Decentralized Q-Learning for Emitter Localization with Consensus Policy Selection

Christopher Peters
Electrical & Computer Engineering
Southern Methodist University
Dallas, TX
peterscl@smu.edu

Mitchell A. Thornton
Darwin Deason Institute for Cyber Security
Southern Methodist University
Dallas, TX
mitch@smu.edu

Eric Larson
Darwin Deason Institute for Cyber Security
Southern Methodist University
Dallas, TX
eclarson@smu.edu

***Abstract*— This study proposes and evaluates a fully decentralized Q-Learning framework for dynamic optimization of unmanned aerial system array geometries to reduce emitter localization error. By combining decentralized multi-agent Q-learning with adjustments to the array geometry based on real-time environmental feedback, signal time-difference of arrival (TDOA), and reduction of the spherical error probable (SEP) of the emitter location, the system develops a consensus policy to optimize sensor positions and converge on the true location of a sensed emitter with minimized localization errors in complex multipath environments. The consensus policy allows scalability to larger drone network sizes, further enhancing localization accuracy without additional training. Simulation results confirm the viability of the system to optimize its geometry to improve the accuracy of localization and show the scalability with a shared policy. Our results demonstrate the opportunity provided through intelligent dynamic repositioning of the UAS array to enhance localization performance in dense urban settings, offering practical approaches for effective performance in active operations of surveillance or search and rescue missions.**

***Keywords*— UAS, UAV Networks, Emitter Localization, LOCA, Decentralized Q-Learning, Consensus Policy**

I. Introduction

Collections of cooperative unmanned autonomous systems (UAS), like the ones represented by the functional architecture shown in Fig. 1, provide the ability to localize unknown-location radio frequency (RF) emitters [1]-[12]. Localization accuracy is essential in search and rescue or military operations, and informed repositioning of mobile UAS receiver systems reduces localization error [13].

Recent work has applied reinforcement learning and distributed machine learning to UAS swarms for mapping, tracking, and target localization [1]-[7]. Other studies have investigated RSS-based and TDOA-based localization strategies, as well as the impact of realistic UAS channel models and air–ground propagation on localization performance [8]-[11]. Our prior work analyzed timing-based emitter localization using multilateration and LOCA under bounded sensor positioning and timing errors [12] and introduced geometry-adaptive UAS arrays for dense multipath environments [13]. Together, these efforts motivate a learning-based approach that exploits geometry-aware localization algorithms and realistic channel effects when controlling a cooperative UAS array.

We study a fully decentralized [14] multi-agent deep Q-learning (MDQL) framework, utilizing a time difference of arrival (TDOA) algorithm, where UAS agents independently optimize their policies to promote geometries that minimize localization error in complex multipath environments. After training, we perform a pairwise comparison analysis of the policies learned by each agent to find a single representative “consensus” policy [15] that can be applied to all agents for system operation. While other approaches optimize policies during training or reach consensus at run-time [16], the consensus representative policy method offers a simple architecture that is scalable to larger drone network sizes without additional training or cross-agent learning parameter sharing.

II. Methodology

A. Localization Algorithm

The UAS array forms a cooperative “drone swarm” that measures received signal strength (RSS) and signal time of arrival from an unknown-location emitter. The emitter location ***E*** is estimated from TDOA measurements among swarm subsets using the Location on a Conic Axis (LOCA) algorithm [17], consistent with the architecture in [12]. LOCA uses TDOA from any subarray of at least four UAS sensors and provides an alternative to traditional hyperbolic ranging. For LOCA, sensor triads lie on the perimeter of a conic with the emitter at the focus as depicted in Fig. 2, whereas hyperbolic ranging relates each TDOA to a single reference sensor. Because LOCA expresses TDOA constraints along a conic axis jointly defined by sensor triads, rather than independent single-reference hyperbolas for each sensor pair, the geometry is well-suited to highly distributed, mobile receiver networks. In an eight-sensor system,

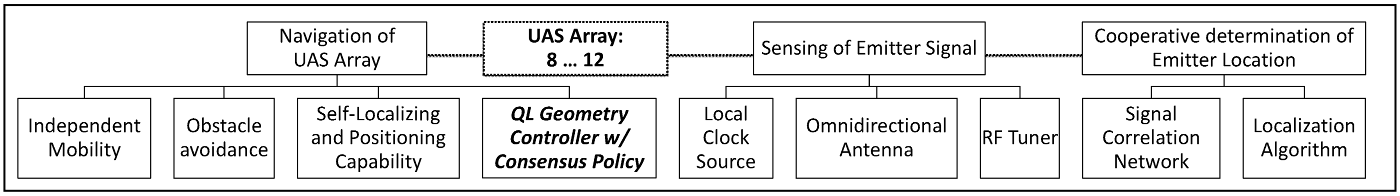
Navigation of UAS Array
UAS Array: 8 … 12
Sensing of Emitter Signal
Cooperative determination of Emitter Location
Independent Mobility
Obstacle avoidance
Self-Localizing and Positioning Capability
QL Geometry Controller w/ Consensus Policy
Local Clock Source
Omnidirectional Antenna
RF Tuner
Signal Correlation Network
Localization Algorithm

Fig. 1. Cooperative UAS representative system block diagram.

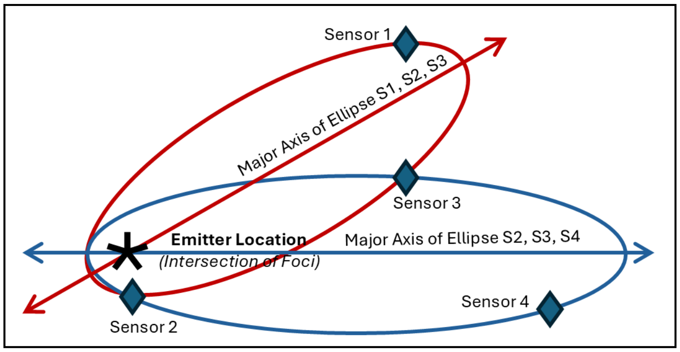

Fig. 2. Location on a Conic Axis (LOCA) representative sensor and emitter geometries: sensors on perimter of conic, emitter found at foci .

this geometry allows combinations of all sensor subsets to generate a cloud of emitter location estimates, and we define the emitter estimate $\hat{\boldsymbol{e}}$ as a probabilistic centroid given by the median of the LOCA solutions $\widehat{\boldsymbol{E}}$ [13]. We approximate the spherical error probable (SEP) radius [18] as the maximum absolute deviation of the subset-based estimates from this centroid. The UAS elements iteratively reposition to reduce SEP until the collective SEP of all subset estimates is below a user-defined goal $\boldsymbol{sep_{min}}$ [13], providing a geometry-driven convergence metric that does not depend on the true (unknown) emitter location.

B. *Decentralized Multi Agent Q-Learning with Policy Consensus*

Our architecture builds on Q-learning localization approaches [2]-[5] in a continuous state space [6]-[7]. Each UAS drone agent independently trains a deep Q-network, $\boldsymbol{Q_\theta(s,a)}$, to approximate Q-values over continuous states $\boldsymbol{s}$ and actions $\boldsymbol{a}$. The system begins by measuring the signal information at an initial UAS location $\boldsymbol{U}$ to create an initial state $\boldsymbol{s_0}$ comprised of the UAS locations, the agents' TDOA measurements, the RSS for each agent, and the collective SEP. We apply greedy action selection to determine a new location for the UAS at each timestamp $\boldsymbol{t}$ given current state $\boldsymbol{s_t}$:

$$a_t = \arg\max_{\mathrm{a}} Q_\theta(s_t, a) \tag{1}$$

using discrete 3-dimensional movement actions scaled to SEP size. Once the agents move to the new location, the new state $\boldsymbol{s_{t+1}}$ is determined. We reward states that satisfy an optimization reward condition $\boldsymbol{r_t}$, and define the target $\boldsymbol{y_t}$ as:

$$y_t = r_t + \gamma \max_a Q_\theta(s_{t+1}, a) \tag{2}$$

with discount factor γ. The Bellman temporal difference error δ_t at time t is then

$$\delta_t = y_t - Q_\theta(s_t, a_t). \tag{3}$$

We update the Q-network parameters by minimizing the mean squared Bellman error (MSE) via backpropagation.

After training of a small set of UAS drone elements, we analyze learned policies to generate a consensus representative policy. The consensus policy [15]-[16] approach assesses the individual policies for similarity using pairwise mean absolute differences and Pearson correlation to identify a single representative policy that is most similar to all agent policies. We apply this representative policy to all agents, replacing the independent policies learned during training. This consensus policy approach allows us to scale the UAS system to any number of elements without additional training. The approach is enabled both by using a network of UAS that have the same hardware architecture and by the use of a localization algorithm that is optimized by a distributed sensor geometry [13].

C. *Reward Methodology*

We evaluate three methods to inform the UAS geometry, utilizing both the SEP determined by the LOCA algorithm and the RSS. The success of each method is measured by its reduction of median localization error, which is only captured as a metric, and is not used as a parameter in the training, since the true target location is unknown in practice. We consider three reward methods: RSS Optimization, SEP Optimization, and combined RSS & SEP Optimization.

1) RSS Optimization: The reward is maximized when the aggregate RSS at a given timestamp is greater than the previous step, as in [6], promoting the UAS geometry to physically converge on the emitter.

2) SEP Optimization: The reward is maximized when SEP decreases [13] relative to a previous timestamp, so that the geometry is optimized towards sep_{min}. This case does not require geometric convergence on the emitter and gives the system the freedom to optimize the geometry according to the localization algorithm.

3) RSS & SEP Optimization: The reward is maximized by equally weighting geometry that both reduces SEP and increases RSS.

Our geometry-optimizing MDQL localization algorithm is summarized in Algorithm 1.

III. Evaluation Approach

The system was simulated using the Actor-Environment Cycle Environment (AECEnv) class provided by PettingZoo [20] in Python, with an observation space defined by Box from Gym [21].

Each simulated UAS utilizes an omnidirectional antenna, contains a self-positioning system with accuracy of 2 cm, and contains a reference clock with 10 ns accuracy synchronized pre-mission [12]. The emitter is stationary and broadcasting a non-CW 5 GHz signal in a simulated 4000 m cubic urban environment, where the building scatterers follow a Rayleigh distribution [19]. Measurement errors for the UAS locations, signal TOA, and RSS levels were injected for the system to simulate real-world multipath and hardware error conditions, using the defined hardware accuracy above and ITU-R P1411 [19], following the approach in [13], creating a complex signal environment.

Each reward method (RSS, SEP, and RSS + SEP) was trained to SEP convergence for 1000 episodes and a UAS size of 8 drones based on the studies in [12][13], with each episode

Algorithm 1: Optimizing UAS Geometry for Localization

```
Initialize target $\boldsymbol{E}$ & UAS locations $\boldsymbol{U_0}$ to random initial positions for step $\boldsymbol{n = 0}$. Set Q-network weights $\boldsymbol{Q_\theta(s_0, a_0) \coloneqq 0}$. Set algorithm goals of maximum steps $\boldsymbol{n}$ to $\boldsymbol{n_{max}}$ and minimum SEP to $\boldsymbol{sep_{\mathrm{min}}}$.
Estimate localization $\boldsymbol{\widehat{E}_0}$ & SEP radius $\boldsymbol{SEP_0}$ using LOCA.
Collect state $\boldsymbol{s_0} = \{\boldsymbol{U_0, TDOA_0, SEP_0, RSS_0}\}$.
Choose reward condition: 1) $\boldsymbol{rss_{\mathrm{max}}}$ or 2) $\boldsymbol{sep_{\mathrm{min}}}$ or 3) $\{\boldsymbol{rss_{\mathrm{max}}}$ and $\boldsymbol{sep_{\mathrm{min}}}\}$
while $\neg\boldsymbol{done}$:
    Select next movement action $\boldsymbol{a_{\mathrm{n}}}$ with greedy strategy (1) using $\boldsymbol{s_n}$ and distance moved scaled by $\boldsymbol{SEP_{n-1}}$.
    Take action $\boldsymbol{a_{\mathrm{n}}}$ and move the agents:
        Move to new UAS locations: $\boldsymbol{U_{\mathrm{n+1}} = U_n + a_{\mathrm{n}}}$.
        Measure emitter signal $\boldsymbol{TDOA_{n+1}}$ & $\boldsymbol{RSS_{n+1}}$ for each UAS.
        Perform LOCA, find $\boldsymbol{\widehat{E}_{n+1}}$, calculate estimate centroid: $\boldsymbol{\hat{e}_{n+1}} = \mathbf{median}(\boldsymbol{\widehat{E}_{n+1}})$, and $\boldsymbol{SEP_{n+1}}$
        Build next state $\boldsymbol{s_{\mathrm{n+1}}} = \{\boldsymbol{U_{n+1}, TDOA_{n+1}, SEP_{n\_1}, RSS_{n+1}}\}$.
        Set reward $\boldsymbol{r_n} = \boldsymbol{+1}$ if condition is met, else $\boldsymbol{r_n} = -1$.
        If $\boldsymbol{n+1 \geq n_{\mathrm{max}}}$ or $\boldsymbol{SEP_{n+1} \leq sep_{\mathrm{min}}}$ then $\boldsymbol{done}$.
    Update the Q-network:
        Compute target $\boldsymbol{y_n}$ (2).
        Compute Bellman error: $\boldsymbol{\delta_n}$ (3).
        Update $\boldsymbol{\theta}$ by minimizing MSE over $\boldsymbol{\delta_n}$ via backpropagation.
    Increment step count $\boldsymbol{n \leftarrow n+1}$.
If $\boldsymbol{done}$, evaluate localization error: $\boldsymbol{\xi} = \|\boldsymbol{E} - \boldsymbol{\hat{e}_n}\|$.
```

having different emitter locations and initial UAS positions. The resultant policies from the training were analyzed, and a single consensus policy for the system was generated that could be applied to any quantity of drone agents.

In Fig. 3, example training metrics for the RSS+SEP method shows the final localization error when the system achieved the minimum SEP at the end of each episode, the average distance moved per UAS to reach SEP convergence, and the number of steps taken per episode to reach SEP convergence. All methods showed similar characteristics, with >75% of all episodes achieving very low localization error when the SEP converged.

IV. Results

The simulated system was first evaluated with the 8-UAS case with the individual (non-consensus) policies developed during the training. The evaluation utilized the same simulation environment as training, including hardware errors and multipath effects, with different emitter locations and initial UAS starting positions. Next, the consensus policy was created and applied to the simulated system, again using the same simulation environment as the training with the same emitter locations and initial UAS starting locations as the independent policy evaluation. For the 8-UAS case, for each of the independent and consensus approaches, we conducted 1500 evaluation episodes. We then scaled the consensus policy to larger UAS swarms by applying the same policy to each additional UAS in the network. For the 9-12 UAS swarm array size cases, we evaluated fewer trials than the 8-UAS case due to faster convergence.

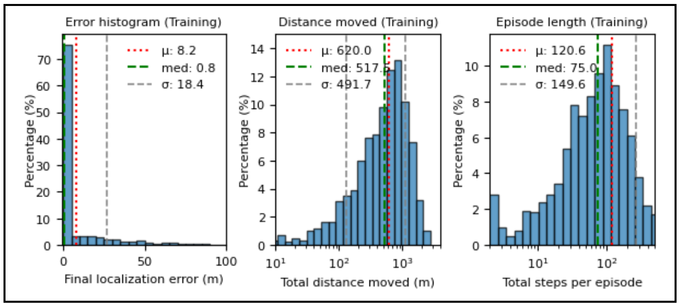

Fig. 3. Training metrics for the UAS Geometry optimizer, rewarded by a combination of SEP and RSS, showing the distribution from all training episodes of the final localization error, the average distance moved per drone, and the number of repositioning steps to convergence.

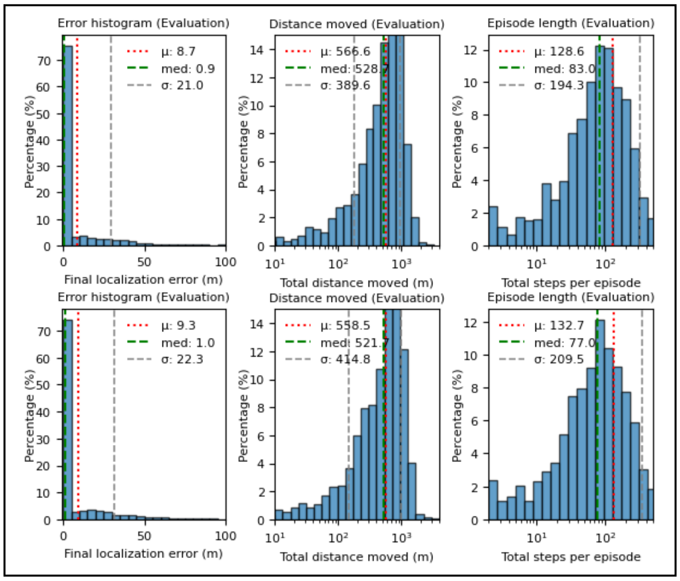

Fig. 4. Evaluation metrics for the UAS Geometry optimizer with per-agent policies (top) and single consensus policy applied to all agents (bottom), rewarded by a combination of SEP and RSS. Performance shows the final localization error, the average distance moved per drone, and the number of repositioning steps to convergence across all evaluation episodes for the given policy approach. Metrics for both policy approaches are comparable.

For the baseline 8-UAS configuration, the localization performance metrics for the independently trained policies (Fig. 4, top) and the learned consensus policy (Fig. 4, bottom) are nearly identical in terms of localization error, per- drone distance moved, and the number of repositioning steps. These results indicate that the consensus policy preserves both convergence behavior and motion cost observed during training. Therefore, for this architecture, the policy-selection procedure does not degrade performance and a single representative policy can effectively replace individually trained agent policies.

After evaluating the 8-UAS case, we evaluated the 9-12 UAS cases. TABLE I. provides the metrics for the individual policies and the shared consensus policy evaluated on the 8 UAS system, and it provides the metrics for scaling the UAS array size under the shared consensus policy. As the swarm increased from 8 to 12 UAS (TABLE I.), the mean localization error drastically reduced from 8 m to nearly 0 m (>99.9% reduction), and the standard deviation reduced from 21 m to 0 m, which is an expected performance improvement with increasing UAS array size [13]. At the same time, the mean per-drone distance moved decreases from 567 m for 8 UAS to only 7.2 m for 12 UAS, and the mean repositioning steps to convergence is reduced from 129 to 2. The intermediate 9–11 UAS cases follow the same trend, with agents simultaneously reducing localization error, tightening the spread of outcomes, and traveling shorter

TABLE I. SHARED CONSENSUS POLICY PERFORMANCE EVALUATION WITH INCREASING UAS QUANTITY

UAS Array Size	**Final Localization Error (m)**			**Per-Drone Distance Moved (m)**			**Number of System Repositioning Steps**		
	Mean (μ)	***Median (med)***	***Standard Deviation (σ)***	***μ***	***med***	***σ***	***μ***	***med***	***σ***
8 (*individual policy*)	9.3	1.0	22.3	559	522	415	133	77	210
8 (*consensus policy*)	8.7	0.9	21.0	567	529	390	129	83	194
9	0.6	0.0	2.5	221	108	287	36	12	64
10	0.1	0.0	0.3	69	1.6	160	10	2	21
11	0.02	0.0	0.1	19.1	1.5	66	3	2	8
12	0.01	0.0	0.1	7.2	1.5	39	2	2	4

distances for convergence. The behavior shows that the approach becomes increasingly advantageous as additional agents are added to the swarm because the consensus policy approach allows scaling the UAS size without retraining agents.

The evolution of localization performance over episode steps (Fig. 5) further highlights the benefit of scaling the array. For all swarm sizes, the system reduces localization error as the geometry is adapted, but larger swarms converge faster and more reliably. Whereas the 8-UAS case often requires tens to over 100 steps to reach a low-error configuration, the 11- and 12-UAS cases typically converge within a few steps, with much lower residual error due to improved geometric diversity and larger volume of localization estimates from LOCA-based TDOA measurements.

The three reward structures (RSS-only, SEP-only, and combined RSS+SEP) also show distinct behaviors. The RSS approach encourages physical convergence on the emitter and can achieve low error, but often requires more motion and steps because it is unaware of malformed geometries that degrade TDOA localization. The SEP-only approach focuses on improving LOCA geometry and often achieves low errors without clustering around the emitter. The mixed RSS+SEP reward provides the best trade-off, consistently reaching the smallest mean error with few steps and limited per-drone distance. Together with the consensus policy scaling results, these results show that combining a geometry-aware metric (SEP) with a physically intuitive metric (RSS) produces swarm behaviors that are both mission-efficient and robust to multipath conditions.

V. CONCLUSION

The agents trained independently and naturally converged to similar behavioral policies due to the shared optimization approach. Through decentralized learning and consensus policy selection, this work enables accurate cooperative localization of emitters with simplified scalability to larger drone networks. The consensus policy maintains performance for the original 8-UAS configuration while providing substantial gains when additional UAS are added, reducing mean localization error by more than two orders of magnitude with fewer steps and less per-drone motion.

Future work will extend to alternate learning methods, incorporate more robust communication models with hardware constraints, and examine nonstationary and multiple emitter scenarios to further evaluate and improve the robustness of the proposed framework for real-world deployment.

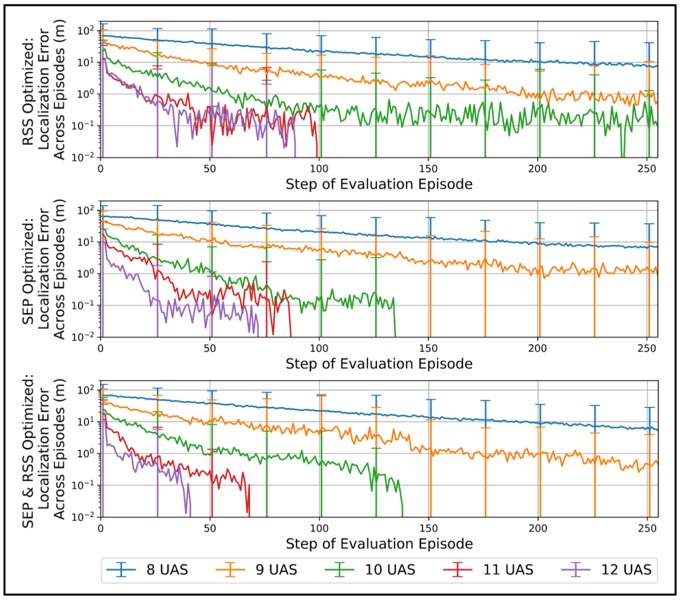

Fig. 5. Comparison of mean localization error with increasing size of UAS "drone swarm" and different optimization reward approaches: RSS Optimization (top), SEP Optimization (Middle), RSS & SEP Optimization (bottom). Results find minimal error and fewest steps to convergence with 12 drones in a mixed RSS+SEP reward structure.

REFERENCES

[1] A. Guerra, F. Guidi, D. Dardari and P. M. Djurić, "Reinforcement Learning for Joint Detection and Mapping Using Dynamic UAV Networks," in IEEE Transactions on Aerospace and Electronic Systems, vol. 60, no. 3, pp. 2586-2601, June 2024, doi: 10.1109/TAES.2023.3300813.

[2] Y. Ding, Z. Yang, Q. -V. Pham, Y. Hu, Z. Zhang and M. Shikh-Bahaei, "Distributed Machine Learning for UAV Swarms: Computing, Sensing, and Semantics," in IEEE Internet of Things Journal, vol. 11, no. 5, pp. 7447-7473, 1 March, 2024, doi: 10.1109/JIOT.2023.3341307.

[3] J. Jyoti and R. S. Batth, "Unmanned Aerial vehicles (UAV) Path Planning Approaches," 2021 International Conference on Computing Sciences (ICCS), Phagwara, India, 2021, pp. 76-82.

[4] S. Wu, "Illegal radio station localization with UAV-based Q-learning," in China Communications, vol. 15, no. 12, pp. 122-131, Dec. 2018, doi: 10.12676/j.cc.2018.12.010.

[5] Y. J. Chen, D. K. Chang and C. Zhang, "Autonomous Tracking Using a Swarm of UAVs: A Constrained Multi-Agent Reinforcement Learning Approach," in IEEE Transactions on Vehicular Technology, vol. 69, no. 11, pp. 13702-13717, Nov. 2020, doi: 10.1109/TVT.2020.3023733.

[6] M. Shurrab, R. Mizouni, S. Singh, and H. Otrok, "Reinforcement learning framework for UAV-based target localization applications," Internet of Things, vol. 23, p. 100867, 2023, doi: 10.1016/j.iot.2023.100867.

[7] X. Chen, C. Fu, and J. Huang, "A Deep Q-Network for robotic odor/gas source localization: Modeling, measurement and comparative study,"

Measurement, vol. 183, p. 109725, 2021, doi: 10.1016/j.measurement.2021.109725.

[8] X. Cheng, F. Shu, Y. Li, Z. Zhuang, D. Wu and J. Wang, "Optimal Measurement of Drone Swarm in RSS-Based Passive Localization With Region Constraints," in IEEE Open Journal of Vehicular Technology, vol. 4, pp. 1-11, 2023, doi: 10.1109/OJVT.2022.3213866.

[9] Y. Dong, F. Li, C. Ma, C. He and Z. J. Wang, "UAV-Based Dynamic Object Tracking with Radio Map," ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, Republic of, 2024, pp. 9166-9170.

[10] C. Yan, L. Fu, J. Zhang and J. Wang, "A Comprehensive Survey on UAV Communication Channel Modeling," in IEEE Access, vol. 7, pp. 107769-107792, 2019, doi: 10.1109/ACCESS.2019.2933173.

[11] D. W. Matolak and R. Sun, "Air-ground channel characterization for unmanned aircraft systems: The near-urban environment," MILCOM 2015 - 2015 IEEE Military Communications Conference, Tampa, FL, USA, 2015, pp. 1656-1660, doi: 10.1109/MILCOM.2015.7357682.

[12] C. Peters and M. A. Thornton, "Cooperative UAS Geolocation of Emitters with Multi-Sensor-Bounded Timing and Localization Error," 2023 IEEE Aerospace Conference, Big Sky, MT, USA, 2023, pp. 1-13, doi: 10.1109/AERO55745.2023.10116023.

[13] C. L. Peters, M. A. Thornton, “Reducing Emitter Localization Error in Urban Environments with Geometry Adaptive UAS Arrays,” *2025 IEEE International Systems Conference (SysCon)*, Montreal, QC, Canada, 2025.

[14] J. Jiang and Z. Lu, “I2Q: A fully decentralized Q-learning algorithm,” in *Advances in Neural Information Processing Systems* 35 (2022): 20469-20481.

[15] D. Shi, J. Tong, Y. Liu, and W. Fan, “Knowledge reuse of multi-agent reinforcement learning in cooperative tasks,” Entropy, vol. 24, no. 4, p. 470, Apr. 2022, doi: 10.3390/e24040470.

[16] Z. Xu, et. al, “Consensus Learning for Cooperative Multi-Agent Reinforcement Learning”, *AAAI*, vol. 37, no. 10, pp. 11726-11734, Jun. 2023.

[17] R. O. Schmidt, "A New Approach to Geometry of Range Difference Location," in IEEE Transactions on Aerospace and Electronic Systems, vol. AES-8, no. 6, pp. 821-835, Nov. 1972.

[18] W. E. Hoover and U. S., “Algorithms for confidence circles and ellipses,” United States National Ocean Service Office of Charting and Geodetic Services, Tech. Rep., 1984, nOAA technical report NOS 107 C&GS 3. [Online] Available: https://repository.library.noaa.gov/view/noaa/23141.

[19] Radiocommunication Sector of International Telecommunication Union, "Propagation data and prediction methods for the planning of short-range outdoor radiocommunication systems and radio local area networks in the frequency range 300 MHz to 100 GHz", Rec. ITU-R P. 1411–12 ITU Recommendation, Aug. 2023.

[20] Farama Foundation, “PettingZoo,” [Online]. Available: https://pettingzoo.farama.org/.

[21] Farama Foundation, “Gym,” [Online]. Available: https://gymnasium.farama.org/.