Vision Is All You Need: A Vision-Only Approach to Dynamics Estimation in Autonomous Navigation

Lazar Milikic (353622), Ahmad Jarrar Khan (353435),
Said Gurbuz (369141), Marko Mekjavic (359986)
CS-503 Project Proposal

***Abstract*—Inspired by biological systems' proficiency in navigating complex environments using primarily visual inputs, we propose a novel, vision-only approach to enable an autonomous drone to navigate highly dynamic environments effectively. Unlike traditional methods that use complex sensor arrays to capture environmental dynamics, our project aims to develop a simpler, cost-effective visual-based model. Specifically, the project investigates whether a drone can solely rely on visual information to estimate the dynamics of objects and make optimal decisions in dynamic settings. We employ a two-phase training process, initially leveraging both visual observations and privileged information about environment dynamics to teach the drone to navigate through gates while evading projectiles. The subsequent phase involves refining a Dynamics Module to predict environmental dynamics from visual inputs alone. This approach could revolutionize the design and deployment of autonomous systems, reducing reliance on expensive sensors and increasing accessibility across various applications.**

I. Introduction

Biological systems excel at navigating complex environments using primarily visual inputs [1]. Inspired by this, our project challenges the conventional multi-sensor approach in autonomous systems, proposing that visual inputs alone can suffice for effective navigation in dynamic environments. Unlike traditional methods that depend on complex sensors to grasp environmental dynamics [2], [3], our project seeks to develop a vision-only-based model for an autonomous drone, consequently reducing system complexity and cost.

A. What is the problem you want to solve?

The core problem addressed by this project is whether an autonomous agent can rely solely on visual information to estimate the dynamics of objects within its environment and make optimal decisions in dynamic settings. To challenge the agent's ability to capture the dynamics, we have designed a specific task and environment: an autonomous drone is assigned to navigate through gates while evading incoming projectiles, depending entirely on visual input to estimate their dynamics. A critical element of our investigation will be exploring the impact of privileged information about environmental dynamics on policy learning.

B. Why is it important that this problem be solved?

Recent advancements in vision-based drones show significant progress. Still, a critical gap persists in effectively capturing environmental dynamics, especially in complex scenarios where drones often operate (e.g., warzone) and where rapid adaptation is crucial for success [4]. Addressing this challenge has profound experimental and practical implications. Enabling drones to navigate highly dynamic environments solely using visual information could revolutionize their design and deployment. This breakthrough would reduce reliance on costly sensors, thereby lowering costs and complexity, and increasing the systems' flexibility and accessibility for various applications.

II. Method and Deliveries

A. How do you solve the problem?

To address the challenge of navigating through gates while evading projectiles, we propose a novel approach that adapts the Rapid Model Adaptation (RMA) training process [5] to capture environmental dynamics from visual signals alone.

Phase 1: Policy Learning

Policy Model: The drone learns a policy π using visual observations combined with privileged information about environmental dynamics. The observation signal is an RGB image, encoded by DINOv2 [6]. This model effectively captures various visual features, reducing the necessity for separate models to handle mid-level visual representations as in previous studies [7], [8]. Privileged information, including coordinates, distances, and velocities of each projectile as well as the target gate, is encoded using an MLP network. Assuming a maximum of P_{max} projectiles, inputs for the transformer policy model to predict the subsequent action include all embedded image patches, vectors for projectiles and the target, and a vector for the previously taken action as per [9]. We employ the transformer architecture as a policy model for its ability to effectively attend to each projectile's dynamics information with corresponding image patches, thereby enhancing the relevance of action predictions.

RL Elements: Due to time and resource constraints, we utilize a simplified high-level action space where the drone decides its velocities (v_x^d, v_y^d, v_z^d) at each timestep. The reward function is structured as follows: a high reward is given for successfully passing through a gate, a significant penalty is applied for being hit by a projectile, and incremental rewards are granted for decreasing the distance to the next gate, with penalties for increasing it. We employ the Proximal Policy Optimization (PPO) algorithm [10], which is well-regarded for its effectiveness in similar contexts, to optimize the policy [5], [7], [8], [9].

Environment: We utilize the Gym Pybullet Drones environment [11] for its simplicity and realistic physics simulations. While alternatives like Flightmare [12] and AirSim [13] provide more detailed realism, their complexity and the time required for mastery make them less suitable for our project's timeline. PyBullet is thus chosen for its straightforward integration with the Gymnasium library, ease of customization in Python, and sufficient realism.

Phase 2: Dynamics Module Learning

In the second phase of our project, we train a Dynamics Module to estimate environmental dynamics from the last N observation-action pairs. This module can employ transformer architectures such as TubeR [14] or TransVOD [15], which extend the foundational DETR architecture [16] to handle 3D data effectively. Contrary to directly predicting raw privileged information, our Dynamics Module is engineered to infer MLP-encoded dynamics information. This approach aligns with findings from the RMA paper, which suggests that direct prediction of privileged information yields suboptimal results in practical applications. During inference, the Dynamics Module's predictions substitute for the direct use of privileged information within the policy model, enhancing its decision-making capabilities in dynamic environments.

B. How will you validate your solution?

Validation focuses on evaluating the drone's success rate in reaching gates and its navigation efficiency by measuring the average time needed to travel between gates under various environmental conditions including wind, scenery changes, and lighting variations. To assess the efficacy of our proposed architecture and the benefits of integrating visual signals with predicted environment dynamics, we plan to train three additional baseline architectures:

- **Baseline 1:** Policy uses only the current RGB signal.
- **Baseline 2:** Policy uses the last N observations.
- **Baseline 3:** Policy employs the privileged information during training and the predicted dynamics during inference. Captured RGB signal is not directly provided to the policy—only used in the Dynamics Module.

Additionally, a **human performance** baseline will be established by having a professional gamer control the drone.

C. Milestones

Milestone 1: The first milestone involves setting up and tuning the environment and the initial policy algorithms by implementing and training the baseline models.

Milestone 2: Once the environment is tuned, we experiment with privileged information to train policies that predict actions based on both visual and privileged data. Verifying the utility of privileged information will lead us to integrate and test the Dynamics Module, completing the training pipeline without direct privileged information. Finally, we also wish to explore the predictive capabilities of the trained Dynamics Module in anticipating future observations.

III. Related Work

Our methodology is deeply influenced by the Rapid Model Adaptation (RMA) framework [5], which pioneers an adaptation module with a two-phase training procedure. Although RMA can integrate visual inputs [17], its application does not maximize the potential of vision in dynamic settings due to task constraints. We have therefore restructured their approach to focus specifically on predicting environmental dynamics and managing a vision-only system, which we believe can better harness the capabilities of vision, especially for our specific task. We retain the foundational two-phase training concept but modify the architecture to capture the dynamics, aligning the system's inductive bias with our project's needs. Furthermore, our work is informed by developments in autonomous drone racing [9], where initial access to privileged information significantly benefits learning policies. Contrary to existing methods that differentiate between teacher (privileged information-based) and student (vision-based) models, our unified policy framework directly incorporates dynamic environmental estimates, allowing for more effective utilization of privileged information.

IV. Discussion

A. Implications and Broader Impact

A successfully implemented Dynamics Module could profoundly impact several fields, including robotics, autonomous systems, and computer vision. If this module can robustly capture environmental dynamics using only visual signals, it could represent a step forward for autonomous technology by improving short-term planning and decision-making in autonomous systems. Additionally, since this method relies on visual inputs alone, this could simplify its deployment, enhancing its scalability and cost-effectiveness. However, the potential applications of such technology, particularly in military systems, pose strong ethical concerns. It is crucial to address these issues responsibly to ensure that the advancements contribute positively to society.

B. Potential Risks and Shortcomings

The main shortcomings arise from using a simplified environment, which may hinder the agent's ability to generalize to complex real-world scenarios with occluded, noisy, or incomplete visual inputs. Additionally, the limited ability to vary environment settings and sceneries could restrict the drone's robustness and generalization in unseen scenarios. Moreover, the computational demands of processing visual inputs could challenge efficiency and real-time performance. Finally, our initial hypothesis—that visual inputs alone can accurately capture environmental dynamics—might be incorrect, potentially leading to sub-optimal outcomes or failure of the approach.

V. Individual Contributions

Lazar Milikic originated the initial project idea of applying RMA principles for the dynamic estimation of objects in environments. Together, Marko Mekjavic and Lazar Milikic directed the project towards the selected task, aligning with the course requirements where dynamic estimation becomes crucial for optimal performance. The task was further refined and developed with feedback and innovative ideas from Said Gurbuz, and ultimately shaped into its final form by Ahmad Jarrar, who finalized the methodology for the task. Marko Mekjavic took on the role of engaging other classmates to join the team. Both Lazar Milikic and Said Gurbuz took charge of writing the project proposal. Ahmad Jarrar was tasked with researching the environments, setting up the simulator, and creating a custom environment featuring projectiles and gates. Ahmad Jarrar and Marko Mekjavic also played key roles in condensing and finalizing the version for submission. Marko Mekjavic created diagrams depicting the architecture and approach, which will be utilized in later milestones (currently not included due to page limit constraints).

References

[1] O. TRULLIER, S. I. WIENER, A. BERTHOZ, and J.-A. MEYER, "Biologically based artificial navigation systems: Review and prospects," *Progress in Neurobiology*, vol. 51, no. 5, pp. 483–544, 1997. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0301008296000603

[2] H. X. Pham, H. M. La, D. Feil-Seifer, and L. V. Nguyen, "Autonomous uav navigation using reinforcement learning," 2018.

[3] W. Koch, R. Mancuso, R. West, and A. Bestavros, "Reinforcement learning for uav attitude control," *ACM Trans. Cyber-Phys. Syst.*, vol. 3, no. 2, feb 2019. [Online]. Available: https://doi.org/10.1145/3301273

[4] J. Xiao, R. Zhang, Y. Zhang, and M. Feroskhan, "Vision-based learning for drones: A survey," 2024.

[5] A. Kumar, Z. Fu, D. Pathak, and J. Malik, "Rma: Rapid motor adaptation for legged robots," 2021.

[6] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[7] A. Sax, B. Emi, A. R. Zamir, L. Guibas, S. Savarese, and J. Malik, "Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies," 2019.

[8] B. Chen, A. Sax, G. Lewis, I. Armeni, S. Savarese, A. Zamir, J. Malik, and L. Pinto, "Robust policies via mid-level visual representations: An experimental study in manipulation and navigation," 2020.

[9] J. Fu, Y. Song, Y. Wu, F. Yu, and D. Scaramuzza, "Learning deep sensorimotor policies for vision-based autonomous drone racing," 2022.

[10] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," 2017.

[11] J. Panerati, H. Zheng, S. Zhou, J. Xu, A. Prorok, and A. P. Schoellig, "Learning to fly – a gym environment with pybullet physics for reinforcement learning of multi-agent quadcopter control," 2021.

[12] Y. Song, S. Naji, E. Kaufmann, A. Loquercio, and D. Scaramuzza, "Flightmare: A flexible quadrotor simulator," in *Proceedings of the 2020 Conference on Robot Learning*, 2021, pp. 1147–1157.

[13] S. Shah, D. Dey, C. Lovett, and A. Kapoor, "Airsim: High-fidelity visual and physical simulation for autonomous vehicles," in *Field and Service Robotics*, 2017. [Online]. Available: https://arxiv.org/abs/1705.05065

[14] J. Zhao, Y. Zhang, X. Li, H. Chen, B. Shuai, M. Xu, C. Liu, K. Kundu, Y. Xiong, D. Modolo *et al.*, "Tuber: Tubelet transformer for video action detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 13 598–13 607.

[15] Q. Zhou, X. Li, L. He, Y. Yang, G. Cheng, Y. Tong, L. Ma, and D. Tao, "Transvod: End-to-end video object detection with spatial-temporal transformers," *arXiv preprint arXiv:2201.05047*, 2022.

[16] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[17] Z. Fu, A. Kumar, A. Agarwal, H. Qi, J. Malik, and D. Pathak, "Coupling vision and proprioception for navigation of legged robots," *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17 252–17 262, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:244896056