Article

Not peer-reviewed version

Multi-Attention Meets Pareto Optimization: A Reinforcement Learning Method for Adaptive UAV Formation Control

Li Zheng [†] , Junjie Zeng [†] , Long Qin , Rusheng Ju [*]

Posted Date: 16 September 2025

doi: 10.20944/preprints202509.1303.v1

Keywords: attention mechanisms; Pareto optimization; multi-agent reinforcement learning; UAV formation control

Article

Multi-Attention Meets Pareto Optimization: A Reinforcement Learning Method for Adaptive UAV Formation Control

Li Zheng [†], Junjie Zeng [†], Long Qin and Rusheng Ju [*]

College of Systems Engineering, National University of Defense Technology (NUDT), Changsha 410073, China
[*] Correspondence: jurusheng@nudt.edu.cn
[†] These authors contributed equally to this work.

Abstract

Autonomous multi-UAV formation control in cluttered urban environments remains challenging due to partial observability, dense and dynamic obstacles, and conflicting objectives (task efficiency, energy use, and safety). Yet many MARL-based approaches still collapse vector-valued objectives into a single hand-tuned reward and lack selective information fusion, leading to brittle trade-offs and poor scalability in urban clutter. We introduce a model-agnostic MARL framework—instantiated on MADDPG for concreteness—that augments a CTDE backbone with three lightweight attention modules (self, inter-agent, and entity) for selective information fusion, and a Pareto optimization module that maintains a compact archive of non-dominated policies to adaptively guide objective trade-offs using simple, interpretable rewards rather than fragile weightings. On city-scale navigation tasks, the approach improves final *team* success by 13–27 percentage points for $N=2$–5 while simultaneously reducing collisions, tightening formation, and lowering control effort. These gains require no algorithm-specific tuning and scale smoothly beyond two agents, underscoring a stronger safety–efficiency trade-off and robust applicability in cluttered, partially observable settings.

Keywords: attention mechanisms; Pareto optimization; multi-agent reinforcement learning; UAV formation control

1. Introduction

With the rapid integration of unmanned aerial vehicles (UAVs) into real-world operations, their roles have expanded from agriculture and logistics to time-critical disaster response and persistent environmental monitoring. In particular, *urban and built-up environments*—characterized by dense buildings, occlusions, and narrow corridors—are becoming key application theaters for UAV swarms, e.g., communication relaying in "urban canyons," cooperative searching among high-rises, and safe navigation through cluttered streets and courtyards [1–3]. While multi-UAV formation can substantially enhance area coverage and resilience, achieving *stable, adaptive, and collision-free* coordination in these city-like scenarios remains challenging.

As the team size grows, multi-UAV control faces intertwined difficulties in perception, decision-making, and real-time execution. Concretely, we highlight three practical challenges. **(1) Selective information use under partial observability.** Each UAV must filter high-dimensional, multi-source inputs (self state, neighbors, and environment entities) to extract salient cues for timely decisions; without targeted filtering, decision latency and credit assignment deteriorate. Traditional rule-based and control-theoretic schemes—e.g., leader–follower and virtual-structure/potential-field designs [4,5], model predictive control (MPC) [6,7], and consensus/distributed optimization [8,9]—offer clarity and guarantees but are sensitive to modeling errors, communication delay, and dynamic clutter typical of urban scenes. **(2) Multi-objective trade-offs.** Formation keeping, obstacle avoidance, task efficiency, and energy economy often conflict; scalarizing them with fixed linear weights masks Pareto structure and yields brittle policies when mission priorities shift. **(3) Scalability and robustness.** In practice, packet loss, sensor noise, and non-stationarity degrade performance; vanilla deep MARL methods (e.g., MADDPG-style CTDE) improve coordination [10–12] yet still lack explicit mechanisms for dynamic information selection and principled multi-objective optimization, causing sharp performance drops as agent count or scene complexity increases [13].

To address these challenges, we present a *reinforcement learning method* that augments a CTDE-style backbone with *multi-source attention* and a *Pareto optimization module*. Specifically, we equip decentralized actors with three lightweight attention branches—self attention for intra-state feature selection, inter-agent attention for targeted neighbor reasoning, and entity attention for salient environment perception—whose outputs are concatenated into an attention-enhanced representation. In training, a vector-valued reward models task progress, energy, formation coherence, and safety; a Pareto module maintains a compact archive of non-dominated solutions and provides adaptive weights for updates, avoiding heavy manual reward tuning while preserving simple, interpretable shaping terms for each objective.In a representative 3D urban-like environment, the proposed modules consistently improve team success, safety, and formation quality across 2–5 UAVs with comparable or lower control effort; detailed results are reported in Section 6.

The main contributions are summarized as follows:

1. We develop a *multi-source attention design* (self/inter-agent/entity) for decentralized actors that selectively fuses critical cues from self, teammates, and urban environment entities, improving coordination efficiency and robustness under partial observability.
2. We introduce a *Pareto optimization module* for vector rewards that approximates the Pareto front during training, enabling adaptive trade-offs across task efficiency, formation coherence, energy, and safety with only simple, objective-wise shaping terms—*not* heavy ad hoc manual weighting.
3. We integrate the above as architecture-agnostic, plug-and-play modules for CTDE-style MARL and validate them in 3D city-like scenes. Across teams of $N = 2$–5 UAVs, inserting our modules into representative MARL backbones increases final *team* success by about $+12$–$+27$ percentage points and reduces collisions by roughly 20–30%, with tighter formation tracking at comparable or lower control effort; the gains persist from two to five agents, indicating effectiveness and scalability in complex urban environments.

The remainder of this paper is organized as follows: **Section 2** reviews related work. **Section 3** details background knowledge.**Section 4 Section 5** provides an in-depth explanation of our proposed algorithm framework, including the implementation details of the graph attention mechanism and the Pareto multi-objective optimization module. **Section 6** presents the detailed experimental setup and the analysis of the results. Finally, **Section 7** concludes the paper and discusses future research directions.

2. Related Work

2.1. Current Research on Multi-UAV Formation Control

Practical deployments are increasingly moving to *urban/built-up* scenes with dense buildings, occlusions, and narrow corridors, where multi-UAV formations support communication relaying, cooperative search, and safe navigation between high-rises. Rule-based and classical control schemes—such as leader–follower and virtual-structure/potential-field designs—remain popular for their clarity and ease of deployment [4,5]. Optimization-theoretic approaches including model predictive control (MPC) and consensus/distributed control offer stronger constraint handling and stability guarantees [6,7,9,14]. Deep reinforcement learning (DRL) has recently shown promise in handling high-dimensional observations and partial observability, and has been explored for formation, trajectory design, and cooperative navigation [11,15,16].

However, when mapped to city-like environments, these lines face three recurring challenges that align with our problem setting. **(i) Selective information use under partial observability:** controllers must extract salient cues from *self*, *neighbors*, and *environmental entities* in high-dimensional, cluttered scenes; fixed-rule filters or hand-crafted interfaces struggle as complexity grows. **(ii) Multi-objective trade-offs:** formation coherence, obstacle avoidance/safety, task efficiency, and energy economy often conflict; scalarizing them with fixed linear weights blurs Pareto structure and leads to brittle behavior when mission priorities shift. **(iii) Scalability and robustness:** communication delays, packet loss, and sensor noise degrade coordination, and performance tends to drop sharply as agent count or urban clutter increases [11]. Within optimization/control methods, even with disturbance observers and estimation filters (e.g., Kalman-consensus and disturbance observers in MPC pipelines [7]), the burden of online optimization and model mismatch in cluttered 3D geometry limits agility. In DRL pipelines, the absence of *explicit* mechanisms for dynamic information selection and principled multi-objective optimization remains a key gap.

2.2. Attention Mechanisms for Collaborative Perception and Decision-Making

Attention has been introduced to enhance multi-agent perception/communication and to focus computation on salient cues in cooperative UAV tasks. For trajectory design and resource assignment, graph attention has improved performance by letting agents emphasize critical neighbors and links [17]. For cooperative encirclement/rounding, multi-head soft attention yields targeted coordination signals [18]. Transformer-style designs with virtual objects have been used for short-range air combat maneuver decision, showing improved decision quality via structured attention to key entities [19]. In adversarial/dangerous settings such as missile avoidance, multi-head attention helps capture dynamic obstacles and threat saliencies [20].

These studies collectively indicate that *multi-source attention* (self/neighbor/entity) can improve collaborative perception and decision quality. At the same time, prior work typically optimizes a *single* scalarized return and does not explicitly couple attention with multi-objective value estimation; as a result, policies may overfit to a particular weight setting and generalize poorly when objective priorities change (e.g., switching from aggressive goal-seeking to safety-first in narrow corridors). Our method targets this gap by pairing lightweight attention branches with vector-valued critics and Pareto-aware training.

2.3. Advances in Multi-Objective Optimization (MOO) and Pareto Methods

Pareto-based multi-objective optimization (MOO) offers a principled way to expose trade-offs among conflicting objectives without collapsing them into a single weighted sum. In UAV-related literature, NSGA-III and variants have been applied to task allocation and planning under complex constraints [21,22]; MOEA/D with adaptive weights has improved solution-set uniformity and has been used for 3D path planning [23]; and dynamic multi-objective resource allocation has been investigated in related communication/energy settings [24,25]. These techniques are effective at *offline* design and static instances, but their computational footprint and lack of tight coupling with the perception–decision pipeline make end-to-end, *online* deployment in cluttered, partially observable environments difficult.

Accordingly, there is a need to *integrate* Pareto reasoning into learning-based coordination rather than treating MOO as an external, offline post-processor. Our approach follows this direction: we retain simple, objective-wise shaping signals (task progress, energy, formation coherence, safety) and train vector-valued critics, while a compact Pareto archive provides adaptive training weights to encourage non-dominated policy updates *within* the MARL loop.

2.4. Summary and Gaps

Summarizing, (i) classical rule/optimization controllers are strong when models and communication are reliable but struggle to *select* salient information and adapt in cluttered urban scenes; (ii) DRL scales to high-dimensional observations yet often relies on ad hoc scalarization, lacking a principled mechanism to balance competing goals; and (iii) existing attention applications improve perception/coordination but are typically optimized for a single weighted objective, limiting robustness when priorities change. This paper addresses these gaps by combining *multi-source attention* (self/inter-agent/entity) for selective information fusion with a *Pareto module* for vector rewards, integrated into a CTDE-style method so that non-dominated trade-offs are discovered *during* training rather than predetermined by fixed weights.

3. Background Knowledge

3.1. Multi-Agent Reinforcement Learning

Multi-Agent Reinforcement Learning (MARL) is a key framework for handling multiple agents that pursue cooperative or competitive goals through sequential decisions in a shared environment. Unlike single-agent reinforcement learning, MARL faces core challenges: environmental non-stationarity, credit assignment, and mutual policy influence among agents. In UAV formation control, these challenges are pronounced. Each UAV must act on local observations, yet their joint actions determine overall system performance.

To address these issues, many MARL paradigms have been proposed. Centralized Training with Decentralized Execution (CTDE) is the mainstream. Its idea is to use global information during training to learn cooperative policies, while at execution each agent relies only on its own observations. This balances model expressiveness and system scalability. A representative algorithm is MADDPG. It combines a centralized critic with decentralized actors and mitigates non-stationarity to some extent.

However, traditional MARL still has clear limits. Agents often lack selective perception of multi-source state information and cannot focus on key cues in complex environments. Most methods also rely on a scalar reward formed by linear weighting of multiple objectives. This cannot capture complex trade-offs among objectives. These limits motivate our use of attention mechanisms and multi-objective optimization to improve MARL for UAV formation control.

3.2. Attention Mechanisms

Attention provides *selective information fusion*: a model assigns higher weights to salient parts of its inputs and suppresses distractions, thereby improving long-range dependency modeling and feature prioritization. Beyond its well-known success in NLP and CV, attention is increasingly used in multi-agent systems (MAS)—including cooperative UAV control—to filter self states, neighbor cues, and environmental entities under partial observability and communication imperfections.

We use the standard scaled dot-product attention as the basic operator:

$$\mathrm{Attn}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V, \tag{1}$$

where *queries* Q encode the current information need, *keys* K index candidate features, and *values* V carry the content to aggregate. The softmax normalizes relevance scores into a distribution and yields a context vector by weighted summation. Multi-head variants apply (1) in parallel and concatenate the outputs for richer feature subspaces.

In the context of UAV formation control, attention is particularly useful for:

- **Selective perception**: highlight task-relevant parts of the local observation (e.g., goal direction, energy, safety margins).
- **Targeted coordination**: focus on the most influential neighbors for collision avoidance and formation keeping.
- **Salient environment awareness**: emphasize nearby obstacles or bottlenecks in cluttered, urban-like scenes.

These properties make attention a natural fit for CTDE-style MARL: it improves the actors' input representations under partial observability and reduces non-stationarity seen by centralized critics. In Section 4 and Section 5 we instantiate this idea via self-, inter-agent-, and entity-attention modules tailored to UAV teams.

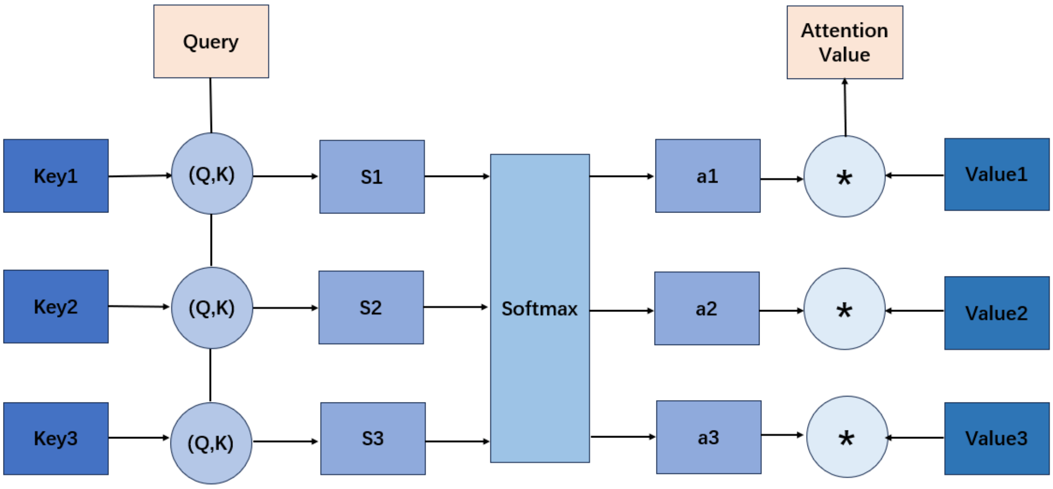

Figure 1. Schematic of scaled dot-product attention: relevance is computed via $QK^{\top} / \sqrt{d_k}$, normalized by softmax, and applied to V to form the context.

3.3. Multi-Objective Optimization

Multi-objective optimization handles trade-offs among conflicting objectives. The core concept is Pareto optimality: a solution set where no objective can be improved without worsening another. The problem is formalized as:

$$\min_{\mathbf{x} \in \mathcal{X}} [f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_k(\mathbf{x})]^T \tag{2}$$

In multi-UAV cooperative control, common objectives include path length, energy consumption, safety, and formation-keeping accuracy. Traditional methods often use a scalarization function to convert multiple objectives into a single one, but this cannot fully capture complex trade-offs. Pareto-based methods provide a set of optimal trade-off solutions and give richer information for decision-making.

In summary, this paper integrates multi-agent reinforcement learning, attention mechanisms, and multi-objective optimization to build a collaborative decision framework for multi-UAV formation control. It addresses multi-objective coordination challenges in dynamic environments.

4. Problem Formulation

We study cooperative multi–UAV formation control in a built urban area. A team departs from randomized starts and moves to a goal region while (i) avoiding collisions with buildings and teammates, (ii) keeping a prescribed formation, (iii) limiting control/energy usage, and (iv) finishing within a time budget. This setting is representative of city–scale sensing and relay missions, where formation coherence benefits coverage and link reliability.

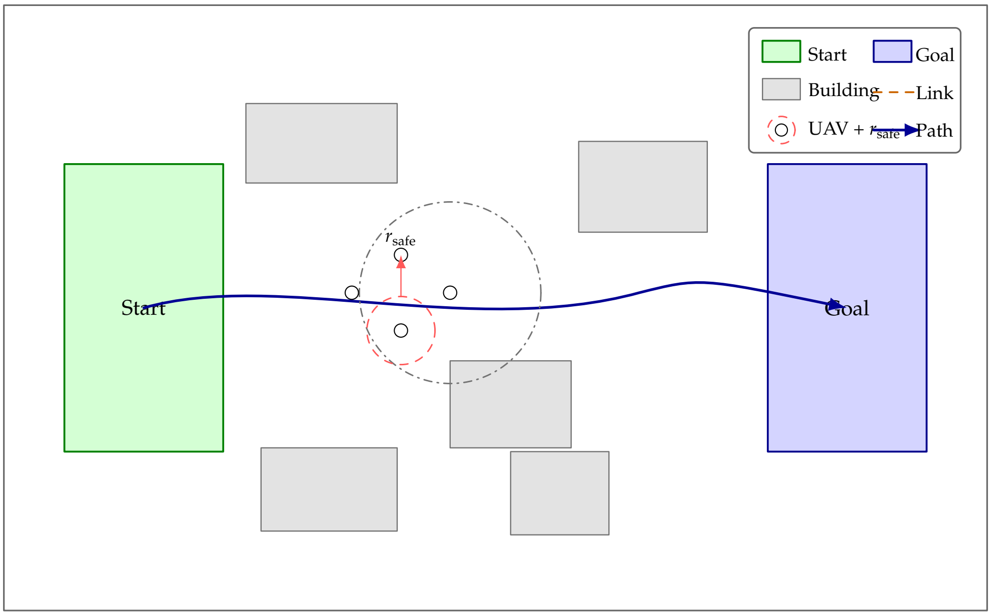

Figure 2. Problem setup (top view). A team departs from a start region, navigates among buildings, maintains a diamond formation with safety radius r_{safe}, uses local sensing and limited neighbor links, and follows a feasible path toward the goal under energy/time budgets.

4.1. Game Model and CTDE Setting

We model the task as a partially observable Markov game (POMG) $\mathcal{G} = (\mathcal{I}, \mathcal{S}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, P, \{R_i\}, \gamma)$ with discrete time $t = 0, \ldots, T$ and step Δt. The agent set is $\mathcal{I} = \{1, \ldots, N\}$. Training follows CTDE: a centralized critic sees global information during learning, whereas execution relies only on local observations.

For the critic, the global state stacks team kinematics, the goal and the formation blueprint, and nearby obstacles:

$$s_t = \big[X_t,\ V_t,\ G,\ \mathcal{F},\ E_t\big], \tag{3}$$

where $X_t = [x_{1,t}, \ldots, x_{N,t}]$ and $V_t = [v_{1,t}, \ldots, v_{N,t}]$ are positions and velocities, G encodes the goal pose/region, $\mathcal{F}$ stores desired slot offsets $\{r_i^\star\}_{i=1}^N$ (formation template), and $E_t = \{b_k\}_{k=1}^M$ lists the M nearest axis–aligned buildings represented by centers and half–sizes.

4.2. Observations and Actions

Each agent observes only local information. To match the three attention branches used later, we organize the local observation into three parts and then fuse them as

$$\tilde{s}_{i,t} = \text{Concat}\big(\text{SelfAtt}(o_{i,t}^{\text{self}}),\ \text{InterAtt}(o_{i,t}^{\text{inter}}),\ \text{EntityAtt}(o_{i,t}^{\text{ent}})\big). \tag{4}$$

Self features. $o_{i,t}^{\text{self}} = \big[x_{i,t},\ v_{i,t},\ g_{i,t},\ e_{i,t},\ d_{i,t}^{\text{goal}}\big]$, where $x_{i,t} \in \mathbb{R}^3$ and $v_{i,t} \in \mathbb{R}^3$ are position and velocity, $g_{i,t}$ is heading, $e_{i,t} \in [0,1]$ is normalized remaining energy, and $d_{i,t}^{\text{goal}} = \|x_{i,t} - x^{\text{goal}}\|_2$ is distance to the goal. **Inter–agent features.** For the K nearest neighbors $\mathcal{N}_i^K$, we use relative kinematics $o_{i,t}^{\text{inter}} = \{\Delta x_{ij,t} = x_{j,t} - x_{i,t},\ \Delta v_{ij,t} = v_{j,t} - v_{i,t},\ d_{ij,t}^{\text{pair}} = \|\Delta x_{ij,t}\|_2\}_{j \in \mathcal{N}_i^K}$ with fixed K (e.g., $K = 4$) recomputed each step. **Entity features.** For the M closest buildings to agent i, $o_{i,t}^{\text{ent}} = \{\Delta x_{ik,t} = c_k - x_{i,t},\ h_k,\ w_k,\ \ell_k\}_{k=1}^M$, where c_k is the building center and (h_k, w_k, ℓ_k) are half–sizes; a small M (e.g., $M = 10$) keeps inference time predictable. Actions are 3-D thrust/velocity commands subject to a magnitude bound,

$$a_{i,t} \in \mathcal{A}_i \subset \mathbb{R}^3, \qquad \|a_{i,t}\|_2 \le a_{\max}. \tag{5}$$

4.3. Vector Reward and Termination

Control is multi-objective. Each agent receives a four-dimensional reward covering task progress, energy, formation coherence, and safety,

$$R_i(s_t, a_t, s_{t+1}) \;=\; \{r_i^{\text{task}},\, r_i^{\text{energy}},\, r_i^{\text{formation}},\, r_i^{\text{safety}}\}, \tag{6}$$

with homogeneous definitions and fixed coefficients across runs (defaults in parentheses). **Task progress and success.** With $\eta_{\text{succ}} = 5.0$ and $c_{\text{prog}} = 0.5$,

$$r_i^{\text{task}} \;=\; \eta_{\text{succ}} \cdot \mathbf{1}\Big\{ d_{i,t+1}^{\text{goal}} \le \varepsilon_{\text{goal}} \Big\} \;-\; c_{\text{prog}} \cdot \big(d_{i,t+1}^{\text{goal}} - d_{i,t}^{\text{goal}} \big), \tag{7}$$

so moving closer to the goal is rewarded each step and entering the $\varepsilon_{\text{goal}}$-ball yields a one-off bonus. **Energy/control.** With $c_{\text{eng}} = 0.01$,

$$r_i^{\text{energy}} \;=\; -\, c_{\text{eng}} \cdot \|a_{i,t}\|_2^2. \tag{8}$$

Formation coherence. With $c_{\text{form}} = 1.0$ and $\varepsilon_{\text{form}} = 1.0$,

$$\text{dev}_t \;=\; \frac{1}{N}\sum_{i=1}^{N} \big\| x_{i,t} - x_{i,t}^{\text{ref}} \big\|_2, \qquad r_i^{\text{formation}} \;=\; -\, c_{\text{form}} \cdot \text{dev}_t \;+ 0.2 \cdot \mathbf{1}\{\text{dev}_t \le \varepsilon_{\text{form}}\}, \tag{9}$$

where the reference slot $x_{i,t}^{\text{ref}} = x_{\text{lead},t} + r_i^{\star}$ follows a (virtual) leader; small average slot error is mildly rewarded. **Safety.** With $c_{\text{col}} = 5.0$, $c_{\text{near}} = 1.0$, and $\delta = 2.0$ m,

$$r_i^{\text{safety}} \;=\; -\, c_{\text{col}} \cdot \mathbf{1}\{\text{collision at } t\} \;-\; c_{\text{near}} \cdot \mathbf{1}\Big\{ \min\big(d_{ij,t}^{\text{pair}},\, d_{i,t}^{\text{obs}}\big) < \delta \Big\}, \tag{10}$$

where $d_{i,t}^{\text{obs}}$ is the distance to the nearest building surface. An episode ends when all agents are in the goal region, upon any collision, or at the horizon T.

4.4. Objective and CTDE Realization

We seek decentralized actors that are Pareto–efficient with respect to the four objectives under a centralized critic. Let the joint policy be $\pi = \{\pi_i\}_{i=1}^{N}$ with per–agent actor $\pi_i(a_{i,t} \,|\, \tilde{s}_{i,t})$. The vector return averaged across agents is

$$\mathbf{J}(\pi) \;=\; \mathbb{E}\Bigg[\sum_{t=0}^{T} \gamma^t \cdot \frac{1}{N} \sum_{i=1}^{N} R_i(s_t, a_t, s_{t+1}) \Bigg], \tag{11}$$

and policies are ordered by Pareto dominance (no worse in all objectives and strictly better in at least one). The centralized critic estimates per–objective values that guide updates,

$$Q^{(\kappa)}(s_t, a_t), \qquad \kappa \in \{\text{task},\ \text{energy},\ \text{formation},\ \text{safety}\}, \tag{12}$$

while the actors consume $\tilde{s}_{i,t}$ constructed in (4). This separation keeps modeling choices (rewards and constraints) and architecture (attention and CTDE) cleanly decoupled.

Table 1. Notation used in the formulation.

Symbol	Description
$\mathcal{I}, N$	Agent set and its size
$t, T, \Delta t$	Time index, horizon, and time step
$s_t, o_{i,t}, a_{i,t}$	Global state, local observation, and action
G	Goal pose/region encoding
$\mathcal{F}, r_i^\star$	Formation template and per–agent slot offset
E_t, M	Set of M nearest buildings (centers and half–sizes)
K	Number of neighbor slots in o^{inter}
R_i	Reward vector in (6)
γ	Discount factor
π_i, π	Decentralized policy and joint policy
$Q^{(\kappa)}$	Per–objective centralized critic in (12)
$a_{\max}, \delta, \varepsilon_{\text{goal}}$	Action bound, safety distance, goal threshold

5. Proposed Method

5.1. Framework Overview

This paper proposes an improved framework for autonomous UAV formation control that is *algorithm-agnostic*: the two core components—a **multi-source attention block** (self / inter-agent / entity) and a **Pareto multi-objective optimization layer** with vector-valued critics—can be plugged into *any* CTDE-style actor–critic MARL algorithm. In this work we *instantiate* the framework with **MADDPG** to provide a concrete realization and fair comparison, but the design applies equally to other backbones (e.g., MATD3/MASAC/MAPPO).

MADDPG, a common MARL method, follows the Centralized Training–Decentralized Execution (CTDE) paradigm: a centralized critic evaluates global information during training, and decentralized actors execute from local observations. Building on this backbone, we introduce attention mechanisms to enhance the handling and fusion of local observations under partial observability. As shown in Figure 3, the overall pipeline still follows CTDE but is adapted to multi-objective formation control.

Specifically, each UAV decides from its local observation, while the Attention Enhancement Layer applies three lightweight modules—self attention, inter-agent attention, and entity attention—to selectively emphasize salient cues. The three outputs are concatenated into an attention-enhanced representation $\tilde{s}_i = \text{Concat}\big(\text{SelfAtt}(o_i^{\text{self}}),\ \text{InterAtt}(o_i^{\text{inter}}),\ \text{EntityAtt}(o_i^{\text{ent}})\big)$. The Actor maps $\tilde{s}_i$ to action a_i, and centralized Critics estimate per-objective values $\{Q^{\text{task}}, Q^{\text{energy}}, Q^{\text{formation}}, Q^{\text{safety}}\}$ in parallel. During training, the Pareto layer maintains a compact archive of non-dominated solutions and provides adaptive signals/weights to guide updates, thereby avoiding brittle manual scalarization. We store $(s, a, \mathbf{r}, s', \text{attention context})$ in an attention-aware replay buffer and optimize a combined loss $L_{\text{total}} = \lambda_1 L_{\text{critic}} + \lambda_2 L_{\text{actor}} + \lambda_3 L_{\text{attention}} + \lambda_4 L_{\text{pareto}}$.

In summary, when instantiated with MADDPG (Figure 3), the attention block improves information saliency under partial observability and the Pareto layer enforces principled multi-objective trade-offs; *more generally*, these two modules augment the chosen MARL backbone with minimal changes and are applicable to a wide range of CTDE actor–critic methods beyond MADDPG.

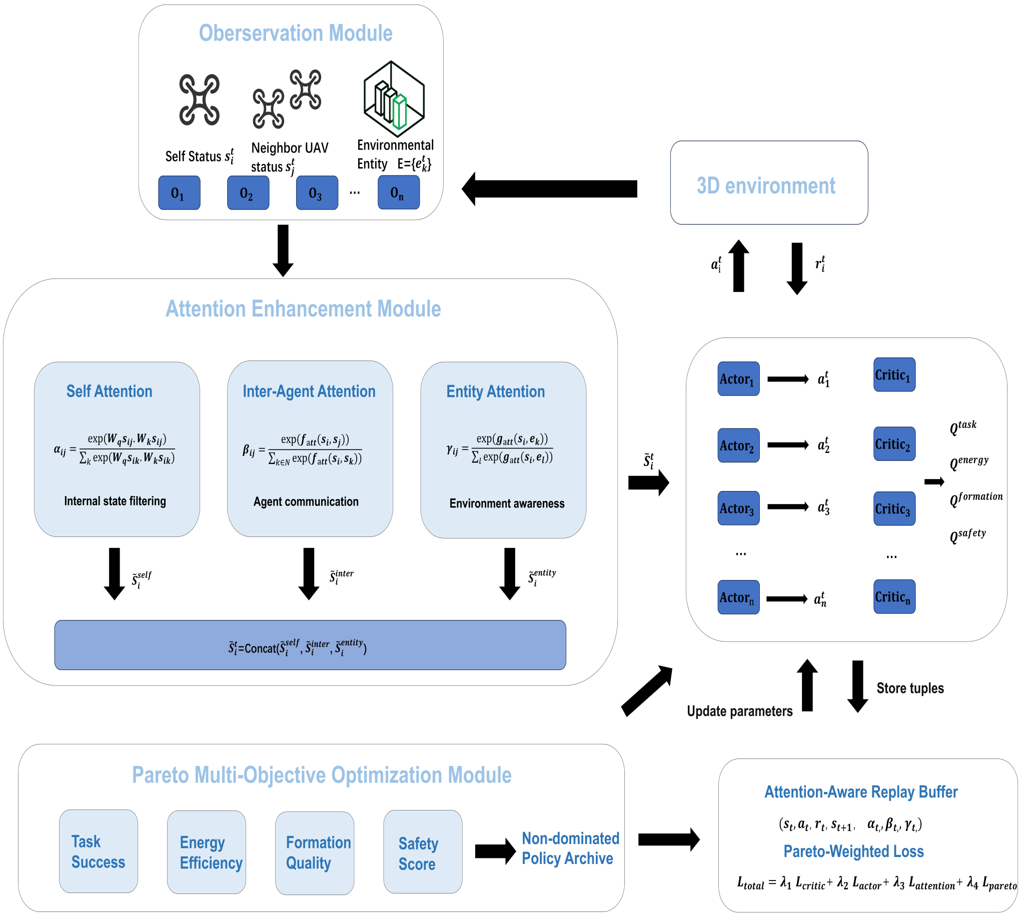
Oberservation Module
Self Status s_i^t
Neighbor UAV status s_j^t
Environmental Entity E={e_k^t}
o_1
o_2
o_3
...
o_n
3D environment
a_i^t
r_i^t
Attention Enhancement Module
Self Attention
$\alpha_{ij} = \frac{\exp(W_q s_{ij} \cdot W_k s_{ij})}{\sum_k \exp(W_q s_{ik} \cdot W_k s_{ik})}$
Internal state filtering
Inter-Agent Attention
$\beta_{ij} = \frac{\exp(f_{att}(s_i, s_j))}{\sum_{k \in N} \exp(f_{att}(s_i, s_k))}$
Agent communication
Entity Attention
$\gamma_{ij} = \frac{\exp(g_{att}(s_i, e_k))}{\sum_l \exp(g_{att}(s_i, e_l))}$
Environment awareness
$\tilde{S}_i^{self}$
$\tilde{S}_i^{inter}$
$\tilde{S}_i^{entity}$
$\tilde{S}_i^t$=Concat($\tilde{S}_i^{self}$, $\tilde{S}_i^{inter}$, $\tilde{S}_i^{entity}$)
$\tilde{S}_i^t$
Actor$_1$
a_1^t
Actor$_2$
a_2^t
Actor$_3$
a_3^t
...
Actor$_n$
a_n^t
Critic$_1$
Critic$_2$
Critic$_3$
...
Critic$_n$
Q^{task}
Q^{energy}
$Q^{formation}$
Q^{safety}
Update parameters
Store tuples
Pareto Multi-Objective Optimization Module
Task Success
Energy Efficiency
Formation Quality
Safety Score
Non-dominated Policy Archive
Attention-Aware Replay Buffer
$(s_t, a_t, r_t, s_{t+1}, \alpha_{t,}, \beta_{t,}, \gamma_{t,})$
Pareto-Weighted Loss
$L_{total} = \lambda_1 L_{critic} + \lambda_2 L_{actor} + \lambda_3 L_{attention} + \lambda_4 L_{pareto}$

Figure 3. Overall framework: plug-and-play attention (self / inter-agent / entity) before each decentralized actor, and a Pareto layer over vector critics during centralized training. We instantiate with a MADDPG backbone for concreteness, but the modules are applicable to other CTDE actor–critic MARL algorithms.

5.2. Attention Mechanism Integration

The attention mechanisms in our framework serve three distinct but complementary purposes: enhancing inter-agent communication, improving environmental perception, and enabling hierarchical decision-making. Each mechanism addresses specific challenges in multi-agent coordination while contributing to the overall system performance. Below we detail the three attention mechanisms integrated in our framework.

Self-Attention Mechanism: The self-attention mechanism processes the internal state representation of each UAV agent to identify the most relevant features for decision-making. Given an agent's state vector $s_i \in \mathbb{R}^d$, the self-attention module computes attention weights $\alpha_{i,j}$ for each state component j:

$$\alpha_{i,j} = \frac{\exp(W_q s_{i,j} \cdot W_k s_{i,j})}{\sum_{k=1}^{d} \exp(W_q s_{i,k} \cdot W_k s_{i,k})} \tag{13}$$

where W_q and W_k are learned query and key transformation matrices. This mechanism enables agents to dynamically prioritize different aspects of their internal state based on the current situation, improving decision quality in complex scenarios through adaptive feature weighting.

Inter-Agent Attention Mechanism: The inter-agent attention mechanism facilitates explicit communication between UAV agents by computing attention weights over neighboring agents' states. For agent i with neighbors $\mathcal{N}_i$, the inter-agent attention weight for neighbor j is computed as:

$$\beta_{i,j} = \frac{\exp(f_{att}(s_i, s_j))}{\sum_{k \in \mathcal{N}_i} \exp(f_{att}(s_i, s_k))} \tag{14}$$

where f_{att} is a neural network that computes the relevance score between agent i and agent j. This mechanism allows agents to selectively focus on the most relevant teammates for coordination, enabling effective formation maintenance and collision avoidance through targeted information exchange.

Entity Attention Mechanism: The entity attention mechanism processes environmental entities such as obstacles, targets, and dynamic elements. Given a set of environmental entities $E = \{e_1, e_2, ..., e_m\}$, the mechanism computes attention weights to determine the relevance of each entity to the current agent:

$$\gamma_{i,k} = \frac{\exp(g_{att}(s_i, e_k))}{\sum_{l=1}^{m} \exp(g_{att}(s_i, e_l))} \tag{15}$$

This mechanism enables agents to dynamically focus on the most relevant environmental features, significantly improving navigation efficiency and obstacle avoidance capabilities by filtering irrelevant sensory input.

Within the CTDE pipeline, attention acts as the front end of representation. Self–attention filters an agent's own kinematics and intent, inter–agent attention highlights the few teammates that matter for the current maneuver, and entity attention foregrounds the most influential obstacles.The fused vector $\tilde{s}_i$ is therefore more structured and temporally stable than raw observations, so the centralized multi–objective critics receive inputs in which progress, formation deviation, energy use, and risk are easier to tease apart.In practice this yields cleaner per–objective value estimates and crisper policy gradients, reducing ambiguity about *which* agent and *which* objective should change.Attention thus strengthens state representation on the actor side and, as a consequence, helps the critics allocate credit with fewer confounding correlations.

5.3. Pareto Multi-Objective Optimization

The Pareto optimization component addresses the inherent multi-objective nature of UAV formation control, where agents must simultaneously optimize competing objectives including task completion, energy efficiency, formation maintenance, and collision avoidance. Traditional reinforcement learning approaches struggle with such multi-objective scenarios due to the difficulty in defining appropriate reward weightings.

Our Pareto-based framework formulates the problem as a multi-objective optimization where the reward function R_i for agent i consists of distinct objective components:

$$R_i = \{r_i^{task}, r_i^{energy}, r_i^{formation}, r_i^{safety}\} \tag{16}$$

each representing critical operational dimensions. The approach maintains a set of non-dominated solutions, enabling exploration of diverse objective trade-offs without manual tuning of reward weights.

The Pareto dominance relationship is defined such that solution x dominates solution y if x is at least as good as y in all objectives and strictly better in at least one objective. The algorithm maintains an archive of non-dominated solutions to guide policy updates:

$$\pi_{\theta_i}^{new} = \arg\max_{\theta_i} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R_i(s_t, a_t, s_{t+1})\right] \tag{17}$$

The above update is carried out under explicit Pareto-optimality constraints, which preserve non-dominated solutions and encourage coverage of diverse trade-offs across objectives. This constraint guarantees that selected policies represent efficient compromises between competing goals.

This approach fundamentally eliminates the need for manual reward engineering while ensuring robust performance across all operational dimensions. The practical significance is particularly valuable in UAV missions where objective priorities dynamically shift during different mission phases, providing adaptive optimization without parameter recalibration.

5.4. Integrated Framework Architecture

The integrated framework incorporates attention mechanisms with Pareto optimization within a modified MADDPG architecture through a hierarchical processing pipeline. This unified structure comprises three principal components: an attention-enhanced observation processor, a multi-objective critic network, and a coordinated actor network, collectively forming the core innovation of our approach.

The attention-enhanced observation processor transforms raw sensory inputs into enriched state representations through sequential attention layers. This module simultaneously applies self-attention to internal states, inter-agent attention to neighboring UAV states, and entity attention to environmental features. The resulting representations are concatenated to form a comprehensive state vector:

$$\tilde{s}_i = \text{Concat}\big(\text{SelfAtt}(s_i), \text{InterAtt}(s_i, s_{\mathcal{N}_i}), \text{EntityAtt}(s_i, E)\big) \tag{18}$$

where each attention module operates according to the mechanisms defined in Section 3.1.2.

The multi-objective critic network extends conventional value estimation by maintaining separate value functions for each operational dimension:

$$Q_{\phi_i}(s,a) = \left[Q^{\text{task}}_{\phi_i}(s,a), Q^{\text{energy}}_{\phi_i}(s,a), Q^{\text{formation}}_{\phi_i}(s,a), Q^{\text{safety}}_{\phi_i}(s,a)\right] \tag{19}$$

This architectural innovation enables distinct value estimation for competing objectives, facilitating precise credit assignment during policy updates under Pareto constraints.

The coordinated actor network synthesizes the attention-enhanced state representations into actions that balance individual objectives with collective coordination requirements. To ensure training stability in decentralized execution, the network architecture incorporates residual connections and layer normalization techniques:

$$a_i \sim \pi_{\theta_i}(\cdot|\tilde{s}_i) \tag{20}$$

A critical feedback loop emerges between these components: attention mechanisms dynamically inform policy decisions, which subsequently reshape attention patterns as environmental conditions evolve. This adaptive interaction enables continuous optimization of mission-specific trade-offs without manual parameter adjustment.

5.5. Training Algorithm and Implementation

Our training methodology extends the MADDPG framework with integrated attention mechanisms and Pareto optimization, creating a robust learning system for multi-UAV coordination. The algorithm maintains the CTDE paradigm, where centralized critics leverage global information during training while decentralized actors operate solely on local observations during mission execution. This architecture preserves the scalability advantages of decentralized systems while benefiting from centralized learning.

A critical innovation lies in the attention-aware experience replay mechanism. Traditional experience tuples (s_t, a_t, r_t, s_{t+1}) are augmented with attention context vectors $(\alpha_t, \beta_t, \gamma_t)$ capturing the instantaneous focus patterns across all three attention mechanisms. This preservation of attention context during replay significantly enhances learning stability and prevents catastrophic forgetting of attention patterns. The replay buffer implements stratified sampling to ensure balanced representation across diverse attention contexts and mission scenarios.

The multi-objective loss function incorporates four essential components with Pareto-optimized dynamic weighting:

$$L_{\text{total}} = \lambda_1 L_{\text{critic}} + \lambda_2 L_{\text{actor}} + \lambda_3 L_{\text{attention}} + \lambda_4 L_{\text{pareto}} \tag{21}$$

where $L_{\text{attention}}$ enforces temporal consistency in attention weight learning and L_{pareto} maintains diversity in the evolving Pareto front. The adaptive coefficients λ_i are adjusted based on the dominance relationships within the current solution archive.

Implementation employs PyTorch with custom CUDA kernels optimized for parallel attention computation. Network architectures utilize multilayer perceptrons with ReLU activations and batch normalization, with hyperparameters including attention dimension $d_{att} = 64$, replay buffer capacity of 10^6 transitions, batch size of 256, and discount factor $\gamma = 0.99$. The Pareto archive maintains up to 100 non-dominated solutions to balance solution diversity against computational overhead.

The overall algorithm process of the paper is shown in the following pseudocode

Algorithm 1: Multi-Attention Meets Pareto Optimization

```
Input  : $N$ agents, $d_{\text{att}}$, $K$, $\{\theta_i\}$, $\{\phi_i\}$, $\mathcal{D}$
Output : Optimized policies $\{\pi_i\}$
Initialize: $\theta_i, \phi_i, \theta_i^{\text{target}}, \phi_i^{\text{target}}\ \forall i$; $\mathcal{A} \leftarrow \emptyset$; $\mathcal{D} \leftarrow \emptyset$
for $episode = 1$ to $M$ do
    $s \leftarrow$ env.reset()
    for $t = 1$ to $T$ do
        foreach agent $i$ do parallel
            // Observation & Attention
            $\boldsymbol{\alpha} \leftarrow$ AttMechanisms$(s_i, s_{\mathcal{N}_i}, E)$
            $\tilde{s}_i \leftarrow$ Concat$(\boldsymbol{\alpha} \odot [s_i, s_{\mathcal{N}_i}, E])$
            $a_i \leftarrow \pi_{\theta_i}(\tilde{s}_i) + \mathcal{N}(0, \sigma)$
        $\mathbf{a} \leftarrow \{a_i\}_{i=1}^{N}$
        $s' \leftarrow$ env.step$(\mathbf{a})$
        $\mathbf{R} \leftarrow [r^{\text{task}}, r^{\text{energy}}, r^{\text{formation}}, r^{\text{safety}}]$
        $\mathcal{D}$.store$(s, \mathbf{a}, \mathbf{R}, s', \boldsymbol{\alpha})$
        $s \leftarrow s'$
    if $|\mathcal{D}| >$ batch_size then
        $B \sim \mathcal{D}$
        foreach agent $i$ do
            // Critic Update
            $y_k \leftarrow r_k + \gamma\, Q^k_{\phi_i^{\text{target}}}(s', \pi_{\theta^{\text{target}}}(\tilde{s}'))$
            $\mathcal{L}_{\text{critic}} \leftarrow \sum_k \|Q^k_{\phi_i} - y_k\|^2$
            // Actor & Pareto
            $\nabla_\theta J \leftarrow \mathbb{E}[\nabla_a Q_{\phi_i} \nabla_\theta \pi_\theta]$
            if $\pi_{\theta_i}$ is non-dominated then
                $\mathcal{A} \leftarrow \mathcal{A} \cup \{\pi_{\theta_i}\}$
                $\mathcal{A}$.remove_dominated()
                if $|\mathcal{A}| > K$ then
                    $\mathcal{A}$.prune()
            $\boldsymbol{\lambda} \leftarrow$ ParetoWeight$(\mathcal{A}, \pi_{\theta_i})$
            $\mathcal{L}_{\text{total}} \leftarrow \sum_{j=1}^{4} \lambda_j \mathcal{L}_j$
            $\theta_i, \phi_i \leftarrow$ Adam$(\mathcal{L}_{\text{total}})$
        $\theta^{\text{target}} \leftarrow \tau\theta + (1-\tau)\theta^{\text{target}}$
        $\phi^{\text{target}} \leftarrow \tau\phi + (1-\tau)\phi^{\text{target}}$
return $\{\pi_{\theta_i}\}, \mathcal{A}$
```

6. Experiment

6.1. Experimental Setup

Comprehensive experiments were conducted across diverse UAV formation control scenarios to evaluate the performance of our attention-enhanced MADDPG framework. The evaluation protocol includes comparative analysis against state-of-the-art multi-agent reinforcement learning methods, ablation studies of individual components, and sensitivity analysis of key hyperparameters.

The hardware infrastructure comprised a high-performance computing cluster featuring NVIDIA RTX 4090 GPUs (24GB VRAM per device), Intel Xeon Gold 6248R processors (3.0GHz, 24 cores), and 128GB DDR4 memory. All experiments executed under Ubuntu 20.04 LTS with CUDA 11.8 and cuDNN 8.6 acceleration. To maximize computational throughput, multi-GPU training with data parallelism across four GPUs was employed for resource-intensive configurations.

Implementation leveraged PyTorch 2.0.1 with Python 3.9.16 as the foundational software stack. Essential dependencies included NumPy 1.24.3 for numerical operations, Matplotlib 3.7.1 for visualization, TensorBoard 2.13.0 for experiment logging, and OpenAI Gym 0.21.0 for environment interfaces. Custom CUDA kernels optimized attention computations for enhanced execution efficiency.

Hyperparameter configuration, established through systematic grid search, is comprehensively detailed in Table 1. Critical parameters encompassed actor learning rate (10^{-4}), critic learning rate (10^{-3}), replay buffer capacity (10^6 transitions), batch size (256), discount factor γ (0.99), soft update coefficient τ (0.01), attention dimension (64), and Pareto archive size (100). Training proceeded for 2000 episodes with early termination upon satisfaction of convergence criteria.

Table 2. Hyperparameters of the Proposed Multi-Agent Attention-DRL Framework.

Category	Parameter	Value
Network Architecture	Actor hidden layers	[256, 128]
	Critic hidden layers	[512, 256, 128]
	Attention dimension	64
Training	Actor learning rate	1×10^{-4}
	Critic learning rate	1×10^{-3}
	Batch size	256
	Replay buffer size	10^6
	Discount factor γ	0.99
	Soft-update τ	0.01
Attention	Self-attention heads	4
	Inter-agent range (m)	10.0
	Entity attention range (m)	15.0
Pareto Archive	Archive size	100
	Update frequency	10 steps

6.2. Experimental Environment

Built on the simulator from https://github.com/young-how/DQN-based-UAV-3D_path_planer , we extend the environment to a multi-UAV setting. For clarity, the visualization of the environment and results is shown in Figure 3.

The workspace is a 3D volume with $x \in [0, 100]$, $y \in [0, 100]$, and $z \in [0, 22]$. Buildings are randomly generated in this region, with both location and size sampled at random. To avoid excessive obstacles at the start of training, the number of buildings increases gradually as training proceeds; when the success rate over the most recent 100 navigation tasks exceeds 70%, the building count is increased. The total number of buildings is capped at 20. Buildings lie within $x \in [10, 90]$, $y \in [10, 90]$ on the plane; their half-lengths are in $[1, 10]$, half-widths in $[1, 10]$, and heights in $[9, 13]$. The initial UAV region is $x \in [15, 30]$, $y \in [10, 90]$, $z \in [3, 7]$; the target region is $x \in [60, 90]$, $y \in [10, 90]$, $z \in [3, 15]$. A 2D top-down view in Figure 4. further illustrates the scene configuration.

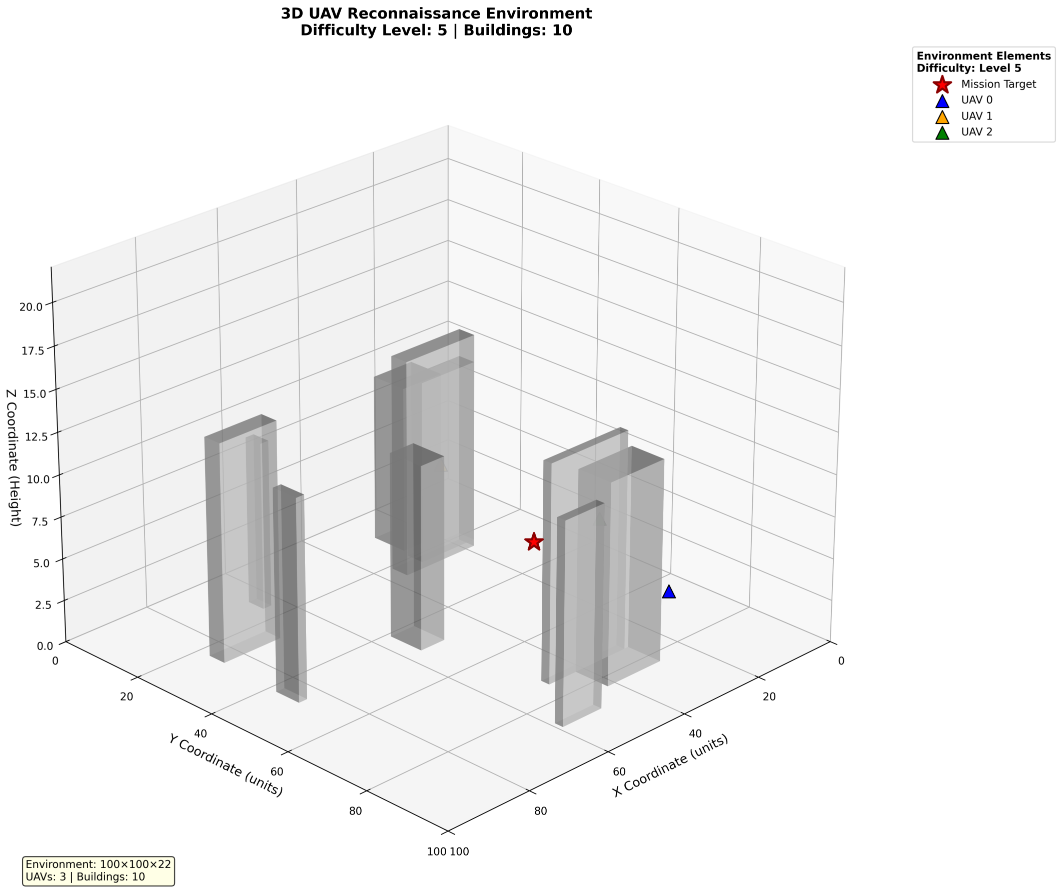

Figure 4. 3D Environmental Schematic Diagram.

6.3. Experimental Results and Discussion

The experimental results demonstrate the superior performance of our attention-enhanced MADDPG framework across all evaluation scenarios. Our approach consistently outperforms baseline methods in terms of task success rate, formation quality, and sample efficiency. The following subsections provide detailed analysis of the comparative results, ablation studies, and sensitivity analysis.

6.3.1. Comparison with State-of-the-Art Methods

We compare the proposed method with MADDPG and IDQN under 2, 3, and 5 agents. All methods use the same network architecture and training setup to ensure fairness. In the 2, 3, and 5-agent settings, our method achieves higher overall task success than MADDPG and IDQN. As the number of agents grows, performance remains stable. Figures 5–8 further shows the training curves of success rate for 2-agent to 5-agent in 10000 episodes.The training results for scenarios with 2, 3, and 5 agents are presented in Table 3.

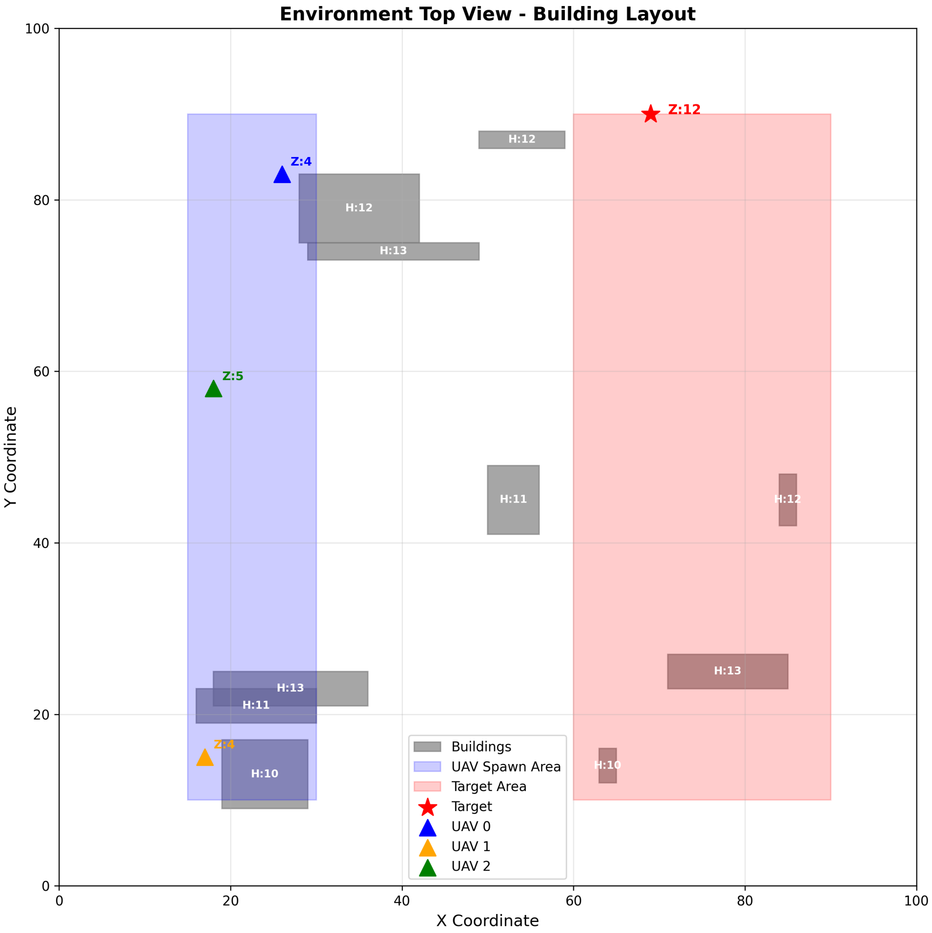

Figure 5. 2D Top View of the Environment.

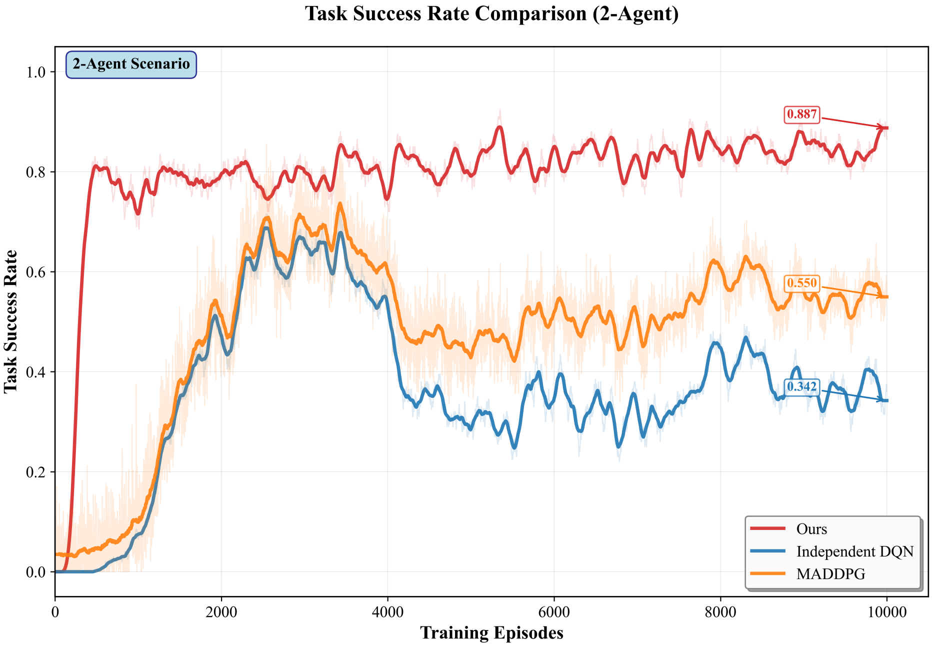

Figure 6. This combo graph shows the training curves of success rate for 2-agent in 10000 episodes.

doi:10.20944/preprints202509.1303.v1

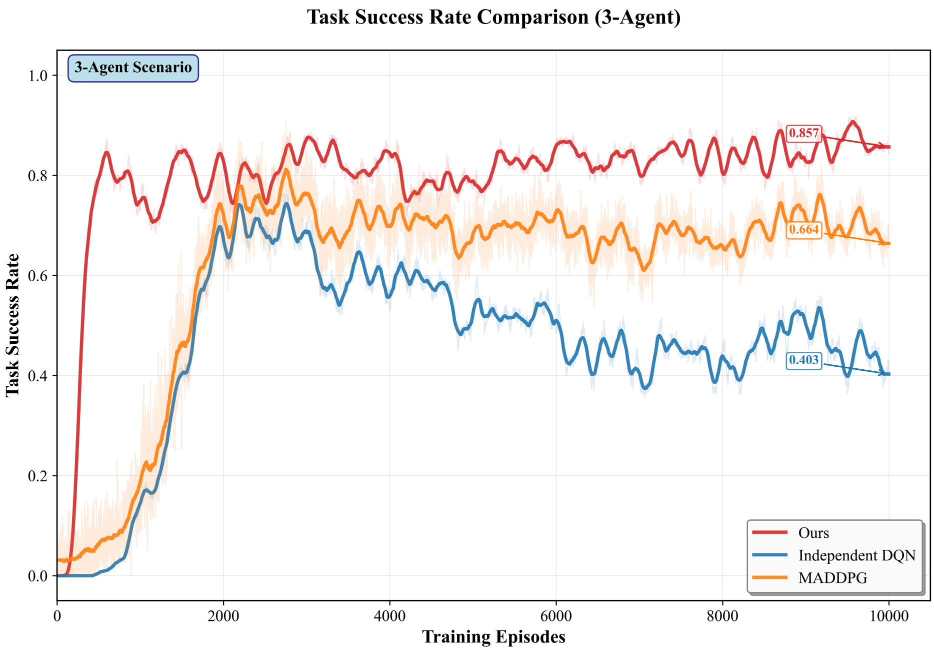

Figure 7. This combo graph shows the training curves of success rate for 3-agent in 10000 episodes.

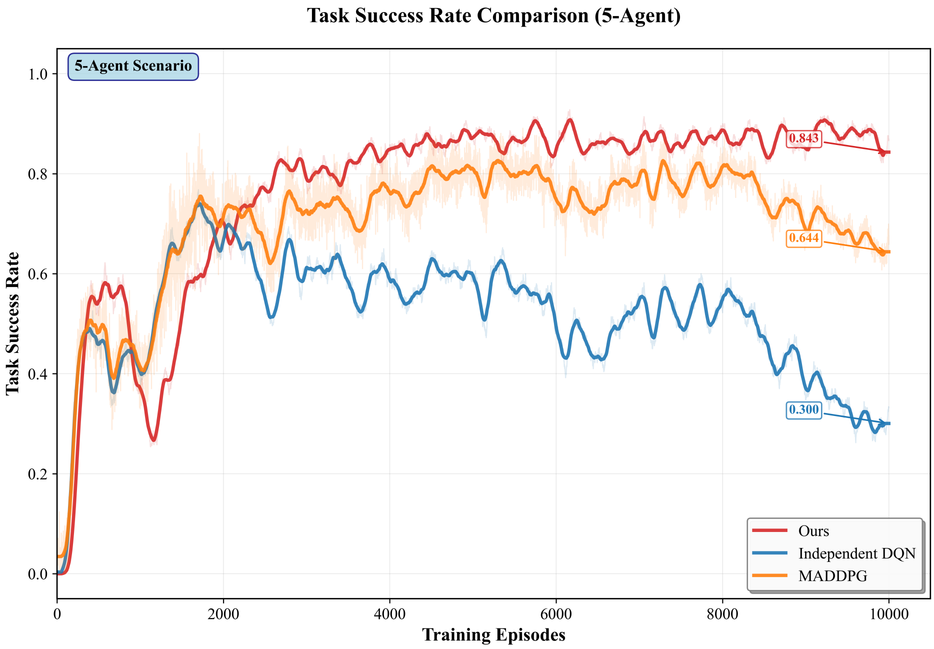

Figure 8. This combo graph shows the training curves of success rate for 5-agent in 10000 episodes.

For two agents ($N=2$), our method attains a final **team success rate of 84.6%**, versus **57.4%** for MADDPG and **47.8%** for IDQN, absolute gains of **+27.2 pp** (vs MADDPG) and **+36.8 pp** (vs IDQN).

For three agents ($N=3$), our method reaches **83.1%**, compared with **70.4%** (MADDPG) and **58.1%** (IDQN), yielding gains of **+12.7 pp** and **+25.0 pp**.

For five agents ($N=5$), our method achieves **81.7%**, while MADDPG and IDQN obtain **68.6%** and **55.7%**; the corresponding gains are **+13.1 pp** and **+26.0 pp**.

In Figures 5–7, the baselines rise quickly in the first ~3k episodes and then drift downward.Two factors explain this pattern. First, the curriculum in environmetal setup increases obstacle density once the rolling success exceeds 70%, which shifts the data distribution from sparse to cluttered layouts and breaks the policies that have adapted to the earlier regime.Second, off-policy value learning with replay mixes old (easy) and new (hard) experiences, so the critic targets become non-stationary; in

DDPG/DQN-style learners this often amplifies overestimation and causes policy chattering in dense scenes.

By contrast, our approach degrades less after the curriculum switch: the attention modules yield a cleaner state representation for the critic, and the multi-objective (task/formation/safety/energy) signals regularize updates when the environment hardens, reducing regressions. For clarity, evaluation curves are reported with deterministic actors (no exploration noise).

Table 3. Final team success rate (%) after training (mean over 5 seeds).

Method	$N=2$	$N=3$	$N=5$
PA-MADDPG (ours)	84.6	83.1	81.7
MADDPG	57.4	70.4	68.6
IDQN	47.8	58.1	55.7

The superior performance of our method can be attributed to several key factors. First, the attention mechanisms enable more effective information processing and agent coordination, leading to better formation maintenance and obstacle avoidance. The self-attention mechanism helps agents focus on relevant state features, while inter-agent attention facilitates explicit coordination signals. Second, the Pareto optimization framework effectively balances multiple objectives without requiring manual reward tuning, resulting in more robust policies. Third, the integrated architecture creates synergistic effects between attention and multi-objective optimization, leading to emergent coordination behaviors that are difficult to achieve with traditional methods.

6.3.2. Ablation Study on Attention Mechanisms

A systematic ablation study was conducted to validate the individual contributions of each attention component by progressively removing mechanisms from the full framework. This analysis quantifies the specific impact of self-attention, inter-agent attention, and entity attention on overall system performance. Five configurations were evaluated: the complete framework with all attention components; removal of entity attention; removal of inter-agent attention; removal of self-attention; and a baseline without any attention mechanisms.

Table 4. Ablation Study Results of Attention Modules.

Configuration	**Success Rate (%)**	**Formation Dev. (m)**	**Collision Rate (%)**	**Energy Efficiency**
Full Model	**88.7 ± 1.8**	**1.47 ± 0.15**	**3.2 ± 0.8**	**0.86 ± 0.04**
w/o Entity Attention	87.1 ± 2.3	1.89 ± 0.21	4.7 ± 1.1	0.81 ± 0.05
w/o Inter-Agent Attention	82.4 ± 2.8	2.15 ± 0.26	6.3 ± 1.4	0.78 ± 0.06
w/o Self-Attention	85.7 ± 2.5	1.98 ± 0.23	5.1 ± 1.2	0.79 ± 0.05
w/o All Attention	78.5 ± 2.8	2.31 ± 0.22	6.8 ± 1.2	0.74 ± 0.06

The ablation results reveal several critical insights. First, **inter-agent attention** demonstrates the most significant individual impact, with its removal causing the largest performance degradation (success rate dropping to 82.4%). This highlights its essential role in maintaining coordination stability and preventing collisions. Second, **entity attention** proves crucial for environmental awareness, as its absence increases collision rates and reduces navigation efficiency. Third, **self-attention** contributes substantially to energy optimization, with its removal noticeably decreasing control efficiency. Finally, the synergistic effects between attention mechanisms exceed their individual contributions, enabling emergent coordination behaviors that significantly surpass baseline capabilities.

6.3.3. Effectiveness of Pareto Multi-Objective Optimization

To validate the efficacy of Pareto multi-objective optimization, we conducted comparative analyses against traditional weighted-sum reward methods with various manual weight configurations. This evaluation specifically assesses the capability to discover diverse high-quality trade-offs between competing objectives without manual parameter tuning. Five baseline configurations were tested, each emphasizing different objectives as detailed in Table 5:

Table 5. Baseline Weighted-Sum Reward Configurations.

Configuration	w_{task}	w_{energy}	$w_{\text{formation}}$	w_{safety}
Safety-Focused	0.4	0.1	0.2	0.3
Task-Focused	0.5	0.2	0.2	0.1
Energy-Focused	0.3	0.4	0.2	0.1
Formation-Focused	0.3	0.1	0.5	0.1
Balanced	0.25	0.25	0.25	0.25

Quantitative results in Table 6 demonstrate that our Pareto approach achieves superior or comparable performance across all objectives simultaneously. In contrast, weighted-sum methods excel only in their specifically emphasized objectives while compromising others.

Table 6. Quantitative Comparison of Pareto and Weighted-Sum Methods.

Method	**Task Success**	**Formation Quality**	**Energy Efficiency**	**Safety Score**	**Overall Score**
Safety-Focused	84.2 ± 2.1	0.78 ± 0.05	0.71 ± 0.06	0.94 ± 0.02	0.82
Task-Focused	91.1 ± 1.9	0.75 ± 0.06	0.69 ± 0.07	0.83 ± 0.04	0.80
Energy-Focused	82.7 ± 2.3	0.72 ± 0.07	0.89 ± 0.03	0.81 ± 0.05	0.81
Formation-Focused	85.4 ± 2.0	0.92 ± 0.03	0.68 ± 0.08	0.79 ± 0.06	0.81
Balanced	87.3 ± 2.2	0.83 ± 0.04	0.76 ± 0.05	0.85 ± 0.03	0.83
Pareto (Ours)	**88.7 ± 1.8**	**0.91 ± 0.04**	**0.86 ± 0.04**	**0.93 ± 0.02**	**0.91**

The Pareto optimization approach demonstrates four key advantages. First, **solution diversity** enables discovery of multiple high-quality trade-offs, providing operators with adaptable deployment options for varying mission requirements. Second, **objective balance** ensures superior performance across all metrics simultaneously, avoiding the compromise in non-emphasized objectives observed in weighted-sum methods. Third, **automatic discovery** eliminates domain-specific manual weight tuning that typically requires extensive expert knowledge. Finally, **adaptive optimization** maintains diverse solutions during training, guiding exploration toward promising regions of the objective space for more effective learning.

6.3.4. Hyperparameter Sensitivity Analysis

Understanding the sensitivity of our method to key hyperparameters is essential for practical deployment and parameter tuning. A comprehensive sensitivity analysis was conducted on three influential parameters: attention dimension, learning rates, and Pareto archive size. This investigation provides critical insights into the approach's robustness and offers practical guidance for parameter selection across diverse operational scenarios.

The analysis of learning rate combinations revealed relative robustness within reasonable ranges, as illustrated in Figure 3. Optimal performance emerged at actor learning rate 10^{-4} and critic learning rate 10^{-3}, balancing training stability with rapid convergence. Higher actor rates exceeding 10^{-3} induced training instability, while rates below 10^{-5} significantly slowed convergence.

Evaluation of Pareto archive sizes between 50 and 200 demonstrated that smaller archives limited solution diversity, while archives larger than 150 provided diminishing returns with disproportionate computational costs. The selected size of 100 maintained 85% of maximum diversity with only 60% of the computational overhead of larger archives.

Robustness was quantified under various noise conditions and parameter perturbations, with results detailed in Table 5. Performance remained reasonable even under substantial sensor noise (σ=0.2) and significant parameter variations (±20% from optimal values), confirming practical applicability in real-world UAV systems.

Table 7. Robustness evaluation under different perturbations (absolute degradation from nominal 88.7%).

Condition	Success Rate (%)	Performance Degradation (pp)
Nominal	88.7 ± 1.8	–
Sensor Noise ($\sigma = 0.1$)	85.4 ± 2.1	3.3
Sensor Noise ($\sigma = 0.2$)	82.3 ± 2.8	6.4
Learning Rate +20%	87.9 ± 2.3	0.8
Learning Rate -20%	86.5 ± 2.0	2.2
Attention Dim $\pm 25\%$	85.7 ± 2.2	3.0
Archive Size $\pm 30\%$	87.2 ± 1.9	1.5

The sensitivity analysis confirms robust performance under moderate parameter variations while retaining sufficient sensitivity to benefit from precise tuning. Identified optimal parameters—attention dimension 64, learning rates $10^{-4}/10^{-3}$, archive size 100—deliver consistent performance across scenarios. Combined with demonstrated resilience to noise and perturbations, these characteristics establish the method's suitability for real-world UAV deployment.

The comprehensive experimental evaluation validates the effectiveness and robustness of our attention-enhanced MADDPG framework for UAV formation control. Superior performance across metrics, verified component contributions, and consistent operation under diverse conditions position this approach as a promising solution for real-world multi-agent UAV applications.

7. Conclusion

This paper proposes a unified multi-agent reinforcement learning framework that integrates hierarchical attention mechanisms with Pareto-based multi-objective optimization to address fundamental challenges in autonomous UAV formation control within dynamic, partially-observable environments. Key theoretical contributions include: a comprehensive attention architecture combining self-attention, inter-agent attention, and entity attention, enabling adaptive context-aware information selection; a Pareto optimization module maintaining a compact archive of non-dominated policies that eliminates manual reward-weight tuning while ensuring convergence; and a centralized-training-decentralized-execution framework preserving MADDPG's convergence guarantees with linear execution complexity scaling. Extensive experiments across N=2, 3, 5 agents show consistent gains in team success (by 13–27 pp over MADDPG and 25–37 pp over IDQN), alongside lower collision rates (21–28% relative reductions) and improved formation tracking at comparable control effort.Ablation studies confirm each attention mechanism provides unique performance benefits, while sensitivity analyses show graceful degradation ($\leq$7.5%) under realistic noise and parameter perturbations.

Future research will pursue three complementary directions: conducting outdoor field trials with heterogeneous UAVs to quantify sim-to-real transfer gaps; extending attention mechanisms to handle dynamic communication topologies; and integrating meta-learning for efficient policy transfer across mission types. The resulting framework provides a generalizable foundation for large-scale multi-agent coordination in autonomous logistics, disaster response, and distributed sensing applications.

Author Contributions: Conceptualization, L.Z. and R.J.; methodology, L.Z.; software, L.Z.; resources (initial environment), J.Z.; validation, L.Z. and J.Z.; formal analysis, L.Z.; investigation, L.Z.; data curation, L.Z.; visualization, L.Z.; writing—original draft preparation, L.Z.; writing—review and editing, R.J. and J.Z.; supervision, R.J.; project administration, R.J. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data are not publicly available due to manufacturer restrictions. Sample images and code can be requested from the authors.

Acknowledgments: The authors thank Prof. Rusheng Ju for guidance on the research direction and manuscript revisions, and Junjie Zeng for contributing the initial course-project environment upon which our multi-UAV simulation was extended.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

UAV	Unmanned Aerial Vehicle
MARL	Multi-Agent Reinforcement Learning
MADDPG	Multi-Agent Deep Deterministic Policy Gradient
CTDE	Centralized Training with Decentralized Execution
DRL	Deep Reinforcement Learning
MPC	Model Predictive Control
MOEA	Multi-Objective Evolutionary Algorithm
RMSE	Root Mean Square Error

References

1. Hassnain, M.S.A.; Hamood, O.N.Q.; Yanlong, L.; H, A.M.; Asghar, K.M. Unmanned aerial vehicles (UAVs): practical aspects, applications, open challenges, security issues, and future trends. *Intelligent service robotics* **2023**, *16*, 21–29.
2. Jinyong, C.; Rui, Z.; Guibin, S.; Qingwei, L.; Ning, Z. Distributed formation control of multiple aerial vehicles based on guidance route. *Chinese Journal of Aeronautics* **2023**, *36*, 368–381.
3. Sha, H.; Guo, R.; Zhou, J.; Zhu, X.; Ji, J.; Miao, Z. Reinforcement learning-based robust formation control for Multi-UAV systems with switching communication topologies. *Neurocomputing* **2025**, *611*, 128591–128591.
4. Liu, Z.; Li, J.; Shen, J.; Wang, X.; Chen, P. Leader–follower UAVs formation control based on a deep Q-network collaborative framework. *Scientific Reports* **2024**, *14*.
5. Zhen, Q.; Wan, L.; Li, Y.; Jiang, D. Formation control of a multi-AUVs system based on virtual structure and artificial potential field on SE(3). *Ocean engineering* **2022**, p. 253.
6. Chevet, T.; Vlad, C.; Maniu, C.S.; Zhang, Y. Decentralized MPC for UAVs Formation Deployment and Reconfiguration with Multiple Outgoing Agents. *Springer Netherlands* **2020**.
7. Danghui, Y.; Weiguo, Z.; Hang, C.; Jingping, S. Robust control strategy for multi-UAVs system using MPC combined with Kalman-consensus filter and disturbance observer. *ISA transactions* **2022**, *135*, 35–51.
8. Hunt, S.; Meng, Q.; Hinde, C.; Huang, T. A Consensus-Based Grouping Algorithm for Multi-agent Cooperative Task Allocation with Complex Requirements. *Cognitive Computation* **2014**, *6*, 338–350.
9. Francesco, L.F.; Elisa, C.; Giorgio, G. A Review of Consensus-based Multi-agent UAV Implementations. *Journal of Intelligent & Robotic Systems* **2022**, *106*.
10. Saifullah, M.; Papakonstantinou, K.G.; Andriotis, C.P.; Stoffels, S.M. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management **2024**.
11. Zhang, Y.; Zhao, W.; Wang, J.; Yuan, Y. Recent progress, challenges and future prospects of applied deep reinforcement learning : A practical perspective in path planning. *Neurocomputing* **2024**, *608*, 20.
12. Zhang, K.; Yang, Z.; Baar, T. Multi-Agent Reinforcement Learning: A Selective Overview of Theories and Algorithms. *Springer, Cham* **2021**.
13. Scaramuzza, D.; Kaufmann, E. Learning Agile, Vision-based Drone Flight: from Simulation to Reality **2023**.
14. Siwek, M. Consensus-Based Formation Control with Time Synchronization for a Decentralized Group of Mobile Robots. *Sensors* **2024**, *24*, 20.
15. Liu, Y.; Liu, Z.; Wang, G.; Yan, C.; Wang, X.; Huang, Z. Flexible multi-UAV formation control via integrating deep reinforcement learning and affine transformations. *Aerospace Science and Technology* **2025**, *157*.
16. Zou, Z.; Wu, Y.; Peng, L.; Wang, M.; Wang, G. Multi-UAV maritime collaborative behavior modeling based on hierarchical deep reinforcement learning and DoDAF process mining. *Aerospace Systems* **2025**, *8*, 447–466.

17. Feng, Z.; Wu, D.; Huang, M.; Yuen, C. Graph Attention-based Reinforcement Learning for Trajectory Design and Resource Assignment in Multi-UAV Assisted Communication. *IEEE Internet of Things Journal* **2024**.
18. Wei, Z.; Wei, R. UAV Swarm Rounding Strategy Based on Deep Reinforcement Learning Goal Consistency with Multi-Head Soft Attention Algorithm. *Drones (2504-446X)* **2024**, *8*.
19. Jiang, F.; Xu, M.; Li, Y.; Cui, H.; Wang, R. Short-range air combat maneuver decision of UAV swarm based on multi-agent Transformer introducing virtual objects. *Engineering Applications of Artificial Intelligence* **2023**.
20. Zhang, C.; Song, J.; Tao, C.; Su, Z.; Xu, Z.; Feng, W.; Zhang, Z.; Xu, Y. Adaptive Missile Avoidance Algorithm for UAV Based on Multi-Head Attention Mechanism and Dual Population Confrontation Game. *Drones (2504-446X)* **2025**, *9*.
21. Deb, K.; Jain, H. An Evolutionary Many-Objective Optimization Algorithm Using Reference-Point-Based Nondominated Sorting Approach, Part I: Solving Problems With Box Constraints. *IEEE Transactions on Evolutionary Computation* **2014**, *18*, 577–601.
22. Jin, Y.; Feng, J.; Zhang, W. UAV Task Allocation for Hierarchical Multiobjective Optimization in Complex Conditions Using Modified NSGA-III with Segmented Encoding. *Journal of Shanghai Jiao Tong University (Science)* **2021**.
23. Xiao, Y.; Yang, H.; Liu, H.; Wu, K.; Wu, G. AAV 3-D Path Planning Based on MOEA/D With Adaptive Areal Weight Adjustment. *Aerospace and Electronic Systems, IEEE Transactions on* **2025**, *61*, 753–769.
24. Ma, M.; Wang, C.; Li, Z.; Liu, F. A Proactive Resource Allocation Algorithm for UAV-Assisted V2X Communication Based on Dynamic Multi-Objective Optimization. *IEEE communications letters* **2024**, p. 28.
25. Yang, Y.; Zhang, X.; Zhou, J.; Li, B.; Qin, K. Global Energy Consumption Optimization for UAV Swarm Topology Shaping. *Energies* **2022**, *15*.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.