2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

Dual-Critic Deep Reinforcement Learning for Push-Grasping Synergy in Cluttered Environment

Jiakang Zhong, Yew Wee Wong, Jiong Jin, Yong Song, Xianfeng Yuan and Xiaoqi Chen

***Abstract*— Robotic push-grasping in densely cluttered environments presents significant challenges due to unbalanced synergy and redundancy between both actions, leading to decreased grasp efficiency. In this paper, a novel double-critic deep reinforcement learning framework is introduced to optimize the push-grasping synergy for robotic manipulation in such environments, aiming to significantly reduce pre-grasping redundancy. This framework incorporates two distinct Deep Q-learning critics: Critic I selects the best course of actions based on the current state derived from visual interpretation, whereas Critic II evaluates the success rate of the current state-action pairing. To further refine the push-grasping synergy, an active double-step learning mechanism is introduced to optimize the training reward function for the pushing action, thereby enhancing its effectiveness through increased intentionality. Simulations show that the proposed framework outperforms contemporary counterparts, notably in grasping success rate and action efficiency. Finally, the framework's generalization and adaptability are demonstrated by conducting real-world experiments using novel objects without the need of retraining.**

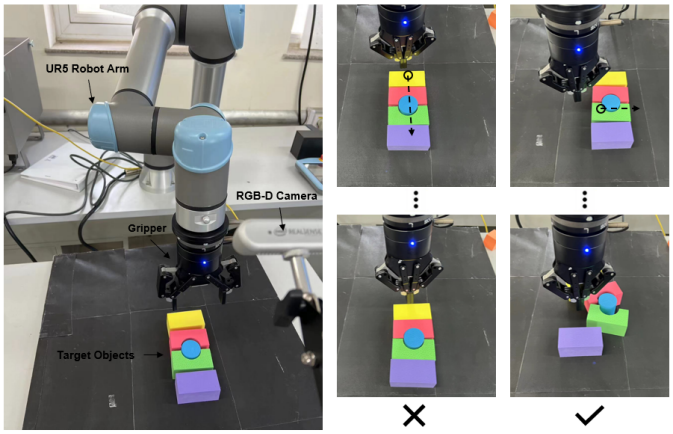

Fig. 1. Ineffective pre-grasping contributes minimally to successful grasping, while this framework amplifies the intentionality of pre-grasping actions for improved grasping performance.

I. INTRODUCTION

Grasping, recognized as a fundamental component of action, holds significant importance in the domain of robotic manipulation. Robotic grasping has emerged as a pivotal research area within robotics where numerous methodologies stemming from this research have found extensive applications in industrial manufacturing and everyday life scenarios [1]. The performance of object grasping has been commendable in non-cluttered environments, emphasizing contact points, dynamics, poses and object shapes [2]. In addition, data-driven methodologies have been employed to map grasp locations under these ideal environment [3], [4]. Nonetheless, the necessity for robots to operate in densely cluttered environments is inevitable and these complex tasks presents ongoing research challenges. In such environments, objects are closely placed, leading to mutual occlusions. This not only diminishes the maneuverable space but also critically impairs robotic grasping efficiency [5], [6].

Based on observations of human behavior, previous researchers introduced additional primitive actions to assist in grasping, such as pushing and dumping. The synergy between pushing and grasping has emerged as a solution for robotic grasping in densely cluttered environments [7]–[10]. These studies primarily categorize the challenges of grasping in clutter into two main categories: 1) goal-agnostic tasks (clutter cleaning) [7], [11], and 2) goal-oriented tasks (specific object search) [12], [13]. Although these works have showcased significant approaches, unresolved issues persist. For instance, there is redundancy and low efficiency in pushing actions, as shown in Fig. 1, and the supportive effect of pushing does not consistently enhance grasping [5].

In this paper, a dual-critic deep reinforcement learning framework for push-grasping synergy is introduced, as illustrated in Fig. 2. This framework possesses the capability to effectively balance the synergies between push and grasp actions, irrespective of whether the tasks are goal-agnostic or goal-oriented. The primary contributions of this study are summarized below:

- We propose a novel dual-critic deep reinforcement learning framework to enhance the accuracy and efficiency of robotic actions by optimizing the synergy between pushing and grasping in a cluttered environment.
- We propose a double steps learning scheme with multi-stage training to improve the intentionality of pushing and address the issue of pre-grasping redundancy.
- We evaluate the performance of the proposed framework both in a simulation environment and in real-world scenarios without the need of additional fine-tuning.

The remainder of the paper is organized as follows: Section II presents related works in the field of pushing-assisted grasp-

J. Zhong, Y. W. Wong, and J. Jin are with the School of Science, Computing and Engineering Technologies, Swinburne University of Technology, Hawthorn, VIC 3122, Australia (email: jiakangzhong@swin.edu.au; yewweewong@swin.edu.au; jiongjin@swin.edu.au).

Y. Song and X. Yuan are with the School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai 264209, China (e-mail: songyong@sdu.edu.cn; yuanxianfeng@sdu.edu.cn).

X. Chen is with Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Xingye Ave, Guangzhou 511442, Guangdong, China (email: xqc@scut.edu.cn).

ing. The dual-critic deep reinforcement learning framework for push-grasping synergy in densely cluttered environment is proposed in Section III. In addition, a double steps learning and multi-stage training algorithm for pushing-grasping synergy is developed. Section IV establishes the performance evaluation metrics and presents the discussion on the results from both simulations and real-world experiments. The paper concludes in Section V, where limitations and potential future works are highlighted.

II. Related work

Grasping techniques primarily categorizes into model analytical methods and data-driven methods [14]. The former constructs 3D models of specific objects to analyse physical properties of grasping process such as force, angle, and location [15], [16]. However, they are dependent to precise models and often struggle in unstructured environments or with unfamiliar objects. Conversely, data-driven methods leverage deep learning methods to train model-agnostic policies to map visual observation and grasping actions [17], [18]. Various datasets of labeled object have been developed, for the subsequent research [19], [20]. Rather than relying on datasets, self-supervised and reinforcement learning algorithms are introduced to derive rewards from environmental interactions [21], [22]. Nevertheless, the assumption of an ideal environment with isolated objects is often deemed essential in these methods. Consequently, learning to grasp in a cluttered environment is perceived as a challenging task.

In response to the challenges posed by a densely cluttered environment, the integration of pre-grasping actions, notably pushing, to facilitate in grasping has been recognized as a predominant strategy. Building on this approach, the Visual Pushing and Grasping (VPG) framework with a parallel PushNet and GraspNet structure, grounded in Deep Q-learning, is introduced in [7]. To enhance grasping efficiency, a composite robotic hand is developed in [8]. On the other hand, action shifting techniques are introduced in [23] to support grasping actions and mitigate the challenge of sparse rewards using a grasping-dependent reward function. In addition, a deep prediction model coupled with Monte Carlo Tree search is employed to predict and plan the sequence of pushing actions in [11], [24]. This prediction model is subsequently enhanced using a vision transformer by [25]. Furthermore, PolicyNMS is introduced, leveraging depth images to modulate grasping, aiming to enhance the success rate of grasping [26]. However, this method does not exploit visual features comprehensively. Manually designed rules, at times, fail under specific conditions, resulting in erroneous directives. It should be noted that such methods predominantly target goal-agnostic tasks, clearing all objects within and from the workspace.

In contrast to goal-agnostic tasks, goal-oriented push-grasping in densely cluttered environment poses greater difficulties, given the requirement to distinguish the sole target object from its environment. Various methods are advanced to address the issue [27], [28]. A critic-policy structure is employed to train a push-grasp strategy, allowing for the grasping of previously concealed target objects [12]. Moreover, based on VPG, the GraspNet is employed as a discriminator to guide the training of the PushNet in [13]. Although aforementioned studies make outstanding contributions on robotic push-grasping in cluttered environment, there are issues with redundancy in pushing actions leading to low action efficiency. To address this problem, a novel dual-critic push-grasping synergy framework is proposed in following Section III.

III. Dual-Critic Deep Reinforcement Learning for Push-Grasping

The process of robotic pushing and grasping manipulation can be formulated as Markov Decision Process (MDP) by the tuple: $< S, A, P, R, \gamma >$. Given a state $s_t \in S$ at time t, the agent (i.e. robot) selects and executes an action $a_t \in A$ according to a value-based policy. Subsequently, the transition function $P(s_{t+1} \mid s_t, a_t)$ produces a new state s_{t+1}, the reward function $R(s, a)$ generates a reward r based on s_t and a_t, and γ represents a discount factor used to sum over future returns across the entirety of the MDP.

In this study, visual images are captured by a stationary RGB-D camera, as illustrated in Fig. 1. These images are converted into RGB-D heightmaps via orthogonal projection, serving to represent the state s_t in each iteration. The action space is then defined as a tuple $(x, y, z, \theta_i, \psi)$, derived from the end effector's position on the gripper. Herein, (x, y, z) denotes the spatial coordinates of the gripper, θ_i designates the rotation angle around the z-axis, and ψ corresponds either the top-down grasping or the horizontal pushing action. As input to the system, the RGB-D heightmaps undergo 16 rotations around the z-axis, with each rotation amounting to $\pi/8$ radians. To standardize input and output actions, the heightmap's rotation angle mirrors the rotation angle θ_i of the gripper. Subsequently, the system generates dense pixel-wise maps of Q values, retaining the same resolution as the original heightmaps. Both grasping and pushing actions are characterized as predefined primitive behaviors based on each pixel's corresponding 3D location (x, y, z) and rotation θ_i.

A. Grasp-Oriented Synergy Framework: Critic I

Conventional push-grasping synergy methodologies employ off-policy Q-learning to train a greedy deterministic policy $\pi(s_t)$ that selects actions by maximizing the action-value function $Q_\pi(s_t, a_t)$, where two fully convolutional networks (FCNs) is modeled to compute the Q maps of both pushing and grasping [7]. Nevertheless, the objective of these two parallel network is to minimize the temporal difference error $\delta_t = \mid Q(s_t, a_t) - y_t \mid$ to a predetermined Q-value, which equates to the maximum Q-value from both networks. Such an approach could lead both networks to perpetually update towards higher-value pushing actions, thereby introducing issues of action redundancy. To address aforementioned issues, a novel function of temporal difference error is proposed with

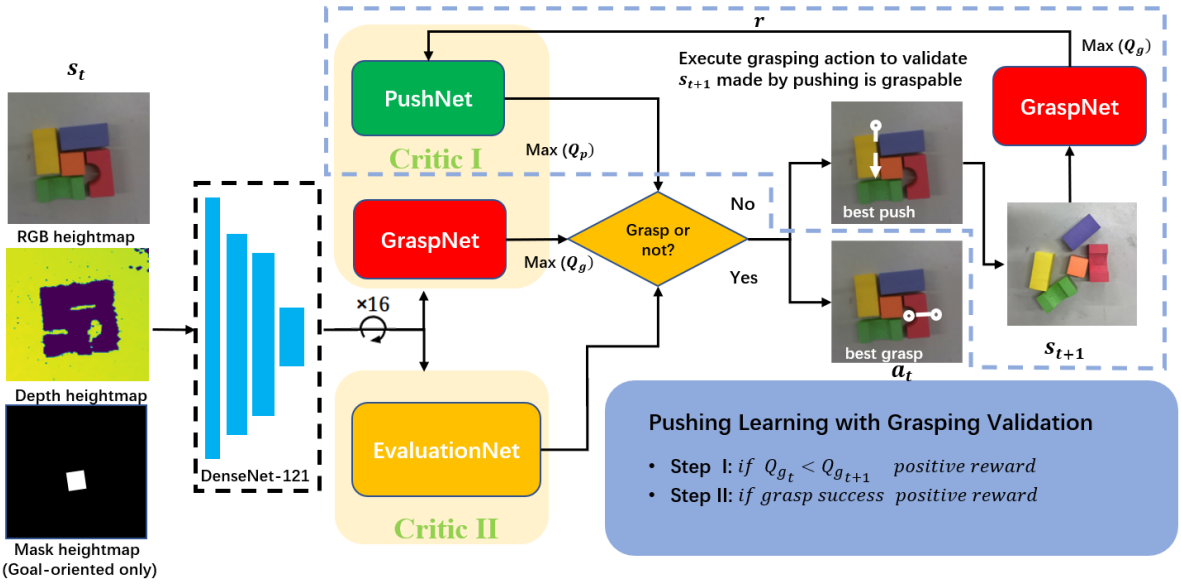

Fig. 2. Overview of framework: An RGB-D image is captured to produce heightmaps which are rotated in 16 orientations and processed via two DenseNet towers for feature extraction. Two distinct critic policies select actions based on the current state and assess the success rate of the state-action pairing, respectively. Pushing actions undergo double steps learning with grasping validation to reduce redundancy of pre-grasping actions.

y_t defined as:

$$y_t = R_{a_t} + \gamma\, Q(s_{t+1}, \arg\max_{a_{t+1}} Q_{grasp}(s_{t+1}, a_{t+1})) \quad (1)$$

where $\arg\max_{a_{t+1}} Q_{grasp}$ represents the maximum value grasping action at time $t+1$. The proposed temporal difference error assists two networks in updating to achieve a higher value grasping action. As a result, both networks evolve towards securing improved grasping actions, enhancing the grasp-oriented intentionality of pushing actions.

As depicted in Fig. 2, heightmaps are inputted into a pretrained feature extractor, DenseNet-121 [29], extracting visual features from both color and depth images. These extracted features serve as inputs for PushNet and GraspNet [7], constituting as Critic I of the proposed framework. PushNet and GraspNet, denoted as ϕ_p and ϕ_g respectively, compute the Q maps based on the current state s_t : $Q(s_t, a_t) = (\phi_p(s_t), \phi_g(s_t))$. Both the push and grasp Q maps are pixel-wise maps, correlating to the entire action space $(224 * 224 * 16)$.

The framework's reward scheme is defined as following: For grasping actions, the reward is contingent upon the outcome of the current action, a_t. A successful grasping attempt yields a reward of 1, while unsuccessful attempts result in a reward of 0. Regarding pushing actions, a double steps learning strategy is employed, with specifics elaborated in Section III-C.

B. Action Evaluation Critic Policy: Critic II

To address the constraints of PolicyNMS [26], the Action Evaluation Critic Policy, ϕ_c, designated as EvaluationNet (Critic II) in Fig. 2, is developed using a three-layer CNN as follows:

$$\phi_c(s_t, a_t) = \begin{cases} P_{success}, & \text{for grasping} \\ 1 - P_{success}, & \text{for pushing} \end{cases} \quad (2)$$

where $P_{success}$ denotes the probability of grasping success. Critic II processes information from the concatenation between the extracted vision features and the target positions acquired by the GraspNet to predict the probability, $P_{success}$, of grasping success. Therefore, this critic assesses the combination of current state, s_t, and the action, a_t, to determine the success rate of the current state-action pairing. From an output perspective, Critic II functions as a binary classifier, ingesting data from visual features combined with maximum value actions. It employs the grasping results, which are obtained during the training process through interactions with the environment, as labels.

In the evaluation process of Critic II, when the selected action is grasping, ϕ_c estimates $1 - P_{success}$ based on the action, a_t, with the highest Q-value and the current state, s_t. This estimation is then compared with a manually set constant, c. If and only if the $1 - P_{success}$ exceeds the constant, c, it indicates that pushing would be the more appropriate action at that juncture. Conversely, if the selected action is pushing, the comparison generated from $P_{success}$ and c, with a higher $P_{success}$ indicates that the objects is more amenable to grasping in its current state. In practice, in order to guide the system toward greater efficiency, the ten largest grasping action are fed into Critic II to ensure that the grasping action will not be easily replaced. This methodology significantly augments the success rate of grasping actions, avoids redundancy of pushing actions and enhances the overarching efficiency of the proposed framework.

C. Double Steps Learning and Multi-Stage Training

A distinct three-stage training scheme is developed for the training of the push-grasping framework:

Stage I: Grasping Training. This stage emphasizes the refinement of GraspNet's capabilities. Stage I initiates by

arranging 10 random objects and fixing the selected action to grasping. The objective is to amplify the Q-value estimation of GraspNet for the current state.

Stage II: Pushing Training with Double Steps Learning. During the second stage, a two-step learning process is employed to accentuate the training of pushing actions. Initially, the selected action is fixed to pushing, with the quantity of random objects increased to 20, introducing heightened complexity and obstruction. Subsequent to a pushing action, Step I leverages the variation in grasping Q-value between adjacent states to generate a reward. This is accomplished by comparing the grasping value Q_g prior to and post pushing (between states s_t and s_{t+1}), to determine the allocation of a push reward defined as follows:

$$R_p = \begin{cases} 0.5, & \text{if } Q_g \text{ is increased} \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

Subsequently, Step II carries out a grasping action and employs the grasping result to assess the efficacy of the preceding push action. The assessment outcomes directly dictate the criteria for the second push reward:

$$R_p = \begin{cases} 0.5, & \text{if validation grasping is successful} \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

Training in two steps based on distinct conditions offers rewards for PushNet. This provides precise direction for the updates of the PushNet and significantly enhances the proficiency of assisting in pushing actions as well as overall action efficiency. Conversely, the outcome of the grasp validation yields a reward for grasping concurrently. As a result, GraspNet undergoes updates in both Stage I and Stage II, increasing the training efficiency.

Stage III: Synergy Actions Training. The objective of the final stage is to balance the numerical outputs of the two networks, ensuring that the Q-value of one network does not predominate during manipulation. Balanced training of the two actions in synergy bolsters their collaborative capabilities.

D. Implementation Details

The proposed dual-critic deep reinforcement learning framework works on both goal-agnostic and goal-oriented tasks. In the case of goal-oriented grasping tasks, an additional goal mask heightmap is incorporated into the framework's inputs. The sole difference of the network's structure pertains to the input channel of the first layer of the feature extractor. The settings of goal-agnostic tasks is designed with the objective of clearing all objects from the workspace. Once cleared, items are either reintroduced randomly or positioned systematically. Conversely, goal-oriented task operates on a continuous cycle in which another goal is determined upon successful grasping of the target object. This process persists until the workspace no longer contains any viable target objects.

All models are trained on an NVIDIA RTX 3090 and an NVIDIA RTX 1660 SUPER. The feature extractor comprises of two parallel 121-layer DenseNets [29] that are pretrained on ImageNet [30]. These DenseNets extract concatenated features from three types of heightmaps, as shown in Fig.2. The PushNet and GraspNet of Critic I (ϕ_p and ϕ_g) shares the same network architecture, which comprises of two 1 × 1 convolutional layers FCNs [31] with ReLU [32] and batch normalization [33], followed by bilinear upsampling. Critic II applies three 3 × 3 convolutional layers CNNs with ReLU and full connection layers prediction. The loss functions of critics are Huber loss and BCEloss, respectively [34]. Prioritized experience replay is also employed to improve training efficiency [35]. An ϵ-greedy exploration strategy initializes at 0.5 and annealed to 0.1. The future discount factor γ is set as a constant value of 0.5. The optimizer is an Adam optimizer with a fixed learning rate on 10^{-4}, weight decay of 2^{-5}, and betas set to [0.9, 0.99]. For Critic II, constant c is set to 0.6. Finally, the training steps is set to 4000 with 1000 for Stage I and Stage II and 2000 for Stage III.

IV. Simulation and Real World Experiments

A series of experiments are conducted to evaluate the proposed methodology against prevailing state-of-the-art approaches. The dual-critic deep reinforcement learning framework introduced in this research is termed as Triple Double Push-Grasping (TDPG), and it leverages double critic policies paired with double steps for learning and updating. The objectives behind these experiments are threefold: 1) Verification of the enhanced efficiency of the PushNet, focusing on the reduction of redundancy and the enhancement of the grasping success rate 2) Assessment of the additional supervisory critic in the enhancement of the framework's overarching performance. 3) Establishment of the method's superior generalization capability when applied to real-world systems.

In this experiment, two type of experimental tasks are established: 1) **goal-agnostic grasping tasks** in which robots are to disperse dense clutter and clear all objects, and 2) **goal-oriented grasping tasks** in which robots are to remove obstructions in order to grasp a designated goal object. Performance evaluations are conducted using four robotic grasping metrics with results averaged over n iterations:

- **Completion**: This metric represents task completion rate over n test runs. The task is deemed completed if the policy manages to lift objects without recording 10 successive unsuccessful attempts in a single trial.
- **Grasp success rate**: This metric calculates the percentage of successful grasps per completion.
- **Action efficiency**: This metric is the ratio of the number of objects to the total actions per completion, assessing the succinctness with which the task is completed.
- **Motion number**: This metric gauges the average number of motions per completion, with a smaller value indicating superior performance.

To assess the effectiveness of TDPG, its performance is benchmarked against three distinct contemporary counterparts:

- **VPG** is a push-grasping synergy method, which generates pushing and grasping actions from visual observation [7].
- **GI** is a goal-oriented variant of VPG, which explores the workspace when targets are invisible and subsequently grasps the targets once they are detected [12].
- **EPG** is an extended version of VPG, which introduces an alternating training process paired with an adversarial training scheme between two actions to optimize grasping efficacy [13].

A. Simulation Experiment Results

The simulation environment setup utilizes a UR5 service robot equipped with an RG-2 gripper in CoppeliaSim [36]. The robot motion planning is CoppeliaSim's internal inverse kinematics module. The simulation runs a total of n=30 times. The objects employed in these simulations encompass 9 distinct 3D blocks with randomly assigned shapes and colors. RGB-D images, with a resolution of 640×480, are generated from the camera using OpenGL without incorporating a noise model.

Goal-Agnostic tasks: Two testing arrangements are conducted: random sets and challenging cases. In the random arrangement, experiments are established at four levels, with 15, 20, 25, and 30 objects randomly positioned in the workspace. The shapes and colors of these objects are selected randomly during each run as illustrated in Fig. 3. To demonstrate the superiority of different components of TDPG, it is divided into two groups in this section: TDPG-base (only Critic I) and TDPG+Critic (Critic I + Critic II). The results in Fig. 5 show that both TDPGs significantly surpass its counterparts in all metrics. Meanwhile, Critic II significantly improves TDPG in terms of grasp success and action efficiency as the number of objects increases. In the ablation experiment, Critic II, serving as an auxiliary component, is added to the VPG, leading to improvements in almost all metrics, especially when the number of objects is substantial. As the number of objects increased, both our method and the control groups face some challenges. However, TDPG consistently maintain a lead of 5 percent in success rate and 10 percent in action efficiency. In various levels of environmental complexity, TDPG exhibits minimal variance in the completion metric. This consistency is attributed to the fact that meaningful pushing actions aided grasping, rather than ejecting objects from the workspace.

The challenging arrangement consisted of 9 designed cases across three levels of complexity, as depicted in Fig. 4. Given the adversarial clutter, pushing becomes imperative for effective grasping since objects are positioned adjacently and encircle one another. Consequently, this setup presents heightened challenges for successful grasping. The results, as shown in Fig. 5, highlight that both TDPGs achieve more pronounced improvements relative to the random arrangement, especially in terms of grasping success rate (averaging a 20 percent lead over the comparison group) and action efficiency (averaging a 10 percent lead over the comparison group). This can be attributed to the fact that, unlike VPG and EPG, TDPG's pushing actions are more targeted and can maximize the advantage of pushing in scenarios with occlusions. Conversely, the pushing actions of VPG and EPG sometimes choose to push objects out of the workspace or push for the sake of the action itself.

Fig. 3. Random arrangement of 15, 20, 25 and 30 objects.

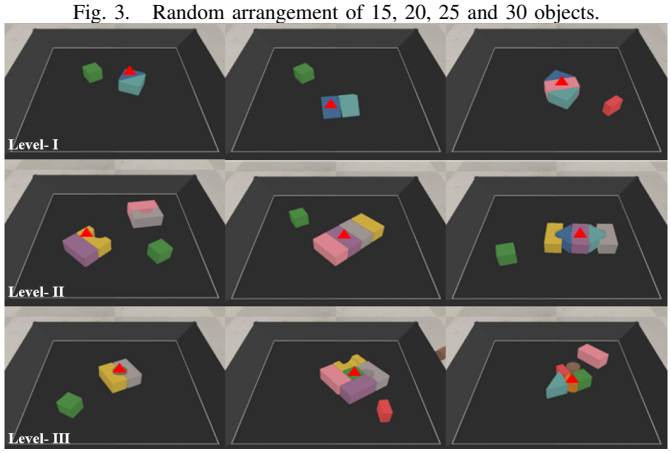

Fig. 4. Level-I cases: One-sided occlusion between objects. Level-II cases: Half occlusion between objects. Level-III cases: Full occlusion between objects. Challenge cases for goal-oriented tasks are consistent across the three levels, with target objects denoted by a red triangle.

Goal-oriented tasks: For the goal-oriented tasks, a target perception is not the primary focus. Consequently, a projection of 3D models in simulation is employed to supply goal masks for all three experimental groups. Experiments were organized into both random and challenging case sets. For the random arrangement, the number of objects is set at 30. The target, in each trial, is chosen randomly from the entire set of objects. On the other hand, the challenging arrangement designates a specific object as the target for each case, as illustrated in Fig. 4. The results of both experimental setups are summarized in Table I. The results indicate that, in challenge cases, TDPG shows significant improvements in terms of both the grasping success rate and the number of actions undertaken. This improved performance can be attributed to the higher efficiency of TDPG's pushing actions when confronted with a distinctly defined target object. The increased grasping success rate demonstrates that pushing actions create more space for successful grasping. Nevertheless, in the randomized setup, the performance metrics of TDPG closely mirror those of the two control groups. Specifically, both the grasping success rate and the number of actions executed by TDPG trail slightly behind those presented by EPG, with deficits of 1.3 percent and 0.16 actions, respectively.

B. Real World Experiment Results

In real-world scenarios, testing is undertaken using an experimental setup that included a UR5 service robot equipped with a ROBOTIQ-85 gripper, as depicted in Fig. 1. The

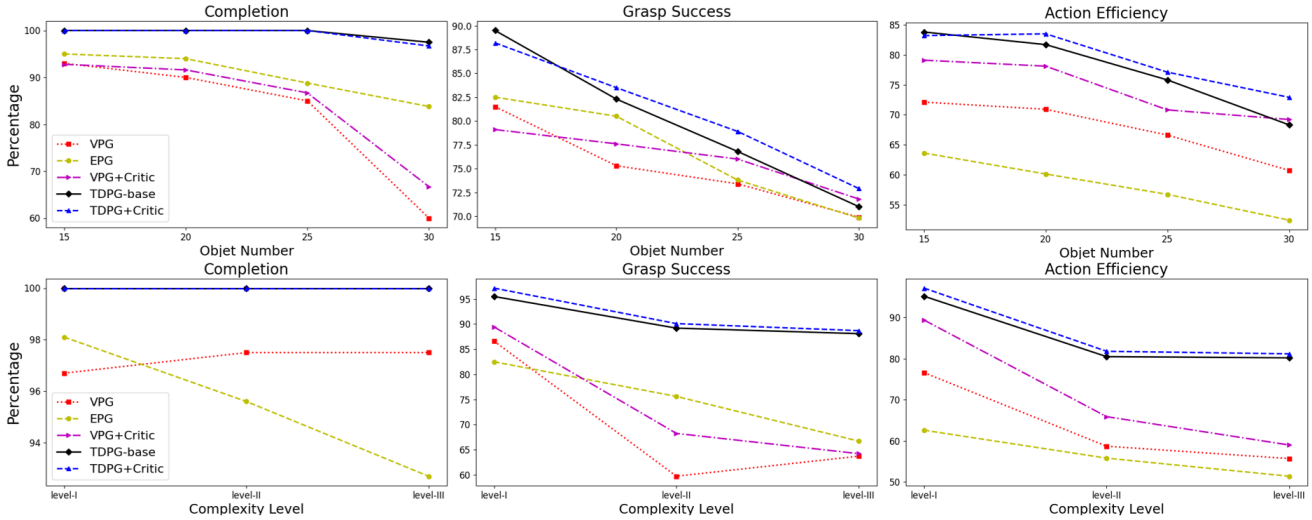

Fig. 5. TDPG comparison results for completion, grasp success, and action efficiency: random arrangement (Top) vs. challenge cases (Bottom).

TABLE I
SIMULATION RESULTS IN GOAL-ORIENTED GRASPING

Method	**Completion**		**Grasp Success**		**Motion Number**	
Arrangement	Random	Cases	Random	Cases	Random	Cases
GI	56.7	95.0	54.6	70.4	5.37	4.33
EPG	59.5	97.7	**66.6**	55.4	**3.00**	5.37
TDPG	**61.6**	**99.0**	65.3	**83.5**	3.16	**2.57**

most proficient version of TDPG is selected for evaluation in this segment. To draw a parallel with VPG [13], toy blocks are employed for case arrangement. Additionally, a diverse assortment of standard office objects is incorporated to assess the model's generalization capabilities when introduced to unfamiliar items. In terms of perception data, RGB-D images with a resolution of 1280 × 720 are sourced from an Intel RealSense D435. This camera is securely stationed on a fixed tripod, strategically positioned to provide an aerial view of the tabletop configuration.

Fig. 6. Object cases (up) and novel objects (down) in real world experiment.

TABLE II
REAL WORLD EXPERIMENT RESULTS IN GOAL-AGNOSTIC GRASPING

Method	**Grasp Success**		**Action Efficiency**	
Arrangement	Cases	Novel Objects	Cases	Novel Objects
VPG	70.8	78.6	57.5	65.8
TDPG	82.3	83.5	70.1	70.4

In the real-world setting, this study did not utilize specialized segmentation models specifically tailored for object grasping, hence, the targeted grasping task was not conducted. Consequently, the real-world experiments concentrate on goal-agnostic grasping tasks and running number is n=10 times. The primary objective is to emphasize the generalization capabilities of the proposed method and the framework is assessed using models trained in a simulated environment. Upon completion of the clearance task on ten separate occasions, findings are documented in Table II. In this experiment, TDPG is combined with both two critics. Table II indicates that TDPG consistently outperforms VPG in real-world scenarios, particularly in metrics related to grasping success rate and action efficiency. Moreover, when juxtaposed with the original findings of the VPG in which the model is trained in a real-world environment, the current experiments and subsequent results illustrate enhanced adaptability.

V. CONCLUSIONS

A novel dual-critic deep reinforcement learning framework is introduced for enhanced push-grasping tasks in densely cluttered environment. This framework refines the traditional Q-learning method to optimize the synergy between the push-grasping actions. Specifically, it trains the pushing policy to pursue a more exercisable statement for grasping and leverages grasping feedback to enhance pushing quality. Empirical evidence shows the proposed method's superiority over other synergy techniques, significantly improving push efficiency by minimizing redundant attempts and promoting optimal grasping conditions. However, an inherent limitation is that optimizing pushes for better grasping might not fully capture intent of grasping actions, potentially limiting imaginative possibilities. Future directions include integrating more advanced learning techniques to further elevate synergy capabilities and transitioning the proposed framework to real-world applications.

REFERENCES

[1] X. Fu, Y. Liu, and Z. Wang, "Active learning-based grasp for accurate industrial manipulation," *IEEE Transactions on Automation Science and Engineering*, vol. 16, no. 4, pp. 1610–1618, 2019.

[2] G. Du, K. Wang, S. Lian, and K. Zhao, "Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review," *Artificial Intelligence Review*, vol. 54, no. 3, pp. 1677–1734, 2021.

[3] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2786–2793.

[4] I. Clavera, D. Held, and P. Abbeel, "Policy transfer via modularity and reward guiding," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 1537–1544.

[5] M. Q. Mohammed, L. C. Kwek, S. C. Chua, A. Al-Dhaqm, S. Nahavandi, T. A. E. Eisa, M. F. Miskon, M. N. Al-Mhiqani, A. Ali, M. Abaker *et al.*, "Review of learning-based robotic manipulation in cluttered environments," *Sensors*, vol. 22, no. 20, p. 7938, 2022.

[6] Z. Pan, A. Zeng, Y. Li, J. Yu, and K. Hauser, "Algorithms and systems for manipulating multiple objects," *IEEE Transactions on Robotics*, 2022.

[7] A. Zeng, S. Song, S. Welker, J. Lee, A. Rodriguez, and T. Funkhouser, "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 4238–4245.

[8] Y. Deng, X. Guo, Y. Wei, K. Lu, B. Fang, D. Guo, H. Liu, and F. Sun, "Deep reinforcement learning for robotic pushing and picking in cluttered environment," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 619–626.

[9] M. Kiatos, I. Sarantopoulos, L. Koutras, S. Malassiotis, and Z. Doulgeri, "Learning push-grasping in dense clutter," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 8783–8790, 2022.

[10] B. Tang, M. Corsaro, G. Konidaris, S. Nikolaidis, and S. Tellex, "Learning collaborative pushing and grasping policies in dense clutter," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6177–6184.

[11] B. Huang, S. D. Han, A. Boularias, and J. Yu, "Dipn: Deep interaction prediction network with application to clutter removal," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 4694–4701.

[12] Y. Yang, H. Liang, and C. Choi, "A deep learning approach to grasping the invisible," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2232–2239, 2020.

[13] K. Xu, H. Yu, Q. Lai, Y. Wang, and R. Xiong, "Efficient learning of goal-oriented push-grasping synergy in clutter," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 6337–6344, 2021.

[14] J. Bohg, A. Morales, T. Asfour, and D. Kragic, "Data-driven grasp synthesis—a survey," *IEEE Transactions on Robotics*, vol. 30, no. 2, pp. 289–309, 2013.

[15] A. Sahbani, S. El-Khoury, and P. Bidaud, "An overview of 3d object grasp synthesis algorithms," *Robotics and Autonomous Systems*, vol. 60, no. 3, pp. 326–336, 2012.

[16] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3629–3635.

[17] C. Goldfeder and P. K. Allen, "Data-driven grasping," *Autonomous Robots*, vol. 31, no. 1, pp. 1–20, 2011.

[18] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 1957–1964.

[19] K. Lai, L. Bo, X. Ren, and D. Fox, "A large-scale hierarchical multi-view rgb-d object dataset," in *2011 IEEE International Conference on Robotics and Automation*. IEEE, 2011, pp. 1817–1824.

[20] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar, "Yale-cmu-berkeley dataset for robotic manipulation research," *The International Journal of Robotics Research*, vol. 36, no. 3, pp. 261–268, 2017.

[21] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 6284–6291.

[22] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke *et al.*, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Conference on Robot Learning*. PMLR, 2018, pp. 651–673.

[23] L. Berscheid, P. Meißner, and T. Kröger, "Robot learning of shifting objects for grasping in cluttered environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 612–618.

[24] B. Huang, T. Guo, A. Boularias, and J. Yu, "Interleaving monte carlo tree search and self-supervised learning for object retrieval in clutter," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 625–632.

[25] S. Yu, D.-H. Zhai, and Y. Xia, "A novel robotic pushing and grasping method based on vision transformer and convolution," *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[26] Y. Yang, Z. Ni, M. Gao, J. Zhang, and D. Tao, "Collaborative pushing and grasping of tightly stacked objects via deep reinforcement learning," *IEEE/CAA Journal of Automatica Sinica*, vol. 9, no. 1, pp. 135–145, 2021.

[27] T. Novkovic, R. Pautrat, F. Furrer, M. Breyer, R. Siegwart, and J. Nieto, "Object finding in cluttered scenes using interactive perception," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 8338–8344.

[28] E. Li, H. Feng, S. Zhang, and Y. Fu, "Learning target-oriented push-grasping synergy in clutter with action space decoupling," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 11 966–11 973, 2022.

[29] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 4700–4708.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[31] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3431–3440.

[32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[33] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on Machine Learning*. pmlr, 2015, pp. 448–456.

[34] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009, vol. 2.

[35] T. Schaul, J. Quan, I. Antonoglou, and D. Silver, "Prioritized experience replay," *arXiv preprint arXiv:1511.05952*, 2015.