Toward high-precision robotic assembly in large workspaces using multimodal reinforcement learning

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

Toward high-precision robotic assembly in large workspaces using multimodal reinforcement learning

Sirui Ren[1], Chao Zeng[2], Chenguang Yang[2], Ning Wang[3*]

[1] College of Automation Science and Engineering,

South China University of Technology, Guangzhou, China

[2] School of Computer Science and Informatics,

University of Liverpool, Liverpool, L69 3BX, U.K.

[3] School of Computing and Digital Technologies,

Sheffield Hallam University, Sheffield, S1 2NU, U.K.

Email: wang.ning@ieee.org

Abstract

Purpose–This study aims to address the challenges of complex contact dynamics, structural constraints, and perceptual uncertainty in robotic peg-in-hole assembly tasks, particularly under large workspace and high-precision requirements.

Design/methodology/approach–We propose a reinforcement learning framework that integrates multimodal perception, self-supervised representation modeling, and hybrid control mechanisms. The framework takes visual images, proprioceptive states, and target pose information as inputs. A self-supervised pretraining phase jointly optimizes image reconstruction and forward prediction to learn structurally aware and temporally consistent latent state representations. Furthermore, a Spectrum Random Masking (SRM) technique introduces frequency-domain perturbations to the visual modality, encouraging stable spectral feature learning and enhancing perception robustness for sim-to-real transfer. During execution, an adaptive impedance control mode is activated when excessive contact force is detected, mitigating insertion impacts and jamming.

Findings–Real-world experiments in extended operational spaces and across diverse peg-hole geometries demonstrate that the proposed method achieves millimeter-level accuracy, high success rates in seen configurations, and strong zero-shot generalization to unseen shapes. These results validate the effectiveness of the framework in ensuring robust and precise robotic assembly across large and variable workspaces.

Originality–This work presents a novel reinforcement learning framework that integrates multimodal perception, spectral feature learning, and hybrid control switching to tackle the challenges of high-precision peg-in-hole assembly. By introducing SRM for robust visual representation and combining it with adaptive impedance control, the framework offers a distinctive solution that enhances sim-to-real transfer and generalization in complex assembly scenarios.

Index Terms

Multimodal Reinforcement Learning, Robotic Assembly, Adaptive Impedance Control, Sim-to-Real Transfer

I. Introduction

Robotic assembly tasks represent a critical aspect of modern manufacturing, spanning diverse applications from simple shaft assembly to complex circuit board integration(Li *et al.* 2021). However, traditional robot programming methods often rely on tedious manual teaching and predefined fixed programs, which are insufficient to meet the increasing demands for automation and intelligent manufacturing. With the rapid advancements in artificial intelligence (AI) technologies, reinforcement learning (RL) has emerged as a promising autonomous learning approach, providing effective solutions to tackle the complex challenges inherent in robotic assembly tasks(Luo *et al.* 2019; Leyendecker *et al.* 2022).

However, RL policies are often constrained by the training environment, making it difficult to directly transfer the learned strategies from simulation to real-world systems, leading to performance degradation during sim-to-real transfer(Valassakis *et al.* 2020). Moreover, most methods rely solely on single-modal state representations, which are insufficient to accurately model the local structural features and real-time contact states during the insertion phase of peg-in-hole tasks, particularly under conditions where the target hole is small or even invisible within the experimental scope, or the hole geometry is complex. These limitations severely restrict the generalization ability of RL-based strategies(Y. Chen, Geng, *et al.* 2023).

To address these issues, this paper proposes a multimodal reinforcement learning strategy that integrates visual information, proprioceptive feedback, and target pose to guide the robot in accomplishing insertion tasks involving diverse structural variations. To overcome the generalization bottleneck of vision-based policies during sim-to-real transfer, we introduce a frequency-domain augmentation mechanism called Spectrum Random Masking (SRM)(Huang *et al.* 2022), which applies spectral perturbations to the visual modality during training in order to mitigate overfitting to specific image distributions and enhance visual robustness during real-world deployment. Furthermore, after the policy-guided contact occurs, the system incorporates a structure-decoupled adaptive impedance controller, which dynamically adjusts the compliance of the end-effector based on real-time force feedback, effectively addressing issues such as contact impact, boundary interference, and insertion failures. The main contributions of this paper are as follows:

1) We propose a multimodal reinforcement learning framework that jointly optimizes a visual encoder through image reconstruction and forward prediction. By generating latent states with structural awareness and temporal consistency, this framework enables the robot to achieve millimeter-level assembly precision even within large workspaces.

2) We employ a structure-decoupled adaptive impedance control mechanism to dynamically adjust the stiffness and damping of the end-effector based on real-time force feedback. The mechanism operates without relying on simulated force information during training, thereby avoiding complex virtual force modeling and enabling streamlined sim-to-real transfer without additional fine-tuning, which enhances the system's adaptability to contact disturbances.

3) We demonstrate that the proposed framework integrates frequency-domain augmentation, multimodal reinforcement learning, and adaptive force control to achieve robust generalization across unseen hole geometries. Experimental results indicate that this hybrid approach effectively compensates for execution errors and contact disturbances, ensuring stable insertion performance in real-world scenarios.

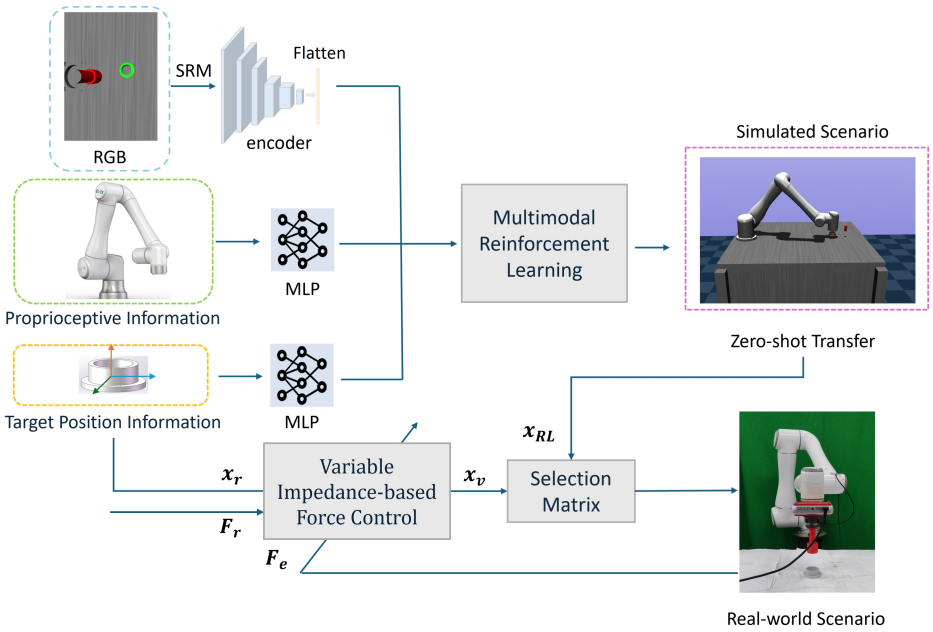

Fig. 1. The overview of our method

II. Related Work

In recent years, reinforcement learning (RL) has emerged as a promising approach for solving complex contact-rich tasks such as robotic insertion and object stacking, due to its ability to optimize policies through interaction without requiring explicit modeling of environment dynamics(Elguea-Aguinaco *et al.* 2023). In precision manipulation scenarios like peg-in-hole assembly, RL has demonstrated strong generalization capabilities and task adaptability, and has been widely applied to high-accuracy, constraint-sensitive robotic control tasks(Lee, Zhu, Zachares, *et al.* 2020; Lee, Zhu, Srinivasan, *et al.* 2019; W. Chen *et al.* 2023; Wu *et al.* 2023).

Several prior works have adopted end-to-end reinforcement learning with visual inputs as the primary modality, enabling autonomous perception and control in robotic assembly tasks (Yadav *et al.* 2024; Liu *et al.* 2023; Xie *et al.* 2022). In addition, demonstration-based approaches have been explored to enhance policy learning by incorporating demonstration trajectories and proprioceptive states, leading to improved task performance (Xiao *et al.* 2024). However, single-modal approaches suffer from inherent limitations: vision-based policies are sensitive to occlusions and viewpoint changes, while proprioception-based policies lack structural and geometric information, often resulting in failures in complex assembly scenarios. To address these limitations, multimodal strategies have been increasingly investigated. Some studies combine visual observations and force feedback within inverse reinforcement learning

frameworks, achieving strong generalization at the cost of manual demonstrations and complex data acquisition (Spector *et al.* 2022; Song *et al.* 2021). More recently, residual reinforcement learning methods that fuse visual and force perception have been proposed to improve insertion accuracy. However, their generalization across diverse geometries remains limited (Zhang *et al.* 2024). In light of these challenges, we propose a multimodal reinforcement learning framework that requires no manual annotations and ensures both structural awareness and temporal consistency. The proposed framework integrates visual encodings, proprioceptive states, and target poses, together with a self-supervised pretraining stage that jointly optimizes visual representations through image reconstruction and forward prediction.

Although multimodal perception strategies exhibit strong adaptability and generalization in simulation, transferring them stably to real-world robotic systems remains highly challenging (Karnan *et al.* 2020). Consequently, significant research efforts have focused on the sim-to-real transfer problem. Existing studies have explored domain adaptation techniques to map simulation and real-world observations into a shared latent space (Bousmalis *et al.* 2018; Jeong *et al.* 2020), while domain randomization has been widely adopted to facilitate robust and safe transfer across domains (Tobin *et al.* 2017). In addition, real-time "Digital Twin" environments have been leveraged to enhance reinforcement learning through imitation-based strategies (Y. Chen, Zeng, *et al.* 2023; Ju *et al.* 2023). To address the specific requirements of robotic assembly, we propose a reinforcement learning framework that integrates frequency-domain augmentation and structure-aware state modeling. Specifically, SRM is applied to the visual modality to introduce frequency perturbations, encouraging robust feature learning and improving generalization (Huang *et al.* 2022). In parallel, a self-supervised scheme jointly optimizes the visual encoder through image reconstruction and forward prediction, yielding latent states with structural awareness and temporal consistency. During execution, an adaptive impedance control module based on force feedback effectively manages contact uncertainty, enabling zero-shot transfer and stable assembly in complex real-world scenarios.

III. Method

We propose a multimodal RL framework for robust peg-in-hole assembly, which is capable of handling scenarios involving various hole geometries. In this section, we provide a detailed description of the implementation of the proposed method.

A. Reinforcement learning

1) Reinforcement Learning Algorithm with Multimodal Fusion: The Deep Deterministic Policy Gradient (DDPG) algorithm is adopted as the foundational reinforcement learning framework for the assembly task due to its suitability for control problems with high-dimensional and continuous action spaces. Within the DDPG algorithm, an actor network learns a policy to maximize the Q-value output of a critic network, with its parameters updated via gradient ascent. The gradient of the policy network is given by(Lillicrap *et al.* 2015):

$$\nabla_{\theta^\pi} J(\theta^\pi) = \mathbb{E}_{s_t \sim \mathcal{D}} \left[\nabla_a Q(s_t, a) \Big|_{a=\pi(s_t)} \nabla_{\theta^\pi} \pi(s_t|\theta^\pi) \right] \tag{1}$$

where θ^{π} represents the policy network parameters, s_t is the state, $\mathcal{D}$ is the experience pool, $\pi(s_t|\theta^{\pi})$ is the action chosen by the policy, and $Q(s_t, \pi(s_t|\theta^{\pi}))$ is the Q-value output. Concurrently, the critic network parameters are optimized by minimizing the Temporal Difference (TD) error:

$$L_{TD} = \frac{1}{N}\sum_{t}\left(Q(s_t, a_t|\theta^{Q}) - \left(r_t + \gamma Q_{\text{target}}(s_{t+1}, \pi_{\text{target}}(s_{t+1}|\theta^{\pi}_{\text{target}})|\theta^{Q}_{\text{target}})\right)\right)^2 \tag{2}$$

where N is the batch size, $Q(s_t, a_t|\theta^{Q})$ is the Q-value for the current state and action, r_t is the reward, γ is the discount factor, and Q_{target} is the Q-value from the target network. To enhance sample efficiency and generalization, we incorporate the Hindsight Experience Replay (HER) mechanism. HER reinterprets failed trajectories by relabeling the original goal g with an achieved future state g' from the same trajectory, transforming the training into a multi-goal learning paradigm. The TD target is updated as(Andrychowicz *et al.* 2017):

$$y_t = r_t + \gamma Q_{\text{target}}(s_{t+1}, \pi_{\text{target}}(s_{t+1}|\theta^{\pi}_{\text{target}})|\theta^{Q}_{\text{target}}) \tag{3}$$

where y_t is the target value based on the relabeled goal g'. This combined DDPG and HER framework is particularly suited for our task, enabling effective learning across diverse assembly targets and task constraints.

2) Multimodal Representation Model: A convolutional autoencoder extracts latent features from 320×240 RGB images (Figure 2). The encoder output is flattened and mapped to a 256-dimensional latent vector z_t, while the symmetric decoder reconstructs the input. The reconstruction loss is:

$$\mathcal{L}_{\text{recon}} = \|\hat{x}_t - x_t\|_2^2 \tag{4}$$

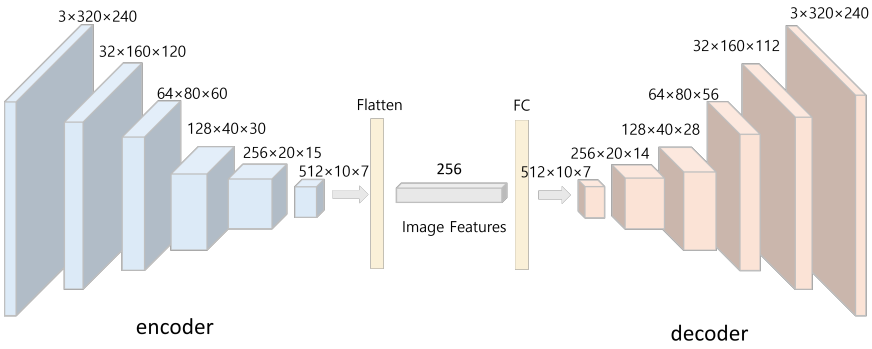

Fig. 2. Convolutional AutoEncoder for extracting visual latent features

To enforce temporal consistency and task-relevant encoding, a forward model predicts the next latent feature z_{t+1} from $\{z_t, s_t^{\text{proprio}}, g_t, a_t\}$. Proprioceptive state s_t^{proprio} and goal g_t are first projected by two-layer MLPs, then concatenated with z_t and a_t. The prediction loss is:

$$\mathcal{L}_{\text{pred}} = \|\hat{z}_{t+1} - z_{t+1}\|_2^2 \tag{5}$$

This design encourages the encoder to capture physically meaningful, geometry-aware features beyond image reconstruction alone (Lesort *et al.* 2018). The total training objective is:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \alpha \cdot \mathcal{L}_{\text{pred}} \tag{6}$$

with $\alpha = 0.3$ in our experiments.

3) Environment: The multimodal observation space incorporates visual information, proprioceptive feedback, and target position. The state and action spaces are defined as follows:

$$\mathbf{S}_t = \left[\mathbf{o}_t^{\text{image}},\ \mathbf{s}_t^{\text{proprio}},\ \mathbf{g}_t\right] \tag{7}$$

$$\mathbf{A}_t = \left[\delta x, \delta y, \delta z\right] \tag{8}$$

Here, $\mathbf{o}_t^{\text{image}}$ denotes the raw image observation captured by the camera at time t, $\mathbf{s}_t^{\text{proprio}} = [x_t, y_t, z_t, \dot{x}_t, \dot{y}_t, \dot{z}_t]$ represents the proprioceptive state including the end-effector position and velocity, and $\mathbf{g}_t = [x^{\text{goal}}, y^{\text{goal}}, z^{\text{goal}}]$ is the target position in Cartesian space. The action space $\mathbf{A}_t$ corresponds to the displacement commands for the robot's end-effector. To simplify the reward design and avoid potential policy bias caused by ill-shaped dense rewards, we employ a sparse reward formulation. The reward function is defined as follows:

$$\text{obs\_dis} = |x_t - x^{\text{goal}}| + |y_t - y^{\text{goal}}| + |z_t - z^{\text{goal}}| \tag{9}$$

$$r_t = \begin{cases} 0 & \text{if obs\_dis} \leq \epsilon, \\ -1 & \text{otherwise} \end{cases} \tag{10}$$

Here, obs_dis denotes the total positional error between the robot's end-effector and the target in Cartesian coordinates, and ϵ is a pre-defined distance threshold (set to 0.02 in our experiments). The reward is only granted when the task is successfully completed within this precision, ensuring that the agent learns from truly successful trials while encouraging exploration otherwise.

B. Sim-to-real Transfer

To improve zero-shot sim-to-real transfer, we adopt domain randomization with two types: observation randomization adds noise to the target pose and initial state to simulate sensing uncertainty, while dynamics randomization perturbs physical parameters to reflect dynamic variability. Following established practices in robust reinforcement learning, both Gaussian and uniform noise are applied during training, which are commonly used to represent stochastic sensor jitter and bounded parameter uncertainties, respectively(Tobin *et al.* 2017). This approach generates diverse scenarios that enhance policy robustness to real-world disturbances. The detailed settings are shown in Table I.

However, image-based policies are sensitive to frequency distribution shifts in sim-to-real transfer, and traditional spatial augmentations (e.g., cropping, rotation) offer limited gains. To address this, we employ SRM(Huang *et al.* 2022), which perturbs frequency components while preserving spatial semantics. By randomly masking frequency

TABLE I

DOMAIN RANDOMIZATION CONFIGURATION

Parameters (unit)	μ	ρ	**Noise distribution**
Initial joint angles j_i (rad)	0	0.05	Uniform
Actions a_t (m)	0	0.01	Gaussian

bands, SRM encourages the policy to learn stable, domain-invariant features, improving robustness to variations in color, texture, and noise in real-world observations.

SRM effectively mitigates overfitting to simulation-specific spectral distributions and plays a critical role in enabling zero-shot transfer under image-based observations. In practice, SRM first maps the image observation o_t to the frequency domain using the Fourier transform, randomly suppresses the magnitude spectrum within a specified radial range, and finally reconstructs the image via inverse Fourier transform. The entire process is formalized as:

$$\text{aug}(o_t) = \mathcal{F}^{-1}\left(\mathcal{M}_{r_1,r_2} \odot \mathcal{F}(o_t)\right) \tag{11}$$

Here, $\mathcal{F}$ and $\mathcal{F}^{-1}$ denote the forward and inverse Fourier transform respectively, and $\mathcal{M}_{r_1,r_2}$ represents a binary mask that randomly erases spectral components within a specified frequency band $[r_1, r_2]$. The symbol $\odot$ denotes element-wise multiplication. This process ensures that the augmented image preserves spatial semantics while suppressing frequency-dependent bias. The SRM-augmented image $\text{aug}(o_t)$ is passed through a visual encoder f_θ to produce a latent feature z_t, which is subsequently used in the Q-value regression loss:

$$\mathcal{L}_Q = \|Q_\theta(f_\theta(\text{aug}(o_t)), a_t) - (r_t + \gamma Q_{\theta'}(f_{\theta'}(o_{t+1}), \mu_{\phi'}(f_{\theta'}(o_{t+1}))))\|_2^2 \tag{12}$$

In practice, SRM is applied to each training batch with a probability of $p = 0.5$, while no augmentation is applied during evaluation. This setup ensures that the learned latent representation z_t is robust to cross-domain spectral shifts while maintaining consistency across training and deployment stages.

C. Force Control

The initial approach relied solely on an RL policy for zero-shot transfer, which failed to control transient contact forces during insertion. This often led to excessive impacts, jamming, or damage, as the policy lacked real-time awareness of contact states and could not dynamically regulate forces in response to pose deviations or environmental disturbances. Furthermore, while incorporating simulated force feedback can improve contact awareness, these signals are often too idealized and fail to capture real-world phenomena, leading to a significant sim-to-real gap that necessitates costly manual tuning.

To address this, we follow an established adaptive impedance control framework as a complementary mechanism (Jung *et al.* 2004). This mechanism combines the high-level strategic guidance provided by the RL policy with the low-level physical compliance of an adaptive impedance module. If the execution of the RL policy causes the contact force to exceed the predefined safety threshold $F_r = 10\,\text{N}$, the system autonomously transit to impedance

control. In this state, the impedance module generates a compliant three-dimensional reference trajectory, denoted as $x_v(t)$, to regulate the contact force and ensure the mechanical safety of the system.

First, the dynamics of the manipulator in Cartesian space is given by:

$$M(x)\ddot{x} + C(x,\dot{x})\dot{x} + F_g(x) = F_\tau + F_e \tag{13}$$

where $M(x)$ and $C(x,\dot{x})$ represent the inertia and Coriolis matrices, $F_g(x)$ is the gravity term, F_τ is the control input, and F_e denotes the external contact force vector.

To regulate the interaction force, we implement a position-based impedance control law. This establishes a dynamic relationship between the force error and the deviation of the generated trajectory $x_v(t)$ from the target x_r:

$$M\ddot{E} + B\dot{E} + KE = F_e - F_r \tag{14}$$

where F_r is the reference force vector and $E = x_r - x_v$ represents the tracking deviation. In the insertion direction, the stiffness K is set to zero to ensure pure force compliance. To compensate for uncertainties in environment location and the sim-to-real gap, an adaptive term $\Omega(t)$ is introduced. The scalar impedance relationship in the contact direction is formulated as:

$$m\ddot{e} + b(\dot{e} + \Omega(t)) = f_e - f_r \tag{15}$$

where $e = x_r - x_v$ is the scalar position error, while f_r and f_e denote the scalar reference force and the measured contact force along the insertion direction, respectively. In a digital implementation, the compensatory term $\Omega(t)$ is updated at each sampling interval λ to eliminate steady-state force bias:

$$\Omega(t) = \Omega(t-\lambda) + \eta \cdot \frac{f_r - f_e(t-\lambda)}{b}, \quad \eta > 0 \tag{16}$$

where η is the adaptation rate. The output of the impedance module is the commanded trajectory $x_v(t)$, which is updated by integrating the velocity correction:

$$x_v(t) = x_v(t-\lambda) + \left(\Omega(t) + \frac{f_e(t-\lambda) - f_r}{b}\right) \cdot \lambda \tag{17}$$

In our experimental setup, the parameters were configured as: $m = 2$, $b = 20$, $\eta = 0.001$, and the initial value of the compensatory term was $\Omega(0) = 0$. The desired insertion position x_r serves as the global target for the task.

IV. Experiments

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

- To what extent can the proposed multimodal fusion framework effectively and reliably perform the peg-in-hole assembly task, particularly under challenging conditions where the target hole is initially occluded or outside the field of view? Furthermore, how well does the framework generalize across insertion tasks involving holes of varying geometries?
- Can the policy trained with SRM achieve zero-shot sim-to-real transfer, enabling successful real-world insertion without any fine-tuning? How does its performance compare with baseline methods in terms of quantitative metrics such as insertion success rate?

- Does the integration of an adaptive force-feedback-based hybrid control strategy enhance the robustness and reliability of the overall assembly process,especially in mitigating task failures caused by alignment errors or excessive contact forces during insertion?

A. Simulation Experiment

In the real-world experimental setup, we utilize a 6-DOF Elite robotic arm, as shown in Figure 3, equipped with a RealSense D455 camera mounted on its end-effector, which captures the spatial relationship between the peg and the hole. A six-axis force/torque sensor ATI Mini45 is installed in-line between the end-effector and the peg to measure interaction forces and torques during the insertion process. To support sim-to-real transfer, a high-fidelity simulation environment is constructed in MuJoCo. For all tested peg-hole configurations, the clearance between the peg and the hole is maintained at 1 mm.

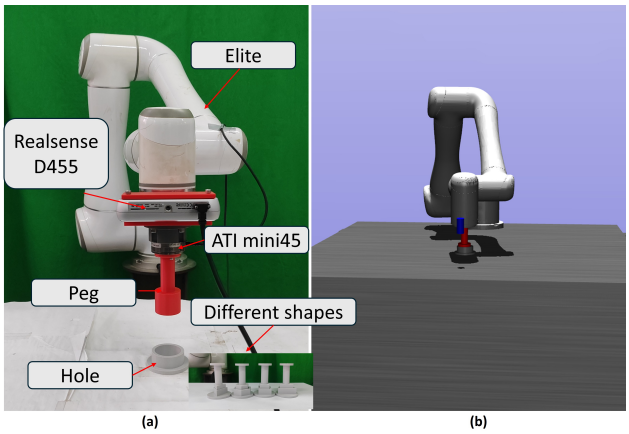

Fig. 3. Real and simulated robotic assembly systems

The policy was trained and evaluated every five training fragments, with cumulative rewards recorded. We conducted a series of ablation studies to assess the contribution of each input modality, training each policy variant under identical conditions except for the state composition. Each training run began with 500 initial random steps to populate the experience replay buffer. Thereafter, a mini-batch of 64 transitions was sampled for network optimization during each environment interaction step. The experience replay buffer capacity was set to 30,000. The learning rates for both the actor and critic networks were uniformly set to actor_lr $= 1 \times 10^{-4}$ and critic_lr $= 1 \times 10^{-4}$. Furthermore, we utilized a discount factor of 0.99 and a soft target update rate of 0.05 for the target networks.

We employed an encoder network to extract structural features from input images. To support this, we collected a total of 150,000 RGB images through randomized exploration in the simulated environment. The dataset included balanced coverage of various peg-hole geometries and images augmented using SRM. In addition to raw image observations, we also recorded proprioceptive signals, target positions, and executed actions, forming a structured dataset. Specifically, each trajectory sample consisted of a tuple $(\text{image}_t, \text{proprio}_t, \text{goal}_t, \text{action}_t, \text{image}_{t+1})$, enabling the model to learn time-consistent latent representations.

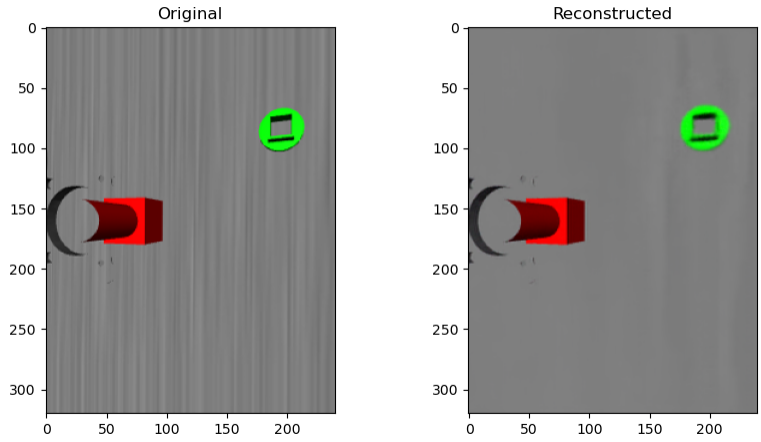

Fig. 4. Reconstruction performance of the visual encoder

The quality of the visual encoding is illustrated in Figure 4, which shows representative examples of original input images and their corresponding reconstructions by the trained autoencoder. During training, the parameters of the visual encoder are initialized with weights pretrained on a shape-diverse dataset through a combination of image reconstruction and forward prediction tasks. The training curves for different modality configurations are illustrated in Figure 5. As shown, the position-only model exhibits the fastest convergence and highest rewards during the simulation training phase. This superior performance in the virtual environment is primarily due to the low-dimensional and highly structured nature of position-based state representations, which allow the agent to directly map task-relevant geometric relationships without the computational burden of high-dimensional visual feature extraction. However, as validated by the real-world results in Table II, this efficiency in simulation comes at the cost of robustness. Such heavy reliance on simplified and noise-free spatial states leads to a significant performance drop during sim-to-real transfer, as the model lacks the ability to handle the complex sensing uncertainties and localization errors inherent in physical environments. In contrast, the vision-only model, though capable of convergence, demonstrates lower training efficiency and significant instability. Without explicit spatial guidance, this model struggles to capture precise geometric relationships in scenarios where the target is partially occluded or invisible, thereby limiting its generalization capability in large and complex workspaces.

The proposed full fusion model achieves a strategic balance by integrating visual, proprioceptive, and target pose information. Although its initial reward growth is more gradual compared to the position-based model, it demonstrates a more stable and sustained learning trajectory. By learning a robust and multimodal representation, this model effectively bridges the sim-to-real gap, ensuring that the policy remains reliable and adaptive under real-world conditions where single-modal information may be insufficient or corrupted.

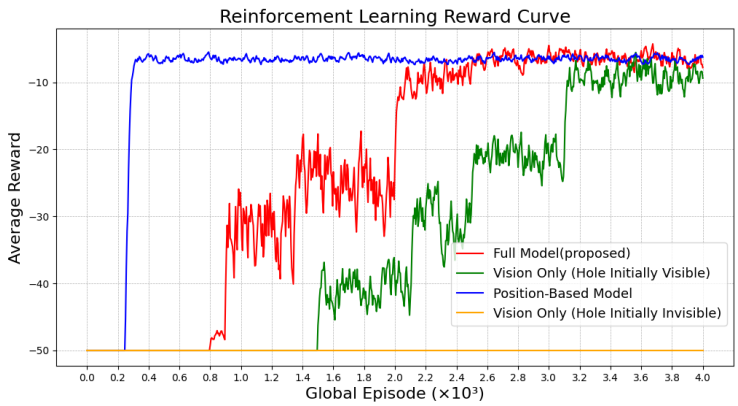

Fig. 5. Ablation study showing the effect of different training configurations on learning a peg-hole insertion policy in simulation

B. Sim-to-real Transfer

Then, we evaluate our method in real-world settings. This subsection analyzes the zero-shot transfer performance of the trained policy, investigates the impact of different input modalities on the transfer success rate, and examines the system's ability to complete the task under high contact uncertainty using adaptive variable impedance control.

To enable zero-shot sim-to-real transfer from simulation to the real robotic system, we incorporate two general yet complementary strategies during training: domain randomization and SRM. Specifically, SRM augmentation is applied to each image batch with a probability of 50%, where the frequency masking radius (r_1, r_2) is randomly sampled from the interval $[0, 0.5]$ (normalized frequency). The range $[0, 0.5]$ corresponds to the entire spectrum from the DC component (0) to the Nyquist frequency (0.5), representing all valid information in the image. No visual augmentation is applied during testing. The SRM augmentation process is visualized in Fig. 6. Meanwhile, we conducted zero-shot sim-to-real transfer experiments under various configurations. Each experimental group was tested 50 times on the real robotic system, and the insertion success rates together with the corresponding performance metrics are summarized in Table II.

Experimental results demonstrate that SRM enhances policy robustness by introducing perturbations in the frequency domain of input images, effectively simulating real-world visual uncertainties such as illumination variation, color distortion, and texture interference. Applying SRM to vision-based policies significantly improves their generalization performance in real-world scenarios. Both the Vision-Only model and the Full Model achieve remarkable gains in real-world insertion success rates across three different hole types, with improvements ranging from 8% to 16%. Meanwhile, as shown in Table II, the introduction of SRM does not come at the cost of execution efficiency. On the contrary, the policy maintains comparable average insertion times while achieving noticeably lower peak insertion forces. This indicates that SRM-trained policies exhibit smoother and more compliant control behaviors during the contact phase, resulting in more stable insertions and reduced mechanical impact on both the robot and the assembled components.

Building upon the improved robustness of the visual modality, we further analyze how the design of sensory

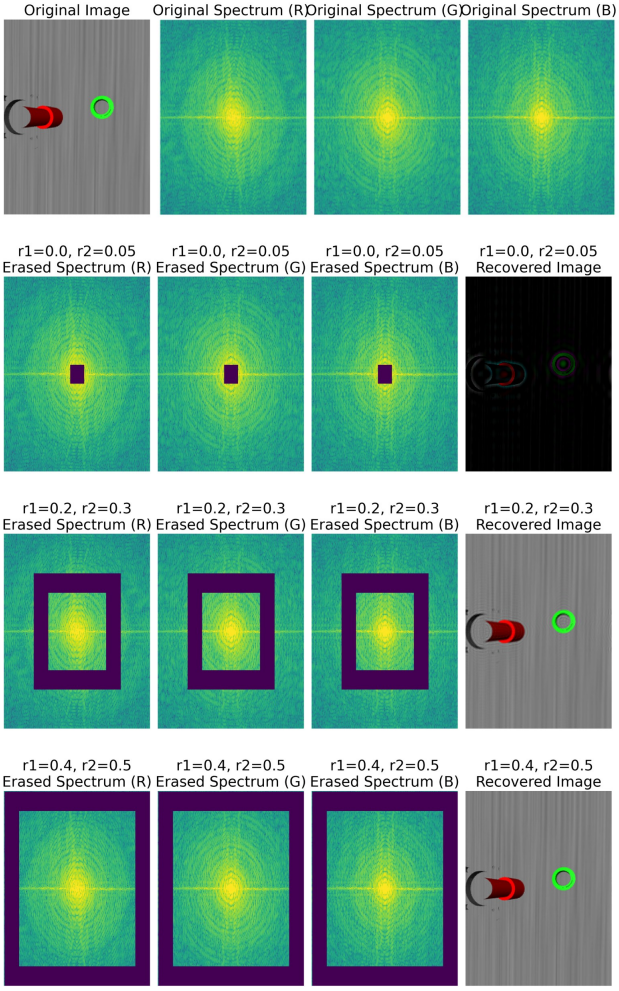

Fig. 6. Visualization of the Spectrum Random Masking (SRM) process

modalities affects policy stability and generalization. The results show that the Full Model strategy consistently achieves the highest success rates across all hole types. Moreover, under the multimodal input configuration, the influence of hole geometry on success rate is substantially reduced. This indicates that the visual modality provides critical structural and geometric information, while the proprioceptive modality enhances real-time control of end-effector motion. These complementary inputs enable the policy to more accurately guide insertion poses, detect failure conditions, and adapt its behavior accordingly.

In contrast, although the Position-Based strategy performs well in simulation, its success rate decreases significantly in real-world scenarios when dealing with asymmetric holes. This is primarily because the policy cannot perceive the specific boundary features of the hole and lacks the ability to correct geometric deviations during the insertion process. The strategy benefits from simplified physical constraints during simulation. However, in real-world settings, execution errors such as control dead zones and actuator latency may cause the end-effector to deviate from the expected trajectory. Although the Vision-Only strategy incorporates visual input, it still struggles

TABLE II

COMPARISON OF REAL-WORLD PERFORMANCE UNDER DIFFERENT POLICY MODALITIES AND HOLE SHAPES

Policy Modality	**Hole Shape**	**SRM Usage**	**Success Rate (%)**	**Average Insertion Time (s)**	**Peak Force (N)**
Full Model	Circle	W/ SRM	88.0	9.5	20.0
	Circle	W/O SRM	80.0	9.0	20.5
	Square	W/ SRM	84.0	9.0	22.0
	Square	W/O SRM	76.0	9.5	23.0
	Triangle	W/ SRM	86.0	8.7	18.5
	Triangle	W/O SRM	76.0	9.8	21.5
Vision Only (Visible)	Circle	W/ SRM	68.0	10.5	23.0
	Circle	W/O SRM	52.0	11.5	25.5
	Square	W/ SRM	68.0	10.5	23.5
	Square	W/O SRM	52.0	11.8	26.0
	Triangle	W/ SRM	68.0	10.5	24.0
	Triangle	W/O SRM	54.0	12.0	26.5
Position-Based	Circle	–	70.0	9.5	23.0
	Square	–	58.0	9.7	23.5
	Triangle	–	54.0	9.3	24.5

to precisely control the end-effector trajectory, resulting in amplified pose deviations and suboptimal performance compared to multimodal fusion strategies. Therefore, modality fusion is a critical factor in improving the success rate and robustness of reinforcement learning-based insertion policies in real-world deployment.

C. Hybrid Control with Adaptive Variable Impedance

Although the introduction of SRM-based frequency domain augmentation and multimodal perception strategies significantly improves the overall success rate of reinforcement learning policies in both simulated and real-world peg-in-hole tasks, further analysis of real-world failure cases reveals that some failures are not caused by incorrect trajectory planning. Instead, they result from the accumulation of execution errors during deployment, which lead to sudden increases in insertion force or edge jamming at the end-effector, ultimately causing task failure.

Thus, we adopt an adaptive variable impedance control mechanism that is decoupled from the policy structure, serving as a low-level compensation module. As illustrated in Figure 7, the process first uses the RL policy to approach the hole, and when the contact force exceeds the predefined threshold, the system switches to impedance control mode.

To validate the effectiveness of the proposed adaptive variable impedance control mechanism in suppressing contact impact and improving insertion stability, we conducted force feedback and trajectory recording experiments on a real robotic platform. Figure 8 (a) and Figure 8 (b) illustrate the variation of three-axis contact forces and the 3D motion trajectory of the end-effector during the insertion process, respectively.

As shown in Figure 8(a), during 0–4 s the end-effector has not yet contacted the hole, and all force components remain near zero. At $t = 4$ s, F_z rises sharply, marking initial contact with the bottom surface. This is accompanied

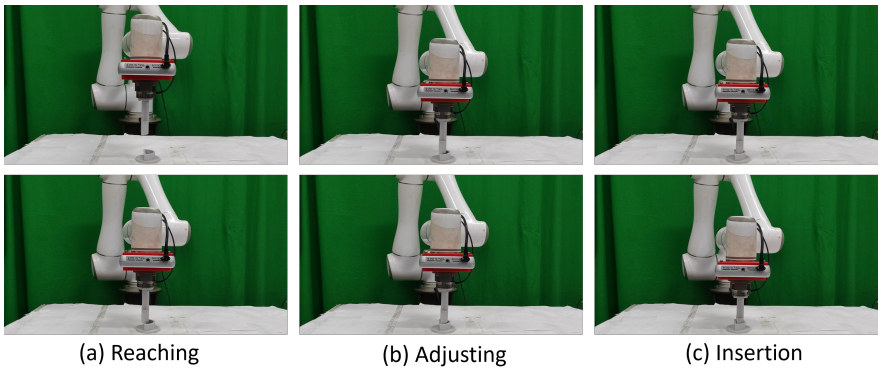

Fig. 7. Snapshots of the insertion process

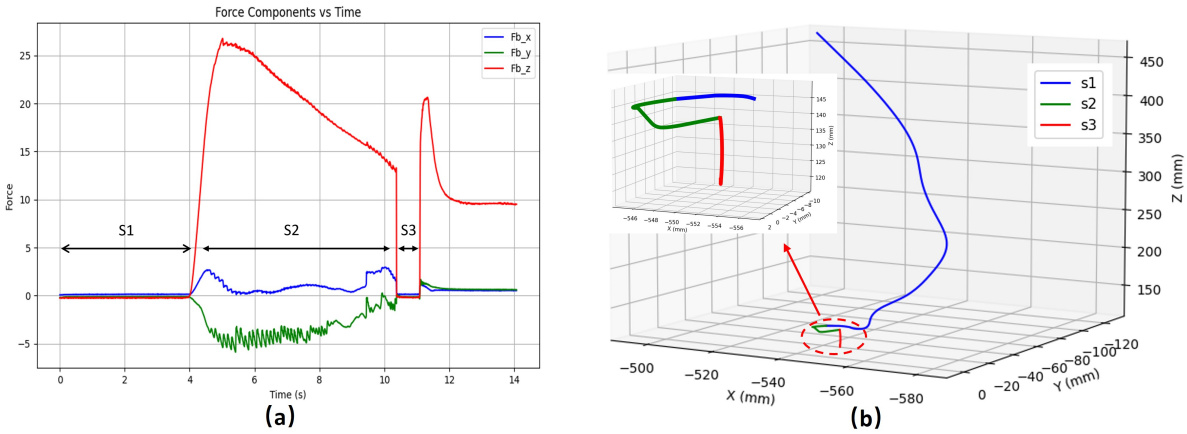

Fig. 8. Plot (a) shows the force measurements recorded during the assembly process, and plot (b) illustrates the corresponding 3D trajectory of the Cartesian motion

by peaks in F_x and F_y, indicating misalignment and boundary collision. Once F_z exceeds the 10 N threshold, the controller activates variable impedance, reducing stiffness and increasing damping. Consequently, F_z decreases and stabilizes at about 10 N.

Figure 8(b) shows the 3D end-effector trajectory. The blue curve (S1) corresponds to RL-guided motion toward the hole. The green curve (S2) shows boundary contact and collision, where excessive F_z triggers impedance control, enabling compliant insertion. The red curve (S3) denotes successful alignment, with forces in all directions converging to zero.

Figure 9 presents the insertion success statistics of our proposed multimodal reinforcement learning strategy across different hole shapes. Each shape was tested 50 times, with the first three shapes seen during training and the last two used for zero-shot transfer evaluation. In most cases, insertions were successfully completed using only the RL policy, while in a few scenarios, the adaptive impedance control module was triggered to correct pose errors or mitigate contact forces. This demonstrates that the multimodal fusion strategy exhibits strong learning capabilities

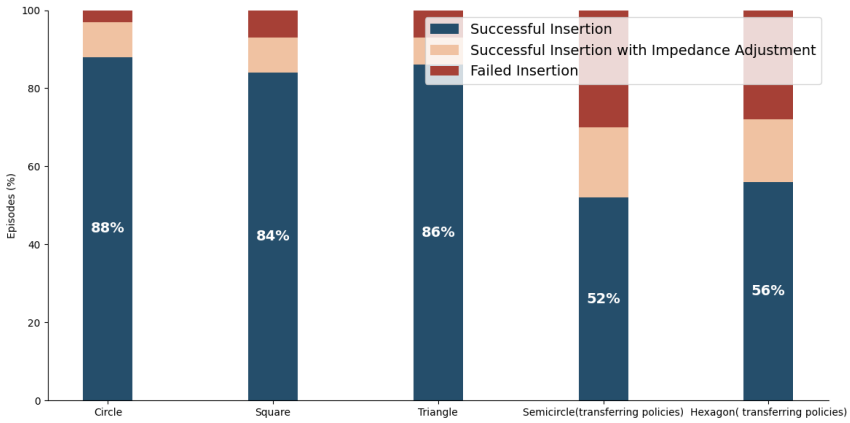

Fig. 9. Statistical results across different hole shapes

and control robustness under familiar structural conditions, enabling stable and reliable insertion operations. For previously unseen shapes (semicircle and hexagon), although the direct insertion success rate of the RL policy alone decreased, the success rates increased to 70% and 72%, respectively, after activating the impedance control module. These results clearly validate the proposed method's transferability to unseen geometries and structural variations.

V. Conclusion

We propose a reinforcement learning framework that integrates multimodal perception, frequency-domain augmentation, and hybrid control strategies to address the challenges of limited generalization and contact uncertainty in complex peg-in-hole robotic assembly tasks. To mitigate the sensitivity of conventional RL policies to visual domain shifts during sim-to-real transfer, we introduce SRM, a frequency-domain augmentation technique that significantly improves the robustness of visual policies in real-world deployment. Experimental results demonstrate that SRM improves the average insertion success rate across different hole shapes by **12%**, while maintaining comparable insertion times and lower peak insertion forces, indicating smoother and more compliant insertion behavior.

To further enhance adaptability under real-time contact conditions, we incorporate an adaptive impedance control module driven by force feedback. This module dynamically adjusts the stiffness and damping of the end-effector when detecting excessive contact force or misalignment, effectively suppressing impact forces and mitigating pose deviation. In zero-shot generalization tests on unseen geometries, the insertion success rate using only the RL policy was **52%** and **56%**, respectively, whereas with the hybrid control mechanism, it increased to **70%** and **72%**. These results highlight that the proposed multimodal-hybrid framework not only generalizes to novel structures but also compensates for execution errors arising from model mismatch and sensor latency.

Moreover, we employ a forward prediction model and a joint reconstruction loss to pretrain the visual encoder via self-supervised learning, thereby enhancing structural awareness and temporal consistency of the latent representations. This leads to a more physically meaningful state space for downstream policy optimization. Nevertheless, the current visual encoder still shows limited capability in capturing previously unseen geometric features, as reflected

by the moderate drop in zero-shot performance. In future work, we plan to explore shape-adaptive perception and representation learning to further improve the generalization of the visual module. Additionally, we expect that integrating complementary sensory modalities such as acoustic and language feedback will further reduce assembly errors and improve system adaptability in unstructured real-world environments.

REFERENCES

Andrychowicz, Marcin *et al.* (2017). "Hindsight experience replay". In: *Advances in neural information processing systems* 30.

Bousmalis, Konstantinos *et al.* (2018). "Using simulation and domain adaptation to improve efficiency of deep robotic grasping". In: *2018 IEEE international conference on robotics and automation (ICRA)*. IEEE, pp. 4243–4250.

Chen, Wenkai *et al.* (2023). "Multimodality driven impedance-based sim2real transfer learning for robotic multiple peg-in-hole assembly". In: *IEEE Transactions on Cybernetics* 54.5, pp. 2784–2797.

Chen, Yuanpei, Yiran Geng, *et al.* (2023). "Bi-dexhands: Towards human-level bimanual dexterous manipulation". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.5, pp. 2804–2818.

Chen, Yuanpei, Chao Zeng, *et al.* (2023). "Zero-shot sim-to-real transfer of reinforcement learning framework for robotics manipulation with demonstration and force feedback". In: *Robotica* 41.3, pp. 1015–1024.

Elguea-Aguinaco, Íñigo *et al.* (2023). "A review on reinforcement learning for contact-rich robotic manipulation tasks". In: *Robotics and Computer-Integrated Manufacturing* 81, p. 102517.

Huang, Yangru *et al.* (2022). "Spectrum random masking for generalization in image-based reinforcement learning". In: *Advances in Neural Information Processing Systems* 35, pp. 20393–20406.

Jeong, Rae *et al.* (2020). "Self-supervised sim-to-real adaptation for visual robotic manipulation". In: *2020 IEEE international conference on robotics and automation (ICRA)*. IEEE, pp. 2718–2724.

Ju, Siwei *et al.* (2023). "Digital twin of a driver-in-the-loop race car simulation with contextual reinforcement learning". In: *IEEE Robotics and Automation Letters* 8.7, pp. 4107–4114.

Jung, Seul *et al.* (2004). "Force tracking impedance control of robot manipulators under unknown environment". In: *IEEE Transactions on Control Systems Technology* 12.3, pp. 474–483.

Karnan, Haresh *et al.* (2020). "Reinforced grounded action transformation for sim-to-real transfer". In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 4397–4402.

Lee, Michelle A, Yuke Zhu, Krishnan Srinivasan, *et al.* (2019). "Making sense of vision and touch: Self-supervised learning of multimodal representations for contact-rich tasks". In: *2019 International conference on robotics and automation (ICRA)*. IEEE, pp. 8943–8950.

Lee, Michelle A, Yuke Zhu, Peter Zachares, *et al.* (2020). "Making sense of vision and touch: Learning multimodal representations for contact-rich tasks". In: *IEEE Transactions on Robotics* 36.3, pp. 582–596.

Lesort, Timothée *et al.* (2018). "State representation learning for control: An overview". In: *Neural Networks* 108, pp. 379–392.

Leyendecker, Lars *et al.* (2022). “Deep Reinforcement Learning for Robotic Control in High-Dexterity Assembly Tasks—A Reward Curriculum Approach”. In: *International Journal of Semantic Computing* 16.03, pp. 381–402.

Li, Guoyuan *et al.* (2021). “Development of a manufacturing system for gear assembly using collaborative robots”. In: *2021 IEEE International Conference on Mechatronics and Automation (ICMA)*. IEEE, pp. 22–27.

Lillicrap, Timothy P *et al.* (2015). “Continuous control with deep reinforcement learning”. In: *arXiv preprint arXiv:1509.02971*.

Liu, Zhenyu *et al.* (2023). “A motion planning method for visual servoing using deep reinforcement learning in autonomous robotic assembly”. In: *IEEE/ASME Transactions on Mechatronics* 28.6, pp. 3513–3524.

Luo, Jianlan *et al.* (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.

Song, Rui *et al.* (2021). “Skill learning for robotic assembly based on visual perspectives and force sensing”. In: *Robotics and Autonomous Systems* 135, p. 103651.

Spector, Oren *et al.* (2022). “Insertionnet 2.0: Minimal contact multi-step insertion using multimodal multiview sensory input”. In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, pp. 6330–6336.

Tobin, Josh *et al.* (2017). “Domain randomization for transferring deep neural networks from simulation to the real world”. In: *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, pp. 23–30.

Valassakis, Eugene *et al.* (2020). “Crossing the gap: A deep dive into zero-shot sim-to-real transfer for dynamics”. In: *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 5372–5379.

Wu, Zheng *et al.* (2023). “Prim-lafd: A framework to learn and adapt primitive-based skills from demonstrations for insertion tasks”. In: *IFAC-PapersOnLine* 56.2, pp. 4120–4125.

Xiao, Ruihong *et al.* (2024). “One-shot sim-to-real transfer policy for robotic assembly via reinforcement learning with visual demonstration”. In: *Robotica* 42.4, pp. 1074–1093.

Xie, Liang *et al.* (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)*. IEEE, pp. 2982–2988.

Yadav, Sudhir Pratap *et al.* (2024). “Learning vision-based robotic manipulation tasks sequentially in offline reinforcement learning settings”. In: *Robotica* 42.6, pp. 1715–1730.

Zhang, Zhuangzhuang *et al.* (2024). “A residual reinforcement learning method for robotic assembly using visual and force information”. In: *Journal of Manufacturing Systems* 72, pp. 245–262.