Impedance Learning-based Adaptive Force Tracking for Robot on Unknown Terrains

Yanghong Li, Li Zheng, *Student Member, IEEE,* Yahao Wang, Erbao Dong, *Member, IEEE,* and Shiwu Zhang, *Member, IEEE,*

***Abstract*—Aiming at the robust force tracking challenge for robots in continuous contact with uncertain environments, a novel adaptive variable impedance control policy based on deep reinforcement learning (DRL) is proposed in this paper. The policy includes a neural network feedforward controller and a variable impedance feedback controller. Based on the DRL algorithm, the iterative network feedforward controller explores and pre-learns the optimal policy for impedance tuning in simulation scenarios with randomly generated terrain. The converged results are then used as feedforward inputs in the variable impedance feedback controller to improve the force-tracking performance of the robot during contact. A simplified dynamic contact model between the robot and the uncertain environment called the "couch model", which satisfies the Lipschiz continuity condition, is developed to provide boundary conditions for the safe transfer of capabilities learned in simulation to real robots. Unlike the exhaustive example that relies on the completeness of the learning samples, this paper gives theoretical proofs of the stability and convergence of the proposed control policy via Lyapunov's theorem and contraction mapping principle. The control method proposed in this paper is more interpretable and shows higher sample utilization efficiency and generalization ability in simulations and experiments.**

***Index Terms*—Cooperating Robots, model learning for control, compliance and impedance control, force tracking**

I. Introduction

With the development of emerging technologies such as collaborative robotic arms, dexterous hands, quadrupedal robots, and wearable robotic exoskeletons, robots are being increasingly applied to contact tasks in various environments [1]–[4]. Complex tasks such as bi-arm collaboration, flexible grasping and human-robot interaction also require contact force stabilization [5]. Unlike soft robots based on material impedance, manipulators require a rigid structure to ensure operational accuracy [6]. Especially facing delicate tasks such as machining and grinding, following the reference trajectory while maintaining the desired force often becomes a contradiction [7].

Research supported by the National Key R&D Program of China (Grant number: 2018YFB1307400) and the Fundamental Research Funds for the Central Universities.

Yanghong Li, Li Zheng, Yahao Wang, and Erbao Dong are associated with the Humanoid Robotics Institute, State Key Laboratory of Precision and Intelligent Chemistry, CAS Key Laboratory of Mechanical Behavior and Design of Materials, Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui, 230026, PRC (*Corresponding author is Erbao Dong and can be contacted at e-mail: ebdong@ustc.edu.cn)

Shiwu Zhang is associated with the Humanoid Robotics Institute, CAS Key Laboratory of Mechanical Behavior and Design of Materials, Department of Precision Machinery and Precision Instrumentation, University of Science and Technology of China, Hefei, Anhui, 230026, PRC

As a classical force control framework, impedance control and force-position hybrid control [8]–[11] have achieved good control results in structured environments. By correct modeling of the robot and the environment, it is possible to control the contact force quickly and stably. The introduction of impedance control on industrial robots [10], [12]–[15] improves the success and efficiency of shaft-hole assembly tasks. Force-position hybrid control helps promote stability in grasping similar objects with dexterous hands [16]. However, impedance control has many problems in the application of robots interacting with unstructured environments. For uncertain environments, the unknown information about the environment, such as stiffness and position, leads to frequent overloading of joint actuators or damage of contact parts during physical interaction [17]–[19]. If it is possible to remain compliant or adjust the impedance parameters in time during the interaction, the contact force interacting with the object can be controlled, thus avoiding the above problems [20]–[22]. The core of force control in uncertain environments is the problem of impedance parameter adjustment. [23]–[26].

The studies of force tracking problems in uncertain environments are mainly categorized into three types [27], [28]. 1) Direct or indirect adjustment of the reference trajectory. The core idea lies in estimating the stiffness and position information of the environment based on the feedback information, and thus adjusting the reference trajectory or impedance gain. By extending the Kalman filter estimation to analyze the stiffness definition, the control gain can come to realize force tracking [29]. Vision has also been used to give a reference trajectory to assist force tracking [30], [31]. Medina [32] accomplished a contact task in an unstructured environment based on an optimal feedback control method. Duan [25] proposed an adaptive variable impedance control method that stabilizes the system by adjusting the gain online based on the force feedback information. A fuzzy logic-based impedance controller [33], [34] can realize the grinding and polishing task. Although these methods have achieved good results in the simulation of control systems, they are only effective for surfaces with linear or nearly linear variations. This is because the real environmental variations are often nonlinear, which leads to force tracking errors that are always present [35]. In addition, the estimation of the environmental information is subject to error, which further affects the dynamic effect of force tracking [36].

2) Humanoid variable impedance regulation. Usually, the robot model can be known accurately enough, but it is difficult to obtain an accurate environment model [17], [19].

The human central nervous system has been proven to be able to effectively adjust its muscle impedance to achieve stable interaction in a nonlinear dynamic environment [37]. To imitate this impedance regulation in humans, scholars have proposed many variable impedance control methods [38]. Yang [39] realized force tracking control without estimating the environmental stiffness by processing electromyography (EMG) signals collected from human muscles to assist robot stiffness adjustment. Based on human-robot interaction technology, Kronander [40] enabled the robot to efficiently learn human impedance adjustment ability. However, the process of humanoid impedance learning often requires a high concentration of the experimenter's energy and a high workload with low transfer efficiency. It is a method based on a priori experience, which is greatly influenced by human state [41].

3) Reinforcement learning-based control methods. Reinforcement learning (RL) is an important area of artificial intelligence, which is based on the principle of finding the optimal policy for a given cost. Especially with the development of Deep Reinforcement Learning (DRL), it has been widely applied to robot control problems, such as wearable exoskeletons [4], [42], bipedal robot walking [43] and dexterous hand grasping [44]. Li and Buchli [42], [45] combined reinforcement learning with impedance control to adjust the impedance gain parameters based on stochastic optimal theory, which realizes trade-offs between system accuracy, energy consumption and impedance, but the learning results lacked interpretability. Although reinforcement learning has been effectively applied to solve a variety of problems, it has rarely been used to deal with force tracking. This is because the process of learning variable impedance control requires a lot of physical interaction with the environment, which may be a security issue. Worse still, the results of learning are often uncertain, i.e., there is no guarantee of convergence of the learning process and overfitting often occurs.

In summary, the first category of methods is based on classical adaptive adjustment mechanisms and always suffers from force-tracking errors when facing higher-order nonlinear dynamical systems [10], [19]. The second category of methods is heavy workload, low efficiency, and cannot target force-sensitive tasks [19], [37]. The third category incorporating DRL is robust to nonlinear changes in the system but needs to address the safe interaction problem and interpretability analysis of the controller [18], [19], [40].

Considering the problems of the above studies, this paper proposes an adaptive impedance learning method to realize the force-tracking task in complex dynamic environments. It combines the idea of reinforcement learning and the advantages of variable impedance control and is divided into an impedance learning part and a variable impedance control part. The tuning policy of impedance parameters in the variable impedance controller will be given by the deep reinforcement learner. In the virtual environment of randomly generated surfaces, Actor-Critic deep reinforcement learning is used to find the optimal policy for impedance adjustment. The Actor network adjusts the impedance parameters based on the value function; the Critic network fits the relationship between the impedance state and the cost function. Based on the data of the digital twin robot interacting with the virtual environment, the network parameters are updated to minimize the cost function (i.e., force error) by backpropagation. The converged impedance learning results are mapped onto the same real robot to achieve a safe transfer of learning outcomes.

The main contributions of this paper are as follows:

1) Regarding the force tracking problem under uncertain environments, a reinforcement learning control policy with interpretability is proposed by combining DRL feedforward and variable impedance feedback, and its stability and convergence analysis are given.
2) Considering the dynamic changes in the equilibrium position of the environment, a simplified contact model of the robot's end-effector with the unstructured environment, called the "couch model", is developed. By satisfying the Lipschiz continuity condition, boundary conditions are set for the safe transfer of impedance learning results from simulation to real terrain.
3) By including the observable state quantities of the dynamic system in the regression vector of the neural network, the overfitting phenomenon caused by other network control policies that rely only on inputs and outputs in uncertain systems is avoided. Simulations in different terrains and real robot experiments validate our method.

The rest of this article is organized as follows. Section II establishes an equivalent model of the dynamic force tracking problem in an unknown environment, and introduces the proposed adaptive impedance learning method. Section III analyzes the stability and convergence of the proposed control algorithm. Sections IV & V carry out a series of simulations and experiments to verify the feasibility, and the control effect is compared with other variable impedance control methods. The advantages of this study over other neural network-based methods are discussed. Section VI concludes the paper.

II. METHODOLOGY

In this study, the main focus is on the force tracking problem during robot contact with the environment with unknown stiffness and terrain information, aiming to determine the optimal impedance parameter adjustment policy. We first model the interaction between a general unstructured environment and a robot. The force tracking problem is then transformed into a problem of finding the optimal policy for Bellman's equation. The intelligent agent is trained to learn to autonomously adjust the impedance parameters to adapt to changes in the environment location. The notations used in this section and the rest sections are shown in Table I.

A. System Modeling and Variable Impedance Control

Consider the workspace vector $x \in \mathbb{R}^m$ of a manipulator in a singularity-free region. To describe the change of contact force from unconstrained to dynamically constrained, the robot is presented by a second-order mass-damping-spring model

$$M(\ddot{x}_{\mathrm{m}} - \ddot{x}_{\mathrm{r}}) + B(\dot{x}_{\mathrm{m}} - \dot{x}_{\mathrm{r}}) + K(x_{\mathrm{m}} - x_{\mathrm{r}}) = f_{\mathrm{m}} - f_{\mathrm{d}} \quad (1)$$

TABLE I
NOTATION IN THIS ARTICLE

Symbol	Meaning
m, n	Degrees of freedom in task space and number of joints of the robot.
M, B, K	Inertia, damping and stiffness gain matrix of the robot.
x_{m}, x_{r}	Measured trajectory and reference trajectory of the robot.
f_{m}, f_{d}	Measured external force and desired force of the robot.
q, τ	Joint angles and joint torques of the robot.
$D(q)$, $g(q)$ $C(q,\dot{q})$	Inertia matrix, gravity vector, Coriolis and centrifugal force matrices under robot joint space.
K_{env}, x_{env}	Stiffness matrix and rest position of the environment.
$I(t)$, $\Delta I(t)$	Impedance parameters and impedance correction action in task space at time t.
$u_p(t)$, $s_p(t)$ $y_p(t)$	Inputs, states and outputs of robot-environment state-space systems at time t of the pth trial.
$u^*(t)$, $s_r(t)$ $y_r(t)$	Unique optimal inputs, reference states and desired outputs at time t
$f(\cdot)$	Nonlinear function of the robot-environment interaction.
$g(\cdot)$, $\boldsymbol{\nu}_p(t)$	Neural network structure and regression vector of DRL controller at time t of the pth trial.
ρ_u, ρ_x	Upper bound on partial derivatives of DRL controller concerning input and state from 0 to t.
u_p, r_p o_p, o'_p	Input (action), reward, observation, and next observation for DRL control policy of the pth trial.
θ, γ, η	Network hyperparameters, discount factor, learning rate.
z, $\mathcal{G}(z)$	Spectral complex variable, transfer function.
$\lambda_{\min}\{\cdot\}$	Minimum eigenvalue of the matrix.
$\|\cdot\|$	Euclidean norm of the vector.
$\|\cdot\|_\lambda$	Upper bounds on Euclidean norms for the product of mathematical expectation of random vector and convergence term from 0 to t [49].

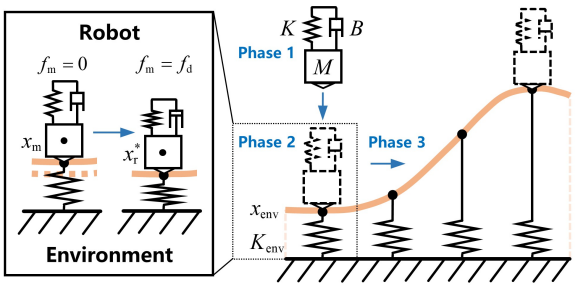

Fig. 1. Couch model for robot contact with dynamic nonlinear environments

where $M \in \mathbb{R}^{m\times m}$, $B \in \mathbb{R}^{m\times m}$, and $K \in \mathbb{R}^{m\times m}$ denote the inertia gain matrix, damping gain matrix, and stiffness gain matrix in the time-varying impedance model, respectively. $x_{\mathrm{m}} \in \mathbb{R}^m$ and $x_{\mathrm{r}} \in \mathbb{R}^m$ are the measured position trajectory and reference position trajectory of the end-effector in Cartesian workspace, respectively. $f_{\mathrm{m}} \in \mathbb{R}^m$ and $f_{\mathrm{d}} \in \mathbb{R}^m$ are the measured and desired contact forces between the robot end and the environment, respectively. Let the force tracking error $\Delta f = f_{\mathrm{m}} - f_{\mathrm{d}}$.

The environment is modeled as a position-varying stiffness model to describe the dynamic changes in the environment.

$$K_{\mathrm{env}}(x_{\mathrm{m}} - x_{\mathrm{env}}) = f_{\mathrm{env}} \tag{2}$$

where K_{env}, x_{env} denote the stiffness and rest position of the environment, respectively.

The force tracking process can be formulated as three phases, as shown in Fig. 1. The first phase is unconstrained free space $(0-t_1)$ where the robot is approaching the environment. The second phase is the beginning of the contact between the robot and the environment, which is usually a rapid collision to a stabilization process $(t_1 - t_2)$. The third phase represents the process in which the environment dynamically changes and the robot tracks the changes and re-stabilizes $(t_2 - t_3)$.

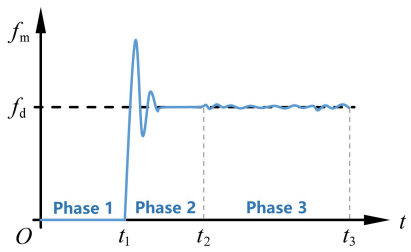

Fig. 2. Schematic diagram of robot-environment contact force in different phases.

As shown in Fig. 2, in most situations, the environmental dynamic changes in the third phase are often unavoidable and nonlinear. Biomechanical studies have shown that human beings, after years of empirical learning, have a well-developed ability to regulate variable impedance [37]. Humans can adapt to nonlinear dynamic changes in the environment by adjusting the stiffness and damping of their skeletal muscles. This is essentially an adaptive variable impedance control based on neural networks. We expect the robot to learn this ability of impedance adjustment and adjust the impedance parameters in real time to track the desired force. To avoid unwanted oscillations, the parameters to be tuned by the controller are $B(t)$ and $K(t)$, which are the row vectors corresponding to the impedance matrix of (1) in the task space.

At each stage of force tracking, the reference trajectory x_{r} and the impedance parameters (stiffness $B(t)$, damping $K(t)$) are provided to the variable impedance controller,

$$I(t) = \left[K(t), B(t), x_{\mathrm{r}}(t)\right]^{\mathrm{T}} \tag{3}$$

where $t \in \{0, 1, ..., N-1\}$ is a discrete-time series and positive integer $N < \infty$.

The controller outputs the end-effector trajectory control command $\alpha(I)$, and α depends on the impedance parameter $I(t)$. A nonredundant manipulator with $n(\geq m)$ joints moving in a singularity-free region of the workspace is chosen to execute the control command. According to the inverse dynamics compensation, the joint torque $\tau \in \mathbb{R}^n$ can be written as

$$\tau = J^{\mathrm{T}}(q)\alpha + F_f\dot{q} + \hat{g}(q) \tag{4}$$

where $q \in \mathbb{R}^n$ denotes the vector of current joint variables, $J(q) \in \mathbb{R}^{n\times m}$ is the end-effector geometric Jacobian matrix, F_f is a positive definite (diagonal) matrix of viscous friction coefficients at the joints and the term $\hat{g}(q)$ is needed to compensate for the static torques due to gravity.

The state change of the environment causes the force error. The designed reinforcement learning feedforward controller will adjust the impedance parameters as the actions

$$\Delta I(t) = [\Delta K, \Delta B, \Delta x_{\mathrm{r}}]^{\mathrm{T}} \tag{5}$$

to generate the feedforward control signal u^f.

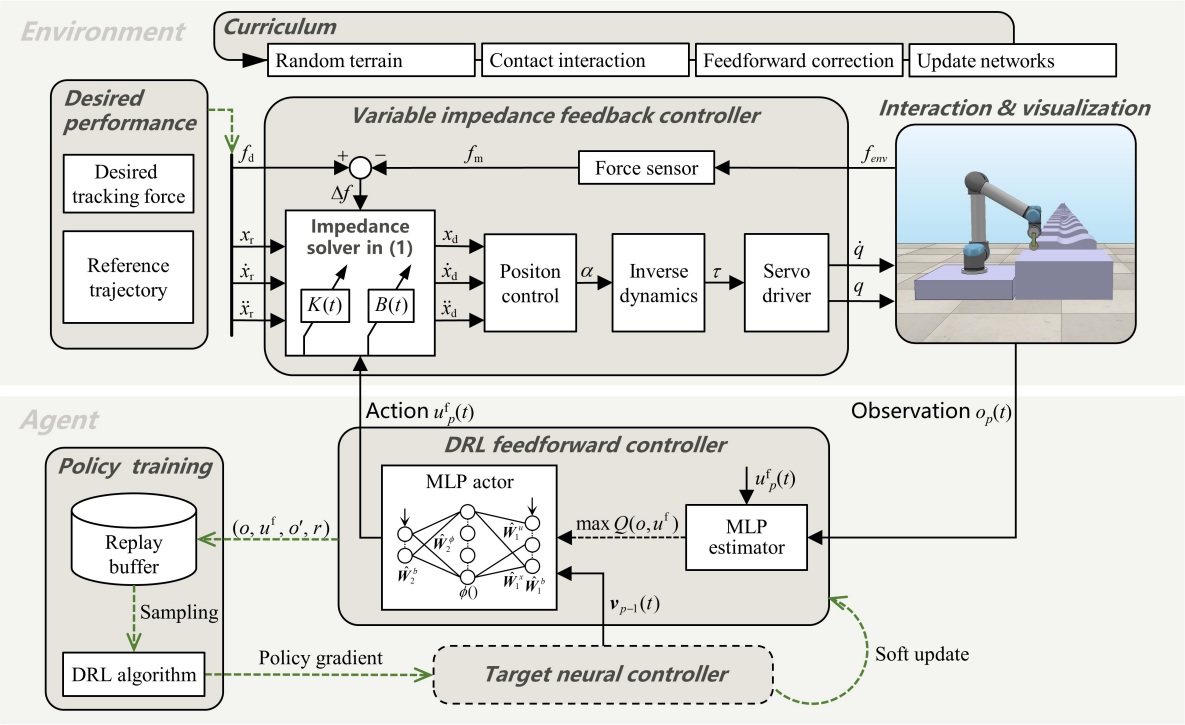

Fig. 3. Overview of the simulation, training, and control method.

The updated impedance parameters will be applied to the robot via (4) to generate new joint torques.

$$I(t+1) = I(t) + \Delta I(t) \tag{6}$$

B. Tracking Problem Formulation

The considered model of robot-environment dynamic interaction is confronted with a more generalized terrain. Dynamic changes in the equilibrium position of the environment make explicit expressions for the contact forces unavailable or difficult to handle. Therefore, we model the interaction process as a Markov decision process in a finite state, described by the following general form of nonlinear state-space discrete-time systems, which was also studied in [42] and [47],

$$\begin{aligned} s_p(t+1) &= h\big(s_p(t), u_p(t)\big) \\ y_p(t) &= K_{\mathrm{C}} s_p(t) \end{aligned} \tag{7}$$

where t is the discrete-time index defined in (3), $p \in \mathbb{N}_+$ represents the trial, $u_p(t) \in \mathbb{R}^m$, $s_p(t) = [x_{\mathrm{m}}, x_{\mathrm{env}}]^{\mathrm{T}}$, and $y_p(t) \in \mathbb{R}^m$ are inputs, states, and outputs of the system at the pth trial, $h(\cdot,\cdot)$ represents the dynamic system consisting of the robot and the environment and is a nonlinear function, x_{env} will be random to characterize the uncertain environmental variations, and K_{C} is the matrix with appropriate dimensions.

Considering the realistic constraints and the existence of solutions, let the system (7) studied in this paper satisfy the following properties and assumptions [47], [48]:

Property 1: $h(\cdot,\cdot)$ is globally Lipschiz continuous on $s_p(t)$ and $u_p(t)$. That is, for any $t \in \{0, ..., N-1\}$, there exists a positive constant l satisfying

$$\begin{aligned} &\big\| h\big(s_1(t), u_1(t)\big) - h\big(s_2(t), u_2(t)\big) \big\| \\ &\quad \le l\big(\|s_1(t) - s_2(t)\| + \|u_1(t) - u_2(t)\| \big) \end{aligned} \tag{8}$$

where $l = \max\{h_u(t), h_s(t)\}$.

Property 2: For any realizable desired force $y_{\mathrm{r}}(t)$ and given reference state $s_{\mathrm{r}}(t) = [x_{\mathrm{r}}, x_{\mathrm{env}}]^{\mathrm{T}}$, there exists a unique $u^*(t)$ that satisfies the following state-space equation

$$\begin{aligned} s_{\mathrm{r}}(t+1) &= h\big(s_{\mathrm{r}}(t), u^*(t)\big) \\ y_{\mathrm{r}}(t) &= K_{\mathrm{C}} s_{\mathrm{r}}(t) \end{aligned} \tag{9}$$

which means that the desired force uniquely determines the reference trajectory based on the environment's stationary state and stiffness coefficient.

Then, the tracking error can be rewritten as

$$e(t) = y_p(t) - y_{\mathrm{r}}(t) = \Delta f \tag{10}$$

with Δf defined in (1). The detailed proof of the existence and uniqueness of Property 2 is given in the literature [48].

Assumption 1: To simplify the analysis, we assume the same initial state for all trials. i.e.

$$\forall s_p(0) = s_{\mathrm{r}}(0) \tag{11}$$

The tracking problem is equivalent to learning the optimal impedance tuning policy from the observed data.

C. DRL for Impedance Learning

Fig. 3 depicts the proposed DRL-based variable impedance control solution. The deep neural network controller is pre-trained in a randomly generated virtual environment. The trained network is used to correct the impedance parameters in real time during the control process. The control signal consists of two parts

$$u(t) = u^f(t) + u^b(t) \tag{12}$$

where $u^f(t)$ is the feedforward control input and $u^b(t)$ is the feedback control input. $u^b(t)$ is determined by the current feedback $e(t)$, in this case $u^b(t) \propto e(t)$. The core of this paper is $u^f(t)$.

1) DRL controller: The neural network controller contains an actor and an estimator. For the pth trial, the actor employs a Multi-Layer Perception (MLP) structure to output the feed-forward signal

$$u_p^f(t) = g\big(\boldsymbol{\nu}_{p-1}(t)\big) \tag{13}$$

where $\boldsymbol{\nu}_{p-1}(t)$ is the regression vector containing the signals recorded during the previous trial. $\boldsymbol{\nu}_{p-1}(t)$ is composed of two components,

$$\boldsymbol{\nu}_{p-1}(t) = \big[u_{p-1}(t), x_{p-1}(t)\big]^{\mathrm{T}} \tag{14}$$

Remark 1: Notably, the observed force errors come from two sources: the position error of the end-effector to the environment and the position error due to the environment change. Dynamic changes in the environment often bring about changes in other orthogonal force information. Therefore, instead of other iterative neural network controllers [46], [47] (which treat the system as a black box and train the network only with the inputs and outputs of the system), we train the network with the input u and the robot state x of the system, which reduces the loss of state information. On the other hand, this makes the feedback controller necessary since it contains information about the output of the system. We will further explain the reasons later in the study.

Our MLP network for iterative impedance learning uses the following structure:

$$u_p^f(t) = W_{2,p}^{\phi}\phi\big(W_{1,p}^{\boldsymbol{\nu}}\boldsymbol{\nu}_{p-1}(t) + W_{1,p}^{b}\big) + W_{2,p}^{b} \tag{15}$$

where $W_{1,p}^{\boldsymbol{\nu}} = [W_{1,p}^{u}, W_{1,p}^{x}]$, $W_{1,p}^{b}$ are the adaptive weight matrices connecting the input, states, and bias between the input layer and the hidden layer, $W_{2,p}^{\phi}, W_{2,p}^{b}$ are the adaptive weight weights and bias matrices between the hidden layer and the output layer, $\phi(\cdot)$ is the nonlinear activation function and hyperbolic tangent is chosen in this paper.

Although (15) has only one hidden layer, it can represent any nonlinear continuous function with assumed accuracy. Importantly, the proposed approach makes it possible to use more complex models.

The estimator in the DRL controller is designed to evaluate the performance of the actor. For a time series t within the pth trial, define the observatons

$$o_p(t) = \big[x_p(t), e_p(t)\big]^{\mathrm{T}} \tag{16}$$

and the immediate reward function

$$r_p(t) = r_{\mathrm{const}} - e_p^{\mathrm{T}}(t)R_e e_p(t) \tag{17}$$

where $r_{\mathrm{const}}(>0)$ is the positive reward, R_e is a diagonal matrix. The total discounted reward or the Q-function ($\gamma < 1$) is chosen as

$$Q\big(o_p(t), u_p(t)\big) = \mathbb{E}\Big[r_p(t) + \sum_{j=1}^{N-t-1}\gamma^j r_p(t+j)\Big] \tag{18}$$

and the Q-function satisfies the Bellman equation

$$Q\big(o_p(t), u_p(t)\big) = r_p(t) + \gamma Q\big(o_p(t+1), u_p(t+1)\big) \tag{19}$$

2) Control training: The dynamic changes in the environment resulted in different instantaneous output errors. For better evaluation of the control performance, the total expected cost J_p over the whole discrete time is used as the evaluation criterion, and the total expected cost is

$$J_{\mathrm{cost}}(\theta_p) = \sum_{t=0}^{N-1}\mathbb{E}\big[Q(t) + \beta\mathcal{H}\big(u_p(t)\big)\big] \tag{20}$$

where $\theta_p = [W_{1,p}, W_{2,p}]^T$ is the generalized network parameter vector, $\beta > 0$ is the temperature coefficient, and $\mathcal{H}(\cdot)$ is the entropy of the neural network controller output. Maximizing the entropy of the policy to better explore the impedance tuning policy. The goal of training is to find the corresponding optimal parameters θ^* that maximize J_{cost}

$$\theta^* = \arg\max_{\theta} J_{\mathrm{cost}}(\theta_p) \tag{21}$$

Then, according to the chain rule, the gradient of the expected cost concerning the controller hyperparameters θ_p is computed as

$$\Delta\theta_p = -\eta\frac{\partial J_{\mathrm{cost}}}{\partial\theta_p} \tag{22}$$

where η is the learning rate.

After several iterations, the network parameters θ_p are updated by back-propagation. A gradient-based policy search is applied to progressively approximate the optimal control parameters θ^*.

III. Control System Analysis

In this section, we qualitatively analyze the stability of the control system and the convergence of the DRL controller.

A. Preliminaries

In order to satisfy Assumption 1 that the initial state starts in the reference state, this requires that the initial state must be stable. Different initial impedance parameters will affect the effectiveness of the control. For this reason, we use spectral analysis to select the initial impedance parameters that satisfy the stabilization condition, at which point the DRL controller does not update the impedance parameters of the system. In the workspace, the reference trajectory error $\Delta\mathcal{X}_{\mathrm{r}}$ can be transformed into the frequency domain [25]

$$M\Delta\mathcal{X}(z)z^2 + B\Delta\mathcal{X}(z)z + K\Delta\mathcal{X}(z) = \Delta\mathcal{F}(z) \tag{23}$$

where z is the complex variable and the corresponding variables are capitalized to indicate the result of the transformation in the frequency domain.

The dynamics between the force tracking error ΔF and the positional perturbation ΔX is described by a second-order system with transfer function $\mathcal{G}(z) = [Mz^2 + Bz + K]^{-1}$. The control system adjusts the robot impedance parameters to generate the position correction $\Delta\mathcal{X}_\mathrm{r}$. $\Delta\mathcal{X}_\mathrm{r}$ modifies the reference trajectory to output the control command $\mathcal{X}_\mathrm{d} = \mathcal{X}_\mathrm{r} + \Delta\mathcal{X}_\mathrm{r} = \mathcal{X}_\mathrm{r} + \Delta\mathcal{F}\cdot\mathcal{G}(z)$. In position servo control mode, then $\mathcal{X}_\mathrm{m} = \mathcal{X}_\mathrm{d}$ holds.

$\mathcal{F}_\mathrm{env}$ is measured by the force sensor, i.e., $\mathcal{F}_\mathrm{m} = \mathcal{F}_\mathrm{env}$. From (2), the force tracking error can be rewritten as

$$\begin{aligned}\Delta\mathcal{F}(z) &= \mathcal{F}_\mathrm{m} - \mathcal{F}_\mathrm{d} = K_\mathrm{env}(\mathcal{X}_\mathrm{d} - \mathcal{X}_\mathrm{env}) - \mathcal{F}_\mathrm{d} \\ &= K_\mathrm{env}\big(\mathcal{X}_\mathrm{r} + \mathcal{G}(z)\Delta\mathcal{F}(z) - \mathcal{X}_\mathrm{env}\big) - \mathcal{F}_\mathrm{d}\end{aligned} \tag{24}$$

Substituting $\mathcal{G}(z) = [Mz^2 + Bz + K]^{-1}$ into (24), it is organized as

$$\Delta\mathcal{F}(z) = \frac{(Mz^2 + Bz + K)[K_\mathrm{env}(\mathcal{X}_\mathrm{r} - \mathcal{X}_\mathrm{env}) - \mathcal{F}_\mathrm{d}]}{Mz^2 + Bz + K - K_\mathrm{env}} \tag{25}$$

Then from (25), the steady-state force tracking error is obtained as follows

$$\Delta f_\mathrm{ss} = \lim_{z\to 0}\Delta\mathcal{F}(z) = \frac{K\cdot[K_\mathrm{env}(\mathcal{X}_\mathrm{r} - \mathcal{X}_\mathrm{env}) - \mathcal{F}_\mathrm{d}]}{K - K_\mathrm{env}} \tag{26}$$

For the force tracking steady-state error to be zero, the reference position should satisfy $K_\mathrm{env}(\mathcal{X}_\mathrm{r} - \mathcal{X}_\mathrm{env}) - \mathcal{F}_\mathrm{d} = 0$ or $K = 0$. However, x_env and K_env of the environment considered in this paper are not known in advance. Any change in the environment information will bring about corresponding changes in the steady-state error. And if setting the stiffness parameter to zero, we have $\Delta f_\mathrm{ss} = 0$ as an equilibrium state.

The same is true for the other dimensions due to the orthogonality and independence of the Cartesian variables in the workspace. Therefore, $K = 0_{n\times n}$ will be exploited in the subsequent system analysis.

B. Lyapunov Stability Analysis

Considering the manipulator in (4), the counterpart of the dynamic model can be described as

$$D(q)\ddot{q} + C(q,\dot{q})\dot{q} + g(q) + \tau_f(\dot{q}) = \tau - J^\mathrm{T} f_\mathrm{m} \tag{27}$$

where $D(q) \in \mathbb{R}^{n\times n}$ is the inertia matrix, $C(q,\dot{q}) \in \mathbb{R}^{n\times n}$ is the nonlinear term matrix of Coriolis and centrifugal forces, $g(q) \in \mathbb{R}^n$ is the gravity vector, $\tau_f \in \mathbb{R}^n$ is the torque due to joint friction, $\tau \in \mathbb{R}^n$ is the driving torque, and the external force $f_\mathrm{m} \in \mathbb{R}^m$ is measured by the force sensor. Usually, the robot's gravity term can be obtained with sufficient accuracy, i.e. $\hat{g}(q) = g(q)$.

Then, some useful properties of the dynamic model (27) for analyzing stability are presented below [6]:

Property 3: $D(q)$ in (27) is symmetric and positive definite. If $d_\mathrm{min}(d_\mathrm{max})$ denotes its minimum (maximum) eigenvalue, then it is

$$0 < d_\mathrm{min} I \le D(q) \le d_\mathrm{max} I \tag{28}$$

with $d_\mathrm{max} < \infty$

Property 4: The matrix $N(q,\dot{q}) = \dot{D}(q,\dot{q}) - 2C(q,\dot{q})$ is skew-symmetric, i.e., for all $\xi \in \mathbb{R}^n$, the following equation holds

$$\xi^T N(q,\dot{q})\xi = 0 \tag{29}$$

Theorem 1: Consider the system (7) with property 2 and the controlled object (27) with Properties 3-4. The initial damping matrix $B(0)$ is chosen as a positive definite matrix, then $\Delta f_\mathrm{ss} \to 0$ is asymptotically stable if

$$\lambda_\mathrm{min}\{\Delta B(t)\} \ge 0,\ \forall t \in \{0,...,N-1\} \tag{30}$$

where $\Delta B(t)$ is the action defined in (5) and $\lambda_\mathrm{min}\{\cdot\}$ denotes the minimum eigenvalue of a matrix.

Proof: As a simplified dynamic tracking process, only viscous friction is considered, i.e., $\tau_f(\dot{q}) = F_f\dot{q}$. Plugging (4) in (27) gives

$$D(q)\ddot{q} + C(q,\dot{q})\dot{q} + J^\mathrm{T} f_\mathrm{m} = J^\mathrm{T}\alpha \tag{31}$$

Let $\dot{e}_q = \dot{q} - \dot{q}_\mathrm{r}$ and $e_x = x_\mathrm{m} - x_\mathrm{env}$ denote the configuration space velocity error and task space position error, respectively, where $\dot{q}_\mathrm{r}$ is the reference angular velocity vector. property 2 ensures the unique existence of the optimal reference trajectory x_r and leads to $\dot{x}_\mathrm{r} = J\dot{q}_\mathrm{r} = \dot{x}_\mathrm{env}$.

Consider the Lyapunov function candidate as follows

$$\mathcal{V} = \frac{1}{2}\dot{e}_q^\mathrm{T} D(q)\dot{e}_q + \frac{1}{2}e_x^\mathrm{T} K_\mathrm{env} e_x \tag{32}$$

Since K_env is positive definite and property 3, we have

$$\mathcal{V} \ge \frac{1}{2}d_\mathrm{min}\|\dot{e}_q\| + \frac{1}{2}\lambda_\mathrm{min}\{K_\mathrm{env}\}\|e_x\| > 0 \tag{33}$$

Based on the preliminary analysis, choosing

$$\alpha = J^{-\mathrm{T}}(D\ddot{q}_\mathrm{r} + C\dot{q}_\mathrm{r}) - B(t)\dot{e}_x \tag{34}$$

where $B(t) = \Delta B(t-1) + B(t-1)$ is the variable impedance parameter in the controller α.

In deriving (34), the inverse of the Jacobian matrix can be computed directly for a nonredundant manipulator $(n = m)$ moving in a singularity-free region of the workspace, whereas the pseudo-inverse can be used in the redundant case $(n > m)$ in conjunction with a suitable term in the null space of the Jacobian matrix describing the internal motion of the manipulator [6]. Folding (34) into (31) gives

$$D\ddot{e}_q = -C\dot{e}_q - J^\mathrm{T}\big(f_\mathrm{m} + B(t)\dot{e}_x\big) \tag{35}$$

Computing the time derivative of (32) and substituting (35) into it. From (2) and property 4, we have

$$\begin{aligned}\dot{\mathcal{V}} &= \frac{1}{2}\dot{e}_q^\mathrm{T}\dot{D}\dot{e}_q + \dot{e}_q^\mathrm{T} D\ddot{e}_q + \dot{e}_x^\mathrm{T} K_\mathrm{env} e_x \\ &= \dot{e}_q^\mathrm{T}(\frac{1}{2}\dot{D} - C)\dot{e}_q + \dot{e}_x^\mathrm{T}\big(K_\mathrm{env}e_x - f_\mathrm{m} - B(t)\dot{e}_x\big) \\ &= -\dot{e}_x^\mathrm{T} B(t)\dot{e}_x\end{aligned} \tag{36}$$

Thanks to the positive definiteness of the initial damping matrix $B(0)$, we have $\big|B(1)\big| = \big|\Delta B(0) + B(0)\big| > \max\big\{\big|\Delta B(0)\big|, \big|B(0)\big|\big\} > 0$, i.e., $B(1)$ is also positive definite. Following mathematical induction, it can be concluded that maintaining $\Delta B(t)$ semi-positive definite yields a positive definite $B(t)$, which implies a negative definite $\dot{\mathcal{V}}$. Combining (33) and using the Lyapunov stability criterion, we get that the system is asymptotically stable and it will always return to force equilibrium state. Thus $f_\mathrm{m} \to f_\mathrm{d}$ is asymptotically stable, i.e., $\Delta f_{ss} \to 0$ is asymptotically stable. This sets the constraints for impedance learning and thus ensures the stability of the DRL variable impedance control system.

C. DRL Controller Convergence Analysis

Although we discussed control system stability in the previous section, it is expected that the DRL controller can explore the optimal state trajectory to the equilibrium state after impedance learning, i.e., converge to the optimal controller. This paper focuses on two convergence properties:

Property 5: Input optimality

$$\lim_{p\to\infty} u_p(t) = u^*(t) \tag{37}$$

Property 6: Tracking zero error property

$$\lim_{p\to\infty} e_p(t) = 0, \forall t \in \{0, ..., N-1\} \tag{38}$$

Before starting the convergence analysis of impedance learning, we introduce the following definition.

Definition 1 [49]: Assuming a set of random vectors $\boldsymbol{v}(t)$ concerning $t \in \{0, ..., N-1\}$, define the λ-norm of $\boldsymbol{v}(t)$ as

$$\|\boldsymbol{v}(t)\|_\lambda = \sup_t a^{-t\lambda}\mathbb{E}\,\|\boldsymbol{v}(t)\| \tag{39}$$

with $a > 1$ and a finite constant $\lambda > 0$, where $\mathbb{E}\,\|\cdot\|$ denotes the Euclidean norm of the mathematical expectation of a random vector.

Now, we can give sufficient conditions for the convergence of our DRL control system.

Theorem 2: Consider the system (7) with Properties 1-2 and Assumption 1 holds, then the DRL feedforward controller (13) with regressor (14) satisfies convergence property 5 if

$$\sup_t \|g_u(t)\| < 1 \tag{40}$$

where

$$g_u(t) = \frac{\partial g\big(u_p(t), x_p(t)\big)}{\partial u_p(t)}, g_x(t) = \frac{\partial g\big(u_p(t), x_p(t)\big)}{\partial x_p(t)} \tag{41}$$

denote the partial derivatives of the DRL controller concerning the input error and the tracking error, respectively.

Proof: For the system (7) with property 2 holding, the error of the pth iteration result relative to (9) is given by

$$\begin{aligned} \Delta s_p(t+1) &= h\big(s_p(t), u_p(t)\big) - h\big(s_{\mathrm{r}}(t), u^*(t)\big) \\ \Delta y_p(t) &= K_c \Delta s_p(t) \end{aligned} \tag{42}$$

where state error $\Delta s_p = s_p - s_{\mathrm{r}} = [x_m - x_r,\ 0]^{\mathrm{T}} = [\Delta x_p,\ 0]^{\mathrm{T}}$ and $\Delta y_p = e_p$ defined in (10).

Let $\Delta u_p = u_p - u^*$, on the assumption of property 1, taking the Euclidean norm for both sides of the state-space equation (42) is given

$$\begin{aligned} \|\Delta x_p(t+1)\| &= \|\Delta s_p(t+1)\| \\ &\le l\big(\,\|\Delta s_p(t)\| + \|\Delta u_p(t)\|\,\big) \end{aligned} \tag{43}$$

From the same initial state of Assumption 1, it follows that $\|\Delta s_p(0)\| = 0$. Noticing the recursive nature of (43) yields

$$\|\Delta x_p(t)\| \le \sum_{i=0}^{t-1} l^{t-i}\,\|\Delta u_p(i)\| \tag{44}$$

Since the feedback control signal $u^b(t)$ is not updated during impedance learning, $\Delta u_p(t) = \Delta u_p^f(t)$. Applying Taylor's formula to the DRL controller (13) with regression vector (14) and taking the mathematical expectation of the Euclidean norm, yields

$$\mathbb{E}\,\|\Delta u_{p+1}(t)\| \le \mathbb{E}\,\|g_u \Delta u_p(t)\| + \mathbb{E}\,\|g_x \Delta x_p(t)\| \tag{45}$$

In view of (44), the inequality (45) can be rewritten as

$$\begin{aligned} \mathbb{E}\,\|\Delta u_{p+1}(t)\| \le\ & \|g_u\|\,\mathbb{E}\,\|\Delta u_p(t)\| \\ &+ \|g_x\|\Big(\sum_{i=0}^{t-1} l^{t-i}\mathbb{E}\,\|\Delta u_p(i)\|\Big) \end{aligned} \tag{46}$$

Applying the λ-norm to (46), i.e., multiplying both sides of the inequality by the same $a^{-k\lambda}$ and then taking the upper bound, gives

$$\begin{aligned} \|\Delta u_{p+1}(t)\|_\lambda \le\ & \sup_t \|g_u\| \cdot \|\Delta u_p(t)\|_\lambda \\ &+ \sup_t \|g_x\| \cdot \sup_t a^{-t\lambda}\Big(\sum_{i=0}^{t-1} l^{t-i}\mathbb{E}\,\|\Delta u_p(i)\|\Big) \end{aligned} \tag{47}$$

Let $a > \max\{1, l\}$, then the last term in (47) can be simplified as

$$\begin{aligned} &\sup_t a^{-t\lambda}\Big(\sum_{i=0}^{t-1} l^{t-i}\mathbb{E}\,\|\Delta u_p(i)\|\Big) \\ &\le \sup_t \Big(\sum_{i=0}^{t-1} a^{-(\lambda-1)(t-i)}\Big) \sup_t \Big(a^{-\lambda i}\mathbb{E}\,\|\Delta u_p(i)\|\Big) \\ &\le \frac{1 - a^{-(\lambda-1)N}}{a^{\lambda-1} - 1}\,\|\Delta u_p(t)\|_\lambda \end{aligned} \tag{48}$$

Substituting (48) into (46) yields

$$\|\Delta u_{p+1}(t)\|_\lambda \le \big(\rho_u + \rho_x \frac{1 - a^{-(\lambda-1)N}}{a^{\lambda-1} - 1}\big)\,\|\Delta u_p(t)\|_\lambda \tag{49}$$

where

$$\rho_u = \sup_t \|g_u(t)\|, \rho_x = \sup_t \|g_x(t)\| \tag{50}$$

If (40) holds, i.e., $\rho_u < 1$, then there always exists a sufficiently large λ such that $\rho_x \frac{1-a^{-(\lambda-1)N}}{a^{\lambda-1}-1} < 1 - \rho_u$, which further gives

$$\|\Delta u_{p+1}(t)\|_\lambda \le \rho\,\|\Delta u_p(t)\|_\lambda \tag{51}$$

holds, where

$$\rho = \rho_u + \rho_x \frac{1 - a^{-(\lambda-1)N}}{a^{\lambda-1} - 1} < 1 \tag{52}$$

Thus, from the contracting mapping principle it follows that

$$\lim_{p\to\infty} \|\Delta u_p(t)\|_\lambda = 0, \forall t \in \{0, ..., N-1\} \tag{53}$$

Noticing $\|\Delta u_p(t)\| \ge 0$ and the finiteness of t, it can be concluded that

$$\lim_{p\to\infty} \|\Delta u_p(t)\| = 0, \forall t \in \{0, ..., N-1\} \tag{54}$$

which implies $\lim_{p\to\infty} u_p(t) = u^*(t)$ holds for all $\forall t \in \{0, ..., N-1\}$ and thus input optimality i.e. property 5 is satisfied.

Remark 2 [47]: Based on the above analysis, to guarantee the convergence of impedance learning, the MLP network in (15) of the DRL feedforward controller so as to satisfy

$$\sup_t \|g_u(t)\| = \sup_t \left\| \left(\left(W_{2,p}^{\phi} \right)^{\mathrm{T}} \circ \phi' \right)^{\mathrm{T}} W_{1,p}^{u} \right\| < 1 \tag{55}$$

where $\phi' = \frac{\partial \phi}{\partial u} \in \mathbb{R}^{h \times m}$ is the activation function derivative and $\circ$ is the element-wise product.

Thanks to the chosen hyperbolic tangent activation function, the range of ϕ' is restricted to $(0, 1)$. Thus, the convergence condition (55) of the DRL controller can be simplified as

$$\sup_t \left\| W_{2,p}^{\phi} W_{1,p}^{u} \right\| < 1 \tag{56}$$

We adopt weight normalization [50] and set a reasonable initial weight matrix to satisfy (56), thus guaranteeing the convergence of the output of the DRL controller during the impedance learning process. This provides a theoretical basis for the convergence of impedance learning, which is no longer an unknown black box.

Corollary 1: Given the nonlinear system (7) with the DRL feedforward controller (13) satisfying theorem 2, it tracks zero error at convergence (property 6).

Proof: Apply the λ-norm to (44) and let $a > \max\{1, L\}$, the derivation along (48) readily yields

$$0 \leq \|\Delta s_p(t)\|_\lambda \leq \frac{1 - a^{-(\lambda-1)N}}{a^{\lambda-1} - 1} \|\Delta u_p(t)\|_\lambda \tag{57}$$

From (53), $\forall \varepsilon > 0$, the λ-norm of the input error for the system that satisfies theorem 2 is less than ε when p is sufficiently large. Thus, by the squeeze theorem, it follows that

$$\lim_{p \to \infty} \|\Delta s_p(t)\|_\lambda = 0, \forall t \in \{0, ..., N-1\} \tag{58}$$

Similarly, since $\|\Delta s_p(t)\| \geq 0$ and t is finite, it implies that

$$\lim_{p \to \infty} \|\Delta s_p(t)\| = 0, \forall t \in \{0, ..., N-1\} \tag{59}$$

Folding (59) into (42), it is easy to obtain that

$$\lim_{p \to \infty} \|e_p(t)\| = \lim_{p \to \infty} \|\Delta y_p(t)\| = 0, \forall t \in \{0, ..., N-1\} \tag{60}$$

which implies that $\lim_{p \to \infty} e_p(t) = 0$ for all $t = 0, ..., N-1$, i.e., property 6 holds. The proof is thus completed.

IV. CoppeliaSim Simulation Studies

In order to verify the proposed control scheme and analysis, a series of simulation studies are conducted and presented in this section. Firstly, the learning procedure for impedance tuning is designed and the stopping criterion is set. Then simulations are performed to test the effectiveness of the proposed control scheme and verify the convergence analysis, which covers most of the actual force-tracking scenarios.

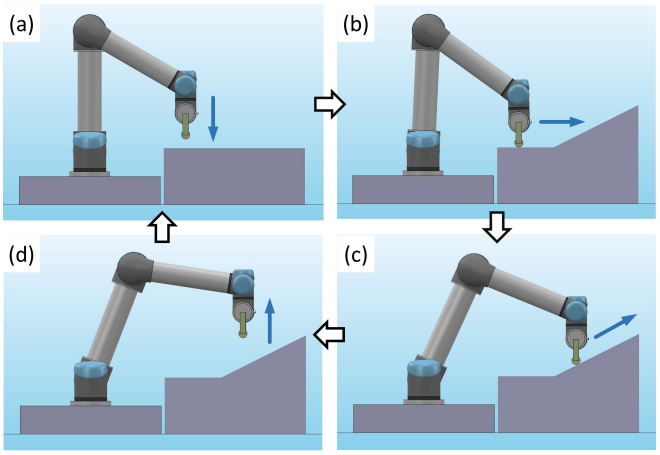

Fig. 4. CoppeliaSim model in different phases. (a) Random terrain, (b) contact interaction, (c) feedforward correction, (d) soft update networks

A. CoppeliaSim Model Setup

We studied the force-tracking control problem using CoppeliaSim, a well-established simulator in the field of robotics. The robot-environment interaction model, as shown in Fig. 4, consists of a manipulator arm with a fixed base and different terrains generated with random polynomial parameters. In this paper, the digital twin model of Universal Robots UR5 is adopted to illustrate the implementation. The model settings, such as kinematics and dynamics parameters, are modeled with reference to the real robot and pre-compensated for tool end force errors. The position and stiffness parameters of the environment are unknown, but the initial position of the environment, i.e., the initial state, is constant. In this study, the impedance parameters of the robot are tuned by the DRL feedforward controller for each sampling time $t \in \{0, ..., N-1\}$. The weight parameters of the neural network model are updated for each training cycle $p \in \mathbb{R}$. The updated DRL feedforward controller replaces the target neural controller for the current training cycle and opens the next round of impedance learning training. The updating rules are based on the principle of gradient descent besides satisfying theorem 1, 2 and corollary 1.

In addition, artificial Gaussian noise is added to state s in CoppeliaSim to simulate sensor noise, with the sensor noise variance set to 30% of the performance tolerance bound. The state s is defined in (16). We also added collision calculations during model contact interactions, which is a consideration to make the simulation more realistic.

B. Simulation Procedure

Algorithm 1 summarizes how our proposed DRL variable impedance controller implements impedance parameter tuning and network updating. In order to normalize the data types, Algorithm 1 normalizes the state parameter s and maps the action parameter ΔI within the action bounds $[-1, 1]$ by the common logarithmic relation. Set the reward decay coefficient $\gamma = 0.99$ and the learning rate $\eta = 0.001$. The built-in actor-critic network is updated at each training cycle until the training is successful. The simulation study consists of

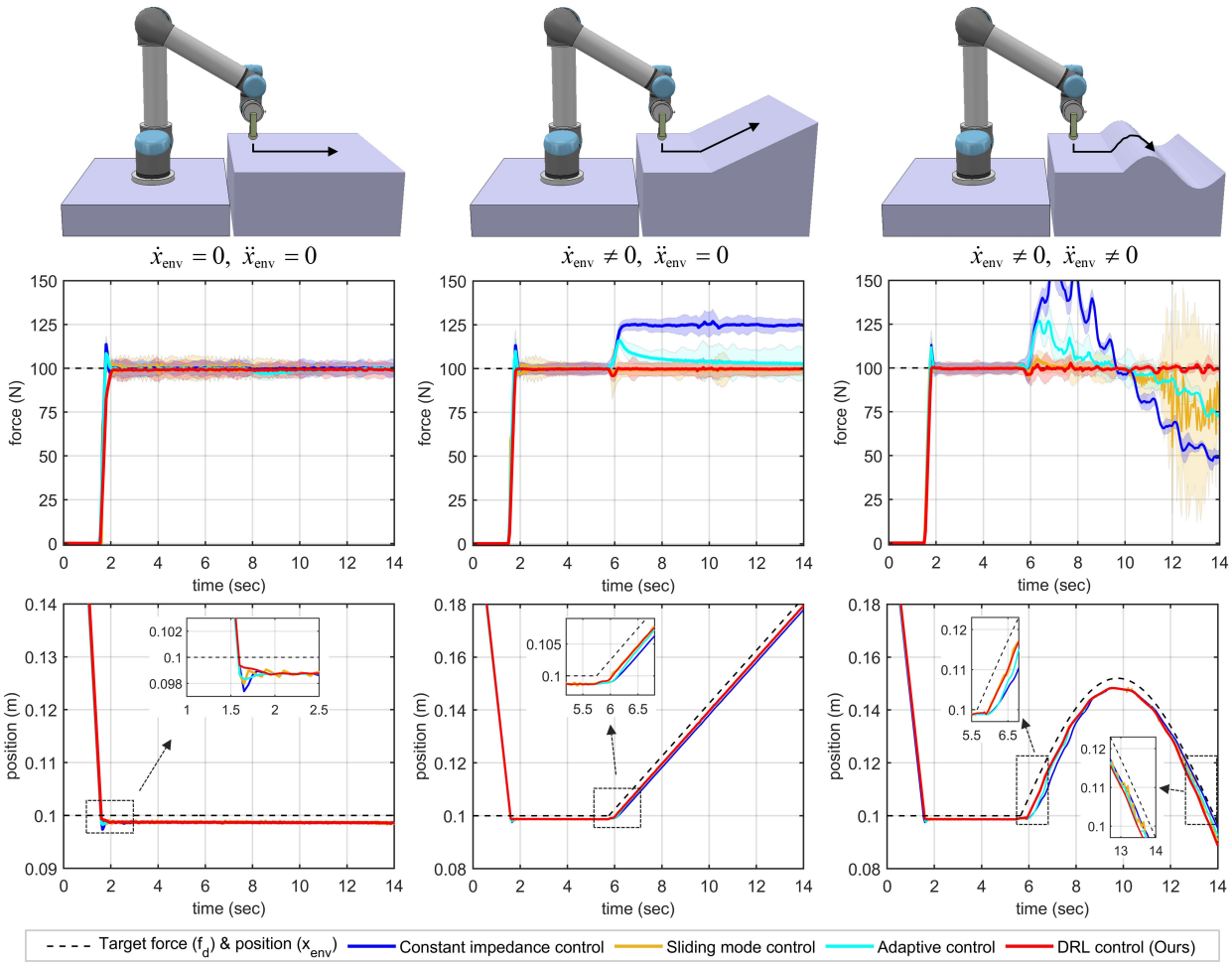

Fig. 5. Statistical comparisons between different control methods. The first row shows typical planar, oblique and sinusoidal environments. The second and third rows of plots reflect the measured force (f_{m}) and the measured position (x_{m}) changes under different control methods, respectively. It can be seen that our method (DRL control) significantly outperforms other methods in force tracking in different environments (especially for nonlinear environments).

TABLE II
COMPARISON BETWEEN VARIOUS VARIABLE IMPEDANCE CONTROL METHODS FOR DIFFERENT SURFACES

Control policy	Plane			Slope surface	Sinusoidal surface
	Setting time (s)	Overshoot (%)	Force error (N)	Force error (N)	Force error (N)
Constant impedance	0.39 ± 0.02	13.6 ± 1.6	$\mathbf{-0.1 \pm 0.7}$	23.8 ± 4.6	24.1 ± 21.0
Sliding mode control [27]	0.61 ± 0.12	9.6 ± 8.0	-0.9 ± 1.6	-0.7 ± 1.2	-17.2 ± 33.7
Adaptive control [25]	0.37 ± 0.03	8.4 ± 2.1	-0.7 ± 0.7	4.6 ± 3.1	9.1 ± 10.9
DRL control (Ours)	$\mathbf{0.33 \pm 0.05}$	$\mathbf{3.8 \pm 2.9}$	-0.4 ± 1.0	$\mathbf{-0.3 \pm 1.1}$	$\mathbf{-0.8 \pm 3.7}$

two main parts: training and pre-tests. 1) Training: The DRL variable impedance controller is applied to explore the optimal policy for tuning the impedance parameters, i.e., impedance learning, in a randomly generated terrain scene. The maximum number of training sessions is set to be 900, the maximum sampling batch is 350, and the sampling interval is 0.05s. Provide that when the average force error for 10 consecutive complete samples is less than 3% of the desired force and the overshoots $\overline{\sigma}\%$ are all less than 75%, the training is done. That is, the current DRL variable impedance controller is considered to have converged to the optimal impedance controller within the tolerance error ε_e. The upper limit of overshooting is set to avoid unacceptable overshooting behavior of the agent. 2) Pre-tests: The tests are mainly employed to illustrate the effectiveness, convergence and robustness of the proposed control method. Firstly, three typical surface terrains are selected to test the force-tracking effect of the DRL variable impedance controller and compared with other control methods to illustrate the effectiveness and advantages. Then, the tracking force error of the controller at different training stages is tested to verify the convergence. Finally, we test for terrain not present in the training dataset to illustrate the robustness of the controller.

C. Simulation Results

1) Force tracking tests: In order to evaluate the control effect of the proposed method, we tested the trained DRL

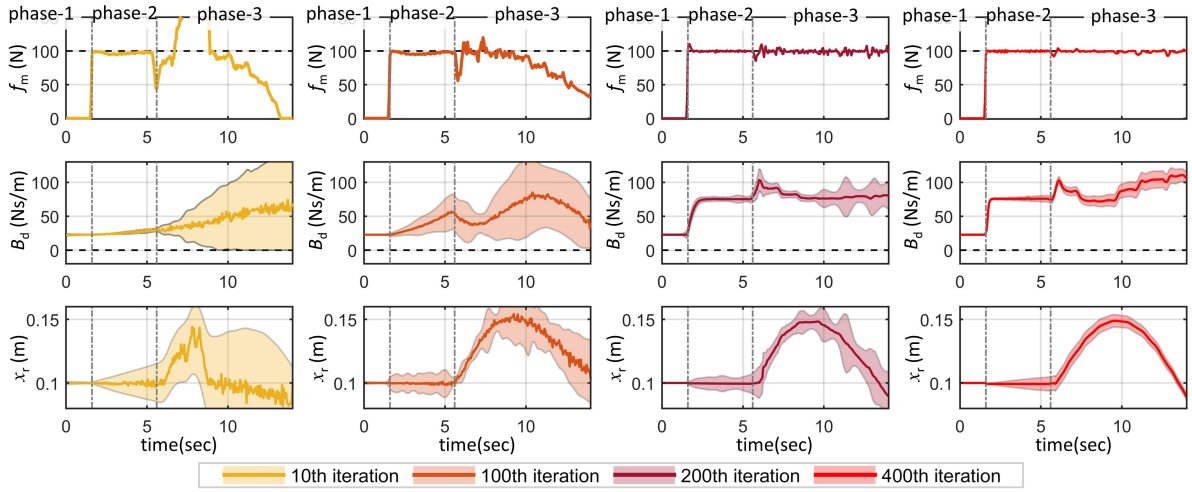

Fig. 7. Impedance parameter evolution and force control process on the sinusoidal plane. The solid line and shading are the desired value and 95% confidence interval of the DRL controller output, respectively. The first row is the contact force variation, the second row is the profile of the desired damping B_{d}, and the third row is the profile of the reference trajectory x_{r}.

Algorithm 1 Impedance Learning for Force Tracking

Input: Target output y_r, tolerance error ε_e
Output: Optimal weights (within tolerance) θ^*

Initialization: Iteration index $p \leftarrow 0$; Initialize the normalized weight network θ_0 by (56), impedance parameters $I(0)$ and state-action value $Q(0)$; Batch index i_{bat}
repeat
  Set Initial States: Time index $t \leftarrow 0$; Initial States $s_p(0) \leftarrow s_r(0)$; Generate x_{env} with random polynomial interpolation
  repeat
    Take Action: Generate feedforward input $u_p^f(t)$ by (15); Compute impedance parameters $I(t)$ by (6); Obtain reward r_p and observations $o_p(t)$, $o_p(t+1)$
    Evaluate: Add input $u_p(t)$, observations $o_p(t)$, $o_p(t+1)$ and reward $r_p(t)$ to the replay buffer; Compute Q function error by (19) and cost $J_{\mathrm{cost}}(\theta_p)$ by (20)
    $t \leftarrow t+1$
  until $t > N$
  if Buffer size $N_p \geq i_{\mathrm{bat}} N$ **then**
    Update Controller: Update network weights θ_p by (21) and (22)
    $\overline{y}_p \leftarrow \frac{1}{N_p}\sum_t y_p(t)$
    $p \leftarrow p+1$
  end if
until $\left\|\overline{y}_p - y_r\right\| < \varepsilon_e$
return $\theta^* \leftarrow \theta_p$

variable impedance control on different terrains for force tracking. We selected three typical terrain scenarios: planes ($\dot{x}_{\mathrm{env}} = 0, \ddot{x}_{\mathrm{env}} = 0$), slope surface ($\dot{x}_{\mathrm{env}} \neq 0, \ddot{x}_{\mathrm{env}} = 0$) and complex surface ($\dot{x}_{\mathrm{env}} \neq 0, \ddot{x}_{\mathrm{env}} \neq 0$). And, the force-tracking effect of our method is compared with classical constant impedance control, adaptive control [25] and sliding mode variable impedance control [27]. Fig. 5 shows the force

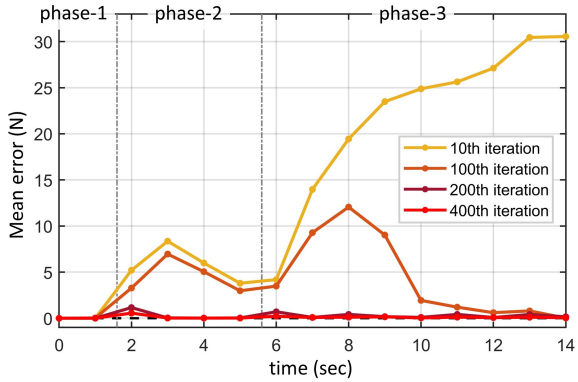

Fig. 6. The mean absolute error of force tracking over the whole time interval at the $10th$, $100th$, $200th$ and $400th$ iterations.

tracking results of different control methods in one complete sampling period.

The above simulation results show that the constant impedance control tracks well in the plane, but is poorly robust to changing terrain ($\dot{x}_{\mathrm{env}} \neq 0$) and suffers from steady-state errors. The sliding mode control is robust to terrain with constant variation ($\ddot{x}_{\mathrm{env}} = 0$) but appears to oscillate when contacting surfaces with specific slopes. This is caused by the oscillation of the state parameters in the neighborhood of the predesigned sliding mode surface. This explains why the sliding mode control in Table II performs well on planar and most of the slopes but less well on sinusoidal surfaces. The adaptive variable impedance control proposed by Duan changes the impedance parameters by force error information, which can gradually approach the target value while ensuring stability. However, it suffers from time lag because the feedback force comes from the current or past moments, and the current impedance parameters do not always meet the requirements, thus failing to track large environmental perturbations in time. In contrast, our proposed variable impedance control based on DRL feedforward not only performs well under structured terrain, but

TABLE III
STATISTICAL RESULTS OF FORCE TRACKING IN DIFFERENT TERRAINS

K_{env} (N·m^{-1})	Metric				
	setting time (s)	over-shoot (%)	force error (N) $f_d = 10$	$= 25$	$= 50$
25000	**0.44**	7.8	**0.25**	**0.44**	**0.75**
2500	0.51	5.3	0.28	0.51	0.87
250	0.53	**5.1**	0.33	0.69	0.85

is also robust to uncertainly varying terrain. Although terrain variations are observable but uncontrollable perturbations, the DRL controller generates feedforward signals to compensate for the perturbations based on the impedance tuning model obtained from previous iterative learning. As shown in Table II, our method reduces the force error on nonlinear surfaces by 62.8% compared to other variable impedance controls, while taking into account the necessary metrics such as setting time, maximum overshoot, and oscillation suppression.

2) Controller convergence: To verify the convergence of the DRL feedforward controller, we tested the tracking force error with different numbers of iterations on the same set of surfaces (containing planar, inclined, and higher-order surfaces) satisfying Properties 1-2. Fig. 6 shows the force tracking results for different number of iterations. It is observed that the force tracking error gradually decreases with the increase in the number of iterations. At the $200th$ iteration, the tracking error is already acceptable.

Taking the sinusoidal surface as an example, Fig. 7 shows the force control and impedance parameter evolution process under different numbers of iterations. As can be seen from the results, in the early stage of learning, the input uncertainty of the DRL feedforward model is large due to the insufficient data collected. Thanks to the setting of the action exploration boundary (the second row in Fig. 7, the damping parameter needs to be kept positive), more useful data can be obtained during the interaction process. With the continuous enrichment of the interaction data, the learned DRL model continuously improves the control policy and tunes the distribution of the impedance parameters to achieve better tracking results. After 400 iterations, the DRL controller has converged to the desired impedance characteristics (fourth column in Fig. 7). As shown in that column, during free space (phase 1), the low damping allows the end to reach the contact surface quickly. After contact with the environment (phase 2), the rapidly increasing high damping allows the end to reach the desired contact force quickly and suppresses oscillations. When the environment changes (phase 3), accurate correction of the reference trajectory and corresponding damping allows the end to maintain stable tracking of the contact force. Note that the reference trajectory at this point is not equivalent to the stationary state of the environment; the difference between them is $K_{\text{env}}^{-1} f_{\text{d}}$.

We conducted ablation studies to show the effect of the choice of regression vector $\boldsymbol{\nu}(t)$ on the convergence of the DRL controller. Fig. 8 illustrates the variation of the average reward along the iterations. The average reward of the other control methods is also added on. DRL control policies that rely only on the inputs and outputs of the system often

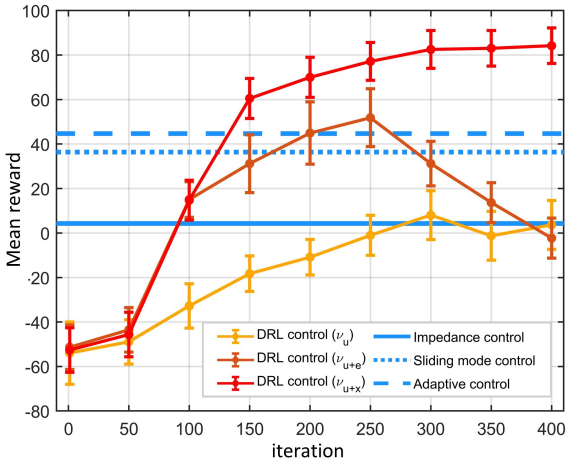

Fig. 8. Comparison of average rewards for DRL control with different regression vectors $\boldsymbol{\nu}_u = u$, $\boldsymbol{\nu}_{u+e} = [u, e]^{\mathrm{T}}$ and $\boldsymbol{\nu}_{u+x} = [u^f + u^b, s]^{\mathrm{T}}$ (Ours). Thanks to the negative correlation between the reward and the square of the force error, the larger cumulative rewards imply the smaller force error, i.e., the better control results. It can be seen that the regression vector consisting of system inputs and states enables the DRL feedforward controller to converge to the optimum asymptotically, effectively avoiding overfitting.

result in non-convergence or overfitting. Higher reward is easily obtained when the magnitude of terrain change is small ($\dot{x}_{\text{env}} \approx 0$). Due to the lack of regression on the system state s, the $\boldsymbol{\nu}_u$-type and $\boldsymbol{\nu}_{u+e}$-type regressors overestimate the input u at this point, which leads to overfitting. On the other hand, our proposed DRL control policy combining neural network feedforward with variable impedance feedback can solve this problem well. Through the impedance learning process, the DRL feedforward controller can obtain the position correction Δx_r by adjusting the impedance parameters to track the terrain change Δx_{env}. Therefore, the system observable state s is necessary for the convergence of the regressors. Theorem 2 provides a theoretical justification for this while the ablation experiments verify Theorem 2 and Corollary 1.

3) Generalization: Further, to evaluate the generalization effect of the proposed DRL controller, we performed the generalizability test on different terrains outside the training set, respectively. The range of environmental stiffness (N/m) is set to $[250, 25000]$ and the range of target force (N) is set to $[10, 50]$. As shown in Table III, the mean force tracking error of the DRL controller is less than 5% of the maximum desired force under different terrains and can quickly stabilize and suppress overshooting. This is because the feedforward controller accurately fits the relationship between the environmental rate of change Δx_{env} and the impedance parameter tuning ΔI during the impedance learning process. Since we model the interaction process as a Markov decision process with the same initial state, then for any realizable desired force, there always exists a unique interaction position on the Lipschitz continuous terrain that makes the feedback force equal to the desired force. Thus, the satisfaction of Assumption 1 and Properties 1-2 gives the DRL controller the ability to generalize to unknown environments.

In summary, simulation studies show that the proposed DRL variable impedance controller can achieve excellent tracking

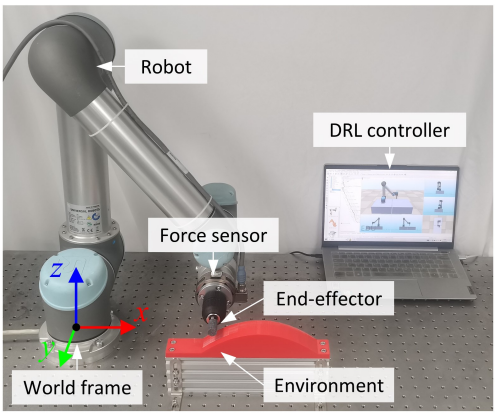

Fig. 9. Experimental hardware setup. DRL controller is integrated in laptop.

results in uncertain environments.

V. Experiments Verification

A. Experiment Setup

In this section, experiments are conducted on the test-bed shown in Fig. 9 to verify the practical effectiveness of the proposed DRL controller. The UR5 robot was chosen to avoid the possible influence on the impedance learning results due to the difference in dynamics parameters between the real robot and the virtual digital model. The robot arm is equipped with an ATI-Gamma force sensor (the nominal capacity of f_z is 400 N and the accuracy is 0.05 N). The self-developed DRL controller was used to control the robot, and the control period was set at 0.05 s.

The experiments included pre-calibration experiments and terrain experiments. a) Calibration experiments. Before performing force-tracking experiments on different terrains, a calibration step is required to compensate for the gravitational and frictional moments of the end tool. Based on the given desired force range, vertical contact experiments without feedforward inputs are performed in the target environment. The contact experiment calibrates the initial impedance parameters in the feedback controller, which in turn yields a range for impedance tuning. On the other hand, the measured position where the measured force is equal to the desired force is set as the initial reference position to meet the assumption 1 condition. The pre-training results obtained from the simulation then need fine-tuning. A random exploration is required for RL. To ensure training safety, fine-tuning is therefore performed in a virtual environment. The maximum episodes are set to 50 and the training time for each episode is determined by a single task cycle. Therefore, the fine-tuning process can always be completed in a limited time. In all controller training, the objective defined by the reward function remains unchanged: tracking expected force in different environments while achieving high robustness and energy optimization. The learning rate η and reward decay coefficient γ also remain unchanged during the process. A large enough boundary I_{inf} is chosen to replace ∞ for implementation convenience. b) Terrain experiments. Variable impedance feedback signals and fine-tuned neural network feedforward signals are integrated into the DRL controller. The experiments comprise three different scenes: a static environment, a quasi-static environment and a dynamic environment. To evaluate the proposed impedance learning-based DRL control strategy, the following three control conditions are compared,

- *Constant impedance control*: Set $B = k_b I$ and $K = k_k I$, where k_b and k_k are constants.
- *Adaptive control*: The velocity-based adaptive control is proposed in [25]. This method is to generate adaptive velocity adjustment from force error.
- *DRL control*: The impedance learning-based DRL control method proposed by us is used.

B. Results and Discussion

Scene 1: Robot machining. Because the material removal by grinding is very small, the workpiece being ground can be considered as a static environment. The designed grinding machining experiments require the end-effector to maintain as constant a force as possible on the surface of the workpiece to guarantee the grinding accuracy. The workpiece is a resin plate fixed to an aluminum alloy with dimensions $280 \times 40 \times 50$ mm and its stiffness is unknown. A grinding head is added to the robot end-effector, and the diameter of the grinding head is 16 mm. After the calibration experiments are completed, three typical terrains are selected for robotic grinding machining tests. To maintain the consistency of the initial state, set $s_p(0) = s_r(0) = [0.15, 0.1]^{\mathrm{T}}$ m. In other words, the initial position of the robot and the environment is constant and the robot is not in contact with the environment at the beginning.

If the contact environment is a flat terrain, it satisfies $\dot{x}_{\text{env}} = 0$, $\ddot{x}_{\text{env}} = 0$. To test the robustness of different control strategies to the environmental stiffness change ΔK_{env}, the suddenly changed environmental stiffness is

$$K_{\text{env}} = \begin{cases} k_{\text{env0}} & , \ 0 \ s \leq t \leq 8 \ s \\ k_{\text{env}}(i) & , \ 8 \ s \leq t \leq 16 \ s \end{cases}$$

where k_{env0} is the initial environmental stiffness, i is the number of experimental groups, and satisfies $k_{\text{env}}(1) < k_{\text{env}}(2) = k_{\text{env0}} < k_{\text{env}}(3) < k_{\text{env}}(4) < k_{\text{env}}(5)$. From Fig. 10(a), it can be seen that the desired force can be tracked after a short overshoot when the environmental stiffness changes. The smaller the change in stiffness, the smaller the overshoot. The three control methods are robust to stiffness changes in the environment.

If the contact environment is a sloped terrain, it satisfies $\dot{x}_{\text{env}} \neq 0$, $\ddot{x}_{\text{env}} = 0$. The robot moves from a flat surface to an sloped surface with an inclination angle of θ. Fig. 10(b) shows the force-tracking results of the three control methods. Due to the fixed impedance parameter, the constant impedance control tracks well on the plane but suffers from steady-state error on the inclined plane. In contrast, the adaptive control can adjust the impedance parameters according to the force error feedback to track the force, but with some hysteresis. The proposed DRL control strategy generates feedforward signals based on orthogonal force information, which effectively removes the hysteresis of the control process. Consistent with the results of the simulation, our method outperforms the other

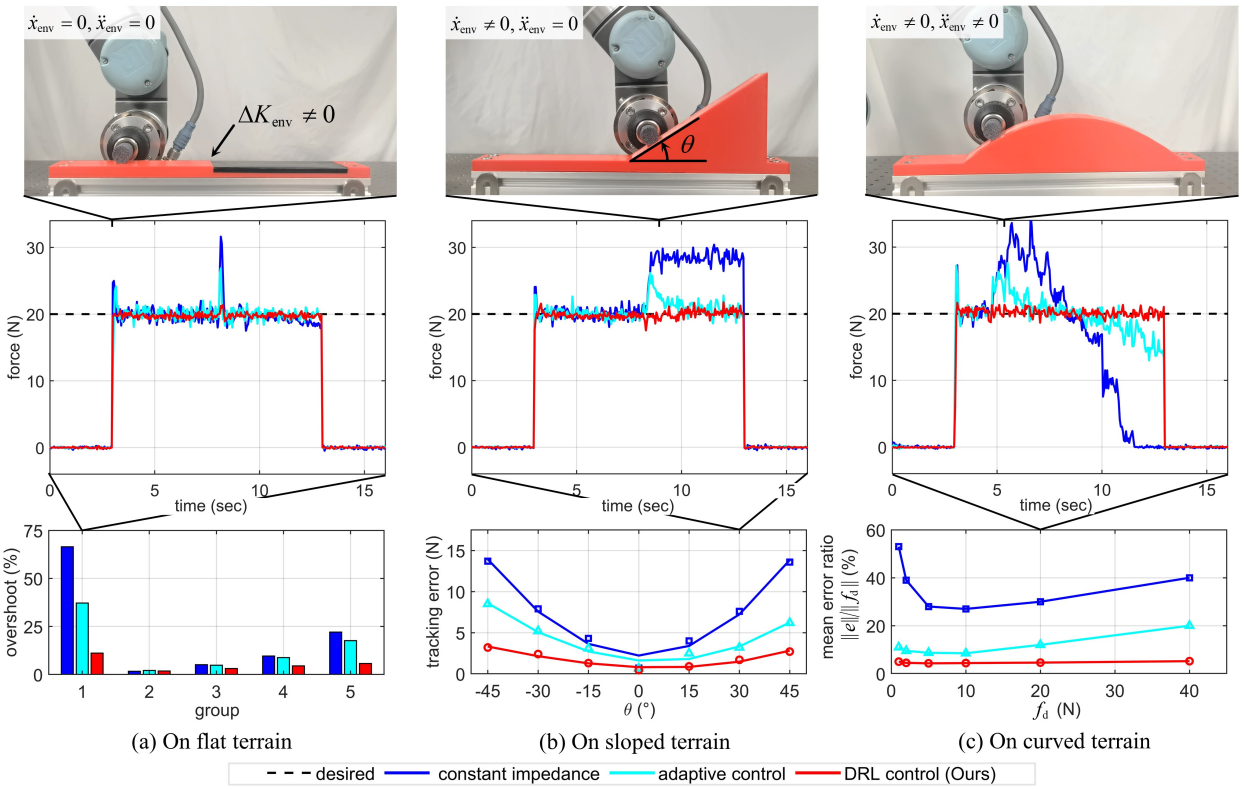

Fig. 10. Results of the robot performing a machining task under constant impedance (blue), velocity-based adaptive control (cyan) and the proposed DRL control method (red). Terrain surface types include a) stiffness-varying planes b) sloped surfaces and c) curved surfaces. (a) Force tracking on flat terrain with different stiffness switching. The switched materials include 1-rubber, 2-resin, 3-glass, 4-aluminum and 5-steel. (b) Force tracking error on oblique terrain with different tilt angles $\theta \in [-45,\ 45]$ °. (c) The ratio of force tracking errors on curved terrain with different desired forces $f_{\mathrm{d}} \in [1,\ 40]$ N.

two control methods in force tracking on inclined surfaces with different inclination angles.

If the contact environment is a curved terrain, it satisfies $\dot{x}_{\mathrm{env}} \neq 0$, $\ddot{x}_{\mathrm{env}} \neq 0$, and other assumptions are the same. Fig. 10(c) illustrates the results of the ratio of the errors relative to the desired force for different desired forces. It can be seen that the average error ratio of the proposed DRL control stays below 5%, which is lower than the other two methods.

From the above three experiments, it can be concluded that the simulation-trained impedance learning method shows a better force-tracking effect in the real environment and effectively realizes the transfer of training results to the real robot.

Scene 2: Robot peeling. We chose three different ingredients (apples, potatoes and cucumbers) as peeling objects. The peeling process removes the surface peel causing a small change in the environment, so the contact surface can be visualized as a quasi-static terrain, i.e., $(x_{\mathrm{env}})_{p+1} \neq (x_{\mathrm{env}})_p$ and $(x_{\mathrm{env}})_{p+1} - (x_{\mathrm{env}})_p \approx 0$. The ends of the objects are fixed to the base. The robot end-effector is a stainless steel knife with a blade length of 50 mm. After the calibration experiments are completed, the unpeeled objects are selected for the robotic peeling experiments.

The task is done if the peeling is continuous and there is no jamming of the knife. Under the same conditions, the lateral sides of the object are selected for three separate peeling tests. Figure 11(a) shows the success rate of each control method under different desired forces. Constant impedance control is difficult to apply to complex surfaces and has the lowest success rate with an average of only 15.6%. Adaptive control and our method effectively improve the peeling success rate (with 57.8% and 73.3%, respectively). The peeling success rate is reduced due to the jamming of the knife caused by too large force or the peel breakage caused by too small force. The three control methods have the best peeling effect when the desired force is 5 N, as shown in Fig. 11(b). In the experiments, the bumps or depressions on the surface of the objects lead to changes in the measured force, which in turn affects the thickness of the peel. In contrast, the proposed control strategy can track the desired cutting force stably. Fig. 11(c) compares the mean and standard deviation of the thickness of the object obtained by cutting at $f_d = 5$ N. It can be seen that the thickness of the peel cut under the proposed DRL control is more uniform, and the standard deviations of its thicknesses are 0.058 mm (apple), 0.033 mm (potato), and 0.033 mm (cucumber) are lower than those of the adaptive control counterparts of 0.106 mm, 0.092 mm, and 0.177 mm.

Scene 3: Human-robot interaction. Robots with compliance control can be used for human-robot interaction. The interaction process is illustrated with examples of shaking hands, combing hair and swabbing skin between a robot and a human. Desired forces $f_d^{\mathrm{handshake}} = [0,\ 0,\ 5]^{\mathrm{T}}$ N, $f_d^{\mathrm{comb}} = [0.1t,\ 0,\ 2]^{\mathrm{T}}$ N, and $f_d^{\mathrm{swab}} = [\sin \pi(t/5+1),\ 0,\ 1]^{\mathrm{T}}$ N are set for handshaking, combing, and swabbing, respectively,

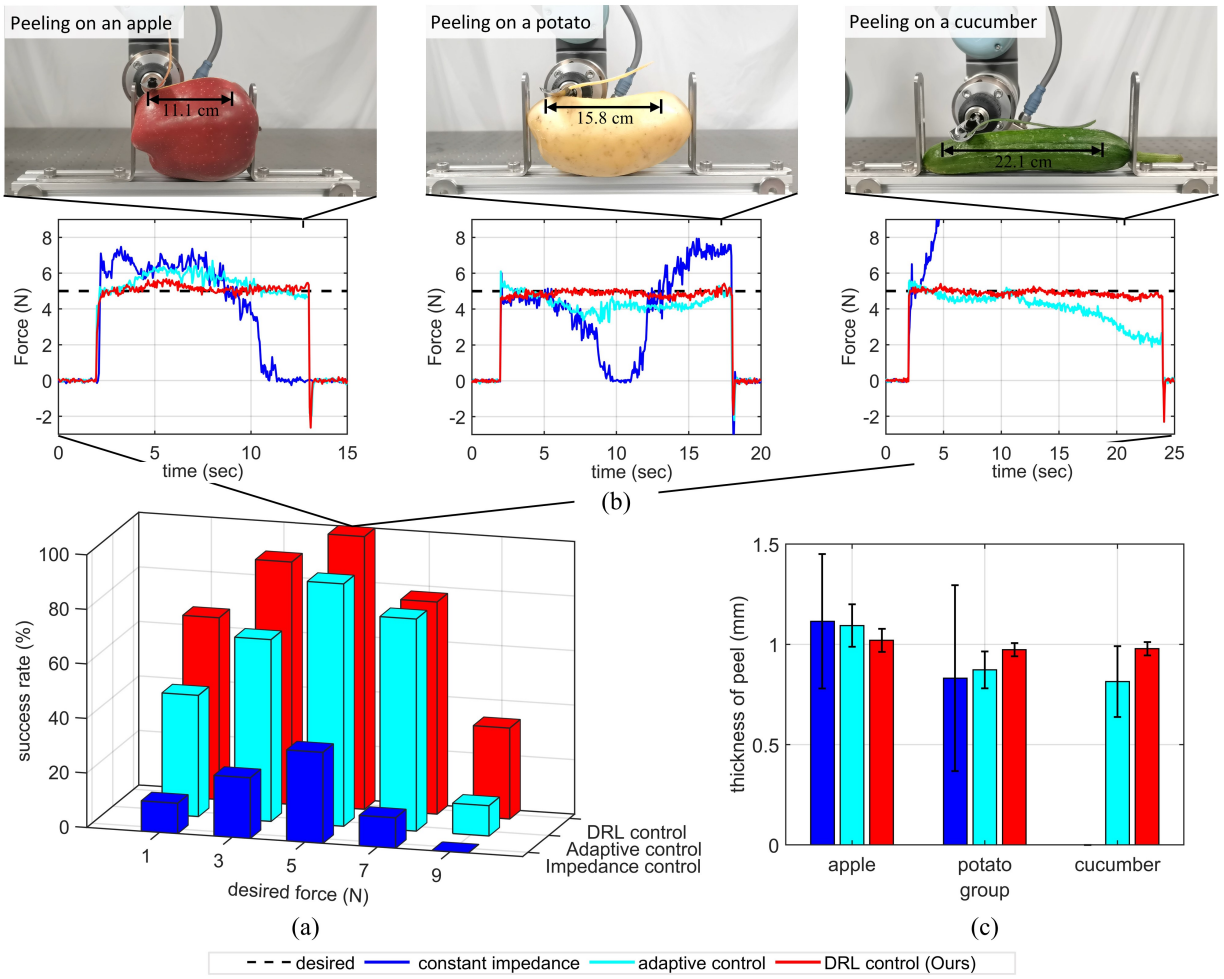

Fig. 11. Results of the robot performing the peeling task under constant impedance (blue), velocity-based adaptive control (cyan) and the proposed DRL method (red). (a) Peeling success rate under different desired forces $f_{\mathrm{d}}(\in [1,\ 9]$ N). (b) Radially measured force of the peeled object under different control methods (f_{d} = 5 N). None of the objects contacted the end-effector at the beginning. (c) Statistics of peeling thickness for different objects (f_{d} =5 N). Constant impedance control for cucumber peeling does not give corresponding results because the task failed in all three tests.

to drive the robot to perform these tasks. Human stiffness is unknown and the epidermis is disturbed by friction and pressure, leading to dynamic changes in surface morphology. Therefore, the same safe position of the arm is selected for the calibration and test experiments. Considering the poor dynamic response of the constant impedance control which may lead to injuries, only the proposed controller's force-tracking performance relative to the adaptive feedback control is compared, as shown in Fig. 12(a). In the handshake test, the robot approaches the human hand at the same speed and generates a force overshoot at the beginning. Based on the force error information, the robot adjusts its impedance to increase flexibility to follow the motion of the human hand. Our method demonstrates a faster response. In the hair combing test, the x-axis is set to gradually increase the force to drive the robot to complete the combing task due to the damping from the hair. The velocity-based adaptive control showed hysteresis during the contact (from 3 s). In five consecutive swabbing tests (starting from 2 s), the skin was stiffened due to the extrusion stack, resulting in five sinusoidal peaks. The proposed DRL control adapts to the stiffness variation law of the skin through the priori impedance learning thus obtaining better force tracking. The statistical results of the force errors for the three tasks are shown in Fig. 12(b), where our method outperforms the adaptive control and reduces the force error by 55.4% on average compared to it. Since the reduction of force tracking error in human-robot interaction may come from two aspects, i.e., the performance improvement of the controller or the human effort, the friendliness of human-robot interaction should also be considered in the experiments. To maintain contact force stabilization, the human can show compliance or impedance to the robot by constantly adjusting his/her state (position or stiffness), but this will bring about constant oscillations in x_m or f_m. To evaluate the human effort J_{human} of driving the robot during the interaction, the following formula [51] is applied

$$J_{\mathrm{human}} = \int_0^{T_0} \left| f_m^{\mathrm{T}}(t)\dot{x}_m(t) \right| \mathrm{d}t$$

where T_0 is the task duration and $\dot{x}_m$ is the interaction speed.

Fig. 12(c) shows that the human effort during human-robot interaction is always less than or equal to that of the adaptive control, which suggests that the reduction of force tracking

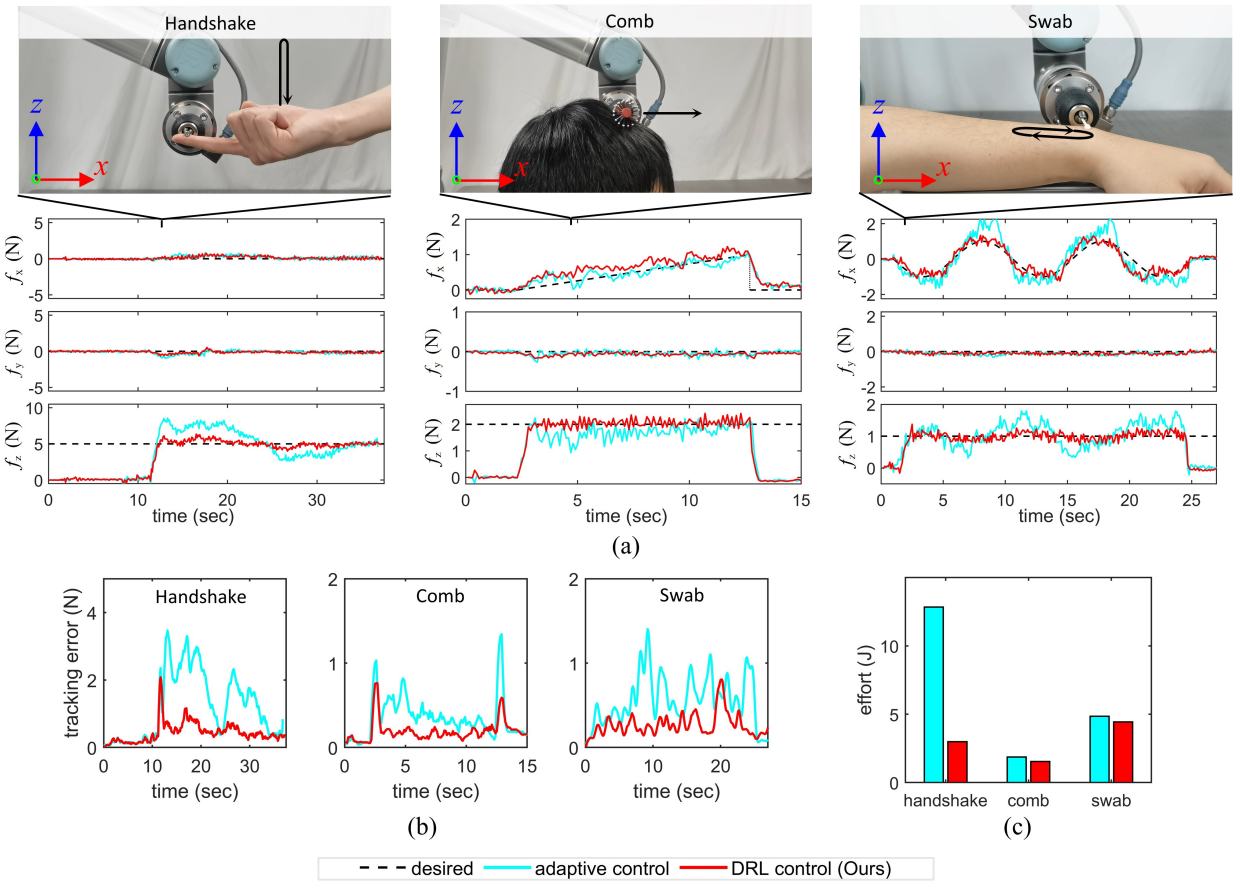

Fig. 12. Results of human-robot interaction under velocity-based adaptive control (cyan) and the proposed DRL method (red). (a) Measured forces for different interaction tasks in the workspace. (b) Statistical results of tracking force errors. (c) Average human effort during human-robot interaction.

error comes entirely from the enhancement of the controller performance to exclude the deliberate human effort. On the other hand, it also suggests that our method can reduce the human effort used to maintain force stabilization, especially on dynamic human-robot interaction tasks (e.g., handshaking). Thus, the proposed DRL controller not only produces better tracking performance, but also reduces the workload of driving the robot and therefore has a better human-robot interaction experience.

In the above three experimental scenarios, static, quasi-static and dynamic environments are included. The proposed controller can be learned in the virtual environment and used in the complex real environment. The experimental results verify the effectiveness of impedance learning feedforward and the feasibility of DRL control. The designer can easily adjust the desired impedance metrics and the robot can track the desired force stably in uncertain environments.

VI. Conclusion

Force tracking is an important issue in the contact manipulation of robots with uncertain environments. In this paper, we propose an interpretable DRL control policy for force tracking in unknown environments, which is a combination of DRL feedforward and variable impedance feedback. Two key properties of the proposed control policy are analyzed: stability and convergence. Sufficient conditions for stability and zero-error convergence of the proposed DRL control are given by establishing a simplified contact model of the robots interacting with uncertain environments.

By satisfying the Lipschiz continuity condition, the impedance learning results in the virtual environment are guaranteed to be safely transferred to the real environment, which equips the robot with impedance adjustment intelligence in uncertain environments. In conjunction with ablation experiments, the effect of the regression vector selection on the convergence of the feedforward network is discussed. The analysis shows that the observable state of the uncertain system is necessary in the learning process, otherwise it may lead to overfitting or underfitting phenomena. This phenomenon is prevalent in other neural network controllers that rely only on input-output.

Then, based on the results of the theoretical analysis, simulation and real robot experiments were carried out under different material terrains and different desired forces. The results show that the proposed DRL control policy not only generalizes to untrained scenarios, but also can perform well in static or dynamic uncertain environments. In future work, we will consider extending the proposed DRL control policy to more kinds of robot interaction tasks with uncertain environments, such as universal grasping and bipedal running.

REFERENCES

[1] A. Ibarguren, P. Daelman and M. Prada, "Control Strategies for Dual Arm Co-Manipulation of Flexible Objects in Industrial Environments," in *2020 IEEE Conference on Industrial Cyberphysical Systems (ICPS)*, Tampere, Finland, 2020, pp. 514-519.

[2] Qin X, Shi H, Gao X, *et al.*, "The adaptive neural network sliding mode control for angle/force tracking of the dexterous hand," *Advances in Mechanical Engineering*, vol.13, 2021.

[3] X. Li, W. Wang and J. Yi, "Foot contact force of walk gait for a quadruped robot," in *2016 IEEE International Conference on Mechatronics and Automation*, Harbin, China, 2016, pp. 659-664.

[4] Z. Li, X. Li, Q. Li, *et al.*, "Human-in-the-Loop Control of Soft Exosuits Using Impedance Learning on Different Terrains," *IEEE Transactions on Robotics*, vol. 38, no. 5, pp. 2979-2993, Oct. 2022.

[5] X. Chen, N. Wang, H. Cheng, *et al.*, "Neural Learning Enhanced Variable Admittance Control for Human–Robot Collaboration," *IEEE Access*, vol. 8, pp. 25727-25737, 2020.

[6] B. Siciliano, L. Villani, "Robot Force Control," *Springer*, Berlin, 2000.

[7] P. Jaroonsorn, P. Neranon, P. Smithmaitrie ,*et al.*, "Robot-assisted transcranial magnetic stimulation using hybrid position/force control". *Advanced Robotics* vol. 34, no. 24, pp. 1559-1570, 2020.

[8] S. Jung and T. C. Hsia, "Force Tracking Impedance Control of Robot Manipulators for Environment with Damping," in *IECON 2007 - 33rd Annual Conference of the IEEE Industrial Electronics Society*, 2007, pp. 2742-2747.

[9] H. F. N. Al-Shuka, S. Leonhardt, W. Zhu, *et al.*, "Active Impedance Control of Bioinspired Motion Robotic Manipulators: An Overview", *Applied Bionics and Biomechanics*, vol. 2018, no. 1, Art. no. 8203054, 2018.

[10] S. Hjorth, D. Chrysostomou, "Human–robot collaboration in industrial environments: A literature review on non-destructive disassembly," *Robotics and Computer-Integrated Manufacturing*, vol. 73, 2022.

[11] R. Mengacci, F. Angelini, M. G. Catalano, G. Grioli, A. Bicchi and M. Garabini, "Stiffness Bounds for Resilient and Stable Physical Interaction of Articulated Soft Robots," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4131-4138, Oct. 2019.

[12] X. Lamy, F. Colledani and P. -O. Gutman, "Identification and experimentation of an industrial robot operating in varying-impedance environments," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 3138-3143.

[13] Z. Jiang, Z. Sun, H. Li, *et al.*, "Designing of a kind of Impedance Controller for Industrial Robots to Assemble Satellites," in *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation*, Control, and Intelligent Systems (CYBER), Honolulu, HI, USA, 2017, pp. 436-441.

[14] P. Yuan, "An Adaptive Feedback Scheduling Algorithm for Robot Assembly and Real-Time Control Systems," in *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, Beijing, China, 2006, pp. 2226-2231.

[15] Wu, J., Ni, F., Zhang, Y., *et al.*, "Smooth transition adaptive hybrid impedance control for connector assembly", *Industrial Robot*, vol. 45 no. 2, pp. 287-299, 2018.

[16] N. Rojas-García, A. Chávez-Olivares, O. Mendoza-Gutiérrez, "Force/position control with bounded actions on a dexterous robotic hand with two-degree-of-freedom fingers," *Biocybernetics and Biomedical Engineering*, vol. 42, 2022.

[17] S. S. Ge,Y. Li and C. Wang, "Impedance adaptation for optimal robot–environment interaction", *International Journal of Control*, vol. 87, pp. 249-263, 2014.

[18] T. Xue, W. Wang, J. Ma, *et al.*, "Progress and Prospects of Multimodal Fusion Methods in Physical Human–Robot Interaction: A Review," *IEEE Sensors Journal*, vol. 20, no. 18, pp. 10355-10370, 2020.

[19] A. Hameed, A. Ordys, J. Możaryn, *et al.*, "Control System Design and Methods for Collaborative Robots: Review," *Applied Sciences*, vol. 13, no. 1, Art. no. 675, 2023.

[20] H. Hu, X. Wang, L. Chen, "Impedance with Finite-Time Control Scheme for Robot-Environment Interaction", *Mathematical Problems in Engineering*, vol. 2020, 2020.

[21] H. Cao, X. Chen, Y. He, *et al.*, "Dynamic Adaptive Hybrid Impedance Control for Dynamic Contact Force Tracking in Uncertain Environments," *IEEE Access*, vol. 7, pp. 83162-83174, 2019.

[22] C., Hongli, "Design of a Fuzzy Fractional Order Adaptive Impedance Controller with Integer Order Approximation for Stable Robotic Contact Force Tracking in Uncertain Environment," *Acta Mechanica et Automatica*, vol.16, no.1, pp. 16-26, 2022.

[23] Li, Z, "A fuzzy adaptive admittance controller for force tracking in an uncertain contact environment," *IET Control Theory*, vol. 15, pp.2158–2170, 2021.

[24] S. Jung, DJ. Jeong, "Admittance Force Tracking Control Schemes for Robot Manipulators under Uncertain Environment and Dynamics," *International Journal of Control, Automation and Systems*, pp. 3753–3763, 2021.

[25] J. Duan, Y. Gan, M. Chen, *et al.*, "Adaptive variable impedance control for dynamic contact force tracking in uncertain environment," *Robotics and Autonomous Systems*, vol. 102, pp. 54-65, 2018.

[26] J. Peng, Z. Yang, T. Ma, "Position/Force Tracking Impedance Control for Robotic Systems with Uncertainties Based on Adaptive Jacobian and Neural Network," *Complexity*, vol. 2019, 2019.

[27] M. Iskandar, C. Ott, A. Albu-Schäffer, *et al.*, "Hybrid Force-Impedance Control for Fast End-Effector Motions," *IEEE Robotics and Automation Letters*, vol. 8, no. 7, pp. 3931-3938, July 2023.

[28] C. Li, Z. Zhang, G. Xia, *et al.*,"Efficient Force Control Learning System for Industrial Robots Based on Variable Impedance Control," *Sensors*, vol. 18, 2018.

[29] L. Roveda, N. Iannacci, F. Vicentini, *et al.*, "Optimal Impedance Force-Tracking Control Design With Impact Formulation for Interaction Tasks," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 130-136, Jan. 2016.

[30] L. Wang and B. Meng, "Adaptive vision-based force/position tracking of robotic manipulators interacting with uncertain environment," in *2019 Chinese Control And Decision Conference (CCDC)*, Nanchang, China, 2019, pp. 5126-5131.

[31] M. Xu, A. Hu and H. Wang, "Image-Based Visual Impedance Force Control for Contact Aerial Manipulation," *IEEE Transactions on Automation Science and Engineering*, vol. 20, no. 1, pp. 518-527, Jan. 2023.

[32] J. R. Medina, D. Sieber and S. Hirche, "Risk-sensitive interaction control in uncertain manipulation tasks," in *2013 IEEE International Conference on Robotics and Automation*, Karlsruhe, Germany, 2013, pp. 502-507.

[33] B. Baigzadehnoe, Z. Rahmani, A. Khosravi, *et al.*, "on position/force tracking control problem of cooperative robot manipulators using adaptive fuzzy backstepping approach," *ISA Transactions*, vol. 70, pp. 432-446, 2017.

[34] Y. Shen, Y. Lu and C. Zhuang, "A fuzzy-based impedance control for force tracking in unknown environment," *Journal of Mechanical Science and Technology*, vol. 36, pp. 5231–5242, 2022.

[35] S. Ito, Y. Ishibashi, P. Huang, *et al.*, "Effect of Robot Position Control Using Force Information: Human versus Robot with Force Sensor," in *International Conference on Information and Education Technology (ICIET)*, Okayama, Japan, pp. 257-261, 2021.

[36] S. Hashimura, H. Sakai, K. Kubota, *et al.*, "Influence of Configuration Error in Bolted Joints on Detection Error of Clamp Force Detection Method," *International Journal of Automation Technology*, vol. 15, no. 4, pp. 396-403, 2021.

[37] B. Vanderborght, A. Albu-Schaeffer, A. Bicchi, *et al.*, "Variable impedance actuators: A review," *Robotics and Autonomous Systems*, vol. 61, no. 12, 2013.

[38] J. Abu-Dakka Fares, S. Matteo, "Variable Impedance Control and Learning—A Review," *Frontiers in Robotics and AI*, vol. 7, 2020.

[39] C. Yang, C. Zeng, P. Liang, *et al.*, "Interface Design of a Physical Human–Robot Interaction System for Human Impedance Adaptive Skill Transfer," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 1, pp. 329-340, Jan. 2018.

[40] K. Kronander and A. Billard, "Learning Compliant Manipulation through Kinesthetic and Tactile Human-Robot Interaction," *IEEE Transactions on Haptics*, vol. 7, no. 3, pp. 367-380, Sept. 2014.

[41] A. Bolotnikova, S. Courtois and A. Kheddar, "Adaptive Task-Space Force Control for Humanoid-to-Human Assistance," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 5705-5712, July 2021.

[42] M. Li, Y. Wen, X. Gao, *et al.*, "Toward Expedited Impedance Tuning of a Robotic Prosthesis for Personalized Gait Assistance by Reinforcement Learning Control," *IEEE Transactions on Robotics*, vol. 38, no. 1, pp. 407-420, Feb. 2022.

[43] J. van den Kieboom and A. J. Ijspeert, "Exploiting natural dynamics in biped locomotion using variable impedance control," in *2013 13th IEEE-RAS International Conference on Humanoid Robots (Humanoids)*, Atlanta, GA, USA, 2013, pp. 348-353.

[44] Z. Zhang, "A Computational Framework for Robot Hand Design via Reinforcement Learning," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Prague, Czech Republic, 2021, pp. 7216-7222.

[45] J. Buchli, E.A. Theodorou, f. Stulp, *et al.*, "Variable Impedance Control - A Reinforcement Learning Approach," *Robotics: Science and Systems*, 2010.
[46] D. Shen, W. Zhang, J. Xu, "Iterative Learning Control for discrete nonlinear systems with randomly iteration varying lengths," *Systems & Control Letters*, vol. 96, pp. 81-87, 2016.
[47] K. Patan, M. Patan, "Neural-network-based iterative learning control of nonlinear systems," *ISA Transactions*, vol. 98, pp. 445-453, 2020.
[48] H. Khalil, "Nonlinear Systems. 3rd Edition," *Prentice Hall*, Upper Saddle River, New Jersey, USA, 2002.
[49] M. Pierallini, F. Stella, F. Angelini, *et al.*, "A Provably Stable Iterative Learning Controller for Continuum Soft Robots," *IEEE Robotics and Automation Letters*, vol. 8, no. 10, pp. 6427-6434, Oct. 2023.
[50] T. Salimans and D. P . Kingma, "Weight Normalization: A Simple Reparameterization to Accelerate Training of Deep Neural Networks," *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 901–909, 2016.
[51] X. Xing, E. Burdet, W. Si, *et al.*, "Impedance Learning for Human-Guided Robots in Contact With Unknown Environments," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3705-3721, Oct. 2023.

Yanghong Li received the B. Eng. degree from the University of Science and Technology of China (USTC), in 2020. He is currently a Ph.D. student at the University of Science and Technology of China.

His research interests include the force control of the manipulator and model learning for control.

Zheng Li received the B. Eng. degree from Dalian University of technology , in 2020. He is currently a Ph.D. student at the University of Science and Technology of China (USTC).

His research interests include the motion planning of the manipulator and deep reinforcement learning.

Yahao Wang received the B. Eng. degree from Anhui Engineering University , in 2018. He is currently a Ph.D. student at the University of Science and Technology of China (USTC).

His research interests include the motion planning of the manipulator.

Erbao Dong (Member, IEEE) received the B.S. and Ph.D. degrees in precision machinery and precision instrumentation from University of Science and Technology of China, Hefei, China in 2005 and 2010, respectively.

He is currently an Associate Professor with the University of Science and Technology of China, Hefei, China. His current research interests include artificial muscles and bionic robots, intelligent mobile robots, dual-arm collaborative robots.

Shiwu Zhang (Member, IEEE) received the B.S. degrees in Mechanical and Electronic Engineering from University of Science and Technology of China, in 1997, and the Ph.D. degree in the Precision Instrumentation and Machinery from USTC in 2003.

He is currently a professor in the Department of Precision Machinery and Precision Instrumentation, USTC. He has been a visiting scholar in University of wollongong, Australia in 2016 and in the Ohio state university, USA in 2012, respectively.