Design of RSSI-Only UAV Path Planning for Search and Rescue: Greedy, Prudent, and DRQN-Based Algorithms in GPS-Denied Mountains

Jiwoong Jeon
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
jjw000628@knu.ac.kr

Hyerim Jeon
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
jhr020528@knu.ac.kr

Jonghyeon Bae
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
bae.jh.dan@gmail.com

Juyeol Park
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
qkrjuyeol@gmail.com

Hoki Baek (Corresponding author)
School of Computer Science and Engineering
Kyungpook National University
Daegu, Korea
neloyou@knu.ac.kr

***Abstract*—This paper proposes three algorithms for received signal strength indicator (RSSI)-only UAV path planning in SAR for GPS-denied mountainous environments. RSSI is attractive for early path planning due to low energy cost, implementation simplicity, and ubiquity, but severe shadowing causes large fluctuations and unstable localization. To cope with this, we design: (i) a Greedy method that chooses the initial heading via the strongest RSSI and proceeds while adjusting using prior branch comparisons; (ii) a Prudent method that performs an initial cross-shaped scan and fork-shaped local searches at each decision point for reliability under heavy shadowing; and (iii) a deep recurrent Q-network (DRQN)-based method that adaptively selects between Greedy and Prudent at each step according to changing shadowing. Simulations show the DRQN approach balances both strategies, matching Prudent's high success rate while maintaining decision-step counts close to Greedy, yielding strong adaptability and robustness in challenging shadowing conditions.**

***Index Terms*—Path planning, Search and rescue, UAV, DRQN**

I. Introduction

In disaster or missing-person search scenarios, rapid search and rescue (SAR) operations are essential for saving lives. Unmanned aerial vehicles (UAVs) are increasingly adopted in SAR missions because their easy deployment and airborne capability enable them to establish a stable Line-of-Sight (LOS) link with ground targets, ensuring wide-area coverage and flexible mobility [1], [2]. However, in disaster scenarios, communication infrastructures are often damaged, or the target may be located in regions that require non-terrestrial network (NTN) connectivity due to the unavailability of terrestrial networks, making GPS-based localization unreliable or even infeasible [2], [3], [4]. Similarly, camera-based detection is advantageous for precise localization once the target's approximate position has been identified. Still, its performance deteriorates in long-range searches or environments such as forests and mountainous areas, where LOS is restricted [2], [5]. To address these challenges, wireless signal–based localization can be leveraged as a supplementary approach to improve the efficiency of UAV-based SAR missions [2]. In particular, the received signal strength indicator (RSSI) offers advantages such as low energy consumption and simplicity, and it can be easily obtained from most wireless communication devices [2], making it a practical signal source for guiding UAVs during the early stages of path planning.

In this study, we focus on search and rescue operations that rely solely on RSSI measurements in GPS-denied mountainous environments and propose three UAV path planning algorithms: Greedy, Prudent, and a deep recurrent Q-network (DRQN)-based algorithm. The Greedy algorithm quickly determines a movement direction, storing and using the most substantial RSSI value, but it becomes vulnerable to severe shadowing loss. The Prudent algorithm performs systematic cross and fork searches to ensure robust localization under high signal variability, though at the cost of longer trajectories. The DRQN-based algorithm dynamically selects between

Greedy and Prudent strategies according to environmental shadowing conditions. By leveraging temporal RSSI patterns through recurrent learning, the DRQN-based approach improves both success rate and efficiency, providing an effective solution for UAV path planning in RSSI-only, GPS-denied search and rescue scenarios.

Several studies have investigated SAR systems that utilize UAVs and wireless signal measurements, as presented in [2], [4]–[7]. Studies [2], [4], [5] focus on localization techniques in environments without GPS or other positional information. In [2], a Kalman filter is applied to preprocess RSSI data, and the position is estimated from the gradient of the filtered RSSI. However, this approach requires a predefined flight path and prior knowledge of the UAV's initial heading. In [4], wireless power transfer (WPT) is used to activate powered-off devices temporarily, and the positions of nearby nodes are determined through trilateration. This work does not address UAV path planning and does not apply to medium- or long-range SAR missions. In [5], a random forest algorithm is used to roughly localize the transmitter by allowing the UAV to move randomly and collect RSSI measurements at multiple positions. Since the UAV estimates the target's position only after collecting signals from various locations rather than updating it sequentially, this method requires a larger number of movements. Studies [6] and [7] estimate relative positions based on RSSI measurements and use GPS data as auxiliary information to refine the absolute localization results.

In contrast, our study focuses on a GPS-denied environment and proposes RSSI-only path planning algorithms that directly guide the UAV toward the target without multilateration or external sensors.

II. The Proposed UAV Path Planning Algorithms

A. Proposed Method 1: Greedy Algorithm

The Greedy algorithm begins with an initialization phase in which the UAV performs exploratory movements and measures RSSI in 8 directions to establish the initial heading that maximizes RSSI.

The Greedy algorithm continuously moves toward the strongest RSSI direction from the current position. At each decision step, the UAV measures signal strength at three candidate directions: left, center, and right relative to its current heading direction. The candidate directions are computed as $\theta_{\text{left}} = (\theta_{\text{current}} - 45^\circ) \mod 360^\circ$, $\theta_{\text{center}} = \theta_{\text{current}}$, and $\theta_{\text{right}} = (\theta_{\text{current}} + 45^\circ) \mod 360^\circ$, where each candidate position is one step away from the current location. The algorithm selects the direction with the highest measured RSSI value among the three candidates and the current position. If the current position yields the strongest signal, the UAV remains stationary, indicating that it has converged to a local maximum.

B. Proposed Method 2: Prudent Algorithm

The Prudent algorithm is a finite state machine composed of three sequential states: Wandering, Phase 1 (Cross Search), and Phase 2 (Fork Search). Unlike the Greedy algorithm, it measure RSSI strength in multiple directions, allowing stable navigation even in environments with severe shadowing loss. The prudent algorithm operates as a finite state machine composed of three sequential states: Wandering, Phase 1 (Cross Search), and Phase 2 (Fork Search). The Wandering state is executed only once at the beginning to detect the initial signal, after which Phase 1 and Phase 2 are repeated alternately throughout the search process.

In the Wandering state, the UAV traverses a square spiral pattern with incrementally increasing leg lengths ($L_0 = 20$ m, incremented by 10 m per cycle) until receiving an RSSI measurement above the detection threshold. Upon detection, the algorithm transitions to Phase 1.

Phase 1 evaluates RSSI at four orthogonal directions $(0^\circ, 90^\circ, 180^\circ, 270^\circ)$ from the detection point. It identifies the pair of adjacent directions (e.g., 0° and 90°) with the highest summed signal strength. Based on this pair, it selects the diagonal direction (e.g., 45°) centered between them as the initial forward direction for Phase 2.

Phase 2 employs a three-way fork pattern: at each step, the UAV measures RSSI at three positions offset by -90°, 0°, and $+90^\circ$ relative to the current diagonal forward direction. It then compares the (left + center) RSSI sum against the (center + right) RSSI sum. This comparison determines the next step. The UAV selects a new diagonal path by shifting 45° left or right toward the stronger signal.

C. Proposed Method 3: DRQN-based Algorithm

The proposed DRQN model incorporates a confidence estimator to balance the advantages and disadvantages of the Greedy and Prudent algorithms. DRQN is designed to make decisions by considering both past observation and the current state.

The SAR problem is formulated as a Markov Decision Process (MDP) defined by the tuple:

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mathcal{R}, \gamma) \tag{1}$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the transition function, and $\mathcal{R} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}$ is the reward function.

1) State: The state vector $s_t \in \mathbb{R}^{11}$ is carefully designed to represent both the current signal conditions and the historical movement behavior of the UAV. It includes the normalized RSSI value at the current step and its temporal difference from the previous step, allowing the agent to recognize whether the signal strength is increasing or decreasing. These features help the network detect short-term signal trends caused by distance or shadowing effects.

In addition to the signal information, the state vector contains several elements describing the UAV's recent motion patterns. It includes the path straightness coefficient to represent how consistent the UAV's movement direction has been, the flip count to measure how often the heading direction has been reversed, and an idle step counter to identify periods when the UAV is stuck or not making progress.

Furthermore, the state vector includes the distance from the best RSSI position recorded so far, the last action taken (*Greedy* or *Prudent*), and a phase indicator inherited from the rule-based Prudent finite state machine (FSM).

Together, these elements enable the DRQN agent to recognize both instantaneous signal variations and accumulated search patterns over time, improving its ability to choose between rapid exploration and cautious adjustment depending on environmental conditions.

2) Actions: The action space is discrete with two options:

$$\mathcal{A} = \{0 : \text{Greedy}, 1 : \text{Prudent}\} \tag{2}$$

Action $a_t = 0$ invokes the Greedy algorithm for one navigation step, while $a_t = 1$ advances the prudent multi-phase FSM by one state transition.

3) Reward function: The reward function balances exploration efficiency, path quality, and convergence speed:

$$r_t = r_{\text{base}} + r_{\text{progress}} + r_{\text{quality}} + r_{\text{terminal}} \tag{3}$$

The base reward incorporates a Greedy-first bonus (0.005 when $a_t = 0$), movement penalty $c_{\text{move}} = 0.01$, NRI penalty $c_z Z_t$ where

$$Z_t = 0.6 f_t + 0.4(1 - \kappa_t) \tag{4}$$

and switch penalty $c_{\text{switch}} = 0.01$ when $a_t \neq a_{t-1}$.

The progress reward provides dense feedback proportional to normalized distance reduction:

$$r_{\text{progress}} = w_{\text{prog}} \cdot \frac{d(p_{t-1}, p_v) - d(p_t, p_v)}{\|p_t - p_{t-1}\|}, \quad w_{\text{prog}} = 0.2 \tag{5}$$

Quality penalties include an escalating idle penalty $c_{\text{idle}} \cdot n_{\text{idle}}$ when $\Delta\rho_g \leq 0$, a stagnation bonus (0.05) for selecting prudent when $n_{\text{idle}} \geq 3$, and a proximity reward when $d_t < 1.5 d_{\text{term}}$:

$$r_{\text{proximity}} = 0.05 \cdot \left(\frac{1.5 d_{\text{term}} - d_t}{0.5 d_{\text{term}}} \right) \tag{6}$$

Terminal rewards provide episodic feedback. Success within $d_{\text{term}} = 30$ m yields:

$$r_{\text{terminal}} = r_{\text{success}} + \beta_{\text{end}} \max(0, 1.5 - 0.01t) \tag{7}$$

where $r_{\text{success}} = 10.0$ and $\beta_{\text{end}} = 1.5$. Timeout after $T_{\max} = 300$ decisions incurs:

$$r_{\text{terminal}} = r_{\text{timeout}} - \min(2.0, 0.02 d_t / s_{\text{step}}) \tag{8}$$

where $r_{\text{timeout}} = -10.0$, scaled by remaining distance to differentiate near-misses from complete failures.

4) Transition Dynamics: The transition function $\mathcal{T}(s_{t+1}|s_t, a_t)$ is determined by the environment dynamics and exhibits stochasticity due to shadowing loss. When an action $a_t \in \{0 : \text{Greedy}, 1 : \text{Prudent}\}$ is selected:

1) The corresponding basic algorithm (Greedy or Prudent) executes one navigation step, updating the UAV's position from p_t to p_{t+1}.
2) A new signal strength RSSI_{t+1} is measured at p_{t+1}, sampled according to the log-distance path loss model with Gaussian shadowing $X_\sigma \sim \mathcal{N}(0, \sigma_s^2)$, where $\sigma_s \in [4, 8, 12]$ dB.
3) The new state s_{t+1} is computed based on the updated position, new RSSI measurement, and historical observation window.

The environment's stochasticity is governed entirely by the shadowing term X_σ.

5) Objective and Episode Termination: The agent's objective is to find a policy $\pi(a_t|s_t)$ that maximizes the expected discounted return:

$$J(\pi) = \mathbb{E}_\pi \left[\sum_{t=0}^{T_{\max}} \gamma^t r_t \right] \tag{9}$$

where γ is the discount factor, balancing immediate and future rewards.

An episode terminates when either of the following conditions is met:

- **Success:** The UAV reaches within 30 m of the target location.
- **Timeout:** The maximum number of decisions (300 steps) is reached.

Terminal rewards are issued only upon episode termination, as defined in the reward function.

III. Deep Recurrent Q-Network for Path Planning

A. Deep Recurrent Q-Network Architecture and Training

The DRQN agent employs a recurrent dueling architecture to process sequential observations and filter shadowing noise. The network consists of: (1) two fully-connected feature extraction layers with layer normalization (128 units each), (2) a single-layer GRU with 256 hidden units to maintain temporal context, and (3) separate value and advantage streams for dueling Q-value decomposition. The feature extractor processes raw state vectors $s_t \in \mathbb{R}^{11}$, producing intermediate representations $\mathbf{h}_2^t \in \mathbb{R}^{128}$. The GRU layer propagates hidden states across timesteps, enabling the network to recognize temporal patterns such as local minima or oscillatory behavior:

$$\mathbf{h}_t^{\text{GRU}} = \text{GRU}(\mathbf{h}_2^t, \mathbf{h}_{t-1}), \quad \mathbf{h}_t^{\text{GRU}} \in \mathbb{R}^{256} \tag{10}$$

where $\mathbf{h}_{t-1}$ is the hidden state vector from the previous timestep, which carries the network's memory of past observations. The dueling streams estimate state value $V(s_t; \theta) \in \mathbb{R}$ and action advantages $\mathbf{A}(s_t, a; \theta) \in \mathbb{R}^2$, with Q-values computed via advantage centering:

$$Q(s_t, a; \theta) = V(s_t; \theta) + \Big(\mathbf{A}(s_t, a; \theta) - \frac{1}{|\mathcal{A}|} \sum_{a'} \mathbf{A}(s_t, a'; \theta) \Big) \tag{11}$$

Training utilizes Double DQN with sequential experience replay, storing complete trajectories (capacity: 100,000). Mini-batches sample $B = 32$ episodes, extracting contiguous subsequences of length $L \sim \text{Uniform}(4, 16)$ with reset hidden

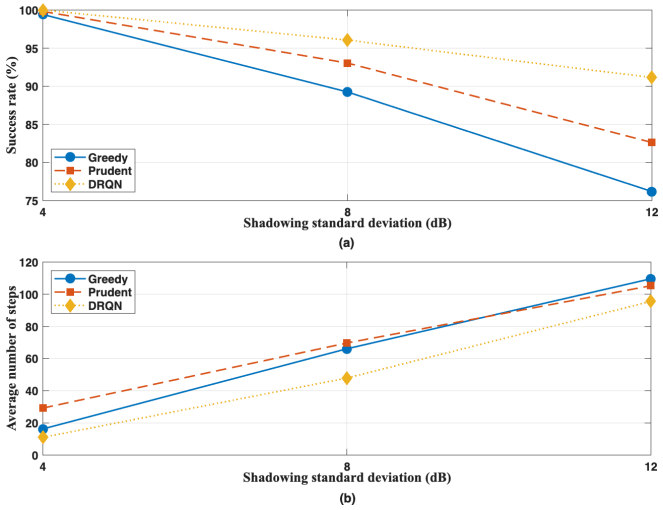

Fig. 1. Simulated results over 10,000 test episodes for $\sigma = \{4, 8, 12\}$ dB comparing the Greedy, Prudent, and DRQN-based algorithms: (a) success rate; (b) average number of steps

states $\mathbf{h}_0 = \mathbf{0}$ for truncated backpropagation. Target Q-values follow:

$$y_t = r_t + \gamma \cdot Q(s_{t+1}, \arg\max_{a'} Q(s_{t+1}, a'; \theta); \theta^-) \tag{12}$$

The network is trained by minimizing the Smooth L1 (Huber) loss, which reduces sensitivity to outliers:

$$\mathcal{L}(Q, y) = \begin{cases} \frac{1}{2}(Q - y)^2 & \text{if } |Q - y| < 1 \\ |Q - y| - \frac{1}{2} & \text{otherwise} \end{cases} \tag{13}$$

with gradient clipping (max-norm 1.0) and Adam optimizer ($\alpha = 5 \times 10^{-5}$). The target network updates via Polyak averaging:

$$\theta^- \leftarrow \tau_{\text{soft}}\theta + (1 - \tau_{\text{soft}})\theta^-, \quad \tau_{\text{soft}} = 0.001 \tag{14}$$

Exploration follows ϵ-Greedy with linear decay from $\epsilon_0 = 1.0$ to $\epsilon_{\min} = 0.05$ over 300,000 steps, combined with episode-based curriculum capping for late-phase stability. After $N_{\text{train}} = 500$ episodes, the learned policy deploys with $\epsilon = 0$:

$$\pi^*(s_t) = \arg\max_{a \in \mathcal{A}} Q(s_t, a; \theta^*, \mathbf{h}_{t-1}) \tag{15}$$

maintaining recurrent state $\mathbf{h}_{t-1}$ for context-aware strategy selection.

IV. Simulations and Results

A. Simulation Environment

We consider a single-UAV single-target search task in a forest-like outdoor environment with log-normal shadowing. We evaluated three shadowing standard deviations, $\sigma \in \{4, 8, 12\}$ dB, with 10,000 test episodes for each σ. At the beginning of each episode, the UAV starts from a fixed geodetic position. The target location is uniformly sampled from an annulus centered at the start of the UAV, with inner and outer radii of 100 m and 120 m, respectively.

Episode seeds are paired across methods so that initial conditions, shadowing realizations, and per-step RSSI generation are identical. We do not use domain randomization across deviations. Instead, we train a separate recurrent policy for each σ. For a given σ, training runs for $5,000$ episodes using recurrent temporal-difference learning with hidden state carried within episodes, ϵ-Greedy exploration over the two macro actions, and periodic target-network updates. All other optimization and replay settings are kept fixed for all deviation levels.

B. Simulation Results

We evaluated the following two metrics for each shadowing deviation σ. The first metric is the success rate, which is the fraction of successful episodes among total test episodes. The second metric is the average number of steps in successful episodes.

1) Success rate: Fig. 1(a) illustrates the success rate of each algorithm, evaluated according to the success criterion described in Section II-C(5).

When $\sigma = 4$, all three algorithms achieve a success rate of over 99%, demonstrating stable path-finding performance under mild shadowing loss. In particular, the DRQN-based algorithm achieves a 100% success rate, indicating that it consistently selects the correct path in all episodes by choosing the Prudent macro action when sudden fluctuations in RSSI signal strength occur. As the shadowing deviation increases, all algorithms exhibit a gradual decline in success rate. When $\sigma = 8$, the Greedy algorithm achieves 89%, the Prudent algorithm achieves 93%, and the DRQN-based algorithm maintains a robust performance with 96% success rate. Under severe shadowing loss with $\sigma = 12$, the success rate of the Greedy algorithm decreases to 76%. This degradation occurs because the Greedy algorithm is highly dependent on its previous decisions; once it deviates due to severe shadowing loss, it tends to continue following an incorrect path, resulting in frequent timeouts. The Prudent algorithm achieves a higher success rate of 82% compared to Greedy, yet its frequent use of cross-search and fork-search phases under severe shadowing loss increases the total number of decisions beyond $T_{\max}$. In contrast, the DRQN-based algorithm achieves 91% success rate by dynamically switching between the Greedy and Prudent strategies, successfully balancing exploration and exploitation to construct an effective path.

2) Average number of steps: Fig. 1(b) shows the average number of decision steps for the three proposed algorithms. The step count increases monotonically with the shadowing deviation σ. For $\sigma = 4$ and 8 dB, the ordering is consistent: Prudent incurs the largest number of steps, followed by Greedy, and then the DRQN-based algorithm. This is expected since Prudent's conservative cross-/fork-search phases lead to longer trajectories, whereas Greedy can be faster but still pays a nontrivial backtracking cost when it commits to a suboptimal heading. The DRQN-based policy reduces such detours by selectively switching macros, yielding fewer steps than Greedy. When $\sigma = 12$ dB, the average step count of

Greedy exceeds that of Prudent. As suggested by Fig. 1(a), under heavy shadowing loss, Greedy more often deviates onto incorrect paths and requires a larger recovery effort to reestablish a correct route, which increases the number of steps among its successful episodes.

V. Conclusion

This study compared three RSSI-based UAV path planning algorithms for SAR in GPS-denied mountainous environments. The Greedy algorithm was fast but sensitive to shadowing, while the Prudent algorithm was stable but slower. The DRQN-based method combined both advantages, achieving high success rates like Prudent with fewer decision steps like Greedy, showing strong adaptability under severe shadowing.

VI. Acknowledgment

This research was supported by the Regional Innovation System & Education(RISE) Glocal 30 program through the Daegu RISE Center, funded by the Ministry of Education(MOE) and the Daegu, Republic of Korea.(2026-RISE-03-001). This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualization program grant funded by the Korea government(MSIT)(IITP-2025-RS-2022-00156389). This work was supported by the SME Technology Innovation Development Program (Market-Responsive Type) grant funded by the Korea government(Ministry of SMEs and Startups)(RS-2025-25455183). This study was supported by the BK21 FOUR project (AI-driven Convergence Software Education Research Program) funded by the Ministry of Education, School of Computer Science and Engineering, Kyungpook National University, Korea (41202420214871).

References

[1] A. Albanese, V. Sciancalepore and X. Costa-Pérez, "First Responders Got Wings: UAVs to the Rescue of Localization Operations in Beyond 5G Systems," IEEE Communications Magazine, vol. 59, no. 11, pp. 28-34, Nov. 2021.

[2] Y. Sun, X. Wen, Z. Lu, T. Lei and S. Jiang, "Localization of WiFi Devices Using Unmanned Aerial Vehicles in Search and Rescue," 2018 IEEE/CIC International Conference on Communications in China (ICCC Workshops), Beijing, China, pp. 147-152, Aug. 2018.

[3] E. Ngo, J. Ramirez, M. Medina-Soto, S. Dirksen, E. D. Victoriano and S. Bhandari, "UAV Platforms for Autonomous Navigation in GPS-Denied Environments for Search and Rescue Missions," 2022 International Conference on Unmanned Aircraft Systems (ICUAS), Dubrovnik, Croatia, pp. 1481-1488, Jun. 2022.

[4] M. Atif, R. Ahmad, W. Ahmad, L. Zhao and J. J. P. C. Rodrigues, "UAV-Assisted Wireless Localization for Search and Rescue," IEEE Systems Journal, vol. 15, no. 3, pp. 3261-3272, Sep. 2021.

[5] V. Acuna, A. Kumbhar, E. Vattapparamban, F. Rajabli and I. Guvenc, "Localization of WiFi Devices Using Probe Requests Captured at Unmanned Aerial Vehicles," 2017 IEEE Wireless Communications and Networking Conference (WCNC), San Francisco, CA, USA, pp. 1–6, Mar. 2017.

[6] I. Prata, A. da Silva Siqueira Almeida, F. C. de Souza, P. F. F. Rosa and A. F. P. dos Santos, "Developing a UAV platform for target localization on search and rescue operations," 2022 IEEE 31st International Symposium on Industrial Electronics (ISIE), Anchorage, AK, USA, pp. 721-726, Jun. 2022.

[7] Z. Liu, Y. Chen, B. Liu, C. Cao and X. Fu, "HAWK: An Unmanned Mini-Helicopter-Based Aerial Wireless Kit for Localization," IEEE Transactions on Mobile Computing, vol. 13, no. 2, pp. 287-298, Feb. 2014.