

Supplementary Information

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

²Higher College of Technology, Abu Dhabi, United Arab Emirates

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

⁴Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

*Correspondence: ekhan@hct.ac.ae

This file contains Supplementary Tables S1–S7, Supplementary Figures S1–S8 and Supplementary Algorithm S1. Every number reported here is regenerated directly from the per-seed evaluation logs released with the article at <https://github.com/adeliusa486/XM-MADRL-eXplainable-Meta-learning-Multi-Agent-DRL>.

Contents

1	Supplementary Tables	2
2	Supplementary Figures	5
3	Supplementary Algorithm	10
4	Supplementary References	11

1 Supplementary Tables

Supplementary Table S1: Positioning of XM-MADRL against representative reinforcement-learning-based UAV methods. ✓, the capability is addressed; –, it is not addressed. “Public code” means that a public code release accompanies the paper.

Method	Multi-agent	Adaptation	Anti-jamming	Interpretable	Public code
Modular RL [1]	–	–	–	–	–
Feature-fusion PPO [2]	–	–	–	–	–
Actor–critic DSA [3]	–	–	✓	–	–
MARL sensing [4]	✓	–	✓	–	–
Distributed anti-jamming [5]	✓	–	✓	–	–
Swarm task assignment [6]	✓	–	–	–	–
Transferable DRL [7]	–	✓	–	–	–
XM-MADRL (this work)	✓	✓	✓	✓	✓

Supplementary Table S2: Notation.

Symbol	Meaning
N, C	Number of UAV agents; number of frequency channels
$s^t \in \mathcal{S}$	Global environment state at step t
$o_i^t \in \Omega_i$	Local observation of agent i
$a_i^t \in \mathcal{A}_i$	Continuous action $[\Delta x, \Delta y, c_1, \dots, c_C]$
P, O	State-transition and observation-emission models
r_i^t, γ	Per-agent reward; discount factor
$\mathcal{G}^t = (V, E^t)$	Communication graph; E^t from the 300 m range
$A^t, \hat{A}^t$	Adjacency matrix and its symmetric normalisation
$h_i^{(\ell)}$	Feature of agent i after ℓ aggregation layers
π_θ, V_ϕ	Policy and centralised value function
$\mathcal{T}, \tau$	Meta-task distribution; a sampled task
α, K	Inner-loop learning rate; inner adaptation steps
P_j, σ	Jammer power; observation-noise level
w_c, w_m, w_e, w_j, w_s	Reward weights (communication, mission, energy, jamming, safety)

Supplementary Table S3: Key hyperparameters. Identical for the proposed model and all baselines except where the component is absent from the baseline. The complete configuration is released as `configs/default.yaml`.

Parameter	Value
UAV agents N / channels C	8 / 6
Hidden size	128
Transformer layers / attention heads	2 / 4
Graph-convolution layers	2
MAPPO clip / γ / λ	0.2 / 0.99 / 0.95
Learning rate (outer)	3×10^{-4}
MAML inner learning rate α / tasks K	10^{-2} / 4
Rollout steps / total environment steps	256 / 6×10^5
Evaluation episodes per seed	30
Seeds (control regime)	{11, 22, 33, 44, 55}
Seeds (hard regime)	{11, 22, 33, 44, 55, 66, 77, 88, 99} plus a tenth seed (100 for PPO and A2C; 111 for XM-MADRL)
Reward weights (w_c, w_m, w_e, w_j, w_s)	1.3 / 1.5 / 0.5 / 1.2 / 2.0
Parameters: XM-MADRL / GRU-PPO / PPO	0.37 M / 0.14 M / 0.04 M
Hardware	Single NVIDIA RTX 4060

Supplementary Table S4: Full control-regime comparison. Full observability with a predictable sweep jammer; mean $\pm$ sample standard deviation over $n = 5$ independent training seeds. Best value per column in bold. MADDPG attains the highest raw mission count only by flying aggressively: it incurs hundreds of collisions and the worst energy use, so it is not a usable policy. PDR, packet delivery ratio; SINR, signal-to-interference-plus-noise ratio.

Method	Mission (%)	PDR (%)	Anti-jam. (%)	SINR (dB)	Energy	Collisions
XM-MADRL (this work)	38.0 ± 4.3	55.1 ± 4.4	97.6 ± 0.2	3.6 ± 0.4	1.3 ± 0.2	0.0 ± 0.0
PPO	38.3 ± 4.9	58.7 ± 1.0	98.3 ± 0.1	3.9 ± 0.1	1.3 ± 0.3	0.0 ± 0.0
A2C	38.0 ± 5.3	58.3 ± 2.6	98.2 ± 0.1	3.9 ± 0.2	1.2 ± 0.3	0.0 ± 0.0
DDPG	39.2 ± 3.9	50.4 ± 7.8	95.4 ± 0.9	3.1 ± 0.8	1.8 ± 0.3	0.1 ± 0.1
MADDPG	58.8 ± 3.1	4.8 ± 0.1	67.3 ± 1.3	-2.2 ± 0.1	5.6 ± 0.0	268.4 ± 44.5

Supplementary Table S5: Control-regime statistical comparison. XM-MADRL versus the strongest baseline (PPO), $n = 5$ independent training seeds per method, two-sided Welch t -tests on seed-level means, Cohen’s d with pooled standard deviation. Only anti-jamming differs significantly, and it does so in favour of PPO. PDR, packet delivery ratio; SINR, signal-to-interference-plus-noise ratio.

Metric	XM-MADRL	PPO	P	Cohen’s d
Mission success (%)	38.0 ± 4.3	38.3 ± 4.9	0.912	−0.07
PDR (%)	55.1 ± 4.4	58.7 ± 1.0	0.139	−1.14
Anti-jamming (%)	97.6 ± 0.2	98.3 ± 0.1	8.0×10^{-4}	−3.65
Mean SINR (dB)	3.6 ± 0.4	3.9 ± 0.1	0.142	−1.13
Energy used	1.25 ± 0.24	1.28 ± 0.27	0.873	−0.10

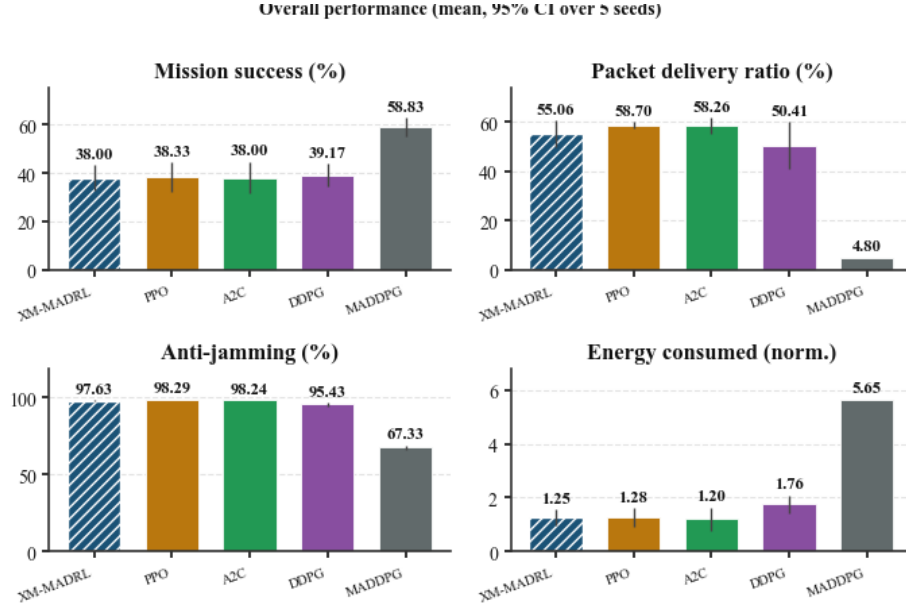
Supplementary Table S6: Few-shot adaptation to unseen jammers. Mission success and packet delivery ratio after K inner gradient steps on held-out adversary behaviours. Meta-training draws only from {sweep, static}; evaluation draws only from the disjoint {reactive, random} family. Values are mean $\pm$ sample standard deviation over $n = 30$ evaluations (5 training seeds $\times$ 6 held-out tasks, each evaluated over 10 episodes). These are the numbers plotted in Fig. 3 of the main article.

Method	$K = 0$	$K = 1$	$K = 3$	$K = 5$	$K = 10$
<i>Mission success (%)</i>					
XM-MADRL (meta-learned init.)	43.1 ± 6.8	57.2 ± 9.7	58.0 ± 11.3	59.6 ± 11.4	59.9 ± 9.9
PPO (fine-tuned)	38.6 ± 3.2	51.0 ± 11.0	49.2 ± 10.3	48.2 ± 10.2	48.5 ± 11.7
<i>Packet delivery ratio (%)</i>					
XM-MADRL (meta-learned init.)	32.0 ± 8.6	6.6 ± 6.2	6.8 ± 6.9	6.0 ± 5.8	6.7 ± 7.6
PPO (fine-tuned)	4.2 ± 1.8	6.3 ± 4.4	6.8 ± 6.7	8.4 ± 8.0	8.6 ± 10.5

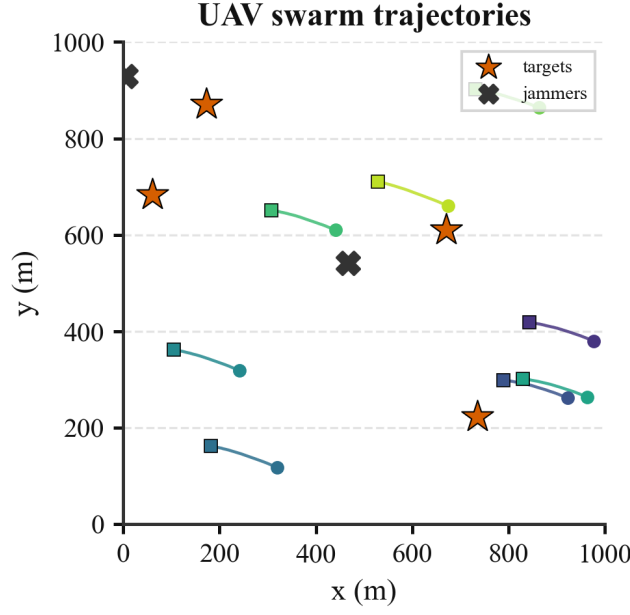
Supplementary Table S7: Abbreviations used in the article.

Abbreviation	Expansion
A2C	Advantage actor-critic
DDPG	Deep deterministic policy gradient
Dec-POMDP	Decentralised partially observable Markov decision process
DRL	Deep reinforcement learning
DSA	Dynamic spectrum access
FOMAML	First-order model-agnostic meta-learning
GAE	Generalised advantage estimation
GCN	Graph convolutional network
GNN	Graph neural network
GRU	Gated recurrent unit
MADDPG	Multi-agent deep deterministic policy gradient
MADRL	Multi-agent deep reinforcement learning
MAML	Model-agnostic meta-learning
MAPPO	Multi-agent proximal policy optimisation
MARL	Multi-agent reinforcement learning
PDR	Packet delivery ratio
PPO	Proximal policy optimisation
RF	Radio frequency
RSSI	Received signal strength indicator
SHAP	SHapley Additive exPlanations
SINR	Signal-to-interference-plus-noise ratio
UAV	Unmanned aerial vehicle

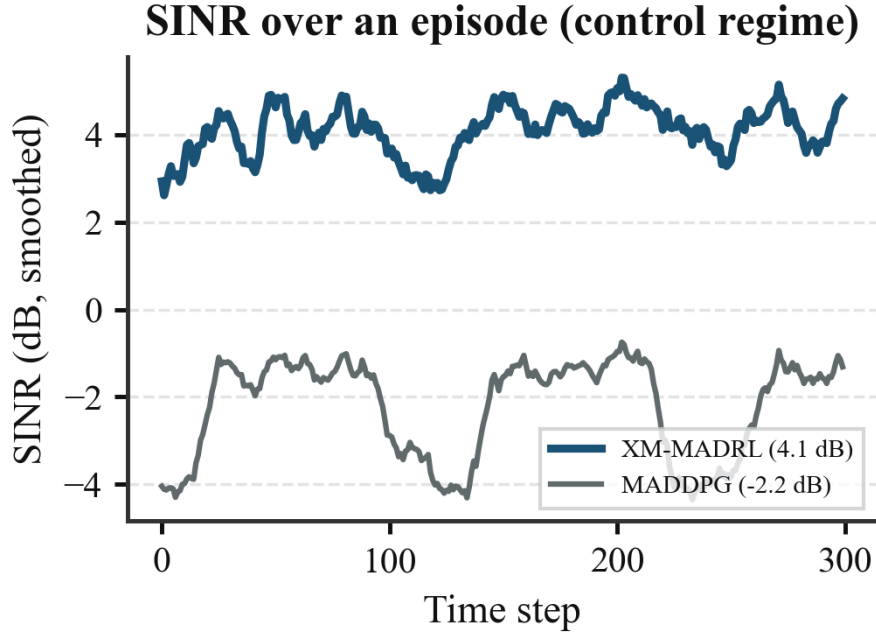
2 Supplementary Figures



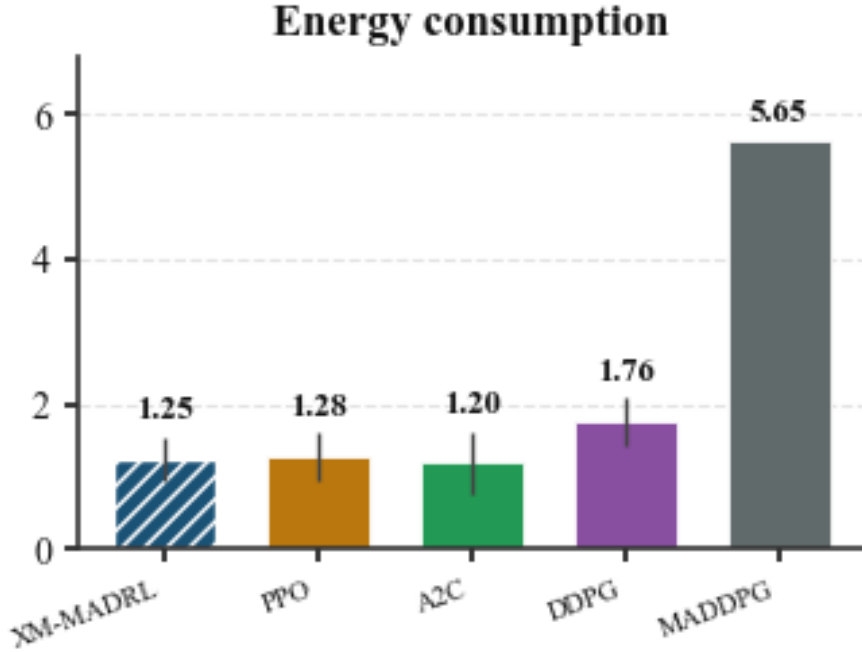
Supplementary Figure S1: Overall performance comparison in the control regime. Full observability with a predictable sweep jammer. Bars are means over $n = 5$ independent training seeds; error bars are 95% confidence intervals of the mean (t -based). The numeric values, with standard deviations, are given in Supplementary Table S4.



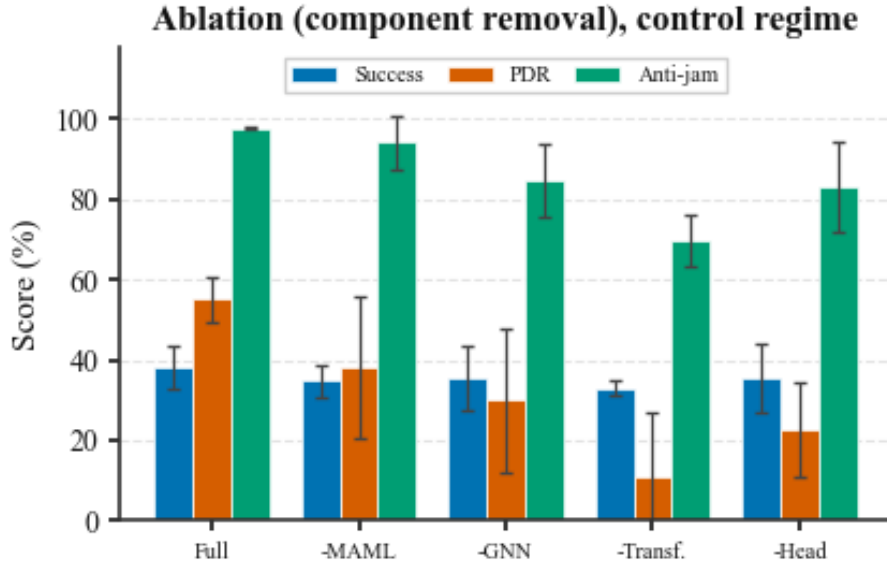
Supplementary Figure S2: Representative XM-MADRL swarm trajectories over one episode. Circles, start positions; squares, end positions; stars, targets; crosses, jammers. The swarm executes compact, collision-free manoeuvres toward the targets while keeping clear of the central jammer.



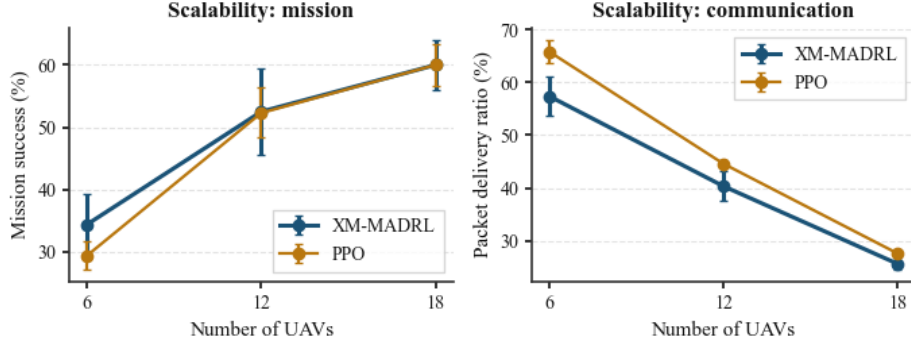
Supplementary Figure S3: Received signal-to-interference-plus-noise ratio over an episode in the control regime. Full observability with a sweep jammer; smoothed. XM-MADRL sustains a stable, positive ratio, whereas the poorly coordinated MADDPG policy is repeatedly driven negative.



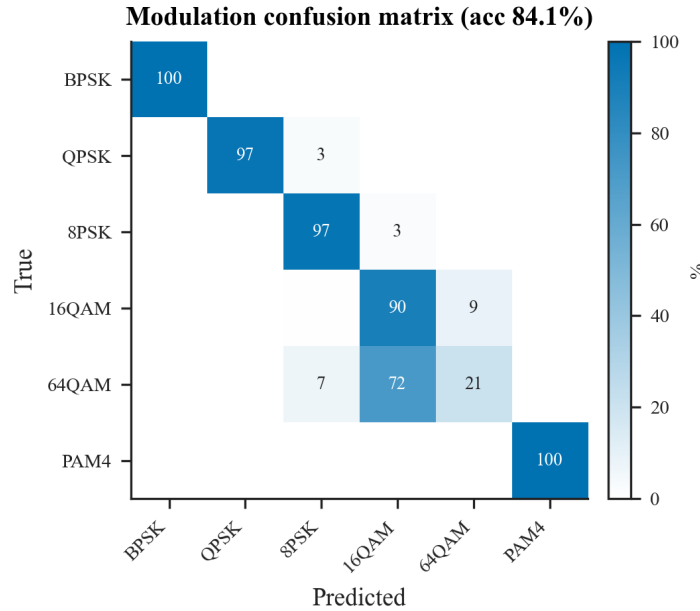
Supplementary Figure S4: Energy consumption per method in the control regime. Bars are means over $n = 5$ independent training seeds; error bars are 95% confidence intervals of the mean. XM-MADRL is on par with the on-policy baselines and far below MADDPG.



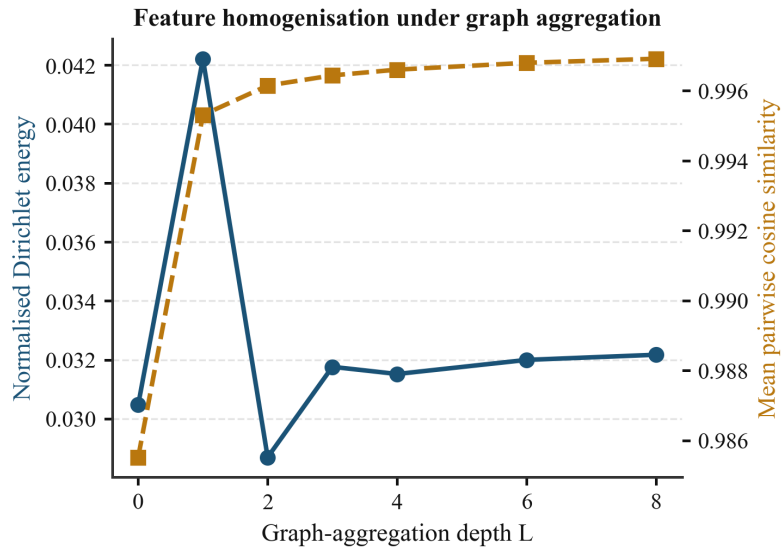
Supplementary Figure S5: Component ablation in the control regime. Effect of removing each component on mission success, packet delivery ratio and anti-jamming. Bars are means over independent training seeds; error bars are 95% confidence intervals of the mean (t -based). The numeric values, with standard deviations and the per-configuration seed counts, are given in Table 3 of the main article.



Supplementary Figure S6: Scalability with swarm size. Mission success and packet delivery ratio for swarms of 6, 12 and 18 UAVs. Points are means over $n = 4$ independent training seeds per method and swarm size; error bars are the standard deviation across those seeds. Mission success improves with swarm size for both methods, whereas communication reliability declines for both as more agents contend for a fixed set of six channels.



Supplementary Figure S7: Confusion matrix for six-class modulation recognition. Row-normalised percentages. This is an auxiliary sensing-validation experiment: the classifier is not part of the reinforcement-learning decision loop and does not feed the policy, the reward or channel selection. Overall accuracy is 84%; residual confusion is concentrated in the dense 16-QAM and 64-QAM constellations, as expected under noise.



Supplementary Figure S8: Feature homogenisation under graph aggregation. Normalised Dirichlet energy (left axis) and mean pairwise cosine similarity (right axis) of the encoder features as a function of graph-aggregation depth L , measured under partial observability. Per-agent features are already highly similar before any aggregation ($L = 0$, cosine similarity 0.986) and become monotonically more similar with depth, which is the empirical basis for driving channel selection from a local spectrum head rather than from the coordination-smoothed representation.

3 Supplementary Algorithm

Supplementary Algorithm S1 XM-MADRL training (first-order MAML wrapped around MAPPO). This is the procedure described in the Methods section of the main article.

- 1: Initialise meta-parameters θ (transformer, graph convolution, decoupled heads, critic)
 - 2: **for** meta-iteration = 1, 2, ... **do**
 - 3: $g \leftarrow 0$
 - 4: **for** each sampled electronic-warfare task $\mathcal{T}_k$ (jammer, mission, noise) **do**
 - 5: $\theta_k \leftarrow \theta$
 - 6: **for** inner step = 1, ..., n **do**
 - 7: Collect rollout with π_{θ_k} ; compute generalised advantage estimates
 - 8: $\theta_k \leftarrow \theta_k - \alpha \nabla_{\theta_k} \mathcal{L}_{\mathcal{T}_k}^{\text{PPO}}(\theta_k)$
 - 9: Collect query rollout; $g \leftarrow g + \nabla_{\theta_k} \mathcal{L}_{\mathcal{T}_k}^{\text{PPO}}(\theta_k)$
 - 10: $\theta \leftarrow \theta - \beta g / K$ ▷ meta-update
 - 11: Apply an additional MAPPO update on the fixed training task
-

4 Supplementary 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., Wei, Y., Wang, D., Jiang, K. & Deng, H. Multi-UAV path planning in GPS and communication denial environment. *Sensors* **23**, 2997 (2023).
- [3] 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).
- [4] 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).
- [5] 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).
- [6] Liu, B. *et al.* Task assignment of UAV swarms based on deep reinforcement learning. *Drones* **7**, 297 (2023).
- [7] Hieu, N. Q. *et al.* Transferable deep reinforcement learning framework for autonomous vehicles with joint radar-data communications. *IEEE Transactions on Communications* **70**, 5164–5180 (2022).