ISSN 2522-9818 (print) *Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)*
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)*

UDC 629.735.3 DOI: https://doi.org/10.30837/2522-9818.2026.1.028

Maksym Yena, Olha Pohudina

MODELING OF ADAPTIVE UAV ROUTE CONTROL BASED ON REINFORCEMENT LEARNING ALGORITHMS

Subject matter is the reward function, action policy, and learning dynamics of the Proximal Policy Optimization (PPO) algorithm in the task of adaptive UAV navigation under dynamic airspace conditions and limited energy resources. **Goal** is to develop a simulation environment and a modified Proximal Policy Optimization (PPO) model for adaptive route management of a single UAV in 2D and 3D environments, considering the distance to the target, collision risk, and energy consumption. **Tasks**: to develop 2D and 3D simulation environments with different obstacle configurations and UAV motion parameters; to formulate a combined PPO reward function that incorporates distance to the target, collisions, and energy consumption; to implement and train PPO, DQN, and A2C algorithms in standardized navigation scenarios; to perform a comparative analysis of algorithm performance using key metrics: path length, number of collisions, reward, and energy consumption; to conduct statistical validation of the results using the t-test and confidence intervals; to analyze the influence of PPO hyperparameters on policy stability and learning convergence in 2D and 3D environments. **Methods**: deep reinforcement learning algorithms (PPO, DQN, A2C); two simulation models (2D and 3D) with randomly generated static obstacles were developed; a combined reward function was formulated, integrating distance-to-target progress, collision penalties, and an energy-related component; model performance was evaluated using average reward, path length, number of collisions, and total energy expenditure; statistical significance was assessed using the t-test and 95% confidence intervals. **Results**: The modified PPO model reduced the number of collisions in the 2D environment by 94,8% and shortened the route length by 94,3% compared to the baseline PPO, while exhibiting higher energy consumption due to more complex avoidance maneuvers. In the 3D environment, similar trends were confirmed, including improved navigation safety, more stable policy behavior, and statistically significant improvements across key metrics ($p < 0,05$). **Conclusions**: A unified 2D/3D simulation environment for adaptive UAV routing and a modified PPO model with a combined reward function were developed. In the 2D environment, the model achieved a ≈94,8% reduction in collisions, a ≈94,3% reduction in path length, and a ≈92,5% increase in average reward compared to the baseline PPO. In the 3D environment, analogous improvements and statistically significant gains ($p < 0,05$) were obtained. A relationship between avoidance aggressiveness and energy consumption was identified, enabling selection of an optimal policy for BVLOS scenarios.

Keywords: adaptive control; Proximal Policy Optimization; reinforcement learning; simulation modeling; routing; 3D navigation.

1. Introduction

The increasing complexity of airspace and the rapid growth in the number of autonomous UAVs create a need for adaptive navigation methods capable of making real-time decisions in the face of dynamic obstacles and limited energy resources [1]. The problem is that existing adaptive routing methods work only in simplified 2D environments and do not account for energy constraints, which is critical for real-world UTM systems [2]. Classic rule-based navigation methods do not provide adaptability, especially in 3D space with a large number of degrees of freedom. Reinforcement learning (RL) methods, in particular Proximal Policy Optimization (PPO), have demonstrated effectiveness in dynamic evasion and trajectory prediction tasks. However, most studies are limited to two-dimensional environments or simplified obstacle models that do not reflect the actual structure of airspace [3].

2. Literature Review and Problem Definition

Among the most promising areas is the use of reinforcement learning algorithms, particularly Proximal Policy Optimization (PPO), which show good results in complex simulation environments [4]. Dai et al. (2023) demonstrate the effectiveness of PPO in safe agent control tasks; however, the main focus of their work is on two-dimensional (2D) spaces. Studies [5, 6] examine the classification of reinforcement learning algorithms and their ability to generalize in simulation models.

There is also a concept of a simulation environment for urban air traffic, but the implementation of adaptive navigation in three-dimensional (3D) space remains partially unresolved [7]. Most recent works [8, 9] highlight the importance of traffic forecasting that accounts for stochastic factors, but do not address their integration with learning algorithms.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

For the most part, it is the combination of several algorithms that demonstrates the effectiveness of Monte Carlo and discrete-event modeling in traffic scenarios; however, these methods do not combine with dynamic collision-avoidance strategies [10]. Particular attention should be paid to studies analyzing the impact of hyperparameters on the effectiveness of agent training in prediction and dynamic collision avoidance tasks [11, 12].

Despite significant progress in UAV navigation, most modern approaches remain limited to two-dimensional models and simplified obstacles that do not account for vertical maneuvers and the multi-level structure of airspace. This reduces the algorithms' ability to generalize behavior in realistic 3D scenarios.

Thus, there is a need to create a simulation model of adaptive UAV route control based on PPO, focused on realistic scenarios in three-dimensional space, with a comparative analysis against classical 2D approaches.

3. Research Objectives and Tasks

The objective of this work is to create a simulation environment and develop a modified Proximal Policy Optimization (PPO) model for adaptive route control of a single UAV in 2D and 3D spaces, taking into account distance to the target, collision risk, and energy consumption. To achieve this goal, the following research tasks must be addressed:

– develop 2D and 3D simulation environments with varying obstacle structures and UAV motion parameters;

– formulate a combined PPO reward function that accounts for proximity to the target, collisions, and energy consumption;

– implement and train PPO, DQN, and A2C algorithms in standardized navigation scenarios;

– conduct a comparative analysis of algorithm performance based on key metrics: path length, number of collisions, average reward, and energy consumption;

– Perform a statistical test of the results using the t-test and confidence intervals;

– Analyze the impact of PPO hyperparameters on policy stability, convergence speed, and navigation characteristics in 2D and 3D environments.

Problem Formulation and Environment Model

The simulation model uses the following basic mathematical relationships: The distance between the agent and the target in 2D and 3D environments is defined by the Euclidean metric:

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2}, \quad (1)$$

where the component $(z_1 - z_2)$ is zero. Using a single metric allows for a fair comparison of the neural network's learning performance when transitioning from a 2D to a 3D navigation problem.

The route length L is defined as the sum of the distances between all consecutive points along the trajectory. It accounts for the sum of the distances between all consecutive points that form the motion trajectory [13]. This approach allows for an objective assessment of the efficiency of the constructed route in a three-dimensional environment.

$$L = \sum_{i=1}^{N-1} \sqrt{(x_{i+1} - x_i)^2 + (y_{i+1} - y_i)^2 + (z_{i+1} - z_i)^2}, \quad (2)$$

where L – route length, N – number of points in the trajectory; x_i, y_i, z_i – the agent's coordinates at point i; x_{i+1}, y_{i+1}, z_{i+1} – coordinates at the next point i+1.

The agent's total reward at step t is defined as:

$$G_t = \sum_{k=0}^{T} \gamma^k r_{t+k}, \quad (3)$$

where G_t – the agent's total discounted reward, starting from step t, r_{t+k} – instant rewards at step $t + k$, $\gamma (0 \le \gamma < 1)$ – discount rate, T – episode length.

Environment Model

A simplified energy consumption model of a multirotor UAV was used to evaluate route efficiency. Instantaneous energy consumption E_t (J) at time step t is calculated as the sum of the energy required for hovering and the linear component proportional to the speed of movement:

$$E_t = (P_t + P_{move} |v_t|) \triangle t, \quad (4)$$

where E_t – instantaneous energy consumption at step t, P_t – UAV hovering power, P_{move} – energy consumption coefficient for horizontal movement, v_t – agent velocity, $\triangle t$ – simulation step time.

The agent's state is generally defined as:

$$s_t = (x_t, y_t, z_t, x_{goal}, y_{goal}, z_{goal}), \quad (5)$$

where x_t, y_t, z_t – current UAV coordinates, v_t – current UAV velocity, $x_{goal}, y_{goal}, z_{goal}$ – target point coordinates. For the two-dimensional case, z_t, $= z_{goal} = 0$.

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

Action space:

– 2D: 5 actions (↑ ↓ ← → stay);

– 3D: discrete displacements ($\Delta x, \Delta y, \Delta z \in \{-1, 0, +1\}$.

In the 3D case, the action space is an extension of the 2D model and includes all possible discrete displacements along the three coordinate axes.

Formalization of the navigation problem

In a two-dimensional environment, the calculation of distance and reward is performed similarly to the three-dimensional case (see Formula 1), assuming $z_{agent} = z_{goal} = 0$.

To evaluate the agent's progress, a variable reward is used depending on the change in distance:

$$R_t = a(d_{t-1} - d_t) - \beta I_{collision} + \gamma I_{goal} + \delta E_t, \qquad (6)$$

where d_t is the Euclidean distance from the agent to the target at step t, defined by formula (1), $I_{collision} \in \{0, 1\}$ – obstacle collision indicator, $I_{goal} \in \{0, 1\}$ – goal achievement indicator, E_t – agent's instantaneous energy expenditure at step t determined by formula (4), α,β,γ,δ – weight coefficients of the reward function components.

This approach allows the agent to receive feedback not only regarding the fact of goal achievement but also regarding the quality of individual actions. The positive component $\alpha(d_{t-1} - d_t)$ encourages movement toward the goal, while penalties for collisions and excessive energy expenditure steer the policy toward a safe and efficient trajectory.

A series of experiments established that optimal ratios between reward components are achieved at values of: $\gamma_{goal} = 50, \gamma_{collision} = -50$, which ensures a balance between convergence speed, learning stability, and the agent's ability to avoid obstacles. The reward configuration also includes a small negative penalty for each action, which incentivizes the agent to minimize path length and avoid unnecessary movements. Thus, the complete reward function at each episode step includes a variable reward for progress, penalties for undesirable events, and a fixed reward for reaching the goal, which ensures the learning of coordinated and goal-oriented agent behavior.

In the implemented model, the agent (UAV) makes movement decisions based on the current state, which includes: the agent's coordinates (x, y), velocity, and the goal's position (x_{goal}, y_{goal}). The action space is discrete and defines the possible directions of movement [14].

In a two-dimensional environment, the agent has access to a fixed set of five discrete actions: move up, down, left, right, or stay in place.

In a three-dimensional environment, the action space is described by a displacement vector $a_t = (\Delta , \Delta y , \Delta z)$, where each coordinate can take the values −1, 0, or +1, but a change is allowed along only one axis per step. Thus, the agent has six permissible directions of movement: forward, backward, left, right, up, and down.

After performing the selected action, the agent transitions to a new state and receives an instant reward, which is calculated using the reward function `$R_t$ `, which accounts for changes in distance to the goal, the presence of collisions, and reaching the end point.

Policy updates are performed using the Proximal Policy Optimization (PPO) algorithm, which takes the preference function into account and applies a clipping mechanism to control policy changes. This prevents excessive updates and maintains the stability of the learning process.

4. Materials and Methods

To update the policy in both environments (2D and 3D), the Proximal Policy Optimization (PPO) algorithm [14] was used, which combines the efficiency of gradient methods with a mechanism for stabilizing updates. The main idea behind PPO is to limit the magnitude of policy updates during each iteration, preventing abrupt changes that could cause learning divergence.

Table 1. *Main PPO training configuration parameters*

Parameter	Value
Number of iterations	4
Frames per second (FPS)	1979 (frames per second)
Total number of steps	65,536
Algorithm	PPO (Proximal Policy Optimization)
Main hyperparameters	clip_range: 0.2; learning_rate: 0.0003; batch_size: 64

One of the components of the PPO algorithm is the critic's value loss function, defined as:

$$L_{value}(\theta) = E_t[(V_\theta(s_t) - V_t^{target})^2], \qquad (7)$$

where $V_t^{target} = r_t + \gamma V(s_{t+1})$ is the target estimate for the value function.

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

The primary objective function of PPO for policy updating is the clipped policy objective:

$$L_{clip}(\theta) = E_t[\min(r_t(\theta)A_t, clip(r_t(\theta), 1-\varepsilon, 1+\varepsilon)A_t)], \quad (8)$$

where $r_t(\theta)$ is the ratio of action probabilities under the new and old policies, A_t is the preference score, and ε is the clip range parameter.

To maintain sufficient policy diversity during training, an entropy bonus is introduced:

$$L_{entropy}(\theta) = \beta_{ent} E_t[H(\pi_\theta(a_t \mid s_t))], \quad (9)$$

where $H(*)$ is the policy entropy, β_{ent} is the coefficient (0.01 is chosen in this work).

All components are combined into a final function:

$$L_{total}(\theta) = L_{clip}(\theta) + c_1 L_{value}(\theta) - c_2 L_{entropy}(\theta), \quad (10)$$

where c_1 = 0.5 and c_2 = 0.01 are the weight coefficients recommended by the authors of PPO and confirmed experimentally.

In all experiments, the clip range parameter ε is set to 0.2, which corresponds to standard recommendations for the PPO algorithm.

This formalization ensures robust policy updates and stable model convergence in both 2D and 3D environments.

5. Research Results

5.1 Simulation Scenarios (2D and 3D)

1. Scenarios: To verify the effectiveness of the PPO algorithm, two experimental simulation scenarios with different spatial complexities were implemented:

Scenario 1 – 2D environment:

– grid size: 20×20;

– number of static obstacles: 30;

– agent start coordinates: (0, 0);

– target coordinates: (19, 19);

– the environment contains randomly placed fixed obstacles;

– goal: teach the agent to minimize the number of steps and avoid collisions.

Scenario 2 – 3D environment:

– cubic space dimensions: 15×15×8;

– obstacles: 50 (including vertical objects);

– starting position: (0, 0, 0);

– target position: (14, 14, 7);

– Objective: to test the stability of PPO as the action space and environment complexity increase.

In each scenario, 5,000 simulation episodes were run. During each run, the following metrics were collected: average reward, episode length, number of collisions, and average time to reach the goal.

5.2 Metrics

1. The average path length is defined as the average value of L, calculated using formula (2) for all simulation episodes.

2. Number of collisions.

3. Average reward per episode.

Visualization allows for a better analysis of the agent's spatial behavior and its interaction with the environment. The figures show:

– The drone's location – red marker;

– Targets – green dots;

– Obstacles – white blocks with an "x" inside.

Figure 1 demonstrates the drone's trajectory in 2D space, emphasizing its ability to navigate around obstacles and reach targets.

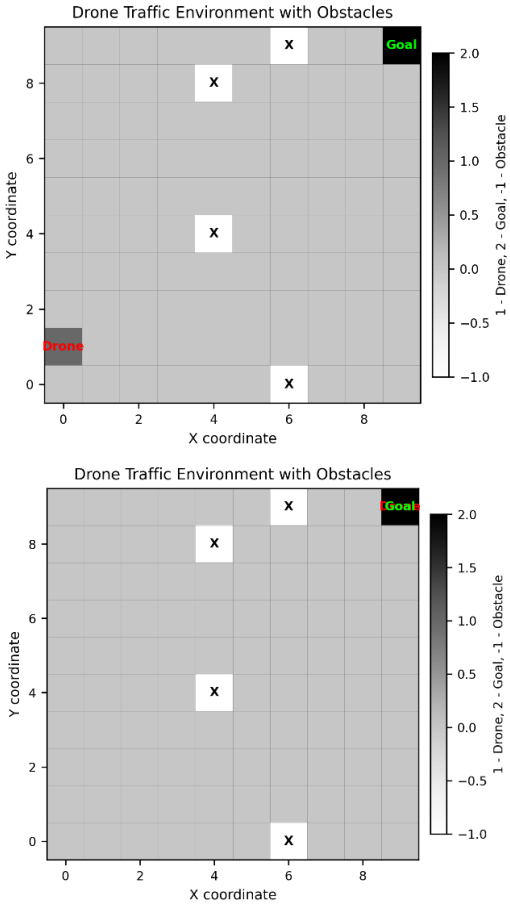

Fig. 1. Drone motion dynamics

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

The visualizations demonstrate the agent's evasion effectiveness and adaptability.

5.3 Results in a 2D environment

Training of the UAV adaptive route control model in a two-dimensional (2D) environment was performed using the Proximal Policy Optimization (PPO) algorithm. The goal was to develop a stable policy capable of effectively avoiding obstacles and reaching targets in an environment with dynamic constraints. The main training configuration parameters are shown in Table 2.

Table 2. *Comparative average metrics for algorithms in a 2D environment (100 episodes*)

Algorithm	Energy_Mean (J)	Mean_Reward	Collisions_Mean (count)	PathLength_Mean (m)
PPO_Custom	4.05	–18.89	4.21	4.63
PPO_base	1.71	–250.32	80.63	80.64
DQN	2.53	–253.64	81.57	81.57
A2C	5.28	–262.55	85.49	85.50

The Impact of the Learning Rate and Batch Size Hyperparameters on PPO Performance

During the training of models using the PPO algorithm, hyperparameters, particularly the learning rate and batch size, have a significant impact on the stability, speed, and quality of policy convergence [15].

A high learning rate (e.g., 0.01–0.1) implies rapid policy updates, which can be effective in simple environments.

A low value (0.0001–0.001) ensures stability but slows down adaptation. Similarly, a large batch size (256–1024) provides smoothness, while a small one (32–64) offers flexibility, though the risk of instability increases;

Thus, the coordinated tuning of PPO hyperparameters, particularly learning_rate and batch_size, plays a critical role in achieving stable and effective agent training. This is especially important when scaling from 2D to 3D environments with increased complexity in dynamics and spatial constraints.

Comparative Analysis Results (2D)

Calculation of percentage changes (PPO_custom vs PPO_base)

1) Energy Consumption

$$\Delta energy = \frac{4.05 - 1.71}{1.71} \times 100\% \approx 136\%$$

PPO_custom uses 136% more energy than the base PPO.

2) Mean Reward Improvement

$$\Delta_{reward} = \frac{-18.89 - (-250.32)}{|-250.32|} \times 100\% \approx 92.5\%$$

The average reward increased by 92.5%.

3) Collision Reduction

$$\Delta_{collisions} = \frac{80.63 - 4.21}{80.63} \times 100\% \approx 94.8\%$$

The number of collisions decreased by 94.8%.

4) Path Length Reduction

$$\Delta_{path} = \frac{80.64 - 4.63}{80.64} \times 100\% \approx 94.3\%$$

Path length decreased by 94.3%

Results of the 2D scenario

In a 2D environment, the modified PPO algorithm (PPO_custom) demonstrated the following improvements compared to the base version of PPO:

1. **+92.5%** increase in average reward (–18.9 vs. –250.3);
2. **–94.8%** reduction in the number of collisions (4.2 vs. 80.6);
3. **–94.3%** reduction in path length (4.6 vs. 80.6).

However, there is **a 136% increase in energy consumption** (4.05 vs. 1.71), which requires further analysis and possible balancing of the reward environment.

5.4 Results in a 3D environment

The key differences between the reinforcement learning algorithm for 3D space and the base version are listed below.

The simulation environment models the drone's navigation process in three-dimensional space, taking into account obstacles and target coordinates. The main components of the model:

– state space – defined by formula (5);

– action space: $a = (\Delta x, \Delta y, \Delta z)$ – discrete displacements in three directions;

– environment: contains randomly generated obstacles that the drone must avoid.

Task Formalization and Key Problem

In three-dimensional space, the agent's progress is evaluated by the change in the Euclidean distance to the target, which is determined by formula (1).

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

This metric is used as the primary indicator of how close the agent is to the target in an environment with three degrees of freedom.

The reward function in a 3D environment retains the same structure as in 2D (Formula 8), where,d_t – current distance to the goal,$I_{collision} \in \{0,1\}$ – collision indicator, $I_{goal} \in \{0,1\}$ – goal attainment indicator, E_t – instantaneous energy expenditure of the agent at step t, α,β,γ,δ – weight coefficients.

This formalization allows for a unified mechanism to evaluate progress in space, penalize dangerous maneuvers, and account for energy constraints. The 3D model architecture includes an extended space of states and actions:

– **State**: (x ,y ,z ,v ,x_{goal},y_{goal},z_{goal})

– **Action**: (Δx , Δy , Δz)

– **Objective**: Minimize the distance to the target while ensuring safety and energy efficiency.

Visualization

This section presents the results of a comparison of PPO performance in 2D and 3D environments.

5.5 Statistical Analysis of Performance

To confirm the model's training effectiveness under varying levels of complexity, a statistical analysis of key metrics was conducted: average reward , collision frequency, and episode duration. Comparisons between 2D and 3D environments were performed using mean values, standard deviation, confidence intervals, and the t-test. The results are visualized in Figures 2 and 3.

Figure 2 presents a comparison of reward distributions across environments. Figure 3 illustrates a training log file with a detailed representation of metric changes over time.

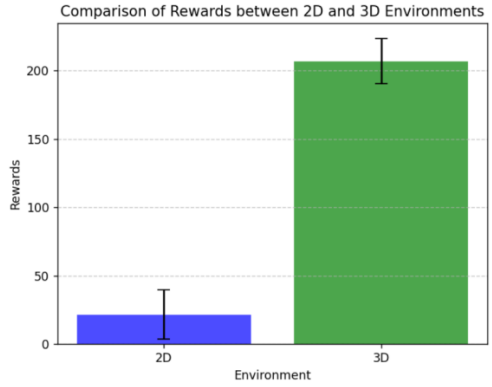

Fig. 2. Comparison of rewards between 2D and 3D spaces

```
Rewards Analysis:
2D - Mean: 21.40, Std: 29.03, 95% CI: (3.40, 39.40)
3D - Mean: 207.00, Std: 26.72, 95% CI: (190.44, 223.56)
T-statistic: -14.87, P-value: 0.0000
```

Fig. 3. Dynamics of reward changes during training depending on the environment type

Analyzing Figure 3, the following conclusions can be drawn:

1. **The average reward value** indicates a noticeable advantage of the 3D environment: in 2D, the model reached 21.40, while in 3D, it reached 207.0. This indicates significantly better policy adaptation in more complex conditions.

2. **The standard deviation** also demonstrates differences: in 2D – 29.03, in 3D – 26.72. The lower variance in 3D indicates more stable agent behavior.

3. **Confidence intervals (95%)** confirm the previous results: a wide interval in 2D (3.40–39.40) indicates high variability, while in 3D (190.44–223.56) it indicates greater statistical confidence.

4. The results of the t-test ($t = -14.87$, $p < 0.001$) indicate a statistically significant difference between the model's performance metrics in 2D and 3D environments, confirming the model's superiority in the more complex three-dimensional space with a significance level of $p < 0.05$.

The obtained data confirm the higher performance of PPO in a 3D environment, which is due to better adaptation to spatial complexity and flexibility of the navigation policy.

Thus, the PPO algorithm demonstrated not only a steady improvement in metrics during training but also a statistically confirmed advantage in the more complex three-dimensional environment. This demonstrates the potential of PPO for real-world air navigation scenarios under dynamic conditions.

Key comparison results

The model in the 2D environment demonstrated faster policy convergence, shorter episodes, and a higher average reward in the early stages of training. The smaller number of degrees of freedom ensured more stable collision dynamics and simpler trajectory behavior.

In the three-dimensional environment, during the initial training stages, the PPO algorithm was characterized by an increased number of collisions and longer episode durations, which is due to the increased

ISSN 2522-9818 (print)
ISSN 2524-2296 (online)
Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

spatial complexity of the task. At the same time, a gradual policy adaptation and a reduction in the number of critical errors were observed during training.

Graphical Comparison

For a more visual analysis, the following metrics were presented in the graphs:

– **reward dynamics**: the graph shows that average rewards in 2D space increase faster compared to 3D;

– **collision frequency**: in 3D space, the collision frequency is higher in the early stages, but the model adapts over time;

– **episode length**: episodes in 3D space are typically longer, which is explained by the more complex trajectory.

Figure 4 illustrates the movement of drones in 2D space. The green dot represents the target's location, the red dots indicate where the drone reached the target, and the black blocks show the locations of obstacles. The graph demonstrates how effectively the algorithm avoids obstacles and reaches targets in a 2D environment.

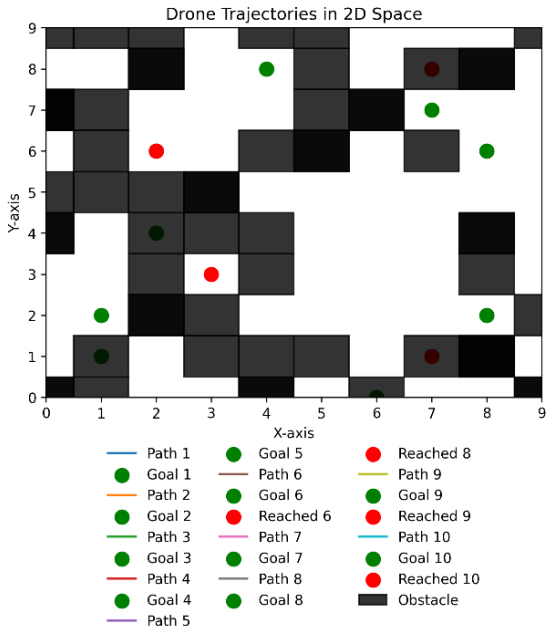

Fig. 4. Drone trajectories in 2D space

Figure 5 demonstrates movement in 3D space. Green dots mark the locations of targets, the red dot indicates where one of the drones reached the target, and the black dots represent obstacles in the space. The additional dimension complicates route planning, but the model demonstrates high adaptability in avoiding obstacles.

The generalized learning dynamics of the model in 2D and 3D environments are presented in Figures 6 and 7.

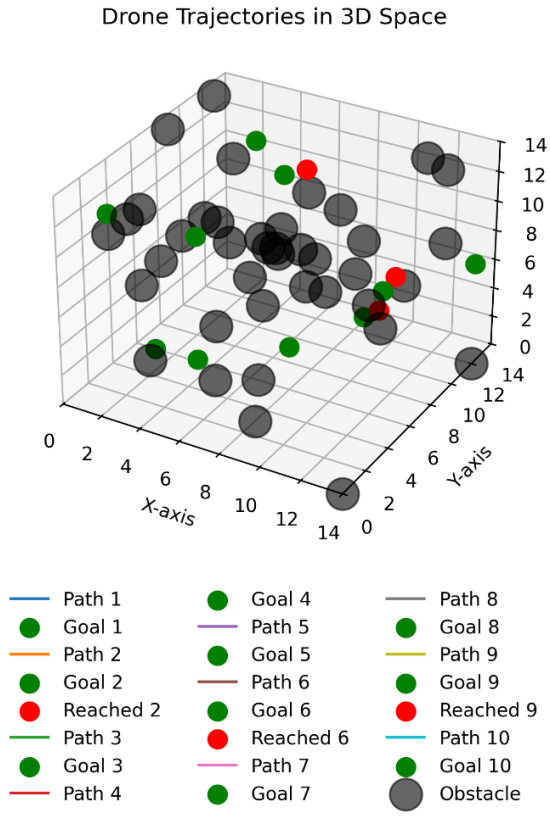

Fig. 5. Drone trajectories in 3D space

Figure 6 shows three main graphs for the 2D environment:

– the increase in the average reward per episode indicates the model's rapid adaptation to a simple topology;

– the decrease in the number of collisions confirms the effectiveness of the developed evasion policy;

– the stable duration of episodes reflects the stability of the agent's behavior.

Figure 7 shows similar graphs for the 3D environment. Here, a slower increase in rewards can be observed, which is due to a more complex trajectory structure and a greater number of obstacles. However, even under these conditions, the algorithm gradually reduces the number of collisions and stabilizes the episode length, indicating successful adaptation to the three-dimensional environment. Such visualization allows for a comprehensive assessment of learning effectiveness in both spaces and confirms the advantage of PPO in solving navigation problems of varying complexity.

Despite its complexity, the 3D approach offers broader practical applicability in real-world conditions, such as urban environments with tall buildings

Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35) *ISSN 2522-9818 (print)*
Innovative technologies and scientific solutions for industries. 2026. No. 1 (35) *ISSN 2524-2296 (online)*

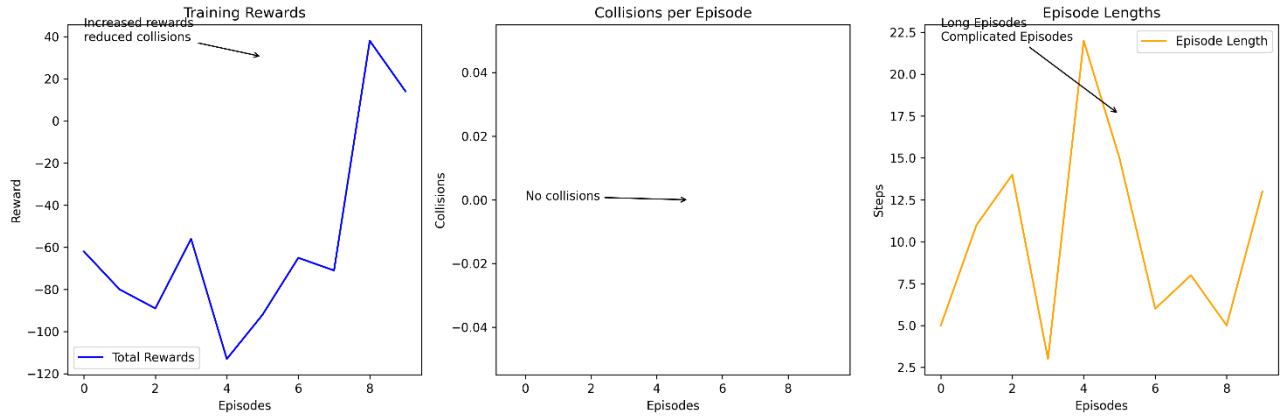

Fig. 6. Training dynamics for 2D space

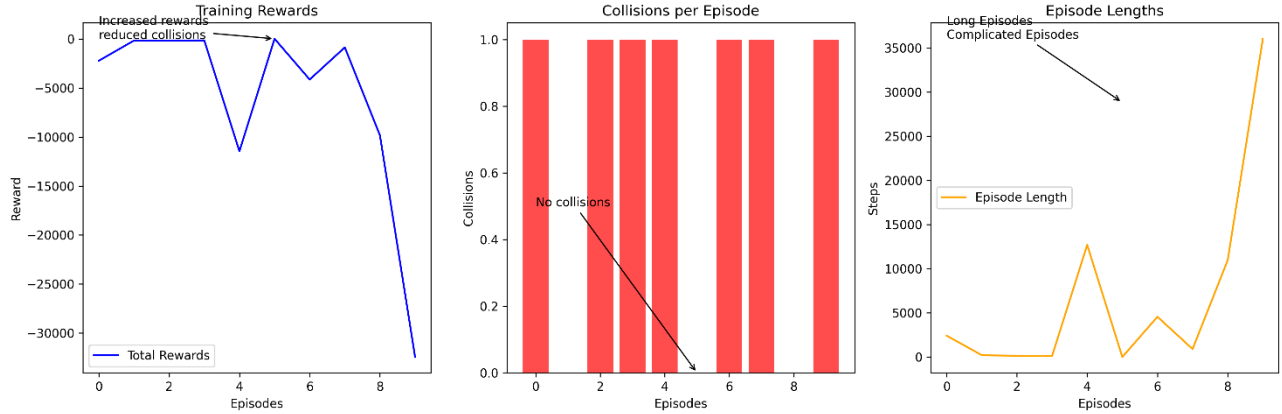

Fig. 7. Learning dynamics for 3D space

6. Discussion of Results

The results of the study confirm the effectiveness of PPO in adaptive routing tasks in both 2D and 3D environments. In 2D, the model achieved rapid convergence due to the lower complexity of the environment, whereas in 3D, greater variability was observed in the early stages. At the same time, as training progressed, the agent demonstrated stable behavior with a gradual decrease in the number of collisions. The modified reward function facilitated effective learning in both cases. A statistical t-test revealed a significant difference between the results, justifying the need to adapt strategies to each type of environment.

For the practical application of the proposed adaptive UAV route control algorithm, it is most appropriate to integrate it into existing Unmanned Traffic Management (UTM) concepts. Specifically:

– **NASA UTM (Unmanned Aircraft System Traffic Management)** aims to create an environment where multiple drones operate Beyond Visual Line of Sight (BVLOS) at altitudes up to 400 feet, coordinated through a suite of data exchange and route planning services. The proposed PPO algorithm with an energy-based reward component can be implemented as a microservice within the Flight Information Management System (FIMS) module to dynamically update routes while accounting for obstacles and limited resources.

– **EU U-Space** regulates a set of automated services (Service Levels 1–4) for the safe access of large drone fleets to airspace. In particular, at the **U3** level **(High-density UTM)**, the algorithm can provide adaptive real-time flight rerouting, reducing the risk of collisions and optimizing battery usage during peak load.

Thus, integrating the proposed modified PPO reward function into these UTM modules will allow for:

1. Ensure **proactive avoidance of obstacles** and airspace restrictions.

2. Dynamically **recalculate routes** based on the current battery charge level and remaining distance to the destination.

3. Optimize **cooperation with other participants** in the UTM ecosystem through standardized exchange protocols (REST, MQTT, AMQP).

This approach contributes to improving the reliability and efficiency of UAS operations in real-world UTM projects.

ISSN 2522-9818 (print) Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)
ISSN 2524-2296 (online) Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)

Limitations of the study. Agent training was conducted over 50,000 steps, which proved sufficient to demonstrate the convergence of the modified algorithm under the selected scenario conditions. However, we acknowledge that this may not be sufficient for more complex real-time navigation tasks (real-world applications). The results obtained should be considered a proof of concept, and further research will involve scaling up training to 10^6 epochs to improve the robustness of the policy.

7. Conclusions

Scientific novelty of the research. The scientific novelty of this work lies in the development of a combined reward function for the PPO algorithm, which simultaneously accounts for the dynamics of approach to the target, collision risk, and UAV energy consumption.

Unlike existing approaches, the proposed model provides a controlled trade-off between navigation safety and energy efficiency.

In addition, a unified 2D/3D simulation environment has been developed, allowing for a fair comparison of reinforcement learning algorithms in spaces of different dimensions. A comparative analysis and statistical verification of the results confirm the effectiveness of the approach and its advantages in three-dimensional navigation scenarios.

Main results of the study. In the course of this study, a simulation environment for UAV navigation in two- and three-dimensional spaces was developed, which takes into account the presence of static obstacles, target coordinates, and energy constraints on movement.

Based on this environment, a modified Proximal Policy Optimization (PPO) model with a combined reward function was implemented, combining incentives for approaching the target, penalties for collisions, and energy costs. A comparative analysis of the effectiveness of the PPO, DQN, and A2C algorithms in standardized 2D and 3D navigation scenarios demonstrated significant advantages of the proposed PPO model. Specifically, in the 2D environment, the number of collisions was reduced by 94.8%, and the route length by 94.3% compared to the baseline PPO version, while the average reward increased by 92.5%.

In three-dimensional space, similar trends toward improved safety and policy stability were confirmed, as evidenced by statistically significant results ($p < 0.05$).

Additionally, a relationship was established between the aggressiveness of obstacle-avoidance strategies and energy consumption, allowing the agent's behavior to be interpreted as a trade-off between navigation safety and energy efficiency. The proposed visualization tools for trajectories and learning dynamics serve as a means of interpreting the agent's behavior and further confirm the convergence stability of the PPO algorithm in environments with varying spatial complexity.

Limitations of the study and prospects for further work. The main limitations of the study are the computational complexity of the reinforcement learning process, which requires significant resources as the environment's dimensionality increases, as well as the limited scalability of the proposed approach to multi-agent scenarios under real-world operating conditions. The results obtained should be considered within the framework of a simulation model with a single agent and static obstacles.

Further research should focus on expanding the model by integrating realistic airspace constraints, including meteorological factors, restricted zones, and variable traffic density, as well as on studying multi-agent interaction using multi-agent reinforcement learning approaches.

A promising direction is the automatic tuning of hyperparameters during training and the combination of the PPO algorithm with evolutionary methods or simulation-based learning methods to improve the stability and generalization ability of the policy.

Conflict of interest

The authors declare that they have no conflicts of interest, including financial, personal, authorial, or any other nature, that could influence the research or the results published in this article.

Funding

The study was conducted without financial support.

Data availability

Data will be provided upon reasonable request.

Use of artificial intelligence

The authors confirm that they did not use artificial intelligence technology in the creation of this paper.

References

1. Debnath, D., Vanegas, F., Sandino, J., Hawary, A. F., Gonzalez, F. (2024), "A review of UAV path-planning algorithms and obstacle avoidance methods for remote sensing applications", *Remote Sensing*, Vol. 16 (21), 4019 p. DOI: https://doi.org/10.3390/rs16214019
2. Martins, F. G., Coelho, M. A. N. (2000), "Application of feedforward artificial neural networks to improve process control of PID-based control algorithms", *Computers & Chemical Engineering*, Vol. 24 (2-7). pp. 853-858. DOI: https://doi.org/10.1016/S0098-1354(00)00339-2
3. Liu, X., Peng, Z.R., Zhang, L.Y. (2019), "Real-time UAV rerouting for traffic monitoring with decomposition based multi-objective optimization", *Journal of Intelligent & Robotic Systems*, Vol. 94, pp. 491–501. DOI: https://doi.org/10.1007/s10846-018-0806-8
4. Almeida, E. N., Campos, R., Ricardo, M. (2022), "Traffic-aware UAV placement using a generalizable deep reinforcement learning methodology", *2022 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6. DOI: https://doi.org/10.48550/arXiv.2203.08924
5. Madani, A., Engelbrecht, A. Ombuki-Berman, B., (2023), "Cooperative coevolutionary multi-guide particle swarm optimization algorithm for large-scale multi-objective optimization problems", *Swarm and Evolutionary Computation*. Vol. 82. 101262 p. DOI: https://doi.org/10.1016/j.swevo.2023.101262
6. Luo, J., Tian, Y., Wang, Z. (2024), "Research on unmanned aerial vehicle path planning". *Drones*, Vol. 8(2). 51 p. DOI: https://doi.org/10.3390/drones8020051
7. Li, C., Lian, J., (2007), "The Application of Immune Genetic Algorithm in PID Parameter Optimization for Level Control System", *Proceedings of the 2007 IEEE International Conference on Automation and Logistics (ICAL)*, Jinan, China, pp. 2670–2674. DOI: https://doi.org/10.1109/ICAL.2007.4338670
8. Yang, F., Lu, Q., Li, R., Xu, Y., Yuan, W., Wu, X. (2023), "Real-time optimal path planning and fast autonomous flight for UAV in unknown environments", *IEEE*. DOI: https://doi.org/10.23919/CCC58697.2023.10240971
9. Li, Q., Li, R., Ji, K., Dai, W. (2015), "Kalman filter and its application", *IEEE*. DOI: https://doi.org/10.1109/ICINIS.2015.35
10. Hooshyar, M., Huang, Y. (2023), "Meta-heuristic algorithms in UAV path planning optimization: A systematic review (2018–2022)", *Drones*, Vol. 7(12), 687 p. DOI: https://doi.org/10.3390/drones7120687
11. Li, H., Zhang, Z.-yu. (2012), "The application of immune genetic algorithm in main steam temperature of PID control of BP network", *Physics Procedia*, Vol. 25, pp. 80-86. DOI: https://doi.org/10.1016/j.phpro.2012.02.013
12. Zhang, M., Liu, Y., Wang, Y., Li, F., Chen, L. (2023), "Real-time path planning algorithms for autonomous UAV", *IEEE*. DOI: https://doi.org/10.1109/CAC57257.2022.10054770
13. Kim, D. H. (2003), "Comparison of PID controller tuning of power plant using immune and genetic algorithms", *The 3rd International Workshop on Scientific Use of Submarine Cables and Related Technologies*, Lugano, Switzerland, pp. 358-363. DOI: https://doi.org/10.1109/CIMSA.2003.1227222
14. Yena, M. (2024), "Optimizing air traffic control: Innovative approaches to collision avoidance in UAV operations", *Integrated Computer Technologies in Mechanical Engineering - 2023 (ICTM 2023)*. pp. 543–553. DOI: https://doi.org/10.1007/978-3-031-60549-9_41
15. Yena, M., & Pohudina, O. (2025), "Integrated simulation model of swarm control and adaptive routeing of UAVS in a changing air environment", *Innovative technologies and scientific solutions for industries*, (4(34), pp. 32-43. DOI: https://doi.org/10.30837/2522-9818.2025.4.032

Received (Надійшла) 23.11.2025
Accepted for publication (Прийнята до друку) 10.02.2026
Publication date (Дата публікації) 30.03.2026

Відомості про авторів / About the Authors

Єна Максим Вікторович – аспірант, Національний аерокосмічний університет «Харківський авіаційний інститут», кафедра «Інформаційних технологій проєктування», Харків, Україна;
Maksym Yena – PhD Student, "Kharkiv Aviation Institute", Department of Information Technology Design, Kharkiv, Ukraine;
e-mail: yenamaxim98@gmail.com
ORCID ID: https://orcid.org/0009-0006-0664-3244
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=59161031800

ISSN 2522-9818 (print) *Сучасний стан наукових досліджень та технологій в промисловості. 2026. № 1 (35)*
ISSN 2524-2296 (online) *Innovative technologies and scientific solutions for industries. 2026. No. 1 (35)*

Погудіна Ольга Костянтинівна – кандидат технічних наук, доцент, Національний аерокосмічний університет «Харківський авіаційний інститут», доцент кафедри інформаційних технологій проєктування, Харків, Україна;
Olha Pohudina – Candidate of Technical Sciences, Associate Professor, National Aerospace University "Kharkiv Aviation Institute", Department of Information Technology Design, Kharkiv, Ukraine;
e-mail: olha.pohudina@poliba.it
ORCID ID: https://orcid.org/0000-0001-5689-2552
Scopus Author ID: https://www.scopus.com/authid/detail.uri?authorId=57204907264

МОДЕЛЮВАННЯ АДАПТИВНОГО УПРАВЛІННЯ МАРШРУТАМИ БПЛА НА ОСНОВІ АЛГОРИТМІВ НАВЧАННЯ З ПІДКРІПЛЕННЯМ

Предмет дослідження – функція винагороди, політика дій та динаміка навчання алгоритму PPO у задачі адаптивної навігації БпЛА в умовах динамічного повітряного простору та обмежених енергетичних ресурсів. **Мета** – створення симуляційного середовища та розроблення модифікованої моделі Proximal Policy Optimization (PPO) для адаптивного управління маршрутом одиночного БпЛА у 2D та 3D просторах із урахуванням відстані до цілі, ризику зіткнень і енергоспоживання. **Завдання:** розробити 2D та 3D симуляційні середовища з різною структурою перешкод і параметрами руху БПЛА; сформувати комбіновану функцію винагороди PPO, що враховує відстань до цілі, зіткнення та енергоспоживання; Реалізувати та навчити алгоритми PPO, DQN і A2C у стандартизованих сценаріях навігації.; провести порівняльний аналіз ефективності алгоритмів за ключовими метриками (довжина маршруту, кількість зіткнень, винагорода, енергоспоживання; виконати статистичну перевірку результатів за допомогою t-тесту та довірчих інтервалів; проаналізувати вплив гіперпараметрів PPO на стабільність політики та збіжність навчання у 2D і 3D середовищах. **Методи:** використано алгоритми глибинного навчання з підкріпленням (PPO, DQN, A2C). Розроблено дві симуляційні моделі (2D та 3D) із випадковими статичними перешкодами. Сформовано комбіновану функцію винагороди, що включає динамічну компоненту зближення до цілі, штрафи за зіткнення та енергетичний термін. Ефективність моделей оцінювалася за середньою винагородою, довжиною маршруту, кількістю зіткнень та енергетичними витратами. Статистичну достовірність перевірено за допомогою t-тесту та 95% довірчих інтервалів. **Результати:** модифікована PPO-модель у 2D середовищі зменшила кількість зіткнень на 94,8% та довжину маршруту на 94,3% у порівнянні з базовою PPO, при цьому спостерігалося збільшення енергоспоживання через складніші маневри ухилення. У 3D середовищі підтверджено аналогічні тенденції: підвищення безпеки навігації, стабілізація політики та статистично значущі покращення ключових метрик ($p < 0,05$). **Висновки:** розроблено уніфіковане 2D/3D симуляційне середовище адаптивної маршрутизації БпЛА та модифіковану PPO-модель з комбінованою функцією винагороди, що враховує зближення до цілі, зіткнення та енергоспоживання. У 2D-середовищі досягнуто зменшення кількості зіткнень на ≈94,8%, скорочення довжини маршруту на ≈94,3% та зростання середньої винагороди на ≈92,5% порівняно з базовою PPO. У 3D-середовищі підтверджено аналогічні тенденції та статистично значущі покращення ($p < 0,05$). Встановлено залежність між агресивністю ухилення та енергоспоживанням, що дозволяє вибирати оптимальну політику для сценаріїв BVLOS.

Ключові слова: адаптивне управління; Proximal Policy Optimization; навчання з підкріпленням; імітаційне моделювання; маршрутизація; 3D-навігація.

Бібліографічні описи / Bibliographic descriptions

Єна М.В., Погудіна О.К. Моделювання адаптивного управління маршрутами БпЛА на основі алгоритмів навчання з підкріпленням. *Сучасний стан наукових досліджень та технологій в промисловості*. 2026. № 1 (35). С. 28–38. DOI: https://doi.org/10.30837/2522-9818.2026.1.028

Yena, M., Pohudina, O. (2026), "Modeling of adaptive UAV route control based on reinforcement learning algorithms", *Innovative Technologies and Scientific Solutions for Industries*, No. 1 (35), P. 28–38. DOI: https://doi.org/10.30837/2522-9818.2026.1.028