退役风机叶片热解过程的 SAC-CPO 温度控制方法

张启亮[1]，蒋德凯[1]，冯俊娥[2]，吴玉锋[3†]，翟梦瑜[3]，陶 然[3]

(1. 华北电力大学 自动化系，河北 保定 071003；2. 山东大学 数学学院，济南 250100；
3. 北京工业大学 材料科学与工程学院，北京 100124)

摘 要: 针对退役风机叶片热解过程中存在的温度跟踪精度不足问题, 提出一种由约束策略优化启发的拉格朗日约束软演员-评论家温度控制方法 (SAC-CPO), 所提出方法由模型驱动基线控制、安全规则映射与约束强化学习残差补偿构成. 首先, 通过热重实验和热解动力学分析构建退役风机叶片热解加热过程数学模型; 然后, 基于软演员-评论家 (SAC) 最大熵框架构建残差强化学习温度控制结构; 在此基础上, 引入约束策略优化算法 (CPO) 思想, 设计安全代价函数与拉格朗日乘子自适应调节机制, 提升温度跟踪精度并有效抑制温度超调. 仿真结果表明, 与已有的 PID 和深度确定性策略梯度热解温度控制方法相比, 所提出 SAC-CPO 温度控制方法在500°C阶跃工况下将温度超调量分别降低 65.7% 和 69.9%, 稳态均方根误差降至8.42°C, 在复合工况下超调分别降低 69.7% 和 63.0%. 此外, 在含参数失配与结构失配的多种工况下, 所提出方法仍然保持较低超调量和良好稳态跟踪精度.

关键词: 温度控制；热解过程；强化学习；退役风机叶片

中图分类号: TP273 **文献标志码:** A

DOI: 10.13195/j.kzyjc.2026.0587

引用格式: 张启亮, 蒋德凯, 冯俊娥, 等. 退役风机叶片热解过程的 SAC-CPO 温度控制方法 [J]. 控制与决策.

A SAC-CPO-based temperature control for the pyrolysis process of end-of-life wind turbine blades

ZHANG Qi-liang[1]，*JIANG De-kai*[1]，*FENG Jun-e*[2]，*WU Yu-feng*[3†]，*ZHAI Meng-yu*[3]，*TAO Ran*[3]

(1. Department of Automation，North China Electric Power University，Baoding 071003，China；2. School of Mathematics，Shandong University，Jinan 250100，China；3. School of Materials Science and Engineering，Beijing University of Technology，Beijing 100124，China)

Abstract: For the pyrolysis of end-of-life wind turbine blades, accurately tracking temperature remains challenging and difficult to ensure explicitly. To address these issues, this paper proposes a Lagrangian-constrained Soft Actor-Critic temperature control method inspired by constraint policy optimization (SAC-CPO). This approach integrates model-driven baseline control, safety-rule mapping, and constrained reinforcement learning for residual compensation. Based on thermogravimetric experiments and kinetic analysis, a mathematical model of the pyrolysis heating process is developed, and a residual reinforcement learning temperature control structure is built on the maximum-entropy framework of the soft actor-critic (SAC) algorithm. Inspired by constrained policy optimization (CPO), a safety cost function and an adaptive adjustment mechanism for the Lagrange multiplier are designed to improve temperature tracking accuracy and reduce temperature overshoot. Simulation results show that, compared with PID and deep deterministic policy gradient methods, SAC-CPO reduces overshoot by 65.7% and 69.9%, respectively, under a 500°C step condition, and by 69.7% and 63.0%, respectively, under composite operating conditions. The steady-state root mean square error is reduced to 8.42°C under the step condition, demonstrating improved tracking accuracy and safer, more stable pyrolysis operation. Moreover, under conditions of parameter and structural mismatch, SAC-CPO maintains low overshoot and achieves excellent steady-state tracking accuracy.

Keywords: temperature control；pyrolysis process；reinforcement learning；end-of-life wind turbine blades

0 引 言

随着早期投运的风电机组逐步进入退役期, 退役风机叶片的规模化处置与资源化利用问题日益突出[1-2]. 风机叶片主要由玻璃纤维与树脂基复合材料

收稿日期: 2026-06-17；**录用日期:** 2026-10-22.
基金项目: 国家重点研发计划项目 (2023YFC3906002)；中央高校基本科研业务费专项资金项目 (2026MS181).
责任编委: 刘宝碇.
†通信作者. E-mail: wuyufeng3r@126.com.

构成, 传统的填埋、焚烧及机械粉碎等处理方式难以兼顾环境友好性与高值化回收需求[3-4]. 相比之下, 热解能够在无氧环境下分解有机树脂基体, 实现纤维与基体分离, 并获得可利用的三相产物, 成为当前具有较大潜力的资源化处理路径之一[5].

当前, 针对退役风机叶片热解动力学, 已有研究采用多种方法开展了系统分析. 文献 [6] 基于 KAS、FWO 与 Miura 等转化率方法, 揭示了活化能随转化率从 177-280 kJ/mol 的阶段性演变. 文献 [7] 引入分布活化能模型与独立平行反应模型, 对退役风机叶片的多阶段降解行为进行数学模拟. 文献 [8] 采用 FWO、KAS 与 Friedman 方法, 获得实际退役叶片约 120–125 kJ/mol 的全过程平均活化能. 通过分析现有文献可知, 退役风机叶片热解呈现多阶段、多机理耦合的强非线性特性, 并存在显著的升温速率热滞后效应. 直接将退役风机叶片热解机理模型用于实时控制将产生较高的计算复杂度. 因此, 基于动力学分析构建面向控制的降阶等效模型具有重要意义.

热解温度是影响树脂基体分解程度、纤维性能保持及能耗水平的关键工艺变量. 目前, 关于热解过程温度控制已经取得了一些研究成果. 文献 [9] 提出基于数学模型的模糊 PD 控制器设计, 实现了热解反应器的温度控制. 文献 [10] 针对生物质微波热解的快速响应, 提出模糊 PID 温度控制方法. 文献 [11] 通过引入深度确定性策略梯度 (DDPG), 构建了热传递与热解行为的代理模型. 然而, 上述方法多依赖经验规则或固定参数整定, 在多阶段非线性热解过程、外部扰动变化及执行器约束条件下, 仍难以同时兼顾快速性、稳定性与安全性, 易出现过冲、振荡及控制性能退化等问题.

软演员-评论家 (SAC) 算法是一种基于最大熵框架的离策略深度强化学习算法[12]. SAC 算法引入熵项以维持策略随机性, 结合双 Q 网络结构, 依托离策略学习提高样本利用效率与训练稳定性[13]. SAC 算法已经广泛应用于电动汽车能源管理、电力系统、跟踪控制及无人机等领域[14-17]. 约束策略优化 (CPO) 算法通过在策略更新过程中引入显式约束条件[18], 并结合拉格朗日乘子实现约束的可行性保证[19]. CPO 算法已经广泛应用于机器人控制、自动驾驶等方向, 在满足系统约束的前提下优化控制性能[20-21].

针对退役风机叶片热解过程中存在的温度跟踪精度不足问题, 本文提出一种由 CPO 思想启发的拉格朗日约束 SAC 方法 (SAC-CPO). 该方法在最大熵强化学习框架下构建安全约束优化机制, 通过安全代价函数和拉格朗日乘子自适应调节实现温度控制, 同时结合物理模型构建残差强化学习结构, 提高温度跟踪性能和抗扰能力. 本文的主要贡献如下:

1) 根据文献 [22] 中热重实验与热解动力学分析, 建立退役风机叶片热解加热过程数学模型;

2) 构建融合模型驱动控制、安全规则映射与残差强化学习的 SAC-CPO 温度控制方法, 通过确定性安全机制与约束策略优化提升热解过程温度控制的安全性和鲁棒性.

1 退役风机叶片热解加热过程建模

本文热解加热过程数学建模是基于文献 [22] 的热重实验与动力学分析结论. 由文献 [22] 可知退役风机叶片的主热解反应发生在 290–500 ℃ 温区, 表观活化能随转化率在 150–233 kJ/mol 范围内演变, 均值约 167 kJ/mol, 反应过程伴随显著的升温速率热滞后效应, 呈现多机理叠加的强非线性特征.

退役风机叶片热解转化率定义[23] 为

$$\alpha = \frac{m_0 - m_t}{m_0 - m_\infty}, \tag{1}$$

其中, m_0为初始质量, m_t为t时刻质量, m_∞为最终残余质量. 非等温热解过程可由 Arrhenius 反应速率方程描述为

$$\frac{d\alpha}{dt} = A\exp(-\frac{E_a}{RT})f(\alpha), \tag{2}$$

其中, A为指前因子, E_a为表观活化能, R为气体常数, T为绝对温度, $f(\alpha)$为反应机理函数.

将热解炉、样品和加热装置等效为带执行时滞的一阶非线性集总热系统. 输入为加热功率百分比u_k, 输出为炉内代表温度T_k. 温度变化由加热输入、散热损失、热解反应吸热和外部扰动共同决定. 效比热采用如下的温度相关 Sigmoid 形式:

$$C_{\rm eq}(T_k) = C_0 + \frac{C_r}{1+\exp[-(T_k - T_c)/\beta_c]}, \tag{3}$$

其中, C_0为基础等效热容系数, C_r为主反应区附加等效热容幅值, T_c为等效热容变化中心温度, β_c为过渡宽度, $C_{\rm eq}$为等效集总热容参数. 散热损失$Q_{loss}(T)$由线性散热和高温非线性散热组成, 可以表示如下:

$$\begin{aligned} Q_{loss}(T) = {} & h_c\max(T - T_{amb}, 0) + \\ & h_n\max(T - T_{amb}, 0)^2, \end{aligned} \tag{4}$$

其中, h_c表示线性散热系数, h_n表示高温非线性散热系数, T_{amb}为环境温度. 线性项描述常规传热损失, 二次项用于近似高温区散热增强. 热解反应吸热项$Q_{\rm rxn}(T_k, \alpha_k)$为

$$\begin{aligned} & Q_{\rm rxn}(T_k, \alpha_k) = \\ & \quad Q_0\exp[-\frac{(T_k - T_Q)^2}{2\beta_Q^2}]\max(1-\alpha_k, 0)^{1/2}, \end{aligned} \tag{5}$$

其中, Q_0为主反应区最大等效吸热强度, 由主反应

区的热平衡确定, T_Q为吸热峰中心温度, 由 DTG 实验中主失重速率峰对应的温区确定, β_Q为吸热峰温度宽度, 由 DTG 主峰宽度确定. 由于树脂基体热解过程涉及多组分复杂断裂与相变, 微观瞬时反应热难以直接解析, 式 (5) 将主热解温区的集中吸热特征表征为以T_Q为中心的高斯修正形式, 引入了剩余反应比例项$max(1-\alpha_{k,0})^{1/2}$, 使吸热功率随转化率递推方程同步衰减, 确保能量与反应进程的一致性. 离散温度递推方程为

$$T_{k+1}=T_k+ T_s[\frac{K_h u_{k-d}/100-Q_{loss}(T_k)-Q_{rxn}(T_k,\alpha_k)}{max[C_{\mathrm{eq}}(T_k),C_{min}]}+d_k], \tag{6}$$

其中, T_s为采样周期, u_{k-d}为考虑执行器时滞后的有效加热功率, d为离散时滞步数, K_h为加热增益, d_k为外部热扰动. 反应速率项可以表示如下:

$$k_{rxn}(T_k)=A\exp[-\frac{E}{R(T_k+273.15)}]. \tag{7}$$

转化率递推方程可以表示如下:

$$\alpha_{k+1}= \mathrm{clip}[\alpha_k+T_s\min(k_{rxn}(T_k)\max(1-\alpha_k,0)^n, \dot{\alpha}_{\max}),0,1], \tag{8}$$

其中, n为表观反应级数, $\dot{\alpha}_{max}$为单步最大转化率变化率. 需要说明的是, 该固定参数模型是面向实时闭环控制的降阶等效处理. 为避免变参数多机理模型带来的计算复杂度提升、满足实时性要求, 模型选取 290-450 ℃ 主反应温区 (对应转化率 0.2-0.8) 作为等效工作区间, 活化能、指前因子与反应级数均基于文献 [22] 中该区间的动力学特征标定, 主工作区内反应速率与实验结果等效精度良好. 构建式 (7) 和式 (8) 所示的标称反应速率递推模型未显式描述高转化率阶段活化能快速升高过程, 由此引入的非线性动态与参数摄动误差, 以及外部热扰动, 构成本文采用残差强化学习进行自适应补偿的依据, 从而在保证计算实时性的同时实现高精度温度跟踪. 基于此, 式 (3)-式 (8) 构成了用于强化学习离散非线性热解环境.

2 SAC-CPO 约束强化学习温度控制

2.1 约束马尔可夫建模与状态设计

将热解温度控制问题建模为约束马尔可夫决策过程, 将热解炉表征为环境, 将温度控制器表征为智能体[24]. 每个采样时刻, 智能体接收归一化状态, 输出连续残差动作, 环境根据热平衡模型更新温度和转化率, 并返回奖励和安全代价. 约束马尔可夫可表示如下:

$$\mathcal{M}=(\mathcal{S},\mathcal{A},P,r,c,\gamma), \tag{9}$$

其中, $\mathcal{S}$为状态空间, $\mathcal{A}$为连续动作空间, P为由热解动力学和扰动共同决定的状态转移, r为奖励函数, c为安全代价函数, γ为折扣因子.

为了同时表征热解过程的温度偏差、动态变化趋势、执行器输入及安全约束状态, 使智能体能够在感知当前控制误差的基础上预测潜在超温风险, 从而提高温度跟踪精度和安全控制能力, 采用 8 维状态向量, 初始状态定义为

$$s_k=[e_k,I_k,T_k,\alpha_k,\dot{T}_k,u_{k-1},\hat{d}_k,T_{max}-T_k]^{\mathrm{T}}, \tag{10}$$

其中, $e_k=T^*-T_k$表示控制误差, 欠温时为正, $I_k=\mathrm{clip}(I_{k-1}+e_kT_s,-I_{max},I_{max})$表示积分项, 采用截断处理, 避免长时间偏差导致积分饱和, $\dot{T}_k=\dfrac{T_k-T_{k-1}}{T_s}$表示温度变化率, $T_{max}-T_k$表示当前温度距离绝对安全上限的裕度. 为减小不同状态量纲差异对训练稳定性的影响, 对状态变量进行尺度归一化和限幅处理. 扰动原始估计$\hat{d}_{raw,k}$为

$$\hat{d}_{raw,k}=\frac{T_{k+1}-T_k}{T_s}- \frac{K_h u_{k-d}/100-Q_{loss}(T_k)-Q_{rxn}(T_k,\alpha_k)}{max[C_e q(T_k),C_{min}]}, \tag{11}$$

式 (11) 为状态转移完成后的估计更新式, k时刻决策使用上一周期计算的$\hat{d}_k$, 执行u_k并测得T_{k+1}后, 由式 (11) 更新得到$\hat{d}_{k+1}$并用于$k+1$时刻决策. 外部扰动根据实际温升速率与模型预测温升速率之间的偏差进行估计, 并通过滤波、死区和限幅处理.

2.2 残差功率控制与安全映射机制

目标温度处维持热平衡所需的稳态功率估计$u_{\mathrm{hold}}(T_k,\alpha_k)$可以表示如下:

$$u_{\mathrm{hold}}(T_k,\alpha_k)=\frac{100[Q_{\mathrm{loss}}(T_k)+Q_{\mathrm{rxn}}(T_k,\alpha_k)]}{K_h}. \tag{12}$$

物理基线功率$u_{base,k}$可以表示如下:

$$u_{base,k}=f_{base}(T_k,\alpha_k,T^*-T_k,I_k,\dot{T}_k,\hat{d}_k). \tag{13}$$

由当前温度、转化率、控制误差、积分项、温度变化率和扰动估计共同决定. 扰动前馈$u_{\mathrm{ff},k}$为

$$u_{\mathrm{ff},k}=G(\hat{d}_k)\frac{C_{\mathrm{eq}}(T_k)\hat{d}_{\mathrm{lead},k}}{K_h}\times 100, \tag{14}$$

其中, $\hat{d}_{\mathrm{lead},k}$表示带提前量的扰动估计, G为非对称前馈增益. 热扰动下前馈用于削减功率, 冷扰动下负扰动前馈等价于补热. 用于限制强化学习残差的实际作用幅度的残差门控g_k为

$$g_k = \mathrm{clip}[g_0 + g_e\phi_e(|e_k|) + g_d\phi_d(|\hat{d}_k|) + g_r\phi_r(|\dot{T}_k|),\ g_{\min},\ g_{\max}], \tag{15}$$

目标附近且扰动弱时残差被压缩, 误差大或扰动明显时残差适当放开. 残差功率命令可表示为

$$\begin{cases} u_{\mathrm{cmd},k} = u_{\mathrm{base},k} - u_{\mathrm{ff},k} + \rho g_k a_k, \\ \bar{u}_k = \mathrm{clip}(u_{\mathrm{cmd},k}, u_{\mathrm{low},k}, u_{\mathrm{high},k}), \\ u_k^r = u_{k-1} + \mathrm{clip}(\bar{u}_k - u_{k-1}, -\Delta u_{\downarrow,k}, \Delta u_{\uparrow,k}). \end{cases} \tag{16}$$

其中, $u_{\mathrm{cmd},k}$为限幅前的候选功率, $\bar{u}_k$为监督安全限幅, u_k^r为方向性限速, ρ为残差幅值系数, g_k为状态相关残差门控系数, $u_{\mathrm{low},k}$与$u_{\mathrm{high},k}$分别为动态功率下限与上限, $\Delta u_{\uparrow,k}$为升功率限速, $\Delta u_{\downarrow,k}$为降功率限速, η_k为低通滤波系数, u_k为最终施加到环境的功率百分比. 最终功率输出u_k为

$$u_k = \mathrm{clip}[(1-\eta_k)u_{k-1} + \eta_k u_k^r, u_{\min}, u_{\max}], \tag{17}$$

其中, η_k为低通滤波系数.

以上控制结构遵循残差强化学习范式[25], 智能体的输出为归一化的非线性补偿残差a_k, 最终加热功率由物理基线、扰动前馈、残差门控、安全上下限、方向性限速与低通平滑共同合成, 强化学习承担基线附近的小幅非线性修正. 总体控制框图如图 1 所示.

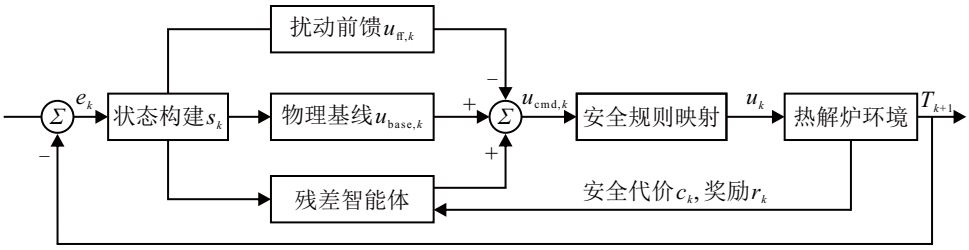

图 1 控制框图

2.3 奖励函数与安全代价设计

奖励函数由多个阶段相关的奖励和惩罚项组成, 目标是在升温速度、目标邻域稳定性、热扰削峰、冷扰恢复、功率平滑和极端安全之间取得平衡, 使控制过程在不同工况下保持一致的性能表现. 奖励函数结构为

$$r_k = r_{\mathrm{tr},k} + r_{\mathrm{prog},k} + r_{\mathrm{safe},k} + r_{\mathrm{dist},k} + r_{\mathrm{sm},k} + r_{\mathrm{hold},k}, \tag{18}$$

其中, $r_{tr,k}$为跟踪目标项, $r_{\mathrm{prog},k}$为升温进度项, $r_{safe,k}$为过温和异常温度惩罚项, $r_{\mathrm{dist},k}$为冷热扰动恢复项, $r_{sm,k}$为功率平滑项, $r_{hold,k}$为目标邻域稳态保持奖励.

本文采用 CPO 思想启发的拉格朗日约束 SAC, 在原始奖励之外计算安全代价, 并用动态乘子λ调节安全约束强度. SAC-CPO 模式在原始奖励之外计算安全代价c_k. 安全代价为连续风险指标, 既考虑当前超温, 也考虑短时预测超温、热扰下高功率、目标附近功率抖振、削峰后欠温和深欠温等风险. 安全代价结构化表达可以表示如下:

$$c_k = \min(\sum_i w_i C_{i,k}^2, c_{max}), \tag{19}$$

其中, c_{max}表示代价上限. 超温风险项$C_{\mathrm{over},k}$为

$$C_{\mathrm{over},k} = \frac{\max[T_k - (T^* + \Delta T_{\max}), 0]}{\Delta T_{\mathrm{scale}}}, \tag{20}$$

其中, ΔT_{max}表示允许的目标温度以上最大过冲裕度. 硬安全上限风险项$C_{\mathrm{hard},k}$为

$$C_{\mathrm{hard},k} = \frac{\max(T_k - T_{\max}, 0)}{T_{\max} - T^*}, \tag{21}$$

其中, T_{max}为系统绝对安全温度上限. $C_{\mathrm{over},k}$和$C_{\mathrm{hard},k}$用于抑制主热解温区附近的温度过冲, 避免树脂异常炭化、玻璃纤维热损伤及能耗增加. 短时预测温度T_{k+m}^{pred}可以表示如下:

$$T_{k+m}^{\mathrm{pred}} = T_k + mT_s\dot{T}_k + \kappa_m \max(\hat{d}_k, 0), \tag{22}$$

目标邻域功率抖振风险$C_{\mathrm{osc},k}$可以表示如下:

$$C_{\mathrm{osc},k} = \mathbf{1}_{|T_k - T^*| < \varepsilon} \frac{\max(|u_k - u_{k-1}| - \Delta u_{\mathrm{tol}}, 0)}{\Delta u_{\mathrm{scale}}}, \tag{23}$$

约束优化目标为

$$\max_{\pi} J_R(\pi),\ \text{s.t. } J_C(\pi) \leqslant \delta \tag{24}$$

约束代价目标为

$$J_C(\pi) = \mathbb{E}_\pi[\sum_{k=0}^{\infty} \gamma^k c_k] \tag{25}$$

用于在尚未越界但趋势已经危险时提前产生约束代价. 拉格朗日修正奖励可以表示如下:

$$\tilde{r}_k = r_k - \lambda c_k, \tag{26}$$

其中, λ为安全代价乘子. 约束代价越大, 有效奖励下降越明显.

2.4 SAC-CPO 约束优化机制

在上述安全代价函数设计基础上, 在 SAC 最大熵框架下引入 CPO 约束思想, 构建拉格朗日约束优化机制. 与 WCSAC 等方法的安全性完全依赖训练

过程不同[26], 本文融合物理基线、安全规则映射与残差强化学习, 由确定性安全层提供基础保障, 学习策略仅承担非线性补偿任务.

SAC-CPO 的约束最大熵目标为

$$J_\lambda(\pi)=\mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k(\tilde{r}_k+\alpha_{\text{ent}}\mathcal{H}(\pi(\cdot|s_k)))], \quad (27)$$

其中, α_{ent}表示熵系数. 拉格朗日乘子更新形式为

$$\lambda_{j+1}=\text{clip}[\lambda_j+\eta_\lambda(\sqrt{\bar{c}_j}-\sqrt{\delta}),\lambda_{\min},\lambda_{\max}], \quad (28)$$

其中, j表示训练回合序号, λ的更新按回合进行, 每个训练回合结束后更新一次, 回合内保持不变. $\bar{c}_j$表示最近一段训练的平均安全代价, δ表示允许的安全代价阈值.

在模型训练过程中, 智能体首先根据当前热解状态s_k输出残差动作a_k, 经由物理基线功率、扰动前馈、安全限幅、方向性限速和低通滤波得到最终加热功率u_k. 随后, 热解环境根据非线性温度模型更新炉温T_{k+1}和转化率α_{k+1}, 并返回奖励r_k与安全代价c_k. 由此得到的交互数据用于策略网络、Q 网络以及拉格朗日乘子的更新.

与 SAC 和 DDPG 相比, 本文提出的 SAC-CPO 具有完全相同的状态向量与动作定义, 以及相同的功率合成链路. 然而, 三者在策略结构与约束机制上均不相同. SAC-CPO 训练超参数见表 1. Physics Baseline 指仅保留物理基线、扰动前馈与安全规则映射、不接入任何学习残差的纯模型控制器, 包含安全限幅、方向性限速与低通滤波环节.

表 1　SAC-CPO 训练超参数设置

参数	取值	参数	取值
Actor网络结构	2×256	经验回放池容量	2×10^5
Critic网络结构	2×256	小批量样本数	256
状态/动作维度	8/1	目标网络平滑系数τ	0.005
Actor学习率	1×10^{-4}	熵权重初值	8×10^{-4}
Critic学习率	3×10^{-4}	熵权重学习率	1×10^{-4}
折扣因子γ	0.99	训练回合数	600
采样周期T_s/s	1	每回合步数	600

3　实验验证

为验证 SAC-CPO 温度控制方法的有效性, 本文将所提方法与已有 PID、模糊 PID、DDPG 等控制方法进行对比[11-13], 并从训练收敛、阶跃响应、复合工况响应及鲁棒性测试等方面进行分析.

设目标温度为T^*, 温度采样序列为$T_k(k=0,1,\ldots,N)$, 本文定义如下控制性能评价指标: (1) 超调量$\sigma=\max_k(T_k-T^*)$, 由于正向超温会加剧风机叶片热解过程中树脂二次裂解与纤维热损伤, 因此超调抑制是本文控制任务的首要目标; (2) 扰动恢复时间, 扰动结束后温度重新进入并保持在±5 ℃ 误差带内所需时间, 度量抗扰恢复能力. (3) 稳态均方根误差$\text{RMSE}_{SS}=\sqrt{\frac{1}{N_s}\sum_{t_k\geqslant120}(T_k-T^*)^2}$. 统计调节阶段之后, 度量调节与稳态阶段的跟踪精度, 不受升温段影响. (4) 最大偏差$e_{max}=max|T_k-T^*|$, 包含负向偏差的双向度量; (5) 总体均方根误差$RMSE_T=\sqrt{\frac{1}{\text{N}+1}\sum_{\text{k}=0}^{\text{N}}(T_k-T^*)^2}$, 统计区间为全程, 升温段误差主导该指标, 主要反映升温快速性. (6) 稳定时间t_s, 为温度首次进入并持续保持在±5 ℃ 误差带内的时刻.

3.1　训练过程与安全约束调节

SAC-CPO 训练过程中的安全代价与拉格朗日乘子变化如图 2 所示. 随着训练难度增加, 安全代价整体升高, 拉格朗日乘子随之自适应增大, 使策略更新更加重视温度超调、功率波动和安全边界约束. 该过程体现了 SAC-CPO 在训练中根据约束违反程度动态调整的重要优化方向.

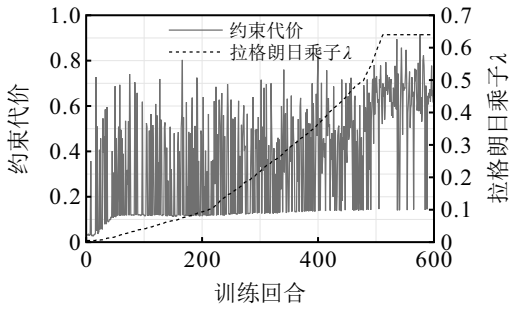

图 2　SAC-CPO 约束代价与拉格朗日乘子变化曲线

3.2　阶跃响应温度跟踪性能分析

目标温度 500 ℃ 的阶跃升温响应曲线如图 3 所示, 不同控制器的性能指标数据如表 2 所示.

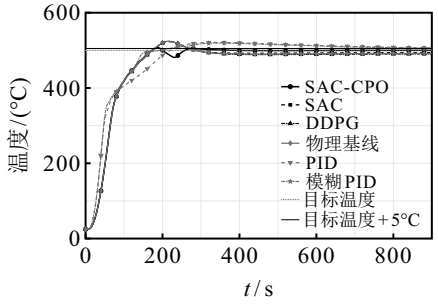

图 3　阶跃响应温度曲线对比

由图 3 与表 2 可知, SAC-CPO 阶跃超调量为 7.27 ℃, 较 SAC 的 23.11 ℃ 和 DDPG 的 24.16 ℃ 分别降低 68.56% 和 69.93%. 表明 CPO 约束机制有效抑制了快速升温阶段的过度加热, 使智能体在接近

表 2 不同控制器阶跃响应性能对比

控制器	超调量/℃	$RMSE_T$/℃	稳定时间/s	$RMSE_{SS}$/℃
SAC-CPO	7.27	110.26	281	8.42
SAC	23.11	110.51	271	11.50
DDPG	24.16	110.72	271	13.51
Physics Baseline	23.84	110.60	271	12.47
PID	21.22	100.80	901	21.55
Fuzzy PID	19.31	98.68	155	13.87

目标温度前能够提前削减功率, 避免热惯性造成温度上冲. SAC-CPO 的稳态 RMSE 为 8.42 ℃, 为所有对比控制器中最低, 说明低超调并不是以牺牲稳态精度为代价获得的. PID 的总体 RMSE 较低, 但稳定时间达到 901s, 说明其到最后没有稳定, 模糊 PID 稳定时间较短, 但稳态 RMSE 仍高于 SAC-CPO, 表明其在快速进入目标附近后仍存在较明显的持续偏差.

3.3 复合工况温度跟踪性能分析

复合工况能够同时检验控制器的热扰过冲抑制、冷扰补热恢复和末端稳定保持能力. 复合工况下温度响应曲线如图 4 所示, 性能指标数据见表 3 和表 4.

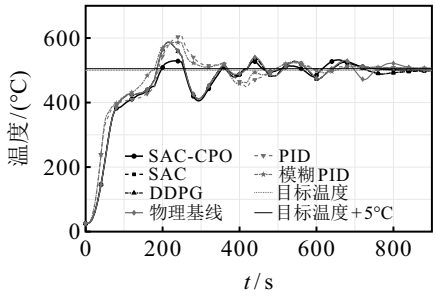

图 4 复合工况温度响应曲线对比

表 3 不同控制器复合工况性能对比

控制器	超调量/℃	$RMSE_T$/℃	稳定时间/s	$RMSE_{SS}$/℃
SAC-CPO	32.24	111.70	460	32.39
SAC	87.20	111.83	827	35.48
DDPG	87.23	111.86	825	35.56
Physics Baseline	87.58	112.03	858	36.21
PID	106.51	102.96	594	37.60
Fuzzy PID	86.28	99.70	901	29.80

表 4 不同控制器复合工况响应性能对比

控制器	最大偏差/℃	恢复时间/s	热扰峰值/℃	冷扰恢复/s
SAC-CPO	94.60	340	29.29	40
SAC	88.03	707	87.20	303
DDPG	88.33	705	87.23	303
Physics Baseline	87.77	738	87.58	438
PID	106.51	474	106.51	174
Fuzzy PID	86.28	516	86.28	171

由图 4 和表 3 可知, SAC-CPO 超调 32.24 ℃, 显著低于 SAC 的 87.20 ℃、DDPG 的 87.23 ℃、PID 的 106.51 ℃ 和模糊 PID 的 86.28 ℃. 较 DDPG 降低 63.02%, 表明约束优化能够在热扰阶段主动限制危险升温趋势.

由表 4 可知, SAC-CPO 的热扰峰值仅为 29.29 ℃, 而 SAC、DDPG 和物理基线均约为 87 ℃, PID 达到 106.51 ℃. 较 SAC 降低 66.42%. 当热扰动导致温度快速上升时, SAC-CPO 的功率输出迅速下调甚至进入较低功率区间, 通过降低外部热输入抵消扰动能量.

在恢复速度方面, SAC-CPO 的整体恢复时间为 340s, 较 SAC 的 707s 缩短 51.91%, 冷扰恢复时间仅为 40s, 较 SAC 的 303s 缩短 86.80%. 然而, SAC-CPO 并非在全部指标上占优. 阶跃工况下其总体 RMSE 为 110.26, 高于 PID 的 100.80 与模糊 PID 的 98.68, 是因为 SAC-CPO 在接近目标前的主动降功率使其升温末段略慢, 是安全导向控制的固有代价. 此外, SAC-CPO 的最大偏差为 94.60 ℃ 高于 SAC 与模糊 PID, 残余偏差主要出现在冷扰区间的负向谷值, 稳态 RMSE 亦略高于模糊 PID, 但模糊 PID 为此付出的超调量为 SAC-CPO 的 2.7 倍. 因此, 本文方法在热扰峰值、超调量与恢复时间上更具优势

3.4 消融实验

为量化 SAC-CPO 各组成机制对控制性能的贡献, 设计 7 个消融变体: (1) 移除 CPO 约束, 退化为普通 SAC 智能体; (2) 固定拉格朗日乘子λ为训练收敛值, 关闭其自适应更新; (3) 移除扰动前馈通道; (4) 移除残差门控; (5) 移除安全限幅与方向性限速; (6) 移除低通滤波; (7) 移除模型驱动物理基线. 各变体分别在标称阶跃工况与复合扰动工况下评估, 结果如表 5、表 6 及图 5、图 6 所示. 消融实验为独立训练运行, 与主实验存在约 1% 的运行间差异, 属正常波动.

表 5 标称阶跃工况消融实验结果

变体	超调量/℃	$RMSE_T$/℃	稳定时间/℃	$RMSE_{SS}$/℃
SAC-CPO	7.18	110.26	283	8.47
(1) 移除CPO约束	23.44	110.55	272	11.83
(2) 固定λ	7.29	110.29	284	8.67
(3) 移除扰动前馈	7.33	110.31	287	8.82
(4) 移除残差门控	7.21	110.27	282	8.45
(5) 移除限幅限速	0.43	78.85	249	4.90
(6) 移除低通滤波	3.48	88.33	167	4.69
(7) 移除物理基线	7.60	115.03	428	33.59

由表 5、表 6 中可知, 各模块对控制性能的贡献

表 6　复合扰动工况消融实验结果

变体	热峰超调/℃	冷谷误差/℃	$RMSE_T$/℃	$RMSE_{ss}$/℃
SAC-CPO	31.01	−21.50	110.61	30.82
(1) 移除CPO约束	87.21	−15.55	111.83	35.50
(2) 固定λ	29.19	−21.34	110.73	31.24
(3) 移除扰动前馈	30.05	−36.36	109.74	27.05
(4) 移除残差门控	33.98	−21.23	110.33	30.03
(5) 移除限幅限速	25.21	−33.66	79.83	21.08
(6) 移除低通滤波	27.95	−31.58	89.05	20.79
(7) 移除物理基线	57.20	−21.04	113.63	39.02

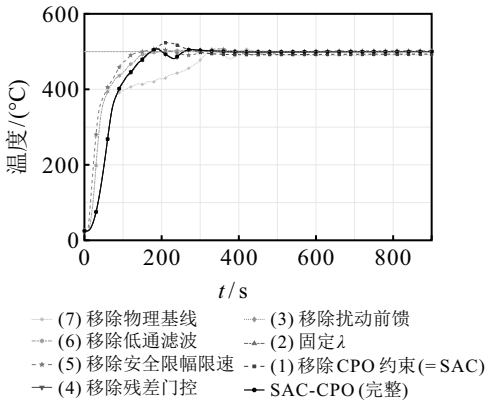

图 5　标称阶跃工况下各变体的温度响应

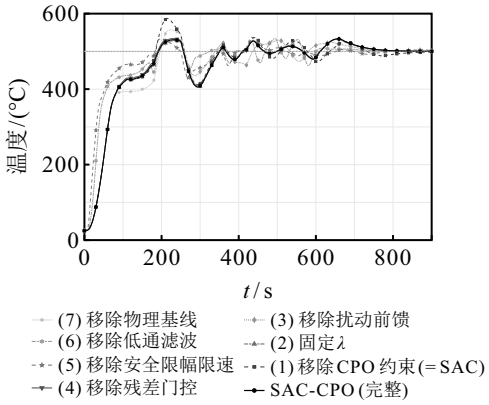

图 6　标准扰动序列下的抗扰响应

层级清晰. 约束机制是超调抑制的首要来源, 移除 CPO 约束后, 标称阶跃工况超调由 7.18 ℃ 升至 23.44 ℃, 复合工况热峰超调由 31.01 ℃ 升至 87.21 ℃, 图 5、图 6 中对应曲线在升温末段与热扰区间均出现明显冲顶. 物理基线是快速性与稳定性的基础. 移除物理基线后, 稳定时间由 283 s 恶化至 428 s, 稳态 RMSE 由 8.47 升至 33.59, 复合工况热峰超调升至 57.20 ℃, 图 6 中其升温段明显迟缓且全程振荡, 因此, 残差结构中的基线粗调与残差精修保证了学习的样本效率与功率量级合理性. 扰动前馈与自适应乘子分别针对特定通道起作用. 移除扰动前馈后, 复合工况冷谷误差由 −21.50 ℃ 恶化至 −36.36 ℃, 热峰变化不大, 说明前馈主要承担冷扰补热功能, 固定 λ在两类工况下性能均略有退化, 验证了乘子自适应调节相对于固定惩罚的增量价值, 移除残差门控各指标小幅退化, 说明了门控的辅助调节作用.

需指出, 移除安全限幅限速与移除低通滤波虽数值上获得更小的超调与 RMSE, 但前者取消了动作级硬安全执行, 超温风险不再受确定性约束, 后者引入高频抖振, 以执行机构磨损为代价. 本文将二者保留为不可学习的确定性安全层, 体现安全优先、性能约束下最优的设计取向. 综上所述, 消融实验表明, 约束机制、物理基线、扰动前馈与自适应乘子对性能各有独立且不可替代的贡献, 所提方法的优异性能源于各模块的协同处理.

3.5　鲁棒性测试

为检验训练后控制器的有效性, 本节进行参数失配鲁棒性测试. 测试中控制器全部内部参数冻结为标称值, 失配仅施加于被控对象一侧, 排除同一模型上循环验证的嫌疑. 失配工况覆盖 7 类 17 种: 加热增益摄动 (+30%、−20%、−30%)、模型时滞失配 (±1、±2 步)、散热系数摄动 (±30%)、反应吸热强度摄动 (±10%)、等效热容摄动 (±20%)、树脂反应项参数摄动 (±20%) 以及目标温度迁移 (450 ℃、550 ℃).

测试结果如表 7 所示, 除稳态 RMSE 列外, 各列为超调量, 负值为欠温. 各工况为独立评估运行, S0 行与表 2 对应数值存在约 1% 的运行间差异, 属随机训练的正常波动. 在 17 种工况中, SAC-CPO 在多数工况下取得最优或并列最优的超调抑制. 由表 7 可以看出, 反应吸热+10% 工况下, SAC-CPO 超调 4.52 ℃, 而 SAC、DDPG 与物理基线为21.8 ～ 22.4℃, 树脂反应项-20% 工况下, SAC-CPO 超调 7.30 ℃, 其余方法约 23 ℃, 目标温度 450 ℃ 迁移工况下, SAC-CPO 基本无超调, 而 SAC、DDPG 超调达 29.6 ～ 33.5 ℃. 上述结果表明, 约束机制使策略在对象失配下不会利用模型误差激进输出, 失配鲁棒性由约束训练过程中内生获得. 在增益下降 30% 工况下, 即使满功率输出也无法驱动样品越过 410 ℃ 附近的吸热峰, 全部 5 种控制器均欠温约 115 ℃, 该工况反映对象物理极限而非控制器缺陷, 在散热+30% 工况下 SAC-CPO 出现 -12.19 ℃ 欠温, 是热侧安全钳位在持续散热增大时偏保守所致. 总体而言, 冻结策略在除物理不可达边界外的全部失配工况下保持温度响应稳定, 且 SAC-CPO 在多数工况下超调最小, 验证了所提方法对参数不确定性的鲁棒性.

3.6　双节点结构失配交叉验证

为进一步检验模型结构失配的影响, 本节构建

表 7 各失配工况下各控制器超调量与稳态 RMSE(°C)

工况	SAC-CPO	SAC	DDPG	物理基线	PID	稳态 RMSE
S0·标称	7.21	24.09	24.16	23.84	21.22	8.7
S1a·K_h+30%	-3.46	-2.27	-2.27	-2.27	46.51	10.2
S1b·K_h−20%	10.99	10.75	10.68	10.64	−58.35	54.1
S1c·K_h−30%	−114.45	−114.87	−115.06	−115.18	−115.19	118.1
S2a·时滞−2步	6.29	22.25	22.88	22.17	21.14	9.1
S2b·时滞-1步	6.83	23.16	23.11	22.73	21.18	9.1
S2c·时滞+1步	7.67	23.89	23.81	23.97	21.27	10.7
S2d·时滞+2步	9.30	23.76	23.74	23.64	21.31	13.6
S3a·散热+30%	−12.19	8.70	8.57	8.72	−10.84	48.3
S3b·散热−30%	−5.82	−4.77	−5.23	−4.87	59.16	12.0
S4a·吸热+10%	4.52	22.41	21.95	21.81	19.77	44.3
S4b·吸热-10%	4.28	5.70	5.13	5.19	21.79	2.4
S5a·热容+20%	6.25	20.89	20.83	21.59	21.00	10.9
S5b·热容−20%	8.89	26.21	25.59	25.66	21.45	11.8
S6a·树脂+20%	7.15	23.66	23.92	23.10	21.31	21.9
S6b·树脂−20%	7.30	23.11	23.67	23.04	21.18	4.9
S7a·目标450	−0.63	29.59	33.45	26.84	−48.20	22.5
S7b·目标550	8.64	23.25	23.28	23.21	21.76	7.6

独立于训练用单节点模型的双节点集总热解模型, 包括炉气节点与样品核心节点, 炉气节点由加热功率直接驱动并与环境换热, 样品核心节点通过气-固换热升温, 树脂吸热反应作用于样品节点. 该模型的热容、换热系数与反应动力学参数均不取自训练模型, 其中炉气等效热容 5.5、加热增益 30、气-固换热系数 0.55、反应吸热强度 13、活化能 121 kJ/mol、反应级数 1.3, 并经可行性校核确保目标物理可达. 控制器不作任何修改, 直接作用于该结构失配对象, 温度测量取样品核心温度. 该设置对应于以另一套独立热力学模型对冻结策略进行交叉测试, 是对集总单节点假设更严格的检验.

交叉验证结果如表 8 所示. 全部 6 种控制器均能将样品核心温度稳定至目标附近, 验证了控制框架在结构失配下的可行性. 由表 8 可以看出, SAC-CPO、SAC、DDPG、物理基线四种方法超调均为 21.8 ∼ 22.0 ℃, 原因是双节点对象的慢样品动态使物理基线主导功率合成, 残差策略修正空间被压缩, 本实验中强化学习残差的增量优势未充分体现, 传统方法对结构失配的适应性明显弱于模型驱动方法, 模糊 PID 稳定时间 901s, 全程未收敛, 常规 PID 收敛缓慢.

表 8 双节点结构失配对象上 6 种控制器性能

控制器	超调/℃	$RMSE_T$/℃	稳定时间/s	$RMSE_{SS}$/℃
SAC-CPO	21.96	162.29	750	74.50
SAC	21.91	162.55	743	75.13
DDPG	21.83	162.72	768	75.51
物理基线	21.76	162.72	730	75.51
PID	−0.73	154.93	845	80.82
模糊PID	−8.62	144.38	901	53.81

4 结论与展望

本文提出了 CPO 约束思想启发的拉格朗日约束 SAC 温度控制方法. 通过引入安全代价函数和拉格朗日乘子自适应调节机制, 实现了温度跟踪性能与运行安全性的协同优化. 仿真结果表明, 在 500 ℃ 阶跃工况下, 该方法将温度超调控制在 7.27 ℃, 相比 DDPG 降低约 69.9%, 稳态 RMSE 降至 8.42 ℃, 显著提高了温度跟踪精度. 未来将进一步考虑能耗、纤维性能保持率及产物品质等多目标优化需求, 提升工程应用能力并在热解实验台架上对所提方法进行实物验证, 研究在线自适应与迁移学习方法以进一步降低对离线模型的依赖.

参考文献 (References)

[1] Martulli L M, Diani M, Sabetta G, et al. Critical review of current wind turbine blades' design and materials and their influence on the end-of-life management of wind turbines[J]. Engineering Structures, 2025, 327: 119625.

[2] 杨双桥, 王福杰, 蒲志龙, 等. 退役风电叶片高质回收再制造利用[J]. 中国科学: 化学, 2026, 56(3): 619-628. (Yang S Q, Wang F J, Pu Z L, et al. High-quality recycling and remanufacturing of decommissioned wind turbine blades[J]. Scientia Sinica Chimica, 2026, 56(3): 619-628.)

[3] You R, Zhao R X, Zhang F B, et al. Research progress on pyrolysis in the recycling of waste wind turbine blades[J]. Journal of Analytical and Applied Pyrolysis, 2026, 196: 107747.

[4] Menna C, De Simone L, Capozzi V. Mechanical recycling of GFRP wind turbine blades: Evaluating the sustainability and economic potential of recycled fibers[J]. Developments in the Built Environment, 2025, 23: 100710.

[5] Wang C H, Qian K Z, Chen D Z, et al. Pyrolysis-NTP assisted oxidation recovery and modification of glass fibers from waste wind turbine blades and application of recycled fibers for H_2S adsorption[J]. Journal of Cleaner Production, 2026, 561: 148413.

[6] Han M, Bai Y H, Ma Y, et al. Thermal decomposition behavior of retired wind turbine blades: Kinetics, pyrolysis product distribution and characterization[J]. Journal of Environmental Management, 2025, 383: 125515.

[7] Yousef S, Eimontas J, Striūgas N, et al. Pyrolysis kinetic behaviour and thermodynamic analysis of waste wind turbine blades (carbon fibres/unsaturated polyester resin)[J]. Energy Sources — Part A: Recovery, Utilization, and Environmental Effects, 2023, 45(4): 10505-10522.

[8] Chen W M, Ye M Y, Li M X, et al. Characteristics, kinetics and product distribution on pyrolysis process for

waste wind turbine blades[J]. Journal of Analytical and Applied Pyrolysis, 2023, 169: 105859.

[9] Kondratenko Y, Kozlov O, Korobko O. Development of intelligent controllers for pyrolysis reactors control systems based on their mathematical models[J]. Collection of Scientific Publications NUS, 2015(6): 66-74.

[10] Bu Q, Cai J, Liu Y Y, et al. The effect of fuzzy PID temperature control on thermal behavior analysis and kinetics study of biomass microwave pyrolysis[J]. Journal of Analytical and Applied Pyrolysis, 2021, 158: 105176.

[11] Zhao P X, Hui Y Z, Qiu Y H, et al. Surrogate model-integrated deep reinforcement learning for temperature and pyrolysis behavior control in the coal coking process[J]. Fuel, 2026, 405: 136770.

[12] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[PP/OL]. 2018, arXiv: 1801.01290.

[13] Huang D J, Sun S Z, Zhao X M, et al. Online energy consumption optimization in electrolytic copper foil electrodeposition *via* incremental soft actor-critic[J]. IEEE Transactions on Industrial Electronics, 2026, 73(7): 10656-10667.

[14] 张思远, 朱晓庆, 阮晓钢, 等. 基于环境反馈机制的四足机器人运动技能学习[J]. 控制与决策, 2024, 39(5): 1461-1468.
(Zhang S Y, Zhu X Q, Ruan X G, et al. Motor skill learning of quadruped robot based on environmental feedback mechanism[J]. Control and Decision, 2024, 39(5): 1461-1468.)

[15] Wang J H, Du C Q, Yan F W, et al. Energy management of a plug-In hybrid electric vehicle using Bayesian optimization and soft actor–critic algorithm[J]. IEEE Transactions on Transportation Electrification, 2025, 11(1): 912-921.

[16] Wei X, Chan K W, Jaafari K A, et al. Real-time resilient power system operation with defender–attacker soft actor-critic reinforcement learning[J]. IEEE Transactions on Industrial Informatics, 2026, 22(2): 684-695.

[17] 秦文龙, 罗贺, 李晓多, 等. 考虑多换电站的多无人机应急电力巡检路径规划方法[J]. 控制与决策, 2025, 40(8): 2391-2399.
(Qin W L, Luo H, Li X D, et al. Multi-UAV emergency power inspection path planning method considering multiple charging stations[J]. Control and Decision, 2025, 40(8): 2391-2399.)

[18] Achiam J, Held D, Tamar A, et al. Constrained policy optimization[PP/OL]. 2017, arXiv: 1705.10528.

[19] Zhu M, Kong M, Wen Y Q, et al. A multi-objective path planning method for ships based on constrained policy optimization[J]. Ocean Engineering, 2025, 319: 120165.

[20] Duan A Q, Yang C G, Zhao J Y, et al. Safe learning by constraint-aware policy optimization for robotic ultrasound imaging[J]. IEEE Transactions on Automation Science and Engineering, 2025, 22: 2349-2360.

[21] Zhao R, Wang K, Li Y, et al. Safe multi-agent deep reinforcement learning for the management of autonomous connected vehicles at future intersections[J]. IEEE Transactions on Parallel and Distributed Systems, 2025, 36(8): 1744-1761.

[22] Tao R, Wu Y F, Geng J N, et al. Thermochemical recycling of waste glass fiber-reinforced polymers: A research based on experiments and quantum chemical calculations[J]. Journal of Analytical and Applied Pyrolysis, 2025, 192: 107277.

[23] Liu H Y, Li P, Zhou Y G, et al. Evolution of characteristic components from retired wind turbine blades in multiple thermal atmospheres: Thermal decomposition characterization and kinetic behavior[J]. Journal of Environmental Management, 2025, 391: 126341.

[24] 季颖, 王建辉. 基于深度强化学习的微电网在线优化调度[J]. 控制与决策, 2022, 37(7): 1675-1684.
(Ji Y, Wang J H. Online optimal scheduling of a microgrid based on deep reinforcement learning[J]. Control and Decision, 2022, 37(7): 1675-1684.)

[25] Johannink T, Bahl S, Nair A, et al. Residual reinforcement learning for robot control[C]. International Conference on Robotics and Automation. Montreal, 2019: 6023-6029.

[26] Yang Q S, Simão T D, Tindemans S H, et al. Safety-constrained reinforcement learning with a distributional safety critic[J]. Machine Learning, 2023, 112(3): 859-887.

作者简介

张启亮 (1991−), 男, 副教授, 博士, 主要研究方向为退役新能源组件热解过程智能优化控制, E-mail: zhangqiliang@ncepu.edu.cn;

蒋德凯 (2001−), 男, 硕士生, 主要研究方向为退役新能源组件热解过程智能优化控制, E-mail: jiangdekai@ncepu.edu.cn;

冯俊娥 (1971−), 女, 教授, 博士, 主要研究方向为逻辑动态系统分析与控制, E-mail: fengjune@sdu.edu.cn;

吴玉锋 (1980−), 男, 教授, 博士, 主要研究方向为城市矿产再生利用技术, E-mail: wuyufeng3r@126.com;

翟梦瑜 (1994−), 女, 研究员, 博士, 主要研究方向为复合固废循环利用系统调控优化, E-mail: zhaimy00@163.com;

陶然 (1998−), 男, 博士生, 主要研究方向为材料与化工, E-mail: taoran98@126.com.