Human-in-The-Loop Sim-to-Real Transfer Policy for Robotic Assembly via Reinforcement Learning

REN, Sirui, ZENG, Chao, LI, Zhiyi, YANG, Chenguang and WANG, Ning

Human-in-the-loop sim-to-real transfer policy for robotic assembly via reinforcement learning

1st Sirui Ren
College of Automation Science and Engineering
South China University of Technology
Guangzhou, China
202321016984@mail.scut.edu.cn

2nd Chao Zeng
Department of Computer Science
University of Liverpool
Liverpool, L69 3BX, U.K.
chaozeng@ieee.org

3rd Zhiyi Li
College of Automation Science and Engineering
South China University of Technology
Guangzhou, China
au_lzy@mail.scut.edu.cn

4th Chenguang Yang
Department of Computer Science
University of Liverpool
Liverpool, L69 3BX, U.K.
cyang@ieee.org

5thNing Wang *
School of Computing and Digital Technologies
Sheffield Hallam University
Sheffield, S1 2NU, United Kingdom
ning.wang@shu.ac.uk
*Corresponding author

Abstract—In complex robotic tasks, reinforcement learning (RL) algorithms have garnered significant attention for their ability to dynamically adapt to environmental changes and optimize control policies. However, challenges such as sparse rewards, poor generalization capability, low sample efficiency, and the sim-to-real transfer gap continue to hinder the widespread industrial application of RL. To address these challenges, we propose a novel reinforcement learning framework that integrates Deep Deterministic Policy Gradient (DDPG), Hindsight Experience Replay (HER), and Behavior Cloning (BC) to efficiently solve robotic assembly tasks in sparse reward environments. Hindsight Experience Replay (HER) addresses sparse rewards by relabeling failed experiences as successes, enabling broader exploration of target states and faster policy learning. Combined with Behavior Cloning (BC), which uses human demonstrations to reduce exploration needs, the proposed approach effectively enhances learning efficiency and generalization in complex robotic tasks. To validate the proposed algorithm, we implemented a benchmark robotic assembly environment in the MuJoCo simulator. Experimental results show that the proposed framework significantly outperforms baseline methods in key metrics, including training speed, task success rate, and assembly efficiency. Furthermore, we developed and validated an online human-in-the-loop correction-based sim-to-real transfer strategy. By leveraging a small amount of human correction data, this strategy effectively bridges the sim-to-real gap, enabling the model to exhibit robust performance and strong generalization in real-world robotic assembly tasks.

Keywords—Reinforcement Learning, Robotic Assembly, Human-in-the-Loop, Sim-to-Real Transfer

I. INTRODUCTION

Robotic assembly is a cornerstone of intelligent manufacturing, yet traditional programming methods are inefficient and lack adaptability, motivating the adoption of reinforcement learning (RL). In practical applications, agents often need to complete multiple steps to accomplish an assembly task. However, during these exploration processes, the environment rarely provides feedback (rewards). Agents only receive reward signals upon completing a full control episode, and these signals often carry limited information, making it challenging to provide effective guidance for the agent. Additionally, traditional reinforcement learning algorithms exhibit limitations in generalization. For instance, in our assembly task, if the position of the gear changes, the agent would need to retrain a new policy, which is evidently impractical for real-world industrial applications. The Hindsight Experience Replay (HER) algorithm has been shown to effectively address a range of similar but not identical tasks[1]. Therefore, in our framework, we integrate the Deep Deterministic Policy Gradient (DDPG) algorithm[2] with HER to transform failure experiences into goal-directed experiences, effectively mitigating the sparse reward problem and enhancing generalization through explicit goal augmentation.

Moreover, assembly tasks typically involve an extremely large exploration space, requiring robots to perform extensive trial-and-error to gain effective experience. This purely exploration-based learning approach is often inefficient and costly[3]. To address this issue, our method incorporates human demonstration data to reduce exploration time and costs. By combining HER and Behavior Cloning (BC), the proposed framework leverages the strengths of both methods, enabling the agent to learn not only from failures but also from successful demonstrations. Given the safety concerns and other challenges associated with collecting data in real-world environments, we prefer to transfer policies trained in simulation to the real world. This inevitably raises the need to address the sim-to-real gap, which includes discrepancies such as dynamic modeling errors, sensor noise, and differences in friction models between

simulation and reality. As previous studies have shown, commonly used sim-to-real transfer methods in deep reinforcement learning (DRL) include Domain Adaptation, Domain Randomization, Inverse Dynamics Models, Progressive Neural Networks (PNNs), and Meta-Reinforcement Learning[4]. To achieve efficient and practical robotic assembly in the real world, this paper proposes a novel framework capable of transferring strategies learned in simulation to reality. The framework is illustrated in Figure 1. The main contributions of this paper are as follows:

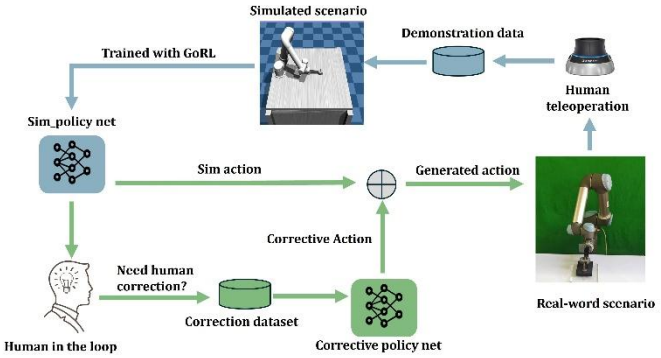

Fig. 1. The overview of our method

1) We redesigned the update strategy formula of DDPG to successfully integrate behavior cloning (BC) with deep deterministic policy gradient (DDPG). This improvement enables the agent to quickly learn key skills from expert demonstration data, significantly accelerating the initial exploration process. By incorporating the behavior cloning term, the agent avoids the inefficiency of random exploration, thereby greatly improving learning efficiency and laying a solid foundation for subsequent reinforcement learning.

2) Building on the improved DDPG, we further introduced Hindsight Experience Replay (HER), a mechanism that transforms failed experiences into successful experiences, to construct a unified framework. This framework effectively alleviates the sparse reward problem by converting failed experiences into goal-oriented successful experiences. Moreover, through explicit goal formulation, the agent is able to better generalize to different task instances. Experimental results demonstrate that this framework significantly outperforms baseline methods under sparse reward conditions and can efficiently complete complex assembly tasks.

3) We proposed a human-in-the-loop online correction method that successfully transfers policies trained in simulation to the real world through a small amount of human intervention data. This method does not require domain-specific knowledge and can implicitly bridge the simulation-to-reality gap. Experiments have shown that this method significantly improves the success rate of assembly tasks in the real environment, providing an efficient and low-cost solution for simulation-to-reality transfer and laying a foundation for future applications in more complex tasks and real-world environments.

II. Related Work

In the field of robotic assembly, recent references have primarily focused on enhancing the flexibility, precision, and intelligence of assembly systems. For instance, the reference [5] employed augmented reality (AR) markers during the human demonstration phase to learn the assembly process. In the robot execution phase, point cloud data and geometric constraints were utilized for object detection and motion planning, resulting in automated assembly. The integration of AR-assisted digital twin (DT) technology [6] enabled operators to intuitively plan tasks at physical workstations, allowing robots to execute assembly actions based on operator guidance and real-time perception. Additionally, the references[7][8] demonstrated successful robotic assembly tasks that incorporated human demonstrations, underscoring the critical importance of human demonstration experience in facilitating complex robotic operations. As artificial intelligence continues to advance, there is an increasing emphasis on reducing dependency on human labor. Robots are expected to autonomously and rapidly adapt to environmental changes when tasks vary, minimizing the time required for adjustments and optimizations. To achieve this, reinforcement learning methods that integrate human expertise have been developed. These approaches leverage human knowledge to accelerate learning processes, improve efficiency, and enhance task success rates.

References such as [9][10] combine reinforcement learning with demonstration data by initializing policies from human actions and modeling trajectories with Dynamic Movement Primitives (DMPs), enabling robots to replicate demonstrations and refine skills through environmental feedback. Behavior Cloning (BC), which maps states to actions via supervised learning, also serves as an effective means of integrating human knowledge. The DAPG framework (Demo-Augmented Policy Gradient) [11] leverages limited demonstrations with policy gradients to acquire dexterous skills, highlighting the potential of demonstration data in complex tasks. Similarly, DDPGfD (Deep Deterministic Policy Gradient from Demonstrations) [12] incorporates demonstration trajectories into the replay buffer with higher priority, combining them with interaction data during updates. While DDPGfD balances both data sources, our method places greater emphasis on demonstrations to guide early training, thereby improving efficiency and reducing reliance on environment interactions.

The combination of human demonstration data can improve learning efficiency and policy quality in the early training stage, but exploration remains difficult under sparse rewards. To address this, [13] proposed Meta Reward Learning (MeRL), which constructs auxiliary reward functions from diverse successful trajectories to provide finer-grained feedback, while [14] introduced Reward-Free Reinforcement Learning (RFRL), where agents first collect trajectories without predefined rewards and later compute near-optimal policies for specified reward functions. In our framework, we employ Hindsight Experience Replay (HER) [1], which provides a more direct implementation while effectively alleviating sparse rewards and enhancing policy adaptability.

In recent years, bridging the simulation-to-reality (sim-to-real) gap has received increasing attention to enable practical

applications of reinforcement learning. Domain adaptation methods [15][16] map simulation and real environments into a shared latent space, while domain randomization techniques [17][18] facilitate safe sim-to-real transfer. In our framework, an online human-in-the-loop correction strategy is employed to efficiently bridge the sim-to-real gap. During both training and deployment, real-time human corrections are collected to form a targeted dataset for policy optimization. Compared with traditional approaches, this strategy better handles dynamically changing tasks, reduces reliance on precise environment modeling, and improves policy robustness and generalization in real-world settings.

III. METHODS

A. Learning Framework

We adopt the Deep Deterministic Policy Gradient[2] integrated with Hindsight Experience Replay (HER)[1] to address sparse reward challenges in robotic control. DDPG, an actor–critic algorithm suitable for high-dimensional continuous tasks, updates the policy network (actor) by maximizing the critic's Q-value:

$$\nabla_{\theta^\pi} J(\theta^\pi) = \mathrm{E}_{s_t \sim D}\left[\nabla_{\theta^\pi} Q\left(s_t, \pi(s_t \mid \theta^\pi)\right)\right] \quad (1)$$

while the critic minimizes the temporal-difference (TD) error:

$$L_{TD} = E[(Q(s_t, a_t) - (r_t + \gamma Q_{target}(s_{t+1}, \pi_{target}(s_{t+1}))))^2] \quad (2)$$

To accelerate convergence, behavior cloning loss is added to the actor updates:

$$LBC = \sum_{i=1}^{M} \| \pi(s_i|\theta_\pi) - a_i \|^2 \quad (3)$$

HER further enhances sample efficiency by redefining goals in failed trajectories. If a trajectory fails to reach the target g but ends at g', HER relabels g' as the new goal, recalculates the reward, and augments training data. This mechanism transforms failures into informative experiences, mitigating the sparsity of reward signals.

B. Simulation Evironment

A MuJoCo-based environment is developed to mimic real-world robot assembly, enabling training in Cartesian space with proprioceptive feedback. The state and action spaces are defined as:

$$S_t = [x_t, y_t, z_t], A_t = [\delta x, \delta y, \delta z] \quad (4)$$

A sparse reward guides the agent:

$$r_t = \begin{cases} 0 & \text{if } |s_x - g_x| + |s_y - g_y| + |s_z - g_z| \le \epsilon, \\ -1 & \text{otherwise} \end{cases} \quad (5)$$

where (s_x, s_y, s_z) and (g_x, g_y, g_z) denote the end-effector and goal positions, respectively, and $\varepsilon = 0.02$ is the success threshold.

Through the combination of DDPG, HER, and the designed environment, the agent efficiently learns insertion behaviors under sparse rewards and improves policy generalization for real-world deployment.

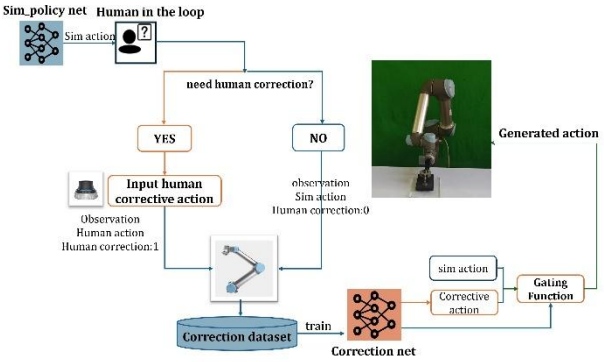

Fig. 2. The pipline of sim-to-real transfer

C. Sim-to-real transfer

Due to the inevitable physical differences between simulation environments and real-world environments—such as discrepancies in dynamics, sensor noise, and variations in friction models—strategies trained solely in simulation often struggle to directly transfer to real-world applications. Traditional sim-to-real approaches often rely on explicit domain knowledge and specialized expertise, such as system identification or domain randomization. However, in many scenarios, such knowledge is either unavailable in advance or difficult to acquire.

To address these limitations, this paper adopts a human-in-the-loop transfer approach. By allowing humans to directly observe and assist robots in executing strategies in the real world, this method implicitly bridges the sim-to-real gap without requiring explicit identification or modeling of the discrepancies. Unlike traditional approaches that rely on predefined models or extensive calibration, the human-in-the-loop method leverages human intuition and adaptability to dynamically compensate for the mismatch between simulation and reality, thereby facilitating a more efficient and flexible transfer process. As illustrated in the Figure 2, our framework initially deploys the policy trained in the simulation environment directly to the real world. A human operator oversees real-time policy execution and intervenes via teleoperation when the robotic arm commits errors or enters unsafe states. During these interventions, an online correction dataset is constructed, which includes the human-provided corrected actions, the robotic arm states before and after the corrections, and the corresponding correction signals.

The neural network is trained with inputs consisting of the simulation policy's action output (sim_action) and the current state observations of the robotic arm. Its outputs include the corrected actions and the probability that human intervention is required. A gating function integrates the corrected actions with the sim_action for deployment to the real robotic arm.

Fig. 3. The gears and shafts of the assembly task

The central idea of the correction network lies in leveraging learning to rectify deficiencies in the existing model rather than learning from scratch. This approach significantly reduces the amount of human correction data required for successful transfer while demonstrating strong generalization capabilities.

IV. Experiments

The experiments conducted in this paper primarily investigate the following questions:

- Can the proposed framework successfully and efficiently complete the assembly task?
- How effective are the human demonstration data and the concept of HER in the assembly task?
- Can the policy trained in the simulation environment be successfully transferred to the real-world environment?

A. Simulation Experiment

We constructed a simulation environment in MuJoCo that closely resembles the physical environment to facilitate subsequent transfer tasks. The simulation environment is shown in the figure below. Both the training and testing stages were conducted on a PC equipped with an Intel 13th Gen Core i9-13900HX CPU @ 2.20 GHz, featuring 24 cores and 32 logical processors, and an NVIDIA GeForce RTX 4060 Laptop GPU. **Robot**: A 6-DOF UR5 robotic arm with a two-finger gripper was used, which features a drag view function that facilitates the collection of human demonstration data. **Object**: As shown in the Figure 3 the assembled gear consists of two parts: the gear and the shaft. The outer diameter of the gear is 30mm, and the inner diameter is 10mm. The diameter of the shaft is 8mm.

Our goal is to train an RL policy that generates actions from the current robotic arm state to autonomously complete assembly tasks. In experiments, the arm's initial pose and the gear position are fixed. The algorithm is trained and evaluated every five episodes, with rewards recorded for performance tracking. To assess component effectiveness, ablation studies are conducted on Hindsight Experience Replay (HER) and Behavior Cloning (BC). Each run begins with 100 random seed steps, after which 128 samples are drawn per step for updates. Actor and critic networks are optimized with learning rates of 1×10^{-3}.

As shown in Figure 4 our proposed algorithm (red, DDPG+HER+BC) achieves higher initial rewards and converges fastest, stabilizing around 100 episodes near -5. The use of demonstration data accelerates early exploration and reduces stochasticity, ensuring faster convergence. Pure DDPG (orange) fails to improve, while DDPG+HER (blue) learns effectively, converging near -5 after about 200 episodes. Using

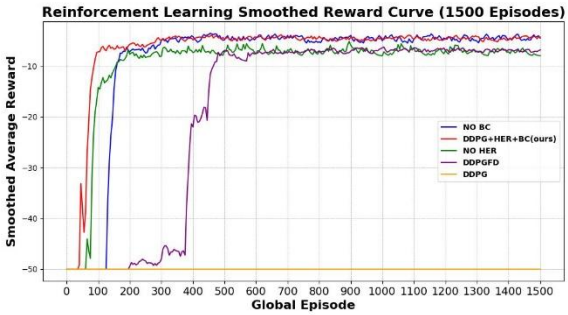

Fig. 4. Training reward curves for the assembly task under different reinforcement learning policies using DDPG

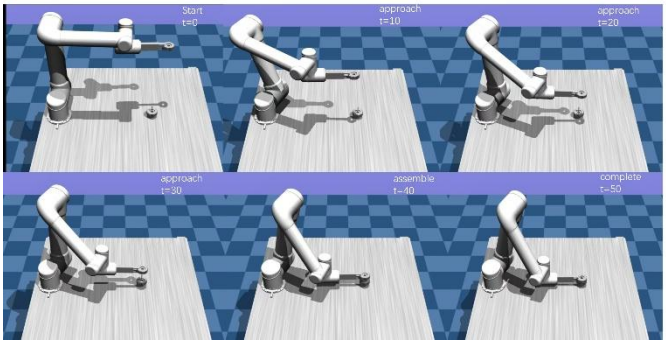

Fig. 5. Snapshots of the robotic assembly process

only demonstrations without HER (green) improves learning but remains less effective than HER-based methods. DDPGfD (purple) benefits from demonstrations only after extended exploration, converging much later.

Snapshots of the assembly process are shown in Figure 5, confirming that the proposed algorithm successfully completes the assembly task. In conclusion, HER is critical for handling sparse rewards, while incorporating human demonstrations further enhances early-stage efficiency and stability.

B. Comparison Study

To assess the generalization capability and robustness of different training strategies, we retained the neural networks obtained at the 900th episode and evaluated each on 100 randomized assembly tasks in the same simulation environment. For each task, the agent was required to complete the assembly within 200 steps. The success rate and the average number of steps were recorded and summarized in Figure 6 and Table Ⅰ.

As shown in Figure 6, the DDPG+HER+BC method achieves the highest success rate with low variance, indicating superior robustness and generalization. DDPG+HER also performs well, suggesting that HER is effective in addressing sparse reward issues. In contrast, DDPGfD yields substantially lower success rates and higher variance, reflecting weaker adaptability under randomized conditions.

In terms of task efficiency, the average number of steps required further highlights the advantage of the proposed method. DDPG+HER+BC requires the fewest steps with the most stable distribution, while DDPG+HER shows slightly slower performance but still outperforms DDPGfD. The latter

requires the greatest number of steps with a wider distribution, indicating reduced efficiency and slower policy learning.

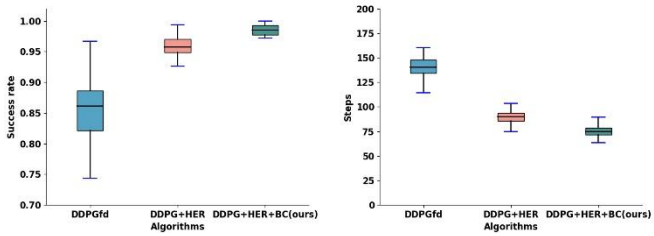

Fig. 6. Average success rate and average number of steps for different methods

TABLE I. PERFORMANCE COMPARISON OF DIFFERENT REINFORCEMENT LEARNING POLICIES

Policy	Training Time(episodes)	Average Reward	Success Rate(%)	Average Steps
DDPG+HER+BC(ours)	120	-5.3	97	75
DDPG+HER	200	-5.5	96	90
DDPG+BC	200	-8.7	94	110
DDPGfD	600	-8.5	65	140
DDPG (baseline)	1500	-50	0	-

Figure 7 further illustrates the trajectory distributions. DDPGfD produces disordered initial trajectories with evident instability and delayed convergence. DDPG+HER improves trajectory stability and convergence speed, while the proposed DDPG+HER+BC method demonstrates the most favorable performance, with rapid convergence in all spatial dimensions and minimal oscillation.

C. Sim-to-real transfer

We collected a total of 60 trajectories, including 309 corrections, to train a correction network. The network is composed of two main components: a feature encoding network and a task module. The feature encoding network consists of two hidden layers, each utilizing ReLU activation functions. The input state has a dimensionality of 6, consisting of both the 3-dimensional goal and the 3-dimensional current state of the robot's end-effector. The correction action output has a dimensionality of 3. Additionally, the network generates a binary probability distribution to predict whether manual intervention is required.

During training, the learning capability of the network was enhanced by jointly optimizing the action regression loss and the intervention classification loss. A gating mechanism was employed to combine the correction actions with the actions generated by the simulation policy, effectively bridging the sim-to-real gap. This approach enables the generation of a correction policy with strong generalization ability, requiring only a small amount of manual correction data.

In the experiments, the gear positions in both the real and simulated environments were fixed at (0.55 m, 0.00 m, 0.05 m) as the task configuration. We conducted 50 experimental trials using the simulation policy and the corrected general policy, respectively. The success rates of the assembly tasks are shown in Table Ⅱ, and the recorded assembly trajectories are shown in Figure 8. From the results, it can be observed that the success rate of the simulation policy is 76%, while the success rate of

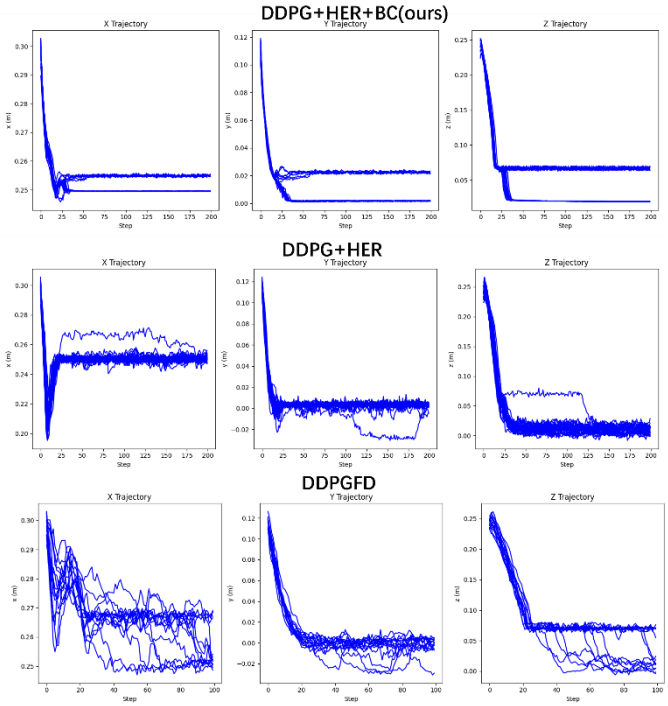

Fig. 7. Trajectory comparison under different training strategies

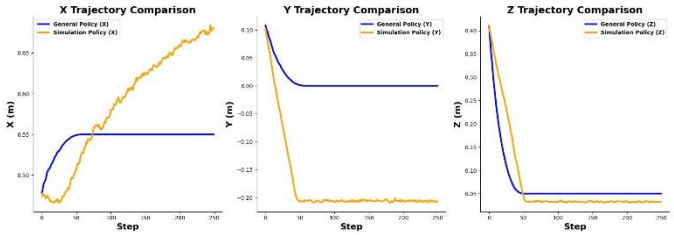

Fig. 8. Position changes using the sim-to-real transfer method

the corrected general policy increases to 94%. Moreover, this method effectively bridges the gap between simulation and reality without requiring a large amount of manual correction data. It can successfully complete the assembly task, with snapshots of the assembly process shown in Figure 9.

V. CONCLUSION

This paper presents a novel framework that integrates Deep Deterministic Policy Gradient (DDPG), Hindsight Experience Replay (HER), and Behavior Cloning (BC) to effectively address robotic assembly tasks in sparse reward environments. The experimental results demonstrate that the proposed framework achieves a success rate of 97% within just 120 episodes. Additionally, the average number of steps required to complete the task is reduced to 75, indicating an improvement in efficiency. In contrast, the baseline DDPG model, trained for 1500 episodes, fails to achieve meaningful success, highlighting the importance of incorporating HER and BC to tackle the sparse reward problem. These results validate the superiority of the proposed approach in improving both learning efficiency and task performance for robotic assembly. Additionally, we

introduce a human-in-the-loop online correction transfer strategy, which bridges the gap between simulation and reality with minimal human intervention data.

Fig. 9. The snapshots of the trained agent performed during the assembly task in the real world

TABLE II. TABLE OF THE RESULTS OF DIFFERENT SIM-TO-REAL TRANSFER METHODS

Transfer Methods	Success rate
Simulation policy	38/50
General policy	47/50

Experimental results demonstrate that this method increases the assembly success rate from 76% to 94%, significantly enhancing the robustness and generalization capability of the model in real-world robotic assembly tasks. This strategy not only effectively addresses the challenges of sparse rewards and limited training data but also provides a scalable solution for transferring policies trained in simulation to real-world applications. Furthermore, the experimental results validate the strong potential of this approach in industrial robotic assembly tasks, substantially improving efficiency and stability.

REFERENCES

[1] M. Andrychowicz, F. Wolski, A. Ray, J. Schneider, R. Fong, P. Welinder, B. McGrew, J. Tobin, O. Pieter Abbeel, and W. Zaremba, "Hindsight experience replay," Advances in neural information processing systems, vol. 30, 2017.

[2] T. Lillicrap, "Continuous control with deep reinforcement learning," arXiv preprint arXiv:1509.02971, 2015.

[3] H. Ravichandar, A. S. Polydoros, S. Chernova, and A. Billard, "Recent advances in robot learning from demonstration," Annual review of control, robotics, and autonomous systems, vol. 3, no. 1, pp. 297–330, 2020.

[4] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: a survey," in 2020 IEEE symposium series on computational intelligence (SSCI), pp. 737–744, IEEE, 2020.

[5] W. Wan, F. Lu, Z. Wu, and K. Harada, "Teaching robots to do object assembly using multi-modal 3d vision," Neurocomputing, vol. 259, pp. 85–93, 2017.

[6] Y. Yin, P. Zheng, C. Li, and K. Wan, "Enhancing humanguided robotic assembly: Ar-assisted dt for skill-based and low-code programming," Journal of Manufacturing Systems, vol. 74, pp. 676–689, 2024.

[7] F. Schirmer, P. Kranz, C. G. Rose, J. Schmitt, and T. Kaupp, "Towards dynamic human–robot collaboration: A holistic framework for assembly planning," Electronics, vol. 14, no. 1, p. 190, 2025.

[8] M. R. Montero, G. Franzese, J. Zwanepol, and J. Kober, "Solving robot assembly tasks by combining interactive teaching and self-exploration," arXiv preprint arXiv:2209.11530, 2022.

[9] X. Sun, J. Li, A. V. Kovalenko, W. Feng, and Y. Ou, "Integrating reinforcement learning and learning from demonstrations to learn nonprehensile manipulation," IEEE Transactions on Automation Science and Engineering, vol. 20, no. 3, pp. 1735–1744, 2022.

[10] Z. Zhang, J. Hong, A. M. S. Enayati, and H. Najjaran, "Using implicit behavior cloning and dynamic movement primitive to facilitate reinforcement learning for robot motion planning," IEEE Transactions on Robotics, 2024.

[11] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," arXiv preprint arXiv:1709.10087, 2017.

[12] M. Vecerik, T. Hester, J. Scholz, F. Wang, O. Pietquin, B. Piot, N. Heess, T. Rothorl, T. Lampe, and M. Ried-¨ miller, "Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards," arXiv preprint arXiv:1707.08817, 2017.

[13] R. Agarwal, C. Liang, D. Schuurmans, and M. Norouzi, "Learning to generalize from sparse and underspecified rewards," in International conference on machine learning, pp. 130–140, PMLR, 2019.

[14] C. Jin, A. Krishnamurthy, M. Simchowitz, and T. Yu, "Reward-free exploration for reinforcement learning," in International Conference on Machine Learning, pp. 4870–4879, PMLR, 2020.

[15] A. Gupta, C. Devin, Y. Liu, P. Abbeel, and S. Levine, "Learning invariant feature spaces to transfer skills with reinforcement learning," arXiv preprint arXiv:1703.02949, 2017.

[16] R. Jeong, Y. Aytar, D. Khosid, Y. Zhou, J. Kay, T. Lampe, K. Bousmalis, and F. Nori, "Self-supervised sim-to-real adaptation for visual robotic manipulation," in 2020 IEEE international conference on robotics and automation (ICRA), pp. 2718–2724, IEEE, 2020.

[17] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in 2017 IEEE/RSJ international conference on intelligent robots and systems (IROS), pp. 23–30, IEEE, 2017.

[18] G. Tiboni, K. Arndt, and V. Kyrki, "Dropo: Sim-to-real transfer with offline domain randomization," Robotics and Autonomous Systems, vol. 166, p. 104432, 2023.