China Instrument and Control Society
Instrumentation
OPEN ACCESS
INSTRUMENTATION, Vol. 12, No. 4, December 2025 https://doi.org/10.15878/j.instr.202500303

Article

Multi-UAV Collaborative Path Planning Method Fusing Multi-Head Attention and SAC

Ziyi Zhu[1*], Ji Huang[1], Wangye Jiang[2]

[1] School of Electronic and Information Engineering, Liverpool John Moores University, Liverpool, UK(email: CITZZHU@ljmu.ac.uk; CITJHUA1@ljmu.ac.uk)

[2] School of Electronic and Information Engineering, Suzhou University of Technology, Suzhou, Jiangsu, China(email:z05122208@szut.edu.cn)

[*] Corresponding author email: CITZZHU@ljmu.ac.uk

Abstract: Aiming at the problem of low convergence efficiency of traditional multi-UAV path planning algorithms in unknown complex environments, this paper proposes a deep reinforcement learning algorithm incorporating the attention mechanism. The method is based on the Soft Actor-Critic (SAC) framework, which introduces a multi-attention mechanism in the Critic network, dynamically learns the dependency relationship between intelligences, and realizes key information screening and conflict avoidance. An environment with multiple random obstacles is designed to simulate complex emergent situations. The results show that the proposed algorithm significantly improves the mission success rate and average reward, significantly extends the survival time and exploration range of the UAVs, and verifies the effectiveness of the attention mechanism in enhancing the efficiency, robustness, and long-term planning capability of multi-UAV collaboration, as compared to the baseline method that does not use attention.

Keywords: Multi-UAV path planning; soft actor-critic; attention mechanism

Citation: Ziyi Zhu, Ji Huang, Wangye Jiang. "Multi-UAV Collaborative Path Planning Method Fusing Multi-Head Attention and SAC." Instrumentation 12, no.4 (December 2025). https://doi.org/10.15878/j.instr.202500303

1 Introduction

With the continuous development of science and technology, UAVs play an important role in various fields, such as military reconnaissance and disaster detection. These scenarios demand require UAVs to have more efficient and stable path planning capabilities under information interference.

Traditional UAV algorithms, operating as white-box models that is highly dependent on communication signals and environmental information, and are ill-suited to deployment in unknown and complex environments[1,2]. Mapping complex environments, especially in unexpected situations, demands significant time, which directly leads to a decrease in detection efficiency.

Deep reinforcement learning is a black-box model that improves flexibility and efficiency in unknown complex environments. For instance, Y. Peng et al. [3] investigated UAV path planning within drone-assisted edge computing networks based on deep reinforcement learning, demonstrating that DRL-based trajectory optimisation can significantly enhance system performance in dynamic environments. Similarly, Y. Lin et al. [4] proposed a DRL-based approach aimed at enhancing the autonomy of unmanned aerial vehicles in complex radio science environments. However, such methods primarily focus on single-UAV scenarios, neglecting robust control issues under multi-UAV collaboration or environmental disturbances. However, such approaches primarily focus on single-UAV scenarios, neglecting issues of multi-UAV collaboration or information interference.

With the continuous iteration and updating of deep reinforcement learning algorithms, many algorithms that can be applied to UAVs have been generated[5,6], offering a novel solution to the persistent challenge of UAV path

planning. The existing MADDPG algorithms are poor in terms of anti-interference and convergence efficiency. [7] integrated the SAC algorithm—grounded in maximum entropy reinforcement learning—into the original MADDPG framework to solve the dynamic path planning problem of UAVs.[8] proposed a multi-intelligent body algorithm (MARDPG) combining recurrent neural network (LSTM) and deep reinforcement learning (DRL) in reinforcement learning algorithm. In the continuous iterative improvement of the attention mechanism, refining network structures emerges as a critical breakthrough point. Various scholars have also attempted to incorporate the attention mechanism into the MADDPG and SAC algorithms[9,10], demonstrating the efficiency of the algorithms after incorporating the attention mechanism.

Therefore, in this paper, we optimise the structure of the Critic network and incorporate the attention mechanism into it, so as to improve the interaction efficiency and stability of the multi-intelligent body and enhance the path planning ability of the UAV.

Integrating the attention mechanism into the multi-UAV path planning system has several key advantages:

Enhanced inter-intelligence collaboration: the attention mechanism is able to learn the interdependence between UAVs, helping the model to understand when it is necessary to act collaboratively and when it is necessary to avoid.

Dynamic information filtering: in complex environments, the attention weights automatically focus on the most critical information for the current decision, ignoring irrelevant factors and improving learning efficiency.

Handling spatial relationships: In a multi-UAV environment, the relative position and state relationships of each UAV are very important, and the attention mechanism is particularly suitable for capturing such spatial relationships.

Avoiding group decision conflicts: By paying attention to the behaviour of other UAVs, each UAV can better predict and avoid potential conflict paths.

Long-term planning capabilities: Attention mechanisms help establish long-term dependencies, especially for complex paths that require multi-step planning.

Interpretability enhancement: Attention weights can be visualised to help understand what the model is focusing on in various situations, making the decision-making process more transparent.

Complementary advantages of multi-head architecture: the two-head attention architecture we implement allows the model to consider state information from different perspectives at the same time, similar to how humans consider multiple factors at the same time to make decisions. This architecture should significantly improve the ability of UAVs to learn complex collaborative strategies during training, especially in scenarios that require obstacle avoidance and mutual coordination.

2 Background

2.1 RL

Ryan Lowe, Yi Wu et al. proposed Multi-Agent Deep Deterministic Policy Gradient (MADDPG) framework to improve the stability of multi-agent systems in dynamic environments through centralized training and decentralized execution[5]. It includes a critic network and an actor network to update value functions and policies.

Critic Network: Each agent's Critic receives global states/actions includes the states and actions of all the agents, which are updated by minimising the Temporal Difference (TD) error. Target Network and Replay Buffer for experience are used to stabilise the training. The error update formula is as follows:

$$\mathcal{L}(\theta)=\mathbb{E}_{s,a,r,s'}[(Q^{*}(s,a|\theta)-y)^{2}],$$

$$\text{where} y=r+\gamma \max_{a'} \bar{Q}^{*}(s',a')$$

Actor network: based on deterministic policy gradients, updated via deterministic policy gradients using gradients estimated by the centralized Critic, relying only on local observations. The deterministic policy gradient theorem gives:

$$\nabla_{\theta_i} J(\boldsymbol{\mu}_i)=\mathbb{E}_{\mathbf{x},a\sim\mathcal{D}}[\nabla_{\theta_i}\boldsymbol{\mu}_i(a_i|o_i)\nabla_{a_i}Q_i^{\boldsymbol{\mu}}(\mathbf{x},a_1,...,a_N)|_{a_i=\boldsymbol{\mu}_i(o_i)}]$$

Describes an offline policy Actor-Critic algorithm Soft Actor-Critic (SAC) based on the maximum entropy reinforcement learning framework, which innovatively proposes a maximum entropy objective function to optimise the original policy, maximise the expected reward and policy entropy, and improve the exploration rate and robustness[6]. SAC employs a double Q-learning technique (utilizing twin Q-functions) to mitigate the overestimation bias commonly encountered in value-based methods, thereby enhancing convergence stability. By addressing key limitations of conventional reinforcement learning approaches—specifically, the tendency to converge to suboptimal local solutions and sample inefficiency—SAC demonstrates superior performance. The following is the objective function:

$$J(\pi)=\sum_{t}\mathbb{E}_{(\mathbf{s}_t,\mathbf{a}_t)\sim\rho_\pi}\left[r(\mathbf{s}_t,\mathbf{a}_t)+\alpha\mathcal{H}(\pi(\cdot|\mathbf{s}_t))\right]$$

In this experiment, we will adopt a multi-agent Actor-Critic algorithm to construct the basic algorithmic framework.

2.2 UAV Application

The article is based on the sac algorithm to solve the UAV path planning problem in dynamic and complex environments, and the reinforcement learning algorithm is practically applied in the UAV scenario, and the experiments prove the efficiency and stability of the sac algorithm in the application of the UAV scenario[7].

The experiments demonstrate the efficiency and stability of the SAC algorithm when applied to UAV scenarios. The details will be introduced in the experiment section.

2.3 Attention Mechanism

The attention mechanism in the Critic network will be similar to the model in [10], and this article confirms the efficiency of the MAAC algorithm compared to the MADDPD algorithm and SAC algorithm in the face of large-scale complex scenarios. MAAC does not rely on global state, supports cooperative, competitive, and mixed reward settings, and is compatible with heterogeneous action spaces (e. g., coexisting with discrete and continuous actions), making it suitable for a wide range of multi-intelligence scenarios. The addition of the attention mechanism enables Critic to dynamically select the attention weights for other intelligences' information, allowing the intelligences to automatically focus on other intelligences that are most relevant to the current task during training. As a result, MAAC can be better adapted to scenarios where UAVs are applied in complex environments than traditional algorithms. The following is the formula for calculating the MAAC:

$$\text{Attention}(Q,K,V)=\text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

3 Attention in DRL for UAVS control

3.1 Description of the Problem

This experiment randomly initializes the initial position (X, Y), velocity v and flight angle Ψ of three UAVs and adds five obstacles and a target location to simulate UAV flight in a complex environment. The environment size is 800*600 pixels, the radius of the UAVs is 10 pixels, the movement speed is 5-30 pixels per step, and the maximum steering rate is π/6 radians per step.

Termination conditions in our experiment:

1. collision with obstacle then mission fails reaches target location and no collision occurs then mission succeeds.

2. Failure if no collision occurs with an obstacle but the specified number of steps is exceeded to reach the target location.

3. Reaching the target location without collision succeeds the mission.

3.2 Environmental Design

According to the actual flight principle and obstacle avoidance scenarios of UAVs, the decision model of UAVs is formalized as 〈S, A, P, R, T〉, which denotes the state space S , the action space A, the state transfer function P, the reward function R, and the termination condition T of each UAV, respectively.

where:

State space S = [x, y, v, ψ]: denotes the position, speed, and heading angle of each UAV.

Action space A [speed_change, turn_rate]: modelling the speed change (-5 to 5) and turn rate (-π/6 to π/6) of the UAVs in terms of continuous quantities.

Reward function R: The target proximity, heading consistency, speed reasonableness, and collision risk are considered to design the reward function. Firstly for each UAV, alignment_reward, speed_reward, and movement_reward are calculated, followed by progress_reward, completion_reward, distance_penalty, task_completion_reward, formation_keeping_reward, collision_penalty, boundary_penalty. Finally normalisation is performed by dividing the total reward by 15.

It can be divided into three parts, the first part is the stand-alone reward, the reward function for directional consistency for each UAV is:

$$3^{\star}\left(\max\left(0,\cos\varphi\right)\right)^2$$

The speed matching function is:

$$2^{\star}\exp\left(-0.5^{\star}\left(\left(\mathbf{v_i}-\mathbf{v_{opt}}\right)/\left(\mathbf{v_{max}}/4\right)\right)\text{^}2\right)$$

And the movement reward.

The second part is the target proximity bonus and the global bonus, where the bonus increases plus 200 for each drone that encounters the target. the third part is the penalty term, where the bonus for any drone that collides with an obstacle or another drone is reduced by 150 as a collision penalty, and the boundary penalty is described using a smooth exponential decay function.

The termination condition T: consists of collision occurrence or mission completion, i. e., (a) either UAV collides with an obstacle or with another UAV (b) either UAV reaches the target.

In Figure 1, three UAVs colored red, yellow, and blue are respectively initialized. A larger green circle represents the destination, along with five randomly generated gray circular obstacles. Within such a simulated environment, the three UAVs will autonomously explore and plan paths until a termination condition is triggered, concluding the episode.

Fig.1 Simulated multi-UAV navigation environment with obstacles and target

3.3 Attention-critic Network

This experiment implements a critic network based on multi-head attention. First, input parameters, a base encoder, and the attention components are initialized. The state s and action a are concatenated and fed into the base encoder via forward propagation to obtain feature representations. These features are then projected into Key, Query, and Value vectors for each attention head. The dot product of the Query and Key matrices is computed, scaled, and then passed through a Softmax function. The resulting attention weights are used to weight the Value vectors, producing features for each head. The features from all heads are concatenated, merged with the original features (if applicable), and finally fed into the dual Q-network to output the Q-value.

3.4 Algorithmic Structure

The algorithm proposed in this paper is based on the Soft Actor-Critic (SAC) framework with several optimizations for the multi-UAV cooperative path planning problem. It mainly consists of an Actor network and two Critic networks with attention mechanisms.

Algorithm 1. Training Procedure of the Proposed Attention-Based SAC Method

Initialize the Critic networks Q1 and Q2, the Actor network, and the target network.
Initialize replay buffer R
For e = 1 to E **do**
Get the initial state of the environment s1
For t_{step} = 1 to T **do**
Selection of action by current policy $a_t = \pi_\theta(s_t)$
Each drone performs the action a_t, gets the reward r_t, sends the action to the environment, and the environment state changes to s_{t+1}
Store(s_t, a_t, s_{t+1}, r_t)in replay buffer R
if Ttotal≥Tmin then
for j=1 to K_{critic} do
Update critic network, compute target Q with attention
Compute critic loss
end for
for j=1 to K_{actor} do
Update actor network
end for
Soft target updates

$$\bar{\psi} = \tau\bar{\psi} + (1-\tau)\psi$$

$$\bar{\theta} = \tau\bar{\theta} + (1-\tau)\theta$$

Ttotal←0
end if
end for
end for

formula of target Q with attention is:

$$Q_{\text{targ}} \leftarrow r_i + \gamma(1-d)(\min_{k=1,2} Q_{\bar{\psi}_k}(o_i', \tilde{a}_i') - \alpha \log \pi_i)$$

critic loss is:

$$\mathcal{L}_{\text{critic}} \leftarrow \frac{1}{2}[\text{MSE}(Q_1, Q_{\text{targ}}) + \text{MSE}(Q_2, Q_{\text{targ}})]$$

4 Experiment Results

We designed a comparative experiment involving two distinct critic network architectures: the first incorporates a multi-head attention mechanism, and the second employs the classical double Q-network (serving as the baseline). This setup enables us to rigorously evaluate the impact of the attention mechanism on the learning performance and navigation capabilities of UAVs in complex environments. The overall architecture of the proposed attention-based actor–critic framework is illustrated in Figure 2.

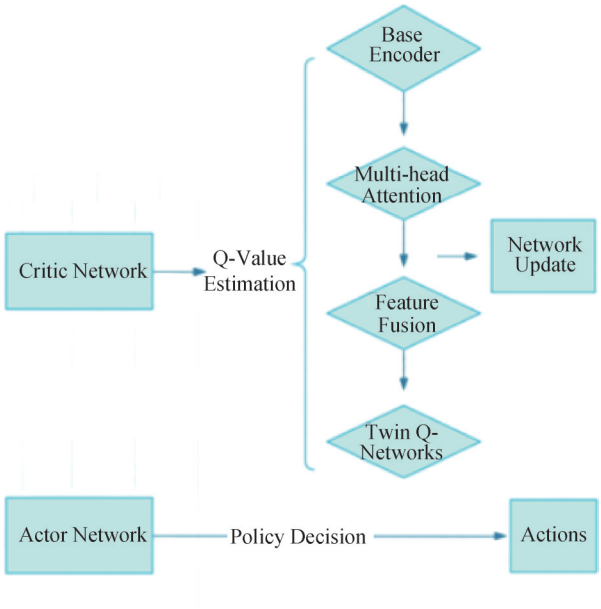

Fig.2. Architecture of the attention-based actor–critic network with multi-head attention and twin Q-networks

In this experiment, to simulate computational constraints under emergent circumstances, the model was trained using a CPU, specifically an AMD Ryzen 7 5700X, with a training time of approximately 25 minutes.

4.1 Parameters of this Experiment

Training parameters: The total number of training episodes is 300, with each episode having a maximum of 300 time steps. During training, a batch size of 128 samples is drawn from the experience replay buffer for learning updates. The discount factor γ is set to 0.98 to emphasize long-term returns, and the target network soft update coefficient τ is 0.01 to ensure stable learning. The neural network uses two hidden layers, each with 256 units. The experience replay buffer has a capacity of 500,000 transitions.

Optimiser configuration: The actor network is optimized with a learning rate of 3×10^{-4}, and the critic network uses the same learning rate of 3×10^{-4}. Training only begins after at least 500 steps of initial random exploration have been collected.

Additional settings: An entropy regularization coefficient α of 0.2 is used to encourage exploration, and a pulse exploration strategy is employed with a probability of 0.2. Reward scaling (0.25) and movement reward (1.5) are applied to shape agent behavior. A positive speed bias of 0.5 promotes forward movement, and exploration noise is set at 0.3 to enhance policy diversity.

4.2 Training Process and Results

We mainly compared the reward and success rate of the procedure with and without attention in each of the 300 epochs

Figure 3 shows the final reward comparison graph, every 20 rounds for detection, blue solid line indicates "with attention", red dotted line indicates "without attention". During the first 100 episodes, both algorithms are in the exploration phase, with their reward functions exhibiting significant fluctuations. In the subsequent 200 episodes, the fluctuations become more stable, demonstrating an overall upward trend. It can be seen that the traditional algorithm is better than the algorithm with the attention mechanism only in the beginning, but later the algorithm with attention is significantly better than the traditional algorithm.

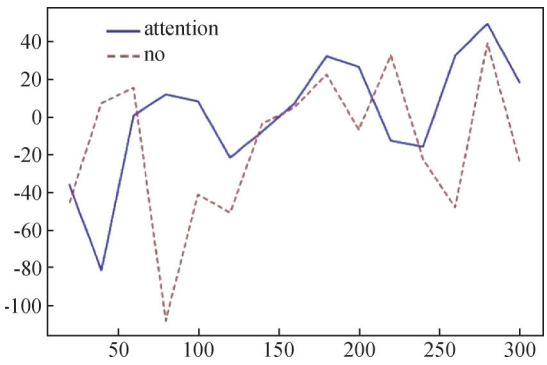

Fig.3. Comparison of cumulative rewards between the attention-based method and the baseline method without attention over 300 training episodes

Figure 4, employing the same plotting scheme as Figure 1, depicts the average number of steps taken by UAVs per episode. The results indicate that UAVs utilizing the attention-based critic consistently achieved significantly longer episode durations than those using the baseline. This enhanced survival time enabled them to navigate more extensively within the environment.

4.3 Result Analysis

Figure 3 shows that in the initial stage, the attention

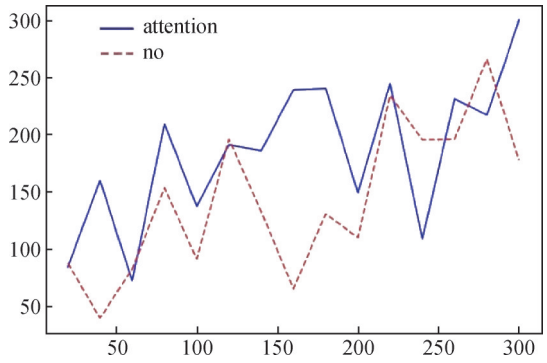

Fig.4 Comparison of the average number of steps per episode achieved by UAVs using the attention-based critic and the baseline critic.

module needs to learn to dynamically allocate drone attention weights, and the computational overhead during the initial learning phase is higher than that of traditional algorithms, which may lead to slower convergence. While the traditional algorithm relies on a simple structure, it is easier to discover feasible paths more readily under stochastic exploration policies in the early stages in the initial stage, which may also lead to the difference in the early stage. In the later stages, the addition of the attention mechanism allows the UAV to better effectively aggregate and prioritize relevant state information from multiple sources and compute safer navigation trajectories more efficiently, leading to higher rewards. Traditional algorithms in comparison are more likely to fall into a local optimum, leading to recurrent collision events.

Figure 4 shows that the UAV with the attention mechanism survives longer in complex environments and obtains more training samples, which also means that the UAV is able to detect the risk of collision earlier and avoid collisions, further expanding the scope of exploration.

Taken together, these results show that the multi-drone algorithm with the attention mechanism can better and faster adapt to unknown complex environments, effectively improving the efficiency in practical applications.

5 Conclusion

This experiment investigates the application of the attention mechanism sac algorithm in multi-UAV path planning. Through the analysis of experimental data, the analysis demonstrates the attention mechanism demonstrates marked superiority in multi-UAV complex path planning tasks through dynamic relationship modelling and directed exploration. Compared with the traditional algorithm, it can achieve accelerated convergence in the sudden-onset disaster scenarios in real scenarios. However, Owing to the computational complexity constraints of the algorithm, it is still

necessary to prune attention heads and refine reward shaping of the algorithm to reduce its computational complexity and recalibrate attention-weight incentives to achieve more efficient exploration.

Author Contribution:

Ziyi Zhu contributed to the conceptualization, methodology, software development, experiment design, data curation, formal analysis, and writing of the original draft.Ji Huang and Wangye Jiang contributed to literature search, formatting, proofreading, and manuscript revision. All authors have read and approved the final manuscript.

Funding Information:

This research received no external funding.

Data Availability:

The authors declare that the main data supporting the findings of this study are available within the paper and its Supplementary Information files.

Conflicts of Interest:

The authors declare no competing interests.

Dates:

Received 29 July 2025; Accepted 15 December 2025; Published online 31 December 2025

References

[1] S. Zhang & J. Liu (**2018**). Analysis and Optimization of Multiple Unmanned Aerial Vehicle-Assisted Communications in Post-Disaster Areas. IEEE Transactions on Vehicular Technology, 67(12)

[2] E. Mohsen, H. Huang, V S. Andrey, W. Ni (**2021**). *Navigation of a UAV Equipped with a Reconfigurable Intelligent Surface for LoS Wireless Communication with a Ground Vehicle. arXiv* :**2110**.09012

[3] Y. Peng, Y. Liu and H. Zhang (**2021**). Deep Reinforcement Learning based Path Planning for UAV-assisted Edge Computing Networks. ***2021*** *IEEE Wireless Communications and Networking Conference (WCNC)*

[4] Y. Lin, S. Zhang, F. Ye, T. Jiang, ang Y. Li (**2021**). *A UAV Path Planning Method Based on Deep Reinforcement Learning.* ***2020*** *IEEE USNC-CNC-URSI North American Radio Science Meeting (Joint with AP-S Symposium)*

[5] Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, Pieter Abbeel, Igor Mordatch (**2017**). Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *arXiv:1706.02275*

[6] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, Sergey Levine (**2018**). Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor. *arXiv: 1801.01290*

[7] Feisheng Yang , Chengliang Fang, and Ruijie Liang(**2024**) UAV Path Planning Based on Maximum Entropy Safe Reinforcement Learning National Natural Science Foundation of China (Number: 62073269)

[8] Y. Xue & W. Chen (**2024**). Multi-Agent Deep Reinforcement Learning for UAVs Navigation in Unknown Complex Environment. *IEEE Transactions on Intelligent Vehicles* , 8(3)

[9] H. Mao, Z. Zhang, Z. Xiao, Z. Gong (**2018**). Modelling the Dynamic Joint Policy of Teammates with Attention Multi-agent DDPG. *arXiv:1811.07029*

[10] Shariq Iqbal, Fei Sha (**2019**). Actor-Attention-Critic for Multi-Agent Reinforcement Learning. *arXiv:1810.02912*