文章编号:1674-2869(2025)05-0565-06

基于改进近端策略优化算法的在线三维装箱方法

徐 虹[1,2],曾祥进*[1,2],华永斌[1,2]
1.武汉工程大学计算机科学与工程学院,湖北 武汉 430205;
2.智能机器人湖北省重点实验室(武汉工程大学),湖北 武汉 430205

摘 要:为解决现有三维装箱算法优化效率低的问题,本文提出了一种改进近端策略优化(PPO)算法的在线三维装箱方法。首先,基于现实装箱的边界约束、支撑约束、重力约束、碰撞约束等条件,在演员-评论家框架中添加可行性掩码预测网络,限制不可行装箱动作点的选取,以满足现实物流过程中的装箱需求。其次,使用长短期记忆网络替换PPO算法神经网络结构中的全连接层,专注学习高奖励值的样本,以便更快速地优化模型。最后,采用两个不同的数据集进行对比实验,其中数据集1采用随机生成的箱子序列,数据集2采用切割库存的箱子序列,保证实验的全面性。实验结果表明,基于改进的PPO算法缩短了强化学习应用于装箱过程中动作节点的盲目搜索时间。在数据集2中,单个箱子平均码放时间缩短了0.3 s,箱子数量增加了2.7个,空间利用率提升了2.2%。本文提出的优化算法能够有效提高三维装箱问题的空间利用率和降低装载时间,为三维装箱问题的工程化应用提供有效的解决方案和参考。

关键词:三维装箱;深度强化学习;长短期记忆网络;近端策略优化算法
中图分类号:TP39 **文献标识码**:A **DOI**:10.19843/j.cnki.CN42-1779/TQ.202410005

An online 3D bin packing method using enhanced proximal policy optimization algorithm

***XU Hong*[1,2], *ZENG Xiangjin*[*1,2], *HUA Yongbin*[1,2]**
1. School of Computer Science and Engineering, Wuhan Institute of Technology, Wuhan 430205, China;
2. Hubei Provincial Key Laboratory of Intelligent Robotics(Wuhan Institute of Technology), Wuhan, 430205, China

Abstract: To address the low optimization efficiency of existing 3D bin packing algorithms, this study proposed an online packing method based on an enhanced proximal policy optimization (PPO) algorithm. First, based on practical packing constraints including boundary limitations, support requirements, gravity effects, and collision avoidance, a feasibility mask prediction network was incorporated into the Actor-Critic framework. This network restricted the selection of infeasible packing action nodes to meet real-world logistics demands. Second, the fully connected layers in the neural network architecture of the PPO algorithm were replaced with long short-term memory (LSTM) networks. This modification enabled focused learning from high-reward samples, thereby accelerating model optimization. Finally, comparative experiments were conducted using two distinct datasets (dataset 1: randomly generated box sequences, dataset 2: cutting-stock derived box sequences) to ensure the comprehensiveness of the experiment. Experimental results demonstrated that the enhanced PPO algorithm reduced blind search time for action nodes in reinforcement learning-based packing. In dataset 2, average placement time per box decreased by 0.3 s, packed box quantity increased by

收稿日期:2024-10-13
基金项目:国家自然科学基金(61502355)
作者简介:徐 虹,硕士研究生。Email:1677535561@qq.com
***通信作者**:曾祥进,博士,副教授。Email:xjzeng21@163.com
引文格式:徐虹,曾祥进,华永斌. 基于改进近端策略优化算法的在线三维装箱方法[J]. 武汉工程大学学报,2025,47(5):565-570.

2.7 units per container and space utilization rate improved by 2.2%. The proposed algorithm effectively enhances spatial efficiency while reducing operational time, providing a viable solution for problems in engineering applications of 3D bin packing .

Keywords: 3D bin packing; deep reinforce learning; long short-term memory network; proximal policy optimization algorithm

近年来,由于物流行业的蓬勃发展,三维装箱问题吸引了学者们的广泛关注和热门研究。三维装箱问题是经典的组合优化问题[1]。该问题的核心目标是在满足特定约束条件下,将一组给定的箱子装入容器内,以实现装载体积最大化[2]。

在处理大规模三维装箱问题时,由于其计算复杂性,通常无法通过精确算法在合理时间内获得最优解。因此,实际应用中多采用构造启发式和元启发式算法[3],目标是在有限时间内寻找到满意的近似最优解。

启发式算法需要大量的设计工作,并且泛化功能受到限制。相比之下,深度强化学习在解决组合优化问题方面表现出显著潜力[4]。Hu等[5]采用深度强化学习(deep reinforce learning,DRL)寻找更好的物品打包顺序,解决离线三维装箱问题。Hu等[6]利用深度强化学习训练最优的包装策略,以最大化包装效率和稳定性。Zhao等[7]采用深度强化学习,利用演员-评论家框架,解决具有约束动作空间的三维装箱问题。Verma等[8]利用深度强化学习解决箱子到达顺序未提前预知的在线三维装箱问题。

传统的三维装箱算法在满足装箱实际约束的同时,需要面临装箱动作点搜索空间大、计算量大的挑战,往往导致算法难以收敛,运行时间过长。在已知序列顺序的离线装箱[9]过程中有较好的效果,而当涉及在线装箱过程,即箱子到达顺序未知的情况下,存在装箱率较低、单个箱子码放时间过长等问题,从而影响整体的装箱表现。

本文采用近端策略优化(proximal policy optimization,PPO)算法[10],使用裁剪函数[11]确保新策略更新不超过预定的范围,避免过大的策略变化导致算法不稳定,保证算法最大程度地收敛。为满足实际应用需求,在演员-评论家框架中添加可行性掩码预测网络,限制不可行装箱动作点的选取。同时为解决强化学习算法优化效率低的问题,采用长短期记忆(long short-term memory,LSTM)网络[12]替换PPO算法神经网络结构中的全连接层,专注学习高奖励值的样本,更快速地优化模型。实验证明,改进后的算法缩短了强化学习应用于装箱过程中动作节点的盲目搜索时间,能较好地实现在线三维装箱。

1 三维装箱问题建模

将三维装箱问题表述为马尔科夫决策过程(Markov decision progress,MDP)[13],分别用(S,A,P,R,γ)五元组来表示:状态集S为当前容器环境信息;动作集A表示当前可行动作的集合;转移概率$P(s'|s,a)$表示在状态s下采取动作a转移到状态s'的概率;奖励集R表示环境反馈的奖励信息;γ为折扣因子,用于调节未来奖励对当前价值的影响,当γ趋近0时,智能体注重当前时刻的奖励,而当γ趋近1时,智能体会考虑未来的奖励,即长期奖励。在实验中,将γ设置为1以便更好地利用未来奖励信息。策略$\pi: S \rightarrow A$表示状态到动作概率的映射,$\pi(a|s)$表示在s状态下采取动作a的策略。三维装箱问题表示如图1所示。

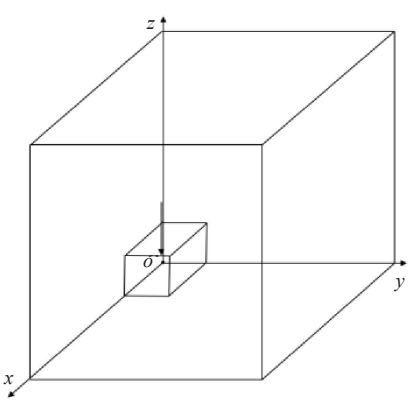

图1 三维装箱示意图

Fig. 1 The schematic diagram of 3D boxing

本文设置4个约束条件,即边界约束、支撑约束、碰撞约束、正交放置约束。边界约束表示货物的各个顶点均在集装箱内部。支撑约束表示货物不可悬空放置,每个货物下表面接触面积至少超过1/2。碰撞约束表示两个货物在xoy、yoz、xoz平面的投影不存在两个投影面相交的情况。正交放置约束表示货物的放置必须与集装箱正交或平行。同时考虑三维装箱的现实因素,箱子有6种不同的放置姿态,如图2所示。

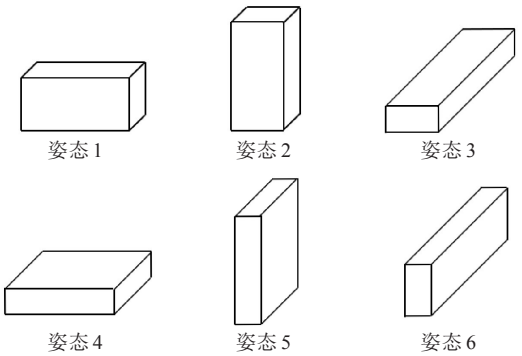

图2　箱子的6种姿态

Fig. 2　The six attitudes of the box

2　改进算法的三维装箱实现

2.1　PPO算法

PPO算法是基于策略梯度的强化学习方法，该算法基于演员-评论家算法的框架实现。演员-评论家算法的框架包含策略网络和价值网络两部分。其中策略网络使用策略函数与环境交互学习如何在给定状态下选择动作，价值网络使用价值函数评估策略网络输出动作的价值，同时指导策略网络下一步的动作。

PPO算法的目标是在与环境交互采样数据后，使用随机梯度上升优化一个“替代”目标函数，从而改进策略。算法在每次策略更新时致力于最小化代价函数，同时通过裁剪函数限制新旧策略之间的差异，确保策略的平滑过渡。目标函数如下：

$$L(\theta)=E_t[\min(r_t(\theta)A_t u(r_t(\theta),1-\varepsilon,1+\varepsilon)A_t)]$$

其中：ε表示截断超参数；r_t表示t时刻新旧策略在样本中的比值；$u(\cdot)$表示截断函数，负责将r_t限制在$[1-\varepsilon,1+\varepsilon]$区间之内，确保样本中新策略更新在预定的范围内，以保证收敛性；A_t为t时刻的优势函数。若优势函数为正数，需要增大新旧策略比值r_t，当$r_t>1+\varepsilon$时，将不提供额外的奖励。如果优势函数是负数，则减少新旧策略比值r_t，但在$r_t<1+\varepsilon$时，不提供额外的奖励，使新旧策略的差异被限制在合理范围内。

价值网络通过最小化价值损失来更新状态价值函数$V(s)$，同时结合策略损失使策略和价值函数协调更新，保证价值估计的精确性。

2.2　状态空间构建和动作空间构建

2.2.1　*状态空间*　将状态空间设置为整个装箱容器，通过降维简化多维问题的复杂性，把三维的装箱空间转化为带高度值的二维平面，第一维表示装箱空间底面的二维平面位置信息，第二维表示每个二维平面所处位置上堆叠箱子的最大高度。每放入1个箱子后，结合装箱位置和箱体长宽高更新此刻平面位置信息以及在该平面处的最大高度，计入状态空间。在环境中，本文建立了用于描述容器内部统一配置情况的高度图，如图3所示，初始高度图为全0状态（左图），放置长宽高分别为4、2、7 m的箱子，高度图更新为右图所示，以二维向量的形式表示离散化空间点上的装箱情况，网格上的数字用于表示当前平面最大堆叠高度。

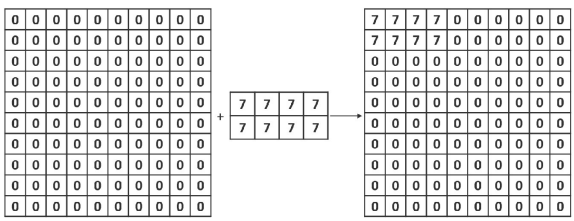

图3　状态空间高度图的更新

Fig. 3　The updation of the state space heightmap

2.2.2　*动作空间*　动作是指将箱子以某种姿态放置到装箱容器中。具体定义为在当前姿态下将待装箱子的后左下角（back-left-bottom，BLB）点放置到容器内部的可行性动作点上。动作空间为箱子按当前姿态放置时容器内部可行性动作点的集合。

2.3　奖励函数的构建

在MDP中，智能体的目标是找到一个将状态映射到行动的策略函数$\pi(s)$。解决MDP问题需要找到使预期累积折扣奖励总和最大化的最佳策略。本文算法致力于最大化装箱利用率，所以将每个装箱的箱子体积作为奖励函数，设计奖励函数$R_1=10\times l\times w\times h/(L\times W\times H)$，其中，$l$、$w$、$h$分别为当前装入箱子的长、宽、高，$L$、$W$、$H$分别为装箱容器的长、宽、高。同时为当前到达的每个箱子合理选择装箱动作点，以便为后续到达的箱子预留更大的剩余空间，设计奖励函数$R_2=D/H+V_{1_2}/(2\times L\times W)$，其中，$D$是当前装入容器中箱子距离容器顶部的高度，$V_{1_2}$是$V_1$平面和$V_2$平面所形成的平面和，$V_1$平面是当前装入容器的前侧最大剩余空间，$V_2$平面是当前装入容器的右侧最大剩余空间，如图4所示。

本文的奖励函数设置为$R=R_1+\omega R_2$，其中ω为权重系数，在实验中设置权重系数$\omega=0.01$，平衡智能体最大化的子剩余空间与当前箱子体积的奖励值比例，以期最大化装箱利用率。

2.4　可行性码放点预测网络

在理想环境下的装箱动作点包括容器内的所有网格点，但在实际物流应用环境下，存在重力约束、边界约束、稳定性约束、碰撞约束等约束条件，

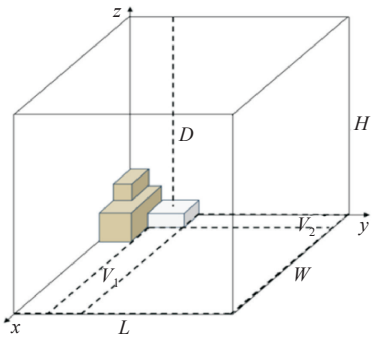

图 4 最大剩余空间的设计

Fig. 4 The design of the maximum remaining space

因此采取的装箱动作必须具有实际意义。在此前提下引入启发式码放点搜索策略,每次装箱成功后,更新容器的状态空间。在提前预知下一个箱子的基础上,可行性码放点预测网络根据此时容器状态空间高度图以及下一个箱子的长宽高信息搜索下一个到达箱子的可行性动作点,与策略网络输出的所有动作概率加权生成可行性动作点概率,智能体根据输出概率随机采取动作。

在线三维装箱算法的核心在于对当前到达箱子码放点的选取,价值网络使用价值函数 $V(s)$ 评估策略网络输出动作的价值,指导策略网络下一步的动作。策略网络输出当前环境中所有动作的概率分布,再经过可行性码放点预测网络过滤掉不可行的动作点,将不可行动作点的概率置为 10^{-3}(相较于0,有利于网络平滑性),智能体根据输出的概率分布随机采取动作。每放置一个箱子后,环境状态更新一次。核心网络框架如图5所示。

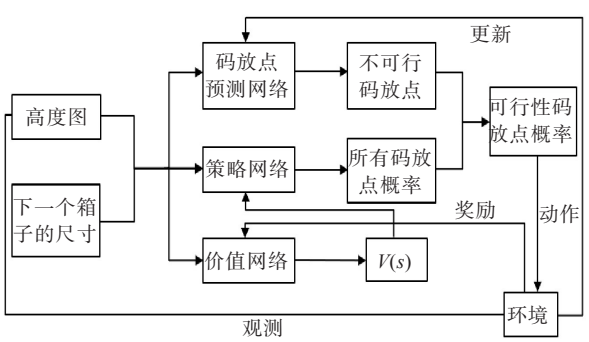

图 5 核心网络框架

Fig. 5 The core network architecture

2.5 改进的PPO算法

为了提升算法的收敛速度同时避免在训练过程中陷入局部最优解,本文对PPO算法进行改进,并引入LSTM网络处理与环境交互时收集的样本信息。LSTM网络利用其记忆单元机制,智能地筛选出需要记忆或遗忘的信息,辅助PPO算法更有效地选择最优动作,并以获得奖励来优化策略。

根据强化学习三维装箱输入的状态信息以及下一个到达箱子的长宽高信息,以当前环境的高度图和下一个箱子的尺寸信息作为输入,当前时刻智能体采取的动作概率分布作为输出,每次执行动作后获得奖励值,通过策略梯度更新网络参数以及自适应优化算法训练得到最终模型。改进的PPO算法网络结构如图6所示。

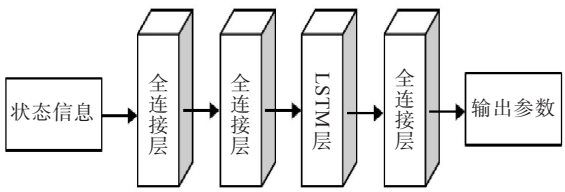

图 6 改进的PPO算法网络结构

Fig. 6 Improved PPO algorithm network structure

3 仿真实验

使用PyTorch搭建网络模型,基于Python语言编码,处理器型号为12th Gen Intel(R) Core(TM) i5-12400KF CPU @2.50 GHz,32 GB。

3.1 训练集和测试集

在仿真环境中,将三维装箱的容器长宽高设置为100 m×100 m×100 m,设计箱子的长宽高不超过容器的1/2,箱子的长宽高在10、20、30、40、50 m之间随机选取,由此可知,共有5×5×5=125种不同尺寸的箱子。训练和测试序列从125种尺寸的箱子中选择生成。

数据集1:从125种尺寸的箱子中随机采样生成货物序列,防止容器不完全填充,设置训练和测试序列的箱子总体积为容器体积的1.2倍。重复上述操作,生成2 000个有效装箱序列。

数据集2:随机采样的缺点是序列的最优性未知。智能体不能提前预知该序列是否会实现成功包装(100%包装),用此基准来衡量包装性能并不可靠。因此使用切割库存[14]的方法生成序列,按顺序将容器“切割”为预定义的125种类型,从该类型生成的有效装箱序列可以完美包装和重新放回容器,以此构建1个有效装箱序列。重复不同的“切割”操作,生成2 000个有效装箱序列。

测试集:分别从数据集1、数据集2中生成50个有效装箱序列作为测试集。对50个有效装箱序列依次进行测试,装箱算法的性能通过空间利用率(space)和容器中装箱的个数(num)进行量化。装箱结果的统计采取50个有效装箱序列装箱效果的平均值。

3.2　参数选取

在实验中，使用固定大小为128个样本的小批量训练方法，通过将数据集划分为大小一致的子集进行迭代训练，以期在模型训练过程中实现更稳定的梯度估计和更快的收敛速度。同时使用具有256个隐藏单元的LSTM网络单元记录箱子的尺寸信息。使用Adam优化器以10^{-3}的初始学习率训练模型，每经过100步衰减0.98倍。裁剪策略超参数为0.2。分别用数据集1和数据集2中的2 000个装箱序列训练模型。设置网络参数和奖励值参数，如表1所示。

3.3　结果分析

表2展示了不同算法在装箱效益方面的结果。在数据集1中，基于内壁点的深度Q网络（deep Q-network，DQN）算法的单个箱子平均码放时间较短，平均装箱率为62.8%。其动作空间为箱子放入容器后与容器壁接触生成的角点的集合，动作空间较小。基于可行性掩码的DQN算法相较于基于内壁点的DQN算法，动作空间范围更加广泛，在状态空间全局搜索码放点，箱子平均码放时间更长，装箱率相较内壁点算法提升了0.9%。DQN算法每次采取最优动作，导致预期奖励总和高于实际值，在实际测试中装箱效果不显著；PPO算法通过输出动作的概率分布，智能体随机采取动作，根据贪心原则，概率大的动作更有机会被采取，平衡了探索与利用，相较于DQN算法，装箱率提升了1.9%。本文算法基于改进的PPO算法同时添加可行性码放点预测模块，与基于可行性掩码的演员-评论家算法[15]相比，装箱个数和装箱率显著提高，装箱个数增加了4.9个，装箱率提升了15.6%。

表1　网络参数和奖励值参数

Table 1　Network parameters and reward value parameters

策略网络学习率	评价网络学习率	样本数	γ	ε	隐藏层个数	隐藏层神经元个数
10^{-3}	10^{-2}	2 000	0.98	0.2	128	256

表2　不同算法的装箱效益

Table 2　The boxing benefits of different algorithms

算法	单个箱子码放时间 / s		装箱个数		装箱率 / %	
	数据集1	数据集2	数据集1	数据集2	数据集1	数据集2
基于内壁点的DQN算法	3.7	3.7	19.1	21.9	62.8	69.4
基于可行性掩码的DQN算法	4.5	4.5	20.2	22.3	63.7	71.2
基于可行性掩码的PPO算法	4.5	4.5	22.9	24.8	65.6	72.9
文献[15]	4.5	4.5	18.2	25.6	50.5	75.2
本文算法	4.2	4.2	23.1	28.3	66.1	77.4

数据集1是随机采样的箱子序列，箱子总体积虽然大于容器体积，但能否完美装入容器的可行性未知，不能很好地作为算法的衡量标准，使用切割库存的数据集2可以较好避免这一问题。在数据集2中，本文算法相较于基于可行性掩码的演员-评论家算法[15]有所提高，平均装箱个数增加了2.7，装箱率提升了2.2%，同时单个箱子码放时间有所缩短。三维装箱效果如图7所示。

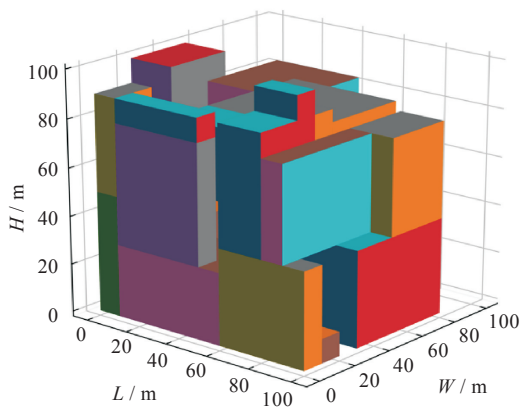

图7　三维装箱效果图

Fig. 7　Three-dimensional packing renderings

上述结果表明，在两个不同的数据集中，与基于内壁点的DQN方法相比，改进的PPO算法在三维装箱过程中更有效地利用了空间资源，提高了箱子的放置数量，并显著增加了装箱率。这是因为基于内壁点的启发式搜索算法动作空间集太小，在算法训练初期由于放置箱子数量较少，生成的动作空间集较小，智能体只能在有限的动作空间集学习策略，容易陷入局部最优，导致算法装箱效率较低。相较于基于可行性掩码的PPO算法和基于可行性掩码的演员-评论家算法[15]，本文算法单个箱子码放时间有所缩短，装箱数量和装箱率

也有较大提升，改进的PPO算法在全局搜索动作点的过程中引入LSTM网络优先学习奖励值大的样本，避免陷入局部最优。综上，本文算法在不同数据集都有更好的适应能力，在平均码放时间、装箱率和装箱个数上都有较大改善。

4 结 论

针对三维装箱问题，本文通过改进PPO算法的强化学习方法，用LSTM网络替换PPO算法神经网络结构中的全连接层，专注学习高奖励值的样本，更快速地优化模型。为简化问题的复杂性，将三维空间转换成带高度值的二维平面。同时引入可行性码放点预测网络搜索动作点，结合装箱体积和最大剩余空间设置奖励函数。实验证明，本文算法缩短了强化学习应用于装箱过程中动作节点的盲目搜索时间，在装箱率以及装箱个数上都有较大提升，能够实现在线三维装箱。

参考文献

[1] CRAINIC T G，PERBOLI G，TADEI R. Extreme point-based heuristics for three-dimensional bin packing[J]. INFORMS Journal on Computing，2008，20（3）：368-384.

[2] 张德富，彭煜，朱文兴，等. 求解三维装箱问题的混合模拟退火算法[J]. 计算机学报，2009，32（11）：2147-2156.

[3] VINOD CHANDRA S S，ANAND H S. Nature inspired meta heuristic algorithms for optimization problems[J]. Computing，2022，104（2）：251-269.

[4] 李明阳，许可儿，宋志强，等. 多智能体强化学习算法研究综述[J]. 计算机科学与探索，2024，18（8）：1979-1997.

[5] HU H Y，ZHANG X D，YAN X W，et al. Solving a new 3D bin packing problem with deep reinforcement learning method[Z/OL]. （2017-08-20）[2025-04-02]. https://doi.org/10.48550/arXiv.1078.05930.

[6] HU R Z，XU J Z，CHEN B，et al. TAP-Net：transport-and-pack using reinforcement learning[J]. ACM Transactions on Graphics，2020，39（6CD）：232.

[7] ZHAO H，SHE Q J，ZHU C Y，et al. Online 3D bin packing with constrained deep reinforcement learning[C]//35th AAAI Conference on Artificial Intelligence. Menlo Park：Association for the Advancement of Artificial Intelligence，2021：741-749.

[8] VERMA R，SINGHAL A，KHADILKAR H，et al. A generalized reinforcement learning algorithm for online 3D bin-packing[Z/OL]. （2020-07-01）[2025-04-02]. https://doi.org/10.48550/arXiv.2007.00463.

[9] 张长勇，刘佳瑜，王艳芳. 求解货物在线装箱问题的融合算法[J]. 科学技术与工程，2021，21（11）：4513-4518.

[10] YU C，VELU A，VINITSKY E，et al. The surprising effectiveness of PPO in cooperative multi-agent games[J]. Advances in Neural Information Processing Systems，2022，35：24611-24624.

[11] HUANG N C，HSIEH P C，HO K H，et al. PPO-Clip attains global optimality：towards deeper understandings of clipping[J]. Proceedings of the AAAI Conference on Artificial Intelligence，2024，38（11）：12600-12607.

[12] 乐恒韬，赵康康，吴松林，等. 基于LSTM网络的机器人异空间手眼标定方法[J]. 武汉工程大学学报，2024，46（5）：574-578.

[13] 高阳，陈世福，陆鑫. 强化学习研究综述[J]. 自动化学报，2004，30（1）：86-100.

[14] GILMORE P C，GOMORY R E. A linear programming approach to the cutting stock problem：part II[J]. Operations Research，1963，11（6）：863-888.

[15] ZHAO H，ZHU C Y，XU X，et al. Learning practically feasible policies for online 3D bin packing[J]. Science China Information Sciences，2022，65（1）：112105.

本文编辑：霍淑利