A Reinforcement Learning Method Based on Hybrid Proximal Policy Optimization for Deformation Control in Machining Titanium Alloy Components

HE Fangzhou, *LIU Changqing*, *TIE Lei*, *XU Yiyun*, *YANG Fan*, *GAO James*, *LI Yingguang**

College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, P. R. China

(Received 29 December 2025; revised 25 May 2026; accepted 18 June 2026)

Abstract: A hybrid action deformation control method based on hybrid proximal policy optimization (HPPO) is proposed for titanium alloy structural components. Existing reinforcement learning algorithms are generally confined to either discrete or continuous action spaces, and thus cannot simultaneously optimize machining sequence and allowance. The proposed method unifies both decision variables—machining sequence as discrete actions and machining allowance as continuous parameters—into a single parameterized hybrid action space. Online deformation force monitoring data serve as state feedback to enable adaptive control under dynamic machining conditions. A dual-layer reward mechanism combining process-level deformation force uniformity with terminal deformation convergence is designed to guide the agent toward synchronized suppression of both local and global deformations. Experimental validation on a Ti6Al4V aviation structural component demonstrates that the proposed method reduces average machining deformation from 0.103 mm to 0.054 mm, with RMSE decreasing from 0.119 mm to 0.071 mm, representing a 47.57% reduction relative to the uncontrolled case. These results confirm the accuracy and effectiveness of the proposed method in real manufacturing environments.

Key words: titanium alloy; hybrid machining process; machining deformation control; hybrid proximal policy optimization (HPPO)

CLC number: V262.3 **Document code:** A **Article ID:** 1005-1120(2026)03-0340-16

0 Introduction

With the continuous advancement of aircraft manufacturing technology, the accuracy requirements for machined large structural components are increasingly stringent. Consequently, machining deformation control has become a critical factor in ensuring both part quality and overall airframe performance. Titanium alloy structural components, serving as key load-bearing parts in aircraft, face severe deformation control challenges during machining processes[1].

Titanium alloy structural components, characterized by high strength, low thermal conductivity, and pronounced work-hardening behavior, are prone to complex deformation patterns during machining processes, including bending, torsion, and combined bending-torsion modes[2]. These complex deformations are manifested not only as displacements in single directions, but also as coupled deformation modes that affect different regions throughout the component, particularly at critical locations. The inherent complexity of machining-induced deformation in titanium alloy components renders traditional control strategies that are inadequate for achieving effective deformation suppression. A

*Corresponding author, E-mail address: liyingguang@nuaa.edu.cn.

How to cite this article: HE Fangzhou, LIU Changqing, TIE Lei, et al. A reinforcement learning method based on hybrid proximal policy optimization for deformation control in machining titanium alloy components[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2026, 43(3): 340-355.

http://dx.doi.org/10.16356/j.1005-1120.2026.03.002

multi-process deformation control method is therefore essential to effectively regulate the evolution of deformation throughout the titanium alloy structural components. The machining sequence and machining allowance are regarded as the core controllable machining processes that affect the component deformation. The machining sequence determines the timing of material removal and the path of load release, thereby dominating the temporal and spatial evolution of the deformation. The machining allowance affects the response of the deformation sensitive areas by changing the stiffness and the distribution of the removed amount. There is a significant coupling between machining sequence and allowance, that is, the same machining allowance will produce completely different deformation responses under different sequences. Meanwhile, the implementation of multi-process collaborative control faces formidable challenges arising from the complexity of the decision space. The machining sequence constitutes a discrete decision variable, where the number of possible sequence combinations grows factorially with the number of machining regions for components featuring multiple zones. Conversely, the determination of machining allowance represents a continuous optimization problem, with the allowance for each region being continuously adjustable within a specified interval. This hybrid action space is characterized by the coexistence of discrete and continuous variables. Traditional optimization methods are ineffective in solving such problems.

To address the deformation control issue in titanium alloy structural components, existing approaches can be broadly classified into mechanism-based and data-driven categories, both of which are reviewed in the following sections.

(1) Mechanism-based methods for deformation control

The deformation control approach based on mechanism models is mainly to predict machining deformation through finite element methods (FEMs) using residual stress measurement results before machining, and optimize machining processes according to predicted deformation. It was mainly studied in the aspect of machining sequence optimization, machining parameter optimization, and finishing allowance optimization.

The machining sequence significantly influences residual stress redistribution, internal stress distribution, and interim component stiffness, all of which are key factors contributing to deformation[3]. Strategic sequence planning ensures gradual stiffness variation. This avoids geometric discontinuities while enabling balanced deformation release. Hao et al.[4] proposed a dual-side collaborative optimization method for machining sequences, effectively reducing deformation in double-sided components. Casuso et al.[5] modeled and simulated final deformation profiles of an aero-engine turbine part under varying sequences to identify the optimal strategy. Cutting-induced residual stress, a primary cause of finishing stage deformation, is governed by cutting forces and temperatures, which is directly affected by machining allowance[6]. Li et al.[7] developed a deformation prediction model linking initial residual stress to finishing allowance, employing a simplex algorithm-based linear programming model to minimize deformation. The initial residual stress distribution varies across different raw components. Consequently, the position of a component within the stock material determines its final internal stress distribution, and divergent stress levels will induce distinct machining-induced deformations[8]. Thus, optimizing the component's positioning within the stock material serves as an effective strategy for deformation control. Cerutti et al.[9] demonstrated that roughing at different stock positions yielded divergent deformation outcomes. Brinksmeier et al.[10] experimentally proved that bearing ring machining positions altered residual stress distribution and subsequent deformation. Nervi[11] identified stock material stress, part geometry, and in-process positioning as deformation determinants. Cutting parameters exerted significant influence on cutting forces and thermal generation during machining processes. These two factors (force and heat) are pivotal determinants of machining-induced residual stress, which interacts synergistically with the material's initial residual stress to define the final component stress state. This stress coupling directly governs component deformation. Consequently, the strategic adjustment of cutting parameters emerged as a viable method for deformation control in precision manufacturing. Yue et al.[12] showed multi-pass strategies reduced cutting forces and tensile residual stresses compared to single-pass. Moganapriya et al.[13] used multi-criteria decision making with grey fuzzy coupled Taguchi method to determine the optimal cutting speed, feed rate, and depth of cut. Chen et al.[14] modeled milling forces under radial depths using instantaneous undeformed chip thickness theory. Huan et al.[15] optimized fixturing by modeling cutting forces, residual stresses, and clamping impacts.

However, mechanism-based methods require precise initial conditions in the numerical modeling process, and it is difficult to take into account the uncertainties of the actual processing procedure, resulting in the difficulty of precisely controlling the processing deformation.

(2) Data-driven methods for deformation control

With the advancement of sensor technologies and the rise of intelligent manufacturing, deformation control methods based on deep learning optimization algorithms have emerged as a promising research direction. Deformation control based on monitoring data during the machining process has gradually become an effective means to improve the quality of numerical control machining. During the machining process, various sensors are employed to acquire physical parameters such as displacement, force and vibration for analyzing the machining state of the workpiece. Consequently, process optimization methods driven by monitoring data enabled more effective control of machining-induced deformation. Li et al.[16] designed an intelligent clamping device that allowed the clamping force to be adjusted during machining for deformation control. Hao et al.[17] proposed a machining sequence optimization method based on online monitoring data to control deformation. According to existing data-driven approaches, the machining process is generally treated as a time-series optimization problem, in which future deformation is predicted based on monitoring data and subsequently adjusted throughout the machining process to achieve deformation control.

Among these, reinforcement learning (RL) stands out, as it learns optimal strategies through interactions with the environment to achieve specific objectives. Owing to its ability to adapt to dynamic conditions, handle high-dimensional state spaces, and optimize long-term performance, RL demonstrates excellent potential in addressing complex decision-making problems in manufacturing processes.

Zhao et al.[18] focused on learning the relationship between different residual stress distributions and deformation control processes, and proposed a machining deformation control RL method based on the meta-invariant feature space. To achieve high-precision machining deformation control and processing efficiency, Lu et al.[19] proposed a new multi-process parameter optimization based on deep RL. In the actual processing of titanium alloys, the above process optimization methods can partially mitigate machining deformations. However, titanium alloy structural components exhibit exceptional manufacturing challenges due to their complex residual stress distribution and severe post-machining distortion, applying traditional deformation control approach is inadequate for achieving precision requirements. Given the pronounced deformation characteristics of titanium alloys, it is essential to address the multi-process machining deformation control problem.

To overcome the limitations of existing methods in controlling machining-induced deformation of titanium alloy structural components, this study proposes a RL-based decision framework with a hybrid action space. The framework unifies machining sequence (discrete actions) and machining allowance (continuous actions) into a single parameterized hybrid action space via a discrete-first-continuous-next strategy, which assigns a continuous parameter range to each discrete action. This design prevents combinatorial explosion as the number of parameters grows, and ensures that each policy decision inherently captures the interactive effects of both variables rather than optimizing them in isolation.

The coupling between machining sequence and machining allowance is addressed from a mechanistic standpoint: Allowance allocation regulates global deformation by controlling the magnitude of residual stress released during finishing, while sequence planning governs local deformation through its influence on structural stiffness evolution during material removal. By integrating continuous allowance optimization with discrete sequence decision-making via the hybrid proximal policy (HPPO) algorithm, the framework enables synchronous suppression of both global and local deformations, as illustrated in Fig.1.

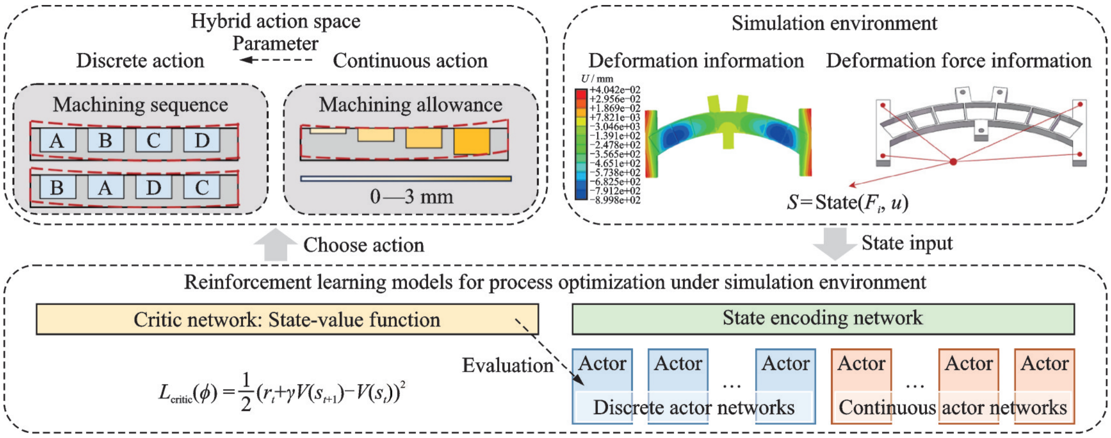


Fig.1 Framework of the proposed method

The main contributions of this work are as follows: (1) In contrast to existing RL-based methods that optimize machining sequence and machining allowance separately within either discrete or continuous action spaces, the proposed framework unifies both decisions into a single hybrid action space, thereby enabling their co-optimization rather than isolated treatment. (2) A dual-layer reward mechanism combining process-level deformation force uniformity with terminal deformation convergence enables simultaneous suppression of local and global deformations within a single training objective.

1 Methodology

By analyzing the geometric features, material properties, and processing techniques of titanium alloy structural components, two deformation control strategies were selected based on machining sequence and machining allowance. The machining sequence directly affects the release path and redistribution of residual stresses inside the component. An improper sequence may lead to uneven stress release, resulting in stress concentration and subsequent deformation; whereas a rational sequence enables gradual and balanced stress relaxation, thereby reducing distortion during material removal. Meanwhile, machining allowance primarily influences the structural stiffness of the component during intermediate stages of machining. Insufficient allowance weakens the load-bearing capacity of the workpiece, making it more susceptible to elastic deformation under the cutting force, while excessive allowance may increase stress accumulation and processing difficulty. Therefore, appropriate machining allowance can ensure the overall stiffness of the workpiece while reducing the risk of deformation during the machining process. By optimizing the processing sequence that controls the release of residual stress while designing appropriate machining allowance to maintain structural stiffness, a synergistic deformation control strategy can be formulated, thereby significantly improving the machining accuracy and geometric stability of titanium alloy structural components.

Based on the RL mechanism within a hybrid dynamic space, we derived a RL-driven strategy for the machining of titanium alloy structural components. By leveraging different configurations of the dynamic space to integrate both discrete and continuous decision-making processes, the proposed approach enables the simultaneous optimization of machining allowance and machining sequence. The method is shown in the Fig.2 and explained below.

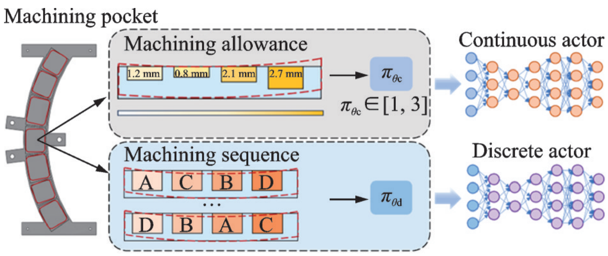


Fig.2 Architecture of the discrete-continuous action network

1.1 Hybrid action reinforcement learning

1.1.1 Reinforcement learning action space

In the field of subtractive manufacturing, adjusting the machining sequence or machining allowance is an effective means to achieve machining deformation control. The core lies in the fact that controlling these two variables can achieve redistribution of the internal stress field within the component. In actual machining, tool paths are usually programmed according to features such as pockets, edges and ribs[20]. This makes the internal machining regions of the component discrete, and these discrete machining regions constitute the component's characteristic machining sequence. After determining the machining regions, the cutting depth of each region can be further determined, which can theoretically be any continuous value, within a preferred range. Based on the above analysis, we define the following parameterized space[21]: The machining sequence is a discrete variable, selected from a finite set $A_d = \{a_1, a_2, \cdots, a_k\}$, where each $a \in A_d$ has a set of corresponding continuous parameter $X_a \subseteq \mathbf{R}$. Thus, a complete machining action is represented as a tuple (a, x), where $a \in A_d$ is the selected discrete action representing the machining sequence, and $x \in X_a$ is the corresponding continuous parameter representing the cutting depth executed together with action a. The entire action space A is the combination of each discrete action and all possible values corresponding to that action

$$A = \bigcup_{a \in A_d} \{(a, x) | x \in X_a\} \tag{1}$$

1.1.2 Reinforcement learning states

With the rapid development of sensors and information technology, a large amount of information in manufacturing processes can be monitored. Recently, the concept of deformation force was proposed and defined, which referred to the force generated by the action of unbalanced internal stress in materials on fixtures. This force can be monitored through force sensors during the machining process[22], providing a data foundation for online decision-making in machining processes. During the machining process, a workpiece undergoes deformation due to the initial residual stress field and the influence of the machining residual stress. However, under clamping constraint conditions, this deformation trend will cause force changes at the clamping points and other constraint points. These changes reflect the deformation forces. According to the principle of virtual work, the effect of external forces (i.e., deformation forces) on virtual displacement is equal to the effect of the residual stress field on virtual displacement. For a certain state v, its equilibrium equation can be expressed as

$$\boldsymbol{K}^v \boldsymbol{\delta}^v = \boldsymbol{F}^v + \boldsymbol{R}^v_{\sigma_0} \tag{2}$$

where $\boldsymbol{F}^v$ and $\boldsymbol{R}^v_{\sigma_0}$ represent the system's external load vector and residual stress equivalent load vector, respectively; $\boldsymbol{K}^v$ is the system stiffness matrix and $\boldsymbol{\delta}^v$ the system's nodal displacement.

Based on the above analysis, deformation force monitoring data is equivalent to the influence of the internal unbalanced stress field of the component under the current state. Therefore, the next machining region and machining allowance can be determined based on the current state's deformation force data, satisfying the Markov property[23]. Therefore, deformation force online monitoring data can serve as feedback from the environment to the agent after a machining action is completed in RL, which constitutes the State.

1.2 HPPO framework

Based on the definition of HPPO in Section 1.1, such an agent can make online decisions on the next machining region (discrete action space) and machining allowance (continuous action space) through deformation force (State), which is a parameterized Markov decision process[24]. The parameterized space is a type of discrete-continuous hybrid action space. The corresponding hybrid actor-critic network architecture contains two parallel actor networks, as shown in Fig. 2. The parallel actors perform action selection and parameter selection respectively: One discrete actor network learns the random strategy $\pi_{\theta d}$ to select discrete actions a representing machining region sequence numbers, and one continuous actor network learns the random strategy $\pi_{\theta c}$ to select continuous parameters $x_{a1}, x_{a2}, \cdots, x_{ak}$ corresponding to all discrete actions, which correspond to the cutting depths of the machining regions. The complete action consists of the selected discrete action a paired with the selected parameter x_a corresponding to that action.

To avoid the over-parameterization problem of action-value functions in parameterized action spaces, the critic network in this framework employs a state-value function network. The state-value function is used to compute a variance-reduced advantage estimator $\hat{A}$. After the policy operates for T time steps, the collected sample data is utilized to calculate the advantage estimator $\hat{A}$, which represents the advantage of the current action relative to the average baseline. The expression is given by

$$\hat{A}_t = -V(s_t) + r_t + \gamma r_{t+1} + \cdots + \gamma^{T-t-1} r_{T-1} + \gamma^{T-t} V(s_T) \tag{3}$$

The HPPO's discrete actor and continuous actor are updated separately at each time step, with the policies $\pi_{\theta d}$ and $\pi_{\theta c}$ of both networks being treated as two independent distributions. To generate discrete actions $\pi_{\theta d}$ with random strategies, the discrete actor outputs k discrete action values $f_{a1}, f_{a2}, \cdots, f_{ak}$, from which the random sampling takes action a. The HPPO's continuous actor network outputs the mean and variance of Gaussian distribution for each parameter to generate random policies $\pi_{\theta c}$ for continuous parameters. During the trajectory sampling process, HPPO runs its policies $\pi_{\theta d}$ and $\pi_{\theta c}$ in the environment for T time steps. Furthermore, the discrete policy $\pi_{\theta d}$ and continuous policy $\pi_{\theta c}$ are updated separately using their respective collected samples to maximize their own clipping surrogate objectives. The objective of the discrete policy $\pi_{\theta d}$ is

$$L_{\mathrm{d}}^{\mathrm{CLIP}}(\theta_{\mathrm{d}}) = \hat{E}_t[\min(r_t^{\mathrm{d}}(\theta_{\mathrm{d}})\hat{A}_t, \mathrm{clip}(r_t^{\mathrm{d}}(\theta_{\mathrm{d}}), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{4}$$

Similarly, the objective of the continuous policy $\pi_{\theta c}$ is

$$L_{\mathrm{c}}^{\mathrm{CLIP}}(\theta_{\mathrm{c}}) = \hat{E}_t[\min(r_t^{\mathrm{c}}(\theta_{\mathrm{c}})\hat{A}_t, \mathrm{clip}(r_t^{\mathrm{c}}(\theta_{\mathrm{c}}), 1-\epsilon, 1+\epsilon)\hat{A}_t)] \tag{5}$$

where the terms $r_t^{\mathrm{c}}(\theta)$ and $r_t^{\mathrm{d}}(\theta)$ represent the importance sampling ratios for the continuous and discrete policies, respectively.

To address the joint optimization problem in hybrid action spaces, this approach jointly trains the discrete and continuous policy networks by formulating the total surrogate objective as the sum of the discrete policy objective and the continuous policy objective

$$L = L_{\mathrm{d}}^{\mathrm{CLIP}}(\theta_{\mathrm{d}}) + L_{\mathrm{c}}^{\mathrm{CLIP}}(\theta_{\mathrm{c}}) \tag{6}$$

Optimizing the machining sequence and machining allowance before the finishing operation of the workpiece is an effective method for achieving machining deformation control. This section establishes an optimization model for machining sequence and machining allowance based on the HPPO model, with the objectives of reducing machining deformation and avoiding local sudden deformation.

In Abaqus finite element software, boundary conditions identical to the actual machining situation were set up, and the birth-death element method was used to execute the actions decided by the intelligent agent, for simulating the process of material removal. Meanwhile, a mask mechanism was introduced, which used a binary vector to mark the status of machining regions, setting machined regions to zero to ensure that the agent did not repeatedly select already machined areas. After initializing the HPPO model, the trajectory data was sampled according to policies $\pi_{\theta d}$ and $\pi_{\theta c}$. The single trajectory sampling procedure was as follows: The workpiece before semi-finishing was taken as the initial state s_0 of the environment in RL, with initial reward $r_0 = 0$. The RL agent made decisions on actions (a_0, x_0) based on the initial state. After material removal, deformation occurred due to unbalanced stress fields. But under clamping constraint conditions, an equivalent deformation force representing the deformation trend was taken as the environment's feedback to the agent, which became s_1. Based on s_1, r_1 was calculated. After T time steps, the machining operations were completed, and a single sequence trajectory was obtained as

$$\tau=\{(s_0,a_0,r_0,s_1),(s_1,a_1,r_1,s_2),\cdots,(s_{T-1},a_{T-1},r_{T-1},s_T)\} \tag{7}$$

To enable the HPPO model to learn optimal strategies for a component of the same geometry under different residual stress field distributions, various predefined stress fields can be applied to the component during the trajectory sampling process. HPPO employs an experience replay mechanism for parameter updates. Once the model had acquired sufficient trajectory data, it began updating the actor and critic networks. Specifically, the trajectory data was utilized to compute the advantage estimator $\hat{A}_t$ for each time step, with both actor networks being updated using their respective PPO clipping objective functions, while gradient clipping techniques were used to prevent gradient explosion. The critic network updated the state-value function $V(s)$ by minimizing the temporal difference error

$$L_{\text{critic}}(\phi)=\frac{1}{2}(r_t+\gamma V(s_{t+1})-V(s_t))^2 \tag{8}$$

where r_t is the immediate reward obtained by the agent after executing an action at time step t and γ the discount factor used to balance the weights of immediate rewards and future rewards; $V(s_t)$ and $V(s_{t+1})$ represent the state value functions at the current time step and the next time step, respectively. The algorithm process is as follows.

Algorithm: HPPO based manufacturing process optimization with mixed action space

Setting:

Initialize actor network θ, critic network φ, experience buffer B;

Set hyperparameters γ, λ, ε, batch_size, max_timesteps;

Begin:

(1) Initialize environment state s, pocket mask;
(2) for $t=1$ to max_timesteps do
(3) if new episode then;
(4) reset state s, pocket depths cs[cao_num] and mask;
(5) end if
(6) //Agent decision and interaction
(7) a_cont, a_disc, log π ← Actor(s, mask);
(8) v ← Critic(s);
(9) s', r, mask$'$, done ← env.step(a_cont, a_disc);
(10) B.store(s, a_cont, a_disc, r, v, log π);
(11) if done then
(12) compute temporal-difference errors and advantages $\hat{A}$;
(13) normalize $\hat{A}$;
(14) if $|B| \geqslant$ buffer_size then
(15) //HPPO training
(16) update critic: $\varphi \leftarrow \varphi - \alpha\nabla \text{MSE}(V\_\varphi, \hat{R})$;
(17) update actor: $\theta \leftarrow \theta - \alpha\nabla L\_\text{PPO}$ (continuous+discrete);
(18) clear B;
(19) end if
(20) reset environment;
(21) end if
(22) $s \leftarrow s'$;
(23) end for
End

1.3 Reward function design

1.3.1 Design of reward function for local deformation

The reward function serves as a core element of RL, acting as a guiding mechanism that directs the agent's learning process and determines the optimization objectives. Since only the intermediate states rather than the final deformation state could be obtained during the machining process, traditional reward mechanisms struggled to effectively guide learning. Therefore, a gradually increasing reward function was designed based on the processing scenarios during the machining, which was explained in this section.

Local deformation is primarily caused by uneven stress distribution and excessive local stress gradients during the machining process, which lead to localized stiffness degradation. Since deformation force can be regarded as a direct response of the current stiffness state, achieving a uniform distribution of deformation forces across the workpiece during machining becomes an effective means to suppress local deformation. To further realize this optimization objective, the reward function was designed to guide the intelligent agent toward decisions that promoted uniform distribution of deformation forces and avoided regions of reduced stiffness. The first element of the reward function is expressed as

$$R_{\text{process}}(s_t, a_t, s_{t+1}) = \frac{1}{N}\sum_{i=1}^{N} r_a e^{-C_p \text{diffs}_i} + r_b \tag{9}$$

$$\text{diffs}_i = | f_i - f_{(i \bmod N)+1} | \tag{10}$$

where r_a is the base reward; C_p the scaling factor controlling the reward decay rate; diffs_i the difference between deformation forces on each surface of the state; f_i the deformation force at each state; and r_b the additional reward, which is assigned based on the relationship between diffs_i and the threshold value.

1.3.2 Design of reward function for global deformation

To avoid machining chatter or tool deflection problems caused by excessively thin component walls, titanium alloy components are typically completed through semi-finishing followed by final finishing operation, during which residual stress release induces intermediate deformation. From the deformation control perspective, the global objective is to ensure that the deformation state after semi-finishing closely approximates the final deformation state after finishing, thereby minimizing unexpected shape deviations at the completion of machining. Assuming that elastic small deformations can be mitigated, aligning these two deformation states becomes a key criterion for global deformation suppression. During the machining process, environmental feedback is represented by deformation forces, which effectively captures the overall deformation tendency of the component under the current geometric state. To avoid reward distribution bias and ensure accurate optimization, the reward function was designed to guide the intelligent agent toward decisions that made the deformation forces observed after semi-finishing converge toward those of the final finishing stage. Accordingly, the second element of the reward function was defined as

$$R_{\text{final}}(s_T) = C_f e^{-\text{mean\_diff}} \tag{11}$$

where mean_diff represents the mean absolute difference between the deformation forces after semi-finishing completion and the final deformation forces after finish machining and C_f the reward amplitude.

In summary, the final reward function consists of both the process reward function and the final reward function. The process reward function depends on the uniformity of deformation force data monitored under different clamping conditions during the machining process, providing timely feedback for machining actions. The final reward function only takes effect at the terminal state of the trajectory, representing the degree of final machining deformation. The overall reward function is expressed as

$$R(s_t, a_t, s_{t+1}) = R_{\text{process}}(s_t, a_t, s_{t+1}) + \lambda R_{\text{final}}(s_T) \tag{12}$$

where λ serves as a weighting coefficient that balances intermediate process incentives against terminal performance requirements. When λ is set to a small value approaching zero, the agent prioritizes process optimization, seeking uniform deformation force distribution throughout intermediate machining stages; conversely, when λ assumes a sufficiently large value, the optimization focus shifts predominantly toward minimizing terminal deformation after the finishing operation. However, relying exclusively on process rewards is insufficient, as uniformity of deformation forces during semi-finishing does not inherently guarantee minimal residual distortion upon completion. On the other hand, terminal-only rewards yield sparse feedback signals, which aggravate the temporal credit assignment problem across sequential material removal decisions and severely retard convergence. The composite reward structure defined in Eq.(12) reconciles these competing objectives, simultaneously alleviating myopic process optimization and sparse-reward inefficiencies.

This dual reward mechanism design ensures uniform variation of deformation forces during machining and strengthens the requirements for final component deformation forces in the finishing operation, enabling the system to balance local optimization with global optimality throughout the entire machining process, ultimately achieving optimal uniform control of component deformation.

2 Training and Validating the Proposed model

2.1 HPPO model training

The training of HPPO machining sequence and machining allowance decision model requires the state of an environment as input. Therefore, for each component with a different residual stress field during the semi-finishing stage, based on the preset residual stress field function, a set of residual stress fields were randomly generated with respect to the component. Additionally, to more realistically simulate the uncertainties in actual machining environments, Gaussian noise with zero mean and 5% standard deviation was introduced when generating the residual stress fields. The first pocket was machined using the machining process strategy decided by the RL random strategy. After the machining was completed, a set of deformation force values were obtained from the floating clamping point, which were recorded as f_1, f_2, f_3, f_4, thereby obtaining the current processing environment state S. The range of the finishing allowance was set between 1 mm and 3 mm, which was determined based on the processing experience. The number of actions of this model was seven, and each action represented the current pocket to be machined and the selection of the machining allowance. The reward function was set based on the deformation force data obtained in the simulated environment and the final deformation data, so that the RL algorithm could converge more easily.

The simulation environment was constructed using the ABAQUS™ software. Each workpiece had three fixed clamping points and four deformation force monitoring points, as shown in Fig.3. This software uses the method of element birth and death to simulate the material removal process. To improve the simulation efficiency, the mesh of the material removed in the finishing stage was refined. In the z direction, each mesh was divided by 0.5 mm, and the remaining of the workpiece was divided by a 2 mm mesh. This mesh accuracy met our requirements. Nevertheless, the simulation is subject to inherent errors originating from material model simplification, in which the Ti6Al4V constitutive model does not account for strain-rate-dependent behavior, and from residual stress field approximation, as the "C"-type layered stress distribution constitutes a simplified representation of the actual measured stress state. Python was used to develop the reinforcement model for interaction with the simulation environment and training. Both the actor and critic networks received as input the deformation forces acquired from four clamping configurations during the machining process. Each network incorporated hidden layers of 256 dimensions. The actor network generated discrete action outputs alongside Gaussian probability distributions for continuous action spaces, whereas the critic network produced action-value estimations.

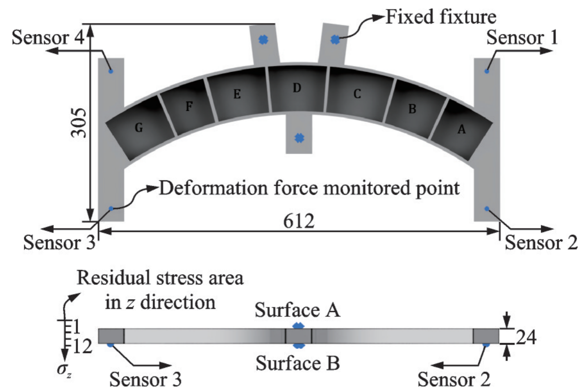


Fig.3 The simulation environment

2.2 Analysis of simulation results

The proposed method was verified by simulation and then validated using actual experiments. Both simulation and experimental results proved the effectiveness of the HPPO based optimization model.

A typical aircraft structural component shown in Fig.3 was selected for verification. The size of the component was 612 mm×305 mm×24 mm, and the material was a Ti6Al4V alloy. The sum of the semi-finishing allowance and the finishing allowance of each pocket was 3 mm, so the optimization problem was to determine the semi-finishing and finishing allowance distribution of each pocket and the machining sequence of the component in order to minimize the final machining deformation.

In the simulation environment, the proposed method was compared with the simulation model where each pocket was removed by 1 mm and the machining sequence was A—G, and the simulation model where each pocket was removed by 2 mm and the machining sequence was A—G. The comparison index were the deformation force and deformation amount after semi-finishing. The component had a total of 7 pockets of 20 mm in depth, which were indicated as A, B, C, D, E, F, and G, and assumed to be machined layer by layer (2 mm between layers), as shown in Fig.3.

In the simulation environment, the influence of machining allowances set at 1 mm and 2 mm respectively on the machining deformation of the component was simulated using a component with the same initial residual stress field. Then, using the trained HPPO hybrid action process decision model, the model was verified using the component with the same initial residual stress field. First, the model selected the first pocket to be processed and its corresponding processing allowance according to the random strategy of the model. After processing, the deformation force obtained at the simulation setting was obtained as the state for the subsequent decision on the processing action. Then, this state was used as the input feedback to the model to determine the pocket and the allowance of the pocket to be processed next. This decision-making process continued until all the pockets were processed.

To verify the robustness of the proposed method, 10 sets of random residual stress fields were generated through random sampling, and the maximum deformation under different process parameters was statistically analyzed and averaged. The averaged results are presented in Table 1. Figs.4(a—c) present the deformation contour maps for the semi-finishing process under various strategies. Compared with machining strategies with 1 mm and 2 mm allowances, the proposed method results in the smallest deformation after semi-finishing.

The deformation force after semi-finishing can reflect the current deformation status of the component. A bigger deformation of the component led to bigger forces acting on the floating clamping points, while a smaller deformation resulted in smaller deformation forces acting on the floating clamping points. Figs.4(d—f) show the deformation force conditions after different machining processes were completed.

Table 1 Mean maximum deformation from multiple simulations under random residual stress fields

Machining technology	1 mm allowance	2 mm allowance	Process decision
Average deformation/mm	0.041 4	0.039 3	0.038 4

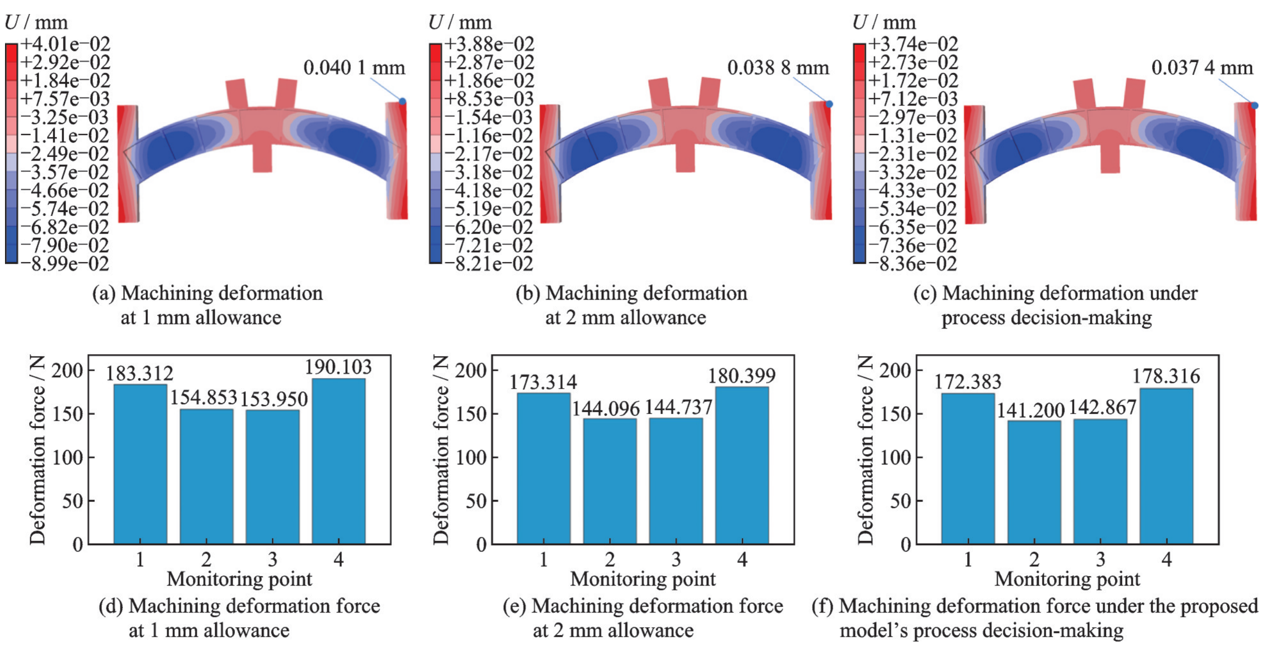
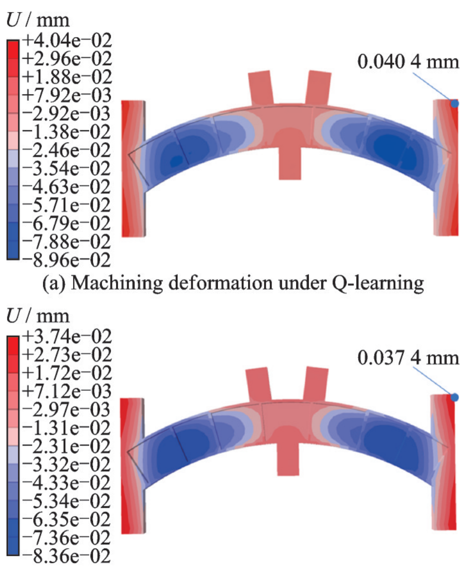


Fig.4 Machining deformation and machining deformation forces under different machining conditions

To verify the effectiveness of the proposed method relative to existing approaches, 10 sets of random residual stress fields were generated via random sampling as algorithm inputs, and the proposed HPPO method was compared against the Q-learning method, which optimized only the machining sequence. The averaged comparison results are summarized in Table 2. Figs.5(a, b) present the deformation contour maps for the semi-finishing process obtained by the two methods, respectively. The comparison results demonstrate that the proposed method achieves significantly lower deformation after semi-finishing than the Q-learning method.

Table 2 Mean maximum deformation of different optimization methods under random residual stress fields

Machining technology	Q-learning	Process decision
Average deformation/mm	0.041 8	0.038 4



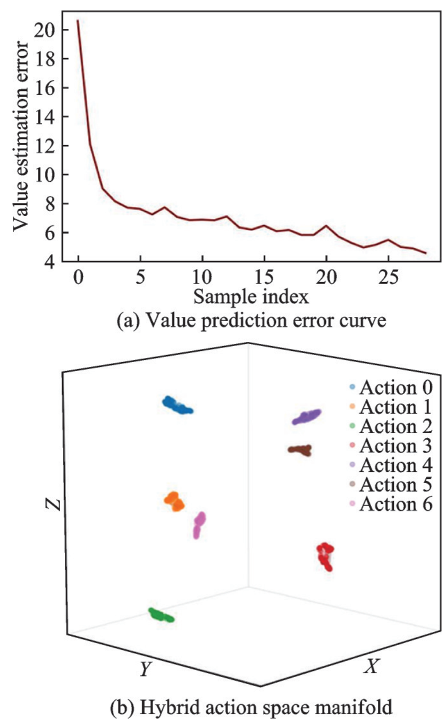
(a) Machining deformation under Q-learning

(b) The proposed method for controlled deformation

Fig.5 Comparison of semi-finishing deformation contours between different optimization algorithms under random residual stress fields

To verify the learning effectiveness and convergence of the HPPO algorithm in the current task of processing deformation control, this study monitored the prediction accuracy of the state value function during the training process. Specifically, we compared the state value V predicted by the critic network. The actual discounted return obtained through sampling from the real environment and calculated as $\hat{A}$. The difference between the two is the core metric for evaluating the predictive ability and stability of the critic network, as shown in Fig.6(a).

To validate the learning effectiveness of HPPO in a hybrid action space, this study analyzed the policy evolution process from the perspective of action space structure, as shown in Fig.6(b). The hybrid action vectors output by the agent at each time step, consisting of discrete strategy selection and continuous parameters, were projected into a low-dimensional space for visualization. The results showed that the actions exhibited clear clustering patterns in the embedding space, with different clusters corresponding to distinct discrete process decisions. Meanwhile, within each cluster, the continuous parameters displayed locally concentrated distributions, indicating that the policy formed hybrid patterns within the space.



(a) Value prediction error curve

(b) Hybrid action space manifold

Fig.6 Explanation of algorithm validity

To further validate the reliability of the proposed method, we utilized the reward mechanism of RL to quantitatively compare the quality of the proposed method with the equal tolerance processing methods. The reward situations under different processing conditions are shown in Fig.7, where steps 1—7 represent the processing process of seven pockets, and the bar charts represent the rewards obtained by the model for each pocket after the current processing action is completed. The proposed model obtained the highest reward, indicating that the decision actions had reliability in controlling the final machining deformation of the parts. This improvement is primarily attributed to the model's implementation of a collaborative optimization strategy for machining sequence and machining allowance allocation based on the Markov decision process (MDP) framework. The optimized process parameters substantially improved the distribution of residual stress of the component and effectively maintained the structural stiffness characteristics of the component during the machining process, and thereby achieved effective control of machining deformation. The equal tolerance processing methods of 1 mm and 2 mm obtained less rewards, indicating that their process decisions did not have reliability in controlling the final machining deformation of the component.

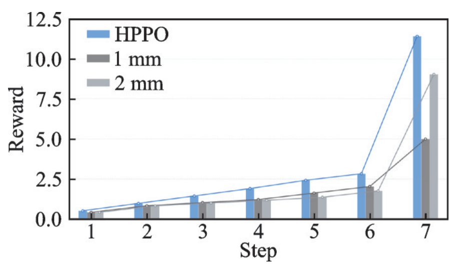


Fig.7 Rewards calculated by RL under different manufacturing processes

2.3 Analysis of experimental results

In the actual experimental environment, to further verify the generalization ability and superiority of the proposed method, a typical aviation structural part made of titanium alloy was selected for validation, as illustrated in Fig.3. The machining process involved seven pockets on the part, identified as A—G, as depicted in Fig.3. The machining process was carried out on a DMG 80P machine tool, with experimental setup involving a combination of three fixed fixtures to ensure six-degree-of-freedom positioning, and four intelligent fixtures with force sensors positioned at the corners of the component, as shown in Fig.8(a). These fixtures provided simply supported boundary conditions for measuring the deformation forces in the z direction, as shown in Fig.8(b). Thus, a total of 56 material removal operations were conducted across the seven pockets, and the machining parameters were set as shown in Table 3. During each material removal operation, four deformation forces were measured using sensors located at the four clamping positions.

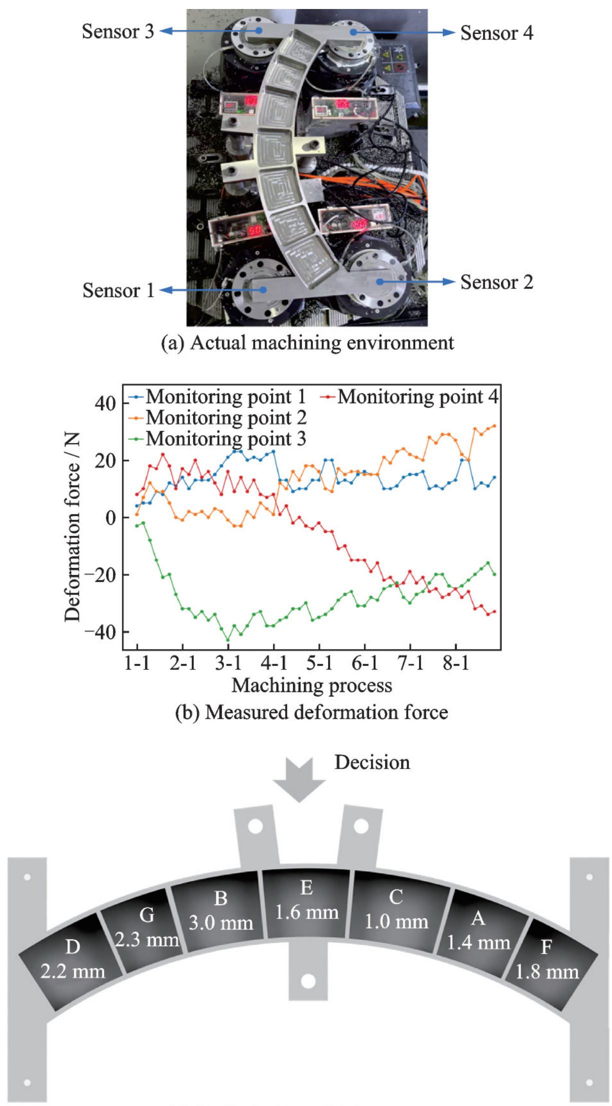


Fig.8 Machining process and deformation force measurement

Table 3 Machining parameters

Feed speed/ ($mm \cdot min^{-1}$)	Spindle speed/ ($r \cdot min^{-1}$)	Milling cutter diameter/mm
800	150	16

During the rough machining, the depth of each pocket was reduced to 16 mm, with material removed in increments of 2 mm per operation, resulting in a total of eight layers being machined from each pocket. In the semi-finishing machining, the machining sequence and machining allowance were optimized based on the proposed model. The optimized machining strategy focused on two key aspects: The finishing allowance and the machining sequence. The initial state of the workpiece was characterized by deformation force data obtained prior to the semi-finishing process. This data was input into the model to determine the finishing allowance for the first pocket. After machining the first pocket, the updated deformation force data was used by the model to decide the finishing allowance for the next pocket. This process was repeated until all pockets were machined and the semi-finishing processing of the component was completed. Based on the optimized finishing allowances and machining sequence determined during the semi-finishing stage, the finishing operation was carried out. The remaining material was machined accordingly, completing the entire machining process.

Fig.8(c) shows the machining sequence and machining allowance margin of the HPPO hybrid action space RL decision-making. And the processing sequence was A—G. After the semi-finishing process was completed, the final finishing process was carried out. Then, the probe of the machine tool was used to detect the deformation of the component after processing.

The corresponding cutting speed was approximately 7.54 m/min. This relatively low spindle speed was intentionally adopted for semi-finishing and finishing, given the poor thermal conductivity of titanium alloys: A reduced speed suppressed heat accumulation and tool wear, thereby reducing deformation caused by cutting heat. The resulting stable cutting forces also ensure reliable deformation force monitoring, aligning with the experiment's core objective of deformation control.

To verify the effectiveness of the proposed method, we conducted a comparative processing experiment. The processing method of the component for comparison was carried out according to the normal procedure, with the processing sequence being A—G. After processing, deformation measurement was conducted on the machine. The measurement points taken were the same as those of the first experiment. The comparison cloud maps of processing deformation is shown in Fig. 9. The corresponding deformation statistics are listed in Table 4.

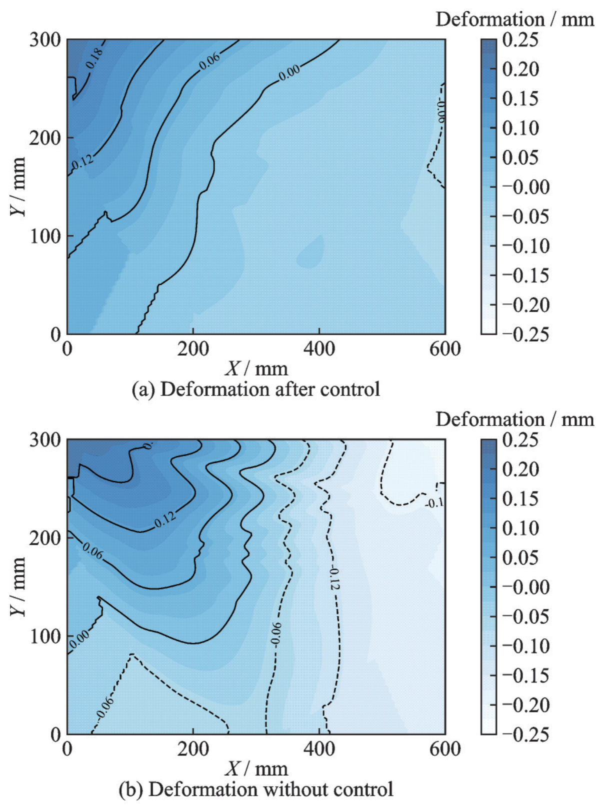


Fig.9 Results of component deformation monitoring during machining

Table 4 Machining deformation statistics

Component	MAE/mm	RMSE/mm
Without control	0.103	0.119
With control	0.054	0.071

From the deformation contour plots and Table 4, it can be observed that after adjusting the machining sequence and machining allowance using the proposed method, the average machining deformation in the pocket section is controlled within 0.054 mm. Compared with the average machining deformation of 0.103 mm in the component without control, the proposed method reduces the average deformation by 47.57%, proving the effectiveness of the proposed method.

3 Conclusions

The proposed HPPO-based hybrid action RL method demonstrated significant advantages in controlling machining deformation for titanium alloy structural components. By integrating both discrete (machining sequence) and continuous (machining allowance) decision variables into a unified RL framework, the method effectively resolved the challenge of the explosion of decision-making combination space in multi-process collaborative optimization.

The current model was trained and validated on a specific type of titanium alloy component with a fixed geometric layout. Its generalization to other materials or more complex geometries requires further investigation. Increasing the number of machining regions inevitably enlarges the state action exploration space, prolongs training time, and increases computational cost in simulation. For components with significantly more machining regions, future work may incorporate surrogate model assisted simulation to accelerate environment interaction and improve computational efficiency.

Meanwhile, the training process relied on simulated environments built in ABAQUS, which despite incorporating noise and variable stress fields, might not fully capture all real-world uncertainties.

In conclusion, a collaborative deformation control method for titanium alloy machining is proposed based on HPPO. The approach integrates discrete machining sequence optimization and continuous machining allowance adjustment into a unified hybrid action space, enabling co-optimization across multiple process parameters. Leveraging area partitioning and online deformation force monitoring, the method enhances the precision and reliability of deformation control for large-scale aerospace structural components. Practical machining validation demonstrates its effectiveness in real manufacturing environments. The main conclusions are:

(1) A hybrid action RL framework based on HPPO is proposed for titanium alloy machining deformation control, unifying discrete machining sequence optimization and continuous machining allowance adjustment into a single parameterized action space. This design prevents combinatorial explosion as the number of parameters grows and ensures that each policy decision inherently captures the interactive effects of both variables rather than optimizing them in isolation.

(2) Experimental validation on a Ti6Al4V aviation structural component confirms that the proposed method reduces average machining deformation from 0.103 mm to 0.054 mm, with RMSE decreasing from 0.119 mm to 0.071 mm, representing a 47.57% reduction relative to the uncontrolled case. These results verify the accuracy and practical effectiveness of the proposed method in real manufacturing environments.

(3) The dual layer reward mechanism, which combines process level deformation force uniformity with terminal deformation convergence, effectively balances local and global deformation suppression throughout the machining process, thereby alleviating myopic process optimization and sparse-reward inefficiencies.

In future research, we will be committed to constructing a more comprehensive multi-parameter collaborative optimization framework. Through more investigation into the coupling mechanisms among various process parameters and their complex nonlinear relationships with machining deformation, we will establish mathematical models and physical mechanism analysis frameworks for multi-parameter collaborative effects. Focus will be placed on the interaction patterns between parameters and dynamic adjustment strategies for optimal parameter combinations, laying a theoretical foundation for achieving more precise deformation prediction and control.

References

[1] CHEN Zhongwei, WU Xian, ZHAO Meng, et al. Experiment of micro-milling of additive manufactured titanium alloy[J]. Journal of Nanjing University of Aeronautics and Astronautics, 2024, 56(3): 457-467. (in Chinese)

[2] CHEN Junsong, LIU Changqing, ZHAO Zhiwei, et al. Inference method for residual stress field of titanium alloy parts based on latent Gaussian process introducing theoretical prior[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2024, 41(2): 135-146.

[3] JOVANI T, CHANAL H, BLAYSAT B, et al. Direct residual stress identification during machining[J].

Journal of Manufacturing Processes, 2022, 82: 678-688.

[4] HAO X Z, LI Y G, NI Y, et al. A collaborative optimization method of machining sequence for deformation control of double-sided structural parts[J]. The International Journal of Advanced Manufacturing Technology, 2020, 110(11): 2941-2953.

[5] CASUSO M, POLVOROSA R, VEIGA F, et al. Residual stress and distortion modeling on aeronautical aluminum alloy parts for machining sequence optimization[J]. The International Journal of Advanced Manufacturing Technology, 2020, 110(5): 1219-1232.

[6] CHEN Y Z, CHEN W F, LIANG R J, et al. Machining allowance optimal distribution of thin-walled structure based on deformation control[J]. Applied Mechanics and Materials, 2017, 868: 158-165.

[7] LI X Y, LI L, YANG Y F, et al. Machining deformation of single-sided component based on finishing allowance optimization[J]. Chinese Journal of Aeronautics, 2020, 33(9): 2434-2444.

[8] BARCENAS L, LEDESMA-OROZCO E, VAN-DER-VEEN S, et al. An optimization of part distortion for a structural aircraft wing rib: An industrial workflow approach[J]. CIRP Journal of Manufacturing Science and Technology, 2020, 28: 15-23.

[9] CERUTTI X, MOCELLIN K. Influence of the machining sequence on the residual stress redistribution and machining quality: Analysis and improvement using numerical simulations[J]. The International Journal of Advanced Manufacturing Technology, 2016, 83(1): 489-503.

[10] BRINKSMEIER E, SÖLTER J, GROTE C. Distortion engineering-identification of causes for dimensional and form deviations of bearing rings[J]. CIRP Annals—Manufacturing Technology, 2007, 56(1): 109-112.

[11] NERVI S. A mathematical model for the estimation of the effects of residual stresses in aluminum plates[D]. St. Louis, USA: Washington University in St. Louis, 2005.

[12] YUE Q B, HE Y, LI Y F, et al. Investigation on effects of single- and multiple-pass strategies on residual stress in machining Ti-6Al-4V alloy[J]. Journal of Manufacturing Processes, 2022, 77: 272-281.

[13] MOGANAPRIYA C, RAJASEKAR R, SANTHOSH R, et al. Sustainable hard machining of AISI 304 stainless steel through TiAlN, AlTiN, and TiAlSiN coating and multi-criteria decision making using grey fuzzy coupled taguchi method[J]. Journal of Materials Engineering and Performance, 2022, 31(9): 7302-7314.

[14] CHEN Y H, LU J, DENG Q L, et al. Modeling study of milling force considering tool runout at different types of radial cutting depth[J]. Journal of Manufacturing Processes, 2022, 76: 486-503.

[15] HUAN H X, ZHANG K, XU J H, et al. Machining deformation control methods and analysis of a thin-walled gear spoke plate[J]. The International Journal of Advanced Manufacturing Technology, 2023, 127(3): 1317-1331.

[16] LI Y G, LIU C Q, HAO X Z, et al. Responsive fixture design using dynamic product inspection and monitoring technologies for the precision machining of large-scale aerospace parts[J]. CIRP Annals—Manufacturing Technology, 2015, 64(1): 173-176.

[17] HAO X Z, LI Y G, NI Y, et al. A collaborative optimization method of machining sequence for deformation control of double-sided structural parts[J]. The International Journal of Advanced Manufacturing Technology, 2020, 110(11): 2941-2953.

[18] ZHAO Y J, LIU C Q, ZHAO Z W, et al. Reinforcement learning method for machining deformation control based on meta-invariant feature space[J]. Visual Computing for Industry, Biomedicine, and Art, 2022, 5(1): 27.

[19] LU F Y, ZHOU G H, ZHANG C, et al. Energy-efficient multi-pass cutting parameters optimisation for aviation parts in flank milling with deep reinforcement learning[J]. Robotics and Computer-Integrated Manufacturing, 2023, 81: 102488.

[20] CICIRELLO V, REGLI W C. Machining feature-based comparisons of mechanical parts[C]//Proceedings of International Conference on Shape Modeling and Applications. Genova, Italy: IEEE, 2001.

[21] XIONG J C, WANG Q, YANG Z R, et al. Parametrized deep Q-networks learning: Reinforcement learning with discrete-continuous hybrid action space[J]. arxiv.org/abs/1810.06394, 2018.

[22] ZHAO Z W, LIU C Q, LI Y G, et al. A new method for inferencing and representing a workpiece residual stress field using monitored deformation force data[J]. Engineering, 2023, 22: 49-59.

[23] DOSHI-VELEZ F. The infinite partially observable Markov Decision Process[C]//Proceedings of the 23rd International Conference on Neural Information Processing Systems. Vancouver, British Columbia, Canada: ACM, 2009: 477-485.

[24] MASSON W, RANCHOD P, KONIDARIS G. Reinforcement learning with parameterized actions[C]//Proceedings of the 30th AAAI Conference on Artifi-

cial Intelligence. Atlanta, USA: AAAI, 2016: 1934-1940.

Acknowledgements This work was financially supported by the Key Projects of the Basic Scientific Research Program (No. JCKY2024605B003) and National Natural Science Foundation of China (No.52575579).

Authors

The first author Mr. **HE Fangzhou** is currently a Ph.D. candidate at College of Mechanical and Electrical Engineering, Nanjing University of Aeronautics and Astronautics (NUAA), China. He received his M.S. degree from Tianjin University, China. His research interest includes data-driven intelligent manufacturing.

The corresponding author Prof. **LI Yingguang** received the B.S. degree in materials engineering and the Ph.D. degree in aeronautics and astronautics manufacturing engineering from NUAA, Nanjing, China, in 1999 and 2004, respectively. Since 2008, he has been a Professor with College of Mechanical and Electrical Engineering, NUAA, Nanjing, China. His research interests include data-driven intelligent manufacturing and advanced composite curing technologies.

Author contributions Mr. **HE Fangzhou** is responsible for research and design. Prof. **LIU Changqing** and Prof. **LI Yingguang** provide guidance on ideas. Prof. **GAO James** is responsible for proofreading the paper. Mr. **XU Yiyun** provides relevant models, designs simulation and experimental plans, and experimental data. Mr. **YANG Fan** provides model data. Mr. **TIE Lei**, Mr. **XU Yiyun**, and Mr. **YANG Fan** jointly completed the experimental verification and paper writing. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: XU Chengting)

基于混合近端策略优化的强化学习钛合金结构件加工变形控制方法

何方舟，刘长青，铁　磊，徐逸云，杨　帆，高晓渝，李迎光

(南京航空航天大学机电学院，南京 210016，中国)

摘要：针对钛合金结构件加工变形控制问题，本文提出一种基于混合近端策略优化(Hybrid proximal policy optimization, HPPO)的混合动作变形控制方法。现有强化学习算法通常局限于离散或连续单一动作空间，难以同时优化加工序列与加工余量。为此，将加工序列作为离散动作、加工余量作为连续参数，统一纳入一个参数化的混合动作空间；并以在线变形力监测数据作为状态反馈，实现动态加工条件下的自适应控制。进一步设计了双层奖励机制，融合过程级变形力均匀性与终端变形收敛性，引导智能体同步抑制局部变形与全局变形。在Ti6Al4V航空结构件上的实验验证表明，该方法将平均加工变形由0.103 mm降至0.054 mm，RMSE由0.119 mm降至0.071 mm，较无控制工况降低47.57%，证实了所提方法在实际制造环境中的准确性与有效性。

关键词：钛合金；混合加工工艺；加工变形控制；混合近端策略优化

研究亮点：

1. 提出基于HPPO的混合动作强化学习框架，首次将离散加工序列优化与连续加工余量调整统一到一个参数化混合动作空间中，实现二者的协同优化而非孤立处理。
2. 设计双层奖励机制，通过过程级变形力均匀性约束与终端变形收敛目标，同步抑制加工过程中的局部变形与全局变形。
3. 将钛合金结构件平均加工变形降低47.57%，显著优于传统无控制加工方法。