

# Quadrotor Navigation using Reinforcement Learning with Privileged Information

Jonathan Lee, Abhishek Rathod, Kshitij Goel, John Stecklein, and Wennie Tabib

**Abstract**—This paper presents a reinforcement learning-based quadrotor navigation method that leverages efficient differentiable simulation, novel loss functions, and privileged information to navigate around large obstacles. Prior learning-based methods perform well in scenes that exhibit narrow obstacles, but struggle when the goal location is blocked by large walls or terrain. In contrast, the proposed method utilizes time-of-arrival (ToA) maps as privileged information and a yaw alignment loss to guide the robot around large obstacles. The policy is evaluated in photo-realistic simulation environments containing large obstacles, sharp corners, and dead-ends. Our approach achieves an 86% success rate and outperforms baseline strategies by 34%. We deploy the policy onboard a custom quadrotor in outdoor cluttered environments both during the day and night. The policy is validated across 20 flights, covering 589 m without collisions at speeds up to 4 m/s.

## I. INTRODUCTION

Traditional navigation approaches decompose perception, planning, state estimation, and control into separate tasks. However, end-to-end learning-based methods, which use a neural network to convert raw sensor observations into actions, are increasingly being deployed for high-speed autonomous flight [1, 2, 3]. The advantages of end-to-end approaches are reduced planning latency and lower computational overhead, which enables deployment on lightweight and low-cost platforms. However, these approaches either require large amounts of expert-labeled data or struggle to find paths in challenging environments that exhibit large obstacles, sharp corners, and dead ends.

To bridge these gaps in the state of the art, we extend the method of Zhang et al. [2] to enable navigation around large obstacles. Like Zhang et al. [2], our policy is reactive at deployment, relying only on depth observations and state estimates. However, by training with ToA maps as privileged information, our policy learns globally-aware navigation behavior without requiring a ToA map at test time. We provide the following contributions:

- 1) an objective function for predicting heading (i.e., yaw) that improves navigation performance relative to state-of-the-art approaches in environments that require changes in orientation (e.g., twisting passageways and sharp corners);
- 2) a method to leverage a time-of-arrival map as privileged information during training, enabling shortest-path navigation without an explicit map at test time;
- 3) approaches to bridge the sim-to-real gap via body rate attitude control and domain randomization; and

The authors are with the Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213 USA. {jlee6@alumni.cmu.edu, {arathod2, kgoel1, jsteckle, wtabib}@andrew.cmu.edu.

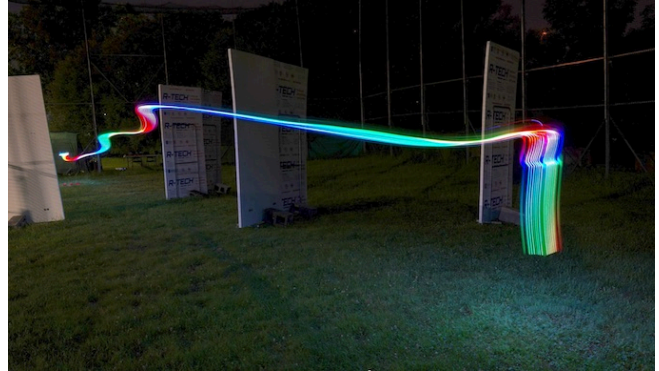


Fig. 1: This paper develops and deploys an end-to-end policy to navigate in challenging environments. The approach outperforms the state of the art by 34%. An example trajectory is captured using long-exposure photography.

- 4) extensive evaluation of the system in photo-realistic simulation and hardware experiments<sup>1</sup> as well as an open-source software release<sup>2</sup>.

## II. RELATED WORK

Loquercio et al. [1] demonstrate high-speed, learning-based navigation by training a student policy in simulation using a privileged, sample-based expert planner. The student learns to predict control points of a polynomial trajectory based on top-ranked expert samples. This approach requires a large amount of expert-labeled data and increases the training overhead when the expert planner is computationally intensive. Kim et al. [4] present an Inverse Reinforcement Learning (IRL) method that similarly relies on expert demonstrations, but instead of directly imitating the expert, a reward function is inferred that explains the observations. Empirically, this approach enables out-of-distribution generalization. However, the learnt policy lacks temporal awareness during inference, making it susceptible to collisions when navigating around large obstacles.

An alternative to relying on expert data is to train a policy by backpropagating reward gradients through a differentiable simulator [5]. Zhang et al. [2] leverage this idea by training a recurrent neural network control policy with a lightweight point-mass dynamics model and rendering engine. This method enables reliable operation at high speeds in several environments. However, the learnt policy does not generalize to scenarios where significant yaw motion is required.

<sup>1</sup>A video of the experiments may be found at <https://youtu.be/RbHJ69o-zUc>.

<sup>2</sup><https://github.com/rislalab/depthnav>

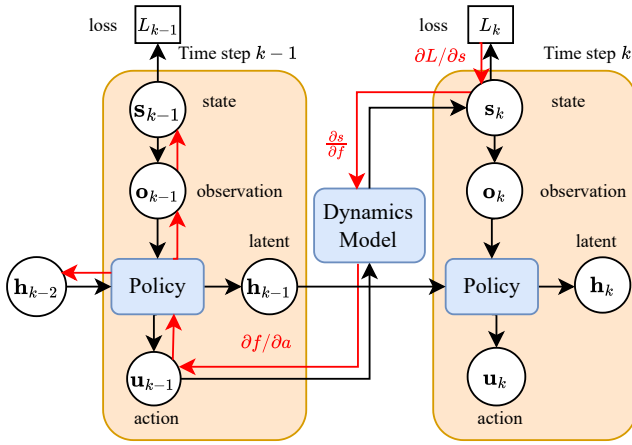


Fig. 2: Differentiable dynamics enables direct policy updates by performing gradient descent on the loss function.

Recent learning-based navigation methods exploit geometric maps as privileged information available during training but absent at inference [3, 6]. YOPO [3] trains a neural policy to generate motion primitive offsets and scores using a Euclidean signed distance field (ESDF) cost map as supervision. However, this approach provides only limited guidance for escaping large convex obstacle regions.

Our approach addresses two limitations of prior work. First, Zhang et al. [2] maintains a fixed heading toward the target, limiting navigation around large obstacles. Our yaw alignment loss directly supervises heading prediction, enabling the robot to yaw while changing direction. Second, as ESDFs encode obstacle distance rather than path direction, Lu et al. [3] provides local avoidance but not globally optimal routing. Our ToA maps address this by guiding the robot along shortest paths. Together, the yaw alignment loss enables reorientation while the ToA gradient teaches the robot which direction to head to make progress towards the goal.

### III. METHODOLOGY

The proposed navigation policy takes a depth image, target information, and the quadrotor state as inputs and uses a neural network to predict a thrust and yaw angle.

#### A. Differentiable Dynamics

Following the approach of Zhang et al. [2], the quadrotor navigation problem is formulated as a Markov decision process with a discrete-time dynamical system. As shown in Fig. 2, observations,  $\mathbf{o}_k$ , are generated at each time step from sensor measurements based on the current state,  $\mathbf{s}_k$ . The policy takes the observation  $\mathbf{o}_k$  and a hidden state  $\mathbf{h}_{k-1}$  as inputs and outputs a new hidden state  $\mathbf{h}_k$  and action  $\mathbf{u}_k$ . The action  $\mathbf{u}_k \in \mathbb{R}^4$  contains the mass-normalized thrust vector  $\mathbf{t}_k \in \mathbb{R}^3$  and a predicted yaw angle  $\psi_k$ , the latter is absent in [2]. The action  $\mathbf{u}_k$  is the control input to the system dynamics  $\mathbf{s}_{k+1} = f(\mathbf{s}_k, \mathbf{u}_k)$ , which determines the next state,  $\mathbf{s}_{k+1}$ . At each time step  $k$ , the agent also receives a loss value  $L_k$  dependent on the current state and action.

Notably, the system dynamics are differentiable with respect to  $\mathbf{s}_k$  and  $\mathbf{u}_k$ , enabling Analytical Policy Gradient

(APG) methods to train the model by backpropagating the loss through the dynamics [5, 7]. This allows for sample efficient training, as even a single sample provides a usable gradient for policy optimization.

While the full quadrotor dynamics model is differentiable, it has been shown that using point-mass dynamics results in reduced training time without sacrificing performance [8]. Therefore, we leverage a point-mass model with a velocity Verlet integration scheme,

$$\mathbf{p}_{k+1} = \mathbf{p}_k + \mathbf{v}_k \Delta t + \frac{1}{2} \mathbf{a}_k (\Delta t)^2 \quad (1)$$

$$\mathbf{v}_{k+1} = \mathbf{v}_k + \frac{\mathbf{a}_k + \mathbf{a}_{k+1}}{2} \Delta t \quad (2)$$

$$\mathbf{a}_{k+1} = \mathbf{t}_{k+1} - [0, 0, g]^\top, \quad (3)$$

where  $\mathbf{p}_k \in \mathbb{R}^3$ ,  $\mathbf{v}_k \in \mathbb{R}^3$ , and  $\mathbf{a}_k \in \mathbb{R}^3$  denote the position, velocity, and acceleration of the point mass at time step  $k$ , respectively. The time step,  $\Delta t$ , is fixed during integration. The constant  $g$  denotes gravity. Since point-mass dynamics do not maintain an explicit orientation, the orientation is determined uniquely by the simulator at every time step. The basis vectors for the body frame are defined such that the body  $z$ -axis points along the predicted thrust vector  $\mathbf{t}_k$  and the  $x$ -axis is aligned with the predicted yaw angle  $\psi_k$ .

#### B. Network Architecture

The network takes as input a depth image, a target velocity, goal importance value, and the robot's state (see Fig. 3). Each input is processed by a dedicated feature extractor and projected into a 192-dimensional vector. These vectors are layer-normalized and summed to form the input to a gated recurrent unit (GRU) [9] cell. The GRU combines this input with the previous hidden state to produce a new hidden state, which encodes relevant features from both current and past observations.

From this latent representation, a linear layer predicts the desired mass-normalized thrust and yaw angle. By maintaining a memory of past observations and actions, the latent representation enables the policy to generate smooth and consistent control outputs.

The state consists of the current velocity and orientation of the robot. The target is defined by a desired velocity vector towards the goal and a goal importance value, defined as the reciprocal of the goal distance. This importance value provides contextual information about the goal's proximity relative to obstacles in the depth image, helping the network distinguish whether the goal lies in front of or behind an obstacle. All input vectors are expressed with respect to the starting frame of the robot. Our network differs from that of [2] in that we use a separate target extractor which includes a goal importance value, we use normalization layers to balance features, and we predict a yaw angle in addition to mass-normalized thrust.

#### C. Loss Functions

This section describes the individual loss (or reward) functions applied at each time step. Together, these terms

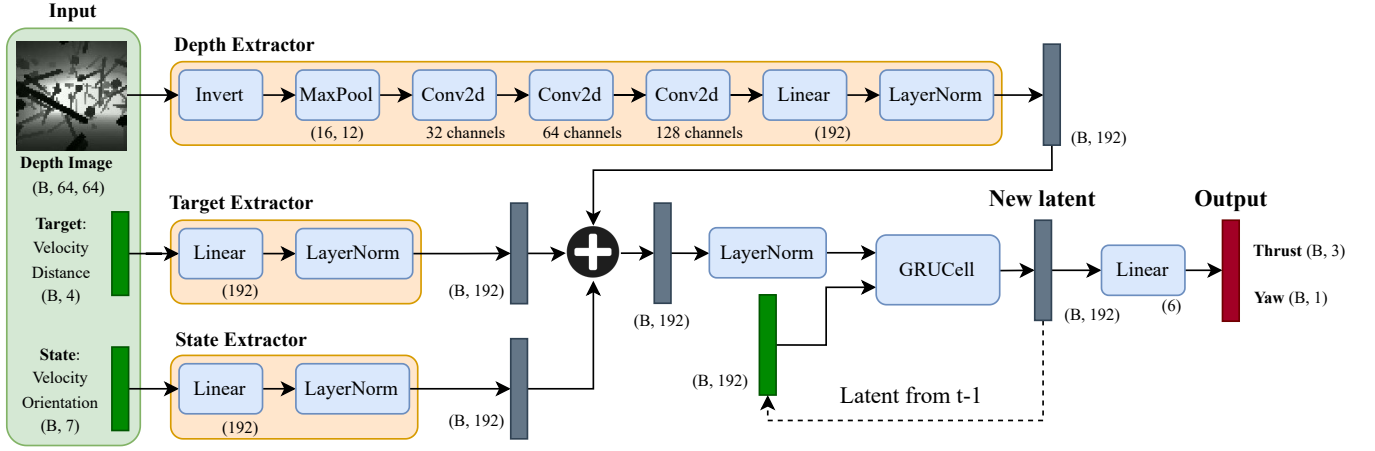


Fig. 3: The end-to-end planning and control architecture is trained as a single neural network. Feature extractors process each input before they are flattened and summed together. A GRUCell helps to maintain consistent action predictions over time.

guide the training of a stable and effective navigation policy. The loss functions developed by Zhang et al. [2] are restated for the sake of completeness. The losses detailed in Sects. III-C.4 and III-C.5 as well as new terms are noted as novel contributions of this work.

1) *Obstacle Avoidance Loss:*

$$L_{\text{clearance}} = \frac{1}{T} \sum_{k=1}^T \beta_1 \ln(1 + e^{\beta_2(d_k - r)}) \quad (4)$$

$$L_{\text{collision}} = \frac{1}{T} \sum_{k=1}^T \|\mathbf{v}_k^c\| \max(1 - (d_k - r), 0)^2 \quad (5)$$

$L_{\text{clearance}}$  is a softplus penalty on the distance  $d_k$  to the nearest obstacle, encouraging the robot to maintain a safe clearance. The scalar  $r$  is the robot radius.  $L_{\text{collision}}$  penalizes the magnitude of the component of velocity  $\mathbf{v}_k^c$  directed toward the closest obstacle. These losses train the policy to keep a safe distance and reduce forward velocity when approaching obstacles. These terms are restated from [2].

2) *Smoothness Loss:*

$$L_{\text{acc}} = \frac{1}{T} \sum_{k=1}^T \|\mathbf{a}_k\|^2 \quad (6)$$

$$L_{\text{jerk}} = \frac{1}{T-1} \sum_{k=1}^{T-1} \left\| \frac{\mathbf{a}_{k+1} - \mathbf{a}_k}{\Delta t} \right\|^2 \quad (7)$$

$$L_{\omega} = \frac{1}{T} \sum_{k=1}^T \|\omega\|^2 \quad (8)$$

Since the policy only predicts a single action at a time, it is important that the sequence of actions over time form a smooth trajectory that can be executed safely. The loss function includes penalties on the  $\ell^2$ -norm of linear acceleration and jerk in the inertial frame, as well as angular velocity in the body frame. Specifically,  $L_{\text{acc}}$  encourages stability near hover at the target (see Sect. V-A), while  $L_{\omega}$  reduces abrupt yaw changes by discouraging large angular accelerations. Equations (6) and (7) are restated from [2], while Eq. (8) is a new penalty term where  $\omega$  is computed

via the Euler angle Jacobian (see Eq. (15)).

3) *Target Velocity Loss:*

$$L_v = \frac{1}{T} \sum_{k=1}^T \text{Smooth L1}(\|\mathbf{v}_k^{\text{set}} - \bar{\mathbf{v}}_k\|, 0) \quad (9)$$

$$L_{v_{\text{max}}} = \frac{1}{T} \sum_{k=1}^T \max(\|\mathbf{v}_k\| - v_{\text{max}}, 0)^2 \quad (10)$$

Equation (9) encourages the policy to track the target velocity  $\mathbf{v}_k^{\text{set}}$ . To allow flexibility for obstacle avoidance while maintaining long-term progress, we use a 2s moving average of the actual velocity,  $\bar{\mathbf{v}}_k$ , in the loss. This smooths short-term deviations and helps the policy stay aligned with the overall velocity objective. Consistent with findings in Zhang et al. [2], this averaging also reduces high-frequency velocity oscillations during rollouts under full rigid-body dynamics. Additionally, we introduce a new penalty term  $L_{v_{\text{max}}}$  for speeds exceeding the maximum target speed  $v_{\text{max}}$ , which helps bound the predicted thrusts and prevents policy divergence.

4) *Yaw Alignment Loss:*

$$L_{\text{yaw}} = -\frac{1}{T} \sum_{k=1}^T \mathbf{x}_k^B \cdot \tilde{\mathbf{v}}_k \quad (11)$$

The yaw alignment loss  $L_{\text{yaw}}$  is defined as the negative inner product between the body  $x$ -axis  $\mathbf{x}_k^B$  and the exponentially weighted moving average of the velocity  $\tilde{\mathbf{v}}_k$ . This loss term, introduced in this work, enables the quadrotor to reorient towards its desired direction of motion to navigate around large obstacles.

5) *Privileged Information via ToA Map:* Existing obstacle avoidance and collision loss terms in prior work [2] are insufficient for navigation around large obstacles. Local collision losses penalize proximity or approach speed to surfaces but do not actively encourage progress or escape, which can cause the robot to become stuck in concave obstacle regions. To address these limitations, we use time-of-arrival (ToA) maps, whose gradients guide the robot's velocity toward the goal while avoiding obstacles (see Fig. 4). Through

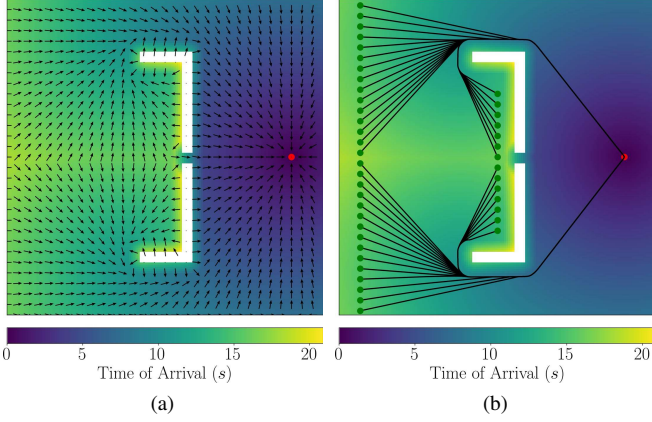


Fig. 4: (a) Heatmap of time-of-arrival (ToA) computed using fast marching method (FMM) and overlaid gradient field. (b) Shortest paths along ToA gradient from starting points (green dots) to the target (red dot) guides robot around concave obstacle regions.

the ToA loss, the policy learns to infer optimal navigation directions from depth alone, requiring no map at inference.

Let  $T(\mathbf{x})$  denote the ToA map, defined as the minimum travel time from a point  $\mathbf{x}$  in free space to the goal position. Travel times are solutions to the Eikonal equation  $|\nabla T(\mathbf{x})|F(\mathbf{x}) = 1$ , which governs how travel time propagates from the goal under a spatially varying speed  $F(\mathbf{x}) > 0$ .

While ToA maps naturally account for obstructions due to obstacles, they do not explicitly consider the risk of being too close to obstacle surfaces, which can produce unrealistic paths through narrow openings. To address this, we define  $F(\mathbf{x})$  as a piecewise continuous cost function (Eq. (12)) that reduces wavefront speed near obstacles and biases the resulting time-optimal paths to maintain safer distances.

$$F(\mathbf{x}) = \begin{cases} md(\mathbf{x}) + (v_{\text{slow}} - mr) & \text{if } d(\mathbf{x}) \leq d_{\text{safe}} \\ 1 & \text{if } d(\mathbf{x}) > d_{\text{safe}} \end{cases} \quad (12)$$

$$m = \frac{d_{\text{safe}} - v_{\text{slow}}}{d_{\text{safe}} - r}, \quad (13)$$

$r$  is the robot radius,  $d(\mathbf{x})$  is the distance to the nearest obstacle,  $d_{\text{safe}}$  is the threshold distance considered safe, and  $v_{\text{slow}}$  is the minimum travel speed near an obstacle, biasing paths away from surfaces (Fig. 4b).

$T(\mathbf{x})$  is computed on a volumetric grid using the Fast Marching Method (FMM) [10] via the *scikit-fmm* library<sup>3</sup>. FMM runs in  $\mathcal{O}(N \log N)$  time, where  $N$  is the number of grid cells, and each ToA map is pre-computed offline once per training environment. No ToA map is required at deployment. Gradients  $\nabla T(\mathbf{x})$  are calculated at grid cell centers via finite differencing and interpolated to provide values in continuous space. The negative gradient then defines the velocity set-point direction  $\mathbf{v}_k^{\text{set}}$  in the target velocity loss (see Eq. (9)), providing a consistent training signal that encourages the robot to make progress toward the goal while avoiding obstacles.

<sup>3</sup><https://github.com/scikit-fmm/scikit-fmm>

TABLE I: Loss function parameters

$\lambda_{\text{acc}}$	$\lambda_{\text{jerk}}$	$\lambda_{\omega}$	$\lambda_v$	$\lambda_{v_{\text{max}}}$	$\lambda_{\text{clearance}}$	$\lambda_{\text{collision}}$	$\lambda_{\text{yaw}}$	$\beta_1$	$\beta_2$
0.01	0.001	0.3	4.0	1.0	6.0	6.0	1.0	2.5	-6.0

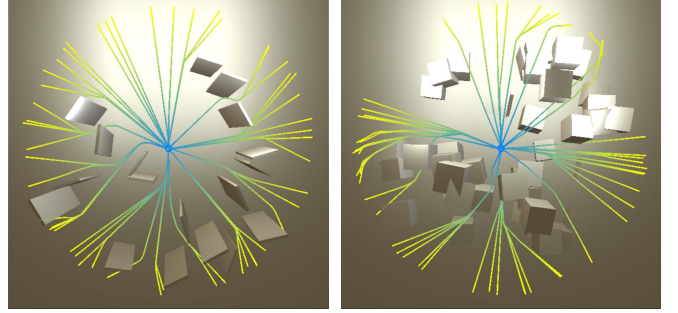


Fig. 5: Top down view of two cylinder shaped training environments with random primitive obstacles and starting points at a fixed radius from the goal (blue). Trajectories illustrate paths following the time-of-arrival map (yellow to blue).

6) *Total Loss*: The total loss is a linear combination of all the loss terms using the coefficients from Table I:

$$L = \lambda_{\text{acc}}L_{\text{acc}} + \lambda_{\text{jerk}}L_{\text{jerk}} + \lambda_{\omega}L_{\omega} + \lambda_vL_v + \lambda_{v_{\text{max}}}L_{v_{\text{max}}} + \lambda_{\text{clearance}}L_{\text{clearance}} + \lambda_{\text{collision}}L_{\text{collision}} + \lambda_{\text{yaw}}L_{\text{yaw}} \quad (14)$$

#### D. Training

The policy is trained entirely in simulation with a customized GPU-accelerated simulator based on VisFly [11], which uses Habitat-Sim [12] to perform rendering and collision checking in training environments that contain randomly generated obstacles (i.e., spheres, cylinders, and cuboids; see Fig. 5). A cylindrical shaped training environment enables multiple agents to share a single environment while observing a variety of ToA gradient directions simultaneously. Environments and ToA maps are pre-generated once to reduce training overhead. We use back propagation through time (BPTT) and temporal gradient decay [2] to accumulate and propagate gradients with a batch size of 16 over 10K iterations. We find that training with an empty environment and no collision loss for the first 500 iterations ensures stable gradients. The policy converges within minutes during this initial stage, while training in the full environment with the complete loss requires several hours to achieve proficiency.

#### E. Body Rate Attitude Control

To bridge the sim-to-real gap between point-mass dynamics in simulation and rigid-body dynamics in reality, we adopt a PD attitude controller [13] that tracks desired thrust  $\mathbf{F}$ , orientation  $\mathbf{R}_d$ , and body rates  $\boldsymbol{\omega}_d$ . The desired rotation  $\mathbf{R}_d$  aligns the body  $\mathbf{z}^B$  axis with  $\mathbf{F}$  and  $\mathbf{x}^B$  with the predicted yaw direction  $\psi_d$ . The desired body rates  $\boldsymbol{\omega}_d$  are computed from ZYX Euler angle rates estimated from consecutive desired attitudes via the Euler angle Jacobian:

$$\boldsymbol{\omega}_d = \begin{bmatrix} 1 & 0 & -\sin(\theta) \\ 0 & \cos(\phi) & \sin(\phi) \cos(\theta) \\ 0 & -\sin(\phi) & \cos(\phi) \cos(\theta) \end{bmatrix} \begin{bmatrix} \dot{\phi} \\ \dot{\theta} \\ \dot{\psi} \end{bmatrix} \quad (15)$$

TABLE II: Domain randomization and parameters

Value	Distribution	Parameters
Gravity	$g \sim \mathcal{N}(\mu, \sigma^2)$	$\mu = 9.81, \sigma = 1.5$
Initial height	$z \sim \mathcal{U}(z_{\min}, z_{\max})$	$z_{\min} = 0.5$ $z_{\max} = 2.5$
Initial $x$	$x = r_{\text{cyl}} \cos \theta$	$r_{\text{cyl}} = 12.5$
Initial $y$	$y = r_{\text{cyl}} \sin \theta$ $\theta \sim \mathcal{U}(0, 2\pi)$	
Target speed	$v_{\text{target}} \sim \mathcal{U}(v_{\min}, v_{\max})$	$v_{\min} = 1$ $v_{\max} = 5$
Velocity noise	$v_{\text{noise}} \sim \mathcal{N}(\mu, \sigma^2) \in \mathcal{R}^3$	$\mu = 0, \sigma = 0.05$
Rotation noise	$r_{\text{noise}} \sim \mathcal{N}(\mu, \sigma^2) \in \mathcal{R}^3$	$\mu = 0, \sigma = 0.02$

$\mathcal{U}$  and  $\mathcal{N}$  denote uniform and normal distributions respectively.  $r_{\text{cyl}}$  denotes the radius of a cylindrical distribution.

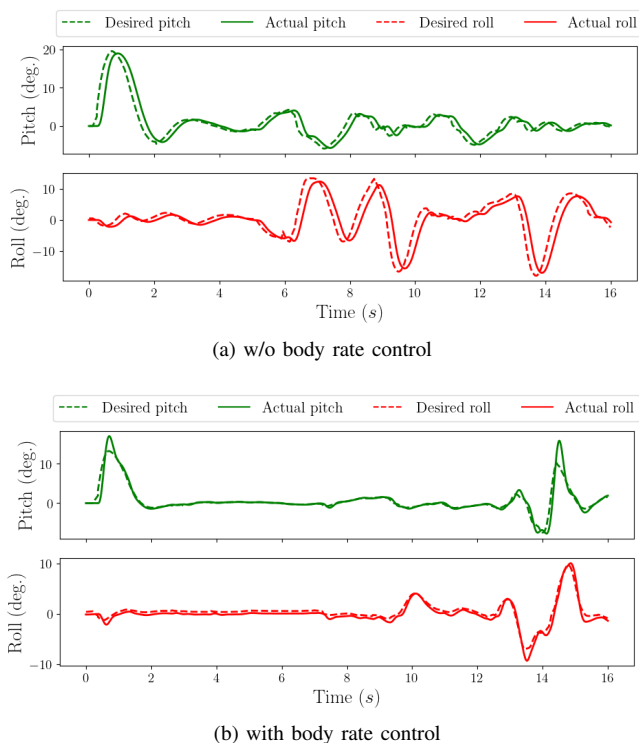


Fig. 6: Comparison of attitude control performance without (a) and with (b) the derivative feedback term,  $\omega_d$ , in the attitude controller.

Including the derivative term  $\omega_d$  significantly improves control response, enabling the attitude controller to achieve the desired orientation with negligible latency in both simulation and real-world flights. In contrast, controllers that rely solely on proportional feedback, such as in [2], exhibit noticeable control lag, which must be accounted for during policy training. We show examples contrasting the control performance in Sect. IV-A.

#### F. Domain Randomization

To ensure platform-agnostic deployment, the policy outputs mass-normalized thrust (in units of  $\text{m/s}^2$ ), which is later converted to motor commands using a parametric quadrotor model. Inaccuracies in parameters such as mass, thrust coefficients, and battery voltage can cause steady-state errors

between expected and actual thrust.

To improve robustness to such discrepancies, we apply domain randomization during training (Table II). We randomize gravity  $g$  across rollouts, forcing the policy to adapt its thrust based on velocity feedback, with smoothness loss discouraging unintended accelerations (see experiments in Sect. V-A). We also randomize initial position, target velocity, and inject noise into state inputs to simulate sensor imperfections and improve generalization.

## IV. SIMULATION EXPERIMENTS

### A. Attitude Control Performance

We perform an ablation study comparing the simulated performance between our attitude control law with body rate control and a controller only tracking desired orientation as in [2]. Figure 6b shows that including a derivative term (i.e., the desired body rate,  $\omega_d$ ) yields negligible control response latency, whereas Fig. 6a exhibits a latency of approximately 200 ms. Such latency can delay evasive maneuvers, increasing the risk of collisions in cluttered or dynamic environments.

### B. Diverse Simulated Environments

We evaluate our planner in both simulated indoor and outdoor environments featuring narrow corridors and large obstacles. The evaluation uses the 11 out-of-distribution scenarios from [14], with depth images rendered by Flightmare [15]. Notably, the policy trained with point-mass dynamics is evaluated under full rigid-body quadrotor dynamics without access to ToA maps. 1,350 simulation trials are conducted across all environments and different start-end location pairs. Further details on the experimental design are available in [14].

We compare our method against two baselines: Back to Newton’s Laws (BNL) [2], trained using the authors’ public codebase<sup>4</sup>, and an ablated version of our model trained without the ToA privileged information (yaw w/o ToA). We do not ablate ToA without yaw prediction, as the ToA gradient may direct the robot toward obstacles outside the depth sensor’s field of view, resulting in unsafe behavior. All approaches are trained using only primitive obstacles and point-mass dynamics. Performance is measured by success rate and failure modes (collisions and timeouts), with success defined as reaching the target without collision within a fixed time limit.

Our method achieves a 86% success rate, 36% higher than the baselines, and has the lowest collision rate across all approaches (Fig. 7). As shown in Fig. 8, The BNL policy struggles to avoid large, flat obstacles due to its fixed heading towards the target. The yaw w/o ToA policy achieves a higher success rate than BNL in environments requiring significant reorientation (e.g., Industry 2), confirming that yaw alignment alone enables the policy to turn around large obstacles; however, without ToA guidance it frequently times out in concave regions. Our full model yaws around large obstacles and learns global navigation cues from the ToA loss, yielding the highest success rate across all environments.

<sup>4</sup><https://github.com/HenryHuYu/DiffPhysDrone>

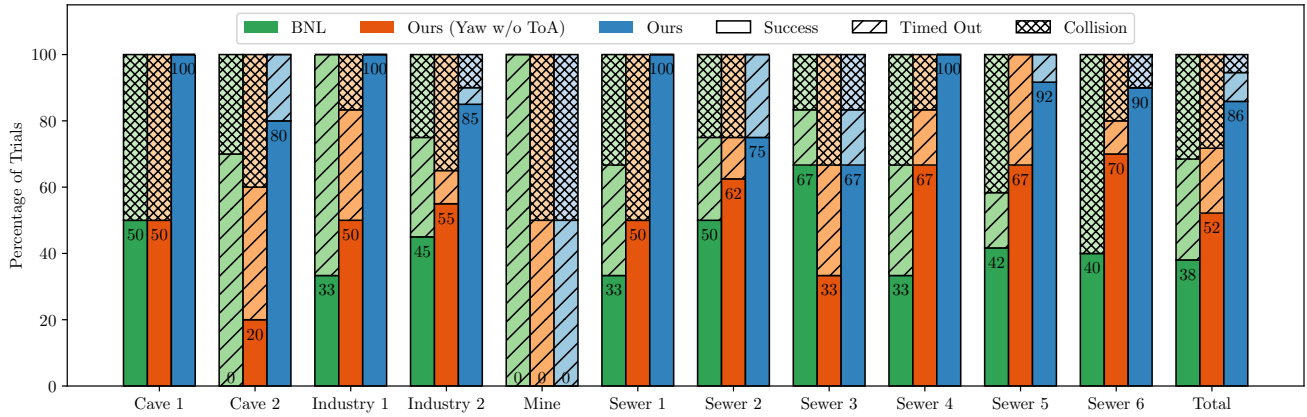


Fig. 7: Planner success rate and failure modes across 11 diverse environments. The proposed method (*Ours*) achieves the highest success rate and lowest collision rate compared to the baseline [2] and the ablated policy trained without privileged information (yaw w/o ToA). The *Mine* environment features a maze-like corridor which results in poor performance from all planners.

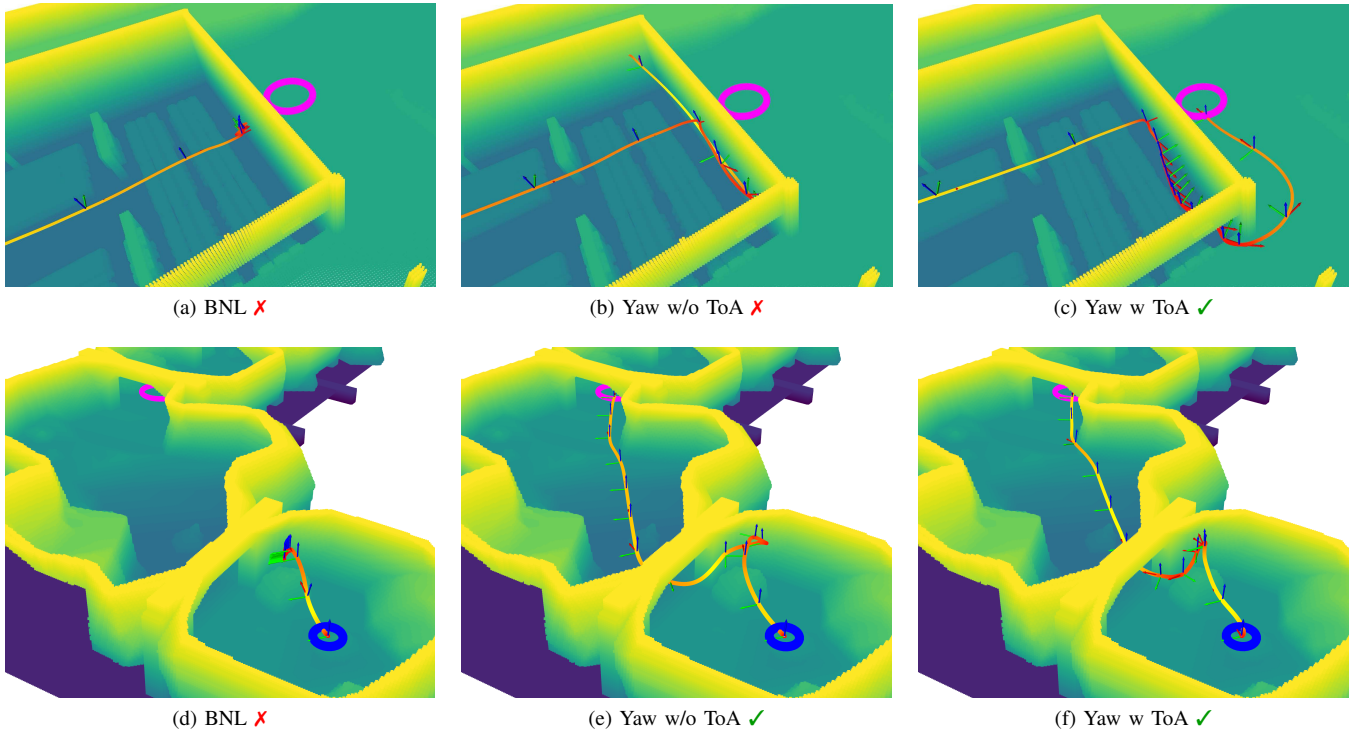


Fig. 8: Trajectories overlaid on ground-truth point clouds, with a cross-section shown for clarity. Trajectories are colored by speed (red to yellow, up to  $v_{\max} = 3$  m/s) with body frames every 2 s. Start and goal are marked in blue and magenta. Successful trials are marked with a  $\checkmark$  in *Industry 2* (top) and *Cave 2* (bottom) environments. ToA maps serve as an inductive bias during training only and are not available during simulation or hardware evaluation.

## V. HARDWARE EXPERIMENTS

A custom quadrotor platform was developed to support hardware experimentation. The vehicle spans 15 cm from rotor to rotor and is equipped with a forward-facing Intel RealSense D456 depth camera, a downward-facing Lightware SF20/C range finder, and a Matrix Vision BlueFox2 global shutter greyscale camera. All onboard computation is performed by an NVIDIA Orin NX module with 16 GB of RAM. The system has a total mass of 1.7 kg. A TBS Lucid H7 flight

controller runs custom Betaflight firmware<sup>5</sup> to provide IMU data at 1000 Hz. For GPS-denied state estimation, we employ a monocular visual-inertial navigation system developed by Yao [16]. The RealSense camera generates depth images at 60 Hz with a resolution of  $640 \times 480$  pixels and the policy runs at 50 Hz.

### A. Domain Randomization Experiments

We conduct an ablation study to evaluate whether domain randomization compensates for modeling errors, comparing

<sup>5</sup><https://github.com/betaflight/betaflight>

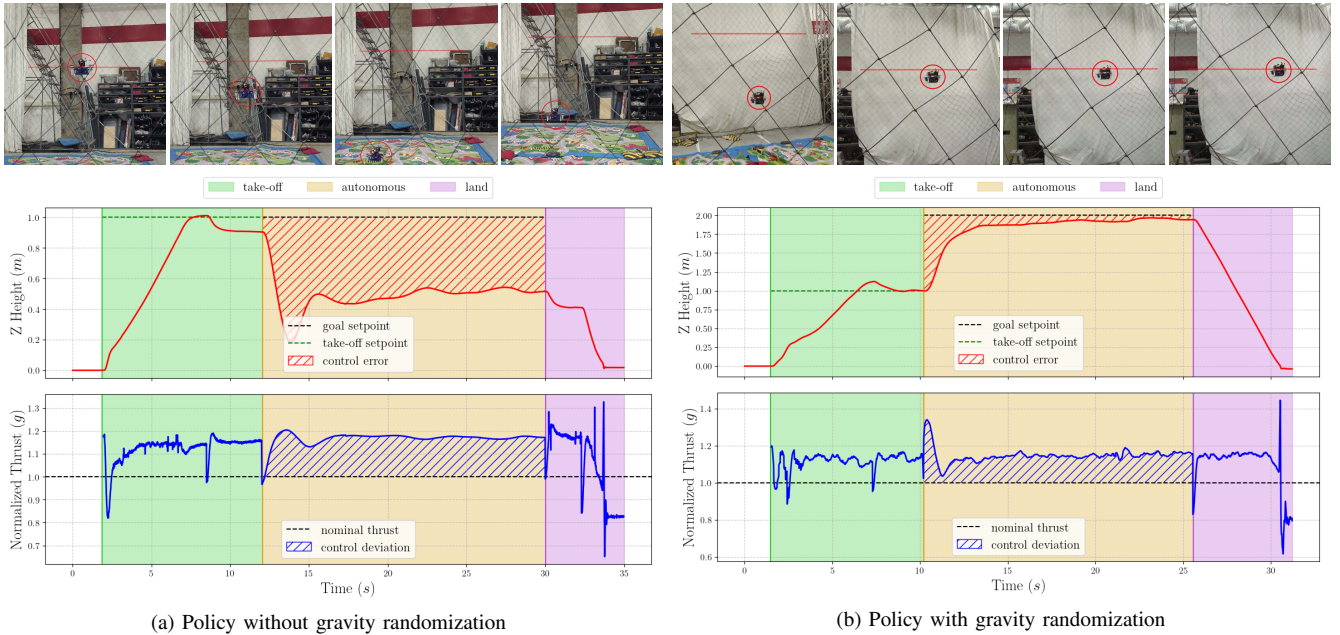


Fig. 9: Hardware ablation comparing policies trained without (a) and with (b) gravity randomization. Top row: flight snapshots with goal setpoint in red. In (a), the start and goal positions coincide, but upon entering autonomous mode, the vehicle quickly loses altitude. With randomization (b), the robot gains altitude and reaches the goal setpoint. Middle row: plot of z-height from motion capture showing significantly reduced altitude error with the gravity randomized policy. Bottom row: plot of normalized thrust predicted by the policy, where the gravity randomized policy initially outputs  $1.3g$ , substantially higher than the policy without gravity randomization, compensating for modeling inaccuracies.

policies trained with and without randomized gravity. A major source of modeling error stems from the thrust-to-RPM mapping, which is empirically derived under benchtop conditions that don't account for in-flight factors like battery voltage drop and frame-induced airflow disturbances. As a result, the quadrotor exhibits steady-state thrust mismatches; for example, it requires sending thrust commands of approximately  $1.15g$  to hover instead of the expected  $1g$  (Fig. 9b).

To address this, we train policies with gravity sampled from a normal distribution (see Sect. III-F), encouraging the policy to learn a closed-loop feedback mechanism. We evaluate both policies in a hover task where the policy is provided a target velocity vector proportional to the distance to a fixed goal setpoint. The non-randomized policy initially outputs exactly  $1g$  of thrust and fails to maintain altitude (Fig. 9a), while the randomized policy increases thrust up to  $1.3g$  achieving stable hover at the goal setpoint (Fig. 9b). These results support our hypothesis that gravity randomization induces adaptive behavior that compensates for thrust modeling inaccuracies during deployment.

### B. Outdoor Navigation Experiments

We deploy the policies in hardware trials in an outdoor flight arena (see Fig. 11) and under a stand of trees with dense foliage and underbrush (see Fig. 10). Table III provides a table of all flights conducted outdoors. The top speed achieved by the system is  $4\text{ m/s}$ . The total length over all trials is  $589\text{ m}$ . No crashes occurred during testing.

TABLE III: Hardware Flight Trials

	Env. #	Flight Time $s$	Path Length $m$	$v_{max}$ $m/s$	Max Speed $m/s$	Avg Speed $m/s$
Outdoor Flight Arena	1	14.4	12.1	2.0	1.5	0.7
	2	13.2	17.9	2.0	2.0	1.2
	3	11.0	16.6	3.0	3.0	1.4
	4	17.3	26.9	3.0	2.8	1.4
	5	13.0	24.1	4.0	3.5	1.7
	6	12.0	24.5	5.0	3.7	1.9
	7	9.3	17.2	3.0	2.9	1.7
	8	11.7	22.3	3.0	2.9	1.8
	9	11.8	24.9	3.0	4.0	2.0
	10	11.1	22.5	3.0	2.8	1.8
	11	11.7	19.6	3.0	2.7	1.5
	12	11.3	20.4	3.0	2.7	1.6
	13	10.0	19.8	3.0	2.9	1.9
	14	9.5	17.9	3.0	2.4	1.9
	15	14.2	21.0	3.0	2.5	1.3
Forest	1	12.7	22.5	3.0	2.8	1.6
	2	31.1	52.9	3.0	2.8	2.2
	3	26.0	74.6	4.0	3.1	2.5
	4	23.7	70.1	5.0	3.8	2.7
	5	20.0	61.3	3.0	3.0	2.4

## VI. CONCLUSION

We developed a simulation and training framework for efficiently learning vision-based navigation policies that map depth and state observations to thrust and heading commands. Trained using simplified point-mass dynamics and privileged information, the policy generalizes to full quadrotor dynamics and reliably navigates through cluttered environments. Domain randomization, especially over gravity, enables robust feedback control, correcting for modeling

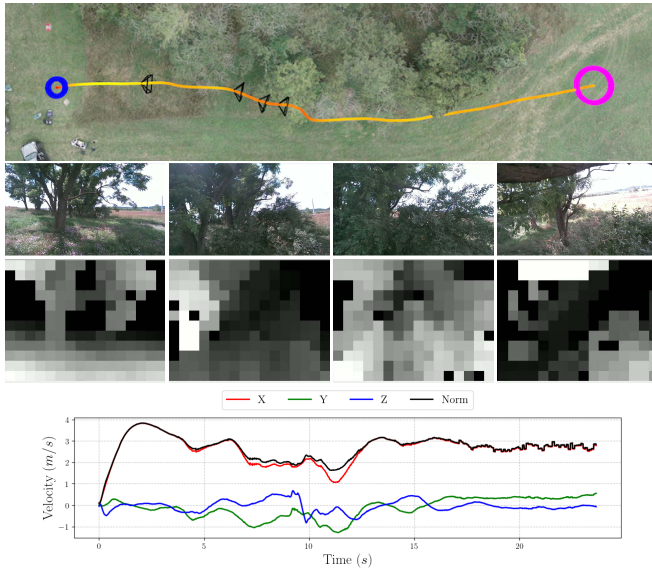


Fig. 10: Outdoor obstacle avoidance test under tree canopy (Table III Forest Flight 4). The policy predicts up to  $30^\circ$  in yaw to navigate through dense underbrush with speeds up to 3.8 m/s. From top to bottom: VINS trajectory overlaid on terrain map, onboard RGB images, policy depth input (inverted and max pooled), and velocity profile.

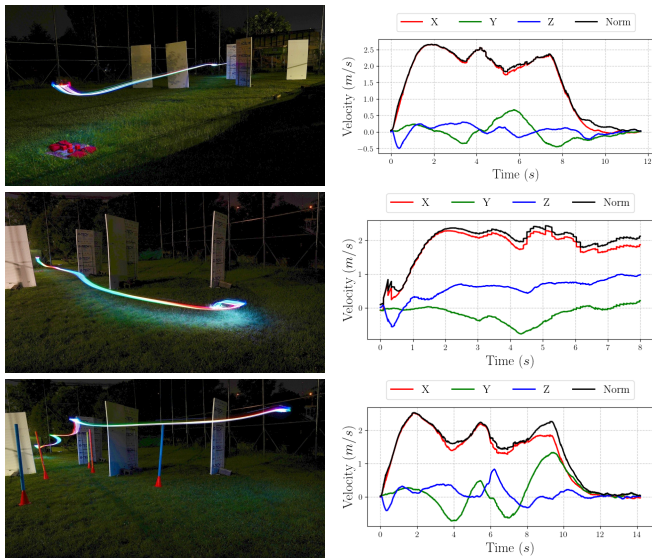


Fig. 11: Outdoor obstacle avoidance tests at night using LED illumination and long-exposure photography.

errors such as a 15% thrust mismatch. Our approach achieves an 86% success rate in simulation and completes 20 hardware flights covering 589 m without collisions. While the policy generalizes from primitive training obstacles to photorealistic simulation and real-world flights without fine-tuning, it has difficulty with backtracking in maze-like environments (e.g., the *Mine* scenario) and exhibits initial yaw oscillations. A promising direction for future work is to explore more expressive learned-memory architectures to enhance spatial reasoning and long-horizon planning without relying on explicit maps. Additionally, incorporating task-specific inputs

and new objective functions could extend the method's use to a wider range of applications.

## ACKNOWLEDGMENTS

The authors would like to thank Ankit Khandelwal for contributions to the codebase and Edsel Burkholder for field testing support. This material is based upon work supported in part by the Army Research Laboratory and the Army Research Office under contract/grant number W911NF-25-2-0153.

## REFERENCES

- [1] A. Loquercio, E. Kaufmann, R. Ranftl, M. Müller, V. Koltun, and D. Scaramuzza, "Learning high-speed flight in the wild," in *Sci. Robot.*, October 2021.
- [2] Y. Zhang, Y. Hu, Y. Song, D. Zou, and W. Lin, "Learning vision-based agile flight via differentiable physics," in *Nat. Mach. Intell.*, June 2025, pp. 954–966.
- [3] J. Lu, X. Zhang, H. Shen, L. Xu, and B. Tian, "You only plan once: A learning-based one-stage planner with guidance learning," in *IEEE Robot. Autom. Letters*, vol. 9, no. 7, May 2024, pp. 6083–6090.
- [4] M. Kim, G. Bae, J. Lee, W. Shin, C. Kim, M.-Y. Choi, H. Shin, and H. Oh, "Rapid: Robust and agile planner using inverse reinforcement learning for vision-based drone navigation," in *Proc. of Robot.: Sci. and Syst.*, June 2025.
- [5] R. Newbury, J. Collins, K. He, J. Pan, I. Posner, D. Howard, and A. Cosgun, "A review of differentiable simulators," *IEEE Access*, vol. 12, pp. 97 581–97 604, July 2024.
- [6] I. Meijer, M. Pantic, H. Oleynikova, and R. Siegwart, "Pushing the limits of reactive navigation: Learning to escape local minima," *IEEE Robot. Autom. Letters*, vol. 10, no. 7, pp. 6792–6799, April 2025.
- [7] D. Freeman, E. Frey, A. Raichuk, S. Girgin, I. Mordatch, and O. Bachem, "Brax: A differentiable physics engine for large scale rigid body simulation," in *Proc. NeurIPS Datasets and Benchmarks Track*, 2021.
- [8] J. Heeg, Y. Song, and D. Scaramuzza, "Learning quadrotor control from visual features using differentiable simulation," *arXiv preprint arXiv:2410.15979*, 2025.
- [9] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.
- [10] J. A. Sethian, "A fast marching level set method for monotonically advancing fronts." *Proc. of the National Acad. of Sci.*, vol. 93, no. 4, pp. 1591–1595, 1996.
- [11] F. Li, F. Sun, T. Zhang, and D. Zou, "Visfly: An efficient and versatile simulator for training vision-based flight," *arXiv preprint arXiv:2407.14783*, 2024.
- [12] M. Savva, A. Kadian, O. Maksymets, Y. Zhao, E. Wijmans, B. Jain, J. Straub, J. Liu, V. Koltun, J. Malik *et al.*, "Habitat: A platform for embodied ai research," in *Proc. of the IEEE Intl. Conf. on Computer Vision*, October 2019.
- [13] A. E. Spitzer, "Dynamical model learning and inversion for aggressive quadrotor flight," Ph.D. dissertation, Carnegie Mellon University Pittsburgh, PA, 2021.
- [14] J. Lee, A. Rathod, K. Goel, J. Stecklein, and W. Tabib, "Rapid quadrotor navigation in diverse environments using an onboard depth camera," in *2024 IEEE Int. Symp. on Saf. Secur. Rescue Robot. (SSRR)*, 2024, pp. 18–25.
- [15] Y. Song, S. Naji, E. Kaufmann, A. Loquercio, and D. Scaramuzza, "Flightmare: A flexible quadrotor simulator," in *Conf. on Robot Learn.* PMLR, 2021, pp. 1147–1157.
- [16] J. W. Yao, "Resource-constrained state estimation with multi-modal sensing," Ph.D. dissertation, Carnegie Mellon University, Pittsburgh, PA, April 2020.