

Robot Navigation with Map-Based Deep Reinforcement Learning

Guangda Chen, Lifan Pan, Yu'an Chen, Pei Xu, Zhiqiang Wang, Peichen Wu,
Jianmin Ji* and Xiaoping Chen

Abstract—This paper proposes an end-to-end deep reinforcement learning approach for mobile robot navigation with dynamic obstacles avoidance. Using experience collected in a simulation environment, a convolutional neural network (CNN) is trained to predict proper steering actions of a robot from its egocentric local occupancy maps, which accommodate various sensors and fusion algorithms. The trained neural network is then transferred and executed on a real-world mobile robot to guide its local path planning. The new approach is evaluated both qualitatively and quantitatively in simulation and real-world robot experiments. The results show that the map-based end-to-end navigation model is easy to be deployed to a robotic platform, robust to sensor noise and outperforms other existing DRL-based models in many indicators.

Index Terms—robot navigation, obstacle avoidance, reinforcement learning, occupancy map.

I. INTRODUCTION

One of the major challenges for robot navigation is to develop a safe and robust collision avoidance policy to navigate from the starting position to desired goal without running into obstacles and pedestrians in unknown cluttered environments. Although numerous methods have been proposed [1], conventional methods are often built upon a set of assumptions that are likely not to be satisfied in practice [2], and may impose intensive computational demand [3]. In addition, conventional algorithms normally involve a number of parameters that need to be tuned manually [4] rather than being able to learn from past experience automatically [5]. It is difficult for these approaches to generalize well to unanticipated scenarios.

Recently, several supervised and self-supervised deep learning approaches have been applied to robot navigation. Giusti et al. [6] used a Deep Neural Network to classify the images to determine which action will keep the quadrotor on the trail. Lei et al. [7] showed the effectiveness of a hierarchical structure that fuses a convolutional neural network (CNN) with a decision process, which is a highly compact network structure that takes raw depth images as input, and generates control commands as network output. Pfeiffer et al. [8] presented a model that is able to learn the complex mapping from raw 2D-laser range findings and a target position to

the required steering commands for the robot. However, there are some limitations that prevent these approaches from being widely applied in a real robotic setting. For instance, a massive manually labeled dataset is required for the training of the supervised learning approaches. Although this can be migrated to an extent by resorting to self-supervised learning methods, their performance is largely bounded by the strategy generating training labels.

On the other hands, deep reinforcement learning (DRL) methods have achieved remarkable success in many challenging tasks [9]–[11]. Different from previous supervised learning methods, DRL based approaches learn from a large number of trials and corresponding rewards instead of labeled data. In order to learn a sophisticated control policy with reinforcement learning, robots need to interact with the environment for a long period to accumulate knowledge about the consequences of different actions. Collecting such interaction data in real world is expensive, time consuming, and sometimes infeasible due to safety issues [12]. For instance, Kahn et al. [5] proposed a generalized computation graph that subsumes value-based model-free methods and model-based methods, and then instantiated this graph to form a navigation model that learns from raw images and is sample efficient. However, It takes hours of destructive self-supervised training to navigate only dozens of meters without collision through a indoor environment. Because of the excessive number of trials required to learn a good policy, training in a simulator is more suitable than experiences derived from the real world. Then we can exploit the close correspondence between a simulator and the real world, to transfer the learned policy.

According to the difference of input data, the existing reinforcement learning-based robot motion planning methods can be roughly divided into two categories: agent-level inputs and sensor-level inputs. And there are different transferability to real world between different input data. As representatives of agent-level methods, Chen et al. [13] train an agent-level collision avoidance policy using DRL, which maps an agent's own state and its neighbors' states to collision-free action. However, it demands perfect sensing. Obviously, this complex pipeline not only requires expensive online computation but makes the whole system less robust to the perception uncertainty.

As for sensor-level inputs, the types of sensor data used in DRL-based navigation mainly include 2D laser range inputs, depth images and color images. The network proposed in [14] outputs control commands based on ten-dimensional laser range inputs and is trained using an asynchronous

This work is partially supported by the 2030 National Key AI Program of China 2018AAA0100500, the National Natural Science Foundation of China (No. 61573386), and Guangdong Province Science and Technology Plan Projects (No. 2017B010110011).

The authors are with the School of Computer Science and Technology, University of Science and Technology of China, Hefei, Anhui, 230026, P.R.China. {cgdsss, lifanpan, an11099, xp816, tt1248163264, wpc16}@mail.ustc.edu.cn, {jianmin, xpchen}@ustc.edu.cn.

* Corresponding author.

DRL algorithm. Similarly, the models introduced in [15] and [12] also derive the steering commands from laser range sensors. 2D laser-based methods are competitive in terms of the transferability to real world because of the smaller discrepancies between their simulated and real domains. However, 2D sensing data is not enough to describe complex 3D scenarios. On the contrary, vision sensors can provide 3D sensing informations. But RGB images suffer from the significant deviation between real-world situations and the simulation environments during training, which leads to quite limited generalization across situations. Compared to RGB images, the depth inputs in simulations exhibit much better visual fidelity due to the textureless nature and, as a result, greatly alleviate the burden of transferring the trained model to real deployments [16]. Based on depth images, Zhang et al. [17] proposed to use successor features to achieve efficient knowledge transfer across tasks in depth-based navigation. Currently, all existing sensor-level works rely on specific sensor types and configurations. While in complex environments, most robots are equipped with different sensors to navigate autonomously and safely [18].

In this paper, we propose an end-to-end model-free deep reinforcement learning algorithm to improve the performance of autonomous decision making in complex environments, which directly maps local probabilistic costmaps to an agent’s steering commands in terms of target position and robot velocity. Compared to previous work on DRL-based obstacle avoidance, our motion planner is based on probabilistic costmaps to represent environment and target position, which enables the learned collision avoidance policy to handle different types of sensor input efficiently, such as the multi-sensor information from 2D/3D range scan finders or RGB-D cameras [19]. And our trained CNN-based policy is easily transferred and executed on a real-world mobile robot to guide its local path planning and robust to sensor noise. We evaluate our DRL agents both in simulation and on-robot qualitatively and quantitatively. Our results show the improvement in multiple indicators over the DRL-based obstacle avoidance policy.

Our main contributions are summarized as follow:

- Formulate the obstacle avoidance for mobile robots as an DRL problem based on a generated costmap, which can handle multi-sensor fusion and is robust to sensor noise.
- Integrate curriculum learning technique to enhance the performance of dueling double DQN with prioritized experience replay.
- Contract a variety of real-world experiments to reveal the high adaptability of our model to transfer to different sensor configurations and environments.

The rest of this paper is organized as follows. The structure of the DRL-based navigation system is presented in Section II. The deep reinforcement learning algorithm for obstacle avoidance based on egocentric local occupancy maps is described in Section III. Section IV presents experimental results, followed by conclusions in Section V.

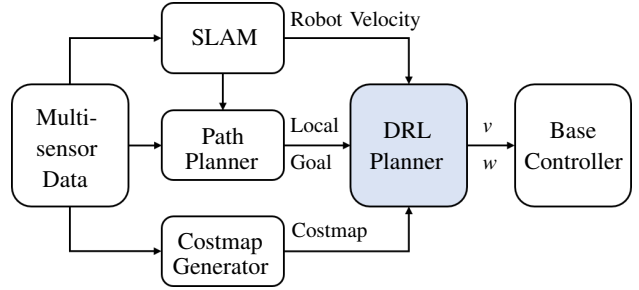


Fig. 1. Block diagram of the DRL-based navigation system for autonomous robots.

II. SYSTEM STRUCTURE

The proposed DRL-based mobile robot navigation system consists of six modules. As shown in Fig. 1, simultaneous localization and mapping (SLAM) module establishes an environment map based on sensor data, and can estimate the position and velocity of the robot in the map simultaneously. When a target position is received, the path planner module generates a path or a series of local goal points from the current position to the target position. In order to cope with the dynamic and complex environments, a safe and robust collision avoidance policy in unknown cluttered environments is required. In addition to the local goal points from the path planning module and robot velocity provided by the positioning module, our local planner also needs the input of the surrounding environment information, which is the egocentric occupancy map produced by the costmap generator module based on multi-sensor data. Generally speaking, Our DRL-based local planner takes the velocity information generated by the SLAM module, the local goal generated by the global path planner and the cost maps from the generator which can fuse multi-sensor information, and outputs the linear velocity v and angular velocity w of the robot. Finally, the output speed command is executed by the base controller module which depends on the specific kinematics of the robot.

III. APPROACH

We begin this section by introducing the problem formulation of the local obstacle avoidance. Next, we describe the key ingredients of our reinforcement learning algorithm and the details about the network architecture of the collision avoidance policy.

A. Problem Formulation

At each timestamp t , given a frame sensing data s_t , a local goal position g_t in the robot coordinate system and the robot linear velocity v_t , angular velocity ω_t , the proposed local obstacle avoidance policy provides an action command a_t as follows:

$$\mathbf{M}_t = f_\lambda(s_t) \quad (1)$$

$$a_t = \pi_\theta(\mathbf{M}_t, g_t, v_t, \omega_t) \quad (2)$$

where \mathbf{M}_t is a local cost map describing the obstacle avoidance task, λ and θ are model parameters. Specifically, the

cost map \mathbf{M}_t is constructed as an aggregate of robot configurations and the obstacle penalty, which will be explained below.

Hence, the robot collision avoidance problem can be simply formulated as a sequential decision making problem. The sequential decisions consisting of observations $\mathbf{o}_t \sim [\mathbf{M}_t, \mathbf{g}_t, \mathbf{v}_t, \boldsymbol{\omega}_t]$ and actions (velocities) $\mathbf{a}_t \sim [v_{t+1}, \boldsymbol{\omega}_{t+1}]$ ($t=0:t_g$) can be considered as a trajectory l from its start position \mathbf{p}_0 to its desired goal \mathbf{p}_g , where t_g is the traveled time. Our goal is to minimize the expectation of the arrival time and take into account that robots do not collide, which is defined as:

$$\begin{aligned} \underset{\pi_\theta}{\operatorname{argmin}} \mathbb{E}[t_g | \mathbf{a}_t = \pi_\theta(\mathbf{o}_t), \\ \mathbf{p}_t = \mathbf{p}_{t-1} + \mathbf{a}_t \cdot \Delta t, \\ \forall k \in [1, N]: \|\mathbf{p}_t - (\mathbf{p}_{obs})_k\| > R] \end{aligned} \quad (3)$$

where \mathbf{p}_{obs} is the position of obstacle and R is the robot radius.

B. Dueling DDQN with prioritized experience replay

Markov Decision Processes (MDPs) provide a mathematical framework to model stochastic planning and decision-making problems under uncertainty. An MDP is a tuple $M = (S, A, P, R, \gamma)$, where S indicates the state space, A is the action space, P indicates the transition function which describes the probability distribution over states if an action a is taken in the current state s , R represents the reward function which illustrates the immediate state-action reward signal, and $\gamma \in [0, 1]$ is the discount factor. In an MDP problem, a policy $\pi(a|s)$ specifies the probability of mapping state s to action a . The superiority of a policy π can be assessed by the action-value function (Q-value) defined as:

$$Q^\pi(s, a) = \mathbb{E}^\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) | s_0 = s, a_0 = a \right] \quad (4)$$

Therefore, the action-value function is the expectation of discounted sums of rewards, given that action a is taken in state s and policy π is executed afterwards. The objective of the agent is to maximize the expected cumulative future reward, and this can be achieved by adopting the Q-learning algorithm which approximates the optimal action-value function iteratively using the Bellman equation:

$$Q^*(s_t, a_t) = R(s_t, a_t) + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}) \quad (5)$$

Combined with deep neural networks, DQN enables reinforcement learning to cope with complex high-dimensional problems. Generally, DQN maintains two deep neural networks, including an online network with parameters θ and a separate target network with parameters θ' . The parameters of the online network are updated constantly by minimizing the loss function $(y_t - Q(s_t, a_t; \theta'))^2$, where y_t can be calculated as follows:

$$y_t = \begin{cases} r_t & \text{if episode ends} \\ r_t + \gamma \max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta') & \text{otherwise} \end{cases} \quad (6)$$

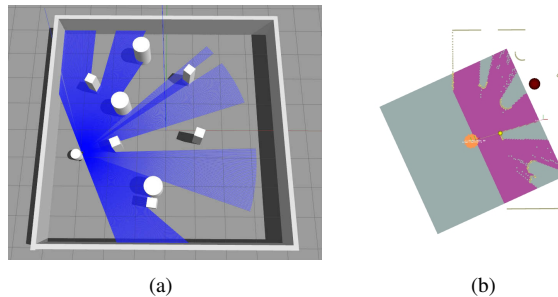


Fig. 2. Gazebo training environments (left) and corresponding occupancy map displayed by Rviz (right).

And the parameters of the target network are fixed for generating Temporal-Difference (TD) targets and synchronized regularly with those of the online network.

Conventional Q-learning is affected by an overestimation bias, due to the maximization step in Equation (6), which would harm the learning process. Double Q-learning [20], addresses this overestimation by decoupling, in the maximization performed for the bootstrap target, the selection of the action from its evaluation. Therefore, if episode not ends, the y_t in the above formula is rewritten as follows:

$$y_t = r_t + \gamma Q(s_{t+1}, \underset{a_{t+1}}{\operatorname{argmax}} Q(s_{t+1}, a_{t+1}; \theta); \theta') \quad (7)$$

In this work, dueling networks [21] and prioritized replay [22] are also deployed for more reliable estimation of Q-values and sampling efficiency of replay buffer respectively. In the following, we describe the details of the observation space, the action space, the reward function and the network architecture.

1) *Observation space*: As mentioned in Section III-A, the observation \mathbf{o}_t consists of the generated costmaps \mathbf{M}_t , the relative goal position \mathbf{g}_t and the robot's current velocity \mathbf{v}_t . Specifically, \mathbf{M}_t represents the cost map images generated from a 180-degree laser scanner or other sensors. The relative goal position \mathbf{g}_t is a 2D vector representing the goal coordinate with respect to the robot's current position. The observation \mathbf{v}_t includes the current translational and rotational velocity of the differential-driven robot.

We use layered costmaps [23] to represent environmental information perceived by multi-sensors. Then though the map generator module, we get the state maps \mathbf{M}_t by drawing the robot configuration (shape) into the layered costmaps. Fig. 2(b) shows an example of generated occupancy maps.

2) *Action space*: The action space is a set of permissible velocities in discrete space. The action of differential robot includes the translational and rotational velocity, i.e. $\mathbf{a}_t = [v_t, w_t]$. In this work, considering the real robots kinematics and the real world applications, we set the range of the translational velocity $v \in [0.0, 0.2, 0.4, 0.6]$ and the rotational velocity in $w \in [0.9, -0.6, -0.3, 0.0, 0.3, 0.6, 0.9]$. Note that backward moving (i.e. $v < 0.0$) is not allowed since the laser range finder can not cover the back area of the robot.

3) *Reward*: The reward function in reinforcement learning implicitly specifies what the agent is encourage to do.

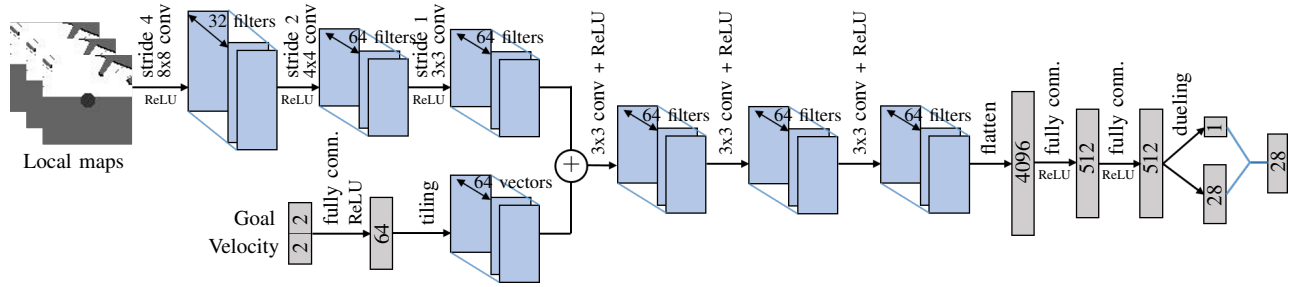


Fig. 3. The architecture of our CNN-based dueling DQN network. This network takes three local maps and a vector with local goal and robot velocity as input and outputs the Q values of 28 discrete actions.

Our objective is to avoid collisions during navigation and minimize the mean arrival time of the robot. A reward function is designed to guide robots to achieve this objective:

$$r_t = (r^g)_t + (r^c)_t + (r^s)_t \quad (8)$$

The reward r at time step t is a weighted sum of three terms: r^g , r^c and r^s . In particular, the robot is awarded by $(r^g)_t$ for reaching its goal:

$$(r^g)_t = \begin{cases} r_{arr} & \text{if } \|\mathbf{p}_t - \mathbf{g}\| < 0.2 \\ \varepsilon(\|\mathbf{p}_{t-1} - \mathbf{g}\| - \|\mathbf{p}_t - \mathbf{g}\|) & \text{otherwise} \end{cases} \quad (9)$$

When the robot collides with other obstacles in the environment, it is penalized by $(r^c)_t$:

$$(r^c)_t = \begin{cases} r_{col} & \text{if collision} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

And we also give robots a small fixed penalty r^s at each step. We set $r_{arr} = 500$, $\varepsilon = 10$, $r_{col} = -500$ and $r^s = -5$ in the training procedure.

4) *Network Architecture*: We define a CNN-based deep convolutional neural network that computes the action-value function for each actions. The input of the network has three local maps \mathbf{M} with 60×60 grey pixels and a four-dimensional vector with local goal \mathbf{g} and robot velocity \mathbf{v} . The output of the network is the Q-values for actions. The architecture of our deep Q-value network is shown in Figure 3. The input costmaps are fed into a 8×8 convolution with stride 4, followed by a 4×4 convolution with stride 2 and a 3×3 convolution with stride 1. The local goal and robot velocity form a four-dimensional vector, which is processed by one fully connected layer, and is then pointwise added to each point in the response map of processed costmap by tiling the output over the special dimensions. The result is then processed by three 3×3 convolutions and three fully connected layers with 512, 512 units respectively, and then fed into the dueling network architecture, after which the network outputs the Q values of 28 discrete actions.

C. Curriculum Learning

Curriculum learning [24] is a learning strategy in machine learning, which starts from easy instances and then gradually handles harder ones. In this work, we use Gazebo simulator [25] to build an environment with multiple obstacles. As the training progresses, we gradually increase the number of obstacles in the environment, and also the distance from the

starting point of the target point gradually increases. This makes our strategy training from easy situation to difficult ones. At the same time, the position of each obstacle and the start and end points of the robot are automatically random during all training episodes. One training scene is shown in Fig. 2(a).

IV. EXPERIMENTS

In this section, we present experiment setup and evaluation in both simulated and real environments. We quantitatively and contrastively evaluate the DQN-based navigation policy to show that it outperforms other related methods in multiple indicators. Moreover, we also performed qualitative tests on real robots, and also integrated our obstacle avoidance policy into the navigation framework for long-range navigation testing.

A. Reinforcement Learning Setup

The training experiments are conducted with a customized differential drive robot in a simulated virtual environment using Gazebo. A 180-degree laser scanner is mounted on the front of the robot as shown in Fig. 2(a). The system parameters are empirically determined in terms of both the performance and our computation resource limits as listed in Table I.

TABLE I
SYSTEM PARAMETERS

Parameter	Value
learning rate	5×10^{-4}
discount factor	0.99
replay buffer size	2×10^5
minibatch size	1024
image size	60×60
episode length	300
initial exploration	1
final exploration	0.1

The implementation of our algorithm is in TensorFlow, and we train the deep Q-network in terms of objective function Eq. 7 with the Adam optimizer [26]. The training hardware is a computer with an i9-9900k CPU and a single NVIDIA GeForce 2080 Ti GPU. The entire training process (including exploration and training time) takes about 10 hours for the policy to converge to a robust performance.

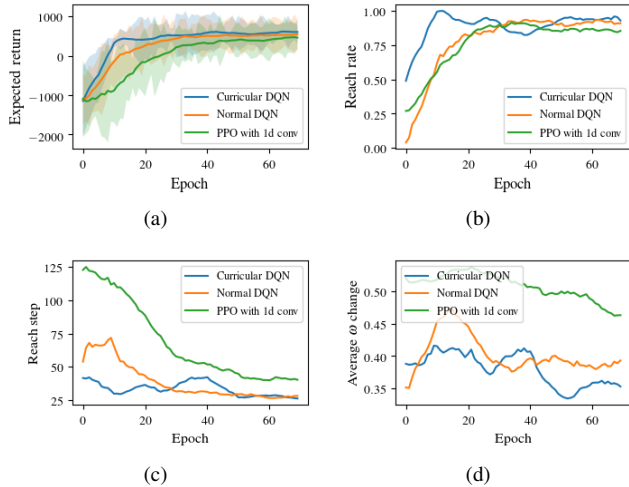


Fig. 4. Indicator curves of different methods during training.

B. Experiments on simulation scenarios

1) *Performance metrics*: To compare the performance of our approach with other methods over various test cases, we define the following performance metrics.

- *Expected return* E_r is the average of the sum of rewards of episodes.
- *Success rate* $\bar{\pi}$ is the ratio of the episodes in which the robot reaching the goals within a certain step without any collisions over the total episodes.
- *Reach step* \bar{s} is the average number of steps required to successfully reach the target point without any collisions.
- *Average angular velocity change* $\nabla\omega$ is the average of the angular velocity changes for each step, which reflects the smoothness of the trajectory.

TABLE II
INDICATORS VALUES OF VARIOUS METHODS

Method	E_r	$\bar{\pi}$	\bar{s}	$\nabla\omega$
PPO with 1d conv	467.87	0.85	40.19	0.46
Normal DQN	547.43	0.91	27.76	0.39
Curricular DQN	617.04	0.94	26.13	0.35

2) *Comparative experiments*: We compare our curricular DQN policy with normal (non-curricular) DQN policy and PPO with one-dimensional convolution network [15] in our tests. As shown in Fig. 4, our DQN-based policy has significant improvement over PPO policy in terms of expected return, success rate, reach step and average angular velocity change, and curricular DQN policy has also a slight improvement on multiple indicators. In the tests with more obstacles, the specific indicators values of various method are shown in the Table II. Fig. 5 shows a test case of our curricular DQN policy in a test scenario.

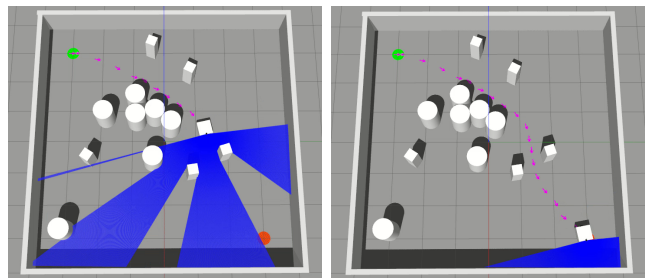


Fig. 5. A test case of our curricular DQN policy in a test scenario, the green dot represents the starting point, the red dot represents the end point, and the robot's trajectory is marked with purple arrows.

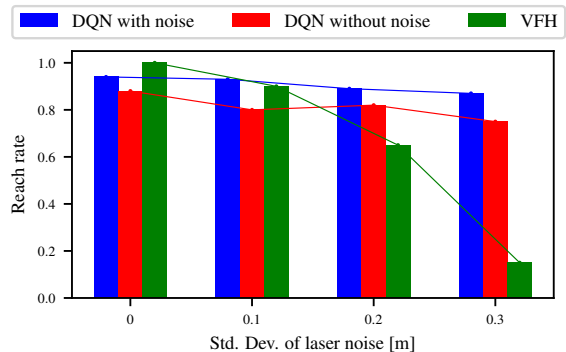


Fig. 6. Reach rate over laser noise. DQN policy trained with sensor noise compared to policy without sensor noise and traditional VFH method.

3) *Robustness to noise*: Fig. 6 depicts that the performances of our DQN-based policy and the traditional vector field histogram (VFH) method [27] vary with the noise error of the laser sensor data in an environment with many obstacles. Results show that our DQN-based policy is resilient to noise, and laser noise heavily influences the reach rate of VFH. This is expected since VFH uses obstacle clearance to calculate its objective function, and such a greedy approach often guides the robot to local minima. More importantly, the learned policy (the DQN policy with noise in the Fig. 6) will work better when using the same noise variance as the test environment during training.

C. Navigation in real-world

To further verify the generalization and effectiveness of our learning policy, we use our robot chassis to do experiments in real-world. As shown in Fig. 7, our robot platform is a differential wheel robot with a Hokuyo UTM-30LX scanning laser rangefinder and a laptop with an i7-8750H CPU and a NVIDIA 1060 GPU. The robot pose and velocity are provided by a particle filter based state estimator. An occupancy map is constructed from laser measurement, from which an egocentric local map is cropped to fixed size $6.0 \times 6.0\text{m}$ and resolution 0.1m at each cycle.

We used paper boxes to build different difficult environments for testing. As shown in Fig. 7, when the robot confronts obstacles, the trained policy succeeds in providing a reactive action command that drives the robot away from



Fig. 7. Our robot chassis with a laptop and a Hokuyo UTM-30LX laser scanner (left). Real test environment(right), including difficult obstacle environments (upper right) and long-range corridor test (bottom right).

obstacles. In the long-range experiments, our robots navigate safely in corridors with obstacles and pedestrians. A video for real and simulated navigation experiments can be found at https://youtu.be/Eq4AjsFH_cU.

V. CONCLUSIONS

In this paper, we propose a model-free deep reinforcement learning algorithm to improve the performance of autonomous decision making in complex environments, which directly maps egocentric local occupancy maps to an agents steering commands in terms of target position and movement velocity. Our approach is mainly based on dueling double DQN with prioritized experience replay, and integrate curriculum learning techniques to further enhance our performance. Finally, both qualitative and quantitative results show that the map-based motion planner outperforms other related DRL-based methods in multiple indicators in simulation environments and is easy to be deployed to a robotic platform.

REFERENCES

- [1] M. Mohanan and A. Salgoankar, "A survey of robotic motion planning in dynamic environments," *Robotics and Autonomous Systems*, vol. 100, pp. 171–185, 2018.
- [2] W. Zhang, S. Wei, Y. Teng, J. Zhang, X. Wang, and Z. Yan, "Dynamic obstacle avoidance for unmanned underwater vehicles based on an improved velocity obstacle method," *Sensors*, vol. 17, no. 12, p. 2742, 2017.
- [3] D. Zhou, Z. Wang, S. Bandyopadhyay, and M. Schwager, "Fast, on-line collision avoidance for dynamic vehicles using buffered voronoi cells," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 1047–1054, 2017.
- [4] C. Rösmann, F. Hoffmann, and T. Bertram, "Integrated online trajectory planning and optimization in distinctive topologies," *Robotics and Autonomous Systems*, vol. 88, pp. 142–153, 2017.
- [5] G. Kahn, A. Villafior, B. Ding, P. Abbeel, and S. Levine, "Self-supervised deep reinforcement learning with generalized computation graphs for robot navigation," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 1–8.
- [6] A. Giusti, J. Guzzi, D. C. Cireşan, F.-L. He, J. P. Rodríguez, F. Fontana, M. Faessler, C. Forster, J. Schmidhuber, G. Di Caro *et al.*, "A machine learning approach to visual perception of forest trails for mobile robots," *IEEE Robotics and Automation Letters*, vol. 1, no. 2, pp. 661–667, 2015.
- [7] L. Tai, S. Li, and M. Liu, "A deep-network solution towards model-less obstacle avoidance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2016, pp. 2759–2764.
- [8] M. Pfeiffer, M. Schauble, J. Nieto, R. Siegwart, and C. Cadena, "From perception to decision: A data-driven approach to end-to-end motion planning for autonomous ground robots," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1527–1533.
- [9] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel *et al.*, "A general reinforcement learning algorithm that masters chess, shogi, and go through self-play," *Science*, vol. 362, no. 6419, pp. 1140–1144, 2018.
- [10] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.
- [11] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.
- [12] L. Xie, S. Wang, S. Rosa, A. Markham, and N. Trigoni, "Learning with training wheels: speeding up training with a simple controller for deep reinforcement learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6276–6283.
- [13] Y. F. Chen, M. Everett, M. Liu, and J. P. How, "Socially aware motion planning with deep reinforcement learning," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 1343–1350.
- [14] L. Tai, G. Paolo, and M. Liu, "Virtual-to-real deep reinforcement learning: Continuous control of mobile robots for mapless navigation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 31–36.
- [15] P. Long, T. Fanl, X. Liao, W. Liu, H. Zhang, and J. Pan, "Towards optimally decentralized multi-robot collision avoidance via deep reinforcement learning," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6252–6259.
- [16] K. Wu, M. Abolfazli Esfahani, S. Yuan, and H. Wang, "Learn to steer through deep reinforcement learning," *Sensors*, vol. 18, no. 11, p. 3650, 2018.
- [17] J. Zhang, J. T. Springenberg, J. Boedecker, and W. Burgard, "Deep reinforcement learning with successor features for navigation across similar environments," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2017, pp. 2371–2378.
- [18] G. Chen, G. Cui, Z. Jin, F. Wu, and X. Chen, "Accurate intrinsic and extrinsic calibration of rgb-d cameras with gp-based depth correction," *IEEE Sensors Journal*, vol. 19, no. 7, pp. 2685–2694, 2018.
- [19] Y. Liu, A. Xu, and Z. Chen, "Map-based deep imitation learning for obstacle avoidance," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 8644–8649.
- [20] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI)*, 2016.
- [21] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, and N. Freitas, "Dueling network architectures for deep reinforcement learning," in *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, 2016, pp. 1995–2003.
- [22] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016.
- [23] D. V. Lu, D. Hershberger, and W. D. Smart, "Layered costmaps for context-sensitive navigation," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2014, pp. 709–715.
- [24] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of the 26th Annual International Conference on Machine Learning (ICML)*. ACM, 2009, pp. 41–48.
- [25] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 3, 2004, pp. 2149–2154.
- [26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [27] I. Ulrich and J. Borenstein, "Vf/sup*: Local obstacle avoidance with look-ahead verification," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, vol. 3, 2000, pp. 2505–2511.