2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Decentralized Communication-Maintained Coordination for Multi-Robot Exploration: Achieving Connectivity and Adaptability

Wei Tang[1], Chao Li[2], Jun Wu[1], Qiuguo Zhu[1†]

***Abstract*— The realm of multi-robot autonomous exploration tasks underscores the critical role of communication in coordinating group activities. This paper introduces an innovative decentralized multi-robot exploration algorithm, meticulously crafted to ensure unbroken communication within robotic groups, a crucial element for effective coordination. The motivation for our work is two-fold: Firstly, seamless communication is vital for coordinating multi-robot autonomous exploration tasks. Secondly, in applications such as disaster rescue operations or military maneuvers, there are numerous scenarios where spatial congregation of multiple robots is imperative for joint task accomplishment. Our approach addresses these challenges through a stringent communication constraint, ensuring that each robot remains in constant communicative contact with the rest of the group. This is realized by employing a decentralized policy that integrates Graph Neural Network (GNN) layers with self-attention mechanism. Such policy network design allows adaptation to different numbers of robots and varied environments. After an initial imitation learning phase, the policy is refined through learning from experiences generated via a tree-search-based lookahead technique. Our experimental analysis validates that the algorithm not only maintains consistent communication links among all group members but also improve the exploration efficiency under the communication constraints. These results highlight the potential of our method in enhancing the effectiveness of robotic group explorations while ensuring robust communication connection.**

I. Introduction

Autonomous exploration is a critical ability for robots in tasks such as disaster rescue and military reconnaissance. Exploration with multiple robots could accelerate the process given proper coordination. There are still a lot of challenges with applying multi-robot exploration in the real-world setting. One of the most serious challenge is to coordinate among robot groups under communication constraints. Most existing work on multi-robot exploration, no matter adopting a centralized or a decentralized architecture, they require certain information to be shared among robot groups. For centralized architecture, one commanding unit will receive information from all robots in the group; for decentralized architecture, each robot needs to either exchange information with peer robots through explicit communication or implicitly observe robots which are in close vicinity. Based on the above observation, to achieve highly efficient coordination during multi-robot exploration, it is crucial to maintain communication among robots. In some previous endeavours on multi-robot autonomous exploration, communication relays are dropped by design onto certain junctures of the environment to enhance the communication network. In this case, the capacity and number of available relays limit the overall exploration performance. Others design mechanisms to aggregate robots periodically to exchange information, which undermines the exploration efficiency.

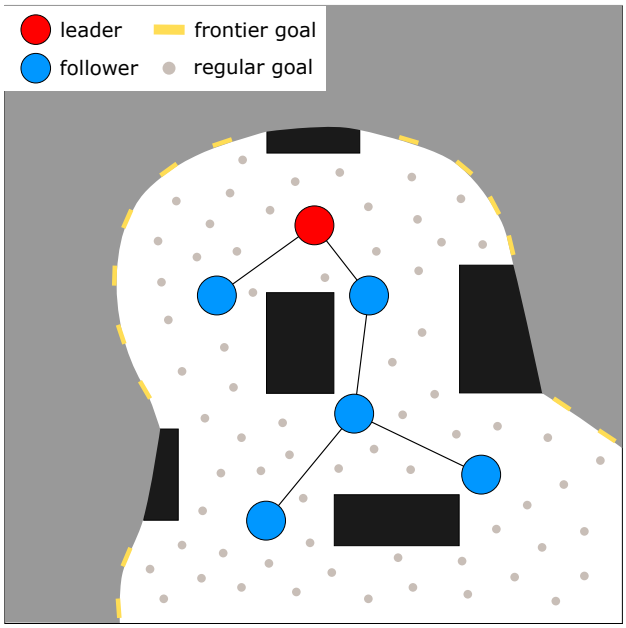


Fig. 1: Multiple robots explore the unknown environment with communication connectivity maintained. Each follower robot chooses exploration goal either from frontier goals or regular goals following the exploration policy in a decentralized fashion. The exploration policy receives pre-processed graph input which consists of local observation and peer information transmitted through communication network.

To address the above mentioned issue, this study introduces a decentralized multi-robot exploration framework that maintains continuous communication among robots by framing the exploration challenge as a Partially Observable Markov Decision Process (POMDP). At the heart of our approach is a novel policy network that integrates Graph Neural Networks (GNNs) with attention mechanisms, enhancing decision-making in dynamic and partially observable environments. The complexities inherent in this problem are illustrated in Figure 1.

Our training methodology, supported by a custom simulation environment for multi-step forward simulation, com-

[1]Wei Tang, [1]Jun Wu, [1]Qiuguo Zhu are with the Department of Control Science and Engineering, Zhejiang University, Hangzhou, China, email:tangwei66, junwuapc, qgzhu@zju.edu.cn
[2]Chao Li is with DeepRobotics Inc., email:lichao@deeprobotics.cn
†Corresponding author.

bines initial imitation learning with subsequent reinforcement learning. This sequential training strategy enables the policy network to rapidly assimilate expert knowledge and further refine decision-making capabilities through environmental interaction.

The framework's versatility is demonstrated through simulation experiments, showcasing its ability to adapt to diverse environments and robot group configurations while maintaining communication and completing exploration tasks effectively.

The remaining of the paper is organized as follows. Section II introduces the past related work. Section III presents the problem setup and notation. Section IV presents the proposed learning-based multi-robot exploration framework. Section V presents the experiments and results analysis, and section VI concludes the work.

II. Related Work

Exploration of Unknown Environments: The problem of robotic autonomous exploration has been approached from different angles. As stated in one of the earliest work on exploration [1], the core question of exploration is: *Given what you know about the world, where should you move to gain as much new information as possible?* Existing methods could be categorized into different groups based on how they deal with this issue. One classical set of works chose next exploration target based on frontiers [1]–[4], i.e. regions on the boundary between open space and unexplored space. These works investigated obtaining frontiers under conditions such as 2d/3d space, robot dynamic constraints and different sensor configurations. In this work, we also treat frontier computation as a basic function block and associate target locations with frontiers. Instead of explicitly computing frontiers, another set of works approach the problem by finding the target location that maximizes the information gain [5]–[7]. Specifically, Bircher et al. propose the Next-Best-View Planner(NBVP) [8]. The method combines the ideas of random exploring tree and maximizing information gain, and shows promising performance while being capable of running online, onboard a robot with limited resources.

Multi-Robot Coordination during Exploration: Exploration with multiple robots could increase the efficiency. In one of the earliest work on multi-robot exploration [9], robots merge local maps from teammates to obtain the global map and target the closest frontier. Lack of coordination on target selection introduces potential conflicts among robots and limits overall efficiency. To tackle this issue, several work propose frontier assignment algorithms [10]–[14]. In [11], [12], robots are assigned based on distance costs relative to frontiers using greedy and Hungarian method. In [14], authors argue that using information gain alone could improve the performances of exploration. Instead of taking advantage of current environment information alone, [10] and [15] introduce the coordination of teammate robots' future target choices by reducing an area's utility if it is chosen as any other robot's target. The above mentioned work all adopt centralized architecture. They have access to shared global map and frontiers but suffer from the single point of failure. On the other hand, distributed exploration systems enjoy better robustness but require complex cooperation mechanisms. Frequent communication between robots can also be a problem. In [16], the robots form an exploration cluster after establishing connection and successfully matching their maps. Each exploration cluster has a leader to collect information and make global assignments. When two clusters meet, a new leader will be selected. However, robots still need to share environmental information between each other. [17] proposed another method in which each robot calculates the cost of the target point relative to other robots when making a decision. If other robots could explore the target more efficiently, the target will not be selected. In [18], each robot computes a Voronoi partition based on their current positions, and only selects those targets located within the robot's own Voronoi region. However, the above methods all require environment information such as maps and frontiers to be transmitted between robots, which is a significant burden for the communication network, especially in communication-constrained environments such as disaster sites.

The integration of learning-based methods in multi-robot exploration has seen significant advancements. A centralized policy network employing Spatial Graph Neural Networks (GNNs) marks a pioneering effort in this domain, demonstrating the potential for GNNs to generalize well to larger maps and robot teams, despite its reliance on a centralized decision-making framework that does not account for communication constraints [19].

Further development in this field introduced Hierarchical-Hops Graph Neural Networks (H2GNN), which utilizes a multi-head attention mechanism and multi-agent reinforcement learning (MARL) to enhance decision-making and cooperation among robots. This method advances the use of MARL in exploration but similarly overlooks the challenges posed by communication constraints, assuming global information availability [20].

NeuralCoMapping's contribution to multi-robot active mapping, combining bipartite graph matching with multiplex GNNs, showcases superior mapping performance and efficiency. However, it operates under a centralized framework with complete information, sidestepping the practicalities of communication limits in decentralized explorations [21].

The exploration of decentralized approaches is evident in Multi-Agent Neural Topological Mapping (MANTM), which employs a novel RL-based Hierarchical Topological Planner (HTP) for efficient cooperative exploration. Despite its promise and reduction in exploration steps, the synchronization requirement for decision-making among robots hints at an implicit assumption of uninterrupted communication, a condition not always guaranteed in field applications [22].

Collectively, these studies highlight the significant advancements in applying GNNs and RL to multi-robot exploration, setting a solid foundation for future research. Despite their remarkable contributions, many of these approaches operate under the assumptions of centralized control or

seamless communication. Our work aims to complement these pioneering efforts by introducing a focus on maintaining communication connectivity within a decentralized exploration framework. This perspective seeks to add another dimension to the robustness and adaptability of multi-robot systems, particularly in scenarios where communication challenges are prevalent.

Multi-Robot Coordination with Communication Constraint: The challenge of decentralized multi-agent path planning, particularly in avoiding collisions while maintaining communication among agents, is addressed by Tuck et al. [23], who propose a novel method tailored for this purpose. On a related note, Yang et al. [24] introduce a centralized optimization framework designed to maintain a minimal set of communication edges, thereby calculating minimal control revisions in relation to a nominal controller. However, Sun et al. [25] approach the problem by incorporating line-of-sight (LOS) connectivity as a nonlinear constraint within a mixed integer problem, a method that, while innovative, faces challenges in real-time computation and scalability with an increasing number of agents.

Further exploring LOS-aware strategies, Shetty et al. [26] and Gao et al. [27] present formation and connectivity control solutions where robot behaviors are significantly influenced by the need to maintain connectivity. This connectivity-oriented design, however, can sometimes hinder the robots' ability to progress toward their primary objectives, especially in scenarios requiring them to spread out across extensive areas.

Addressing the practical aspect of maintaining communication in exploration tasks, especially in environments with communication restrictions (e.g., subterranean settings), Saboia et al. [28] propose an innovative networking solution. A key component of this solution involves autonomously deploying relay radios to enhance multi-robot exploration performance, showing the critical importance of communication maintenance. The empirical results from their experiments provide compelling evidence of the benefits of such communication strategies in challenging exploration contexts.

III. Problem Setup and Notation

Define $Env \subset R^2$ as the work space to be explored. It is initially-unknown, continuous, and bounded. The interior areas of the Env that belong to obstacles are denoted by Env_o. Thus, the free, traversable area is denoted by $Env_{free} = Env/Env_o$. A team of m mobile robots $R = r_1, r_2, ..., r_m$ is deployed in Env. Each mobile robot is equipped with finite-range sensors able to perceive the surrounding space(e.g., laser range scanners or depth cameras) and is able to navigate freely within Env_{free}. Within the team, there is one leader robot, and the other robots are follower robots. The leader robot is assumed to have better communication and computation capabilities so that it either receives remote command from human or make exploration decisions autonomously. The follower robots could not communicate with remote human operator, and make local exploration decisions to maintain group communication connectivity among the team. As time evolves in discrete steps $t \in 1, 2, ..., T$, where T denotes the last step of the exploration mission. Denote v_i^t as the pose of robot i at time step t and $p_i^t \subset Env_{free}$ as the area within Env_{free} perceived by robot i at v_i^t. At time step $\bar{t}$, the area within Env_{free} that has been perceived by the team of robots is denoted by:

$$M^{\bar{t}} = \bigcup_{i=1,2,...,m} \bigcup_{t=0,1,...,\bar{t}} p_i^t$$

In terms of communication constraint, we first adopt the following definitions, then formally define the communication graph:

1) **Pairwise Direct Communication constraint:** The distance between the robots is within a maximum communication range, ρ (i.e., $\|x_{i,t} - x_{j,t}\|_2 \leq \rho$).
2) **Pairwise Multi-Hop Communication constraint:** there exists at least one path between the pair of robots, each node on the path is a robot in the group. For each edge along the path, the two robots on the ends of the edge satisfy the Pairwise Direct Communication constraint.
3) **Group Communication Maintained constraint:** For each pair of robots in the team, they satisfy either the Pairwise Multi-Hop Communication constraint or the Pairwise Direct Communication constraint.

In the context of our communication-maintained multi-robot exploration task, the above communication-related constraints could be expressed using the notion of communication graph. Formally, we define the communication graph as $G_c = (V, E)$, where V represents the set of vertices, with each vertex corresponding to a robot in the exploration team, and E denotes the set of edges. An edge $(v_i, v_j) \in E$ exists if and only if the corresponding robots i and j are within the maximum communication range ρ, thus satisfying the Pairwise Direct Communication constraint. Under the graph structure representation, when the group communication maintained constraint is satisfied, the communication graph is connected. This facilitates the planning of exploration paths that ensure continuous group connectivity under the defined communication constraints.

The communication-maintained multi-robot exploration task is to plan for each robot the sequence of poses to reach sequentially, so as the time evolves, at some time step T, M^T equals Env_{free}. And in the meanwhile, the communication graph is connected throughout the exploration process.

In this work, we focus on coordinating the exploration decisions for teams comprising of single leader robot and follower robots. In reality, the leader robot may receive a human command through remote teleoperation or operate autonomously following an leader exploration policy. The main problem we try to resolve in this work is to design decentralized exploration policy for the follower robots. The goal is to increase the exploration efficiency under the hard constraint of maintaining communication connectivity.

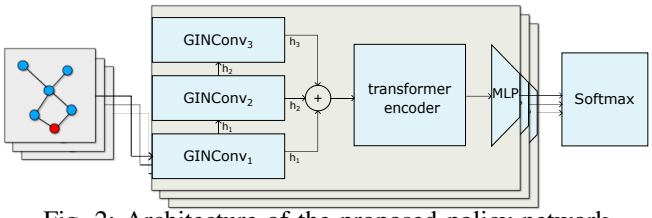


Fig. 2: Architecture of the proposed policy network.

IV. Methods

A. Multi-agent Partially Observed Markov Decision Process

Formally, communication-maintained decentralized coordination for multi-robot exploration could be formulated as a decentralized multi-agent extension of a Partially Observed Markov Decision Process(POMDP) [29], given as a tuple

$$(\mathcal{S}, \mathcal{A}^N, \mathcal{O}, \mathcal{R}, \mathcal{T}, \rho, \mathcal{N}, \gamma)$$

, where agents are follower robots in the team. N is the number of follower robots. $s_i \in S_i$ is the state of the i-th follower robot. At each step, a follower robot $i \in \mathcal{N}$ receives observation $o_i \in O_i$, which consists of a range-limited local observation and information transmitted from other robots in the group. and takes an action $a_i \in A$ given by policy $\pi_\theta(a_i|o_i)$, where θ represents the set of parameters. Each robot in the group executes actions in a decentralized and asynchronous manner.

The action of a follower robot is defined as selecting the next exploration goal, towards which it will plan and execute a path. In the team, robots do not operate synchronously. When a follower robot reaches its last exploration goal, it immediately makes a new action to select and move to its next goal. The observation of i-th follower robot consists of its current potential exploration goals, the estimated exploration gain for each potential exploration goals and other robots' current paths. The reward design, integral to our Partially Observable Markov Decision Process (POMDP) formulation, is specified as follows:

- A penalty of -0.5 is assigned if the execution of the current robot's action results in the disconnection of the communication graph, emphasizing the importance of maintaining connectivity.
- A slight negative reward of -0.1 is applied for each action step to encourage efficient exploration by minimizing the number of steps taken.
- A positive reward is assigned proportionally to the area newly discovered during the execution of an action, thereby encouraging the exploration of previously uncharted territories. Specifically, this reward is calculated as the ratio of the newly explored area to half of the sensor's coverage area.

Within the decentralized multi-robot exploration framework proposed, the rewards—both the penalty for communication disconnection and the positive incentive for new area exploration—are globally determined, depends on states of the team's robots. This framework facilitates the development of decentralized policies for follower robots through a combination of imitation learning and reinforcement learning methodologies. The policy network and its training process are elaborated in the subsequent section.

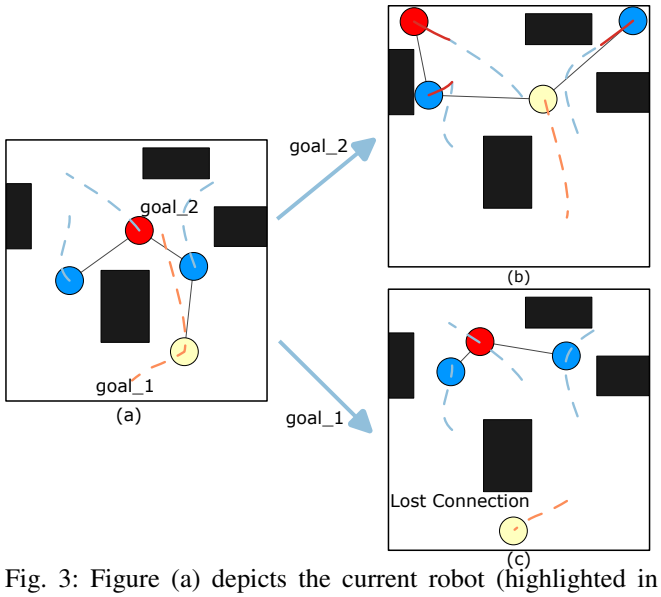


Fig. 3: Figure (a) depicts the current robot (highlighted in yellow) alongside two potential exploration paths delineated by orange dotted lines, leading to goals 1 and 2. Figures (b) and (c) present the outcomes of a one-step forward simulation. In (b), the path to goal 2, being longer than the remaining paths of other robots, prompts an estimation of their future positions by moving away from the group's centroid, with the estimated path segments marked in dark red.

B. Pre-processing of Inputs and Policy Network Architecture

The large state space of the multi-robot group exploration problem motivates a deep RL approximate solution. Employing a centralized training with decentralized execution (CTDE) strategy, we designed a policy network for the follower robots. This network processes the robot's observation data to determine its exploration goal. In this work, the decision-making process for robots in the team is asynchronous. After a robot arrives at its previously exploration goal, it proceeds to make the next decision. At this moment, other robots in the team are in the process of moving to their last exploration goals. Since the robotic team maintains communication connectivity throughout the process, robots can exchange information to aid in decision-making. The exchanged information includes each robot's current location, the remaining exploration path from the last decision, and maps of newly explored area.

The following section introduces the pre-processing of inputs and the architecture of the policy network. The observation data for a follower robot includes its own exploration map, potential exploration goals at the current moment, and the information communicated from other robots in the team as described above. This work extends our previous efforts [30], leveraging the same frontier detection module. For the purpose of maintaining connection, we now also uniformly sample regular goal points within the explored areas.

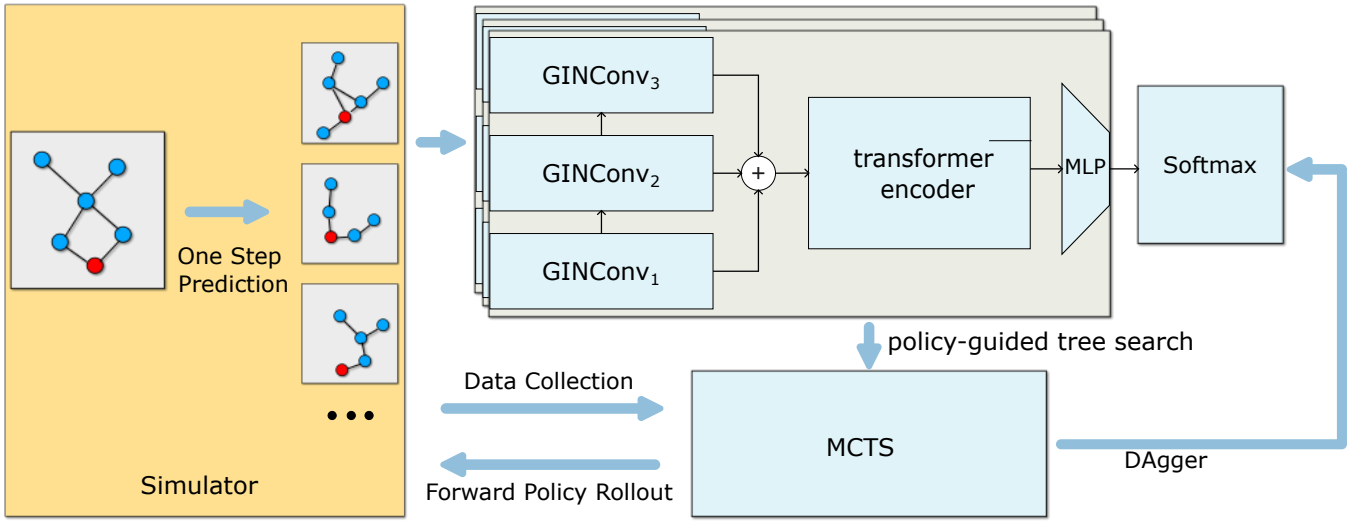


Fig. 4: The training pipeline employs MCTS-based reinforcement learning, where MCTS utilizes the simulator for forward policy rollouts to estimate the cumulative future rewards of potential actions. These estimates inform the training of the policy network, following the DAgger imitation learning algorithm's methodology for integrating expert behavior.

Collectively, these elements form the potential exploration goals.

The pre-processing of observation data involves, for each potential exploration goal of the current robot, integrating the remaining paths of other robots in the team and using a simulator for forward prediction. This predicts the team's relative formation and map exploration status upon the current robot completing that potential exploration goal. The process of the forward prediction is depicted in Figure 3. The estimated future formation and map exploration information can be encoded into the input graph for Graph Neural Network layers, denoted as G_i for potential exploration target T_i. Let $x = [r, \theta, E]$ denote the feature vector of a graph vertex. The components of x are defined as follows:

$$r = \frac{r_{\text{follower}}}{r_{\max}},$$
$$\theta = \theta_{\text{follower}} - \theta_{\min},$$
$$E = \frac{\text{explored area}}{\text{sensor coverage area}}.$$

In this context, r and θ are normalized polar coordinates of the follower robot, with the leader robot serving as the origin of the polar coordinate system. E denotes the ratio of the newly explored area of the follower robot to the sensor coverage area, which is crucial for enhancing exploration efficiency.

The set of G_i corresponding to all potential exploration goals constitutes the input to the policy network.

The architecture of the proposed policy network is optimized for processing observation data from robots to select exploration goals. It is illustrated in Figure 2. Its structure comprises:

- **Graph Isomorphism Network (GIN) Layers:** Three GINConv layers form the core of the network, each employing a Multi-Layer Perceptron (MLP) for feature transformation. The MLPs transform input dimensions to a hidden dimension (`dim_h`), incorporating batch normalization and ReLU activations to enhance feature extraction.
- **Pooling and Concatenation:** Features from each GIN layer are pooled (using sum pooling) and concatenated, creating a unified feature vector for subsequent processing.
- **Transformer Encoder:** The pooled features are then processed by a Transformer encoder, comprising multiple layers with `dim_h*3` model dimension. This configuration allows for capturing long-range dependencies within the data.
- **Linear Layers:** Post-Transformer, two linear layers further refine the features. The first maintains the dimension at `dim_h*3`, and the second reduces it to a single value indicative of the exploration goal.
- **Softmax Operator:** Following the linear layers, a softmax operator is applied to the logits from different graph inputs. This operator outputs the probability of selecting a particular potential exploration goal, enabling a probabilistic decision-making process among the available goals.

C. Training pipeline of the policy network

Due to the large state space associated with this problem, initiating training directly through reinforcement learning (RL) poses significant challenges. Thus, the training of the policy network employs a hybrid approach that combines imitation learning and reinforcement learning. In the imitation learning phase, for input graphs obtained through forward prediction in the pre-processing stage, an expert policy computes logit values. These logits are then processed through a softmax function to derive decisions.

The expert policy is defined as follows:

1) If the forward-predicted graph indicates the current robot and the leader robot are connected, the logit is

Algorithm 1 Hybrid Training Process with DAgger and MCTS for Policy Networks

Initialize the policy network.
while policy network has not converged **do**
  Perform a rollout with the current policy network in a simulated environment.
  for each step in the rollout **do**
    Explore future actions using MCTS:
      Select a node for expansion based on the UCB criterion.
      if the depth of the selected node is less than `expand_max_depth` **then**
        Expand the tree by adding child nodes to the selected node.
      end if
      Execute a policy rollout from the selected node up to `rollout_limit` depth, accumulating rewards to calculate `sum_of_reward`.
      Backpropagate the `sum_of_reward` through the path, updating node values.
      Perform `mcts_iteration_num` iterations of MCTS exploration and backpropagation.
  end for
  Update the policy network by treating the values of explored child nodes as target logits in supervised learning.
end while

set to 0.

2) If the forward-predicted graph shows the current robot and the leader robot are not connected, the logit is equal to the minimum movement distance required for the current robot to reconnect with the leader robot.

Following the imitation learning phase, the training utilizes the DAgger (Dataset Aggregation) algorithm, with the policy network's performance evaluated based on the returns from testing environments. Once the episode return reaches convergence, the process proceeds with reinforcement learning training.

The expert policy employed during imitation learning is limited to considering communication graph connection from a single forward prediction step without optimizing the exploration efficiency. Given the deterministic system transition model within the constraints of the multi-robot cooperative exploration environment and the availability of a highly efficient simulation environment, a model-based RL method akin to that used in AlphaZero [31] is applied. The reinforcement learning training pipeline is depicted in Figure 4.

In the reinforcement learning phase detailed in Algorithm 1, our approach leverages an adaptation of the Monte Carlo Tree Search (MCTS) technique for enhanced strategic planning and decision-making in multi-robot exploration. Through a sequence of simulated rollouts confined within a specified lookahead window, this methodology facilitates the examination of diverse exploration strategies. The MCTS process, as outlined in the algorithm, employs the Upper Confidence Bound (UCB) criterion to judiciously select and expand nodes, fostering an optimal balance between exploring novel actions and exploiting known beneficial ones.

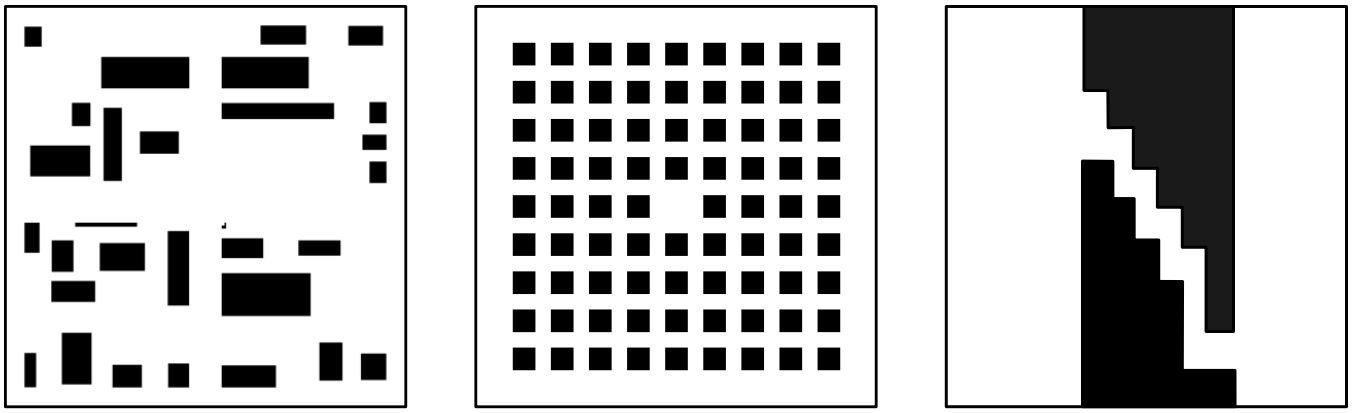
Within each simulation step, the MCTS algorithm iteratively refines the exploration decision-making process by assessing the potential rewards of future actions. This is achieved through a structured procedure of node expansion, tree backpropagation, and iterative exploration, culminating in a comprehensive dataset of action-reward pairings. These pairings serve as the foundation for the supervised update of the policy network, significantly enhancing the robots' decision-making capabilities. The iterative reinforcement process underscored by Algorithm 1 is instrumental in advancing the policy network's proficiency in making well-informed decisions, ultimately leading to the development of more effective exploration strategies and the improvement of the collective performance of the multi-robot system.

V. Evaluation

The effectiveness of the proposed method is validated in simulation. The simulator is developed on top of the one used in our previous work [30] with additional support for the multi-agent POMDP formulation. Notably, as our proposed method preserves communication connectivity, it allows for the assumption that each robot's individually explored map is shared among all robots within the group. As mentioned previously, the formulation of the policy network input enables the policy to adapt to environments with varying size. For the training, environments with fixed dimensions and randomly generated obstacles are employed. One sample setup of the training environment is shown in Figure 5a. The environments are split into training and test sets, each containing 50 different environments. The reported performance metrics were obtained by averaging results from 10 runs for each test environment. For the leader robot, a simple greedy policy was employed, always choosing the closest frontier goal to explore.

A. Training performance of the proposed policy network and its variants

Given the innovative nature of our problem formulation, direct comparisons with existing methods are not immediately applicable. Therefore, the evaluation concentrates on internal comparisons, contrasting different configurations of our proposed system to underscore the significance of each component. Baseline variations include omitting the transformer encoder from the policy network and modifying the policy input pre-processing technique, such as using the last exploration goals instead of predicted future poses for robots nearing the completion of their paths. These comparative studies underscore the efficacy of our design. For the comparative analysis, experiments were conducted with 5 robots. The accumulated reward during training is depicted in Figure 6, with the results averaged over three random seeds to account for variability, showcasing both



(a) sampled training map (b) map with evenly-placed obstacles (c) map with corridor

Fig. 5: Environment for testing the exploration methods.

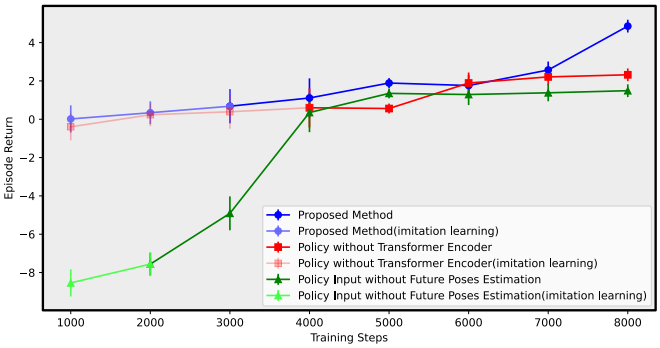


Fig. 6: Training curve showing episode returns for 5 robots comparing different system configurations.

mean and standard deviation. Evaluation metrics comprise the communication connectivity ratio during exploration and the task makespan—defined as the elapsed time or the largest distance traveled by robots in the group. Results for the final policy version are detailed in Table I. These results showed that the attention mechanism in the transformer encoder helps with learning complex and better policy. And the estimation of the future poses is critical for the system both in the imitation learning phase and for the final policy performance. The comparison between the proposed system and the system with imitation learning only shows the model-based reinforcement learning phase greatly enhances the efficiency while maintaining the connectivity of communication graph.

Configuration	Communication Ratio	Makespan
Proposed System	96%	833
No Transformer Encoder	93%	875
No Future Poses Estimation	89%	945
Imitation Learning Only	95%	1203

TABLE I: Comparison of communication connectivity and makespan for three system configurations.

B. Validation Across Different Numbers of Robots

We further validate the adaptability of our proposed policy network by training with mixed number of robots. The results are reported in Table II. It showed the proposed policy network architecture could adapt to different number of robots with reasonable performance, which is important for ease of deployment.

Number of Robots	Communication Ratio	Makespan
3 Robots	95.3%	1242
4 Robots	95.5%	1046
5 Robots	95.4%	853
6 Robots	94.5%	832
7 Robots	94.7%	790
8 Robots	94.5%	784

TABLE II: Performance comparison across different numbers of robots.

C. Adaptability to Different Environmental Sizes and Special Unseen Environments

Finally, we deploy the trained policy network in various environments of different sizes. The outcomes confirm that the graph feature design enable the policy network to naturally work with environments with varying sizes, highlighting its robust adaptability. Also, experiments conducted with special unseen environments depicted in 5b and 5c further proved the robustness and adaptability of the proposed system.

VI. Discussion and Future Work

In this study, we have proposed a novel policy network design for facilitating multi-robot grouped exploration tasks, with a specific focus on maintaining communication connectivity. Through the integration of imitation learning and reinforcement learning in a mixed training paradigm, our approach has demonstrated significant effectiveness across various simulated scenarios. The adaptability of our method to different numbers and configurations of robots underscores its potential for a wide range of exploration tasks.

The results of this work lay the groundwork for further exploration into more versatile multi-robot tasks. Future research will extend the learning methodology to accommodate a broader spectrum of multi-robot operations, aiming to

enhance the autonomy and efficiency of robotic systems in complex environments.

Conclusively, our work contributes a significant advancement in the development of intelligent, autonomous multi-robot systems. As we progress towards real-world applications, we anticipate new insights and developments that will further the capabilities of multi-robot systems in performing coordinated exploration tasks with an emphasis on maintaining communication connectivity.

ACKNOWLEDGMENT

This work was supported by the "Leading Goose" R&D Program of Zhejiang (Grant No. 2023C01177), the National Key R&D Program of China (Grant No. 2022YFB4701502), the Key R&D Project on Agriculture and Social Development in Hangzhou City (Asian Games) (Grant No. 20230701A05), and the Key Research Project of Zhejiang Lab (Grant No. 2021NB0AL03).

REFERENCES

[1] B. Yamauchi, "A frontier-based approach for autonomous exploration," in *Proceedings 1997 IEEE International Symposium on Computational Intelligence in Robotics and Automation CIRA'97.'Towards New Computational Principles for Robotics and Automation'*. IEEE, 1997, pp. 146–151.

[2] C. Dornhege and A. Kleiner, "A frontier-void-based approach for autonomous exploration in 3d," *Advanced Robotics*, vol. 27, no. 6, pp. 459–468, 2013.

[3] L. Heng, A. Gotovos, A. Krause, and M. Pollefeys, "Efficient visual exploration and coverage with a micro aerial vehicle in unknown environments," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2015, pp. 1071–1078.

[4] T. Cieslewski, E. Kaufmann, and D. Scaramuzza, "Rapid exploration with multi-rotors: A frontier selection method for high speed flight," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 2135–2142.

[5] F. Bourgault, A. A. Makarenko, S. B. Williams, B. Grocholsky, and H. F. Durrant-Whyte, "Information based adaptive robotic exploration," in *IEEE/RSJ international conference on intelligent robots and systems*, vol. 1. IEEE, 2002, pp. 540–545.

[6] W. Tabib, M. Corah, N. Michael, and R. Whittaker, "Computationally efficient information-theoretic exploration of pits and caves," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 3722–3727.

[7] S. Bai, J. Wang, F. Chen, and B. Englot, "Information-theoretic exploration with bayesian optimization," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 1816–1822.

[8] A. Bircher, M. Kamel, K. Alexis, H. Oleynikova, *et al.*, "Receding horizon" next-best-view" planner for 3d exploration," in *2016 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2016, pp. 1462–1468.

[9] B. Yamauchi, "Frontier-based exploration using multiple robots," in *Proceedings of the second international conference on Autonomous agents*, 1998, pp. 47–53.

[10] R. Simmons, D. Apfelbaum, W. Burgard, D. Fox, *et al.*, "Coordination for multi-robot exploration and mapping," in *Aaai/Iaai*, 2000, pp. 852–858.

[11] A. Solanas and M. A. Garcia, "Coordinated multi-robot exploration through unsupervised clustering of unknown space," in *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, vol. 1. IEEE, 2004, pp. 717–721.

[12] K. M. Wurm, C. Stachniss, and W. Burgard, "Coordinated multi-robot exploration using a segmentation of the environment," in *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2008, pp. 1160–1165.

[13] A. Bautin, O. Simonin, and F. Charpillet, "Minpos: A novel frontier allocation algorithm for multi-robot exploration," in *International Conference on Intelligent Robotics and Applications*. Springer, 2012, pp. 496–508.

[14] J. Faigl and M. Kulich, "On determination of goal candidates in frontier-based multi-robot exploration," in *2013 European Conference on Mobile Robots*. IEEE, 2013, pp. 210–215.

[15] W. Burgard, M. Moors, C. Stachniss, and F. E. Schneider, "Coordinated multi-robot exploration," *IEEE Transactions on robotics*, vol. 21, no. 3, pp. 376–386, 2005.

[16] D. Fox, J. Ko, K. Konolige, B. Limketkai, D. Schulz, and B. Stewart, "Distributed multirobot exploration and mapping," *Proceedings of the IEEE*, vol. 94, no. 7, pp. 1325–1339, 2006.

[17] C. Wei, K. V. Hindriks, and C. M. Jonker, "Dynamic task allocation for multi-robot search and retrieval tasks," *Applied Intelligence*, vol. 45, no. 2, pp. 383–401, 2016.

[18] J. Hu, H. Niu, J. Carrasco, B. Lennox, and F. Arvin, "Voronoi-based multi-robot autonomous exploration in unknown environments via deep reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 14 413–14 423, 2020.

[19] E. Tolstaya, J. Paulos, V. Kumar, and A. Ribeiro, "Multi-robot coverage and exploration using spatial graph neural networks," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 8944–8950.

[20] H. Zhang, J. Cheng, L. Zhang, Y. Li, and W. Zhang, "H2gnn: hierarchical-hops graph neural networks for multi-robot exploration in unknown environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3435–3442, 2022.

[21] K. Ye, S. Dong, Q. Fan, H. Wang, L. Yi, F. Xia, J. Wang, and B. Chen, "Multi-robot active mapping via neural bipartite graph matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 839–14 848.

[22] X. Yang, Y. Yang, C. Yu, J. Chen, J. Yu, H. Ren, H. Yang, and Y. Wang, "Active neural topological mapping for multi-agent exploration," *IEEE Robotics and Automation Letters*, vol. 9, no. 1, pp. 303–310, 2023.

[23] V. Tuck, Y. V. Pant, S. A. Seshia, and S. S. Sastry, "Dec-los-rrt: Decentralized path planning for multi-robot systems with line-of-sight constrained communication," in *2021 IEEE Conference on Control Technology and Applications (CCTA)*. IEEE, 2021, pp. 103–110.

[24] Y. Yang, Y. Lyu, and W. Luo, "Minimally constrained multi-robot coordination with line-of-sight connectivity maintenance," *arXiv preprint arXiv:2303.04271*, 2023.

[25] X. Sun, D. Ding, *et al.*, "Optimal coverage control of multi-agent systems in constrained environments with line-of-sight connectivity preservation," in *2020 39th Chinese Control Conference (CCC)*. IEEE, 2020, pp. 4616–4621.

[26] A. Shetty, T. Hussain, and G. Gao, "Decentralized connectivity maintenance for multi-robot systems under motion and sensing uncertainties," *NAVIGATION: Journal of the Institute of Navigation*, vol. 70, no. 1, 2023.

[27] Z. Gao and G. Guo, "Velocity free leader-follower formation control for autonomous underwater vehicles with line-of-sight range and angle constraints," *Information sciences*, vol. 486, pp. 359–378, 2019.

[28] M. Saboia, L. Clark, V. Thangavelu, J. A. Edlund, K. Otsu, G. J. Correa, V. S. Varadharajan, A. Santamaria-Navarro, T. Touma, A. Bouman, *et al.*, "Achord: Communication-aware multi-robot coordination with intermittent connectivity," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 184–10 191, 2022.

[29] C. Boutilier, "Planning, learning and coordination in multiagent decision processes," in *TARK*, vol. 96. Citeseer, 1996, pp. 195–210.

[30] W. Tang, C. Xue, C. Li, and Q. Zhu, "Towards coordinated multi-robot exploration under bandwidth-constrained conditions," in *2022 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*. IEEE, 2022, pp. 180–187.

[31] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, *et al.*, "Mastering chess and shogi by self-play with a general reinforcement learning algorithm," *arXiv preprint arXiv:1712.01815*, 2017.