

Reinforcement Learning Framework for Adaptive Control of Quadrotor UAVs

An internship report submitted in partial fulfillment of the
requirements for the award of the degree of

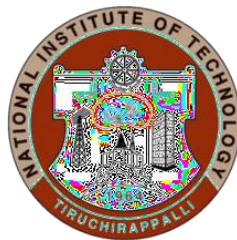
B.Tech

in

Instrumentation and Control Engineering

By

Shreehan S Kate (Roll No. 110122101)



**INSTRUMENTATION AND CONTROL ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
TIRUCHIRAPPALLI – 620015**

OCTOBER 2025

BONAFIDE CERTIFICATE

This is to certify that the report titled **Reinforcement Learning Framework for Adaptive Control of Quadrotor UAVs** is a bonafide record of the work done during the **SUMMER INTERNSHIP** by

Shreehan S Kate (Roll No. 110122101)

At **National Institute of Technology, Tiruchirappalli** during the period of **06/05/2025 till 21/07/2025**

in partial fulfillment of the requirements for the award of the degree of **Bachelor of Technology in Instrumentation and Control Engineering** of the **NATIONAL INSTITUTE OF TECHNOLOGY, TIRUCHIRAPPALLI**, during the year 2024-25.

Coordinator – Summer Internship

Head of the Department

Internship Review Presentation held on _____

ABSTRACT

Nonlinearity, underactuated configuration, and vulnerability to uncertainties and disturbances make the Quadrotor Unmanned Aerial Vehicles (UAVs) a complex control problem. Feedback linearization is a nonlinear control technique that leverages nonlinear transformations to obtain a linearized model of the original system. This transformation facilitates the application of linear control strategies. Fixed-gain PID controllers often suffer from performance degradation caused by parameter variations and external disturbances. This study proposes an adaptive PID controller, leveraging the Deep Reinforcement Learning (DRL) framework for altitude and attitude control. The Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is employed to tune the PID controller gains in an online, model-free environment. The TD3 agent learns control policies that dynamically modify PID gains in response to observed system behavior through continuous interaction with the environment. The performance of the TD3-driven PID controller is assessed through extensive simulations performed on the feedback-linearized quadrotor model under varying altitude trajectories and external disturbances, followed by a comparative analysis between the proposed controller, fixed-gain PID and other commonly known linear control strategies to evaluate improvements in robustness and tracking.

Keywords: Deep Reinforcement Learning, Model Predictive Control, Quadrotor dynamics, PID Controller, Linear Quadratic Regulator, Feedback linearization (FL), Actor-Critic Networks,

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to all those who have supported me throughout the course of my internship, and who have contributed to the successful completion of this report.

First and foremost, I would like to extend my heartfelt thanks to Prof. Dr. K Srinivasan, my mentor, for providing me with the opportunity to work under his expert guidance. His vast knowledge, insightful feedback, and constant encouragement have been invaluable to me throughout this internship.

I would also like to acknowledge Prof. Dr. Shuprajhaa T, Assistant Professor, Department of ECE, School of Engineering, Amrita Vishwa Vidhyapeetham, Coimbatore under whose guidance I carried out my research. Her support in technical matters, research methodologies, and intellectual guidance was pivotal to my learning and personal growth during the internship. I appreciate her patience, direction, and collaborative approach, which greatly enriched my experience.

My sincere thanks go to RECAL and NITT Batch of 1974, the generous organization that provided the research fellowship for this internship. Their financial support made this valuable experience possible, enabling me to dedicate myself fully to the research work and gain hands-on experience in the field. I am truly grateful for their contribution to my academic and professional development.

I would also like to take this opportunity to express my gratitude to the Department of Instrumentation and Control Engineering at the National Institute of Technology, Tiruchirappalli. The department's stimulating academic environment, rigorous coursework, and the opportunity to engage in research projects like this one have played an essential role in shaping my academic and career aspirations.

Finally, I wish to thank my family and friends for their love, encouragement, and unwavering support throughout the internship.

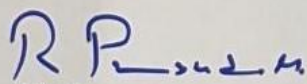


**Dr. RAM AND THAILA RAMANUJAM FOUNDATION'S
TOP MERIT CERTIFICATE – 2025**

This is to certify that **Mr. Shreehan S Kate**, 110122101 III year of **Instrumentation and Control Engineering Department**, NIT Tiruchirappalli has been presented a Top Merit Certificate jointly by **Dr. Ram and Thaila Ramanujam Foundation** and **1974 Batch Alumni**, in recognition of his proposal titled, '**Adaptive reinforcement learning framework for predictive control in autonomous quadrotor systems**' under the supervision of **Dr. K. Srinivasan**.

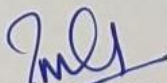
This certificate serves as a Testament to his academic merit, and potential to make significant contributions. We appreciate Mr. Shreehan S Kate for his interest in submitting the research proposal and wish him continued success in his research endeavors.

Awarded on 17th of March, 2025.



Chairpersons

1974 Batch Scholarship Committee



Dean (ID&AR)



Director, NITT



Department of Instrumentation and Control Engineering

National Institute of Technology
(Formerly Regional Engineering College)
Tiruchirappalli – 620015, Tamil Nadu, India

Website:

https://www.nitt.edu/home/academics/departments/ice/faculty/Dr_K.Srinivasan

Dr .K. Srinivasan

Professor

Email : srinikkn@nitt.edu

INTERNSHIP CERTIFICATE

This is to certify that Mr. Shreehan S Kate, Third year, B.E. student from the Department of Instrumentation and Control Engineering, "National Institute of Technology", Trichy has successfully completed his (ONLINE) internship program from 06.05.2025 to 21.07.2025 in the Department of Instrumentation and Control Engineering. His internship focused on Reinforcement Learning for Controller Design . His Conduct during the Internship period is very good.

Dr .K. Srinivasan

TABLE OF CONTENTS

Title	Page No.
ABSTRACT	i
ACKNOWLEDGEMENTS	ii
ACCEPTANCE LETTER.....	iii
INTERNSHIP CERTIFICATE	iv
TABLE OF CONTENTS.....	v
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER 1 INTRODUCTION.....	8
CHAPTER 2 METHODS & MATERIALS.....	9
2.1 Nonlinear Quadrotor UAV Model	9
2.2 Feedback Linearization	12
2.3 Formulation of Control Problem.....	13
2.4 Proposed Actor-Critic Architecture	16
2.5 Environment setup and Reward Function Design.....	18
CHAPTER 3 RESULTS.....	20
3.1 Training	21
3.2 Tracking Performance	22
CHAPTER 4 DISCUSSION & CONCLUSIONS.....	23
LEARNING OUTCOMES	25
REFERENCES.....	26

LIST OF TABLES

2.1	Quadrotor Parameters.....	5
2.2	Actor Network.....	5
2.3	Critic Network.....	5
2.4	Training Parameters	5
2.5	Time Domain Metrics	5
2.6	Performance Metrics	5

LIST OF FIGURES

2.1	Diagram of the UAV	4
3.1	Tracking Performance of the UAV	6

CHAPTER 1

INTRODUCTION

Quadrotor Unmanned Aerial Vehicles (UAVs) have gained popularity in applications like last-mile delivery, environmental monitoring, infrastructure inspection, and aerial surveillance because of their small size, high manoeuvrability, and ability to take off and land vertically. Due to their intrinsically nonlinear and underactuated dynamics, high sensitivity to modelling uncertainties, and external disturbances like wind gusts or shifting payloads, quadrotors pose significant control challenges despite their versatility. For stable and accurate operation under a variety of dynamic conditions, these features call for sophisticated control strategies that can adjust in real time.

Among the widely adopted control strategies, the Proportional-Integral-Derivative (PID) controller remains a cornerstone in industrial and academic applications, valued for its conceptual simplicity and reliable performance in well-understood environments. However, its effectiveness is highly dependent on the appropriate tuning of the controller gains (K_P , K_I , K_D) which are traditionally set through heuristic or offline methods such as Ziegler-Nichols or Cohen-Coon. These approaches lack the flexibility to adjust to real-time changes in system dynamics, leading to degraded performance under varying conditions.

Linear Quadratic Regulator (LQR) is another foundational optimal control strategy widely used in both theoretical and practical control applications due to its elegant formulation and guaranteed stability under ideal conditions. LQR computes a state feedback gain matrix that minimizes a quadratic cost function, balancing state deviations and control effort. Like MPC, LQR requires the careful selection of weight matrices Q and R , which directly influence the controller's aggressiveness and energy usage. While LQR performs well in linear, time-invariant systems with known dynamics, its lack of built-in constraint handling and sensitivity to model inaccuracies can limit its effectiveness in complex, real-world environments. Furthermore, static weight tuning fails to accommodate dynamic operating conditions, often necessitating gain scheduling or adaptive extensions for robust performance.

Model Predictive Control (MPC), in contrast, offers a more sophisticated framework that optimizes control inputs over a prediction horizon while incorporating system constraints and future reference trajectories. However, even MPC relies on manually chosen cost function weights - namely Q , R , and ΔR , that govern the trade-off between tracking accuracy, control effort, and input smoothness. Improperly tuned weights can result in overly aggressive or overly conservative control behavior, particularly when the system operates outside nominal conditions.

To overcome the limitations of fixed-parameter control schemes, this project explores the integration of Deep Reinforcement Learning (DRL) with PID frameworks to enable adaptive control. Specifically, the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm is employed due to its suitability for continuous control tasks and its improved stability over earlier actor-critic methods. TD3 enables real-time, model-free adaptation by learning optimal control parameters through continuous interaction with the environment and reward-driven feedback. By embedding TD3 into the PID design loop, the resulting controller can dynamically adjust their parameters: PID gains, leading to enhanced performance in terms of tracking accuracy, and robustness.

CHAPTER 2

METHODS & MATERIALS

This chapter introduces the Nonlinear Quadrotor Model, Feedback Linearization and subsequent PID and MPC controller design, and the process for designing the actor-critic network, reward function design and training the TD3 Agent.

2.1 Nonlinear Quadrotor UAV model

To design robust and adaptive controllers for altitude and attitude tracking, an accurate dynamic model of the quadrotor UAV is essential. The quadrotor system considered in this project is six degrees of freedom (6-DOF) nonlinear dynamical system, modeled using Newton-Euler equations. It comprises four rotors in a symmetric cross configuration, with opposite pairs rotating in counter directions to balance angular momentum and produce control torques. This arrangement enables the generation of control torques for pitch, roll, and yaw by varying the relative speeds of individual rotors.

Coordinate Frames and States

The system is described using two frames of reference:

- The **Inertial Frame (E-frame)**, fixed to the earth.
- The **Body Frame (B-frame)**, attached to the quadrotor's center of mass.

The translational position of the UAV is denoted by $\xi = [x, y, z]^T$ in the inertial frame, and the rotational orientation is given by Euler angles $\eta = [\phi, \theta, \psi]^T$, corresponding to roll, pitch, and yaw. The body and inertial frames are captured using a rotation matrix and a transformation matrix, which relate linear and angular velocities between the two frames.

Transformation between these frames uses:

- **Rotation Matrix R**: relates body-frame velocities to inertial-frame velocities.
- **Transformation Matrix T**: relates angular velocity ω to Euler angle rates $\dot{\eta}$.

Equations of Motion

The translational dynamics are given by:

$$\ddot{x} = \frac{1}{m} (\cos \phi \sin \theta \cos \psi + \sin \phi \sin \psi) u_4$$

$$\ddot{y} = \frac{1}{m} (\cos \phi \sin \theta \sin \psi - \sin \phi \cos \psi) u_4$$

$$\ddot{z} = -g + \frac{1}{m} \cos \phi \cos \theta u_4$$

The rotational dynamics (attitude) are:

$$\ddot{\phi} = \frac{\dot{\theta}\psi(I_y - I_z)}{I_x} + \frac{J_r \Omega_d}{I_x} + \frac{l}{I_x} u_1$$

$$\ddot{\theta} = \frac{\dot{\phi}\psi(I_z - I_x)}{I_y} + \frac{J_r \Omega_d}{I_y} + \frac{l}{I_y} u_2$$

$$\ddot{\psi} = \frac{\dot{\phi}\dot{\theta}(I_x - I_y)}{I_z} + \frac{1}{I_z} u_3$$

Where:

- m is the mass of the UAV.
- g is gravitational acceleration.
- I_x, I_y, I_z are moments of inertia along respective axes.
- l is arm length.
- J_r is rotor inertia.
- $\Omega_d = -\Omega_1 + \Omega_2 - \Omega_3 + \Omega_4$ account for the gyroscopic disturbances present in the model

Control Inputs: The four control inputs are generated by varying rotor speeds

$$u_1 = b(\Omega_4^2 - \Omega_2^2), u_2 = b(\Omega_3^2 - \Omega_1^2)$$

$$u_3 = d(\Omega_4^2 + \Omega_2^2 - \Omega_3^2 - \Omega_1^2), u_4 = b(\Omega_1^2 + \Omega_2^2 + \Omega_3^2 + \Omega_4^2)$$

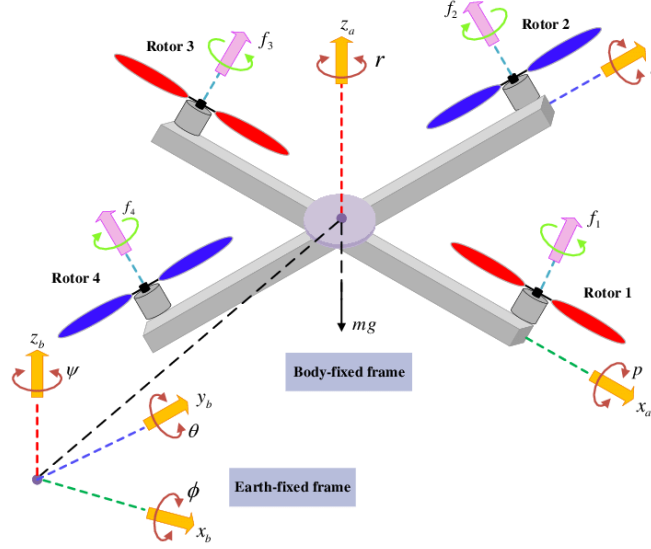


Figure 2.1: Diagram of the UAV

Physical parameters	Value
Mass (m)	0.650 kg
inertia along x axis (I_x)	0.0075 kgm ²
inertia along y axis (I_y)	0.0075 kgm ²
inertia along z axis (I_z)	0.013 kgm ²
thrust coefficient (b)	3.13e-5 Ns ²
drag coefficient (d)	7.5e-7 Nms ²
rotor inertia (J_r)	6e-5 kgm ²
arm length (l)	0.23 m

Table 2.1: Quadrotor Parameters

Table 1 lists the parameters used to model the quadrotor on a simulation platform. This nonlinear quadrotor model forms the foundational basis for the design, analysis, and simulation of advanced control strategies. However, the complexity introduced by its nonlinear, coupled dynamics makes direct application of classical linear controllers impractical. To overcome this, feedback linearization is applied in the subsequent section to systematically cancel nonlinear terms and decouple the dynamics. This transformation results in a linearized model for each control axes: roll, pitch, yaw, and altitude, facilitating the application of linear control techniques such as PID, Linear Quadratic Regulator (LQR) and Model Predictive Control (MPC).

2.2 Feedback Linearization

The nonlinear nature of quadrotor dynamics presents significant challenges in controller design. Traditional linear control strategies are not directly applicable due to strong coupling and state-dependent nonlinearities in the equations of motion. To address this, **Feedback Linearization (FBL)** is employed to transform the nonlinear system into an equivalent, decoupled linear system through input–output redefinition and cancellation of nonlinear terms. Feedback linearization involves expressing the control inputs u_1, u_2, u_3, u_4 as nonlinear functions of a new set of virtual control inputs v_1, v_2, v_3, v_4 which represent the desired second derivatives of the system's outputs (roll, pitch, yaw, and altitude). The physical control inputs u_1, u_2, u_3, u_4 which are functions of rotor thrusts, are then expressed in terms of these virtual control inputs by inverting the nonlinear dynamic equations.

The resulting expressions are:

$$\begin{aligned} u_1 &= I_x \left(-\dot{\theta}\dot{\psi} \frac{I_y - I_z}{I_x} - \dot{\theta}\Omega_d \frac{J_r}{I_x} + v_1 \right) \\ u_2 &= I_y \left(-\dot{\phi}\dot{\psi} \frac{I_z - I_x}{I_y} - \dot{\phi}\Omega_d \frac{J_r}{I_y} + v_2 \right) \\ u_3 &= I_z \left(-\dot{\phi}\dot{\theta} \frac{I_x - I_y}{I_z} + v_3 \right) \\ u_4 &= \frac{m}{\cos \phi \cos \theta} (g + v_4) \end{aligned}$$

The resulting linear system is as follows:

$$\ddot{\phi} = v_1, \quad \ddot{\theta} = v_2, \quad \ddot{\psi} = v_3, \quad \ddot{z} = v_4$$

By applying these equations, the nonlinear couplings and gyroscopic effects are effectively cancelled. This results in a decoupled linear system where each of the four virtual inputs v_1 through v_4 directly governs the second derivative of the corresponding output states. This transformation enables the application of classical linear control laws (e.g., PID or MPC) on each axis independently, simplifying controller synthesis and analysis.

The linearized dynamics of the quadrotor are expressed in state-space form as shown below:

$$y = C(x), \quad \dot{x} = Ax + Bv$$

Where A, B, C are the canonical forms of the feedback-linearized UAV model.

2.3 Formulation of the Control problem

Proportional-Integral-Derivative Control

Following feedback linearization, each output of the quadrotor system—roll ϕ , pitch θ , yaw ψ , and altitude z is treated as a decoupled linear subsystem. For each of these, an independent PID controller is designed. PID controllers are widely used in industry due to their intuitive structure.

Each PID controller computes a **virtual control input** $v_i(t)$, which serves as the desired second-order acceleration for the corresponding output state.

The standard PID control law is given by:

$$v_i(t) = K_{p_i} \cdot e_i(t) + K_{i_i} \cdot \int e_i(t)dt + K_{d_i} \cdot \frac{de_i(t)}{dt}$$

Where:

- $e_i(t) = y_i^{\text{ref}}(t) - y_i(t)$ is the tracking error for the i^{th} output,
- $K_{p_i}, K_{i_i}, K_{d_i}$ are the proportional, integral, and derivative gains for the i^{th} controller,
- $v_i(t)$ is the output of the PID controller, used as a virtual control input.

These gains are traditionally tuned offline using methods like Ziegler-Nichols or trial-and-error. However, in dynamic or uncertain environments, **fixed gains** may lead to suboptimal performance. This project overcomes that limitation by using a **Twin Delayed Deep Deterministic Policy Gradient (TD3)** agent to **adaptively tune the PID gains** online. The TD3 agent observes system performance in real time and adjusts the gains to maintain optimal control under changing conditions such as wind disturbances or payload variation.

Model Predictive Control

Model Predictive Control (MPC) is an advanced optimal control strategy that computes future control actions by solving an optimization problem over a finite prediction horizon. It explicitly incorporates system dynamics and constraints, making it particularly suitable for UAV systems that must maintain safety and performance under input and state limitations.

The feedback-linearized system is modeled in discrete-time state-space form for each output channel, which forms the prediction model

$$x_{k+1} = Ax_k + Bu_k$$

$$y_k = Cx_k$$

Where:

- x_k is the state vector at time step k,
- u_k is the control input (virtual control input),
- y_k is the system output,
- A, B, and C are the system matrices derived via feedback linearization

MPC Cost Function

At each control interval, the MPC controller solves the following quadratic cost function to minimize tracking error and control effort:

$$J = \sum_{i=0}^N \left[(y_i - y_i^{ref})^T Q (y_i - y_i^{ref}) + \Delta u_i^T R \Delta u_i + u_i^T \Delta R u_i \right]$$

Where:

- Q is the weight matrix penalizing output tracking error,
- R penalizes changes in control input: $\Delta u_i = u_i - u_{i-1}$
- ΔR penalizes absolute control magnitude,
- N is the prediction horizon.

Constraints Handling

MPC allows incorporation of actuator constraints and rate limits naturally into its formulation:

$$u_{min} \leq u_k \leq u_{max}$$

$$\Delta u_{min} \leq \Delta u_k \leq \Delta u_{max}$$

These constraints prevent the optimizer from issuing commands that exceed actuator limits or cause abrupt motion, ensuring safe and feasible control signals.

Linear Quadratic Regulator

Linear Quadratic Regulator (LQR) is a widely used optimal control strategy that determines the optimal state feedback gain to minimize a cost function that balances control effort and state deviation. LQR is well-suited for systems with known linear dynamics and provides a systematic way to design stable and efficient controllers for UAV systems under nominal operating conditions

State Feedback Control

LQR relies on full-state feedback, where the control input is computed based on the measured or estimated system states:

$$u_k = -Kx_k$$

Where:

- x_k is the state vector at time step k ,
- K is the optimal state feedback gain matrix computed via LQR.

The system dynamics are modeled in discrete-time state-space form:

$$x_{k+1} = Ax_k + Bu_k$$

Where A and B are the system matrices derived from feedback linearization.

LQR Cost Function

LQR optimizes a quadratic cost function to minimize the cumulative weighted state error and control effort:

$$J = \sum_{k=0}^{\infty} (x_k^T Q x_k + u_k^T R u_k)$$

Where:

- $Q \geq 0$ is the state weighting matrix (penalizes deviation from the desired state),
- $R > 0$ is the input weighting matrix (penalizes control effort),
- The balance between Q and R governs the optimality of the controller.

The optimal gain matrix K is computed by solving the Discrete-Time Algebraic Riccati

Equation (DARE):

$$P = A^T P A - A^T P B (R + B^T P B)^{-1} B^T P A + Q$$

$$K = (R + B^T P B)^{-1} B^T P A$$

Selecting optimal Q and R matrices directly influences controller performance:

- Large Q: prioritizes tracking performance, may lead to higher control effort.
- Large R: suppresses control effort, can cause sluggish or slower behavior.

Limitations and Extensions

Standard LQR assumes perfect model knowledge and unconstrained inputs. In dynamic and uncertain environments, its performance may degrade. To address this, adaptive or gain-scheduled LQR and integration with observers (e.g., Kalman filters) are employed.

A key enhancement in this project is the use of a **TD3 agent** to adaptively optimize the PID gains. Instead of relying on fixed, hand-picked values, the agent learns to adjust these weights online based on system behavior. The reward function incentivizes precise tracking with minimal control effort and smooth transitions. Over time, the TD3 agent develops a policy that dynamically tunes the cost function for optimal performance, making the proposed strategy **adaptive** and **robust to disturbances**. In the further sections, the proposed actor-critic architecture along with performance benchmarking with fixed-gain PID, linear MPC and standard LQR designs will be discussed.

2.4 Proposed Actor-Critic Architecture

The actor-critic framework forms the backbone of the adaptive PID tuning strategy in the proposed quadrotor UAV control architecture. This section details the structural design, and operational aspects of the actor-critic reinforcement learning approach, with a focus on the implementation of the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm. The TD3-driven architecture enables dynamic adaptation of PID gains for robust altitude and attitude control without relying on an explicit system model.

Actor-critic methods combine the strengths of both policy-based and value-based reinforcement learning:

- Actor Network: Learns and updates a policy that maps observed states to specific actions.
- Critic Network: Evaluates the quality of actions taken by the actor by estimating expected cumulative rewards (Q-values).

This project adopts the TD3 algorithm for the design of adaptive PID control architecture.

TD3 addresses two major challenges in actor-critic methods: function approximation errors and overestimation bias of Q values.

- **Clipped Double Q-learning:** Utilizes two critic networks, selecting the minimum estimated Q-value to mitigate overestimation bias.

$$y = r + \gamma \min_{i \in \{1,2\}} Q'_{\theta_i}(s', \pi_{\phi'}(s') + \epsilon), \quad \epsilon \sim \text{clip}(\mathcal{N}(0, \sigma), -c, c)$$

- **Delayed Policy Update:** Updates the actor network less frequently than the critics, contributing to stable policy evolution.

$$\mathcal{L}(\theta_i) = E_{(s,a,r,s') \sim \mathbb{D}} \left[(Q_{\theta_i}(s, a) - y)^2 \right], \quad i \in \{1,2\}$$

- **Target Policy Smoothing:** Adds noise to policy targets during critic updates, improving robustness by regularizing value estimates.

$$\nabla_{\phi} J(\phi) = E_{s \sim \mathbb{D}} \left[\nabla_a Q_{\theta_1}(s, a)|_{a=\pi_{\phi}(s)} \cdot \nabla_{\phi} \pi_{\phi}(s) \right]$$

- **Soft Target Updates:** Uses weighted averaging for updating target networks, thus preventing abrupt changes and promoting smoother training convergence.

$$\theta'_i \leftarrow \tau \theta_i + (1 - \tau) \theta'_i, \quad \phi' \leftarrow \tau \phi + (1 - \tau) \phi'$$

These mechanisms jointly strengthen the reliability and adaptability of the actor-critic controller, allowing for real-time PID gain adjustment in response to varying operational conditions. The actor-critic architectures are summarized in the table 1 and 2.

Table 2.2: Actor Network Architecture

Layer	Nodes	Activation Function
Input Layer	16	N/A
First Hidden Layer	256	ReLU + LayerNorm
Second Hidden Layer	256	ReLU + LayerNorm
Output Layer	12	Tanh + Scaling

Table : Critic Network Architecture (Per Q-Function)

Layer	Nodes	Activation Function
State Input Layer	16	N/A
Action Input Layer	12	N/A
Hidden Layer (State)	128	ReLU + LayerNorm
Hidden Layer (Action)	128	ReLU + LayerNorm
Merged Layer	128	ReLU
Output Layer	1	Identity

2.5 Environment Setup and Reward Function Design

The environment consists of a 16-dimensional observation space and a 12-dimensional action space. The observation space includes the elements required for PID adaptation: the tracking error and its integral and derivative signals and the corresponding output states for each of the four axes (ϕ , θ , ψ , z). These observations are modeled into a vector of size 16x1, with no upper or lower bounds to ensure the RL agent explores the entire observation space. The actions are formulated as a continuous vector of dimension 12x1, corresponding to the three PID gains for each of the control axes. Each gain value is bound within the range to ensure practical feasibility. The TD3 agent receives the states defined as observation space. The actor network maps the states to actions, facilitating continuous online tuning. Table 3 lists the parameters used for the training of the proposed actor-critic network.

Table 2.4: Training Parameters

Hyperparameter	Value
Total number of episodes	1000
Maximum steps per episode	100
Actor Learning Rate	0.0005
Critic Learning Rate	0.001
Gradient Threshold	1
Sample Time (s)	0.1
Mini-Batch Size	256
Experience Buffer Length	2000000
Exploration Noise Mean	0
Exploration Noise Standard Deviation	0.15
Standard Deviation Decay Rate	0.003

This work adopts a Linear Quadratic Gaussian (LQG)-inspired cost formulation to define the reward signal used for training the reinforcement learning agent. LQR serves as an evaluation metric that balances tracking accuracy and control effort. In this research, a TD3-based agent is employed to adaptively tune the PID gains for altitude and attitude control. The reward function is constructed as the negative of the LQR=G cost, thereby encouraging the agent to minimize both tracking error and control effort. This formulation guides the agent toward optimal control policies that balance performance and efficiency.

$$R = - \sum_{k=0}^T (w_{\text{error}} \cdot e_k^2 + w_{\text{action}} \cdot u_{k-1}^2)$$

Weights w_{error} and w_{action} are assigned to tracking error and control input terms, respectively. A higher weight imposes a greater penalty on the corresponding term in the cost function, thereby encouraging the agent to minimize that component more aggressively during training.

Figure 2 illustrates the block diagram of a closed-loop control system designed for attitude and altitude tracking of the quadrotor UAV. This system utilizes four individual PID controllers, each responsible for one of the four state variables to achieve the desired flight performance.

The TD3 agent learns from variations in tracking errors induced by external disturbances and stochastic reference signals, thereby enhancing the robustness of the control system. The actor network maps observed system states to control actions in real time, enabling online tuning of the PID gains. These gains are then applied to the PID controllers, ensuring adaptive and efficient tracking under uncertain operating conditions.

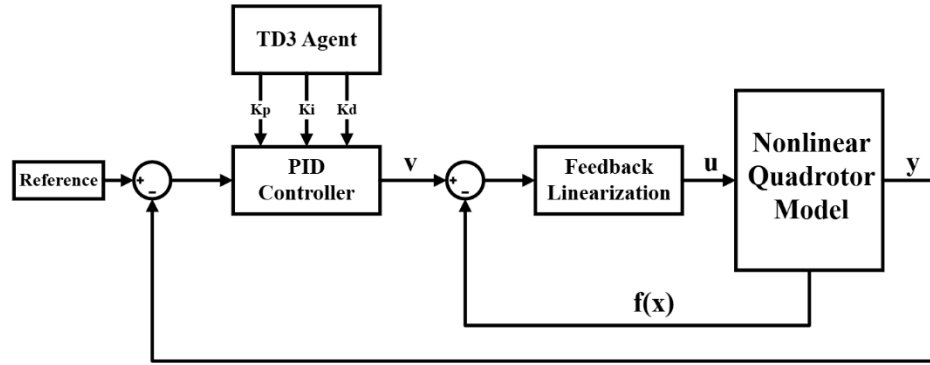


Figure 2.2: Block diagram of the control system.

During the exploration phase, uncontrolled state trajectories can result in excessive accumulation of integral error, leading to high-variance gradients and instability in the process of convergence. An early stopping criterion during the training process is introduced to tackle the problem of divergence. The proposed algorithm defines a process dependent constraint $y \in [y_{\min}, y_{\max}]$ on the system output y . During the training process, if the system outputs explode or go out of bounds, the episode is terminated immediately, and a penalty is assigned to the cost, which encourages the agent not to pursue actions that result in unsafe behavior. The early stopping criterion ensures that the TD3 agent is trained in the absence of unstable state trajectories and eventually learns a control policy that caters to the system's operational boundaries. In addition, a reward bonus is provided upon successful completion of the episode and a low Integral Square error (ISE) for all the state variables.

CHAPTER 3

RESULTS

3.1 Training

This section presents the results obtained from training and simulations. The reinforcement learning (RL)-driven PID control scheme was implemented on a feedback-linearized quadrotor UAV model, incorporating the nonlinear dynamics described in Section II. The control strategy outlined in Chapter 2 was employed to conduct the simulations. The nonlinear UAV model, feedback linearization equations, and PID control architecture were developed in MATLAB and Simulink. The RL training environment, including the actor-critic architecture used for the Twin Delayed Deep Deterministic Policy Gradient (TD3) agent, was designed using MATLAB's Reinforcement Learning Designer App. The training was conducted using the parameters described in Section IV.B. The progression of the TD3 agent's training is illustrated in Fig. 3.

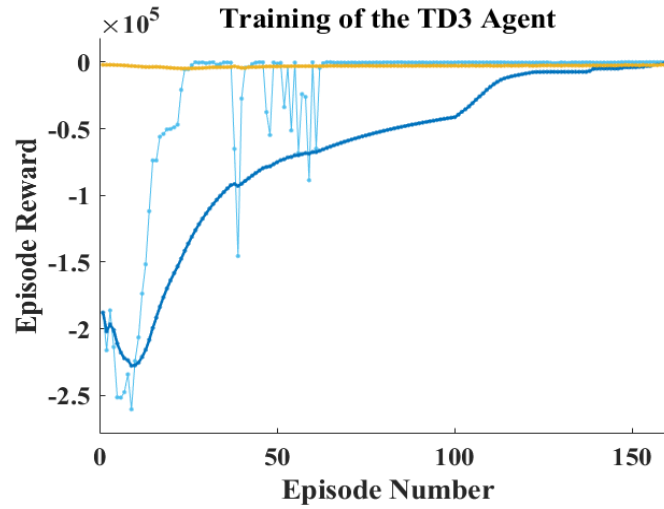


Figure 3.1: Training Progress

The training process incorporated a threshold to stop training. The training is stopped when the average reward over a finite number of episodes exceeds a particular threshold. The performance of the adaptive PID controller is compared with the fixed-gain PID, Linear MPC, and LQR control schemes. The feedback-linearized quadrotor UAV was made to track an altitude step change of 1 metre for altitude tracking and the remaining attitude variables were meant to drive to zero radians. In addition, the system was simulated to track a sine and a square wave input to validate

the robustness of the altitude controller. The response of the simulated system is documented below in Figure 3.1. The time-domain metrics of both the controllers is listed in Table 4

3.2 Tracking Performance

The proposed control architecture effectively enables accurate attitude and altitude tracking. System performance was evaluated using standard performance metrics, including Root Mean Square Error (RMSE), Integral of Squared Error (ISE), Integral of Absolute Error (IAE), and time-domain specifications such as rise time, settling time, and overshoot. Optimal tracking behavior is achieved by maximizing the reward function, as detailed in Chapter 2. The response of the simulated system is documented below in Figure 3.1. The time-domain metrics of both the controllers is listed in Table 3.1

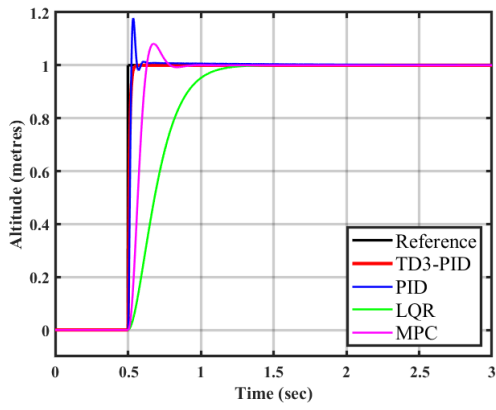
Table 3.1: Time-domain metrics

Time Domain Parameters	PID	LQR	MPC	TD3-PID
Settling time (sec)	1.0	0.6008	0.2548	0.0354
Rise Time (sec)	0.0161	0.3527	0.0828	0.0195
Overshoot (%)	17.84	0.0013	8.1578	0
Peak Time (sec)	0.036	1.297	0.1766	12
Peak	1.1783	1.00	1.0816	1.00
Settling min	0.9027	0.9005	0.9031	0.9057
Settling max	1.1783	1.00	1.0186	1.00

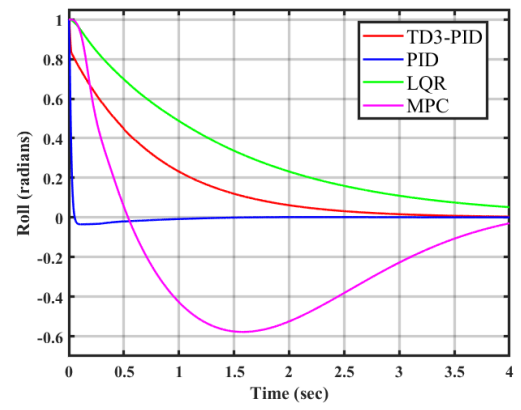
Table 3.2: Performance Metrics

Performance Metric	PID	LQR	MPC	TD3-PID
RMSE	$2.86 * 10^{-6}$	$1.25 * 10^{-17}$	$4.16 * 10^{-17}$	$1.97 * 10^{-10}$
IAE	0.08148	0.2138	0.075	0.03
ISE	0.0905	0.1357	0.051	0.05067

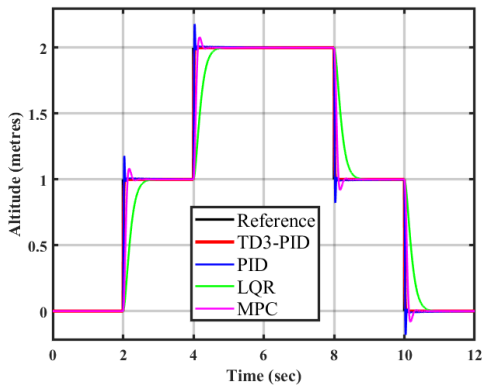
Table 3.2 indicates that a significant improvement in performance metrics across the altitude axes has been achieved. Reduced IAE and ISE values demonstrate improved tracking capabilities of the proposed controller. It also achieves faster response and lower overshoot compared to the other control schemes. The controller ensures that stability and robustness are maintained across a range of operating conditions and external disturbances. This dynamic gain adjustment capability significantly enhances transient response without compromising steady-state accuracy. Overall, the proposed framework demonstrates improved adaptability.



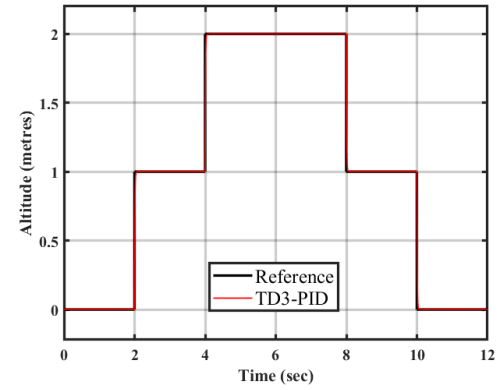
a) Altitude (step)



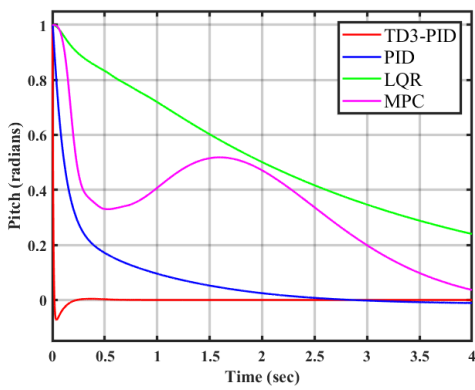
b) Roll



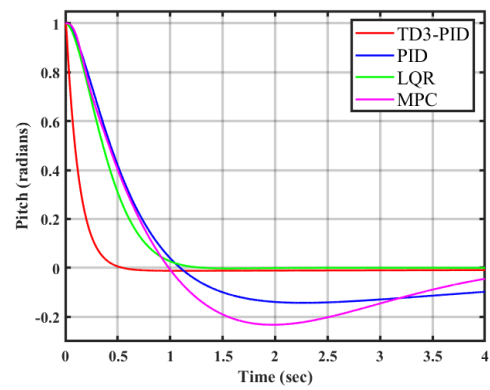
c) Square wave



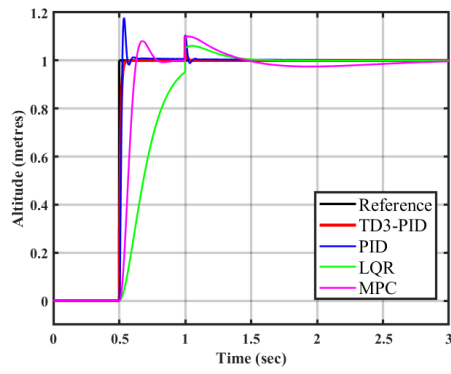
d) square wave tracking of the proposed controller



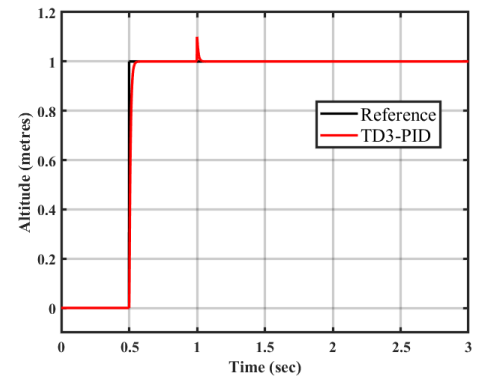
e) Pitch



f) Yaw



g) Disturbance rejection



h) Disturbance rejection of the proposed controller

Figure 3.1: Tracking performance of the UAV

CHAPTER 4

DISCUSSION, CONCLUSIONS AND FUTURE WORKS

By combining adaptive PID control with actor-critic reinforcement learning and feedback linearization, this study demonstrates improved tracking performance in quadrotor UAVs under dynamic and uncertain conditions. While traditional PID controllers offer simplicity and robustness, their fixed gain structure limits adaptability to varying flight scenarios and external disturbances. This limitation is addressed by deploying a Twin Delayed Deep Deterministic Policy Gradient (TD3) agent that dynamically tunes the PID parameters in real time, enabling adaptive control across a range of nonlinear operating conditions.

In order to apply classical control strategies, the nonlinear quadrotor dynamics are first converted into a feedback-linearized form. The control problem is simplified by this linearization, but it is still susceptible to disturbances and model uncertainties. As a result, a DRL-based gain tuning scheme is used, in which the TD3 agent optimises PID gains in response to ongoing environmental feedback. Even in the presence of disturbances, this interaction enables the agent to learn subtle behaviours in tracking both attitude and altitude commands.

For simplicity and low computational complexity, the PID structure is kept, but the addition of DRL significantly increases adaptability, particularly during brief flight phases. Measured by improvements in standard performance indices like RMSE, IAE, and ISE, the TD3-based control scheme performs better than a traditional fixed-gain PID controller in terms of tracking precision.

However, this study is not without limitations. The performance of the TD3 agent is contingent on training environment diversity and reward function design. Additionally, while the modular nature of the control scheme supports extension to other control architectures, real-time onboard deployment still requires further investigation regarding computational efficiency and hardware constraints. Nevertheless, this methodology highlights the potential of actor-critic networks in adaptive control of underactuated, nonlinear aerial systems. The demonstrated performance gains position this approach as a promising candidate for robust and intelligent UAV control in real-world scenarios, with possibilities to extend towards other robotics and aerospace platforms requiring robustness and precision in uncertain environments.

The proposed adaptive control architecture will be implemented on Parrot Mambo Minidrones in subsequent work, taking advantage of their easily accessible SDK and embedded control platform compatibility. The TD3 policy will be optimised for lightweight inference due to the limited computational resources onboard, either through embedded Python/C implementations or code generation from Simulink. Depending on the control loop configuration, Bluetooth or Wi-Fi will be used to enable real-time communication between the agent and the drone. Reducing latency in the transmission of control signals and state estimation, compensating for onboard sensor noise, and preserving flight stability under hardware limitations are important areas of focus during deployment. The proposed controller's small size and modular design make it ideal for Mambo's restricted environment, allowing adaptive control strategies to be validated in practice in a secure and expandable experimental setup. This deployment will be a first step towards more extensive hardware integration in increasingly autonomous UAV platforms.

LEARNING OUTCOMES

1. **Advanced Control Systems Design:** Gained hands-on experience in designing and implementing adaptive PID controllers for nonlinear, underactuated systems like quadrotors..
2. **Application of Reinforcement Learning in Control:** Applied the TD3 algorithm to enable real-time adaptive tuning of PID gains, enhancing controller robustness under varying flight conditions.
3. **Feedback Linearization of Nonlinear Dynamics:** Performed mathematical transformations to linearize the inherently nonlinear dynamics of quadrotor models, enabling the application of classical control techniques.
4. **Simulation and Validation in MATLAB/Simulink:** Built and validated the complete control pipeline in MATLAB/Simulink, simulating dynamic flight scenarios and disturbance conditions to evaluate controller performance.
5. **Performance Evaluation and Benchmarking:** Used quantitative metrics such as RMSE, IAE, and ISE to benchmark the adaptive controller against conventional fixed-gain PID, Linear MPC and LQR strategies.
6. **Integration of Control Theory and Machine Learning:** Bridged concepts from nonlinear control and deep reinforcement learning to tackle real-world control challenges in aerial robotics.
7. **Preparation for Hardware Deployment:** Developed insights into real-time implementation challenges, with a roadmap for deploying the control architecture on Parrot Mambo Minidrones.
8. **Research and Technical Writing Skills:** Documented experimental procedures, controller architectures, and results clearly and concisely, with emphasis on reproducibility and clarity.
9. **Critical Thinking and Problem Solving:** Identified limitations of traditional control methods and proposed data-driven adaptive alternatives suitable for complex environments.
10. **Professional Collaboration and Ethics:** Maintained transparency in reporting, collaborated effectively with academic mentors, and upheld ethical standards in algorithm testing and result interpretation.

REFERENCES

1. Alexandru BÂRSAN, "Position Control of a Mobile Robot through a PID controller", ACTA UNIVERSITATIS CIBINIENSIS – TECHNICAL SERIES Vol. 71 2019
2. Hans Oersted, Yudong Ma Review of PID Controller Applications for UAVs, <https://arxiv.org/abs/2311.06809>
3. Baha Zarrouki, Verena Klöos, Nikolas Heppner, Simon Schwan, Robert Ritschel and Rick Voßwinkel "Weights-varying MPC for Autonomous Vehicle Guidance: a Deep Reinforcement Learning Approach", 2021 European Control Conference (ECC)
4. K.-S. Hwang, S.-W. Tan, M.-C. Tsai, Reinforcement learning to adaptive control of nonlinear systems, IEEE Trans. Syst. Man Cybern. B 33 (3) (2003) 514–521, <http://dx.doi.org/10.1109/TSMCB.2003.811112>.
5. M. Srivastava, S. Indu, and R. Sharma, "Design and flight testing of LQR attitude control for quadcopter UAV," *arXiv preprint arXiv:2404.12261*, Apr. 2024.
6. M. Reich, "Error-state LQR formulation for quadrotor UAV trajectory tracking," *arXiv preprint arXiv:2501.15768*, Jan. 2025
7. H. H. Bilgic, M. A. Sen, and M. Kalyoncu, "Tuning of LQR controller for an experimental inverted pendulum system based on The Bees Algorithm," *J. Vibroengineering*, vol. 18, no. 6, pp. 3684–3694, Sep. 2016.
8. T. Shuprajhaa, Shiva Kanth Sujit, K. Srinivasan "Reinforcement learning based adaptive PID controller design for control of linear/nonlinear unstable processes " *Applied Soft Computing*, Volume 128, 2022, 109450, ISSN 1568-4946
9. Bartomeu Rubí and Bernardo Morcego and Ramon Perez, "A Deep Reinforcement Learning Approach for Path Following on a Quadrotor" in 2020 European Control Conference (ECC), Saint
10. Ding, Y.; Ren, X.; Zhang, X.; Liu, X.; Wang, X. Multi-Phase Focused PID Adaptive Tuning with Reinforcement Learning. *Electronics* 2023, 12, 3925.
- Hajian, A. Z. and Howe, R. D. (1997). Identification of the mechanical impedance at the human finger tip. *Journal of Biomechanical Engineering* 119, 109–114
11. Z. Guan, T. Yamamoto, Design of a reinforcement learning PID controller, in: 2020 International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–6,
12. V. van Veldhuizen, "Autotuning PID control using Actor-Critic Deep Reinforcement Learning," B.Sc. thesis, Univ. of Amsterdam, Amsterdam, The Netherlands, 2020.
13. Nathan P. Lawrence, Michael G. Forbes, Philip D. Loewen, Daniel G.

- McClement, Johan U. Backström, R. Bhushan Gopaluni, Deep reinforcement learning with shallow controllers: An experimental application to PID tuning, *Control Engineering Practice*, Volume 121, 2022, 105046, ISSN 0967-0661, Larsen, J. V., Overholt, D., and Moeslund, T. B. (2013). The actuated guitar: A platform enabling alternative interaction methods. In *SMAC stockholm music acoustics conference 2013 SMC sound and music computing conference 2013: Sound science, sound experience* (Logos Verlag Berlin), 235–238
14. U. Alejandro-Sanjines, A. Maisincho-Jivaja, V. Asanza, L. L. Lorente-Leyva, and D. H. Peluffo-Ordóñez, “Adaptive PI Controller Based on a Reinforcement Learning Algorithm for Speed Control of a DC Motor,” *Biomimetics*, vol. 8, no. 5, p. 434, Sep. 2023
 15. Subbareddy Chitta and Ramakalyan Ayyagari, "On the Nonlinear Control of a Class of Cruise Missiles," *International Journal of Control, Automation, and Systems*, vol. 23, no. 3, pp. 788-797, 2025.
 16. A. E. Taha, M. F. Rahmat, J. Mu'azu, and M. Musa, “Feedback linearization and sliding mode control design for quadrotor’s attitude and altitude,” 2019 IEEE 1st International Conference on Mechatronics, Automation and Cyber-Physical Computer System, Owerri, Nigeria, Apr. 2019, pp. 205–210.
 17. Fujimoto, S., van Hoof, H., & Meger, D. (2018). “Addressing Function Approximation Error in Actor-Critic Methods.” *Proceedings of the 35th International Conference on Machine Learning*, 2018.
 18. Gheorghe Bujgoi and Dorin Sendrescu, "Tuning of PID Controllers using Reinforcement Learning for Nonlinear Systems Control" (2024), doi: 10.20944/preprints202403.0914.v1
 19. Chowdhury, M. A., & Lu, Q. (2022). "A Novel Entropy-Maximizing TD3-Based Reinforcement Learning for Automatic PID Tuning." *IEEE Transactions on Systems, Man, and Cybernetics: Systems*. <https://doi.org/10.1109/TSMC.2022.3179646>
 20. R. Olfati-Saber, "Nonlinear Control of Underactuated Mechanical Systems with Application to Robotics and Aerospace Vehicles", Ph.D. dissertation, Dept. of Electrical Eng. and Computer Science, Massachusetts Institute of Technology, Cambridge, MA, 2001.
 21. P. Mukherjee and S. L. Waslander, “Direct adaptive feedback linearization for quadrotor control,” in *Proc. AIAA Guidance, Navigation and Control Conf.*, Minneapolis, MN, USA, Aug. 2012, pp. 1–10, doi:10.2514/6.2012-4917.
 22. V. P. Tran, M. A. Mabrok, S. G. Anavatti, I. R. Petersen, and M. A. Garratt, "Robust fuzzy Q-learning-based strictly negative imaginary tracking controllers for the uncertain quadrotor systems," **IEEE Transactions on Cybernetics**, vol. 53, no. 8, pp. 5306–5317, Aug. 2023.
 23. Zachary Dydek, Anuradha Annaswamy and Eugene Lavretsky in "Combined

Composite Adaptive Control of a Quadrotor UAV in the Presence of Actuator Uncertainty" AIAA Guidance, Navigation, and Control Conference, August 2010, Toronto, Ontario, Canada, <https://doi.org/10.2514/6.2010-7575>

24. Mohd Ariffanan Mohd Basri , A. R. Husain, and K. A. danapalasingam, ``Stabilization and Trajectory Tracking Control for Underactuated Quadrotor Helicopter Subject to Wind-Gust Disturbance," *Sadhana* Vol. 40, Part 5, August 2015, pp. 1531–1553 Indian Academy of Sciences