基於強化學習之無人機飛行控制

黃彥皓*、任境梧、董祐聖、吳柏慶、甯勇捷

摘要

隨著無人機技術與影像辨識精度的提升，無人機的應用已從軍事擴展至民生領域。本研究使用強化學習控制無人機，通過物理模擬驗證其性能，透過改進獎勵函數，提升無人機的飛行軌跡的準確性，實現更加穩定且高效的飛行控制。具體而言，研究模擬了直線前進與固定路徑的飛行，並使用 TRPO 與 PPO 方法，將目標點與模擬飛行的位置進行誤差計算，實驗結果顯示，驗證優化超參數與獎勵機制，TRPO 可以更優化飛行路徑，進一步提升飛行性能。

關鍵字：無人機、強化學習、獎勵函數、模擬飛行。

黃彥皓(通訊作者)，國立臺東大學綠能與資訊科技學系 助理教授，E-mail: yhhuang@nttu.edu.tw
任境梧，國立臺東大學綠能與資訊科技學系 學生，E-mail: 11022120@gm.nttu.edu.tw
董祐聖，國立臺東大學綠能與資訊科技學系 學生，E-mail: 11022126@gm.nttu.edu.tw
吳柏慶，國立臺東大學綠能與資訊科技學系 學生，E-mail: 11122121@gm.nttu.edu.tw
甯勇捷，國立臺東大學綠能與資訊科技學系 學生，E-mail: 11122130@gm.nttu.edu.tw

Drone Motion Control Based on Reinforcement Learning

Yan-Hao Huang *, Ching-Wu Jen , Yo-Sheng Tung , Bo-Cing Wu , Yong-Jie Ning

Abstract

The application of Unmanned Aerial Vehicles (UAVs) has significantly expanded from military operations to civilian sectors, driven by advancements in drone technology and improved image recognition precision. This study investigates the application of reinforcement learning for UAV control, with performance validation conducted through physical simulations. We propose an enhanced reward function to improve the accuracy of UAV flight trajectories, thereby achieving more stable and efficient flight control. Specifically, the research simulates flight scenarios involving straight-line and fixed-path trajectories. The Trust Region Policy Optimization (TRPO) and Proximal Policy Optimization (PPO) algorithms are employed to compute the error between target points and simulated flight positions. Experimental results demonstrate that optimizing hyperparameters and reward mechanisms, particularly with TRPO, leads to superior flight path optimization and enhanced overall flight performance.

Keywords: uav, reinforcement learning, reward function, simulated flight.

Yan-Hao Huang (Corresponding Author), Assistant Professor, Department of Green Energy and Information Technology, National Taitung University, E-mail: yhhuang@nttu.edu.tw

Ching-Wu Jen, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11022120@gm.nttu.edu.tw

Yo-Sheng Tung, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11022126@gm.nttu.edu.tw

Bo-Cing Wu, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11122121@gm.nttu.edu.tw

Yong-Jie Ning, Student, Department of Green Energy and Information Technology, National Taitung University, E-mail: 11122130@gm.nttu.edu.tw

壹、簡介

Unmanned Aerial Vehicles (UAVs): An Emerging Technology for Logistics (Rana et al., 2016)中顯示了無人機在多項領域中的應用。無人機技術已從早期軍事用途發展至多樣化的民生應用。在初期，無人機主要被用於娛樂和新聞媒體領域，如航空飛行拍攝和影片錄製，為相關產業提供了更廣闊的發展前景(Kaufman & Somaiya, 2013)。

在農業實際應用中，無人機被用於農產品的採收和害蟲防治(Mogili & Deepak, 2018) (Costa et al., 2012)0 ，透過無人機的巡視與監視，掌握農產品的生長、成熟度，部分特定水果可利用無人機進行自動採收，從而提升採收效率。在於害蟲防治方面，更可以通過無人機進行高精度的空中噴灑，相較於人工方式，無人機作業範圍更廣，並能減少藥物浪費，通過無人機均勻噴灑，更能夠確保每株植株都受到了一定的保護力。。

傳統無人機導航系統依賴於全球定位系統（GPS）或人工控制，在特定環境中面臨諸多挑戰。例如，在室內空間、城市峽谷或茂密森林等環境中，GPS 訊號易受干擾或完全無法接收，導致無人機無法正常運作。在這些環境中，無人機需克服訊號障礙，不僅要求其具備卓越的飛行操控能力，還需即時處理並避開障礙物，在缺乏 GPS 訊號的情況下保持穩定飛行。此外，連線異常等突發狀況亦會導致無人機無法正確執行任務，甚至造成嚴重的損壞與財產損失。

這些問題限制了無人機的操作範圍，阻礙了其在多種環境下的應用。因此，環境限制下的無人機控制技術是未來發展的重要方向。近年來，人工智慧技術的快速發展為無人機導航提供了全新的解決方案。先進的深度影像技術和機器視覺顯著提升了無人機的環境感知能力，使其能夠有效識別環境深度與障礙物(Li et al., 2019)。結合強化學習方法，無人機可在訓練過程中學習如何應對不同環境狀態，並通過獎勵函數的計算輸出最佳控制策略。

本研究旨在應用強化學習優化無人機的飛行方法，使其在複雜現實環境中實現更穩定、高效的飛行控制。我們著重於無人機的常見基礎動作，並設計最佳獎勵函數以提升強化學習效果。我們的模擬模型基於 Jacopo Panerati 等人開發的 gym-pybullet-drones 平台(Panerati et al., 2021)。該平台基於 Bullet 物理引擎，構建了專為四軸飛行器控制設計的 OpenAI Gym 環境。此整合平台提供簡易的介面，方便用戶將強化學習整合至虛擬飛行環境中。該物理引擎基於真實碰撞模型和空氣動力學理論構建物理仿真模型，使我們能夠模擬接近現實環境的狀況，在高擬真環境中推導四軸無人機的控制策略，實現精確的飛行控制。通過引入位置狀態並在仿真環境中反覆模擬測試與訓練，我們調整獎勵函數，結合多種連續型強化學習方法，使無人機高效完成預定的多項基礎飛行動作。我們設計了多種獎勵函數與終止方法，並呈現了這些模型在不同飛行動作中的測試結果，用於評估系統性能和穩定性。實驗結果表明，相較於其他連續型強化學習方法，TRPO 在我們設計的多項測試中表現更優。

貳、文獻探討

隨著深度強化學習（Deep Reinforcement Learning, DRL）在機器學習與智能控制領域的廣泛應用，許多研究致力於提升策略學習的穩定性與樣本效率。為克服傳統策略梯度方法在訓練過程中面臨的高變異性與不穩定性問題，許多研究提出了一系列改良型演算法，其中最具代表性的便是 Trust Region Policy Optimization（TRPO）與 Proximal Policy Optimization（PPO）(Liu et al., 2019)。

其中，TRPO 是由 Schulman 等人(Schulman et al., 2015) 提出的一種具理論基礎的強化學習演算法，旨在透過限制策略變動範圍，以實現策略效能的單調提升。該方法透過在每次策略更新時引入信賴域限制（如 Kullback-Leibler 散度），確保策略更新不會偏離原始策略過遠，進而避免策略崩潰的風險。實驗結果顯示，TRPO 在多項模擬任務中皆展現穩健的表現，如機器人游泳、跳躍、行走以及 Atari 遊戲的圖像輸入控制任務。即使進行多項近似簡化以利於實作，TRPO 仍能在大多數情況下維持策略效能的穩定提升，並且對超參數的敏感性較低，顯示其良好的泛化能力與實用性。

然而，早期 TRPO 方法無法有效利用離線策略數據，需要大量線上互動，這限制了其在許多現實世界應用中的性能。因此，後續有許多人不斷改進優化 TRPO。例如，Wenjia Meng 等人(Meng et al., 2021) 提出了一種離線策略 TRPO 方法，稱為 off-policy TRPO，該方法通過單調地優化一個使用線上和離線數據的替代目標函數（surrogate objective function），來保證策略的單調改進。實驗結果顯示，與其他使用離線數據的信任區域策略方法相比，該方法表現出更優越的性能。

而在 TRPO 的基礎上，Schulman 等人(Schulman et al., 2017) 更進一步提出了 PPO，旨在兼顧演算法的穩定性與實作效率。PPO 採用 Actor-Critic 架構，其核心思想為在策略更新過程中引入「裁剪機制」（Clipped Objective），以限制新舊策略間的變動幅度。此方法簡化了 TRPO 中需進行的二階優化與限制條件，進而降低實作難度並加快訓練速度。PPO 被廣泛應用於各種高維連續控制任務中，包含機器人操控、自主導航與無人機控制等領域，並已被證實能有效解決非線性與高維度控制問題。

Gu, Y 等人(Gu et al., 2021)更提出帶有策略反饋的 PPO（PPO-PF），他們具體的定義了策略網路和價值網路的損失函數，以確保其策略更新滿足信任區域的無偏估計。實驗結果表明，與 PPO 相比，PPO-PF 具有更快的收斂速度、更高的獎勵和更小的獎勵方差。

綜上所述，TRPO 與 PPO 作為現代強化學習中的代表性演算法，不僅在理論上具備收斂與穩定性的優勢，在實務應用上也被證實能有效解決非線性與高維度控制問題。因此，本研究主要使用 TRPO 與 PPO，通過設計任務、改進獎勵值並比較兩者的性能，驗證本研究提出的獎勵值設計之可行性。

參、研究方法

一、研究步驟

本研究採用 gym-pybullet-drones 模擬平台(Panerati et al., 2021)，因其具有簡單的接口，可便捷地獲取無人機的各項狀態並設定四軸的轉速。相較於其他無人機模擬軟體，如 Unreal Engine 4 上的 Airsim (Smyth et al., 2018)和 ROS (Li et al., 2019) ，gym-pybullet-drones 安裝更為容易，功能亦更為強大，且對硬體設備的要求較低。狀態蒐集方面，主要以無人機的三維位置為狀態空間，用於計算與獎勵函數中預定位置的偏移距離，以提升位置變化的穩定性。動作輸出方面，採用 PPO (Schulman et al., 2017) 和 TRPO (Schulman et al., 2015) 等連續動作空間的強化學習方法，通過不斷更新動作函數以及對應的獎勵和狀態，在達到預設的獎勵閾值後，將用於 Bullet 物理引擎的測試。測試結果將以圖表呈現無人機的三維位置、三維速率、加速度、角速率、角加速度以及四軸馬達轉速隨時間的變化。

本研究利用 PyBullet 模擬真實環境中的無人機空氣動力學和碰撞運動學，並使用 Gym 提供的強化學習環境標準 API，設置和運行訓練環境，通過 Gym PyBullet Drones 擴展包連接 Bullet 引擎。

二、模型訓練環境建置

模型訓練環境包括代理的允許動作、環境狀態以及獎勵或懲罰機制。代理作為學習者和決策者，通過策略和學習演算法的互動，學習在環境中完成特定任務 (Luo et al., 2024)。策略將環境狀態映射到動作的機率分佈，而學習演算法根據經驗迭代改進策略，以最大化長期累積獎勵 (Kalidas et al., 2023)。

本研究使用 Gym 提供的強化學習環境，並將無人機動作直接導入動作類別。動作空間採用連續動作，允許各軸馬達具有不同轉速，以實現三維空間的自由移動。狀態空間選擇無人機的三維位置，並通過計算不同時間的預期位置，將其設定為目標點。獎勵函數設計為目標點與無人機實際位置的差值，即無人機在接近目標點時將獲得更高的獎勵。

三、TRPO (Trust Region Policy Optimization)

TRPO (Trust Region Policy Optimization) (Schulman et al., 2015) 是一種基於策略迭代改良的強化學習方法。其透過限制策略函數的變動範圍，實現穩定的策略更新。TRPO 主要採用信賴域方法，約束策略函數的更新幅度，以維持新策略與舊策略目標函數的接近程度。TRPO 中的信賴域有助於防止過大的策略函數變動，從而提升訓練的穩定性。

除了傳統的最大化目標函數方法外，TRPO 在更新策略函數時，為避免偏離正確的回報獎勵函數，引入信賴域作為約束條件，防止更新幅度過大。最後，採用拉格朗日乘數法解決最大化目標函數與信賴域約束條件的問題。

信賴域方法是TRPO改良策略迭代的重要創新。其核心概念是在每次迭代中定義一個包含當前舊策略的信賴域。當新函數位於此區域內時，該新目標函數的模型被視為可接受。最終，在此信賴域內進行優化，以持續提高目標函數的獎勵值。信賴域的定義採用 KL 散度，其在 TRPO 中扮演限制策略函數變動的主要角色，控制策略更新的幅度，避免新策略過度偏離舊策略。

拉格朗日對偶則用於同時處理信賴域與梯度 g 的更新方向，將約束條件的 KL 散度海森矩陣納入目標函數，以求得同時滿足信賴域限制與最大化目標函數的解。

$$\theta_{k+1} = \theta_k + \sqrt{\frac{2\delta}{g^T H^{-1} g}} H^{-1} g \quad (1)$$

為減輕直接求解海森矩陣H^{-1}的計算負擔，TRPO 利用共軛梯度法近似求解海森矩陣 $x \approx H^{-1}g$，透過H_x迭代找出最接近解答的 x，取代計算海森矩陣的負擔。

四、PPO (Proximal Policy Optimization)

PPO(Proximal Policy Optimization) (Schulman et al., 2017) 是一種利用策略梯度法的強化學習方法。其主要目標是控制策略更新，提升訓練的穩定性，避免新策略與舊策略的目標函數變化過大，穩定性能波動。同時，保持穩定性、性能和簡化演算法設計。

PPO 方法透過與環境的互動收集數據，並使用隨機梯度更新目標函數。與傳統策略函數利用梯度下降法收斂類似，PPO 的目標是最大化累積回饋。然而，PPO 引入了新的剪裁技術，以控制策略更新的範圍。PPO 採用一階導數最佳化方法，解決 TRPO 二次近似的複雜度，並透過設計剪裁機制防止過大的更新。當新目標函數與舊目標函數的模型適合程度差距達到設定的剪裁值時，禁止新策略對舊策略的過度偏離。

PPO 主要使用裁剪函數限制更新幅度。透過裁剪函數 clip，將新舊策略機率比率限制在一定範圍內([1 - ϵ, 1 + ϵ]，ϵ 為常數)平衡更新步長，防止策略更新中過大的幅度變化導致學習過程的不穩定。

其更新策略網路的方法是透過最大化裁剪後的目標函數進行更新：

$$LCLIP(\theta) = Et[\min(rt(\theta)At, clip(rt(\theta), 1-\epsilon, 1+\epsilon)At)] \quad (2)$$

$rt(\theta)$為新舊策略獎勵比，At 為優勢估計，ϵ 為裁剪範圍。透過新舊策略獎勵比與優勢估計，兩參數選擇隨機動作中，相較平均行動更符合所需對應狀態的動作機率進行增強或減弱。最後，加入 clip 函數的剪裁機制，使新策略與舊策略的變動幅度不會過大。

其中，優勢估計(Advantage Estimation)：

$$A(s,a) = Q(s,a) - V(s) \quad (3)$$

透過同一狀態下指派隨機動作得到的行動 Q 值與使用當前目標函數的動作，兩者的回報相減。若為正值，表示在狀態 s 下選擇行動 a 相較於其他可能動作的平均值（即當前策略目標函數）表現更佳，應增強 a 動作的選擇概率。總結來說，優勢估計用於評估特定狀態下，行動相對於平均策略動作的優劣，以增強或減少該動作出現的機率。

肆、結果分析與探討

一、直線前進的飛行動作

本研究設計一目標點，其座標為 [變動的x值, 0, 1]，其中變動的x值為時間 t 與設定速度的乘積。透過計算當前狀態與目標點間的偏移量，並以此偏移量作為回報函數的輸入。獎勵值計算公式為若偏移量平方小於 2，則獎勵值為 2 − (偏移量平方)，若偏移量平方大於等於 2，則獎勵值為 0。

$$\Delta = \sqrt{\left(x_{current} - x_{target}\right)^2 + \left(y_{current} - y_{target}\right)^2 + \left(z_{current} - z_{target}\right)^2} \qquad (4)$$

偏移量 Δ 根據各軸的當前位置與目標點之間的差值計算：。

$$R = \max(0, 2 - \Delta^2) \qquad (5)$$

若偏移量 Δ 大於 $\sqrt{2}$，則獎勵為 0。訓練步數 (time_steps) 設定為 1e7，每 100 次訓練記錄回報函數，每 1000 次輸出結果。

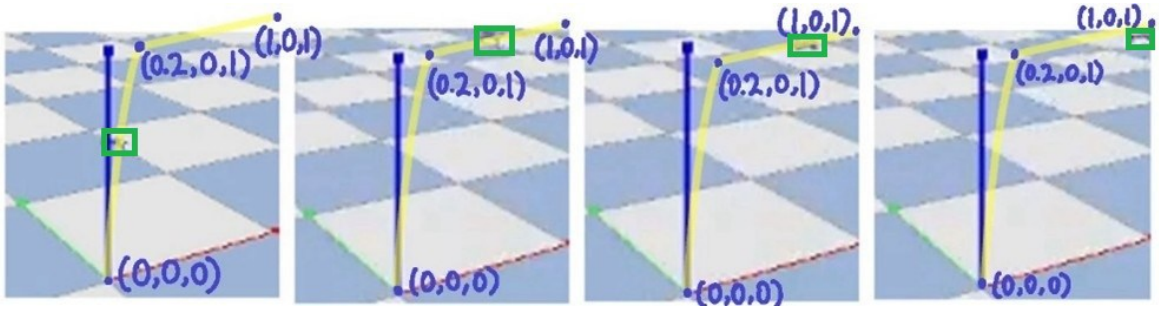

圖 1 在 bullet 中模擬飛行圖

圖 1 以黃線呈現無人機在 Bullet 物理模擬環境中的飛行軌跡並展示了模擬中，隨時間變化之無人機的確切位置(以綠色方框標示)。如圖所示，實驗目標為使無人機以 0.1 單位/秒的速度沿 x 軸正向移動，並同時達到 z = 1 的目標高度。因此，在 z < 1 的階段，無人機同時向 x 軸和 z 軸正向移動。一旦達到 z = 1 附近，無人機則主要沿 x 軸方向前進，並維持 z 軸的穩定性。

實驗初始位置為 [0, 0, 0]，座標系定義如下：x 軸（紅線）為右方正向、y 軸（綠線）為前方正向、z 軸（藍線）為上方正向。

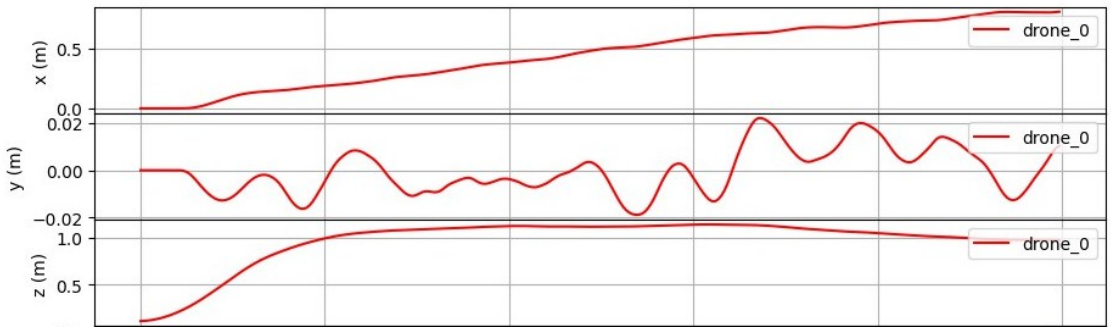

圖 2 TRPO 直線飛行的軌跡圖

在此次實驗中，我們對直線飛行動作進行了測試，評估無人機在進行直線移動時的表現。目標是驗證無人機是否能夠準確地沿著 x 軸正向飛行，並且在 y 軸和 z 軸上的控制也能夠達到預期的穩定效果。

圖 2 顯示了應用 TRPO 演算法後，相對於出發點的三軸位置隨時間之變化。從結果可見，無人機能穩定沿 x 軸前進，與目標點的偏移在可接受範圍內。在 y 軸方面，位移波動小，顯示控制穩定；而 z 軸高度在達到目標後亦能維持穩定，符合預期。

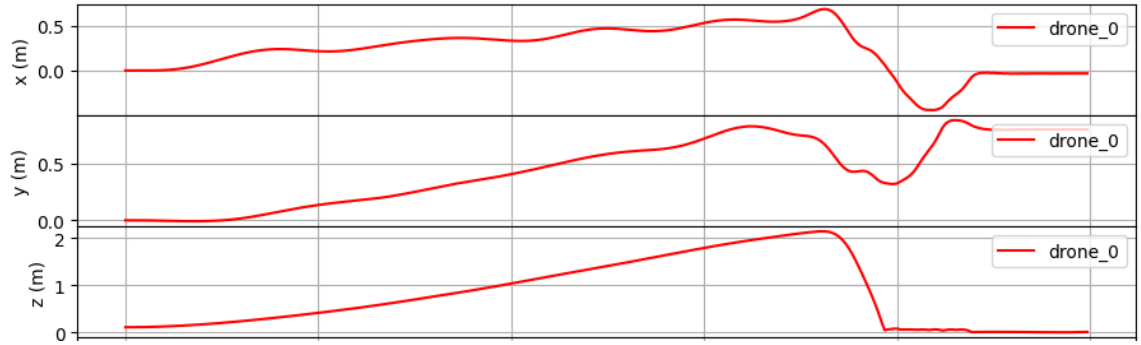

圖 3 應用 PPO 直線飛行的軌跡圖

圖 3 則為 PPO 演算法應用於相同任務的表現結果。無人機在前 7 秒內能沿 x 軸正確前進，但其後出現明顯偏移。在 y 軸方向控制表現不佳，出現劇烈擺動，顯示穩定性不足；在 z 軸方面，雖一開始能上升至目標高度，但無法長時間穩定維持，整體飛行表現不如 TRPO。

二、正方形飛行軌跡

本研究延續直線前進飛行的設計概念，規劃一組固定的正方形飛行軌跡。目標點座標為[變動的 x 值, 變動的 y 值,1]，以此方法來模擬正方形飛行的軌跡。接下來，我們計算目前狀態與目標點之間的偏移量，並以此偏移量來計算回報函數。獎勵值的計算公式為 2－（偏移量的平方），如果此值小於 0，則獎勵值為 0。

以下為設定正方形軌跡飛行的方法：將飛行分為四個階段，模擬正方形四條邊的移動：

(一)在方形路徑的第一段，目標點的 (x ,y)值從 (0,0) 變化到(1,0)。
(二)在第二段，目標點的 (x ,y)值從 (1,0) 變化到(1,1)。
(三)在第三段，(x ,y)值從 (1,1) 變化到(0,1)。
(四)在第四段，目標點的(x ,y)值從 (0,1) 變化到(0,0)。

如**圖 4** 所示，黃色軌跡代表無人機的實際飛行路徑並展示了模擬中，隨時間變化之無人機的確切位置(以綠色方框標示)，無人機地面起飛，先垂直上升至[1,0,1] ，再沿著邊長為 1 的正方形路徑飛行。

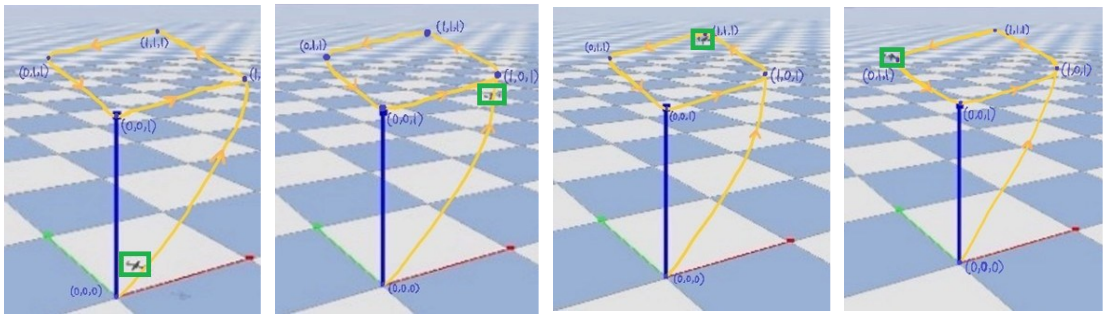

圖 4 在 bullet 中沿正方形飛行圖

實驗初始位置為 [0, 0, 0]，座標系定義如下：x 軸（紅線）為右方正向、y 軸（綠線）為前方正向、z 軸（藍線）為上方正向。

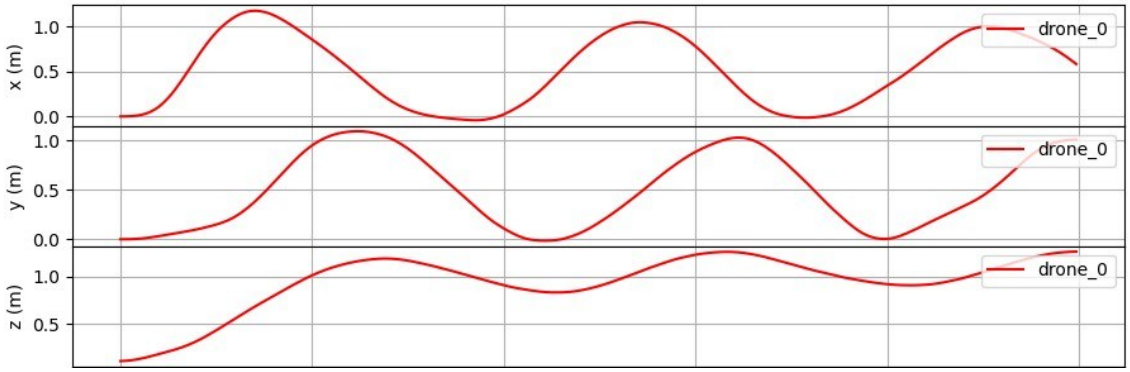

圖 5 TRPO 繞行正方形航線的飛行軌跡圖

由於 TRPO 在直線飛行任務中展現穩定表現，因此我們進一步測試其在更複雜的正方形飛行任務中的適應性。此實驗旨在觀察 TRPO 是否能於多方向變化中維持有效控制。

圖 5 顯示 TRPO 演算法控制下的飛行結果。無人機在 x 軸變化的第一與第三階段皆能於一秒內精準到達目標點；而在第二與第四階段（需維持 x 軸，改變 y 軸）中，也成功保持在原定 x 軸位置，穩定完成目標點遷移。

y 軸飛行表現亦類似，於需移動的階段能平順前進，於需維持的階段則穩定停留，呈現良好飛行能力。z 軸則於第 2 秒達到目標高度 z=1 並於後續時間內穩定保持，僅出現輕微抖動，整體控制表現穩定。

表 1 正方形航線的偏移值

時間	實際位置	目標位置	偏移量	偏移距離平方
2	[0.89,0.95,1.01]	[1,1,1]	[0.11,0.05,0.01]	0.0147
2.5	[0.46,1.09,1.15]	[0.5,1,1]	[0.04,0.09,0.15]	0.0322
3	[0.08,0.92,1.15]	[0,1,1]	[0.08,0.08,0.15]	0.353
3.5	[-0.03,0.50,1.05]	[0,0.5,1]	[0.03,0,0.05]	0.0034
4	[0.03,0.11,0.91]	[0,0,1]	[0.03,0,0.05]	0.0211
4.5	[0.42,0.003,0.83]	[0.5,0,1]	[0.08,0.003,0.17]	0.035309
5	[0.89,0.17,0.90]	[1,0,1]	[0.11,0.17,0.1]	0.051
5.5	[1.02,0.53,1.08]	[1,0.5,1]	[0.02,0.03,0.08]	0.0077

$MSE = \frac{1}{N}\sum_{i=1}^{N}(\text{實際位置} - \text{目標位置})^2$=0.025089

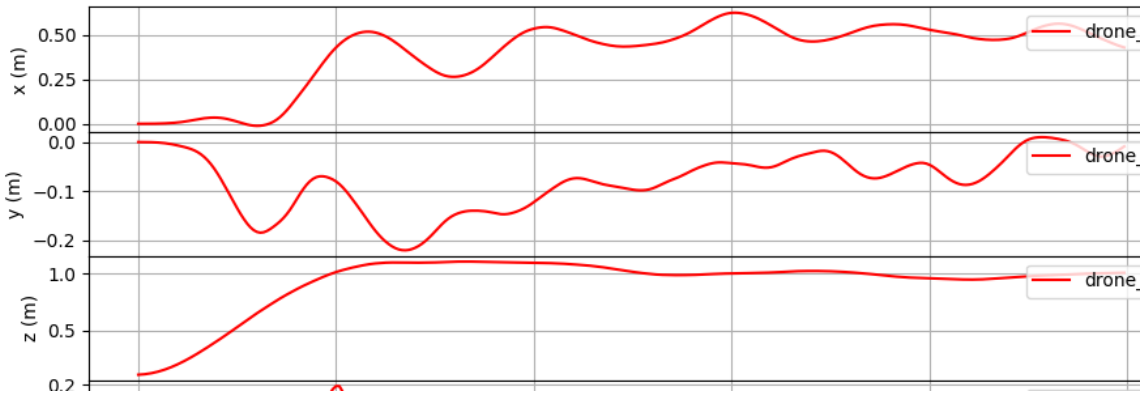

圖 6 應用 PPO 繞行正方形航線的飛行軌跡圖

從**圖 6** 中，我們可以看出，PPO 繞行正方形航線時，於第一與第三階段（x 軸移動）中，無法準確飛抵目標點；於第二與第四階段（需維持 x 軸，改變 y 軸）中，亦出現明顯偏移，未能維持原位。y 軸飛行亦顯示不穩定，前進與停留階段均無法穩定控制，飛行方向與幅度變動明顯。惟 z 軸表現相對穩定，無人機在 2 秒時，提升至指定點 z=1 上，並持續維持在此目標點中，在後續的 8 秒行動中，沒有出現太多飛離目標點的動作產出。

三、討論

從**圖 2** 可觀察到，TRPO 在直線飛行任務中的 x 軸表現達到預期目標，顯示在速度驅動控制下，無人機能夠沿直線路徑穩定前進，並準確抵達目標位置。這項結果驗證了我們所設計之直線飛行控制策略的有效性。雖然 y 軸上仍存在些微偏移，但偏差幅度相對有限，顯示 y 軸方向的控制亦具備一定準確度。此偏移可能與飛控系統中微小誤差、外部環境干擾或演算法設計參數相關。未來可針對 y 軸控制機制進一步優化，進一步分析控制信號的時序圖，並探討 TRPO 如何通過調整控制律（例如 PID 控制）來減小誤差，以降低偏移程度。z 軸表現則展現出良好的穩定性。無人機在達到預定高度後，能有效保持高度變化於最小範圍內，顯示 z 軸控制策略已趨成熟，具備維持飛行高度的能力。TRPO 的信賴域優化方法有效避免了策略崩潰，提高了學習的穩定性。

圖 3 顯示，PPO 在直線飛行任務中，於前 7 秒內能在 x 軸方向精確移動，顯示其具備學習正確移動方向的能力。然而在最後 3 秒中，x 軸出現明顯偏移，可能與 PPO 所採用的裁剪機制（clipping）無法有效限制策略更新幅度有關，導致策略不穩定。在 y 軸上，飛行表現不佳，可能是獎勵函數設計偏向強化 x 軸動作，使得 y 軸移動行為未被充分鼓勵。另一可能原因為觀測狀態空間資訊不足，例如未納入偏航角數據，導致 y 軸方向控制準確度降低。未來，我們將嘗試增加 y 軸移動的獎勵項，並擴充觀測狀態空間資訊。z 軸方面則表現出高度不穩定的現象，雖於初期能逐步貼近目標高度，後期仍快速偏離。與 y 軸問題相似，可能源於獎勵設計未能有效懲罰偏離行為，未來應加入更多懲罰項以限制垂直方向誤差並調整獎勵函數以提高 z 軸穩定性，相較 TRPO，PPO 明顯更不適用於本研究所提之方法。

圖 5 顯示，TRPO 在正方形繞行任務中於三個軸向皆表現穩定，x 與 y 軸無論在移動或靜止階段，皆能依循正確策略完成飛行控制，z 軸則穩定維持在目標高度 z=1。本結果說明 TRPO 能夠有效學習並執行多方向動作序列，展現其於複雜飛行任務中的良好適應能力與穩定性。

相對地，**圖 6** 顯示 PPO 在此任務中表現不如預期。x 與 y 軸均無法完成正確的繞行動作，甚至於 y 軸出現反向移動現象，顯示其對多方向控制的適應性不足。唯一表現穩定的是 z 軸，無人機能維持在目標高度 z=1 附近，除了輕微抖動外，未出現明顯高度偏移。本結果說明 PPO 無法有效學習並執行多方向動作序列，其於複雜飛行任務中的適應能力與穩定性極差。

綜合而言，TRPO 在本研究中於直線與正方形繞行任務中皆展現出良好的學習能力與控制表現，而 PPO 雖能初步學習直線飛行方向，於複雜路徑控制上仍顯不足，顯示未來仍需針對策略更新穩定性與獎勵設計進行優化。

伍、結論

本研究分別採用兩種處理連續動作的強化學習方法(TRPO、PPO) 對無人機飛行控制模型進行訓練與測試，並分析其於不同飛行任務中的表現。實驗結果顯示，TRPO 在直線與正方形軌跡的飛行任務中展現出更為穩定且高效的學習能力，而 PPO 雖在初期 x 軸直線飛行表現尚可，但整體飛行穩定性與策略精確度皆顯不足，尤其在複雜軌跡控制上表現欠佳。

我們推測，PPO 容易陷入區域最小值的情況，可能源於其策略裁剪（clipping）機制雖能控制策略更新幅度，但在較為嚴格的控制場景中，可能抑制了策略的有效探索與學習。而 TRPO 所採用的 KL 散度限制，能夠更有效地控制新舊策略間的變動幅度，使策略更新更加穩定，避免因更新步幅過大而導致策略崩潰。這一穩定性正是 TRPO 在本研究中表現優異的關鍵原因。

PPO 優勢在於其演算法結構相對簡單，運算資源消耗較低，但為了提升其在本模型中的適應性，未來可嘗試調整裁剪範圍，或透過增加探索範圍與動作雜訊，進一步提升策略學習能力與表現。

綜合實驗結果，TRPO 於本研究中的無人機飛行控制任務中，無論是在策略收斂速度、飛行穩定性，或任務完成準確度方面，均優於 PPO，顯示其更適合應用於本模型的飛行控制訓練。未來研究將朝向更複雜的飛行軌跡設計發展，除了直線與正方形路徑外，也將嘗試螺旋、曲線、動態路徑等多變環境下的控制挑戰。同時，將針對現有演算法與獎勵函數進行進一步優化，並引入更多動態參數，如轉速與傾角控制模擬，期望提升無人機於各種複雜環境中維持機體穩定性與完成指定任務的能力。

引用文獻

Rana, K., Praharaj, S., & Nanda, T. (2016). Unmanned aerial vehicles (UAVs): An emerging technology for logistics. *International Journal of Business and Management Invention, 5*(5), 86-92.

Kaufman, L., & Somaiya, R. (2013). Drones offer journalists a wider view. *The New York Times*.

Mogili, U. R., & Deepak, B. (2018). Review on application of drone systems in precision agriculture. *Procedia Computer Science, 133*, 502–509.

Costa, F. G., Ueyama, J., Braun, T., Pessin, G., Osório, F. S., & Vargas, P. A. (2012). The use of unmanned aerial vehicles and wireless sensor network in agricultural applications. Paper presented at the *2012 IEEE International Geoscience and Remote Sensing Symposium,* 5045–5048.

Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., & Freeman, W. T. (2019). Learning the depths of moving people by watching frozen people. Paper presented at the *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,* 4521–4530.

Panerati, J., Zheng, H., Zhou, S., Xu, J., Prorok, A., & Schoellig, A. P. (2021). Learning to fly—a gym environment with pybullet physics for reinforcement learning of multi-agent quadcopter control. Paper presented at the *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),* 7512–7519.

Liu, B., Cai, Q., Yang, Z., & Wang, Z. (2019). Neural trust region/proximal policy optimization attains globally optimal policy. *Advances in Neural Information Processing Systems, 32*.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. Paper presented at the *International Conference on Machine Learning,* 1889–1897.

Meng, W., Zheng, Q., Shi, Y., & Pan, G. (2021). An off-policy trust region policy optimization method with monotonic improvement guarantee for deep reinforcement learning. *IEEE Transactions on Neural Networks and Learning Systems, 33*(5), 2223–2235.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv Preprint arXiv:1707.06347*.

Gu, Y., Cheng, Y., Chen, C. P., & Wang, X. (2021). Proximal policy optimization with policy feedback. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 52*(7), 4600–4610.

Smyth, D. L., Glavin, F. G., & Madden, M. G. (2018). Using a game engine to simulate critical incidents and data collection by autonomous drones. Paper presented at the *2018 IEEE Games, Entertainment, Media Conference (GEM),* 1–9.

Luo, F., Xu, T., Lai, H., Chen, X., Zhang, W., & Yu, Y. (2024). A survey on model-based reinforcement learning. *Science China Information Sciences, 67*(2), 121101.

Kalidas, A. P., Joshua, C. J., Md, A. Q., Basheer, S., Mohan, S., & Sakri, S. (2023). Deep reinforcement learning for vision-based navigation of UAVs in avoiding stationary and mobile obstacles. *Drones, 7*(4), 245.