Reinforcement Learning Approaches for Cooperative Multi-UAV Localization of Emitters

Christopher Peters
Darwin Deason Institute for Cyber Security
Southern Methodist University
Dallas, TX
peterscl@smu.edu

Michael Watts
Darwin Deason Institute for Cyber Security
Southern Methodist University
Dallas, TX
mcwatts@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*— Cooperative unmanned aerial vehicle (UAV) arrays can improve RF emitter localization by actively changing sensor geometry, but the controller choice is often difficult to decide because prior comparisons vary the estimator, sensing model, or motion assumptions at the same time. This work compares four reinforcement-learning geometry controllers against a non-learning baseline under a common localization approach. Each controller receives the same signal observations, operates under the same motion model, and is evaluated using the same localization estimator in the same dense-multipath error model. The controller evaluation compares independent versus centralized training decentralized execution models as well as off-policy versus on-policy models. In Monte Carlo simulations with randomized initial UAV and emitter positions, learned controllers achieve larger error reductions and more stable localization performance than the greedy baseline: median localization error decreases by 60% and the uncertainty radius by 90% within 7-12 steps. Further, the study shows that while learned geometry control improves localization convergence, centralized training is not always necessary for this real-time, continuous feedback localization problem.**

I. MOTIVATION

A cooperative UAV array provides a more adaptable sensing geometry for RF emitter localization over that of fixed arrays or fixed-wing collection platforms which have limited field of view or limited elevation diversity [1], [2]. In cluttered RF environments with time-based localization, multipath and timing effects can increase uncertainty and cause a static or poorly conditioned array to converge slowly [3] or to settle around an inaccurate estimate.

Prior work has shown that UAV repositioning can improve localization by reducing uncertainty in the estimator [4]-[9]. However, there are limited studies that compare how the controller improves the estimation process without simultaneously changing multiple system elements: the estimator, the reward approach, the sensing model, the controller, or the environment. In this work, we instead vary only the controller while holding fixed the localization approach and environment, allowing this comparison to focus on how reinforcement-learning (RL) architectures provide an advantage over a non-learning baseline and how the choice in centralized/decentralized training/execution and off/on-architectures impacts the operational outcome.

II. LOCALIZATION AND CONTROL

The localization system uses a cooperative set of UAV sensors that reposition their geometries as they localize an unknown-location 5 GHz emitter. The system collects received signal strength (RSS) and time-difference-of-arrival (TDOA) measurements from the emitter at each localization snapshot. For the estimator, Location on a Conic Axis (LOCA) [10], [11] produces a candidate emitter location from sensor subsets [12]. Aggregating subset estimates forms a solution cloud which is filtered to remove outliers, and the remaining distribution is summarized by a median emitter estimate and a spherical error probable radius (SEP) [13], providing an internal uncertainty metric without requiring knowledge of the true emitter location.

The sensing model includes UAV self-position error, local clock timing error, and range/SNR-dependent time-of-arrival uncertainty. The dense multipath environment includes mixed line-of-sight and non-line-of-sight contributions modeled by [14]. During policy evaluation, a position-only Kalman filter [15] refines the stationary emitter estimate across repositioning snapshots.

III. CONTROLLER COMPARISON

The RL controllers are selected to span independent versus centralized-training decentralized execution (CTDE) approaches and off-policy versus on-policy learning. Centralized approaches are generally acceptable for multi-agent problems [6] to further encourage system cooperation. Further, on-policy approaches are typically used to improve stationarity [7] in dynamic environments.

A Deep Q-Network (DQN) [16] is an independent, off-policy value-based method using replay and target networks. Asynchronous Advantage Actor Critic (A3C) [17] is an independent, on-policy actor-critic method using short policy rollouts. Multi-agent Deep Deterministic Policy Gradient (MADDPG) [18] uses decentralized actors trained with a centralized critic over joint observations and actions. Multi-agent Proximal Policy Optimization (MAPPO) [19] uses a shared actor and centralized value function with PPO-style updates over longer rollouts. Table 1 categorizes the RL controllers based on their primary architectural elements.

In addition to the RL approach, we compare a non-learning

Table 1: Reinforcement learning architecture summary.

Architecture	Off-Policy	On-Policy
Independent	DQN	A3C
CTDE	MADDPG	MAPPO

controller that greedily selects the action that most reduces the real-time SEP uncertainty metric (SEPG) [1]. Each controller maintains the same observation states, including normalized relative UAV geometry, per-agent RSS and TDOA features, the current location estimate, and SEP. The UAV geometry control actions are 3-D displacements. The system is rewarded based on uncertainty reduction, penalized by the number of steps to localization convergence, and during training, improvement in localization error.

IV. EXPERIMENTAL SIMULATION AND RESULTS

Experiments use eight identical UAV agents in a Gymnasium-simulated [20] 3-D, 2 km dense multipath volume. The RL policies follow standard implementations, with hyperparameters tuned and optimized for this localization scenario. Each policy is trained and evaluated for 1000 Monte Carlo episodes using the same sensing, motion, and environmental error models [14], where initial UAV and emitter locations are randomized for every episode. The evaluation metrics are final localization error, error reduction, repositioning steps to convergence, median movement distance per UAV, and final SEP, summarized in Table 2. Figure 1 compares the stepwise SEP reduction, error reduction, and localization-error interquartile range (IQR) as a measure of controller stability across episodes.

All controllers reduce SEP within the first several repositioning steps, confirming that geometry control improves localization relative to static sensing. SEPG reduces uncertainty early but plateaus because it optimizes the immediate SEP metric without planning through longer-term geometry changes. MAPPO improves more slowly and produces larger error IQR, suggesting that the longer on-policy rollouts are less effective in this real-time, continuous

Table 2: Model performance during policy evaluation.

Parameters	SEPG	DQN	A3C	MADDPG	MAPPO
Final error (m)	24	17	16	17	25
Error reduction	38%	61%	61%	59%	34%
Steps to convergence	11	8	12	7	19
Distance moved (m)	349	312	316	431	850
Final SEP (m)	6	5	5	5	6

feedback geometry-control setting.

DQN, A3C, and MADDPG provide the strongest final localization performance. A3C reaches the lowest final median localization error but with a larger convergence step count, consistent with steadier incremental policy updates. MADDPG achieves low error with fewest steps, but requires much longer distance moved by the UAVs. DQN provides the best overall efficiency, with high error reduction in few steps and shorter distance traveled.

These results indicate that cooperative RL can improve UAV geometry for emitter localization, but centralized training does not always provide the best tradeoff. In this setting, the estimator provides a compact shared uncertainty signal through the solution cloud and SEP metric. As a result, independent policies can learn useful cooperative behavior without requiring a centralized critic. The best practical tradeoff is therefore not just final error, but must consider system resources in convergence steps and movement cost. Under these conditions, DQN gives the strongest efficiency, A3C gives steady and stable improvement, and MADDPG trades additional movement for faster convergence. Overall, the comparison shows that learned geometry control provides a measurable improvement over a policy-free controller for cooperative RF emitter localization in dense multipath. Further, the study concludes that the structure of the localization problem matters for RL choice: when the estimator provides frequent uncertainty feedback, independent off-policy learning can be sufficient to produce fast and efficient geometry adaptation.

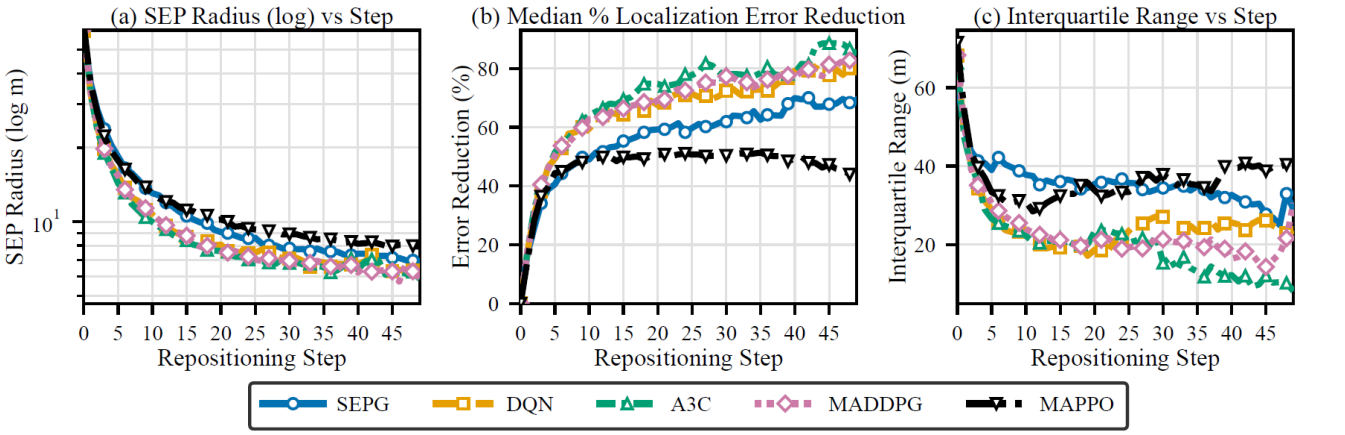

Figure 1. a) Comparison of SEP radius convergence across different policies; b) relative improvement in localization accuracy across repositioning steps; c) uncertainty in localization estimates across episodes.

V. REFERENCES

[1] 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.

[2] W. Wang, et al., “Optimal Configuration and Path Planning for UAV Swarms Using a Novel Localization Approach,” Applied Sciences, vol. 8, no. 6, Art. 973, 2018, doi: 10.3390/app8060973.

[3] S. Aditya, et al., “A Survey on the Impact of Multipath on Wideband Time-of-Arrival-Based Localization,” IEEE Signal Processing Magazine, vol. 35, no. 4, pp. 59–89, 2018, doi: 10.1109/MSP.2018.2818159.

[4] A. Guerra, et al., "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.

[5] 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.

[6] 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.

[7] D. Wei, L. Zhang, and Q. Liu, “UAV Swarm Cooperative Dynamic Target Search: A MAPPO-Based Discrete Optimal Control Method,” Drones, vol. 8, no. 6, Art. 214, 2024, doi: 10.3390/drones8060214.

[8] B. Li, D. Yan, D. Tang, and Z. Wang, “Multi-UAV Roundup Strategy Based on Deep Reinforcement Learning with Curriculum Experience Learning,” Expert Systems with Applications, vol. 238, Art. 121663, 2024, doi: 10.1016/j.eswa.2023.121663.

[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] 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.

[11] R. O. Schmidt, "Least Squares Range Difference Location," IEEE Transactions on Aerospace and Electronic Systems, vol. 32, no. 1, pp. 234-242, 1996

[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] 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. https://repository.library.noaa.gov/view/noaa/23141.

[14] 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.

[15] D. Kim, J. Ha; K. You. “Adaptive extended Kalman filter based geolocation using TDOA/FDOA”. Int. J. Control Autom. 2011, 4, 49–58.

[16] V. Mnih, K. Kavukcuoglu, D. Silver. et al. Human-level control through deep reinforcement learning. Nature 518, 529–533 (2015). https://doi.org/10.1038/nature14236.

[17] R. Lowe, et al. "Multi-agent actor-critic for mixed cooperative-competitive environments.” Advances in neural information processing systems 30 (2017).

[18] V. Mnih, et al. "Asynchronous methods for deep reinforcement learning." International conference on machine learning. PmLR, 2016.

[19] Yu, Chao, et al. "The surprising effectiveness of ppo in cooperative multi-agent games." Advances in neural information processing systems 35 (2022): 24611-24624.

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