Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.074398

ARTICLE

A PPO-Based DRL Approach for Scalable Communication in Civilian UAV Networks

Chu Thi Minh Hue[1] **and Nguyen Minh Quy**[2,*]

[1]Faculty of Software Engineering, FPT University, Hanoi, Vietnam
[2]Faculty of Information Technology, Hung Yen University of Technology and Education, Hungyen, Vietnam
*Corresponding Author: Nguyen Minh Quy. Email: minhquy@utehy.edu.vn

Received: 10 October 2025; Accepted: 19 January 2026; Published: 12 March 2026

ABSTRACT: Nowadays, Unmanned Aerial Vehicles (UAVs) are making increasingly important contributions to numerous applications that enhance human quality of life, such as sensing and data collection, computing, and communication. However, communication between UAVs still faces challenges due to high-dynamic topology, volatile wireless links, and strict energy budgets. In this work, we introduce an improved communication scheme, namely Proximal Policy Optimization (PPO). Our solution casts hop–by–hop relay selection as a Markov decision process and develops a decentralized Proximal Policy Optimization framework in an actor–critic form. A key novelty is the design of the reward function, which jointly considers the delivery ratio, end-to-end delay, and energy efficiency, enabling flexible prioritization in dynamic environments. The simulation results across swarms of 20–70 UAVs show that, the proposed framework enhances delivery ratio to 5% over a Deep Q-Network baseline (reaching ≈ 80% at 70 nodes), reduces latency by about 2–3 ms in medium-to-dense settings (from ~ 43 to 35–36 ms), and attains comparable or slightly lower total energy consumption (typically 0.5%–2% lower). The results indicate that the proposed communication scheme, adaptive and scalable learning-based UAV scenarios, pave the way for re-world UAV deployments.

KEYWORDS: Reinforcement learning; proximal policy optimization (PPO); UAV; 6G

1 Introduction

Unmanned Aerial Vehicles (UAVs) have emerged as one of the potial solutions for countless of civilian applications, such as disaster monitoring [1], search and rescue [2], smart agriculture [3], delivery and logistics [4], environment sensing [5], and traffic monitoring [6]. Thanks to their high mobility, flexible deployment, and advanced sensing capabilities, UAVs play an important role in rapid data collection and wide-area coverage when traditional communication networks are unavailable. Furthermore, the integration of UAVs within future 6G networks is expected to revolutionize wireless connectivity through space-air-ground integrated networking, enabling ubiquitous, ultra-reliable, and intelligent services for next-generation applications such as autonomous transportation, massive Internet of Things (IoT), and global emergency response [7]. This vision positions UAVs as core participants for realizing large-scale, resilient, and adaptive communication infrastructures in the 6G era [8]. Fig. 1 illustrates a range of UAV-enabled applications such as disaster monitoring, search and rescue, delivery logistics, environmental sensing, smart agriculture, and traffic surveillance. This highlights the interactions among UAVs, ground stations, and wireless links, clarifying the process by which UAVs collect, relay, and transmit information. When multiple

UAVs operate collaboratively, they form a flying ad hoc network (FANET), where each UAV not only carries payloads but also acts as a communication relay.

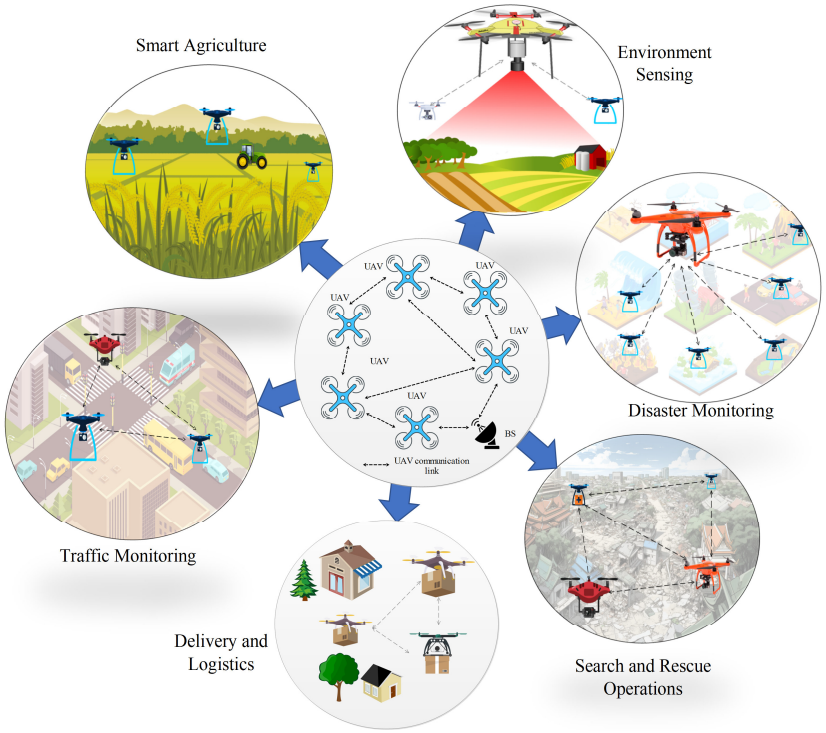

Figure 1: Several typical civilian UAV-enabled systems.

FANETs are crucial to ensure reliable coordination, data exchange, and route planning among UAVs. However, the management of FANETs faces challenges due to the high mobility, dynamic topology, limited onboard energy, and unstable wireless links. These factors pose challenges in designing efficient communication solutions for UAVs. Conventional communication schemes lack the flexibility and adaptability required to be effective under highly dynamic environments, and often rely on static metrics, periodic control messages, or greedy forwarding decisions, which result in high routing overhead, frequent link failures, and reduced packet delivery performance in UAV swarms. As a result, there is a strong demand for intelligent routing mechanisms that can adapt in real time to changes in network topology and communication quality.

Recently, Reinforcement Learning (RL) has emerged as a promising paradigm for intelligent network management, enabling agents to learn optimal routing policies through interaction with the environment [9]. Unlike static rule-based routing, RL-based protocols can adapt to environmental uncertainty such as fluctuating link quality or changing UAV positions. Among RL approaches, actor–critic methods, which combine policy learning (actor) with value estimation (critic), provide a powerful framework for sequential decision-making in complex, dynamic networks than traditional Q-learning methods. In this study, we propose a routing protocol based on Proximal Policy Optimization (PPO) tailored for FANETs in civilian applications. Our key contributions can be summarized as follows:

- Designing FANET communication process as a Markov Decision Process (MDP), defining suitable state, action, and reward representations that capture UAV mobility, link quality, and energy constraints.
- Proposing an actor–critic PPO framework to learn adaptive routing policies that maximize packet delivery while minimizing delay and energy consumption.

- Evaluating the proposed protocol in a simulation environment and comparing it with baseline protocols. Results demonstrate significant improvements in terms of packet delivery ratio, end-to-end delay, and energy efficiency.

The rest of the paper is organized as follows. Section 2 presents related works on routing in FANETs and reinforcement learning-based approaches. Section 3 presents the proposed solution and actor-critic framework. Section 4 demonstrates the experimental results along with related analyses. The potential research directions are outlined in Section 5, and conclusions are presented in Section 6.

2 Related Work

The communication in FANETs faces challenges such as fast topology changes, intermittent connectivity, and stringent energy constraints. Traditional communication protocols offer low overhead but degrade under high mobility due to stale neighborhood information and myopic link metrics. Recent surveys emphasizing the need for learning-based approaches that adapt to temporal dynamics and uncertainty in air–ground links [10]. In this context, RL has emerged as a potential solution to couple routing/forwarding with mobility, power control, and spectrum decisions in a closed-loop manner.

Value-based RL: One of the most typical techniques of the approach is the DQN scheme. The study [11] demonstrated a multi-agent imitation learning approach for differentiated UAV services, showing that data-driven policies improve deployment decisions over handcrafted heuristics in mobile edge computing (MEC) scenarios. While effective, such methods can be brittle under policy shifts and partial observability due to a lack of explicit trust-region safeguards, which are typical of modern policy-gradient methods.

Multi-agent RL: To improve reliability under adversarial environments, the work in [12] proposed a multi-agent RL framework for UAV swarm communications that jointly optimizes relay selection and power allocation against jamming. The study evidences the gains of centralized training with decentralized execution (CTDE) in highly dynamic air-to-air links, a setting analogous to multi-relay selection in FANET routing. Beyond robustness, multi-agent Deep Reinforcement Learning (DRL) has been leveraged to plan trajectories and allocate MEC resources jointly. The study in [13] developed a multi-agent DRL scheme for distributed trajectory optimization under differentiated services, achieving latency and energy improvements by co-optimizing UAV motion and task processing. These results motivate routing formulations that elevate next-hop selection to part of a larger control surface spanning mobility and computation.

Energy efficiency and reliable relaying: Energy-aware formulations are pivotal for battery-limited UAVs. The work [14] addresses an energy-efficient 3D access problem and demonstrates that carefully designed access/placement policies reduce the total UAV count while meeting service constraints. Additionally, the study [15] introduces reliable and energy-efficient relaying via collaborative beamforming, demonstrating that learning-aided controllers can balance rate and energy under rapidly fluctuating channels. These works inform routing cost functions that trade off delivery ratio, delay, and propulsion/radio energy in FANETs.

Policy-gradient methods and PPO: In the domain of continuous control, reinforcement learning agents based on value functions are often susceptible to training instability and high variance. To mitigate these shortcomings, trust-region policy optimization (TRPO) and PPO algorithms were developed. These methods constrain the magnitude of policy updates by employing a clipped surrogate objective function, which substantially enhances training stability and improves sample efficiency. Illustrating this in a networking context, the study in [16] utilized PPO to autonomously optimize transmission periods for IoT edge devices. Their research demonstrated robust convergence and a favorable trade-off between latency and energy consumption, even under non-stationary traffic conditions.

3 Proposed Framework

3.1 RL and Actor-Critic Framework

In this subsection, we will introduce the RL and the actor-critic scheme, as well as the proposed framework to enhance system performance.

3.1.1 Reinforment Learning

RL is a widely adopted branch of artificial intelligence designed to address decision-making problems that evolve over time under uncertainty. In this framework, a learning entity interacts continuously with its surrounding environment by taking actions and observing their outcomes. The learning objective is to optimize a long-term performance measure, commonly defined as the cumulative expected return. Through repeated interaction and experience, the learner gradually develops an optimal or near-optimal decision-making strategy, known as a policy, that maximizes the total reward over an extended horizon. The fundamental components that characterize the RL framework are summarized as follows:

- **Agent:** The autonomous learner responsible for selecting actions based on observed information and its current decision policy.
- **Environment:** The external system that responds to the agent's actions by changing its condition and issuing evaluative feedback.
- **State (S):** A description or observation capturing the current condition of the environment as perceived by the agent.
- **Action (A):** A set of possible decisions or controls that the agent may apply to influence the environment.
- **Reward (R):** A scalar feedback signal that quantifies the immediate benefit or penalty resulting from an action.
- **Policy (π):** A mapping that defines the agent's action-selection behavior for each possible state.

The interaction process proceeds sequentially over discrete time steps. At time instant t, the agent observes the current state S_t and selects an action A_t according to its policy π. The environment then reacts to this action, transitioning to a subsequent state S_{t+1} and generating a corresponding reward R_{t+1}. This cycle of observation, action, and feedback continues until a terminal condition is reached or the learning episode terminates. RL algorithms are commonly categorized based on their learning paradigm, particularly into value-based, policy-based, actor–critic, and model-based approaches [17].

- **Value-based** methods aim to approximate a value function (e.g., Q-function) that estimates the reward return of taking an action in a given state. The optimal policy is derived implicitly by selecting actions that maximize this value. Representative algorithms include Q-learning and Deep Q-Networks (DQN). While effective in discrete action spaces, these methods face limitations when extended to a continuous domain.
- **Policy-based** methods directly learn a parameterized policy without relying on an intermediate value function. Optimization is performed through gradient ascent on the expected return, with well-known examples such as REINFORCE, Trust Region Policy Optimization (TRPO), and PPO. These methods are suitable for continuous action spaces and expressive policy representations, but they often suffer from high variance and require substantial data for convergence.
- **Actor–Critic** methods integrate the advantages of the previous two paradigms, where the actor represents the policy and the critic evaluates it through a value function. This structure improves the stability and efficiency of policy updates, as demonstrated by algorithms such as Advantage Actor–Critic and Deep Deterministic Policy Gradient (DDPG). Nevertheless, balancing the learning dynamics between actor and critic remains a challenge.

- **Model-based** methods differ fundamentally in that they attempt to learn or assume an explicit model of the environment's dynamics, which can then be exploited for planning and decision making. Techniques such as Dyna-Q exemplify this approach.

3.1.2 Actor-Critic Framework

The Actor-Critic framework represents a foundational architecture in reinforcement learning, designed to integrate the advantages of both policy-based and value-based methods. In this approach, the learning agent maintains two interdependent components: an actor, which determines the policy for selecting actions, and a critic, which evaluates those actions through a value estimation process. This dual structure allows the agent to refine its policy in a more stable and data-efficient manner compared to purely value-based or policy-based schemes [18].

Fig. 2 illustrates the standard architecture of the actor–critic framework in reinforcement learning. This model comprises three main components: the *Actor*, the *Critic*, and the *Environment*. Each component has a distinct role but operates in a tightly coupled feedback loop, ensuring both policy exploration and value estimation are jointly optimized. During operation, the *Actor* is represented by a parameterized policy, denoted as $\pi_\theta(a|s)$, which determines the action a_t to be executed given the current state s_t. The selected action is then applied to the *Environment*, which generates a reward signal r_t and transitions to the next state s_{t+1}. These feedback signals from the environment are subsequently used to evaluate the performance of the current policy and guide its refinement. The *Critic* acts as an evaluator by estimating the state value function $V_w(s)$.

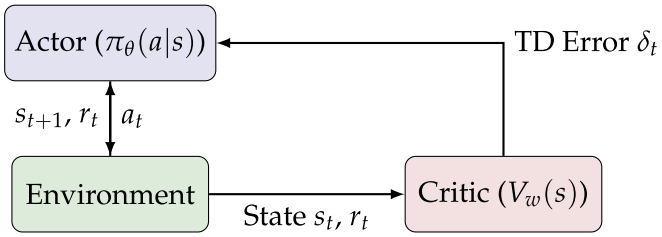

Figure 2: Standard structure of the actor–critic model.

$$V_w(s) \approx V^\pi(s) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty} \Upsilon^k r_{t+k} \,\middle|\, s_t = s\right] \tag{1}$$

At each timestep t, the temporal-difference (TD) error is computed as follows:

$$\delta_t = r_t + \Upsilon V_w(s_{t+1}) - V_w(s_t) \tag{2}$$

where Υ is the discount factor that reflects the importance of future rewards. The TD error δ_t serves as a feedback signal to the actor, enabling it to adjust the policy direction to maximize the long-term expected return. The critic parameters are updated using gradient descent to minimize this error:

$$w = w + \beta_c \, \delta_t \, \nabla_w V_w(s_t) \tag{3}$$

where β_c denotes the learning rate of the critic. Meanwhile, the actor updates its policy parameters following the policy gradient theorem:

$$\theta = \theta + \beta_a \, \delta_t \, \nabla_\theta \log \pi_\theta(a_t|s_t) \tag{4}$$

where β_a is the actor's learning rate. Through this mechanism, the TD error acts as a reinforcement signal that aligns the actor's policy toward actions yielding higher estimated returns, while penalizing suboptimal choices.

Fundamentally, the actor is responsible for *exploration* and *decision-making*, whereas the critic ensures *evaluation* and *stability* during the learning process. This reciprocal relationship forms a synergistic learning mechanism: the actor improves its policy based on the critic's feedback, and the critic enhances its value estimation using the actor's updated behavior. Such two-way interaction accelerates convergence and reduces variance compared to methods that rely solely on either policy gradients or value functions.

While the Actor-Critic framework provides a robust foundation for policy learning, it remains sensitive to large or unstable policy updates, which may lead to divergence during training. To address this limitation, advanced algorithms such as Proximal Policy Optimization (PPO) have been developed to stabilize learning by introducing constrained updates to the policy network.

3.1.3 Proximal Policy Optimization

PPO is a refinement of the Actor-Critic framework designed to improve training stability and sample efficiency [19]. The key innovation of PPO lies in its use of a clipped surrogate objective function, which restricts the magnitude of policy updates between iterations. This ensures smoother learning dynamics and prevents destructive policy shifts during the training process. This constraint prevents destructive updates that can occur in conventional policy gradient methods, effectively balancing exploration and stability without requiring complex second-order optimization. PPO defines the clipped objective as:

$$L^{\text{CLIP}}(\theta) = \mathbb{E}_t \left[\min \left(r_t(\theta) A_t, \text{clip}\big(r_t(\theta), 1- \epsilon, 1+ \epsilon \big) A_t \right) \right], \tag{5}$$

where $r_t(\theta) = \dfrac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{\text{old}}}(a_t|s_t)}$ represents the probability ratio between the new and old policies, A_t is the estimated advantage function, and ϵ is a small positive constant (typically in the range $[0.1, 0.3]$) that bounds the policy update.

3.2 PPO-Based RL for FANET Routing

3.2.1 Problem Setting and Design Goals

We address hop–by–hop relay selection in FANETs under rapid topology changes, fluctuating link quality, and on-board compute limits. The objective is to maximize packet delivery ratio while controlling end-to-end delay and control overhead, using a decentralized actor–critic agent at each UAV. The proposed implementation follows a proximal policy optimization (PPO) scheme with (i) generalized advantage estimation (GAE), (ii) true mini-batch updates with candidate masking/padding, (iii) a shared-encoder backbone, (iv) normalized advantages, and (v) learning-rate scheduling, as realized in the provided code base.

3.2.2 System and Network Model

Each node i observes a compact *state* vector and a set of *per-candidate* neighbor features. The improved implementation instantiates an 8-D state and a 7-D candidate descriptor, enabling robust selections under

varying neighbor counts via masking/padding during inference and training. Formally, at time t, the local state s_t^i aggregates position/velocity proxies, residual energy, queue/load surrogates, and link-quality statistics; each candidate j contributes features such as normalized distance, normalized bitrate/signal to interference plus noise ratio proxies, progress-to-destination, and relative mobility. (Feature construction and normalization are implemented in the utility routines for state and candidate extraction.)

We formulate, for each node i, a discounted MDP $\mathcal{M}_i = (\mathcal{S}_i, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \Upsilon)$. The *state* $s_t^i \in \mathbb{R}^8$ summarizes local kinematics/resources and a permutation-invariant summary of neighbor descriptors; the *action* a_t^i selects the next hop among feasible neighbors using a state-dependent mask (invalid/padded candidates are pruned before softmax). Transitions are induced by wireless/Media Access Control (MAC) outcomes and mobility; the reward couples terminal outcomes with a per-step shaping tied to link quality (Section 3.2.4). Advantage estimates and PPO updates follow the clipped surrogate with entropy regularization.

3.2.3 Policy, Value Function, and Learning Signals

The actor implements a parametric policy $\pi_\theta(a \mid s)$ that scores each candidate by concatenating the shared state embedding with per-candidate features; logits are masked for invalid entries and then normalized by a categorical distribution. The critic approximates the state value $V_w(s)$. The improved implementation realizes both actor and critic on top of a shared encoder (two hidden layers with LayerNorm and ReLU), followed by task-specific heads; weights are orthogonally initialized for stable optimization. At decision time, the system pads candidate sets to a common length and carries a Boolean mask *through* the forward pass to ensure permutation invariance and numerical stability. To estimate A_t, PPO-Proposed uses the generalized advantage estimator (GAE) [20], which mitigates variance while preserving bias in Eq. (6).

$$A_t = \sum_{l=0}^{\infty} (\Upsilon\lambda)^l \delta_{t+l} \tag{6}$$

where $\lambda \in [0,1]$ regulates the bias–variance trade-off. The combination of GAE and the clipped surrogate objective significantly improves the stability and consistency of policy updates across different tasks. PPO-Proposed is implemented in an Actor-Critic form, where the actor updates the policy according to the clipped objective, and the critic minimizes the value function loss in Eq. (7).

$$L^{\text{value}}(w) = \mathbb{E}_t\left[\left(V_w(s_t) - V_t^{\text{target}}\right)^2\right] \tag{7}$$

The final objective function is a weighted sum of the clipped policy loss, value loss, and an entropy regularization term that encourages policy exploration.

$$L^{\text{total}} = L^{\text{CLIP}}(\theta) - c_1 L^{\text{value}}(w) + c_2 S[\pi_\theta] \tag{8}$$

where $S[\pi_\theta]$ denotes the entropy of the policy, and c_1, c_2 are weighting coefficients. Through this formulation, PPO-Proposed achieves a robust trade-off between sample efficiency, learning stability, and implementation simplicity.

PPO-Proposed uses GAE with parameters $(\Upsilon, \lambda) = (0.99, 0.95)$ to reduce variance, with advantages normalized per-trajectory before batching; the critic regresses to returns $R_t = A_t + V_w(s_t)$. Mini-batch updates apply the clipped surrogate, squared value loss, and an entropy bonus with coefficients $(\epsilon, c_1, c_2) = (0.2, 0.5, 0.02)$ and gradient clipping at $\|g\|_2 \le 0.5$.

3.2.4 Reward Function

We implement event-driven shaping: successful delivery results in a strong positive terminal reward, while failure results in a negative terminal reward. Ongoing forwarding earns a small base reward, minus a per-step cost, and is further enhanced by a link-quality bonus, as shown in Eq. (9).

$$r_t = \begin{cases} 2.0, & \text{delivered (terminal)} \\ 0.1 + 0.3\,\ell_t - \texttt{step\_cost}, & \text{in progress} \\ -2.0, & \text{failure (terminal)} \end{cases} \tag{9}$$

where $\ell_t \in [0,1]$ is a normalized indicator that reflects current link quality—effectively representing delay and communication reliabqity. The configuration exposes `step_cost` (default 0.01 in the base class). The base reward (0.1) and the positive link-quality term (weighted by 0.3) encourage the policy to select routes with both low delay and high delivery probability, while `step_cost` penalizes long or inefficient forwarding paths and is proportional to per-hop energy consumption. Terminal states (successful delivery or failure) receive strong positive (+2.0) or negative rewards (−2.0), respectively, aligning the learning objective with maximizing delivery ratio. All coefficients are tunable and can be adapted to prioritize delay, energy, or reliability as desired. This structure ensures that the agent optimizes for high delivery success, low latency, and energy efficiency in a unified reward signal. One-step TD targets and advantages are then formed for PPO-Proposed/GAE as implemented in the update routines.

3.2.5 Protocol Operations

Data plane (online selection). Given neighbors, the agent builds s_t, extracts/pads candidate features, applies the mask, samples $a_t \sim \pi_\theta(\cdot|s_t)$, and forwards the packet. The decision record stores $(s_t, \phi_{ij}, \text{mask}, a_t, \log \pi_{\text{old}}, V_w(s_t))$ for learning.

Control plane (feedback and update). Upon feedback, the reward has been computed with link-quality fallback estimation, appends the sample, and triggers PPO updates once a trajectory or buffer threshold is reached. Updates run in mini-batches with dynamic padding.

Fig. 3 describes the PPO-Proposed routing pipeline used for hop-by-hop relay selection. At each decision instant, the *Actor* implements a stochastic policy $\pi_\theta(a \mid s)$ that maps the current observation s_t (from the *Environment*) to an action a_t (next hop). Executing a_t yields a reward r_t and the successor state s_{t+1}. The *Critic* estimates the state value $V_w(s)$ and together with r_t, supplies $V_w(s_t)$ and $V_w(s_{t+1})$ to the *Advantage (GAE)* module to form A_t (via TD residuals). The advantage A_t then enters the *Clipped Objective* $L^{\text{CLIP}}(\theta)$, which also receives the current policy probability $\pi_\theta(a_t \mid s_t)$ and the *Old Policy* $\pi_{\theta_{\text{old}}}(a_t \mid s_t)$ to construct the ratio $r_t(\theta) = \pi_\theta / \pi_{\theta_{\text{old}}}$; an entropy term $S[\pi_\theta]$ regularizes exploration. Gradients with respect to θ flow from $L^{\text{CLIP}}(\theta)$, while gradients with respect to w originate from the *Value Loss* $L^{\text{value}}(w)$, which regresses $V_w(s_t)$ toward a TD-based target. Both gradient streams meet at the *Parameter Update* block, producing updated parameters (θ, w) that are applied back to the actor and critic, respectively. A dashed arrow indicates the snapshot operation $\theta_{\text{old}} \leftarrow \theta$ used to refresh the behavior policy for the next PPO-Proposed epoch. Overall, the diagram highlights the closed-loop interaction between data collection (Actor–Environment), evaluation (Critic/GAE), trust-region-like policy improvement (clipped surrogate with entropy), and synchronized parameter updates.

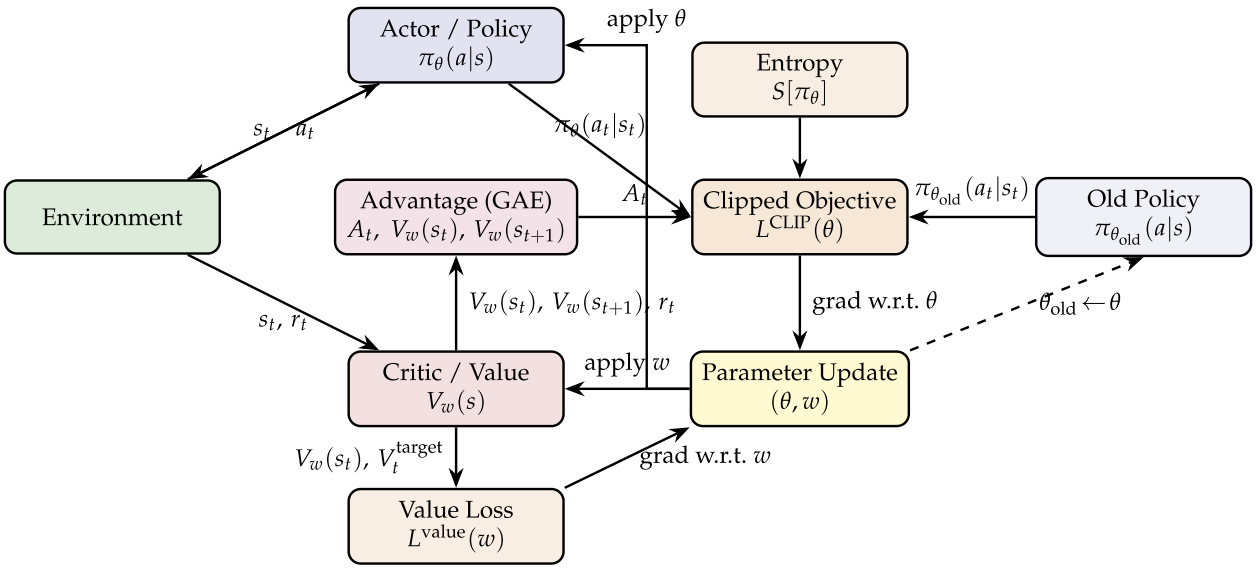

Figure 3: PPO-proposed model.

Algorithm 1 outlines the workflow of the proposed PPO-based RL protocol. It summarizes the state construction, relay selection based on learned policy, reward evaluation, sample collection, and policy parameter updates performed at each UAV node during network operation.

Algorithm 1: PPO-proposed algorithm

Inputs: actor π_θ, critic V_w, buffer $\mathcal{B}$, mask operator mask$(\cdot)$
Hyperparameters: γ, λ, ε, c_1, c_2, batch, update_every, epochs
procedure RELAY-SELECTION ($\mathcal{N}_i(t)$)
    Build state s_t and per-candidate features $\{\phi_{ij}\}$; pad to $N_{\max}$ and construct mask M *(padding/mask)*
    $a_t \sim$ Cat (softmax(`mask_logits` $(\pi_\theta(s_t, \phi), M)$)); forward packet
    Record $(s_t, \phi, M, a_t, \log \pi_{\text{old}}, V_w(s_t))$ for feedback
end procedure
procedure FEEDBACK (o_t, id$_j$, ℓ_t)
    Compute reward r_t via event-driven shaping; estimate v_{t+1}; push sample to $\mathcal{B}$
    if $|\mathcal{B}| \geq$ update_every **then** PPO-UPDATE($\mathcal{B}$); clear $\mathcal{B}$
    end if
endprocedure
procedure PPO-UPDATE ($\mathcal{B}$)
    Build trajectories, compute A_t using GAE(Υ, λ); normalize advantages
    for epoch = 1:epochs **do**
        for mini-batch $\mathcal{M} \subset \mathcal{B}$ **do**
            Pad candidate tensors and masks to $\max_{x \in \mathcal{M}} |\mathcal{N}(x)|$
            Compute ratio = $\exp(\log \pi_\theta - \log \pi_{\text{old}})$; $L_\pi = -\min(\text{ratio}A, \text{clip}(\text{ratio}, 1\pm\epsilon)A)$
        $L_V = \|V_w - \text{return}\|_2^2$, $L = L_\pi + c_1 L_V - c_2 \mathbb{E}$; update with grad clip $\|g\|_2 \leq 0.5$
        end for
    end for
end procedure

4 Results and Discussion

4.1 Experimental Parameters

To validate the effectiveness of the proposed approach, simulation experiments are conducted using the DroNETworkSimulator platform[1]. The number of unmanned aerial vehicles (UAVs) varies between 20 and 70 to examine scalability. These UAVs are randomly deployed within a two-dimensional area measuring 2000 m^2 × 2000 m^2. The complete set of simulation settings is summarized in Table 1.

Table 1: Experimental parameters.

Parameters	Values
Simulation Area	2000 m × 2000 m × 50 m
Number of UAVs	[20–70]
UAV Speed	12 m/s
Transmission Range	200 m
Protocol	DQN, PPO-Proposed
Battery model	Energy linear
Energy initial	1200 KJ
MAC Layer	802.11n
Mobility model	Random Waypoint Mobility
PPO-Proposed hyperparameters	Υ = 0.99, λ = 0.95, ϵ = 0.2, c_1 = 0.5, c_2 = 0.02, batch = 64, update_every = 1024, epochs = 10

The considered scenarios reflect practical applications such as smart agriculture monitoring and search-and-rescue operations in smart city environments. Accordingly, the UAV cruising velocity is configured at approximately 12 m/s, while the maximum inter-UAV communication range is limited to around 220 m. Based on this setup, we consider the performance of the DQN and the PPO-Proposed protocols. We use common metrics to evaluate, including end-to-end delay, packet delivery ratio (PDR), and energy consumption. Although real-world UAV missions typically follow more structured or task-driven trajectories, we adopt the Random Waypoint (RWP) model as a standard and widely used baseline for FANET evaluations. RWP induces continuous topology variations, providing a suitable stress-test environment for assessing routing robustness under highly dynamic link conditions, which aligns with the primary objective of this study. Moreover, prior UAV networking literature shows that routing performance trends remain consistent across mobility models when the focus is on link dynamics rather than precise flight kinematics. Future work will incorporate mission-oriented mobility patterns to further strengthen the generalizability of the results.

4.2 Simulation Results

Fig. 4 presents the latency as the number of UAVs increases. The DQN baseline shows that the latency decreases from approximately 46 ms at 20 UAVs to ≈38–39 ms at 70 UAVs as connectivity improves. While PPO-Proposed provides the lowest delay across all settings, from approx. 43 ms to 35–36 ms. The improvement is rather small in sparse networks but shows an uptrend when the number of increasing UAVs ranges from 50 to 70 nodes, indicating more efficient forwarding under higher contention.

[1] https://github.com/Andrea94c/DroNETworkSimulator

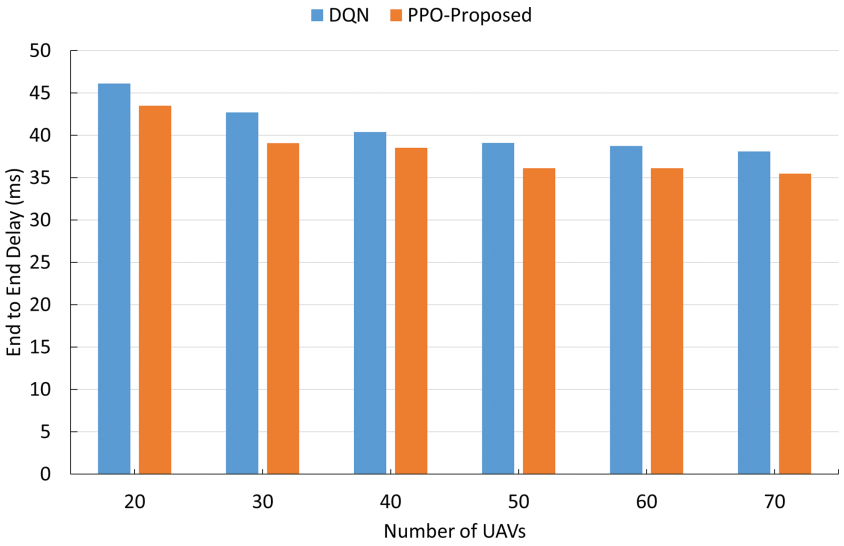

Figure 4: End-to-end delay.

These results were achieved thanks to PPO's clipped updates and advantage estimation, which stabilize policy improvement alongside link-quality features and action masking, thereby reducing retransmissions and queueing along the selected paths.

Fig. 5 demonstrates the packet delivery ratio as the number of increasing UAVs. Both solutions benefit from higher node density. While DQN improves from 65% at 20 nodes to about 76% at 70 nodes, the PPO-Proposed achieves a higher PDR rate, increasing from 66% at 20 nodes to 80% at 70 nodes. The results also show lower achieved performance in sparse networks and an upward trend as the number of UAVs increases, particularly between 50 and 70 nodes, reflecting superior robustness under increased contention and faster topology changes. These results were achieved thanks to PPO-Proposed's clipped policy updates and advantage estimation, which stabilize the learning process. Together with link-quality features and action masking, these promote reliable next-hop choices in FANETs.

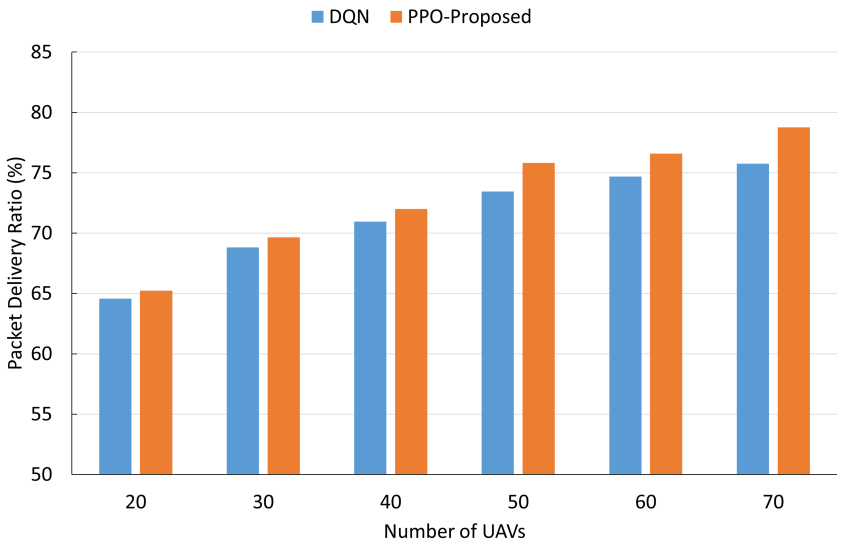

Figure 5: Packet delivery ratio.

Fig. 6 compares the total energy consumption as the size of the increasing UAV swarm. Both schemes exhibit an almost linear increase with the number of UAVs: the **DQN** baseline rises from roughly 440 KJ at 20 nodes to just above 2000 KJ at 70 nodes. The proposed solution also has a similar trend but improved across most settings (~430 KJ at 20 nodes and ~2010 KJ at 70 nodes). These results were achieved thanks to PPO-Proposed clipped policy updates and GAE-stabilized forwarding decisions, which, together with action masking and link-quality features, reduce retransmissions and route repairs.

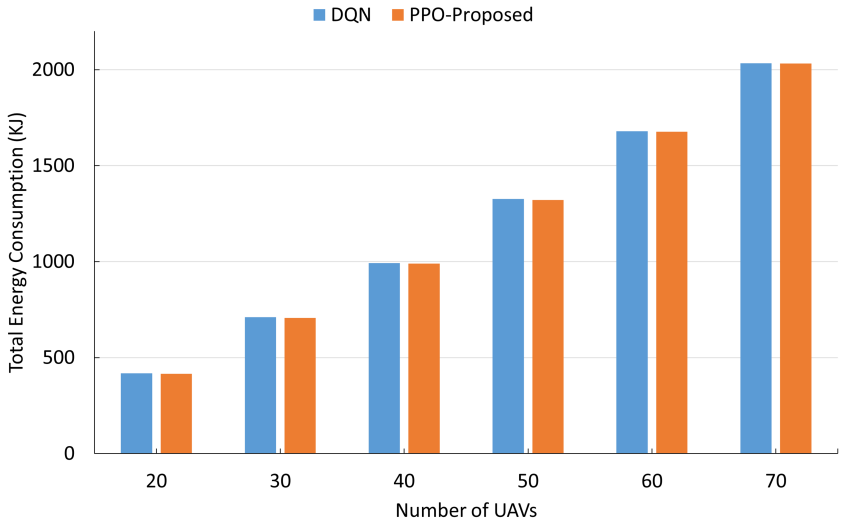

Figure 6: Energy consumption.

5 Future Research Directions

The above analyses demonstrate that civilian UAVs will increasingly contribute to countless human tasks. Besides, it also shows problems. We determine several challenges and potential research directions, as follows.

- **Model Realism and 6G Integration:** Future research should aim to couple PPO agents with the non-terrestrial networks and cell-free architectures envisioned for 6G. This requires explicit modeling of aerial-specific interference patterns, dynamic spectrum sharing, and mobility-induced Doppler effects [21].
- **Multi-Agent Coordination:** The cooperative routing problems, such as resource optimization in relay selection, power control, and medium access, can be modelled as a multi-agent PPO problem under the CTDE paradigm. This approach would extend recent advances in anti-jamming and trajectory planning to packet-level relaying policies for UAV swarms [12].
- **Safety- and Regulation-Aware RL:** Deployable FANETs must operate in strict compliance with airspace regulations such as no-fly zones, altitude corridors, and communication safety protocols. Applying constrained PPO variants, such as those employing Lagrangian or primal-dual surrogate objectives, alongside shielded exploration techniques, can enforce hard constraints on link loads, interference budgets, and geofences while preserving policy improvement guarantees [21].
- **Energy and Lifecycle Modeling:** Beyond communication overhead, propulsion energy is the primary factor in the operational budget of UAVs. A significant opportunity lies in the co-optimization of trajectory-aware routing policies with energy-saving mechanisms, such as duty cycling and wake-sleep schedules [14].

- **Edge Intelligence and Privacy Preservation:** Civilian applications such as disaster response and smart agriculture demand on-board learning with stringent latency and privacy requirements. Federated or on-device PPO frameworks, augmented with periodic policy distillation, can minimize backhaul dependency and mitigate data exfiltration risks while ensuring continuous adaptation to local conditions [16].
- **Enhancing Robustness and Generalization:** To develop policies that are resilient to unforeseen conditions, domain randomization across mobility, traffic, and channel models should be systematically paired with risk-sensitive objectives. This combination is critical for mitigating high-impact tail events, such as congestion bursts and catastrophic link failures.
- **Simulations to Real-world Applications:** Validation through flight trials incorporating real radio hardware is indispensable. Such experiments are necessary to verify sample efficiency, assess stability under real-time topological churn, and confirm compatibility with standard routing stacks, thereby closing the critical sim-to-real gaps [22].

6 Conclusion

This work presented a reinforcement learning–driven routing strategy for FANETs, referred to as PPO-Proposed. The proposed scheme enhances the relay node selection mechanism by jointly considering multiple network performance indicators, including channel reliability, remaining energy, and communication latency, within a Proximal Policy Optimization (PPO) learning framework. Simulation-based evaluations indicate that the proposed protocol achieves a higher packet delivery ratio while simultaneously reducing end-to-end delay compared to the baseline DQN-based routing approach. These performance gains highlight the effectiveness of the proposed method in dense and highly mobile UAV network environments. In the future, we will explore the incorporation of advanced RL paradigms, such as multi-agent RL and federated learning, to further improve network scalability, robustness, and data privacy. Such enhancements are expected to facilitate practical deployment in civilian scenarios and emerging 6G-enabled Internet of Things applications, including intelligent transportation systems, emergency management, and large-scale urban sensing.

Acknowledgement: None.

Funding Statement: This research is funded by Hung Yen University of Technology and Education under grant number UTEHY.L.2026.05.

Author Contributions: The authors confirm contributions to the paper as follows: study conception and design: Chu Thi Minh Hue and Nguyen Minh Quy; data collection: Chu Thi Minh Hue; analysis and interpretation of results: Nguyen Minh Quy; draft manuscript preparation: Chu Thi Minh Hue and Nguyen Minh Quy. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Not applicable.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Xu W, Wang C, Xie H, Liang W, Dai H, Xu Z, et al. Reward maximization for disaster zone monitoring with heterogeneous UAVs. IEEE/ACM Trans Netw. 2024;32(1):890–903. doi:10.1109/tnet.2023.3300174.
2. Qi S, Lin B, Deng Y, Chen X, Fang Y. Minimizing maximum latency of task offloading for multi-UAV-assisted maritime search and rescue. IEEE Trans Veh Technol. 2024;73(9):13625–38. doi:10.1109/tvt.2024.3384570.

3. Akbari M, Syed A, Kennedy WS, Erol-Kantarci M. Constrained federated learning for AoI-limited SFC in UAV-aided MEC for smart agriculture. IEEE Trans Mach Learn Commun Netw. 2023;1:277–95. doi:10.1109/tmlcn.2023.3311749.
4. Chen J, Wan P, Xu G. Cooperative learning-based joint UAV and human courier scheduling for emergency medical delivery service. IEEE Trans Intell Trans Syst. 2025;26(1):935–49. doi:10.1109/tits.2024.3486789.
5. Li CM, Wu LC, Wang PJ. Integrated environment sensing and green communication for non-terrestrial network. IEICE Trans Commun. 2025;E108-B(7):851–8. doi:10.23919/transcom.2024ebp3166.
6. Kumar VDA, Ramachandran V, Rashid M, Javed AR, Islam S, Al Hejaili A. An intelligent traffic monitoring system in congested regions with prioritization for emergency vehicle using UAV networks. Tsinghua Sci Technol. 2025;30(4):1387–400. doi:10.26599/tst.2023.9010078.
7. Xu J, Xu X, Cui G, Bilal M, Gu R, Dou W, et al. A privacy-preserving auction for task offloading and resource allocation in UAV-assisted MEC. IEEE Trans Mobile Comput. 2025;25(2):2611–26. doi:10.1109/tmc.2025.3609202.
8. Toka L, Konrad M, Pekar A, Biczók G. Integrating the skies for 6G: techno-economic considerations of LEO, HAPS, and UAV technologies. IEEE Commun Mag. 2024;62(11):44–51. doi:10.1109/mcom.003.2400120.
9. Nazib RA, Moh S. Reinforcement learning-based routing protocols for vehicular Ad Hoc networks: a comparative survey. IEEE Access. 2021;9:27552–87. doi:10.1109/access.2021.3058388.
10. Mansoor N, Hossain MI, Rozario A, Zareei M, Rodriguez-Arreola A. A fresh look at routing protocols in unmanned aerial vehicular networks: a survey. IEEE Access. 2023;11:66289–308. doi:10.1109/access.2023.3290871.
11. Wang X, Ning Z, Guo S, Wen M, Guo L, Poor HV. Dynamic UAV deployment for differentiated services: a multi-agent imitation learning based approach. IEEE Trans Mobile Comput. 2023;22(4):2131–46. doi:10.1109/tmc.2021.3116236.
12. Lv Z, Xiao L, Du Y, Niu G, Xing C, Xu W. Multi-agent reinforcement learning based UAV swarm communications against jamming. IEEE Trans Wirel Commun. 2023;22(12):9063–75. doi:10.1109/twc.2023.3268082.
13. Ning Z, Wang X, Song Q, Guo L, Wen M, Guo S, et al. Multi-agent deep reinforcement learning based UAV trajectory optimization for differentiated services. IEEE Trans Mobile Comput. 2024;23(5):5818–34. doi:10.1109/tmc.2023.3312276/mm1.
14. Gong H, Huang B, Jia B. Energy-efficient 3-D UAV ground node accessing using the minimum number of UAVs. IEEE Trans Mobile Comput. 2024;23(12):12046–60. doi:10.1109/tmc.2024.3405494.
15. Zheng X, Sun G, Li J, Liang S, Wu Q, Jin M, et al. Reliable and energy-efficient communications via collaborative beamforming for UAV networks. IEEE Trans Wirel Commun. 2024;23(10):13235–51. doi:10.1109/mwc.001.2100677.
16. Lee GH, Park H, Jang JW, Han J, Choi JK. PPO-based autonomous transmission period control system in IoT edge computing. IEEE Int Things J. 2023;10(24):21705–20. doi:10.1109/jiot.2023.3293511.
17. Arulkumaran K, Deisenroth MP, Brundage M, Bharath AA. Deep reinforcement learning: a brief survey. IEEE Signal Proc Mag. 2017;34(6):26–38. doi:10.1109/msp.2017.2743240.
18. Konda VR, Tsitsiklis JN. Actor-critic algorithms. SIAM J Control Optim. 2003;42(4):1143–66.
19. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv:1707.06347. 2017.
20. Schulman J, Moritz P, Levine S, Jordan M, Abbeel P. High-dimensional continuous control using generalized advantage estimation. In: Proceedings of the 4th International Conference on Learning Representations (ICLR); 2016 May 2–4; San Juan, Puerto Rico.
21. Geraci G, Garcia-Rodriguez A, Azari MM, Lozano A, Mezzavilla M, Chatzinotas S, et al. What will the future of UAV cellular communications be a flight from 5G to 6G. IEEE Commun Surv Tutor. 2022;24(3):1304–35. doi:10.1109/comst.2022.3171135.
22. Cao P, Lei L, Cai S, Shen G, Liu X, Wang X, et al. Computational intelligence algorithms for UAV swarm networking and collaboration: a comprehensive survey and future directions. IEEE Commun Surv Tutor. 2024;26(4):2684–728. doi:10.1109/comst.2024.3395358.