

Multi-Task Reinforcement Learning for Quadrotors

Jiaxu Xing, Ismail Geles, Yunlong Song, Elie Aljalbout, and Davide Scaramuzza

Abstract—Reinforcement learning (RL) has shown great effectiveness in quadrotor control, enabling specialized policies to develop even human-champion-level performance in single-task scenarios. However, these specialized policies often struggle with novel tasks, requiring a complete retraining of the policy from scratch. To address this limitation, this paper presents a novel multi-task reinforcement learning (MTRL) framework tailored for quadrotor control, leveraging the shared physical dynamics of the platform to enhance sample efficiency and task performance. By employing a multi-critic architecture and shared task encoders, our framework facilitates knowledge transfer across tasks, enabling a single policy to execute diverse maneuvers, including high-speed stabilization, velocity tracking, and autonomous racing. Our experimental results, validated both in simulation and real-world scenarios, demonstrate that our framework outperforms baseline approaches in terms of sample efficiency and overall task performance.

Video: <https://youtu.be/HfK9UT1OVnY>.

Index Terms—Reinforcement Learning, Machine Learning for Robot Control, Aerial Systems: Perception and Autonomy

I. INTRODUCTION

REAL world quadrotor applications typically involve multiple tasks and skills. For example, in search and rescue scenarios or inspection, quadrotors are required to perform a range of specific tasks within a single mission, such as tracking moving objects, maintaining stable hover positions, and precisely following designated paths or targets. To meet these diverse demands, a generalist control policy that can effectively manage these tasks can greatly enhance the versatility and adaptability of quadrotors, making them more effective in real-world applications. However, developing a generalist controller for multiple tasks is a challenging problem since different tasks often have different objectives and state spaces. For instance, in quadrotor control, the objective for hovering is to stabilize the vehicle by reducing its velocity to zero, whereas quadrotor racing requires maximizing speed while avoiding collisions with gates. These tasks inherently conflict with their goals and demand different observations and strategies.

In this work, we tackle the multi-task quadrotor control problem using deep reinforcement learning (RL), which offers the advantage of automatically optimizing parametric controllers through trial and error. RL is particularly effective

Manuscript received: September 3, 2024; Revised November 6, 2024; Accepted November 26, 2024.

This paper was recommended for publication by Editor Giuseppe Loianno upon evaluation of the Associate Editor and Reviewers' comments.

The authors are with the Robotics and Perception Group, Department of Informatics, University of Zurich, and Department of Neuroinformatics, University of Zurich and ETH Zurich, Switzerland (<http://rpg.ifi.uzh.ch>). This work was supported by the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101120732 (AUTOASSESS) and the European Research Council (ERC) under grant agreement No. 864042 (AGILEFLIGHT).

Digital Object Identifier (DOI): see top of this page.

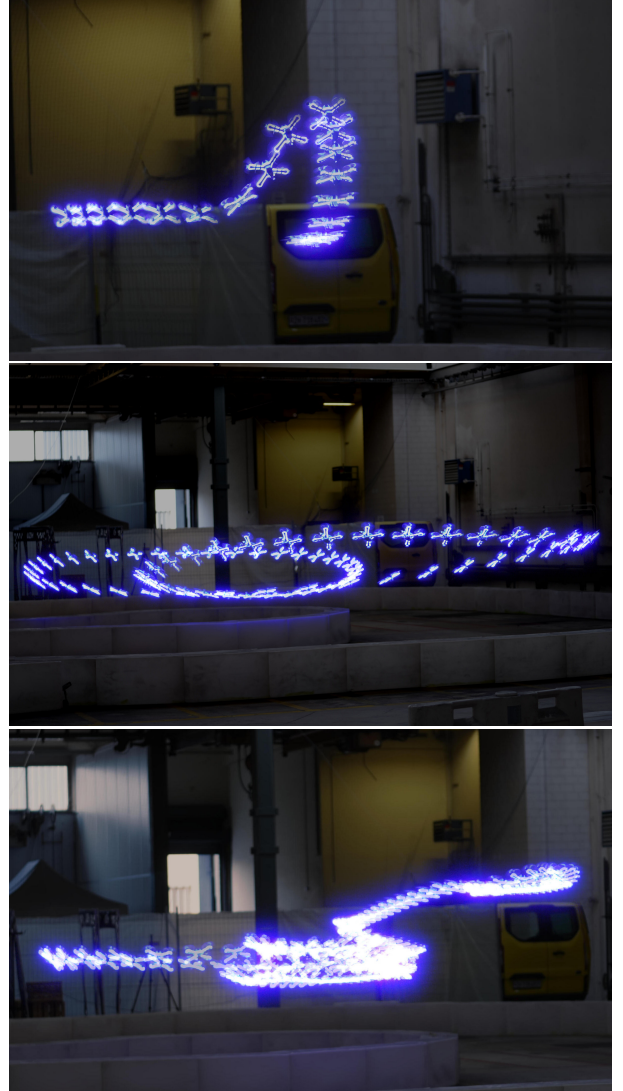


Fig. 1: The proposed approach performs three distinct tasks for quadrotor control in the real world. The resulting *single MTRL policy* can (*Top*) stabilize the quadrotor from high speed, (*Middle*) autonomously race through a fixed track, and (*Bottom*) track randomly generated velocities.

in handling highly non-linear dynamical systems and non-convex, non-differentiable objectives—challenges that are typically difficult for conventional optimization-based methods such as Model Predictive Control (MPC) [1], [2]. RL has demonstrated significant success in different domains of agile quadrotor control, ranging from time-optimal drone racing to obstacle avoidance [1], [3], [4].

State-of-the-art RL approaches have demonstrated specialization with great performance, even reaching the human-champion level in single-task scenarios such as drone rac-

ing [5]. However, these specialized policies often struggle to perform novel, out-of-distribution tasks and require re-training from scratch when faced with even minor changes in task configuration [6]. Consequently, the commonly used task-specific training approach is unsuitable for multi-task scenarios, making developing a multi-task RL policy for quadrotors a significant challenge. A natural solution to MTRL problems is to train a network jointly on all tasks to uncover shared structures that can improve efficiency and performance beyond what individual task solutions can achieve. However, learning multiple tasks simultaneously often poses a challenging optimization problem involving conflicting objectives [7], sometimes resulting in poorer overall performance and data efficiency compared to individual task learning.

Despite these challenges, previous work on MTRL has demonstrated the potential of integrating shared task structures to perform various manipulation skills, such as lifting and pick-and-place, particularly in fixed-base manipulation scenarios [8]–[11]. While MTRL has shown promise in manipulation on a simulation benchmark [12], its application for quadrotors remains largely unexplored.

Contributions

In this paper, we propose the first MTRL framework for learning various quadrotor control tasks efficiently. A key advantage of MTRL is its ability to share knowledge across different tasks, thereby enhancing learning efficiency. Although the reward objectives of RL may vary between tasks, the underlying physical dynamics of the quadrotor, such as mass, inertia, and other physical properties, remain constant throughout the learning process. Leveraging these invariant conditions, our framework efficiently utilizes information sharing to learn multiple tasks within a single policy.

The reinforcement learning framework employs a multi-critic setup, and we propose a shared task encoder for observations containing dynamical information while handling task-specific observations using task-specific encoding networks. By sharing common information and isolating task-specific data, we have developed a high-performance policy capable of executing maneuvers such as stabilization, velocity tracking, and racing.

We validate our MTRL policy in both simulation and real-world settings. We demonstrate that our framework outperforms baseline approaches in terms of sample efficiency and overall task performance. We believe this advancement is an important step toward developing a generalist quadrotor control policy, which could enable robots to handle diverse tasks in real-world scenarios more effectively.

II. RELATED WORKS

A. Reinforcement Learning for Quadrotor Control

In recent years, Reinforcement Learning (RL) has gained significant attention for the control of quadrotors. The work [13] is one of the first successful applications of RL to quadrotor control by tracking waypoints and recovering from challenging initial conditions. In [14], the performance gains of using RL for low-level control over classical methods are demonstrated. Besides obstacle avoidance [15], [16],

drone racing represents an important benchmark task for agile quadrotor flight, where the impact of reinforcement learning is very significant [1], [5], even outperforming state-of-the-art model-based control [1]. By optimizing several aspects of the training framework, in [17] it is demonstrated RL policies may be trained within seconds. Several works focus on vision-based flight with reinforcement learning, such as [18], which uses a CAD model to produce discrete velocity commands directly from pictures. In [19], it is shown how RL bootstrapped imitation learning [3] benefits vision-based agile flight. In [5], the combination of vision-based state estimation and RL-based control enabled surpassing human world champions in drone racing. In [20], RL is used to learn drone racing from pixels without explicit state estimation.

B. Multi-task Reinforcement Learning for Robotics

In [8], a large-scale collective robotic learning system, can rapidly acquire diverse skills by sharing exploration, experience, and representations across tasks, improving overall performance and capabilities. The work [9] demonstrates using MTRL to perform complex robotic skills like in-hand manipulation autonomously, significantly reducing the need for human resets during training sessions. In [21], modularization in network design is explored to facilitate MTRL, improving both sample efficiency and the performance of robotic tasks by dynamically configuring network modules according to the task requirements. In [22], a sequential multi-task learning scenario where a robot incrementally learns various tasks is demonstrated, using experiences from previous tasks to reduce the need for relearning and improve efficiency. Recently multi-task world models in proposed in [11], leveraging language model embeddings as task representations for model-based reinforcement learning of multiple robotic tasks. However, most existing MTRL research focuses on manipulation tasks with varying underlying dynamics across tasks, limiting opportunities for knowledge sharing. In contrast, our work leverages the consistent dynamics of the quadrotor, enabling us to propose a novel framework that enhances knowledge sharing for more efficient multi-task learning.

III. METHODOLOGY

A. Notation

In this manuscript, we define two reference frames. The first W is the fixed world frame with its z -axis aligned with gravity. The second frame B is the quadrotor body frame. These reference frames are illustrated in Fig. 2. Vectors and matrices are represented as bold quantities, with capital letters denoting matrices. Vectors include a suffix indicating the frame in which they are expressed and their endpoint. For example, the quantity \mathbf{p}_{WB} represents the position of the body frame B relative to the world frame W . The rotation matrix that transforms a vector from frame B to W is denoted by \mathbf{R}_{WB} .

B. Quadrotor Dynamics

Let \mathbf{p}_{WB} , \mathbf{q}_{WB} , and \mathbf{v}_{WB} represent the position, orientation, and linear velocity of the quadrotor, respectively, expressed in the world frame W . Let $\boldsymbol{\omega}_B$ denote the angular

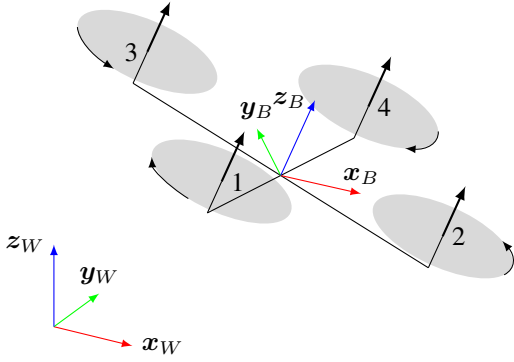


Fig. 2: Diagram of quadrotor model with the world and body frames and propeller numbering convention.

velocity of the quadrotor expressed in the body frame B . Additionally, let $c = \sum_i c_i$ represent the body's collective thrust, where c_i is the thrust produced by the i -th motor, and let $c = [0 \ 0 \ c]^T$ denote the collective thrust vector. Here, m represents the mass of the quadrotor, and \mathbf{g}_W is the gravity vector. Finally, let \mathbf{J} be the diagonal moment of inertia matrix, and $\boldsymbol{\tau}_B$ the body's collective torque. The quadrotor's dynamic model is

$$\dot{\mathbf{x}} = \begin{bmatrix} \dot{\mathbf{p}}_{WB} \\ \dot{\mathbf{q}}_{WB} \\ \dot{\mathbf{v}}_{WB} \\ \dot{\boldsymbol{\omega}}_B \end{bmatrix} = \begin{bmatrix} \mathbf{v}_{WB} \\ \frac{1}{2}\Lambda(\boldsymbol{\omega}_B) \cdot \mathbf{q}_{WB} \\ \mathbf{q}_{WB} \odot \mathbf{c}/m + \mathbf{g}_W \\ \mathbf{J}^{-1}(\boldsymbol{\tau}_B - \boldsymbol{\omega}_B \times \mathbf{J} \cdot \boldsymbol{\omega}_B) \end{bmatrix}. \quad (1)$$

C. Policy Learning

1) *Problem Formulation*: In the MTRL setting, we have N tasks $\mathcal{T} = \{T_1, \dots, T_i, \dots, T_N\}$, where each task T_i has a specific reward function r_i . Each task is defined by a Markov Decision Process (MDP) $\mathcal{M}_i = (\mathcal{S}_i, \mathcal{A}_i, \mathcal{P}_i, \mathcal{R}_i, \gamma_i)$, where \mathcal{S}_i is the state space, \mathcal{A}_i is the action space, \mathcal{P}_i is the transition probability, \mathcal{R}_i is the reward function, and γ_i is the discount factor. In the MTRL setting, the goal is to learn a policy π that maximizes the expected return $J(\pi)$ across all tasks. Given uniformly sampled tasks, the optimization objective is defined as

$$J(\pi) = \frac{1}{N} \sum_{i=1}^N \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma_i^t r_i(s_t, a_t) \right]. \quad (2)$$

In our work, our multi-task configurations are primarily distinguished by the *reward settings* \mathcal{R} as model dynamics \mathcal{P} remain identical due to the unchanging physical properties of the drones. In the following sections, we detail the tasks considered in this work and the policy learning approach.

2) *Autonomous Racing*: The autonomous racing task can be formulated as an optimization problem that aims to minimize the time required for an agile quadrotor to navigate through a predefined sequence of gates [23], as shown in Fig. 5. In this task, we use the observations $\mathbf{o} = [\mathbf{p}, \tilde{\mathbf{R}}, \mathbf{v}, \boldsymbol{\omega}, a_{\text{prev}}, \delta \mathbf{p}_1, \delta \mathbf{p}_2]$, where $\mathbf{p} \in \mathbb{R}^3$ denotes the drone's position, $\tilde{\mathbf{R}} \in \mathbb{R}^6$ is a vector comprising the first two columns of \mathbf{R}_{WB} [24], $\mathbf{v} \in \mathbb{R}^3$ and $\boldsymbol{\omega} \in \mathbb{R}^3$ denote the linear and angular velocity of the drone, a_{prev} represents the previous action from the actor policy, $\delta \mathbf{p}_1 \in \mathbb{R}^{12}$ represents

the relative position differences of the four upcoming gate corners on the race track with respect to the drone agent, with each corner specified by a 3D position in the world frame. Similarly, $\delta \mathbf{p}_2 \in \mathbb{R}^{12}$ represents the relative position differences of the corners from the next gate to the gate after that. Here $\delta \mathbf{p}_1$ represents the difference of the 4 corners of the next gate to pass between the current quadrotor position. And $\delta \mathbf{p}_2$ represents the positional difference of the corners between the next gate to pass and the gate after the next gate to pass on the race track. The RL policy training rewards are adjusted based on [1]. The reward at time t , denoted as r_t , is defined as the sum of various components

$$r_t^{\text{racing}} = r_t^{\text{prog}} + r_t^{\text{perc}} + r_t^{\text{act}} + r_t^{\text{br}} + r_t^{\text{pass}} + r_t^{\text{crash}}, \quad (3)$$

where r_t^{prog} encourages progress towards the next gate to be passed [1], r_t^{perc} encodes perception awareness by adjusting the quadrotor's attitude such that the optical axis of its camera points towards the next gate's center, r_t^{act} penalizes action changes from the last time step, r_t^{br} penalizes body rates and consequently reduces motion blur, r_t^{pass} is a binary reward that is active when the robot successfully passes the next gate, r_t^{crash} is a binary penalty that is only active when a collision happens, which also ends the episode. The reward components are formulated as follows

$$\begin{aligned} r_t^{\text{prog}} &= \alpha_1 (d_{\text{Gate}}(t-1) - d_{\text{Gate}}(t)), \\ r_t^{\text{perc}} &= \alpha_2 \exp(\alpha_3 \cdot \delta_{\text{cam}}^4), \\ r_t^{\text{act}} &= \alpha_4 \|\mathbf{u}_t - \mathbf{u}_{t-1}\|, \\ r_t^{\text{br}} &= \alpha_5 \|\boldsymbol{\omega}_{B,t}\|, \\ r_t^{\text{pass}} &= \alpha_6 \quad \text{if robot passes the next gate,} \\ r_t^{\text{crash}} &= \alpha_7 \quad \text{if robot crashes (gates, ground).} \end{aligned} \quad (4)$$

3) *Stabilization from High Speed*: In the stabilization task, the quadrotor is expected to recover to the static status, given randomized poses and high initial velocities. Here, successful stabilization is defined as achieving near-zero velocities around a predefined height z_d , starting from random initial positions and velocities. In this setting, the quadrotor is initialized with a random position, orientation, and linear and angular velocities. The observation contains $\mathbf{o}_h = [\mathbf{p}, \tilde{\mathbf{R}}, \mathbf{v}, \boldsymbol{\omega}, a_{\text{prev}}, \ddot{\mathbf{p}}_{WB}, z_d]$. Here $\ddot{\mathbf{p}}_{WB}$ is defined as the acceleration of the quadrotor agent in the world frame, and z_d represents a predefined constant height in meters to stabilize at. The reward function $r_t^{\text{stabilize}}$ is defined as

$$r_t^{\text{stabilize}} = r_t^{\text{height}} + r_t^{\text{attitude}} + r_t^{\text{velocity}} + r_t^{\text{br}} + r_t^{\text{act}} + r_t^{\text{success}}. \quad (5)$$

Here r_t^{height} rewards the quadrotor for maintaining a constant height, r_t^{attitude} rewards the quadrotor for maintaining a constant orientation, r_t^{velocity} and r_t^{angular} penalizes the non-zero linear and angular velocities, r_t^{act} penalizes the non-smooth actions, and r_t^{success} is a discrete reward when the quadrotor is stabilized. All the rewards are formulated here using the \mathcal{L}_2 norm multiplied by a corresponding constant coefficient. The

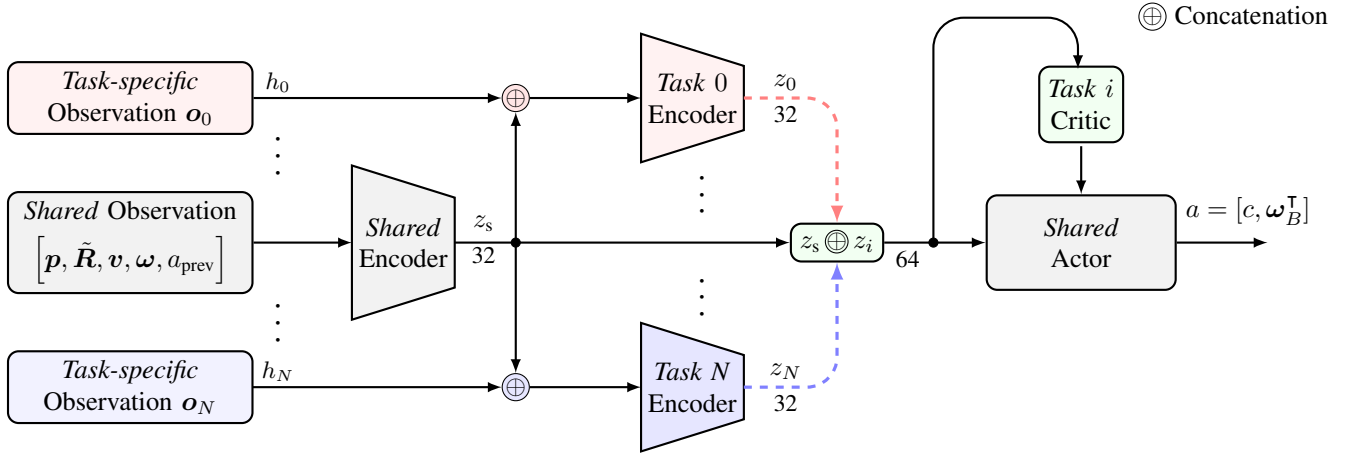


Fig. 3: Our MTRL framework utilizes a shared encoder for observations related to the quadrotor dynamics across all tasks. The embedding output from the shared encoder is then merged with the task-specific observation (e.g., the gate observation from the racing task and the desired velocity from the tracking task) to create a task-specific embedding. The policy uses both the concatenated embedding (64) from the shared embedding (32) and the task-specific embedding (32) to generate control commands. A separate critic function is used for each task, which is not employed during deployment.

detailed rewards are defined as

$$\begin{aligned}
 r_t^{\text{height}} &= \beta_1 \|z_t - z_d\|, \\
 r_t^{\text{attitude}} &= \beta_2 \|\mathbf{R}_t\|, \\
 r_t^{\text{velocity}} &= \beta_3 \|\mathbf{v}_t\|, \\
 r_t^{\text{br}} &= \beta_4 \|\boldsymbol{\omega}_{B,t}\|, \\
 r_t^{\text{act}} &= \beta_5 \|\mathbf{u}_t - \mathbf{u}_{t-1}\|, \\
 r_t^{\text{success}} &= \beta_6 \quad \text{if robot remains hovering.}
 \end{aligned} \tag{6}$$

During training, we applied a curriculum to gradually increase the difficulty of the tasks. We increase the initial speed of the quadrotor in the x , y , and z axes by 10% for every 100,000 data samples. The curriculum will stop once the predefined upper limits in each direction are reached.

4) *Velocity Tracking*: In the velocity tracking task, the quadrotor is required to track a randomly generated velocity profile. The observation contains $\mathbf{o}_v = [\mathbf{p}, \tilde{\mathbf{R}}, \mathbf{v}, \boldsymbol{\omega}, a_{\text{prev}}, \mathbf{v}_d, \dot{\mathbf{p}}_{WB}]$, where \mathbf{v}_d represent the desired linear velocity. The reward function r_t^{tracking} is defined as

$$r_t^{\text{tracking}} = r_t^{\text{velocity}} + r_t^{\text{br}} + r_t^{\text{act}}, \tag{7}$$

where r_t^{velocity} rewards the quadrotor for tracking the desired velocity, r_t^{br} penalizes the quadrotor for non-zero angular velocities, and r_t^{act} penalizes the quadrotor for excessive actions. All the rewards are formulated here using the \mathcal{L}_2 norm multiplied by a corresponding constant coefficient,

$$\begin{aligned}
 r_t^{\text{velocity}} &= \lambda_1 \|\mathbf{v}_t - \mathbf{v}_d(t)\|, \\
 r_t^{\text{br}} &= \lambda_2 \|\boldsymbol{\omega}_{B,t}\|, \\
 r_t^{\text{act}} &= \lambda_3 \|\mathbf{u}_t - \mathbf{u}_{t-1}\|.
 \end{aligned} \tag{8}$$

During training, we implemented a curriculum to progressively increase the difficulty of the tasks. The desired speed of the quadrotor in the x , y , and z axes was increased by 1m s^{-1} for every 100,000 data samples. This gradual increase will continue until the predefined upper-speed limits in each direction are achieved.

D. Multi-task Learning Framework

Since the control or navigation tasks are typically not contact-rich for a fixed quadrotor platform, we generally do not expect changes in the dynamics equations (see Eq. 1), whether during different phases of a single task or across various tasks. Many previous works also present this valid assumption, from agile drone racing to obstacle avoidance [1], [4], [5]. Hence, in the quadrotor setting, we can utilize the consistent physical property to simplify the multi-task scenario by assuming an identical transition probability across all tasks. This assumption serves as the primary motivation for our proposed information-sharing structure in the MTRL setup, allowing us to share data samples across tasks to efficiently learn an encoding network for the observation information related to the transition dynamics. To leverage the shared information across tasks, we propose a multi-task learning framework that consists of shared and task-specific modules.

As we aim to have a single policy capable of solving multiple tasks, our actor needs to be common to all tasks. As shown in the previous sections, the observation space in different tasks shares some common features, namely the position, orientation, linear and angular velocities, and the previous action. The selection of the shared features is based on the fact that these represent the essential dynamical properties outlined in Eq. 1, which remain consistent across different tasks. Other features can be task-specific and have different dimensions, which would lead to different tasks having different policy input sizes. To distinguish observations from various potential tasks, we use an identifier based on the task-specific observation length, which allows us to incorporate one-hot encoding into the task-specific observations. To overcome this, we propose a feature encoding architecture including a shared encoder network among different tasks to extract the shared features from the observation. Meanwhile, we use task-specific encoders to generate the task-specific policy inputs. Although the task-specific observations' dimensions vary from task to

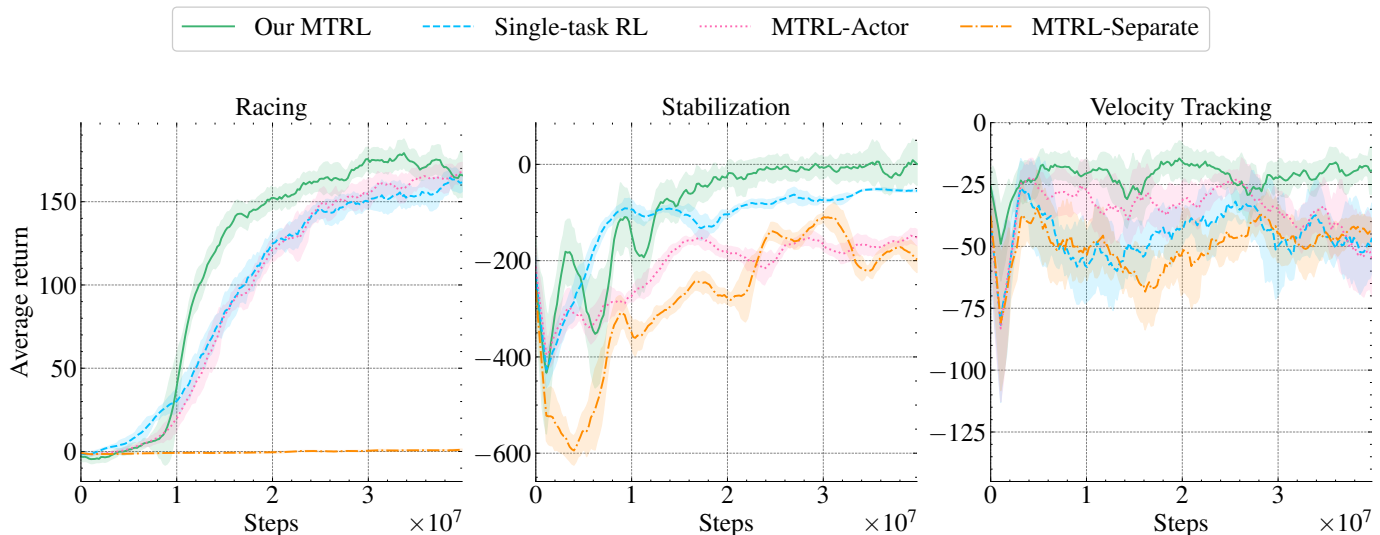


Fig. 4: Overview of the average return comparison of different tasks. It is clearly shown that our proposed MTRL approach achieves a higher average return within the same number of training steps compared to single-task RL baselines. Notably, single-task RL policies still perform comparably to the MTRL approach when only the actor network is shared.

task, the task-specific encoder is designed to map the specific observation to the same latent space.

For the MTRL part, we employ a shared actor policy with multiple individual critic networks, each corresponding to a specific task. The actor policy takes the concatenated latent features from the shared encoder and the task-specific encoder as input and outputs the action, namely the collective thrust and the body rates [4]. The overall architecture is shown in Fig. 3. We train the shared actor policy to maximize the expected return across all tasks, while the individual critic networks are trained to evaluate the value function for each task. In contrast, the task-specific critic networks take directly the full observation as the input and output of the value function for each task. Based on [25], we went for the choice not sharing the policy feature encoder with the critic networks, as it has been shown to improve the performance of the policy.

IV. EXPERIMENTS

Using the individual tasks described in the previous section, we evaluate the performance of the proposed MTRL approach. Our experiments are designed to answer the following research questions: (i) How sample-efficient is our MTRL approach compared to the single-task RL baselines? (ii) How does the MTRL policy’s performance compare to the single-task RL policies? (iii) How do different knowledge-sharing strategies affect the MTRL performance? (iv) Does the result transfer to a real-world scenario?

A. Training Configurations

For the policy training, we employ a policy network consisting of a two-layer MLP, each layer containing 256 neurons, with a final layer outputting a 4-dimensional vector using a *tanh* activation function. In our experiments, for the shared dynamic encoder, we use a three-layer MLP with 19 neurons in the input layers and 128 neurons in the hidden layers to generate a 32-dimensional latent embedding. For the task-specific

encoder, we use a three-layer MLP with task-dependent input dimensions and 128 neurons in the hidden layer to generate a 32-dimensional latent embedding. Table I shows the detailed task reward parameters for the MTRL training. In our setting, we optimized our hyperparameters solely based on the single-task performance. For MTRL training, we employ model-free reinforcement learning approach using Proximal Policy Optimization (PPO) [26]. For the quadrotor platform used for both training and deployment, we detail the information regarding components and physical parameters in Table II. We use the Flightmare simulator [27] for policy training in simulation.

B. Baselines

In our experiments, we compare our approach with the following baselines: (i) **Single-task RL**: We train a separate policy for each task using the same RL algorithm as the MTRL approach. (ii) **MTRL-Actor**: For this baseline, we keep the *Actor* as the only network shared among different tasks. And all the other encoder networks and critic networks are different among all tasks. We train a shared actor policy with multiple individual critic networks, each corresponding to a specific task. (iii) **MTRL-Separate**: To ablate the design choice of fusing shared and task-specific observations, we train a shared encoder and task-specific encoders, but the task encoder does

TABLE I: Reward parameters for MTRL training.

Param.	Value	Param.	Value	Param.	Value
α_1	0.5	β_1	-2e-3	λ_1	-2e-4
α_2	0.025	β_2	-2e-4	λ_2	-1.2e-3
α_3	-1.0	β_3	-4e-5	λ_3	-1e-4
α_4	-2e-4	β_4	-1e-5		
α_5	-5e-4	β_5	-1e-4		
α_6	-5	β_6	10		
α_7	-10				

TABLE II: Overview of the drone parameters for both simulation and real-world experiments.

Param.	Agilicious Platform
Thrust-to-weight Ratio	5.78
Mass [kg]	0.6
Maximum Thrust [N]	20.00
Arm Length [m]	0.15
Inertia [g m^2]	[2.50, 2.51, 4.32]
Motor Time Constant [s]	0.033

not receive the shared observation as input. The observation encoders are different for each task, and the shared actor policy takes the latent embedding directly as input to output the action. For all of the approaches, we performed 10 runs of the same training configurations using different random seeds, and we report the average evaluation metrics in this section. Since the primary contribution of our approach is to enhance learning efficiency and task performance of the RL policies using MTRL, we have chosen not to include baselines that rely solely on classical control approaches in our comparison.

C. Sample Efficiency Analysis

To evaluate the sample efficiency of our MTRL approach, we compare the return curves of the MTRL approach with the aforementioned baselines. All the policies are trained using the same hyperparameters and the same number of training steps. The number of training steps in our setting is 40M, where all the policies' performances converge. As shown in Fig. 4, the MTRL approach outperforms the single-task RL baselines in all three different tasks. The MTRL approach achieves a higher average return in the same number of training steps than single-task RL baselines. Notably, the single-task RL policies still perform closely to the MTRL approach when only sharing the actor network. This highlights the necessity of information sharing in our framework. However, although information sharing is beneficial, the MTRL-Separate cannot fly at all in the racing task. This is likely because if we do not fuse the shared information with task-specific information, it becomes difficult when the task requirements are conflicting, e.g. racing and stabilization from high speed. The policy will then prefer learning the rather easier task. Hence, this highlights the importance of fusing the shared and task-specific information in the MTRL framework.

D. Individual Task Performance of MTRL Policy

In this section, we showcase the individual task performance of our MTRL policy. We evaluate the performance of the MTRL policy in three different tasks: stabilization from high speed, racing, and velocity tracking. The MTRL policy is trained using the same hyperparameters as the single-task RL policies.

1) *Racing Performance*: In the racing task, we evaluate the performance of the MTRL policy by racing on a predefined race track. The race track contains 6 gates, and the quadrotor is required to pass these gates in a fixed order. The policy is evaluated in 64 different starting positions in uniformly sampled starting positions. Fig. 5 visualizes one sample rollout that successfully completes the race track.

2) *Stabilization Performance*: As shown in Fig. 6 (a), we evaluate the high-speed stabilization task by randomly initializing the quadrotor with various positions, orientations, and velocities. The initial speeds in the x and y directions are randomly sampled from $[-20, 20] \text{ m s}^{-1}$, while the velocity in the z direction is sampled from $[-4, 4] \text{ m s}^{-1}$. To prevent the quadrotor from crashing into the ground when the initial z -direction velocity is high and negative, we adjust the initial height accordingly to ensure the quadrotor will not crash within one second, even if no control input is applied. From the results, we observe that the MTRL policy can successfully stabilize the quadrotor in the hovering condition from high speed task within seconds, even when the initial conditions are challenging; the maximum initial speed from our experiments is 82.57 km h^{-1} . This demonstrates the robustness of the MTRL policy in stabilization from high-speed tasks.

3) *Velocity Tracking Performance*: In the tracking task, we evaluate the performance of the MTRL policy by tracking a randomly generated velocity command, where we apply random walks in the acceleration space. We randomize the velocity references in x and y directions up to 54 km h^{-1} and 18 km h^{-1} in z direction. As shown in Fig. 6 (b), the MTRL policy can successfully track the challenging commanded velocities, where they can go up to 50 km h^{-1} in a very short time, and we did not even include the non-holonomic constraints of the quadrotor to generate velocity command.

E. Quantitative Analysis

1) *Evaluation Metrics*: For the autonomous racing task, we use three evaluation metrics: success rate (SR), mean-gate-passing-error (MGE), and lap time (LT). SR is the ratio of completed laps to total trials. MGE measures the distance between the drone's position and the gate center when passing through; here, the inner gate size used for experiments is 1.5m. LT indicates the duration of completing a full race track and flying through all gates. For the stabilization task, we employ two evaluation metrics, namely t_{half} and t_{full} , to evaluate the time usage of the RL controller to stabilize the quadrotor. Here

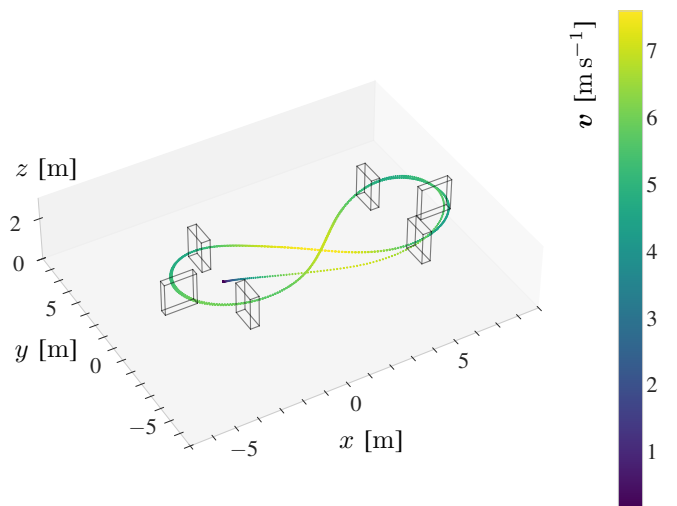


Fig. 5: Illustration of one racing policy rollout. The policy successfully completes a Figure-8 race track, which consists of six gates, with a 100% success rate.

TABLE III: Individual task performance after 20M and 40M training samples. For the stabilization task, we evaluate the time to reduce the velocity to half the initial velocity (t_{half}) and to hovering (t_{full}). For velocity tracking, we report the tracking error e_v . For racing, we report the success rate (SR), the mean gate passing error (MGE), and the lap time (LT).

Methods	20M Samples						40M Samples					
	SR [%]	Racing MGE [m]	Stabilization LT [s]	Stabilization t_{half} [s]	Stabilization t_{full} [s]	Tracking e_v [m s ⁻¹]	SR [%]	Racing MGE [m]	Stabilization LT [s]	Stabilization t_{half} [s]	Stabilization t_{full} [s]	Tracking e_v [m s ⁻¹]
Single-task RL	95	0.224	6.23	0.50	4.20	2.15	100	0.163	5.83	0.41	3.53	1.18
MTRL-Actor	94	0.231	6.09	0.61	4.57	2.22	100	0.159	5.88	0.39	3.29	1.44
MTRL-Separate	0	crash	crash	0.59	4.29	2.04	0	crash	crash	0.91	4.12	1.93
Ours	100	0.187	5.92	0.41	3.77	1.69	100	0.152	5.80	0.35	3.26	1.21

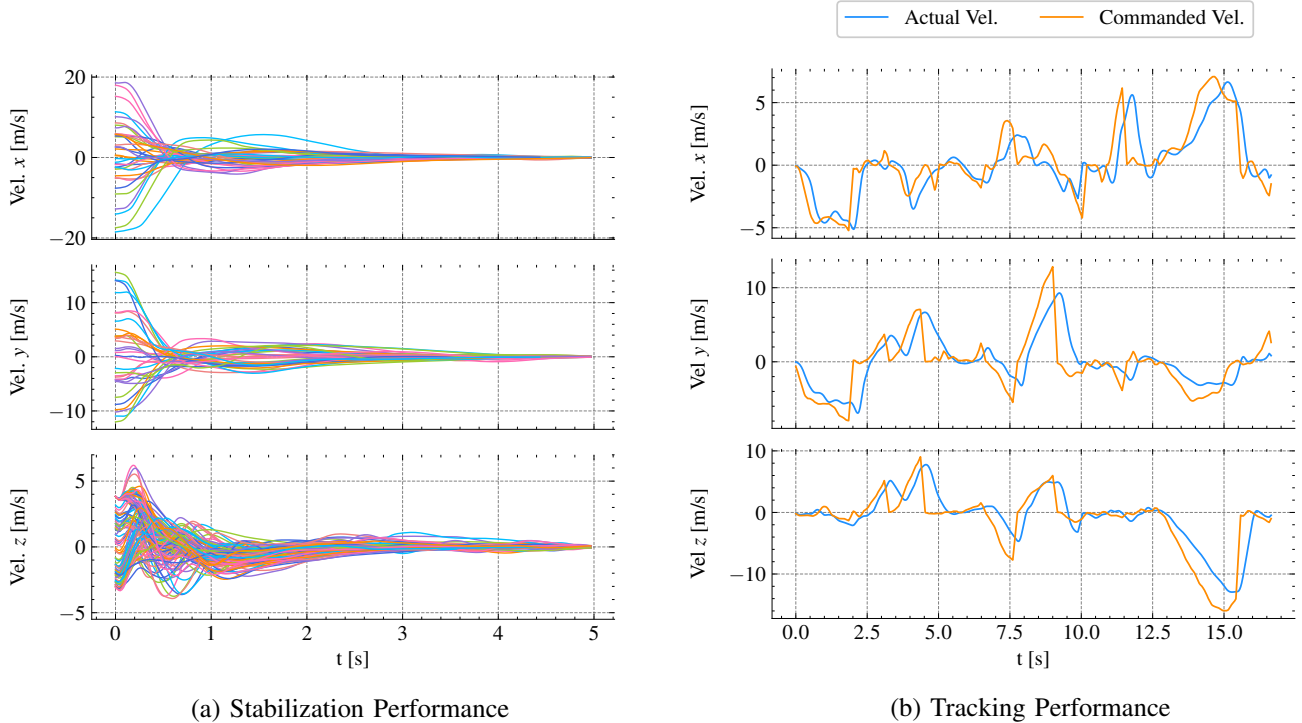


Fig. 6: Visualizations of the MTRL policy on individual task performance. (a) The MTRL policy successfully stabilizes the quadrotor in a hover within seconds, even from high-speed tasks and challenging initial conditions. (b) The MTRL policy successfully tracks the commanded velocity, even for challenging trajectories with speeds reaching up to 50 km h⁻¹.

t_{half} is the time usage to reduce the actual velocity to half of the initial velocity, and t_{full} is the time usage to control a quadrotor to hover condition; in our experiment, we determine this when the quadrotor’s linear velocity is smaller than 0.5m s⁻¹. For the velocity tracking task, we simply use a tracking error metric e_v , which computes the averaged velocity difference over time.

2) *Analysis*: In Table III, we present a quantitative comparison of our approaches with the baselines. To demonstrate the effectiveness of the policy trained with different numbers of samples, we list the performance of the MTRL policies in two different timesteps, namely 20M and 40M. First of all, for the policy trained with 20M steps, our MTRL approach demonstrates a much better task performance than all of the baseline approaches. This strongly demonstrates the sample efficiency of our approach. Secondly, when the policy is trained till convergence, the policy’s performance of our approach is still not worse than learning individually in all the tasks and all the metrics. This further indicates the claim of our approach: our MTRL framework could learn multiple tasks efficiently

without trading off performance.

F. Real World Performance

To demonstrate the effectiveness of our policy improvements, we conducted validation tests in real-world scenarios. We utilized an Agilicious quadrotor platform [28] with the identical properties presented in I with state estimation provided by a VICON motion capture system to feed accurate inputs to the policy. For low-level control, the BetaFlight2 firmware was employed to track the commanded collective thrusts and body rates. We conducted five individual runs for each task, varying the starting conditions in each run. Remarkably, our MTRL policy achieved a 100% success rate across all tasks in these real-world experiments, as illustrated in Fig. 1. These results clearly indicate that our policy effectively transfers to and performs reliably in real-world scenarios. The detailed evaluation results are presented in Table IV. The metrics for each task indicate that our policy consistently delivers stable performance both in simulation and the real world using identical configurations.

TABLE IV: Comparison of our MTRL policy’s task performance between simulation and real world.

Methods	Racing			Stabilization		Tracking
	SR [%]	Error [m]	Time [s]	t_{fast} [s]	t_{slow} [s]	e_v [m s ⁻¹]
Simulation	100%	0.152	5.80	0.15	1.03	0.99
Real-world	100%	0.189	5.94	0.17	1.19	0.95

G. Discussion

Our experiments thoroughly evaluated the proposed MTRL framework, addressing its critical performance aspects. First, we demonstrated that the MTRL approach significantly improves sample efficiency and task performance compared to single-task RL baselines. After training for half the amount of total samples (20M), our MTRL method reduced stabilization time by 18% (t_{half}) and gate passing error (MGE) by 16% compared even to individual RL methods. Furthermore, even when trained to convergence, our MTRL approach consistently performs better than, or as well as other baseline methods. Notably, it achieved a 6.7% reduction in gate passing error in racing tasks compared to the next best approach. In contrast, the MTRL-Separate baseline without the integration of shared and task-specific information resulted in a 0% success rate in the racing task, as it could not even complete the task. This failure underscores the importance of properly integrating shared and task-specific elements within our MTRL framework. Finally, the deployment of our MTRL policy in real-world scenarios confirmed its robustness, with performance closely matching that observed in simulations.

V. CONCLUSIONS

In this work, we introduced the first Multi-Task Reinforcement Learning (MTRL) framework specifically designed for quadrotor control, addressing the challenges posed by diverse task requirements in real-world scenarios. By leveraging the shared physical dynamics of the quadrotor and employing a novel multi-critic setup with a shared task-agnostic observation encoder, our approach successfully integrates information across different tasks while maintaining high performance. The experimental results, both in simulation and real-world applications, demonstrated the effectiveness and efficiency of our method, particularly in enhancing sample efficiency when learning different tasks while maintaining strong task performance. This advancement paves the way for more versatile quadrotor control systems capable of performing a wide range of tasks within a single mission, thereby significantly contributing to the broader application of quadrotors in critical areas like search and rescue and infrastructure inspection.

REFERENCES

- [1] Y. Song, A. Romero, M. Mueller, V. Koltun, and D. Scaramuzza, “Reaching the limit in autonomous racing: Optimal control versus reinforcement learning,” *Science Robotics*, p. adg1462, 2023.
- [2] J. Xing, G. Cioffi, J. Hidalgo-Carrió, and D. Scaramuzza, “Autonomous power line inspection with drones via perception-aware mpc,” in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2023, pp. 1086–1093.
- [3] J. Xing, L. Bauersfeld, Y. Song, C. Xing, and D. Scaramuzza, “Contrastive learning for enhancing robust scene transfer in vision-based agile flight,” in *2024 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2024.
- [4] E. Kaufmann, L. Bauersfeld, and D. Scaramuzza, “A benchmark comparison of learned control policies for agile quadrotor flight,” in *International Conference on Robotics and Automation (ICRA)*, 2022.
- [5] E. Kaufmann, L. Bauersfeld, A. Loquercio, M. Müller, V. Koltun, and D. Scaramuzza, “Champion-level drone racing using deep reinforcement learning,” *Nature*, vol. 620, no. 7976, pp. 982–987, Aug 2023.
- [6] H. Wang, J. Xing, N. Messikommer, and D. Scaramuzza, “Environment as policy: Learning to race in unseen tracks,” *arXiv preprint arXiv:2410.22308*, 2024.
- [7] B. Liu, X. Liu, X. Jin, P. Stone, and Q. Liu, “Conflict-averse gradient descent for multi-task learning,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 878–18 890, 2021.
- [8] D. Kalashnikov, J. Varley, Y. Chebotar, B. Swanson, R. Jonschkowski, C. Finn, S. Levine, and K. Hausman, “Scaling up multi-task robotic reinforcement learning,” in *Conference on Robot Learning*, 2022.
- [9] A. Gupta, J. Yu, T. Z. Zhao, V. Kumar, A. Rovinsky, K. Xu, T. Devlin, and S. Levine, “Reset-free reinforcement learning via multi-task learning: Learning dexterous manipulation behaviors without human intervention,” in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6664–6671.
- [10] C. D’Eramo, D. Tateo, A. Bonarini, M. Restelli, and J. Peters, “Sharing knowledge in multi-task deep reinforcement learning,” *International Conference on Learning Representations*, 2020.
- [11] E. Aljalbout, N. Sotirakis, P. van der Smagt, M. Karl, and N. Chen, “Limt: Language-informed multi-task visual world models,” *arXiv preprint arXiv:2407.13466*, 2024.
- [12] T. Yu, D. Quillen, Z. He, R. Julian, K. Hausman, C. Finn, and S. Levine, “Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning,” in *Conference on robot learning*, 2020.
- [13] J. Hwangbo, I. Sa, R. Siegwart, and M. Hutter, “Control of a quadrotor with reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 2, no. 4, pp. 2096–2103, 2017.
- [14] N. O. Lambert, D. S. Drew, J. Yaconelli, S. Levine, R. Calandra, and K. S. Pister, “Low-level control of a quadrotor with deep model-based reinforcement learning,” *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4224–4230, 2019.
- [15] G. Zhao, T. Wu, Y. Chen, and F. Gao, “Learning speed adaptation for flight in clutter,” *IEEE Robotics and Automation Letters*, 2024.
- [16] Z. Huang, Z. Yang, R. Krupani, B. Şenbaşlar, S. Batra, and G. S. Sukhatme, “Collision avoidance and navigation for a quadrotor swarm using end-to-end deep reinforcement learning,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2024.
- [17] J. Eschmann, D. Albani, and G. Loianno, “Learning to fly in seconds,” *IEEE Robotics and Automation Letters*, 2024.
- [18] F. Sadeghi and S. Levine, “CAD2RL: Real single-image flight without a single real image,” in *Robotics: Science and Systems (RSS)*, 2017.
- [19] J. Xing, A. Romero, L. Bauersfeld, and D. Scaramuzza, “Bootstrapping reinforcement learning with imitation for vision-based agile flight,” *Conference on Robot Learning*, 2024.
- [20] I. Geles, L. Bauersfeld, A. Romero, J. Xing, and D. Scaramuzza, “Demonstrating agile flight from pixels without state estimation,” in *Proceedings of Robotics: Science and Systems*, 2024.
- [21] R. Yang, H. Xu, Y. Wu, and X. Wang, “Multi-task reinforcement learning with soft modularization,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4767–4777, 2020.
- [22] A. Xie and C. Finn, “Lifelong robotic reinforcement learning by retaining experiences,” in *Conference on Lifelong Learning Agents*, 2022.
- [23] Y. Song, M. Steinweg, E. Kaufmann, and D. Scaramuzza, “Autonomous drone racing with deep reinforcement learning,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2021.
- [24] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, “On the continuity of rotation representations in neural networks,” in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.
- [25] M. Andrychowicz, A. Raichuk, P. Stańczyk, M. Orsini, S. Girgin, R. Marinier, L. Hussenot, M. Geist, O. Pietquin, M. Michalski, *et al.*, “What matters in on-policy reinforcement learning? a large-scale empirical study,” *arXiv preprint arXiv:2006.05990*, 2020.
- [26] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal policy optimization algorithms,” *arXiv*, 2017.
- [27] Y. Song, S. Naji, E. Kaufmann, A. Loquercio, and D. Scaramuzza, “Flightmare: A flexible quadrotor simulator,” in *Conference on Robot Learning*, 2020.
- [28] P. Foehn, E. Kaufmann, A. Romero, R. Penicka, S. Sun, L. Bauersfeld, T. Laengle, G. Cioffi, Y. Song, A. Loquercio, and D. Scaramuzza, “Agilicious: Open-source and open-hardware agile quadrotor for vision-based flight,” *Science Robotics*, vol. 7, no. 67, 2022.