MAC: Memetic Actor-Critic Framework for UAV Path Planning

Jingeun Kim
Dept. IT Convergence, Gachon University
Gyeonggi-do, Republic of Korea
wlsrms27@gachon.ac.kr

Seung-Yeol Hong
Dept. of Computer Science, Kwangwoon University
Seoul, Republic of Korea
ghdqkxk159@naver.com

Yong-Hyuk Kim
Dept. of Computer Science, Kwangwoon University
Seoul, Republic of Korea
yhdfly@kw.ac.kr

Yourim Yoon
Dept. Computer Engineering, Gachon University
Gyeonggi-do, Republic of Korea
yryoon@gachon.ac.kr

Abstract

Unmanned Aerial Vehicles play a pivotal role in maritime Search and Rescue (SAR) missions, yet generating optimal flight paths remains a significant challenge. Traditional meta-heuristic algorithms, such as genetic algorithms (GA), often suffer from premature convergence to local optima, while reinforcement learning (RL) approaches typically struggle with low sample efficiency and slow initial convergence. To address these limitations, this paper proposes a hybrid framework termed Memetic Actor-Critic (MAC). The MAC framework integrates the decision-making policy of Actor-Critic (AC) with the global search capabilities of a memetic algorithm. Specifically, we introduce a problem-specific local refinement mechanism that utilizes state-value estimations to refine the actor's policy, enabling precise exploitation of high-value regions beyond immediate rewards. The proposed method was thoroughly evaluated using a realistic SAR scenario constructed from oceanographic particle simulation data collected from the East Sea of South Korea. Experimental results demonstrate that MAC significantly outperforms standalone GA, MA, AC, and other variants in terms of both solution quality and convergence stability, proving its effectiveness in balancing global exploration and local exploitation.

CCS Concepts

• **Computing methodologies** → *Genetic algorithms*; *Planning and scheduling*; • **Theory of computation** → *Reinforcement learning*.

Keywords

Memetic algorithm, reinforcement learning, uav path planning, search and rescue

ACM Reference Format:
Jingeun Kim, Seung-Yeol Hong, Yong-Hyuk Kim, and Yourim Yoon. 2026. MAC: Memetic Actor-Critic Framework for UAV Path Planning. In *Genetic and Evolutionary Computation Conference (GECCO Companion '26), July 13–17, 2026, San Jose, Costa Rica.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3795101.3805326

GECCO Companion '26, San Jose, Costa Rica

ACM ISBN 979-8-4007-2488-6/2026/07
https://doi.org/10.1145/3795101.3805326

1 Introduction

Unmanned aerial vehicles (UAVs) have demonstrated significant potential in conducting search and rescue missions [6]. Various sectors, including agriculture [9], marine remote sensing [11], environmental monitoring [1], disaster response [2], and search and rescue [5], actively utilize UAVs due to their flexibility, cost-effectiveness, and autonomous capabilities. However, despite these advantages, achieving efficient and optimal path planning remains a critical challenge for successful UAV navigation.

To address this, many researchers have studied autonomous path planning algorithms using meta-heuristic algorithms [3], reinforcement learning [10], and deterministic algorithms [8]. Meta-heuristic algorithms, such as the genetic algorithm (GA), demonstrate superior performance in global path planning by effectively exploring large solution spaces. Although GA has demonstrated good performance in UAV path planning, it often suffers from slow convergence or stagnation in local optima, failing to explore globally optimal paths. Conversely, reinforcement learning (RL) has shown remarkable adaptability and real-time decision-making capabilities in dynamic UAV environments. Nevertheless, the application of RL is frequently hindered by low sample efficiency. Specifically, tabular RL approaches like Q-learning inherently rely on iterative trial-and-error interactions, which often results in inefficient exploration during the initial phase, particularly in complex environments lacking global search guidance.

To address these issues, we propose a hybrid framework termed Memetic Actor-Critic (MAC), which integrates reinforcement learning with a memetic algorithm. Unlike existing memetic RL frameworks that primarily rely on fitness-based heuristics for local search [4], MAC introduces a value-consistent refinement strategy. This approach directly aligns evolutionary operators with the critic's learned value landscape. Consequently, the local refinement capability effectively ameliorates the slow convergence inherent in RL agents, while the population-based evolutionary mechanism promotes diverse global exploration, thereby preventing premature convergence to local optima.

The proposed method, MAC, employs an Actor-Critic (AC) approach as the core decision-making process for UAV path planning. To further enhance performance, we adopt a population-based strategy that generates multiple agents and applies genetic operators to improve behavioral diversity. Additionally, a problem-specific local refinement mechanism is incorporated to ensure efficient and precise path planning.

2 Problem Formulation

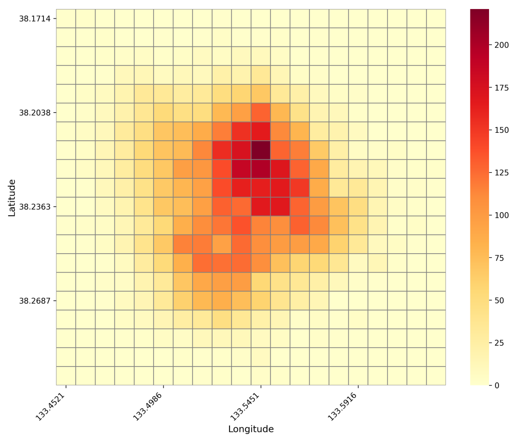

Figure 1: Visualization of the target probability distribution map within the maritime search environment

2.1 Environment

The search area is modeled as a discretized grid map $\mathcal{G}$ with dimensions $H \times W$, where $H = 20$ and $W = 20$, as shown in Figure 1. The environment is represented by a probability heatmap matrix $M \in \mathbb{R}^{H \times W}$, derived from oceanographic particle simulations.

In this study, particles are used to represent the potential locations of the search target. A total of 3,000 particles are initially deployed in the vicinity of the estimated accident location to model the uncertainty of the target's initial position. The drift trajectories of these particles are subsequently predicted using machine-learning-based models, which capture the effects of ocean currents and environmental factors over time.

The simulation data spans a specific temporal window from 18:00:00 to 20:00:00 on October 19, 2021, with a discrete time-step of $\Delta t = 1800$ seconds (30 minutes) in the East Sea of South Korea. To construct a robust search map that accounts for the drift uncertainty over this duration, the particle distributions from consecutive time steps are temporally integrated during the optimization phase. Consequently, each cell (i, j) contains a value $v_{i,j}$ representing the cumulative number of target particles, serving as a static probability density function of the missing target's presence within the designated time frame. Finally, in the precise evaluation phase, we utilized the raw, time-variant particle data instead of the cumulative heatmap. Unlike the static model used for optimization, this dynamic environment reflects the real-time drift of targets, allowing for a more precise assessment of the UAV's search performance.

We defined five distinct starting points. Four starting cells closest to each vertex of the data space and one cell located at the geometric center of the space were selected, resulting in a total of five instances.

2.2 Constraints

We assume the use of a specific UAV, specifically modeled after the DJI Matrice 300 RTK, for SAR operations. The UAV's movement is defined as a sequence of waypoints (vertices), denoted as a path $\mathcal{P} = \{w_0, w_1, \ldots, w_N\}$, where each waypoint w_k corresponds to the center coordinates of a grid cell.

The transition between consecutive waypoints is restricted to the Moore neighborhood. At each step, the UAV can move horizontally, vertically, or diagonally to an adjacent cell. This discrete action space ensures that the distance between w_k and w_{k+1} corresponds to a single movement unit within the grid structure.

The operation is strictly constrained by the UAV's battery capacity and flight speed. We define the maximum flight distance D_{max} based on the average search speed V_{avg} and the maximum endurance time T_{endure}:

$$D_{max} = V_{avg} \times T_{endure} \tag{1}$$

To align this physical constraint with the discretized grid environment, we convert the maximum flight distance into a maximum node-visiting capacity, denoted as $T_{capacity}$. Therefore, the length of the generated path is constrained so that the total number of movements does not exceed $T_{capacity}$. In this study, the UAV specifications are based on the DJI Matrice 300 RTK, which provides a maximum flight endurance of $T_{endure} = 55$ min and an average flight speed of $V_{\text{avg}} = 1.38$ m/s. Under these conditions, T_{capacity} is set to 75.

2.3 Objective Function

The primary objective of the path planning algorithm is to maximize the expected number of detected target particles. We assume that the UAV may fail to detect a target particle even if it is located within the UAV's search area.

Let A_{total} denote the total size of the search area, which corresponds to the dimensions of the heatmap. The UAV's search capacity, denoted as A_{drone}, is calculated based on its flight speed V_{avg}, detection width W_{cov}, and the operational time duration T_{op}:

$$A_{total} = H \times W \tag{2}$$

$$A_{drone} = V_{avg} \times W_{cov} \times T_{op} \tag{3}$$

The coverage probability P_{cov} is calculated using Equation 4, following [7]. This value represents the depletion rate of the target probability in a cell after a UAV visit:

$$P_{cov} = 1 - \exp\left(-\frac{A_{total}}{A_{drone}}\right) \tag{4}$$

When the UAV visits a cell (i, j) containing $v_{i,j}$ particles, it receives the reward. To simulate the probability of detection and environmental depletion for subsequent visits, the value of the cell is updated as follows:

$$v_{i,j}^{new} = v_{i,j}^{old} \times (1 - P_{cov}) \tag{5}$$

The optimization problem is formulated to maximize the cumulative reward collected along the path $\mathcal{P}$:

$$\mathcal{P}^* = \underset{\mathcal{P}}{\operatorname{argmax}} \sum_{t=0}^{|\mathcal{P}|-1} R(s_t) \tag{6}$$

where $R(s_t)$ is the number of collected particles at state s_t at time t.

3 Memetic Actor-Critic Algorithm

We employ the Actor-Critic reinforcement learning framework to identify the optimal path $\mathcal{P}$. This architecture is characterized by the simultaneous training of a policy matrix (Actor) and a value matrix (Critic). Here, both representations are implemented as explicit matrices rather than neural networks, as the search environment is defined on a compact 20×20 discrete map, where matrix-based representations are computationally more efficient and faster to update. Unlike conventional single-agent reinforcement learning, we initialize a population of N agents, adopting a population-based strategy to maintain solution diversity.

Each agent maintains its own distinct policy and value matrices, which are iteratively updated as episodes progress. To evolve the population, we apply an elitism strategy, preserving the top 20% of agents for the next generation. Simultaneously, parents are selected from the population to generate $N/2$ offspring through crossover and mutation operators. Subsequently, a local search mechanism is applied to refine the policy matrix by incorporating weights derived from the value matrix. A detailed description of each component is provided in the following subsections.

3.1 Initialization

Each agent is initialized with a policy matrix $\pi \in \mathbb{R}^{(H \times W) \times 8}$, where each row corresponds to a unique grid state and each column represents one of the eight possible movement actions. The value function is represented as a flattened vector $V \in \mathbb{R}^{H \times W}$, where each entry corresponds to a state value, initialized to zero for implementation convenience.

3.2 Genetic Operator

A total of $N/2$ offspring were generated from N parents using genetic operators, including crossover and mutation. Specifically, a two-dimensional quadrant-based crossover was employed, where a randomly selected row and column defined the crossover point. Each offspring inherited the first and third quadrants from one parent and the second and fourth quadrants from the other parent, with the complementary configuration applied to generate the paired offspring. This crossover was performed with 100% probability.

For mutation, each individual was perturbed with a predefined mutation probability of 20%. In the policy representation, zero-mean Gaussian noise was added to the action probability vector of each state, followed by clipping negative values to zero and renormalizing to preserve a valid stochastic policy. Similarly, Gaussian noise was applied to the critic value of each state to encourage local exploration in the value space.

3.3 Value-Informed Local Refinement

To accelerate convergence and enhance the quality of the policy, we employ a task-specific local refinement strategy. Unlike conventional local search mechanisms that rely on immediate fitness scores, it explicitly leverages the landscape of the critic's learned value function to bias the actor's policy toward promising regions, thereby bridging the gap between global evolution and local value estimation.

Table 1: Comparison of performance results across instances for 30 runs

Instance	Greedy	GA	MA	RL-AC	AC+GA	**MAC**	p-value[1]
I-1	8206.00	7382.80	8367.03	8446.63	**8461.30**	8460.37	3.42e-01
I-2	7366.00	7226.53	8113.80	8237.77	8267.70	**8318.50**	2.56e-11
I-3	8374.00	7455.70	8513.57	8474.30	8477.07	**8546.07**	3.92e-18
I-4	7935.00	7415.23	8354.27	8425.00	8431.53	**8443.93**	9.64e-06
I-5	8097.00	7399.83	8289.77	8404.30	8427.27	**8435.83**	2.07e-02

[1] One-tailed t-test to verify that MAC outperforms AC+GA.

For every state in the grid, the algorithm performs a one-step greedy look-ahead to evaluate the value estimates of all valid neighboring states. The action leading to the neighbor with the highest estimated value is identified as the locally optimal direction. We then refine the policy by injecting a fixed positive bias, referred to as the strength, into the probability of this optimal action. In this study, the strength is set to 0.8. Following this perturbation, the action probability distribution is re-normalized to ensure it remains a valid stochastic policy.

3.4 Experimental Setup

All experiments were performed on a workstation equipped with an Intel Core i7-7700K CPU (4.20 GHz) and 64 GB of RAM. To ensure a fair comparison, the identical search space was applied across all algorithms. We implemented various baseline algorithms under these conditions to objectively evaluate the performance of the proposed method, MAC. Specifically, the comparison included AC+GA (the proposed method without local search), AC (without evolutionary computation), MA and GA (without reinforcement learning), and a greedy search algorithm as a baseline. It is important to note that the proposed framework is designed as an offline planning system by design. In maritime SAR operations, flight paths are typically generated prior to deployment using simulation data, where solution reliability and robustness are more critical than real-time responsiveness.

4 Result

Table 1 presents the comparative results of the average performance across five different instances, calculated from 30 independent runs for each algorithm. Except for instance I-1, the proposed MAC algorithm outperformed all comparative methods, achieving the best performance in the remaining cases. This indicates that the MAC architecture, which integrates an RL-based policy, genetic algorithm, and local search, effectively enhances search performance. We conducted a one-tailed t-test to verify the efficacy of MAC. The results confirm that MAC performed significantly better than the baselines in most instances. In the case of I-1, although MAC showed a marginally lower average performance compared to AC+GA with a difference of only 0.83, it demonstrated consistently superior performance across all other instances. In the case of I-1. the performance gap between the greedy search and the AC+GA is merely 255.3. This narrow margin indicates that the instance presents a relatively simple landscape where near-optimal solutions are readily accessible. For this reason, the global search capacity of AC+GA was already sufficient to bridge this narrow gap and reach the global optimum.

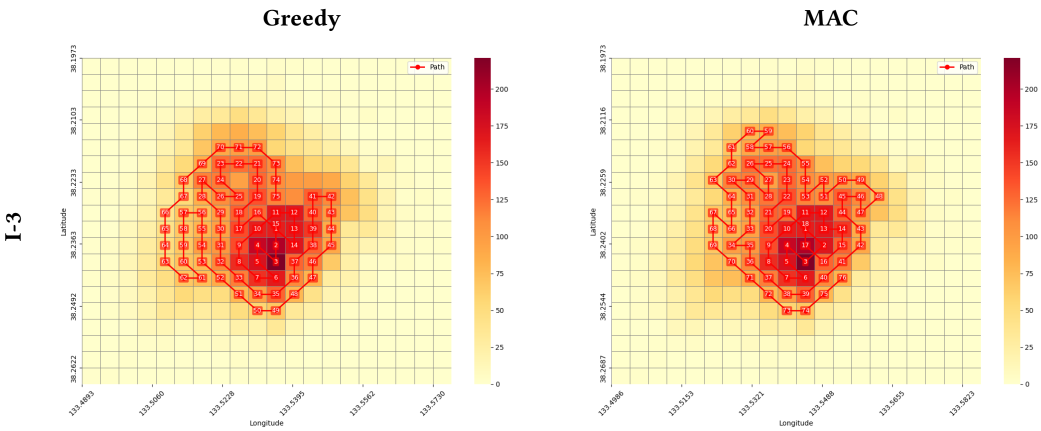

Figure 2: Comparison of UAV path planning results obtained by Greedy and MAC on I-3.

Figure 2 presents a comparison of UAV trajectories for I-3. The background heatmap represents the particle density distribution where darker red regions indicate higher density. The red lines depict the flight paths planned by the greedy algorithm and MAC. It is observed that MAC consistently navigates to and covers the high-density regions more effectively than the greedy algorithm.

The results demonstrate that the proposed method, MAC, exhibits superior capability in locating densely clustered particles. Unlike comparative algorithms, MAC robustly avoids local optima characterized by low-density cells and consistently converges to the global optimum.

5 Conclusion

In this paper, we proposed a novel framework for UAV path planning by combining the Actor-Critic method with a Memetic Algorithm. The proposed method, MAC, demonstrated that the integration of evolutionary computation significantly enhances the performance of Reinforcement Learning, particularly through its value-informed local refinement.

However, this study presents several challenges that remain to be addressed. First, the current system employs a single UAV, which inherently limits the detection rate; additionally, the absence of constraints such as collision avoidance may simplify the problem complexity compared to multi-agent scenarios. Second, while the current experiments utilized five instances derived from a single dataset with a fixed grid size, future research is required to optimize and validate the framework across varying map scales and resolutions to ensure broader scalability. Beyond maritime SAR, the proposed MAC framework may be applicable to a broader class of spatiotemporal search and coverage problems characterized by reward depletion and uncertainty, such as environmental monitoring and target tracking. These results suggest that MAC can serve as a general framework for integrating evolutionary search with value-based reinforcement learning.

Acknowledgments

This research was part of a project titled "Establishing a smart response platform for maritime accidents" (RS-2022-KS221629) funded by the Korea Coast Guard and supported by the Korea Institute of Marine Science & Technology Promotion. This work was also supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2026-25490713).

References

[1] Murat Bakirci. 2024. Evaluating the impact of unmanned aerial vehicles (UAVs) on air quality management in smart cities: A comprehensive analysis of transportation-related pollution. *Computers and Electrical Engineering* 119 (2024), 109556.

[2] Tiziana Calamoneri, Federico Corò, and Simona Mancini. 2024. Management of a post-disaster emergency scenario through unmanned aerial vehicles: Multi-depot multi-trip vehicle routing with total completion time minimization. *Expert Systems with Applications* 251 (2024), 123766.

[3] Shijie Fan, Ruichen Wang, Yang Song, David Crosbee, and Kang Su. 2025. Nature-Inspired Adaptive Differential Evolution: A Unified Meta-Heuristic Framework for Complex Engineering Optimisation and UAV Path Planning. *Results in Engineering* (2025), 106530.

[4] Felix Grumbach, Nour Eldin Alaa Badr, Pascal Reusch, and Sebastian Trojahn. 2023. A memetic algorithm with reinforcement learning for sociotechnical production scheduling. *IEEE Access* 11 (2023), 68760–68775.

[5] Haesang Jeong, Choong-Ki Kim, Seung-Yeol Hong, Jong-Hwui Yun, Do-Youn Kim, and Yong-Hyuk Kim. 2026. Genetic-algorithm-based heatmap optimization of maritime search and rescue unit deployment considering time-step dynamics. *Ocean Engineering* 347 (2026), 124029.

[6] Yan Jin, Yan Liao, Ali A Minai, and Marios M Polycarpou. 2006. Balancing search and target response in cooperative unmanned aerial vehicle (UAV) teams. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* 36, 3 (2006), 571–587.

[7] IAMSAR Manual. 2016. International aeronautical and maritime search and rescue manual. *International Maritime Organization, IMO Publishing, London* (2016).

[8] Shuguang Sun, Haolin Wang, Yanzhi Xu, Tianguang Wang, Ruihua Liu, and Wantong Chen. 2024. A Fusion Approach for UAV Onboard Flight Trajectory Management and Decision Making Based on the Combination of Enhanced A* Algorithm and Quadratic Programming. *Drones* 8, 6 (2024), 254.

[9] Dashuai Wang, Minghu Zhao, Zhuolin Li, Sheng Xu, Xiaohu Wu, Xuan Ma, and Xiaoguang Liu. 2025. A survey of unmanned aerial vehicles and deep learning in precision agriculture. *European Journal of Agronomy* 164 (2025), 127477.

[10] Meng Xi, Huiao Dai, Jingyi He, Wenjie Li, Jiabao Wen, Shuai Xiao, and Jiachen Yang. 2024. A lightweight reinforcement-learning-based real-time path-planning method for unmanned aerial vehicles. *IEEE Internet of Things Journal* 11, 12 (2024), 21061–21071.

[11] Chenjie Zhao, Mengwei Bao, Jinhao Liang, Maohan Liang, Ryan Wen Liu, and Guangyao Pang. 2025. Multi-sensor fusion-driven surface vessel identification and tracking using unmanned aerial vehicles for maritime surveillance. *IEEE Transactions on Consumer Electronics* (2025).