2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

STAGE: Structure-Adaptive Graph-Encoded Multi-Agent Policy Gradient for Moving Target Search in Uncertain Topological Networks

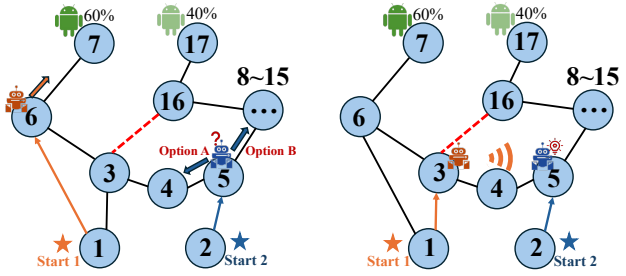
Qihang Peng[1†], Lizhou Zhu[2†], Lekai Chen[3], Hongliang Guo[4] and Chih-Yung Wen[1]

***Abstract*— This paper investigates the multi-robot efficient search (MuRES) problem in uncertain topological networks. One unique characteristic of the studied problem is that the topology of the underlying network is uncertain, posing great challenges to canonical MuRES solutions which presumes a fixed network topology. To address the challenge, this paper proposes the STructure-Adaptive Graph-Encoded policy gradient (STAGE) algorithm for moving target search. STAGE comprises two main components: (1) the bi-scale graph attention network (GAT) encoder, which fuses a k-hop local GAT with a distance-augmented long-range GAT to enable the encoder to capture both local and long-range network structural changes; and (2) the entropy-regularized counterfactual policy gradient module, which employs a structure-aware centralized critic to estimate both the team returns and the network structure information, and train the decentralized actors via counterfactual marginalization with entropy regularization. Extensive simulation results and physical experiment demonstrate the feasibility and superiority of STAGE for solving MuRES in uncertain topological environments.**

I. Introduction

Multi-robot efficient search (MuRES) has been deemed as a prominent application of multi-robot systems, attracting sustained attention from both academia and industry over the past several decades. From the academic perspective, MuRES serves as an important research direction of multi-robot collaboration, situated at the intersection of key fields such as multi-agent reinforcement learning, swarm intelligence, and operations research. From the practical standpoint, MuRES can be instantiated across diverse real-world scenarios such as search and rescue, patrolling and surveillance, and environmental exploration.

Research in MuRES has seen significant progress in recent years, and a brief literature review will be provided in Section II. Here, we would like to articulate that one common research gap is that almost all existing MuRES solutions presume a constant network topology, which might not be the case in many real world multi-robot search scenarios. For example, in the fire rescue use case, the previously passable edge might become blocked because of the falling furniture.



(a) Canonical Methods (b) STAGE Method

Fig. 1. An illustrative example: Two robots start at Node 1 and Node 2 separately, to search for one target residing in either Node 7 (60%) or Node 17 (40%). Edge $(3, 16)$ is uncertain, whose traversability is revealed only when a robot reaches Node 3 or Node 16. In the canonical search methods, Robot 1 would take path $(1 \to 6 \to 7)$ to Node 7 with highest target existence probability, while Robot 2 takes path $(2 \to 5)$ first and then chooses between: *Option A*, probing via $(5 \to 4 \to 3)$ to test Edge $(3, 16)$ potentially unlocking a fast route to Node 17; or *Option B*, committing to the long but reliable route $(5 \to 8 \to 9 \to 10 \to ... \to 16 \to 17)$. However, in STAGE, Robot 1 first moves to Node 3 to help test the vulnerable edge and broadcast the result to Robot 2. Robot 2 then conditionally selects the high-value route if passable or the safe route otherwise. This coordinated information gathering reduces expected capture time under topology uncertainty.

When facing uncertain network topology, researchers often deem the environment as completely unknown, and treat it as a multi-robot exploration problem. However, the strategy has two obvious drawbacks: (1) completely discarding the graph's prior knowledge as well as the target dynamics reduces the search efficiency; (2) multi-robot exploration primarily optimizes mapping/coverage rather than accelerating search efficiency, *i.e.*, the task objective is different. Conversely, simply assuming an unchanging topology and applying standard MuRES algorithms is also inefficient: although prior maps enable modeling target dynamics and crafting efficient strategies, real search and rescue settings frequently suffer from topology damage. Potential changes in the graph structure create discrepancies with the MuRES solution, which can substantially reduce the efficiency of search strategies or even cause complete task failure.

To address the aforementioned challenges, we propose structure-adaptive graph-encoded policy gradient (STAGE), a multi-agent reinforcement learning based algorithm tailored for MuRES problem in uncertain topological environments. The STAGE algorithm consists of two main modules: (1) the bi-scale GAT encoder, which fuses a k-hop local graph attention network (GAT) with a distance-augmented long-range GAT. The augmented graph is updated whenever newly

†These authors contributed equally to this work.
[1]Department of Aeronautical and Aviation Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, China. qihang1.peng@connect.polyu.hk, cywen@polyu.edu.hk. Corresponding author: Chih-Yung Wen.
[2]Glasgow College, University of Electronic Science and Technology of China (UESTC), Chengdu, China, 611731.
[3]School of Information and Communication Engineering, University of Electronic Science and Technology of China (UESTC), Chengdu, China, 611731.
[4]College of Computer Science, Sichuan University (SCU), Chengdu, China, 610064.

observed edges shorten shortest-path distances, enabling the encoder to capture both neighborhood and long-range structural changes; and (2) the entropy-regularized counterfactual policy gradient, in which a structure-aware centralized critic learns both team return and graph-information signals, and decentralized actors are trained via counterfactual marginalization with entropy regularization. The comparison of STAGE algorithm with canonical MuRES method is illustrated in Fig. 1.

The paper's main contributions are summarized as follows: (1) We present STAGE, a new MuRES solution explicitly tailored for uncertain topological environments. (2) We introduce distance-augmented GAT, which effectively captures long-range structural changes while alleviating oversmoothing common to global graph networks. (3) We integrate entropy regularization into the actor update, a simple yet effective mechanism to promote agent's exploration and stabilize the learning process.

II. Literature Review

This section provides an overview of multi-robot efficient search (MuRES). We first outline typical MuRES configurations and then summarize prevailing MuRES methodologies.

A. MuRES Configurations and Variants

According to the actual search tasks and robotic platform, MuRES admits multiple configurations, yielding several variants: (1) Environmental modelling: the search environment can be modeled as a graph [1]–[6], a grid map [7]–[12], or continuous space [13]–[19]. (2) Robot characteristics: based on specific robot platform, robot teams may be homogeneous [3] or heterogeneous [7], [15]. The equipped sensors led to different sensing models include circular range [8], [9], or angled line-of-sight [17], [20]. In graph settings, sensing is often abstracted as detection on the current node [4] or within k-hop neighborhoods [2]. (3) Target motion dynamics: target can be stationary [7], [21], stochastically moving [2], [10] or adversarial/evasive [6], [13], [22]. (4) Task-related objectives: The task-related objectives depend on the urgency of search. For time-sensitive tasks, the objective can be maximize the probability within the time budget [9], [11]. Others may concentrate on minimizing the capture time expectation to improve the average performance [3], [23]. Notably, multi-robot exploration is related but pursues different objectives, see [24], due to space constraints we omit related literature.

B. MuRES Methodologies

Existing MuRES algorithms can be divided into three distinct categories:

1) Planning Methods: In this category, the MuRES is formulated as an optimization problem and then utilize an off-the-shelf solver to obtain optimal or near optimal solutions [2], [4], [14], [23]. The principal advantage is modeling flexibility: objectives and constraints can be tailored to the specific search scenario. However, the planning methods require comprehensive prior information to establish optimization problem which is usually impossible in real search tasks.

2) Heuristic Swarm Intelligence: In this group, researchers are often inspired from nature, and focus on the integration between robots to form cooperative algorithms [12], [13], [16], [18], [25], [26], such as grey wolf optimizer(GWO) [13], Bird Flocking [25], PSO [26], and anti-optimization [12]. The heuristic mechanisms focus on the cooperation mechanism within the robots, thereby usually being highly efficient. However, the simplification of search problem might lead to local optima.

3) Reinforcement Learning Methods: RL-based methods are another stream of MuRES solutions [3], [5], [8], [11], [15], [22], [27]–[29]. RL algorithms can also be divided into value-based methods, such as DQN [30], DRL [5], and QPLEX [9] and policy-based methods such as PG [3], PPO [15], and MAAC [11]. Reinforcement learning (RL) stands out as a promising approach for solving the MuRES problem due to its ability to reduce reliance on explicit domain modeling and its rapid execution capabilities.

Research Gap: Existing MuRES algorithms typically assume the environment is known and immutable, which conflicts with many practical search scenarios. When environment topology changes are expected, one possible approach is to treat the map as entirely unknown and reformulate the task as multi-robot exploration. However, completely discarding prior graph information significantly reduces search efficiency. To address this challenge, we propose structure-adaptive graph-encoded policy gradient (STAGE), a multi-agent reinforcement learning based algorithm tailored for MuRES problem in uncertain topological environments.

III. Problem Formulation and Backgrounds

This section presents the MuRES problem formulation under uncertain environments, and the GAT background.

A. MuRES Problem Formulation

The MuRES problem studied in this paper is to coordinate a team of N robots to search for a moving target in a discrete environment. To formally describe MuRES in uncertain environments, we specify the following configurations:

1) Environmental modeling: The search environment is modeled as an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is the node set and $\mathcal{E}$ is the edge set. The graph is assumed unit-cost, where robots traverse any adjacent edge in one time step, a convention widely adopted in the MuRES literature [2]–[4], [11]. A subset $\mathcal{E}_v \subset \mathcal{E}$ denotes vulnerable edges whose traversability is uncertain prior to deployment. The status of any edge $(u_1, u_2) \in \mathcal{E}_v$ (traversable or destroyed) is revealed only when a robot reaches an incident node u_1 or u_2.

2) Robot/target dynamics: Robot team is assumed to be homogeneous with identical motion primitives and sensing range. At time t, robot i's position is represented as $p_t^{(i)}$ and selects an action $a_t^{(i)}$ to move to its neighbor node (or stay) according to its own policy $\pi^{(i)}$. The target is a non-adaversarial, moving independently based on its own motion dynamics: $\mathbb{P}[e_{t+1}|e_t] = \Gamma(e_t, e_{t+1})$. The sensor range

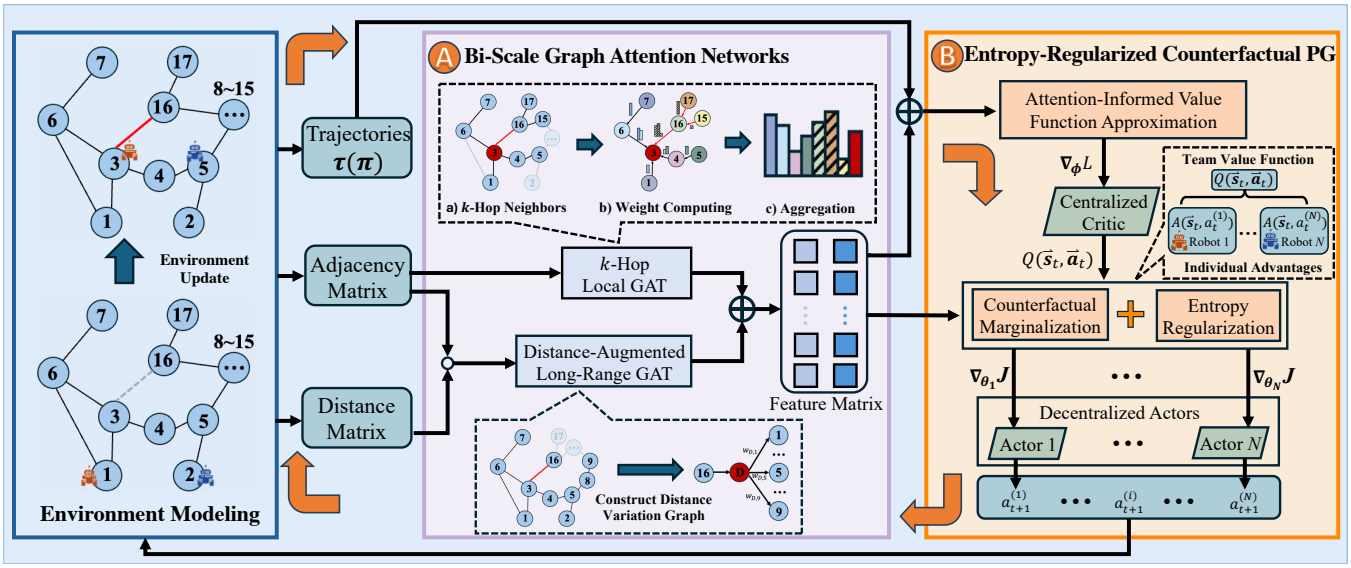


Fig. 2. The STAGE Framework: (1) the bi-scale graph attention networks (GAT) encoder, which fuses a k-hop local GAT with a distance-augmented long-range GAT. The augmented graph is updated whenever newly observed edges shorten shortest-path distances; and (2) the entropy-regularized counterfactual policy gradient, which combines a structure-aware centralized critic learns both team return and graph-information signals, and decentralized actors are trained via counterfactual marginalization with entropy regularization.

is restricted to the same node, which means a capture occurs once the target and any robot are on the same node.

3) MuRES objective: The MuRES objective here is to minimize expected capture time:

$$\min_{\boldsymbol{\pi}} \mathbb{E}[t_{\text{cap}}^{\boldsymbol{\pi}}], \tag{1}$$

where, $t_{\text{cap}}^{\boldsymbol{\pi}}$ is expected capture time when robots follow joint policy $\boldsymbol{\pi} = \{\pi^{(1)}, \pi^{(2)}, ..., \pi^{(N)}\}$.

B. Graph Attention Network

GAT applies self-attention over node neighborhoods to aggregate information, producing embeddings that adapt to local graph structure [31], [32]. For a Node i with neighbors $\mathcal{V}_i$, the attention coefficients α_{ij} are computed as:

$$\alpha_{i,j} = \text{softmax}_{j\in\mathcal{V}_i}\big(\text{LeakyReLU}(a^\top[Wh_i \| Wh_j])\big), \tag{2}$$

where W and a are learnable parameters and $\|$ denotes concatenation. Then the node representation aggregates neighbor messages as:

$$h_i' = \sigma\big(\sum_{j\in\mathcal{V}_i} \alpha_{ij}\, Wh_j\big). \tag{3}$$

GAT further employs multi-head attention mechanism to improve stability and expressiveness. With uncertain graph, GAT can adapt to local topology changes by reweighting neighborhoods. However, stacked k-hop GAT layers primarily capture local context and suffer from over-smoothing, reducing sensitivity to long-range structural changes.

IV. STAGE: Structure-Adaptive Graph-Encoded Policy Gradient

In this section, we present the methodology of structure-adaptive graph-encoded policy gradient (STAGE), a MARL algorithm tailored for the MuRES problem in uncertain topological environments, as outlined in Fig. 2. STAGE first employs a bi-scale GAT encoder to capture both local neighborhood and long-range structural changes. It then applies an entropy-regularized counterfactual policy gradient, wherein the attention-informed value function approximation jointly evaluates team return and graph information, the counterfactual marginalization provides agent-level credit assignment, and the entropy regularization sustains robots' exploration.

A. Bi-Scale GAT Encoder

To remain sensitive to both nearby topology changes and long-range shortcuts revealed by newly resolved edges, the bi-scale GAT encoder comprises two complementary branches: a k-hop local GAT aggregates neighborhood changes, and a distance-augmented long-range GAT integrates long-range graph structure variation.

1) k-Hop Local GAT: As the connectivity status of vulnerable edges is uncertain prior to robot team, we maintain the currently revealed topology $\mathcal{G}_t(\mathcal{V}, \mathcal{E}_t)$. We then stack k GAT layers on $\mathcal{G}_t$ to obtain a local embedding $h_t^{\text{loc}}(v)$ for each node v that depends only on nodes within k hops of v in $\mathcal{G}_t$. Intuitively, each GAT layer computes the attention coefficients $\alpha_{u,v}$ based on Eq. (2) over the current neighbor set and aggregates messages based on Eq. (3).

When a vulnerable edge (v_p, v_q) is verified to be traversable at time step $t+1$, we update the edge set to $\mathcal{E}_{t+1} = \mathcal{E}_t \cup (v_p, v_q)$, and expand the neighbor set of v_p and v_q. Consequently, for any node v within k hops distance of v_p or v_q would change which can be expressed as:

$$h_{t+1}^{\text{loc}}(v) \neq h_t^{\text{loc}}(v), \text{ if } d_{\mathcal{G}_{t+1}}(v, v_p) \text{ or } d_{\mathcal{G}_{t+1}}(v, v_q) \leq k \tag{4}$$

where $d_{\mathcal{G}_{t+1}}(v, v_p)$ refers to the distance of node v and

node v_p in graph $\mathcal{G}_{t+1}$. The k-hop local GAT captures the local structural variation from the newly revealed edge by re-aggregating messages on the updated graph $\mathcal{G}_{t+1}$.

2) Distance-Augmented Long-Range GAT: Focusing only on local structure variation is insufficient, since a newly revealed edge can substantially reduce shortest-path distances for distant nodes. To inject non-local signals without deep stacks GAT, we construct a distance variation map that routes influence from the newly confirmed edge's endpoints to other nodes whose shortest-path distances decreased.

Suppose a vulnerable edge (v_p, v_q) is confirmed traversable at time t, we recompute shortest paths only source from node v_p and v_q on $\mathcal{G}_t$. For brevity, we detail the construction for v_p as an example and define the distance reduction as follows:

$$\Delta_{v_p}(v) = d_{\mathcal{G}_{t-1}}(v, v_p) - d_{\mathcal{G}_t}(v, v_p). \tag{5}$$

With the distance drop function, we can then define the distance variation node set for v_p as: $v \in \mathcal{V}_{v_p}$ for all $\Delta_{v_p}(v) > \delta$. Then, we add a mediator dummy node v_d, for every node v we add directed edges $v_p \to v_d$ and $v_d \to v$. This augmentation process ensures v_p becomes a two-hop neighbor for every node v in $\mathcal{V}_{v_p}$, allowing a 2-layer GAT to propagate the shortcut's effect to v in one forward pass. In practice, we can simplify computation by evaluating distance changes only at fork nodes that can alter search strategies. Since attention computations on this distance variation graph are identical to those of the k-hop local GAT, we omit them here and denote the resulting long-range embedding feature by h_t^{lr}.

Then we fuse the long-range GAT embedding h_t^{lr} with the local GAT embedding h_t^{loc} via a learnable gate, the fuse weight is as follows:

$$\kappa_v = \sigma\big(w_g^\top [h_t^{\text{loc}}(v); h_t^{\text{lr}}(v)] + b_g\big), \tag{6}$$

where w_g and b_g are learnable parameters. On this basis, we could obtain the updated node feature representation as:

$$h_t(v) = \text{LayerNorm}\big((1-\kappa_v)h_t^{\text{loc}}(v) + \kappa_v h_t^{\text{lr}}(v)\big), \tag{7}$$

In this way, the resulting node embeddings capture both local and global topology variations. Notably, the bi-scale GAT encoder inherent objective is also to optimize the MuRES objective under topological uncertainty, its parameters are updated via backpropagation from both the centralized critic or the decentralized actor losses, rather than being fixed or precomputed. To decouple representation learning for value estimation and policy, we employ distinct bi-scale GAT encoders for each network.

B. Entropy-Regularized Counterfactual Policy Gradient

Given the bi-scale graph encoder, the node-embedding matrix captures both local and long-range topological variations. On this basis, we introduce an entropy-regularized counterfactual policy gradient to train the multi-robot search policy which contains a centralized critic and decentralized actors.

Algorithm 1: STAGE

Input: Max episodes: $E_{\max}$; target model: Γ; learning rates: α_c, α_a, α_g; max detection time: T; regularization parameter: β;
Output: policy $\pi^{(i)}(\boldsymbol{\theta}_i^*)$ and graph encoder ξ^a
Init: counter $\leftarrow 1$;

while *counter* $\le E_{\max}$ **do**
  Initialize target position e_0, and robot positions $p_0^{(i)}$ for each robot, and $t \leftarrow 0$;
  while *target_captured==false and* $t < T$ **do**
    Update embeddings h_t^c, h_t^a based on Eq. (7);
    Obtain robot actions: $a_t^{(i)} \sim \pi^{(i)}\big(.|s_t^{(i)}; \boldsymbol{\theta}_i\big)$ $\forall i \in \{1, 2, \ldots, N\}$;
    Update target position $e_{t+1} = \Gamma(e_t)$, robot positions $p_{t+1}^{(i)}$ for each robot, and reward $\widetilde{r}_t$;
    if *Edge* (v_p, v_q) *confirmed traversable* **then**
      Update $\mathcal{E}_{t+1} = \mathcal{E}_t \cup (v_p, v_q)$;
    $t \leftarrow t + 1$;
  Update $\boldsymbol{\phi}$ and $\boldsymbol{\xi}^c$ based on Eq. (9);
  foreach $i \in \{1, 2, \ldots, N\}$ **do**
    Update $\boldsymbol{\theta}_i$ and $\boldsymbol{\xi}^a$ based on Eq. (13);
  counter $\leftarrow$ counter $+ 1$;
Final.

1) Attention-Informed Value Function Approximation: We first formalize the attention-informed value function:

$$Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\phi}) = f_{\boldsymbol{\phi}}([\boldsymbol{h}_t || \boldsymbol{h}_{t+1} || \psi(t)]). \tag{8}$$

Here, $\boldsymbol{h}_t = ||_i^N h_t(p_t^{(i)})$ concatenates the encoder embeddings of the robots' current positions under the current revealed graph structure $\mathcal{G}_t$ which parameterized by ξ_c for critic, and ξ_a for actors respectively. While the joint action is implicitly encoded by the transition $\boldsymbol{h}_t \xrightarrow{\boldsymbol{a}_t} \boldsymbol{h}_{t+1}$. The term $\psi(t)$ denotes a time embedding, as the robot may make different decision at same node when time varying. Then the associated Bellman operator for the attention-informed centralized value can be expressed as:

$$Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t; \boldsymbol{\phi}) = \sum_{\boldsymbol{s}_{t+1}} [\widetilde{r}_t + \gamma \sum_{\boldsymbol{a}_{t+1}} Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}; \boldsymbol{\phi}) \times \pi(\boldsymbol{s}_{t+1}, \boldsymbol{a}_{t+1}; \boldsymbol{\theta})], \tag{9}$$

where $\widetilde{r}_t = r_t + \lambda r_t^{\text{info}}$ augments the task reward with a graph-information signal and γ refers to the discount factor. In practice, we directly applied one-step Temporal Difference (TD) learning for updates.

2) Counterfactual Marginalization: Updating each policy directly from the team value induces a centralized–decentralized mismatch and may cause 'lazy' agents due to inter-agent coupling. To address this, we adopt an entropy-regularized counterfactual actor update. We begin with counterfactual marginalization, whose baseline for agent

i can be expressed as:

$$b_t^{(i)}(\boldsymbol{s}_t, \boldsymbol{a}_t^{(-i)}) = \sum_{u_t^{(i)}} \pi^{(i)}(u_t^{(i)}|s_t^{(i)}) Q_{\text{tot}}^{\boldsymbol{\pi}}\big(\boldsymbol{s}_t, (u_t^{(i)}, \boldsymbol{a}_t^{(-i)}); \boldsymbol{\phi}\big), \quad (10)$$

where $\boldsymbol{a}_t^{(-i)}$ denotes the joint action of all robots except i. This baseline is the conditional expectation of the team value over only agent i's action with teammates held fixed. Then, the per-agent policy gradient is as follows:

$$\nabla_{\boldsymbol{\theta}_i} J(\boldsymbol{\theta}_i) = \mathbb{E}_{\tau^{(i)} \sim \pi^{(i)}} \Big[\sum_{t=0}^{H} A^{(i)}(\boldsymbol{s}_t, \boldsymbol{a}_t) \times \nabla_{\boldsymbol{\theta}_i} \log \pi(a_t^{(i)}|s_t^{(i)}; \boldsymbol{\theta}_i) \Big], \quad (11)$$

where $A^{(i)}(\boldsymbol{s}_t, \boldsymbol{a}_t) = Q_{\text{tot}}^{\boldsymbol{\pi}}(\boldsymbol{s}_t, \boldsymbol{a}_t) - b_t^{(i)}(\boldsymbol{s}_t, \boldsymbol{a}_t^{(-i)})$ represent the counterfactual advantage. Since $b_t^{(i)}$ has zero mean under $\pi^{(i)}$, subtracting it leaves the gradient unbiased while reducing variance, as proved in COMA [33]. Note that although the actor network also consumes bi-scale GAT features, this does not alter the derivation above, the counterfactual advantage and unbiasedness properties remain intact.

3) Entropy Regularization: To prevent premature collapse and promote exploration, we augment each actor's objective with entropy regularization:

$$J_{\mathcal{H}}(\pi^{(i)}) = J(\pi^{(i)}) + \beta \mathbb{E}_{s^{(i)}} \big[\mathcal{H}\big(\pi(.|s^{(i)})\big) \big], \quad (12)$$

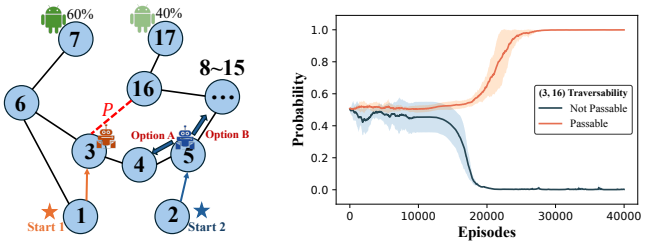
where the entropy at state $s^{(i)}$ is defined as $\mathcal{H}\big(\pi(.|s_t^{(i)})\big) = \sum_{a_t} \pi(a_t^{(i)}|s_t^{(i)}) \log \pi(a_t^{(i)}|s_t^{(i)})$ and the term β represents regularization coefficient that controls the exploration–exploitation trade-off. During training, β is often annealed, progressively shifting the objective's emphasis from exploration to exploitation. Combining entropy regularization with the counterfactual baseline yields the entropy-regularized counterfactual policy gradient:

$$\nabla_{\boldsymbol{\theta}_i} J_{\mathcal{H}}(\pi^{(i)}) = \mathbb{E}_{\tau^{(i)} \sim \pi^{(i)}} \Big[\sum_{t=0}^{H} \big(A^{(i)}(\boldsymbol{s}_t, \boldsymbol{a}_t) - \beta \log \pi(a_t^{(i)}|s_t^{(i)}; \boldsymbol{\theta}_i) \big) \nabla_{\boldsymbol{\theta}_i} \log \pi(a_t^{(i)}|s_t^{(i)}; \boldsymbol{\theta}_i) \Big], \quad (13)$$

By maximizing the entropy-regularized objective, the policy is encouraged to avoid premature convergence to a specific strategy, instead promoting the exploration of diverse and potentially optimal strategies.

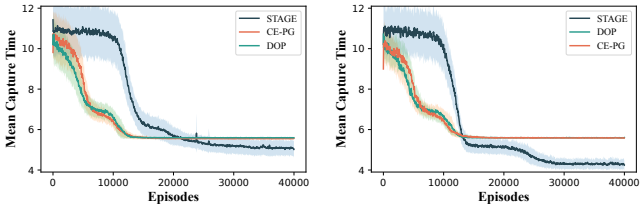
V. Simulation Results and Analysis

This section evaluates the performance of STAGE against other state-of-the-art MuRES solutions. For MuRES baselines, we selected (1) cross-entropy regularized policy gradient (CE-PG) in [3]; (2) deep Q Learning (DQN) in [30]; (3) distributional reinforcement learning based searcher (DRL) in [5]; (4) signal mediated coordination (SMC) in [22]. Additionally, we include three representative multi-agent reinforcement learning (MARL) algorithms: (1) off-policy multi-agent decomposed policy gradients (DOP) in [34]; (2) factored multi-agent centralised policy gradients (FACMAC) in [35]; (3) probability factorized multi-agent actor-critic (PF-MAAC) in [11]. Note that FACMAC and DOP are not specified for the MuRES problem, we make necessary adaptations to meet the unique challenge from MuRES. All algorithms are implemented in Python 3.9, and evaluated on a 10-core Apple M1 Pro machine with the 64-bit version of Mac-OS system and 32GB RAM. The source code of STAGE as well as algorithms' parameter configuration are publicly available at the anonymized repository [1].



(a) Simple Use Case (b) Probability of Option A

Fig. 3. Simple use case and decision adaptation. (a) Scenario with a vulnerable edge $(3, 16)$ whose traversability is revealed only when any robot arrive at Nodes 3 or 16. (b) Learning curve of the probability that Robot 2 selects *Option A* upon arriving at Node 5.



(a) Learning Curve ($P = 0.5$) (b) Learning Curve ($P = 0.8$)

Fig. 4. Learning Curve Comparison between STAGE with representative MuRES solutions in simple use case with different passable probabilities.

A. Performance Comparison in the Simple Case

In this subsection, we evaluate the performance of STAGE in the simple use case as illustrated in Fig. 3(a). The setup has already been described in Fig. 1, for brevity, we omit details here. We first examine Robot 2's decision at Node 5 when $t = 1$, with the edge $(3, 16)$ being either passable or blocked, to assess whether the robot team can detect environmental changes and adapt search strategy accordingly. Then, we compare STAGE with representative baselines under different passability probabilities for edge $(3, 16)$. Given that the primary objective of this simulation is to demonstrate STAGE's core functionality, an excessive number of comparisons may dilute the focus of the analysis. Here we selected the most representative baselines: (1) CE-PG, as a canonical RL-based MuRES solution, (2) DOP, as a representative MARL algorithms.

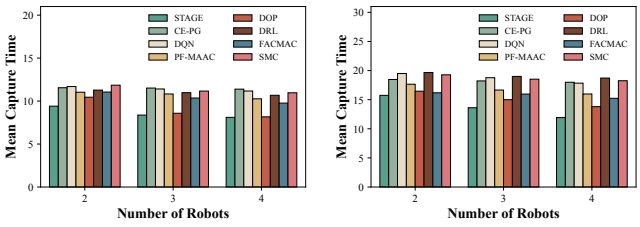
Fig. 3(b) represents the probability that Robot 2 chooses Option A at Node 5. We can see that when the edge $(3, 16)$ is confirmed passable by Robot 1, Robot 2 learns to choose option A to reduce the mean capture time, conversely, when the edge $(3, 16)$ is confirmed blocked, Robot 2 alternates to choose Option B. Fig. 4 presents learning curves of

[1] https://github.com/Anonymous-0205/STAGE.

STAGE compared with other baselines. We observe that STAGE learns more slowly at the beginning but achieves substantially better final performance. The main reason is that our STAGE algorithm employs the bi-scale GAT, which introduces higher representational complexity in the early stage of learning both local and global structures under uncertainty. This requires the agent to simultaneously learn structural representations and optimize the policy, leading to slower initial performance. Once the graph representations are formed, STAGE adapts to the revealed topology and outperforms the baselines. In contrast, canonical MuRES methods can not adapt their search strategies under uncertain environment. With connectivity probabilities unknown to robot team, these methods often learn to adopt a more conservative strategy. Consequently, when the actual passable probability of edge $(3, 16)$ is increased to 0.8, see Fig. 4(b), STAGE's advantage becomes more pronounced.

B. Evaluation in Canonical MuRES Environments

This subsection presents the performance comparison of STAGE against state-of-the-art methods in two canonical MuRES environments, OFFICE and MUSEUM, both widely used in the MuRES literature [2]–[4], [22]. The target is initialized on a biased distribution and then moves randomly. To ensure reliable evaluation, all algorithms are trained through 10 independent runs, and the search process is repeated 1,000 times for each trained policy.



(a) OFFICE Performance (b) MUSEUM Performance

Fig. 5. Performance comparison between STAGE and state of the arts in OFFICE and MUSEUM.

Fig. 5 shows the averaged capture time of STAGE and each baseline algorithm under different robot team sizes. From the figure, we can see that the STAGE achieves the lowest mean capture time, indicating superior performance among all compared algorithms in both environments. The main reason is that STAGE has the ability to explicitly handle environmental uncertainty and to adapt its search strategy under varying graph connectivity. Additionally, we observe that except STAGE, MARL algorithms (especially DOP and FACMAC) demonstrate superior performance compared to other baselines, particularly when the deployed number of robots is sufficient. The underlying reason is that the MARL based algorithm takes advantage of credit assignment, which makes the robots cooperation more efficient.

C. Ablation Study

We introduce the main modules of STAGE in Section IV. In this subsection, we conduct the ablation study to isolate the contribution of each module in STAGE by evaluating STAGE variants with the following elements removed in turn: (1) bi-scale GAT encoder, (2) long-range GAT, (3) counterfactual marginalization module, and (4) entropy regularization. We compare the performance in the simple use case with different passable probabilities for edge $(3, 16)$.

TABLE I
STAGE'S ABLATION STUDY

STAGE Variants	$p = 0.3$	$p = 0.5$	$p = 0.7$
STAGE	**5.481**	**4.982**	**4.529**
No Bi-Scale GAT Encoder	5.756	5.475	5.244
No Long-Range GAT	5.769	5.401	5.205
No Counterfactual Marginalization	5.701	5.962	6.191
No Entropy Regularization	5.510	5.456	5.092

Table I shows the performance comparison of different STAGE variants across various scenarios. The main findings are as follows: (1) the bi-scale GAT encoder is pivotal for handling structural uncertainty, removing it produces a significant performance drop in all test scenarios. (2) Counterfactual marginalization is critical for robot cooperation, without which, robot teams may not be able to form an effective cooperation mechanism. (3) Entropy regularization provides a consistent benefit by sustaining exploration and reducing premature convergence.

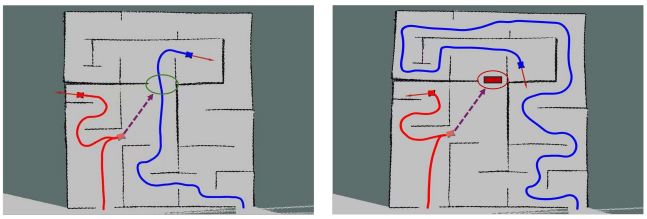
VI. EXPERIMENTS WITH A REAL MULTI-ROBOT SYSTEM

In this section, we deploy STAGE on a real multi-robot platform to validate its feasibility and core functionalities under topological uncertainty. The testbed employs S30 six-wheeled differential-drive robots equipped with an onboard LiDAR (VLP-16), a LiDAR-based odometry system, and a depth sensing camera. In our implementation, STAGE functions as the global task-allocation layer that issues high-level way point directives, while low-level motion planning and control are handled by the S30's built-in autonomous navigation module.

(a) S30 Robot (b) Search Environment (c) RViz Map

Fig. 6. The autonomous robot, the constructed environment, and the RViz visualization used in the demonstrative scenario.

The initial conditions are illustrated in Fig. 6(c): two robots start at Node 1 and Node 4 perspectively, and the target is positioned at either Node 10 (60%) or Node 16 (40%). Fig. 7 demonstrates the final path results generated by the STAGE over 10 independent training runs. From the results, we observe that the STAGE has ability to adapt search strategy based on different graph structure: when the edge $(8, 15)$ is confirmed passable by Robot 1, Robot 2 proceeds to Node 8 to take the shortest path, conversely, when the edge is confirmed blocked, Robot 2 switches to



(a) Scenario with Edge Passable (b) Scenario with Edge Blocked

Fig. 7. STAGE search paths under varying environment topological

a longer, conservative route. The complete video of STAGE has been uploaded together with the manuscript.

VII. Conclusion and Future Work

This paper introduces STAGE, a structure-adaptive graph-encoded policy gradient method for MuRES in uncertain topological environments. STAGE first incorporates a bi-scale GAT encoder, combining k-hop local attention with a distance-augmented long-range GAT to capture both the local and global structure variation. And then employ the entropy-regularized counterfactual policy gradient to learn decentralized search strategy. To the best of our knowledge, STAGE is the first MuRES solution explicitly tailored for uncertain topological environments. Extensive simulation results indicate the feasibility and superiority of STAGE for solving MuRES under topological uncertainty. In the future, we plan to improve the GAT encoder to more powerful GNN such as Transformer. Additionally, we plan to extend STAGE from discrete graphs to continuous state and action spaces to enable more general multi-robot coordination.

Acknowledgments

This work was supported by the Research Center for Unmanned Autonomous Systems, Hong Kong Polytechnic University, Hong Kong, SAR, China, and the National Natural Science Foundation of China (Grant No. 62576229).

References

[1] Y.-S. Li and K.-S. Tseng, "Multi-robot search in a 3D environment with intersection system constraints," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 5963–5969.

[2] B. A. Asfora, J. Banfi, and M. Campbell, "Mixed-integer linear programming models for multi-robot non-adversarial search," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6805–6812, 2020.

[3] H. Guo, Z. Liu, R. Shi, W.-Y. Yau, and D. Rus, "Cross-entropy regularized policy gradient for multirobot nonadversarial moving target search," *IEEE Transactions on Robotics*, vol. 39, no. 4, pp. 2569–2584, 2023.

[4] G. Hollinger, S. Singh, J. Djugash, and A. Kehagias, "Efficient multi-robot search for a moving target," *The International Journal of Robotics Research*, vol. 28, no. 2, pp. 201–219, 2009.

[5] H. Guo, Q. Peng, Z. Cao, and Y. Jin, "DRL-Searcher: A unified approach to multirobot efficient search for a moving target," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 35, no. 3, pp. 3215–3228, 2023.

[6] N. M. Stiffler and J. M. O'Kane, "Asymptotically-optimal multi-robot visibility-based pursuit-evasion," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 10 366–10 373.

[7] M.-R. Ling, J.-W. Huo, J.-L. Wang, and Y. Zhou, "A heterogeneous robot collaborative search method for radioactive sources," *Annals of Nuclear Energy*, vol. 213, p. 111145, 2025.

[8] X. Cao, M. Li, Y. Tao, and P. Lu, "HMA-SAR: Multi-agent search and rescue for unknown located dynamic targets in completely unknown environments," *IEEE Robotics and Automation Letters*, vol. 9, no. 6, pp. 5567–5574, 2024.

[9] X. Kong, J. Yang, X. Chai, and Y. Zhou, "An advantage duPLEX dueling multi-agent Q-learning algorithm for multi-UAV cooperative target search in unknown environments," *Simulation Modelling Practice and Theory*, vol. 142, p. 103118, 2025.

[10] M. Li, Y. Tao, X. Cao, and P. Lu, "DAPTP: Distributed awareness planner using time potential for dynamic target search," *Robotics and Autonomous Systems*, p. 105010, 2025.

[11] Q. Peng, H. Guo, Z. Zhang, C.-Y. Wen, and Y. Jin, "PF-MAAC: A learning-based method for probabilistic optimization in time-constrained non-adversarial moving target search," *Swarm and Evolutionary Computation*, vol. 92, p. 101785, 2025.

[12] M. Morin, I. Abi-Zeid, and C.-G. Quimper, "Ant colony optimization for path planning in search and rescue operations," *European Journal of Operational Research*, vol. 305, no. 1, pp. 53–63, 2023.

[13] H. Tang, W. Sun, A. Lin, M. Xue, and X. Zhang, "A GWO-based multi-robot cooperation method for target searching in unknown environments," *Expert Systems with Applications*, vol. 186, p. 115795, 2021.

[14] C. Huang, B. Du, and M. Chen, "Multi-UAV cooperative online searching based on voronoi diagrams," *IEEE Transactions on Aerospace and Electronic Systems*, vol. 60, no. 3, pp. 3038–3049, 2024.

[15] Y. Chen and J. Xiao, "Target search and navigation in heterogeneous robot systems with deep reinforcement learning," *Machine Intelligence Research*, vol. 22, no. 1, pp. 79–90, 2025.

[16] X. Lin, F. Gao, and W. Bian, "A high-effective swarm intelligence-based multi-robot cooperation method for target searching in unknown hazardous environments," *Expert Systems with Applications*, vol. 262, p. 125609, 2025.

[17] J. Xiao, P. Pisutsin, and M. Feroskhan, "Collaborative target search with a visual drone swarm: An adaptive curriculum embedded multi-stage reinforcement learning approach," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 36, no. 1, pp. 313–327, 2023.

[18] P. Gokul, K. Harikumar, and J. Senthilnath, "A dynamic area approximation-based stochastic multi-UAV target search with noisy measurements," in *2024 International Conference on Unmanned Aircraft Systems (ICUAS)*. IEEE, 2024, pp. 718–723.

[19] M. Wang, B. Xin, M. Jing, and Y. Qu, "A priority-based multi-robot search algorithm for indoor source searching," *IEEE Transactions on Automation Science and Engineering*, vol. 22, pp. 10 457–10 469, 2025.

[20] R. Ghods, W. J. Durkin, and J. Schneider, "Multi-agent active search using realistic depth-aware noise model," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 9101–9108.

[21] M. Fan, H. Liu, G. Wu, A. Gunawan, and G. Sartoretti, "Multi-UAV reconnaissance mission planning via deep reinforcement learning with simulated annealing," *Swarm and Evolutionary Computation*, vol. 93, p. 101858, 2025.

[22] Q. Peng, H. Guo, B. Li, C.-Y. Wen, and Y. Jin, "SMC-Searcher: Signal mediated coordination for decentralized multi-robot adversarial moving target search," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2025.

[23] H. Xiao, R. Cui, D. Xu, and Y. Li, "MPC-based cooperative multiagent search for multiple targets using a bayesian framework," *Journal of Field Robotics*, vol. 41, no. 8, pp. 2630–2649, 2024.

[24] C. Wang, C. Yu, X. Xu, Y. Gao, X. Yang, W. Tang, S. Yu, Y. Chen, F. Gao, Z. Jian *et al.*, "Multi-robot system for cooperative exploration in unknown environments: A survey," *arXiv preprint arXiv:2503.07278*, 2025.

[25] Y. Shen, C. Wei, Y. Sun, and H. Duan, "Bird flocking inspired methods for multi-UAV cooperative target search," *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2023.

[26] J. T. Ebert, F. Berlinger, B. Haghighat, and R. Nagpal, "A hybrid pso algorithm for multi-robot target search and decision awareness," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 11 520–11 527.

[27] W. Sheng, H. Guo, W.-Y. Yau, and Y. Zhou, "PD-FAC: Probability density factorized multi-agent distributional reinforcement learning for multi-robot reliable search," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8869–8876, 2022.

[28] J. Lindsay, M. Seto, and R. Bauer, "Collaboration of marine robots towards dynamic target localization and tracking," in *OCEANS 2024-Halifax*. IEEE, 2024, pp. 01–10.

[29] C. Yu, A. Velu, E. Vinitsky, J. Gao, Y. Wang, A. Bayen, and Y. Wu, "The surprising effectiveness of PPO in cooperative multi-agent games," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 611–24 624, 2022.

[30] X. Qin, X. Li, Y. Liu, R. Zhou, and J. Xie, "Multi-agent cooperative target search based on reinforcement learning," in *Journal of Physics: Conference Series*, vol. 1549, no. 2. IOP Publishing, 2020, p. 022104.

[31] P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Lio, Y. Bengio *et al.*, "Graph attention networks," in *International Conference on Learning Representations*, 2018.

[32] Z. Yu, H. Guo, C.-M. Chew, A. H. Adiwahono, J. Chan, B. W. T. Shong, and W.-Y. Yau, "Multi-robot reliable navigation in uncertain topological environments with graph attention networks," *IEEE Robotics and Automation Letters*, vol. 10, no. 5, pp. 5082–5089, 2025.

[33] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson, "Counterfactual multi-agent policy gradients," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 32, 2018.

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

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