ORIGINAL RESEARCH **Open access**

Pareto-Guided Multi-Agent Reinforcement Learning for UAV Formation in Urban Environments

James Mwangi[1*], Grace Wanjiru[1], Samuel Kiprop[2], Cynthia Atieno[3], Peter Odhiambo[4]

Abstract

Achieving autonomous formation control for multiple UAVs in dense urban landscapes poses persistent challenges stemming from limited visibility, crowded, shifting obstacles, and competing demands (operational efficiency, energy expenditure, and safety). A significant number of MARL-based techniques continue to collapse multi-dimensional reward signals into a single, manually calibrated scalar and lack selective information integration, leading to fragile compromises and weak generalization in city-like clutter. This work presents a model-agnostic MARL framework—concretely realized with MADDPG—that supplements a CTDE backbone with three streamlined attention modules (self-focused, inter-agent, and entity-based) serving as selective fusion mechanisms, together with a Pareto optimization component that curates a compact repository of non-dominated policies to dynamically steer trade-offs among objectives via straightforward, interpretable reward components rather than delicate weighting schemes. Tested in city-scale navigation scenarios, the approach elevates final team success by 13–27 percentage points for N = 2–5 while concurrently reducing collision counts, improving formation precision, and decreasing control expenditure. These benefits demand no method-specific calibration and hold steady across all evaluated team sizes (N = 2–5), underlining a superior safety–efficiency balance and dependable performance in obscured, cluttered conditions.

Keywords Attention mechanisms, Pareto optimization, Multi-agent reinforcement learning, UAV formation control

*Correspondence:
James Mwangi
james.mwangi@uonbi.ac.ke
[1] Department of AI for Maize and Sorghum Pest Outbreak Prediction, Faculty of Agriculture, University of Nairobi, Nairobi, Kenya
[2] Department of Machine Learning for Rain-fed Agriculture Risk Assessment, Faculty of Agriculture, Moi University, Eldoret, Kenya
[3] Department of Farm Data Fusion and Decision Support Systems, Faculty of Agriculture, Kenyatta University, Nairobi, Kenya
[4] Department of Smart Soil Salinity Monitoring, Faculty of Agriculture, Egerton University, Njoro, Kenya

Introduction

With UAVs rapidly entering operational practice, their applications now extend well beyond agriculture and freight delivery into time-pressured disaster relief and sustained environmental observation. Dense urban and built-up zones—marked by tall structures, frequent occlusions, and constricted passageways—constitute increasingly vital deployment theaters for UAV collectives, including communication bridging inside “urban canyons,” collaborative reconnaissance around skyscrapers, and secure traversal amid congested streets and enclosed courtyards [1-3]. While multi-UAV formations can markedly boost spatial coverage and system resilience, delivering stable, flexible, and collision-free coordination in such city-mimicking scenarios remains a formidable challenge.

With growing team sizes, multi-UAV control grapples with coupled perception, choice-making, and runtime constraints. We pinpoint three operational bottlenecks. (1) Discriminating information uptake amid partial observability. Individual UAVs must triage high-volume, heterogeneous input streams (proprioceptive data, neighboring agents, and

This is an open access article distributed under the terms of the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License (CC BY-NC-SA 4.0), which permits non-commercial use, distribution, reproduction, and adaptation in any medium or format, provided that appropriate credit is given to the original author(s) and source, a link to the license is provided, and any changes made are indicated. Adapted material must be distributed under the same license.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

environmental features) to isolate decision-critical signals; without targeted filtering, response delays mount, and credit attribution degrades. Established rule-driven and control-theoretic paradigms—such as leader–follower and virtual structure/potential field constructs [4, 5], model predictive control (MPC) [6, 7], and consensus-based/distributed optimization [8, 9]—furnish lucidity and formal assurances but prove susceptible to modeling imprecision, communication latency, and the fluid clutter characteristic of metropolitan settings. (2) Balancing competing objectives. Formation maintenance, collision prevention, mission throughput, and energy thrift routinely stand in opposition; collapsing these into rigid linear combinations conceals the underlying Pareto geometry and yields inflexible policies when operational priorities evolve. (3) Scaling and resilience. Under real-world impairments—packet dropout, sensor disturbance, environmental non-stationarity—effectiveness erodes; off-the-shelf deep MARL formulations (exemplified by MADDPG-class CTDE) boost multi-agent harmonization [10-12] yet omit a dedicated apparatus for dynamic information triage and rigorous multi-criteria optimization, triggering steep performance decay as agent counts or scene intricacy rise [13].

To address these gaps, we devise a reinforcement learning scheme that enriches the Centralized Training with Decentralized Execution (CTDE) substrate with multi-source attention and a Pareto-driven optimization layer. In particular, we arm distributed actors with three lean attention pathways—self-directed attention for intra-state feature discrimination, inter-agent attention for selective teammate reasoning, and entity-focused attention for context-salient environmental parsing—whose results are fused into a joint attention-informed representation. Throughout training, a vector-valued reward tracks task advancement, energy usage, formation fidelity, and safety margins; a Pareto module curates a concise set of non-dominated trade-off solutions and yields adaptive weighting signals for policy updates, circumventing arduous manual reward engineering while preserving simple, transparent shaping components per objective. Within a representative 3D urban-analog setting, the proposed enhancements uniformly improve team success, safety metrics, and formation precision across 2–5 UAVs, while maintaining or reducing control effort; comprehensive measurements are presented in Section 6.

Vis-à-vis antecedent MARL designs for UAV formation, our contribution bifurcates into: (i) a category-specific, single-hop multi-source attention interface (self/inter-agent/entity) operating as a lightweight triage mechanism instead of a deep message-propagation encoder, thereby safeguarding real-time responsiveness at $O(K+M)$ per-agent per-step computation; and (ii) a Pareto module schooled via vector-valued critics that substitutes delicate hand-engineered scalarization, laying bare robust compromise surfaces spanning task progression, safety, formation tightness, and energy expenditure. This pairing produces pronounced gains in cluttered, partially obscured urban contexts and slots into CTDE backbones as a plug-and-play augmentation.

The principal contributions are itemized below:

· We construct a multi-source attention architecture (self/inter-agent/entity) for decentralized actors that discriminatively amalgamates pivotal cues from the self, collaborators, and the urban setting, thereby improving coordination throughput and dependability under incomplete observability.

· We devise a Pareto optimization module for vector-based rewards that approximates the Pareto frontier throughout training, facilitating adaptive compromise among task throughput, formation cohesion, energy consumption, and safety using only simple, per-objective shaping terms—eliminating the labor-intensive manual balancing.

· We embed the foregoing as architecture-agnostic, insertable modules for CTDE-type MARL and corroborate them in 3D city-emulating environments. Across formations of N = 2–5 UAVs, incorporating our modules into canonical MARL substrates raises final team success by roughly +12–+27 percentage points and lowers collision incidence by about 20–30%, coupled with more faithful formation tracking at similar or reduced control cost; the improvements endure for N = 2–5, demonstrating consistent benefit across tested scales; computational scaling is examined in Section 6.9.

Apart from these core challenges, two complementary investigative threads merit mention, though they lie adjacent to our focal scope: (i) real-time communication frameworks with blockchain-anchored identity verification and auditable trails for

© 2023 The Author(s). | Published by Agro Intelligence Research Press

mission-grade coordination, and (ii) AI-infused swarm intelligence/control architectures oriented toward higher-order autonomy and deliberative decision-making. Contemporary scholarship documents headway on both dimensions [14, 15]; accordingly, we situate our contribution within the low-latency control stratum and subsequently elaborate on how these avenues may synergize with our approach.

Notational conventions used herein are: Unmanned Aerial Vehicle (UAV); Multi-agent reinforcement learning (MARL); Centralized Training with Decentralized Execution (CTDE); multi-objective optimization (MOO); Root Mean Square Error (RMSE); Multi-Agent Deep Deterministic Policy Gradient (MADDPG) [16]; (Independent) Deep Q-Network (IDQN) [17].

The structure of the remainder runs as follows: Section 2 surveys related literature. Section 3 recaps foundational concepts. Section 4 formalizes the problem and CTDE instantiation. Section 5 details our method (attention branches and Pareto optimization). Section 6 reports experiments, ablation studies, and robustness evaluations. Section 7 offers conclusions.

Related Work

Current research on Multi-UAV formation control

Field deployments are increasingly migrating into urban/built-up settings replete with tightly packed buildings, occluded sightlines, and slender corridors, within which multi-UAV formations underpin communication relay chains, joint search operations, and hazard-free transit among high-rise edifices. Rule-based and classical regulatory schemes—notably leader–follower architectures and virtual structure/potential field paradigms—retain wide currency thanks to their conceptual straightforwardness and operational simplicity [4, 5]. Optimization-rooted strategies, encompassing model predictive control (MPC) and consensus-flavored and distributed regulation, provide stronger constraint enforcement and theoretical stability guarantees [6, 7, 9, 18]. Deep reinforcement learning (DRL) has recently emerged as a compelling approach for handling rich observation spaces and incomplete observability, with exploratory applications in formation maintenance, trajectory crafting, and cooperative guidance [11, 19, 20].

Transplanting these paradigms into city-analogous domains, however, reveals three persistent impediments that mirror our problem framing. (i) Selective signal extraction under partial observability: control policies must pluck decision-salient features from self-referenced, neighbor-referenced, and environment-referenced data streams within voluminous, visually congested scenes; rigid-rule filters or manually engineered interfaces buckle as environmental complexity escalates. (ii) Multi-criteria tension: formation tightness, obstacle avoidance/safety margins, operational tempo, and energetic frugality routinely compete; flattening them via static linear coefficients erases the Pareto landscape and incubates brittle responses when mission imperatives shift. (iii) Scale tolerance and dependability: communication lag, dropped packets, and transduction noise corrode multi-agent coherence, and effectiveness typically plummets when agent population or spatial density mounts [11]. Even within optimization/control formalisms augmented by disturbance observers and estimation filters (e.g., Kalman-consensus loops and disturbance monitors within MPC workflows [7]), the computational burden of real-time optimization, coupled with model disparity across tangled 3D geometry, curtails nimbleness. Across DRL workflows, the absence of purpose-built provisions for dynamic information discrimination and methodical multi-objective resolution constitutes an unmet need.

Attention mechanisms for collaborative perception and decision making

The incorporation of attention into multi-agent systems has sharpened perception and inter-agent communication, directing computational resources toward the most informative elements in cooperative UAV missions. In the domains of trajectory generation and resource allocation, graph-based attention mechanisms have yielded measurable gains by enabling individual agents to weight the importance of specific neighbors and edges [21]. For tasks involving cooperative enclosure or herding behaviors, soft attention operating across multiple heads produces targeted coordination cues [22]. Architectures inspired by Transformer models and augmented with virtual entities have been applied to short-range air

© 2023 The Author(s). | Published by Agro Intelligence Research Press

combat maneuver selection, where structured attention directed at key environmental and adversarial entities demonstrably improves decision quality [23]. When operating under hostile or high-risk conditions—missile evasion being a representative case—multi-head attention proves useful for registering dynamic obstacles and extracting threat-relevant saliency patterns [24].

Taken as a body of evidence, these works point toward the capacity of multi-source attention (spanning self, neighbor, and entity streams) to strengthen collaborative perception and the caliber of resulting decisions. That said, the bulk of earlier research optimizes against a single scalarized performance metric and refrains from tightly integrating attention with estimation of multi-dimensional value; the consequence is that learned behaviors can become excessively tailored to one particular weighting configuration and transfer poorly when the relative importance among objectives changes (for example, when moving from assertive goal pursuit to safety-dominant traversal within constricted passages). Our contribution directly confronts this lacuna by marrying compact attention pathways with vector-valued critics and Pareto-conditioned training.

Advances in Multi-Objective Optimization (MOO) and pareto methods

Approaches to multi-objective optimization rooted in Pareto principles furnish a theoretically sound apparatus for surfacing the trade-offs among conflicting desiderata, without resorting to monolithic weighted aggregation. In UAV-oriented scholarship, NSGA-III along with its derivatives has been brought to bear on task distribution and strategic planning under elaborate constraint sets [25, 26]; the MOEA/D framework employing adaptive weight adjustment has been shown to enhance the evenness of solution coverage and has found application in three-dimensional route planning [27]; additionally, dynamic multi-objective resource apportionment has received attention in adjacent communication and power-management domains [28, 29]. While these methodologies prove potent for offline design workflows and static scenario snapshots, their computational expense and loose coupling with the online perception–decision loop obstruct end-to-end, real-world deployment in cluttered, partially occluded operational theaters.

What follows is a recognized need to weave Pareto-style reasoning into the fabric of learning-driven coordination, rather than treating MOO as a disconnected, post hoc batch processor. Our work advances along precisely this vector: we retain per-objective shaping components that remain individually straightforward (task advancement, energy outlay, formation tightness, safety margins) and train critics that output vectors rather than scalars, while a compactly maintained Pareto archive contributes adaptive weighting signals during training, thereby promoting non-dominated policy refinements from within the ongoing MARL optimization loop.

Real-time communication and secure coordination (Blockchain-Enabled)

For formations of multiple UAVs, real-time communication must simultaneously satisfy demands for authentication, message integrity, and the ability to audit interactions—all within dynamic network topologies that may be intermittently connected. Protocols grounded in blockchain technology (distributed ledger technology, DLT) deliver tamper-evident record-keeping, decentralized identity management, and consensus at the mission level, capabilities that can facilitate task handoff, forensic investigation of faults, and higher-order orchestration when the supporting infrastructure cannot be fully trusted [14]. That being acknowledged, classical Byzantine Fault Tolerant-style consensus mechanisms and the attendant duplication of ledger state incur significant latency and bandwidth penalties, positioning them more naturally at the granularity of event-driven control functions (role designation, cryptographic key rotation, mission-level replanning) rather than within the tight per-timestep actuation cycle. Our method occupies a complementary niche: the attention and Pareto modules execute within the latency-sensitive control stratum, whereas DLT-backed services can be layered above to furnish authenticated messaging and mission governance—potentially informed by attention-derived saliency estimates that indicate which links or occurrences are sufficiently critical to warrant immutable recording.

AI-driven swarm intelligence and AI-Based control

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Looking beyond the boundaries of MARL proper, an extensive literature on AI-driven swarm intelligence and AI-centric control investigates paradigms drawn from biological coordination, MPC enriched with learning components, and layered decision hierarchies, all aimed at boosting autonomy, behavioral flexibility, and the quality of decentralized decision-making [15]. These frameworks typically concern themselves with behaviors at higher levels of abstraction—parceling out tasks, reconfiguring formations, negotiating roles—and serve as complements to low-level functions such as obstacle avoidance and formation-keeping. Our Pareto-attentive CTDE formulation is designed to slot in as the reactive execution layer within such hierarchical constructs, receiving goal designations or safety envelopes from higher-tier swarm planners; in the reverse direction, vector-valued critics are positioned to supply interpretable, per-objective signals upward whenever mission priorities undergo adjustment.

Summary and gaps

Bringing the foregoing strands together, the landscape can be characterized as follows: (i) classical rule-based and optimization-driven controllers deliver robust performance when system models and communication links are dependable, yet they encounter difficulty in isolating salient information and flexibly adapting within tangled urban geometries; (ii) deep reinforcement learning accommodates voluminous, high-dimensional observation streams but habitually leans on improvised scalarization techniques, thereby missing a principled mechanism for reconciling objectives that stand in opposition; and (iii) existing deployments of attention improve aspects of perception and coordination but are generally tuned under a single composite objective, which leaves them brittle when the relative emphasis among priorities changes. The present paper works to close these identified gaps through the synthesis of multi-source attention (self/inter-agent/entity) for targeted information fusion and a Pareto module that handles vector-valued rewards, the whole integrated within a CTDE-aligned methodology so that non-dominated trade-off surfaces are uncovered during the training process rather than being preordained through static weight assignments.

Background Knowledge

Multi-agent reinforcement learning

Multi-agent reinforcement learning (MARL) is a foundational formalism for reasoning about collectives of agents that pursue either shared or competing objectives through sequences of interdependent decisions within a common world model. Relative to the single-agent formulation, MARL introduces a set of core complications: the environment is rendered non-stationary from the perspective of any one learner, responsibility for observed outcomes must be apportioned across agents (the credit assignment problem), and the policies being learned exert mutual influence on one another’s update trajectories. Within the specific context of UAV formation flight, these complications manifest with particular intensity. Each airborne platform is constrained to act on locally available sensory information, yet the aggregate effect of their combined actions is what ultimately governs system-wide measures of success.

A variety of MARL organizational schemes have been advanced to meet these challenges. Among them, Centralized Training with Decentralized Execution (CTDE) has emerged as the dominant template. The central insight is to exploit globally scoped information during the training phase—when the luxury of off-board computation and unrestricted data sharing can be assumed—to cultivate cooperative policies. At the same time, each agent is restricted to its own partial view of the world at execution time. This division achieves a workable compromise between the expressive power needed for coordinated behavior and the scalability and communication constraints of real deployment. A widely recognized concrete algorithm in this class is MADDPG [16], which couples a centralized critic (permitted access to all agents’ observations and actions during training) with decentralized actors that map local observations to actions, thereby attenuating, though not eliminating, the effects of environmental non-stationarity.

Yet traditional MARL formulations continue to exhibit significant shortcomings. Agents operating within these frameworks are not typically endowed with mechanisms for selectively attending to the multiple distinct streams of state information available to them. Consequently, they cannot zero in on the most decision-relevant features when embedded in complex,

© 2023 The Author(s). | Published by Agro Intelligence Research Press

high-clutter environments. Furthermore, the preponderance of existing methods collapses the evaluation of behavior into a scalar reward signal produced by linearly combining numerous objectives. This practice is fundamentally incapable of representing the intricate compromise structures that exist among those objectives. These dual limitations provide the direct impetus for our introduction of attention mechanisms and multi-objective optimization as enhancements to MARL tailored for UAV formation control.

Attention mechanisms

The core function of attention lies in weighted information aggregation: the mechanism amplifies input components deemed most relevant while attenuating extraneous signals. This property directly benefits long-range dependency modeling and the prioritization of salient features. Although its transformative impact originated in natural language processing and computer vision, attention is now being deployed with increasing regularity across multi-agent systems (MAS)—cooperative UAV control included—to filter an agent's own state variables, signals received from neighboring teammates, and descriptors of environmental entities under conditions marked by limited observability and unreliable communication channels.

Our foundational operator is the standard scaled dot-product attention, formulated as:

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

In this expression, queries Q represent the agent's current informational demand, keys K serve as an index over the pool of candidate features, and values V contain the actual content to be combined. Applying softmax to the relevance scores yields a normalized weighting distribution, and the weighted sum produces the output context vector. When multiple such attention heads are executed concurrently, and their results are concatenated, the model gains access to several distinct representational subspaces.

For the specific case of UAV formation control, attention confers particular advantages in the following respects:

· Selective perception: drawing focus toward the portions of the local observation that carry decision-relevant content (e.g., the direction toward the goal, onboard energy status, proximity-based safety margins).

· Targeted coordination: zeroing in on those neighbors whose relative states most strongly influence collision avoidance and formation shape maintenance.

· Salient environment awareness: elevating the prominence of proximate obstacles or spatial constrictions encountered in dense urban-like terrain.

These attributes position attention as an especially well-suited complement to CTDE-structured MARL: on the actor side, it enhances the quality of input representations despite partial observability, while on the critic side, it helps mitigate the non-stationarity that centralized value estimators must contend with. Sections 4 and 5 provide the concrete realization of this design philosophy through three specialized attention branches—self, inter-agent, and entity—tailored to UAV team operations.

As illustrated schematically in **Figure 1**, the scaled dot-product attention mechanism calculates relevance through , converts these scores into a distribution via softmax, and uses the resulting weights to aggregate values V. In our instantiation, the queries encode whatever information the agent currently needs to extract. In contrast, keys and values are drawn from one of three sources: the agent's own feature set, the features of neighboring agents, or the features of

© 2023 The Author(s). | Published by Agro Intelligence Research Press

environmental entities. **Figure 1**, therefore, represents the generic computational primitive from which the three typed attention branches for UAV teams are derived.

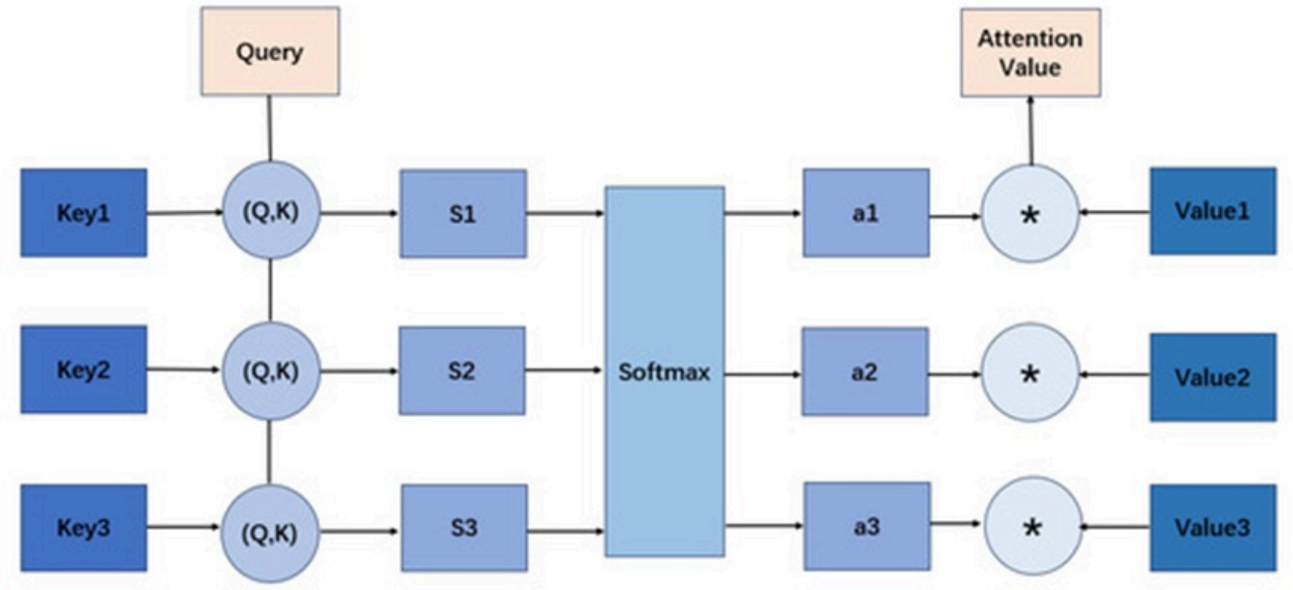


Figure 1. Schematic of scaled dot-product attention: relevance is computed via , normalized by softmax, and applied to V to form the context.

Multi-objective optimization

The discipline of multi-objective optimization concerns itself with navigating compromises across objectives that are fundamentally at odds with one another. Its organizing concept is Pareto optimality: a set of solutions in which no individual objective can be improved without degrading another. The formal statement of the problem is:

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

In cooperative multi-UAV control, the objectives that commonly arise include trajectory length, energy cost, collision safety, and the accuracy with which the formation is maintained. Traditional solution strategies often collapse the vector of objectives into a single scalar using a predefined function. Still, this reduction cannot faithfully capture the nuanced structure of the underlying trade-offs. Pareto-oriented alternatives, by contrast, deliver a collection of optimal compromise solutions and thereby supply richer information to guide decision-making.

In summary, this paper synthesizes multi-agent reinforcement learning, attention mechanisms, and multi-objective optimization into an integrated collaborative decision framework for multi-UAV formation control. The resulting system directly addresses the challenges of multi-objective coordination in dynamic, partially obscured environments.

Problem Formulation

The setting under investigation is cooperative formation flight by a multi-UAV team operating within a constructed urban district. The team is initialized from randomly sampled starting locations and must navigate collectively to a designated goal zone while meeting four simultaneous requirements: (i) avoiding any collision with buildings or with fellow team members, (ii) adhering to a prescribed spatial formation, (iii) keeping control effort and energy expenditure within bounds,

© 2023 The Author(s). | Published by Agro Intelligence Research Press

and (iv) completing the transit before a time limit expires. This scenario captures the essential characteristics of city-scale sensing and communication relay assignments, where formation quality directly influences both area coverage and inter-UAV link reliability **(Figure 2)**.

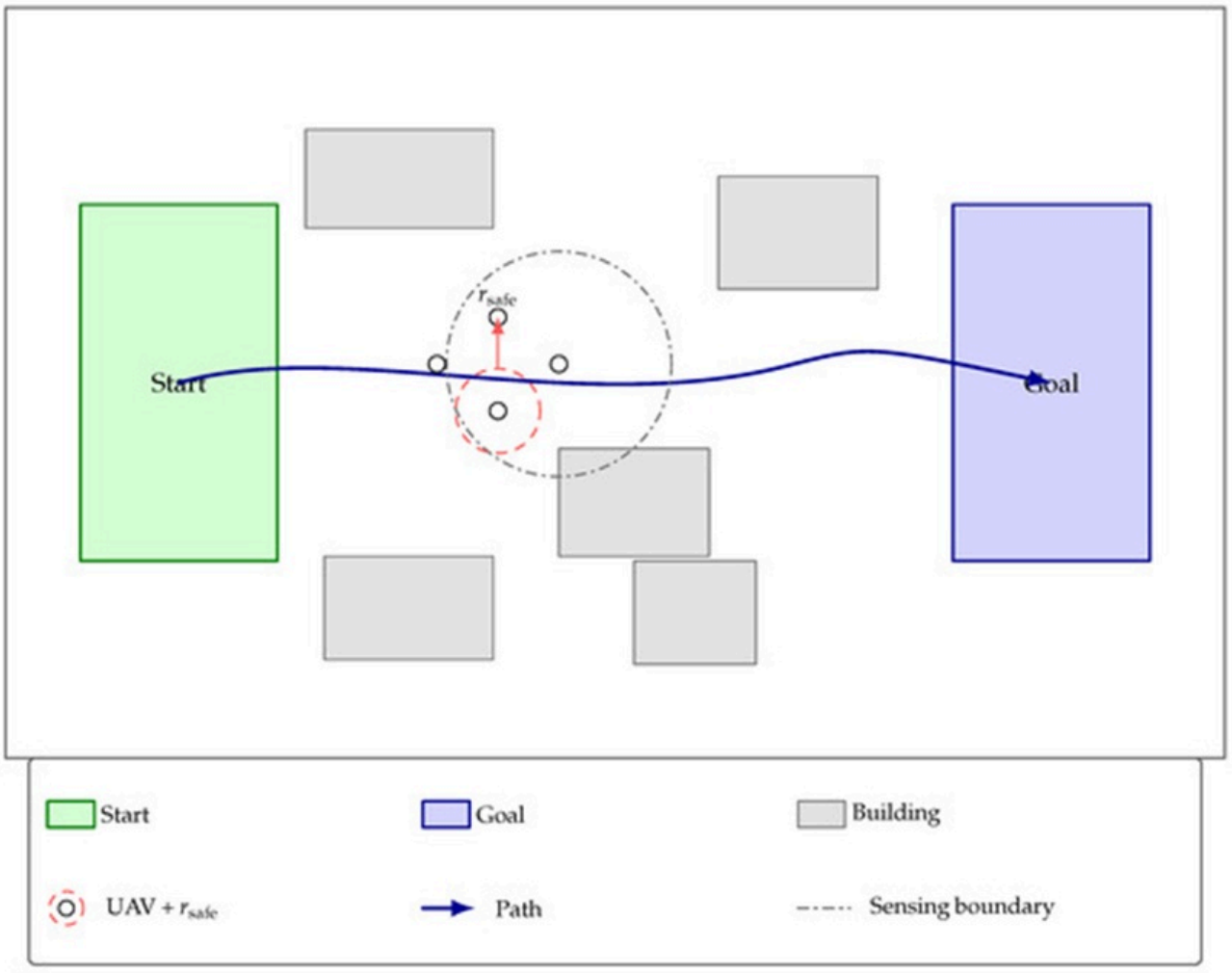


Figure 2. Problem setup (top view). Workspace: $x \in [0,100]$ m, $y \in [0,100]$ m, $z \in [0,22]$ m. Buildings are within $x \in [10,90]$, $y \in [10,90]$; half-length/width $\in$[1,10] m, height $\in$[9,13] m (cap 20). Initial region: $x \in [15,30]$, $y \in [10,90]$, $z \in [3,7]$; target region: $x \in [60,90]$, $y \in [10,90]$, $z \in [3,15]$. Example radii for illustration: sensing radius ≈15 m, safety radius $r$safe=2 m. The legend is placed below the map to avoid any overlap with the environmental elements.

Game Model and CTDE setting

The task is formalized as a partially observable Markov game (POMG) $\mathcal{G}$ specified through:

$$\mathcal{G} = \left(\mathcal{I}, \mathcal{S}, \{O_i\}_{i=1}^{N}, \{A_i\}_{i=1}^{N}, P, \{R_i\}_{i=1}^{N}, \gamma \right). \quad (3)$$

With the time index advancing as $t = 0,\ldots,$ in increments of Δt. The collection of agents is denoted $\mathcal{I}=\{1,\ldots,N\}$. The learning architecture adheres to the CTDE principle: a centralized critic enjoys access to globally aggregated information during the training phase. At the same time, each agent must rely exclusively on its local observation stream at execution time.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

The global state assembled for the critic concatenates the kinematic variables of the entire team, the goal specification and formation blueprint, and a representation of nearby obstacles:

$$s_t = [X_t, V_t, G, \mathcal{F}, E_t], \quad (4)$$

Where $\mathrm{X_t} = [\mathrm{x_{1,t}}, \ldots, \mathrm{x_{N,t}}]$ and $\mathrm{V_t} = [\mathrm{v_{1,t}}, \ldots, \mathrm{v_{N,t}}]$ gather the position and velocity vectors of all agents, G encapsulates the goal location/region, $\mathcal{F}$ records the desired formation slot offsets $\{\mathrm{r}_i^*\}_{\mathrm{i=1}}^{\mathrm{N}}$ (the spatial template for the formation), and $\mathrm{Et} = \{\mathrm{bk}\}_{\mathrm{k=1}}^{\mathrm{M}}$ lists the M most proximal axis-aligned buildings, each represented by its center coordinates and half-dimensions.

Observations and actions

The information accessible to each agent is strictly local. To interface cleanly with the three attention branches that will be introduced later, the local observation is organized into three distinct segments and then fused as:

$$\tilde{s}_{t,t} = \mathrm{Concat}\left(\mathrm{SelfAtt}\left(o_{t,t}^{\mathrm{self}}\right), \mathrm{InterAtt}\left(o_{t,t}^{\mathrm{inter}}\right), \mathrm{EntityAtt}\left(o_{t,t}^{\mathrm{ent}}\right)\right). \quad (5)$$

Self features. $o_{\mathrm{i,t}}^{\mathrm{self}}$ =[xi,t, vi,t, gi,t, ei,t, $d_{\mathrm{i,t}}^{\mathrm{goal}}$], where $x_{i,t}\in\mathbb{R}^3$ and $v_{i,t}\in\mathbb{R}^3$ capture the agent's own position and velocity, gi,t records its heading angle, $e_{i,t}\in[0,1]$ provides a normalized indicator of remaining energy, and $=\|xi,t-x^{\mathrm{goal}}\|2$ gives the Euclidean distance separating the agent from the goal. Inter-agent features. Drawn from the K nearest teammates $\mathcal{N}_{\mathrm{i}}^{\mathrm{K}}$, the inter-agent observation comprises relative kinematic quantities: $o$interi,t= {Δxij,t=xj,t−xi,t, Δvij,t=vj,t−vi,t, $d$pairij,t=‖Δxij,t‖2}$j\in\mathcal{N}Ki$, with the neighborhood size K (e.g., K=4) and re-evaluated at every timestep. Entity features. For the M buildings closest to agent i, the entity observation is defined as i, $o_{\mathrm{i,t}}^{\mathrm{ent}} = \{\Delta x_{ik,t} = c_k - x_{i,t},\ h_k,\ w_k,\ \ell_k\}_{\mathrm{k=1}}^{\mathrm{M}}$, where ck denotes the geometric center of the building and (hk,wk, ℓk) specify its half-extents along each axis; keeping M small (e.g., M=10) ensures that inference latency remains predictable. The action space consists of 3-D thrust or velocity commands, each clipped to satisfy a magnitude constraint:

$$a_{t,t} \in \mathcal{A}_t \subset \mathbb{R}^3,\ \|a_{t,t}\|_2 \le a_{\max}. \quad (6)$$

Vector reward and termination

The control problem is inherently multi-objective. Every agent receives a reward vector with four components, corresponding respectively to task advancement, energy consumption, formation quality, and safety:

$$R_t(s_t, a_t, s_{t+1}) = \left\{r_t^{\mathrm{task}}, r_t^{\mathrm{energy}}, r_t^{\mathrm{formation}}, r_t^{\mathrm{safety}}\right\}, \quad (7)$$

With identical definitions and coefficient values held constant across all experimental runs (defaults indicated parenthetically). Task progress and success. Using ηsucc=5.0 and $c$prog=0.5:

$$r_t^{\mathrm{task}} = \eta_{\mathrm{succ}} \cdot \mathbf{1}\left\{d_{t,t+1}^{\mathrm{goal}} \le \varepsilon_{\mathrm{goal}}\right\} - c_{\mathrm{prog}} \cdot \left(d_{t,t+1}^{\mathrm{goal}} - d_{t,t}^{\mathrm{goal}}\right), \quad (8)$$

© 2023 The Author(s). | Published by Agro Intelligence Research Press

So that each step yielding progress toward the goal incurs a proportional reward, and the act of entering the εgoal-ball triggers a lump-sum bonus—energy/control. Using $c$eng=0.01:

$$r_t^{\text{energy}} = -c_{\text{eng}} \cdot \|a_{t,t}\|_2^2. \quad (9)$$

Formation coherence. Using $c$form=1.0 and εform=1.0:

$$\text{dev}_t = \frac{1}{N}\sum_{t=1}^{N} \| x_{t,t} - x_{t,t}^{\text{ref}} \|_2,\ r_t^{\text{formation}} = -c_{\text{form}} \cdot \text{dev}_t + 0.2 \cdot \mathbf{1}\{\text{dev}_t \le \varepsilon_{\text{form}}\}, \quad (10)$$

Where the reference position for agent i, denoted $x_{\text{i,t}}^{\text{ref}}$=$x_{\text{lead,}}$+$r^{\star}{}_i$, is computed relative to a (virtual) team leader; a small average deviation from the assigned slot is lightly positively reinforced. Safety. Using $c$col=5.0, $c$near=1.0, and δ=2.0 m:

$$r_t^{\text{safety}} = -c_{\text{col}} \cdot \mathbf{1}\{\text{collision at } t\} - c_{\text{near}} \cdot \mathbf{1}\left\{\min\left(d_{tj,t}^{\text{pair}}, d_{t,t}^{\text{obs}}\right) < \delta\right\}, \quad (11)$$

Where $d_{\text{i,t}}^{\text{obs}}$ measures the shortest distance from the agent to the surface of the nearest building. An episode is brought to a close when all agents occupy the goal region, when any collision event occurs, or when the horizon T is exhausted.

Objective and CTDE realization

The aim is to produce decentralized actors that achieve Pareto efficiency with respect to the four specified objectives under the guidance of a centralized critic. Let the joint policy be written as $\pi = \{\pi i\}_{i=1}^{N}$, where $\pi_i(a_{i,t} \mid \tilde{s}_{i,t})$ denotes the per-agent actor mapping from local observations to actions. The vector-valued return, averaged across the team, is expressed as:

$$\mathbf{J}(\pi) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t \cdot \frac{1}{N}\sum_{t=1}^{N} R_t(s_t, a_t, s_{t+1})\right], \quad (12)$$

And the relative quality of policies is determined by the Pareto dominance criterion (no objective value is worse, and at least one is strictly better). The centralized critic generates per-objective value estimates that direct the policy update process:

$$Q^{(\kappa)}(s_t, a_t),\ \kappa \in \{\text{task, energy, formation, safety}\}, \quad (13)$$

while the actors consume $\tilde{\text{s}}_{\text{i,t}}$ as constructed according to (5). Maintaining this separation ensures that modeling considerations (reward definitions and constraint specifications) remain cleanly independent of architectural design choices (attention modules and CTDE structure).

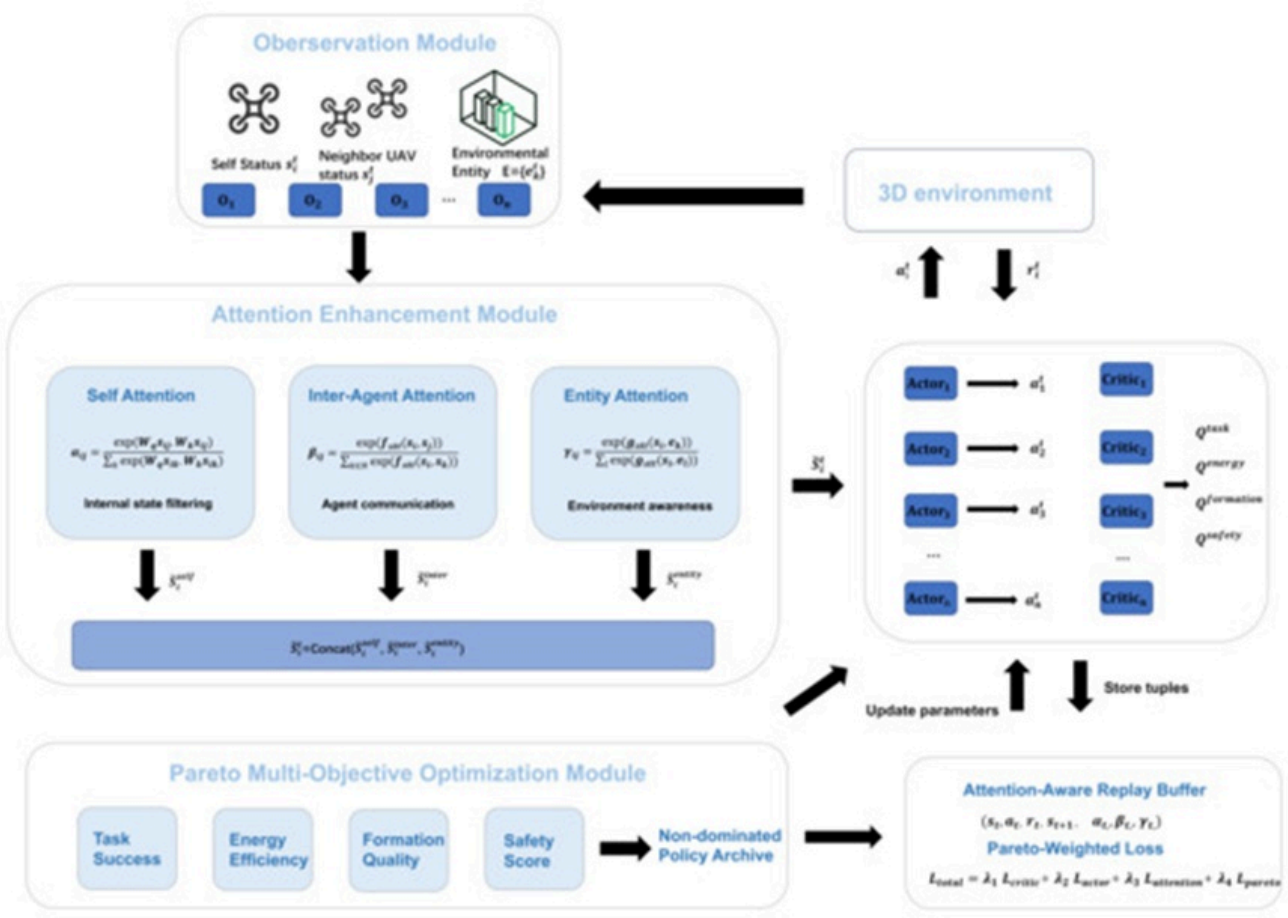
Proposed Method

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Framework overview

We introduce an enhanced architecture for autonomous UAV formation control that is designed to be backbone-agnostic: the attention-based front end and the Pareto optimization layer can be coupled with any standard CTDE actor–critic algorithm (such as MADDPG, MATD3, MASAC, or MAPPO) through a minimal integration interface—the observation encoder preceding each actor is replaced by the attention block, while the Pareto layer is applied on top of per-objective critics or advantage functions. To ensure a concrete instantiation and a level playing field for evaluation, this study implements and tests the design specifically with MADDPG; comprehensive benchmarking across multiple backbones is reserved for future exploration, given the substantial computational and hyperparameter-tuning demands of long-duration urban navigation tasks. In the present work, MADDPG serves as the concrete substrate for realizing and fairly assessing the framework, but the design principle extends without modification to other backbones (e.g., MATD3/MASAC/MAPPO). It is important to note that, unlike deep GNN or Transformer stacks, our attention block is purposefully typed and limited to a single hop to remain within the computational budgets of onboard hardware; this design choice is elaborated in Section 5.3.

MADDPG [16], a representative MARL algorithm, embodies the Centralized Training–Decentralized Execution (CTDE) philosophy: a centralized critic, trained with access to global state information, guides the learning of decentralized actors that make decisions based solely on local observations during execution. Starting from this base architecture, we embed attention mechanisms to improve the processing and fusion of local observations under conditions of partial observability. As illustrated in **Figure 3**, the overall information flow adheres to the CTDE paradigm but is specifically adapted to handle multi-objective formation control.



© 2023 The Author(s). | Published by Agro Intelligence Research Press

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 apply to other CTDE actor–critic MARL algorithms.

In detail, each UAV makes its decision based on its local observations, which are first processed by the Attention Enhancement Layer, which deploys three lightweight processing branches. The resulting three outputs are concatenated together into a single attention-enriched representation:

$$\tilde{s}_t = \text{Concat}\left(\text{SelfAtt}\left(o_t^{\text{self}}\right), \text{InterAtt}\left(o_t^{\text{inter}}\right), \text{EntityAtt}\left(o_t^{\text{ent}}\right)\right). \quad (14)$$

The Actor module maps this enhanced representation $\tilde{s}_i$ to the corresponding action a_i, while centralized Critics produce per-objective value estimates $\{Q^{\text{task}}, Q^{\text{energy}}, Q^{\text{formation}}, Q^{\text{safety}}\}$ simultaneously. During training, the Pareto layer maintains a compact archive of non-dominated policy solutions and provides adaptive signals or weighting coefficients to guide learning updates, effectively replacing brittle manual scalarization. Experience tuples of the form (s,,′,attention context) are stored in an attention-aware replay buffer, and training proceeds by minimizing a composite loss function $L_{\text{total}}=\lambda 1 L_{\text{critic}}+\lambda 2 L_{\text{actor}}+\lambda 3 L_{\text{attention}}+\lambda 4 L_{\text{pareto}}$.

In summary, within the MADDPG-based realization **(Figure 3)**, the attention module improves the saliency of information extracted under partial observability, and the Pareto module enforces a principled approach to multi-objective compromise; viewed more generally, these two modules augment the selected MARL backbone with minimal structural intrusion and can be transferred across a broad family of CTDE actor–critic methods beyond MADDPG alone.

Attention mechanism integration

The trio of attention mechanisms incorporated into our framework fulfills three separate but mutually reinforcing functions: facilitating richer communication between agents, sharpening environmental perception, and enabling hierarchical decision-making. Each mechanism is designed to address particular bottlenecks in multi-agent coordination while jointly enhancing the overall system. We describe each of the three integrated attention mechanisms in turn below.

Self-attention mechanism

The purpose of the self-attention mechanism is to operate on each UAV agent's internal state representation and identify the features most relevant to the current decision context. For an agent with state vector $si \in \mathbb{R}d$, the self-attention module derives an attention weight αi for every state component j according to:

$$\alpha_{t,j} = \frac{\exp\left(W_q s_{t,j} \cdot W_k s_{t,j}\right)}{\sum_{k=1}^{d} \exp\left(W_q s_{t,k} \cdot W_k s_{t,k}\right)} \quad (15)$$

Where Wq and Wk denote learned matrices for query and key projection, respectively, this mechanism enables an agent to dynamically emphasize or de-emphasize different dimensions of its own state depending on situational demands, leading to more informed decisions in complex scenarios through context-sensitive feature weighting.

Inter-agent attention mechanism

The inter-agent attention mechanism establishes an explicit channel for information exchange among UAVs by computing attention weights across the states of neighboring agents. For agent i with neighbor set $\mathcal{N}i$, the attention weight assigned to a particular neighbor j is calculated as:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$\beta_{i,j} = \frac{\exp\left(f_{att}\left(s_i, s_j\right)\right)}{\sum_{k \in N_i} \exp\left(f_{att}\left(s_i, s_k\right)\right)} \quad (16)$$

Where $fatt$ represents a neural network tasked with scoring the relevance between agent i and agent j, this mechanism empowers agents to selectively attend to those teammates whose information is most valuable for coordination, thereby supporting reliable formation keeping and collision avoidance through directed information sharing.

Entity attention mechanism

Processes the set of environmental entities, including obstacles, target locations, and other dynamic elements. Given a collection of environmental entities E={e1,e2,…,em}, the mechanism computes attention weights that determine how relevant each entity is to the agent currently making a decision:

$$\gamma_{i,k} = \frac{\exp\left(g_{att}\left(s_i, e_k\right)\right)}{\sum_{l=1}^{m} \exp\left(g_{att}\left(s_i, e_l\right)\right)} \quad (17)$$

Through this mechanism, agents can dynamically focus on the environmental features that matter most, substantially improving navigation efficiency and obstacle-avoidance performance by filtering out irrelevant sensory clutter.

Interpreting attention as information priority

Throughout this paper, the softmax-based weights produced by each attention branch are interpreted directly as indicators of information priority. For the inter-agent branch specifically, given a query qi derived from agent i and per-neighbor key/value pairs $\{\mathrm{k_j}, \mathrm{v_j}\}_{\mathrm{j} \in \mathcal{N}_{\mathrm{i}}(\mathrm{t}),}$

$$\beta_{ij,t} = \mathrm{softmax}_j\left(\frac{q_i^{\top} k_j}{\sqrt{d_k}}\right),\ m_{i,t}^{\mathrm{inter}} = \sum_{j \in N_i(t)} \beta_{ij,t}\, v_j, \quad (18)$$

The constraint $\sum j\beta ij$,=1 holds over the neighbors present. A comparatively higher value of means that neighbor j contributes proportionally more to the aggregated message $m_{i,t}^{inter}$ and therefore occupies a more elevated position in the priority ordering for decision-making at time t. An analogous definition applies to γik in the entity branch, ranking environmental entities by their relative importance; for self-attention, we obtain feature-group weights wi that reweight internal cues (goal bearing, safety margins, energy levels, etc.). At test time, when packets are lost or links become unavailable, the entries corresponding to masked elements are removed, and the weights over the remaining elements are renormalized; in the extreme case where no neighbor is accessible, a learned default embedding is substituted. The actor consumes the fused representation

$$\tilde{s}_{i,t} = \mathrm{Concat}\left(m_{i,t}^{\mathrm{self}},\ m_{i,t}^{\mathrm{inter}},\ m_{i,t}^{\mathrm{ent}}\right), \quad (19)$$

So the notion of “priority” directly influences which cues dominate the policy input. When needed, we also quantify the concentration of attention via the attention entropy (t)=−$\sum j\beta ij,t$log$\beta ij,t$ and the Top-1 share max$j\beta ij,t$.

Positioned within the CTDE pipeline, attention serves as the representational front end. Self-attention filters the agent’s own kinematics and internal intent indicators, inter-agent attention singles out the few teammates whose state is most

© 2023 The Author(s). | Published by Agro Intelligence Research Press

informative for the ongoing maneuver, and entity attention pushes the most critical obstacles into the foreground. The resulting fused vector $\tilde{s}i$ thus exhibits a more organized structure and greater temporal coherence than the raw observations, enabling the centralized multi-objective critics to operate on inputs in which task progress, formation error, energy usage, and collision risk can be more readily separated. The practical consequence is cleaner per-objective value estimation and sharper policy gradient signals, reducing ambiguity about which agent bears responsibility and which objective should be adjusted. Attention, therefore, strengthens the representational quality on the actor side and, as an indirect benefit, helps the critics perform credit assignment with fewer confounding correlations.

Design contrast to Stacked GNN/transformer front ends

Our front-end design employs three typed, single-hop attention branches (self/inter-agent/entity) that function as feature selectors and feed into a CTDE actor–critic structure. This configuration is distinct from a stacked GNN or Transformer encoder. The principal points of differentiation are as follows:

(i) Receptive field and depth. Our architecture performs only one-hop aggregation over the K nearest teammates and the M nearest entities at each timestep and does not stack multiple layers; by comparison, stacked GNNs propagate information over L hops (depth L), and stacked Transformers conduct global token mixing at every layer. Our single-hop restriction preserves local responsiveness and caps computational latency.

(ii) Heterogeneous typing vs. homogeneous mixing. We process the self, inter-agent, and entity information streams through separate, dedicated branches, then concatenate the results rather than merging all tokens into a single homogeneous processing block. This design avoids the overhead of learning type-specific embeddings or positional encodings, thereby reducing the overall computational footprint.

(iii) Compute/latency scaling. For each agent at each timestep, the cost of our method scales linearly with $K+M$ (fixed fan-in). Stacked GNNs scale with both the number of edges and depth (L $|E|$), with $|E| \approx NK$), and stacked Transformers scale quadratically in the number of tokens per layer (L $T2$), with $T \approx 1+K+M$).

(iv) Communication modeling. At test time, we mask unavailable links and renormalize the attention distribution accordingly; gradients are not backpropagated through an explicit learned communication channel. Many stacked designs assume either reliable broadcast communication or learning the underlying graph topology.

(v) Training signal. Our vector-valued critics, together with the Pareto archive, handle multi-objective trade-offs natively; most stacked architectures optimize a scalarized reward unless explicitly customized for multi-objective settings.

In essence, while the attention operator is used, the architecture that emerges is a typed, single-layer, locally scoped selector engineered for real-time UAV control, not a deep message-passing or global token-mixing stack.

With fan-in limits (K, M) held constant, our typed single-hop attention contributes $O(K+M)$ per-agent computation per control step on the actor side; consequently, test-time team cost grows linearly with N. Centralized critics impose computational requirements only during the training phase and do not affect deployment-time latency. This matches the characterization in **Table 1**. In practice, bounded values of K and M, masking of absent links, and mixed-precision inference collectively ensure that the actor-side latency remains predictable as N increases.

Table 1. Architectural contrast between PA-MADDPG and stacked GNN/Transformer front ends.

Aspect	PA-MADDPG (Proposed Method)	GNN-Oriented MARL (Layered)	Transformer-Oriented RL (Layered)
Information aggregation scope	One-hop KNN neighbors with M entities; separated typed branches	Multi-layer message propagation with L-hop	Self- and cross-attention across all tokens with global

© 2023 The Author(s). | Published by Agro Intelligence Research Press

	combined by concatenation	coverage	interaction at each layer
Computational cost per step	$O(K+M)$ using fixed fan-in	$O(L\|E\|)$, where $\|E\| \approx NK$	$O(L\,T2)$, with $T \approx 1+K+M$
Runtime and memory demand	Lightweight and suitable for on-board deployment	Moderate to high; increases with L and N	Large overhead due to quadratic token scaling
Support for heterogeneous data	Explicitly typed (self/inter/entity) without embedding types	Depends on engineered-type edges	Relies on token-type and positional embeddings
Communication assumption	Masked KNN with normalized attention	Static or learned graph structures, commonly assuming reliable communication links	Typically assumes full global broadcasting
Optimization signal	Vectorized critics with Pareto-based multi-objective learning	Commonly optimized with scalar rewards	Commonly optimized with scalar rewards
Main objective	Fast local decision-making in real time	Rich multi-hop relational reasoning	Capturing global context and long-distance dependencies

Pareto multi-objective optimization

The component built around Pareto optimization tackles what is inherently a multi-dimensional problem in UAV formation flight: agents must concurrently satisfy aims that are frequently at cross-purposes—mission completion, thrifty energy use, formation integrity, and impact prevention. Standard reinforcement learning pipelines falter in such multi-criteria regimes because determining how to weight the various reward components appropriately is a persistent challenge.

By recasting the learning problem within a multi-objective optimization frame, our Pareto-based formulation expresses the reward function Ri for agent i as a collection of separate, semantically distinct objective terms:

$$R_i = \left\{ r_i^{task}, r_i^{energy}, r_i^{formation}, r_i^{safety} \right\} \quad (20)$$

Each of which maps to a dimension of operational concern. Rather than painstakingly calibrating reward coefficients, the method maintains a running set of non-dominated solutions, enabling the exploration of a wide spectrum of trade-off configurations across the objectives.

The following rule defines the dominance criterion of Pareto: solution x is considered to dominate solution y whenever x matches or betters y on every objective and registers a strict improvement on at least one. The algorithm stores a curated archive of mutually non-dominating solutions and uses this repository to steer the course of policy updates:

$$\pi_{\theta_i}^{new} = \arg \max_{\theta_i} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R_i\,(s_t, a_t, s_{t+1}) \right] \quad (21)$$

The procedure that enacts the above update does so under constraints that explicitly enforce Pareto optimality—constraints that safeguard the retention of non-dominated solutions while encouraging the archive to span a diverse set of trade-off profiles across the objective space. What this constraint guarantees is that the policies retained correspond to efficient bargaining outcomes among the operational aims in tension.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

A direct practical consequence of this architecture is that the burdensome effort of manual reward design is rendered unnecessary, while strong performance is sustained across all operational dimensions simultaneously. The benefit is especially meaningful in UAV mission settings, where the priority ordering of objectives is liable to shift as the mission transitions from one phase to another; the framework delivers adaptive optimization without the need to revisit or recalibrate parameters.

Integrated framework architecture

The integrated framework knits together attention mechanisms and Pareto optimization within an augmented MADDPG structure that is laid out as a hierarchical processing sequence. Three principal building blocks make up this unified design—a processor that leverages attention to enrich observations, a critic network that handles multiple objectives, and an actor network that produces coordinated behaviors—and these blocks collectively embody the core contribution of our work.

Raw sensory streams are transformed into enriched state representations by the attention-enhanced observation processor through successive attention operations. Operating in parallel within this module are three distinct attention mechanisms: self-attention focused on the agent's internal state, inter-agent attention on neighboring UAVs' states, and entity attention on environmental features. The resulting representations are then concatenated into a single comprehensive state vector:

$$\tilde{s}_i = \mathrm{Concat}\left(\mathrm{SelfAtt}\left(s_i\right), \mathrm{InterAtt}\left(s_i, s_{\mathcal{N}_i}\right), \mathrm{EntityAtt}\left(s_i, E\right)\right) \quad (22)$$

Wherein each attention module operates according to the specifications outlined in Section 3.1.

The multi-objective critic network extends the conventional paradigm of value estimation by allocating independent value functions to each operational dimension:

$$Q_{\phi_i}\left(s, a\right) = \left[Q_{\phi_i}^{\mathrm{task}}\left(s, a\right), Q_{\phi_i}^{\mathrm{energy}}\left(s, a\right), Q_{\phi_i}^{\mathrm{formation}}\left(s, a\right), Q_{\phi_i}^{\mathrm{safety}}\left(s, a\right)\right] \quad (23)$$

What this architectural decision affords is the capacity for separate value estimation along axes that are frequently in opposition, which in turn makes possible a finer-grained assignment of credit when policy update steps are performed subject to Pareto constraints.

The coordinated actor network ingests the attention-enriched state representations and generates actions that reconcile the demands of individual objective satisfaction with the requirements of collective team coordination. To protect the stability of training within a decentralized execution regime, the design of this network weaves in residual connections and layer normalization:

$$a_i \sim \pi_{\theta_i}\left(\cdot \middle| \tilde{s}_i\right) \quad (24)$$

A vital feedback cycle arises between these components: attention mechanisms continuously shape the policy decisions being made. In contrast, those very decisions feed back to reconfigure attention patterns as environmental circumstances evolve. Through this mutually adaptive dynamic, the system continuously hones mission-specific compromises without recourse to manual parameter intervention.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Pareto archive, normalization, and adaptive weighting

Per-objective normalization

Denote by $\mathbf{J}^{(e)} = \left[\mathrm{J}^{(e)}_{\mathrm{task}}, \mathrm{J}^{(e)}_{\mathrm{energy}}, \mathrm{J}^{(e)}_{\mathrm{formation}}, \mathrm{J}^{(e)}_{\mathrm{safety}}\right]$ the episodic returns collected over the four objectives for episode e (evaluated in a manner aligned with **Eq. 12**) and aggregated at the team level). The scalar magnitudes of these returns vary considerably across objectives, a discrepancy we address by normalizing each objective through an online affine transformation whose parameters are derived from a running min–max envelope computed over the union of the current archive $\mathcal{A}$ and a FIFO buffer $\mathcal{B}$ that stores returns from recent episodes:

$$\tilde{J}_k^{(e)} = \frac{J_k^{(e)} - m_k}{M_k - m_k + \varepsilon}, \tilde{J}_k^{(e)} \in [0,1], \quad (25)$$

Where mk and Mk denote the running minimum and maximum for each objective, respectively, updated using an exponential moving average with a decay factor of 0.99 over the set $\{ \mathbf{J}^{(e)} \in \mathcal{B} \} \cup \mathcal{A}$. The small constant $\varepsilon = 10^{-8}$ is included to avoid division by zero. This monotonic transformation maintains the Pareto ordering while normalizing the scales to support consistent weighting and diversity measurement.

Archive update and pruning

The archive $\mathcal{A}$ we maintain is bounded in size and consists of non-dominated normalized vectors residing within $[0,1]^4$. Upon obtaining a candidate vector $\tilde{J}$ from the current policy, we evaluate it against the archive using the criterion of ϵ-dominance, adopting a modest per-objective margin of ϵk=0.01 to suppress excessive cycling:

$$\tilde{\mathbf{u}} \prec_\varepsilon \tilde{\mathbf{v}} \iff (\forall k: \tilde{u}_k \le \tilde{v}_k + \epsilon_k) \wedge (\exists k: \tilde{u}_k < \tilde{v}_k - \epsilon_k), \quad (26)$$

Where maximization is understood to be implemented by flipping signs on objectives that are conventionally minimized, if no archive member $\tilde{a} \in \mathcal{A}$ is found to ϵ-dominate the candidate $\tilde{J}$, we proceed to insert $\tilde{J}$ into the archive and concurrently purge any pre-existing points that are now ϵ-dominated by $\tilde{J}$. To keep the archive within its capacity limit $|\mathcal{A}| \le K\mathrm{max}$ (with the default set at $K\mathrm{max}$=100; consult **Table 2**), we resort to pruning according to the NSGA-II crowding distance metric computed in normalized space: when capacity is exceeded, the point with the smallest crowding distance is repeatedly removed; in cases where distances tie, the point bearing the older timestamp is discarded, thereby favoring the retention of fresher coverage of the Pareto front.

Table 2. Hyperparameters of the proposed multi-agent attention-DRL framework.

Category	Parameter	Value
Network Architecture	Actor network hidden layers	[256, 128]
Network Architecture	Critic network hidden layers	[512, 256, 128]
Network Architecture	Attention embedding size	64
Training	Actor learning rate	1×10−4
Training	Critic learning rate	1×10−3
Training	Mini-batch size	256

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Training	Replay buffer capacity	10^6
Training	Discount coefficient (γ)	0.99
Training	Soft update rate (τ)	0.01
Attention	Number of self-attention heads	4
Attention	Inter-agent interaction radius (m)	10.0
Attention	Entity attention radius (m)	15.0
Pareto Archive	Maximum archive capacity	100
Pareto Archive	Update interval	10 steps

Adaptive scalarization pareto weight ($\mathcal{A}$,π)

We construct a vector of per-objective weights $\lambda \in \Delta^3$ (confined to the simplex) with the explicit intent of directing greater emphasis toward those objectives on which the current policy's performance falls short of what the archive shows to be achievable. The utopia reference point is defined as $z_k^* = \max_{\tilde{a} \in \mathcal{A}} \tilde{a}k$. The current policy's expected normalized returns, denoted $\bar{J}$, are estimated either from a dedicated validation batch or from a brief evaluation rollout. From these, we form nonnegative gap measures $g_k = \max\left\{0,\ z_k^* - \bar{J}_k\right\}$ and compute a softmax distribution parameterized by a temperature τw:

$$\lambda_k = \frac{\exp\left(g_k/\tau_w\right)}{\sum_j \exp\left(g_j/\tau_w\right)},\ \tau_w = 0.15. \qquad (27)$$

To ensure against the pathology of concentrating all weight on a single objective, the distribution obtained from the gaps is blended with a uniform prior uk=1/4:

$$\lambda_k^{\text{final}} = (1-\rho)\ u_k + \rho\,\lambda_k,\ \rho \uparrow 0.9 \text{ linearly over training.} \qquad (28)$$

These final weights are subsequently plugged into the multi-objective loss as , where each per-objective loss $L_{\text{total}} = \sum_k \lambda_k^{\text{final}}\ L_k$ L_k is specified in Section 5. Since the critics are trained on normalized target quantities, the λ weights operate across dimensions that are already brought onto a comparable footing.

Computation frequency and overhead

Every 10 gradient steps, we trigger the ArchiveUpdate procedure (consistent with the "Update frequency" row of **Table 2**), derive a fresh λ vector at each such invocation, and recompute the running statistics (mk,k) from the merged pool $\mathcal{A} \cup \mathcal{B}$. All of these operations are confined to $\mathbb{R}4$, making their computational cost entirely insignificant relative to the forward and backward passes through the actor and critic networks.

Archive-size sensitivity (How Kmax Matters)

© 2023 The Author(s). | Published by Agro Intelligence Research Press

To quantify the diversity of solutions retained in the archive, we measure the hypervolume (HV) in the normalized space 〖[0,1]〗^4, with the reference point fixed at $r = \mathbf{0}$, approximating the integral via Monte Carlo sampling of 〖10〗^5 points. The sensitivity experiments reported in Section 6.3 sweep Kmax across the range {50, 100, 150, 200} and demonstrate that the configuration Kmax=100 preserves roughly 85% of the peak hypervolume attainable with much larger archives while consuming only around 60% of their computational budget; shrinking Kmax below this value causes a marked contraction in HV, with the most visible degradation appearing as sparse coverage along the safety and energy dimensions. Based on this evidence, we set Kmax = 100 as the standard value and expose it as an adjustable parameter for practitioners.

Training algorithm and implementation

The training methodology developed here extends MADDPG by weaving attention mechanisms and Pareto-driven optimization directly into its fabric, assembling a learning system capable of robust multi-UAV coordination. Throughout, the CTDE separation is maintained: centralized critics exploit globally aggregated information during training, while decentralized actors depend solely on their local observation streams at execution time. The architecture thus combines the scalability properties of decentralized decision-making with the learning efficiency provided by centralized information.

A particularly noteworthy addition is the attention-aware replay mechanism. Instead of storing only the standard components $(\mathrm{s_t}, \mathrm{a_t}, \mathrm{r_t}, \mathrm{s_{t+1}})$, experience tuples are augmented to include attention context vectors (αt, βt, γt) that capture, at the moment of collection, the instantaneous distribution of focus produced by each of the three attention branches. Retaining this attentional context across replay operations contributes substantially to learning stability and helps prevent the network from overwriting earlier attention behaviors with new but less effective patterns. Stratified sampling within the replay buffer further ensures that diverse attention contexts and mission types are represented in balanced proportions during training.

The loss function, which must accommodate multiple objectives simultaneously, combines four essential terms under the governance of dynamically adjusted weights derived from Pareto optimization:

$$L_{\text{total}} = \lambda_1 L_{\text{critic}} + \lambda_2 L_{\text{actor}} + \lambda_3 L_{\text{attention}} + \lambda_4 L_{\text{pareto}} \quad (29)$$

Within this expression, Lattention enforces temporal consistency in learning attention weights, while Lpareto preserves diversity across the evolving set of non-dominated solutions. The adaptive coefficients λi shift in accordance with the dominance structure observed within the current solution archive.

The implementation uses PyTorch as its framework, supplemented by custom CUDA kernels written to accelerate the parallelized attention computations. Network architectures take the form of multilayer perceptrons with ReLU nonlinearities and batch normalization layers; the principal hyperparameter settings include an attention dimensionality of $datt$=64, replay storage for up to 10^6 transitions, minibatch draws of size 256, and a discount factor γ=0.99. The Pareto archive accommodates a maximum of 100 non-dominated solutions, a limit chosen to balance the breadth of the retained trade-off surface against the processing overhead incurred during archive maintenance.

Information shared between agents travels over a logical neighborhood graph. To reproduce the effects of imperfect communication channels during evaluation, we subject the system to Bernoulli packet loss (occurring with probability $p$loss), random delays simulated by replaying neighbor states that have been held in a short sliding-window history (delay values drawn uniformly from [0,Dmax] ms), and time-varying connectivity produced by probabilistically severing links at rate q. Entries that cannot be retrieved are omitted and substituted with zero vectors (equivalently, masked away within the attention operations). Additional stress-testing includes injecting Gaussian-distributed noise into the feature vectors, occasionally fabricating a phantom obstacle, or constraining the neighbor fan-in parameter K to a reduced value.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

The complete algorithmic specification of the proposed method is presented in **Algorithm 1**.

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

Algorithm 1. Multi-Attention Meets Pareto Optimization.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Experiment

Experimental setup

A comprehensive experimental campaign was conducted across a range of UAV formation control scenarios to assess the effectiveness of our attention-augmented MADDPG framework. The evaluation design includes direct comparisons against established multi-agent reinforcement learning methods, systematic ablation of individual architectural components, and sensitivity testing over critical hyperparameter ranges.

The computational resources consisted of a high-performance cluster built around NVIDIA RTX 4090 GPUs (each equipped with 24 GB of VRAM), Intel Xeon Gold 6248R CPUs (operating at 3.0 GHz across 24 physical cores), and 128 GB of DDR4 RAM. All experiments were run on Ubuntu 20.04 LTS, with hardware acceleration provided by CUDA 11.8 and cuDNN 8.6. For the most computationally demanding configurations, training throughput was increased through data parallelism across four GPUs.

The software stack was anchored by PyTorch 2.0.1 and Python 3.9.16. Supporting packages included NumPy 1.24.3 for numerical processing, Matplotlib 3.7.1 for figure generation, TensorBoard 2.13.0 for logging experimental trajectories, and OpenAI Gym 0.21.0 for the environment abstraction layer. Hand-tuned CUDA kernels further accelerated attention-specific computations.

Hyperparameter assignments, determined through an exhaustive grid search, are catalogued in full within **Table 3**. The parameters bearing the greatest influence on performance include the actor step size (10−4), the critic step size (10−3), the replay storage limit (106 transitions), the minibatch cardinality (256), the discount factor γ (0.99), the Polyak averaging coefficient τ (0.01), the attention embedding dimension (64), and the Pareto archive bound (100). In the absence of explicit notation to the contrary, each experiment is trained for 10,000 episodes without early termination or manual tuning. Reported curves trace the moving average over a trailing window of 100 episodes, and unless stated otherwise, results are aggregated over 5 independently seeded runs and presented as mean ± one standard deviation. Every training figure adopts the same horizontal range of 10,000 episodes.

Table 3. 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
$\mathcal{F}$, $r^{\star}_i$	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 (7)
γ	Discount factor
π_i, π	Decentralized policy and joint policy
$Q(\kappa)$	Per the objective centralized critic in (13)
$a_{\max}, \delta, \varepsilon_{\text{goal}}$	Action bound, safety distance, goal threshold

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Baseline selection and scope

For our benchmarking, we select two CTDE reference algorithms positioned at opposite poles of the complexity axis: (i) MADDPG, representing the off-policy deterministic actor–critic family with centralized critics and decentralized actors, and (ii) IDQN, representing value-based methods. This dyadic comparison enables us to cleanly separate the contributions of our two proposed modules—the typed, single-hop attention front end and the Pareto-driven multi-objective layer—while holding training protocols and hyperparameter schedules constant across all compared conditions.

Broadening the baseline to encompass additional actor–critic variants, such as MATD3, MASAC, or MAPPO, would be technically feasible but is not undertaken in the present study for three reasons grounded in practical considerations. (1) Attribution clarity versus confounding factors. We aim to measure the marginal gain that the proposed modules yield when embedded in a fixed CTDE template; introducing a heterogeneous mix of backbones would simultaneously vary whether learning is on-policy or off-policy, whether policies are stochastic or deterministic, and how entropy regularization and advantage estimation are handled—all of which would obscure the source of any observed differences. (2) Wall-clock and tuning budget. Each configuration reported here trains for 10,000 episodes, across team sizes N=2,3,5 and under multiple random seeds. Adding MATD3, MASAC, or MAPPO—each of which requires its own stability-sensitive hyperparameter adjustments, such as target smoothing magnitudes or entropy temperature schedules—would multiplicatively increase the total computational cost and demand careful per-N retuning. (3) Focus on real-time feasibility. Our design targets execution with minimal latency, relying on single-hop attention with bounded fan-in; stacked architectures and on-policy algorithms would significantly increase either the sample complexity or the per-step computational load, a direction that falls outside the scope of the contribution under investigation.

To support reproducibility and to assist researchers who may wish to extend our work, we supply implementation notes and configuration details both in the main text and in the accompanying code release, including guidance on how to interface our attention and Pareto modules with MATD3, MASAC, and MAPPO (specifying loss formulations, target construction, and the points at which adaptive weights are incorporated). A subsequent study operating under expanded runtime allowances will include these broader baseline comparisons.

Experimental environment

Starting from the simulator provided at https://github.com/young-how/DQN-based-UAV-3D_path_planer (retrieved on 27 October 2025), we modify and scale the environment to support multiple UAVs operating simultaneously. A visual depiction of the environment, along with representative results, appears in **Figure 4**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

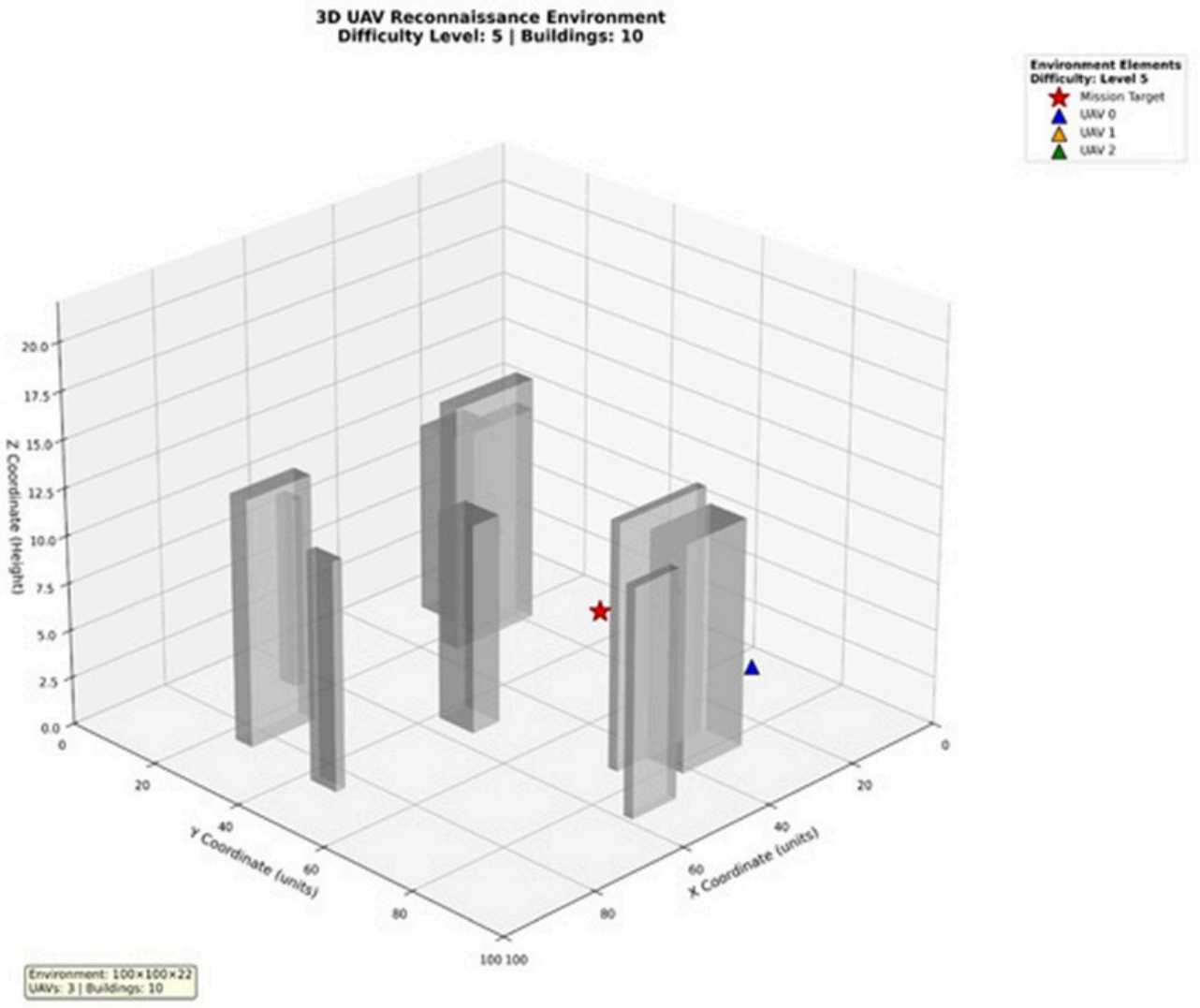


Figure 4. Three-dimensional environmental schematic diagram.

The operational volume is a 3D rectangular region bounded by $x \in [0,100]$, $y \in [0,100]$, and $z \in [0,22]$. Structures are placed throughout this volume via stochastic generation, with both spatial coordinates and dimensional parameters drawn randomly. To avoid overwhelming the learning process with excessive obstacles during the earliest training episodes, the number of buildings is ramped up incrementally: each time the success rate computed over the trailing 100 navigation episodes exceeds 70%, additional structures are introduced. A hard ceiling of 20 buildings is enforced. Within the horizontal plane, buildings are restricted to the footprint defined by $x \in [10,90]$, $y \in [10,90]$; their half-lengths span the interval [1,10], half-widths span [1,10], and heights occupy the range [9,13]. UAVs are initialized within the box $x \in [15,30]$, $y \in [10,90]$, $z \in [3,7]$; the destination zone corresponds to $x \in [60,90]$, $y \in [10,90]$, $z \in [3,15]$. **Figure 5** supplements this description with a 2D overhead view that clarifies the spatial arrangement.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

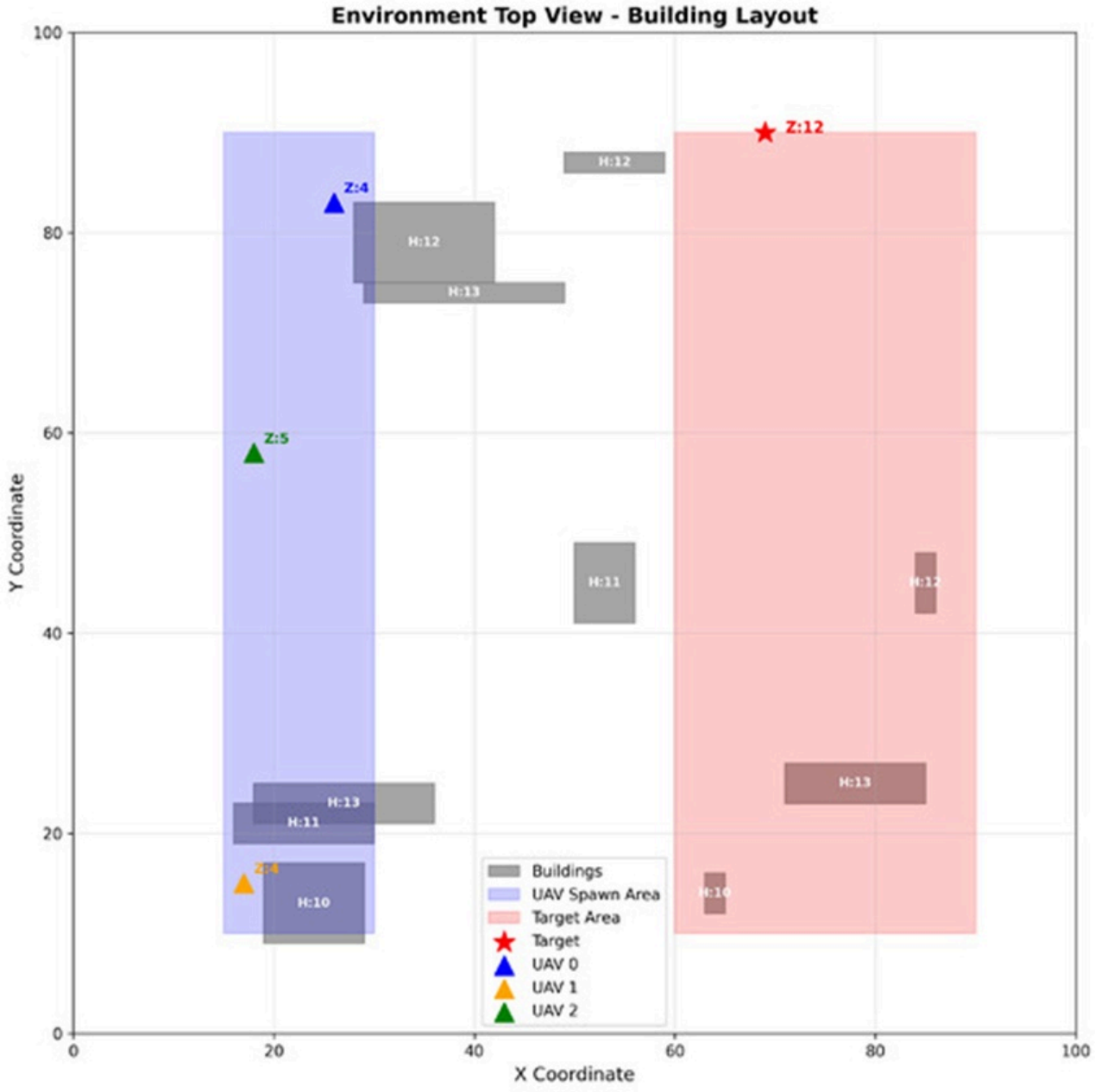


Figure 5. Two-dimensional top view of the environment.

Because stacked GNN or Transformer encoders incur steep penalties in both latency and memory consumption, we confine our empirical comparisons to widely used CTDE baselines (MADDPG, IDQN) and to controlled ablation studies of our attention and Pareto components under a uniform training protocol. To maintain emphasis on the constraints of on-board execution, we present an architectural juxtaposition against stacked GNN and Transformer front ends (Section 5.3; **Table 1**) rather than benchmarking against deep stacks that would exceed the real-time budget we target.

Evaluation metrics and units

Four primary metrics are tracked, each supplied with a formal operational characterization and the corresponding units. Except where explicitly stated otherwise, all rates are computed at the level of a complete team episode and are reported

© 2023 The Author(s). | Published by Agro Intelligence Research Press

as the 100-episode moving average at the conclusion of training or evaluation; tables list the mean ± one standard deviation of this summary quantity computed across independent random seeds.

Team success rate (%)

An episode e is judged to be successful if and only if every agent in the team reaches the goal region before the horizon T is reached, and no collision event of any kind occurs during the episode:

$$\mathbf{1}_{\text{succ}}^{(e)} = \mathbf{1}\left\{\forall i \in \mathcal{I}:\ \tau_i^{(e)} \le T \ \wedge\ \text{no collision in } e\right\}. \quad (30)$$

The team's success rate is obtained by averaging this indicator over an evaluation set $\mathcal{E}$ (for instance, the final 100 episodes):

$$\text{Succ. rate} = \frac{1}{|\mathcal{E}|}\sum_{e\in\mathcal{E}} \mathbf{1}_{\text{succ}}^{(e)} \times 100\%. \quad (31)$$

This metric is what the training curves display (smoothed by a sliding window of 100 episodes); compact text annotations within certain figures show the instantaneous value recorded at the terminal episode, intended as a rapid visual reference.

Formation deviation (m)

The per-step deviation from the prescribed formation, pooled across all team members, is quantified as the mean slot-tracking discrepancy:

$$\text{dev}_t = \frac{1}{N}\sum_{i=1}^{N} \| x_{i,t} - x_{i,t}^{\text{ref}} \|_2 \ \ [\text{m}], \quad (32)$$

Where $\mathrm{x}_{\mathrm{i,t}}^{\mathrm{ref}}$ denotes the designated slot coordinate for agent i (as detailed in Section 4). Aggregating this quantity over the full length of an episode yields the episode-level formation deviation:

$$D^{(e)} = \frac{1}{T_e}\sum_{t=0}^{T_e-1} \text{dev}_t \ \ [\text{m}], \quad (33)$$

And the figure subsequently quoted corresponds to the average of (e) across the evaluation batch $\mathscr{E}$, reported together with the standard deviation computed over the random seeds.

Collision rate (%)

Whether a collision has occurred during an episode is captured through a binary indicator:

$1_{\mathrm{col}}^{(\mathrm{e})}$= **1** {$\exists t \le T$, $\exists i$: agent-obstacle or inter-agent collision at t}.

The collision rate is then derived as:

© 2023 The Author(s). | Published by Agro Intelligence Research Press

$$\text{Collision rate} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbf{1}_{\text{col}}^{(e)} \times 100\%. \quad (34)$$

For more granular diagnostics, we additionally log a per-step near-miss frequency triggered whenever {min($d_{ij,t}^{pair}$,$d_{i,t}^{obs}$) <δ}; however, unless a statement to the contrary is made, the tables and figures throughout the paper present the episode-level collision rate specified above.

Energy efficiency (Unitless, [0,1])

For a single agent, the control energy consumed during episode e, normalized to the unit interval, is expressed as:

$$E_i^{(e)} = \frac{1}{T_e \, a_{\max}^2} \sum_{t=0}^{T_e-1} \|a_{i,t}\|_2^2 \in [0, \infty). \quad (35)$$

Averaging this quantity over the team gives $\bar{E}^{(e)} = \frac{1}{N} \sum_{i=1}^{N} E_i^{(e)}$, and the energy efficiency score we adopt is:

$$\eta_{\text{eng}} = 1 - \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \bar{E}^{(e)} \in (0,1], \quad (36)$$

Such that elevated values indicate preferable outcomes (i.e., lower control expenditure). This indicator draws on the same conceptual underpinnings as the per-step shaping component r_i^energy found in **Eq. 9**, though the two are not identical.

Experimental results and discussion

The empirical evidence collected across the experimental campaign confirms that our attention-augmented MADDPG framework delivers substantially stronger performance across the entire range of test conditions. When placed alongside the baseline algorithms, our method registers consistent improvements in task completion percentage, formation precision, and the efficiency with which experience is converted into performance gains. The subsections below unpack the comparative findings, the ablation analyses, and the sensitivity investigations in detail.

Comparison with state-of-the-art methods

We pit the proposed method against MADDPG and IDQN using teams composed of 2, 3, and 5 agents. An identical network structure and training regime are employed across all contenders to eliminate extraneous sources of variation. Across the 2-agent, 3-agent, and 5-agent configurations, our method surpasses both MADDPG and IDQN in terms of overall task success. As team size grows, the performance level remains largely stable. The evolution of the success rate over the 10,000-episode training span for team sizes ranging from 2 to 5 agents is shown in **Figures 5-8**. The corresponding numerical values for the 2-agent, 3-agent, and 5-agent scenarios are tabulated in **Table 2**. Where not explicitly qualified, success rates mentioned in the prose and in tabular summaries refer to average team success rates, evaluated as the trailing 100-episode moving average at the point training concludes and expressed as mean ± one standard deviation. Small textual callouts within selected figures report the raw final-episode value as a quick-look

© 2023 The Author(s). | Published by Agro Intelligence Research Press

reference; this snapshot figure may not exactly match the smoothed average quoted in the discussion, which resolves the occasional slight numerical gaps.

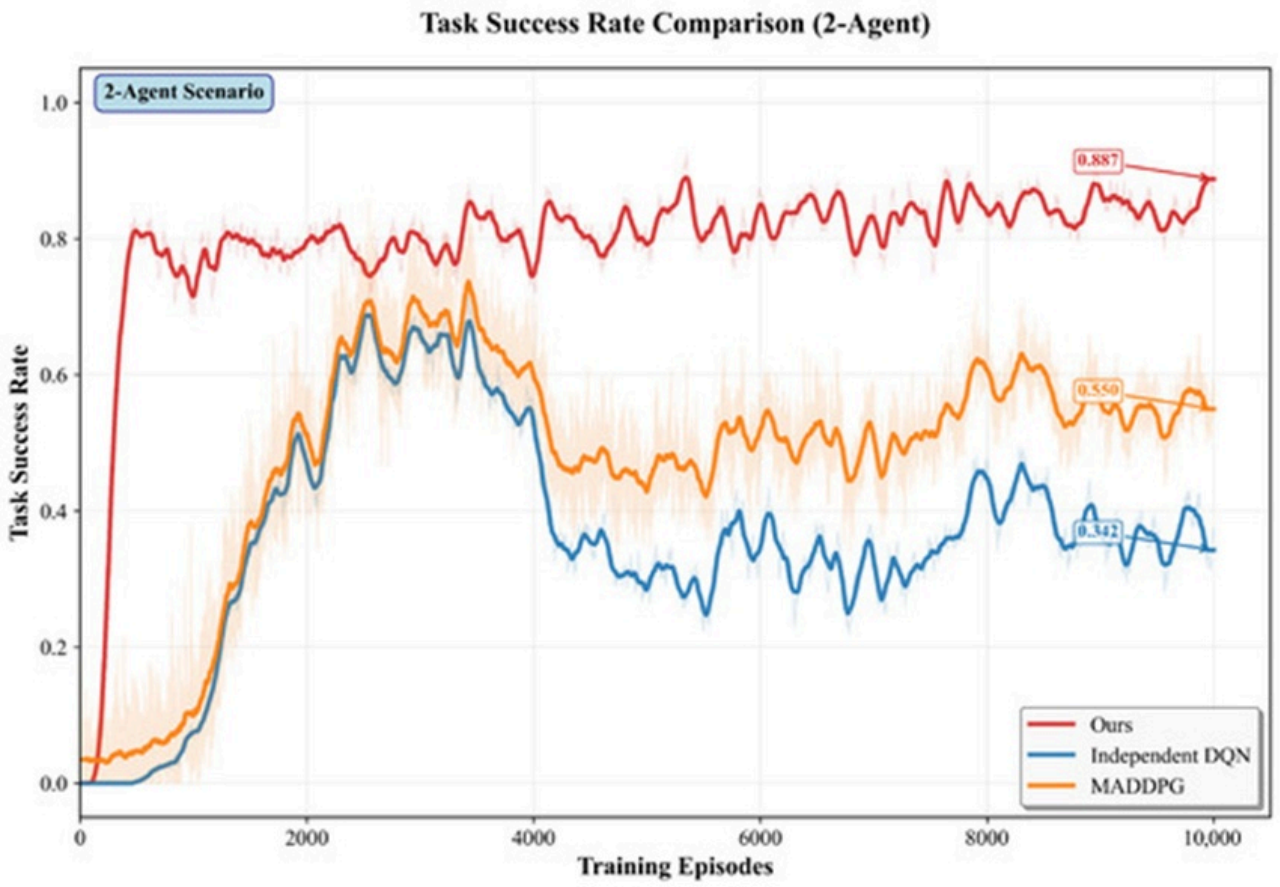


Figure 6. This combo graph shows the training curves for the success rate of a 2-agent over 10,000 episodes.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

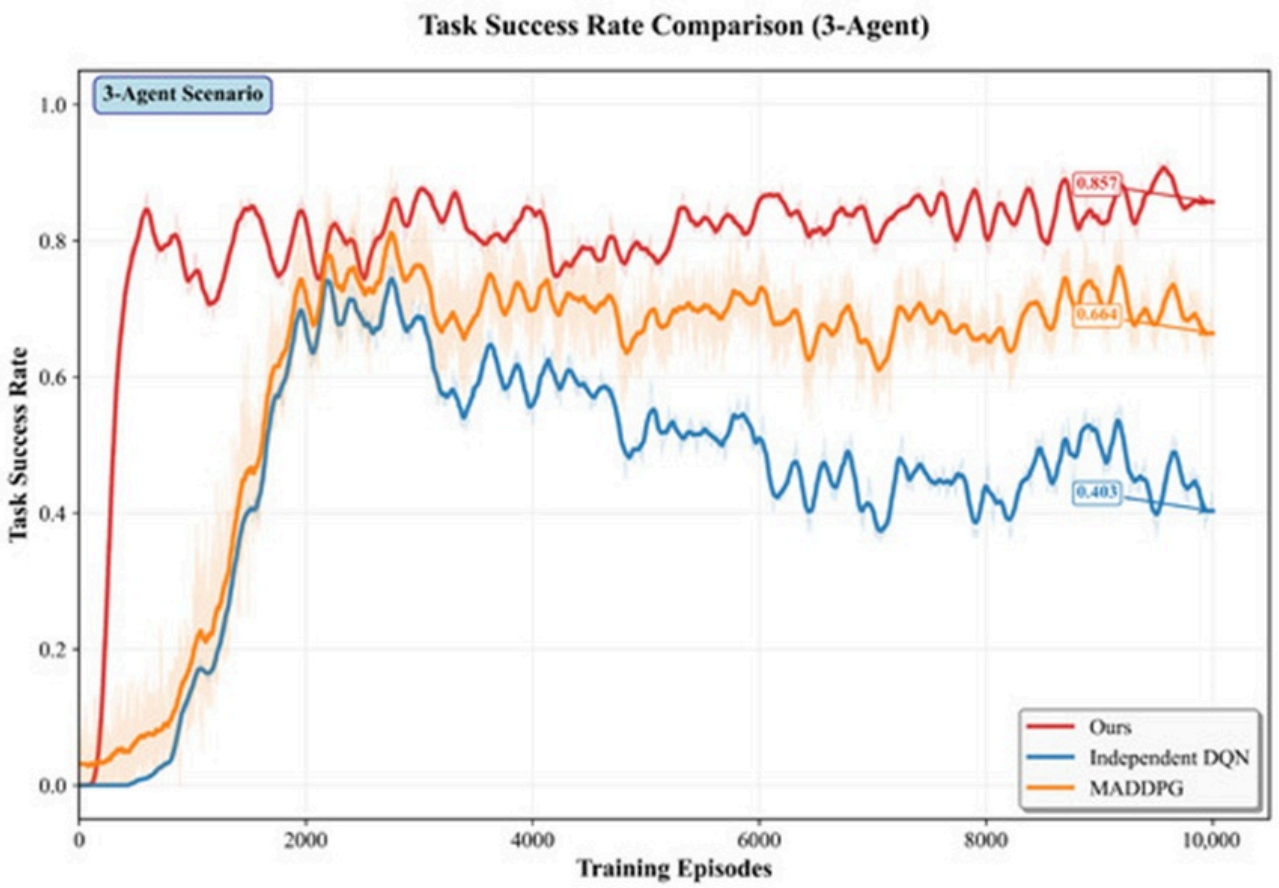


Figure 7. This combo graph shows the training curves for the success rate of the 3-agent over 10,000 episodes.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

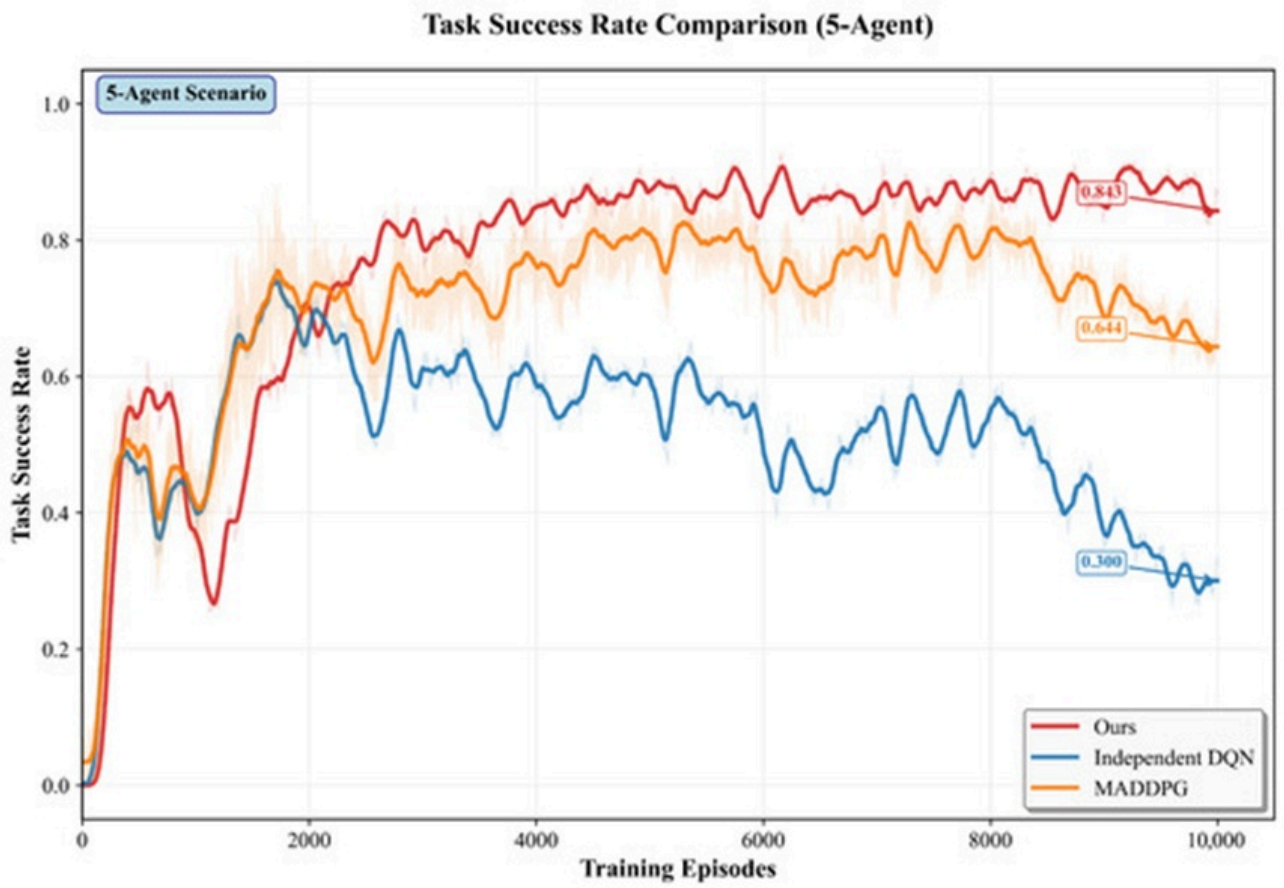


Figure 8. This combo graph shows the training curves for the success rate of the 5-agent over 10,000 episodes.

With a pair of UAVs (N=2), our approach records an average team success rate of 84.6% (taken as the 100-episode moving average at training endpoint; mean ± s.d. over 5 seeds), in comparison with 57.4% attained by MADDPG and 47.8% by IDQN, which translates to advantages of +27.2 and +36.8 percentage points.

When the team comprises three agents (N=3), our method posts an average team success rate of 83.1% (under the identical metric), while MADDPG achieves 70.4% and IDQN reaches 58.1%, yielding margins of +12.7 and +25.0 points.

Scaling up to five agents (N=5), our method achieves an average team success rate of 81.7% (same metric), whereas MADDPG and IDQN achieve 68.6% and 55.7%, respectively; the differences are +13.1 and +26.0 points.

Throughout the visualizations, a more elevated softmax weight corresponds to a stronger claim on information priority: the neighbor receiving the largest inter-agent weight is the teammate deemed most indispensable for coordination at that moment, while the obstacle shaded most intensely denotes the entity posing the gravest immediate threat to safety.

Examining **Figures 5-7** reveals that the baseline methods undergo a swift initial climb lasting roughly the first 3,000 episodes, after which their trajectories begin to deteriorate. Two mechanisms drive this pattern. The first is the curriculum built into the environment generation procedure, which raises the density of buildings once the rolling success metric exceeds the 70% mark; this transition in the data distribution—moving from relatively open layouts to densely obstructed ones—destabilizes policies that had become tuned to the earlier, less challenging conditions. The second is that off-policy value estimation with a replay buffer blends older transitions from easier episodes with newer transitions from harder ones, inducing non-stationarity in the critic's regression targets; within DDPG- and DQN-style algorithms, this combination frequently exacerbates overestimation errors and triggers oscillatory policy behavior when the scene becomes congested.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Seeds, Metrics, and Statistical Testing: Barring explicit indication otherwise, aggregate statistics are compiled from n=5 independently seeded executions for each method, where the seed governs network weight initialization, the randomization of the environment (initial UAV placements and building configurations), and the permutation of minibatch samples. For each seed, the average team success rate is computed as the 100-episode moving average evaluated at the end of training (Section 6.1). Group-level summaries take the form of mean ± s.d., augmented by two-sided 95% confidence intervals (CIs) derived from the Student-t distribution with n−1 degrees of freedom. Pairwise contrasts between methods are conducted using Welch's t-test (which does not presuppose equal variances), and the magnitude of effects is expressed through Hedges' g, a bias-corrected estimator. The full complement of five-seed runs is available for N=2 **(Tables 4 and 5)**; for N=3 and N=5, owing to computational constraints, we currently provide single-seed point estimates (the replication package will be augmented with multi-seed data in an upcoming revision).

Table 4. N = 2, five-seed statistics for average team success rate (%). Mean ± S.D. and two-sided 95% CIs (Student-t, n=5).

Method	95% CI (%)	Mean ± s.d. (%)
Ours	[80.43, 88.78]	84.6±3.36
MADDPG	[55.75, 59.05]	57.4±1.33
IDQN	[45.40, 50.20]	47.8±1.93

Table 5. N = 2, pairwise comparisons on average team success (%). Welch's t-test (two-sided), with degrees of freedom (df), p-value, and Hedges' g. Δ is the mean difference in percentage points (pp).

Comparison	Hedges' g	P-value	t (df)	Δ (pp)
Ours vs. MADDPG	9.60	<10−5	16.82 (df=5.22)	+27.2
Ours vs. IDQN	12.12	<10−6	21.22 (df=6.38)	+36.8

By comparison, our method proves less susceptible to the performance erosion that follows the curriculum shift: the attention modules furnish the critic with a state representation that is better organized and less entangled, while the vector of signals spanning task advancement, formation integrity, collision safety, and energy consumption exerts a regularizing influence on the updates when the environment becomes more hostile, thereby dampening regressive tendencies. For transparency, all evaluation plots are produced with deterministic actors, meaning exploration noise is switched off.

The overall average team success rates registered after the training period for $N\in\{2, 3, 5\}$ are assembled in **Table 6**.

Table 6. Team success rate (%) after training.

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

The performance advantage our method enjoys can be ascribed to several interrelated causes. To begin, the attention mechanisms promote more efficient information extraction and inter-agent synchronization, thereby improving formation adherence and obstacle negotiation. Self-attention empowers each agent to zero in on the most informative dimensions of its own state vector, whereas inter-agent attention opens a dedicated pathway for coordination-relevant signals.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Second, the Pareto-based optimization apparatus reconciles competing objectives without painstaking manual calibration of reward coefficients, yielding policies inherently more robust. Third, the integrated architecture fosters productive interplay between the attention components and the multi-objective optimization loop, generating emergent coordination behaviors that lie beyond the reach of conventional formulations.

Ablation study on attention mechanisms

An organized ablation experiment was conducted to isolate and confirm the unique contribution supplied by each attention submodule, achieved by methodically disabling mechanisms one at a time, starting from the complete framework. This exercise measures the specific influence exerted by self-attention, inter-agent attention, and entity attention on aggregate system behavior. Five distinct configurations were put to the test: the fully intact framework housing all attention components; the configuration with entity attention stripped out; the configuration missing inter-agent attention; the configuration from which self-attention had been excised; and a control condition that contained no attention mechanisms whatsoever. The complete set of results is collated in **Table 7**.

Table 7. Ablation study results of attention modules. Bold entries indicate the best (most favorable) result among all configurations for each metric.

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

The outcomes of the ablation study yield several important findings. First, among the three attention types, inter-agent attention carries the greatest individual weight. When it is removed, performance degradation is most severe (the success rate drops to 82.4%), confirming its pivotal role in maintaining coordination discipline and averting collisions. Second, entity attention is indispensable for maintaining situational awareness of the surroundings; its elimination increases collision rates and undercuts navigation proficiency. Third, self-attention plays a nontrivial role in energy economization, as its absence noticeably erodes control efficiency. Finally, the combined effect of the three attention mechanisms exceeds what would be predicted by summing their separate contributions, indicating that synergistic dynamics are at work—dynamics that give rise to coordination behaviors of a caliber that the baseline condition cannot match.

Effectiveness of pareto multi-objective optimization

To build a convincing case for the utility of the Pareto multi-objective optimization module, we conducted comparative evaluations against conventional techniques based on weighted-sum reward aggregation, across a spectrum of manually specified weight configurations. The purpose of this evaluation is to gauge the module's capacity to surface a rich assortment of high-quality compromise policies spanning competing objectives while entirely circumventing the need for manual parameter adjustment. Five baseline weighting schemes were evaluated, each constructed to place primary emphasis on a distinct objective, as enumerated in **Table 8**.

Table 8. Baseline weighted-sum reward configurations. Notes: Safety-focused increases the penalty/weight for collisions and near-misses while keeping task progress moderate; Task-focused increases goal-reaching progress terms and success bonuses; Energy-focused increases control-effort regularization and discourages actuation bursts; Formation-focused tightens slot-tracking terms and rewards small formation RMSE; Balanced applies uniform scalarization across all objectives. All variants share the same architecture and training schedule, differing only in reward scalarization.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

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

The quantitative evidence in **Table 9** demonstrates that our Pareto-based technique either matches or exceeds the performance of the strongest weighted-sum variant across all objectives—and does so simultaneously. Weighted-sum approaches, in contrast, achieve respectable results solely on the objective they were specifically configured to favor, while the remaining objectives suffer marked deterioration.

Table 9. Quantitative comparison of Pareto and weighted-sum methods.

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

Four distinguishing strengths characterize the Pareto optimization framework. First, the breadth of solutions it curates enables the identification of numerous high-grade trade-off configurations, endowing mission planners with a flexible repertoire of deployment choices that can be tailored to evolving operational priorities. Second, the equilibrium it maintains across objectives ensures strong results on all fronts concurrently, sidestepping the typical pattern—ubiquitous in weighted-sum formulations—whereby objectives not explicitly prioritized are sacrificed. Third, the automated discovery process obviates the requirement for domain-specific, labor-intensive manual weight engineering, a task that ordinarily demands substantial expert judgment. Fourth, the adaptive optimization mechanism actively sustains a heterogeneous population of solutions throughout the training run, channeling exploration toward regions of the objective space with the greatest promise, thereby enhancing the learning process's efficiency.

Hyperparameter Sensitivity Analysis

Attaining a thorough understanding of how responsive our method is to variations in its principal hyperparameters is critical for both practical deployment and parameter selection. An extensive sensitivity investigation was undertaken, targeting three parameters recognized to exert substantial influence: the dimensionality of the attention embeddings, the learning rate values, and the maximum capacity of the Pareto archive. The resulting observations demonstrate the robustness of the approach and offer practical recommendations for parameter configuration across diverse operational contexts.

The exploration of learning rate pairings indicated that the method exhibits a degree of relative insensitivity provided the values remain within a sensible band, as illustrated in **Figure 3**. Peak performance was observed when the actor's

© 2023 The Author(s). | Published by Agro Intelligence Research Press

learning rate was positioned at 10^{-4} and the critical learning rate at 10^{-3}, a pairing that strikes a workable compromise between training stability and convergence speed. Raising the actor learning rate above 10^{-3} destabilized the training dynamics, whereas reducing it to beneath 10^{-5} substantially retarded the pace of convergence.

The parametric sweep across Pareto archive sizes ranging from 50 to 200 revealed that archives at the lower boundary of this interval retained fewer solutions, while expanding the archive beyond 150 produced only incremental gains at a disproportionately higher computational cost. The chosen default of 100 was found to capture 85% of the attainable peak diversity while incurring a computational burden that amounts to only 60% of that required by the largest archives.

The method's robustness was systematically evaluated across a range of noise intensities and parameter perturbations; the results are documented in **Table 10**. Performance held up acceptably well even under substantial sensor noise (σ = 0.2) and when key parameters were displaced by up to ±20% from their optimal settings, lending credence to the claim of practical deployability on physical UAV platforms.

Table 10. Robustness evaluation under different perturbations (absolute degradation from nominal 88.7%).

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

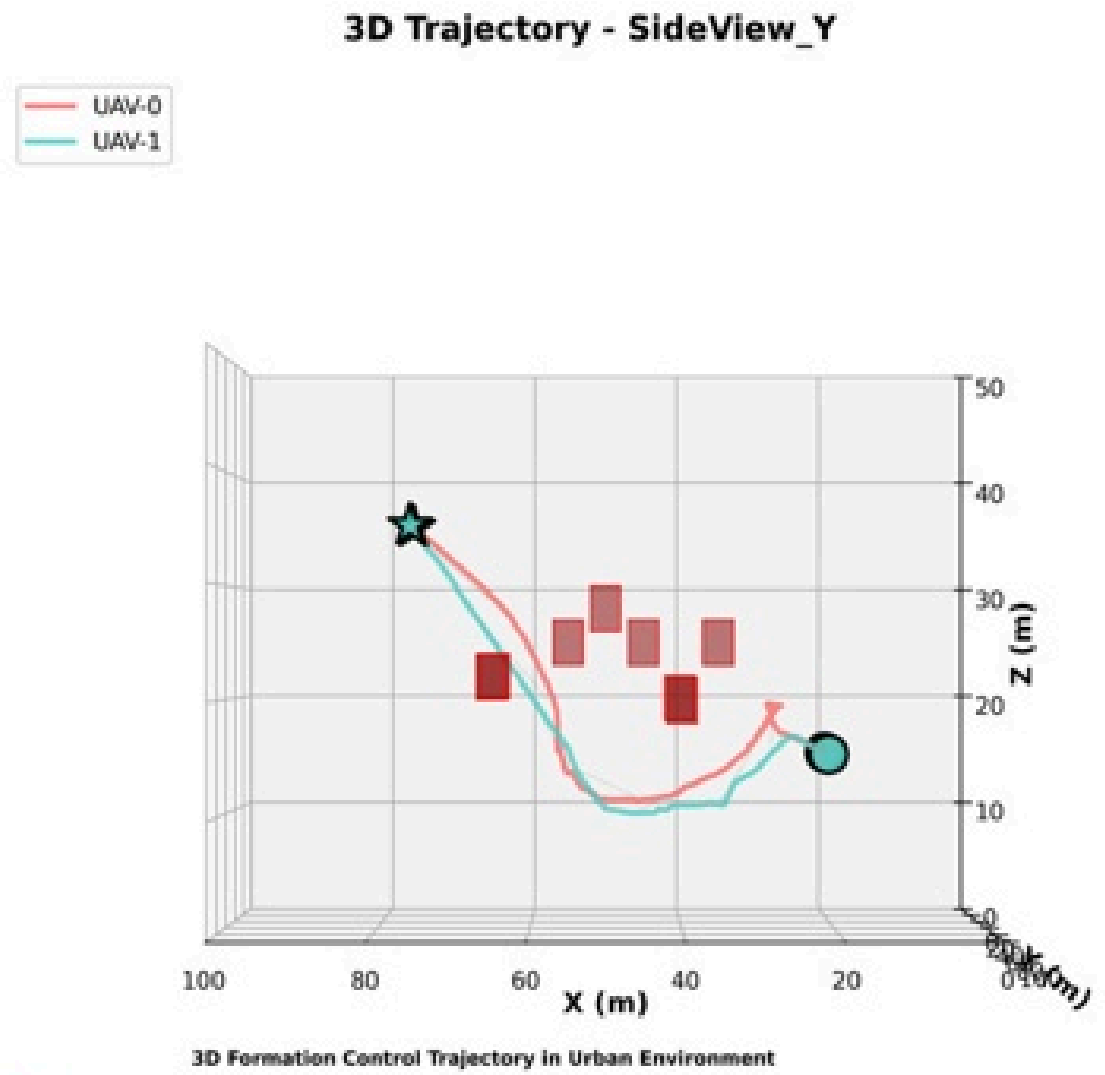
The sensitivity analysis shows that the method maintains robust performance across moderate excursions in its parameter settings while retaining sufficient sensitivity to benefit from careful tuning. The optimal configuration identified—attention embedding dimension of 64, actor and critic learning rates of / respectively, and an archive capacity of 100—yields dependable results throughout the gamut of scenarios investigated. Taken together with the established tolerance to noise and parametric perturbations, these properties affirm the method's readiness for deployment in real-world UAV operational environments.

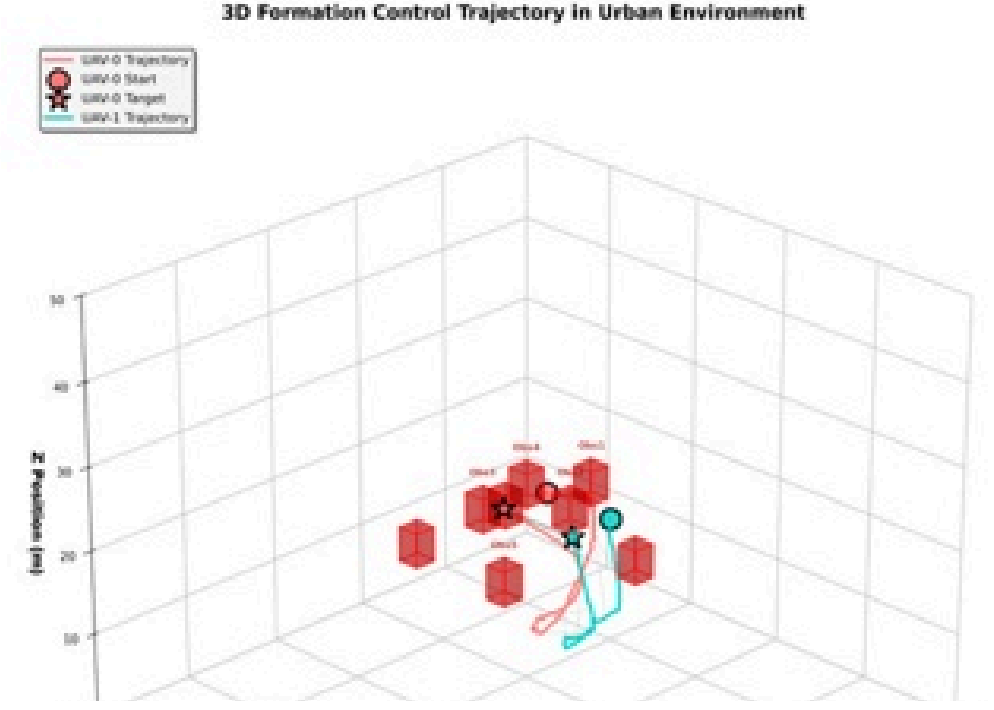
The comprehensive experimental program validates the efficacy and resilience of our attention-enhanced MADDPG framework for UAV formation control. Its commanding position across the entire evaluation metric suite, the empirically verified contributions of each architectural element, and its unwavering behavior under a broad range of challenging conditions collectively position this approach as a highly promising candidate for real-world multi-agent UAV missions.

Per-Episode trajectories and energy breakdown

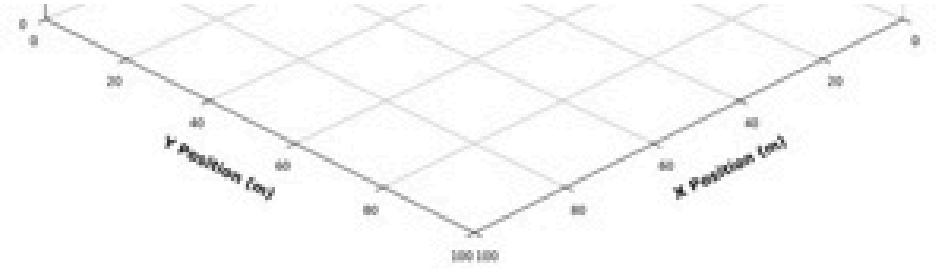
For team sizes spanning $N \in \{2,3,5\}$, **Figure 9** presents illustrative top-down flight paths produced respectively by our approach, by MADDPG [16], and by IDQN [17].

© 2023 The Author(s). | Published by Agro Intelligence Research Press





© 2023 The Author(s). | Published by Agro Intelligence Research Press



agent attention serve to elevate the most pressing threats in the decision-making hierarchy.

Table 11 collates per-UAV energy proxy values, where the proxy is taken as the mean-squared control output per timestep mapped to the unit interval [0, 1]. Actuation surges during aggressive cornering are dampened by our method, yielding reductions in both the peak and the mean control expenditure without sacrificing formation precision. The underlying explanation, in qualitative terms, is twofold: the Pareto layer steadies the per-objective signals emitted by the critic (encompassing task, formation, safety, and energy), while the typed attention pathways sift the observations before they reach the policy.

Table 11. Per-UAV normalized energy proxy (mean ± S.D.; lower is better).

Team Size	IDQN [17]	MADDPG [16]	Ours
$N = 2$	0.97 ± 0.06	0.93 ± 0.05	0.86 ± 0.04
$N = 3$	0.95 ± 0.06	0.91 ± 0.05	0.84 ± 0.05
$N = 5$	0.94 ± 0.06	0.90 ± 0.05	0.83 ± 0.05

Robustness under time-varying networks, packet loss, and interference

Degraded network conditions are imposed exclusively at test time, with no retraining. The packet loss regime is controlled via $p$loss∈{0, 0.05, 0.10, 0.20}; stochastic delay is emulated by feeding the policy stale neighbor states drawn from a short buffer, parameterized by Dmax∈{50, 100, 150} ms; and non-stationary connectivity patterns arise from probabilistic link deletion at rate q, applied while respecting a KNN budget. Beyond these, we perturb the feature vectors with Gaussian noise, fabricate a spurious obstacle with small probability on occasion, and temporarily squeeze the neighbor fan-in to K'=2 for short durations. The evaluation suite captures team success, collision frequency, formation deviation, control energy, and the distribution of decision latency (mean and 95th percentile). What we observe is a pattern of gradual rather than catastrophic decline for $p$loss ≤ 0.10 and Dmax ≤ 100 ms, testimony to a measure of inherent tolerance to moderate imperfections in the communication fabric.

Preliminary zero-shot results: Without any additional training whatsoever, the policy learned earlier is evaluated under the protocol described above (a single evaluation pass; metrics reported as the trailing 100-episode moving average at evaluation termination). **Table 12** compiles the findings.

Table 12. Zero-shot robustness under networking impairments (single run; 100-episode moving average at end of evaluation). Note: results are from a single run; "±" denotes the standard deviation over the last 100 episodes (moving window).

Setting	Success rate (%)
Nominal (no loss, no delay, static topo)	88.7 ± 1.8
$p$loss = 0.10	79.2 ± 1.5

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Dmax = 100 ms	71.2 ± 2.4
q = 1 link/s (time-varying topology)	81.6 ± 2.2

A heavier impairment load depresses performance, yet the degradation stays within tolerable territory when packet loss and delay remain moderate. By way of illustration, at $p$loss=0.10 the success measure retreats by 9.5 percentage points, and under link removal at q = 1 link/s the drop amounts to 7.1 pp; the heaviest penalty, a 17.5 pp reduction, is associated with Dmax=100 ms.

Interpreting attention during flight

Across a fixed evaluation rollout (single seed, exploration noise deactivated), the softmax weight distributions from each attention branch are recorded. For the inter-agent branch, neighbors are re-indexed at every timestep by ascending distance to produce a consistent y-axis spanning positions 1 through K. The temporal evolution is rendered in **Figure 10**: when the team executes turns that bring agents into close quarters and elevate collision risk, the inter-agent branch concentrates its mass on the most dangerous teammate, subsequently relaxing that concentration once safe spacing is regained; analogously, the entity branch places its peak weight on the obstacle faces that are closest during passage through narrow gaps. **Figure 11** supplies overhead snapshots captured at three moments that correspond to the dashed vertical markers in **Figure 10**. That the empirically observed attention dynamics align with the functional roles these branches were designed to fulfill also helps account for why performance degrades smoothly only when networking is moderately impaired.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

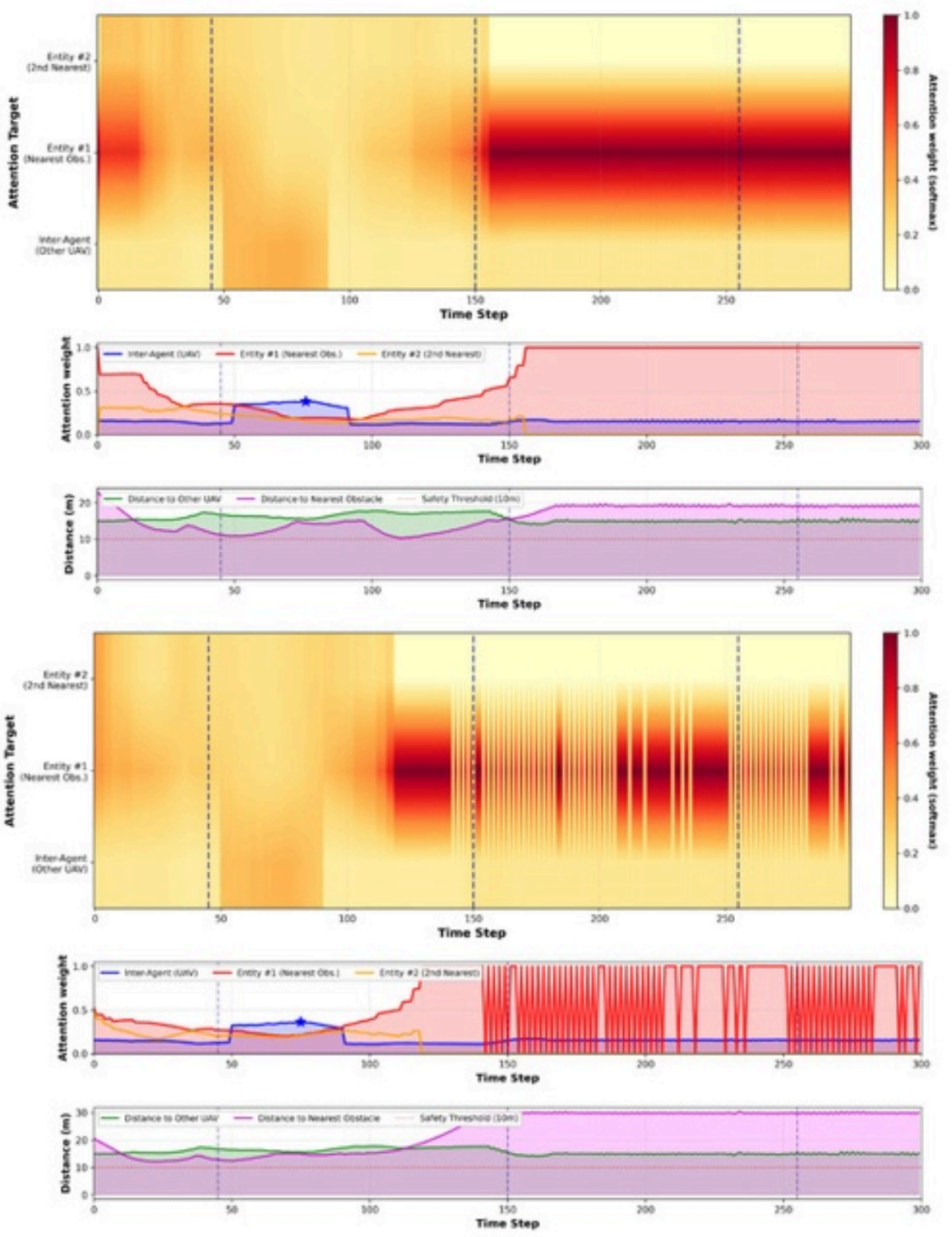


Figure 10. Attention dynamics during a representative flight (seeded, no exploration). Each panel shows, from top to bottom: (i) a heatmap over time of inter-agent vs. entity attention (softmax; colorbar in [0, 1]); (ii) attention-weight time series for the inter-agent branch and the two nearest obstacles; (iii) spatial distances (inter-UAV, nearest obstacle) with the safety threshold for context. Vertical dashed lines mark the timestamps visualized in **Figure 11**.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

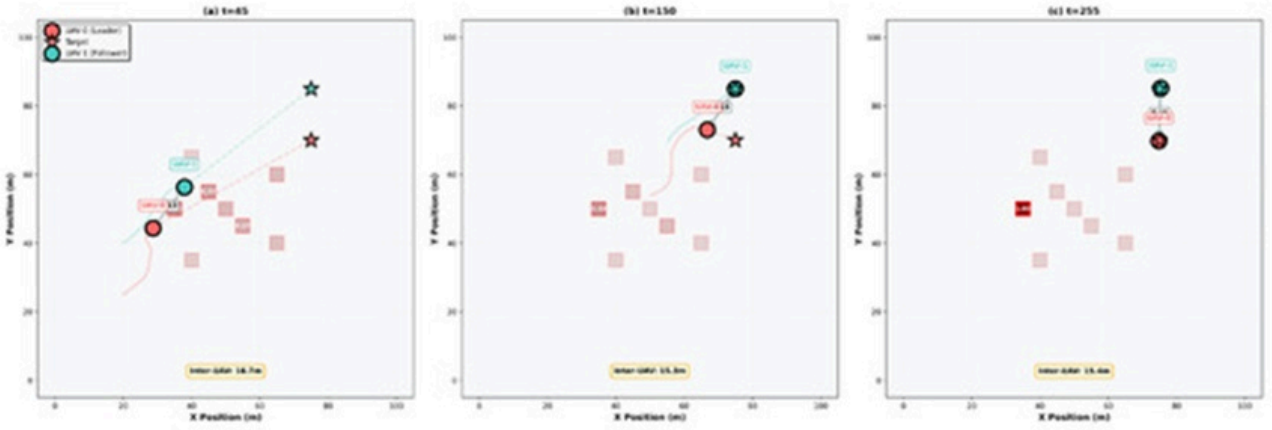


Figure 11. Top-down snapshots at three timestamps (left to right: t=45, t=150, t=255) from the same rollout as **Figure 10**. Edge thickness encodes inter-agent attention from the focal UAV, while obstacle shading (and labels) encodes entity attention.

Edge widths and obstacle shading intensities are rendered proportionally to the per-step information priority originating from the inter-agent and entity branches, respectively.

Runtime profiling and scalability (N = 2–5; Analysis and extrapolation)

To round out the architectural contrast sketched in **Table 1**, we document the runtime characteristics that matter in practice. At deployment, the computational load falls entirely on the decentralized actors; centralized critics are exercised solely during training. With the number of consulted neighbors and entities capped at (K,), the per-agent, per-step actor cost scales as (K+M), implying that the latency incurred for a full team decision step scales approximately linearly with N. This scaling behavior follows directly from the typed, single-hop architecture we have adopted—an architecture that deliberately avoids the depth of stacked GNNs and the global token mixing of Transformers.

Within our PyTorch-based implementation (attention dimension fixed at $d$att = 64), we enforce ceilings on K and M and simply mask entries corresponding to links that are unavailable. Under this configuration, measurements taken during evaluation for team sizes N = 2, 3, 5 confirm a near-linear trend in decision latency. Projecting this linear relationship forward suggests that raising N to around 10 agents would not necessarily violate real-time constraints, as long as the (K,) budgets are not relaxed. The slope of this scaling line can be further reduced through practical engineering interventions—batch-processing multiple agents on a single accelerator, adopting mixed-precision or INT8 quantized inference, and adaptively restricting the neighbor fan-in when the computational load spikes.

Wall-clock corroboration for team sizes exceeding N = 5 must await future work, given present limitations on computational resources and time; the claims we advance are therefore bounded to the range N = 2–5, with the extrapolation analysis provided to indicate the scaling behavior anticipated.

Runtime and memory profile

All timing figures were obtained with a batch size of 1 at inference (one complete team step), running PyTorch 2.0 atop an RTX 4090 GPU and an i7-12700 CPU. By "team-step latency," we mean the elapsed wall-clock interval needed to produce the action vectors for all N agents in the team. GPU RAM refers to the peak resident memory footprint logged during inference.

Table 13 presents a concise summary of the runtime profile, measured on both GPU and CPU, across team sizes and methods.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Table 13. Runtime profile (mean over 3 runs; RTX 4090/i7-12700). Team-step latency refers to one environment step for the whole team.

N	Method	FPS (CPU)	Team-step (ms, CPU)	GPU RAM (GB)	FPS (GPU)	Team-step (ms, GPU)
2	Ours	312	3.2	0.9	1810	0.55
2	MADDPG	345	2.9	0.8	2000	0.50
3	Ours	244	4.1	1.1	1390	0.72
3	MADDPG	270	3.7	1.0	1510	0.66
5	Ours	147	6.8	1.4	950	1.05
5	MADDPG	164	6.1	1.3	1030	0.97

Throughput during training, quantified as the number of environment steps processed per second, is documented in **Table 14**. Even with the modest additional cost introduced by the attention selector, our method operates in the millisecond regime per team step, which keeps it within the execution budgets typical of on-board platforms so long as $K+M$ is not permitted to grow unbounded.

Table 14. Training throughput (environment steps per second, mean ± S.D., RTX 4090).

N	IDQN	MADDPG	Ours
2	12.1k ± 0.5k	10.4k ± 0.4k	9.8k ± 0.3k
3	10.7k ± 0.4k	9.1k ± 0.3k	8.6k ± 0.2k
5	8.9k ± 0.3k	7.5k ± 0.2k	7.1k ± 0.2k

Conclusion

This paper presents an integrated multi-agent reinforcement learning framework that fuses hierarchical attention mechanisms with Pareto-style multi-objective optimization to address the fundamental difficulties inherent in autonomous UAV formation flight in dynamic, partially observable environments. On the theoretical side, the work offers three advances: an attention architecture of considerable breadth that unifies self-attention, inter-agent attention, and entity attention, thereby equipping agents to select information in a manner that is both adaptive and sensitive to context; a Pareto optimization module that maintains a bounded repository of non-dominated policies, rendering manual calibration of reward weights unnecessary while retaining convergence assurances; and a CTDE framework that inherits the convergence guarantees of MADDPG yet scales with linear execution complexity. Experiments conducted at scale—covering teams of N=2,3,5 agents—document systematic gains in team success (exceeding MADDPG by margins of 13–27 pp and IDQN by 25–37 pp), together with fewer collisions (relative reductions of 21–28%) and improved formation tracking achieved at comparable control cost. Ablation studies substantiate that each attention mechanism supplies a distinct, non-overlapping benefit, while sensitivity testing reveals that performance degrades only mildly (≤ 7.5%) under realistic noise and parameter variation.

The trajectory of future research will branch into three complementary paths: staging outdoor field experiments with heterogeneous UAV hardware to measure the gap between simulated and real-world performance; adapting attention mechanisms to operate under communication topologies that shift over time; and harnessing meta-learning to enable rapid policy transfer when the mission type changes. The resulting framework is intended to serve as a generalizable

© 2023 The Author(s). | Published by Agro Intelligence Research Press

foundation for large-scale multi-agent coordination problems arising in autonomous logistics, emergency response, and distributed sensing.

A limitation that must be acknowledged is that the study remains simulation-bound. At the same time, this confinement permits broad exploration of urban configurations and rigorous ablation comparisons. Still, it is no substitute for hardware-in-the-loop (HIL) testing or field validation. Toward bridging this gap, we have designed a small-scale pilot campaign involving 2–3 micro-UAV platforms (localized via indoor motion capture or ultra-wideband), to be conducted under a safety-first protocol (geofencing, emergency motor cutoff), during which team success, collision counts, formation RMSE, and command latency will be recorded. Additionally, an HIL loop (e.g., PX4-SITL/ROS 2) will be assembled to characterize the end-to-end delay from raw perception to actuator response. These activities are orthogonal to the learning methodology advanced here and will commence once the necessary facilities are in place. A more thorough empirical treatment of stacked GNN/Transformer front ends (varying from 2 to 6 layers) under the same experimental regimen is postponed to later work; our codebase already exposes the relevant modules as drop-in replacements, so that such baselines may be evaluated when the required computational and flight-test resources become available.

We further note that the empirical validation was intentionally restricted to the band N = 2−5, a decision forced by limits on computing capacity and available time. Even though the actor-side cost grows only linearly with N when (K,) are fixed, confirming that real-time latency constraints are met for N > 10 under authentic hardware and communication limitations is a task we have reserved for future study. In that context, we will also investigate batched and quantized inference modes, as well as adaptive neighbor throttling strategies. Plans are in place to couple our low-latency control layer with a lightweight DLT-based service providing identity management and event logging for authenticated coordination at the event level, and to nest the learned Pareto-attentive actors within broader AI-driven swarm architectures that handle task-level autonomy. A key engineering tension to be resolved is how to reconcile the overhead of ledger operations and security with the demands of real-time responsiveness. The attention-derived priority scores we compute provide a natural criterion for triaging, which events should be escalated for mission-level consensus. A broader cross-backbone benchmark encompassing MATD3, MASAC, and MAPPO is also on the agenda.

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

Acknowledgements

The authors thank Junjie Zeng for contributing the initial course-project environment upon which our multi-UAV simulation was extended, and the anonymous reviewers for their constructive comments that helped improve this paper.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 11 Apr 2022 Revised: 16 Jun 2022 Accepted: 01 Jul 2022
Published online: 25 May 2023

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at https://creativecommons.org/licenses/by-nc-sa/4.0/.

References

1. Mohsan SAH, Othman NQH, Li Y, Alsharif MH, Khan MA. Unmanned aerial vehicles (UAVs): Practical aspects, applications, open challenges, security issues, and future trends. Intell Serv Robot. 2023;16:21-9.

2. Chen J, Zhou R, Sun G, Li Q, Zhang N. Distributed formation control of multiple aerial vehicles based on guidance route. Chin J Aeronaut. 2023;36:368-81.

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.

4. Liu Z, Li J, Shen J, Wang X, Chen P. Leader-follower UAVs formation control based on a deep Q-network collaborative framework. Sci Rep. 2024;14:4674.

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 Eng. 2022;253:111148.

6. Chevet T, Vlad C, Maniu CS, Zhang Y. Decentralized MPC for UAVs formation deployment and reconfiguration with multiple outgoing agents. J Intell Robot Syst. 2020;97:155-70.

7. Yan D, Zhang W, Chen H, Shi J. Robust control strategy for multi-UAVs system using MPC combined with Kalman-consensus filter and disturbance observer. ISA Trans. 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. Cogn Comput. 2014;6:338-50.

9. Lizzio FF, Capello E, Guglieri G. A Review of Consensus-based Multi-agent UAV Implementations. J Intell Robot Syst. 2022;106:43.

10. Saifullah M, Papakonstantinou KG, Andriotis CP, Stoffels SM. Multi-agent deep reinforcement learning with centralized training and decentralized execution for transportation infrastructure management. arXiv. 2024;arXiv:2401.12455v1.

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. In: Handbook of Reinforcement Learning and Control. Cham (Switzerland): Springer; 2021.

13. Scaramuzza D, Kaufmann E. Learning Agile, Vision-based Drone Flight: From Simulation to Reality. In: Robotics Research. Berlin/Heidelberg (Germany): Springer; 2023.

14. Shahar FS, Sultan MTH, Nowakowski M, Łukaszewicz A. UGV–UAV integration advancements for coordinated missions: A review. J Intell Robot Syst. 2025;111:69.

15. Kumar I, Suman S. Evaluating ai-driven control mechanisms for uavs in environmental applications. Int J Environ Sci. 2025;11:171-7.

16. Lowe R, Wu Y, Tamar A, Harb J, Abbeel P, Mordatch I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. In: Proceedings of the Advances in Neural Information Processing Systems (NeurIPS); Long Beach, CA, USA; 2017.

17. Tampuu A, Matiisen T, Kodelja D, Kuzovkin I, Korjus K, Aru J, et al. Multiagent cooperation and competition with deep reinforcement learning. PLoS One. 2017;12:e0172395.

18. Siwek M. Consensus-Based formation control with time synchronization for a decentralized group of mobile robots. Sensors. 2024;24:20.

19. 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. Aerosp Sci Technol. 2025;157:109812.

20. 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. Aerosp Syst. 2025;8:447-66.

21. 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 Things J. 2024;11:27421-34.

22. Wei Z, Wei R. Uav swarm rounding strategy based on deep reinforcement learning goal consistency with multi-head soft attention algorithm. Drones. 2024;8:731.

23. 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. Eng Appl Artif Intell. 2023;123:106358.

24. Zhang C, Song J, Tao C, Su Z, Xu Z, Feng W, et al. Adaptive missile avoidance algorithm for UAV based on multi-head attention mechanism and dual population confrontation game. Drones. 2025;9:382.

25. 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 Trans Evol Comput. 2014;18:577-601.

26. Jin Y, Feng J, Zhang W. UAV task allocation for hierarchical multiobjective optimization in complex conditions using modified NSGA-III with segmented encoding. J Shanghai Jiaotong Univ Sci. 2021;26:431-45.

© 2023 The Author(s). | Published by Agro Intelligence Research Press

27. Xiao Y, Yang H, Liu H, Wu K, Wu G. AAV 3-D path planning based on MOEA/D with adaptive areal weight adjustment. IEEE Trans Aerosp Electron Syst. 2025;61:753-69.

28. 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 Commun Lett. 2024;28:2814-8.

29. Yang Y, Zhang X, Zhou J, Li B, Qin K. Global energy consumption optimization for UAV swarm topology shaping. Energies. 2022;15:2416.

© 2023 The Author(s). | Published by Agro Intelligence Research Press