

Local Path Planning among Pushable Objects based on Reinforcement Learning

Linghong Yao, Valerio Modugno, Andromachi Maria Delfaki,
 Yuanchang Liu, Danail Stoyanov, and Dimitrios Kanoulas

Abstract—In this paper, we introduce a method to tackle the problem of robot local path planning among pushable objects – an open problem in robotics. In particular, we simultaneously train multiple agents in a physics-based simulation environment, utilizing an Advantage Actor-Critic algorithm coupled with a deep neural network. The developed online policy enables these agents to push obstacles in ways that are not limited to axial alignments, adapt to unforeseen changes in obstacle dynamics instantaneously, and effectively tackle local path planning in confined areas. We tested the method in various simulated environments to prove the adaptation effectiveness to various unseen scenarios in unfamiliar settings. Moreover, we have successfully applied this policy on an actual quadruped robot, confirming its capability to handle the unpredictability and noise associated with real-world sensors and the inherent uncertainties in unexplored object-pushing tasks.

I. INTRODUCTION

Mobile robots have gained great capabilities in the past decade, such that they are now able to autonomously navigate efficiently and safely even in clutter environments, by avoiding static or dynamically moving obstacles [1]. Although, the navigation problem where objects can be moved around to free space –also known as Navigation Among Movable Obstacle (NAMO)– is still an open problem. This concept mirrors the human instinct to shift, for instance, furniture or other objects blocking their way in densely furnished areas, suggesting that robots could similarly optimize their routes by strategically moving obstacles to clear a path towards their destination. The applications in robotics are highly relevant in several scenarios, including tasks with regular maintenance in factories where unused containers and boxes might block access points, domestic service robots navigating through furniture-cluttered homes, or robots conducting inspections in subterranean environments obstructed by rocks and debris. The capacity for effective obstacle manipulation can greatly enhance autonomous navigation efficiency in these settings.

The problem when movable objects need to be moved around, even in simplified versions, is proved to be NP-hard [2]. To solve global path planning among movable obstacles, previous studies have explored iterative and recursive algorithms [3], but often relying on certain simplifications,

¹The authors are with the Department of Computer Science and Mechanical Engineering, University College London, Gower Street, WC1E 6BT, London, UK. d.kanoulas@ucl.ac.uk

This work was supported by the UKRI FLF [MR/V025333/1] (RoboHike) and EPSRC [EP/P012841/1]. For the purpose of Open Access, the author has applied a CC BY public copyright license to any Author Accepted Manuscript version arising from this submission.

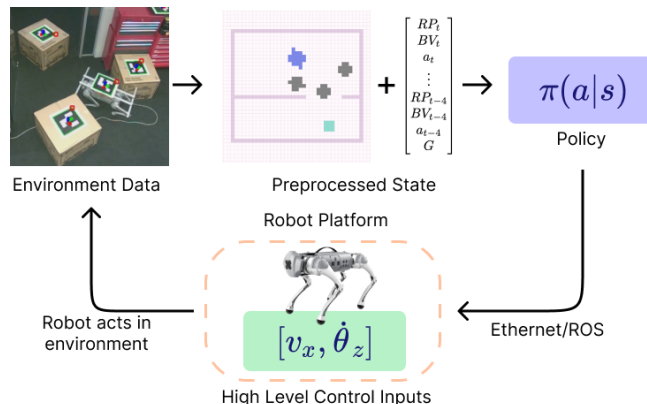


Fig. 1. Visual sensors capture the surroundings, and this data is then processed into a specific state representation. This processed information is inputted into a previously trained policy network. The network then generates a strategic action directive for the robot, enabling it to navigate and address the task of local path planning.

such as having prior knowledge about the environment, planning tasks offline which is exponential to the number of local obstacles [4], and limiting movements to axial-aligned object pushes. On the other hand, local path planning in a movable object setting has been minimally studied in the past, with only a few considering sensor inaccuracies when dealing with unexpected object dynamics [5], [6], [7], usually based on traditional optimization methods, that require intensive fine tuning and handcrafted design choices.

In this paper, we propose a novel approach to overcome the aforementioned limitations, by employing deep Reinforcement Learning (DRL) as depicted in Fig. 1. In particular, we utilize a neural network for policy-based RL, to allow an online agent complete local path plans without the constraint of previous work for axis-aligned pushing, as well as allowing uncertainties in sensor inputs and obstacle dynamics. Our solution targets the keyhole problem [8], where a mobile robot aims to traverse from one disjointed area to another by pushing obstacles through narrow passages. We acknowledge that other forms of object manipulation exist, such as pulling or lifting, but additional robotic mechanisms such as arms are needed, while just pushing remains NP-Complete. The goal of the trained policy is local path planning that can be integrated into other global navigation methods, such as A* [9]. Our approach is based on Advantage Actor-Critic [10], leveraging the advanced capabilities of the NVIDIA Isaac Gym physics engine [11] for simulating and training parallel agents. We showcase outcomes for policies that are adept at navigating through both familiar environments with new

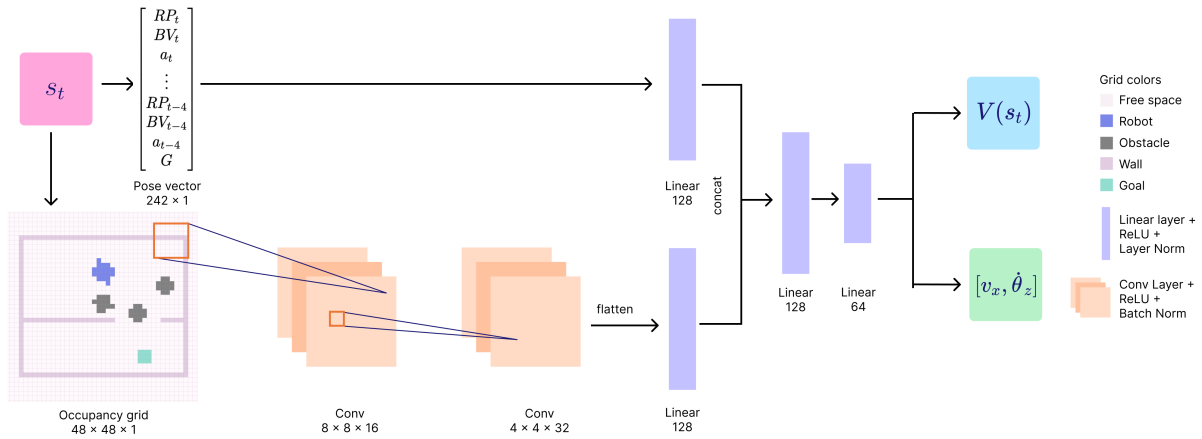


Fig. 2. Our approach utilizes a deep neural network for policy-making. The state representation, s_t , includes both a vector and a grid. The vector, of length 242, contains data on the agent’s current position, the corners of the object, the previous action taken, and the destination. The grid is a 48×48 matrix, with each cell semantically annotated. Initially, the vector is processed through a linear unit, while the grid undergoes processing through two convolutional units followed by a linear unit. The results from both processes are then merged and further processed through two additional linear units. Subsequently, the network splits, employing two distinct sets of weights to generate both the estimated value and the proposed action.

movable obstacle placements, and entirely unknown environments with unseen movable obstacles. Moreover, we validate our findings through practical experiments using a Unitree Go1 quadruped robot, illustrating the policy’s effectiveness in dealing with sensor inaccuracies and varying dynamics of real-world obstacles.

In Sec. II, we discuss the literature on the NAMO problem, while in Sec. III we state the problem formulation in the reinforcement setting, and how we implement training in simulation. In Sec. IV we present our results both in simulation and with real robot experiments. Lastly, Sec. V concludes the paper and points to future directions for solving NAMO with RL.

II. RELATED WORK

Path planning, both globally or locally, for obstacle avoidance has been heavily studied in the past [12], and while several interesting techniques use reinforcement learning [13] we will not extend further in this related work. In contrary, we will briefly review methods that deal with movable obstacles, especially in the local path planning case.

Global navigation among movable objects/obstacles is a topic that has been studied from the 1980’s [14] – an NP-Complete problem even in the simple case of moving square blocks in the plane [2]. A series of papers by Stilman et al., such as [8], [3], considered the problem as a graph planning one in which disjointed free spaces (nodes in the graph) can be connected when obstacles can be moved around. In other works, similar setups were solved using RRTs and adaptive heuristics [15] or axis-aligned obstacle movements [16], [17]. The problem has been studied in more generic ways, e.g., axis-aligned object manipulations via non-linear optimization [4]. In such solutions, all computation is offline (exponential to the number of objects) with prior knowledge of the environment, including the movability of the objects themselves. In contrary, there are also methods that could online re-plan pushing actions in unknown environments [18] or by pick-and-place on humanoids [5] using traditional

path planning techniques. Hierarchical RL was used first by Levihn et al. [19] to deal with uncertain sensory information, while later in non-axial manipulation of obstacles developed in [6], a physics-based RL framework in unexpected obstacle behaviors such as rotation was handled. Compared to those, our method runs in constant time complexity and therefore completes similar tasks five times faster, while we are able to solve harder non-linear problems too. More recently, tactile sensing is used for negotiation of unknown objects [7], while curriculum learning is used in [20] to solve the global navigation among movable obstacles problem. Further extensions to NAMO, such as socially aware obstacle placement, have also been examined in [21], [22], [23], using classic search-based approaches.

In this paper, we utilize deep RL to deal with the local path planning problem in narrow spaces. In a similar setup, Xia et al. [24] used deep RL to deal with collisions with pushable objects, rather than actual planning interactions with those. Given all the aforementioned methods, the real novelty of our approach is:

- We propose a deep RL policy that can solve local path planning among pushable objects, with non-axial-aligned pushing and constant computational complexity.
- We demonstrate that the proposed policy is able to work for unseen object positions in known environments, and generalizes to unseen object positions in unknown environments.
- We show that reliable sim-to-real transfer is possible to handle sensor noises and uncertain object dynamics.

III. METHODS

In a standard episodic Reinforcement Learning (RL) scenario, an agent interacts with its environment in discrete steps. At each step, the agent observes a state s_t and chooses an action a_t based on a policy distribution $\pi(a_t|s_t; \theta)$, with θ representing the parameters of a function approximator. Following the action, the agent is presented with a new

state s_{t+1} and a scalar reward r_t . The goal in a policy-based framework is to adjust θ to enhance the expected total reward $R_t = \sum_{k=0}^{\infty} \gamma^k r_{t+k}$, where $\gamma \in (0, 1]$ serves as the discount factor. Within Advantage Actor-Critic techniques, the approach involves calculating both a policy $\pi(a_t|s_t; \theta)$ that dictates action selection and a value function $V(s_t; \mathbf{w})$ that predicts the expected reward $\mathbb{E}_{\pi}[R_t|s_t = s]$ if the agent follows policy π from state s_t . The value function, or critic, is adjusted through the parameters \mathbf{w} , while the policy, or actor, is modified using the parameters θ .

We developed a deep neural network to serve as a function approximator, modifying its parameters through the process of stochastic gradient descent. The update rules for the parameters are as follows:

$$\theta \leftarrow \theta + \alpha \nabla_{\theta} \log \pi(a_t|s_t; \theta) A(s_t, a_t; \mathbf{w}) \quad (1)$$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha_w \nabla_{\mathbf{w}} V(s_t; \mathbf{w}) A(s_t, a_t; \mathbf{w}) \quad (2)$$

Here, α and α_w denote the learning rates for the policy and value function updates, respectively. The term $A(s_t, a_t; \mathbf{w})$ calculates the n-step advantage for a given state-action pair (s_t, a_t) , incorporating a lookahead of k steps, where $A(s_t, a_t; \mathbf{w}) = \sum_{i=0}^{k-1} \gamma^i r_{t+i} + \gamma^k V(s_{t+k}; \mathbf{w}) - V(s_t; \mathbf{w})$.

The architecture combines the actor and critic parameters θ and \mathbf{w} through a shared weights strategy as depicted in Fig. 2, enhancing the stability of the learning process. Furthermore, we integrate an entropy term $\nabla_{\theta} H(\pi(s_t; \theta))$ into the update equations Eqs. (1) and (2), following the guidance of prior research to regularize the learning phase. Additionally, we implement a clipped surrogate objective [25]:

$$L^{CLIP}(\theta) = \mathbb{E}_t \left[\min(r_t(\theta) \hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon) \hat{A}_t) \right]$$

where ϵ represents a hyper-parameter that needs adjustment. Such objective further refines the stability and efficiency of policy updates during training.

A. Problem Formulation

We address the problem of local path planning, aiming to link two disjointed, neighboring free spaces divided by obstacles that can be pushed around. An agent is introduced to operate within this environment across multiple timesteps, taking actions in a continuous space defined by forward velocity v_x and angular velocity θ_z . At every timestep, the agent earns a reward r_t from the environment, with the goal being to optimize the total return $\sum_{k=0}^{\infty} \gamma^k r_{t+k}$.

The setting is a confined area featuring a narrow path blocked by various pushable objects; the agent/robot needs to move from one place to another by locally pushing objects around. The episode concludes when the agent either reaches the objective or surpasses the maximum duration allowed for an episode. We presuppose the availability of certain preliminary inputs for defining the agent's state: a semantically annotated, coarse occupancy grid accessible through LiDAR or RGB-D sensors coupled with semantic segmentation techniques [26]; bounding boxes around obstacles detected via object recognition software [27]; and data on the agent's instantaneous condition derived from the robot's

TABLE I
REWARDS AT EACH TIMESTEP

reward	description	weight
goal	1 if reach goal, 0 otherwise	10
progress	$[-1, 1] \propto$ velocity towards goal	1
dist	$[0, 1] \propto$ distance to goal	0.1
wall collision	-1 if collision with wall	0.2
object collision	-1 if collision with object	0.1
vel effort	$[-1, 0] \propto v_{actual} - v_{target} $ target velocity	0.05
rot effort	$[-1, 0] \propto \dot{\theta}_{actual} - \dot{\theta}_{target} $	0.1
vel offset	$[-1, 0] \propto v_{actual} - v_{target} $	0.2
rot offset	$[-1, 0] \propto \dot{\theta}_{actual} - \dot{\theta}_{target} $	0.1
time	-1	1

built-in sensors, acknowledging some degree of inaccuracy in all sensory data. Utilizing these initial inputs, we formulate the state of the agent to encompass the target location G , the coordinates of pushable obstacles BV_t , details on the robot's present status (position, velocity, rotation, angular velocity) RP_t , the most recent action a_t , and a semantically annotated occupancy map. To impart temporal context to the agent, we incorporate a sequence of prior states concerning RP_t , BV_t , and a_t , echoing the approach used in previous research [28]. Our choice to integrate the last 5 frames balances effectively between operational efficiency and computational demands. The comprehensive state consists of the occupancy map and a vector detailing G , $RP_{t-4:t}$, $BV_{t-4:t}$, and $a_{t-4:t}$, as illustrated in Fig. 2. For the policy to remain functional and adaptable, we employ a straightforward control scheme based on a unicycle model, which is characterized by two movement parameters: v_x (linear velocity) and $\dot{\theta}_z$ (angular velocity). This policy is designed to be versatile, applicable across various mobile robotic platforms, as demonstrated in the experimental section.

The goal for the agent is to optimize the total cumulative reward. Rewards at each step are outlined in Table I. A significant reward is allocated for successfully completing the objective, with no reward given for incomplete tasks. Incremental positive rewards are awarded each timestep for actions that advance the agent toward the target (*progress*), or for maintaining proximity to the goal (*dist*). Conversely, minimal negative rewards are assigned in relation to the effort required for movement (*vel effort*, *rot effort*), collision with walls (*wall collision*), and interactions with objects while pushing (*object collision*). Additionally, the agent is penalized for significant deviations between intended and executed actions, often caused by object collisions or sudden changes in action (*vel offset*, *rot offset*).

B. Implementation

Our agent's training is conducted in a simulated environment using NVIDIA Isaac Gym [11], a platform that supports the concurrent training of multiple agents on a single GPU, promoting both stability and efficiency in policy-based learning strategies [10]. We adopt the advantage actor-critic method, utilizing a deep neural network that integrates two distinct components of the state space, as depicted in Fig. 2. This state space is comprised of both vector and grid elements, each normalized within the range of $[-1, 1]$. The vector portion undergoes processing in a linear block, which

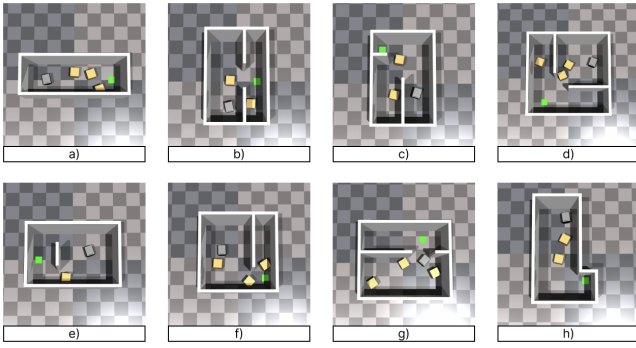


Fig. 3. Fixed maps and random obstacle positions. Agents (gray) need to push object (yellow) to reach the goal (green). The maps include corridors (a,b), mid (c,d), side (e,f), and diagonal (g,h) doorways.

includes a single perceptron layer with 128 units, followed by ReLU activation and a normalization layer. Meanwhile, the grid data is processed through two convolutional blocks, each equipped with convolutional filters, ReLU activation, and batch normalization. Specifically, the first convolutional block utilizes 16 filters sized 8×8 , and the second employs 32 filters of 4×4 . After flattening the output, it is directed through a linear block containing 128 units. Following the concatenation of the two processed streams, the combined data is further processed through two linear blocks, one with 128 units and another with 64 units, leading to the final stage where two separate linear layers generate the value estimate $V(s_t)$ and the proposed action $[v_x, \hat{\theta}_z]$.

C. Scene Generation and Curriculum Training

In Fig. 3, we present eight distinct map layouts created within Isaac Gym, each designed to encapsulate a broad spectrum of local Navigation Among Movable Obstacles path planning among pushable objects scenarios. These configurations include tight corridors (maps *a* and *b*), entryways with adjacent spaces (maps *c* and *d*), entrances flanked by walls (maps *e*, *f*, and *i*), and diagonal doorways (maps *g* and *h*). Agents are placed in specifically designed rooms, each measuring approximately 6×6 square meters, a dimension that accommodates mobile robots ranging from 0.5 to 1 meter in length, enabling their navigation through narrow paths approximately 1 to 2 meters wide. The challenge is heightened by obstacles placed close to these narrow paths, complicating the robot’s ability to traverse them. Robots and their target destinations are randomly positioned within designated zones, with a small probability (5%) that robots may spawn anywhere on the map, ensuring they have the opportunity to explore every area to some extent.

We opt for pushable obstacles (represented as boxes) approximately 60cm^3 in size, limiting the number to a maximum of 5 per room. This quantity was selected as it populates the room with sufficient diversity without overcrowding it, which would complicate the generation of object positions. Random generation of obstacle locations would simplify most scenarios, minimizing the need for the agent to interact with obstacles. To counteract this, obstacles are placed with a skewed probability towards more challenging positions. Specifically, for each obstacle i , there’s a λp_i

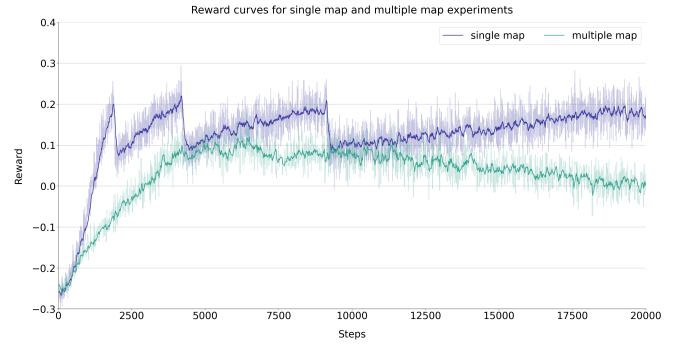


Fig. 4. The progression of rewards through policy update iterations is depicted with the single map scenario in purple and the multi-map scenario in green. The phenomenon of curriculum learning is evident through the steep declines in the purple line, indicating the increasing difficulty of tasks; this effect is mitigated in the multi-map scenario, where the curve is more gradual since not all tasks advance through the curriculum at the same time. While the single-map approach tends toward a steady state of balance, the multi-map approach demonstrates a need for early termination due to its varied progression.

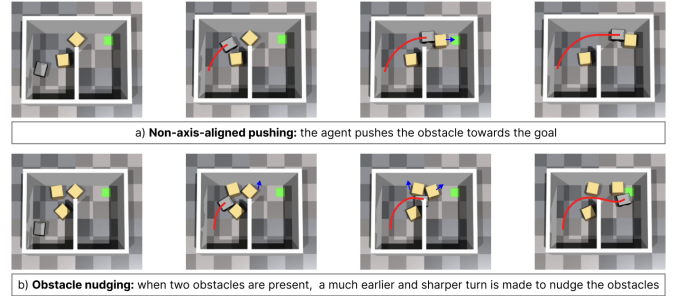


Fig. 5. Qualitative simulated results on single map setting: the agent adapts to different obstacle positions and finds efficient paths.

chance it will appear at a random spot within the room, with an arbitrary orientation. With a $\lambda(1 - p_i)$ likelihood, it is positioned in strategically difficult locations (e.g., blocking a narrow path or situated close to such a path) with slight random adjustments. An obstacle is omitted, represented as zeros in the input data, with a $1 - \lambda$ probability. The probability p_i for each object ranges from 0.2 to 0.6, and λ is incrementally increased from 0.2 in set steps. Through the random placement of pushable objects, we compel the agent to discover solutions for varying obstacle configurations during each attempt. This process encourages the agent to develop an understanding of how different obstacles relate to one another, including determining the sequence in which to move the obstacles.

The parameter λ serves as a mechanism to modulate the number of objects, thus adjusting the complexity of the Navigation Among Movable Obstacles path planning among pushable objects challenge. This methodology introduces a curriculum training strategy, where the difficulty escalates as the agent becomes proficient at simpler tasks. A completion rate threshold (e.g., 90%) determines when the difficulty level increases. Starting from $\lambda = 0.2$, the agent is trained until it achieves a 90% completion rate, at which point λ is incremented by 0.2. This cycle repeats until λ equals 1, ensuring that 4 to 5 objects are consistently present at the commencement of each scene. The impact of this curriculum

learning approach throughout the training is illustrated in Fig. 4.

D. Domain Randomization

Intensive domain randomization is employed across both state and action spaces, incorporating Gaussian white noise to reflect the anticipated variability and noise in real-world sensor readings and object movements. This approach is designed to ensure that the developed policy can effectively adapt to and manage the unpredictability inherent in real-world scenarios. By integrating noise into the state space inputs, we aim to mimic the sensor inaccuracies robots might face, such as variations in the detection of obstacle and robot positions over time. It’s important to note that noise is added separately to the vector and grid states, mirroring the often uncorrelated nature of real-world sensor noise sources. Likewise, we introduce Gaussian noise to the model’s action outputs to replicate unforeseen robot dynamics. The adoption of domain randomization techniques in our training methodology reduces the discrepancy between simulation and real-world applications, thereby enhancing the policy’s reliability in the presence of noisy data and unpredictable dynamics. This preparation ensures policy robustness, a claim we will substantiate in the subsequent section.

IV. EXPERIMENTAL VALIDATION

A. Simulation

Initially, we evaluate the effectiveness of our trained policy through both numerical and observational assessments in a simulated environment. The training is conducted under two distinct scenarios: a single room and multiple rooms. In the single-room scenario, the agent’s training involves navigating through randomly placed objects, as depicted in Fig. 5. Conversely, the multiple-room scenario introduces the agent to eight varied room configurations. Our findings indicate that within the single-room setup, the policy successfully adapts to novel obstacle arrangements within a familiar setting, showcasing efficient strategies and a minimal rate of failures. Meanwhile, in the context of multiple rooms, the policy demonstrates adaptability to new obstacle placements across unfamiliar settings, albeit with slightly less refined behaviors.

For both experimental setups, we implement identical training configurations. The methodology for scene creation and the introduction of variability are outlined in Sec. III, highlighting the random allocation of agent, target, and pushable object locations, the adoption of a progressively challenging curriculum, and the application of domain randomization to both inputs and outputs. The optimization process leverages the ADAM algorithm [29], enhanced with l2 regularization to improve stability [30], and incorporates gradient clipping to mitigate the risk of exploding gradients, a common issue in scenarios involving off-policy learning, approximate function representation, and value bootstrapping (referred to as the deadly triad) [31]. An adaptive learning rate strategy is employed, guided by a predefined KL divergence threshold. The training protocol specifies a horizon of

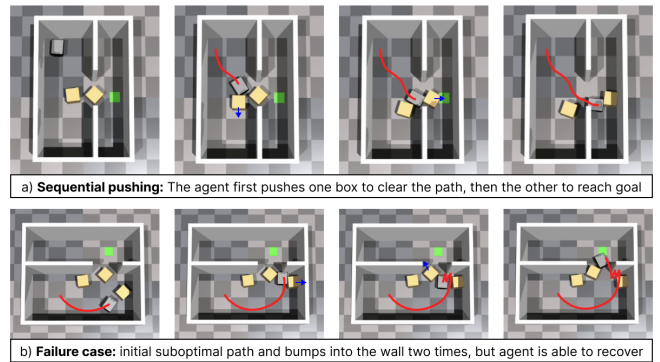


Fig. 6. Observational findings from simulations in various map configurations reveal that: a) the agent strategically maneuvers around obstacles in sequence, optimizing its path with only slight deviations; b) the agent effectively adapts to unforeseen changes in dynamics.

TABLE II
POLICY PERFORMANCE ON SINGLE MAP WITH VARYING λ

λ	objects	completion rate	time taken	objects pushed
0	0	99.9	6.80	0
0.2	0.7	98.5	7.13	0.34
0.4	1.3	98.3	7.68	0.64
0.6	2.1	97.4	8.09	1.00
0.8	2.8	94.5	8.14	1.37
1	3.7	91.0	8.69	1.92

50 actions, with policy updates occurring every 20 physics simulation steps (equivalent to 333 milliseconds). Episodes are capped at a duration of 45 seconds, translating to 2700 physics simulation steps. The training process engages 4,000 parallel environments, with each update drawing on a mini-batch of 2,000 samples. This regimen is executed on an NVIDIA RTX 3080 GPU, spanning 20,000 policy update iterations.

In our initial experiment using a singular map layout, the policy undergoes training within a consistent map (Fig. 5) with varied starting points for the agent, destination, and pushable obstacles. The evaluation is conducted over 1,000 unique scenarios within the identical map setup but featuring previously unseen obstacle placements. This evaluation process is replicated six times, each instance adjusting the λ parameter to increment the obstacle count within the scene, thus escalating the complexity of the challenges faced. Fig. 5 showcases exemplary outcomes from this experiment. Observations from the upper section indicate the agent’s capability to maneuver obstacles in a non-linear manner along its path. The lower segment illustrates the agent executing a tighter maneuver near a doorway, opting to slightly displace obstacles to facilitate passage, a strategy necessitated by the potential blockage of the goal (indicated in green) if the obstacles were pushed directly into the doorway. These actions underscore the adaptability and strategic planning of the agent, surpassing the limitations of axis-aligned obstacle manipulation commonly seen in previous research.

The numerical outcomes of this study are detailed in Table II, with each row indicating a progressive increase in challenge level corresponding to higher λ values. It’s important to highlight that the actual count of objects present

TABLE III
COMPARISON BETWEEN POLICY PERFORMANCE ON TRAINING MAPS
AND TEST MAP, WITH $\lambda = 0.8$

Experiment	completion rate	time taken (s)	objects pushed
Training (8 maps)	79.8	10.99	1.12
Testing (unseen map)	54.3	13.05	1.51

in each scenario doesn't linearly correlate with λ , due to the stochastic nature of object placement and the occasional lack of available space for additional objects. As the scenarios grow more complex with an increased presence of objects, we note a decline in the success rate, primarily when the agent encounters situations where pathways or the goal itself become obstructed, rendering progress impossible. In the table's final column, we document the average number of objects displaced during successful task completions. Interestingly, this figure typically represents only half of the average total object count in the scene, suggesting the agent prioritizes moving only those obstacles that directly impede its route. Across the single map experiments, the agent exhibits consistent and effective behavior, adeptly adapting to new obstacle arrangements and achieving a solution in 91% of instances, even under the most demanding conditions.

In our subsequent experiment, we assess the ability of a singular policy, trained across multiple map configurations, to adapt not only to novel obstacle placements but also to entirely new environments. The training encompasses maps labeled *a* through *h*, as depicted in Fig. 3, followed by testing the policy in 1,000 scenarios within an unfamiliar map layout (the same layout utilized in the single map experiment). Maintaining stable learning progress in this multifaceted context proves to be challenging. Despite the implementation of various regularization strategies, achieving consistent policy convergence is not always possible. As illustrated in Fig 4, the policy's performance fails to match that observed in the single map scenario. Consequently, we resort to early stopping, selecting the most effective policy version identified during the training phase for further application.

Within the context of training across various maps, the agent develops adaptable strategies suitable for diverse scenarios, such as the methodical displacement of obstacles and the ability to dynamically respond to unforeseen events and inaccuracies in sensory input. An instance of this is demonstrated in Fig. 6-a, where the agent executes non-linear, ordered obstacle movement (preferring to move one object before another). This figure underscores the agent's proficiency in fine-tuning its approach by making only minor deviations from its intended path, a feat difficult to achieve under the limitations of axis-aligned movement restrictions. Fig 6-b showcases the agent's capability to navigate out of less favorable situations. Challenges in accurately identifying the doorway due to the coarse resolution of input images may lead the agent to mistakenly contact a wall adjacent to a doorway, as depicted in frame 2. Nevertheless, the agent effectively maneuvers out of such predicaments by reversing and then proceeding forward again, as illustrated

in frames 3 and 4. Although this approach results in a policy that may not be as refined as that developed under a single map framework, it evidentially highlights the agent's resourcefulness in rectifying its course from disadvantageous positions.

The numerical outcomes are detailed in Table III, assessing the agent's efficacy within familiar settings (the eight original maps) against novel obstacle configurations and entirely new environments. The agent manages an approximate success rate of 80% across all eight maps in the most challenging scenarios, while it secures a 54% success rate in a novel map setting. Although the performance in the unfamiliar environment is comparatively lower, it nonetheless evidences a notable capability. Enhancing performance further could likely be achieved through increased randomization in training data and leveraging prioritized experience replay to re-engage with more complex path planning problems [32].

B. Failure Cases

In scenarios involving a single map, the predominant cause of failure is typically the agent pushing an object to a position from which the path becomes irreversibly obstructed. Conversely, the multi-map approach exhibits a higher incidence of failures, often characterized by the agent making superfluous maneuvers or occasionally becoming stuck in a corner, as depicted in Fig. 6. These issues are likely the result of inadequate feature discernment by the network, which is trained on data exhibiting high inter-correlation, coupled with the limited detail available in the input imagery. Enhancements in feature extraction could be achieved by expanding the variety of training environments and utilizing experience replay techniques to reduce data correlation. Similarly, an increase in the resolution of the input images could further mitigate these issues.

C. Robot Experiments

This section details our results using an actual quadruped robot (Unitree Go1), outfitted with aluminum extensions to aid in pushing obstacles. We utilize cardboard boxes as pushable objects, each measuring approximately $50 \times 50 \times 50 \text{ cm}^3$. Due to the interaction between the boxes and the carpet, the robot is limited to pushing a single box at a time. To monitor the positions of both the obstacles (edges) and the robot (orientation), two overhead cameras equipped with ArUco Markers are employed. This setup allows us to generate a 2D grid that forms the basis of the state space. While the specifics of obstacle recognition and mapping are not covered within this document, we demonstrate the robot's capability to navigate through sensor discrepancies and unpredictable obstacle movements. The decision to apply a neural network trained in a singular map environment is due to its demonstrated reliability and stability, underscoring the feasibility of applying the developed policy in real-world scenarios, characterized by variable robot dynamics and sensor data.

Our evaluations indicate that the robot is adept at maneuvering through confined spaces that include pushable obsta-



Fig. 7. Left to right: $[t=0]$ Green dot is the goal. The robot: $[t=4]$ avoids collision with object 3; $[t=8]$ moves box 2 to create small opening; $[t=10]$ stops pushing object 2 and rotates towards box 1; $[t=12]$ pushes object 1 out of the door. $[t=23]$ Goal reached. The robot performs non-axis-aligned pushing.

cles, as illustrated in Fig. 7. The objective for the robot is to locate a designated green point, effectively addressing the Navigation Among Movable Obstacles (local path planning challenge by circumventing collision with box 3 (by time $t = 4$), strategically rotating and slightly moving box 2 to make way (by time $t = 10$), and displacing box 1 from the entrance until the target is reached (by time $t = 23$). It was commonly noted that the robot employs pushing actions along curved paths, utilizing minimal adjustments to alter an obstacle’s orientation, frequently resulting in the most efficient method for clearing narrow passages.

V. CONCLUSIONS

In our study, we introduced a deep reinforcement learning strategy enabling a robot to execute non-linear obstacle manipulation to address local path planning among pushable objects, maintaining constant computational complexity. This approach was validated both in simulated environments and through practical implementation on a physical robot, accounting for sensor inaccuracies. Looking ahead, our goals include refining the learning process for agents operating across diverse map layouts. Furthermore, we plan to explore strategies enabling agents to navigate around obstacles of uncertain movability and novel geometries by integrating these variables into their training regimen.

REFERENCES

- [1] K. Cai, C. Wang, J. Cheng, C. W. De Silva, and M. Q.-H. Meng, “Mobile Robot Path Planning in Dynamic Environments: A Survey,” *arXiv preprint arXiv:2006.14195*, 2020.
- [2] E. D. Demaine, M. L. Demaine, and J. O’Rourke, “PushPush and Push-1 are NP-hard in 2D,” *12th CCCG*, pp. 211–219, 2000.
- [3] M. Stilman and J. Kuffner, “Planning Among Movable Obstacles with Artificial Constraints,” *IJRR*, vol. 27, no. 11-12, pp. 1295–1307, 2008.
- [4] S. K. Moghaddam and E. Masehian, “Planning Robot Navigation among Movable Obstacles (NAMO) through a Recursive Approach,” *Journal of Intelligent & Robotic Systems*, vol. 83, pp. 603–634, 2016.
- [5] Y. Kakiuchi, R. Ueda, K. Kobayashi, K. Okada, and M. Inaba, “Working with Movable Obstacles using On-line Environment Perception Reconstruction using Active Sensing and Color Range Sensor,” in *IEEE/RSJ IROS*, 2010, pp. 1696–1701.
- [6] J. Scholz, N. Jindal, M. Levihn, C. L. Isbell, and H. I. Christensen, “Navigation Among Movable Obstacles with Learned Dynamic Constraints,” in *IEEE/RSJ IROS*, 2016, pp. 3706–3713.
- [7] S. Armleder, E. Dean-Leon, F. Bergner, J. R. Guadarrama Olvera, and G. Cheng, “Tactile-Based Negotiation of Unknown Objects during Navigation in Unstructured Environments with Movable Obstacles,” *Advanced Intelligent Systems*, p. 2300621, 2-24.
- [8] M. Stilman and J. J. Kuffner, “Navigation Among Movable Obstacles: Real-Time Reasoning in Complex Environments,” *World Scientific IJRR*, vol. 2, no. 04, pp. 479–503, 2005.
- [9] P. E. Hart, N. J. Nilsson, and B. Raphael, “A Formal Basis for the Heuristic Determination of Minimum Cost Paths,” *IEEE Transactions on Systems Science and Cybernetics*, vol. 4, no. 2, pp. 100–107, 1968.
- [10] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous Methods for Deep Reinforcement Learning,” in *ICML*, 2016, pp. 1928–1937.
- [11] V. Makoviychuk *et al.*, “Isaac Gym: High Performance GPU-Based Physics Simulation for Robot Learning,” *arXiv preprint arXiv:2108.10470*, 2021.
- [12] L. Liu, X. Wang, X. Yang, H. Liu, J. Li, and P. Wang, “Path Planning Techniques for Mobile Robots: Review and Prospect,” *Expert Systems with Applications*, vol. 227, p. 120254, 2023.
- [13] X. Shen, “Visual Local Path Planning Based on Deep Reinforcement Learning,” in *IEEE 2nd International Conference on Electrical Engineering, Big Data and Algorithms (EEBDA)*, 2023, pp. 1616–1621.
- [14] G. Wilfong, “Motion Planning in the Presence of Movable Obstacles,” in *4th Annual Symp. on Computational Geometry*, 1988, pp. 279–288.
- [15] D. Nieuwenhuisen, A. Stappen, and M. H. Overmars, “An Effective Framework for Path Planning Amidst Movable Obstacles,” in *Springer Algorithmic Foundation of Robotics VII*, 2008, pp. 87–102.
- [16] J. v. d. Berg, M. Stilman, J. Kuffner, M. Lin, and D. Manocha, “Path Planning Among Movable Obstacles: a Probabilistically Complete Approach,” in *Springer Algorithmic Foundation of Robotics VIII*, 2009.
- [17] V. S. Raghavan, D. Kanoulas, D. G. Caldwell, and N. G. Tsagarakis, “Reconfigurable and Agile Legged Wheeled Robot Navigation in Cluttered Environments With Movable Obstacles,” *IEEE Access*, 2022.
- [18] H. N. Wu, M. Levihn, and M. Stilman, “Navigation Among Movable Obstacles in Unknown Environments,” in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1433–1438.
- [19] M. Levihn, J. Scholz, and M. Stilman, “Hierarchical Decision Theoretic Planning for Navigation Among Movable Obstacles,” in *Springer Algorithmic Foundations of Robotics X*, 2013, pp. 19–35.
- [20] H.-C. Wang, S.-C. Huang, P.-J. Huang, K.-L. Wang, Y.-C. Teng, Y.-T. Ko, D. Jeon, and I.-C. Wu, “Curriculum Reinforcement Learning From Avoiding Collisions to Navigating Among Movable Obstacles in Diverse Environments,” *Robotics and Automation Letters*, vol. 8, no. 5, pp. 2740–2747, 2023.
- [21] B. Renault, J. Saraydaryan, and O. Simonin, “Towards S-NAMO: Socially-aware Navigation Among Movable Obstacles,” in *Springer Robot World Cup*, 2019, pp. 241–254.
- [22] K. Ellis, H. Zhang, D. Stoyanov, and D. Kanoulas, “Navigation Among Movable Obstacles with Object Localization using Photorealistic Simulation,” in *IEEE/RSJ IROS*, 2022.
- [23] K. Ellis *et al.*, “Navigation Among Movable Obstacles via Multi-Object Pushing Into Storage Zones,” *IEEE Access*, 2023.
- [24] F. Xia, W. B. Shen, C. Li, P. Kasimbeg, M. E. Tchappin, A. Toshev, R. Martín-Martín, and S. Savarese, “Interactive Gibson Benchmark: A Benchmark for Interactive Navigation in Cluttered Environments,” *IEEE RA-L*, vol. 5, no. 2, pp. 713–720, 2020.
- [25] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, “Proximal Policy Optimization Algorithms,” *arXiv preprint arXiv:1707.06347*, 2017.
- [26] A. Rosinol, M. Abate, Y. Chang, and L. Carlone, “Kimera: an Open-Source Library for Real-Time Metric-Semantic Localization and Mapping,” in *IEEE ICRA*, 2020, pp. 1689–1696.
- [27] B. Wen, J. Tremblay, V. Blukis, S. Tyree, T. Müller, A. Evans, D. Fox, J. Kautz, and S. Birchfield, “BundleSDF: Neural 6-DoF tracking and 3D reconstruction of unknown objects,” in *CVPR*, 2023.
- [28] V. Mnih *et al.*, “Human-level Control through Deep Reinforcement Learning,” *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
- [29] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” in *3rd ICLR*, 2014.
- [30] J. Farebrother, M. C. Machado, and M. Bowling, “Generalization and Regularization in DQN,” *arXiv preprint arXiv:1810.00123*, 2018.
- [31] R. S. Sutton and A. G. Barto, *RL: An Introduction*. MIT press, 2018.
- [32] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, “Prioritized Experience Replay,” *arXiv preprint arXiv:1511.05952*, 2015.