

2020

How to train your quadrotor: a framework for consistently smooth and responsive flight control via reinforcement learning

Siddharth Mysore, Bassel Mabsout, Kate Saenko, Renato Mancuso. 2020. "How to Train your Quadrotor: A Framework for Consistently Smooth and Responsive Flight Control via Reinforcement Learning.." CoRR, Volume abs/2012.06656, <https://arxiv.org/abs/2012.06656> <https://hdl.handle.net/2144/43028>

"Downloaded from OpenBU. Boston University's institutional repository."

How to Train your Quadrotor: A Framework for Consistently Smooth and Responsive Flight Control via Reinforcement Learning

SIDDHARTH MYSORE* and BASSEL MABSOUT*, Boston University

KATE SAENKO, Boston University & MIT-IBM Watson AI Lab

RENATO MANCUSO, Boston University

We focus on the problem of reliably training Reinforcement Learning (RL) models (agents) for stable low-level control in embedded systems and test our methods on a high-performance, custom-built quadrotor platform. A common but often under-studied problem in developing RL agents for continuous control is that the control policies developed are not always smooth. This lack of smoothness can be a major problem when learning controllers as it can result in control instability and hardware failure.

Issues of noisy control are further accentuated when training RL agents in simulation due to simulators ultimately being imperfect representations of reality — what is known as the *reality gap*. To combat issues of instability in RL agents, we propose a systematic framework, ‘REinforcement-based transferable Agents through Learning’ (RE+AL), for designing simulated training environments which preserve the quality of trained agents when transferred to real platforms. RE+AL is an evolution of the Neuroflight infrastructure detailed in technical reports prepared by members of our research group. Neuroflight is a state-of-the-art framework for training RL agents for low-level attitude control. RE+AL improves and completes Neuroflight by solving a number of important limitations that hindered the deployment of Neuroflight to real hardware. We benchmark RE+AL on the NF1 racing quadrotor developed as part of Neuroflight. We demonstrate that RE+AL significantly mitigates the previously observed issues of smoothness in RL agents. Additionally, RE+AL is shown to consistently train agents that are flight-capable and with minimal degradation in controller quality upon transfer. RE+AL agents also learn to perform better than a tuned PID controller, with better tracking errors, smoother control and reduced power consumption. To the best of our knowledge, RE+AL agents are the first RL-based controllers trained in simulation to outperform a well-tuned PID controller on a real-world controls problem that is solvable with classical control.

CCS Concepts: • **Computing methodologies** → **Reinforcement learning; Control methods**; • **Computer systems organization** → *Embedded systems; Robotics*.

Additional Key Words and Phrases: neural networks, continuous control, quadrotor

1 INTRODUCTION

Low-level control of physical systems such as unmanned aerial vehicles (UAVs) is usually implemented with simple PID controllers. Such controllers are straightforward to use, but cannot learn from experience or adapt to the environment. Recently, Reinforcement Learning (RL) has been applied to low-level control in embedded systems, promising a general way to train agents to map control commands to motor outputs. Due to safety concerns, it is not always possible or advisable to train RL agents on the platforms they are meant to be deployed on, therefore training is done in simulation. However, simulations are not perfect representations of reality — this is termed as the “reality gap”. The reality gap can often result in aberrant behavior when models are transferred from simulators to their real-world counterparts.

*Both authors contributed equally to this research.



Fig. 1. The NF1 platform and the pilot’s view on the heads up display while flying over the treeline with with our RE+AL controller on-board.

A prevalent issue with learning and deploying low-level controllers trained in simulation is that the trained agents, despite strong performance in simulation, often present with a number of stability issues resulting from over-actuation (and consequent efforts to compensate) [14, 19, 21]. Despite their efforts, prior works [11, 12, 15, 19, 21] found this issue difficult to mitigate reliably.

Members of our research-group previously developed Neuroflight, described in the technical report [14] and described in greater detail in [13]. Neuroflight is a framework for training and deploying learned controllers on UAVs based on the GymFC [15] training environment. While Neuroflight demonstrates the viability of learning low-level control for high-performance quadrotors, the authors found that the trained agents (i) did not consistently transfer with acceptable performance in real flight, despite exhibiting good performance in simulation – with only one out of dozens of agents trained in simulation proving controllable on the real drone – and (ii) were often unstable and non-smooth, resulting in increased power consumption and excessive strain and wear on the motors. By analyzing their methods and results, we hypothesize that the root of the issue is the reward structure – the basis of optimization in RL – which is highly tuned for good performance in simulation, but fails to capture the problem in a way that transfers to real-world observations and dynamics, likely due to the inherent limitations to the fidelity of simulation.

Building on the groundwork laid by Neuroflight, our work presents a systematic approach for effectively and consistently developing RL agents that can handle high-performance low-level control needs while achieving smoother outputs. We achieve this through a two-fold approach:

- (i) We develop a new reward structure that is designed to be more intuitive and allows for faster and more repeatable training of agents that transfer reliably to a real drone.
- (ii) By carefully studying the behavior of the firmware in receiving and responding to control signals, we construct a more fitting state representation for our problem and also generate training signals that better approximate the RC-commands that might come from a pilot.

We present a full-stack implementation where we adapt the simulation environment, the training, and compilation pipeline to produce flight-capable control firmware. All agents trained with our method offer a significant reduction in high-frequency oscillations in the motor-control signals when deployed on the real drone (when compared to the baseline method), with the peak oscillatory frequency of approximately 130Hz, down from the previous 330Hz, and at significantly lower amplitudes. Nonetheless, they maintain comparably low tracking errors, with an average error of 4.2 deg/s. Our training structure also yields a 10x speedup in the time required to train transferable agents, which, on our machines, reduced training time from nearly 9 hours to under 50 minutes.

The remainder of the paper is organized as follows. In Section 2, we provide a basic overview of reinforcement learning, covering the core problem structure and optimization framework. Section 3 reviews related work which also tackles the problem of continuous control and details the learning pipeline for the Neuroflight framework, which we use as our primary baseline. Section 4 describes our proposed approach, RE+AL, for achieving smoother and more transferable RL agents. Section 5 presents performance metrics for agents trained with RE+AL and compares them with the original Neuroflight method on both simulated and real-world flights. Finally, we conclude with Section 6.

2 BACKGROUND

Reinforcement Learning represents a class of machine learning algorithms that attempt to develop an optimal states-to-actions mapping intended to maximize a numerical reward. The sequential decision making problem in RL is typically formulated as a Markov Decision Process (MDP), where actions influence both the immediate and eventual rewards. Additionally, state transitions are assumed to obey the Markov property – i.e. the response of the system at time t to any state and action should only be reliant on the current state s_t and action a_t , which can be formalized as:

$$P(s_{t+1}|s_t, a_t, s_{t-1}, a_{t-1}, \dots, s_0, a_0) = P(s_{t+1}|s_t, a_t)$$

Extending dynamical systems theory, the problem of RL can be formalized as “the optimal control of incompletely-known Markov decision processes” [30].

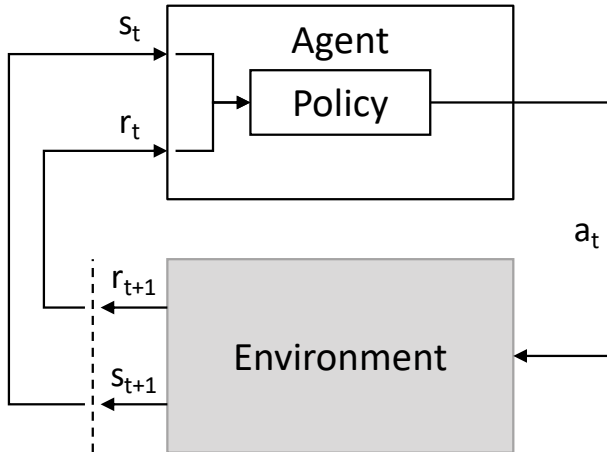


Fig. 2. General structure of the reinforcement learning interaction loop.

There are four key elements of RL problem formulations: the (i) agent, (ii) environment, (iii) reward signal, and (iv) policy, as shown in Fig. 2. The learner (or agent) in RL settings are tasked with learning what to do in order to best achieve their goals through interaction with the training environment and developing an understanding of which actions to take in any given state. At any given time instance t , the agent interacting with the environment is in state $s_t \in S$, and can take an action $a_t \in A(s_t)$ (an action in the valid action space A of s_t), for which a numerical reward, r_t , is received. Typically $A(s_i) = A(s_{j \neq i})$, however, for some problems this may not always hold – for the purposes of this paper, we assume $A(s) = A \forall s \in S$, i.e. that the action space is constant. The reward signal is a numerical performance measure that indicates the quality of an action in the environment, given the current state. The signal serves to penalize objectively bad actions while

rewarding good ones to reinforce good behavior in an agent - hence ‘reinforcement learning’. The agent’s goal is to maximize its expected return — the total expected reward that it receives:

$$R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$$

where $\gamma \in [0, 1]$ is a discount factor introduced to control the relative weight of current and future rewards in the estimation of expected total reward received. The policy, π , is defined as a mapping of states to probability of selecting an action given the state - i.e. $\pi(s) = \{P(a|s) \mid a \in A\}$.

RL seeks to develop a policy π_* that maximizes the value function of a state, $v_\pi(s)$ or equivalently the action-value function $q_\pi(s, a)$ for agents acting under a policy π :

$$\begin{aligned} \pi_* &= \operatorname{argmax}_\pi \sum_s v_\pi(s) \\ &= \operatorname{argmax}_\pi \sum_s q_\pi(s, a^*) \quad \text{s.t. } a^* = \operatorname{argmax}_a q_\pi(s, a) \end{aligned}$$

where

$$\begin{aligned} v_\pi(s) &= \mathbb{E}_\pi [R_t | s_t = s] \\ q_\pi(s, a) &= \mathbb{E}_\pi [R_t | s_t = s, a_t = a] \end{aligned}$$

It is important to note however that, while the RL agents seek to maximize the total expected reward, this is equivalently achieved (and easier optimized) when focusing on maximizing the total expected reward of any individual action, and thus the focus of RL agents is to learn to take the best action possible for any given input state.

A number of tools have been developed in order to tackle the optimal policy problem. Policies may be optimized by on- or off-policy techniques. On-policy RL utilizes the same policies for exploring the state and action space of an environment as they do in optimization. Conversely off-policy RL utilize a derivative of the target policy (the policy being optimized) for exploration, thus allowing for the use of distributed deployment and batch optimization techniques. For low-dimensional state spaces with limited actions, the optimal policy can be solved using Monte Carlo methods or approximately solved (in a less data-intensive way) using temporal difference (TD) learning techniques such as SARSA [23], Q-learning [33], or double Q learning [10], with TD-learning techniques having been demonstrated as being optimal in the limit [30]. Additionally, policy gradient methods have been developed, such as REINFORCE [34], which represent the policy in a parametric form and utilizes gradient ascent techniques to improved the policy incrementally. Policy gradient techniques form the basis of many of the deep reinforcement learning techniques which have gained popularity in recent years [20, 25, 26, 31]. One of the key issues that deep RL helps addressing, which classical RL was ill-equipped for, is the application of RL techniques to continuous control.

Deep Reinforcement Learning. Applied broadly, policy gradient methods define policies π_θ by a set of parameters, θ , and apply iterative gradient ascent to maximize policy performance. Given $J(\theta) \propto \tau_\theta$, where J is a performance measure on θ — which in RL is typically proportional of the estimated temporal-difference τ_θ — we have:

$$\begin{aligned} \tau_\theta &= R_t - v_{\pi_\theta}(s_t) \\ &\approx r_t + \gamma \max_{a'} q_{\pi_\theta}(s_{t+1}, a') - \max_{a''} q_{\pi_\theta}(s_t, a''). \end{aligned}$$

Parameters in policy gradient algorithms are updated proportionally to the gradient on the policy performance with respect to parameters, $\nabla J(\theta)$:

$$\theta_{t+1} = \theta_t + \alpha \nabla J(\theta_t)$$

where α modulates the rate of change of θ .

Deep reinforcement learning methods extend classical policy gradient methods by representing RL policies with artificial neural networks. The networks are in turn parameterized by their internal network weights and biases. This innovation allowed RL to capitalize upon the significant representative capacity of neural networks for modeling arbitrary functions to tackle more challenging tasks, with works in the field demonstrating the ability to train agents to play video games, solve various controls problems and even besting humans in complex games such as chess and go [28, 29], and even achieve ‘master’ level play in games like Starcraft [32].

While initial developments in deep RL were built to work with discrete action spaces, extensions were made to deep RL architectures to allow for the development of RL algorithms for learning action policies in continuous state and action spaces. This change to the policy network structure allowed for RL techniques to be developed and deployed on continuous control tasks and has been demonstrated to have a wide range of applications – most notably a number of complex controls tasks where developing a controller following classic controls theory is highly non-trivial [8, 18, 19, 25, 26].

3 PRIOR WORK – LEARNING QUADROTOR CONTROL

One significant issue with deep learning techniques, or indeed machine learning techniques, is that the learned models could over-fit to the training domain. This has been observed to be a significant problem in deep RL, with agents often memorizing correct behavior in a given state, as opposed to learning a systematic model that would allow it to generalize to novel inputs [24, 35, 36]. In problems of control, this leads to noticeable aberrant behavior when agents are transferred between domains [14, 15, 19, 21], which can even cause mechanical failures [19, 21].

A number of recent works address the specific problem of quadrotor flight control with controllers trained with RL. Molchanov et al. [21] train neural networks policies for low-level attitude control using the Proximal Policy Optimization (PPO) [26] algorithm. In their work, they define the RL cost function, i.e. their reward signal, as a weighted sum over position, velocity, and angular velocity errors, with additional penalties on the drone’s acceleration and orientation. They investigate the sensitivity of their RL agents as a function of the weight parameters and provide a performance metric over tracking errors and oscillations in the quadrotor’s frame. They note that, while some agents allowed for stable flight, others might result in issues ranging from visible oscillations to an inability to take off. They do not however investigate the underlying causes of the oscillations, nor do they report on control signal behavior during tests, and their effect on the motors. Related work by Hwangbo et al. [11], also presents a full pipeline for training RL agents for quadrotor flight control. Similar to Molchanov et al., they construct their reward signal additively over state tracking errors. Unlike Molchanov et al., to achieve stable flight, they employ a varied exploration strategy for exploring the state-space. They also make stronger assumptions on prior knowledge of the quadrotor dynamics and augment training with a re-tuned Proportional Differential (PD) controller. While the authors comment on the stability or lack thereof of their quadrotor in various flight conditions, they do not offer a quantitative analysis of the observed phenomena. Our work builds upon the preliminary results we obtained in the Neuroflight framework, for which a technical report is available in Koch et al. [14]. Neuroflight is also described in greater detail in [13]. The tools developed through the Neuroflight work are summarized in Sections 3.1 and 3.2.

3.1 Neuroflight

The Neuroflight framework introduced by Koch et al. [14] was designed to be a full-stack solution for training, testing and compiling neural-network-based flight controllers into firmware that can be flashed on off-the-shelf micro-controllers (see Fig. 3 for an overview). The training environment, called GymFC and introduced in [15], is a simulator built using Gazebo [16], to provide an OpenAI Gym-compatible [6] interface for training RL agents. GymFC allows simulated controllers to interact with simulated quadrotors safely in a virtual environment, making it a good training and testbed for our agents. With their original work, Koch et al. provide a feasibility study on the use of controllers trained using deep RL techniques for deployment on real drones. They benchmarked a number of the then state-of-the-art deep RL algorithms on the task of low-level drone attitude control and, like Molchanov et al. [21], were able to successfully demonstrate the viability of the PPO algorithm on this task. They further establish that agents trained with PPO were able to attain faster convergence to rotational angular rates taken from real flight tests than a tuned PID controller both in the real world and in simulation.

3.2 Neuroflight tuning and improvements

As identified by Koch et al. [14, 15], one of the key issues with the control policies learned by the RL agents with the Neuroflight framework is the lack of smoothness and stability in the control output. These instabilities were previously observed both in simulated tests as well as when transferred to the quadrotors for real-world flight. In an extension to their prior work [13], Koch et al. further developed the simulation and RL training sub-systems in an effort to improve set-point tracking and flight stability. Their work focused primarily on improving model fidelity for the quadrotor within the simulator, and tuning the neural network construction and reward-signal for the deep RL algorithm employed. We briefly discuss these improvements to help contextualize our work, which aims to address some of the remaining key issues present in the Neuroflight policy training framework.

While initial versions of the simulated quadrotor training environment were built with a generic off-the-shelf quadrotor model, Koch et al. developed a higher-fidelity model of the target quadrotor platform. By building a custom Gazebo model for the simulator which better represented the shape, weight, dimensions and thrust forces of the NF1 quadrotor, Koch et al. are able to more accurately model both the quadrotor construction and the motors' properties. To better simulate the physical forces acting on the quadrotor, Koch et al. also changed the internal physics engine in Gazebo to work with DART [17] instead of the standard ODE [2] engine typically employed by Gazebo. In our work, described in Section 4, we employ these same improvements to the simulator.

The problem of control stability also often arises because stability is not a criterion typically optimized for in RL. To address these issues, Koch et al. engineer their reward signals to encourage a smooth control response. Their new reward signal is comprised of five components: (i) a penalty for high motor acceleration, r_a ; (ii) an in-band-reward for maintaining low motor activation while staying within an error threshold, r_b ; (iii) a penalty proportional to the state error of the quadrotor, r_e ; (iv) a penalty for attempting to over-actuate the motors, r_o ; (v) a penalty for not actuating the motors, r_n .

These five reward components are composed additively to give the final state reward at time t , r_t :

$$r_t = r_a + r_b + r_e + r_o + r_n \quad (1)$$

with the five individual reward components computed as:

$$r_a = -100 \cdot \max_i (|y_t^{(i)} - y_{t-1}^{(i)}|) \quad (2)$$

$$r_b = 1000 \cdot \left(1 - \frac{1}{4} \sum_i y_t^{(i)}\right) \mathbb{1}_{\mathbf{e}_t < \epsilon} \quad (3)$$

$$r_e = \|\mathbf{e}_{t-1}\| - \|\mathbf{e}_t\| \quad (4)$$

$$r_o = -1 \times 10^9 \cdot \sum_i \max(y_t^{(i)} - 1, 0) \quad (5)$$

$$r_n = -1 \times 10^9 \cdot \mathbb{1}_{\|\bar{\phi}_t\| > 0} \cdot \sum_i \mathbb{1}_{y_t^{(i)} = 0} \quad (6)$$

where $\mathbb{1}_z$ is an indicator function on criterion z , $y_t^{(i)} \in [0, 1]$ is the current motor-control signal output by the RL agent for each motor $i \in \{1, 2, 3, 4\}$, $\bar{\phi}_t$ is the desired angular roll, pitch and yaw velocity set-point, ϕ_t is the current angular roll, pitch and yaw velocity, and $\mathbf{e}_t = \bar{\phi}_t - \phi_t$. While this reward signal structure was shown to consistently achieve smooth control in simulation, the performance of the agents in real world flight was observed to still present with high-frequency oscillations in the motor control signals, contributing to significant motor heating and power drain.

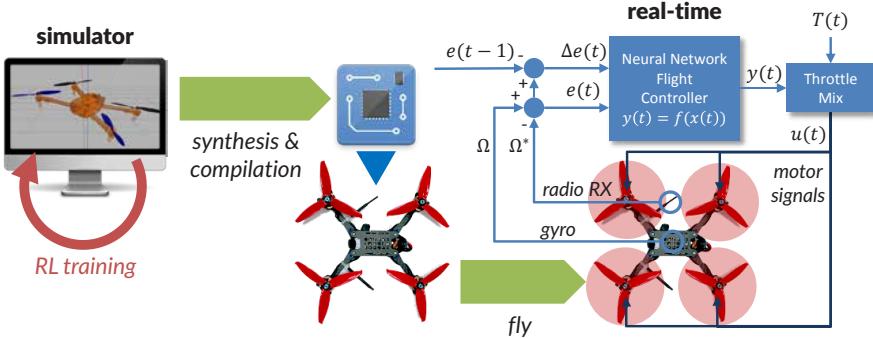


Fig. 3. An overview of the Neuroflight framework [14]

4 RE+AL: REINFORCEMENT-BASED TRANSFERABLE AGENTS THROUGH LEARNING

Prior work in the field has established the viability of deep RL in continuous controls task and the works of Koch et al. [13] provided us with a baseline upon which to build. There are a number of crucial pending problems with controllers trained on the Neuroflight pipeline, namely the lack of reproducibility and failure to preserve smooth control in transfer. We posit that the main issue is the over-engineering of the reward signal and that many of the other problems with their agents are consequential of this. To combat this issue, we developed a new reward scheme to better reflect desired network behavior, without being as closely tied to the environment dynamics, while also modifying the training environment to encourage a better exploration of the state-space.

The reward structure introduced by Koch et al., which was summarized in Section 3.2, allowed for the successful training of PPO agents. However we found that the reward structure was far too fine-tuned for the very specific network architecture and simulator design, with minor changes in either causing the agents’ training to break down. Despite the highly engineered rewards, the final agent(s) did not reliably transfer to the real drone. To mitigate this, the authors had to select potentially flight-capable agents manually through trial-and-error with basic flight tests. Furthermore, even the best agent presented by Koch et al. [13, 14] was noisy in its control actuation, with high-frequency oscillations, resulting in significant motor heating and power consumption. We believe this reflects a reward structure fundamentally ill-suited to represent the control problem at hand.

In the remainder of this section, we will discuss the systematic process by which we developed ‘REinforcement-based transferable Agents through Learning’, or RE+AL, a new regimen for training control policies. While we demonstrate the viability of our methods on a quadrotor drone, the underlying principles of our design should apply to a broader spectrum of control problems.

4.1 Constructing and Composing Multi-Objective Rewards

When developing agents in simulation with the explicit intention of transferring trained agents to a different domain, it is important to consider what aspects of the RL agent’s performance can be reliably translated between the domains and how the performance can be captured by the reward signal. We need to recognize that what is intuitive for us as the *human* designers is not always what contributes to good training and/or transfer, and vice versa. If the reward design is meant to both be intuitive and good for training, we need to identify the core performance metrics and represent them in a smooth and continuous function of the agents’ behavior to allow for more stable numerical optimization.

When considering the problem of training a low-level controller for drones, we require three key performance traits: (i) a low tracking error – the agents need to respond accurately to input control signals; (ii) smoothness – the control signal should have minimal unnecessary actuation to preserve the life of the hardware components; and (iii) maintaining a minimum output – all motors need to output at least enough power to keep the drone airborne.

Koch et al. identify and attempt to account for these same performance metrics. Equation 4 attempts to encourage low tracking error. Equation 2 penalizes acceleration in the motor, thus promoting smooth control signals. Equations 3, 5, and 6 provide feedback on the quality of the motor signal and drive the policy to maintain a low, non-zero level of actuation. The main problem with these reward signals however is that they are both adversarial and on vastly different scales – with some reward signals being several orders of magnitude larger than others – and this often results in destructive interference in behavior optimization [5].

It is important to keep in mind that the RL formulation used requires a single scalar reward signal. Koch et al., as do many others [4, 11, 12, 21, 27], compose the different reward signals additively (see Equation 1). Composing rewards additively can be effective, but it is our hypothesis that rewards should only be composed additively if none of the objectives represented by the reward signals conflict with each other and if they are all of the same polarity (all positive or all negative). Otherwise, signals of opposite polarity cancel each other out when they ought not. With this control problem, there is always a conflict: tracking errors need to be minimized, which requires motor actuation, however this conflicts with the goal of minimal actuation. One might attempt to weigh the contributions of each reward component, but this brings the requirement of a lot of fine-tuning on the reward coefficients to promote the right behavior. We find this approach tedious and to be the culprit for poor reproducibility. Instead, we propose that rewards be normalized in the $[0, 1]$ interval and composed multiplicatively.

With multiplicative composition and normalized reward components, the importance of each component reward signal is preserved. If any one reward component falls too low, the overall reward signal is lowered regardless of the quality of the other components. In other words, the agent is not allowed to offset its poor performance on one objective by being better on another. This type of composition encourages agents to learn to jointly optimize across all the reward components. It also has the advantage of being scale-invariant as the relative scales of the different reward components do not cause any of them to overshadow the others. It is however sensitive to translation, where padding a reward can change its “importance”. We use this property to our advantage to specify which rewards we value more than others. Instead of simply taking the product across the rewards however, we take their geometric mean [9] $g(r)$ defined as:

$$g(r) = \left(\prod_k^K \min(1, r_k + \epsilon) \right)^{K^{-1}} \quad (7)$$

where $r \in \mathbb{R}^K$ and ϵ is a small value meant to ensure that all components being multiplied are non-zero. In practice, we use $\epsilon = 0.001$. With the normalized rewards being on the range of $[0, 1]$, simply composing them multiplicatively could drastically reduce their numerical value upon composition, thus reducing the estimated value of a state and possibly resulting in an under-estimation of the value. The geometric mean on a K -dimensional vector takes the K^{th} root on the vector, thus preserving the scale of the vector’s components. Applied to our reward signals, this prevents policies from too quickly learning to underestimate the value of some states, consequently encouraging a better exploration of the state space during training. To the best of our knowledge, we are the first to propose multiplicative reward composition as a deliberate design choice in representing reward signals for better learning in RL.

To work with our multiplicative composition, we define a new set of reward signals to reflect the controller performance traits identified earlier in this section. We begin by reflecting each performance trait as a penalty on each of the performance criteria:

- (i) Error penalty defined by the difference between the set point and the actual angular rate of the drone

$$p_s = \|\phi_t - \bar{\phi}_t\|_4; \quad (8)$$

- (ii) Smoothness penalty on each motor, i , which penalizes change in motor usage

$$p_c^{(i)} = |y_t^{(i)} - y_{t-1}^{(i)}|; \quad (9)$$

- (iii) A thrust penalty on each motor, i , which is intended to keep motor usage at μ , the average power used for maintaining stable thrust for maintaining altitude, while penalizing it for over actuating

$$p_u^{(i)} = |y_t^{(i)} - \mu|. \quad (10)$$

The networks we train are designed specifically for attitude control. Thrust control is handled separately by a thrust mixer which mixes a pilot’s thrust input with the motor actuation required for the desired rotational velocities. We observed that an average 34% thrust was required for stable hovering when flying with a tuned PID controller. Treating this as a baseline for the minimum thrust required during flight, we utilized a value $\mu = 0.34$ in training to better match the simulated dynamics with the expected minimum thrust during flight.

The penalties are cast as positive rewards, p^+ , by reflecting it around 0 and clipping it to $[0, 1]$, i.e.:

$$p^+(p) = \min(1, \max(0, 1 - p)). \quad (11)$$

This results in the reward components being defined as:

$$r_s = p^+(p_s/\beta) \quad (12)$$

$$r_u = g(p^+(p_u)) \quad (13)$$

$$r_c = g(p^+(p_c)) \quad (14)$$

where p_s is scaled and bounded by β , such that everything above β deg/s would be considered a 0. In practice, we observed that $\beta = 300$ served as a good upper bound for our drone in training. Finally, the composite reward signal, r_t , used in training is composed as:

$$r_t = g([r_s, r_u, r_c]). \quad (15)$$

4.2 State space

Given knowledge of the reward signals, it is important to define the state space under which the agent operates in a way that contains all the relevant information that contributes to the reward signals.

Koch et al. [13, 14] originally composed their state vector, s_t as:

$$s_t = [\mathbf{e}_t, \mathbf{e}_t - \mathbf{e}_{t-1}]$$

to capture the current tracking error and change in angular velocity, i.e. the angular acceleration, of the drone.

We instead define our state vector, s_t as:

$$s_t = [\mathbf{e}_t, \phi_t, \phi_t - \phi_{t-1}, \mathbf{y}_t]. \quad (16)$$

While compressing information about the current and desired angular velocities into a single error quantity does allow for a reduction in state complexity, it is predominantly only useful when one can assume that control responses are always the same given any error, regardless of the current underlying state of the system. We found that this is not the case in practice, with the required control response varying depending on the angular velocity of the drone. In light of this, we chose to additionally include current angular velocity and acceleration in our state space, instead of leaving the networks to learn the disambiguation on their own. We also chose to include the last action taken by the agent as a part of the state space. When penalizing an agent for having significant changes in actions taken between time steps (see Equation 9), it is important for the agent to be aware of what action was last taken in order to effectively learn to remain smooth.

4.3 Goal generation

The original Neuroflight training framework provided training goal signals, i.e. the desired set-point angular velocities for agents, as a sequence of step inputs. Over a single training episode¹, the goal was changed once from 0 deg/s to a randomly selected value within the quadrotor’s flight envelope, held constant for a few seconds and returned to zero. This was not a good reflection of the types of signals received during actual flight, as human pilots would typically perform continuous maneuvers, rarely holding at a specific target velocity. We further hypothesized that the long periods of consistent signal levels might be conditioning the policy’s estimations of state transition probabilities to be biased towards minimal change.

¹An episode is defined as a contiguous sequence of interactions between the agent and the environment across several simulated seconds.

In order to get more realistic set-point changes in target angular velocities during training, we opted to procedurally generate set-points according to a Perlin noise[22] function. Goal generation built on Perlin noise allows us to have constantly changing random set points but preserves smoothness between subsequent timesteps. Furthermore, agents are able to explore more of the state space in individual episodes, instead of being conditioned for stable goal inputs. In our experience, this contributed to agents learning faster and seemingly to generalize better, likely due to not over-estimating the transition probabilities between self-same states.

We further augment the base Perlin noise function by multiplying it with a slower Perlin noise function so that the environment can be conditioned to expose the drone to a balance of steady as well as more aggressive maneuvers. To aid in the agents’ ability to generalize over non-continuous inputs.

For a Perlin noise function, $\mathcal{P}(t, o)$ defined on time, t and the number of desired octaves, o , we define our set-point generation function, $\bar{\phi}(t)$, as:

$$\bar{\phi}(t) = \mathcal{P}(t, 4) \times \mathcal{P}(t, 1)^2 \quad (17)$$

The first Perlin component, with an octave count of 4, generates a sequence of values whose rate of change is neither too smooth nor noisy. The second component biases the set-point to be closer to zero, which is intended to help the function to more closely resemble real-life controls. Indeed, pilots do not constantly change their command inputs, nor do they do it at a constant rate.

During training episodes, we monitor the agent’s tracking error and terminate episodes early if errors grow too large. This is done both to indicate to agents when their behavior is unacceptably bad, and also to avoid wasting training time on unrecoverable states.

4.4 Specific Network Architecture

In this work, we use a standard PPO agent, defined by and trained with open-source code provided by OpenAI baselines [7]. RL algorithms are often sensitive to the specific dynamics of the environments they are trained in [8, 19]. We specifically use PPO as it has already been proven to be a viable training algorithm for this specific problem and drone. The policy network we train is a fully-connected network with 2 hidden layers, and 64 artificial neurons per hidden layer. The output layer of the policy network has 4 neurons — one for each of the motor command signals.

4.5 Compilation Toolchain

At the end of training, the TensorFlow [3] graph representing the trained agent is frozen and saved as a TensorFlow checkpoint. The checkpoint is then optimized and compiled using TFcompile, a compilation tool offered by TensorFlow, into a self-contained executable, which is then linked with a custom version of the Betaflight firmware [1]. Finally, the firmware is compiled to generate a hex file ready to be flashed on the quadrotor. Typically, Betaflight is configured to work with a Proportional Integral Differential (PID) controller but we replace this control block with a call to our trained neural network agents. Aside from the replaced controller, the Betaflight firmware offers routines for (i) reading the remote control commands for the desired angular velocity, (ii) reading and filtering the angular velocity from the on-board gyroscope sensor, and (iii) writing the motor control signals to the electronic speed controller (ESC).

Due to the relatively small size of the neural networks used, the final optimized neural network is 12KB large and operates at approximately 730Hz on an ARM Cortex-M4 processor clocked at 216MHz.

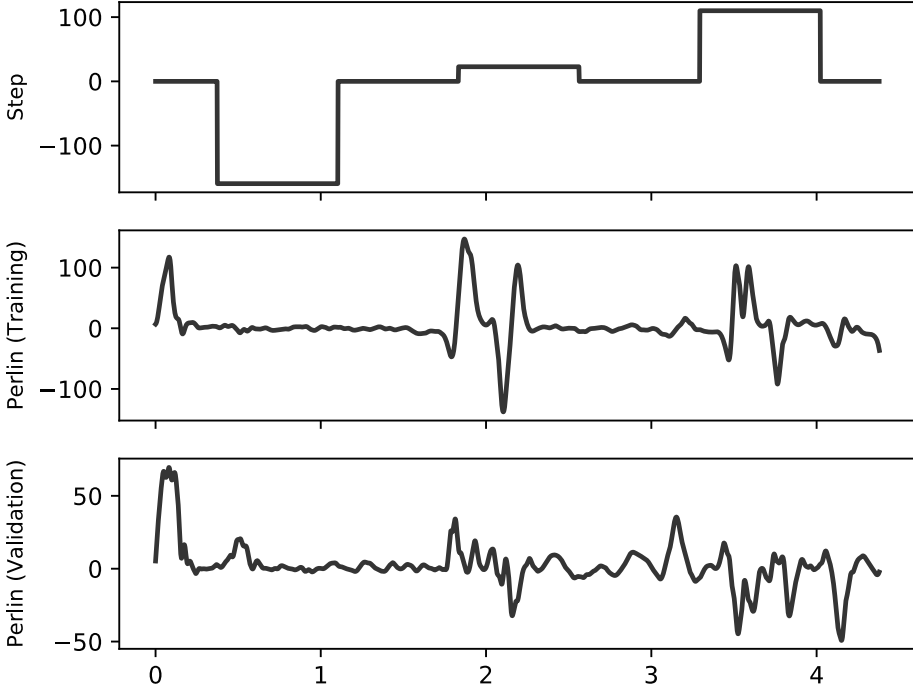


Fig. 4. Neuroflight’s step set-point generator compared to our Perlin based one – We make use of a more aggressive procedural generator during training to encourage agents to explore more of the state space, and validate on a gentler signal, one that is meant to reflect more typical inputs from a pilot.

5 EVALUATION

Our primary goal is to be able to reliably and consistently train agents for low-level attitude control. We therefore define our evaluation criteria predominantly on the quality of controller transfer to the real drone as observed in real-world flight tests. As in Section 4.1, performance metrics are defined on control input tracking error, $\mathbf{e}_t = \phi_t - \bar{\phi}_t$ and smoothness, $\Delta y_t = y_t - y_{t-1}$. Before deeming any agent to be potentially flight-worthy, however, it is important to verify their performance in training and simulation.

In this section, we discuss the performance trends achieved in simulation during training and the performance of the controllers transferred to the physical drone. Our experiments show that agents trained using our pipeline consistently yield good tracking performance both in simulation and in the actual test flights. While tracking performance is comparable to the baseline established by Koch et al. [14], our controllers present with significantly less strain on the motors thanks to the absence of high-frequency control oscillations.

We use the techniques proposed by Koch et al. [13] as the baseline for our evaluation by comparing our agents against the best agents produced by their methods. We also compare our agents against PID controllers tuned by the Ziegler-Nichols method [37] to further demonstrate the relative efficacy of RE+AL in developing good controllers. The PID tuning was performed independently for simulated and real control in recognition of the differences in simulated dynamics from real-world physics. This also provides the most fair comparison of agents trained by our method against controllers specifically tuned for optimality in their respective domains.

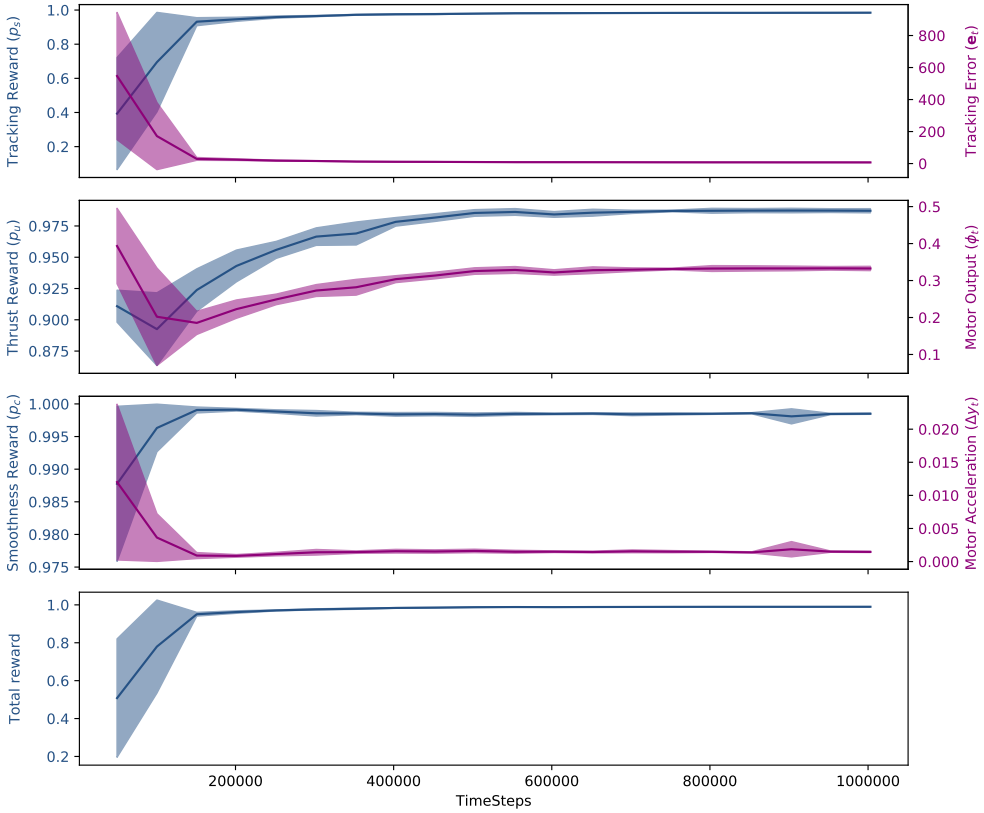


Fig. 5. Tracking RE+AL training progress - The performance of agents in simulation begins to plateau after approximately 200,000 timesteps. Despite this however, there are still minor improvements to the agents’ behaviors that are learned. We practice a form of early stopping to terminate training after the average reward in simulation falls below a threshold in the interest of decreasing the number of samples required before reaching what we consider a high performance agent. For our drone, this seemed to happen by approximately 1 million timesteps – testing agents trained for longer periods of time revealed that, while they achieved better performance in simulation, their performance in transfer was often worse, we suspect this is because the agents overfit to the specific dynamics of the simulation. We also show how the behavior of our trained agents develops to satisfy the three performance criteria outlined in Section 4.1 and their respective rewards.

5.1 Training Performance

We begin by analyzing the trends observed in training in order to verify the performance and reproducibility of agents in simulated flights. Throughout the course of training, checkpoints are logged at regular intervals of 50,000 training steps (corresponding to 50,000 interactions between the agent and the simulated environment). The performance of the agents at each checkpoint is then validated in simulated flight sequences while the errors, rewards and simulated motor behavior are logged. Each training cycle begins with a randomly initialized training policy and a different random seed.

Fig. 5 shows the typical trends in training for our RL agents. Note that tracking error and motor acceleration are minimized, as expected, while average motor actuation is driven to 34%, exactly as intended by our reward signal design presented in Section 4.1. We observe that the performance of

the agents in simulation tends to plateau after approximately 500,000 timesteps. We allow the agent to train slightly longer so as to further reinforce good behavior. We also observed that allowing agents to continue training until they reach 1,000,000 timesteps typically resulted in a late-stage improvement in tracking errors, bringing the simulated-environment tracking error below an average of 10 deg/s. We practice early-stopping, a process by which training of ML models is stopped even if performance appears to be improving in training, to prevent over-fitting to the training domain and improve the chances for better transfer to the test domain. We found that stopping the training at approximately 1 million timesteps – i.e., just after agents typically learn to reduce tracking error in training below 10 deg/s – offered a good balance of tracking, stability and policy transfer.

The training trends of our agents contrast starkly with agents trained by Koch et al. [14] in three significant ways: (i) we achieve a 10x reduction in required training time, which in practice corresponds to training for 50 minutes on our machines as opposed to 9 hours; (ii) the performance of our agents remains relatively stable even after plateauing instead of experiencing a sudden, inexplicable drop in performance half-way through training, and (iii) as we will demonstrate in Section 5.2, agents trained by our method can simply be transferred to the drone at the end of training without requiring an additional search over the checkpoints to identify ‘good’ agents. Hence our training exhibits good repeatability.

Table 1. Comparing Tracking Errors

Agent	Mean Absolute Error (deg/s)				Current (Amps)
	Roll	Pitch	Yaw	Overall	
Simulated Validation					
PID	10.31	17.21	6.75	11.41	-NA-
Neuroflight	8.85	6.65	6.48	7.30	-NA-
<i>RE+AL</i> (ours)	7.00 ± 2.37	7.80 ± 3.50	5.76 ± 1.51	6.75 ± 2.13	-NA-
Test-Flight					
PID	6.68	4.54	3.80	5.01	8.07
Neuroflight	6.26	5.42	3.87	5.19	22.87
<i>RE+AL</i> (ours)	6.01 ± 1.80	4.52 ± 1.18	2.64 ± 0.18	4.20 ± 1.14	5.86 ± 3.10

Table 1 demonstrates the high accuracy of our agents in simulation, with mean average errors of 6.75deg/s and a relatively low, and almost imperceptible variance of 2.13deg/s. These metrics further speak to the repeatability of agents trained by our method and their apparent insensitivity to random seeds.

Fig. 6 provides a temporal close-up on the tracking behavior of 3 independently trained agents in simulation to allow readers a better sense of how individual agents perform on simulated control tasks. *RE+AL* agents also generally exceed the set-point tracking performances of both a tuned PID and the best Neuroflight agent (provided by Koch et al.).

5.2 Test Performance

Upon completing training, the TensorFlow graphs representing each agent was compiled and deployed onto the drone, with which our pilot conducted a series of flight tests. Real-world flight tests focused on exploring the controllers’ viable flight envelope, responsiveness, smoothness and power consumption. We compare our results with the best agent from the original Neuroflight framework. As Koch et al. noted in [13], they were only able to get one agent to successfully transfer

a viable controller from simulation to reality while maintaining reasonable tracking performance. We were able to acquire a copy of this agent from the authors for comparison with agents trained by our method. Due to over-heating of the motors, however, our tests with the Neuroflight agent had to be limited in order to prevent hardware failures.

We note from Table 1 that our method produces agents that are comparable to the best agent previously developed by Neuroflight. As the Neuroflight training pipeline only produced one flight-worthy agent, we are unable to present a measure of variance for the Neuroflight agent.

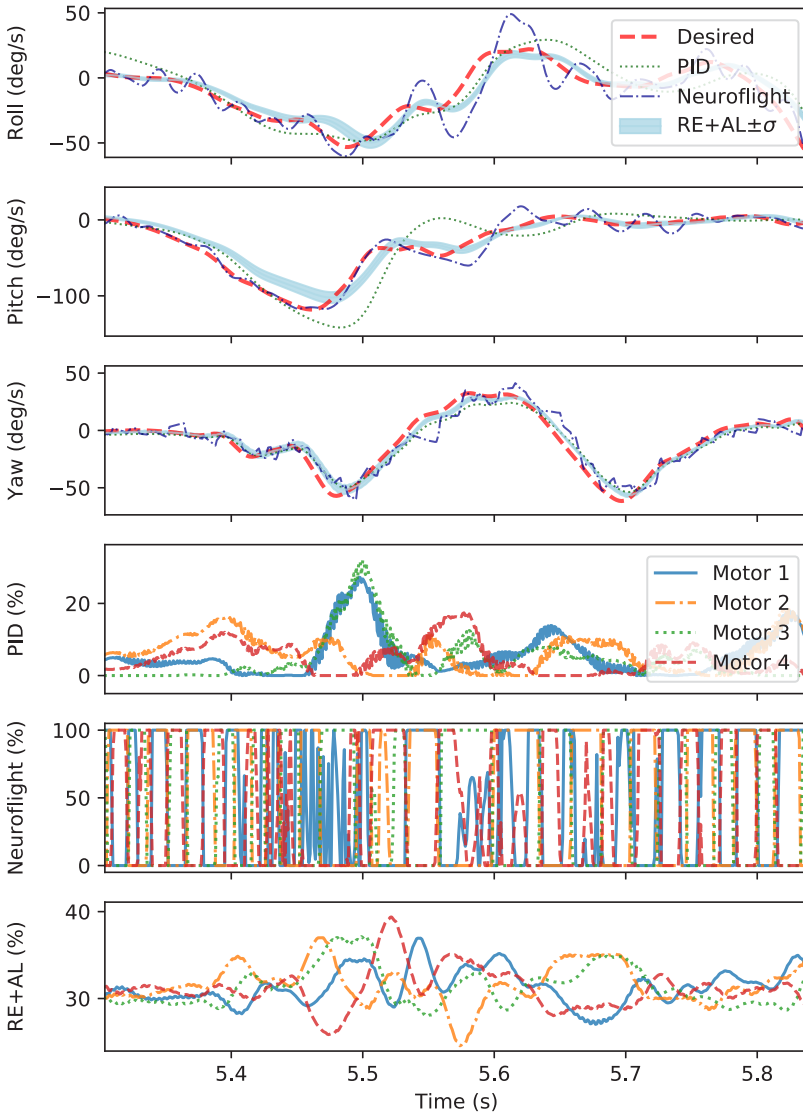


Fig. 6. Zooming in to analyze stability - By zooming in on the agents' behaviors over 0.5s of simulated time, we show that RE+AL control signals are very smooth and stable in simulation and our agents present with practically no oscillatory behavior.

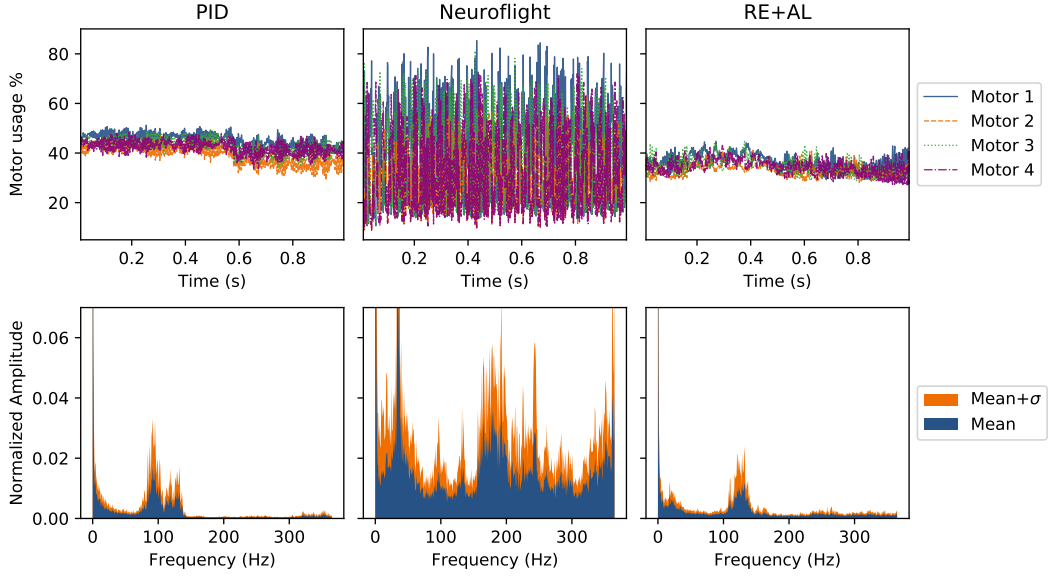


Fig. 7. Fourier transform of control signals - Each column shows a slice of the motor usage during similar real flight maneuvers (top), in comparison to the mean frequency spectrum resulting from a Fourier transform on the control signals of the quad-rotor’s 4 motors (bottom). This visualization is presented on the same scale to highlight the significant difference in control signal noise between RE+AL and the best agent trained by the previous Neuroflight training pipeline. Note that the RE+AL agents offer significantly smoother control, smoother even than the PID controller, while offering better tracking. This is further reflected in the lower power consumption of RE+AL agents, as shown in Table 1.

Crucially however, we are able to do this *consistently*, with every agent trained by our method serving as a viable candidate for transfer to the real drone. Our method also produces agents that have significantly smoother control policies when transferred to the drone, as shown in Fig. 7.

Fig. 7 shows the average Fourier transform of the motor control signals across all our RE+AL agents compared against the best Neuroflight agent. Our agents have shown to produce smoother motor-control signals, with a peak output frequency of approximately 130Hz, compared to Neuroflight’s 330Hz. Note as well that the magnitude of the RE+AL agents’ high frequency peak is significantly smaller than that of Neuroflight. In practice, this smoothness results in noticeably reduced heat generation in the motors, allowing us to fly for longer and without risking hardware failures due to motor burnout. A sample flight log is presented in Fig. 8. The improved control output is further highlighted in the current draw during operation, with the motors of the RE+AL agents drawing, on average, 5.86 ± 3.10 Amps during flight tests, while the Neuroflight agent drew an average of 22.87 Amps on a similar hover test. This is also notably better than the average current draw of the PID controller. This implies that, not only are RE+AL agents capable of delivering smooth, stable and reliable control; but that they are also able to offer longer flight times, better even than a classical controller. To the best of our knowledge, our work is the the first to clearly demonstrate the improved practical utility of a neural-network controller over classical control on a problem where conventional wisdom might suggest that classical control techniques would get us closest to optimal control.

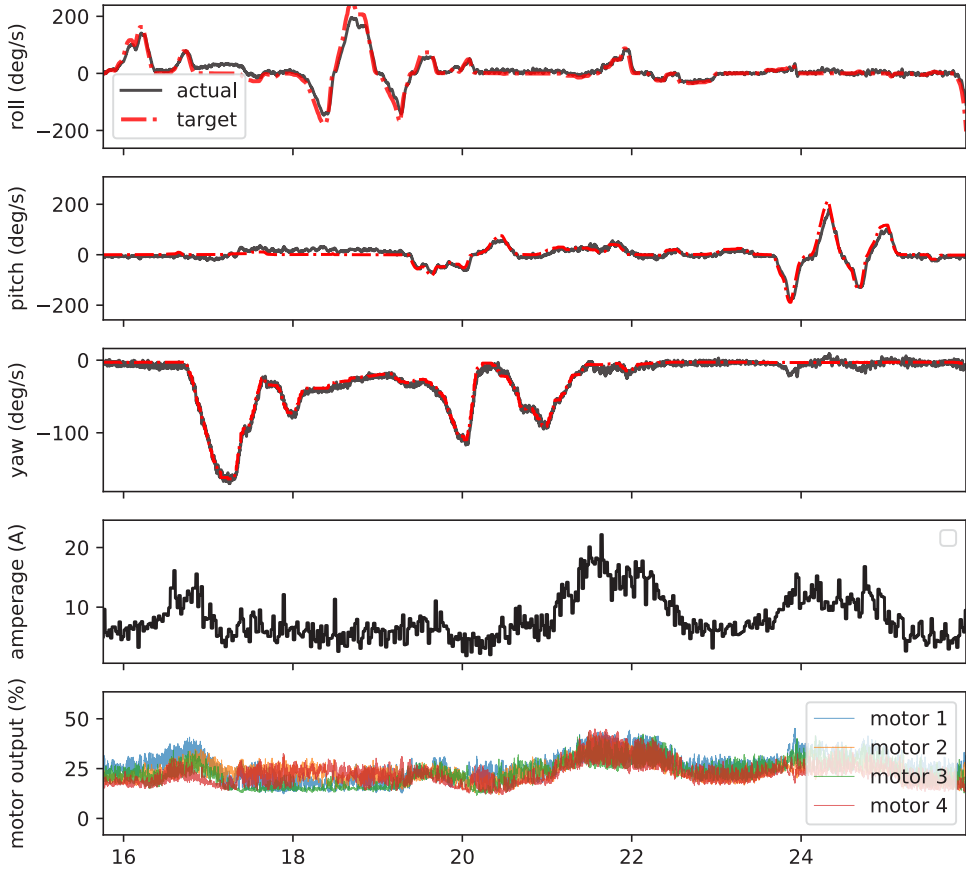


Fig. 8. We show here a sample flight with a typical RE+AL agent – While set-point tracking is generally successful, we observe that the motor control signals are noisier than in simulation but not enough to cause excessive spikes in the quadrotor’s state response or power consumption.

5.3 Discussion

Based on our results, it is clear that RE+AL makes significant strides in improving the general viability of RL-based control on the task at hand. It is interesting to note from Table 1 that both Neuroflight and RE+AL perform better in real flights (at least in tracking control set-points), as compared to tests in simulation – a clear sign of the reality gap at work. What this implies is that albeit the GymFC simulator makes for a good representation of the dynamical environment, it is not without its limitations. This also further justifies our decision to practice early stopping during training. Though not reported here due to incomplete data, preliminary findings suggest that allowing RE+AL agents to train longer in simulation did not offer substantial benefits in transfer to the real quadrotor – in fact, we observed that actuation would actually get noisier. This strongly suggests that further training would simply cause the agent to over-fit to the simulator, resulting in reduced transferability.

Our RE+AL method is not without its drawbacks. A persistent issue observed both in training and testing is that the agents seem to be overly biased towards achieving good yaw tracking – sometimes at the expense of good tracking along roll and pitch. This was also found to occasionally manifest as sudden perturbations in the quadrotor’s angular roll or pitch velocities when a large change in angular yaw velocity was requested. While our pilot did not consider this to be a major issue and was able to quickly recover from any flight irregularities that resulted from this behavior, the fact remains that these occasional control glitches do exist in some agents. The nature of this work also makes it difficult to compare RE+AL with control techniques not based on the Neuroflight framework as training for real-world tasks requires tailored simulation environments and reward signals. As more generalized frameworks for training and testing RL-based control on hardware platforms become available, it would be interesting to evaluate RE+AL more thoroughly against alternative techniques such as those proposed by Hwangbo et al. [11] and Molchanov et al. [21].

6 CONCLUSION

Our work introduces REinforcement-based transferable Agents through Learning (RE+AL), a framework for designing simulated training environments which encourage the training of smoother and more transferable controllers. RE+AL is built on a philosophy of improving the value of information provided to an RL agent during training while maintaining interpretability. Through the systematic design of descriptive yet simple reward signals, state space and environment-interaction paradigm, RE+AL facilitates the consistent training of RL agents with good control performance and reliable transferability to real hardware. RE+AL is tested on the NF1 quadrotor platform, and we verified that 100% of all agents trained with RE+AL were flight-worthy, offering good RC-control tracking and smooth actuation of the quadrotor’s motors. RE+AL lays the groundwork for future work in developing controllers for more complex controls tasks, and also for incorporating additional learning techniques to further close the reality-gap.

REFERENCES

- [1] [n.d.]. Betaflight. <https://betaflight.com/>
- [2] [n.d.]. Open Dynamics Engine. <https://www.ode.org/>
- [3] Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, et al. 2016. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
- [4] Axel Abels, Diederik M Roijers, Tom Lenaerts, Ann Nowé, and Denis Steckelmacher. 2018. Dynamic weights in multi-objective deep reinforcement learning. *arXiv preprint arXiv:1809.07803* (2018).
- [5] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. Concrete Problems in AI Safety. *arXiv:1606.06565 [cs.AI]*
- [6] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. 2016. OpenAI Gym. *CoRR abs/1606.01540* (2016). *arXiv:1606.01540*
- [7] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. 2017. OpenAI Baselines. <https://github.com/openai/baselines>.
- [8] Yan Duan, Xi Chen, Rein Houthoofd, John Schulman, and Pieter Abbeel. 2016. Benchmarking Deep Reinforcement Learning for Continuous Control. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48* (New York, NY, USA) (*ICML’16*). JMLR.org, 1329–1338.
- [9] Philip J. Fleming and John J. Wallace. 1986. How not to lie with statistics: the correct way to summarize benchmark results. *Commun. ACM* 29 (1986), 218–221.
- [10] Hado V. Hasselt. 2010. Double Q-learning. In *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta (Eds.), Curran Associates, Inc., 2613–2621.
- [11] Jemin Hwangbo, Inkyu Sa, Roland Siegwart, and Marco Hutter. 2017. Control of a Quadrotor With Reinforcement Learning. *IEEE Robotics and Automation Letters* 2 (2017), 2096–2103.
- [12] Alex Kendall, Jeffrey Hawke, David Janz, Przemyslaw Mazur, Daniele Reda, John-Mark Allen, Vinh-Dieu Lam, Alex Bewley, and Amar Shah. 2019. Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 8248–8254.

- [13] William Koch. 2019. *Flight Controller Synthesis Via Deep Reinforcement Learning*. Ph.D. Dissertation. Department of Computer Science, Boston University. arXiv:1909.06493 [cs.LG]
- [14] William Koch, Renato Mancuso, and Azer Bestavros. 2019. Neuroflight: Next Generation Flight Control Firmware. *CoRR* abs/1901.06553 (2019). arXiv:1901.06553
- [15] William Koch, Renato Mancuso, Richard West, and Azer Bestavros. 2018. Reinforcement Learning for UAV Attitude Control. *ACM Transactions on Cyber-Physical Systems* (2018).
- [16] Nathan Koenig and Andrew Howard. [n.d.]. Design and use paradigms for gazebo, an open-source multi-robot simulator. In *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)(IEEE Cat. No. 04CH37566)*, Vol. 3. IEEE, 2149–2154.
- [17] Jeongseok Lee, Michael Grey, Sehoon Ha, Tobias Kunz, Sumit Jain, Yuting Ye, Siddhartha Srinivasa, Mike Stilman, and C. Liu. 2018. DART: Dynamic Animation and Robotics Toolkit. *Journal of Open Source Software* 3, 22 (2018), 500. <https://doi.org/10.21105/joss.00500>
- [18] Timothy P. Lillicrap, Jonathan J. Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. 2016. Continuous control with deep reinforcement learning. *International Conference on Learning Representations* (2016). arXiv:1509.02971
- [19] A. Rupam Mahmood, Dmytro Korenkevych, Gautham Vasan, William Ma, and James Bergstra. 2018. Benchmarking Reinforcement Learning Algorithms on Real-World Robots. *Conference on Robot Learning* (2018). arXiv:1809.07731
- [20] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602* (2013).
- [21] Artem Molchanov, Tao Chen, Wolfgang Hönig, James A. Preiss, Nora Ayanian, and Gaurav S. Sukhatme. 2019. Sim-to-(Multi)-Real: Transfer of Low-Level Robust Control Policies to Multiple Quadrotors. *International Conference on Intelligent Robots and Systems* (2019). arXiv:1903.04628
- [22] Ken Perlin. 1985. An Image Synthesizer. *SIGGRAPH Comput. Graph.* 19, 3 (July 1985), 287–296. <https://doi.org/10.1145/325165.325247>
- [23] Gavin A Rummery and Mahesan Niranjan. 1994. *On-line Q-learning using connectionist systems*. Vol. 37.
- [24] Fereshteh Sadeghi, Alexander Toshev, Eric Jang, and Sergey Levine. 2017. Sim2Real View Invariant Visual Servoing by Recurrent Control. *CoRR* abs/1712.07642 (2017). arXiv:1712.07642
- [25] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust Region Policy Optimization. In *Proceedings of the 32nd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 37)*, Francis Bach and David Blei (Eds.). PMLR, Lille, France, 1889–1897.
- [26] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal Policy Optimization Algorithms. *CoRR* abs/1707.06347 (2017). arXiv:1707.06347
- [27] Christian R Shelton. 2001. Balancing multiple sources of reward in reinforcement learning. In *Advances in Neural Information Processing Systems*. 1082–1088.
- [28] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.
- [29] David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. 2017. Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815* (2017).
- [30] Richard S. Sutton and Andrew G. Barto. 1998. *Introduction to Reinforcement Learning* (1st ed.). MIT Press, Cambridge, MA, USA.
- [31] Richard S. Sutton, David McAllester, Satinder Singh, and Yishay Mansour. 1999. Policy Gradient Methods for Reinforcement Learning with Function Approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems (Denver, CO) (NIPS'99)*. MIT Press, Cambridge, MA, USA, 1057–1063.
- [32] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.
- [33] Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning* 8, 3-4 (1992), 279–292.
- [34] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8, 3-4 (1992), 229–256.
- [35] Amy X. Zhang, Nicolas Ballas, and Joelle Pineau. 2018. A Dissection of Overfitting and Generalization in Continuous Reinforcement Learning. *CoRR* abs/1806.07937 (2018).
- [36] Chiyuan Zhang, Oriol Vinyals, Rémi Munos, and Samy Bengio. 2018. A Study on Overfitting in Deep Reinforcement Learning. *CoRR* abs/1804.06893 (2018).
- [37] J. G. Ziegler and N. B. Nichols. 1993. Optimum Settings for Automatic Controllers. *Journal of Dynamic Systems, Measurement, and Control* 115, 2B (06 1993), 220–222.