Applied Intelligence
https://doi.org/10.1007/s10489-024-05417-x

Task attention-based multimodal fusion and curriculum residual learning for context generalization in robotic assembly

Chuang Wang[1] · Ze Lin[1] · Biao Liu[1] · Chupeng Su[1] · Gang Chen[1] · Longhan Xie[1]

Accepted: 21 March 2024

Abstract

In the domain of flexible manufacturing, Deep Reinforcement Learning (DRL) has emerged as a pivotal technology for robotic assembly tasks. Despite advancements in sample efficiency and interaction safety through residual reinforcement learning with initial policies, challenges persist in achieving context generalization amidst stochastic systems characterized by large random errors and variable backgrounds. Addressing these challenges, this study introduces a novel framework that integrates task attention-based multimodal fusion with an adaptive error curriculum within a residual reinforcement learning paradigm. Our approach commences with the formulation of a task attention-based multimodal policy that synergizes task-centric visual, relative pose, and tactile data into a compact, end-to-end model. This model is explicitly designed to enhance context generalization by improving observability, thereby ensuring robustness against stochastic errors and variable backgrounds. The second facet of our framework, curriculum residual learning, introduces an adaptive error curriculum that intelligently modulates the guidance and constraints of a model-based feedback controller. This progression from perfect to significantly imperfect initial policies incrementally enhances policy robustness and learning process stability. Empirical validation demonstrates the capability of our method to efficiently acquire a high-precision policy for assembly tasks with clearances as tight as 0.1 mm and error margins up to 20 mm within a 3.5-hour training window-a feat challenging for existing RL-based methods. The results indicate a substantial reduction in average completion time by 75% and a 34% increase in success rate over the classical two-step approach. An ablation study was conducted to assess the contribution of each component within our framework. Real-world task experiments further corroborate the robustness and generalization of our method, achieving over a 90% success rate in variable contexts.

Keywords Robotic assembly · Residual reinforcement learning · Multimodal fusion · Curriculum · Task attention · and Context generalization

1 Introduction

Assembly maintains its pivotal role in the manufacturing landscape, yet it is still predominantly labor-intensive. Although industrial robots are widely used in modern industrial systems because of their ability to perform complex tasks, the integration of robotic systems into these assembly processes presents significant challenges. A comprehensive examination of progress in this domain is detailed in [1]. The principal challenges, as delineated by [2], encompass: (1) the requirement for assembly systems to efficiently navigate an extensive spectrum of interrelationships and environmental conditions, (2) the imperative for precise manipulation using universally applicable equipment, and (3) the necessity to achieve consistent reliability and high success rates within industrial contexts. The task of peg-in-hole insertion

✉ Gang Chen
gangchen@scut.edu.cn

✉ Longhan Xie
melhxie@scut.edu.cn

Chuang Wang
wangchuangwofl@163.com

Ze Lin
201930362071@mail.scut.edu.cn

Biao Liu
437147672@qq.com

Chupeng Su
15914467363@163.com

[1] Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, Guangzhou, People's Republic of China

Published online: 03 April 2024

represents a quintessential benchmark in robotic assembly, categorizing it as a task rich in physical contact (contact-rich) within the realm of robotic manipulation.

In addressing the challenges posed by robotic assembly, the formulation of effective policies, both manually designed and those based on learning algorithms, is essential. Robotic assembly systems often incorporate vision-based methods for broad localization and force-based techniques for precision adjustments, as explored in studies by [3, 4]. However, the real-world application of these two-tiered strategies encounters difficulties in terms of design intricacy and operational effectiveness, often necessitating specialized expertise. Deep Reinforcement Learning (DRL) has been proposed as a potential solution to these issues, with its ability to handle complex tasks via reward mechanisms that are easier to design, as discussed in [2, 5]. Nevertheless, the requirement of DRL for proactive exploration and extensive interaction poses significant challenges, thereby constraining its broader application, as observed by [6].

Residual Reinforcement Learning (RL) has emerged as a viable strategy to address the concerns of unsafe exploration and large data requirements typical of DRL, by building upon existing solutions for directed exploration, as demonstrated by [7]. This approach focuses on refining these existing solutions and has shown potential in various integrative methods within the residual RL framework, including few-shot learning [8] and meta-learning [9]. The residual policy aims to enhance system robustness by learning from variability and supports the transferability of acquired skills across different scenarios by adapting a simple initial policy, a concept further explored by [10]. However, the adaptation of systems to new contexts with substantial stochastic positional errors and diverse environmental conditions remains a challenge for both tactile-based [11] and vision-based [12] residual learning methods. As illustrated in Fig. 1, tactile sensors, while offering localized contact state data, may lack robustness against significant errors due to their limited range [11]. Conversely, vision sensors provide a broader range of sensing but may struggle with variable backgrounds comprising task-independent features and lack precision in pose estimation [12]. Additionally, the learning of a residual policy to address substantial uncertainties in initial strategies is inefficient in low-clearance and contact-rich tasks by random exploration. The extensive error range can provide misleading guidance, complicating the identification of small clearance holes.

The primary aim of our research is to develop efficient methods for acquiring context-generalizable assembly skills, leveraging multimodal fusion and residual Reinforcement Learning (RL). Our objectives include enhancing the robustness of the residual policy in the face of random errors and fluctuating backgrounds, as well as improving the sample efficiency of residual learning amidst the large range of stochastic pose errors. To accomplish these objectives, our approach integrates multimodal data to bolster observability and employs a specially designed policy, encompassing a curriculum and attention, to steer the residual agent's exploration and observation. This paper introduces a novel hybrid framework for assembly manipulation tasks, merging a model-based controller with a context-adaptive residual policy. The model-based controller, grounded in modified Cartesian compliance control and a partial model, is geared towards improving the safety and efficiency of the residual policy's exploration. The residual policy aims to heighten robustness in context generalization by assimilating high-dimensional multimodal information into a succinct fusion model, thereby simplifying the design and enhancing the operational efficiency of the model-based controller. To address the inefficient exploration caused by uncertainties linked to the initial policy during the learning phase, we propose an adaptive curriculum for residual learning. This curriculum offers structured guidance and imposes constraints for exploration within a set context, progressively increasing the uncertainty of the initial policy and decreasing its reliance as the agent's policy gains reliability, thus effectively tackling exploration challenges stemming from potentially misleading initial strategies. Furthermore, we confront the challenge of learning invariant features across various contexts in real robot learning by harnessing initial policy guidance and introducing a task-focused attention mechanism for observation. This mechanism allows the agent to concentrate on the immediate task while minimizing the impact of the variable environment, thereby enabling context generalization through the guidance of learning invariant features. Experimental results show that our proposed framework outperforms existing methods in both learning cost and strategy performance. Ablation experiments demonstrate that residual learning can efficiently learn multimodal strategies to improve robustness under larger random errors, curriculum residual learning has a significant impact on learning efficiency, and task-focused observation can improve generalization. Real robot experiments show that our method has considerable application potential.

Our key contributions are as follows:

1. We present a framework for assembly tasks based on residual RL, combining a Cartesian compliance control and trajectory with a residual multimodal fusion policy. This innovative framework is adept at efficiently learning the robust policy on real robots and low-cost adapting to diverse contexts.
2. We construct a task-focused attention mechanism for high-dimensional multimodal sensory data to be integrated by a streamlined neural network architecture. This design significantly aids in the formation of a robust residual policy capable of managing substantial random errors and adapting to various environmental settings.

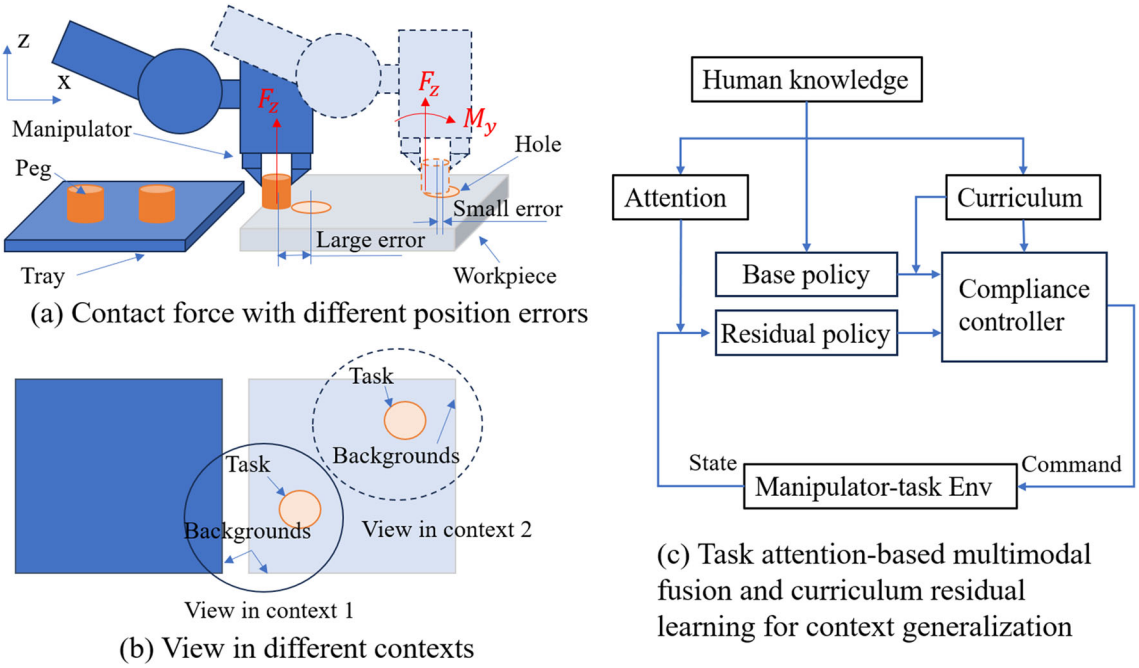


Fig. 1 The proposed framework for context generalization with large position error and variable backgrounds. (a) Different from the point or line contact with a small error, a large error will result in a region with only normal contact force, which provides an undifferentiated contact state. (b) A wide view consists of task-related information, such as the edge of the hole, and background information, such as the edge of the tray and workpiece. (c) This work proposes a framework of Task attention-based multimodal fusion and curriculum residual learning for context generalization

3. We further the application of adaptive curriculum learning within the realm of residual learning by an automated system dynamically modulating the uncertainty inherent in the initial policy. This breakthrough enhances the learning efficiency and stability for robust residual policy.
4. Through a range of comparative and ablation studies, the efficacy of our proposed framework is rigorously validated. We also perform extensive evaluations to establish its robustness in a variety of assembly tasks and under different environmental conditions.

2 Related work

This section summarizes the background and foundations related to our approach. The related works of robotic assembly tasks through multimodal information and DRL are summarized, and the concepts and methods of vision attention mechanism and curriculum learning are discussed.

2.1 Robotic assembly tasks with multimodal information and deep reinforcement learning

Robotic assembly, integral to flexible manufacturing and contact-rich manipulation, has undergone significant research developments in recent decades. The integration of multimodal information has been pivotal in augmenting the robustness and reducing the noise sensitivity in stochastic systems, as evidenced by the literature [1]. Noteworthy in this context is the contribution of [13], who implemented haptic feedback in admittance control for adaptive response to environmental disturbances, and [14], who introduced an innovative compliant control method using Gaussian Mixture Models (GMM) to assimilate learning from the manual assembly. Further, the works of [3] and [4] have enriched this domain by integrating vision and tactile information to enhance robustness and precision in manipulation tasks. Despite these advancements, the conventional two-step process of coarse vision-based positioning followed by fine force-based alignment remains suboptimal. This limitation is analogous to the challenges in human assembly learning, where acquiring skills from expert demonstrations or predefined policies is complex and often inadequate. Furthermore, the costs associated with reprogramming policies or collecting expert data escalate with the growing diversity of assembly tasks.

Deep Reinforcement Learning (DRL) has emerged as a promising method for learning optimal perception and control policies through iterative interactions. Various studies have employed DRL-based techniques to fine alignment and

insertion policies in assembly tasks, presupposing the correction of pose errors by vision systems and the implementation of safe interaction controllers [15–19]. For the integration of RL and compliance controllers, other methods have investigated approaches such as variable impedance control and learning-based force control frameworks to dynamically adapt to different stages of assembly [20–23]. Residual RL has been proposed to utilize a foundational policy, guiding and constraining the exploration of agents in initial learning stages, thus improving exploration and learning efficiency [8, 24]. However, these methods have not fully integrated visual information in the agent, necessitating the provision of precise assembly positions (error less than 2 mm) for initial policy through specialized, costly demonstrations or vision systems.

The application of multimodal information through DRL has been explored in recent studies. Some research has focused on learning latent state spaces and incorporating these compact representations into RL frameworks to enhance sample efficiency and generalization [5, 25]. Others have distinguished between perception and policy, addressing the sim-to-real gap and generalization to novel object shapes through domain randomization and impedance control [26–29]. Direct visual-haptic fusion has also been explored, with policies initially trained in simulated environments and subsequently fine-tuned or transferred to real-world settings [2, 30, 31]. Although sim-to-real approaches can reduce real robot learning time, the disparity between simulated and real environments or different object shapes presents challenges in real-world tasks, leading to considerable performance declines during generalization. The dependency on high-quality simulation environments restricts their practical manufacturing applicability.

Our work diverged from these methodologies by introducing a residual RL approach that leverages both visual and force data for completing the three stages of peg-in-hole assembly tasks. Our robust multimodal residual policy simplifies the design complexity and parameter tuning of the initial policy, facilitating swift adaptation to new contexts through modifications to the initial policy's waypoints. Guided by the initial policy, we directly train the residual policy in real-world settings and repurpose it across various contexts, effectively bypassing generalization gaps and promising improved performance metrics, such as success rate, insertion efficiency, and operational delicacy. This strategy aligns with the goals of [32], while also addressing the limitations of suboptimal teaching data highlighted by [23].

2.2 Curriculum learning for precise tasks

Curriculum Learning (CL) has been recognized as a potent methodology for augmenting the sampling efficiency of Reinforcement Learning (RL). It achieves this by progressively increasing task complexity and strategically guiding exploration during the learning phase, as discussed in literature [33–35]. The work of Dong et al. [36] exemplifies the effectiveness of CL in progressively more complex insertion tasks (wall→corner→U→hole). By systematically escalating the complexity of these tasks, they achieved enhanced data efficiency and facilitated broader generalization across various objects. Similarly, Jin P et al. [31] further advanced CL by tailoring task difficulty to correlate with sensory data inputs, thereby dividing the training process into pure visual policy learning with larger peg-hole clearance and continued vision-force policy learning with narrowed peg-hole clearance. In terms of exploration curriculum within a single task, Luo et al. [37] employed a curriculum that incrementally increased task complexity and gradually relaxed safety constraints, successfully training an RL agent for a dynamic insertion task. This approach began with a non-randomized and static socket pose, maximizing the likelihood of task completion through random exploration. Hermann et al. [38] tackled exploration challenges by introducing an adaptive curriculum generation algorithm that modulated difficulty via varying the sampling of initial states from demonstration trajectories. This method initiated task engagement with simpler conditions and maintained rewards within a desirable success rate range by initially sampling states from the end of trajectories, gradually moving towards the start. However, despite these innovations, existing implementation of task curricula often requires a sequence of simplified tasks and modifications to the task environment during learning, necessitating significant human intervention and potentially incurring costs in real-world insertion task training. Moreover, existing methods for exploration tend to focus on modulating the onset and step size of exploration without offering comprehensive guidance.

Diverging from these existing approaches, our research delves into the application of adaptive curriculum learning within the context of residual learning, specifically targeting imperfect trajectories and compliance control. By adaptively adjusting trajectory error and controller stiffness, our methodology streamlines the learning process for complex assembly operations, reducing the dependency on manually crafted curricula and minimizing the need for human intervention. This innovative curriculum strategy first enables agents to learn contact force feedback-based force control with small errors and then to learn vision-based error compensation, providing more robust guidance and demonstrating practical utility for robotic assembly tasks. This approach represents an improvement in the autonomous learning of precision tasks, contributing substantially to the field of robotic assembly and machine learning.

2.3 Task-centric attention mechanisms for context generalization

Context generalization is a paramount consideration in robotic task execution, especially when a robot is expected to function in diverse locations and under various background conditions. Existing methods can be divided into those that increase the similarity between training and testing environments, and those that explicitly aim to handle differences [39]. A common strategy to increase the similarity between training and testing environments involves training an end-to-end policy in a wide array of contexts, to explicitly improve adaptability [40]. To achieve generalization across different backgrounds or tasks, techniques such as data augmentation, domain randomization, and task distribution have been utilized during training. These techniques focus on extracting invariant features that are consistent regardless of variations between training and testing environments [41]. Nonetheless, applying these techniques in real-world robot learning faces two major obstacles [42]. Firstly, developing a policy that generalizes across multiple contexts often requires additional human input for context reconfiguration and extensive data collection. Secondly, the arbitrary visual representations learned may not effectively contribute to control tasks. Recent advancements in generalization for robotic systems in unfamiliar environments have highlighted the effectiveness of incorporating scene priors [43]. Among various strategies, the use of keypoint-based representations has emerged as a notable approach. These representations focus on encoding task-specific, simplified information, often termed 'attention' [25]. While this method has shown promise by decoupling perception from planning and control to enhance sample efficiency, it inherently depends on learning across diverse contexts. In one hand, the process of identifying keypoints typically involves human labeling, a task that is not only labor-intensive but also a bottleneck for scalability [44, 45]. In the other hand, selecting the appropriate key points to represent the state of the task-robot and the reliance on human-labeled data for keypoints identification continues to be a significant challenge.

Our research diverges notably from pure learning-based and knowledge-based invariant feature acquisition methods by introducing a unique approach that harnesses coarse task knowledge to direct attention. Central to this approach is the use of rectangular boxes, steered by a foundational policy, to strategically focus attention, representing the application of Regions of Interest (ROIs) as a mechanism for attention. This technique enables Reinforcement Learning (RL) to focus on the task's most crucial, contact-dense segments, thereby honing the feature learning scope to key elements. Additionally, integrating an eye-in-hand camera system grants the agent an advantageous perspective for task execution. The employment of ROIs for attention is particularly efficacious to obviate the need for manual keypoints identification and training across diverse contexts. This innovation signifies a substantial leap forward in robotic task learning, offering a viable and efficient resolution to the challenges of context generalization and attention allocation in robotic systems.

3 Problem statement

This study addresses the intricacies of robotic assembly tasks, marked by stochastic position errors and dynamic background changes. We frame the problem within a stochastic control system context. This system includes the robot and its interactive configuration space with the task. The observation state of the robot at a given time t, represented as $Y(t)$, is determined through measurements that incorporate an error component $N(t)$. The robot's control actions are denoted as $u(t)$. Challenges arise due to uncertainties in grasping and motion control, introducing a random disturbance $V(t)$ in the assembly process. Consequently, the system's dynamics are modeled as follows:

$$\begin{aligned} \dot{X}(t) &= A(t)X(t) + B(t)u(t) + V(t) \\ Y(t) &= C(t)X(t) + N(t) \end{aligned} \tag{1}$$

where, $X(t)$ indicates the state of the robot, starting from the initial condition $X(t_0) = X_0$. The matrices $A(t)$, $B(t)$, and $C(t)$ represent time-variant system parameters.

The manipulation process, particularly in contact-rich scenarios, is primarily challenging due to uncertainties in geometry, pose, and dynamics. To tackle this, the process is segmented into free-space motion and contact-rich manipulation, with the latter being further divided into deterministic and stochastic components. While a model-based controller may be adequate for free-space motion and the deterministic part of contact-rich interactions, the stochastic disturbances in dexterous contact-rich manipulation necessitate a learning-based residual policy.

The control architecture is formalized as follows:

$$u_t = C_h\left(\pi_h\left(t\right), \pi_\theta\left(s_t\right)\right) \tag{2}$$

where the action u_t is the output of the hybrid controller $C_h(*)$, integrating a hand-designed policy $\pi_h(t)$and a learning-based policy $\pi_\theta(s_t)$.

To effectively manage stochastic systems with significant random errors and diverse backgrounds, it is imperative to efficiently develop a policy $\pi_\theta(s_t)$ in one context and reuse it in multiple contexts. Given the random nature of localization errors across different contexts, the policy must demonstrate robustness. Furthermore, system design should account for variable backgrounds to ensure consistent deployment

performance across various contexts, highlighting the importance of policy generalization.

4 Method

In this work, we introduce a practical framework for robotic assembly manipulation, which is schematically represented in Fig. 2. The framework is structured around three pivotal components: (1) Hybrid Policy Architecture (HPA): We propose an innovative HPA that integrates model-based planning for contact-rich scenarios as Manipulation Primitive (MP), ensuring effective motion control and safe interactions. To address estimation and motion errors, the HPA is augmented with a residual policy, enhancing its reliability and performance. (2) Task Attention-based Multimodal Residual Policy (TA-MRP): As the core of the residual policy, we employ a task attention-based multimodal model for stochastic policy. This model leverages task-focused attention mechanisms and multimodal sensory inputs to bolster its resilience against significant random position errors and diverse background conditions. (3) Adaptive Curriculum Residual RL (ACRRL): To optimize the training of the residual policy, we employ a novel Adaptive Curriculum Residual Reinforcement Learning strategy. This method utilizes a curriculum generation policy that systematically orchestrates the RL agent's exploration, progressively enhancing its ability requirement to the error of the initial policy. The detailed introductions of the three components are in the following sections.

4.1 Hybrid policy architecture

In this study, we present a hybrid policy architecture that synergistically combines residual reinforcement learning (RL) with parallel position and force control. This architecture is designed to facilitate efficient manipulation skill learning and reconfiguration on a rigid robot. Our approach distinguishes between contact-rich and contact-free segments of the manipulation task, basing the skill design on the applicability of traditional feedback control. Specifically, the contact-rich segment is further divided into a portion addressable by conventional feedback control and a residual segment managed by a parameterized residual policy, as shown in Fig. 3. We propose a contact-rich motion planning strategy to ensure safe interactions and provide initial guidance for the manipulation task.

4.1.1 Cartesian compliance control for rigid robot

We have developed a compliance control system for a rigid robot equipped with a wrist-mounted 6D force/torque sensor and a kinematic solver. The compliance control aims to

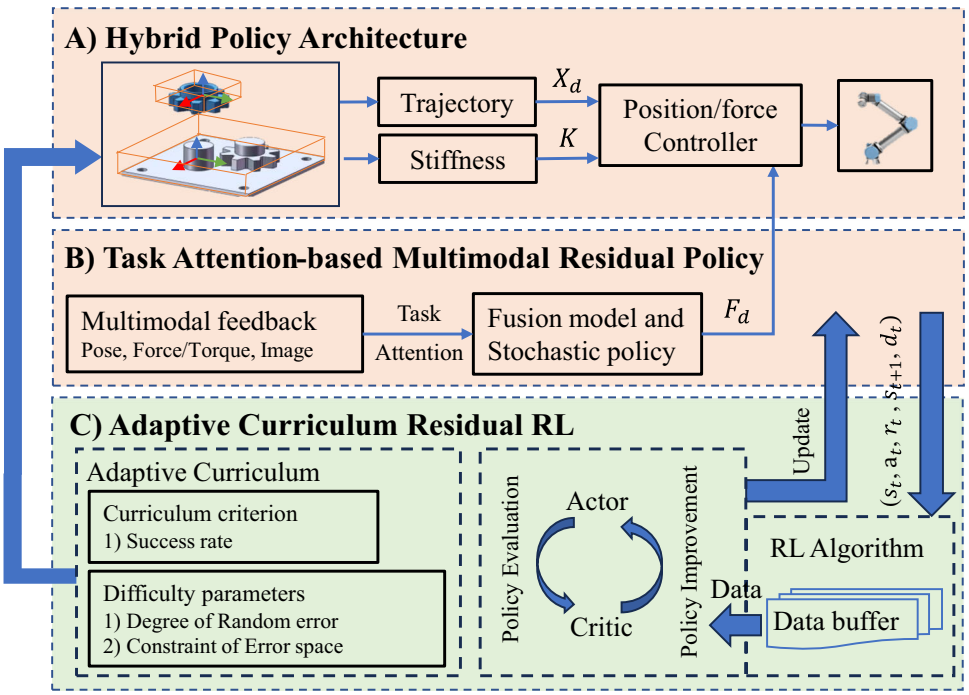


Fig. 2 System overview of the proposed method. A) Hybrid policy architecture combines model-based planning with learning-based residual policy for assembly manipulation. B) Task attention-based multimodal fusion encodes and fuses task attention-based multiple information modalities to generate the stochastic residual force policy. C) Curriculum residual reinforcement learning automatically adjusts the exploration guidance for efficient and stable learning

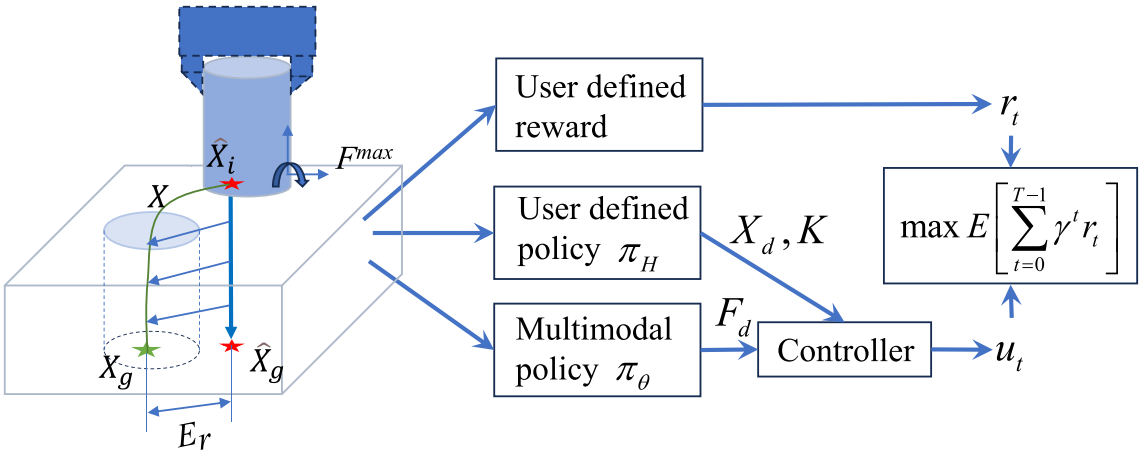


Fig. 3 Hybrid policy for contact-rich manipulation in a peg-in-hole task. The trajectory and low-stiffness controllers acted as coarse guidance and constraints. The residual policy is used to compensate for their accuracy

modulate the dynamic behavior of the robot during interaction, characterized by the relationship between the external force and the robot's motion. This relationship is modeled as a mass-spring-damper system, described by the following equation:

$$F = K(X_d - X) + B(\dot{X}_d - \dot{X}) + M(\ddot{X}_d - \ddot{X}) \tag{3}$$

where M, B, and K represent the virtual mass, damping, and stiffness matrices, respectively. F and F_d are the measured interaction force and the desired force. X and X_d denote the current and desired poses, respectively.

To minimize overshoot and ensure smooth motion, we assume that the command X_d remains static over short intervals, allowing us to neglect $\dot{X}_d$ and $\ddot{X}_d$. We introduce a virtual force F_d into the model to manage the dynamic interaction, leading to:

$$F - F_d = K(X - X_d) + B\dot{X} + M\ddot{X} \tag{4}$$

Employing the virtual force-driven spring-mass-damping modal and robot kinematics, we utilize a modified Cartesian parallel position and force controller as the low-level control for robot learning, generating velocity commands. The control law for joint velocities $\dot{q}$ is expressed as:

$$\dot{q} = \frac{J^{-1}M^{-1}}{s + M^{-1}B}\left[K\left(X_d - X\right) - \left(F_d - F\right)\right] \tag{5}$$

where Jacobian matrix J provides the relation between end-effector and joint velocities. Desired poses and forces X_d and F_d govern the behavior, with the stiffness matrix K balancing the six-dimensional tracking error for position/orientation and force/torque. The inertia matrix M and damping matrix B influence the response speed and stability.

4.1.2 Hybrid policy

For the efficient derivation of policies suitable for complex assembly tasks, we define a hybrid policy-based MP that fuses model-based planning with model-free learning for control architecture in (2), as follows:

$$\dot{q} = f_c\left(\tau(t), K, \pi_\sigma(s)\right) \tag{6}$$

where f_c represents the parallel position and forces controller. The trajectory $\tau(t)$ produces a time series of desired poses X_d. The stiffness matrix K pertains to the compliance controller. They work as the hand-designed policy $\pi_h(t)$ to deal with the known part $A(t)$, $B(t)$, and $C(t)$ of the system in (1). $\pi_\theta(s_t)$ generates a force profile F_d in both translation and rotation based on the state s_t to handle the measurement error $N(t)$ and random disturbance $V(t)$ in (1).

We utilize partial model-based planning to define the hand-designed policy $\pi_h(t)$, which includes the trajectory $\tau(t)$ and the stiffness K. Initially, we assume that a partial model of the geometry and contact requirements is available, as shown in Fig. 3. This model enables us to identify the estimated pre-assembly point $\hat{X}_i$ and target assembly point $\hat{X}_g$ for manipulation. The contact safety requirement F^{max} serves as a safety constraint for both the robot and the product. The error range E_r is the distance between the real target assembly point X_g and $\hat{X}_g$. Subsequently, we propose contact-rich motion planning to generate the trajectory and stiffness. A linear trajectory $\tau(t)$ is generated using the pre-assembly point $\hat{X}_i$ and the target assembly point $\hat{X}_g$. The relationship between the trajectory deviation and the force is derived from (2), as demonstrated in (7). Taking into account the geometric constraints and the estimation error E_r, a deviation workspace W is formed as shown in (8). Using the half contact safety requirement F^{max} as the action space range of

the agent, the safe interaction constraint is achieved through the stiffness matrix K, as indicated in (9).

$$K\,(X_d - X) = (F_d - F) \tag{7}$$

$$W = E_r + \left(\hat{X}_i - \hat{X}_g\right) \tag{8}$$

$$K = 1/2 * F^{max} diag(W)^{-1} \tag{9}$$

The virtual desired force F_d is capable of compensating for trajectory errors and geometric constraints by generating an offset $X_d - X$. Additionally, the contact force F should meet the contact safety requirement F^{max}. Therefore, the defined K aims to balance position tracking with the exploration capabilities and safety of the residual policy, which generates F_d to account for trajectory discrepancies and variable force policies.

4.2 Task attention-based multimodal residual policy

This section presents the TA-MRP, as depicted in Fig. 4, designed to address uncertainties inherent in planned initial policy and variations in the operational environment. TA-MRP leverages preliminary task information and planned motion guidance to enhance the agent's observability over random errors and maintain focus on the task, thereby facilitating context generalization. The state representation is preprocessed with obvious prior knowledge to encapsulate only the task-relevant information and cover the full task's properties. A compact neural network for multimodal information fusion and stochastic policy is designed to learn features directly from the raw sensory inputs and generate actions.

4.2.1 Knowledge-based task attention for data preprocessing

To effectively encode the state of the robot and task, s_t, our approach integrates three distinct types of sensor data as the observation space for the policy: the gray-scale image from the eye-in-hand I_t^{eih}, proprioceptive data about the tool's pose X_t, and tactile feedback from the wrist-mounted force-torque sensor F_t, collectively described as follows:

$$s_t = [I_t^{eih}, X_t, F_t] \tag{10}$$

Guided by the initial policy, the feedback data undergoes preprocessing to concentrate on the task at hand. For pose information, we adopt a strategy that sets the target pose $\hat{X}_g$ as the task's 'origin' and expresses spatial information in terms of the relative pose Xr_t, in line with the methodology presented in [37]. This representation of relative pose is invariant to the manipulation task's pose, facilitating context generalization. The tactile sensor data F_t is intrinsically task-focused, providing direct feedback on the contact state at the

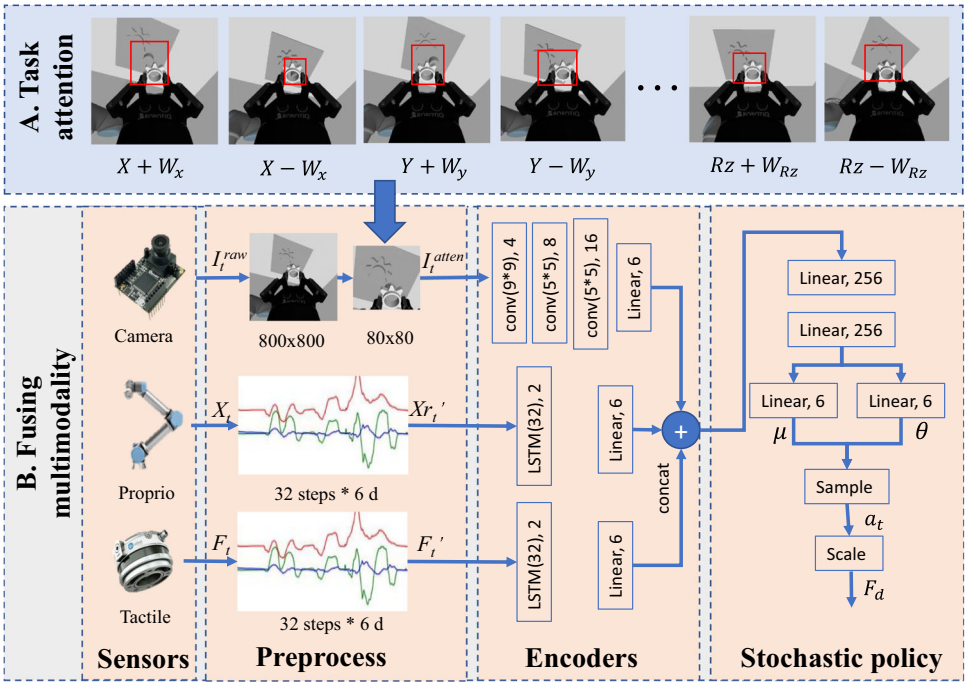


Fig. 4 Task attention-based multimodal stochastic policy. A) Task attention generates ROI for practical and efficient context generalization learning. B) Multimodal fusion module obtains, preprocesses, encodes, and fuses three modalities to generate stochastic force policy

specific interaction point under the guidance of the initial policy, independent of environmental variations. To enhance tactile perception in localizing the contact, we maintain a history of the last 32 readings of relative poses Xr'_t and tactile data F'_t, capturing the force distribution over the object's surface.

The raw image I_t^{eih} is cropped to a defined region of interest (ROI), centering the image on the task at hand. This process effectively isolates the task-relevant visual information, resulting in a focused image I_t^{atten}, as shown in Fig. 4. Given the limitations of position error manageable by the hand-designed controller and visual region related to the task, we focus on a restricted camera field of view to bolster policy robustness. The vision input I_t^{eih} offers a comprehensive view of the environment, contributing to observability but potentially compromising generalization. To counteract this, we decompose the context into subcomponents, such as the end-effector, table, background, gear, and task board. With prior knowledge of the task, we define an ROI $[(x_{min}, y_{min}), (x_{max}, y_{max})]$ that encapsulates the task's critical features while excluding visual noise [46]. For example, the ROI for the gear assembly task includes the peg-in-hole and gear mesh but ignores the table. Additionally, considering the uncertainty of the hand-designed controller, vision provides essential information about the pose of the assembled objects, particularly during the search and alignment phases. Therefore, the adjusted ROI must encompass the necessary features within the maximum allowable error range E_r across n search directions. For instance, in the gear assembly task, the robot needs to be aligned in the X, Y, and R_z directions before the insertion operation, and the ROI is determined by the intersection of the ROIs within the E_r range at pre-assembly point $\hat{X}_i$ across the three alignment directions. The formula for determining the task-centric attention mechanisms' ROI is as follows:

$$ROI = [min(x^i_{min}, y^i_{min}), max(x^i_{max}, y^i_{max})], i = 0, ..., n \tag{11}$$

4.2.2 Multimodal fusion and stochastic policy

Distinct domain-specific models are employed to capture the unique characteristics of each sensory modality [5]. For physical contact state identification, we utilize a 2-layer Long-Short-Term Memory (LSTM) network with 64 hidden nodes to encode the time series data from touch and proprioception [47], resulting in a 6-dimensional feature vector. The visual feedback is processed through a 3-layer Convolutional Neural Network (CNN), which converts the 80×80 gray-scale image into a corresponding 6-dimensional feature vector. The convolutional layers use progressively increasing kernel sizes and stride parameters to extract relevant features from the image.

The extracted feature vectors from each modality are concatenated into an 18-dimensional vector and passed through a fusion module, which integrates the multimodal information. A stochastic policy, represented by a neural network, then processes the fused feature vectors to establish a parametric probability distribution. The policy network comprises two fully connected layers with 256 units each, which output the mean μ and standard deviation θ for a 6-dimensional Gaussian distribution. Actions a_t are sampled from this distribution and correspondingly mapped to the predefined range of the desired force F_d, which constitutes the residual policy for robot control.

$$a_t = [F_d] = \pi_\theta\,(s_t) \tag{12}$$

4.3 Adaptive curriculum residual reinforcement learning

In tackling contact-rich problems, we model the compensation for the hand-designed policy as a Markov decision process (MDP). Deep reinforcement learning (DRL) emerges as an apt approach for training a deep model in such tasks by maximizing cumulative rewards. The policy is trained within a structured environment enriched with information pertinent to the hand-designed policy and task-associated rewards, as illustrated in Fig. 5. An adaptive curriculum generation in Residual RL for guidance and constraint is designed to refine the efficiency of residual learning and enhance the robustness of the residual policy.

4.3.1 Curriculum learning for the robust residual policy

Within the residual RL framework, the initial policy serves as guidance for the RL agent, which in turn compensates for uncertainties. To ensure robustness against the variability of the initial policy and the compliance controller across different contexts, we introduce errors into the initial policy to induce sufficient variability during training. This is achieved by generating initial and goal points with an error component sampled from a uniform distribution within an error range E_r, as shown in (13). The resulting error-infused guidance necessitates a broader search range for the residual agent. As defined in (9), the search range is governed by the stiffness matrix K, which moderates the balance between the initial policy and the TA-MRP. For smaller K value results in a larger trajectory deviation for a given force, effectively increasing the search step size. The exploration range for pose error in workspace W is defined as Ex_r.

$$error \sim U(-1, 1) * E_r \tag{13}$$

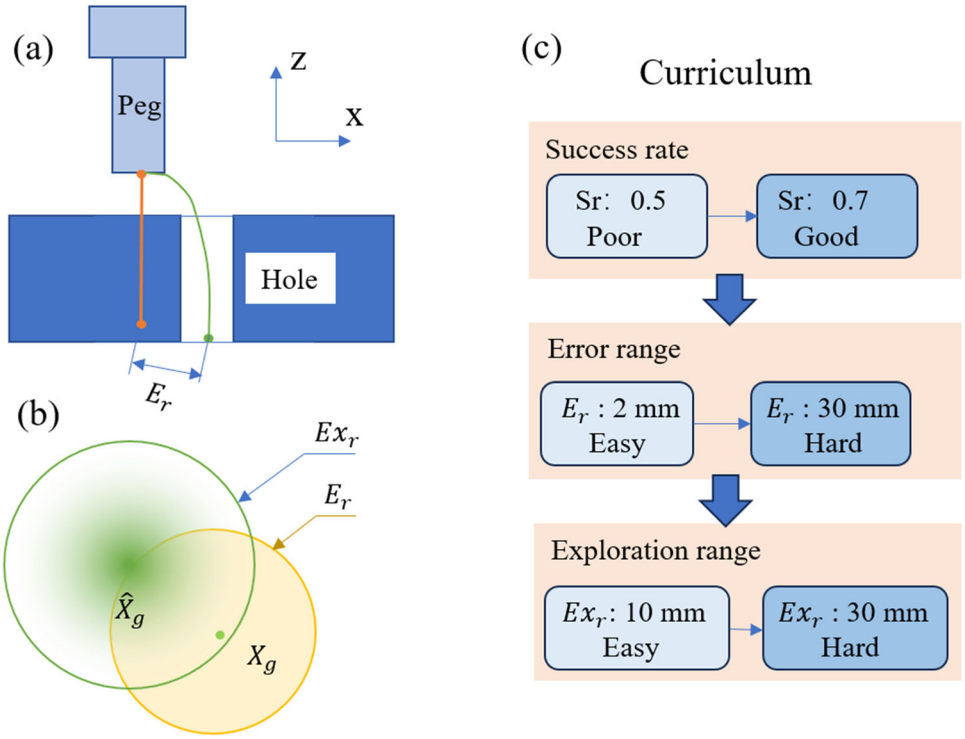


Fig. 5 The error and exploration in the residual learning. (a) The real goal pose is known in the structured training environment, which provides necessary information for reward and curriculum. (b) The estimated goal pose is evenly distributed within the margin of error. However, the agent conducts random exploration similar to Gaussian distribution within a certain range around the estimated pose. (c) Random error and exploration range determine the robustness of residual policy, which directly impacts the task's difficulty

$$\hat{X}_i = X_i + error \tag{14}$$

$$\hat{X_g} = X_g + error \tag{15}$$

As depicted in Fig. 5 (a) and (b), the challenge in learning TA-MRP is tied to the guidance and constraints imposed by the initial policy. Large errors may cause the exploration center to deviate significantly from the true position, and an excessively large step size may hinder the precision required for tasks like insertion. It is crucial to maintain an initial exploration range that exceeds the estimated pose error of X_g to ensure the agent can locate the target by searching. Our approach introduces a curriculum that incrementally escalates the task's difficulty, as shown in Fig. 5 (c). Initially, the error and exploration range are minimal, providing the most accessible configuration and effective guidance through prior knowledge. During training, we progressively increase the difficulty using coefficients $\delta \subset [0, 1]$. Finally, training with realistic levels of uncertainty is vital to develop a robust policy.

Following the principles outlined by [38], we define ACGRL to refine our policy training. While progressively enhancing the difficulty level is beneficial, excessive dwelling on a limited error range can impede training progress. Conversely, escalating the error range too rapidly may preclude the policy from receiving any reward. Our ACGRL strategy modulates the degree of misleading guidance (variance in position and orientation error E_r) and the exploration area (Ex_r constrained by the stiffness matrix K) based on the current success rate, adjusting the challenge to keep the agent's performance within a target success rate window $[\alpha, \beta]$ (16).

$$\delta = \delta + \varepsilon \cdot 1_{(sr>\alpha)} - \varepsilon \cdot 1_{(sr<\beta)} \tag{16}$$

4.3.2 Reward and context generalization TA-MRP learning

The optimal behavior is characterized by rapid and smooth task completion. In structured training environments, dense and sparse rewards are formulated using normalized ground truth distances and contact forces. The reward function is weighted by w_1, w_2, and w_3 to promote the desired behavior.

$$r_t(s_{t+1}) = -w_1 \left\| \frac{X - X_g}{W} \right\| - w_2 \left\| \frac{F}{F^{max}} \right\| + w_3 r_{succ} \tag{17}$$

$$r_{succ}, d = \begin{cases} 100, 1 & \text{if } X - X_g < E_{th} \\ 0, 0 & \text{otherwise.} \end{cases} \tag{18}$$

where W and F^{max} represent the estimated exploration range and maximum contact force, respectively, which normalize the reward terms for task-invariant weight ratios. The success reward r_{succ} and the signal d are contingent upon the distance between the current pose X and the ground truth target pose X_g being less than a threshold E_{th}. $||*||$ denotes the Euclidean norm.

In this work, we augment the soft actor-critic (SAC) off-policy algorithm with multimodal actors and critics to achieve context generalization in residual TA-MRP learning. The multimodal SAC interacts with the environment and trains the policy as outlined in Algorithm 1. The algorithm initializes the learning architecture (lines 1-5), which includes defining the hand-designed part of the skill, determining the uncertainty range, obtaining the ROI, and designing the reward function. Lines 6-10 detail the learning of the residual TA-MRP through environmental interaction, policy updates via SAC, and curriculum generation based on the success rate. In the interaction, we collect the stack position and force/torque time series and vision as states, compute the reward, and store them in the replay buffer R. The episode is terminated upon successful insertion or maximum steps. In the training process, we train the multimodal actor and critic by sampling from R. In particular, the multimodal fusion and the stochastic policy are trained together by interaction without any additional pre-training effort.

Algorithm 1 Learning context-generalized TA-MRP with ACRRL)

1: INITIALIZE THE LEARNING ARCHITECTURE
2: Define the hand-designed part of the skill with (14)
3: Determine the error range E_r
4: Obtain the ROI for E_r with (11)
5: Design reward for learning objective with (17) and (18)
6: Choose curriculum parameters for E_r with (14) and (16)
7: LEARN RESIDUAL TA- MRP
8: **For** n in max-episodes
9: Interact with environment
10: Update the policy with SAC
11: Generate curriculum with (16)
12: **For end**

5 Experiment

We propose an innovative manipulation primitive framework tailored for robotic assembly tasks. This framework integrates a learning-based residual policy with a model-based feedback controller, to learn and adapt efficiently across various contexts. To improve the context generalization of the policy and sample efficiency of residual RL, especially under the uncertainties of a low-cost context setup, we introduce several methodological enhancements. In the following, we evaluate the performance of our task attention-based multimodal policy and the curriculum residual learning approach through peg-in-hole and gear assembly tasks. Our evaluation comprises a comparative analysis with existing methods, an ablation study to assess the contribution of each enhancement, and a validation of the framework in real-world tasks.

5.1 Simulation experiments

5.1.1 Experiment setup

Hardware and software The simulation experiments are conducted using an Nvidia Titan RTX GPU and an Intel i9-7940X CPU. We employ Gazebo version 11, utilizing the Open Dynamics Engine (ODE) for physics simulation, to create a virtual environment for the robot and assembly tasks. The Robot Operating System (ROS) serves as the middleware facilitating communication between the learning algorithms, control systems, and simulation modules.

Task and manipulator The assembly tasks involve a task board secured to a worktable and a robot arm equipped with either a peg or gear, as shown in Fig. 6. The peg-in-hole task requires precise insertion, while the gear assembly task demands the alignment of gear teeth with a mating wheel on a shaft. These tasks necessitate tolerances below 0.1 mm and 0.03 rad, respectively, which exceed the capabilities of most contemporary assembly robots, particularly those using commercial-off-the-shelf (COTS) components. [48]. Moreover, the learned strategies must be transferable across

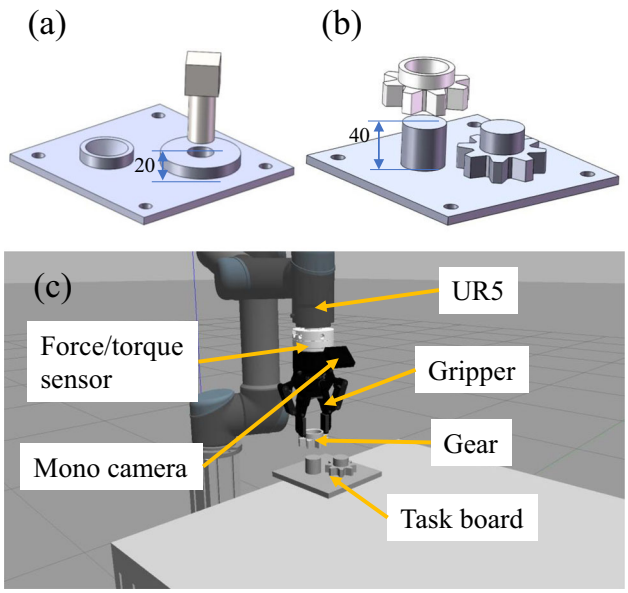


Fig. 6 Simulation assembly task and robot. (a) and (b) show the geometric information of the assembly task. (c) is the manipulator for the assembly task

various scenarios. To simulate real-world inaccuracies, we introduce pose errors for the initial strategies along the x, y, and z axes within a range of ±20 mm, and rotational errors about the z-axis within ±0.1 rad. The context variability is further reflected by modifications in product variables and the unstructured nature of the environment, such as different assembly positions or alterations in surrounding objects.

Evaluation metrics To evaluate the sample efficiency and context generalization of the proposed method, the following metrics are presented: success rate (SR), completion time (CT), and contact force (CF). Cumulative reward (CR), error range (Er), and exploration range (Exr).

Initial policy and reward design According to the geometry information, operation requirement, and estimated localization of the fixed task board, the trajectory is generated with the starting point, target point, and required time. Maximum contact force F^{max} is set as 10 N. The hybrid policy updates the pose and force commands to the controller at 5 Hz. The controller receives the inputs, including reference commands and feedback, and generates the target joint velocity for the robot at 120 Hz. Each episode lasts a maximum of 120 steps. We updated the policy nets for 200 gradient steps per episode. The reward weights w_1, w_2, and w_3 are set as 1, 0.8, and 1 determined by a preliminary experiment to balance operation speed and smoothness.

5.1.2 Comparative study

In our comparative analysis, we assess the performance of our proposed manipulation primitive framework against established methods in the domain of industrial assembly tasks, a simulated peg-in-hole task. The comparison encompasses both classical approaches and contemporary reinforcement learning (RL)-based strategies. Baseline 1 employs a two-step method, as outlined in [49–51], which divides the gear assembly process into reach, search, and insertion phases. These phases are tackled using a visual servo and a spiral search complemented by contact state detection. Baseline 2 represents a Vanilla multimodal RL approach without a base policy, drawing parallels with the works of [2, 19, 30]. Baseline 3 integrates vision-based RL with force control, akin to the methods described in [26–28]. Baseline 4 leverages a keypoint-based vision representation fused with tactile feedback, similar to [45].

5.1.3 Experiment result

Our method exhibits remarkable efficiency in the simulated peg-in-hole task, as depicted in Fig. 7. Benefiting from an initial policy that provides directional guidance, our policy demonstrates a notable probability of completing the peg insertion through random exploration alone. The ACRRL mechanism enables the residual agent to initially concentrate on the insertion task, even with a minimal error range. Upon reaching a predefined success rate threshold, the trained policy swiftly adjusts to increasing positional errors. The TA-MRP integrates visual and force feedback, equipping the agent to handle significant random errors and achieve precise localization. After 200 training episodes, the policy's performance, evaluated over 20 test trials, is summarized in Table 1. The learned TA-MRP attains a 100% success rate, with an average completion time of 2.41 seconds and average contact forces of 12.14 N, 6.29 N, and 3.23 N in the X, Y, and Z directions.

In contrast, the two-step method is hindered by its reliance on intricate policy design and extensive parameter tuning,

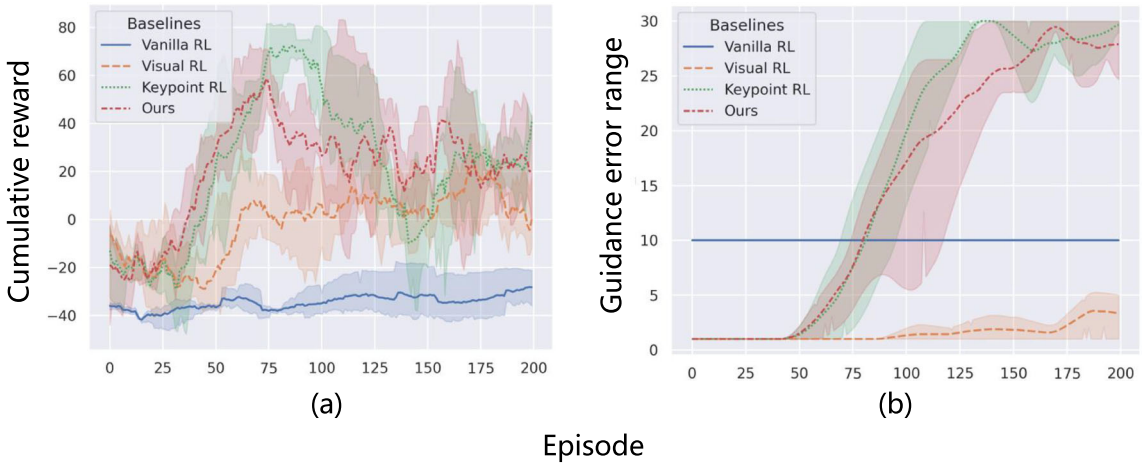


Fig. 7 Learning curves of the policies with three kinds of observations. (a) is the cumulative reward in each episode. (b) is the guidance error range of the fixed policy

Table 1 Comparison of execution with three strategies

Task	Success rate	Completion time (s)	Contact force-x (N)	Contact force-y (N)	Contact force-z (N)
Two-step	0.533	19.000 ± 5.751	13.21 ± 4.53	13.22 ± 4.02	9.78 ± 2.074
Vanilla RL	-	-	-	-	-
Visual RL	-	-	-	-	-
Keypoint RL	0.96	4.27 ± 2.97	**8.02 ± 4.37**	**5.13 ± 2.58**	**3.07 ± 0.87**
Ours	**1.0**	**2.41 ± 1.93**	12.14 ± 7.97	6.92 ± 4.05	3.23 ± 1.17

resulting in suboptimal success rates and efficiency, as indicated in Table 1. The search-based policy is notably slower, taking approximately 15.65 seconds-four times the duration required by learning-based approaches. The low success rate is due to the bumps and tight clearance, which makes it difficult for robots to detect holes and tend to get stuck on uneven surfaces. Vanilla RL, lacking the base policy's guidance, struggles with tasks that involve small-clearance assembly, maintaining a consistently low Cumulative Reward (CR), and demonstrating the difficulty of hole location and insertion via exploration alone. Approaches that rely solely on visual RL face challenges in task learning when vision is the only state space input. Although these methods can complete the task through exploration and achieve a moderate CR, they fall short in the error curriculum, primarily due to the difficulty in extracting precise positional information from pixel data through learning. The keypoint visual representation-based fusion method matches our framework's efficiency and assembly performance. However, our method offers the advantage of not requiring the intricate design and training processes associated with keypoint visual representation.

In summary, our proposed framework not only demonstrates superior performance in terms of efficiency and success rate but also simplifies the implementation process by eliminating the need for complex feature engineering inherent in other methods.

5.1.4 Ablation study

To demonstrate the efficacy of our proposed framework's components, we conducted an ablation study during both the training and deployment stages in the gear assembly task.

Evaluating robustness to position error with multimodal sensory input We assessed the impact of integrating multiple sensory modalities and temporal data on policy robustness. Three observation types were compared: (1) Proprioceptive-Tactile (Proprio-Tactile), which includes the robot's current 6-dimensional position and Euler angles along with force/torque data, which were encoded using a Multilayer Perceptron (MLP); (2) Temporal Proprioceptive-Tactile (Time Series P-T), which extends the Proprio-Tactile data over the last 32 time steps; and (3) Temporal Proprioceptive-Tactile with Vision (Time Series P-T-V), which adds current visual information to the Time Series P-T data. We trained each policy variant under an increasing error range and a constant exploration range of 10 mm. The final error range is compared after 200 episodes.

Assessing curriculum residual learning efficiency We examined the influence of fixed policy uncertainty on residual learning by considering position errors of 10, 20, and 30 millimeters. Specifically, we evaluated the curriculum's effectiveness in scenarios with significant uncertainty (using a 30 mm error as a case study) and compared it against traditional residual learning approaches. The adaptive curriculum adjusted the guidance error range by 0.5 mm and the exploration constraint based on the current success rate, within the bounds of [0.5, 0.7]. Three baseline curriculums were tested: (1) Curri-g2-c30, which starts with an error range of 2 mm and maintains a 30 mm exploration range; (2) Curri-g2-c10, which begins with a 2 mm error and a 10 mm exploration range, increasing both parameters once a 10 mm error range is reached; and (3) Curri-g2-c2, which scales both parameters starting from 2 mm.

Validating context generalization with task attention We trained policies using different Regions of Interest (ROIs) to validate the designed attention mechanism, as shown in Fig. 8, represented by ROI-800*800, ROI-300*300, ROI-200*200, and ROI-100*160. In the ROI-x-y, the x and y represent the height and width of the ROI. The policies trained with ROIs of ROI-800*800 and ROI-300*300 were tested in varying contexts: one with the task board placed at four different table locations (ROI-800-l and ROI-300-l), and another with varying backgrounds by introducing additional objects into the scene (ROI-800-b and ROI-300-b). The learned policies were evaluated on context generalization capabilities, measured by average cumulative reward.

5.1.5 Experiment result

The cumulative reward, success rate, error range, and exploration range in three repeated experiments are smoothed by convolution and estimated by Seaborn to show the learning process, where the estimated means and bootstrap confidence intervals of the variables are plotted as line and error bars.

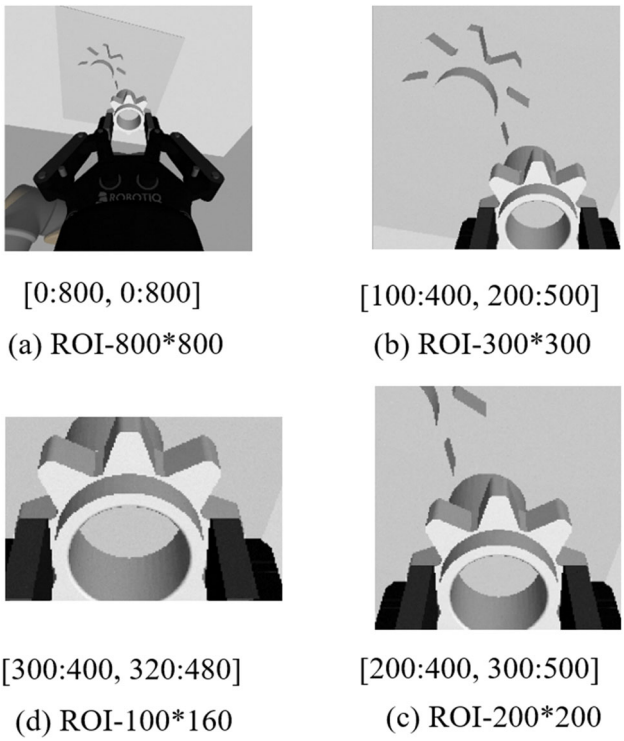


Fig. 8 Four kinds of ROI for task attention. (a) represents the full image without attention. (b) represents the view with suitable task attention. In (c) and (d), the restricted view makes important assembly features invisible

Cumulative reward is used to qualify the performance of trained policy in other contexts, which consists of success, completion time, and contact force.

Robustness with multiple sensors Multimodal sensory input enhances the robustness of residual policies to initial pose errors, as shown in Fig. 9. The residual policies with three types of observations can be learned to solve the gear insertion task with a random pose error of 2 mm. However, as the error increases, the widely used policy with only proprioception and wrist force/torque information performs worse, with an uncertainty tolerance of only 5 mm. The time series baseline behaves better but is limited. The full multimodal model with vision achieves the best performance, which is robust under 10 mm random error. A continuous increase in random error causes only a small decrease in cumulative reward and a rapid recovery. Using a fixed policy, the multimodal residual policy can be learned end-to-end with 100 episodes.

Curriculum learning for large position error Curriculum learning is effective for managing large position errors with the adaptive curriculum outperforming traditional residual learning, particularly in early learning stages, as shown in Fig. 10. Comparing the common residual learning in three types of errors, we observe that the large uncertainty of the initial policy will significantly reduce the effect of residual learning. With the error of 10 mm, the cumulative reward and success rate increase rapidly, reaching 70 and 0.9 with 300 episodes. The error bars show that the learning performance fluctuates slightly. However, when the error increases to 20 mm, the learning performance decreases, with the cumulative reward and success rate reaching 20 and 0.6 after 300 episodes. And the learning performance fluctuates significantly even without convergence. The larger errors, 30mm, cause further performance degradation. The remaining learning benefits from the guidance of the gradually increasing difficulty. Compared with the 30mm-nocurri, curri-g2-c30 increased the uncertainty from 2mm with a

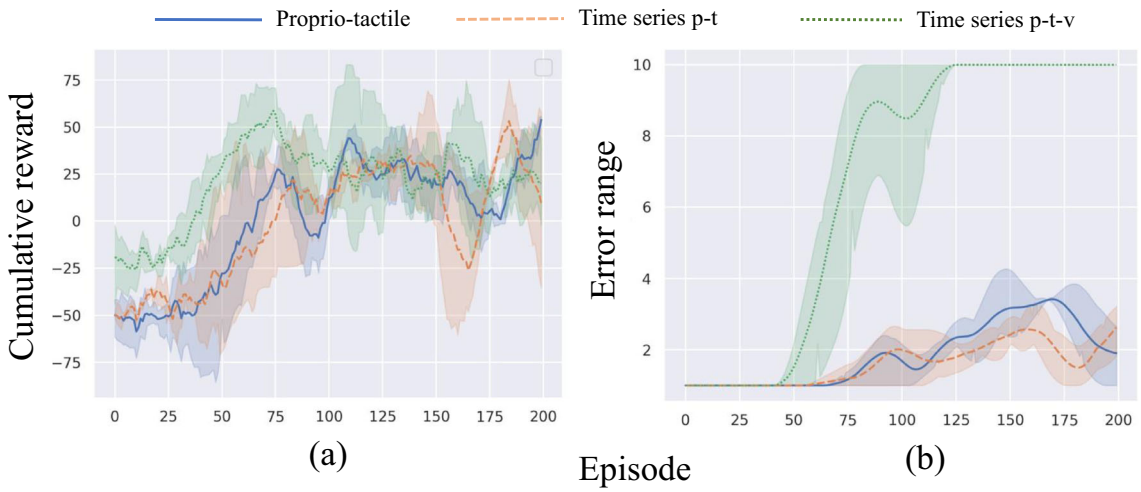


Fig. 9 Learning curves of the policies with three kinds of observations. (a) is the cumulative reward in each episode. (b) is the random error range added to the hand-designed policy

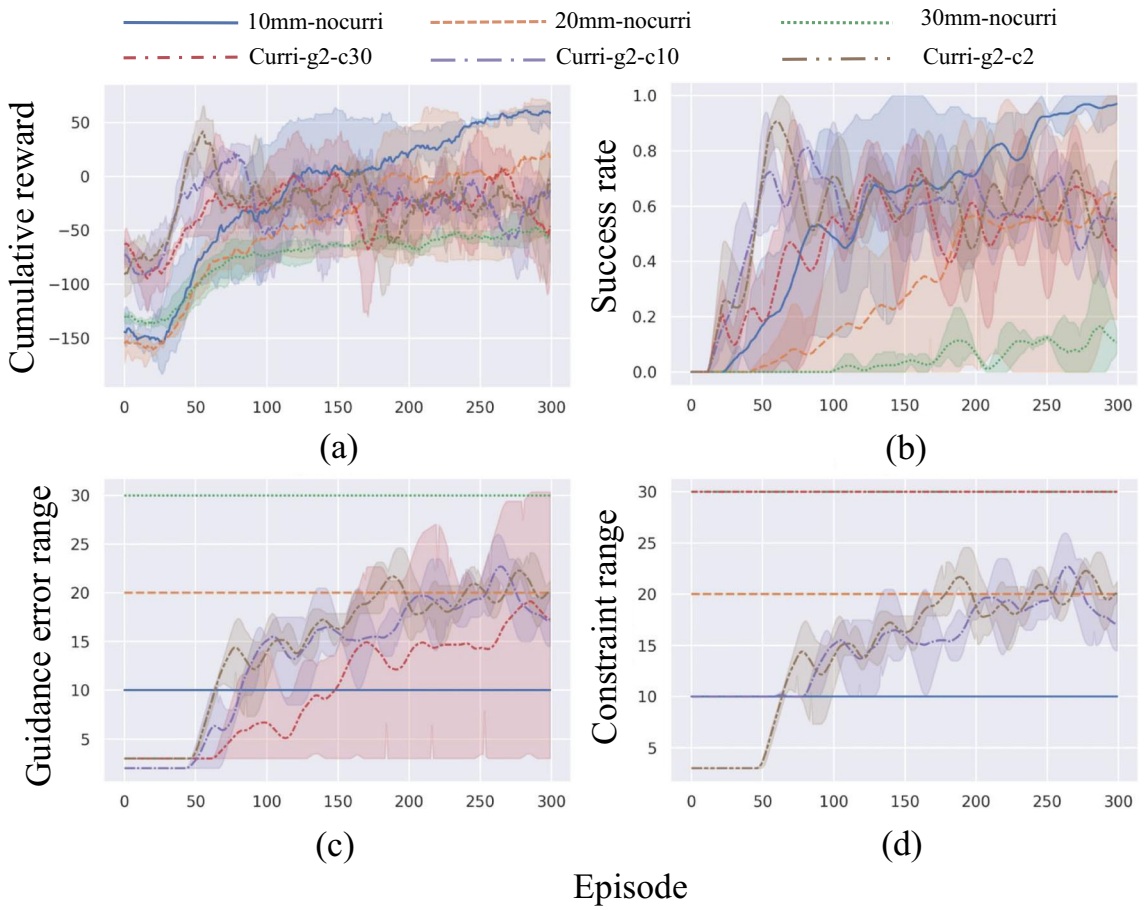


Fig. 10 Learning curves of the policies with different curriculums. (a) is the cumulative reward in each episode. (b) is the success rate during the last 15 episodes. (c) is the guidance error range of the fixed policy. (d) is the constrained exploration range of the fixed policy

constant constraint of 30mm according to the increasing success rate, as shown in Fig. 10 (c) and (d). More precise guidance can accelerate the cumulative reward and success rate. curri-g2-c10, which starts the constraint curriculum from 10 mm, further accelerates the learning. However, a stronger constraint, curri-g2-c2, had similar behavior. Comparing curri-g2-c10 and 20mm-nocurri, experiments show that the former achieves better and more stable cumulative reward and success rate in the early stages, but similar final performance.

Context generalization with task attention Appropriate task attention is crucial for both learning efficiency and context generalization. Policies trained with suitable ROIs demonstrate robust performance across different backgrounds and locations, while those without proper attention show significant performance degradation (Figs. 11 and 12). Task attention that is not well designed significantly affects the convergence and final performance. For the attention ROI-100*160, which does not cover the main task features, the guidance curriculum reaches only 7 mm, similar to no vision. ROI-200*200, which covers more features but not all in maximum error, increases the robustness in a limited uncertainty range. ROI-300*300 and ROI-800*800, which cover enough features, are robust in a larger range and more efficient in the learning process. In particular, ROI-800*800, which includes more background, has the fastest convergence speed. However, the redundant environmental information will prevent context generalization due to false feature dependence. The average cumulative reward of the policy with attention changes slightly, from 61.19 to 54.73 in different backgrounds and from 69.23 to 46.46 in different locations. However, the baseline without task attention decreases significantly, from 61.12 to -23.06 in different backgrounds and from 59.51 to -14.65 in different locations. Furthermore, the decrease in performance is related to the degree of contextual variability. For backgrounds, the magnitude of reward attenuation is related to the size of the object in the visual field. For locations, the magnitude of reward attenuation is related to the distance between training and test locations.

In summary, these results underscore the importance of multimodal sensory input, adaptive curriculums, and task

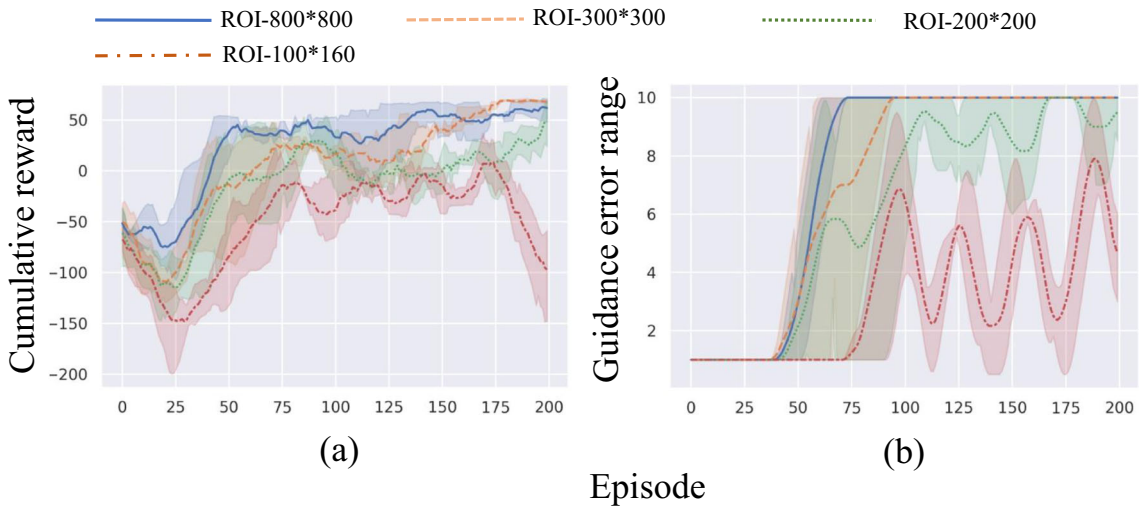


Fig. 11 Learning curves of the policies with four kinds of ROI. (a) is the cumulative reward of each episode. (b) is the guidance error range of the fixed policy

attention in enhancing the robustness and generalizability of policies for robotic tasks.

5.2 Real experiment for the comprehensive evaluation

5.2.1 Experiment setup

In this real-world evaluation, we focus on peg-in-hole and gear-insertion tasks to assess the sample efficiency and context generalization capabilities of our approach. Similar to our simulation experiments, we utilize a 3D-printed task board secured to a table and employ a UR5 robot arm equipped with a two-finger gripper for task execution. The setup includes a wrist-mounted force/torque sensor and a USB camera for tactile and visual data acquisition. After adjusting the Region of Interest (ROI) to the actual camera's field of view and using the goal pose as the original point for task-centric observation, our learning strategy integrates three types of sensory modalities with a fusion model to learn the residual uncertainties. Our adaptive curriculum based on the residual learning framework controls the error and exploration range, targeting a success rate between 0.5 and 0.7, with adjustments made in 0.5 mm increments.

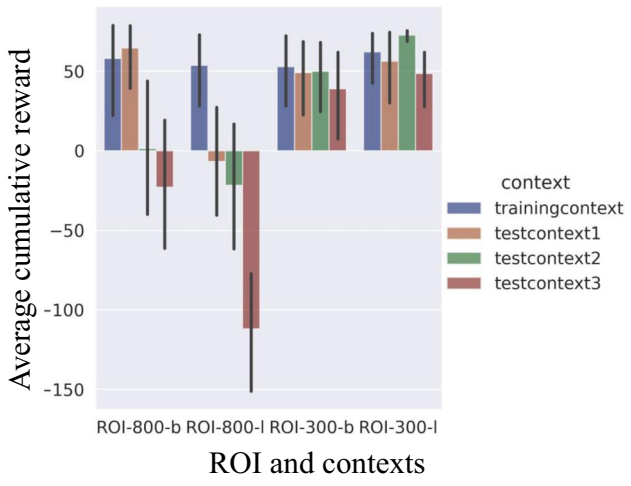


Fig. 12 Average cumulative reward of policies with different task attention (ROI) in different backgrounds and locations

This work directly trains the policy on a real robot for 300 episodes. Each episode encompasses the full operation cycle-grasp, transfer, and assembly, as depicted in Fig. 13 (a) and (b). The robot autonomously executes the assembly phase using the learned residual and a hand-designed policy, while other operation parts rely solely on the hand-designed policy. Post-training, we assess the policy across different contexts to gauge its reliability and generalization capabilities by placing task boards at different locations with fixtures, where positions are determined by demonstration as shown in Fig. 13 (c) and (d). The success rate (SR), completion time (CT), and cumulative reward (CR) are measured over 20 episodes using a trajectory that simulates a series of positional errors.

Understanding why an input leads to a particular output is critical to both safety and performance when using the models in a real robotic system. It explains how system inputs trigger specific responses and why a model may not generalize to new situations. Specifically, it provides a deeper understanding of how data-driven models work and allows us to thoroughly evaluate and improve them according to the task. In this work, the input and output are visualized to explain the behavior. In addition, important regions of the

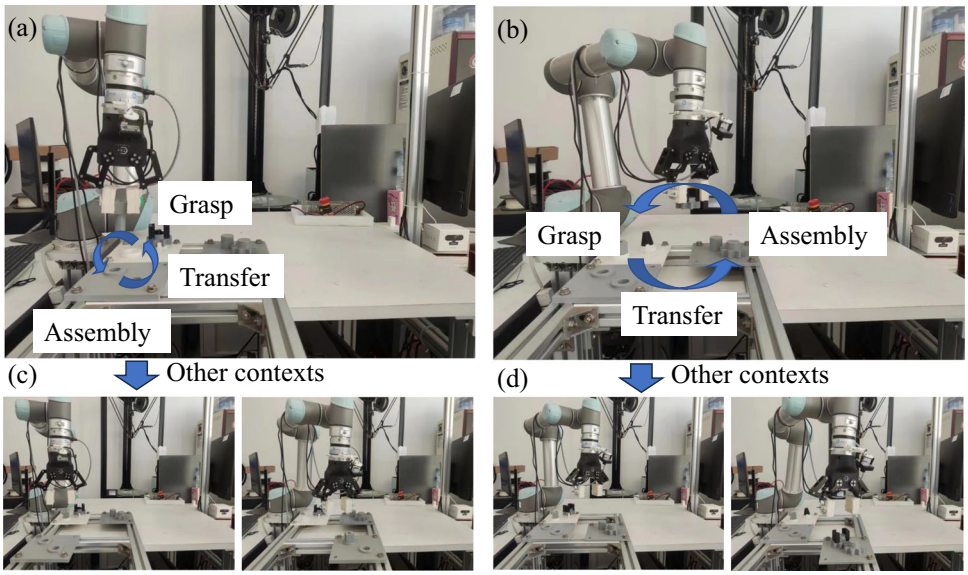


Fig. 13 Experimental platform of the precision assembly system. (a) and (b) are the manipulator and tasks in the training context. (c) and (d) are the tasks in the test context 1 and 2

image corresponding to each decision of interest are visualized in high-resolution detail.

5.2.2 Experiment result

The learning curves (Fig. 14) illustrate the sample efficiency of our method on the real robot, highlighting the slight reduction due to real-world noise factors. For example, in the gear task, the policy overcomes dynamic uncertainties within the first 50 episodes, achieving a success rate of 0.7. The curriculum's dynamic adaptation is evident as the error margin and success rate fluctuate, ultimately reaching a 20 mm error range after 300 episodes, or 3.5 hours, confirming the efficacy of multimodal policies and curriculum residual learning in precise insertion tasks despite significant pose errors. The similar learning efficiency in peg-in-hole and gear insertion tasks with different geometries shows the stable capability of our learning method.

Table 2 and Fig. 15 present the performance of learned policies across different contexts and error margins, demonstrating an admirable success rate of over 0.9 for tasks with guidance errors up to 15 mm. This stability across contexts underscores the robustness of our RL assembly strategy, facilitated by focused visual and tactile information. The comparison of learned policy with different geometries shows that the peg-in-hole task with simpler geometries has less completion time than gear insertion. Video of experiments is shown in https://github.com/WangChuang-163/Task-attention-based-multimodal-residual-RL/blob/main/README.md

Figure 16 visualizes the behavior over one episode, mapping observations and actions to a normalized range. The peg-insertion task has three phases: a search phase before contact (Yellow), a search phase after initial contact (Red), and an insertion phase (Green). The task is solved in about 30 steps, where the state of the image, the relative pose and contact force, and the action of the desired force F_d have a clear response to the phases. Before contact (Yellow), the desired force is generated based on visual servo for rough alignment. For errors in the x-negative and y-positive directions, the agent generates the desired force in the x-negative and y-positive directions. The learned important regions of the image, apart from the grasped gears, are initially the shaft for larger deviations, and then more attention is paid to the gears of the task board when it is obscured, which explains why a model can generalize to new situations. After the first contact with the surface (Red), F_d is drastically reduced in the x and y directions and the contact force is maintained in the z direction, thus improving the local search for the fine alignment. Then, when the pin is properly aligned (Green), F_d is increased to apply force to insert the pin against the friction of insertion and to complete the task faster.

6 Discussion

This research introduces a compact model that integrates visual and force information through curriculum-based residual reinforcement learning (RL), drawing inspiration from human perceptual and learning mechanisms in manipulation tasks. Our approach demonstrates the capability of a simple initial policy to be enhanced by a residual policy that compensates for contact dynamics and positional inaccuracies, facilitating efficient learning and transferability across

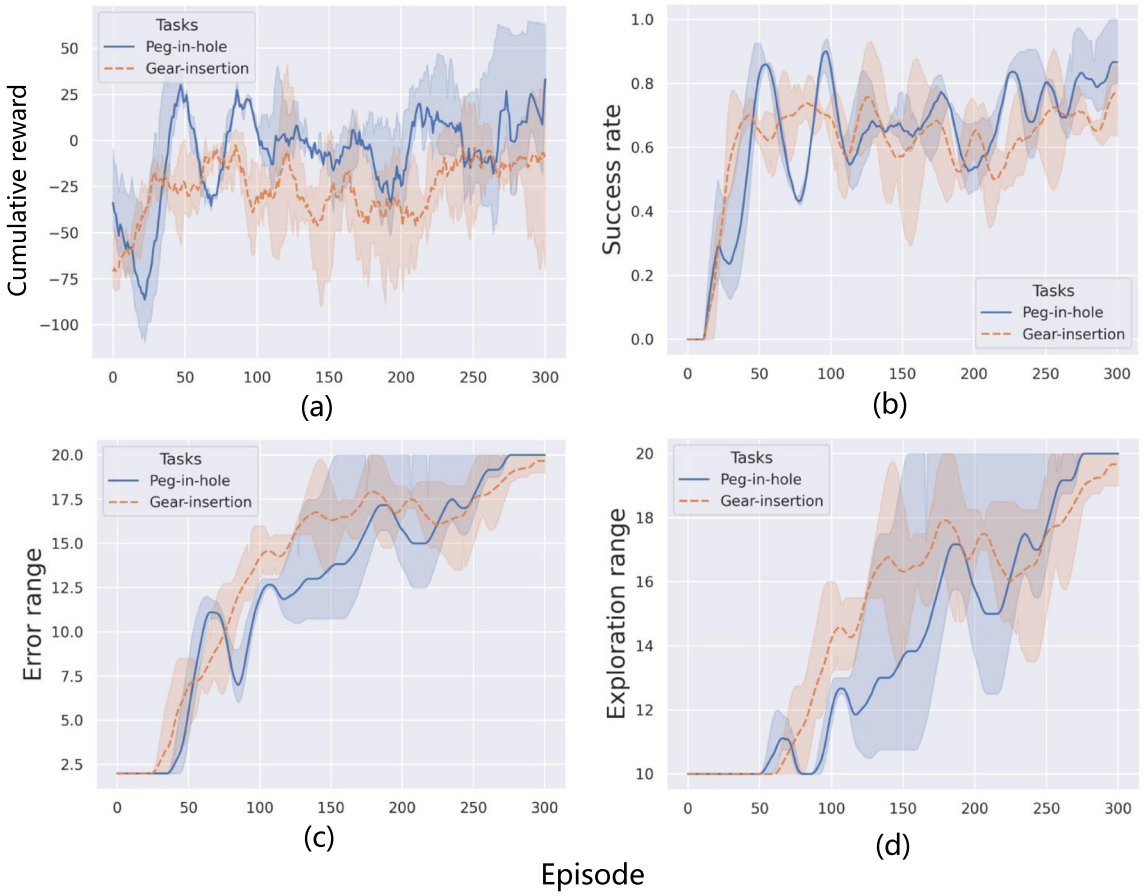


Fig. 14 Learning curves of the proposed method on the real robot. (a) is the cumulative reward in each episode. (b) is the success rate during the last 15 episodes. (c) is the guidance error range of the fixed policy. (d) is the constrained exploration range of the fixed policy

Table 2 The performance of learned policies

	Context	Perf.	Error			
			5 mm	10 mm	15 mm	20 mm
Peg-in-hole	Training	SRI	1.00	1.00	1.00	0.80
		COT	5.81*s*	6.11*s*	7.13*s*	8.08*s*
		ACR	78.37 ± 5.86	74.49 ± 6.65	65.53 ± 9.80	53.49 ± 18.9
	Unseen1	SRI	1.00	1.00	0.90	0.75
		ACR	80.25 ± 6.47	74.38 ± 9.36	66.11 ± 10.49	53.47 ± 20.98
	Unseen2	SRI	1.00	1.00	1.00	0.80
		ACR	79.07 ± 5.56	75.78 ± 7.88	67.20 ± 13.92	53.59 ± 19.99
Gear-insertion	Training	SRI	1.00	1.00	0.95	0.70
		COT	6.91*s*	7.04*s*	7.45*s*	7.75*s*
		ACR	63.84 ± 5.99	61.52 ± 7.11	54.92 ± 9.11	53.58 ± 10.26
	Unseen1	SRI	1.00	1.00	0.90	0.85
		ACR	57.17 ± 10.99	58.45 ± 7.60	52.48 ± 9.30	45.63 ± 15.77
	Unseen2	SRI	1.00	1.00	0.90	0.85
		ACR	61.53 ± 11.70	62.57 ± 8.66	48.33 ± 17.33	46.63 ± 17.51

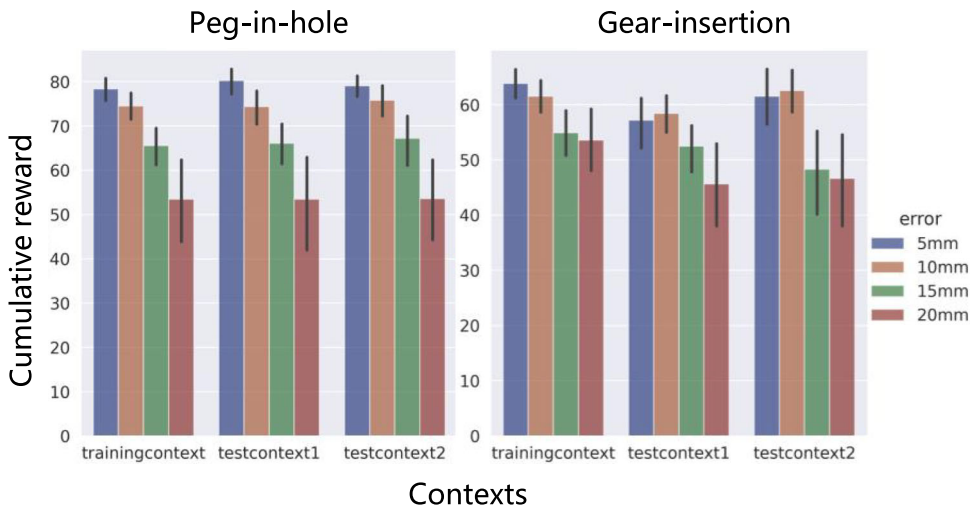


Fig. 15 Average cumulative reward of policies with different task attention (ROI) in different backgrounds and locations

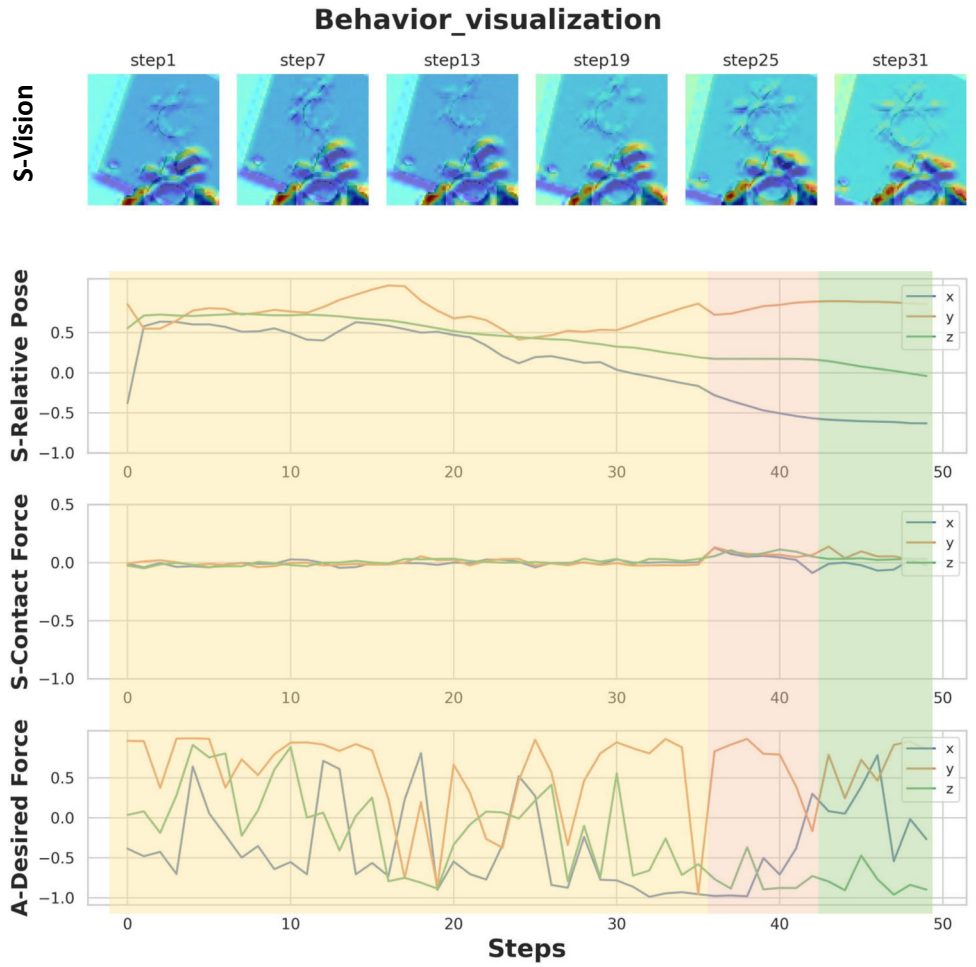


Fig. 16 Behavior visualization. The behavior over one episode is visualized, mapping observations and actions to a normalized range. The three phases are discussed, including a search phase before contact (Yellow), a search phase after initial contact (Red), and an insertion phase (Green)

contexts. The potential of residual RL for improving sampling efficiency and ensuring safer interactions has been corroborated by [8]. The integration of multimodal information is crucial for executing precise tasks with large random errors, as also noted by [5]. Our work expands on these by employing multimodal information to inform residual policy learning. Utilizing the structured guidance of a fixed policy, we have simplified the exploration process, resulting in a more rewarding learning transition and enabling the efficient acquisition of a model with multimodal fusion tailored for error compensation. In scenarios necessitating precise detection and secure contact as discussed in [4], we observed that the process is inherently slow, with increased pose errors amplifying the cost steps of the spiral search and visual servo. Mechanisms of leveraging force and visual features for alignment and search policies require design and tuning efforts for specific tasks, where performance relies on expert experience. The process of designing and tuning parameters is also inherently an interactive learning process. By interactively learning perception and control with the end-to-end policy, we circumvent the need for extensive design efforts and facilitate the extraction of salient features to develop a coherent policy.

The significance of a well-structured curriculum for mastering precise tasks with large errors has been echoed in the work of [31]. Our approach based on residual learning offers more effective guidance, centering the exploration near the actual target location and imposing stringent constraints to minimize the search space and incremental steps. The optimal initial policy facilitates exploration by enhancing the likelihood of successful search and insertion amidst random explorations. Nonetheless, the introduction of error randomization within realistic bounds is imperative to the development of a robust policy. Our adaptive curriculum is meticulously designed to modulate task difficulty by varying the error and exploration range, thereby optimizing residual learning in the face of substantial uncertainties. Furthermore, the attention mechanism can interpret generalization learning as agents selectively focusing on invariant features while ignoring variable features [25]. This mechanism is particularly critical when dealing with the problem that a narrow field of vision may compromise observability and diminish robustness to uncertainties, while a wide field of vision may introduce irrelevant background information, adversely affecting performance in different contexts. By employing an initial policy, the agent is required to explore and learn within a confined domain. Task attention, informed by prior knowledge, assists the agent in maintaining focus on the essential elements of the task, facilitating generalization across varying contexts without the need for data augmentation or domain randomization. Learning coarse pose information directly from pixel information shows similar sample efficiency to that of prior visual representations [31], which are mainly due to high-quality samples guided by the initial policy.

Our experimental validation using peg-in-hole and gear assembly tasks substantiates the efficiency of our policy learning and its successful transferability to other contexts. This suggests that our approach is theoretically capable of satisfying the demands of high-precision assembly tasks within flexible manufacturing environments. Moreover, we posit that this method holds promise for application in simulation-to-reality transfers or meta-reinforcement learning to enhance sample efficiency and generalization capabilities.

Notwithstanding the progress made, our approach is subject to several noteworthy limitations that merit further discussion. For instance, the attention Region of Interest (ROI) must exclude background distractions while capturing adequate task features to strike a balance between feature invariance and observability. The restricted robustness and context generalization can be attributed to the limitations inherent in hand-designed ROIs. Although curriculum residual learning stabilizes the learning process and enhances sampling efficiency, the number of episodes needed to learn a stable policy is directly proportional to the level of uncertainty. Learning a robust policy requires data distribution of the whole error space. These findings highlight the need to balance the engineering efforts expended on manual controller design against the learning costs. Our method trains the policy within a structured environment, where the reward and success criteria are predicated on distance metrics. Designing conditions for initiating and concluding episodes becomes challenging when the agent operates in an unstructured environment.

7 Conclusion and future work

The proposed method integrates a model-based policy and residual learning for skill formulation, facilitating learning and reconfiguration for context generalization. With a simple hand-designed initial policy, task attention-based multimodal observation and curriculum residual learning enable efficient learning of the robust residual policy. The state representation to encapsulate only the task-relevant information and compact neural network for fusion and stochastic policy is designed to enhance the robustness and context generalization, reducing reliance on high-precision initial strategies. Adaptive curriculum generation in residual RL for guidance and constraint is designed to refine the learning efficiency of the robust policy, mapping directly from the raw sensory inputs to residual actions. Adjusting the waypoints of the initial policy permits rapid adaptation to novel contexts, which has been validated through our experiments with high-precision peg-in-hole and gear assembly tasks requiring tight

clearances and large pose errors. In future work, we plan to incorporate stable pre- and post-conditions recognition with multimodal data from the agent learning process. Further investigation into the integration of a vision model with our task attention-based multimodal residual policy is expected to yield improvements in robustness and adaptability, particularly in unstructured environments. In addition, research into sim-to-real for the task attention-based multimodal residual policy is needed to further improve the training cost.

Acknowledgements This document is the results of the National Key Research and Development Program of China (Grant No. 2021YFB3301400), the research project funded by the National Natural Science Foundation of China (Grant No. 52075177 and No.52305105), Research Foundation of Guangdong Province (Grant No. 2019A050505001 and 2018KZDXM002), Guangzhou Research Foundation (Grant No. 202002030324 and 201903010028), Zhongshan Research Foundation (Grant No. 2020B2020).

Author Contributions Chuang Wang: Conceptualization, Methodology, Software, Writing - original draft. Ze Lin: Software, Investigation. Biao Liu: Visualization, Investigation. Chupeng Su: Writing - review and editing. Gang Chen: Writing - review and editing, Supervision, Funding acquisition. Longhan Xie: Supervision, Funding acquisition. All authors read and approved the final manuscript.

Data availability and access The data that support the findings of this study are available from the corresponding author upon reasonable request.

Declarations

Conflicts of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Jiang J, Yao L, Huang Z, Yu G, Wang L, Bi Z (2022) The state of the art of search strategies in robotic assembly. J Industrial Information Integration 26:100259
2. Li J, Pang D, Zheng Y, Guan X, Le X (2022) A flexible manufacturing assembly system with deep reinforcement learning. Control Eng Prac
3. De Gregorio D, Zanella R, Palli G, Pirozzi S, Melchiorri C (2018) Integration of robotic vision and tactile sensing for wire-terminal insertion tasks. IEEE Trans Automation Sci Eng 16(2):585–598
4. Zhao D-Y (2021) Sun F Wang Z, Zhou Q: A novel accurate positioning method for object pose estimation in robotic manipulation based on vision and tactile sensors. Int J Adv Manufac Technol 116:2999–3010
5. Lee MA (2020) Zhu Y, Zachares P, Tan M, Srinivasan K, Savarese S, Fei-Fei L, Garg A, Bohg J: Making sense of vision and touch: Learning multimodal representations for contact-rich tasks. IEEE Trans Robotics 36(3):582–596
6. Hou Z, Yang W, Chen R, Feng P, Xu J (2022) A hierarchical compliance-based contextual policy search for robotic manipulation tasks with multiple objectives. IEEE Trans Industrial Inform 19(4):5444–5455
7. Johannink T, Bahl S, Nair A, Luo J, Kumar A Loskyll M, Ojea JA, Solowjow E, Levine S (2019) Residual reinforcement learning for robot control. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 6023–6029
8. Kulkarni P, Kober J Babuka R, Santina CD (2021) Learning assembly tasks in a few minutes by combining impedance control and residual recurrent reinforcement learning. Adv Intell Syst 4
9. Hao P Lu T, Cui S, Wei J, Cai Y Wang S (2022) Meta-residual policy learning: Zero-trial robot skill adaptation via knowledge fusion. IEEE Robotics Automation Lett 1–1
10. Staessens T, Lefebvre T, Crevecoeur G (2022) Adaptive control of a mechatronic system using constrained residual reinforcement learning. IEEE Trans Industrial Electron 69(10):10447–10456
11. Shi Y, Chen Z Liu H, Riedel S, Gao C Feng Q, Deng J, Zhang J (2021) Proactive action visual residual reinforcement learning for contact-rich tasks using a torque-controlled robot. In: 2021 IEEE International conference on robotics and automation (ICRA), IEEE, pp 765–771
12. Schoettler G, Nair A, Luo J, Bahl S, Ojea JA, Solowjow E, Levine S (2020) Deep reinforcement learning for industrial insertion tasks with visual inputs and natural rewards. In: 2020 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 5548–5555
13. Song JM (2021) Chen Q Li Z: A peg-in-hole robot assembly system based on gauss mixture model. Robotics Comput Integr Manuf 67
14. Lin H-I (2020) Design of an intelligent robotic precise assembly system for rapid teaching and admittance control. Robotics Comput Integr Manuf 64:101946
15. Apolinarska AA, Pacher M, Li H, Cote N, Pastrana R, Gramazio F, Kohler M (2021) Robotic assembly of timber joints using reinforcement learning. Automation Construction, 103569
16. Inoue T, De Magistris G, Munawar A, Yokoya T, Tachibana R (2017) Deep reinforcement learning for high precision assembly tasks. In: 2017 IEEE/RSJ International conference on intelligent robots and systems (IROS), IEEE, pp 819–825
17. Xu J, Hou Z, Wang W, Xu B, Zhang K, Chen K (2019) Feedback deep deterministic policy gradient with fuzzy reward for robotic multiple peg-in-hole assembly tasks. IEEE Trans Industrial Inform 15:1658–1667
18. Ren T, Dong Y, Wu D, Chen K (2018) Learning-based variable compliance control for robotic assembly. J Mechan Robotics
19. Song R (2021) Li F Quan W, Yang X, Zhao J: Skill learning for robotic assembly based on visual perspectives and force sensing. Robotics Auton Syst 135:103651
20. Luo J, Solowjow E, Wen C, Ojea JA, Agogino AM, Tamar A, Abbeel P (2019) Reinforcement learning on variable impedance controller for high-precision robotic assembly. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 3080–3087
21. Beltran-Hernandez CC, Petit D, Ramirez-Alpizar IG, Nishi T, Kikuchi S, Matsubara T, Harada K (2020) Learning force control for contact-rich manipulation tasks with rigid position-controlled robots. IEEE Robotics Automation Lett 5(4):5709–5716
22. Bogdanovic M, Khadiv M, Righetti L (2019) Learning variable impedance control for contact sensitive tasks. IEEE Robotics and Automation Lett 5:6129–6136
23. Chen C, Zhang C, Pan Y-D (2023) Active compliance control of robot peg-in-hole assembly based on combined reinforcement learning. Appl Intell
24. Liu Q, Ji Z, Xu W, Liu Z, Yao B, Zhou Z (2023) Knowledge-guided robot learning on compliance control for robotic assembly task with predictive model. Expert Syst Appl 234:121037
25. Yasutomi AY (2023) Ichiwara H, Ito H, Mori H, Ogata T: Visual spatial attention and proprioceptive data-driven reinforcement learning for robust peg-in-hole task under variable conditions. IEEE Robotics Automation Lett 8:1834–1841

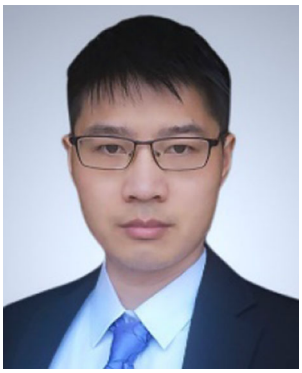
26. Xie L, Yu H, Zhao Y, Zhang H, Zhou Z, Wang M, Wang Y, Xiong R (2022) Learning to fill the seam by vision: Sub-millimeter peg-in-hole on unseen shapes in real world. In: 2022 International conference on robotics and automation (ICRA), pp 2982–2988
27. Shi Y, Yuan C, Tsitos AC, Cong L, Hadjar H, Chen Z, Zhang J-W (2023) A sim-to-real learning-based framework for contact-rich assembly by utilizing cyclegan and force control. IEEE Trans Cognitive Develop Syst 15:2144–2155
28. Zhang Z, Wang Y, Zhang Z, Wang L, Huang H, Cao Q (2024) A residual reinforcement learning method for robotic assembly using visual and force information. J Manuf Syst
29. Ahn K, Na M-W, Song J-B (2023) Robotic assembly strategy via reinforcement learning based on force and visual information. Robotics Auton Syst 164:104399
30. Chen W, Zeng C, Liang H, Sun F, Zhang J (2023) Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly. IEEE Trans Cybernet
31. Jin P, Lin Y, Song Y, Li T, Yang W (2023) Vision-force-fused curriculum learning for robotic contact-rich assembly tasks. Front Neurorobotics 17
32. Abu-Dakka FJ, Nemec B, Kramberger A, Buch AG (2014) Norbert: Solving peg-in-hole tasks by human demonstration and exception strategies. Ind Robot 41:575–584
33. Wang X, Chen Y, Zhu W (2021) A survey on curriculum learning. IEEE Trans Pattern Anal Mach Intell 44:4555–4576
34. Li X, Li J, Shi H (2023) A multi-agent reinforcement learning method with curriculum transfer for large-scale dynamic traffic signal control. Appl Intell 53:21433–21447
35. Cui F, Di H, Huang H, Ren H, Ouchi K, Liu Z, Xu J (2022) Multi-source inverse-curriculum-based training for low-resource dialogue generation. Appl Intell 53:13665–13676
36. Dong S, Jha D.K, Romeres D, Kim S, Nikovski D, Rodriguez A (2021) Tactile-rl for insertion: Generalization to objects of unknown geometry. In: 2021 IEEE International conference on robotics and automation (ICRA), 6437–6443
37. Luo J, Sushkov O, Pevceviciute R, Lian W, Su C, Vecerik M, Ye N, Schaal S, Scholz J (2021) Robust multi-modal policies for industrial assembly via reinforcement learning and demonstrations: A large-scale study. arXiv preprint arXiv:2103.11512
38. Hermann L, Argus M, Eitel A, Amiranashvili A, Burgard W, Brox T (2020) Adaptive curriculum generation from demonstrations for sim-to-real visuomotor control. In: 2020 IEEE International conference on robotics and automation (ICRA), IEEE, pp 6498–6505
39. Kirk R, Zhang A, Grefenstette E, Rocktaeschel T (2023) A survey of zero-shot generalisation in deep reinforcement learning. J Artif Intell Res 76:201–264
40. Ballou A, Reinke C, Alameda-Pineda X (2022) Variational meta reinforcement learning for social robotics. Appl Intell 53:27249–27268
41. Wang K, Kang B, Shao J, Feng J (2020) Improving generalization in reinforcement learning with mixture regularization. Adv Neural Inform Process Syst 33:7968–7978
42. Finn C, Tan X.Y, Duan Y, Darrell T, Levine S, Abbeel P (2015) Deep spatial autoencoders for visuomotor learning. In: 2016 IEEE International conference on robotics and automation (ICRA), 512–519
43. Zhou K, Guo C, Zhang H (2022) Improving indoor visual navigation generalization with scene priors and markov relational reasoning. Appl Intell 52:17600–17613
44. Huang X, Chen D, Guo Y, Jiang X, Liu Y (2023) Untangling multiple deformable linear objects in unknown quantities with complex backgrounds. IEEE Trans Automation Sci Eng
45. Sundaresan P, Grannen J, Thananjeyan B, Balakrishna A, Laskey M, Stone K, Gonzalez J, Goldberg K (2020) Learning rope manipulation policies using dense object descriptors trained on synthetic depth data. In: 2020 IEEE International conference on robotics and automation (ICRA), 9411–9418
46. Mnih V, Heess N, Graves A et al (2014) Recurrent models of visual attention. Adv Neural Inform Process Syst 27
47. Wang S, Fan Y, Jin S, Takyi-Aninakwa P, Fernandez C (2023) Improved anti-noise adaptive long short-term memory neural network modeling for the robust remaining useful life prediction of lithium-ion batteries. Reliability Eng Syst Safety, 108920. https://doi.org/10.1016/j.ress.2022.108920
48. Luo J, Solowjow E, Wen C, Ojea JA, Agogino AM, Tamar A, Abbeel P (2019) Reinforcement learning on variable impedance controller for high-precision robotic assembly. In: 2019 International conference on robotics and automation (ICRA), IEEE, pp 3080–3087
49. Lee D-H, Choi M-S, Park H, Jang G-R, Park J-H, Bae J-H (2022) Peg-in-hole assembly with dual-arm robot and dexterous robot hands. IEEE Robotics Automation Lett 7(4):8566–8573
50. Haugaard R, Langaa J, Sloth C, Buch A (2021) Fast robust peg-in-hole insertion with continuous visual servoing. In: Conference on Robot Learning, PMLR, pp 1696–1705
51. Stevsic S, Christen S, Hilliges O (2020) Learning to assemble: Estimating 6d poses for robotic object-object manipulation. IEEE Robotics Automation Lett 5(2):1159–1166

Chuang Wang received the B.S. degree in School of Mechanical and Power Engineering from Zhengzhou University, China, in 2017. He is currently pursuing the Ph.D. degree with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His research interests include robotic manipulation, compliance control, deep reinforcement learning, assembly robot.

Ze Lin is currently pursuing the B.S. degree in Intelligent Manufacturing Engineering with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His current research interests include robotic manipulation, machine vision and assembly robot.

Biao Liu received the B.S. degree in measurement and control technology and instrument from Liaoning Technical University, Liaoning, China, in 2015. And the M.S. degree in automation in control science and engineering from the Guangdong University of Technology, Guangdong, China, in 2019. He is currently pursuing the Ph.D. degree in the South China of Technology University, Guangdong, China. His current research interests include bipedal robot, robot control, exoskeleton.

Chupeng Su received the B.E. degree in Mechanical and Electronic Engineering from Zhongkai University of Agriculture and Engineering, China, in 2017. He is currently pursuing the Ph.D. degree with the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology (SCUT), China. His research interests include the production scheduling, deep reinforcement learning, as well as smart factory.



Gang Chen received the bachelor's and master's degrees in mechanical engineering from Shanghai Jiao Tong University, Shanghai, China, in 2012 and 2015, respectively and his Ph.D. degree in mechanical and aerospace engineering from the University of California, Davis, Davis, CA, in 2020. He was a research fellow with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore from 2020 to 2021. He is currently an associate professor at the Shien-Ming Wu School of Intelligent Engineering, South China University of Technology, China. His research interests include machine learning, formal methods, control, signal processing and fault diagnosis.

Longhan Xie received a B.S. degree and a M.S. degree in mechanical engineering in 2002 and 2005, respectively, from Zhejiang University. He received a Ph.D. degree in mechanical and automation engineering in 2010 from the Chinese University of Hong Kong. From 2010 to 2016, he was an Assistant Professor and Associate Professor in the School of Mechanical and Automotive Engineering at the South China University of Technology. Since 2017, he has been a professor in Shien-Ming Wu School of Intelligent Engineering at the same university. His research interests include biomedical engineering and robotics. He is a member of ASME and IEEE.