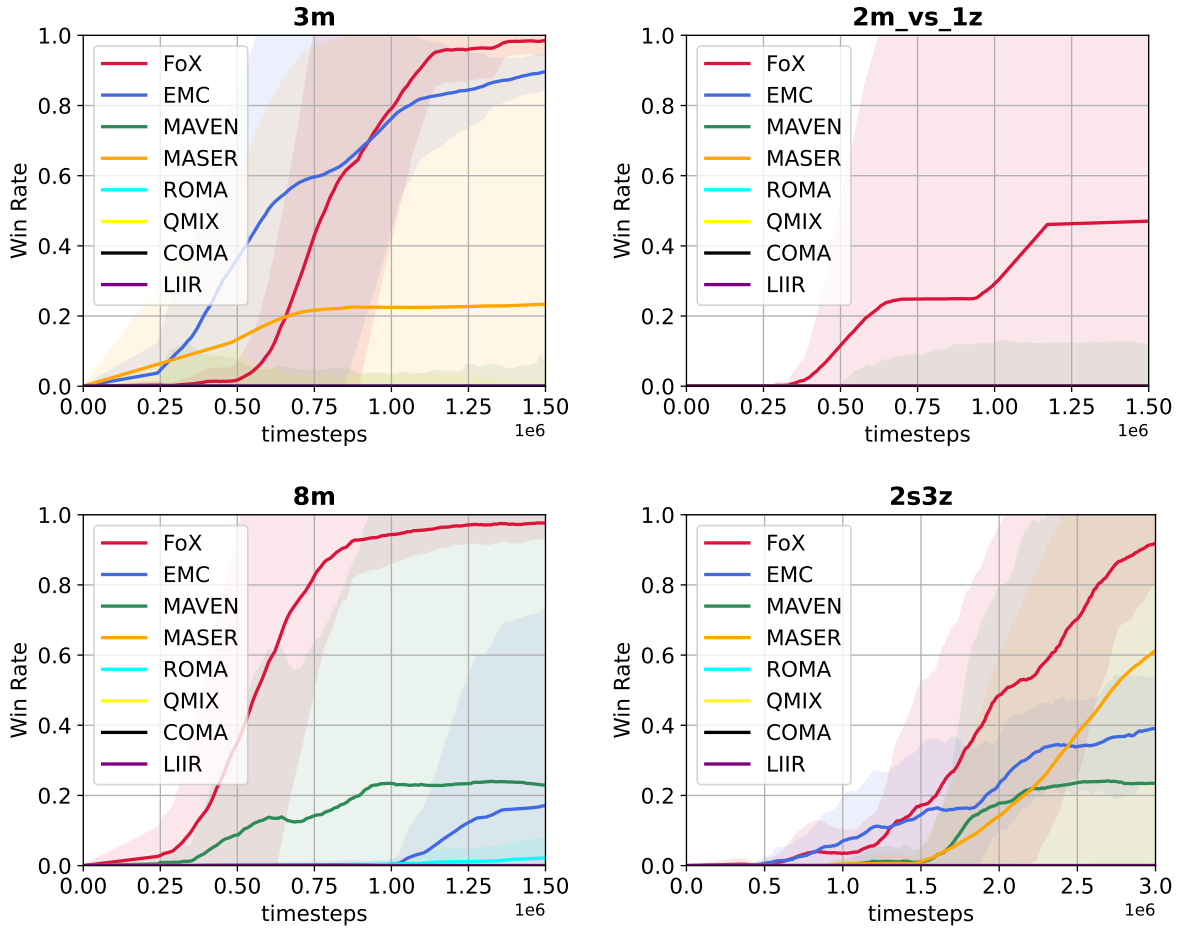


Figure 9: Performance results on SMAC(sparse)

Additionally, we compared FoX with MAVEN [50] and LIIR [42] to assess its performance against works that specifically address exploration, along with COMA [10] as a policy-gradient method. All experiments of the baselines were conducted with the author-provided codes. As both FoX and EMC

can be implemented on QMIX or QPLEX [28], we conducted our experiments with QMIX-based FoX and EMC for a 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. This highlights the effectiveness of FoX in tackling the challenges of sparse reward environments and its robustness across diverse scenarios in SMAC.

6.2 Performance on GRF

In Google Research Football (GRF) [21], the agents' observation contains crucial information such as the location of players, opponents, and the ball. Our implementation introduces a termination condition in the episode, triggering a negative reward for agents when the ball crosses into the other half of the court. Agents receive rewards solely upon episode termination, where a reward of +100 is granted if they score and -1 otherwise. Within the GRF framework, we conduct a comprehensive comparison of the FoX framework against prominent baselines, including QPLEX [28], CDS [51], MAVEN [50], and QMIX [7]. Additionally, we explore variations of QMIX by implementing the visitation count method, denoted as QMIX+JC for QMIX with joint observation visitation counts and QMIX+IC for individual observation visitation counts.

In our experiments, we meticulously evaluate the performance of FoX across three academy scenarios: *Academy_3_vs_1_with_keeper*, *Academy_counterattack_hard*, and *Academy_corner*. Figure 7 provides a visual representation of the performance of FoX alongside the baseline algorithms in these GRF scenarios. The results showcase FoX's exceptional performance in all three scenarios, consistently outperforming the other baselines. Particularly noteworthy is FoX's superiority over other visitation count baselines, reinforcing the notion that formation-based state equivalence efficiently reduces the search space in MARL. These findings underscore the efficacy of FoX in addressing the challenges posed by GRF scenarios and emphasize its robustness in optimizing exploration and learning in complex environments.

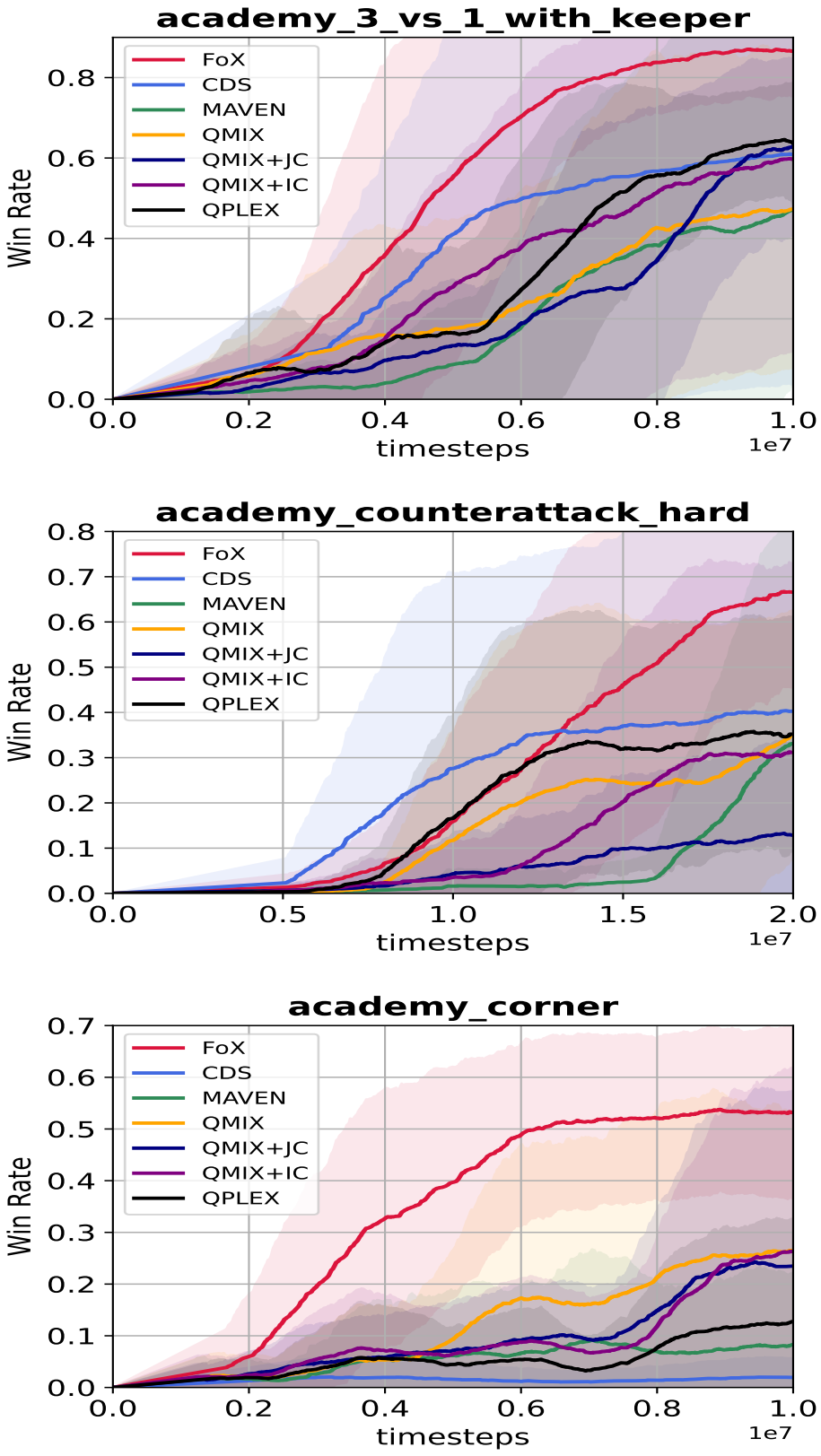


Figure 10: Performance results on GRF

VII Ablation Study

In this section, we analyze the exploration behavior and tendency of intrinsic rewards of FoX. Furthermore, we evaluate the individual 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* environment with a default setting of $\beta_1 = 0.01$ and $\beta_2 = 0.01$ in our ablation studies.

7.1 Component Evaluation

To comprehensively assess the individual contributions of each component within the FoX framework, we systematically deactivate specific elements, offering valuable insights into the impact of each component on the intrinsic rewards r^{exp} and r^{aware}, which are integral to formation-aware exploration. The Component Evaluation graph, vividly depicted in Figure 11, serves as a visual testament to the nuanced effects of various components, including intrinsic rewards, gradient flipping (GF), Q-function $Q^{\mathcal{F}}$ with formation information, and shared Q learning.

Upon careful examination of Figure 11, the FoX variants, namely FoX-r^{exp} and FoX-r^{aware}, exhibit a discernible decrease in performance when compared to the complete FoX algorithm. This marked reduction underscores the pivotal role played by intrinsic rewards in enhancing the overall performance of FoX. Moreover, the exclusion of $Q^{\mathcal{F}}$ and Shared Q from the FoX framework leads to a significant performance reduction, underscoring the indispensability of these components in bolstering the algorithm's efficacy. The inclusion of Q-function $Q^{\mathcal{F}}$ equips the FoX framework with formation-specific information, while the shared Q learning fosters collaborative learning among agents, both of which contribute significantly to the observed improvements.

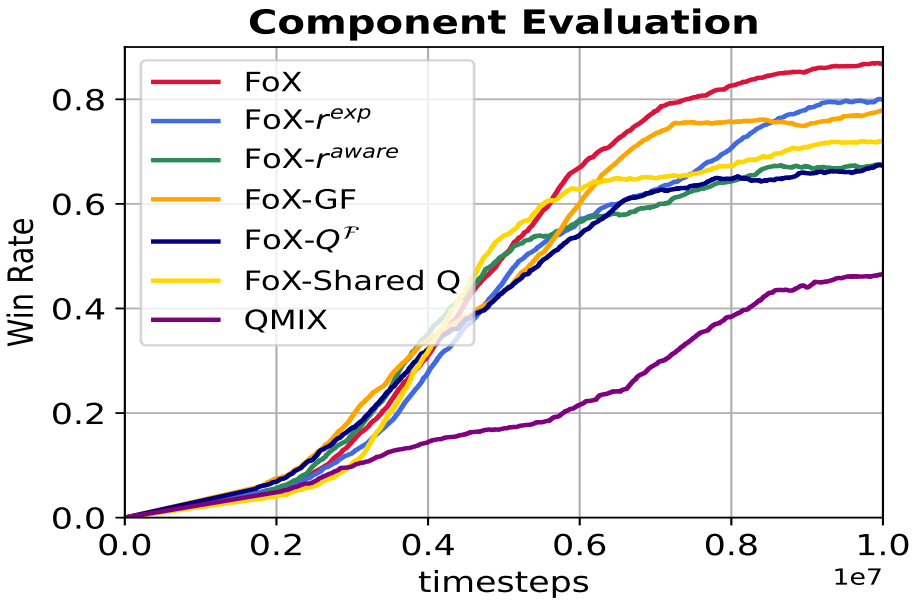


Figure 11: Component evaluations on GRF

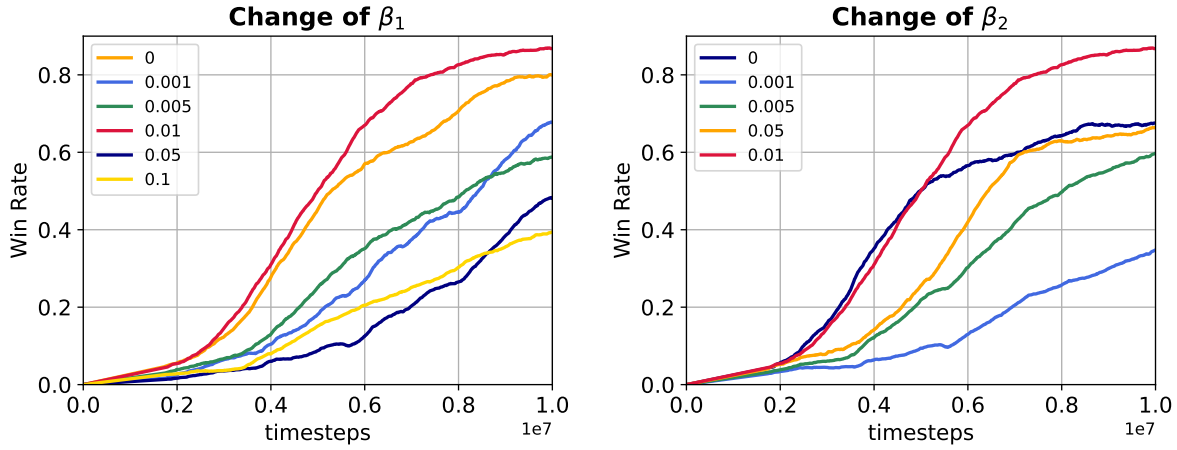


Figure 12: Effect of intrinsic rewards on GRF

The final component is the impact of gradient flipping (GF), a crucial mechanism designed to minimize irrelevant information from individual trajectories during latent variable extraction. FoX-GF, representing the variant without gradient flipping, shows a notable decrease in performance, providing compelling evidence for the efficacy of the gradient flipping mechanism. This outcome underscores the importance of gradient flipping in enabling the latent variable z to selectively capture formation-relevant information from the agent's trajectory τ, thereby enhancing the overall performance of the FoX framework.

7.2 The Effect of Intrinsic Rewards

The change in the β_1 graph, depicted in Figure 12, provides insightful observations into the efficacy of formation-aware exploration within the context of the *Academy_3_vs_1_with_keeper* environment. Notably, the performance improvement associated with $\beta_1 = 0.01$ compared to $\beta_1 = 0$ underscores the effectiveness of incorporating formation-awareness in exploration. However, it's imperative to note that an overly aggressive exploration strategy, as indicated by β_1 values exceeding 0.01, can inadvertently impede learning by slowing down the exploration of diverse formations. Striking a balance in β_1 is essential to ensure optimal exploration without impeding the learning process.

On the other front, scrutinizing the reward changes corresponding to $\beta_2 > 0.01$ reveals another facet of the formation-aware exploration's impact. In this context, the agents are deliberately guided to be cognizant of the formation. This deliberate encouragement of formation-awareness proves to be instrumental in mitigating the information bottleneck arising from partial observability. By fostering awareness of formations formulated by other agents, the learning process receives a significant boost. This nuanced approach not only aids in effective exploration but also facilitates a more informed learning process in dynamic and complex environments.

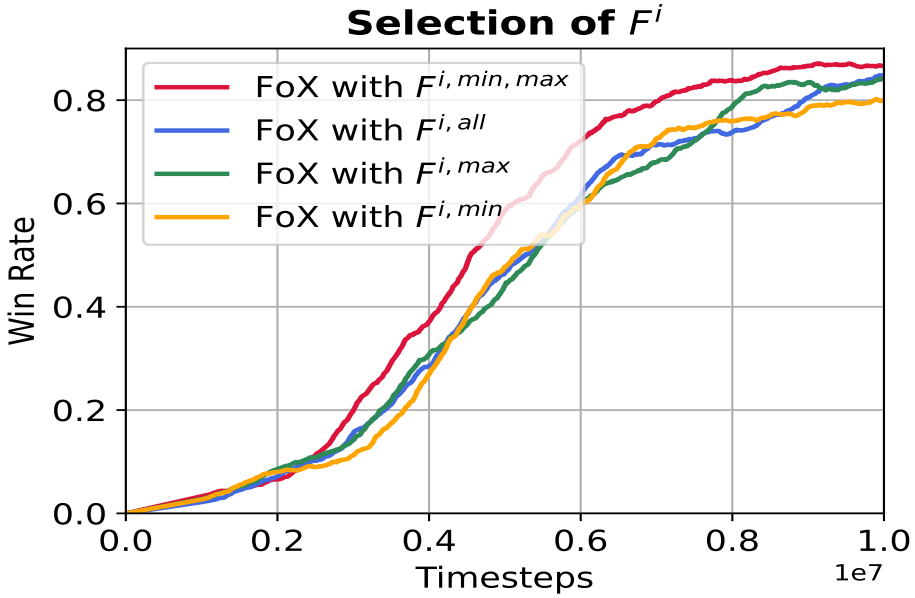


Figure 13: Effect of formation selection on GRF

7.3 Formation Selection

According to Figure 13, 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 6.

7.4 Exploration path on GRF

In Figure 14, we gain valuable insights into the exploration trajectory of FoX as the agent gradually becomes cognizant of its current formation. The depiction commences at the initialization phase ($t = 0$), where formations are crafted beyond the immediate sight range of the agents. Notably, owing to partial observability, these formations remain elusive to the agents' awareness. As we progress to time steps $t = 10$ and $t = 20$, the persistent exploration pattern indicates agents' endeavors to discover formations within their awareness scope.

A pivotal transition becomes evident at $t = 30$, marking the point where agents successfully achieve a formation that aligns with their awareness capabilities. This achievement triggers a notable reduction in exploration, and agents shift their focus towards maximizing r^{ext}, the reward emanating from interactions with the environment. The nuanced exploration dynamics captured in Figure 9 underscore the adaptive nature of FoX, showcasing how it efficiently guides agents to explore and comprehend formations within the constraints of partial observability, ultimately enhancing their decision-making and learning capabilities over time.

Figure 14: Exploration path on GRF *Academy_counterattack_hard*

7.5 Intrinsic Reward Analysis

In the comprehensive examination presented in this section, we delve into the intricate dynamics of intrinsic rewards, namely r^{exp} and r^{aware}, as the learning process unfolds. Figure 10 serves as a visual guide to deciphering the nuanced behaviors of these intrinsic rewards over distinct phases of training. Commencing at the initial stage ($t = 0$), where agents grapple with formulating a viable winning strategy, the intrinsic rewards exhibit noteworthy patterns. The agents, finding themselves without a clear-cut winning strategy, embark on an exploration journey, resulting in the visitation of diverse formations.

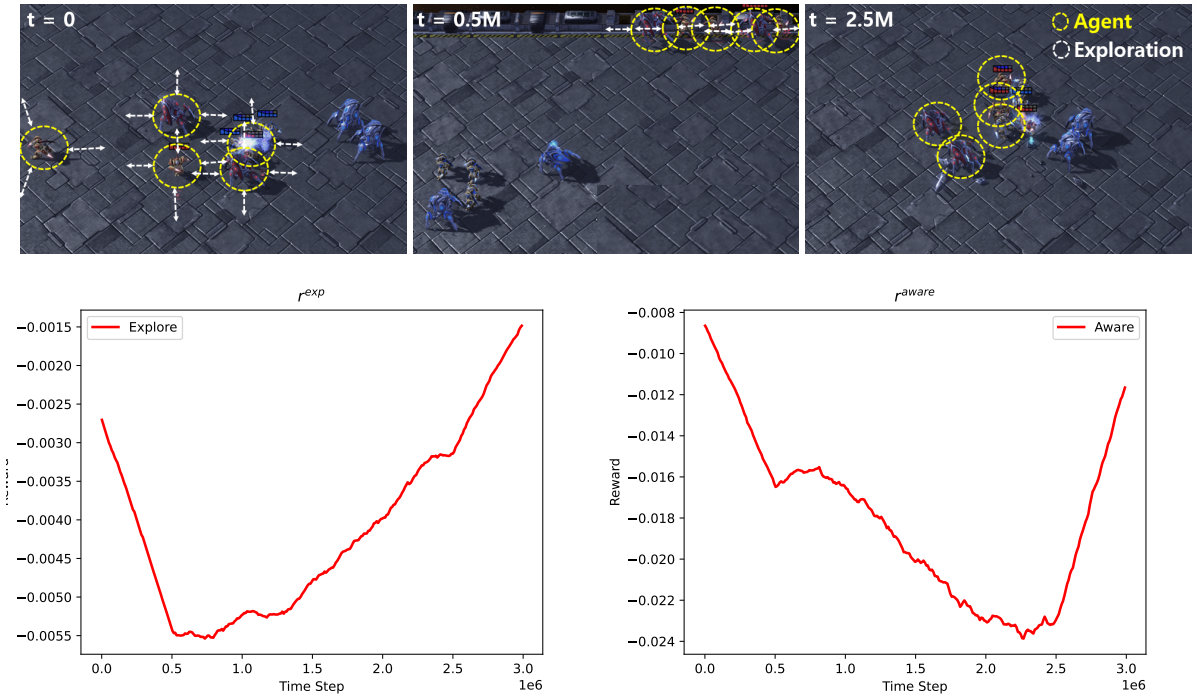


Figure 15: Agent behaviors according to intrinsic rewards

However, a counterintuitive trend surfaces as r^{exp} experiences a decline, primarily attributed to recurrent instances of states where all ally agents face elimination.

As the training progresses to $t = 0.5M$, a pivotal turning point unfolds, marked by a discernible ascent in r^{exp}, signifying a notable improvement in the agents' ability to navigate the combat scenarios. Concurrently, r^{aware} demonstrates a contrary trajectory, showcasing a decline as agents delve into the exploration of diverse formations.

A crucial juncture emerges at $t = 2.5M$, where the agents have successfully assimilated winning strategies. This phase is characterized by a convergence of both r^{exp} and r^{aware} towards zero, indicating that the agents, having acquired optimal combat strategies, now prioritize maximizing the external reward r^{ext} from interactions with the environment. The intricate interplay between these intrinsic rewards captures the evolving nature of the learning process, shedding light on how FoX enables agents to traverse a trajectory from exploration-driven diverse formation visits to a refined strategy-driven approach, ultimately enhancing their combat effectiveness.

VIII Conclusion

In Multi-Agent Reinforcement Learning (MARL), the inherent challenges lie in agents having to make decisions based on partial observations, a limitation that becomes increasingly pronounced as the number of agents rises, leading to exponential growth in the exploration space. The complexity is compounded by the fact that an agent's observation is not only contingent on its own actions but is also influenced by the actions of other agents. In navigating these challenges, this paper proposes a novel approach to address the intricate dynamics among agents, leveraging the concept of formations. Formations serve as condensed representations that encapsulate information about the relationships between agents and the holistic state of the environment, effectively compressing and consolidating pertinent details.

Furthermore, the formulation of formations proves particularly relevant in the context of incomplete observability, where the agent's observation may lack comprehensive information about the actions and states of fellow agents. To mitigate this, the paper advocates for the exploration of diverse formations, introducing a nuanced strategy that enhances the efficiency of exploration techniques in MARL. The essence of this approach lies in guiding agents to be well-aware of their current formation, fostering a more informed decision-making process.

Moreover, the notion of Formation-aware exploration introduced by the FoX algorithm demonstrates state-of-the-art performance in two benchmark environments: the StarCraft Multi-Agent Challenge (SMAC) and the Google Research Football (GRF) environment. The algorithm's prowess in these settings underscores its efficacy in addressing the formidable challenges posed by partial observability and the expansive exploration space inherent in MARL scenarios.

References

[1] T. Chu, J. Wang, L. Codecà, and Z. Li, "Multi-agent deep reinforcement learning for large-scale traffic signal control," *IEEE Transactions on Intelligent Transportation Systems*, vol. PP, 03 2019.

[2] X. Wang, L. Ke, Z. Qiao, and X. Chai, "Large-scale traffic signal control using a novel multi-agent reinforcement learning," 08 2019.

[3] O. Vinyals, I. Babuschkin, W. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. Agapiou, M. Jaderberg, and D. Silver, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, 11 2019.

[4] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Qt-opt: Scalable deep reinforcement learning for vision-based robotic manipulation," 06 2018.

[5] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proceedings of the Tenth International Conference on Machine Learning*, 1993, pp. 330–337.

[6] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *NIPS*, 2017.

[7] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[8] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 5872–5881. [Online]. Available: https://proceedings.mlr.press/v80/zhang18n.html

[9] J. Gupta, M. Egorov, and M. Kochenderfer, "Cooperative multi-agent control using deep reinforcement learning," 11 2017, pp. 66–83.

[10] J. N. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

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

[12] S. Sukhbaatar, A. Szlam, and R. Fergus, "Learning multiagent communication with backpropagation," 05 2016.

[13] G. Ostrovski, M. Bellemare, A. Oord, and R. Munos, "Count-based exploration with neural density models," 03 2017.

[14] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, "Curiosity-driven exploration by self-supervised prediction," in *Proceedings of the 34th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. PMLR, 06–11 Aug 2017, pp. 2778–2787. [Online]. Available: https://proceedings.mlr.press/v70/pathak17a.html

[15] Y. Burda, H. Edwards, D. Pathak, A. Storkey, T. Darrell, and A. A. Efros, "Large-scale study of curiosity-driven learning," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=rJNwDjAqYX

[16] H. Tang, R. Houthooft, D. Foote, A. Stooke, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "#exploration: A study of count-based exploration for deep reinforcement learning," 11 2016.

[17] T. Wang, J. Wang, Y. Wu, and C. Zhang, "Influence-based multi-agent exploration," *ArXiv*, vol. abs/1910.05512, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:204509513

[18] N. Jaques, A. Lazaridou, E. Hughes, Çaglar Gülçehre, P. A. Ortega, D. Strouse, J. Z. Leibo, and N. de Freitas, "Social influence as intrinsic motivation for multi-agent deep reinforcement learning," in *International Conference on Machine Learning*, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:60440666

[19] L. Zheng, J. Chen, J. Wang, J. He, Y. Hu, Y. Chen, C. Fan, Y. Gao, and C. Zhang, "Episodic multi-agent reinforcement learning with curiosity-driven exploration," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 3757–3769. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/1e8ca836c962598551882e689265c1c5-Paper.pdf

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

[21] K. Kurach, A. Raichuk, P. Stańczyk, M. Zając, O. Bachem, L. Espeholt, C. Riquelme, D. Vincent, M. Michalski, O. Bousquet *et al.*, "Google research football: A novel reinforcement learning environment," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 04, 2020, pp. 4501–4510.

[22] J. Foerster, I. A. Assael, N. de Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/c7635bfd99248a2cdef8249ef7bfbef4-Paper.pdf

[23] Y. Yang, J. Hao, B. Liao, K. Shao, C. Guangyong, W. Liu, and H. Tang, "Qatten: A general framework for cooperative multiagent reinforcement learning," *arXiv preprint arXiv:2002.03939*, 2020.

[24] S. Iqbal and F. Sha, "Actor-attention-critic for multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2961–2970.

[25] C. Yu, A. Velu, E. Vinitsky, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of ppo in cooperative, multi-agent games," *arXiv preprint arXiv:2103.01955*, 2021.

[26] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 5872–5881. [Online]. Available: https://proceedings.mlr.press/v80/zhang18n.html

[27] K. Son, D. Kim, W. J. Kang, D. E. Hostallero, and Y. Yi, "Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2019, pp. 5887–5896.

[28] J. Wang, Z. Ren, T. Liu, Y. Yu, and C. Zhang, "{QPLEX}: Duplex dueling multi-agent q-learning," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=Rcmk0xxIQV

[29] Y. Wang, B. Han, T. Wang, H. Dong, and C. Zhang, "Dop: Off-policy multi-agent decomposed policy gradients," in *International Conference on Learning Representations*, 2020.

[30] B. Peng, T. Rashid, C. Schroeder de Witt, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson, "Facmac: Factored multi-agent centralised policy gradients," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 208–12 221, 2021.

[31] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/8d8818c8e140c64c743113f563cf750f-Paper.pdf

[32] R. Houthooft, X. Chen, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "Vime: Variational information maximizing exploration," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/abd815286ba1007abfbb8415b83ae2cf-Paper.pdf

[33] A. Nair, B. McGrew, M. Andrychowicz, W. Zaremba, and P. Abbeel, "Overcoming exploration in reinforcement learning with demonstrations," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6292–6299.

[34] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu, "Reward-free exploration for reinforcement learning," in *Proceedings of the 37th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, H. D. III and A. Singh, Eds., vol. 119. PMLR, 13–18 Jul 2020, pp. 4870–4879. [Online]. Available: https://proceedings.mlr.press/v119/jin20d.html

[35] M. Machado, M. Bellemare, and M. Bowling, "Count-based exploration with the successor representation," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 5125–5133, 04 2020.

[36] A. Strehl and M. Littman, "An analysis of model-based interval estimation for markov decision processes," *Journal of Computer and System Sciences*, vol. 74, pp. 1309–1331, 12 2008.

[37] M. Bellemare, S. Srinivasan, G. Ostrovski, T. Schaul, D. Saxton, and R. Munos, "Unifying count-based exploration and intrinsic motivation," in *Advances in Neural Information Processing Systems*, D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Eds., vol. 29. Curran Associates, Inc., 2016. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2016/file/afda332245e2af431fb7b672a68b659d-Paper.pdf

[38] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel, "Vime: Variational information maximizing exploration," 05 2016.

[39] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, J. Dy and A. Krause, Eds., vol. 80. PMLR, 10–15 Jul 2018, pp. 1861–1870. [Online]. Available: https://proceedings.mlr.press/v80/haarnoja18b.html

[40] S. Han and Y. Sung, "A max-min entropy framework for reinforcement learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 25 732–25 745. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/d7b76edf790923bf7177f7ebba5978df-Paper.pdf

[41] B. Eysenbach, A. Gupta, J. Ibarz, and S. Levine, "Diversity is all you need: Learning skills without a reward function," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SJx63jRqFm

[42] Y. Du, L. Han, M. Fang, J. Liu, T. Dai, and D. Tao, "Liir: Learning individual intrinsic reward in multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds., vol. 32. Curran Associates, Inc., 2019. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2019/file/07a9d3fed4c5ea6b17e80258dee231fa-Paper.pdf

[43] L. Wang, Y. Zhang, Y. Hu, W. Wang, C. Zhang, Y. Gao, J. Hao, T. Lv, and C. Fan, "Individual reward assisted multi-agent reinforcement learning," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 17–23 Jul 2022, pp. 23 417–23 432. [Online]. Available: https://proceedings.mlr.press/v162/wang22ao.html

[44] T. Gupta, A. Mahajan, B. Peng, W. Boehmer, and S. Whiteson, "Uneven: Universal value exploration for multi-agent reinforcement learning," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 3930–3941. [Online]. Available: https://proceedings.mlr.press/v139/gupta21a.html

[45] I.-J. Liu, U. Jain, R. A. Yeh, and A. Schwing, "Cooperative exploration for multi-agent deep reinforcement learning," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 6826–6836. [Online]. Available: https://proceedings.mlr.press/v139/liu21j.html

[46] B. Liu, Z. Pu, Y. Pan, J. Yi, Y. Liang, and D. Zhang, "Lazy agents: A new perspective on solving sparse reward problem in multi-agent reinforcement learning," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 21 937–21 950. [Online]. Available: https://proceedings.mlr.press/v202/liu23ac.html

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

[48] T. Wang, T. Gupta, A. Mahajan, B. Peng, S. Whiteson, and C. Zhang, "{RODE}: Learning roles to decompose multi-agent tasks," in *International Conference on Learning Representations*, 2021. [Online]. Available: https://openreview.net/forum?id=TTUVg6vkNjK

[49] J. Jeon, W. Kim, W. Jung, and Y. Sung, "Maser: Multi-agent reinforcement learning with subgoals generated from experience replay buffer," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 041–10 052.

[50] A. Mahajan, T. Rashid, M. Samvelyan, and S. Whiteson, "Maven: Multi-agent variational exploration," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[51] C. Li, T. Wang, C. Wu, Q. Zhao, J. Yang, and C. Zhang, "Celebrating diversity in shared multi-agent reinforcement learning," in *Advances in Neural Information Processing Systems*, M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, Eds., vol. 34. Curran Associates, Inc., 2021, pp. 3991–4002. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2021/file/20aee3a5f4643755a79ee5f6a73050ac-Paper.pdf

[52] F. A. Oliehoek and C. Amato, *A Concise Introduction to Decentralized POMDPs*, 1st ed. Springer Publishing Company, Incorporated, 2016.

[53] R. Bellman, "Dynamic programming," *Science*, vol. 153, no. 3731, pp. 34–37, 1966.

[54] K. Aggarwal and H. K. Verma, "Hash rc6 — variable length hash algorithm using rc6," in *2015 International Conference on Advances in Computer Engineering and Applications*, 2015, pp. 450–456.

[55] N. Senavirathne and V. Torra, "Rounding based continuous data discretization for statistical disclosure control," *Journal of Ambient Intelligence and Humanized Computing*, 09 2019.

[56] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[57] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[58] Y. Burda, H. Edwards, A. Storkey, and O. Klimov, "Exploration by random network distillation," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=H1lJJnR5Ym

Acknowledgements

I would like to express my sincere gratitude to all those who have contributed to the completion of this master's thesis.

First and foremost, I would like to extend my thanks to Professor Seungyul Han, who guided me with dedication and sincerity from the beginning to the end. I had many shortcomings, but thanks to the professor's knowledge and guidance based on that knowledge, I was able to grow as a researcher. Moreover, the professor's attitude towards the work demonstrated in the process of completing the master's thesis, and the perseverance shown, allowed me to grow personally as well. Without a doubt, Professor Han had the most significant impact on completing the master's thesis. Being able to receive guidance from the professor was a great fortune for me. Thank you.

I am also thankful to the members of my thesis committee, Professor Jeong hwan Jeon and Professor Gi-soo Kim, for their insightful comments and thoughtful suggestions, which greatly enriched the quality of this work.

Next, I would like to express my gratitude to my family who supported me from afar. There were many changes in the direction I took until my master's graduation, and during this process, my father, mother, and sister always believed in me and provided encouragement. Knowing that they are proud of me has been a great help during difficult times.

Lastly, I want to express my gratitude to the people in the research lab who spent my master's life closely. Especially, I want to thank Junghyuk Yeom, who has been with me from the beginning to the end of the master's program as a fellow student. It seems that having fellow students who can help and support each other has allowed us to successfully conclude the master's program. Along with Junghyuk, I extend my thanks and encouragement to all the people in our research lab who shared both joys and hardships.

I hope that the heartfelt contributions of everyone who assisted in this challenging yet rewarding journey are well reflected. Thank you.