

Multi-Task Multi-Agent Reinforcement Learning via Skill Graphs

Guobin Zhu^{1,2}, Rui Zhou¹, Wenkang Ji², Hongyin Zhang³, Donglin Wang³ and Shiyu Zhao²

Abstract—Multi-task multi-agent reinforcement learning (MT-MARL) has recently gained attention for its potential to enhance MARL’s adaptability across multiple tasks. However, it is challenging for existing multi-task learning methods to handle complex problems, as they are unable to handle unrelated tasks and possess limited knowledge transfer capabilities. In this paper, we propose a hierarchical approach that efficiently addresses these challenges. The high-level module utilizes a skill graph, while the low-level module employs a standard MARL algorithm. Our approach offers two contributions. First, we consider the MT-MARL problem in the context of unrelated tasks, expanding the scope of MTRL. Second, the skill graph is used as the upper layer of the standard hierarchical approach, with training independent of the lower layer, effectively handling unrelated tasks and enhancing knowledge transfer capabilities. Extensive experiments are conducted to validate these advantages and demonstrate that the proposed method outperforms the latest hierarchical MAPPO algorithms. Videos and code are available at <https://github.com/WindyLab/MT-MARL-SG>

Index Terms—Multi-robot systems, Multi-agent reinforcement learning, Multi-task reinforcement learning,

I. INTRODUCTION

MULTI-AGENT reinforcement learning (MARL) is promising for decision-making in multi-robot systems and has achieved notable success in areas like games [1], [2] and robotic control [3]–[5]. However, its performance tends to be task-specific, requiring retraining for new tasks due to weak generalization [6]. Therefore, it is important to study how MARL can adapt flexibly to multiple tasks.

In recent years, numerous studies on multi-task reinforcement learning (MTRL) have emerged [7]–[9]. These works leverage inter-task correlations to enhance performance across tasks. Existing approaches can be broadly categorized into two types. The first focuses on learning multiple tasks simultaneously, where shared attributes/knowledge (*e.g.* data, model parameters) among tasks are utilized to accelerate learning [10]. The second emphasizes sequential task learning, where previously acquired knowledge is retained and reused to facilitate learning when encountering related new tasks [11], [12]. Both approaches inherently involve knowledge transfer. A common strategy is

Manuscript received: March, 22, 2025; Revised xxx, xxx; Accepted xxx, xxx. This paper was recommended for publication by xxx upon evaluation of the Associate Editor and Reviewers’ comments. This work was supported by the STI 2030-Major Projects (No.2022ZD0208804), National Natural Science Foundation of China (No.62473017, No.62473320). (Corresponding author: Shiyu Zhao.)

¹School of Automation Science and Electrical Engineering, Beihang University, Beijing, China.

²WINDY Lab, Department of Artificial Intelligence, Westlake University, Hangzhou, China.

³MiLAB Lab, Department of Artificial Intelligence, Westlake University, Hangzhou, China.

E-mail: {zhugb, zhr}@buaa.edu.cn, {jiwenkang, zhanghongyin, wangdonglin, zhaooshiyu}@westlake.edu.cn.

Digital Object Identifier (DOI): see top of this page.

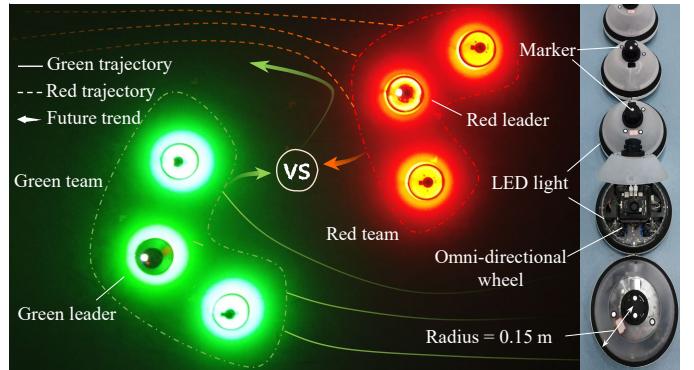


Fig. 1: Real-world experiment. The red and green robot teams, moving in flocking, engage in combat. After defeating the red team, the green team continues flocking motion.

to distill knowledge from multiple expert policies into a single network through supervised learning, enabling the integrated network to adapt to diverse tasks [13], [14].

Although MTRL has demonstrated effectiveness by leveraging inter-task knowledge [15], [16], it still has two notable limitations when applied to complex tasks. First, MTRL assumes that tasks share common attributes. If tasks are unrelated (*e.g.* adversarial and cooperative), knowledge transfer becomes infeasible, rendering conventional MTRL methods ineffective. Second, current MTRL rarely accounts for multi-agent settings. Multi-agent systems involve inter-agent interactions and often operate under partial observability, making MT-MARL fundamentally different from a direct extension of single-agent MTRL.

Coincidentally, hierarchical reinforcement learning (HRL) offers a promising approach. HRL is designed to decompose complex, long-term tasks into a hierarchy of subtasks, with the high-level policy deciding which low-level skill to execute [17]–[20]. In multi-agent multi-task settings, HRL presents two main advantages. First, it can be naturally integrated with MTRL [11], [20]. In a hierarchical framework, the high-level policy’s selection of low-level skills effectively serves as a mechanism for transferring knowledge of modular skills. Since the tasks underlying these skills may be either related or unrelated, the hierarchical approach is well-suited for multi-task scenarios. Second, HRL simplifies complex multi-agent settings by breaking problems down in a layered, step-by-step manner, thereby reducing both environmental and task complexity [21], [22].

However, existing HRL approaches still have two limitations. The first is that standard HRL cannot select multiple low-level strategies concurrently; in other words, HRL’s capacity for knowledge transfer is limited. The second is that most HRL studies consider only several dozen agents [23], and as this number increases, the curse of dimensionality becomes a

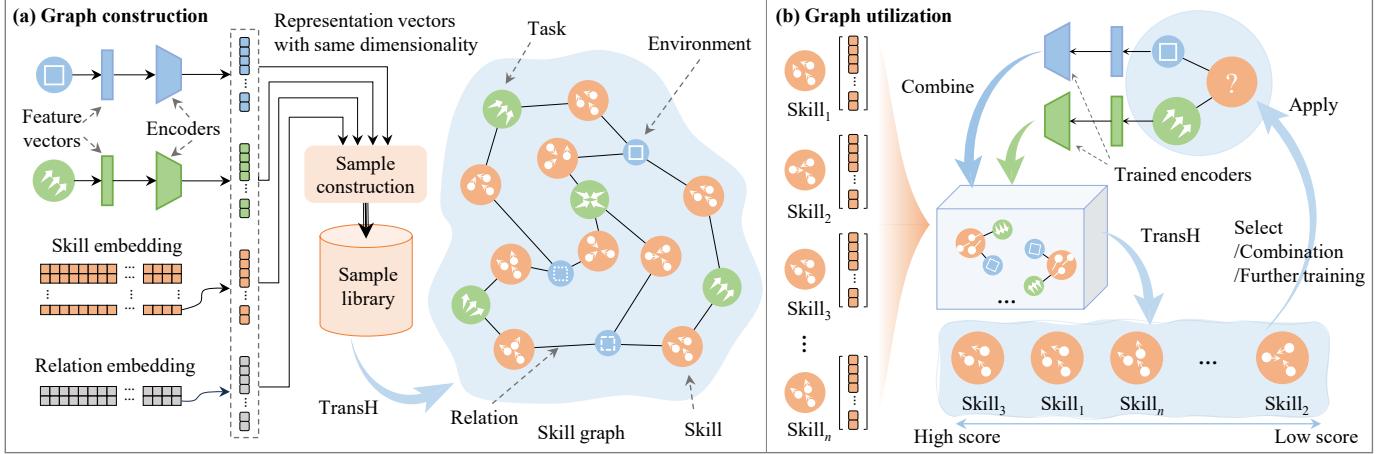


Fig. 2: Method overview. (a) is the construction of the skill graph and (b) is the utilization of the skill graph. In each embedding layer and representation vector, each small square represents a numerical value.

significant issue. In light of these limitations, there is a need for more general MT-MARL methods.

This paper proposes an effective hierarchical approach to address multi-task multi-agent problems. We considered the unrelated task settings, a scenario that has been rarely considered in previous work. On the high level, our approach uses a skill graph [24], inspired by the concept of knowledge graph (KG) [25]. KG conveniently describes entities and their relationships and enables effective knowledge transfer when queried. Thus, our method consists of two steps. First, we treat environments, tasks, and skills in RL as entities and use knowledge graph embedding (KGE) [26] to learn representations of these entities and their relationships, thereby constructing a skill graph. Second, when queried by new environments or tasks, all skills in the graph are evaluated via a scoring mechanism, and the final skill is selected, combined, or further refined from the high-scoring ones. Such a skill graph can handle unrelated tasks and enhance effective knowledge transfer. On the low level, our approach employs the MADDPG algorithm [27] with a local critic instead of a centralized one. This modification better aligns with robots' local perception and supports large-scale tasks. Finally, both simulation and real-world experiments verify the effectiveness of the proposed method.

In summary, the main contributions are as follows: 1) Our approach can handle multiple unrelated tasks, including both adversarial and cooperative tasks. This unrelated case has been rarely considered in prior work since no transferable knowledge exists. The proposed approach can handle the unrelated case due to its hierarchical structure combined with a skill graph. 2) Our approach enables skill selection, combination, and further learning through a scoring mechanism, offering effective knowledge transfer compared to standard HRL. The scoring mechanism inherently offers a more robust and flexible approach to standard HRL since it always produces a list of scores, meaning that high-scoring skills that rank at the top are always obtained. 3) We conducted experiments on a real robotic swarm platform to validate the effectiveness of the proposed method.

II. PRELIMINARIES

A. Knowledge Graph

A KG is a multi-relational graph composed of entities (nodes) and relations (different types of edges) [28]. Mathematically, a KG is defined as $\mathcal{G} = (\mathcal{E}, \mathcal{R}, \mathcal{T})$, where \mathcal{E} is the set of all entities contained in the KG. \mathcal{R} represents the set of all relations. \mathcal{T} is the set of factual triples, where a standard binary fact is a triple (h, r, b) , h and b mean the head entity and the tail entity, and r is the relation between the head entity and the tail entity.

Such a structure conveniently stores and links diverse knowledge with a unified specification. When encountering a new query, the KG can rapidly retrieve the target entity, while its associated entities can also be easily identified. Therefore, the KG facilitates efficient knowledge transfer [29].

B. Knowledge Graph Embedding

Although the KG provides a convenient representation of structured knowledge, its symbolic nature poses operational challenges. KGE addresses this issue by mapping entities and relations onto a continuous vector space (*i.e.*, a representation space), thereby facilitating operations while preserving the inherent structure of the KG.

The typical KGE process consists of three key steps [28]. First, entities and relations are represented, typically as vectors or matrices. Second, a scoring function is defined to assess the plausibility of a given fact triple, where observed facts in the KG generally receive higher scores than unobserved ones. Third, the representations of entities and relations are learned by solving an optimization problem that maximizes the plausibility of all facts. In essence, the KG is considered constructed once the representations of entities and relations have been learned.

Several KGE algorithms have been proposed, such as TransE [30] and TransH [26]. In this study, we employ the TransH model because it supports one-to-many, many-to-one, and many-to-many relationships. For example, in a one-to-many relationship, a single head entity may correspond to multiple

tail entities. In the context of RL, this corresponds to training multiple skills within a single environment.

III. METHODOLOGY

As illustrated in Fig. 2, the proposed method comprises two main components. The first constructs the skill graph using knowledge graph (KG) construction methods with the TransH model. The second involves utilizing the skill graph, where the TransH model scores all skills in the graph to infer the most suitable one.

A. Graph Construction

As described in Section II-B, the construction of the skill graph consists of three steps. The first step involves representing entities and relations. In the skill graph, entities include the environment (e), task (t), and skill (s), while relations denote the correspondence from the environment to skill ($r_{e \rightarrow s}$) and from task to skill ($r_{t \rightarrow s}$). To map both entities and relations into a unified representation vector space, there are two sub-steps (see Fig. 2(a)). First, we use feature vectors (different from representation vectors) to describe e and t . Then, we use an environment encoder and a task encoder (MLP networks with Leaky-ReLU activation) to map their respective feature vectors into the representation space. Second, we employ two embedding layers (initialized orthogonally) to represent skills and relations, respectively. Each embedding layer is a matrix where rows correspond to the number of relations/skills, and columns to the dimensions of their representation vectors. In this way, entities ($e/t/s$) and relations (r) are represented within the same vector space.

The second step involves defining the scoring function. We adopt the TransH model as our scoring function, which is expressed as $\mathbb{S} = \exp(-\lambda \|(\mathbf{h} - \mathbf{w}_r^T \mathbf{h} \mathbf{w}_r) + \mathbf{d}_r - (\mathbf{b} - \mathbf{w}_r^T \mathbf{b} \mathbf{w}_r)\|)$, where λ is a normal constant. \mathbf{h} and \mathbf{b} denote the head entity and tail entity representation vectors, respectively. \mathbf{w}_r and \mathbf{d}_r denote the relation-specific hyperplane and translation vector, respectively. The scoring function measures the plausibility of a triple (h, r, b) .

The third step involves learning the representation vectors. This is a supervised learning process that comprises two phases: sample construction and representation learning. In the terminology of KGE, samples are categorized as positive, negative, and soft samples [26]. Positive samples correspond to correct facts. During the skill collection phase (see Section IV), we know which skills are trained under specific environments and tasks, allowing us to construct factual triples $(e, r_{e \rightarrow s}, s)$ and $(t, r_{t \rightarrow s}, s)$; these triples serve as positive samples. Negative samples represent incorrect facts, which may arise from errors in the entities (e.g. $(e, r_{e \rightarrow s}, e')$, $(t, r_{t \rightarrow s}, t')$) or the relations (e.g. $(e, r_{t \rightarrow s}, s)$, $(t, r_{e \rightarrow s}, s)$). Soft samples refer to triples with varying degrees of plausibility: positive samples have a plausibility of 1, negative samples 0, and soft samples fall between 0 and 1. For example, for two similar environments e_1 and e_2 , if e_1 corresponds to skill s_1 , then $(e_1, r_{e \rightarrow s}, s_1)$ is a positive sample and $(e_2, r_{e \rightarrow s}, s_1)$ is a soft sample. Similarly, $(t_1, r_{t \rightarrow s}, s_1)$ is positive and $(t_2, r_{t \rightarrow s}, s_1)$ is soft. In this

Algorithm 1 Construction and Utilization of Skill Graph

- 1: **Construction:**
 - 2: Initialize encoders of e and t , embeddings of s and r
 - 3: Construct positive $(e/t, r_{e/t \rightarrow s}, s)$, negative $(e/t, r_{e/t \rightarrow s}, e'/t')$, $(e/t, r_{t/e \rightarrow s}, s)$ and soft $(e_2/t_2, r_{e/t \rightarrow s}, s_1)$
 - 4: **for** iteration = 1 to M **do**
 - 5: Randomly sample a mini-batch from the library
 - 6: Map e and t to representation space using encoders
 - 7: Update encoder and embeddings by minimizing \mathcal{L}
 - 8: **end for**
 - 9: **Utilization:**
 - 10: Map $e_{\text{new}}, t_{\text{new}}$ to representation space using encoders
 - 11: Combine the representation vectors of $e_{\text{new}}, t_{\text{new}}, r, s$
 - 12: Compute $\mathbb{S} = \mathbb{S}_{(t, r_{t \rightarrow s}, s)} \cdot \mathbb{S}_{(e, r_{e \rightarrow s}, s)}$ for each combination
 - 13: Infer the most suitable skill based on scores
-

manner, positive, negative, and soft samples can form a sample library.

With the samples prepared, the next step is to learn representations. We use the following loss function:

$$\mathcal{L} = (\mathbb{S}_{\text{posi}} - 1)^2 + (\mathbb{S}_{\text{nega}} - 0)^2 + \text{ReLU}(\mathbb{S}_{\text{soft}} - 1 + \delta), \quad (1)$$

where, $\delta = \sum_j k_j(p_{j,1} - p_{j,2}) / \sum_j k_j$ measures the similarity between entities 1 and 2, where k_j represents the weight coefficient and p_j denotes the j -th attribute of the feature vector. We mainly adjust k_j to differentiate between various types of entities. The subscript of \mathbb{S} indicates the sample type. As can be seen from loss (1), we hope that the positive sample score is 1, the negative sample score is 0, and the soft sample score is no more than $1 - \delta$. The value of \mathbb{S} depends on the representation vectors of entities and relations, which are obtained through the encoder and embedding layers. Therefore, minimizing \mathcal{L} is to optimize the parameters of encoders and embedding layers. The result of constructing the skill graph is to obtain these networks.

B. Graph Utilization

With the constructed skill graph, we can identify the most suitable skill when facing a new environment e_{new} and task t_{new} . The process involves three steps (see Fig. 2(b)). First, map the feature vectors of e_{new} and t_{new} to the representation space using their trained encoders. Second, combine their representation vectors with the relation and skill representation vectors from the graph. Third, score each combination using $\mathbb{S} = \mathbb{S}_{(t, r_{t \rightarrow s}, s)} \cdot \mathbb{S}_{(e, r_{e \rightarrow s}, s)}$ and TransH model to generate a sequence of scores. The resulting scores indicate how well each skill in the graph matches the query e_{new} and t_{new} .

If the score is close to 1 ($\alpha_{\text{high}} < \mathbb{S} \leq 1$), it indicates that the query pair has been previously encountered; thus, the highest-scoring skill is directly applied. If the score is in the intermediate range ($\alpha_{\text{low}} < \mathbb{S} \leq \alpha_{\text{high}}$), it suggests a reasonable relationship between the skill and the query pair, and the robot's action is derived through $\mathbf{a} = \sum_j w_j \mathbf{a}_j$, where j is the skill index for this score range, w_j is the normalized score for skill j . Here, normalized scores mean $\sum_j w_j = 1$. If all skill scores are low ($0 \leq \mathbb{S} \leq \alpha_{\text{low}}$), it indicates that the

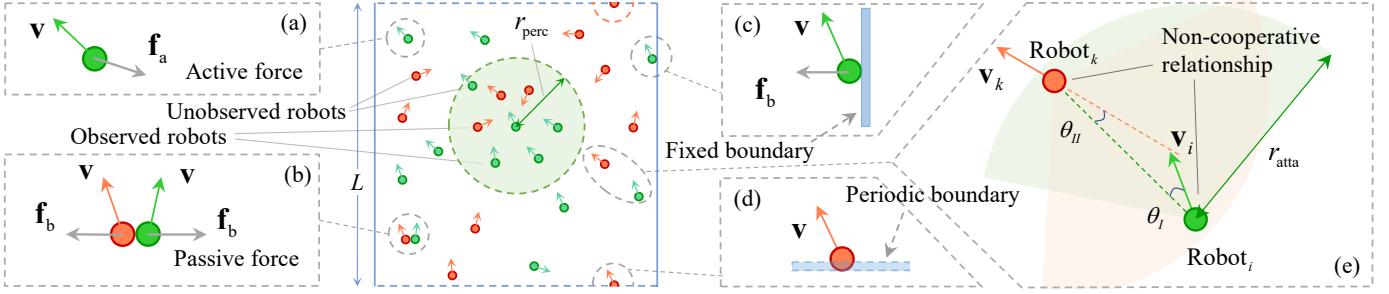


Fig. 3: Motion environment illustration. (a)(b) show the active and passive force, (c)(d) shows the fixed and periodic boundaries, and (e) shows the relative motion relationship, where $\theta_I = \arccos(\mathbf{p}_{ki} \cdot \mathbf{v}_i / (\|\mathbf{p}_{ki}\| \cdot \|\mathbf{v}_i\|))$, $\theta_{II} = \arccos(\mathbf{p}_{ki} \cdot \mathbf{v}_k / (\|\mathbf{p}_{ki}\| \cdot \|\mathbf{v}_k\|))$, The light green and red sectors represent $\theta_I \leq k_I \pi$, $\theta_{II} \leq k_{II} \pi$ and we choose $k_I = k_{II} = 0.4$.

query pair falls outside the skill distribution in the graph. In this case, the highest-scoring skill is further optimized using RL and then applied to the query pair.

Throughout the construction and utilization of the skill graph, it does not require task-related dependencies and can acquire the most suitable skills through selection, combination, or further learning, thereby enhancing knowledge transfer. This approach offers an effective solution for multi-task learning. The complete process is outlined in Algorithm 1.

IV. SKILL COLLECTION AND FEATURE CONSTRUCTION

This section presents skill collection and feature vector construction through specific environments and task examples. It first introduces the environment, then describes adversarial and cooperative tasks—highlighting the proposed method’s ability to handle unrelated tasks—and finally outlines the MARL algorithm and feature vector construction.

A. Environment Description

The environment is shown in Fig. 3, where the robot is represented as a disk. Robots of the same/different colors represent cooperative/non-cooperative relationships. Taking the green side as an example, consider N homogeneous robots, defined as $\mathcal{A} = \{1, \dots, N\}$. Each robot’s state is $\mathbf{x} = [\mathbf{p}^T, \mathbf{v}^T]^T$, where, \mathbf{p} and \mathbf{v} are the position and velocity, respectively. The robot’s movement is driven by the active and passive force (see Fig. 3(a)(b)). The active force, \mathbf{f}_a , is a self-generated force, which is the output of the actor network. The passive force, \mathbf{f}_b , is an elastic force following Hooke’s Law. Thus, the robot’s dynamics is $\dot{\mathbf{p}}_i = \mathbf{v}_i$, $\dot{\mathbf{v}}_i = (\mathbf{f}_a + \mathbf{f}_b)/m_i$, $i \in \mathcal{A}$, where m_i is the mass of robot i , and the velocity is bounded by $v_{\min} \leq \|\mathbf{v}_i\| \leq v_{\max}$.

The robot’s perception range is a circular area with radius r_{perc} . In the observation model, perception depends on both Euclidean and topological distances [31], where the topological component limits the number of perceivable individuals. We combine these factors and set a threshold of $n_h = 6$ [31].

All robots move within a bounded region of length L . We consider two types of boundaries: a fixed boundary (see Fig. 3(c)), where robots cannot pass through and will be repelled upon contact, and a periodic boundary (see Fig. 3(d)), where robots can exit the environment from one side and reappear on the opposite side at the same velocity.

B. Adversarial Task

The adversarial task involves two groups of robots, where each side aims to defeat the other. The task ends when one group is eliminated. The adversarial task setup mirrors many biological group behaviors [32], with the attack process illustrated in Fig. 3(e). From the green side’s perspective, a red robot is considered attacked if the following conditions are met: 1) the green robot is behind the red robot (the light red area); 2) the green robot’s velocity is approximately directed towards the red robot (the light green area); and 3) the distance between them is less than the attack radius r_{att} .

In our setup, each robot has a health value hp . When a robot is attacked by an enemy (non-cooperating robot), its hp decreases by Δh , with a maximum of n_o enemies attacking simultaneously. Additionally, if a robot is not under attack, its hp increases by $0.1\Delta h$ up to the initial health value hp_{\max} . A robot is considered dead when $hp \leq 0$.

1) *Reward*: The reward can be expressed as $r_a = r_{\text{surv}} + r_{\text{situ}}$, with default values $r_{\text{surv}} = r_{\text{situ}} = 0$. Here, r_{surv} represents the survival reward. If a robot eliminates an enemy, $r_{\text{surv}} = k_{\text{surv}}$; if it is eliminated, $r_{\text{surv}} = -k_{\text{surv}}$. r_{situ} represents the situational reward if a robot meets the conditions to attack an enemy, $r_{\text{situ}} = k_{\text{situ}}$. Both k_{surv} and k_{situ} are positive constants.

2) *Action*: The action is a two-dimensional vector with components along the x and y axes. This vector indicates the active force \mathbf{f}_a exerted on the robot.

3) *Observation*: Based on the setup, each robot’s observation vector is designed to include its own state and the relative states of teammates and enemies. We set the maximum number of teammates and enemies to $n_h/2$. For example, the robot i ’s observation is $\mathbf{o}_i = [\mathbf{x}_i^T, \mathbf{x}_{ji,1}^T, \dots, \mathbf{x}_{ji,n_h/2}^T, \mathbf{x}_{ki,1}^T, \dots, \mathbf{x}_{ki,n_h/2}^T]^T$, where, $\mathbf{x}_{ji} = \mathbf{x}_j - \mathbf{x}_i$, $j \in \mathcal{N}_i$, $\mathbf{x}_{ki} = \mathbf{x}_k - \mathbf{x}_i$, $k \in \mathcal{O}_i$, $\mathcal{N}_i/\mathcal{O}_i$ are the sets of robot i ’s teammates/enemies.

C. Cooperative Task

The cooperative task involves a single group of robots, for which we select flocking—a behavior widely observed in nature [32], [33]. Flocking is typically governed by three heuristic rules: *cohesion*, *separation*, and *alignment*. A robot is attracted to or repelled by neighbors when their distance is too large or too small, respectively, and its velocity tends to align with the average of its neighbors. Ultimately, the inter-robot distance stabilizes at a certain value.

1) *Reward*: Similar to the three flocking rules, we designed three reward criteria to achieve the same effect. $r_{\text{attr}} = -k_{\text{attr}}(d_{ij} - d_{\text{ref}})$ if $d_{ij} > d_{\text{ref}}$, and 0 otherwise. $r_{\text{repl}} = -k_{\text{repl}}(d_{\text{ref}} - d_{ij})$ if $d_{ij} < d_{\text{ref}}$, and 0 otherwise. $r_{\text{alig}} = -k_{\text{alig}} \left\| \frac{1}{N_i} \sum_{j \in \mathcal{N}_i} \left(\frac{\mathbf{v}_j}{\|\mathbf{v}_j\|} - \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|} \right) \right\|$, where k_{attr} , k_{repl} , k_{alig} are positive constants, $d_{ij} = \|\mathbf{p}_i - \mathbf{p}_j\|$ is the distance between i and j , d_{ref} is the final stable value, and N_i is the number of robot i 's neighbors. Therefore, the total reward is expressed as $r_f = r_{\text{attr}} + r_{\text{repl}} + r_{\text{alig}}$.

2) *Action*: Flocking involves only the robot's movement; therefore, the action is the same as in the adversarial task.

3) *Observation*: Each robot's observation vector is designed to include its own state and the relative states of its neighbors. For example, the observation of robot i is denoted as $\mathbf{o}_i = [\mathbf{x}_i^T, \mathbf{x}_{ji,1}^T, \dots, \mathbf{x}_{ji,n_h}^T]^T$.

D. MARL Algorithm

This paper uses a local-critic version of MADDPG (local-MADDPG) to train the low-level skills. MADDPG employs an actor-critic architecture [27]. In this paper, both the actor and critic are MLPs, with Leaky-ReLU in hidden layers, Tanh for the actor's output, and no activation for the critic's output. As the robots are homogeneous, a single actor and critic can be shared among cooperative robots. Specifically, the adversarial task involves two critics and actors, while flocking involves one of each.

The centralized critic in [27] is $Q(\mathbf{x}_1, \dots, \mathbf{x}_N, \mathbf{a}_1, \dots, \mathbf{a}_N)$, where, $\mathbf{a}_i = \mu(\mathbf{o}_i), i \in \mathcal{A}$ and μ represents the actor. We modified it to $Q(\mathbf{o}_i, \mathbf{a}_i)$; thus, the critic only inputs the robot i 's observation and action. This mimics the local perception in biological swarms, while the local critic enables efficient handling of large-scale tasks. For more details on MADDPG, please refer to [27].

E. Feature Construction

In this paper, feature vectors serve as numerical representations of entities (*i.e.*, environments or tasks), capturing their attributes. An environment can be represented as $e = (y, L)$, where y indicates the boundary type, with $y = 0$ representing a periodic boundary and $y = 1$ representing a fixed boundary. An adversarial task can be represented as $t = (v_{\max}, v_{\min}, \Delta h, n_o, r_{\text{atta}})$, and a flocking task can be represented as $t = (v_{\max}, v_{\min}, d_{\text{ref}}, r_{\text{perc}})$. Once these feature vectors are constructed, they are encoded as representation vectors through environment or task encoders. By combining the skill and relation embedding layers, the skill graph can then be constructed.

V. SIMULATION EXPERIMENTS

This section delineates three components: the collection of low-level skills, the application of the skill graph, and a direct comparative analysis between the skill graph and HRL. The analysis aims to elucidate the advantages of integrating hierarchical structures with the skill graph, thereby contrasting this approach with the standard HRL approach. Moreover, HRL adopts the latest MAPPO algorithm [34]. We carefully

considered alternative baselines, but most existing methods are not directly applicable to the unrelated tasks in our study. Thus, they do not offer a fair basis for comparison.

A. Low-level Skill Collection

1) *Experimental setup*: The number of robots for both teams in the adversarial task and for the flocking task is set to $N = 50$. In the adversarial task, the red team adopts an auxiliary strategy $\mathbf{f}_a = k_p(\mathbf{p}_c - \mathbf{p}) + k_v(\mathbf{v}_c - \mathbf{v})$, where \mathbf{p}_c and \mathbf{v}_c are the position and velocity of the nearest green robot. The green team adopts the local-MADDPG strategy. In the flocking task, to prevent swarm splitting due to local perception, several robots are designated as leaders with fixed movements (see Fig. 4(d)(f)). These leaders are otherwise identical to regular agents. The task parameters are set as follows: $r_{\text{perc}} = 3$ m, $m_i = 1$ kg, $v_{\max} = 1$ m/s, $v_{\min} = 0$, $L = 4 \sim 6$ m, $\Delta h = 1$, $h_{\text{pmax}} = 80$, $n_o = 3$, $r_{\text{atta}} = 0.3$ m, $d_{\text{ref}} = 0.4 \sim 0.7$ m, $k_{\text{situ}} = 1$, $k_{\text{surv}} = 5$, $k_{\text{attr}} = 1 \sim 2.5$, $k_{\text{repl}} = 15$, $k_{\text{alig}} = 2$. Hyperparameters for local-MADDPG are listed in Table I.

2) *Results analysis*: The experimental results are shown in Fig. 4. In the adversarial scenario, the green team's performance depends on whether its local-MADDPG policy is trained. When untrained, the red team—guided by an auxiliary strategy—successfully maneuvers behind and eliminates it (see Fig. 4(a)). Conversely, a trained green team can outmaneuver and defeat the red team. (see Fig. 4(b)). In the flocking task (see Fig. 4(c)-(f)), the robot team maintains nearly equal inter-agent distances once a stable formation is reached. Notably, under periodic boundary conditions, Reynolds' three rules are well preserved. These results demonstrate that the local-MADDPG algorithm successfully collects low-level skills.

B. Skill Graph Application

1) *Experimental setup*: Suppose we have 8 flocking subtasks with feature vectors as follows: floc_1: (1,0,0.4,2), floc_2: (1,0,0.8,2), floc_3: (1,0,0.4,3), floc_4: (1,0,0.8,3), floc_5: (1,0,0.4,4), floc_6: (1,0,0.8,4), floc_7: (1,0,0.4,5), floc_8: (1,0,0.8,5). Additionally, there are 8 adversarial subtasks with feature vectors: adve_1: (1,0,1,2,0.3), adve_2: (1,0,1,3,0.3), adve_3: (1,0,1,4,0.3), adve_4: (1,0,1,5,0.3), adve_5: (1,0,1,2,0.4), adve_6: (1,0,1,3,0.4), adve_7: (1,0,1,4,0.4), adve_8: (1,0,1,5,0.4). Moreover, there are 2 environments: fixed: (1,6) and periodic: (0,6). This yields a total of 32 skills (*i.e.*, $(8+8) \times 2 = 32$). Using these tasks, environments, and skills, we can construct samples and subsequently construct a skill graph. The training parameters for the skill graph are set as follows: representation vector size = 96, hid size = 256, hid layers = 3, $M = 500$, batch size = 256, $k_{1,t} = 0$, $k_{2,t} = 0$, $k_{3,t} = 0$, $k_{4,t} = 3$, $k_{5,t} = 1$, $k_{1,e} = 0.95$, $k_{2,e} = 0.05$, $\lambda = 3$, $\alpha_{\text{high}} = 0.95$, $\alpha_{\text{low}} = 0.85$.

To illustrate the application of the skill graph, we designed a scenario as shown in Fig. 5(a)-(c) where two groups of 50 robots each move along predetermined paths. When the groups do not encounter each other, they perform a flocking task with $d_{\text{ref}} = 0.4$ m (stage 1). Upon encountering—defined as a robot from one group coming within r_{perc} of the opposing group's center—they engage in an adversarial task (stage 2). After one

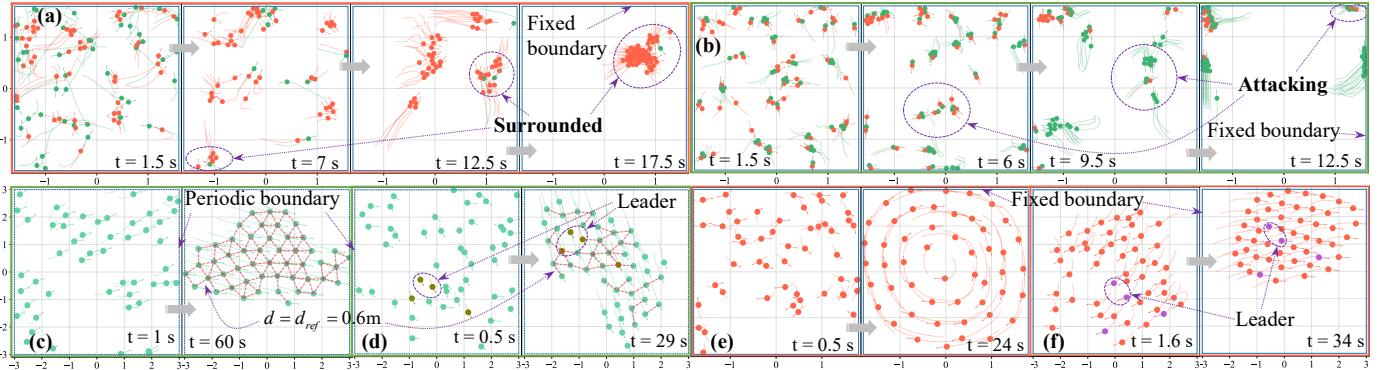


Fig. 4: Low-level skill collection. (a) and (b) illustrate an adversarial task where the green team uses local-MADDPG, and the red team employs an auxiliary strategy with $k_p = 1$ and $k_v = 2$. In (a), local-MADDPG is untrained, leading to the green team’s defeat. In (b), after training, the green team prevails over the red team. (c)(d), (e)(f) illustrate the flocking behavior with local-MADDPG strategy under periodic and fixed boundaries, respectively.

TABLE I: Hyperparameters for local-MADDPG and MAPPO

local-MADDPG		MAPPO [34]	
Adve/Floc	Adve/Floc	Adve/Floc	Adve/Floc
eps	1200/1800 critic lr	1e-3/1e-3	- 2400/2000 -
eps len	200/200 actor lr	1e-4/1e-4	- 350/200 -
buf size	5e5/5e5 gamma	0.99/0.99	- 1e5/1e5 -
batch	512/512 explr rate 0.1/0.1	- 2048/2048 ppo epoch 20/20	
hid size	128/64 noise 0.8/0.8	- 256/128 clip para 0.1/0.2	
hid layers 2/3	soft-upd 0.01	- 3/3 gae gamm 0.95/0.95	

group is eliminated, the remaining group resumes the flocking task (stage 3) with a different $d_{\text{ref}} = 0.6$ m. In such a scenario, the skill graph needs to infer (decide) the most suitable skill for the robot team in each stage.

2) *Results analysis:* The skill graph’s application is illustrated in Fig. 5. In stage 1 (see Fig. 5(a)), both robot teams perform the flocking task with $d_{\text{ref}} = 0.4$ m. The input environment + task query is $(1,6) + (1,0,0.4,3)$, and the output (see Fig. 5(d)) shows that the skill “floc_3_fixed” achieves the highest score (0.97), exceeding the threshold $\alpha_{\text{high}} = 0.95$, and is therefore selected for stage 1. Similarly, the query is $(1,6) + (1,0,1,3,0.3)$ in stage 2, and the highest-scoring skill “adve_2_fixed” is chosen (see Fig. 5(e)). In stage 3, the query is $(1,6) + (1,0,0.6,3)$. Although this query does not match any fundamental skill in the graph, the graph provides the two most relevant skills, “floc_4_fixed” and “floc_3_fixed” (see Fig. 5(f)). Then, the robot’s action is a weighted combination of these two skills. This process clearly demonstrates the skill graph’s selection and combination capabilities.

Fig. 5(g)-(j) further demonstrates how the skill graph handles unfamiliar queries—not entirely unseen, but rather cases that may resemble previously encountered ones. For instance, when the query input is $(1,6) + (1,0,1,3)$, the output’s highest score is below α_{low} (see Fig. 5(g)). In this case, the highest-scoring skill, “floc_4_fixed”, is further trained and used for the query pair. This approach results in a skill that requires less training time and higher sampling efficiency compared to training from scratch (see Fig. 5(h)), enabling robots to quickly adapt to unfamiliar environments and tasks. In summary, the skill graph demonstrates excellent knowledge transfer capabilities.

C. Comparative Experiments

1) *Experimental setup:* The setup of the skill graph is described in Section V-B; here, we focus on the HRL setup. It consists of two layers, both using MAPPO. The low-level MAPPO shares the same observation, action, and reward structure as in Section IV, with hyperparameters listed in Table I. The high-level MAPPO is set as follows. *Observation:* For the green side, the observation vector is designed as $\mathbf{o}_g = [\mathbf{x}_{rg}^T, n_g, n_r]^T$, where, $\mathbf{x}_{rg} = \mathbf{x}_r - \mathbf{x}_g$ represents the relative position between team centers, and n_g and n_r denote the number of robots on the green and red teams, respectively. *Action:* The action space is discrete, matching the number of available low-level skills; the actor outputs the index of the selected skill. *Reward:* A -1 reward is given if the selected skill does not match the expected one. All other high-level MAPPO hyperparameters follow those used in the flocking task (Table I), except for the number of episodes (set to 200) and the episode length (set to 300).

To implement the comparison, we define two metrics. The first is the *decision success rate*, denoted as $\rho_{\text{succ}} = n_{\text{succ}} / n_{\text{total}}$, where, n_{succ} and n_{total} represent the numbers of successful and total decisions, respectively. we aim for the method to achieve as high a success rate as possible. The second is *generalization* which is a qualitative metric. During testing, we will vary certain parameters and observe changes in ρ_{succ} . Ideally, an effective method should exhibit stable performance despite these parameter variations.

It is crucial to clarify that our comparison involves two schemes: Scheme 1 (a high-level skill graph + a low-level local-MADDPG) and Scheme 2 (a high-level MAPPO + a low-level MAPPO). Despite differences in low-level algorithms, this comparison validly demonstrates the advantages of the skill graph for two reasons: First, the defined metrics do not depend on the low-level algorithm and exclusively evaluate high-level performance. Second, both frameworks are modular; although replacing Scheme 2’s low-level component with the local-MADDPG is feasible, such a hybrid configuration would be less natural.

2) *Results analysis:* Table II summarizes detailed ρ_{succ} metric, where “dim” indicates the number of low-level skills, and “base” represents the case with identical training and

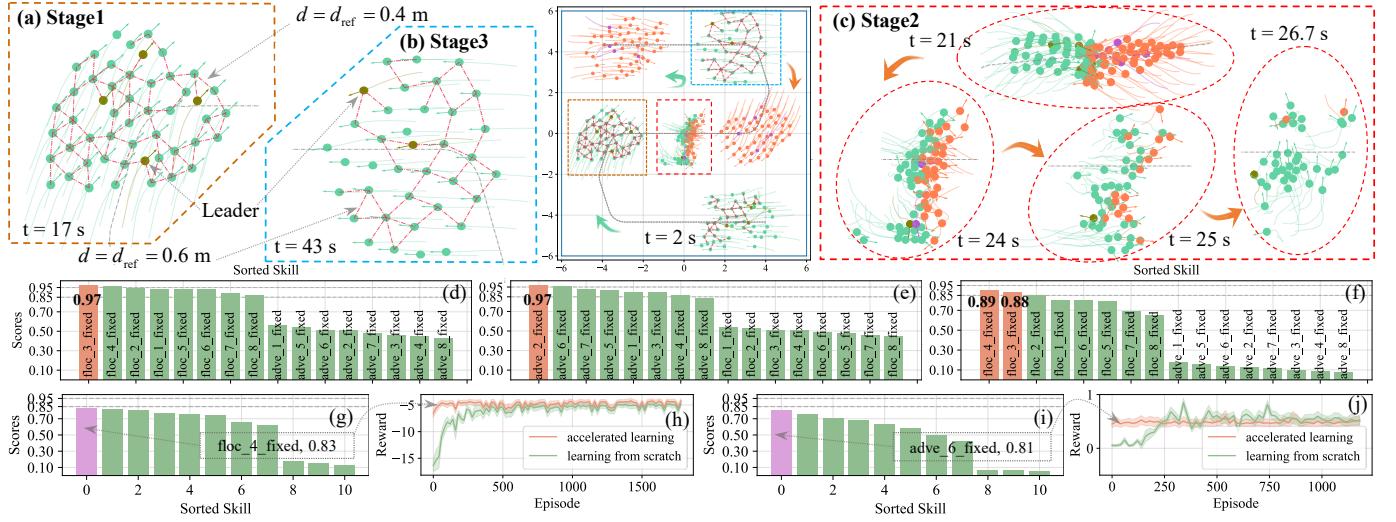


Fig. 5: Skill graph application. In (a)(b), the red lines on the green side indicate that the inter-robot distance is approximately d_{ref} . In (c), the green side uses local-MADDPG strategy, while the red side employs an auxiliary strategy with $k_p = 1$ and $k_v = 2$. (d)(e)(f) show the top 16 skill scores for the three stages, with red highlighting the skills that can be directly reused. (g) displays the top 11 skills for the query pair $(1,6)+(1,0,1,3)$, while (h) illustrates the reward curves for further training of the highest-scoring skills versus training from scratch. Similarly, (i)(j) correspond to $(1,6)+(1,0,1,3,0,8)$.

TABLE II: ρ_{succ} of high-level MAPPO

case	base	init num*1.1	init num*0.9	leader v*1.4	leader v*0.6
dim=2	95.2(6.8)%	93.2(8.2)%	93.9(6.9)%	93.8(7.9)%	95.1(6.3)%
dim=4	94.6(7.4)%	93.1(8.2)%	90.9(9.1)%	93.2(8.7)%	94.3(7.7)%
dim=6	94.4(7.2)%	93.0(8.3)%	89.7(8.0)%	92.9(8.5)%	94.1(7.1)%
dim=8	93.7(7.7)%	91.4(8.8)%	88.3(8.8)%	92.6(8.7)%	93.1(8.4)%
dim=10	93.2(7.5)%	90.1(7.8)%	88.0(9.2)%	92.1(8.6)%	93.0(7.7)%
dim=12	92.5(7.6)%	90.7(7.9)%	86.8(9.5)%	90.7(8.8)%	92.1(8.0)%

testing parameters. Cols 2-5 (different training and testing parameters) represent variations in the initial number (“init num”) and leader speed (“leader v”), which aim to assess the generalization ability of the MAPPO strategy. The table demonstrates two key findings: First, ρ_{succ} decreases as the number of available skills increases. Second, ρ_{succ} is sensitive to parameter changes, indicating poor generalization.

Many parameters can highlight hierarchical MAPPO’s weak generalization, but we report only two due to space limits. Its underperformance stems from two reasons. First, the high-level strategy depends on low-level skills during training, creating a tight coupling where many variations in low-level parameters can impact high-level decision-making. Second, the high-level strategy only selects or rejects low-level skills, which is too rigid and overlooks the inherent rationality of those skills. In contrast, the skill graph has three advantages. First, its construction is independent of low-level skills, ensuring that changes in low-level parameters do not affect it, leading to good generalization. Second, the skill graph selects skills based on their scores without requiring a perfect score, (*i.e.* 1), resulting in $\rho_{\text{succ}} = 100\%$. Third, the skill graph can combine or further optimize high-scoring skills, a capability that hierarchical MAPPO lacks, *i.e.* the skill graph exhibits better knowledge transfer than hierarchical MAPPO.

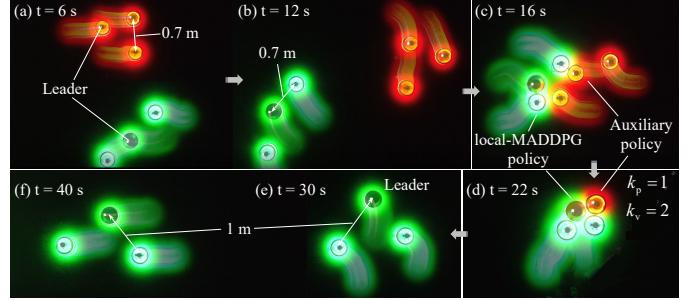


Fig. 6: Snapshots of the real-world experiment. (a)(b), (c)(d), and (e)(f) correspond to stage 1, stage 2, and stage 3, respectively.

VI. REAL-WORLD EXPERIMENTS

The real experimental setup is shown in Fig. 1 and Fig. 6, where six omnibot robots [35] are divided into red and green teams, moving within a 5×5 m area. The experiment follows the procedure illustrated in Fig. 5(a)-(c). The robot structure is depicted in Fig. 1, and its state is tracked using a motion capture system. The robot skill library is mostly consistent with that mentioned in Section V-B, with three differences. First, the flocking’s d_{ref} is adjusted from 0.4/0.8 m to 0.7/1.3 m. Second, the boundary length is reduced from 6 m to 5 m. Third, training episodes for flocking are increased from 1200 to 3000. We set d_{ref} to 0.7 m for stage 1 and 1 m for stage 3. All other parameters match those in the Simulation.

In stages 1, 2, and 3, the query pairs are $(1,5) + (1,0,0.7,3)$, $(1,5) + (1,0,1,3,0,3)$, and $(1,5) + (1,0,1,3)$, respectively, and the skill graph outputs skills “floc_3_fixed”, “adve_2_fixed”, and “floc_4_fixed” and “floc_3_fixed” for robot team. As shown in Fig. 6, both the red and green teams first form flocking with $d_{\text{ref}} = 0.7$ m, then enter the adversarial phase, where the green team defeats the red team. Finally, the green team transitions to a flocking with $d_{\text{ref}} = 1$ m. The entire process demonstrates the practical feasibility of the method presented in this paper.

VII. CONCLUSION

This paper proposed an effective hierarchical approach for MT-MARL, with a skill graph at the upper layer. The skill graph demonstrated three advantages: it did not require related attributes among low-level tasks, offered superior knowledge transfer compared to standard hierarchical methods, and was trained independently of low-level skills with good generalization. Simulations and real-world experiments were conducted to validate the advantages of the proposed approach.

It is worth noting that the fact that the high-level skill graph does not require lower-level tasks to be related does not imply that it can transfer knowledge in entirely unseen scenarios—these are two separate issues. When the latter occurs, the scores of all skills in the library tend to be low due to the absence of transferable knowledge, and new skills must be learned from scratch. This also represents a potential direction for future research.

REFERENCES

- [1] O. Vinyals, *et al.*, “Grandmaster level in starcraft ii using multi-agent reinforcement learning,” *nature*, vol. 575, no. 7782, pp. 350–354, 2019.
- [2] Z.-J. Pang, R.-Z. Liu, Z.-Y. Meng, Y. Zhang, Y. Yu, and T. Lu, “On reinforcement learning for full-length game of starcraft,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 4691–4698.
- [3] T. Fan, P. Long, W. Liu, and J. Pan, “Distributed multi-robot collision avoidance via deep reinforcement learning for navigation in complex scenarios,” *The International Journal of Robotics Research*, vol. 39, no. 7, pp. 856–892, 2020.
- [4] C. De Souza, R. Newbury, A. Cosgun, P. Castillo, B. Vidolov, and D. Kulić, “Decentralized multi-agent pursuit using deep reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4552–4559, 2021.
- [5] D. Johnson, G. Chen, and Y. Lu, “Multi-agent reinforcement learning for real-time dynamic production scheduling in a robot assembly cell,” *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 7684–7691, 2022.
- [6] F. Zhang, C. Jia, Y.-C. Li, L. Yuan, Y. Yu, and Z. Zhang, “Discovering generalizable multi-agent coordination skills from multi-task offline data,” in *Proceedings of the International Conference on Learning Representations*, 2023, pp. 1–24.
- [7] Y. Zhang and Q. Yang, “A survey on multi-task learning,” *IEEE transactions on knowledge and data engineering*, vol. 34, no. 12, pp. 5586–5609, 2021.
- [8] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, “Deep decentralized multi-task multi-agent reinforcement learning under partial observability,” in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 2681–2690.
- [9] M. Hessel, H. Soyer, L. Espeholt, W. Czarnecki, S. Schmitt, and H. Van Hasselt, “Multi-task deep reinforcement learning with popart,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019, pp. 3796–3803.
- [10] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, “Conflict-averse gradient descent for multi-task learning,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2021, pp. 18878–18890.
- [11] C. Tessler, S. Givony, T. Zahavy, D. Mankowitz, and S. Mannor, “A deep hierarchical approach to lifelong learning in minecraft,” in *Proceedings of the AAAI conference on Artificial Intelligence*, 2017, pp. 1553–1561.
- [12] B. Liu, L. Wang, and M. Liu, “Lifelong federated reinforcement learning: A learning architecture for navigation in cloud robotic systems,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4555–4562, 2019.
- [13] E. Parisotto, J. L. Ba, and R. Salakhutdinov, “Actor-mimic: Deep multitask and transfer reinforcement learning,” in *Proceedings of the International Conference on Learning Representations*, 2016, pp. 1–16.
- [14] A. A. Rusu, *et al.*, “Policy distillation,” *arXiv preprint arXiv:1511.06295*, 2015.
- [15] Z. Xu, K. Wu, Z. Che, J. Tang, and J. Ye, “Knowledge transfer in multi-task deep reinforcement learning for continuous control,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2020, pp. 15146–15155.
- [16] L. Sun, H. Zhang, W. Xu, and M. Tomizuka, “PaCo: Parameter-compositional multi-task reinforcement learning,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2022, pp. 21495–21507.
- [17] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, “Hierarchical reinforcement learning: A comprehensive survey,” *ACM Computing Surveys (CSUR)*, vol. 54, no. 5, pp. 1–35, 2021.
- [18] R. Gieselmann and F. T. Pokorný, “Planning-augmented hierarchical reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5097–5104, 2021.
- [19] S. Christen, L. Jendele, E. Aksan, and O. Hilliges, “Learning functionally decomposed hierarchies for continuous control tasks with path planning,” *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 3623–3630, 2021.
- [20] J. Andreas, D. Klein, and S. Levine, “Modular multitask reinforcement learning with policy sketches,” in *Proceedings of the International Conference on Machine Learning*, 2017, pp. 166–175.
- [21] K. Lee, S. Kim, and J. Choi, “Adaptive and explainable deployment of navigation skills via hierarchical deep reinforcement learning,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023, pp. 1673–1679.
- [22] R. Wang, D. Zhao, A. Gupta, and B.-C. Min, “Initial task allocation in multi-human multi-robot teams: An attention-enhanced hierarchical reinforcement learning approach,” *IEEE Robotics and Automation Letters*, vol. 9, no. 4, pp. 3451–3458, 2024.
- [23] K. Zhang, Z. Yang, and T. Başar, “Multi-agent reinforcement learning: A selective overview of theories and algorithms,” *Handbook of reinforcement learning and control*, pp. 321–384, 2021.
- [24] H. Zhang, *et al.*, “RSG: Fast learning adaptive skills for quadruped robots by skill graph,” *arXiv preprint arXiv:2311.06015*, 2023.
- [25] S. Ji, S. Pan, E. Cambria, P. Marttinen, and S. Y. Philip, “A survey on knowledge graphs: Representation, acquisition, and applications,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 2, pp. 494–514, 2021.
- [26] Z. Wang, J. Zhang, J. Feng, and Z. Chen, “Knowledge graph embedding by translating on hyperplanes,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2014, pp. 1112–1119.
- [27] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. Pieter Abbeel, and I. Mordatch, “Multi-agent actor-critic for mixed cooperative-competitive environments,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2017, pp. 6379–6390.
- [28] Q. Wang, Z. Mao, B. Wang, and L. Guo, “Knowledge graph embedding: A survey of approaches and applications,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 29, no. 12, pp. 2724–2743, 2017.
- [29] J. Xu, M. He, and Y. Jiang, “A novel framework of knowledge transfer system for construction projects based on knowledge graph and transfer learning,” *Expert Systems with Applications*, vol. 199, p. 116964, 2022.
- [30] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, “Translating embeddings for modeling multi-relational data,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2013, pp. 2787–2795.
- [31] J. Li, L. Li, and S. Zhao, “Predator-prey survival pressure is sufficient to evolve swarming behaviors,” *New Journal of Physics*, vol. 25, no. 9, p. 092001, 2023.
- [32] M. Dorigo, G. Theraulaz, and V. Trianni, “Swarm robotics: Past, present, and future [point of view],” *Proceedings of the IEEE*, vol. 109, no. 7, pp. 1152–1165, 2021.
- [33] G. Vásárhelyi, C. Virág, G. Somorjai, T. Nepusz, A. E. Eiben, and T. Vicsek, “Optimized flocking of autonomous drones in confined environments,” *Science Robotics*, vol. 3, no. 20, p. eaat3536, 2018.
- [34] C. Yu, *et al.*, “The surprising effectiveness of PPO in cooperative multi-agent games,” in *Proceedings of the Advances in Neural Information Processing Systems*, 2022, pp. 24611–24624.
- [35] Z. Ma, *et al.*, “Omnibot: A scalable vision-based robot swarm platform,” in *Proceedings of the IEEE International Conference on Control & Automation (ICCA)*, 2024, pp. 975–980.