

Explainable graph-based meta-reinforcement learning for cognitive UAV swarms under partial observability and adaptive jamming

Arshad Ali

Islamic University of Madinah

Muhammad Shaharyar Ramzan

Superior University

Eraj Khan

ekhan@hct.ac.ae

Higher College of Technology

Adeel Ahmad

Islamic University of Madinah

Gahangir Hossain

University of North Texas

Article

Keywords: Multi-agent reinforcement learning, Meta-learning, Graph neural networks, Explainable artificial intelligence, Unmanned aerial vehicles, Adaptive jamming

Posted Date: October 9th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11088012/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Explainable graph-based meta-reinforcement learning for cognitive UAV swarms under partial observability and adaptive jamming

Arshad Ali¹, Muhammad Shaharyar Ramzan³, Eraj Khan^{2*},
Adeel Ahmad¹, Gahangir Hossain⁴

¹Faculty of Computer and Information Systems, Islamic University of
Madinah, Madinah, Saudi Arabia.

^{2*}Higher College of Technology, Abu Dhabi, United Arab Emirates.

³Department of Information Technology, Faculty of Computer Science
and IT, Superior University, Lahore, Pakistan.

⁴Anuradha and Vikas Sinha Department of Data Science, University of
North Texas, Denton, TX, USA.

*Corresponding author(s). E-mail(s): ekhan@hct.ac.ae;

Contributing authors: a.ali@iu.edu.sa; shaharyarramzan56@gmail.com;
443057803@stu.iu.edu.sa; gahangir.hossain@unt.edu;

Abstract

Multi-agent deep reinforcement learning frameworks for cognitive unmanned aerial vehicle (UAV) swarms increasingly bundle graph coordination, meta-learning and explainability, yet it is rarely shown when that machinery actually helps. We present XM-MADRL, an explainable meta-learning graph framework for UAV swarms in contested electromagnetic environments, together with a controlled study that isolates when its components pay off. Under full observability against a predictable jammer, coordination and meta-adaptation give no measurable benefit: the framework merely ties a vanilla proximal policy optimisation (PPO) baseline. Under partial observability against an adaptive, reactive jammer, the same framework is better on every spectrum metric, raising packet delivery ratio sixfold (16.8% versus 2.5%; Cohen’s d of 4.45), anti-jamming from approximately 0% to 11.1%, and mean signal-to-interference-plus-noise ratio from -3.7 to -1.3 dB, all significant after Holm–Bonferroni correction (adjusted P below 0.0001) over ten seeds. A recurrent control that adds memory but not communication fails as badly as PPO, and a component ablation shows the gain is

not due to graph message passing. Meta-learning improves zero-shot generalisation to unseen jammers (packet delivery 32.0% versus 4.2%). Coordination and meta-adaptation are conditional capabilities, not universal improvements.

Keywords: Multi-agent reinforcement learning, Meta-learning, Graph neural networks, Explainable artificial intelligence, Unmanned aerial vehicles, Adaptive jamming

1 Introduction

Unmanned aerial vehicles (UAVs) have become a transformative technology for surveillance, reconnaissance, spectrum monitoring and electronic warfare [1, 2]. Cognitive UAV networks are expected to sense their environment, learn from experience and adapt autonomously in dynamic, adversarial, spectrum-congested conditions [3]. Deep reinforcement learning has emerged as a powerful tool for such autonomy, with strong results in navigation [1, 4], dynamic spectrum access [5, 6] and anti-jamming [7, 8].

Four challenges nonetheless remain only partially addressed. Many deep reinforcement learning controllers degrade as the swarm grows and inter-agent communication becomes essential [9]. Policies trained in one setting transfer poorly to unseen jamming patterns and mission conditions, so expensive retraining is required [10]. Many anti-jamming methods assume simplified, static adversaries [8]. Finally, learned policies are opaque, which undermines trust in mission-critical operations [11]. These challenges are usually studied in isolation, whereas a deployable cognitive UAV swarm must satisfy them jointly. Addressing them jointly is non-trivial because the underlying objectives interact and even conflict: aggressive spectrum exploration improves throughput but drains energy; tight formation aids coordination but raises collision risk; and larger, more expressive models improve capacity but are harder to train and to interpret.

Research on intelligent UAV systems spans four largely separate strands. *Navigation and control.* Classical model-based designs struggle with the nonlinear, uncertain dynamics of real flight, so learned controllers have become central: Choi *et al.* [1] decompose flight control into modular sub-policies; Xu *et al.* [4] fuse raw sensor streams with visual features in a feature-fusion PPO controller that exceeds 90% success in navigation without satellite positioning; and Fu *et al.* [12] apply reinforcement learning to visual servoing under field-of-view constraints. These methods advance single-agent competence but do not model inter-agent communication, meta-level adaptation or interpretability. *Cognitive spectrum access and anti-jamming.* Dong *et al.* [5] and Wang *et al.* [6] learn continuous spectrum-access policies with actor-critic methods, and Liu *et al.* [13] use hierarchical reinforcement learning for efficient multichannel sensing. For multiple UAVs, Jiang *et al.* [3] show that decentralised multi-agent learning approaches centralised performance for cooperative sensing and access. On the adversarial side, Chen *et al.* [8] propose a distributed actor-critic anti-jamming scheme and Zheng *et al.* [7] design anti-jamming radar waveforms. A recurring limitation is the assumption of a static or simplified jammer, and the spectrum objective is typically optimised in isolation from navigation and mission execution. *Swarm coordination*

and mission planning. Liu *et al.* [9] extend multi-agent deterministic policy gradients with local communication for task assignment, Xu *et al.* [2] address multi-target electronic reconnaissance under threat, Souto *et al.* [14] plan energy-efficient urban paths with tabular methods that scale poorly in high-dimensional state spaces, and Hieu *et al.* [10] study transferable policies for joint radar–communication without integrating them into cooperative swarm control. *Signal understanding.* Situational awareness depends on recognising emissions from few samples: Wu *et al.* [15] improve few-shot radar-emitter identification with a relation network, and O’Shea *et al.* [16] established deep modulation recognition on over-the-air data. Learned models also improve UAV channel estimation [17] and small-UAV detection [18], while energy-optimisation surveys highlight efficiency as a first-class concern [19]. Meta-learning [20] is promising for rapid adaptation but is rarely coupled to multi-agent decision-making. Supplementary Table S1 positions the present work against representative studies from these four strands.

Two further lines are directly relevant. A long tradition in multi-agent learning lets agents learn *what* to transmit rather than hand-designing messages, from broadcast averaging in CommNet [21] to attention-addressed messages in TarMAC [22] and learned gating of *when* to speak in IC3Net [23]. Coordination-graph methods learn *which* agents should be connected rather than assuming a fixed graph [24], and graph-attention policies jointly optimise UAV trajectories and resource assignment [25]. These methods are typically evaluated without an adaptive adversary, so the conditions under which communication is worth its cost remain largely untested. Separately, recurrence is the standard remedy for partial observability, from deep recurrent Q-networks [26] onward, and centralised training with decentralised execution underpins modern actor–critic multi-agent learning [27, 28]. Memory and communication are, however, distinct mechanisms: memory integrates an agent’s *own* history, whereas communication imports *other* agents’ observations. Existing studies rarely separate the two. Meanwhile, learning-based anti-jamming has moved toward intelligent, reactive adversaries that sense the spectrum and concentrate power where it hurts most [29], and explainable reinforcement learning has matured into a substantial literature [30] alongside cautionary evidence that visual attribution can mislead if it is not validated [31]. Contemporary UAV-swarm studies combine energy harvesting with multi-agent anti-jamming communication [32], control a swarm under electronic warfare with a cognitive-intent dual-stream architecture [33], and apply hierarchical role-based learning to heterogeneous swarms for radiation-source localisation [34]. These works establish that multi-agent reinforcement learning is viable for UAV electronic-warfare tasks and report performance on their respective benchmarks.

The gap we address is that these lines are largely disjoint: graph coordination is rarely stress-tested against an adaptive jammer, and anti-jamming rarely uses multi-agent graph coordination under partial observability. Our aim is therefore complementary to proposing yet another architecture. We hold the framework fixed and vary the operating regime, in order to measure when graph communication and meta-adaptation contribute at all. A subtler difficulty, which we identify and resolve here, is that the very mechanism used for swarm coordination, namely graph message passing, can degrade a capability it is not meant to touch. Naive graph aggregation

over-smooths the per-agent radio-frequency features required for channel selection, a phenomenon the graph-learning community has formalised and quantified through the normalised Dirichlet energy [35, 36], and a decoupled policy resolves it.

We propose XM-MADRL, an explainable meta-learning multi-agent framework that addresses these challenges in a single design (Fig. 1). Our contributions are fourfold. First, we integrate multi-modal transformer fusion [37], graph-convolutional swarm communication [38], multi-agent proximal policy optimisation [28], model-agnostic meta-learning [20] and Shapley-value explainability [11] into one cognitive-swarm controller. Second, we report an ablation-driven architectural insight: graph aggregation over-smooths the *local* spectrum-sensing features required for channel selection, and a *decoupled* policy with a coordinated navigation head and a local spectrum head restores communication performance without sacrificing coordination. Third, we release a fully reproducible multi-agent electronic-warfare benchmark in which every result is produced from logged runs with five or ten independent random seeds and full statistical testing. Fourth, we evaluate against PPO, advantage actor-critic (A2C), deep deterministic policy gradient (DDPG) and multi-agent DDPG (MADDPG) baselines, together with a component ablation, a scalability analysis, a modulation-recognition study and a Shapley-value interpretability analysis.

The study is organised around three falsifiable hypotheses, each tested by a dedicated experiment. **H1**: when agents have sufficient *local* information (full observability, predictable jammer), inter-agent coordination and meta-adaptation provide little benefit over a strong single-agent baseline. **H2**: under *partial* observability with an adaptive jammer, the framework provides a substantial and statistically significant benefit over single-agent baselines, and that benefit is not explained by recurrent memory alone. **H3**: meta-learning improves zero-shot generalisation and few-shot adaptation to *unseen* jammer behaviours.

2 Results

2.1 Framework and experimental design

XM-MADRL treats the swarm as a decentralised, partially observable multi-agent system (Fig. 1). Each UAV fuses radio-frequency, radar, visual and navigation observations with a transformer encoder, shares the fused representation with neighbours through a graph convolution, and acts through two heads: a navigation head driven by the coordinated representation and a channel-selection head driven by the agent’s own radio-frequency observations. Policies are trained with multi-agent proximal policy optimisation, meta-learned with first-order model-agnostic meta-learning, and explained post hoc with Shapley additive explanations. Full specifications are given in Methods.

We evaluate in two operating regimes that differ only in observability and in the sophistication of the adversary. The *control* regime gives every agent full observability and a predictable sweep jammer; it tests H1. The *hard* regime masks the radar and visual observation blocks and replaces the adversary with a reactive jammer that each step jams the channels the swarm is currently using; it tests H2 and H3. Every method, proposed or baseline, is trained for an identical budget of 6×10^5 environment steps

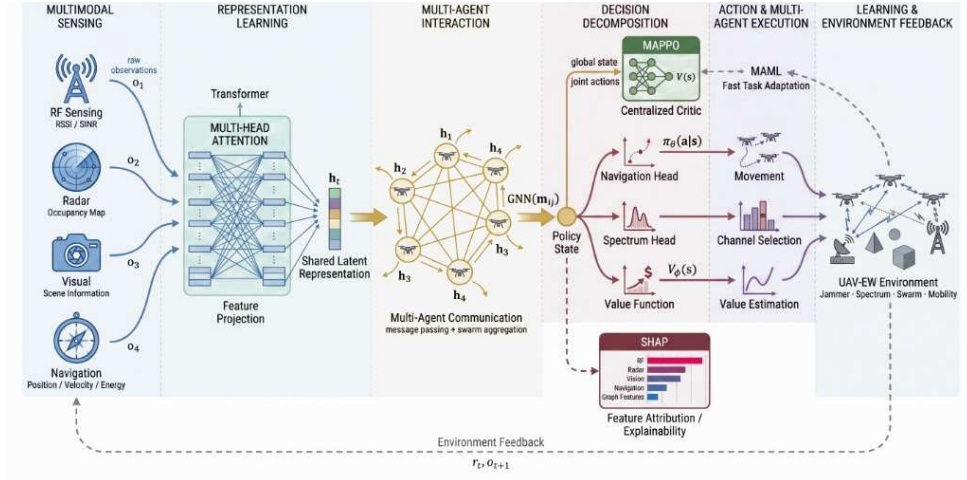


Fig. 1 The XM-MADRL framework. Multi-modal observations are fused by a transformer encoder and shared across the swarm by a graph convolutional network for coordinated navigation, while channel selection is driven by each agent’s *local* radio-frequency observations through a decoupled spectrum head (green “local-RF skip”). Policies are trained with multi-agent proximal policy optimisation, meta-learned with first-order model-agnostic meta-learning for fast adaptation, and explained with Shapley additive explanations.

and evaluated on 30 held-out episodes per seed. The control regime uses five seeds and the hard regime ten, except where a smaller number is stated in the relevant table. The control regime is not a weaker version of the main experiment: it is the condition that establishes when the framework’s machinery is *not* needed, and it is what makes the hard-regime result interpretable.

2.2 Coordination and meta-adaptation are inert under full observability

Under full observability with a predictable jammer, XM-MADRL is statistically indistinguishable from the strongest baseline on the metrics that matter for the mission (Table 1a). It ties PPO on mission success ($P = 0.912$, two-sided Welch t -test, $n = 5$ seeds per method), on energy use ($P = 0.873$) and on flight safety, where both methods reach zero collisions, and its packet-delivery-ratio ($P = 0.139$) and signal-to-interference-plus-noise-ratio ($P = 0.142$) differences are likewise not significant (Supplementary Table S5). It does not outperform vanilla PPO on this benchmark, and on anti-jamming it is in fact slightly but reliably lower (97.6% versus 98.3%; $P = 8.0 \times 10^{-4}$, $d = -3.65$), which we report as measured. The off-policy MADDPG baseline records a higher raw mission count only by flying aggressively, incurring hundreds of collisions and the worst energy use, and is therefore not a usable policy (Supplementary Table S4 and Supplementary Fig. S1). The swarm’s qualitative behaviour is nonetheless sound: a representative episode shows compact, collision-free manoeuvres toward the targets while keeping clear of the central jammer (Supplementary Fig. S2), XM-MADRL sustains a stable positive signal-to-interference-plus-noise ratio

Table 1 Performance in the two operating regimes. Mean $\pm$ sample standard deviation across n independent training seeds; each seed contributes the mean of 30 held-out evaluation episodes. Panel **a**, control regime (full observability, sweep jammer). Panel **b**, hard regime (partial observability, reactive jammer). PDR, packet delivery ratio; SINR, signal-to-interference-plus-noise ratio; XM-noGNN, XM-MADRL with graph message passing removed; GRU-PPO, recurrent single-agent control with memory but no inter-agent communication. Statistical comparisons for panel **b** are given in Table 2.

Method	n	Mission success (%)	PDR (%)	Anti-jamming (%)	SINR (dB)
a Control regime (full observability, sweep jammer)					
XM-MADRL	5	38.0 ± 4.3	55.1 ± 4.4	97.6 ± 0.2	3.6 ± 0.4
XM-noGNN	4	35.4 ± 4.9	30.0 ± 11.2	84.8 ± 5.6	1.0 ± 1.2
PPO	5	38.3 ± 4.9	58.7 ± 1.0	98.3 ± 0.1	3.9 ± 0.1
A2C	5	38.0 ± 5.3	58.3 ± 2.6	98.2 ± 0.1	3.9 ± 0.2
b Hard regime (partial observability, reactive jammer)					
XM-MADRL	10	39.2 ± 6.9	16.8 ± 4.0	11.1 ± 4.1	-1.3 ± 0.6
XM-noGNN	5	37.7 ± 5.7	19.3 ± 1.4	13.7 ± 1.7	-0.9 ± 0.2
PPO	10	37.2 ± 4.2	2.5 ± 2.1	0.0 ± 0.1	-3.7 ± 0.6
A2C	10	32.6 ± 3.2	4.1 ± 2.1	0.4 ± 0.5	-3.4 ± 0.6
GRU-PPO	6	34.2 ± 2.0	1.3 ± 0.0	0.0 ± 0.0	-4.1 ± 0.0

throughout an episode whereas the poorly coordinated MADDPG policy is repeatedly driven negative (Supplementary Fig. S3), and energy use is on par with the on-policy baselines and far below MADDPG (Supplementary Fig. S4). H1 is supported: when every agent can already sense its objective and the adversary is stationary, neither inter-agent coordination nor meta-adaptation has anything to exploit.

2.3 Coordination wins decisively under partial observability and adaptive jamming

The control-regime tie inverts in the hard regime (Table 1b and Fig. 2). XM-MADRL attains a packet delivery ratio of 16.8% against 2.5% for PPO, a sixfold improvement; an anti-jamming rate of 11.1% against approximately 0% for every baseline; and a mean signal-to-interference-plus-noise ratio of -1.3 dB against -3.7 dB. Table 2 establishes significance. The three communication and spectrum metrics, which are the framework’s design objective, are all significant after Holm–Bonferroni correction with very large effect sizes (two-sided Welch t -tests, $n = 10$ seeds per method; packet delivery ratio $P = 1.5 \times 10^{-7}$, Cohen’s $d = 4.45$; anti-jamming $P = 1.2 \times 10^{-5}$, $d = 3.85$; signal-to-interference-plus-noise ratio $P = 2.2 \times 10^{-8}$, $d = 4.21$). Mission success is nominally higher (39.2% versus 37.2%) but the difference is not statistically significant ($P = 0.444$, $d = 0.35$), so we make no claim on that metric. H2 is supported on the communication and spectrum objectives.

2.4 The gain is not recurrent memory, and not graph message passing

Under partial observability a memoryless agent is inherently handicapped, so the advantage might be attributed to recurrence rather than to coordination. The

Table 2 Statistical comparison in the hard regime. XM-MADRL versus PPO, $n = 10$ independent training seeds per method. Values are seed-level means. P values are from two-sided Welch t -tests on seed-level means; P_{Holm} is adjusted across the five-metric family by the Holm–Bonferroni step-down procedure at $\alpha = 0.05$; Cohen’s d uses the pooled standard deviation. The three communication and spectrum metrics remain significant after adjustment; mission success and energy do not. PDR, packet delivery ratio; SINR, signal-to-interference-plus-noise ratio.

Metric	XM-MADRL	PPO	P	P_{Holm}	Cohen’s d
Mission success (%)	39.2	37.2	0.444	0.868	0.35
PDR (%)	16.8	2.5	1.5×10^{-7}	5.8×10^{-7}	4.45
Anti-jamming (%)	11.1	0.02	1.2×10^{-5}	3.7×10^{-5}	3.85
Mean SINR (dB)	−1.28	−3.74	2.2×10^{-8}	1.1×10^{-7}	4.21
Energy used	1.33	1.19	0.434	0.868	0.36

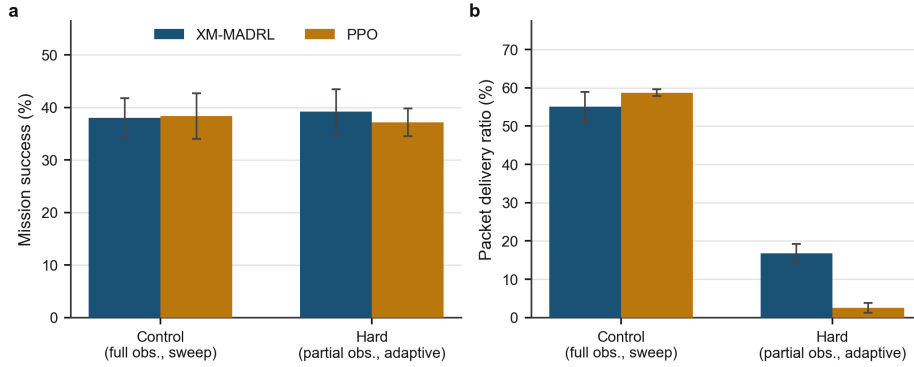


Fig. 2 The regime inversion. XM-MADRL ties PPO under full observability against a predictable sweep jammer and wins decisively under partial observability against an adaptive, reactive jammer. **a**, Mission success. **b**, Packet delivery ratio. Bars show means over independent training seeds ($n = 5$ in the control regime, $n = 10$ in the hard regime); error bars are 95% confidence intervals of the mean across seeds.

GRU-PPO control, a recurrent single-agent policy with memory but no inter-agent communication, rules this out: its packet delivery ratio (1.3%) and anti-jamming rate (0%) are indistinguishable from memoryless PPO (Table 1b, $n = 6$ seeds). Memory alone does not recover the missing state.

A component ablation locates the source of the gain more precisely, and in doing so corrects our initial hypothesis. In the control regime, the transformer fusion and the local channel head are the decisive components (Table 3 and Supplementary Fig. S5): removing the transformer collapses packet delivery, and removing the local channel head sharply degrades both packet delivery and anti-jamming, which directly validates the decoupled design. In the hard regime, removing graph message passing (XM-noGNN, $n = 5$ seeds) did *not* degrade performance; if anything it was marginally higher (packet delivery 19.3% versus 16.8%, anti-jamming 13.7% versus 11.1%, signal-to-interference-plus-noise ratio −0.9 dB versus −1.3 dB; no significant difference, $P \approx 0.10$). The advantage over PPO and GRU-PPO is therefore

Table 3 Component ablation in the control regime. Mean $\pm$ sample standard deviation over n independent training seeds, under the same protocol as Table 1a. The full model uses the decoupled channel head; the MAML, GNN and transformer rows remove one component from that architecture, and the “without local channel head” row is the earlier single-head architecture, which routes channel selection through the coordinated representation and therefore isolates the contribution of the decoupled design. PDR, packet delivery ratio; SINR, signal-to-interference-plus-noise ratio; MAML, model-agnostic meta-learning; GNN, graph neural network.

Configuration	n	Mission (%)	PDR (%)	Anti-jam. (%)	SINR (dB)
XM-MADRL (full)	5	38.0 ± 4.3	55.1 ± 4.4	97.6 ± 0.2	3.6 ± 0.4
without MAML	5	34.8 ± 3.3	38.4 ± 14.2	94.3 ± 5.5	1.8 ± 1.5
without GNN	4	35.4 ± 4.9	30.0 ± 11.2	84.8 ± 5.6	1.0 ± 1.2
without transformer	5	33.0 ± 1.5	10.6 ± 13.3	69.7 ± 5.2	-1.5 ± 1.6
without local channel head	5	35.5 ± 6.9	22.7 ± 9.5	83.1 ± 8.9	0.1 ± 1.2

attributable to the multi-modal fusion encoder and the decoupled spectrum head rather than to the communication graph. We report this disconfirming ablation rather than omit it, and we have revised the attribution to match the evidence.

2.5 Meta-learning improves generalisation to unseen jammers

We meta-train on one family of adversaries (sweep and static) and evaluate on a disjoint, held-out family (reactive and random). Figure 3 compares the meta-learned XM-MADRL initialisation against a jointly trained PPO fine-tuned from its own weights with an identical inner-loop optimiser; the underlying numbers are given in Supplementary Table S6. XM-MADRL leads on mission success at every number of inner adaptation steps K , and its zero-shot ($K = 0$) packet delivery ratio on unseen adversaries is 32.0% against 4.2% for PPO, a nearly eightfold generalisation gap. Once both policies are allowed to adapt, the packet-delivery advantage narrows and PPO catches up, so the meta-learning benefit is concentrated in the zero-shot and few-shot regime. H3 is supported for zero-shot generalisation and for few-shot mission success.

2.6 Decisions are attributable to the expected sensing modalities

Shapley additive explanations attribute the learned policy’s decisions to the six sensing modalities (Fig. 4). Channel-selection decisions are dominated by radio-frequency signal-to-interference-plus-noise ratio and received signal strength, which matches domain expectation and provides a transparent, auditable rationale for autonomous actions. Because the attribution module is analysis-only and does not affect training, removing it leaves task performance unchanged while eliminating decision transparency.

2.7 Supporting analyses

Three further analyses are reported in full in the Supplementary Information. Across swarm sizes of 6, 12 and 18 agents ($n = 4$ seeds per size), mission success improves

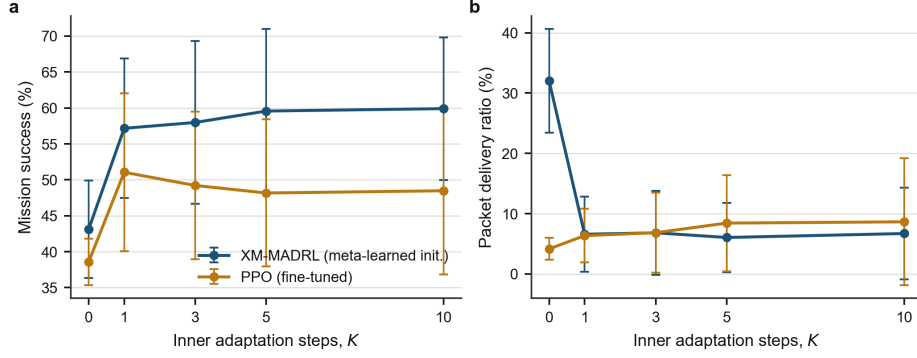


Fig. 3 Few-shot adaptation to unseen jammers. Meta-training draws adversaries only from the sweep and static families; evaluation draws only from the disjoint, held-out reactive and random families. **a**, Mission success. **b**, Packet delivery ratio. XM-MADRL, initialised by first-order model-agnostic meta-learning, adapts to a higher mission-success ceiling than PPO fine-tuning at every number of inner gradient steps K ; its packet-delivery advantage is concentrated at $K = 0$. Points are means over $n = 30$ evaluations (5 training seeds $\times$ 6 held-out tasks, each evaluated over 10 episodes); error bars are one sample standard deviation across those 30 evaluations.

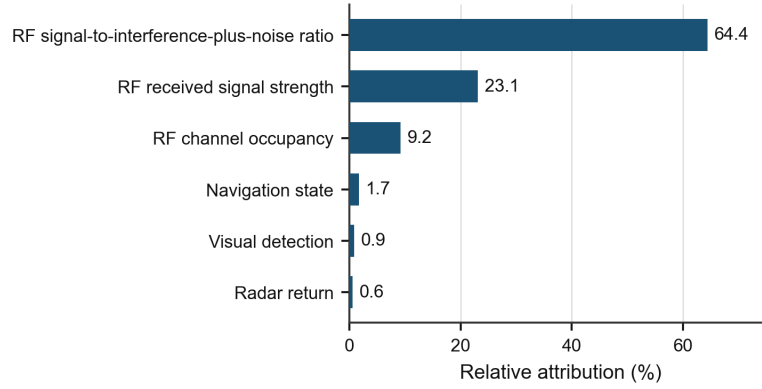


Fig. 4 Attribution of policy decisions to sensing modalities. Normalised Shapley feature-group attribution for the learned policy, computed with KernelSHAP over 40 held-out on-policy observations and averaged across instances. Per-feature absolute Shapley values are summed within each of the six sensing modalities and normalised to sum to 100%. RF, radio frequency.

with swarm size for both XM-MADRL and the strongest on-policy baseline, and XM-MADRL remains on par at every size, whereas communication reliability declines for both methods as more agents contend for a fixed set of six channels (Supplementary Fig. S6). An auxiliary sensing-validation experiment, which is not part of the reinforcement-learning decision loop, classifies six modulation types with 84% accuracy, with residual confusion concentrated in the dense 16-QAM and 64-QAM constellations as expected under noise (Supplementary Fig. S7). Finally, direct measurement of the normalised Dirichlet energy and pairwise cosine similarity of the

encoder features shows that per-agent features are already highly similar before aggregation under partial observability (cosine similarity ≈ 0.99), and become more similar with each additional aggregation layer (Supplementary Fig. S8), which is the empirical basis for driving channel selection from a local spectrum head.

3 Discussion

Synthesising the two regimes yields a practitioner-facing guideline. Graph coordination and meta-adaptation are not free: they add parameters, communication and training cost, and our control regime shows that they can be entirely inert. Their value materialises when observations are *partial*, so that agents must share information to recover missing state, and when the adversary is *non-stationary*, so that a fixed policy is insufficient and adaptation is rewarded. Systems operating only in benign, fully observed conditions gain little from this machinery; systems facing contested, partially observed spectra gain a great deal. Most reinforcement learning studies in this area report that a new architecture outperforms a standard baseline and stop there. The more useful question, and the one a practitioner actually faces, is when the added machinery earns the complexity it costs. Our answer is sharply conditional, and the condition is measurable in advance from the deployment environment rather than only after training.

Three points merit emphasis. The decoupled channel head is the single most impactful design choice for spectrum performance: it more than doubles packet delivery relative to routing channel decisions through the graph, confirming that local radio-frequency sensing must not be over-smoothed by neighbour aggregation. This is a methodological finding that transfers to any multi-agent system combining graph communication with local sensing, and it is consistent with the over-smoothing phenomenon formalised in the graph-learning literature [35, 36]. Second, a candid reading of the control regime is that on a fully observable cooperative task the sophisticated architecture *matches* rather than beats a well-tuned PPO; exposing this at equal budget is more useful than an inflated comparison. Third, the practical differentiator in that regime is interpretability: XM-MADRL matches the baselines while yielding auditable Shapley rationale that opaque policies cannot, which is decisive in mission-critical electronic-warfare deployments. These conditional advantages do translate into raw gains once the regime demands them.

Several limitations bound the scope of these findings. All evidence comes from a fast, reproducible simulator rather than a photorealistic or hardware-in-the-loop testbed; the propagation, jamming and mobility models are deliberately simple, so absolute values should not be read as predictions of field performance. Absolute performance under adaptive jamming remains low for every method we evaluated: XM-MADRL delivers 16.8% of packets where the strongest baseline delivers 2.5%, so the contribution is a large relative improvement over a failing baseline rather than a deployable communication system. The component ablations were run only in the control regime, where the machinery is inert; we do not report a full hard-regime ablation sweep because the non-meta variants are roughly an order of magnitude more expensive per run, although the hard-regime XM-noGNN control reported above is

the single most informative of them. The modulation-recognition study is an auxiliary sensing experiment that does not feed the reinforcement-learning loop. Finally, we study a single simulator and a single family of adversaries; whether the regime-conditional pattern reported here generalises to other environments, larger swarms, or adversaries that themselves learn remains open.

We also note three threats to validity and how they are addressed. To avoid an unfair-budget confound, every method is trained for an identical 6×10^5 -step budget and evaluated on held-out seeds, results are averaged over five to ten seeds with confidence intervals and significance tests, and all reported numbers are regenerated directly from logged runs, which eliminates transcription error. The metric set (mission success, packet delivery ratio, signal-to-interference-plus-noise ratio, anti-jamming, energy and collisions) jointly captures the multi-objective nature of the task rather than a single proxy, and the strongest baseline is used for pairwise significance testing. The environment is a simulation; while it models the salient dynamics of the domain, including fading, adaptive jamming, spectrum contention and navigation without satellite positioning, transfer to hardware is left to future work.

In summary, the central finding is regime-conditional. Under full observability with a predictable jammer, graph coordination and meta-adaptation are inert and the full framework merely ties a vanilla PPO baseline; we report that parity openly. Under partial observability with an adaptive jammer, the same framework is better on every communication and spectrum metric, raising packet delivery from 2.5% to 16.8%, anti-jamming from approximately 0% to 11.1%, and mean signal-to-interference-plus-noise ratio from -3.7 dB to -1.3 dB, all significant after Holm–Bonferroni correction over ten seeds. Two controls make the attribution specific: a recurrent single-agent policy with memory but no inter-agent communication fails as badly as the memoryless baseline, which rules out recurrence, and a component ablation shows that the gain is not attributable to graph message passing either. A held-out evaluation on adversary behaviours never seen during meta-training isolates the contribution of the meta-learned initialisation. Taken together, these results argue that coordination and meta-adaptation should be understood as conditional capabilities rather than universal improvements, and that reporting the regime in which a multi-agent architecture is evaluated is as important as reporting its score. Future work targets hardware-in-the-loop validation, larger swarms, and adversaries that adapt by learning rather than by fixed rule.

4 Methods

4.1 Problem formulation

We model the cognitive UAV swarm as a decentralised partially observable Markov decision process $\langle \mathcal{N}, \mathcal{S}, \{\mathcal{A}_i\}, P, \{\Omega_i\}, O, R, \gamma \rangle$ with $\mathcal{N} = \{1, \dots, N\}$ agents. The global state $s^t \in \mathcal{S}$ collects all UAV positions, velocities and energies, the target positions and discovery flags, and the jammer positions and currently jammed channels. At time t , agent i draws an observation $o_i^t \in \Omega_i$ through the emission model $O(o_i | s, i)$, where $o_i^t = [o_i^{\text{rf}}, o_i^{\text{radar}}, o_i^{\text{vis}}, o_i^{\text{nav}}]$ comprises per-channel radio-frequency features (received signal strength, signal-to-interference-plus-noise ratio and occupancy), radar range

507 and bearing to the nearest target, a visual field-of-view indicator, and navigation state
 508 (position, velocity and energy). Under partial observability the radar and visual blocks
 509 are masked, so the swarm must recover that information through communication.
 510 Each agent selects a continuous action $a_i^t = [\Delta x, \Delta y, c_1, \dots, c_C]$, where the first two
 511 components control movement and $\arg \max_k c_k$ selects the communication channel.
 512 Agents act simultaneously, the environment transitions by $P(s^{t+1} | s^t, \mathbf{a}^t)$ under UAV
 513 kinematics, jammer dynamics and stochastic fading, and every agent receives a shared
 514 multi-objective reward that balances communication, mission, energy, anti-jamming
 515 and safety,

$$516 \quad 517 \quad r_i^t = w_c R^{\text{comm}} + w_m R^{\text{miss}} - w_e E^{\text{energy}} - w_j P^{\text{jam}} - w_s P^{\text{safe}}, \quad (1)$$

518
 519 discounted by γ . Because execution is decentralised, agent i conditions only on infor-
 520 mation available locally, namely its own observation and the messages aggregated
 521 over its communication neighbourhood $\mathcal{G}^t$, while the critic may exploit the joint rep-
 522 resentation during centralised training. The objective is the joint policy $\pi_\theta = \{\pi_i\}$
 523 that maximises $\mathbb{E}[\sum_t \gamma^t N^{-1} \sum_i r_i^t]$. The five terms of equation (1) are: communi-
 524 cation quality R^{comm} , a logistic function of channel signal-to-interference-plus-noise
 525 ratio mapped to packet delivery ratio; mission reward R^{miss} , a shaped approach to
 526 and detection of targets; energy E^{energy} , the movement and transmission cost; jam-
 527 ming penalty P^{jam} , incurred for selecting a jammed channel and scaled by jammer
 528 proximity; and safety penalty P^{safe} , an inter-agent collision penalty. The weights
 529 $(w_c, w_m, w_e, w_j, w_s)$ were selected by grid search on validation tasks and held fixed
 530 across all methods and seeds for a controlled comparison. Supplementary Table S2
 531 lists the full notation.

532

533 4.2 Network architecture

534

535 Each sensing modality is projected to a token and processed by a transformer
 536 encoder [37], whose self-attention $\text{Attn}(Q, K, V) = \text{softmax}(QK^\top / \sqrt{d_k})V$ lets the
 537 model attend to the most informative observations, producing a per-agent feature h_i .
 538 The swarm then forms a communication graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with edges between UAVs
 539 within communication range, and a two-layer graph convolution [38]

$$540 \quad 541 \quad H' = \sigma\left(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} H W\right) \quad (2)$$

542

543 aggregates neighbour features into coordinated representations for navigation. Because
 544 graph aggregation over-smooths the local radio-frequency features an agent needs
 545 in order to choose an unjammed channel, the policy is decoupled: the navigation
 546 head consumes the coordinated transformer-plus-graph feature, whereas the channel-
 547 selection head consumes the agent’s own radio-frequency observations directly,

548

$$549 \quad \mu_i^{\text{move}} = f_{\text{nav}}(h_i^{\text{gnn}}), \quad \mu_i^{\text{chan}} = f_{\text{spec}}(o_i^{\text{rf}}). \quad (3)$$

550

551 Policies are optimised with multi-agent proximal policy optimisation [28] using a
 552 centralised critic and the clipped surrogate objective. To enable rapid adaptation

to unseen electronic-warfare tasks, training is wrapped in first-order model-agnostic meta-learning [20]: for each task $\mathcal{T}$, an inner update $\theta' = \theta - \alpha \nabla_{\theta} \mathcal{L}_{\mathcal{T}}(\theta)$ is followed by a meta-update that aggregates adapted gradients across a batch of sampled tasks. The complete training procedure is given as Supplementary Algorithm S1.

Per step, the transformer processes $T = 4$ modality tokens per agent at $O(NT^2d)$ cost, the dense graph convolution costs $O(N^2d)$ for a swarm of N agents, and the decoupled heads are $O(Nd)$. For the swarm sizes considered ($N \leq 20$) the dense graph operation is negligible and the networks remain computationally compact.

4.3 Simulator and environment

All evidence comes from a fast, fully vectorised multi-agent electronic-warfare simulator implemented in NumPy; absolute numbers are not claimed to transfer to hardware, but the model is specified in full for reproducibility. The default scenario uses $N = 8$ UAVs, $C = 6$ channels, 4 targets and 2 jammers in a 1000×1000 m arena, with time step $\Delta t = 0.1$ s and a 300-step horizon. UAVs move at up to 25 ms^{-1} ; the communication graph connects agents within 300 m; the collision radius is 15 m; targets are detected within 120 m. Per-step energy is 8×10^{-4} per unit distance plus 2×10^{-4} per transmission from a unit budget. The per-agent observation concatenates, for each channel, received signal strength, signal-to-interference-plus-noise ratio and occupancy ($3C$ values); the range and bearing to the nearest target (2); a visual in-field-of-view flag and confidence (2); and normalised position, velocity and remaining energy (5). Under partial observability the radar and visual blocks are masked to zero.

The link model draws Rayleigh fading (scale 1), adds co-channel interference (2.0 per co-channel agent) and jammer interference $50P_j/(1 + (d/200)^2)$ for jammer power P_j at range d , plus noise $1 + \mathcal{U}(0, \sigma) \cdot 5$, where σ denotes the observation-noise level throughout. The signal-to-interference-plus-noise ratio is $\text{SINR} = 10 \log_{10}(\text{signal}/(\text{interference} + \text{jamming} + \text{noise}))$ and the packet delivery ratio follows the logistic link $\text{PDR} = \text{sigmoid}(\text{SINR} - 3)$ with $\text{sigmoid}(x) = 1/(1 + e^{-x})$. Each episode randomises UAV, target and jammer positions and the initially jammed channels; evaluation uses a fixed, held-out seed offset so that training and test scenarios never coincide. The environment supports variable swarm size for the scalability analysis. It is deliberately lightweight rather than photorealistic, so that the full study runs on a single consumer machine while still exercising every capability the framework targets.

Four jammer behaviours are implemented. The sweep jammer advances the jammed channels by one every T_j steps; the static jammer holds them fixed; the random jammer redraws them every $T_j/2$ steps; and the reactive jammer, the adaptive adversary of the hard regime, counts at each step the channels the swarm is currently using and jams the n_j most used, so that a naive fixed spectrum policy is directly punished. In all modes the jammers also drift spatially by a Gaussian random walk. The hard regime uses the reactive jammer with power $P_j = 1.2$ and noise level $\sigma = 0.10$; the control regime uses the sweep jammer with $P_j = 1.0$ and $\sigma = 0.05$.

4.4 Meta-learning task distribution

Meta-training uses first-order model-agnostic meta-learning wrapped around multi-agent proximal policy optimisation. Each meta-iteration samples 4 tasks, takes one inner adaptation step per task (inner learning rate 0.01, rollout 256), and applies the averaged adapted gradients to the meta-parameters. A task is a draw from the environment distribution in which the following are randomised independently: number of targets $\mathcal{U}\{2, \dots, 6\}$, number of jammers $\mathcal{U}\{1, \dots, 3\}$, jammer power $\mathcal{U}(0.7, 1.4)$, sweep period $\mathcal{U}\{10, \dots, 29\}$, observation-noise level $\mathcal{U}(0.02, 0.12)$, and all UAV, target and jammer positions and initially jammed channels. For the generalisation study the jammer *behaviour* itself is split: meta-training draws only from {sweep, static} and evaluation only from the disjoint, held-out set {reactive, random}, so that the test adversaries are never seen during training. The split is verified from the `train_jammer_pool` field recorded in every evaluation log. The adaptation study draws 6 held-out tasks for each of 5 training seeds, takes $K \in \{0, 1, 3, 5, 10\}$ inner proximal-policy-optimisation updates on each task, and evaluates over 10 episodes after each K , giving $n = 30$ independent task-by-seed evaluations per point.

4.5 Baselines and training protocol

We compare against PPO and A2C (on-policy) and DDPG and MADDPG (off-policy, with replay and target networks), plus GRU-PPO, a recurrent single-agent control that adds memory but no inter-agent communication, and XM-noGNN, the proposed model with graph message passing removed. For a fair comparison, every method is trained for the same budget of 6×10^5 environment steps and evaluated over 30 held-out episodes. Held-out evaluation episodes use seeds disjoint from training. The control regime uses the five seeds {11, 22, 33, 44, 55}; the hard regime uses ten seeds, namely {11, 22, 33, 44, 55, 66, 77, 88, 99} plus a tenth seed (100 for PPO and A2C, 111 for XM-MADRL). Seed counts smaller than the nominal five or ten are stated per row in Tables 1 and 3. They arise from runs that were either not launched or terminated with a non-zero exit status in the original compute campaign, whose launcher logs are released with the code; no run was dropped on the basis of its result.

Models are implemented in PyTorch and trained on a single NVIDIA RTX 4060. The swarm uses $N = 8$ agents and $C = 6$ channels, a hidden size of 128, 2 transformer layers with 4 attention heads, and 2 graph-convolution layers. Multi-agent proximal policy optimisation uses a clip parameter of 0.2, discount $\gamma = 0.99$ and generalised-advantage-estimation parameter $\lambda = 0.95$, with an outer learning rate of 3×10^{-4} and rollouts of 256 steps. The meta-learning inner learning rate is $\alpha = 10^{-2}$ with $K = 4$ sampled tasks. The reward weights are $(w_c, w_m, w_e, w_j, w_s) = (1.3, 1.5, 0.5, 1.2, 2.0)$. The XM-MADRL policy has 0.37 M parameters, against 0.14 M for GRU-PPO and 0.04 M for PPO. On a single CPU core the full eight-agent forward pass takes 2.1 ms per step (0.26 ms per agent), and a single 6×10^5 -step run completes in tens of minutes to a few hours on one consumer CPU or GPU depending on the meta-learning and recurrence settings. Supplementary Table S3 lists all hyperparameters, the full configuration is released as `configs/default.yaml`, and Supplementary Table S7 expands every abbreviation used in this article.

4.6 Statistical analysis

The independent experimental unit is a seed, that is, a complete, independently initialised training run. Each trained policy is evaluated on 30 deterministic episodes; the per-episode metrics are aggregated to a single per-seed mean, and all statistical tests operate on these seed-level means, never on individual episodes. The headline comparison uses $n = 10$ seeds per method; the control regime and the component ablations use $n = 5$; the GRU-PPO control completed $n = 6$. We report the mean $\pm$ sample standard deviation, with 95% confidence intervals of the mean shown in figures. Between-method comparisons use two-sided Welch t -tests, which do not assume equal variances, between XM-MADRL and the strongest baseline, together with Cohen’s d computed with the pooled standard deviation. Welch’s test is used in preference to Student’s t -test because the seed-level variances differ substantially between methods (Table 1); with $n \leq 10$ per group the seed-level means are approximately normal by construction, each being an average over 30 evaluation episodes. Because several metrics are tested jointly, P values are adjusted across the predefined five-metric family by the Holm–Bonferroni step-down procedure at $\alpha = 0.05$, and both raw and adjusted P values are reported. All statistics are recomputed from the released per-seed logs by the analysis script `stats.py`.

4.7 Explainability protocol

Attributions use KernelSHAP [11]. We collect 60 on-policy observations as a background set, summarise them with k -means ($k = 10$), and explain 40 held-out observations with 100 coalition samples each. The scalar model output explained is the agent’s action strength, defined as movement magnitude plus selected-channel confidence, scored per agent with an identity adjacency so that attributions are local. The resulting per-feature absolute Shapley values are summed within each of the six sensing modalities, normalised, and then averaged across the explained instances. Following cautionary evidence that attribution can mislead if it is not validated [31], we state this protocol explicitly and interpret attributions only at the level of sensing modalities rather than individual pixels or scalars.

4.8 Ethics and dual-use considerations

This study did not involve human participants, human tissue or animals, so no ethical approval or informed consent was required. The work studies electronic-warfare scenarios, specifically adaptive jamming and anti-jamming, purely in simulation and for the defensive purpose of improving the resilience of cooperative UAV communication in contested spectrum. No hardware, real radio-frequency emissions, operational jamming systems or field experiments are involved, and the released code targets a synthetic benchmark rather than any deployable weapon or countermeasure. The authors have considered the potential for dual-use research of concern and judge the risk to be low: the contribution is a methodological study of *when* coordination helps, not an operational jamming capability. No generative language model was used to produce the scientific content, the experimental design, the analyses or the results reported in this manuscript.

Data availability

All code, configurations, result logs and random seeds required to reproduce every reported result are openly available at <https://github.com/adeliusa486/XM-MADRL-eXplainable-Meta-learning-Multi-Agent-DRL>. The repository includes the simulation environment, all model and baseline implementations, the experiment launchers for both regimes, the analysis scripts, and the per-seed metric logs backing every table and figure in this article and in the Supplementary Information. A fast end-to-end verification script (`smoke_test.py`) is provided. No other datasets were generated or analysed during the current study.

References

- [1] Choi, J., Kim, H. M., Hwang, H. J., Kim, Y.-D. & Kim, C. O. Modular reinforcement learning for autonomous UAV flight control. *Drones* **7**, 418 (2023).
- [2] Xu, Y., Zhang, M. & Jin, B. Pursuing benefits or avoiding threats: Realizing regional multi-target electronic reconnaissance with deep reinforcement learning. *IEEE Access* **11**, 63972–63984 (2023).
- [3] Jiang, W., Yu, W., Wang, W. & Huang, T. Multi-agent reinforcement learning for joint cooperative spectrum sensing and channel access in cognitive UAV networks. *Sensors* **22**, 1651 (2022).
- [4] Xu, Y., Wei, Y., Wang, D., Jiang, K. & Deng, H. Multi-UAV path planning in GPS and communication denial environment. *Sensors* **23**, 2997 (2023).
- [5] Dong, L., Qian, Y. & Xing, Y. Dynamic spectrum access and sharing through actor-critic deep reinforcement learning. *EURASIP Journal on Wireless Communications and Networking* **2022**, 48 (2022).
- [6] Wang, X., Teraki, Y., Umehira, M., Zhou, H. & Ji, Y. A usage aware dynamic spectrum access scheme for interweave cognitive radio network by exploiting deep reinforcement learning. *Sensors* **22**, 6949 (2022).
- [7] Zheng, Z., Li, W. & Zou, K. Airborne radar anti-jamming waveform design based on deep reinforcement learning. *Sensors* **22**, 8689 (2022).
- [8] Chen, Y., Niu, Y., Chen, C., Zhou, Q. & Xiang, P. A distributed anti-jamming algorithm based on actor-critic countering intelligent malicious jamming for WSN. *Sensors* **22**, 8159 (2022).
- [9] Liu, B. *et al.* Task assignment of UAV swarms based on deep reinforcement learning. *Drones* **7**, 297 (2023).
- [10] Hieu, N. Q. *et al.* Transferable deep reinforcement learning framework for autonomous vehicles with joint radar-data communications. *IEEE Transactions*

	<i>on Communications</i> 70 , 5164–5180 (2022).	737
		738
[11]	Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. In <i>Advances in Neural Information Processing Systems 30 (NIPS)</i> , 4765–4774 (2017).	739
		740
		741
		742
[12]	Fu, G., Chu, H., Liu, L., Fang, L. & Zhu, X. Deep reinforcement learning for the visual servoing control of UAVs with FOV constraint. <i>Drones</i> 7 , 375 (2023).	743
		744
		745
[13]	Liu, S., Wu, J. & He, J. Dynamic multichannel sensing in cognitive radio: Hierarchical reinforcement learning. <i>IEEE Access</i> 9 , 25473–25481 (2021).	746
		747
		748
[14]	Souto, A., Alfaia, R., Cardoso, E., Araújo, J. & Francês, C. UAV path planning optimization strategy: Considerations of urban morphology, microclimate, and energy efficiency using Q-learning algorithm. <i>Drones</i> 7 , 123 (2023).	749
		750
		751
[15]	Wu, Z., Du, M., Bi, D. & Pan, J. IRelNet: An improved relation network for few-shot radar emitter identification. <i>Drones</i> 7 , 312 (2023).	752
		753
		754
[16]	O’Shea, T. J., Roy, T. & Clancy, T. C. Over-the-air deep learning based radio signal classification. <i>IEEE Journal of Selected Topics in Signal Processing</i> 12 , 168–179 (2018).	755
		756
		757
		758
[17]	Al-Gburi, A., Abdullah, O., Sarhan, A. Y. & Al-Hraishawi, H. Channel estimation for UAV communication systems using deep neural networks. <i>Drones</i> 6 , 326 (2022).	759
		760
		761
		762
[18]	Dewangan, V., Saxena, A., Thakur, R. & Tripathi, S. Application of image processing techniques for UAV detection using deep learning and distance-wise analysis. <i>Drones</i> 7 , 174 (2023).	763
		764
		765
		766
[19]	Abubakar, A. I. <i>et al.</i> A survey on energy optimization techniques in UAV-based cellular networks: From conventional to machine learning approaches. <i>Drones</i> 7 , 214 (2023).	767
		768
		769
		770
[20]	Finn, C., Abbeel, P. & Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. In <i>Proceedings of the 34th International Conference on Machine Learning (ICML)</i> , Vol. 70, 1126–1135 (2017).	771
		772
		773
[21]	Sukhbaatar, S., Szlam, A. & Fergus, R. Learning multiagent communication with backpropagation. In <i>Advances in Neural Information Processing Systems 29 (NIPS)</i> , 2244–2252 (2016).	774
		775
		776
		777
[22]	Das, A. <i>et al.</i> TarMAC: Targeted multi-agent communication. In <i>Proceedings of the 36th International Conference on Machine Learning (ICML)</i> , 1538–1546 (2019).	778
		779
		780
		781
		782

783 [23] Singh, A., Jain, T. & Sukhbaatar, S. Learning when to communicate at scale in
784 multiagent cooperative and competitive tasks. In *Proceedings of the International*
785 *Conference on Learning Representations (ICLR)* (2019).
786

787 [24] Duan, W., Lu, J. & Xuan, J. Group-aware coordination graph for multi-agent
788 reinforcement learning. In *Proceedings of the Thirty-Third International Joint*
789 *Conference on Artificial Intelligence (IJCAI-24)*, 3926–3934 (2024).
790

791 [25] Feng, Z., Wu, D., Huang, M. & Yuen, C. Graph-attention-based reinforcement
792 learning for trajectory design and resource assignment in multi-UAV-assisted
793 communication. *IEEE Internet of Things Journal* **11**, 27421–27434 (2024).
794

795 [26] Hausknecht, M. & Stone, P. Deep recurrent Q-learning for partially observable
796 MDPs. Preprint at <https://arxiv.org/abs/1507.06527> (2015).
797

798 [27] Lowe, R. *et al.* Multi-agent actor-critic for mixed cooperative-competitive envi-
799 ronments. In *Advances in Neural Information Processing Systems 30 (NIPS)*,
800 6379–6390 (2017).
801

802 [28] Yu, C. *et al.* The surprising effectiveness of PPO in cooperative multi-agent games.
803 In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 24611–
804 24624 (2022).
805

806 [29] Yang, J. *et al.* Agent-based anti-jamming techniques for UAV communications
807 in adversarial environments: A comprehensive survey. Preprint at <https://arxiv.org/abs/2508.11687> (2025).
808

809 [30] Milani, S., Topin, N., Veloso, M. & Fang, F. Explainable reinforcement learning:
810 A survey and comparative review. *ACM Computing Surveys* **56**, 1–36 (2024).
811

812 [31] Atrey, A., Clary, K. & Jensen, D. Exploratory not explanatory: Counterfactual
813 analysis of saliency maps for deep reinforcement learning. In *Proceedings of the*
814 *International Conference on Learning Representations (ICLR)* (2020).
815

816 [32] Li, Y., Zhao, T., Wu, Z., Lin, Y. & Zhang, Y. Energy-harvesting-assisted UAV
817 swarm anti-jamming communication based on multi-agent reinforcement learning.
818 *Drones* **10**, 294 (2026).
819

820 [33] Chen, Y. & Niu, J. Intelligent unmanned aerial vehicle swarm control under elec-
821 tronic warfare: A cognitive-intent dual-stream reinforcement learning framework.
822 *Drones* **10**, 342 (2026).
823

824 [34] Sun, Y. *et al.* Hierarchical role-based multi-agent reinforcement learning for UHF
825 radiation source localization with heterogeneous UAV swarms. *Drones* **10**, 54
826 (2026).
827
828

[35] Cai, C. & Wang, Y. A note on over-smoothing for graph neural networks. Preprint at https://arxiv.org/abs/2006.13318 (2020).	829
	830
	831
[36] Rusch, T. K., Bronstein, M. M. & Mishra, S. A survey on oversmoothing in graph neural networks. Preprint at https://arxiv.org/abs/2303.10993 (2023).	832
	833
	834
[37] Vaswani, A. <i>et al.</i> Attention is all you need. In <i>Advances in Neural Information Processing Systems 30 (NIPS)</i> , 5998–6008 (2017).	835
	836
	837
[38] Kipf, T. N. & Welling, M. Semi-supervised classification with graph convolutional networks. In <i>Proceedings of the International Conference on Learning Representations (ICLR)</i> (2017).	838
	839
	840
Acknowledgements. This research received no external funding. All figures in this article were drawn by the authors from the released result logs with Matplotlib 3.10.9 (https://matplotlib.org); models were implemented in PyTorch 2.5.1 (https://pytorch.org) and the simulator in NumPy 1.26.4 (https://numpy.org).	841
	842
	843
	844
	845
Author contributions. Ar.A. and Ad.A. conceived the study. Ad.A. designed the methodology, implemented the software, performed the formal analysis, curated the data and produced Figures 1 to 4. Ar.A. and E.K. validated the results. Ad.A., E.K. and M.S.R. carried out the investigation. Ar.A. provided resources and administered the project. Ar.A. and G.H. supervised the work. Ad.A. wrote the main manuscript text. Ar.A. and E.K. reviewed and edited the manuscript. All authors reviewed and approved the final manuscript. Two authors share the initials A.A., so Arshad Ali is written Ar.A. and Adeel Ahmad is written Ad.A. The other initials are M.S.R. for Muhammad Shaharyar Ramzan, E.K. for Eraj Khan and G.H. for Gahangir Hossain.	846
	847
	848
	849
	850
	851
	852
	853
	854
Competing interests. The authors declare no competing interests.	855
	856
Additional information. Supplementary information accompanies this paper. It contains Supplementary Tables S1–S7, Supplementary Figures S1–S8 and Supplementary Algorithm S1.	857
	858
	859
Correspondence and requests for materials should be addressed to E.K.	860
	861
	862
	863
	864
	865
	866
	867
	868
	869
	870
	871
	872
	873
	874

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)