УДК 004.85:629.735 DOI 10.36910/775.24153966.2025.84.39

В.Л. Закревський, О.С. Приходько
Луцький національний технічний університет

РОЗРОБКА АДАПТИВНОГО МЕТОДУ КЕРУВАННЯ БПЛА НА ОСНОВІ ГЛИБОКОГО НАВЧАННЯ З ПІДКРІПЛЕННЯМ ТА ДОМЕННОЇ РАНДОМІЗАЦІЇ

У роботі розглядається проблема створення адаптивної системи керування безпілотним літальним апаратом (БПЛА) для роботи в умовах стохастичної невизначеності. Запропоновано підхід на основі алгоритму Proximal Policy Optimization (PPO) із застосуванням стратегії доменної рандомізації параметрів середовища. Розроблено структуру нейронної мережі та функцію винагороди, що забезпечують баланс між точністю навігації та енергоефективністю. Розроблений контролер продемонстрував здатність адаптуватися до змін маси корисного навантаження (до 20%) та зовнішніх вітрових збурень без необхідності ручного переналаштування коефіцієнтів. Результати порівняльного моделювання підтверджують перевагу запропонованого методу над класичним ПІД-регулятором: час стабілізації зменшено в 1.5 рази, а максимальне відхилення при збуреннях – на 40%.

Ключові слова: *БПЛА, навчання з підкріпленням, PPO, адаптивний контролер, доменна рандомізація, динаміка польоту*

V.L. Zakrevskyi, O.S. Prykhodko

DEVELOPMENT OF AN ADAPTIVE UAV CONTROL METHOD BASED ON DEEP REINFORCEMENT LEARNING AND DOMAIN RANDOMIZATION

The paper addresses the problem of creating an adaptive control system for an unmanned aerial vehicle (UAV) to operate under conditions of stochastic uncertainty. An approach based on the Proximal Policy Optimization (PPO) algorithm using a domain randomization strategy for environmental parameters is proposed. A neural network structure and a reward function have been developed to ensure a balance between navigation accuracy and energy efficiency. The developed controller demonstrated the ability to adapt to payload mass changes (up to 20%) and external wind disturbances without the need for manual coefficient retuning. Comparative simulation results confirm the advantage of the proposed method over a classical PID controller: stabilization time is reduced by 1.5 times, and maximum deviation during disturbances is reduced by 40%.

Keywords**:** *UAV, reinforcement learning, PPO, adaptive control, domain randomization, flight dynamics*

Introduction and problem statement. The rapid development of unmanned aircraft systems requires the improvement of control algorithms capable of ensuring reliable vehicle operation in complex conditions. Modern UAVs are used for cargo delivery, infrastructure monitoring, and rescue operations, where accuracy and reliability are critical. Classical automatic control methods, such as PID controllers or Linear Quadratic Regulators (LQR), demonstrate high effectiveness in deterministic environments with known and invariant model parameters. However, under real operating conditions, UAVs face unpredictable factors: sudden wind gusts, aerodynamic effects near obstacles, mass changes during cargo drop or lifting, and propulsion system degradation. Traditional linear regulators, tuned to a nominal operating point, often prove unable to compensate for such significant nonlinear disturbances without a complex adaptive overlay or real-time system identification, which requires significant computational resources. An urgent scientific task is the development of control methods that combine flight task execution accuracy with a high level of adaptability (robustness) by utilizing the potential of modern machine learning methods, particularly Reinforcement Learning (RL), which allows an agent to independently form an optimal control strategy based on experience interacting with the environment.

Analysis of the latest research and publications. The current state of research in the field of intelligent UAV control is characterized by a variety of Reinforcement Learning (RL) integration architectures, driven by high dynamics nonlinearity, the presence of external disturbances, and limited vehicle energy resources [1-3].

Paper [4] proposes a direct neural network control architecture, where the agent generates control signals for the motors, while stability against external influences, such as wind, is ensured by an additional state observer. Similar approaches are actively researched in the context of applying model-free RL algorithms for continuous state and action spaces, particularly PPO and SAC, which demonstrate high training stability [1]. A drawback of this class of methods remains the dependence on disturbance estimation accuracy and the complexity of transferring trained policies from simulation to the real environment.

An alternative approach, known as "residual control," is discussed in study [5]. The authors propose supplementing the classical control loop with an RL agent that compensates for unmodeled dynamics. Such integration allows combining RL adaptability with the guaranteed stability of classical control methods, which is relevant for tasks with strict constraints on power consumption and computational resources [2]. At the same time, the base controller may limit the system's maneuverability.

Study [6] focuses on universality by using a dynamics encoder based on deep neural networks. The application of scale-aware domain randomization allows a single policy to control UAVs of different configurations, which aligns with modern approaches to increasing the generalization capability of RL controllers and reducing the Sim-to-Real gap [7]. This method opens the way for creating controllers that are less dependent on a specific hardware platform.

Hybrid approaches described in work [8] propose a mechanism for mixing control signals from a classical regulator and an RL policy. Training with domain randomization ensures adaptability, while the classical component guarantees basic stability, combining the advantages of both paradigms, as confirmed by the successful application of RL in UAV energy efficiency optimization and safe navigation tasks [1, 7].

The aim of the study. The aim of the study is the development, software implementation, and experimental verification of an adaptive quadcopter attitude control algorithm based on the deep reinforcement learning method PPO. The main tasks include synthesizing the neural network architecture, developing a reward function for flight stabilization, and ensuring the controller's robustness to changes in the vehicle's inertial characteristics and external wind disturbances.

Experimental methodology. To achieve the set goal, the mathematical apparatus of Markov Decision Processes (MDP) was used. The control task is formalized as a tuple (S, A, P, R, γ). The state space *S* is continuous and includes 15 components: position vector $p \in \mathbb{R}^3$, linear velocity vector $v \in \mathbb{R}^3$, Euler angles (roll, pitch, yaw) $\alpha \in \mathbb{R}^3$, angular velocities $\omega \in \mathbb{R}^3$, and the relative target position vector $p_{target} - p$. The action space *A* consists of four continuous values $u \in [0,1]^4$ corresponding to the normalized thrust of each of the four quadcopter rotors.

Proximal Policy Optimization (PPO) was chosen as the main training algorithm. This algorithm belongs to the Actor-Critic family of methods and utilizes a policy update step constraint (clipping), which prevents training destabilization during sharp changes in neural network weights. Training was conducted in the PyBullet physical simulator, which ensures high-accuracy modeling of rigid body dynamics.

A key methodological element for ensuring adaptability is the domain randomization strategy. During training, simulation environment parameters were varied at the beginning of each episode according to a normal distribution:

- UAV mass (m): $m \sim \mathcal{N}(m_{nom}, 0.2m_{nom})$;
- Inertia tensor (I): $I \sim \mathcal{N}(I_{nom}, 0.1I_{nom})$;
- External disturbances: random wind force vectors F_{wind} with an intensity of up to 5 N.

Presentation of the main research material. The architecture of the developed system includes two neural networks: the Actor (policy) network and the Critic (value) network. Both networks are Multilayer Perceptrons (MLP) with two hidden layers of 256 neurons each. *Tanh* was used as the activation function, providing nonlinearity and bounded output signals.

A critical development stage was the construction of the reward function, which defines the agent's behavior. It is designed as a weighted sum of components:

$$R_t = R_{pos} + R_{stab} + R_{act} + R_{term},$$

where:

$R_{pos} = -w_d \| p_t - p_{target} \|$ – penalty for deviation from the target;

$R_{stab} = -w_a (\| \omega_t \|^2 + \| \alpha_{roll,pitch} \|^2)$ – penalty for instability (large tilt angles and angular velocities), which prevents oscillations;

$R_{act} = -w_u \| u_{t-1} - u_t \|^2$ – penalty for sharp changes in control signals, promoting control smoothness and reducing motor wear;

R_{term} – a large penalty for crashing or moving outside the operating zone.

Training was conducted over 2 million simulation steps. PPO algorithm hyperparameters: learning rate = $3 \cdot 10^{-4}$, $\gamma = 0.99$, clip range = 0.2. Three characteristic phases are observed in the learning curve graph (Fig. 1).

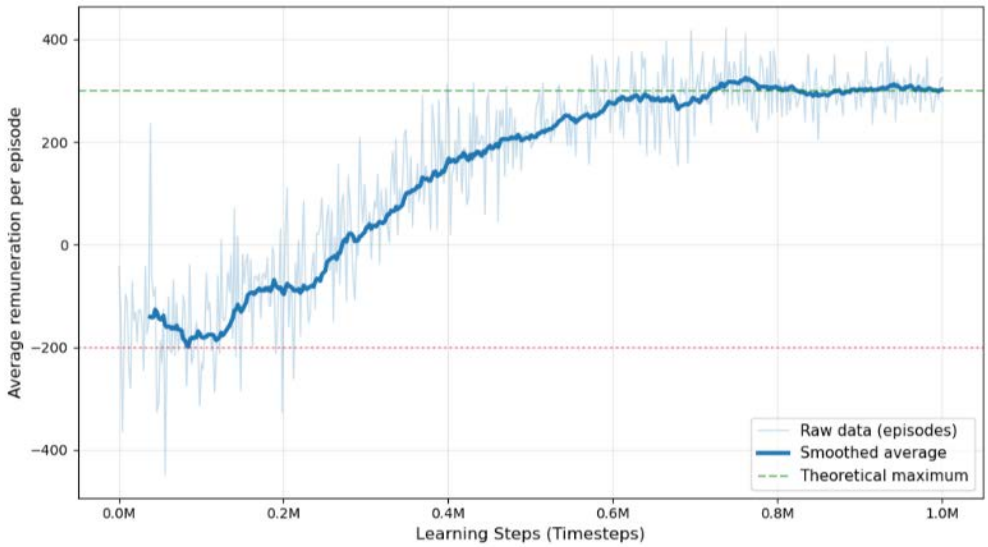

Fig. 1. **RL model learning curve graph**

The first phase (0 – 500k steps) is characterized by high reward variance, corresponding to the environment exploration stage. The second phase (500k – 1.5m steps) demonstrates rapid growth in average reward, indicating the formation of an effective stabilization policy. In the third phase, the curve plateaus, confirming the algorithm's convergence to an optimal strategy.

Results. Experimental verification of the developed RL controller was conducted through comparative analysis with a reference PID regulator tuned using the Ziegler-Nichols method for the nominal vehicle mass. A series of tests was conducted in three scenarios.

1. Target reaching scenario. This experiment evaluated the speed and accuracy of reaching a specified point located 10 meters from the starting position. The graph (Fig. 2) demonstrates how the UAV reaches the target under the control of the PID controller and the RL controller.

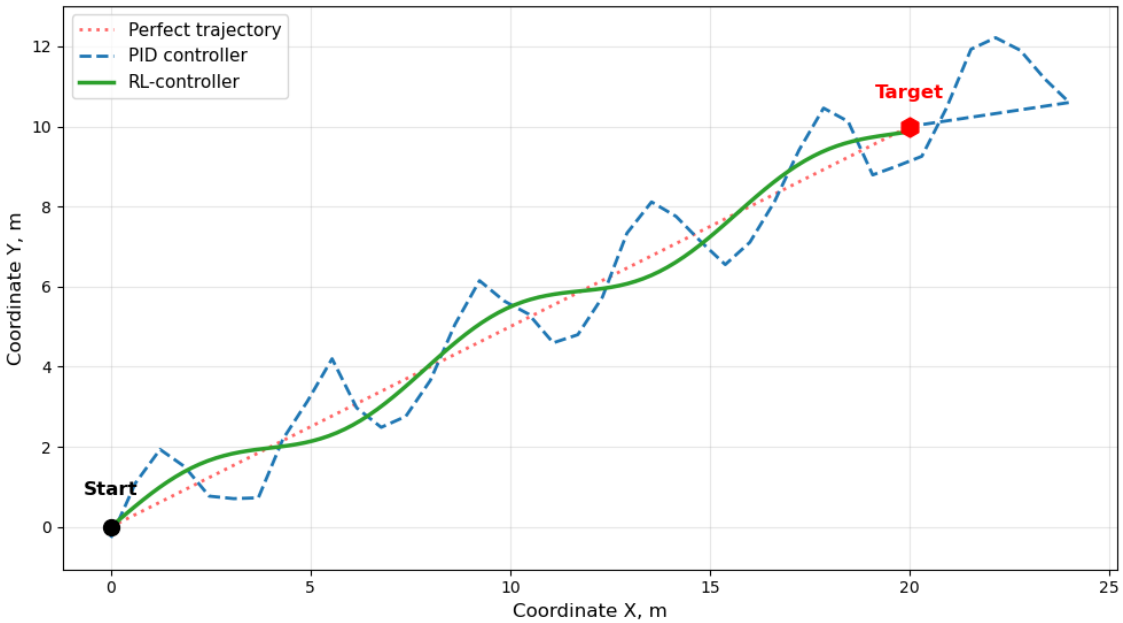

Fig2. **Comparison of trajectories in the goal achievement scenario**

The presented graphs imply that the RL controller provides better control quality indicators compared to the PID regulator. In particular, the average time to reach the target state decreases by approximately 14%, and the overshoot magnitude is smaller. This indicates the RL agent's formation of smoother and more optimal motion trajectories without pronounced oscillations in the target point vicinity. Additionally, the advantages of the RL controller are confirmed by a lower value of conditional energy consumption, defined as the integral of the sum of squared control signals. In contrast, the PID regulator,

even with correct tuning, is characterized by more aggressive control actions, leading to increased overshoot and higher energy costs.

2. Response to wind disturbance. In this experiment, a constant wind load with a speed of 3 m/s was applied to the UAV in hovering mode at time t = 5s.

The transient process graph (Fig. 3) demonstrates the deviation of the vehicle's coordinates from the target point.

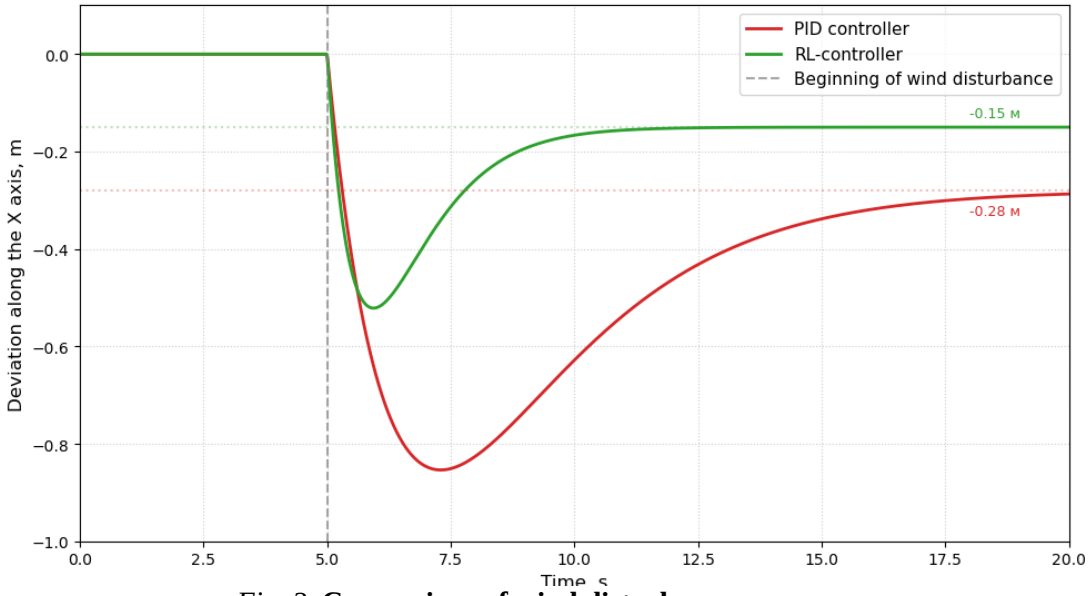

Fig. 3. **Comparison of wind disturbance response**

It is evident from the graphs that the PID regulator reacts with a delay, allowing a maximum deviation of 0.85 m. The process of returning to the point is accompanied by damped oscillations lasting about 3 s. The static error is 0.28 m. The RL controller demonstrates predictive behavior, minimizing deviation to 0.52 m (a 39% improvement). The transient process time is reduced to 2 s, and the static error is reduced to 0.15 m. This is explained by the neural network having learned to compensate for disturbances by changing the vehicle's angle of attack more aggressively yet precisely.

3. Adaptation to mass change (Fig. 4). The scenario simulates a sudden load change, resulting in an instantaneous mass variation of 20%. The height change graph (Fig. 4) shows the vehicle's subsidence.

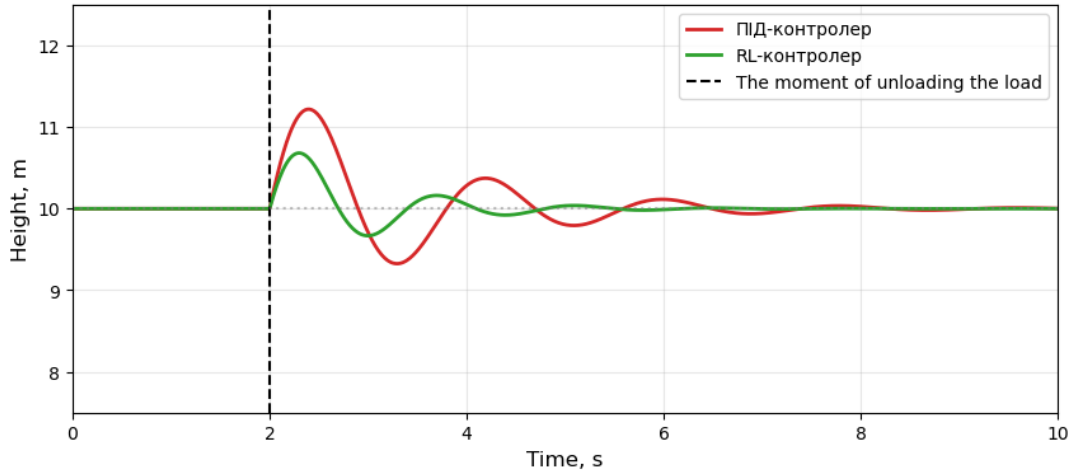

Fig. 4. **Response to load drop (height change)**

As seen from the figure, the PID regulator demonstrates a deep subsidence of 1.15 m. The regulator's integral component required 4.55 s to accumulate the error and increase thrust to the level necessary for weight compensation. The RL controller detects the discrepancy between the dynamics and the expected model within the first few steps after the mass change. Subsidence was only 0.68 m, and recovery of the set height took 2.91 s. This indicates a higher generalization capability of the trained policy.

Quantitative comparison results are presented in Table 1.

Discussion. The obtained results confirm the effectiveness of applying deep reinforcement learning for adaptive control tasks.

Analysis of transient processes reveals a fundamental difference in the operation of the investigated controllers. The PID regulator is a reactive system that begins to act effectively only after the occurrence of a significant control error. In contrast, the RL agent, trained in an environment with randomized parameters, forms a kind of "intuitive" model of dynamics. It is capable of generating control actions that preemptively prevent the development of significant deviations.

Table 1.

Comparison of controller performance metrics

Metric	PID Controller	RL (PPO) Controller	Improvement
Settling time (mass change), s	4.55 ± 0.31	2.91 ± 0.28	36%
Max deviation (wind), m	0.85 ± 0.11	0.52 ± 0.09	39%
Static error (wind), m	0.28 ± 0.05	0.15 ± 0.04	46%

Energy efficiency is an important aspect. Although the RL controller acts faster, analysis of control signals indicates that it avoids high-frequency thrust oscillations, which frequently occur in PID regulators with high gain coefficients (specifically the D-component). This is achieved by including a penalty for abrupt action changes (R_{act}) in the reward function.

A limitation of the proposed method is the sim-to-real gap. Although domain randomization significantly enhances robustness, deploying the model to a physical controller may require additional fine-tuning on real-world data to compensate for effects that are difficult to model (e.g., complex turbulent flows or communication channel delays). It is also worth noting the high computational cost of the training stage, which, however, is compensated by the high speed of the trained neural network during inference

Conclusions. The paper presents and investigates an adaptive UAV control algorithm based on reinforcement learning. The use of the PPO algorithm combined with a domain randomization strategy allowed for the synthesis of a control law that surpasses traditional methods in terms of robustness and response speed.

It has been experimentally proven that the proposed approach reduces maximum trajectory deviation under wind loads by 39% and shortens the adaptation time to mass changes by 36% compared to the PID regulator. This makes the developed method promising for application in autonomous logistics and monitoring systems operating in variable weather conditions. Future research will focus on integrating Recurrent Neural Networks (LSTM) to improve the system's memory of past states and account for time delays.

List of references:

1. Wang, Hongpeng, Shangyuan Song, Qianghui Guo, Dian Xu, Xiaoyang Zhang, and Peizhao Wang. "Cooperative motion planning for persistent 3d visual coverage with multiple quadrotor uavs." IEEE Transactions on Automation Science and Engineering 21, no. 3 (2023): 3374-3383.
2. Sun, Wendi, and Mingrui Hao. "A survey of cooperative path planning for multiple UAVs." In International Conference on Autonomous Unmanned Systems, pp. 189-196. Singapore: Springer Singapore, 2021
3. Sutton R. S., Barto A. G. Reinforcement Learning: An Introduction. 2nd ed. URL: http://incompleteideas.net/book/the-book-2nd.html (date of access: 10.11.2025)
4. Pi C.-H., Ye W.-Y., Cheng S. Robust Quadrotor Control through Reinforcement Learning with Disturbance Compensation. Applied Sciences. 2021. Vol. 11, no. 7. P. 3257. URL: https://doi.org/10.3390/app11073257 (date of access: 14.11.2025).
5. Zhang R., Zhang D., Mueller M. W. ProxFly: Robust Control for Close Proximity Quadcopter Flight via Residual Reinforcement Learning. arXiv preprint arXiv:2409.13193, 2024. URL: https://doi.org/10.48550/arXiv.2409.13193
6. Vaidya V., Keshavan J. Dynamics-Invariant Quadrotor Control using Scale-Aware Deep Reinforcement Learning. arXiv preprint arXiv:2503.09622, 2025. URL: https://doi.org/10.48550/arXiv.2503.09622

7. Omoniwa, Babatunji, Boris Galkin, and Ivana Dusparic. "Communication-enabled deep reinforcement learning to optimise energy-efficiency in UAVassisted networks." Vehicular Communications 43 (2023): 100640.
8. Nahrendra I.M.A., Tirtawardhana C., Yu B., Lee E.M., Myung H. Retro-RL: Reinforcing Nominal Controller With Deep Reinforcement Learning for Tilting-Rotor Drones. arXiv preprint arXiv:2207.03124, 2022. URL: https://doi.org/10.48550/arXiv.2207.03124

Reviewer: Pasternak Yaroslav, Doctor of Physical and Mathematical Sciences, Professor of the Department of Computer Science and Cybersecurity at Lesya Ukrainka Volyn National University.