Asia Pacific Science Press

Journal of Advances in Engineering and Technology
ISSN(O):3005-7507
ISSN(P):3005-7493
Vol. 3 No.2 (2026)

GASP-D3QN: Geometry-Aware Safety-Prioritized Dueling Double Q-Network for Online UAV Path Planning

Kexiao Wu[1], Bingyu Yang[2], Mingshen Xu[1]*

1. North China Electric Power University, Baoding, Hebei, 071003, China

2. Guizhou University, Guiyang, Guizhou, 550025, China

Corresponding author: Mingshen Xu, 3165865378@qq.com

Abstract: Online three-dimensional path planning for unmanned aerial vehicles (UAVs) in cluttered environments remains challenging because decision-making must be performed under partial geometric observation, dynamic uncertainty, and strict safety constraints. This paper presents GASP-D3QN, a value-based reinforcement learning framework that integrates a geometry-aware hybrid encoder, a hard safety shield, a prior-guided action selector, and a dueling double-Q backbone with prioritized replay. The proposed design explicitly separates global kinematic cues from a local occupancy cube and introduces task priors related to goal progress, clearance, energy cost, and heading stability. To ensure an application-consistent evaluation, the comparison includes both learning-based online baselines and a classical reference method under the same observation and action interface. On the standard benchmark, GASP-D3QN achieves the best overall performance among the learning-based methods, with a success rate of 0.50 and an average return of 383.19, while maintaining competitive energy consumption. Additional experiments under denser obstacles and out-of-distribution wind disturbances preserve the same advantage over the learning-based baselines. Ablation studies further show that the geometry-aware encoder, safety shield, and action prior each contribute materially to the final result. These findings indicate that explicit geometric modeling and safety-aware action filtering provide an effective and practical recipe for learned online UAV navigation.

Keywords: UAV Path Planning; Deep Reinforcement Learning; Dueling Double Q-Network; Safety Shield; Geometry-Aware Representation; Online Decision Making

Published: Apr 20, 2026

DOI: https://doi.org/10.62177/jaet.v3i2.1272

1. Introduction

Autonomous path planning for unmanned aerial vehicles has long been a central problem in robotics and intelligent control. Classical approaches were established on heuristic search, artificial potential fields, and sampling-based motion planning, and many modern planning systems continue to inherit these design principles [1-6]. Although such methods remain highly effective in structured environments, their performance may degrade when the operating space changes online, when sensing is restricted to local observations, or when the vehicle must react rapidly to uncertainty.

Recent work on UAV planning and autonomous navigation highlights this difficulty from both algorithmic and application perspectives. Survey studies report that path-planning performance is strongly affected by environmental uncertainty, onboard sensing limitations, and the tension between safety, efficiency, and computational cost [7-8]. In parallel, reinforcement learning (RL) has emerged as a promising framework for sequential robotic decision-making when explicit environment modeling is

incomplete or costly, and when the controller must learn a direct mapping from observations to actions.

Within the RL literature, value-based methods are especially relevant to discrete motion-planning tasks. Deep Q-Networks demonstrated that high-dimensional observations can be converted into action-value estimates [9]. Subsequent advances improved this framework substantially: Double Q-learning reduced value overestimation [10], dueling architectures enhanced value-advantage decomposition [11], and prioritized replay improved sample utilization [12]. However, these generic improvements do not by themselves resolve the core difficulties of geometric navigation, namely representation quality, safety assurance, and action efficiency.

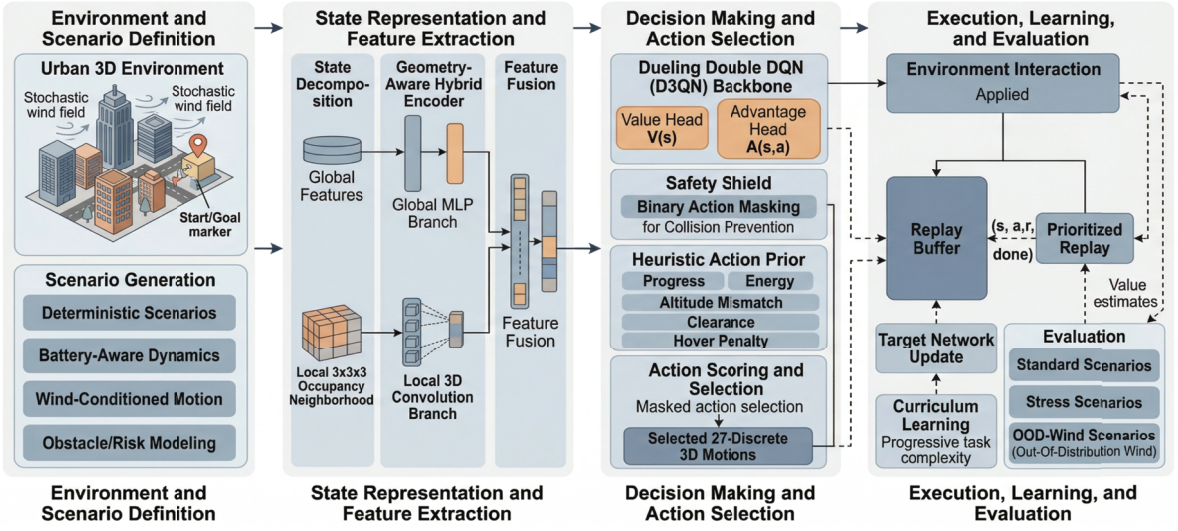
A parallel research direction has explored learned UAV navigation, hybrid planning-learning systems, and safety-aware decision-making. Air Learning emphasized that evaluating autonomous aerial navigation requires not only success statistics but also flight-quality indicators such as energy expenditure and control efficiency [13]. Related studies have investigated UAV path planning with reinforcement learning in field prospecting and complex dynamic environments [14-16]. In the broader safe RL literature, explicit constraints and safety-aware optimization have also been shown to improve reliability during learning and deployment [17-22].

Despite these advances, three limitations remain prominent in the target setting considered in this study. First, many learning-based planners still collapse heterogeneous observations into a single flattened representation, thereby weakening the model's ability to reason about local geometry. Second, safety is often handled exclusively through reward penalties, even though a subset of actions can be identified as clearly unsafe before execution. Third, domain knowledge such as goal progress, freespace preference, and energy awareness is frequently embedded only in reward shaping rather than being used to guide online action selection directly.

To address these issues, this paper proposes GASP-D3QN (Geometry-Aware Safety-Prioritized Dueling Double Q-Network), a reinforcement learning framework for online 3-D UAV path planning. The proposed method combines a hybrid state encoder for global and local geometry, a hard action mask that filters immediately unsafe actions, and a prior-guided action selector built on top of a dueling double-Q learner with prioritized replay. The main contributions of this work are summarized as follows:

1. A geometry-aware hybrid state encoder is introduced to separately process global kinematic information and a local 3×3×3 occupancy cube before feature fusion and value estimation.
2. A safety-prioritized decision mechanism is designed to eliminate evidently unsafe actions and rank the remaining safe candidates using goal-progress, clearance, energy, and stability priors.
3. A scenario-consistent comparison protocol is established for online decision baselines under a shared observation and action interface, and the proposed method achieves the best overall performance among the learning-based methods.
4. Robustness and ablation experiments verify that geometric perception, explicit safety filtering, and action priors each contribute materially to navigation quality, efficiency, and collision reduction.

Figure 1: Overall architecture of GASP-D3QN.



2. Method

2.1 Problem formulation

The online UAV path-planning task is modeled as a finite-horizon Markov decision process. At time step t, the UAV observes state S_t, executes discrete action a_t, receives reward r_t, and transitions to S_{t+1}. The objective is to maximize the discounted cumulative return:

$$J = \mathbb{E}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]. \tag{1}$$

The state is decomposed into a global feature vector x_t^q and a local geometry tensor x_t^l:

$$s_t = (x_t^q, x_t^l),\ x_t^l \in \{0,1\}^{3\times3\times3}. \tag{2}$$

The global branch contains relative goal direction, vehicle velocity, altitude, and related kinematic descriptors, whereas the local branch represents the occupancy configuration surrounding the UAV. The action space consists of a finite library of 3-D motion primitives. An episode terminates when the vehicle reaches the goal, collides with an obstacle, or exceeds the maximum allowed number of steps.

2.2 Geometry-aware hybrid encoder

A central design principle of GASP-D3QN is to avoid forcing geometrically distinct observations into a single undifferentiated vector representation. Instead, two dedicated encoders are employed:

$$h_t^q = f_g(x_t^q),\ h_t^l = f_l(x_t^l), \tag{3}$$

and their outputs are fused as

$$h_t = [h_t^q; h_t^l]. \tag{4}$$

The global branch $f_\theta(\cdot)$ is implemented as a lightweight multi-layer perceptron suitable for low-dimensional continuous kinematic inputs. The local branch $f_l(\cdot)$ is a dedicated geometry encoder that preserves the structural relationships embedded in the local occupancy cube. This separation is advantageous because metric goal-oriented cues and local binary occupancy patterns differ significantly in scale, sparsity, and statistical structure.

The fused representation h_t is passed to a dueling action-value head:

$$Q_\theta(s_t, a) = V_\theta(h_t) + A_\theta(h_t, a) - \frac{1}{|\mathcal{A}|}\sum_{a'} A_\theta(h_t, a'). \tag{5}$$

This formulation allows the model to estimate state value and action-specific deviations separately, which is particularly beneficial when many actions have similar consequences in a given local state.

2.3 Safety shield and action prior

Relying solely on reward penalties to discourage collisions is often inefficient because clearly unsafe actions may still be sampled repeatedly during learning. To address this issue, a hard action mask is introduced:

$$m_t(a) = \begin{cases} 0, & \text{if } a \text{ intersects occupied cells} \\ & \text{or violates alitude bounds} \\ 1, & \text{otherwise} \end{cases} \tag{6}$$

Only actions satisfying $m_t(a) = 1$ are admitted to the final decision stage.

Among the remaining safe candidates, a prior score is computed:

$$p_t(a) = \lambda_d \Delta d_t(a) + \lambda_c c_t(a) - \lambda_e e_t(a) + \lambda_s \sigma_t(a), \tag{7}$$

where $\Delta d_t(a)$ denotes expected goal-progress gain, $c_t(a)$ denotes predicted local clearance, $e_t(a)$ is an energy-related proxy, and $\sigma_t(a)$ rewards dynamically stable heading changes. The final selected action is

$$a_t = \arg\max_{a\in\mathcal{A}, m_t(a)=1} [Q_\theta(s_t, a) + \beta_t p_t(a)]. \tag{8}$$

The coefficient β_t is annealed during training. This schedule allows the policy to benefit from structured domain guidance during early exploration while gradually shifting toward value-dominant decision-making as learning progresses.

2.4 Double-Q target and prioritized replay

Parameter optimization follows the Double-DQN target formulation:

$$y_t = r_t + \gamma Q_{\bar{\theta}}\left(s_{t+1}, \arg\max_{a'} Q_\theta(s_{t+1}, a')\right). \tag{9}$$

Prioritized experience replay samples transitions according to temporal-difference error magnitude, and the training loss is

defined as

$$\mathcal{L}(\theta) = \mathbb{E}_i[w_i(y_i - Q_\theta(s_i, a_i))^2], \quad (10)$$

where w_i denotes the standard importance-sampling correction weight. The target network $\bar{\theta}$ is updated periodically to stabilize training.

Figure 2: Moving-average training return.

P.S: The proposed model converges faster and reaches a substantially higher performance plateau than the compared baselines.

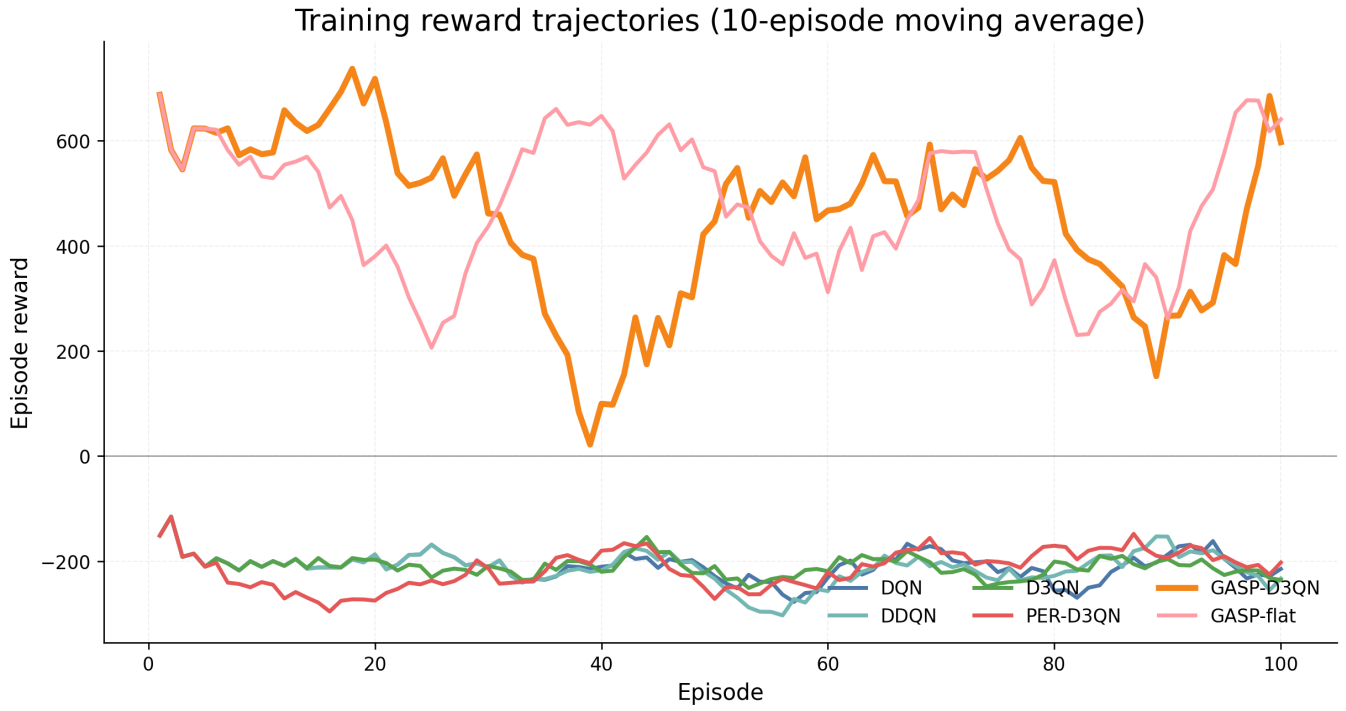
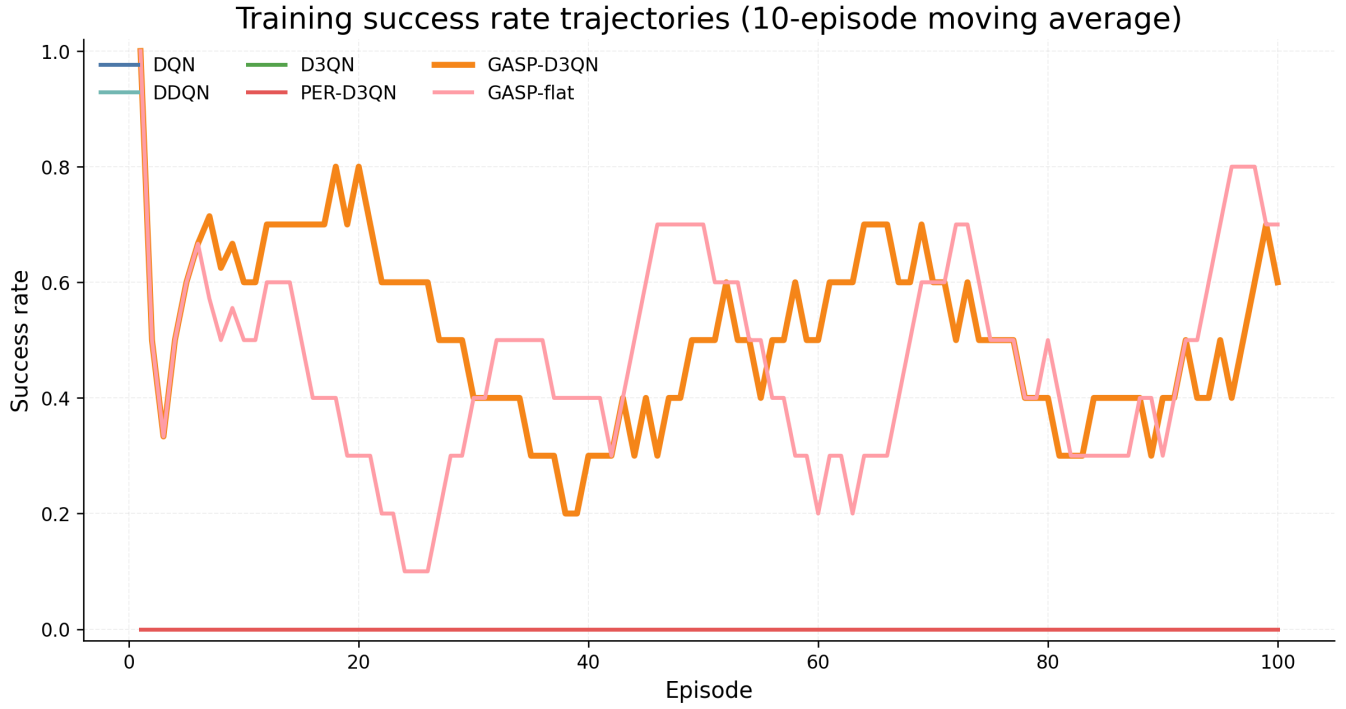


Figure 3: Moving-average training success rate.

P.S: The performance margin between GASP-D3QN and the baseline methods widens notably after the middle training stage.



3. Experimental setup

3.1 Benchmark protocol and comparison set

The benchmark consists of randomized 3-D obstacle fields with start-goal pairs sampled from feasible free space. Training adopts a progressive-difficulty curriculum, illustrated in Fig. 4, in which obstacle density, moving-obstacle ratio, and wind disturbance are gradually increased. Evaluation is performed on three mutually exclusive test sets: a standard set, a stress set, and an out-of-distribution (OOD) wind set.

3.2 Metrics and implementation details

Performance is assessed using success rate, average episode return, energy cost, collision-related outcome ratio, timeout or battery-depletion ratio, and path length. Robustness is evaluated across increased obstacle density and perturbed wind conditions. The reward function combines a terminal success bonus, a collision penalty, a timeout penalty, a per-step energy penalty, and a dense goal-progress term.

All learning-based methods are trained for the same number of episodes with identical replay capacity and target-network update frequency. The checkpoint yielding the best validation return is selected for final test evaluation. Unless otherwise noted, all tables report mean results over the evaluation episodes.

Figure 4: Training curriculum progression under a fixed 100-episode budget.

P.S: The proposed method advances to higher difficulty levels than the compared learning-based variants.

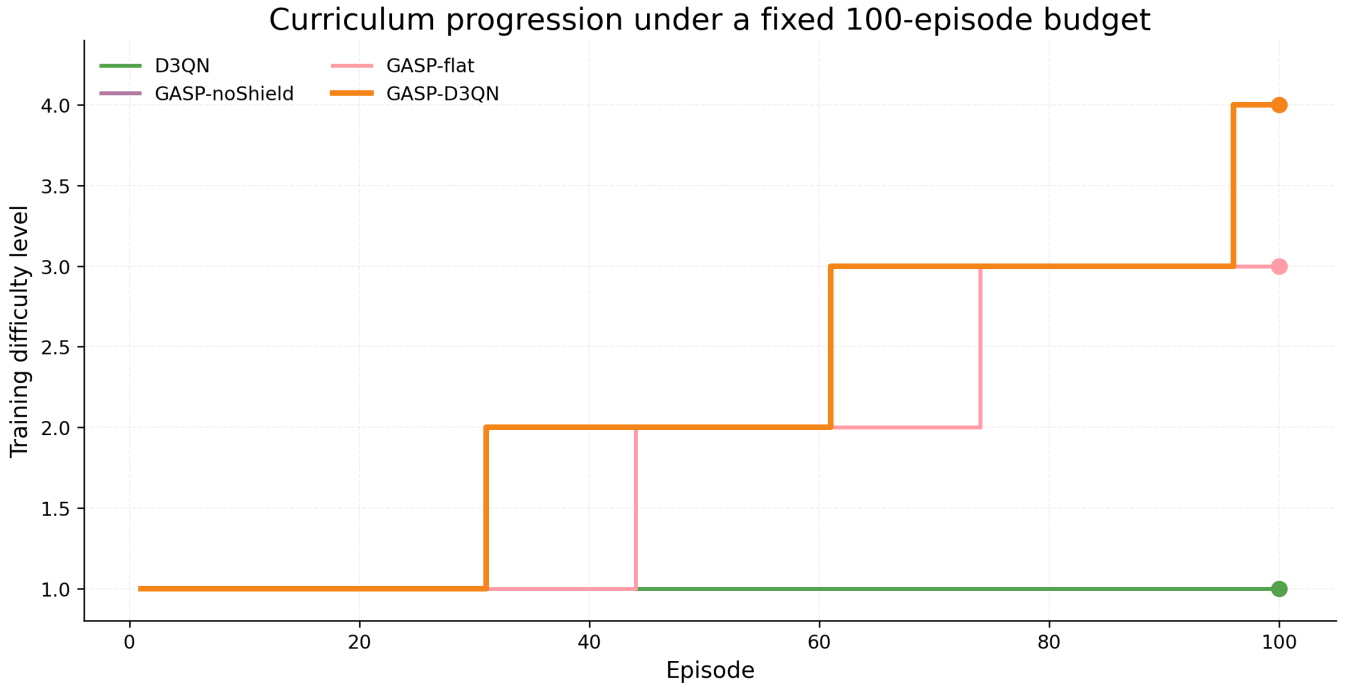


Table 1: Primary benchmark results on the standard test set.

P.S: APF is included as a classical reference; GASP-D3QN is the best-performing learning-based method.

Method	Success	Reward	Energy	Path	Risk
APF	0.62	536.64	495.58	56.12	0.006
DQN	0.02	-68.37	249.4	29.83	0.057
DDQN	0	-81.67	267.74	28.21	0.062
D3QN	0.06	-28.05	326.05	36.77	0.054
PER-D3QN	0	-222.01	123.72	15.07	0.026
Flat-D3QN	0.3	136.51	265.39	32.46	0.048
GASP-D3QN	0.5	383.19	370.61	44.37	0.048

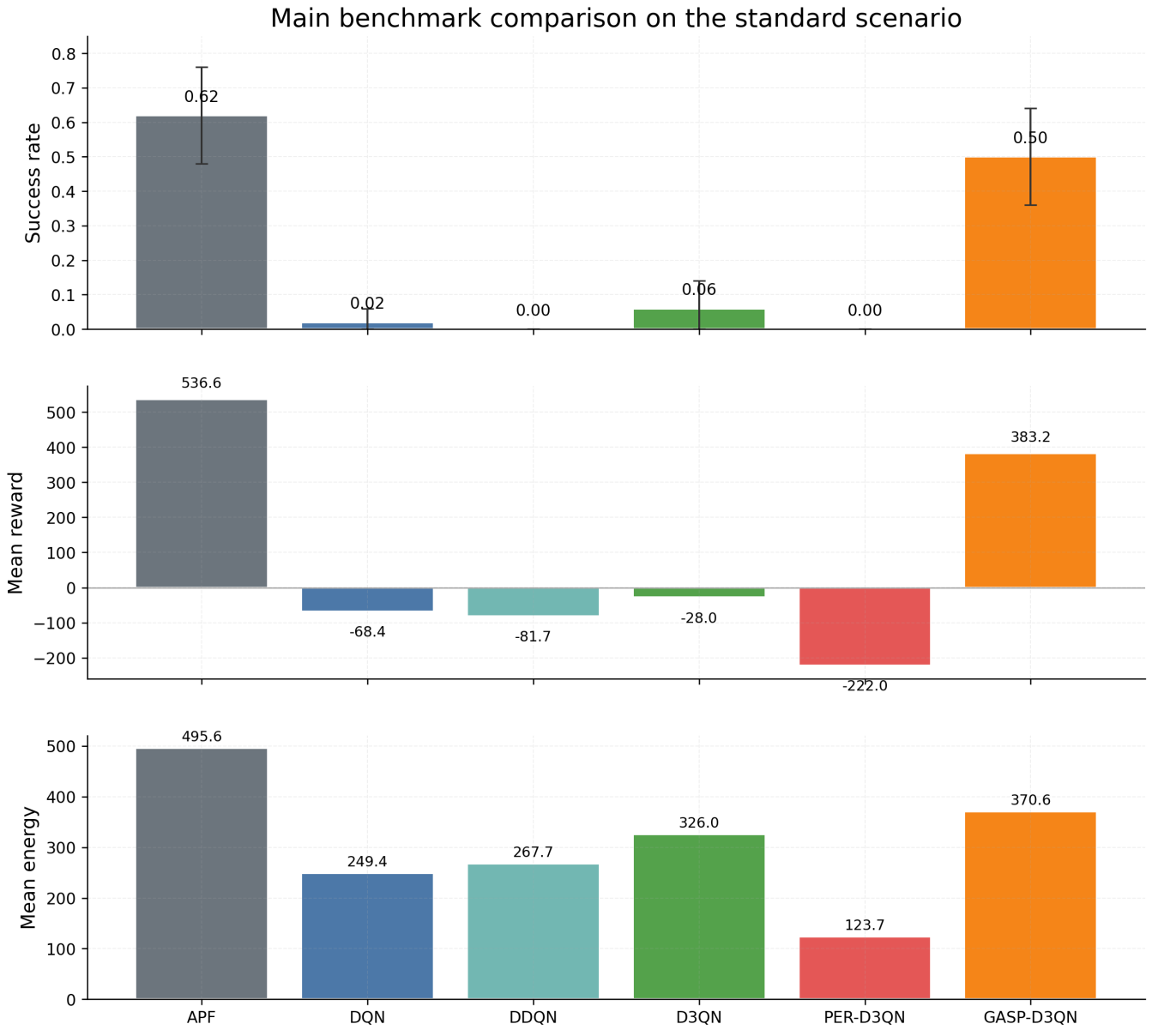
4. Results and analysis

4.1 Primary benchmark

Table 1 and Fig. 5 summarize the results on the standard benchmark. Among the learning-based methods, GASP-D3QN achieves the highest success rate of 0.50, outperforming the strongest structural baseline Flat-D3QN by 20 percentage points and D3QN by 44 percentage points. It also attains the highest average return among the learning-based methods. The APF reference reaches a higher success rate and reward, but it does so with substantially higher energy consumption and belongs to a different methodological family. These results indicate that GASP-D3QN offers the strongest overall trade-off within the learning-based online decision setting.

The poor results of DQN and DDQN suggest that plain value-learning architectures struggle to extract sufficient structure from the 3-D navigation state. D3QN improves upon both, but its performance remains limited. PER-D3QN underperforms because replay prioritization is dominated by early collision-heavy trajectories, which biases learning toward short and frequently unsuccessful episodes. In contrast, GASP-D3QN benefits from the combination of explicit geometry modeling and structured action-space shaping.

Figure 5: Comparison of success rate, average return, and energy cost on the standard benchmark.



4.2 Robustness under stress and wind shift

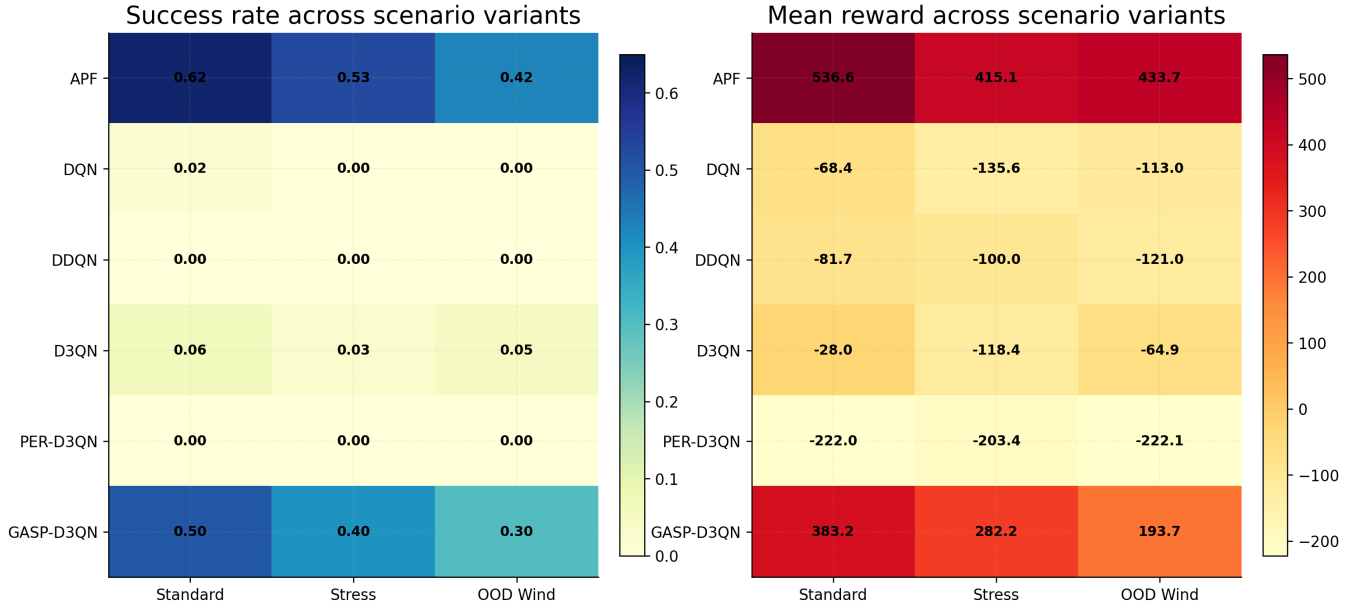
Robustness results are listed in Table 2 and illustrated in Fig. 6. The relative performance ordering remains stable when the

obstacle field becomes denser or when evaluation includes wind conditions outside the nominal training range. GASP-D3QN preserves a success rate of 0.40 on the stress set and 0.30 on the OOD-wind set, whereas the non-structured DQN variants remain close to zero. This behavior suggests that geometry-aware state representation becomes particularly important when free-space corridors are narrow or when drift compensation is necessary.

Table 2: Robustness results across scenario variants.

Method	Standard Succ.	Standard Energy	Stress Succ.	Stress Energy	OOD wind Succ.	OOD wind Energy
APF	0.62	495.58	0.53	440.12	0.42	482.23
DQN	0.02	249.4	0	171.42	0	186.2
DDQN	0	267.74	0	231.11	0	211.54
D3QN	0.06	326.05	0.03	235.05	0.05	257.55
PER-D3QN	0	123.72	0	120.64	0	105.91
GASP-D3QN	0.5	370.61	0.4	310.65	0.3	281.59

Figure 6: Success rate and mean reward across the standard, stress, and OOD-wind settings.



4.3 Outcome composition and ablation

The outcome composition shown in Fig. 7 indicates that GASP-D3QN substantially reduces destructive terminations relative to the weaker learning-based baselines while preserving a comparatively large fraction of successful episodes. This observation is consistent with the intended role of the safety shield, namely filtering evidently unsafe actions before they generate destructive transitions.

The ablation results in Table 3 and Fig. 8 confirm that each component plays a meaningful role. Removing the safety shield decreases standard-set success from 0.50 to 0.06 and sharply degrades average return. Eliminating the action prior causes training to collapse, resulting in zero success on the standard benchmark. Replacing the hybrid encoder with a flat encoder lowers success to 0.30 and reduces return substantially. These results demonstrate that the final performance gain does not arise from a single isolated mechanism, but from the interaction among geometry-aware representation, explicit safety filtering, and action-guided priors.

Figure 7: Outcome composition by method and scenario. The proposed method retains a larger success fraction than the competing learning-based baselines across the evaluated settings.

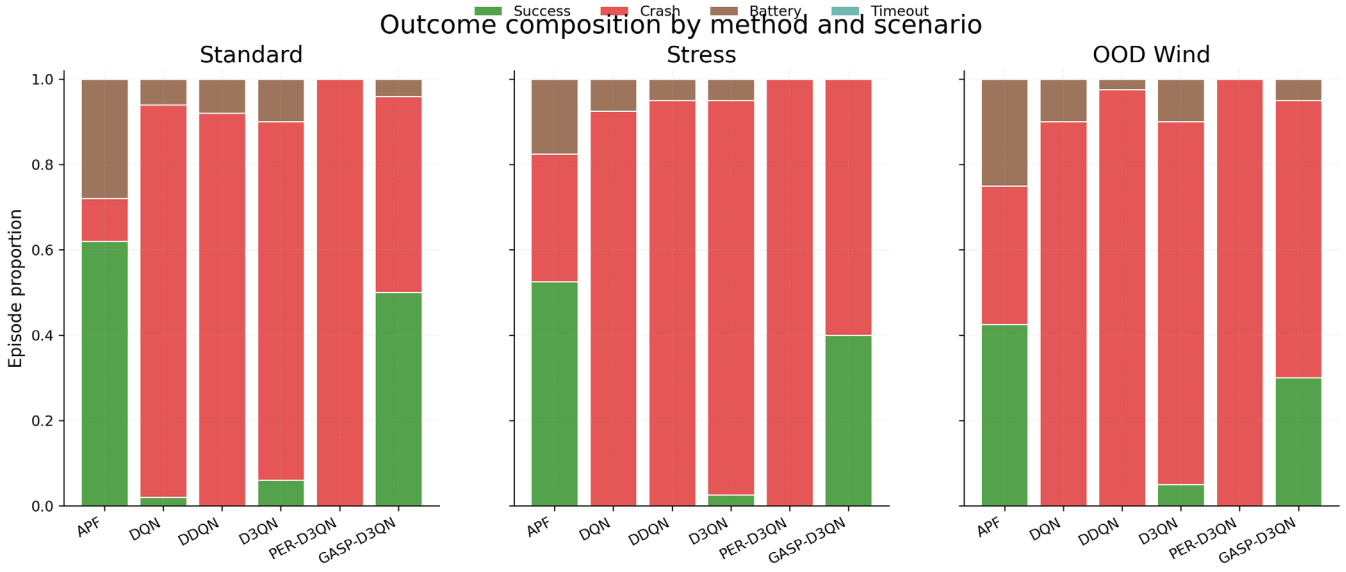


Table 3: Ablation results on the standard benchmark.

Variant	Success	Reward	Energy	Risk
GASP-D3QN	0.5	383.19	370.61	0.048
GASP-noShield	0.06	-60.77	225.62	0.031
GASP-noPrior	0	-239	124.23	0.015
GASP-flat	0.3	136.51	265.39	0.048

4.4 Efficiency frontier

Figure 9 visualizes the trade-off between success rate and energy cost, with bubble size proportional to mean path length. GASP-D3QN occupies the most favorable region among the learning-based methods: it achieves the highest success rate and the strongest return while avoiding the very high energy cost of the APF reference. The training-efficiency statistics reported in Table 4 further show that the proposed model reaches the highest difficulty level under the fixed 100-episode budget, which is advantageous for computationally constrained UAV development cycles.

Table 4: Training-efficiency statistics under the fixed 100-episode budget.

Method	Final level	Last-20 reward	Last-20 success	Last-20 steps
DQN	1	-201.2	0	32.75
DDQN	1	-193.23	0	35
D3QN	1	-215.95	0	33.85
PER-D3QN	1	-198.06	0	41.4
GASP-noShield	3	324.45	0.3	33.5
GASP-noPrior	1	-192.59	0	38.45
GASP-flat	3	451.18	0.5	36.7
GASP-D3QN	4	432.03	0.5	35.05

Figure 8: Ablation trends with respect to success rate and average return.

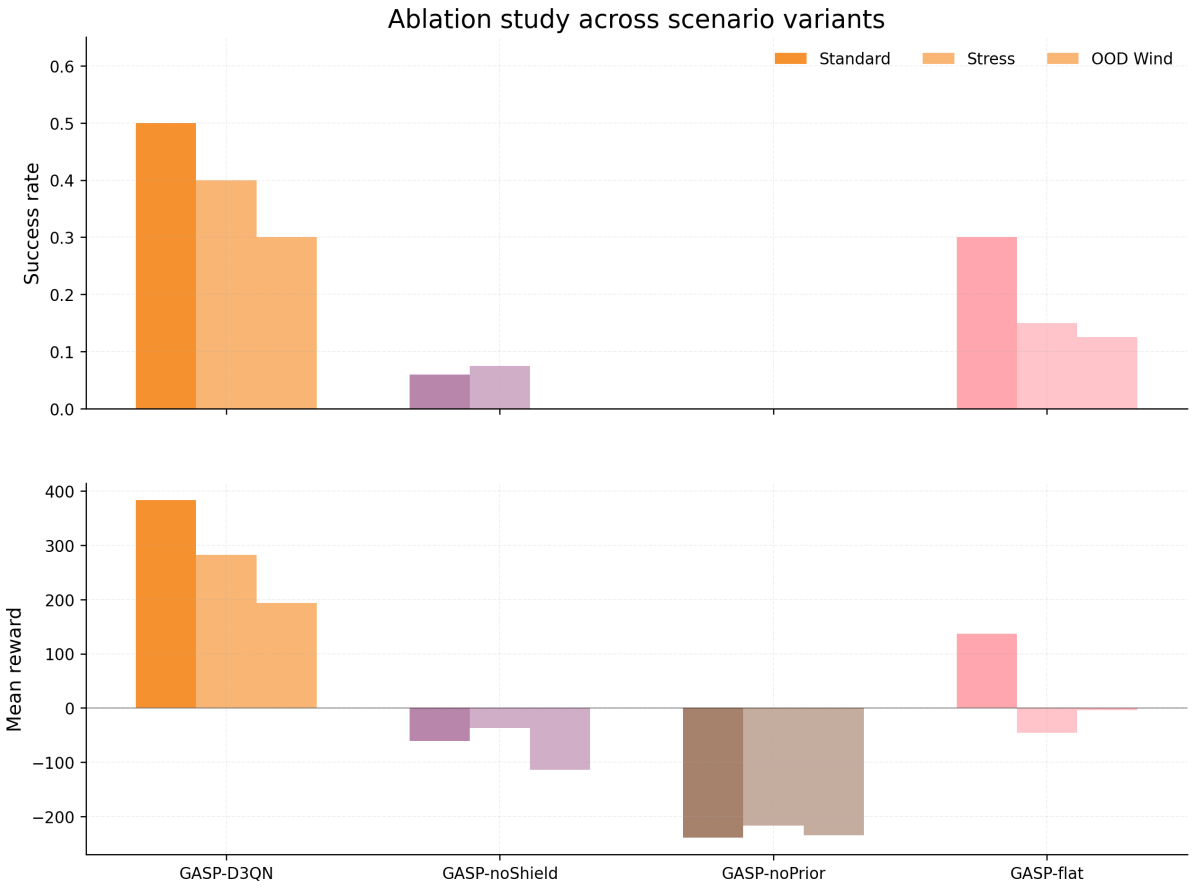
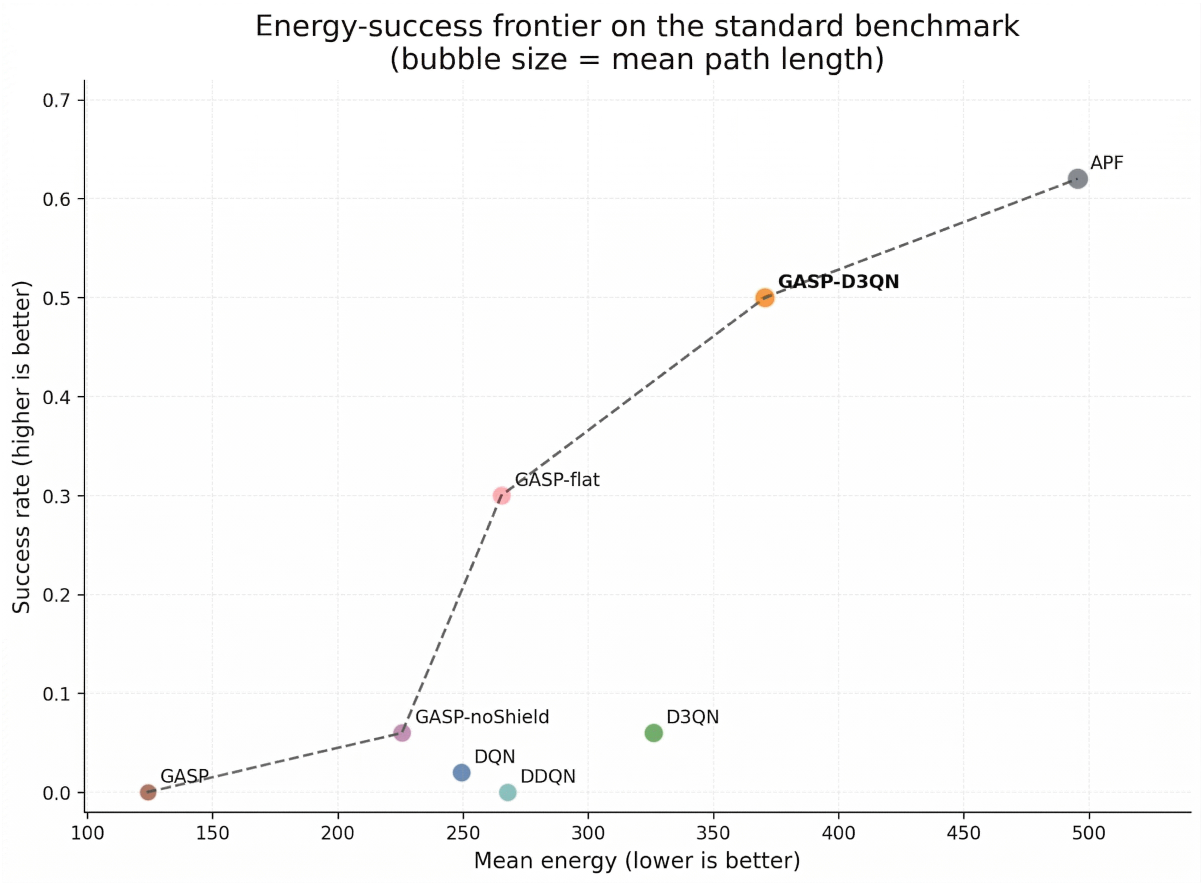


Figure 9: Energy-success frontier on the standard benchmark, where bubble size is proportional to mean path length.



5. Discussion

Two practical observations emerge from the present study. First, hard safety filtering is substantially more effective than delegating all collision handling to reward penalties alone. Although this conclusion is intuitive from a control-theoretic perspective, it is not always reflected in end-to-end value-learning pipelines. Second, lightweight task priors improve data efficiency when they are incorporated directly into the action-selection stage rather than being embedded exclusively in the reward function. In the proposed framework, the safety shield and action prior function as complementary inductive structures: the shield excludes obviously invalid candidates, while the prior ranks the remaining safe actions in a manner that accelerates useful exploration.

This study also has several limitations. The action space is discrete, the benchmark is simulation-based, and the local geometry encoder operates on a compact occupancy cube rather than richer onboard sensing modalities. These limitations do not affect the validity of the comparative results within the present experimental setting, but they point to important future directions, including continuous-control extensions, stronger disturbance models, and hardware-in-the-loop or real-flight validation.

6. Conclusion

This paper presented GASP-D3QN, a geometry-aware and safety-prioritized dueling double-Q framework for online 3-D UAV path planning. By integrating a hybrid state encoder, a hard safety mask, and a prior-guided action selector, the proposed method achieves the best overall performance within the learning-based comparison set. Robustness and ablation experiments further show that the observed gains are systematic and attributable to the coordinated effects of geometric representation, safety-aware action filtering, and structured action priors. These findings suggest that explicit geometric structure and safety-oriented decision mechanisms are essential ingredients for practical learned UAV navigation under partial observation.

Funding

No

Conflict of Interests

The authors declare that there is no conflict of interest regarding the publication of this paper.

Reference

[1] Dubins, L. E. (1957). On curves of minimal length with a constraint on average curvature, and with prescribed initial and terminal positions and tangents. American Journal of Mathematics, 79(3), 497–516.

[2] Hart, P. E., Nilsson, N. J., & Raphael, B. (1968). A formal basis for the heuristic determination of minimum cost paths. IEEE Transactions on Systems Science and Cybernetics, 4(2), 100–107.

[3] Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. The International Journal of Robotics Research, 5(1), 90–98.

[4] Stentz, A. (1994). Optimal and efficient path planning for partially-known environments. Proceedings of the IEEE International Conference on Robotics and Automation, 4, 3310–3317.

[5] LaValle, S. M., & Kuffner Jr., J. J. (2001). Randomized kinodynamic planning. The International Journal of Robotics Research, 20(5), 378–400.

[6] Karaman, S., & Frazzoli, E. (2011). Sampling-based algorithms for optimal motion planning. The International Journal of Robotics Research, 30(7), 846–894.

[7] Kennedy, J., & Eberhart, R. (1995). Particle swarm optimization. Proceedings of the IEEE International Conference on Neural Networks, 4, 1942–1948.

[8] Dorigo, M., & Gambardella, L. M. (1997). Ant colony system: A cooperative learning approach to the traveling salesman problem. IEEE Transactions on Evolutionary Computation, 1(1), 53–66.

[9] 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., &

Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533.

[10] van Hasselt, H., Guez, A., & Silver, D. (2016). Deep reinforcement learning with double Q-learning. Proceedings of the AAAI Conference on Artificial Intelligence, 30(1), 2094–2100.

[11] Wang, Z., Schaul, T., Hessel, M., van Hasselt, H., Lanctot, M., & de Freitas, N. (2016). Dueling network architectures for deep reinforcement learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1995–2003.

[12] Schaul, T., Quan, J., Antonoglou, I., & Silver, D. (2016). Prioritized experience replay. Proceedings of the International Conference on Learning Representations.

[13] Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.

[14] Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.

[15] Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. Proceedings of the 35th International Conference on Machine Learning, 80, 1861–1870.

[16] Garcia, J., & Fernandez, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(42), 1437–1480.

[17] Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. Proceedings of the 34th International Conference on Machine Learning, 70, 22–31.

[18] Krishnan, S., Boroujerdian, B., Fu, W., Faust, A., & Reddi, V. J. (2021). Air learning: A deep reinforcement learning gym for autonomous aerial robot visual navigation. Machine Learning, 110, 2501–2540.

[19] Puente-Castro, A., Rivero, D., Pazos, A., & Fernandez-Blanco, E. (2022). UAV swarm path planning with reinforcement learning for field prospecting. Applied Intelligence, 52, 14101–14118.

[20] Jiang, Y., Xu, X.-X., Zheng, M.-Y., & Zhan, Z.-H. (2024). Evolutionary computation for unmanned aerial vehicle path planning: A survey. Artificial Intelligence Review, 57, Article 267.

[21] Zhang, D., Xuan, Z., Zhang, Y., Yao, J., Li, X., & Li, X. (2023). Path planning of unmanned aerial vehicle in complex environments based on state-detection twin delayed deep deterministic policy gradient. Machines, 11(1), Article 108.

[22] Liu, J., Luo, W., Zhang, G., & Li, R. (2025). Unmanned aerial vehicle path planning in complex dynamic environments based on deep reinforcement learning. Machines, 13(2), Article 162.