Safe Learning for Multi-Robot Mapless Exploration

Wenxing Liu[1], *Member, IEEE*, Hanlin Niu[1], *Member, IEEE*, Kefan Wu[2], Wei Pan[3], *Member, IEEE*, Ze Ji[4], *Member, IEEE* and Robert Skilton[1], *Member, IEEE*

***Abstract*—When using deep reinforcement learning (DRL) to perform multi-robot exploration in unknown environments, the training model may produce actions that lead to unpredictable system behaviors due to the complexity and unpredictability of the surroundings. Therefore, ensuring safe exploration with DRL becomes critical. To tackle this issue, we propose a multi-agent framework that utilizes the formation scheme based on intermediate estimator compensation (IEC) to address the uncertainties introduced by DRL to ensure safe exploration. The convergence of the proposed scheme is verified via the Lyapunov method in the presence of tracking errors. An actor-critic-based DRL method is proposed for each mobile robot to deal with collision avoidance tasks. To enhance the efficiency of obtaining the DRL training model, a consensus-based training policy is introduced. The proposed safe learning framework successfully addresses uncertainties introduced by DRL while ensuring mapless exploration in both simulations and real-world experiments. The experimental video is available at: https://youtu.be/ce99FxKFFtY, and the code can be accessed at: https://github.com/ukaea/SLMRME.**

***Index Terms*—Multi-robot exploration, Deep reinforcement learning, Intermediate estimator compensation, Formation, Collision avoidance, Lyapunov methods.**

I. Introduction

Deep reinforcement learning (DRL) has attracted widespread attention since it can be applied to various tasks that are too complex for traditional methods to achieve [1], [2]. DRL allows agents to learn optimal behaviors autonomously in environments where explicit programming or rules-based approaches are impractical or infeasible [3]. One of the key strengths of DRL is its ability to learn optimal policies directly from raw sensory data, eliminating the need for explicit system modeling [4], [5]. Additionally, DRL is able to handle high-dimensional and continuous state-action spaces, making it suitable for real-world applications that require precise and adaptive decision-making [6]. Many works [7], [8] on DRL have introduced adaptive and intelligent decision-making for multi-agent systems. Unlike traditional rule-based or optimization-based approaches, DRL enables agents to learn from experience and generalize to complex, dynamic settings [9], [10]. However, DRL may introduce uncertainties when applied in unknown environments, as the training model may produce actions that lead to unpredictable system behaviors. These uncertainties pose significant challenges, particularly in the context of multi-robot efficient and safe exploration, where collision avoidance and cooperative behavior are paramount. To address these challenges, we propose an IEC-based formation scheme to tackle uncertainties introduced by DRL, enabling multiple mobile robots to conduct safe exploration tasks autonomously without relying on pre-existing maps.

In this paper, we aim to design a safe learning cooperative framework for multi-robot exploration that addresses uncertainties introduced by DRL while ensuring collision avoidance, as illustrated in Fig. 1. The proposed framework demonstrates a robust and scalable solution that enhances the safety and effectiveness of multi-robot exploration, making DRL approaches more feasible for practical deployment in unknown environments. The main contributions of the paper can be summarized as follows:

1) An actor-critic-based DRL method is presented for each mobile robot to address collision avoidance tasks. To enhance the training efficiency of the DRL method, a multi-robot consensus-based training policy is developed. This policy reduces the number of required training steps while maintaining the same level of training reward.
2) An IEC-based formation scheme is proposed to handle uncertainties introduced by the DRL method to ensure safe exploration. The convergence of the proposed scheme is verified via the Lyapunov method in the presence of tracking errors.
3) The proposed framework has been successfully implemented in both simulated and real-world multi-agent mapless safe exploration scenarios with wheeled mobile robots. Under the proposed framework, each robot effectively preserves the formation shape for safe exploration, despite the uncertainties introduced by DRL.

The structure of this paper is organized as follows: Section II reviews the related work. Section III presents the preliminaries. Section IV elucidates the convergence analysis of the proposed cooperative framework. Simulations and real-world experiments are provided in Section V to demonstrate the feasibility of the proposed cooperative framework. Section VI concludes this paper.

This work has been funded by the EPSRC Energy Programme [grant number EP/W006839/1]. To obtain further information on the data and models underlying this paper, please contact PublicationsManager@ukaea.uk. For the purpose of open access, the author(s) has applied a Creative Commons Attribution (CC BY) licence (where permitted by UKRI, 'Open Government Licence' or 'Creative Commons Attribution No-derivatives (CC BY-ND) licence' may be stated instead) to any Author Accepted Manuscript version arising.

[1]Wenxing Liu, Hanlin Niu and Robert Skilton are with Remote Applications in Challenging Environments (RACE), United Kingdom Atomic Energy Authority, Culham, UK. *(Corresponding author: Hanlin Niu. hanlin.niu@ukaea.uk).*

[2]Kefan Wu is with the Lakeside Labs GmbH, Klagenfurt, Austria.

[3]Wei Pan is with the Department of Computer Science, The University of Manchester, Manchester, UK.

[4]Ze Ji is with the School of Engineering, Cardiff University, Cardiff, UK.

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3581872

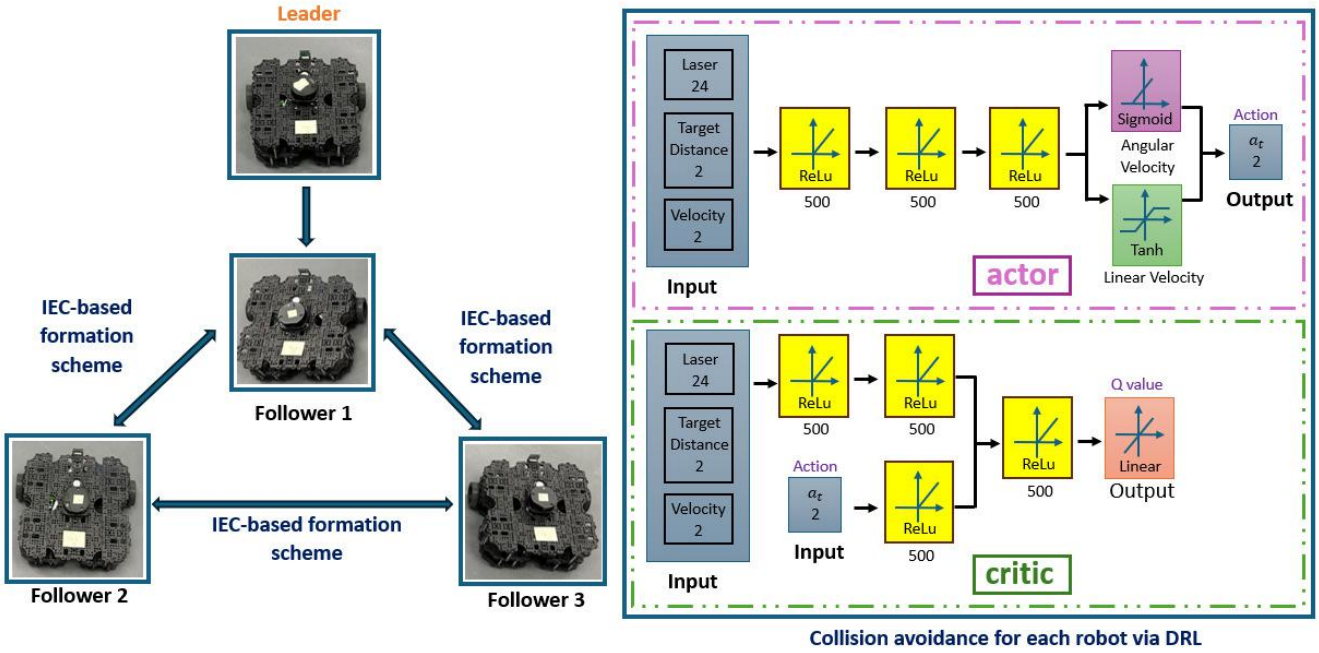


Fig. 1. A safe learning cooperative framework for multi-robot exploration. For each mobile robot, an actor-critic-based DRL method is presented to deal with collision avoidance tasks. An IEC-based formation scheme is proposed to tackle uncertainties introduced by the DRL method to ensure safe exploration.

II. Related Work

In this section, we review the work related to the proposed framework. The first part discusses safe learning in reinforcement learning, while the second part focuses on collision avoidance in multi-agent systems.

Safe Learning in Reinforcement Learning: Recently, there has been an increased focus on safe learning when applying the reinforcement learning approach [11], [12]. The aim is to develop policies that achieve high performance in terms of task objectives while satisfying safety constraints, which ensures that the agent operates reliably and avoids failures in complex environments [13], [14]. A model-based reinforcement learning algorithm that ensures safety through Gaussian process models was proposed in [15] to enable provably stable policy optimization without risking safety-critical failures. A learning-based model predictive control approach with high-probability safety guarantees was proposed in [16] to ensure safe trajectory constraints during reinforcement learning. In [17], a reinforcement learning architecture that enforces safety constraints by correcting unsafe actions from a neural network was introduced to enable safe policy execution in real-world robotic tasks. A model-based reinforcement learning algorithm that penalizes unsafe trajectories and provides guarantees for avoiding unsafe states was proposed in [18] to obtain decent rewards with fewer violations in safety. However, an essential foundation in the aforementioned works is their model-based nature, which assumes that the structure of the constraints or the system dynamics is explicitly known. This could present a limitation when addressing model-free problems.

Model-free safe reinforcement learning algorithms [19], [20] have demonstrated significant success in systems with continuous state spaces and action spaces. A reinforcement learning algorithm that separates task optimization and safety constraint satisfaction through distinct policies was proposed in [21] to preemptively identify unsafe zones. A policy search algorithm called constrained policy optimization was introduced in [22] to enable safe and effective policy learning in high-dimensional control tasks. In [23], a multi-timescale method called reward constrained policy optimization was developed to guide policies toward constraint satisfaction. Nevertheless, the aforementioned works primarily focus on addressing specific, task-dependent challenges. While these methods are effective within their designated robotic tasks, they often lack adaptability to diverse real-world scenarios. In contrast, our framework emphasizes high-level tasks by modeling each mobile robot as an autonomous agent, offering greater generality and a broader scope of applications.

Collision Avoidance in Multi-agent Systems: Formation control has been extensively applied in multi-agent systems [24], [25], particularly to facilitate collision avoidance tasks [26]–[28]. For instance, a formation control scheme was proposed in [29] for stochastic second-order multi-agent systems to deal with obstacle avoidance problems under directed topology. In [30], an adaptive formation tracking control protocol with an obstacle avoidance mechanism was introduced to address the obstacle avoidance problem in multi-vehicle systems. By using an artificial potential function, a distributed control algorithm was developed in [31] for obstacle avoidance and formation control of multiple rectangular agents. In [32], an enhanced artificial potential field algorithm, combined with a finite-time consistent formation control algorithm, was used to improve rapid obstacle avoidance control for unmanned aerial vehicle clusters operating in complex environments. A

significant assumption in the aforementioned studies is the requirement for prior knowledge of obstacle positions, which may present a potential limitation in unknown or complex environments.

Learning-based methods [33]–[35] have demonstrated significant advantages in collision avoidance tasks. The application of learning-based methods in these tasks aims to improve decision-making intelligence, enabling the handling of random and unpredictable scenarios [36]. In [37], a deep residual reinforcement learning approach on the basis of the soft actor-critic framework, incorporating the artificial potential field method as a prior controller, was proposed to address the issue of low data utilization arising from the extensive episode experience data. A novel integrated approach was presented in [38] to address motion planning and decision-making for autonomous vehicle lane-change maneuvers. In [39], a local attention-based safety DRL decision-making approach was developed for an ego vehicle to attend to varying social vehicle states in complex traffic environments and effectively manage the influence of surrounding vehicles. However, the aforementioned studies considered only a single robot, which may be viewed as a limitation.

When addressing multi-agent collision avoidance problems, learning-based methods [40]–[42] have been extensively studied, demonstrating promising results in dynamic and unknown environments. A learning-based collision-avoidance policy was introduced in [43] to enable multiple nonholonomic mobile robots to navigate toward their target positions in complex and rich environments. In [44], an improved DRL controller was proposed to handle the problem of decentralized collision avoidance. A learning-based end-to-end framework was developed in [45] to generate a reactive collision avoidance policy aiming at accomplishing efficient distributed multi-agent navigation. In [46], a hybrid algorithm of force-based motion planning and DRL was illustrated to deal with the distributed motion planning problem in dynamic and dense environments. Nevertheless, learning-based methods optimize expected rewards rather than enforcing hard safety constraints, which may lead to uncertainties or unsafe actions. Our cooperative framework incorporates multiple robots and leverages an IEC-based formation scheme with DRL to autonomously perform safe exploration and inspection tasks without relying on pre-existing maps. This approach enhances the safety and effectiveness of mapless exploration.

III. Preliminaries

A. Graph Theory

The interaction topology of a multi-agent system with one leader and P followers can be described by an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. $\mathcal{V}$ stands for a vertex set $\mathcal{V} = \{0, 1, 2, \cdots, P\}$ and an edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$. The edge $(i, j) \in \mathcal{E}$ means that the i^{th} and j^{th} agents can share information with each other. If $(i, j) \in \mathcal{E}$, $a_{ij} > 0$, where a_{ij} stands for the strength of the connection. Let the Adjacency matrix $\mathcal{A} = [a_{ij}] \in \mathbb{R}^{(P+1)\times(P+1)}$ of $\mathcal{G}$ be

$$[a_{ij}] = \begin{cases} a_{ij}, & i \neq j,\ (i, j) \in \mathcal{E}, \\ 0, & i \neq j,\ (i, j) \notin \mathcal{E}. \end{cases} \tag{1}$$

As a result, the Laplacian matrix $\mathcal{L}$ of $\mathcal{G}$ can be defined as $\mathcal{L} = \mathcal{D} - \mathcal{A}$, where $\mathcal{D} = diag\{d_{00}, \cdots, d_{PP}\} \in \mathbb{R}^{(P+1)\times(P+1)}$ and $d_{ii} = \sum_{j=0, j\neq i}^{P} a_{ij}$. The interaction topology among the followers is considered to be undirected. The Laplacian matrix $\mathcal{L}$ with one leader and P followers can be divided into the following structure [47]:

$$\mathcal{L} = \begin{bmatrix} \mathbf{0} & \mathbf{0}_{1\times P} \\ \mathcal{L}_2 & \mathcal{L}_1 \end{bmatrix},$$

where $\mathcal{L}_1 \in \mathbb{R}^{P\times P}$ and $\mathcal{L}_2 \in \mathbb{R}^{P\times 1}$. It can be deduced that $\mathcal{L}_1 > 0$ and $\mathcal{L}\mathbf{1}_{P+1} = 0$, where $\mathbf{1}_{P+1} = [1, \cdots, 1]^T \in \mathbb{R}^{P+1}$.

B. Problem Formulation

Consider the following nonlinear multi-robot system with P follower robots and one leader robot labeled by 0, the system with uncertainties can be modeled as follows:

$$\begin{cases} \dot{x}_i(t) = h(x_i(t)) + u_i(t) + f_i(t), \ \ i = 1, 2 \cdots, P \\ \dot{x}_0(t) = h(x_0(t)) + w_0(t), \end{cases} \tag{2}$$

where $x_i(t) \in \mathbb{R}^m$ is the position of each robot, and $u_i \in \mathbb{R}^m$ denotes control input. $h(\cdot) \in \mathbb{R}^m$ stands for the nonlinear function of the system. $f_i(t) \in \mathbb{R}^m$ and $w_0(t) \in \mathbb{R}^m$ represent the uncertainties in the followers and leader. The uncertainties can be interpreted as those introduced by DRL in the multi-robot system. Denote the target distance between the leader and i^{th} follower as $d_i \in \mathbb{R}^m$ which is set initially, and $d = [d_1^T, \cdots, d_P^T]^T$ which represents the target formation configuration.

To derive the main results of this paper, we make the following assumptions:

Assumption 1: There exists at least one spanning tree in graph $\mathcal{G}$, whose root can be a leader.

Assumption 2: The nonlinear function $h(\cdot)$ satisfies the Lipschitz condition with a Lipschitz constant $\eta > 0$, i.e.,

$$\|h(x_i) - h(x_0)\| \leq \eta \|x_i - x_0\|. \tag{3}$$

Assumption 3: $f_i(t)$ is first-order derivable. Furthermore, $\dot{f}_i(t)$ is bounded and continuous, i.e., there exists a positive constant θ such that

$$\|\dot{f}_i(t)\| \leq \theta \tag{4}$$

Assumption 4: The uncertainty in the leader is bounded and continuous, i.e., there exists a positive constant w such that

$$\|w_0(t)\| \leq w \tag{5}$$

This paper addresses the safe learning problem for multi-robot navigation and exploration. Initially, we proposed an actor-critic DRL algorithm for each robot to avoid the obstacles incorporated into the mission. Uncertainties will be introduced in the system described in (2) if DRL is implemented in robots. Additionally, the tracking error of each robot should be guaranteed in a neighborhood around the origin. Hence, an IEC-based formation protocol is designed to address the uncertainties introduced by the DRL method. To sum up, the main objectives of this work can be demonstrated as follows: i) How to construct the DRL training strategy for each robot to cope with collision avoidance. ii) How to design

an IEC-based formation protocol for the robots to deal with uncertainties introduced by DRL. iii) How to implement the proposed framework in both simulated and real scenarios with wheeled mobile robots.

IV. Methodology

In this section, we introduce a cooperative framework that leverages the IEC-based formation scheme to address uncertainties arising from the DRL process to ensure safe learning. We begin by introducing the DRL approach tailored for each mobile robot. To enhance training efficiency, we employ a consensus-based training method. Following this, we present the IEC-based formation scheme to ensure safe learning. Section E outlines this innovative framework, which enables mapless exploration while ensuring safe learning.

A. DRL Setup

1) Observation Space and Action Space: For every robot, the observation space s is a 28-dimensional vector comprising the angular velocity V_a, linear velocity V_l, the distance between the robot and the goal, the angular difference between the robot's orientation and the goal and 24 laser readings. The 24 laser readings are normalized by the maximum detection range. The action space a is a vector of dimension 2 comprising two distinct velocities: the angular velocity V_a and the linear velocity V_l.

2) Reward Design: The reward function should guide each mobile robot to successfully attain its goal position while guaranteeing collision avoidance during the target-reaching process, as described in Fig. 2. Denote the reward function for the robot k as

$$r_k = r_k^d + r_k^a + r_k^e + r_k^n + r_k^o \tag{6}$$

where r_k is the total reward of the k^{th} robot, r_k^d and r_k^a stand for the distance and arrive rewards of the k^{th} robot. Linear and angular punishment rewards of the robot k are denoted by r_k^e and r_k^n. r_k^o describes the collision reward of the k^{th} robot.

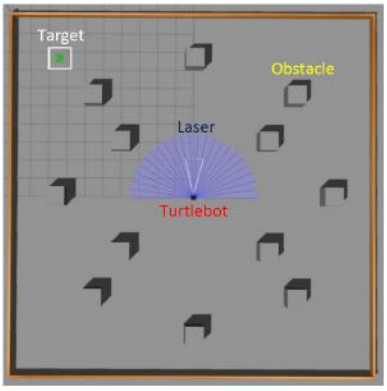


Fig. 2. Training scenario for each robot.

We design the distance reward r_k^d as

$$r_k^d = D_k^p - D_k^c \tag{7}$$

where D_k^p and D_k^c represent the distance between the k^{th} robot to the target with the previous and current action. If $D_k^p < D_k^c$, the robot will receive a negative distance reward since it moves further to the goal.

The arrive reward r_k^a is given by

$$r_k^a = \begin{cases} r_a & \textit{if } \ D_k^c < \varepsilon \\ 0 & \textit{otherwise} \end{cases} \tag{8}$$

where ε is the distance threshold and $r_a > 0$ denotes a large arrive reward.

The linear punishment reward r_k^e is defined as follows:

$$r_k^e = \begin{cases} r_e & \textit{if } \ V_l < \tilde{V}_l \\ 0 & \textit{otherwise} \end{cases} \tag{9}$$

where r_e is a negative reward, V_l and $\tilde{V}_l$ represent the linear velocity and minimum linear velocity threshold of the robot. The linear punishment reward encourages the robot to actively explore the environment.

Similarly, we can define the angular punishment reward r_k^n as

$$r_k^n = \begin{cases} r_n & \textit{if } \ |V_a| > \bar{V}_a \\ 0 & \textit{otherwise} \end{cases} \tag{10}$$

where r_n is a negative reward, V_a and $\bar{V}_a$ stand for the angular velocity and maximum angular velocity threshold of the robot. The angular punishment reward limits the excessive rotational speed of the robot, thereby promoting effective exploration of its surroundings.

Denote the distance between the robot k and the obstacles as D_k^o. The collision reward r_k^o can be written as

$$r_k^o = \begin{cases} r_o & \textit{if } \ \bar{D} \le D_k^o < 2\bar{D} \\ 2r_o & \textit{if } \ D_k^o < \bar{D} \\ 0 & \textit{otherwise} \end{cases} \tag{11}$$

where r_o is a negative collision reward, $\bar{D}$ is the laser threshold limit. Consequently, r_o incurs a penalty if D_k^o exceeds the laser threshold. By incorporating r_o into the total reward, the robot is stimulated to perform obstacle avoidance.

3) Network Structure: The actor network processes its input through three dense layers with ReLU activation functions [48], each comprising 500 nodes, as depicted in Fig. 1. This input vector consists of 28 elements, including parameters such as angular velocity, linear velocity, the distance between the robot and its goal, the angular difference between the robot and its goal and 24 laser readings. The actor network's output is a 2-element vector representing the angular velocity V_a and the linear velocity V_l. This output also serves as an input to the critic network, as shown in Fig. 1. The critic network's other input matches the elements of the actor network's input vector. The critic network's output, generated using a linear activation function, provides a Q value that quantitatively evaluates the actor network's performance.

In each iteration, every robot explores the environment by incorporating random noise into its actions, yielding a total reward denoted as r_k. During each training session, the observation space, action space, total reward, and next

observation space are stored and recorded in the replay buffer. Once the buffer reaches full capacity, the training model is able to randomly sample from its contents to facilitate learning.

B. Consensus-based Training

To enhance the training efficiency of DRL, a multi-agent training method on the basis of the consensus protocol is designed. Accordingly, this section presents the consensus strategy that supports this training process.

Let z_k be the row vector of the training parameter for each robot. After a single consensus step, we denote the updated training parameter as $\hat{z}_k$. The training protocol based on the consensus technique can be described as

$$\hat{z}_k = z_k + u_k^c \tag{12}$$

$$u_k^c = \sum\nolimits_{j=0}^{P} a_{kj}(z_j - z_k) \tag{13}$$

Substituting (13) to (12), we can obtain the consensus strategy as follows

$$\begin{aligned} \hat{z}_k &= z_k + \sum\nolimits_{j=0}^{P} a_{kj}(z_j - z_k) \\ &= z_k - \sum\nolimits_{j=0}^{P} l_{kj} z_j \\ &= \mathcal{C}_k\left(z_j, l_{kj}\right) \end{aligned} \tag{14}$$

where $\mathcal{C}_k$ stands for the consensus scheme of the robot k, and l_{kj} is the entry of the Laplacian matrix $\mathcal{L}$. Then, we can summarize the training parameter update for all robots as a compact form

$$\begin{aligned} \hat{z} &= ((I_{P+1} - \mathcal{L}) \otimes I_m) z \\ &= \mathcal{C}\left(z, \mathcal{L}\right) \end{aligned} \tag{15}$$

where $\mathcal{C}$ represents the consensus scheme for all robots. The entry of the matrix $I_{P+1} - \mathcal{L}$ stands for the weight of each robot. As observed from (15), the weights influence how information is shared among robots, thereby affecting the efficiency of training convergence. By iteratively calculating (15), the consensus approach facilitates the convergence of all robots to their weighted average [49].

C. Consensus-based DRL

Building upon the consensus training approach presented in the previous section, we now introduce the consensus-based DRL method. This policy reduces the number of required training steps while maintaining the same level of training reward. Define the actor and critic training parameters as τ^k and ζ^k for robot k. Let $g_\zeta(s_t^k, a_t^k)$ and $\pi_\tau(s^k)$ be the linearized value function related to ζ and linearized policy function related to τ. Their gradients are denoted by $\nabla_\zeta g(s_t^k, a_t^k)$ and $\nabla_\tau \pi(s_t^k)$. From [50], their training process can be expressed as

$$\varrho_t^k = r_{t+1}^k + \mu g_\zeta(s_{t+1}^k, a_{t+1}^k) - g_\zeta(s_t^k, a_t^k) \tag{16}$$

$$\zeta_{t+1}^k = \hat{\zeta}_t^k + \psi \varrho_t^k \nabla_\zeta g(s_t^k, a_t^k) \tag{17}$$

$$\tau_{t+1}^k = \hat{\tau}_t^k + \omega \nabla_\tau \pi(s_t^k) \nabla_a g_\zeta(s_t^k, a^k)|_{a^k = \pi_\tau(s^k)} \tag{18}$$

where ω and ψ are the actor and critic learning rates. ϱ_t^k is the temporary difference error and μ represents the discounted factor.

By implementing the consensus protocol (14), the update of the critic training parameter based on consensus for robot k can be described as

$$\hat{\zeta}_t^k = \zeta_t^k + \sum\nolimits_{j=0}^{P} a_{kj}(\zeta_t^j - \zeta_t^k) \tag{19}$$

$$\zeta_{t+1}^k = \hat{\zeta}_t^k + \psi \varrho_t^k \nabla_\zeta g(s_t^k, a_t^k) \tag{20}$$

Combining (19) and (20), it can be obtained that

$$\zeta_{t+1}^k = \zeta_t^k + \sum\nolimits_{j=0}^{P} a_{kj}(\zeta_t^j - \zeta_t^k) + \psi \varrho_t^k \nabla_\zeta g(s_t^k, a_t^k) \tag{21}$$

Let $q_t^k = \nabla_\tau \pi(s_t^k) \nabla_a g_\zeta(s_t^k, a^k)|_{a^k = \pi_\tau(s^k)}$. In the same way, we can get the update of the actor training parameter based on consensus for robot k as follows

$$\tau_{t+1}^k = \tau_t^k + \sum\nolimits_{j=0}^{P} a_{kj}(\tau_t^j - \tau_t^k) + \omega q_t^k \tag{22}$$

Considering all robots at iteration t, the critic and actor training parameters update with consensus are demonstrated as the following compact forms

$$\begin{aligned} \zeta_{t+1} &= ((I_{P+1} - \mathcal{L}) \otimes I_m)\zeta_t + \psi \Xi_t \\ &= \mathcal{C}\left(\zeta_t, \mathcal{L}\right) + \psi \Xi_t \end{aligned} \tag{23}$$

$$\begin{aligned} \tau_{t+1} &= ((I_{P+1} - \mathcal{L}) \otimes I_m)\tau_t + \omega Q_t \\ &= \mathcal{C}\left(\tau_t, \mathcal{L}\right) + \omega Q_t \end{aligned} \tag{24}$$

where $\Xi_t = [(\varrho_t^0 \nabla_\zeta g(s_t^0, a_t^0))^T, \cdots, (\varrho_t^P \nabla_\zeta g(s_t^P, a_t^P))^T]^T$ and $Q_t = [(q_t^0)^T, \cdots, (q_t^P)^T]^T$. The proposed DRL for obstacle avoidance with the consensus approach is illustrated in Algorithm 1.

Compared with reinforcement learning methods that utilize shared experience pools, the advantage of the proposed consensus-based distributed reinforcement learning method is that it protects the privacy of each agent. In real-world applications, directly utilizing shared experience pools to train a reinforcement learning model may not be feasible due to privacy concerns. To address this, our proposed consensus approach involves sharing training weights between agents rather than utilizing shared experience pools, as weights are simply numerical representations and do not contain sensitive information. This method of sharing weights is far more secure and trustworthy compared to directly using shared experience pools, as it ensures the protection of individual privacy while still enabling collaborative learning.

D. Leader-follower IEC-based Formation Scheme

In this section, an IEC-based formation protocol is illustrated to tackle uncertainties introduced by DRL. First, an intermediate observer is introduced to estimate the position and uncertainty of each follower robot. The intermediate variable for the i^{th} follower robot is defined as

$$\gamma_i(t) = f_i(t) - \iota x_i(t) \tag{25}$$

where ι is a positive scalar gain.

Algorithm 1 DRL for Obstacle Avoidance with Consensus Approach

Initialize fixed behavior policy χ, initial observation s_0^k, maximum time step $\tilde{T}$, actor and critic training parameters τ_0^k and ζ_0^k, exploration noise $\mathcal{O}_t$, discounted factor μ, actor and critic learning rates ω and ψ of each agent.
Acquire initial observation s_1^k of each agent.
Reset the environment.
while $t = 1, \tilde{T}$ **do**
  Select action $a_t^k = \chi(s_t^k) + \mathcal{O}_t$ on the basis of the exploration noise $\mathcal{O}_t$ and the fixed behavior policy χ.
  Execute action a_t^k, compute reward r_{t+1}^k and obtain new observation s_{t+1}^k.
  Store transition $(s_t^k, a_t^k, r_t^k, s_{t+1}^k)$ in the replay buffer.
  Sample a mini batch $(s_t^{\bar{k}}, a_t^{\bar{k}}, r_t^{\bar{k}}, s_{t+1}^{\bar{k}})$ from transitions in the replay buffer.
  Compute the temporal difference error:
$\varrho_t^k = r_{t+1}^k + \mu g_\zeta(s_{t+1}^k, a_{t+1}^k) - g_\zeta(s_t^k, a_t^k)$.
  Update the critic training parameter:
$\zeta_{t+1}^k = \hat{\zeta}_t^k + \psi \varrho_t^k \nabla_\zeta g(s_t^k, a_t^k)$.
  Update the actor training parameter:
$\tau_{t+1}^k = \hat{\tau}_t^k + \omega \nabla_\tau \pi(s_t^k) \nabla_a g_\zeta(s_t^k, a^k)|_{a^k=\pi_\tau(s^k)}$.
  Execute the consensus-based training approach on the actor and critic training parameters:
$\hat{\tau}_{t+1}^k = \mathcal{C}_k\left(\tau_{t+1}^j, l_{kj}\right)$.
$\hat{\zeta}_{t+1}^k = \mathcal{C}_k\left(\zeta_{t+1}^j, l_{kj}\right)$.
end while

It can be observed from (2) that

$$\begin{aligned}\dot{\gamma}_i(t) &= \dot{f}_i(t) - \iota \dot{x}_i(t) \\ &= \dot{f}_i(t) - \iota(h(x_i(t)) + u_i(t) + \gamma_i(t) + \iota x_i(t)).\end{aligned} \tag{26}$$

Let $\hat{x}_i(t)$, $\hat{\gamma}_i(t)$, and $\hat{f}_i(t)$ be the estimation of $x_i(t)$, $\gamma_i(t)$, and $f_i(t)$, respectively. The intermediate estimator of the uncertainties can be designed as

$$\begin{cases}\dot{\hat{x}}_i(t) = h(\hat{x}_i(t)) + u_i(t) + \hat{f}_i(t) + \Gamma(x_i(t) - \hat{x}_i(t)) \\ \dot{\hat{\gamma}}_i(t) = -\iota\hat{\gamma}_i(t) - \iota(h(\hat{x}_i(t)) + u_i(t) + \iota\hat{x}_i(t)) \\ \hat{f}_i(t) = \hat{\gamma}_i(t) + \iota\hat{x}_i(t),\end{cases} \tag{27}$$

where $\Gamma > 0$ denotes the scalar estimation gain. Hence, the distributed IEC-based formation protocol for each follower under the intermediate estimator (27) is proposed as

$$u_i(t) = -K\delta_i - \bar{w}\varphi(\delta_i) - \hat{f}_i, \;\; i = 1, 2\cdots, P, \tag{28}$$

where

$$\varphi(\delta_i) = \begin{cases}\dfrac{\delta_i}{\|\delta_i\|}, & \text{when } \|\delta_i\| \neq 0 \\ 0, & \text{when } \|\delta_i\| = 0.\end{cases} \tag{29}$$

$\bar{w} = w + \sqrt{2}\eta\|d\|$. K is the scalar positive control gain, and δ_i represents the formation error defined as

$$\delta_i = \sum_{j=1}^{P} a_{ij}((x_i - d_i) - (x_j - d_j)) + a_{i0}(x_i - d_i - x_0). \tag{30}$$

Denote the tracking error ξ_i of the system as $\xi_i = x_i - x_0 - d_i$. The estimation errors of x_i and γ_i are defined as $\xi_i^x = x_i - \hat{x}_i$ and $\xi_i^\gamma = \gamma_i - \hat{\gamma}_i$. From (2), (26), and (27), it can be obtained that

$$\begin{cases}\dot{\xi}_i = h(x_i) - h(x_0) - K\delta_i - \bar{w}\varphi(\delta_i) - w_0 + \iota\xi_i^x + \xi_i^\gamma \\ \dot{\xi}_i^x = h(x_i) - h(\hat{x}_i) + \xi_i^\gamma + (\iota - \Gamma)\xi_i^x \\ \dot{\xi}_i^\gamma = -\iota(\xi_i^\gamma + h(x_i) - h(\hat{x}_i) + \iota\xi_i^x) + \dot{f}_i.\end{cases} \tag{31}$$

Let $x = [x_1^T, \cdots, x_P^T]^T$, $\hat{x} = [\hat{x}_1^T, \cdots, \hat{x}_P^T]^T$, $\gamma = [\gamma_1^T, \cdots, \gamma_P^T]^T$, $\hat{\gamma} = [\hat{\gamma}_1^T, \cdots, \hat{\gamma}_P^T]^T$, $\xi = [\xi_1^T, \cdots, \xi_P^T]^T$, $\xi^x = [(\xi_1^x)^T, \cdots, (\xi_P^x)^T]^T$, $\xi^\gamma = [(\xi_1^\gamma)^T, \cdots, (\xi_P^\gamma)^T]^T$, and $\delta = [\delta_1^T, \cdots, \delta_P^T]^T$. From the definition of the Laplacian matrix, we can infer that

$$\delta = (\mathcal{L}_1 \otimes I_m)\xi, \tag{32}$$

where $I_m \in \mathbb{R}^{m\times m}$ is the identity matrix. The compact form of (31) can be rewritten as

$$\begin{cases}\dot{\xi} = H(x) - H(x_0) - K\delta - \bar{w}\Phi(\delta) - \mathbf{1}_P \otimes w_0 + \iota\xi^x + \xi^\gamma \\ \dot{\xi}^x = H(x) - H(\hat{x}) + \xi^\gamma + (\iota - \Gamma)\xi^x \\ \dot{\xi}^\gamma = -\iota(\xi^\gamma + H(x) - H(\hat{x}) + \iota\xi^x) + \dot{f}.\end{cases} \tag{33}$$

where $\mathbf{1}_P = [1, \cdots, 1]^T$, $H(x_0) = \mathbf{1}_P \otimes h(x_0)$, $H(x) = [h(x_1)^T, \cdots, h(x_P)^T]^T$, $H(\hat{x}) = [h(\hat{x}_1)^T, \cdots, h(\hat{x}_P)^T]^T$, and $\Phi(\delta) = [\varphi(\delta_1)^T, \cdots, \varphi(\delta_P)^T]^T$.

Theorem 1: Under Assumptions 1-4 and IEC-based formation protocol (28), the error system (33) is uniformly ultimate bounded if the gains K, Γ, and ι are designed to satisfy the following conditions:

$$K > \frac{\sqrt{2}\eta\|\mathcal{L}_1 \otimes I_m\|}{\lambda_{min}(\mathcal{L}_1^2)} + \frac{1}{\iota} + \frac{\iota}{2}, \tag{34}$$

$$\Gamma > \eta + \iota^2 + \frac{3}{2}\iota + \frac{1}{\iota} + 1, \tag{35}$$

and

$$0 < \iota < \frac{1}{\eta^2 + 1}, \tag{36}$$

where $\lambda_{min}(\mathcal{L}_1^2) > 0$ denotes the minimum eigenvalue of the matrix $\mathcal{L}_1^2$.

Proof: Consider the Lyapunov function:

$$V = V_1 + V_2 + V_3 \tag{37}$$

where $V_1 = \frac{1}{2}\xi^T(\mathcal{L}_1 \otimes I_m)\xi$, $V_2 = \frac{1}{2}(\xi^x)^T\xi^x$, and $V_3 = \frac{1}{2}(\xi^\gamma)^T\xi^\gamma$.

Taking the derivation of V_1, from (32) and (33), we can get

$$\begin{aligned}\dot{V}_1 &= \xi^T(\mathcal{L}_1 \otimes I_m)\dot{\xi} = \delta^T\dot{\xi} \\ &= \delta^T(H(x) - H(x_0) - K\delta - \bar{w}\Phi(\delta) \\ &\quad - \mathbf{1}_P \otimes w_0 + \iota\xi^x + \xi^\gamma) \\ &= -K\delta^T\delta + \delta^T(H(x) - H(x_0)) \\ &\quad - \delta^T(\bar{w}\Phi(\delta) + \mathbf{1}_P \otimes w_0) + \delta^T(\iota\xi^x + \xi^\gamma).\end{aligned} \tag{38}$$

Since the nonlinear function $h(\cdot)$ satisfies Assumption 2, it can be concluded that

$$\begin{aligned}\|H(x)-H(x_0)\| &= (\sum_{i=1}^{P}\|h(x_i)-h(x_0+d_i)\\ &\quad +h(x_0+d_i)-h(x_0)\|^2)^{\frac{1}{2}}\\ &\le (\sum_{i=1}^{P}2(\|h(x_i)-h(x_0+d_i)\|^2\\ &\quad +\|h(x_0+d_i)-h(x_0)\|^2))^{\frac{1}{2}}\\ &\le \sqrt{2}\eta(\sum_{i=1}^{P}(\|\xi_i\|^2+\|d_i\|^2))^{\frac{1}{2}}\\ &\le \sqrt{2}\eta(\|\xi\|+\|d\|).\end{aligned} \tag{39}$$

Combining with (32), we can imply that

$$\begin{aligned}\delta^T(H(x)-H(x_0)) &\le \|\delta^T\|\|(H(x)-H(x_0))\|\\ &\le \sqrt{2}\eta\|\delta^T\|(\|\xi\|+\|d\|)\\ &\le \sqrt{2}\eta\|\mathcal{L}_1\otimes I_m\|\xi^T\xi+\sqrt{2}\eta\|\delta\|\|d\|.\end{aligned} \tag{40}$$

Based on average inequality, we have

$$\iota\delta^T\xi^x \le \frac{\iota}{2}(\delta^T\delta+(\xi^x)^T\xi^x), \tag{41}$$

and

$$\delta^T\xi^\gamma \le \frac{1}{\iota}\delta^T\delta+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma. \tag{42}$$

Furthermore, from (5) and the definition of φ, it can be obtained that

$$\begin{aligned}-\delta^T(\bar{w}\Phi(\delta)+\mathbf{1}_P\otimes w_0) &= -\bar{w}\sum_{i=1}^{P}\|\delta_i\|-\sum_{i=1}^{P}\delta_i^T w_0\\ &\le -\bar{w}\sum_{i=1}^{P}\|\delta_i\|+\sum_{i=1}^{P}\|w_0\|\|\delta_i\|\\ &= (\|w_0\|-\bar{w})\sum_{i=1}^{P}\|\delta_i\|\\ &\le -\sqrt{2}\eta\|\delta\|\|d\|.\end{aligned} \tag{43}$$

Substituting (40), (41), (42), and (43) to (38), it follows that

$$\begin{aligned}\dot{V}_1 &\le -K\delta^T\delta+\sqrt{2}\eta\|\mathcal{L}_1\otimes I_m\|\xi^T\xi+\frac{\iota}{2}(\delta^T\delta+(\xi^x)^T\xi^x)\\ &\quad+\frac{1}{\iota}\delta^T\delta+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma\\ &= (-K+\frac{1}{\iota}+\frac{\iota}{2})\delta^T\delta+\sqrt{2}\eta\|\mathcal{L}_1\otimes I_m\|\xi^T\xi+\frac{\iota}{2}(\xi^x)^T\xi^x\\ &\quad+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma\\ &\le -((K-\frac{1}{\iota}-\frac{\iota}{2})\lambda_{min}(\mathcal{L}_1^2)-\sqrt{2}\eta\|\mathcal{L}_1\otimes I_m\|)\xi^T\xi\\ &\quad+\frac{\iota}{2}(\xi^x)^T\xi^x+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma,\end{aligned} \tag{44}$$

Next, we compute the derivation of V_2. From (33), it follows that

$$\begin{aligned}\dot{V}_2 &= (\xi^x)^T\dot{\xi}^x\\ &= (\xi^x)^T((H(x)-H(\hat{x}))+\xi^\gamma+(\iota-\Gamma)\xi^x)\\ &= (\xi^x)^T(H(x)-H(\hat{x}))+(\xi^x)^T\xi^\gamma+(\iota-\Gamma)(\xi^x)^T\xi^x.\end{aligned} \tag{45}$$

Similarly to (39), we can get

$$\|H(x)-H(\hat{x})\| \le \eta\|\xi^x\| \tag{46}$$

Hence, we have

$$\begin{aligned}(\xi^x)^T(H(x)-H(\hat{x})) &\le \|(\xi^x)^T\|\|(H(x)-H(\hat{x}))\|\\ &\le \eta(\xi^x)^T\xi^x.\end{aligned} \tag{47}$$

According to the average inequality, we have

$$(\xi^x)^T\xi^\gamma \le \frac{1}{\iota}(\xi^x)^T\xi^x+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma. \tag{48}$$

Putting (47) and (48) into (45), it can be implied that

$$\begin{aligned}\dot{V}_2 &\le (\iota-\Gamma)(\xi^x)^T\xi^x+\eta(\xi^x)^T\xi^x+\frac{1}{\iota}(\xi^x)^T\xi^x+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma\\ &= -(\Gamma-\eta-\iota-\frac{1}{\iota})(\xi^x)^T\xi^x+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma.\end{aligned} \tag{49}$$

Then, we turn to the derivation of V_3. It can be obtained from (33) that

$$\begin{aligned}\dot{V}_3 &= (\xi^\gamma)^T\dot{\xi}^\gamma\\ &= (\xi^\gamma)^T(-\iota(\xi^\gamma+H(x)-H(\hat{x})+\iota\xi^x)+\dot{f})\\ &= -\iota(\xi^\gamma)^T\xi^\gamma-\iota(\xi^\gamma)^T(H(x)-H(\hat{x}))-\iota^2(\xi^\gamma)^T\xi^x\\ &\quad+(\xi^\gamma)^T\dot{f}.\end{aligned} \tag{50}$$

In light of (47) and (48), we have

$$\begin{aligned}-\iota(\xi^\gamma)^T(H(x)-H(\hat{x})) &\le \iota\|(\xi^\gamma)^T\|\|(H(x)-H(\hat{x}))\|\\ &\le \iota\eta\|\xi^\gamma\|\|\xi^x\|\\ &\le \frac{\iota^2\eta^2}{4}(\xi^\gamma)^T\xi^\gamma+(\xi^x)^T\xi^x.\end{aligned} \tag{51}$$

From average inequality, we can get

$$-\iota^2(\xi^\gamma)^T\xi^x \le \frac{\iota^2}{4}(\xi^\gamma)^T\xi^\gamma+\iota^2(\xi^x)^T\xi^x, \tag{52}$$

and

$$(\xi^\gamma)^T\dot{f} \le \frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma+\frac{1}{\iota}\|\dot{f}\|^2 \tag{53}$$

Adding (51), (52), and (53), it can be seen from (4) that

$$\begin{aligned}\dot{V}_3 &\le -\iota(\xi^\gamma)^T\xi^\gamma+\frac{\iota^2\eta^2}{4}(\xi^\gamma)^T\xi^\gamma+(\xi^x)^T\xi^x+\frac{\iota^2}{4}(\xi^\gamma)^T\xi^\gamma\\ &\quad+\iota^2(\xi^x)^T\xi^x+\frac{\iota}{4}(\xi^\gamma)^T\xi^\gamma+\frac{1}{\iota}\|\dot{f}\|^2\\ &\le -(\frac{3}{4}\iota-\frac{\iota^2\eta^2+\iota^2}{4})(\xi^\gamma)^T\xi^\gamma+(\iota^2+1)(\xi^x)^T\xi^x+\frac{P\theta^2}{\iota}.\end{aligned} \tag{54}$$

Hence, it follows from (44), (49), and (54) that

$$\begin{aligned}\dot{V} &= \dot{V}_1 + \dot{V}_2 + \dot{V}_3 \\ &\leq -((K - \frac{1}{\iota} - \frac{\iota}{2})\lambda_{min}(\mathcal{L}_1^2) - \sqrt{2}\eta\|\mathcal{L}_1 \otimes I_m\|)\xi^T\xi \\ &- (\Gamma - \eta - \iota^2 - \frac{3}{2}\iota - \frac{1}{\iota} - 1)(\xi^x)^T\xi^x - (\frac{1}{4}\iota - \frac{\iota^2\eta^2 + \iota^2}{4}) \\ &(\xi^\gamma)^T\xi^\gamma + \frac{P\theta^2}{\iota} \\ &\leq -\frac{2((K - \frac{1}{\iota} - \frac{\iota}{2})\lambda_{min}(\mathcal{L}_1^2) - \sqrt{2}\eta\|\mathcal{L}_1 \otimes I_m\|)}{\|\mathcal{L}_1 \otimes I_m\|}V_1 \\ &- 2(\Gamma - \eta - \iota^2 - \frac{3}{2}\iota - \frac{1}{\iota} - 1)V_2 - (\frac{1}{2}\iota - \frac{\iota^2\eta^2 + \iota^2}{2})V_3 \\ &+ \frac{P\theta^2}{\iota} \\ &\leq -\tilde{a}V + \frac{P\theta^2}{\iota}.\end{aligned} \tag{55}$$

where

$$\begin{aligned}\tilde{a} = \min\{&\frac{2((K - \frac{1}{\iota} - \frac{\iota}{2})\lambda_{min}(\mathcal{L}_1^2) - \sqrt{2}\eta\|\mathcal{L}_1 \otimes I_m\|)}{\|\mathcal{L}_1 \otimes I_m\|}, \\ &2(\Gamma - \eta - \iota^2 - \frac{3}{2}\iota - \frac{1}{\iota} - 1),\ (\frac{1}{2}\iota - \frac{\iota^2\eta^2 + \iota^2}{2})\}\end{aligned} \tag{56}$$

Denote the bounded set Ω as

$$\Omega = \{(\xi, \xi^x, \xi^\gamma) \mid \xi^T(\mathcal{L}_1 \otimes I_m)\xi + \|\xi^x\|^2 + \|\xi^\gamma\|^2 < \frac{2P\theta^2}{\tilde{a}\iota}\} \tag{57}$$

if $(\xi, \xi^x, \xi^\gamma) \in \bar{\Omega}$, where $\bar{\Omega}$ is the supplementary set of Ω, we can get from (55) that

$$\dot{V} < 0. \tag{58}$$

That is to say, when $t \to \infty$,

$$V \leq \frac{P\theta^2}{\tilde{a}\iota}. \tag{59}$$

Hence, the error system (33) is uniformly ultimately bounded. This completes the proof. ■

Moreover, the following corollary can be obtained on the basis of the conclusion of Theorem 1.

Corollary 1: Under Assumptions 1-4 and IEC-based formation protocol (28), if the gains K, Γ, and ι are designed to satisfy conditions (34), (35), and (36), the error system (33) converges to zero exponentially if $f_i(t)$ is constant.

Proof: Choosing the Lyapunov candidate

$$V = \frac{1}{2}\xi^T(\mathcal{L}_1 \otimes I_m)\xi + \frac{1}{2}(\xi^x)^T\xi^x + \frac{1}{2}(\xi^\gamma)^T\xi^\gamma. \tag{60}$$

Since $f_i(t)$ is constant, we have $\dot{f} = 0$. Similarly to the analysis in Theorem 1, based on (55), it can be obtained that

$$\dot{V} \leq -\tilde{a}V, \tag{61}$$

where $\tilde{a}$ is defined in (56). That is to say, the error system (33) converges to zero exponentially for constant $f_i(t)$ with the exponential convergence rate $\tilde{a}$. This completes the proof. ■

E. Safe Multi-robot DRL Exploration with IEC-based Formation Scheme

This section introduces the proposed framework, which enables multi-robot safe DRL exploration under an IEC-based formation protocol, effectively addressing uncertainties introduced by DRL. In this work, safe learning is accomplished through a combination of DRL and the IEC-based formation protocol. The DRL algorithm ensures safety by integrating obstacle avoidance directly into the reward function. The reward penalizes collisions while encouraging safe navigation. Since DRL introduces uncertainties due to its obstacle avoidance process, we introduce an IEC-based formation protocol to provide an additional layer of control. This protocol functions as a corrective mechanism, compensating for errors and uncertainties in DRL-based decision-making while maintaining formation constraints. By leveraging this approach, the system maintains stability and safety despite the inherent unpredictability of DRL policies. The convergence of the Lyapunov function proves that the proposed framework effectively mitigates unsafe behaviors and ensures robust navigation.

After being trained with consensus-based DRL, the training model for each agent will be saved and loaded for both the leader and followers for the proposed framework to handle obstacle avoidance tasks. In every iteration, the distance to the obstacles D_k^o for each robot is obtained. If D_k^o is above the laser threshold limit $2\bar{D}$, each robot will perform formation-based exploration. Otherwise, the loaded training model will be triggered to deal with obstacle avoidance tasks. In this process, an action is chosen on the basis of the pre-trained model obtained from Algorithm 1. Then, this action is executed and a reward will be computed to get a new observation. Once the DRL obstacle avoidance algorithm is triggered, uncertainties will be introduced in the multi-robot system, which can be handled by the proposed IEC-based formation protocol (28). Algorithm 2 details the proposed safe multi-robot DRL exploration with the IEC-based formation scheme, which enables collision-free mapless exploration while maintaining a focus on safe learning.

V. Experiments and Results

This section validates the feasibility and effectiveness of the proposed cooperative framework. The experiments were conducted with TurtleBot3 Waffle Pi robots in both simulated and real-world environments.

A. Training Details

1) Simulation: The proposed framework was trained in Robot Operating System (ROS) [51] and Gazebo [52]. For each robot, the maximum laser range is 3.5 m. The actor learning rate ω and the critic learning rate ψ are both fixed at 0.0001. The discounted factor μ has a fixed value of 0.9. Action exploration follows the ϵ-greedy exploration strategy [53], with ϵ initialized at 0.9 and a decay rate of 0.99995.

Considering a multi-robot exploration task which includes one leader and three followers with dimension $m = 2$. The position of each robot can be denoted as $x_i(t) =$

Algorithm 2 Safe Multi-robot DRL Exploration with IEC-based Formation Scheme

Choose one leader robot labeled with 0 and P follower robots labeled with $\{1, \cdots, P\}$.
Initialize the target distance d between the leader and each follower, and the initial position of each robot.
Initialize the connections between each robot.
Initialize the laser threshold limit $\bar{D}$ and maximum real time step $\bar{T}$.
Load the consensus-based DRL pre-trained model obtained from Algorithm 1.
Reset the environment and get observation state s_1^k.
Set the control gains K, Γ, and ι
if Assumptions 1-4 are satisfied **then**
  if (34), (35), and (36) are satisfied **then**
    while $t = 1, \bar{T}$ **do**
      Get the position of each robot.
      Get the distance to the obstacles D_k^o.
      if $D_k^o > 2\bar{D}$ **then**
        Calculate formation error δ_k for robot k.
        if $\|\delta_k\| \neq 0$ **then**.
          $\varphi_k(\delta_k) = \frac{\delta_k}{\|\delta_k\|}$.
        else
          $\varphi_k(\delta_k) = 0$.
        end if
        Design the fault estimator (27).
        Establish the formation controller (28).
      else
        Select action $a_t^k = \chi(s_t^k)$ based on the pre-trained model in Algorithm 1.
        Execute action a_t^k, compute reward r_{t+1}^k and get new observation s_{t+1}^k.
        $s_t^k = s_{t+1}^k$.
      end if
    end while
  else
    Back to step 5.
  end if
else
  Back to step 1.
end if

$[x_{ix}(t), x_{iy}(t)]^T$, $i \in \mathcal{V}$, where $\mathcal{V} = \{0, 1, 2, 3\}$. We select the target formation shape as a 10 m square. The interaction topology is introduced in Fig. 1.

In the dynamic of each robot, we set the nonlinear function h as

$$h(x_i(t)) = 0.02[\sin(\frac{x_{ix}(t)}{2}) + 2, \cos(\frac{x_{iy}(t)}{2}) + 2]^T, \quad i \in \mathcal{V} \tag{62}$$

The initial position of each robot is set by

$$x_0(0) = [-3, -18]^T,\ x_1(0) = [-8, -18]^T,$$
$$x_2(0) = [-13, -18]^T,\ x_3(0) = [-18, -18]^T.$$

All the initial values in the position and fault estimators are set to zero. The gains are chosen as $K = 6$, $\Gamma = 9$, and $\iota = 0.3$, which satisfy the conditions in Theorem 1.

2) Real Environment: To confirm the validity of the proposed cooperative framework, four TurtleBot3 Waffle Pi robots were used in the real-world scenario, as shown in Fig. 3.

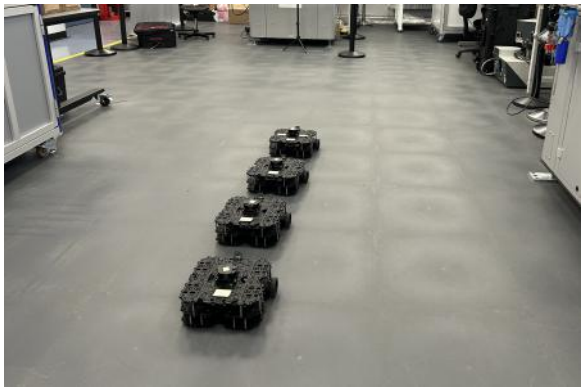


Fig. 3. Real world scenario with four TurtleBot3 Waffle Pi robots.

We set the target formation shape as a 1.25 m square according to the actual space limitations. Then we set the nonlinear function h as

$$h(x_i(t)) = 0.02[\sin(\frac{x_{ix}(t)}{2}) + 2, 0.01]^T, \quad i \in \mathcal{V} \tag{63}$$

The initial position of each robot is set by

$$x_0(0) = [0, 0]^T,\ x_1(0) = [-0.5, 0]^T,$$
$$x_2(0) = [-1, 0]^T,\ x_3(0) = [-1.5, 0]^T.$$

All the initial values in the position and fault estimators are set to zero. The gains are chosen as $K = 5$, $\Gamma = 6$, and $\iota = 0.4$, which satisfy the conditions in Theorem 1.

B. Simulation Results

1) Evaluation of Consensus-based DRL: This section evaluates the performance of DRL for obstacle avoidance tasks with the consensus approach. The consensus-based training environment is shown in Fig. 4. As a baseline for comparison, we removed the consensus approach and applied the same DRL approach to train a single robot. The benchmark is established by the average reward obtained, calculated as the mean reward collected over a predefined batch size. Notably, the consensus-based DRL algorithm demonstrates a faster training process relative to the baseline. As can be seen from Fig. 5, with the consensus algorithm, each robot achieves a training reward of 25 within approximately $33,000$ training steps, while the baseline approach needs roughly $74,000$ steps to reach an equivalent reward. Thus, compared to single-robot DRL, the consensus approach decreases the training steps needed to attain the same reward, which enhances the training efficiency of the DRL method.

Fig. 6 illustrates the training time per step of the consensus-based DRL with four robots compared to the single-agent baseline approach. The average training time per step of four robots with the consensus method is around 0.43 s, which is slightly higher than the 0.27 s observed in the single-agent baseline. However, without the consensus strategy, training four robots independently would require 0.27 s per step per

robot, leading to a total time of 1.08 s per step. This is significantly higher than the 0.43 s per step required with the consensus-based approach. Thus, the consensus method not only accelerates convergence but also reduces the overall training time, making it a more efficient solution for multi-robot learning.

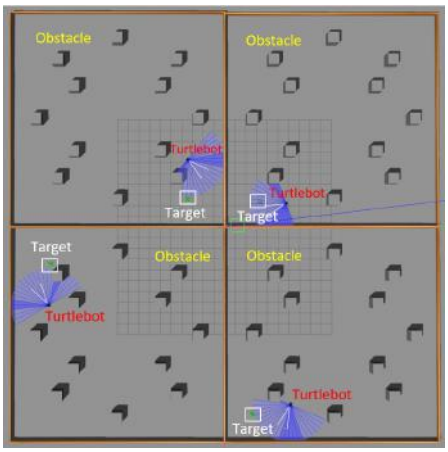


Fig. 4. Consensus-based DRL in Gazebo. The cubes are obstacles and the green squares represent target positions.

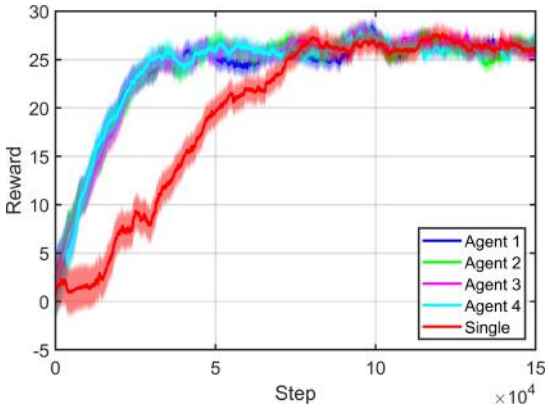


Fig. 5. Average reward graph of the consensus-based DRL with four robots compared to the single-agent baseline approach (Red line).

2) Evaluation of Randomly Placed Obstacles: Fig. 7 confirms the feasibility of the proposed cooperative framework with randomly placed obstacles. Under the IEC-based formation scheme, the proposed multi-robot framework successfully avoids randomly placed obstacles while maintaining the formation shape for safe exploration. For baseline trajectory comparison, pure formation control is applied to four mobile robots in the same environment without obstacles. As can be seen from Fig. 8, the proposed framework successfully addresses uncertainties while ensuring collision-free mapless exploration. In comparison to the pure formation strategy baseline, the uncertainties introduced by randomly placed obstacles are effectively handled by the proposed multi-robot framework under the IEC-based formation scheme, thereby validating our theoretical analysis.

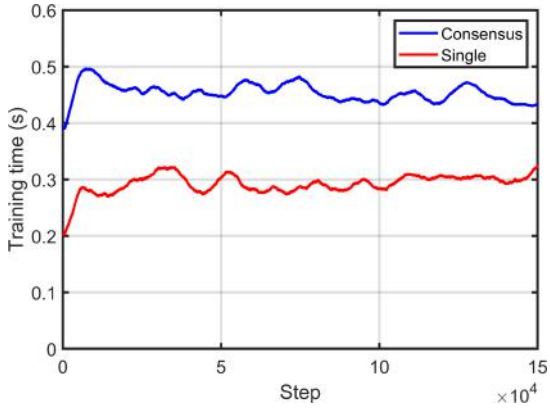


Fig. 6. Training time per step of the consensus-based DRL with four robots compared to the single-agent baseline approach (Red line).

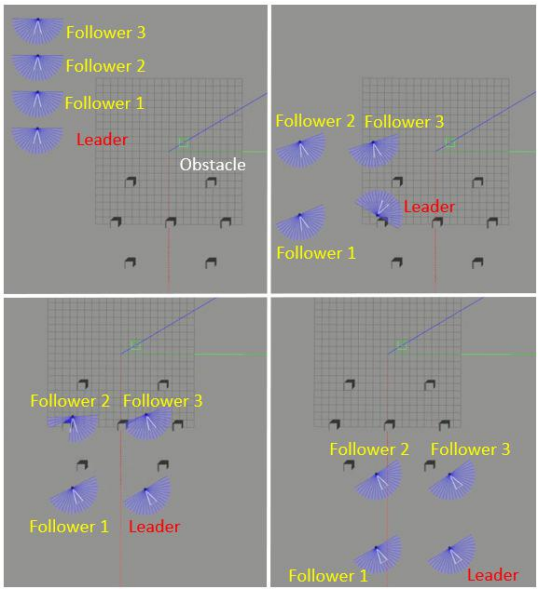


Fig. 7. The proposed framework with randomly placed obstacles.

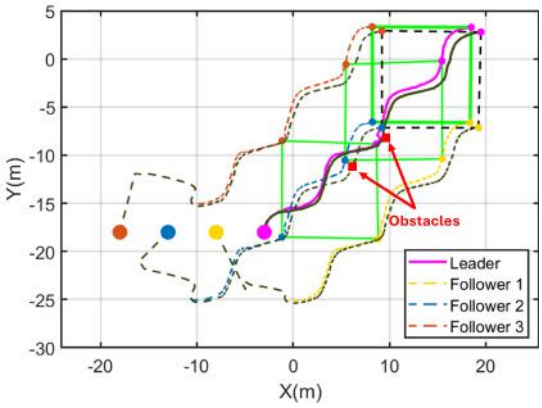


Fig. 8. Formation trajectories of the proposed framework with randomly placed obstacles compared with the pure formation strategy baseline (Black dotted line).

The tracking error $\|\xi_i\|$ of each follower robot in the

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3581872

presence of randomly placed obstacles is illustrated in Fig. 9. Although the tracking error for each follower robot increases when encountering obstacles, these tracking errors converge to a bounded set regardless of the uncertainties introduced by the randomly placed obstacles, which demonstrates the feasibility of safe exploration.

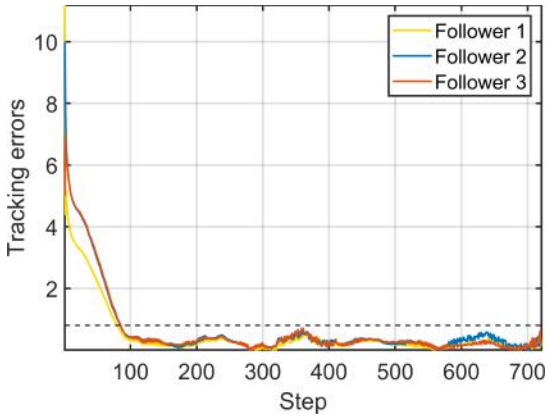


Fig. 9. Tracking errors of the followers with randomly placed obstacles. The dashed lines represent the theoretical bounds.

To evaluate the effectiveness of the proposed IEC-based formation protocol, we conduct a comparative experiment by removing the IEC component and employing only the traditional formation protocol under identical conditions. Fig. 10 depicts the tracking error $\|\xi_i\|$ of each follower robot when using a traditional formation protocol in the presence of randomly placed obstacles. The results indicate that, without the IEC-based formation protocol, the tracking errors of the followers fail to converge to a bounded set due to the uncertainties introduced by the obstacles. This observation highlights the efficacy of the proposed framework in handling such uncertainties.

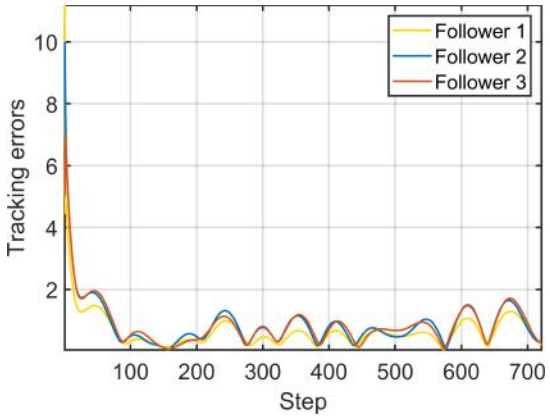


Fig. 10. Tracking errors of the followers with randomly placed obstacles using a traditional formation protocol.

3) Evaluation of Unexpected Obstacles: To demonstrate that the proposed multi-robot framework operates independently of maps for exploration, we introduce unexpected obstacles within the ongoing exploration process, as shown in Fig. 11. Although the global map suddenly changes, robots are able to successfully detect and avoid unexpected obstacles while maintaining the formation shape for safe exploration.

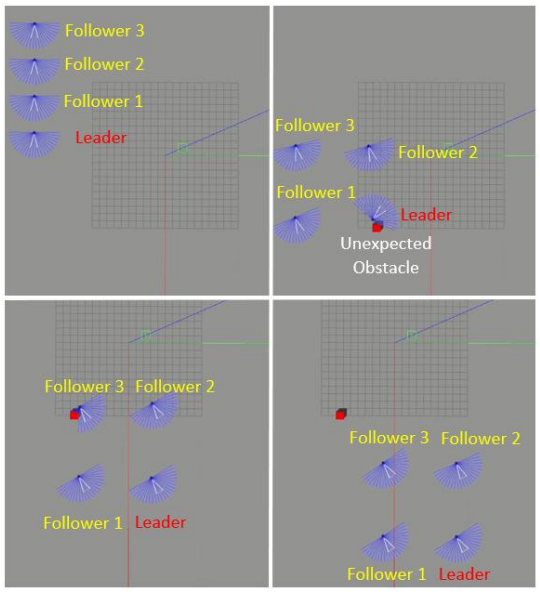


Fig. 11. The proposed framework with unexpected obstacles.

The tracking error $\|\xi_i\|$ of each follower robot in the presence of unexpected obstacles is demonstrated in Fig. 12. Compared with Fig. 9, the tracking error for each follower robot rises sharply upon encountering unexpected obstacles. Nonetheless, these tracking errors ultimately converge to a bounded set despite the uncertainties introduced by the unexpected obstacles, which confirms the validity of safe exploration.

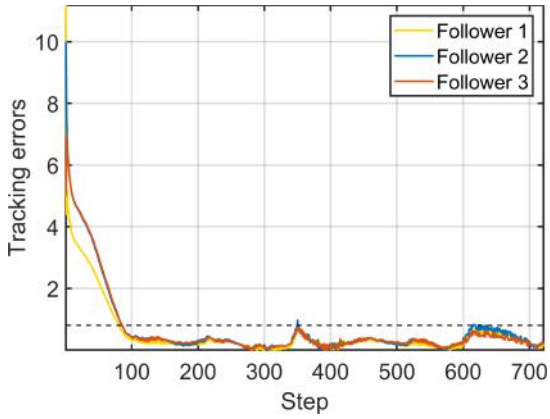


Fig. 12. Tracking errors of the followers with unexpected obstacles. The dashed lines represent the theoretical bounds.

C. Real-world Experiments

1) Evaluation of Randomly Placed Obstacles: The proposed framework has been implemented on four TurtleBot3 Waffle Pis for real-world evaluation. The objective is to deploy multi-robot safe exploration despite the uncertainties posed by randomly placed obstacles. As shown in Fig. 13, four

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3581872

TurtleBots successfully detect the randomly placed obstacle, as indicated by the presence of laser scans surrounding it. When any robot encounters the randomly placed obstacle during environment exploration, it triggers the DRL algorithm, enabling the robot to initiate obstacle avoidance and introducing uncertainties into the multi-robot system. Fig. 14 illustrates the formation trajectories of each robot with randomly placed obstacles in the real environment. The proposed cooperative framework successfully enables multi-robot safe exploration and effectively handles the uncertainties introduced by the randomly placed obstacle, which validates the effectiveness of Algorithm 2.

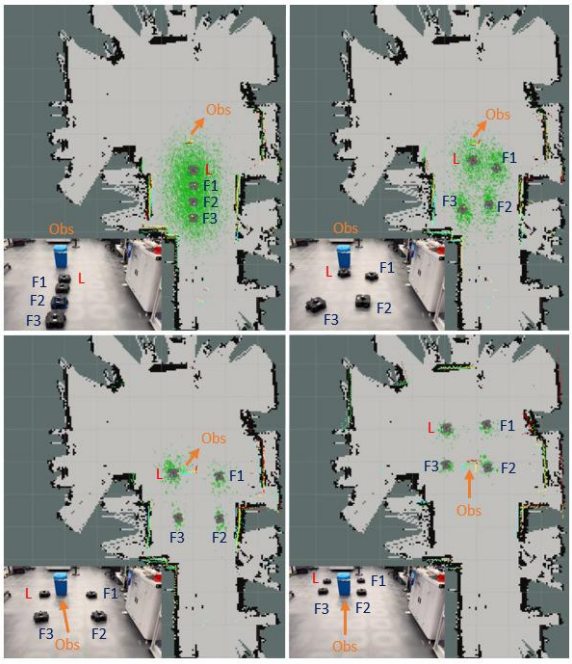


Fig. 13. The proposed framework with randomly placed obstacles in the real environment. L stands for Leader. F1, F2 and F3 represent Follower 1, Follower 2 and Follower 3, respectively.

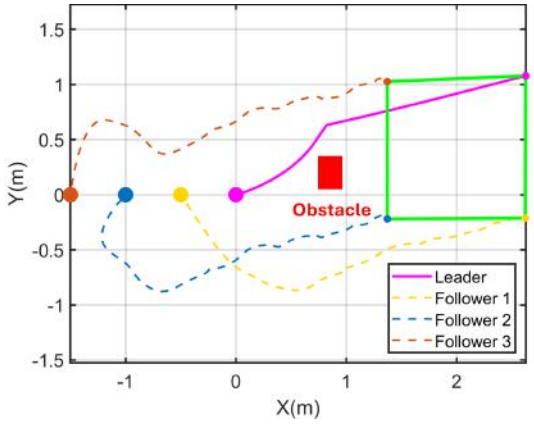


Fig. 14. Formation trajectories of the proposed framework with randomly placed obstacles in the real environment.

The real-world tracking error $\|\xi_i\|$ of each follower robot in the presence of randomly placed obstacles is presented in Fig. 15. In comparison with Fig. 9, the initial tracking error gradient for each follower is higher due to real-world uncertainties such as communication delay or friction. While the tracking error for each follower robot increases upon encountering obstacles, these errors ultimately converge to a bounded set regardless of the uncertainties introduced by the randomly placed obstacle, demonstrating the applicability of safe exploration.

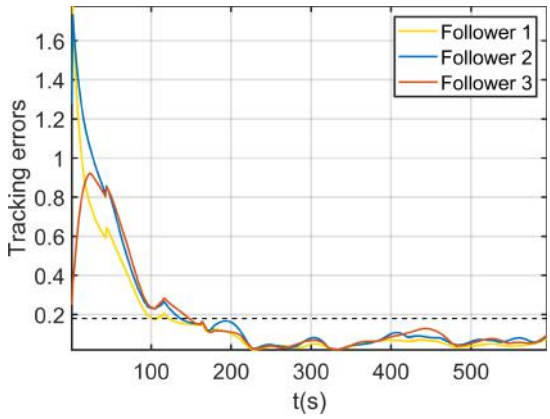


Fig. 15. Tracking errors of the followers with randomly placed obstacles in the real environment. The dashed lines represent the theoretical bounds.

To assess the performance of the proposed IEC-based formation protocol in the real world, a comparative analysis is conducted by eliminating the IEC component and relying solely on the conventional formation protocol under the same experimental conditions. Fig. 16 illustrates the real-world tracking error $\|\xi_i\|$ of each follower robot when using a traditional formation protocol in the presence of randomly placed obstacles. In the absence of the IEC-based formation protocol, the real-world tracking errors of the follower robots do not converge within a bounded region, primarily due to real-world uncertainties. This result underscores the effectiveness of the proposed framework in mitigating such challenges.

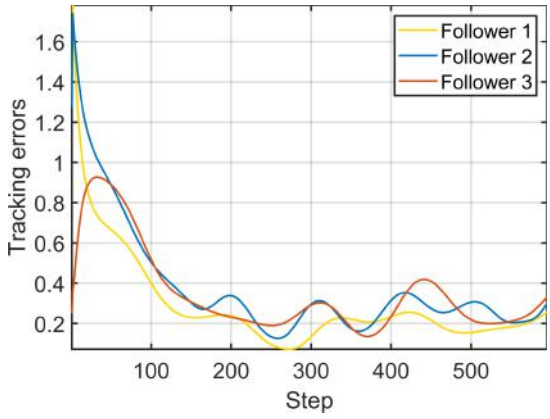


Fig. 16. Tracking errors of the followers with randomly placed obstacles in the real world using a traditional formation protocol.

2) Evaluation of Unexpected Obstacles: The performance of the cooperative framework with the unexpected obstacle in the real-world environment is validated in Fig. 17. In

This article has been accepted for publication in IEEE Transactions on Vehicular Technology. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TVT.2025.3581872

contrast to Fig. 13, four TurtleBots initially can not detect the unexpected obstacle until it is randomly placed in their path by the operator. Once the obstacle appears within the proximity defined by the laser threshold, the DRL algorithm is activated by the robot, which introduces uncertainties into the multi-robot system. Fig. 18 shows the formation trajectories of each robot with unexpected obstacles in the real environment. Each mobile robot successfully maintains the formation shape for safe exploration under the proposed framework, effectively managing uncertainties caused by unexpected obstacles.

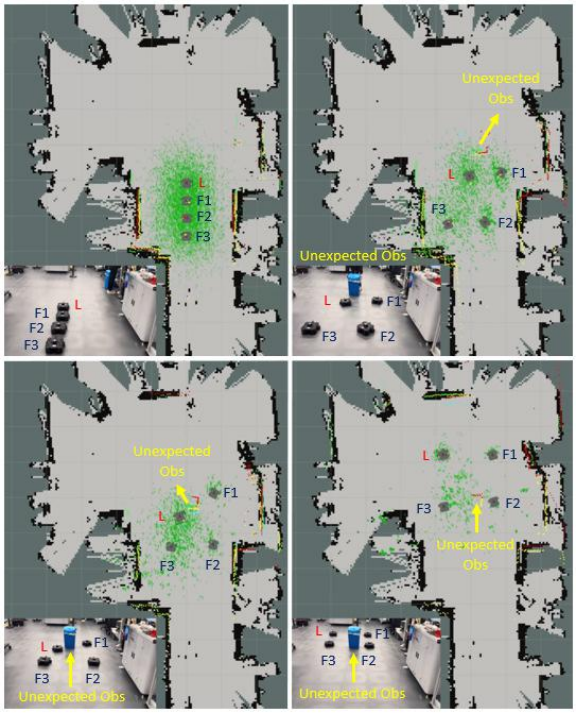


Fig. 17. The proposed framework with unexpected obstacles in the real environment. L stands for Leader. F1, F2 and F3 represent Follower 1, Follower 2 and Follower 3, respectively.

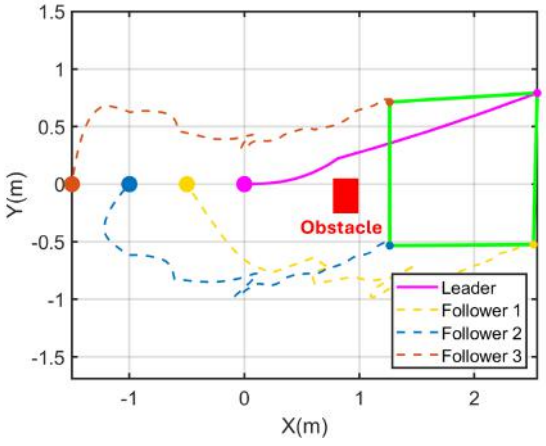


Fig. 18. Formation trajectories of the proposed framework with unexpected obstacles in the real environment.

The real-world tracking error $\|\xi_i\|$ of each follower robot in the presence of unexpected obstacles is presented in Fig. 19. Compared to Fig. 15, the tracking error for each follower robot exhibits an additional peak within the first 200 steps due to external uncertainties introduced by the unexpected obstacles, which impact the tracking error of each follower robot. Despite uncertainties introduced by the unexpected obstacle in the real world, the tracking errors still converge to a bounded set, affirming the robustness of the proposed framework.

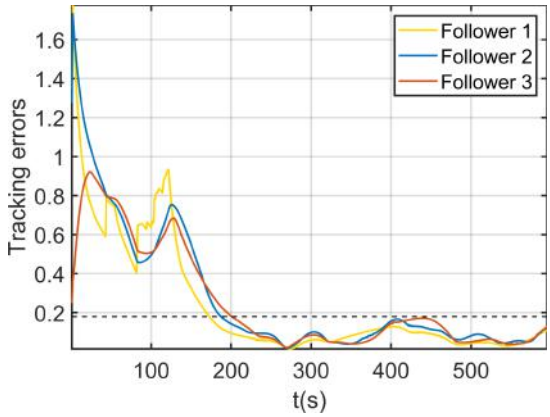


Fig. 19. Tracking errors of the followers with unexpected obstacles in the real environment. The dashed lines represent the theoretical bounds.

VI. Conclusion

In this work, a safe learning framework for multi-robot mapless exploration with the IEC-based formation scheme is presented. An actor-critic-based DRL method is implemented for each mobile robot to manage collision avoidance tasks. To enhance the efficiency of the DRL training process in the multi-robot system, a consensus-based training policy is proposed, reducing the required training steps without compromising training rewards. Additionally, an IEC-based formation scheme is developed to manage robots with uncertainties. A compensation function and an intermediate estimator are introduced in the controller to ensure that the tracking error of each follower robot converges to a bounded set with the aim of safe exploration. Simulations and real-world experiments are provided to validate the feasibility and effectiveness of the proposed framework. Future research will aim to refine this approach to tackle increasingly complex scenarios. In particular, investigating adaptive weight adjustments or DRL-based weight tuning could further improve the consensus process in dynamic environments.

Acknowledgment

The authors would like to thank Dr. Salvador Pacheco-Gutierrez and Dr. Kirsty Hewitson for the computational resource support from Robotics and Artificial Intelligence Collaboration (RAICo). We are grateful to Jeff Slater for his assistance with the deployment of the real-world experiments. We also thank Marc Torrance and John Jukes for their valuable support with both hardware and software.

REFERENCES

[1] X. Dong and G. Hu, "Time-varying formation tracking for linear multi-agent systems with multiple leaders," *IEEE Transactions on Automatic Control*, vol. 62, no. 7, pp. 3658–3664, 2017.

[2] K. Wu, J. Hu, Z. Ding, and F. Arvin, "Distributed bearing-only formation control for heterogeneous nonlinear multi-robot systems," *IFAC-PapersOnLine*, vol. 56, no. 2, pp. 3447–3452, 2023.

[3] W. Liu, H. Niu, W. Pan, G. Herrmann, and J. Carrasco, "Sim-and-real reinforcement learning for manipulation: A consensus-based approach," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3911–3917.

[4] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare, J. Pineau *et al.*, "An introduction to deep reinforcement learning," *Foundations and Trends® in Machine Learning*, vol. 11, no. 3-4, pp. 219–354, 2018.

[5] Z. Zhu and H. Zhao, "A survey of deep rl and il for autonomous driving policy learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 043–14 065, 2021.

[6] Y. Zhang, W. Zhao, J. Wang, and Y. Yuan, "Recent progress, challenges and future prospects of applied deep reinforcement learning: A practical perspective in path planning," *Neurocomputing*, vol. 608, p. 128423, 2024.

[7] X. Bi, M. He, and Y. Sun, "Mix q-learning for lane changing: A collaborative decision-making method in multi-agent deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, 2025.

[8] L. Li, R. Zhu, S. Wu, W. Ding, M. Xu, and J. Lu, "Adaptive multi-agent deep mixed reinforcement learning for traffic light control," *IEEE Transactions on Vehicular Technology*, vol. 73, no. 2, pp. 1803–1816, 2023.

[9] J. Hao, T. Yang, H. Tang, C. Bai, J. Liu, Z. Meng, P. Liu, and Z. Wang, "Exploration in deep reinforcement learning: From single-agent to multiagent domain," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[10] C. Wang, J. Wang, Y. Shen, and X. Zhang, "Autonomous navigation of uavs in large-scale complex environments: A deep reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 2124–2136, 2019.

[11] E. Marchesini, D. Corsi, and A. Farinelli, "Exploring safer behaviors for deep reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 7, 2022, pp. 7701–7709.

[12] H. Li, Z. Wan, and H. He, "Constrained ev charging scheduling based on safe deep reinforcement learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 3, pp. 2427–2439, 2019.

[13] S. Gu, L. Yang, Y. Du, G. Chen, F. Walter, J. Wang, and A. Knoll, "A review of safe reinforcement learning: Methods, theory and applications," *arXiv preprint arXiv:2205.10330*, 2022.

[14] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, pp. 411–444, 2022.

[15] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in neural information processing systems*, vol. 30, 2017.

[16] T. Koller, F. Berkenkamp, M. Turchetta, and A. Krause, "Learning-based model predictive control for safe exploration," in *2018 IEEE conference on decision and control (CDC)*. IEEE, 2018, pp. 6059–6066.

[17] T.-H. Pham, G. De Magistris, and R. Tachibana, "Optlayer-practical constrained optimization for deep reinforcement learning in the real world," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6236–6243.

[18] G. Thomas, Y. Luo, and T. Ma, "Safe reinforcement learning by imagining the near future," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 859–13 869, 2021.

[19] Y. Yang, Y. Jiang, Y. Liu, J. Chen, and S. E. Li, "Model-free safe reinforcement learning through neural barrier certificate," *IEEE Robotics and Automation Letters*, vol. 8, no. 3, pp. 1295–1302, 2023.

[20] L. Zhang, Q. Zhang, L. Shen, B. Yuan, X. Wang, and D. Tao, "Evaluating model-free reinforcement learning toward safety-critical tasks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 12, 2023, pp. 15 313–15 321.

[21] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery rl: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.

[22] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.

[23] C. Tessler, D. J. Mankowitz, and S. Mannor, "Reward constrained policy optimization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=SkfrvsA9FX

[24] C. Wei, Z. Ji, and B. Cai, "Particle swarm optimization for cooperative multi-robot task allocation: a multi-objective approach," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2530–2537, 2020.

[25] C. Wei, K. V. Hindriks, and C. M. Jonker, "Dynamic task allocation for multi-robot search and retrieval tasks," *Applied Intelligence*, vol. 45, pp. 383–401, 2016.

[26] Y. Wu, J. Gou, X. Hu, and Y. Huang, "A new consensus theory-based method for formation control and obstacle avoidance of uavs," *Aerospace Science and Technology*, vol. 107, p. 106332, 2020.

[27] B. Wang, S. G. Nersesov, and H. Ashrafiuon, "Robust formation control and obstacle avoidance for heterogeneous underactuated surface vessel networks," *IEEE Transactions on Control of Network Systems*, vol. 9, no. 1, pp. 125–137, 2022.

[28] K. Wu, J. Hu, Z. Li, Z. Ding, and F. Arvin, "Distributed collision-free bearing coordination of multi-uav systems with actuator faults and time delays," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 9, pp. 11 768–11 781, 2024.

[29] G. Wen, C. P. Chen, and Y.-J. Liu, "Formation control with obstacle avoidance for a class of stochastic multiagent systems," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5847–5855, 2017.

[30] X. Ge, Q.-L. Han, J. Wang, and X.-M. Zhang, "A scalable adaptive approach to multi-vehicle formation control with obstacle avoidance," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 6, pp. 990–1004, 2021.

[31] T. Nguyen, H. M. La, T. D. Le, and M. Jafari, "Formation control and obstacle avoidance of multiple rectangular agents with limited communication ranges," *IEEE Transactions on Control of Network Systems*, vol. 4, no. 4, pp. 680–691, 2016.

[32] Y. Liu, C. Chen, Y. Wang, T. Zhang, and Y. Gong, "A fast formation obstacle avoidance algorithm for clustered uavs based on artificial potential field," *Aerospace Science and Technology*, p. 108974, 2024.

[33] Y. Han, I. H. Zhan, W. Zhao, J. Pan, Z. Zhang, Y. Wang, and Y.-J. Liu, "Deep reinforcement learning for robot collision avoidance with self-state-attention and sensor fusion," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 6886–6893, 2022.

[34] D. Wang, T. Fan, T. Han, and J. Pan, "A two-stage reinforcement learning approach for multi-uav collision avoidance under imperfect sensing," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3098–3105, 2020.

[35] A. Singla, S. Padakandla, and S. Bhatnagar, "Memory-based deep reinforcement learning for obstacle avoidance in uav with limited environment knowledge," *IEEE transactions on intelligent transportation systems*, vol. 22, no. 1, pp. 107–118, 2019.

[36] Y. Zhu, R. Mottaghi, E. Kolve, J. J. Lim, A. Gupta, L. Fei-Fei, and A. Farhadi, "Target-driven visual navigation in indoor scenes using deep reinforcement learning," in *2017 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2017, pp. 3357–3364.

[37] S. Wen, Y. Shu, A. Rad, Z. Wen, Z. Guo, and S. Gong, "A deep residual reinforcement learning algorithm based on soft actor-critic for autonomous navigation," *Expert Systems with Applications*, vol. 259, p. 125238, 2025.

[38] P. Hang, C. Lv, C. Huang, J. Cai, Z. Hu, and Y. Xing, "An integrated framework of decision making and motion planning for autonomous vehicles considering social behaviors," *IEEE transactions on vehicular technology*, vol. 69, no. 12, pp. 14 458–14 469, 2020.

[39] S. Li, K. Peng, F. Hui, Z. Li, C. Wei, and W. Wang, "A decision-making approach for complex unsignalized intersection by deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, 2024.

[40] K. Zhu, B. Li, W. Zhe, and T. Zhang, "Collision avoidance among dense heterogeneous agents using deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 8, no. 1, pp. 57–64, 2022.

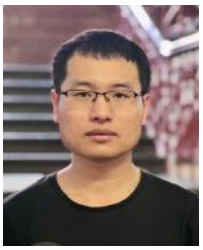
[41] Z. Sui, Z. Pu, J. Yi, and S. Wu, "Formation control with collision avoidance through deep reinforcement learning using model-guided demonstration," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 6, pp. 2358–2372, 2020.

[42] Y. F. Chen, M. Liu, M. Everett, and J. P. How, "Decentralized non-communicating multiagent collision avoidance with deep reinforcement learning," in *IEEE international conference on robotics and automation (ICRA)*, 2017, pp. 285–292.

[43] P. Long, T. Fanl, X. Liao, W. Liu, H. Zhang, and J. Pan, "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6252–6259.

[44] N. Thumiger and M. Deghat, "A multi-agent deep reinforcement learning approach for practical decentralized uav collision avoidance," *IEEE Control Systems Letters*, vol. 6, pp. 2174–2179, 2021.
[45] P. Long, W. Liu, and J. Pan, "Deep-learned collision avoidance policy for distributed multiagent navigation," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 656–663, 2017.
[46] S. H. Semnani, H. Liu, M. Everett, A. De Ruiter, and J. P. How, "Multi-agent motion planning for dense and dynamic environments via deep reinforcement learning," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3221–3226, 2020.
[47] C. Godsil and G. F. Royle, *Algebraic graph theory*. Springer Science & Business Media, 2013, vol. 207.
[48] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.
[49] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Transactions on automatic control*, vol. 49, no. 9, pp. 1520–1533, 2004.
[50] W. Liu, H. Niu, I. Jang, G. Herrmann, and J. Carrasco, "Distributed neural networks training for robotic manipulation with consensus algorithm," *IEEE transactions on neural networks and learning systems*, vol. 35, no. 2, pp. 2732–2746, 2022.
[51] M. Quigley, K. Conley, B. Gerkey, J. Faust, T. Foote, J. Leibs, R. Wheeler, and A. Y. Ng, "Ros: an open-source robot operating system," in *ICRA workshop on open source software*, vol. 3, no. 3.2. Kobe, Japan, 2009, p. 5.
[52] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 3. IEEE, 2004, pp. 2149–2154.
[53] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

Wenxing Liu (Member, IEEE) received her B.Eng. degree (First Class Honours) in Mechatronic Engineering from the University of Manchester, Manchester, UK, in 2019. She was awarded her Ph.D. in Electrical and Electronic Engineering by the University of Manchester in 2023. She is currently a Robotics Research Engineer under the robot learning theme at Remote Applications in Challenging Environments (RACE), UK Atomic Energy Authority, Culham, UK.
Her research interests include deep reinforcement learning, distributed control, and their applications in robotic manipulators and mobile robots.



Hanlin Niu (Member, IEEE) received the Ph.D degree in Aeronautical Engineering from Cranfield University in 2018. From July 2017 to January 2020, he worked as a Research Associate in Robotics at Cardiff University. From January 2020 to December 2022, he worked as a Research Associate in Robotics at the University of Manchester, being involved in the Robotics and AI in Nuclear project, one of the four big robotics and AI projects funded by the Engineering and Physical Sciences Research Council. From December 2022 to now, Hanlin works at UK Atomic Energy Authority and leads the autonomous robot inspection theme and robot learning theme of the Magnetic Fusion Research Programme. From October 2024 to now, Hanlin works on the locomotion algorithm of the legged robot as a senior robotics researcher in the Oxford Robotics Institute, University of Oxford. His research interests include deep reinforcement learning, imitation learning, locomotion, manipulation, guidance, navigation and control algorithms.

Kefan Wu received B.Sc degree in Applied Mathematics from Lanzhou University in 2016, M.Sc degree in Applied Mathematics from Wuhan University in 2019, and Ph.D degree in Electrical and Electronic Engineering from the University of Manchester in 2023. He is currently a senior researcher at Lakeside Labs GmbH, Klagenfurt, Austria.
His research interests include networked control systems and swarm intelligence.

Wei Pan (Member, IEEE) received the Ph.D. degree in control engineering from Imperial College London, London, U.K., in 2017. He is currently an Associate Professor with the Department of Computer Science, The University of Manchester, Manchester, U.K. He was an Assistant Professor with TU Delft, Netherlands and Project Leader with DJI, China. His research interests lie in machine learning for robotics. Dr. Pan is currently an Associate Editor for IEEE Transactions on Robotics and IEEE Robotics and Automation Letters.

Ze Ji (Member, IEEE) received the Ph.D. degree from Cardiff University, Cardiff, U.K., in 2007. He is a Reader with the School of Engineering, Cardiff University, UK. Prior to his current position, he was working in industry (Dyson, Lenovo, etc) on autonomous robotics. His research interests are broad, including autonomous robot navigation, robot manipulation, robot learning, computer vision, simultaneous localization and mapping (SLAM), acoustic localization, and tactile sensing. He is on the editorial boards of several journals, including IEEE/ASME Transactions on Mechatronics.

Robert Skilton (Member, IEEE) is a robotics and AI expert specializing in robotics for extreme and safety-critical environments. He is the UKAEA Robotics Fellow and Head of Robotics Research and Technology at the UK Atomic Energy Authority, where he leads national and international programs in fusion energy and nuclear robotics. Dr. Skilton holds a Ph.D. in generative AI for visual inspection, an M.Sc. in Cybernetics, and a B.Sc. in Computer Science. He has led major research and development (R&D) collaborations, including the UK-Japan LongOps programme, and contributed to the development of robotics strategies for future clean energy systems. He is a Chartered Engineer, Fellow of the Institution of Engineering and Technology (IET), and active in IEEE technical committees.