航測及遙測學刊　第三十卷　第 4 期　第 255 – 267 頁　民國 114 年 12 月
Journal of Photogrammetry and Remote Sensing
Volume 30, No.4, 2025, pp. 255 – 267
DOI：10.6574/JPRS.202512_30(4).0002

UAV Path Determination for Visual Navigation using Deep Reinforcement Learning

Pei-Hsuan Huang [1*]　Chao-Hung Lin [2]

Abstract

Traditional path planning algorithms for Unmanned Aerial Vehicles (UAVs) primarily optimize for geometric metrics such as path length and energy efficiency. However, in GPS-denied environments, where external positioning is unreliable, the quality of visual localization is paramount for mission success. This study introduces a novel Deep Reinforcement Learning (DRL) framework designed to co-optimize the UAV path for both geometric efficiency and visual localization robustness. Specifically, our method integrates the density of matched image feature points, extracted from post-processed aerial imagery, directly into the planning process, ensuring the generated trajectory passes through visually rich areas that enhance navigation accuracy. To tackle the path planning challenge and address issues related to sparse rewards and unstable training, we employ an advanced DRL architecture: Noisy Dueling Double DQN with Prioritized Experience Replay (Noisy D3QN with PER). This integration leverages Double DQN to refine value estimation, Dueling DQN to improve generalization, PER to enhance sample efficiency, and Noisy Networks to promote robust and efficient exploration. The proposed framework is implemented within a simulated 2.5D environment with a customized reward function that considers both UAV state parameters and terrain features. Experimental results demonstrate that the method generates efficient, visually coherent, and dynamically smooth trajectories. Crucially, it enables path inference for multiple independent missions from various starting points after a single training session, achieving superior computational efficiency compared to traditional geometric planners. This highlights the potential of integrating visual features into a reinforcement learning-based UAV path planning to significantly enhance visual localization performance in complex environments.

Keywords: Reinforcement Learning, Unmanned Aerial Vehicle, Deep Q Network

1. Introduction

In recent years, drone warfare has emerged as a prominent trend in modern military operations, largely due to the numerous advantages offered by Unmanned Aerial Vehicles (UAVs). UAVs facilitate the execution of high-risk missions without placing personnel in direct danger, thereby significantly reducing battlefield casualties. Moreover, their modular architecture and compact design allow for rapid deployment across diverse terrains and tactical scenarios, enhancing operational flexibility. Given that many missions are conducted within adversarial territories, robust and reliable navigation systems are critical to ensuring both positional accuracy and overall mission success. UAVs primarily depend on the Global Positioning System (GPS) for positioning and navigation. However, GPS signals are inherently vulnerable to jamming, spoofing, and other electronic countermeasures frequently employed by hostile forces. As a result, the failure or disruption of UAV navigation systems may lead to mission deviation, target misidentification, and potential loss of the platform.

To address scenarios in which GPS signals are unavailable, such as during wartime or in GPS denied environments, visual navigation has emerged as a viable alternative to support UAV localization and sustain mission continuity. This method typically utilizes aerial imagery captured by UAV cameras, alongside reference orthoimages to perform image matching. By extracting feature conjugate points through image matching, the Exterior Orientation Parameters (EOP), include the camera's spatial position and orientation, can be estimated via spatial resection. This enables accurate localization in the absence of GPS.

To extract reliable and repeatable feature correspondences, recent advances in deep learning-based image matching models have played a pivotal role. SuperPoint (Detone *et al.*, 2018) is a self-supervised interest point detector and descriptor that jointly learns keypoint locations and descriptors using a fully convolutional neural network, enabling robust and efficient feature extraction in aerial imagery. LightGlue (Lindenberger *et al.*, 2023), a lightweight graph-matching framework, complements SuperPoint

[1] Master Student, Department of Geomatics, National Cheng Kung University
[2] Professor, Department of Geomatics, National Cheng Kung University
[*] Corresponding Author, E-mail: hpei254@gmail.com

Received Date: Jul. 15, 2025
Revised Date: Sep. 26, 2025
Accepted Date: Oct. 16, 2025

by establishing reliable matches between feature points through attention-based architecture and adaptive filtering strategies, making it well-suited for real-time UAV visual localization. These models significantly improve the accuracy and robustness of feature point detection and matching under varying perspectives and illumination conditions, thus enhancing the reliability of visual-based UAV positioning.

When operating without GPS support, path planning becomes a crucial element in guiding the UAV to its target area. In such cases, pre-mission path planning must incorporate image feature points, as they directly influence the robustness and accuracy of visual localization. Additionally, in navigating realistic three-dimensional environments, obstacle avoidance becomes an essential constraint that must be strictly observed.

While traditional methods like Dijkstra's algorithm and A* are widely used for computing optimal paths in static environments, they often struggle with dynamic, high-dimensional spaces characteristic of complex UAV missions. Accordingly, Reinforcement Learning (RL) has emerged as a powerful paradigm for adaptive decision-making through environment interaction. The Deep Q-Network (DQN) framework (Mnih *et al.*, 2015) addresses the limitations of traditional tabular Q-Learning by integrating deep neural networks to approximate value functions, enabling learning in complex environments without exhaustive state enumeration. In the domain of path planning for autonomous systems, DRL is increasingly applied due to its adaptability to complex, constraints-rich environments (Wu *et al.*, 2023; Yao *et al.*, 2022). Specifically for UAV navigation, researchers have leveraged DRL to optimize multi-objective functions, considering factors such as distance, energy consumption, and collision penalties (Wu *et al.*, 2023). However, achieving reliable performance requires mitigating inherent DQN challenges, including Q-value overestimation, unstable learning, and inefficient exploration, which remain critical areas of research.

This research aims to design a UAV path planning system with multiple objectives, including maximizing the density of matched image feature points. Higher feature-point density enhances the chances and precision of UAV localization in signal-denied environments. To this end, an extended DQN framework is employed for path planning, followed by postprocessing with cubic Bézier curves to generate smooth trajectories suitable for real-world UAV deployment.

To address the aforementioned challenges and enhance training stability, several advanced DQN extensions have been introduced and widely studied in UAV applications. Double DQN (Van Hasselt *et al.*, 2016) mitigates the overestimation bias prevalent in standard DQN by decoupling action selection and evaluation. Dueling DQN (Wang *et al.*, 2016) further improves generalization and learning stability by decomposing the Q-value into state-value and advantage components, a structure proven effective in complex UAV scenarios (Yao *et al.*, 2022; Huang and Li, 2023). Furthermore, Prioritized Experience Replay (PER) (Schaul *et al.*, 2016) enhances sample efficiency by focusing training on high-Temporal-Difference error experiences, a technique successfully integrated into path planning for autonomous surface vehicles (Zhu *et al.*, 2021).

Building upon this foundation, our study adopts an advanced hybrid architecture: Noisy Dueling Double DQN with Prioritized Experience Replay (Noisy D3QN with PER). This integration harnesses the stability advantages of the Dueling architecture, alleviates Q-value overestimation through Double DQN, and achieves improved sample efficiency via PER. Crucially, we incorporate Noisy Networks (Fortunato *et al.*, 2018) to replace the conventional ϵ-greedy strategy. By injecting parametric noise into the network weights, this approach fosters state-dependent exploration, which is superior for consistent and adaptive agent behavior in environments characterized by sparse rewards, a methodology relevant to UAV path planning (Villanueva and Fajardo, 2019).

Since our objective is to assist UAVs in visual navigation and localization, the reward function incorporates the density of matched feature points, UAV altitude, and obstacle constraints. This reward design enables the agent to generate feasible and reliable paths. Furthermore, our proposed DQN-based approach supports multi-UAV by generating individual paths from different starting points to a shared destination. This flexibility addresses the limitation of traditional algorithms that must regenerate the path for each new UAV configuration. The evaluation time of the Noisy D3QN with PER method also outperforms traditional approaches, reducing computational resources and improving efficiency.

2. Methodology

This research separates the workflow into three parts, the dataset preprocessing of feature points' density map, the comprehensive deep Q network model for path finding, and the post-processing for smoothing the path, shown as Figure 1 below.

2.1 Dataset and Pre-Processing

The dataset comprises two principal components. The first component is a simulated three-dimensional environment based on the DSM of National Cheng Kung University. This environment serves as the training backdrop for UAV pathfinding and provides the foundation for acquiring image matching pairs. The second component consists of matched feature points, which are essential components guiding the UAV's pathfinding process. The matched feature points were generated utilizing SuperPoint (Detone *et al.*, 2018) for keypoint detection and LightGlue (Lindenberger *et al.*, 2023) for matching, applied to UAV acquired images alongside their corresponding reference satellite images, culminating in a robust dataset of matched points that serves as a vital visual guidance mechanism during model training. An illustrative example of a single image matching is presented in Figure 2.

Given the restricted field of view inherent to UAV imagery and the uneven distribution of feature points, directly employing raw matched pairs as model input are not suitable. To ameliorate this issue, a density map was devised to statistically depict the spatial distribution of feature points across the DSM domain. The density map construction commenced with spatial clipping of all matched points to the bounds of the DSM. Spanning an area of 1 km × 1 km with a resolution of 10 meters. Each matched point was assigned to a specific grid cell according to its easting and northing coordinates. Instead of merely counting the distribution of points within each grid cell, a Gaussian kernel density estimation was applied to render a smoother representation of each point's contribution. Each point imparted weighted values w to its neighboring grid cells, employing a Gaussian function centered at the respective point's location, characterized by a predetermined standard deviation σ=1.5 and effective radius of ±3σ. The weight for each cell was computed using the following equation:

$$w = e^{-\frac{dx^2+dy^2}{2\sigma^2}} \quad \text{..} \quad (1)$$

2.2 DQN Pathfinding

During the navigation process, the UAV must acquire a DSM of the target area, alongside a set of matching pairs of captured images within that region. This preprocessing of path planning plays a crucial role in optimizing the UAV's navigation route, enabling it to adapt to its mobility constraints while simultaneously avoiding potential collisions with the terrain or obstacles, such as buildings. This approach not only enhances the safety of subsequent tasks associated with the UAV's operational objectives but also increases the number of identifiable matching pairs, which is a primary goal of the mission. In this study, we employed Deep Q-Networks (DQN) (Mnih *et al.*, 2015) as the foundational framework. This approach synergizes the principles of the Q-learning algorithm with the strengths of deep learning to effectively address the UAV path planning challenge within a simulated three-dimensional environment.

In our implementation, the DQN designates the UAV as the agent, which engages in interactions within the environment. For each state within the environment, upon executing an action, the UAV receives a Q-value as a reward from the neural network. This Q-value serves as an indication of the potential value associated with the current state-action pair, subsequently guiding the UAV to transition to the corresponding next state. The decision-making process persists until the agent successfully reaches the designated target, a scenario referred to as an episode.

In conventional reinforcement learning applications within simpler environments, the Q-value of each state can be represented using a Q-table. However, in more complex environments, DQN utilizes a deep neural network to supplant the traditional Q-table, thereby amplifying the advantages of deep reinforcement learning. The architecture of the deep neural network consists of multiple layers of neurons, where each layer is interconnected with neurons from preceding layers. This structural characteristic enables the DQN to learn from its experiences, facilitating the optimization of policy learning. Consequently, the transformation of state representations into a value-based function emerges as the output, enhancing the overall effectiveness of the policy optimization process. Figure 3 illustrated the whole process of our adopted comprehensive DQN methods.

The key innovations of DQN are the experience replay and the target network. Each interaction that the agent has with the environment generates a series of experiences represented as $(s, a, r, s\,\acute{}\,)$. This notation indicates the current state, the action taken by the agent, the reward received for that action, and the subsequent state following that action, respectively. These past transitions are stored in a replay buffer. During training, random mini-batches are sampled from this buffer in a process known as experience replay. This approach reduces data correlation, enhancing sample efficiency and stabilizing the learning process.

Another significant innovation in DQN is the introduction of the target network. The target network has the same neural network architecture as the primary network that is being trained. However, it operates as a separate entity, generating stable Q-values that serve as targets for calculating the loss in relation to the main network. Functioning effectively as a ground truth, the parameters of the target network are updated periodically from the main network, rather than being updated with every training iteration. This periodic update helps maintain a fixed target for the main network, preventing rapid fluctuations during the learning process.

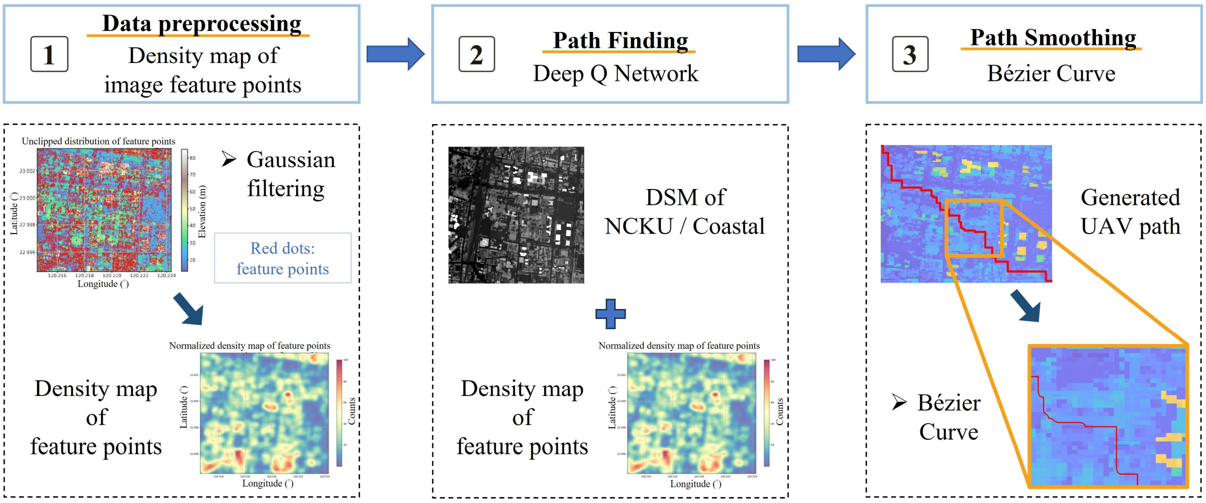

Figure 1 Overview of the path planning workflow

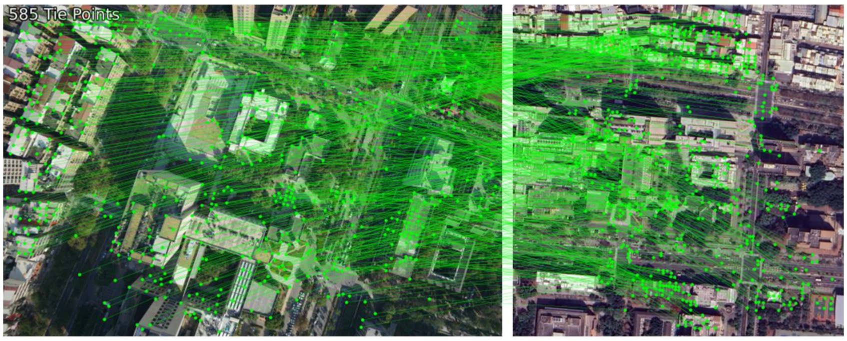

Figure 2 Example of a single image matching result

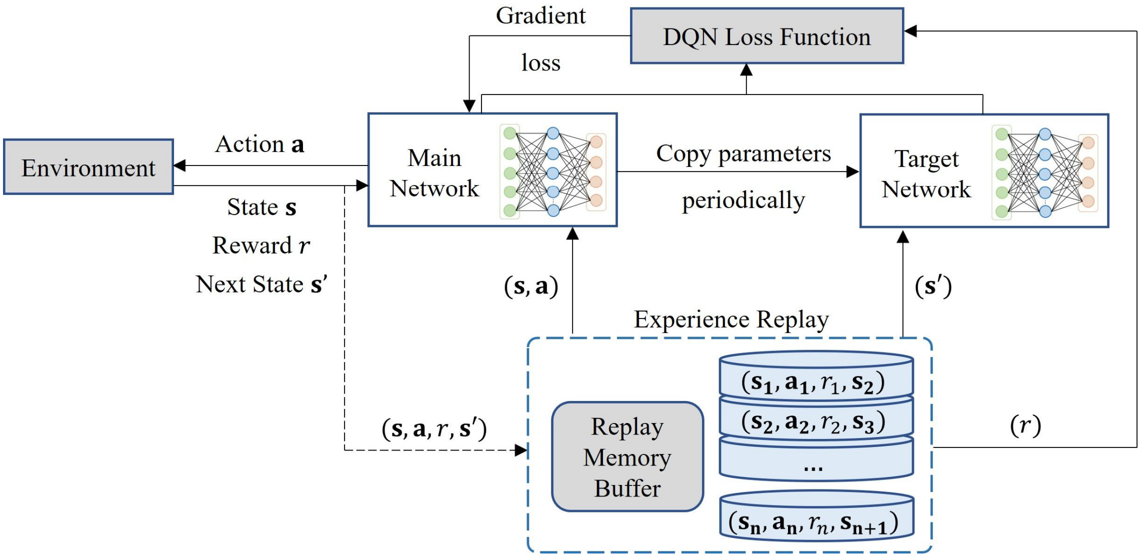
Gradient
loss
DQN Loss Function
Environment
Action **a**
State **s**
Reward r
Next State **s′**
Main Network
Copy parameters
periodically
Target Network
$(\mathbf{s}, \mathbf{a})$
$(\mathbf{s}')$
Experience Replay
$(\mathbf{s}, \mathbf{a}, r, \mathbf{s}')$
Replay Memory Buffer
$(\mathbf{s}_1, \mathbf{a}_1, r_1, \mathbf{s}_2)$
$(\mathbf{s}_2, \mathbf{a}_2, r_2, \mathbf{s}_3)$
...
$(\mathbf{s}_n, \mathbf{a}_n, r_n, \mathbf{s}_{n+1})$
(r)

Figure 3 Workflow of the proposed DQN

To address the inherent limitations of conventional DQN, we adopt an integrated framework that incorporates four established enhancements: Double DQN (Van Hasselt *et al.*, 2016), Dueling DQN (Wang *et al.*, 2016), Noisy Networks (Fortunato *et al.*, 2018), and prioritized experience replay (PER) (Schaul *et al.*, 2016). First, the Double DQN algorithm effectively mitigates the overestimation bias associated with Q-values updates by decoupling the action selection and evaluation processes. In this dual-network architecture, the main network is utilized for action selection, while the target network is responsible for action evaluation. This refinement promotes more accurate value estimations during learning. Second, the Dueling DQN framework distinguishes between the state value function and the advantage function, thereby enhancing the agent's capacity to evaluate states effectively, even when the action choices exhibit minimal variation. This facilitates accelerated convergence by enabling the network to focus on the relative advantages of actions within specific states.

Third, Noisy Networks introduce adaptive exploration strategies by substituting traditional fixed exploration methods, such as epsilon-greedy strategy, with learnable stochastic noise incorporated into the network weights. The incorporation of adaptive noise enhances the exploration-exploitation balance, facilitating more efficient learning. Finally, PER enhances learning efficiency and stability by emphasizing experiences with high temporal-difference (TD) errors. This allows the agent to concentrate on transitions that provide the most significant informative value for policy updates. The collective implementation of these enhancements strengthens both the robustness and performance metrics of the baseline DQN, ultimately leading to superior decision-making capabilities in reinforcement learning environments.

2.3 Network Structure and Reward Function Design

We define the input as a DSM representing a virtual three-dimensional environment. Our network architecture is based on an integrated framework that combines noisy dueling double deep Q-network (D3QN) with PER, as illustrated above. The input consists of two channels: the DSM and a corresponding density map that encodes the spatial distribution of image feature matches. To control the UAV agent, we define six discrete actions that correspond to movement in the cardinal directions of 2.5D space: forward, backward, left, right, up, and down. These actions are treated independently and are not combined into compound vectors. This design prevents ambiguity in the output Q-values, ensuring that each predicted action value clearly corresponds to a single direction on the X, Y, or Z axis. This separation is similar to a classification task, where each class (i.e., direction) is distinctly modeled, which improves both convergence and decision accuracy. The network's output is a set of Q-values for each action, calculated through a dueling architecture that separately estimates the state value and the advantage function. The inclusion of noisy linear layers injects learnable stochasticity into the network, promoting more effective exploration during training. With this architecture, the model is better equipped to evaluate the optimal navigation action in complex 2.5D environments. The schematic diagram of the entire network architecture is shown in Figure 4.

In accordance with the established network architecture, we have designed a bespoke reward function aimed at optimizing the path planning capabilities of the agent within a virtual three-dimensional environment. This reward mechanism is meticulously designed to achieve a harmonious balance among multiple objectives, thereby augmenting both navigation efficacy and the quality of visual localization.

The UAV gent is trained using the DQN framework, and the total reward function R is defined as

$$R = r_{obs1} + r_{obs2} + r_{step} + r_{goal} + r_h + r_{match} \quad \ldots(2)$$

where each term is delineated as follows: r_{obs1}denotes the obstacle avoidance for out of the map searching area, r_{obs2}denotes the observation penalty for colliding with obstacles, r_{step}denotes the step-reward balance, r_{goal}denotes the incentives for goal-oriented navigation, r_h denotes constraints related to excessive flying height, and r_{match} denotes the value associated with matched feature points. This comprehensive approach to reward structuring thus promotes the development of robust navigation strategies while ensuring the agent's adaptability to dynamic operational scenarios.

The function imposes penalties for collisions and for navigating into prohibited areas, thereby promoting effective obstacle avoidance strategies. Concurrently, it provides incentives for minimizing the trajectory length through the reduction of steps required to attain the target. Additionally, the reward structure is strategically formulated to diminish the Euclidean distance between the agent and the designated objective, ensuring that the agent makes consistent progress toward its goal. Significantly, to enhance positioning accuracy, the reward function further integrates the count of matched feature pairs as a positive reinforcement signal. This mechanism encourages the agent to prioritize regions rich in visual correspondences. Through this multi-faceted design, the agent is effectively motivated to assimilate and execute proficient and reliable navigation strategies amidst complex spatial constraints, thereby facilitating superior performance in pathfinding tasks.

2.4 Bézier Curve

To enhance the generated path, Bezier curve (Bezier, 1972) smoothing is used during post-processing to refine the initially rough trajectory. This technique results in a final path that is smoother and more continuous, making it more suitable for real-world UAV deployment. Such refinement is necessary because abrupt directional changes can be impractical due to the constraints of flight dynamics.

3. Experimental Results and Analysis

3.1 Preprocessing Results - Density Map of Image Matching Pairs

In Section 2.1, we provide a comprehensive overview of the dataset and the preprocessing methodologies employed, particularly highlighting the generation of matching pairs of feature points. Utilizing the Gaussian kernel process delineated in Equation 1, we transformed these matching pairs into a density map. Recognizing the significance of matching pairs of feature points in UAV path planning, we proceeded to clip the density map to produce an alternative version that omits a minor subset of the matching pairs. The two resulting datasets are designated as "Unclip" and "Clip," respectively. This methodological step aims to assess the model's sensitivity to variations in matching pairs and to elucidate the role of these pairs in enhancing UAV path planning efficacy. The results arising from this preprocessing phase are compared and presented in Figure 5.

3.2 Simulation Environment Setting

The specification of parameters is a critical component in the training process of reinforcement learning algorithms. Table 1 provides a summary of the essential hyperparameters employed during the training phase. The experiments were conducted within a 2.5D simulation environment that encompassed a 1 km × 1 km area of the National Cheng Kung University campus. A DSM was utilized, configured at a spatial resolution of 10 meters, which resulted in a voxel grid with dimensions of 100 × 100 × 10; the elevation values along the z-axis were normalized by dividing the original height by 10. The UAV commenced its trajectory from the initial coordinates [1,1,3] and was tasked with reaching the target coordinates [98,98,4]. The goal was established as being within a one-pixel radius of the target location, ensuring precision in the UAV's navigation objectives.

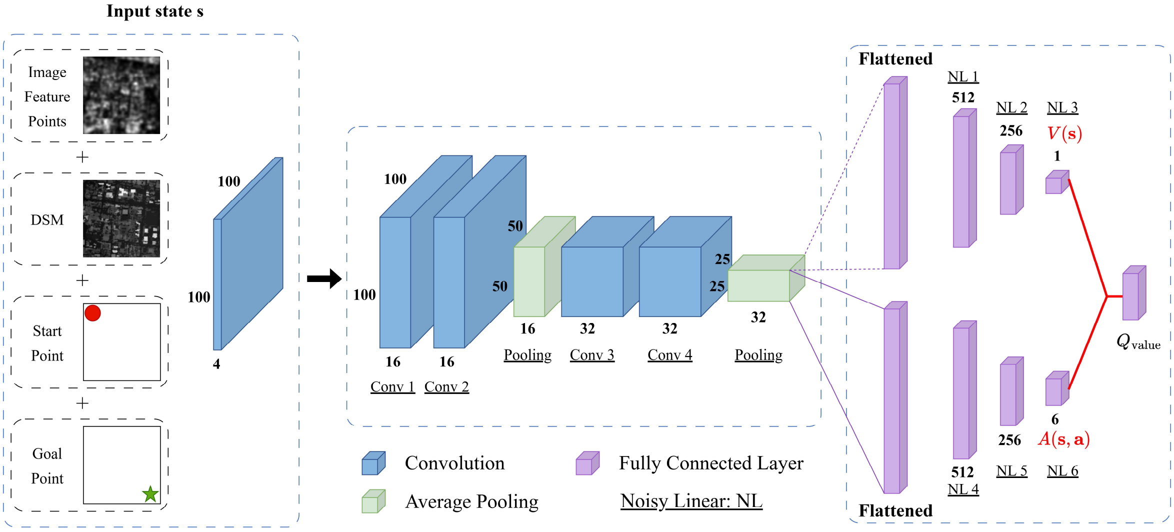

Figure 4 Proposed enforcement learning neural network structure

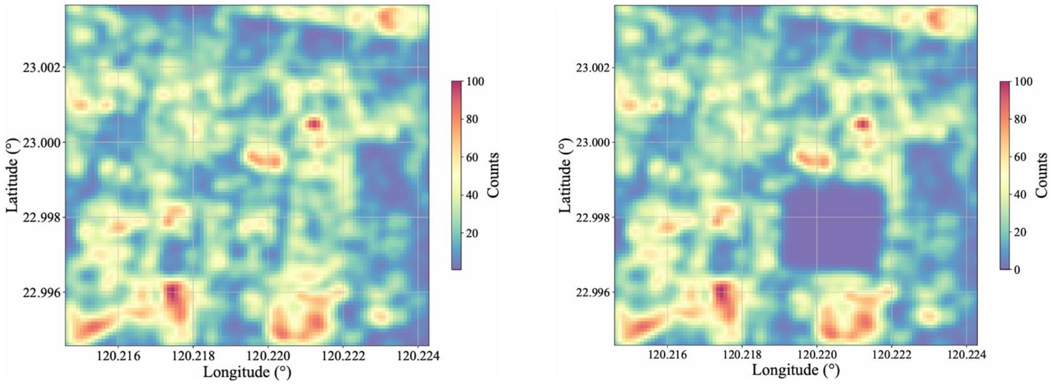

Figure 5 Left: Normalized density map. Right: Normalized clipped density map

3.3 Training Results and Analysis

Figure 6 illustrates the top view of the trajectory generated by DQN algorithm, utilizing the original unclip image feature points dataset. The path is visualized over a DSM background, with the UAV trajectory highlighted in red. The starting point is indicated in green, while the goal point is represented in blue.

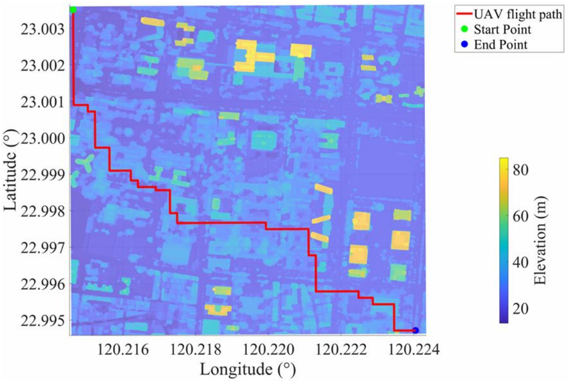

Figure 6 Topview of original DQN-generated path

Figure 7 and Figure 8 depict the UAV's trajectory in both a 2.5D view and a side perspective, respectively. These visualizations provide a comprehensive view of the UAV's movement in three-dimensional space. Notably, the side view emphasizes that the generated path navigates effectively through the terrain, successfully avoiding obstacles while maintaining an optimal altitude. This altitude maintenance is significant as it contributes to reducing the UAV's energy consumption during flight.

The pathway design intentionally targets areas characterized by higher densities of image feature points, which have been normalized and inverted. These feature points are critical for effective visual localization and navigation tasks. To evaluate the algorithm's performance, we conducted an experiment comparing the results obtained from the original dataset against those derived from the clipped dataset. This methodological step aimed to assess the model's robustness by observing whether the path generation mechanism would actively avoid the clipped region, which now contains a reduced number of feature points. The deliberate removal of these features simulates a scenario where the UAV's visual navigation ability might be compromised. This experiment is thus designed to elucidate the role of these feature points in enhancing UAV path planning efficacy and to quantify the model's sensitivity to variations in feature point distribution.

Despite the original path traversing areas with a higher density of feature points, the clipped path also navigates through the terrain by selecting an alternative route that, while different, still encompasses regions of relatively high feature point density. This experiment demonstrates the algorithm's capability to produce a path that is both efficient and feasible, showcasing its robustness in complex path planning scenarios. Figure 9 overlays the UAV paths on the density maps corresponding to both the original and clipped image feature points datasets. The left side of the figure illustrates the path generated using the unaltered density map, while the right side presents the path derived from the clipped density map. This visual comparison effectively highlights the impact of data clipping on the UAV's path planning and navigational efficacy.

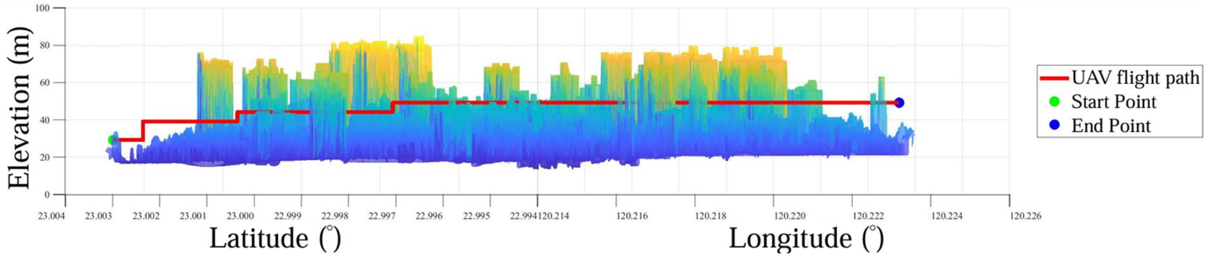

Figure 7 2.5D view of original DQN-generated path

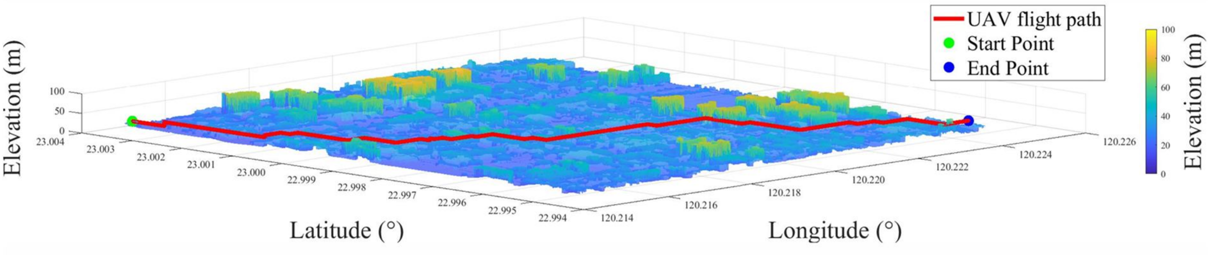

Figure 8 Side view of original DQN-generated path

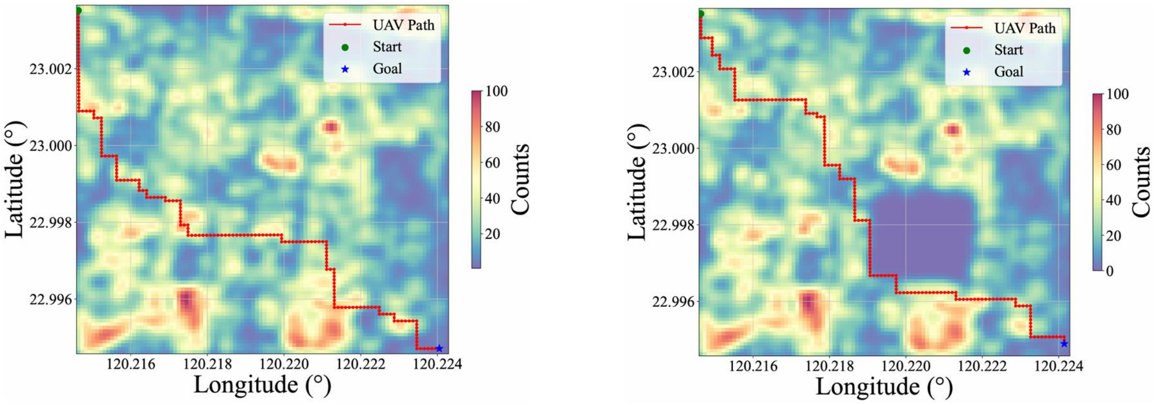

Figure 9 The comparison between two different density maps. Left: Generated path over the density map of original image feature points data. Right: Generated path over the density map of clipped image feature points data

Additionally, Table 2 outlines the training results for both the original and clipped datasets. This table includes key metrics such as path length and the average value of image feature points for each dataset. The findings indicate that the clipped data results in a slightly longer path length of 196.50 grid size compared to 196.00 grid size for the original data, accompanied by a lower average value of image feature points in the clipped data. These results suggest that while the clipping of data may yield a longer path, it still preserves a sufficient number of matching pairs. Furthermore, the results demonstrate a preference in route generation: the system tends to forgo a shorter, more direct path to the destination in favor of traversing areas with a higher average value of image feature points. This suggests that the generated route prioritizes maximizing the opportunity for successful UAV visual localization by seeking regions with richer visual features.

Table 2 Comparison of path planning metrics using original and clipped data

	Original data	Clipped data
Path length	196.00	196.50
Average image feature points value	77.10	73.63

3.4 Postprocessing Results- Path Smoothing

The properties of Bézier curves facilitate effective path smoothing, which is particularly beneficial for refining the trajectories generated by the Noisy D3QN with PER methods. The model's configuration for determining the action direction of the UAV imposes constraints that can impair its performance, often resulting in waypoints characterized by abrupt 90-degree turns. Such sharp turns are not conducive to the UAV's motion characteristics, posing considerable challenges in real-world applications. To address this issue, we employed third-order (cubic) Bézier curves to enhance the smoothness of the generated waypoints, thereby reducing the abruptness of the turning angles. At first glance, the adjustments made through the smoothing process may appear subtle; however, a more detailed examination reveals a measurable refinement in path continuity and turning angle realism. Accordingly, Figure 10 provides a closer view of the Bézier-smoothed trajectory, specifically highlighting modifications in regions where the UAV was previously required to execute sharp turns. The application of the Bézier curve effectively mitigates these infeasible turns, yielding a more natural and dynamically viable flight path for the UAV.

3.5 Comparison with Traditional DQN Algorithms

To rigorously evaluate the effectiveness of the proposed DQN-based reinforcement learning methodologies and to analyze the impacts of reward function design, we conducted a series of simulation training and testing experiments employing four distinct algorithms: DQN, dueling DQN (DDQN), dueling doubling DQN (D3QN), and noisy D3QN with prioritized experience replay (PER) algorithm. The comparative performance assessment was grounded on four critical dimensions: average steps and path length, reward, image feature point values, and training duration. Each algorithm was subjected to a total of 1000 training episodes.

Figure 11 and Figure 12 depict the trajectories of average steps and path lengths per episode, respectively. At the onset of training, all algorithms display considerable fluctuations attributable to random exploration dynamics. Notably, the noisy D3QN with PER demonstrates a marked ability to converge rapidly and stabilize around the 300-episode mark, resulting in both shorter and more consistent paths. The determination of convergence is primarily based on the stabilization of the average number of steps required per episode. During the initial phase, data fluctuation is expected due to random environment exploration. However, the Noisy D3QN with PER shows a frequent reduction in the magnitude and frequency of fluctuations after a few hundred episodes. While minor, occasional spikes may still occur as the agent explores previously unknown states, the path length and steps quickly return to a stable range. In contrast, the other algorithms, DQN, DDQN, and D3QN, exhibit protracted training periods and erratic performance spikes, indicating instability in the policy update process.

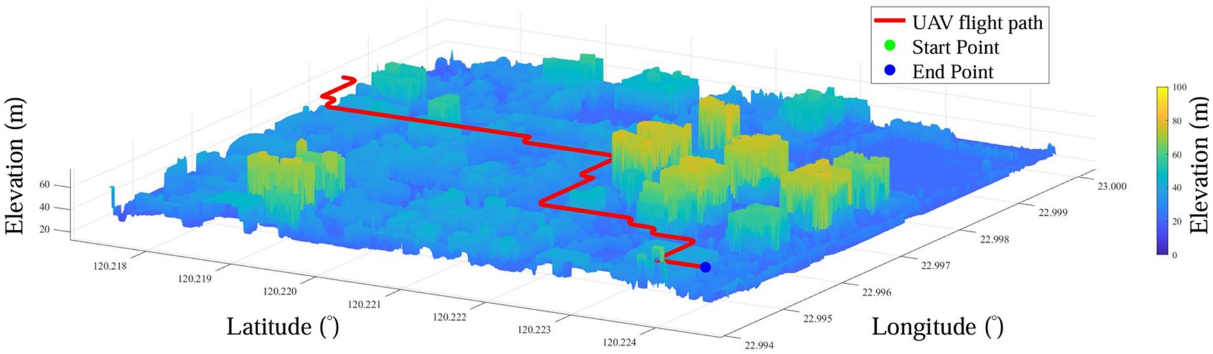

Figure 10 The detail of the Bézier-smoothed trajectory

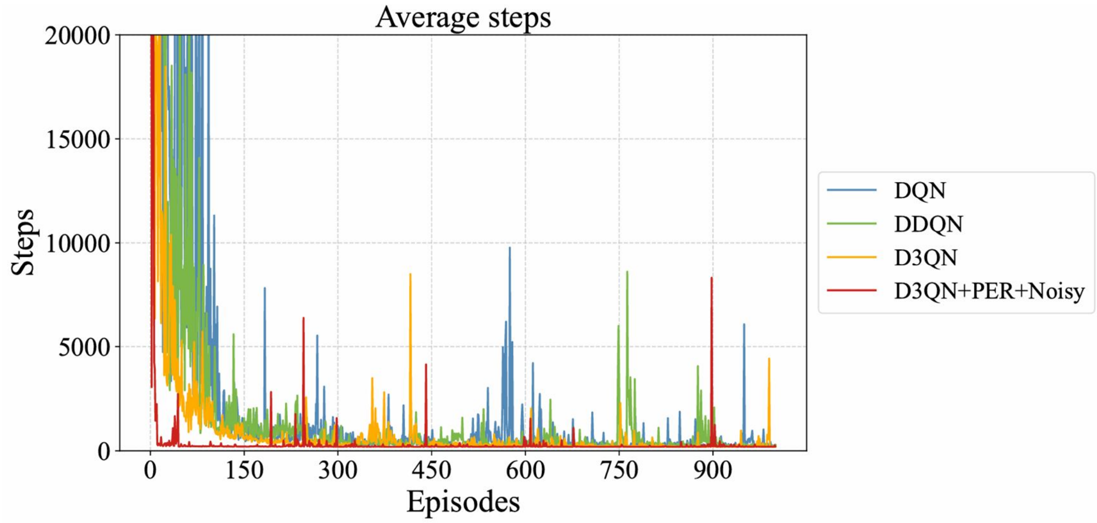

Figure 11 Average steps of generated path over episodes of the noisy D3QN with PER, D3QN, DDQN, DQN

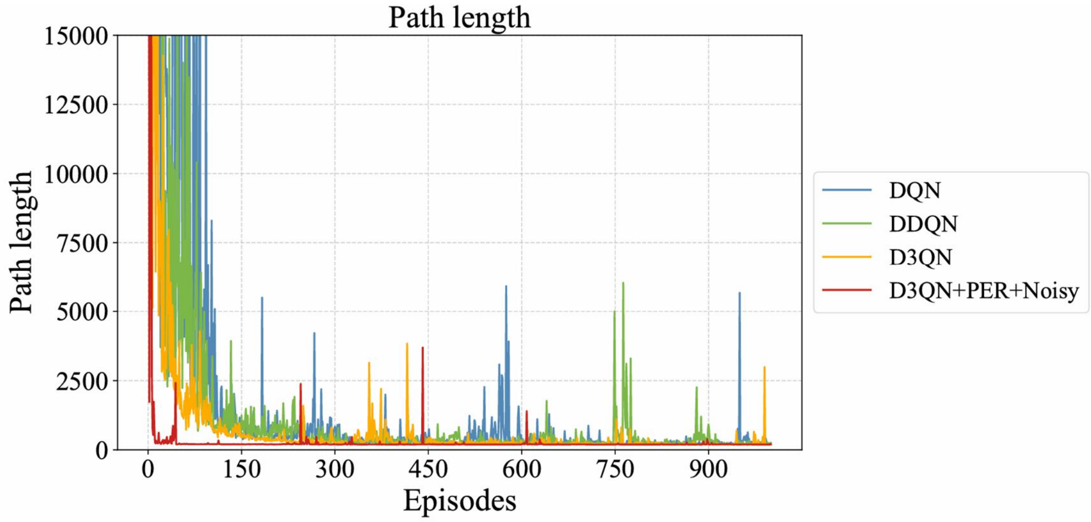

Figure 12 Path length of generated path over episodes of the noisy D3QN with PER, D3QN, DDQN, DQN

The enhanced convergence exhibited by the noisy D3QN with PER can be attributed to the synergistic interplay between PER and noisy networks. PER facilitates the prioritization of informative experiences during the replay phase, whereas Noisy Nets innovate by supplanting the traditional epsilon-greedy strategy with adaptive stochastic exploration methods. This dual combination fosters more effective policy refinements over time, culminating in reduced step counts and expedited paths toward objective attainment.

Further analysis is presented in Figure 13 and Figure 14, which outline the average reward per step alongside the total reward per episode. The DQN algorithm, lacking a dueling architecture, frequently suffers from overestimation bias, resulting in unstable and overly optimistic Q-value estimations. While DQN might yield elevated per-step rewards, its total reward trajectory is characterized by wide fluctuations and a lack of coherence. Conversely, the noisy D3QN with PER exhibits more stable total rewards over time, signifying more effective long-term strategic planning, despite registering marginally lower average per-step returns.

Table 3 encapsulates essential performance metrics, affirming that this algorithm surpasses its counterparts across several indicators, demonstrating fewer average steps, shorter path lengths, a higher number of feature points, and expedited training times. As detailed, the Noisy D3QN with PER substantially eclipses the performance of the other algorithms across several crucial benchmarks. It achieves the lowest average steps of 396.91 and path length of 196.00, while concurrently capturing the highest number of feature points of 77.10, highlighting its efficacy in navigation and visual positioning. Additionally, it registers the fastest training time per episode of 6.56 seconds, emphasizing its computational efficiency. Although it recorded a negative total reward of 13119.49, this figure represents a significant improvement over that of DQN, illustrating a more favorable balance between penalties and rewards throughout the learning phase.

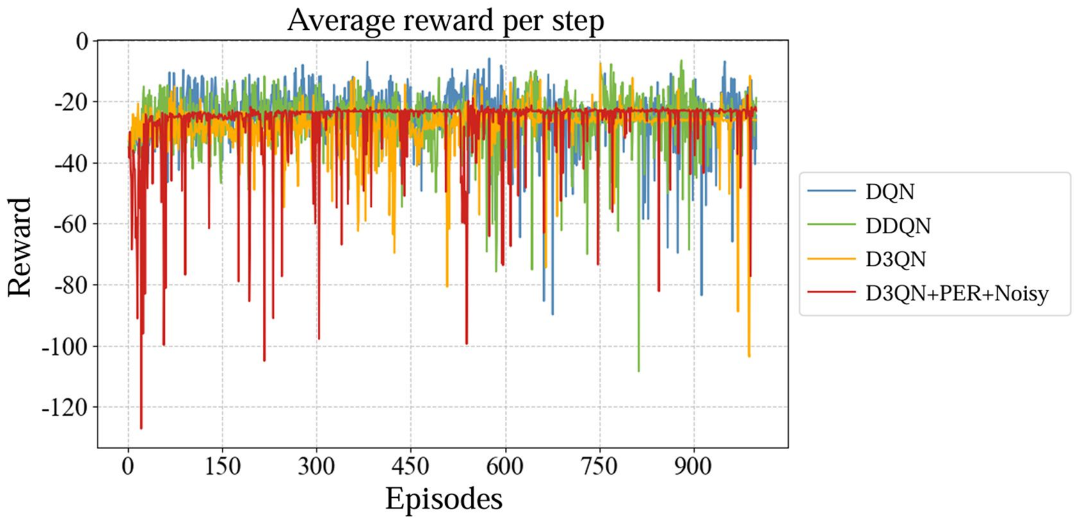

Figure 13 Average reward of generated path over episodes of the noisy D3QN with PER, D3QN, DDQN, DQN

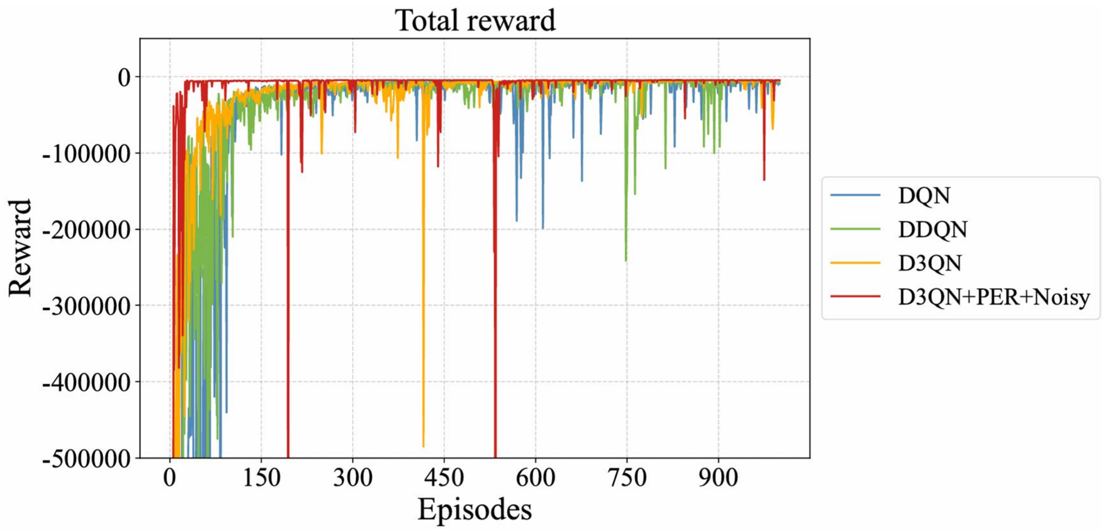

Figure 14 Total reward of generated path over episodes of the noisy D3QN with PER, D3QN, DDQN, DQN

Table 3 Performance comparison between Noisy D3QN with PER, D3QN, DDQN, DQN

Performance parameters	Noisy D3QN with PER	D3QN	DDQN	DQN
Total episodes	1000	1000	1000	1000
Average steps (per episode)	396	1309	2455	3843
Path length (single episode)	196.0	199.0	211.5	220.0
Total reward (single episode)	-13119	-39135	-71175	-115647
Average feature points (single episode)	77	71	70	75
Average training time (s)	6.56	9.75	17.07	22.74
Total training time (s)	6558.25	9754.52	17072.39	22735.01

In conclusion, the results of this investigation underscore that the integration of prioritized experience replay (PER) and noisy networks into the D3QN framework considerably enhances the learning efficiency of the agent, the quality of path planning, and its ability to perform visual localization. This evidence highlights the promising potential of advanced DQN variants for practical applications in UAV path planning within complex three-dimensional environments.

4. Conclusions

Path planning in 2.5D environments is inherently complex due to irregular terrain and the need for algorithms that manage dynamic obstacles and localization constraints. Traditional algorithms like Dijkstra, A*, and basic DQN methods are limited in such contexts, especially in terms of efficiency and adaptability. This study proposes an enhanced DQN-based framework incorporating Double DQN, Dueling Networks, Prioritized Experience Replay, and Noisy Nets. By using DSM data, image feature point density, and start/end positions as input, the system generates obstacle-aware and energy-efficient UAV trajectories suitable for visual localization. Bézier curve smoothing further improves trajectory feasibility. The framework is tailored for GPS-denied environments, enabling robust UAV localization through optimized image feature alignment. Results show improved training stability and adaptability across diverse terrain conditions. Moreover, it supports path inference from various starting points, enabling efficient individual path generation for multi-UAV navigation without frequent replanning—making it well-suited for real-time and resource-constrained scenarios. This work demonstrates that the future of robust autonomous navigation lies in learning-based planners that treat path quality and localization uncertainty as first-class citizens, moving beyond purely geometric optimality.

References

Bezier, P., 1972. Numerical Control: Mathematics and Applications, John Wiley and Sons, London.

Detone, D., Malisiewicz, T., and Rabinovich, A., 2018. SuperPoint: Self-supervised interest point detection and description, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Salt Lake City, UT, USA, pp. 337-349, DOI: 10.1109/CVPRW.2018.00060.

Fortunato, M., Azar, M.G., Piot, B., Menick, J., Hessel, M., Osband, I., Graves, A., Mnih, V., Munos, R., Hassabis, D., Pietquin, O., Blundell, C., and Legg, S., 2018. Noisy networks for exploration, Proceedings of the 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings, Vancouver, Canada, DOI: 10.48550/arXiv.1706.10295.

Huang, Z., and Li, T., 2023. Path planning for UAV assisted IOT data collection with dueling DQN, Proceedings of the 42nd Chinese Control Conference (CCC), pp. 6227–6232, IEEE, Tianjin, China, DOI: 10.23919/CCC58697.2023.10240175.

Lindenberger, P., Sarlin, P.E., and Pollefeys, M., 2023. LightGlue: Local feature matching at light speed, Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, DOI: 10.1109/ICCV51070.2023.01616.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., and Hassabis, D., 2015. Human-level control through deep reinforcement learning, Nature, 518(7540): 529–533, DOI: 10.1038/nature14236.

Schaul, T., Quan, J., Antonoglou, I., and Silver, D., 2016. Prioritized experience replay, Proceedings of the 4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings, Caribe Hilton, San Juan, Puerto Rico, DOI: 10.48550/arXiv.1511.05952.

Van Hasselt, H., Guez, A., and Silver, D., 2016. Deep reinforcement learning with double Q-learning, Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pp. 2094–2100, AAAI Press, Phoenix, Arizona, USA, DOI: 10.1609/aaai.v30i1.10295.

Villanueva, A., and Fajardo, A., 2019. UAV navigation system with obstacle detection using deep reinforcement learning with noise injection, Proceedings of the International Conference on ICT for Smart Society (ICISS), pp. 1–6, Bandung, Indonesia, DOI: 10.1109/ICISS48059.2019.8969798.

Wang, Z., Schaul, T., Hessel, M., Van Hasselt, H., Lanctot, M., and De Frcitas, N., 2016. Dueling Network Architectures for Deep Reinforcement Learning, Proceedings of the 33rd International Conference on Machine Learning (ICML), Vol.48, New York, USA.

Wu, X., Huang, S., and Huang, G., 2023. Deep reinforcement learning-based 2.5D multi-objective path planning for ground vehicles: Considering distance and energy consumption, Electronics, 12(18):3840, DOI: 10.3390/electronics12183840.

Yao, J., Li, X., Zhang, Y., Ji, J., Wang, Y., Zhang, D., and Liu, Y., 2022. Three-dimensional path planning for unmanned helicopter using memory-enhanced dueling deep Q network, Aerospace, 9(8):417, DOI: 10.3390/aerospace9080417.

Zhu, Z., Hu, C., Zhu, C., Zhu, Y., and Sheng, Y., 2021. An improved dueling deep double-Q network based on prioritized experience replay for path planning of unmanned surface vehicles, Journal of Marine Science and Engineering, 9(11): 1267, DOI: 10.3390/jmse9111267.

無人機視覺導航路徑規劃使用深度強化學習網路

黃珮瑄[1*] 林昭宏[2]

摘要

傳統無人飛行載具 (UAV) 路徑規劃側重於優化路徑長度與能源效率等多種指標。然而，在無 GPS 環境中，視覺定位的品質至關重要。本研究使用深度強化學習 (DRL) 框架並引入優先經驗回放的噪聲雙決策深度 Q 網絡 (Noisy D3QN with PER)，影像匹配的特徵點整合到路徑規劃中，共同優化路徑的幾何效率與視覺穩健性。此架構能有效解決稀疏獎勵和訓練不穩定問題，提高價值估計準確性與探索效率。實驗結果顯示，不僅產生高效、動態平滑且視覺連續的軌跡，並且在單次訓練後能從多個起始點推斷路徑。除了計算效率的改進，亦能提升複雜環境中視覺定位的性能與穩定性，以提高導航精度。

關鍵詞：強化學習、無人機、深度 Q 網路(DQN)

[1] 國立成功大學測量及空間資訊學系 碩士生
[2] 國立成功大學測量及空間資訊學系 教授
[*] 通訊作者, E-mail: hpei254@gmail.com

收到日期：民國 114 年 07 月 15 日
修改日期：民國 114 年 09 月 26 日
接受日期：民國 114 年 10 月 16 日