DOI：10.13973/j.cnki.robot.210008

基于深度强化学习的机器人操作行为研究综述

陈佳盼[1]，郑敏华[1,2]

（1. 北京交通大学机械与电子控制工程学院，北京 100044；
2. 载运工具先进制造与测控技术教育部重点实验室（北京交通大学），北京 100044）

摘 要：通过梳理、总结前人的研究，首先对深度学习和强化学习的基本理论和算法进行介绍，进而对深度强化学习的流行算法和在机器人操作领域的应用现状进行综述。最后，根据目前存在的问题及解决方法，对深度强化学习在机器人操作领域未来的发展方向作出总结与展望。

关键词：深度学习；强化学习；机器人操作；深度强化学习；机器人学习

中图分类号：TP242.6 **文献标识码**：A **文章编号**：1002-0446(2022)-02-0236-21

A Survey of Robot Manipulation Behavior Research Based on Deep Reinforcement Learning

CHEN Jiapan[1]，ZHENG Minhua[1,2]

(1. *School of Mechanical, Electronic and Control Engineering, Beijing Jiaotong University, Beijing* 100044, *China*;
2. *Key Laboratory of Vehicle Advanced Manufacturing, Measuring and Control Technology* (*Beijing Jiaotong University*), *Ministry of Education, Beijing* 100044, *China*)

Abstract: By summarizing previous studies, the basic theories and algorithms of deep learning and reinforcement learning are introduced firstly. Secondly, the popular DRL (deep reinforcement learning) algorithms and their applications to robot manipulation are summarized. Finally, the future development directions of applying DRL to robot manipulation are forecasted according to the current problems and possible solutions.

Keywords: deep learning; reinforcement learning; robot manipulation; deep reinforcement learning; robot learning

1 引言（Introduction）

随着机器人技术的发展，机器人被广泛应用于医疗、工业、国防以及家庭服务等领域。机器人在人工示教、遥操作以及复杂编程等传统方法的基础上，经过训练后具备一定的操作技能，并且在结构化环境下可以快速准确地完成任务[1-2]。然而，在智能化时代，机器人面对的往往是复杂多变的非结构化环境，传统的机器人技术会面对一些难题，比如机器人不具备处理未知环境的能力、开发时间长以及专业技能需求高等[3]。在一些情况下机器人仅能完成固定工作且不能泛化到新任务[4]。为了使机器人技能泛化到新环境中，机器人需要不断地与环境交互和学习，提高应对复杂环境的能力[5]。

随着人工智能（artificial intelligence，AI）的发展，AI 赋予了机器人强大的学习能力，使机器人学习更快并且缩减了机器人操作技能的开发时间，机器人的学习能力在一定程度上甚至能达到人类的水平[6-7]。在 AI 背景下，机器学习为机器人领域带来了新的机遇[8]，尤其是强化学习（reinforcement learning，RL）。RL 是机器人与环境不断交互，进而不断强化自身决策能力的过程。RL 不仅可以有效地解决复杂编程的问题，而且已经在机器人操作领域得到了广泛应用。深度学习与强化学习结合形成的深度强化学习进一步提升了机器人学习操作技能的能力。深度强化学习（deep reinforcement learning，DRL）将深度学习的感知能力和强化学习的决策能力相结合，可以直接根据输入信息控制机器人的行为，赋予了机器人接近人类的思维方式，是机器人获得操作技能非常重要的方法。机器人技能学习是使机器人通过交互数据，从行为轨迹中自主获取和优化技能，并应用于类似的任务[9]。机器人操作技能作为机器人与外界交互的重要技能之一，对机器人的发展应用具有重要意义。

近年来，机器人操作行为的研究已经成为机器人领域的研究趋势和热点[10-11]。但 RL 应用于机器人操作行为的研究存在数据特征提取困难和机器人

基金项目：中央高校基本科研业务费专项资金（2019RC028）；国家自然科学基金（51705016）.
通信作者：郑敏华，mhzheng@bjtu.edu.cn 收稿/录用/修回：2021-01-04/2021-04-02/2021-07-07

缺乏感知能力等问题。因此，深度学习与强化学习的结合必不可少。

本文首先对基于深度强化学习的机器人操作行为研究进行了概述，然后介绍了深度学习和强化学习的核心概念和算法模型、深度强化学习的流行算法及原理以及深度强化学习在机器人操作领域的实际应用以及存在的问题，最后对深度强化学习在机器人操作领域的应用研究进行展望和总结。

2　概念和术语（Concepts and terminology）

2.1　深度学习

深度学习侧重于对事物的感知和表达，其核心思想是通过多层网络结构和非线性变换，将低层次数据特征映射为易于处理的高层次表示，以发现数据之间的联系和特征表示。深度学习使用多层结构抽象表征数据特征以构建计算模型，足够复杂的结构可以处理高维度的原始数据。深度学习的模型主要有深度信念网络（deep belief network）、卷积神经网络（convolutional neural network，CNN）、循环神经网络（recurrent neural network）等。

CNN 是前馈神经网络，经典的 CNN 由一个或多个卷积层和顶端的全连接层组成。CNN 使用反向传播算法训练模型，在图像处理方面应用广泛。对 CNN 进行改进的典型工作如下：Krizhevsky 等[12]提出 AlexNet 深度卷积神经网络，该网络引入了全新的深层结构，并采取随机丢弃部分隐藏神经元的方法抑制过拟合现象；Simonyan 等[13]通过增加网络层数，提出了 VGG-Net 模型，图像识别准确率进一步提升；Lin 等[14]通过增加卷积模块，利用多层感知卷积层提取图像特征，大大降低了图像识别错误率。研究表明，CNN 图像识别具有良好的性能，为基于视觉的机器人操作研究工作提供了技术保证。

将深度学习方法应用到机器人操作领域具有一定的挑战性，其中包括状态估计中存在噪声干扰、奖励函数难以确定、连续行为空间难以处理等[15]。但是仍有研究人员在基于深度学习的机器人操作领域进行了深入研究：杜学丹等[16]提出了基于深度学习算法的机械臂抓取方法，在 Universal Robot 5 机械臂上验证了方法的有效性和鲁棒性。伍锡如等[17]运用 CNN 进行图像处理以定位目标，并通过六轴柔性工业分拣机器人验证了模型的识别精度可达 98%。除此之外，深度学习已经成功应用在机器人推动目标物[18]、操作 3 维物体模型[19]和操作容器倾倒液体[20]等任务。

然而，基于深度学习训练的机器人模型不具备行为决策能力和对未知环境的适应能力，因此强化学习的应用不可或缺。

2.2　强化学习

2.2.1　强化学习算法原理

强化学习算法的原理是智能体不断与环境交互，理解最佳的行为方式，最终学习到最优的行为策略。智能体与环境的交互过程如图 1 所示。

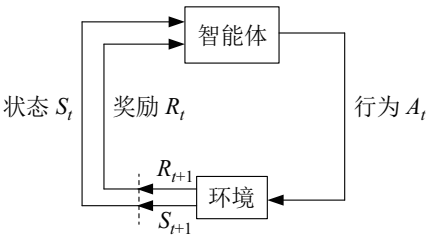

图 1　智能体与环境的交互过程

Fig.1　Interaction process between agent and environment

智能体与环境的交互过程是智能体在 t 时刻，采取行为 A_t 并作用于环境。然后环境从 t 时刻状态 S_t 转变到 $t+1$ 时刻状态 S_{t+1}，同时奖励函数对 A_t 进行评价得到奖励值 R_t。智能体根据 R_t 不断优化行为轨迹，最终学习到最优的行为策略。

标准的强化学习算法包括 4 个要素：行为策略函数、奖励函数、价值函数和环境模型[21]。行为策略函数是智能体的行为准则，其本质是映射，将环境状态集合 S 映射到行为集合 A 的概率分布函数或者概率密度函数，指导智能体选择最佳行为；奖励函数是智能体的评价标准，通常对好的行为给予正奖励，反之给予负奖励；价值函数用来描述当前状态的好坏，包括状态价值函数和行为价值函数；环境模型负责感知环境的变化，不同于真实的环境。

整个强化学习过程可简化为马尔可夫决策过程（Markov decision process，MDP），假定所有状态均具备马尔可夫性[22]。MDP 可用一个五元组 $\langle S,A,P,R,\gamma\rangle$ 表示，各元素意义如下：

S 是有限状态集合，即智能体在环境中探索到的所有可能状态。s 表示智能体在当前时刻的状态，s' 表示智能体在下一时刻的状态。

A 是有限行为集合，即智能体根据环境状态采取的所有可能行为的集合。a 表示智能体在当前时刻采取的行为。

P 是状态转移函数，定义如下：

$$P_{ss'}^{a}=P\{S_{t+1}=s'|S_t=s,A_t=a\} \tag{1}$$

R 是奖励函数，即智能体基于状态 S_t 采取动作 A_t 后，在 $t+1$ 时获得的期望奖励，公式表示如下：

$$R_s^a = E(R_{t+1}|S_t = s, A_t = a) \tag{2}$$

γ 表示折扣因子，即未来的奖励在当前时刻的价值比例。

在 MDP 中，价值函数包括状态价值函数和动作价值函数。状态价值函数表示在策略 π 的指导下当前时刻状态 s 所获得的期望回报，定义如下：

$$v_\pi(s) = E_\pi(G_t|S_t = s) \tag{3}$$

动作价值函数表示在策略 π 的指导下，根据状态 s，采取行为 a 所获得的期望回报，定义如下：

$$q_\pi(s,a) = E_\pi(G_t|S_t = s, A_t = a) \tag{4}$$

其中 G_t 表示折扣奖励，定义如下：

$$G_t = R_{t+1} + \gamma R_{t+2} + \cdots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \tag{5}$$

最优状态价值函数和最大动作价值函数分别用式 (6) 和式 (7) 表示，定义如下：

$$v_*(s) = \max_\pi v_\pi(s) \tag{6}$$

$$q_*(s,a) = \max_\pi q_\pi(s,a) \tag{7}$$

在 MDP 中，通过选取最大动作值函数求解最优行为策略，定义如下：

$$a = \arg\max_{a \in A} q_*(s,a) \tag{8}$$

2.2.2 强化学习算法分类

(1) 无模型（model-free）算法和基于模型（model-based）的算法

无模型强化学习算法是智能体通过与环境交互产生的样本数据，直接优化动作，而不是拟合模型。该算法以最小化偏差的方式与动态环境进行交互，保证算法渐近收敛，最终获得最优解。但是，无模型算法在样本数据收集方面非常昂贵，相对简单、低维度的行为也可能需要百万级数据，高维度的复杂行为需要花费更多的时间和精力。除此之外，无模型算法对超参数（比如学习率）非常敏感，微调参数后才能达到较好的结果。

基于模型的强化学习算法是智能体根据其与环境交互产生的数据，训练并拟合模型，然后智能体基于模型优化行为准则。在基于模型的算法中，智能体可以推断未知的环境状态，提前计算状态转移概率和未来期望奖励，提高了样本效率。然而，该算法对未知的、复杂的动态环境难以精确地建模。因此，模型可能存在严重的偏差，且不能保证算法最优解渐近收敛，导致难以产生有效的行为策略。

总之，无模型的和基于模型的强化学习方法没有绝对的好坏之分，不同的任务设计需求对应不同的算法类型。基于深度强化学习，机器人操作行为研究多采用无模型强化学习方法。

(2) 基于价值（value-based）的算法和基于策略（policy-based）的算法

在基于价值的强化学习算法中，动作选择策略固定不变，如 ε 贪婪策略 [6]。特定状态下，动作的选择方式相同。动作价值函数对智能体的行为进行评价，价值越大则动作越好，最佳动作定义如下：

$$a = \arg\max_a q(s,a) \tag{9}$$

基于价值的算法计算所有可能的动作值，但只选取价值最大的动作。但是，该算法计算量大，存在振荡不收敛现象，而且难以解决连续动作空间的问题。Gu 等 [23] 利用值函数处理连续动作空间任务，但是过多的假设条件导致算法鲁棒性差，对于不同的任务泛化能力较弱。

在基于策略的强化学习算法中，动作选择策略动态变化，导致智能体可以学习到不同的行为策略。基于策略的算法中，智能体直接输出下一时刻可能动作的概率，然后根据概率选择行为。该算法将策略参数化，将累积回报的期望值作为目标函数，并通过梯度策略法求解目标函数的最优解 [24]。目标函数定义如下：

$$J(\theta) = E(G_t|\pi_\theta) \tag{10}$$

总之，基于策略的算法能直接优化目标、提高模型的稳定性和可靠性；基于价值的算法能间接估计价值函数，稳定性差但样本效率较高。针对深度强化学习在机器人操作行为中的研究和应用，并且考虑到真实机器人的实际训练，基于策略的算法应用更为广泛。

3 深度强化学习（Deep reinforcement learning）

3.1 深度强化学习概述及分类

深度强化学习可以直接根据输入的原始数据进行动作选择，是一种更加接近人类思维方式的人工智能算法 [25]。深度学习通过学习深层的非线性网络结构和数据集的本质特征，实现函数的逼近 [26]。智能体在与环境交互的过程中，利用强化学习通过不断试错和最大化累积奖励来生成最优的行为策略 [21]。近年来，深度强化学习 [27] 已经成功应用到围棋 [28-31]、视频游戏 [32-38] 和多智能体 [39] 等领域。

许多公司及机构，如 DeepMind 公司、OpenAI 公司及加州大学伯克利分校等，基于深度强化学习对机器人行为的研究做出了突出贡献。DeepMind 公司成功将深度强化学习算法应用到连续动作领域，比如机器人操作和运动等[40]。Heess 等[41] 基于分布式近端策略优化算法，使用前向传播的简单奖励函数，在多种具有挑战性的地形和障碍物上，成功训练了多个虚拟人物完成跑酷任务。OpenAI 公司提出了新型的近端策略优化算法[42]，成功训练多腿机器人相互玩游戏，并指导机器人不断适应彼此策略中的增量变化[43]。加州大学伯克利分校提出策略搜索算法，该算法迭代拟合局部线性模型以优化连续的动作轨迹[44]，并且训练机器人成功完成了拧瓶盖任务[45]。总之，上述研究工作极大地促进了机器人领域的发展，为机器人实现智能化提供了强大的技术支撑。

下面对主流的深度强化学习算法进行归纳和对比，见表 1，其中算法的对比结果均来源于算法首次被提出的工作，且改进型算法的优点均是相对于原版算法。表 1 所有算法名称均为缩写形式，详细的介绍见 3.2 节。

3.2　典型深度强化学习算法

3.2.1　深度 Q 网络算法

强化学习的经典 Q 学习算法和神经网络结合形成 DQN（deep Q network）算法[46]，有效地解决了 Q 学习算法计算效率低和数据内存受限的问题。DQN 算法的亮点是目标网络和经验池，结构示意图如图 2 所示。DQN 算法的网络结构由目标网络和估计网络组成，这 2 个网络的结构相同但参数不同。估计网络具有最新的网络参数，计算当前状态－动作对的价值，并且定期更新目标网络的参数，使其计算目标 Q 值。双网络结构打破了数据之间的相关性，使得 DQN 算法具有更好的泛化性。经验回放部分储存了智能体历史行为信息，打破了经验池中数据相关性和非静态分布问题。

DQN 算法的更新方式如下：

$$Q(s,a) \leftarrow Q(s,a)+\alpha\left(r+\gamma\max_{a'}Q(s',a')-Q(s,a)\right) \tag{11}$$

s' 和 a' 分别表示下一时刻的状态和动作，r 和 γ 分别表示行为奖励和折扣因子。

DQN 算法的损失函数为

$$L(\theta)=E\left[\left(r+\gamma\max_{a'}Q(s',a';\theta)-Q(s,a;\theta)\right)^2\right] \tag{12}$$

DQN 算法的行为策略为

$$a_t=\arg\max_{a}Q(\boldsymbol{\phi}(s_t),a;\theta) \tag{13}$$

$\boldsymbol{\phi}(s_t)$ 表示状态的特征向量，θ 为网络参数。

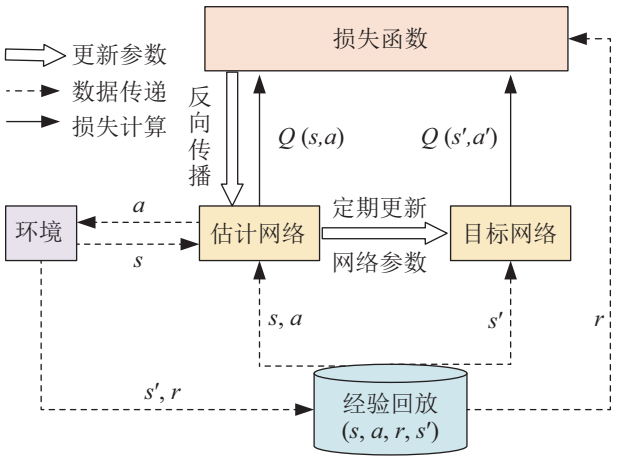

图 2　DQN 算法结构示意图

Fig.2　Structure of DQN algorithm

DQN 算法一般通过贪婪策略的最大化原则选择 Q 值，但这会导致算法过度估计，并且产生较大的偏差。为了解决策略过度估计价值问题，van Hasselt 等[47] 提出 Double DQN（DDQN）算法，该算法解耦动作价值计算和动作选择。相比于 DQN 算法，DDQN 算法首先通过估计网络选择最大 Q 值对应的动作，然后利用该动作在目标网络中计算目标 Q 值。DDQN 算法有效地解决了 DQN 算法过度估计的问题。Wang 等[48] 提出 Dueling DQN 算法，优化了 DQN 算法结构。该算法将网络分成价值函数和优势函数 2 部分。价值函数仅与状态有关，与具体采取的动作无关；优势函数与状态和动作都有关。Dueling DQN 算法将价值函数和优势函数的线性组合作为最终输出，相对于 DQN 算法具有较强的表达能力。Bellemare 等[49] 提出了 C51（categorical 51-atom DQN）算法，直接对价值的分布进行建模，采用 KL 散度（Kullback-Leibler divergence）作为损失函数。不同于 DQN 算法直接对价值期望建模，C51 算法对价值的分布进行建模，其优势是提高了学习稳定性，而且近似分布也降低了基于非平稳策略开展学习所造成的影响。不同于 DQN 算法直接对价值期望结果进行建模，C51 算法对价值分布进行建模，其优势是提高了学习稳定性，其延伸版本是 Dabney 等[50] 提出的 QR-DQN（quantile regression DQN）算法。该算法通过损失函数拟合价值的分位数，进行反向传播计算，其稳定性和精度均优于传统的 DQN 算法。

表 1 深度强化学习算法比较

Tab.1 Comparison of DRL algorithms

算法名称		无模型算法		基于模型的算法	特点	优点	实验环境
		基于策略	基于价值				
DQN	原版[46]		√		目标网络，经验池	减小数据相关性	Atari 2600 游戏
	改进 1 Double DQN[47]		√		解耦动作价值计算和动作选择	解决了价值过高估计	Atari 2600 游戏
	改进 2 Dueling DQN[48]		√		价值函数，优势函数	速率快且收敛效果好	Atari 2600 游戏
	改进 3 C51[49]		√		直接对价值期望建模	提高了学习稳定性	Atari 2600 游戏
	改进 4 QR-DQN[50]		√		分布式，分位数回归	提高了稳定性和精度	Atari 2600 游戏
AC	原版[51]	√	√		集成值函数估计和策略搜索算法	相比值函数算法更快	Atari 2600 游戏
	改进 1 A3C[52]	√	√		异步，多线程	解决 AC 算法难以收敛的问题	Atari 2600 游戏、TORCS 3D 赛车[53]、MuJoCo 仿真器[54]
	改进 2 SAC[55]	√	√		熵正则化	增加探索，加快学习速度	OpenAI Gym 平台[56]
DDPG	原版[40]	√	√		DQN+AC	稳定性高，连续动作空间中表现良好	MuJoCo 仿真器
	改进 1 TD3[57]	√	√		双 Q 网络，“演员”延迟更新，策略平滑化	解决 Q 值高估问题，鲁棒性强	OpenAI Gym 平台
	改进 2 MA-BDDPG[58]			√	动态概率决定数据类型	减小模型数据的负面作用	MuJoCo 仿真器
TRPO	原版[59]	√	√		回报函数单调递增	策略优化性能渐进提高	Atari 2600 游戏 MuJoCo 仿真器
	改进 1 PPO[42]	√	√		重要性采样机制，剪切项	在复杂高维度空间中性能优异	MuJoCo 仿真器
	改进 2 ME-TRPO[60]			√	TRPO + 环境模型	限制智能体在未知环境中迭代	MuJoCo 仿真器
	改进 3 SLBO[61]			√	采用多步 L2 范数损失函数训练动力学模型	少样本、高性能	MuJoCo 仿真器
HER[62]			√		附加目标数据和目标奖励	利用次优数据限制探索	OpenAI Gym 平台
I2As[63]				√	想像力机制，预测信息确定当前动作	准确、高效	Sokoba、MiniPacman
MBMF[64]				√	初始化无模型算法学习器	速度快，鲁棒性强	MuJoCo 仿真器
MBVE	原版[65]			√	Q 值短期预测和长期预测结合	Q 值预测更加准确	MuJoCo 仿真器
	改进 STEVE[66]			√	多个候选目标集成获取最优目标，加权求 Q 值	降低一个数量级样本复杂度	MuJoCo 仿真器
SimPLe[67]				√	直接操作原始像素	减少样本数据	Atari 2600 游戏

3.2.2 演员－评论家算法

演员－评论家（actor-critic，AC）算法[51]的原理是演员部分通过探索环境生成动作集合，然后根据动作概率函数选择动作；评论家部分负责评估演员的动作；演员根据评论家的评分优化动作概率函数，最终指导智能体选择最优动作。演员－评论家

算法的结构示意图如图 3 所示。

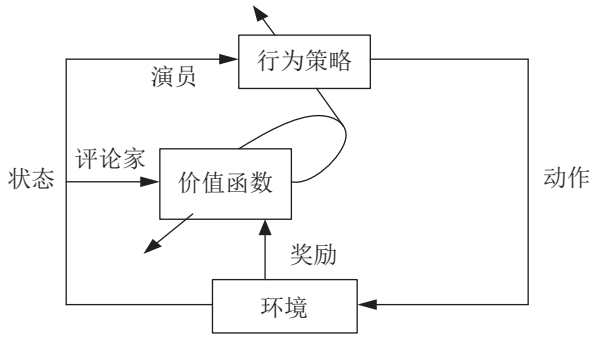

图 3　AC 算法结构示意图 [51]

Fig.3　Structure of AC algorithm [51]

AC 算法采用时序差分误差为评估点，演员的策略函数更新方式为

$$\theta = \theta + \alpha \nabla_{\theta} \log \pi_{\theta}(s_t, a_t) \delta(t) E(t) \tag{14}$$

$\delta(t)$ 为时序差分误差，$E(t)$ 为状态的资格迹。

$$\delta(t) = R_{t+1} + \gamma V(s_{t+1}) - V(s_t) \tag{15}$$

$$E(t) = \gamma \lambda E_{t-1}(s) + 1, \quad S_t = s = \begin{cases} 0, & t < k \\ (\gamma\lambda)^{t-k}, & t \geqslant k \end{cases} \tag{16}$$

$V(s_t)$ 为 t 时刻的状态价值，$\lambda, \gamma \in [0,1]$。

评论家根据估计动作值和实际动作值的均方差更新网络参数，损失函数为

$$L = \frac{1}{n} \sum_{i=1}^{n} \delta_i^2 \tag{17}$$

A3C（asynchronous advantage actor-critic）算法相对于 AC 算法的改进包括训练框架异步化、网络结构优化以及评估点优化等 [52]。A3C 算法借鉴 DQN 算法的经验回放并利用多线程的方法，使多个智能体同时与环境进行交互，并将智能体的学习数据汇集到公共空间，以使得每个智能体都能从公共空间采样并共享数据，指导自身的策略学习，解决了 AC 算法难以收敛的问题。

SAC（soft actor-critic）算法是当今最有效的无模型算法之一，通过最大熵方法保证算法的稳定性和有效性 [55]。在 SAC 算法中，演员同时最大化期望和策略分布的熵，并且在选取最优行为的同时保证行为策略的随机性。相对于 AC 算法，SAC 算法高效且稳定，对不同环境具有更强的鲁棒性。

3.2.3　深度确定性策略梯度算法

Lillicrap 等 [40] 在确定性策略梯度（deterministic policy gradient，DPG）算法 [68] 基础上，借鉴 DQN 算法和 AC 算法的思想，提出了 DDPG（deep deterministic policy gradient）算法，其结构示意图如图 4 所示。

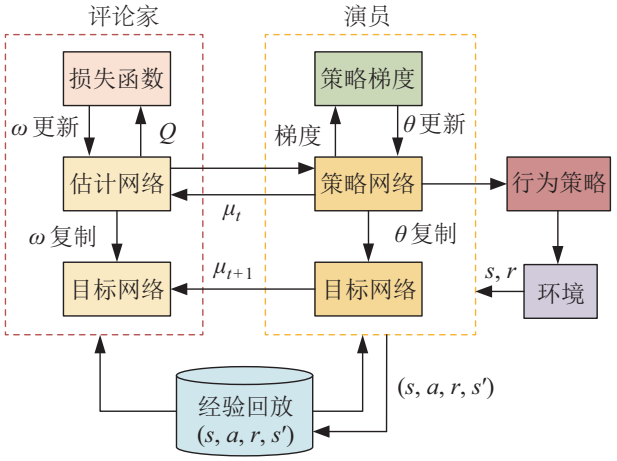

图 4　DDPG 算法结构示意图

Fig.4　Structure of DDPG algorithm

DDPG 算法的网络结构分为评论家模块和演员模块，包含了 4 个神经网络。评论家模块采用时序差分误差（temporal difference error，TD-error）的方式更新网络参数 ω，并且定期复制 ω 到目标网络。演员模块采用 DPG 算法的方式更新网络参数 θ，并且行为策略根据其策略网络的输出结果来选择动作作用于环境。DDPG 算法与 DQN 算法、AC 算法相比具有良好的稳定性，而且能够处理连续动作空间任务。但是 DDPG 算法对超参数的变化很敏感，需要经过长时间的参数微调才能实现较好的算法性能，而且评论家的 Q 函数存在高估 Q 值的问题，会导致行为策略学习不充分，收敛到非最优状态。

双延迟深度确定性策略梯度（twin delayed DDPG，TD3）算法 [57] 对 DDPG 算法的优化包括 3 部分：采用双 Q 网络的方式解决了评论家中 Q 函数高估 Q 值的问题；通过延迟演员的策略更新使得演员的训练更加稳定；利用目标策略平滑化的方法在演员的目标网络计算 Q 值的过程中加入噪声，使网络准确且鲁棒性强。

MA-BDDPG（model-assisted bootstrapped DDPG）算法 [58] 将 DDPG 算法中的经验池分为传统经验池和想像经验池，想像经验池数据来自动力学模型生成的随机想像转换。训练前，智能体计算当前状态行为序列 Q 值的不确定性。Q 值的不确定性越大，则智能体从想像经验池中采集数据的概率越大。该方法通过扩充训练数据集显著加快了训练速度。

3.2.4 信赖域策略优化算法

TRPO（trust region policy optimization）算法[59]针对策略梯度算法[69]中难以选择合适的更新步长的问题，对策略进行改善，使得回报函数单调递增。策略梯度算法中，神经网络直接输出策略函数，根据状态选择动作，其回报函数定义为

$$\eta(\tilde{\pi})=E_{\tau|\tilde{\pi}}\left[\sum_{t=0}^{\infty}\gamma^{t}(r(s_t))\right] \tag{18}$$

τ 为智能体的状态－行为序列。

在 TRPO 算法中，回报函数定义为旧策略与新旧策略回报差之和，定义如下：

$$\eta(\tilde{\pi})=\eta(\pi)+E\left[\sum_{t=0}^{\infty}\gamma^{t}A_{\pi}(s_t,a_t)\right] \tag{19}$$

$$A_{\pi}(s,a)=Q_{\pi}(s,a)-V_{\pi}(s) \tag{20}$$

$\tilde{\pi}$ 为新策略，π 为旧策略，$Q_{\pi}(s,a)$ 表示单个动作值函数，$V_{\pi}(s)$ 表示所有动作值函数关于动作概率的平均值，$A_{\pi}(s,a)$ 大于 0 表明策略 π 生成的动作优于平均动作。

对式 (19) 进行变换：

$$\eta(\tilde{\pi})=\eta(\pi)+\sum_{s}\rho_{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(a|s)A^{\pi}(s,a) \tag{21}$$

其中 $\rho_{\tilde{\pi}}(s)=P(s_0=s)+\gamma P(s_1=s)+\gamma^2P(s_2=s)+\cdots$。

TRPO 算法的最终目标是优化式 (21) 中 $\sum_{s}\rho_{\tilde{\pi}}(s)\sum_{a}\tilde{\pi}(a|s)A^{\pi}(s,a)$，将其进行简化得到：

$$\max_{\theta}E\left[\frac{\pi_{\theta}(a|s_n)}{\pi_{\theta_{\text{old}}}(a|s_n)}A_{\theta_{\text{old}}}(s,a)\right] \tag{22}$$

$$\text{s.t. } D_{KL}^{\max}(\theta_{\text{old}},\theta)\leqslant\delta \tag{23}$$

TRPO 算法首先通过蒙特卡洛方法估计 Q 值，然后根据平均 Q 值得到目标和约束的估计，最后采用共轭梯度和线搜索方法近似解决约束优化问题[70]。TRPO 算法保证了策略优化过程中性能渐进提高。但是，由式 (19)～(23) 的推导可知，其计算量较大，并且策略与值函数之间参数不共享。

针对 TRPO 算法存在的问题，Schulman 等[42]提出了 TRPO 1 阶近似形式的改进型算法，近端策略优化（proximal policy optimization，PPO）算法。

PPO 算法同样有 AC 架构形式，采用重要性采样机制重复利用样本数据，提高了样本效率，限制了采样网络和训练网络的分布相差程度。PPO 算法在目标函数中增加剪切项，将策略更新限制在规定区间内。PPO 算法使用了 1 阶近似形式，相比较 TRPO 算法的 2 阶泰勒展开，在复杂的高维度空间中具有更好的性能，保证了精度和训练速度。

相对于 TRPO 算法，ME-TRPO（model-ensemble trust-region policy optimization）算法[60]采用集成神经网络解决环境中数据不稳定性的问题，并交替进行模型学习和策略学习，对复杂任务具有良好的适应性。

SLBO（stochastic lower bound optimization）算法[61]相对于 ME-TRPO 算法，保证了单调性的提高，并使用 L2 范数损失函数训练动力学模型。少样本训练时，SLBO 算法在多项 MuJoCo 仿真器任务中的性能优于 SAC 算法、TRPO 算法等。

3.2.5 其他深度强化学习算法

HER（hindsight experience replay）算法主要解决了稀疏奖励导致强化学习困难的问题。HER 算法通过附加目标奖励和价值函数，使得智能体到达的每个状态均有目标，且每个目标均对应一套稀疏奖励函数。智能体可以利用失败的探索经历进行动作限制，提高了样本利用率。除此之外，HER 算法将目标数据附加到经验池中，重塑了经验池数据结构。Andrychowicz 等[62]基于 HER 算法进行了机械臂推动、滑动、抓取并放置方块等 3 个任务。实验表明，HER 算法可以结合任意的离线策略算法，并且效果优于原版算法。

I2As（imagination-augmented agents）算法基于模型想像力增强的思想，解决了基于模型的算法存在的模型不准确导致行为预测存在误差的问题。I2As 算法在环境模型想像过程中加入模型训练，并且模拟人类的思维方式，使智能体根据已掌握的知识进行信息预测，以选择当前的动作，增强了无模型智能体的性能。Weber 等[63]在 Sokoban 和 MiniPacman 游戏中对 I2As 算法进行验证，结果表明 I2As 算法在各个模式中的性能均优于标准无模型算法和基于模型的算法。但是，想像过程中加入模型训练增加了模型复杂度，导致参数增多，算法收敛较慢。

MBMF（model-based RL with model-free fine-tuning）算法[64]将神经网络和模型预测相结合，使智能体高效地利用数据学习行为轨迹。MBMF 算法采用基于模型算法的控制器对无模型算法的学习器进行初始化，加快了学习速率。除此之外，MBMF 算法仅训练一次就能通过修改奖励函数，将模型应用到不同的任务，无需重新训练智能体模型。

MBVE（model-based value expansion）算

法[65]引入环境模型，在预测目标 Q 值前，先使环境模型与环境进行多次迭代，然后进行行为价值的预测。因此，目标 Q 值的预测融合了基于环境模型的短期估计和基于目标网络的长期预估，预测结果更加准确。

STEVE（stochastic ensemble value expansion）算法[66] 针对 MBVE 算法的模型累积误差和 Q 值估计误差问题，提出在不同环境中直接展开特定的步数，然后通过计算每一步的不确定性，动态调整 MBVE 算法的权重，以得到更优的 Q 值估计。

SimPLe（simulated policy learning）[67] 是一种完全基于视频预测模型的方法，直接处理每一帧图片。在训练过程中，生成的视频预测器根据每帧像素预测下一帧图像的动作，以自我监督的方式将环境模型的观察像素作为学习策略的监督信号。

4　深度强化学习在机器人操作中的应用（Applications of DRL to robot manipulation）

机器人操作行为是机器人与外界交互的首要条件，机器人只有具备了类似人类的思维方式，才能自主地与外界环境交互。研究表明，基于深度强化学习，机器人能够根据交互信息学习到行为策略，并根据行为策略和环境的状态表征选择合适的操作行为。传统机器人操作研究的局限性表现为：动态环境具有不可预测性、机器人仅在固定位置完成任务且不具备自主学习的能力、机器人技术开发时间长等。部分传统机器人依靠多种传感器采集机器人工作过程信息，传感器信息的融合过程不仅会导致信息丢失，而且会严重压缩信息质量。机器学习技术在计算机视觉领域广泛应用，这使得越来越多的机器人将视觉信号作为输入控制信号[71]。基于视觉的机器人操作系统是从图像中提取视觉特征信息来控制机器人运动，直接根据输入信息，输出机器人的行为。相对于基于多传感器数据融合的机器人操作系统，以基于深度强化学习的视觉信息作为输入的机器人操作系统能够直接将状态信息映射到行为空间，高效且精确。以机器人抓取为例，机器人需要通过视觉采集物体的空间位姿，计算出最佳的抓取位置和方向。因此，基于深度强化学习、以视觉信息为输入的机器人操作行为研究成为机器人操作领域的主流方向。

本节主要讨论基于深度强化学习、以视觉信息为输入的机器人操作行为研究。针对不同的目标物属性，总结了以刚性物体和非刚性物体为操作目标的研究工作；针对不同的模型训练场景，总结了在模拟环境和真实环境中训练模型的研究工作；针对不同的奖励函数设计方式，总结了以稀疏奖励和塑性奖励为行为评价标准的研究工作；针对不同质量的示范数据，总结了示范和次优示范的研究工作，包括模仿学习在其中的应用；针对模型迁移到新任务需要进行大量微调或重新训练的问题，对元强化学习在机器人操作行为中的应用进行了总结。深度强化学习在机器人操作领域的应用总结与对比见表 2，其中列举的工作为基于深度强化学习的机器人操作行为的近期研究成果和被引量较高的工作。

表 2　DRL 在机器人操作领域的应用总结与对比

Tab.2　Summary and comparison of applications of DRL to robot manipulation

类别	代表工作及核心框架	特点	优点	应用
刚体	Zeng 等[73] DQN	推动与抓取 协同策略	推动为抓取 创造空间	清理物体
	Kalashnikov 等[72] QT-Opt	多个机器人 收集数据	视觉闭环 控制系统	清理形状各异 的生活物品
	Jiang 等[81] DDPG	非对称 AC 结构 辅助任务分支	算法收敛速率 快且鲁棒性高	目标点精确定位
	Popov 等[82] DPG+A3C	分布式	提高了 数据效率	堆叠物体
	Zakka 等[94] Form2Fit	形状匹配函数	可泛化应用 于新对象	组装套件
	Khansari 等[96] 有监督的评论家网络	动作图像表 示抓取建议	准确性高	清理形状各异 的生活用品
非刚体	Tsurumine 等[105] DPN	策略更新与 深度网络结合	数据效率高 稳定性强	翻转手帕，叠衣服
	Wu 等[106] SAC	编码拾取和放 置的条件关系	从零开始训练	展开手帕

表 2（续）

类别	代表工作及核心框架	特点	优点	应用
模拟训练	Rusu A 等[112] A3C	渐进网络 辅助模型迁移	减小领域差异	抓取物体
	Peng 等[116] RDPG+HER	领域随机化 辅助模型迁移	弱化领域关系	操作摆放物体
	Hundt 等[119] SPOT	迁移模型适应 长期多步骤任务	模型迁移无 需额外微调	堆叠物体
真实训练	Mahmood 等[122] TRPO	更新延迟	机器人直接训练	目标点追踪
	Finn 等[123] MPC	视频预测+MPC	适应新物体	抓取物体
	Yahya 等[124] DAGPS	分布式，异步式	本体训练 泛化性强	开门
稀疏奖励	Riedmiller 等[126] SACX	调度辅助控制	可靠性高	抓取物体
	Andrychowicz 等[62] HER	加入目标状态	将稀疏问题转化 为非稀疏问题	抓取物体， 目标点追踪
	Li 等[130] Q 学习	关系图结构 课程学习	逐步稀疏奖励	堆叠物体成塔
塑性奖励	Liu 等[131] PPO	基于 OCM 设置奖励函数	瞬时动作奖励	抓取物体
	Zuo 等[132] DDPG	视觉模块 辅助智能体	非线性 奖励函数	目标定位
示范	Rajeswaran 等[139] DAPG	加入成功开 门示范数据	加快算法收敛	开门
	Zhu 等[140] DAPG	加入专家 示范数据	学习时间降低	旋转阀门， 翻转物体
	Gupta 等[88] GPS	手持机械手 示范期望动作	降低样本 复杂度	抓取锤子 敲打木块
	Zhu 等[141] IL+PPO	处理像素 级数据	加快算法收敛， 增强泛化能力	堆叠物体， 倾倒液体
	Berscheid 等[144] SAC	对比损失法	利用目标状态 定义对象位姿	抓取物体 摆造型
次优示范	Mandlekar 等[149] IRIS	适应于目标的 低层控制器+高 层目标选择机制	利用次优示 范数据限制 无效探索	抓取物体
	Gao 等[151] NAC	优化动作 价值函数	降低次优数据 的负面影响	目标点追踪
元强化学习	Goyal 等[166] PixL2R	映射像素 到奖励	快速适应 新环境	对指定物体进 行旋转、推动 等交互操作
	Wu 等[167] SAC	只需一个 视频演示	少量样本 微调模型	抓取容器 倾倒物体
	Yu 等[170]	开源模拟基准	提高模型泛化能力	50 种操作任务

4.1 刚性目标和非刚性目标

大多数机器人的操作目标主要是刚性物体，当机器人操作刚性物体时，物体不会发生形变或者形变可忽略不计。最近几年，家庭辅助机器人的产量逐年递增，机器人被大规模地应用于现实生活当中，而且人们对家庭辅助机器人的性能要求越来越高。家庭辅助机器人的操作对象主要集中在非刚性物体，如衣服、毛巾等。由于机器人操作非刚性物体会导致结构发生变化，非刚性物体的精确建模异常困难，因此，基于不同的操作对象属性选择不同的深度强化学习方法至关重要。

4.1.1 刚性目标

一般机器人操作刚性物体的首要任务是抓取物体，基于视觉的抓取流程主要分为 3 个阶段：感

知、规划和行动。机器人首先通过视觉感知环境，然后确定被操作物体的抓取位置，最后规划到达抓取位置的行为轨迹[72]。

针对在杂乱、物体排列紧密的环境中机器人难以抓取物体的问题，Zeng 等[73] 利用深度相机采集工作空间信息，基于 DQN 算法训练机器人完成推动和抓取物体的协同操作任务，如图 5 所示。Berscheid 等[74] 基于 Zeng 的工作提出了动作最佳位姿算法，减轻了智能体对稀疏奖励的需求，加强了数据学习。相对于 Zeng 等的工作而言，Berscheid 等的工作能够将模型泛化到未知的形状不规则的生活物品，并且抓取效率达到 90% 以上，如图 6 所示。

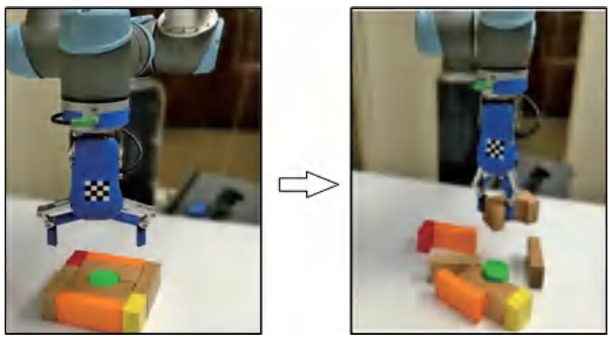

图 5　基于 DQN 算法的机器人操作[73]

Fig.5　Robot manipulation based on DQN algorithm[73]

图 6　机器人移动并抓取木块的实验过程[74]

Fig.6　The experimental process of robot shifting and grasping block[74]

在进一步的研究中，Kalashnikov 等[72] 提出了 QT-Opt 框架，通过多个机器人收集抓取操作数据，经过训练后的机器人抓取未知、形状各异物体的成功率高达 96%，并且可以灵活地应对动态变化，实验配置如图 7 所示。QT-Opt 框架的成果得益于基于视觉的闭环控制系统，其他基于闭环抓取的研究也取得了不错的效果[75-79]。

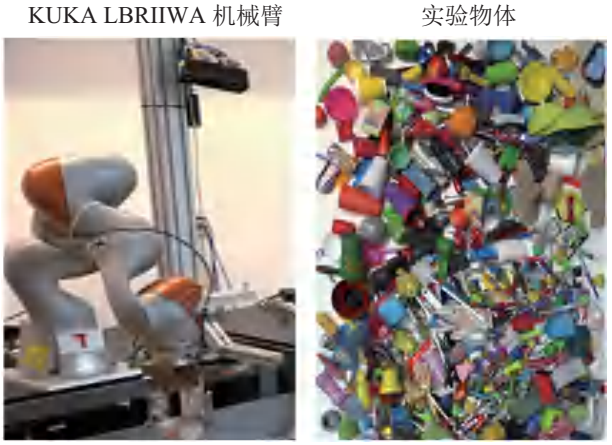

图 7　基于 QT-Opt 框架的机器人操作[72]

Fig.7　Robot manipulation based on QT-Opt framework[72]

针对模型对未知环境适应性差的问题，Mahler 等[80] 利用 Dex-Net4.0 行为策略，以 300 次/h 的平均抓取速度清理 25 个未知物品，可靠性达到 95%，证明了模型在未知环境中的高度适应能力。针对基于视觉数据开展学习时效率低的问题，Jiang 等[81] 设计了基于 DDPG 的机器人操作算法，构造了非对称的 AC 结构，并在演员网络结构中加入了辅助任务分支。机器人在操作实验中通过行为策略成功学习了操作技能，而且非对称 AC 结构和辅助任务分支可以有效提高学习效率和抓取策略性能。

同样，针对算法集中式训练中存在的数据收集率低的问题，Popov 等[82] 基于 DPG 算法，融合 A3C 分布式思想，将训练16 台机器人所用的时间成功缩短到仅 10 h，大大提高了数据收集率。

下列工作均针对机器人操作刚性物体的情形开展研究。Jang 等[83] 使用表示学习（representation learning）方法学习以物体为中心的数据表示并指导智能体完成复杂的任务；Nair 等[15] 提出了以假想目标为条件的策略学习；Li 等[84] 利用强化学习策略指导机械手进行抓取等操作，成功降低了视觉反馈的复杂性；Fang 等[85] 使用面向任务的抓取网络，成功操作工具完成任务且准确率达到 70%～80%，使得机器人握持工具完成操作任务成为可能。除了抓取操作外，深度强化学习已成功应用于灵活机械手操作[86-89]、开门[90-92]、装配[93-95] 等操作任务。其中，Nagabandi 等[89] 提出了基于模型的深度强化学习算法学习多指灵巧手的操作技能方法，仅花费 4 h 即可成功掌握操作技能，如写字。Zakka 等[94] 提出了通用化装配框架（Form2Fit），将装配任务表述为形状匹配问题，并学习了一种通用的匹配函数。该函数对初始条件具备较强的鲁棒性，可以处理新的套件组合并泛化到新的对象。

在机器人操作领域中，大部分机器人动作以矢量形式编码，但是 Khansari 等[96] 提出以动作图像表示抓取建议，并利用深度卷积网络提取抓取任务的局部特征，推断抓取质量。相同的网络结构，以动作图像训练的模型相比于以矢量形式编码的模型，抓取成功率提高了 31%。针对在杂乱环境中抓取特定目标物的问题，Murali 等[97] 提出了 6 自由度抓取方法。行为策略根据目标物的位置进行路径规划并完成抓取任务。在无法立即抓取目标物的情况下，碰撞检测模块可以推理出抓取顺序检索目标物。在真实机器人操作任务中，该方法相对于基准方法成功率提高了 17.6%，使抓取策略具备一定的推理能力。

4.1.2 非刚性目标

针对衣物等非刚性目标的研究方法主要有 2 种：对非刚性目标进行显式建模，不对目标进行建模而直接采用视觉伺服系统[98]。传统非刚性物体的研究是在模拟环境下对布料进行建模，进而寻找最优轨迹[99-101]，但是传统模型难以推广到未知物体。相比显式建模的方法，基于视觉伺服系统，机器人可以根据启发式方法识别衣物的理想抓取点，然后执行任务轨迹[102-104]。本节主要介绍如何解决非刚体的配置空间大、学习效率低和缺乏非刚性目标模拟基准等问题。

Matas 等[98] 设计了基于 DDPG 的任务不可知算法（task agnostic algorithm），利用 RGB 相机采集工作空间信息，完成毛巾折叠和悬挂等任务，如图 8 所示。该方法成功将模型部署到真实环境，在一定程度上解决了非刚体配置空间大的问题。策略编码的方式大大提高了机器人的学习效率和稳定性。

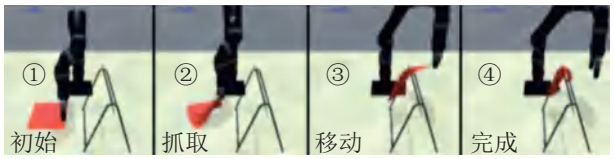

图 8 基于 DDPG 算法的机器人操作毛巾过程[98]

Fig.8 Robot manipulates towel based on DDPG algorithm

Tsurumine 等[105] 将深度神经网络与平稳策略更新相结合，提出了基于动态策略编码的深度强化学习算法 DPN（deep P-network），最终以较少的样本完成了翻转手帕和叠衣服等任务，提高了样本利用率和学习稳定性。同样，Wu 等[106] 提出可迭代的动作空间，对拾取和放置可变形物体的条件关系进行编码，并学习以随机拣选点为条件的策略。该方法加快了机器人操作可变形物体的学习过程，并且是首项从零开始训练可变形物体的工作。

针对非刚体研究缺乏模拟基准的问题，Seita 等[107] 开发了具有 1D、2D 和 3D 变形结构的模拟基准，并将目标条件集成到机器人操作的模型体系结构。该结构根据目标和环境的深层特征推测机器人操作行为的轨迹。Lin 等[108] 提出的处理可变形对象的开源模拟基准 SoftGym 为非刚体的研究提供了便利。

以上研究工作的操作物体具有明显的视觉特征，研究人员能够通过视觉直接采集物体信息。但是对于白色透明物体，3D 传感器难以准确估计物体深度。针对此问题，Sajjan 等[109] 提出了一种深度学习方法（ClearGrasp）。该方法使用深度卷积网络从单个 RGB-D 图像推断物体表面法线、透明表面的遮挡和遮挡边界，并以此估计透明物体的 3D 几何形状，以进行机器人操作。该项工作丰富了机器人操作目标的属性。

深度强化学习在机器人操作非刚体目标的实验中需要指导机器人完成大量的数据采样，但现如今机器人模拟器中缺乏对非刚性物体的精确模型，导致非刚体的操作研究并不广泛，因此基于深度强化学习的非刚体操作任务的研究亟待解决。

4.2 模拟训练和真实训练

基于深度强化学习的机器人操作行为研究的挑战之一是采样效率低。机器人的训练方式有模拟训练和真实训练。在模拟环境中训练机器人是一种有效的方式，不存在安全问题且采样效率高。但由于领域差异，仅用模拟环境下的数据集训练的模型不能保证实体机器人正常工作。在真实环境中，直接通过机器人进行数据收集有利于机器人实际开发，但采样效率低、存在安全隐患。因此，不同的训练环境决定了机器人不同的开发时间和实际效果。

4.2.1 模拟训练

模拟训练的模型主要借助迁移学习转移到真实环境中完成测试和验证。迁移学习可以有效地解决模拟环境和真实环境的领域差异，该方法与深度强化学习模型的结合可以提高模型的可靠性和鲁棒性[110]。迁移学习将模型从模拟环境迁移到真实环境，其本质是知识的迁移再利用。本节内容主要涉及解决模型迁移困难的问题和减少迁移后模型微调的工作量。

Lin 等[111] 在迁移学习的框架下验证了视觉预训练能够有效地提高学习操作对象的泛化能力和样本利用率。通过迁移学习的方法可以有效地解决模型迁移的困难。Rusu 等[112] 通过 A3C 算法在模拟环境训练模型，并通过渐进网络[113] 弥补模型迁移

的差距，成功地指导 Joco 机械臂完成了抓取任务。Gupta 等[114] 基于以轨迹为中心的强化学习算法，利用时变线性高斯策略指导机器人完成目标定位和目标移动等任务，并通过领域自适应[115] 成功完成了模型转移。Peng 等[116] 通过组合递归确定性策略梯度（recurrent deterministic policy gradient，RDPG）与 HER，指导 7 自由度机械手完成了物体布置任务，并通过领域随机化[117] 将模型成功部署到了真实机器人。Haarnoja 等[10] 将 SQL （soft Q-learning）应用到 Sawyer 机器人的推动和堆叠等任务中，并且验证了其鲁棒性优于 NAF （normalized advantage function）[23] 和 DDPG 算法。

通常领域自适应法需要大量未标记的真实世界数据，领域随机化法弱化了建模能力。针对这 2 种方法的共性问题，James 等[118] 提出随机化到规范化的适应网络，不使用现实世界数据克服视觉现实差异。将该训练模型转移到真实世界中，实现了 70% 的未知物体抓取成功率，几乎是领域随机化法的 2 倍，真实数据的使用量比最先进的系统[72] 减少了 99%，样本利用率较高。虽然在模拟环境中训练模型的速度快、成本低，但将模型迁移到真实环境要进行大量的微调工作，且难以保证模型精度。

针对在模拟环境中训练模型存在的问题，Hundt 等[119] 设计了 SPOT（schedule for positive task）框架，直接在真实机器人上加载模拟训练的模型，而无需进行额外的实际微调。同时，该项工作是首次将模拟环境下训练的模型应用于真实环境中的长期多步骤机器人操作任务。同样，Pedersen 等[120] 利用 CycleGAN 算法[121] 缩小模拟环境和真实环境的差距。在将深度强化学习代理直接转移到真实环境中并且不在现实环境中进行模型微调的情况下，对训练阶段对象的抓取成功率能达到 83%，对未知物体的抓取成功率与训练阶段相同。可见，深度强化学习结合 CycleGAN 算法的实验效果优于基准算法，并且具有更强的鲁棒性。

4.2.2　真实训练

机器人进行真实训练有利于实际开发，且不存在模型迁移的问题。本节涉及的研究工作均为在真实环境中对模型进行训练。

最典型最先进的真实训练机器人操作的案例是 Kalashnikov 等[72] 提出的 QT-Opt 框架。该算法侧重于抓取任务的长期推理并分解抓取任务为多个动作序列。实验过程是对 7 个 KUKA LBR IIWA 机械臂进行为期 4 个月的训练，做出 58 万次抓取尝试。实验最终实现了 96% 的抓取成功率，代价大但精度高。Mahmood 等[122] 采用 TRPO 算法直接训练 UR5 机械臂，证明了减少延迟可以有效解决参数设置带来的敏感问题，并利用该方法进行了高度可靠且可重复的实验，表明了基于深度强化学习开发真实机器人的可能性。Finn 等[123] 提出了结合深度动作条件的视频预测模型和使用完全未标记训练数据的模型预测控制（model-predictive control，MPC）的方法。通过 10 个 7 自由度机械臂对数百个对象进行 5 万次的操作，并基于此制作数据集。实验结果表明，直接训练真实机械臂能够完成推动和抓取动作，并且可以处理训练过程中未出现的物体。Yahya 等[124] 将引导策略搜索（guided policy search，GPS）[125] 方法扩展为分布式和异步版本的引导策略搜索（distributed asynchronous GPS，DAGPS），并指导 4 个机器人执行开门任务的策略学习。实验结果表明，利用 DAGPS 方法指导机器人成功地完成了协作任务，并且具有较好的泛化性。虽然在真实环境中训练的模型不需要迁移，但是需要强大的设备支持和长时间训练。

4.3　稀疏奖励和塑性奖励

在机器人操作任务中，奖励函数主要负责评价机器人行为的好坏。如果行为有助于实现目标任务，则给予正的奖励值；反之，给予负的奖励值。奖励函数一般包括 2 类：稀疏奖励和塑性奖励。稀疏奖励函数一般是二元函数，机器人完成目标才能获得相应的奖励值；塑性奖励函数一般是某些变量的函数，机器人行为的奖励值随着变量动态变化。2 类奖励函数各有其不足之处，就稀疏奖励函数而言，机器人在训练初期完成目标的次数太少或者完成目标的轨迹太长，导致奖励空间中负奖励样本数远远高于正奖励样本数，严重影响机器人学习；就塑性奖励函数而言，奖励函数将算法逐步引导至奖励函数增大的决策空间，但在复杂任务中函数定义复杂，难以实际应用。

4.3.1　稀疏奖励

在机器人的操作任务中，稀疏奖励函数只需要根据任务完成的标准定义二元奖励函数。本节涉及的研究工作均为机器人完成动作时才获得稀疏奖励，或者采用逐步稀疏奖励方式评价操作行为。

Zeng 等[73] 定义 UR5 机械臂成功推动方块得到 +1 奖励值，其他情况奖励值为 0。Berscheid 等[74] 利用 Franka 机械臂执行杂乱环境下的抓取任务，奖励函数定义为二元函数。这 2 项工作均实现了较高的抓取成功率，整体性能均符合设计要求。针对此类问题，奖励函数可以设计为机器人末端执行器和

方块之间的距离函数，即塑性奖励函数。但仅通过视觉很难获得实验环境中方块的随机位置，因此在目标位置的动态变化情况下，采取稀疏奖励函数定义任务成功与否更为合适。Riedmiller 等[126] 提出了调度辅助控制（scheduled auxiliary control，SACX）方法，SACX 方法能够在存在多个稀疏奖励信号的情况下，从头开始学习复杂行为，并且主动调度和辅助策略的执行，使智能体对环境进行充分探索。实验结果表明，SACX 方法生成的策略具有高度可靠性，其行为多样化且鲁棒性强。Andrychowicz 等[62] 提出 HER 机制，该机制不需要复杂的奖励函数过程设计，将智能体每个回合到达的目标状态存入经验池中，大大扩展了经验池中完成任务的经验数量，将稀疏问题转化成了非稀疏问题，并且实验效果优于仅使用稀疏奖励机制的方法。此外，许多工作已经成功解决了稀疏奖励函数存在的问题，并且成功完成了操作任务，比如堆叠 2 个方块[127] 以及目标定位[128-129] 等。对于堆叠任务，Li 等[130] 使用深度强化学习、关系图结构和课程学习的方法将多个方块堆叠成塔，并且通过逐步稀疏的奖励方式评价机器人的操作行为。该方法在数据利用率方面提高了几个数量级，并且效果超过了利用演示数据的相关方法。

4.3.2 塑性奖励

塑性奖励函数一般根据智能体当前状态与目标状态之间的关系构造变量函数，以使得智能体在任意状态都能获得奖励值。塑性奖励函数可基于目标属性或者任务状态（例如末端执行器与目标点的距离）进行构造，也可通过算法自主学习。

Liu 等[131] 提出一种基于对象配置匹配（objects configuration matching，OCM）的奖励函数设计方法，采用 PPO 算法的改进型算法学习和优化技能策略，奖励函数由对象目标配置和对象当前配置进行构造，实验结果表明，改进型 PPO 算法性能优于 PPO 算法和 A3C 算法。Zuo 等[132] 基于 DDPG 算法，实现了 OWI-535 机械臂的目标点定位。奖励函数为目标点与机械臂末端抓取器之间的距离函数。只要机械臂的动作是在靠近目标点，那么赋予该动作正的奖励；反之，动作得到负的奖励。该实验结果超过了人类亲自控制机械臂到达目标点的精度。

以上 2 种奖励函数的形式均为手工设计，对于一些复杂的任务，手工设计的奖励函数效果不太理想。针对此问题，Singh 等[133] 没有设计奖励函数，而是提供机器人成功完成任务的演示。机器人执行动作后，请求查询并根据得到的标签判断任务是否成功完成。Zhu 等[134] 基于 SAC 算法，设计了 VICE（variational inverse control with events）框架，算法训练了能够区分成功和失败的判别器，判别信号指导智能体学习。最近，直接通过人工监督（比如等级[135] 和行为偏好[136]）的方式学习奖励功能的研究取得了很大的进步，但存在奖励低估和高估的问题。针对此问题，Xu 等[137] 将奖励函数学习方法与正确－未标注（positive-unlabeled）学习方法相结合，解决了算法过拟合问题和奖励学习的欠拟合问题。Wu 等[138] 设计了一种从机器人高维观测数据中学习奖励函数的方法，以自我监督的方式估算任务进程奖励。与基准方法相比，该方法学习的操作策略具有更好的操作性能和更快的收敛性。

4.4 示范和次优示范

示范数据是专业人士成功演示任务完成的行为轨迹。在一定程度上，示范数据能够降低样本复杂度，提高机器人的行为效率。模仿学习（imitation learning，IL）和深度强化学习的融合方法利用示范数据提高了机器人的自学能力和学习速率。但是，次优示范数据对机器人学习行为策略有负面影响。因此，应用示范数据的同时要采取合适的措施优化可能存在的次优数据。

4.4.1 示范

在机器人训练的数据集中加入成功示范能够有效地降低样本的复杂度，且已经得到了有效证明。同样，模仿学习方法结合示范数据与深度强化学习，提高了机器人学习操作技能的能力和速率。

Rajeswaran 等[139] 基于示范强化策略梯度（demo augmented policy gradient，DAPG）算法，加入机械手成功开门示范数据，解决了从零开始训练的低效问题，相对于从零开始学习，速率提高了近 30 倍。Zhu 等[140] 同样基于 DAPG 算法，加入人类示范数据。机器人成功地完成了旋转阀门和翻转物体等操作，学习时间减少了一半。Gupta 等[88] 通过手持机械手完成期望动作的示范，然后利用 GPS 算法训练智能体执行操作任务，效果优于未加入示范的情况。Zhu 等[141] 使用 3D 动作控制器收集专家演示任务动作数据，同时训练模仿学习与 PPO 模型。该项工作采取端到端模型，直接以图像数据为输入，以机器人关节速度为输出，在抓取操作任务中（堆叠物体，倾倒液体等），其实验效果优于仅基于 PPO 训练的模型。Chen 等[142] 提出 BAIL（best-action imitation learning）方法学习策略函数，并选择高性能动作。模仿学习通过高性能动作训练策略网络，收敛速度和性能均得到提高。Gupta

等[143]提出解决多阶段、长时间任务的机器人操作任务策略学习方法，其中模仿学习阶段生成目标条件的分层策略，强化学习阶段微调任务执行策略。该策略的性能明显优于从零开始的分层强化学习模型和模仿学习模型，并且能够指导机器人完成复杂的操作任务。Berscheid 等[144]以自我监督的训练方式使机器人掌握基本操作动作（抓取和预抓取），并整合了模仿学习的方法，使用专家演示的目标状态定义对象位姿。在此融合算法的指导下，机器人能够完成极具挑战性的操作任务，比如从多个物体中选择目标物体，抓取木块搭建房子模型等。从示范数据中学习行为策略在一定程度上能够降低样本的复杂度并提高算法性能[145-147]。

4.4.2　次优示范

示范数据在一定程度上可以加快智能体的学习速度，但是非专业的示范不能保证数据准确性，且存在一定的瑕疵。尽管如此，许多研究工作证明了利用次优示范数据仍可提高机器人的学习速率。除此之外，本节介绍了收集示范数据的新颖方法。

Xiang 等[148]提出将任务奖励和目标导向奖励相结合的方法，并利用少量不完美的专家演示进行训练指导，形成的 AC 算法相对于部分基准深度强化学习算法具有更高的采样效率，大大降低了样本复杂度。更重要的是，该算法在稀疏奖励和延迟奖励任务中取得了实质性进展，极大地促进了机器人自动获取操作技能的研究进展。Mandlekar 等[149]提出了新型框架 IRIS（implicit reinforcement without interaction at scale），其中包括目标低层控制器和高层目标选择机制。目标低层控制器模仿了简短的演示序列，高层目标选择机制为目标低层控制器选择目标并有选择性地组合部分次优示范。实验表明，IRIS 框架可以从大型操作数据集（如 RoboTurk[150]）中恢复策略性能，并显著优于其他强化学习算法。Gao 等[151]提出了统一的强化学习规范算法，归一化演员－评论家（normalized actor-critic，NAC）。NAC 算法有效地规范化了动作价值函数，降低了示范数据中不存在的动作价值，对次优数据具有很强的鲁棒性。Brown 等[152]设计了一种从次优示范中学习奖励功能的 T-REX（trajectory-ranked reward extrapolation），并且效果优于专家示范。在次优示范问题上，许多工作通过演示学习[153]、逆强化学习[154-155]等方法进行研究，性能指标相对于无示范情况均有所提高。

事实证明，示范数据能在一定程度上降低样本的复杂度且加快实验进程，尤其是来自于相关专家的成功的任务示范。虽然非专业人士示范的次优数据能够采取优化方法进行学习利用，但技术要求仍较高。针对此问题，Kilinc 等[156]提出示范不再需要专家而是通过强化学习自动产生的思想。首先从操作任务中解耦出一个运动任务，然后通过虚拟环境模拟器进行学习，得到的目标轨迹被称为模拟运动演示（simulated locomotion demonstration，SLD）。基于 DDPG 算法，SLD 作为行为策略的辅助奖励，机器人在堆叠对象和非刚体操作任务中达到了 100% 的完成率，其他基准算法远远不能达到。SLD 方法很好地解决了示范过程中动作不规范、机器损耗等问题。

目前，示范数据收集工作仅限于自上而下的运动和开环执行，存在效率低和有效数据占比低的问题。针对此问题，Song 等[157]提出了一种低成本的硬件接口，用于收集各种环境中的示范数据。实践表明，该数据成功训练了基于深度强化学习的闭环抓取模型，并且此模型成功转移到了真实机器人。Young 等[158]针对动觉教学和遥操作方式收集专家数据所存在的局限性，设计了简化数据收集的模仿界面，并且允许数据传输到机器人。该方法在未知的物体推动和堆叠实验环境中分别达到了 87% 和 62% 的成功率，为机器人操作领域的演示数据收集工作提供了参考。

4.5　元强化学习

基于深度强化学习的机器人操作技能模型在面对新环境和新任务时，适应性较差，往往需要大量的微调工作，甚至需要被重新训练。元强化学习的本质是利用少量样本数据微调模型以使行为策略匹配新环境和新任务。其原理是将基础强化学习方法中手动设定的超参数设定为元参数，然后学习和调整元参数以指导底层强化学习。元强化学习的理念来源于元学习，元学习可以使机器人学习评价自己的行为，并且可以根据经验和少量样本使机器人快速适应新任务或新环境。元强化学习分为基于模型的元强化学习和基于优化的元强化学习。基于模型的元强化学习与强化学习不同之处在于，除当前时刻状态外，前一时刻的奖励和行为也被行为策略考虑在内，以便学习状态、奖励和行为之间的动态关系，及时调整行为策略。该类算法包括元学习 RNN 权重[159-160]，元学习神经目标函数并指导梯度更新[161]，元学习基于环境和智能体参数分布的最优更新规则[162]等。基于优化的元强化学习主要是学习更新模型参数的方法，以便在新任务上实现更好的泛化性能，其中包括元学习概率编码器[163]、

元学习参数化损失函数[164]、元 Q 学习[165] 等。

Goyal 等[166] 提出 PixL2R（pixels and language to reward）模型，直接将像素映射到奖励，使机器人学习操作行为策略。实验表明，PixL2R 在稀疏奖励和密集奖励设置中，显著提高了策略学习的样本利用率。Wu 等[167] 设计了采样效率高的元强化学习算法，该算法只根据一个视频演示就可以执行新任务。首先通过行为克隆学习引导智能体学习任务编码器和条件策略。然后结合真实机器人经验和演示数据学习 Q 函数，进而评估机器人采样数据完成策略学习。在机器人的抓取和放置等操作任务中，该算法性能远超行为克隆和强化学习算法。此外，元逆强化学习算法在机器人操作领域的成功应用拓展了机器人操作的研究方向[168-169]。

元强化学习的研究促进了机器人操作领域的发展，最近 Yu 等[170] 提出了包含 50 个不同的机器人操作任务的用于元强化学习和多任务学习的开源模拟基准，为机器人的训练提供了支撑，且证明了元强化学习在机器人操作领域的重要作用。

对机器人操作任务而言，算法设计固然重要，操作策略对末端执行器的适应性问题也不可忽视。大部分机器人操作研究仅限于单个末端执行器装置，机器人需要进行大量训练才能将策略推广到新的末端执行器。为此，Xu 等[171] 设计了 AdaGrasp，学习单一抓取策略方法以适应不同配置的手爪。在训练过程中，AdaGrasp 学习使用大量手爪，获得在不同抓取任务中如何使用不同手爪的通用知识。实验表明 AdaGrasp 提高了系统的多功能性和适应性。该项技术的实现可以帮助人们轻松使用多种工具完成不同任务，并快速适应未知的抓取工具。

5 挑战和未来展望（Challenges and future prospects）

基于深度强化学习的机器人操作研究，其发展动力主要源自于如何将深度强化学习的能力最大化。机器人只有具备了自主思考的能力，才能像人类一样在面对未知环境时探求最优的行为方式。虽然基于深度强化学习的算法已经解决了多种多样的机器人操作任务，并使机器人在执行操作任务时具备自主学习的能力，但是仍然存在很多技术挑战。几乎没有机器人操作问题可以被严格地定义为马尔可夫决策过程，而是表现为部分可观性和非平稳性，这是实验效果并不如预期的原因之一。对于多步骤任务而言，它涉及到很多动作和行为，不可能对每个动作或者行为都设置奖励函数。因此，机器人为了达到目标状态会表现出大量不合理的动作。此外，如何平衡“探索”和“利用”一直是机器人操作领域难以解决的问题，许多工作通过机器人在实验过程中的表现，人为限定“探索”和“利用”的界限，但是机器人难以自主学习如何去平衡“探索”和“利用”。基于深度强化学习的机器人操作研究仍然在如下几个方面存在挑战：

1) 非刚性物体难以被精确建模。机器人在重复性高、环境设定简单和操作不确定度小等场景下表现较好，但所操作的对象大部分为刚性物体。而非刚性物体（如毛巾、衣服等）的姿态和形状是不断变化的，难以构造其精确的物理模型，因此机器人在操作这类物体时难以选择抓取点，或者抓取点被遮盖住，直接导致机器人操作任务失败。针对此挑战，设计多指灵巧机械手和开发非刚性物体的开源模拟平台可能成为机器人操作柔性物体的有效解决方法。具体而言，多指灵巧机械手的手指可以相互配合，类似于人类手指，并且每个手指可以单独动作。对于柔性物体的姿态和形状发生变化的情况，开发强大的模拟平台实时对物体形状进行建模以确定理想的操作位置是有效解决途径。

2) 模型难以从模拟环境迁移到真实环境。大部分机器人操作研究首先在模拟场景中训练行为模型，然后将其迁移到真实机器人测试。虽然模拟场景有训练速度快、采样效率高和不损耗机器人设备等优点，但是由于模拟场景和真实场景存在物理约束和环境差异等领域差异，模型迁移后需要进行大量微调工作才能工作，往往效果不佳。因此，根据实验需求，将二者的领域差异因素加入到模拟环境中，可使行为策略具备更强的鲁棒性。在算法层面上，开发高级深度强化学习算法并从动力学角度加入真实场景中的摩擦力、光照、噪声等干扰因素，也可以成为一种有效减小领域差异的途径。

3) 不同环境不同任务间模型可移植性差。在某个特定环境和特定任务中训练的行为策略移植到新环境中往往需要微调以适应环境；移植到新任务中往往不具备直接完成任务的能力，甚至需要重新开始训练。针对此挑战，元强化学习可成为有效解决方法。元强化学习可根据经验和少量样本数据，改进行为策略适应新任务、新环境。此外，多任务学习可以通过多个任务间共享结构实现不同任务间的模型移植[172]。

4) 人为设计的奖励函数难以评价多步操作任务。大部分机器人操作任务采用稀疏奖励评价机器人是否完成任务，这种方式简单直接但是难以用于

多步骤操作任务。大部分塑性奖励函数根据机器人末端执行器与目标点的距离判断行为的好坏，这种方式适用于目标点定位任务，但对于复杂的操作行为往往不适用。让机器人学习如何评价自身行为可以有效解决人为设计奖励函数的问题。逆强化学习可以指导机器人根据已知的行为策略学习奖励函数，一定程度上可为操作算法提供思路。此外，还可根据任务需求以自我监督的方式估算任务奖励或者人为辅助奖励。

5) 机器人采样效率低、学习速率慢。该问题是深度强化学习的共性问题之一。深度强化学习的本质是指导机器人自主探索、学习行为策略，这导致机器人面临巨大的探索空间，学习效率低。针对该问题，可利用已有的操作知识引导机器人进行有效探索，缩小探索空间。专家示范数据可丰富机器人采样数据集，引导机器人向靠近目标的方向探索。此外，模仿学习与强化学习的结合可以提高机器人学习速率。但是，专家示范方法存在对示范者要求高和成本大的问题。因此，可通过深度强化学习指导机器人自动生成行为示例，以此作为示范数据，丰富机器人的数据集[156]。

基于深度强化学习的方法不仅使机器人具备了自主学习操作技能的能力，而且为机器人真正实现智能化提供了技术支持，但仍需长期关注和研究它所面临的挑战。随着人机共融理念的流行，机器人操作领域的研究也需要考虑环境中人的影响，即如何实现人机协作。实现灵活复杂的人机协作任务对机器人硬件和软件都有较高要求。任务执行中需实时采集、分析交互者的行为数据，进而反馈并指导机器人的行为。因此，机器人还需具备推理和预测能力，以配合人的操作行为，共同完成任务。

6　结论（Conclusion）

对深度强化学习算法的原理及其在机器人操作领域的应用现状进行了详细的论述。基于深度强化学习的机器人技术打破了传统方法中复杂编程及示教编程的壁垒，并赋予了机器人自主学习操作技能的能力。在基于深度强化学习算法的机器人操作任务中，深度强化学习算法训练的行为策略可指导机器人探索行为空间，使机器人面对未知环境时具备一定的决策能力。在此算法下，机器人面对未知环境时能自动调整行为策略以适应不同的操作任务。目前，深度强化学习已经在机器人操作领域取得了显著的成功，但仍面临非刚性物体建模困难、模型迁移性差、数据效率低等挑战。总之，深度强化学习的创新与进步促进了机器人操作行为的研究进程，为机器人真正实现智能化提供了技术保证。

参考文献（References）

[1] 刘乃军，鲁涛，蔡莹皓，等. 机器人操作技能学习方法综述[J]. 自动化学报，2019，45(3)：458-470.
Liu N J, Lu T, Cai Y H, et al. A review of robot manipulation skills learning methods[J]. Acta Automatica Sinica, 2019, 45(3): 458-470.

[2] 倪自强，王田苗，刘达. 基于视觉引导的工业机器人示教编程系统[J]. 北京航空航天大学学报，2016，42(3)：562-568.
Ni Z Q, Wang T M, Liu D. Vision guide based teaching programming for industrial robot[J]. Journal of Beijing University of Aeronautics and Astronautics, 2016, 42(3): 562-568.

[3] Rozo L, Jaquier N, Calinon S, et al. Learning manipulability ellipsoids for task compatibility in robot manipulation [C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 3183-3189.

[4] Broquere X, Sidobre D, Nguyen K. From motion planning to trajectory control with bounded jerk for service manipulator robots[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2010: 4505-4510.

[5] Richter M, Sandamirskaya Y, Schöner G. A robotic architecture for action selection and behavioral organization inspired by human cognition[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2012: 2457-2464.

[6] Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[7] Silver D, Hubert T, Schrittwieser J, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play[J]. Science, 2018, 362(6419): 1140-1144.

[8] Zhou L, Pan S, Wang J, et al. Machine learning on big data: Opportunities and challenges[J]. Neurocomputing, 2017, 237: 350-361.

[9] 秦方博，徐德. 机器人操作技能模型综述[J]. 自动化学报，2019，45(8)：1401-1418.
Qin F B, Xu D. Review of robot manipulation skill models[J]. Acta Automatica Sinica, 2019, 45(8): 1401-1418.

[10] Haarnoja T, Pong V, Zhou A, et al. Composable deep reinforcement learning for robotic manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 6244-6251.

[11] Peters J, Vijayakumar S, Schaal S. Reinforcement learning for humanoid robotics[C]//IEEE-RAS International Conference on Humanoid Robots. Piscataway, USA: IEEE, 2003: 1-20.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2012: 1097-1105.

[13] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[DB/OL]. (2014-04-10) [2021-04-02]. https://arxiv.org/abs/1409.1556.

[14] Lin M, Chen Q, Yan S. Network in network[DB/OL]. (2014-03-04) [2021-04-02]. https://arxiv.org/abs/1312.4400.

[15] Nair A, Bahl S, Khazatsky A, et al. Contextual imagined goals for self-supervised robotic learning[C]//Conference on Robot Learning. Cambridge, USA: JMLR, 2020: 530-539.

[16] 杜学丹，蔡莹皓，鲁涛，等．一种基于深度学习的机械臂抓取方法[J]．机器人，2017，39(6)：820-828,837.
Du X D, Cai Y H, Lu T, et al. A robotic grasping method based on deep learning[J]. Robot, 2017, 39(6): 820-828,837.

[17] 伍锡如，黄国明，孙立宁．基于深度学习的工业分拣机器人快速视觉识别与定位算法[J]．机器人，2016，38(6)：711-719.
Wu X R, Huang G M, Sun L N. Fast visual identification and location algorithm for industrial sorting robots based on deep learning[J]. Robot, 2016, 38(6): 711-719.

[18] Agrawal P, Nair A V, Abbeel P, et al. Learning to poke by poking: Experiential learning of intuitive physics[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2016: 5074-5082.

[19] Krainin M, Curless B, Fox D. Autonomous generation of complete 3D object models using next best view manipulation planning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2011: 5031-5037.

[20] Schenck C, Fox D. Visual closed-loop control for pouring liquids[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 2629-2636.

[21] Sutton R S, Barto A G. Reinforcement learning: An introduction[M]. Cambridge, USA: MIT Press, 2018.

[22] Kaelbling L P, Littman M L, Moore A W. Reinforcement learning: A survey[J]. Journal of Artificial Intelligence Research, 1996, 4: 237-285.

[23] Gu S, Lillicrap T, Sutskever I, et al. Continuous deep Q-learning with model-based acceleration[J]. Proceedings of Machine Learning Research, 2016, 48: 2829-2838.

[24] 万里鹏，兰旭光，张翰博，等．深度强化学习理论及其应用综述[J]．模式识别与人工智能，2019，32(1)：67-81.
Wan L P, Lan X G, Zhang H B, et al. A review of deep reinforcement learning theory and application[J]. Pattern Recognition and Artificial Intelligence, 2019, 32(1): 67-81.

[25] 刘全，翟建伟，章宗长，等．深度强化学习综述[J]．计算机学报，2018，41(1)：1-27.
Liu Q, Zhai J W, Zhang Z C, et al. A survey on deep reinforcement learning[J]. Chinese Journal of Computers, 2018, 41(1): 1-27.

[26] Bottou L, Chapelle O, DeCoste D, et al. Large-scale Kernel machines[M]. Cambridge, USA: MIT Press, 2007: 321-359.

[27] 赵冬斌，邵坤，朱圆恒，等．深度强化学习综述：兼论计算机围棋的发展[J]．控制理论与应用，2016，33(6)：701-717.
Zhao D B, Shao K, Zhu Y H, et al. Review of deep reinforcement learning and discussions on the development of computer Go[J]. Control Theory & Applications, 2016, 33(6): 701-717.

[28] Silver D, Huang A, Maddison C J, et al. Mastering the game of Go with deep neural networks and tree search[J]. Nature, 2016, 529(7587): 484-489.

[29] Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge[J]. Nature, 2017, 550(7676): 354-359.

[30] Silver D, Hubert T, Schrittwieser J, et al. Mastering chess and shogi by self-play with a general reinforcement learning algorithm[DB/OL]. (2017-12-05) [2021-04-02]. https://arxiv.org/abs/1712.01815.

[31] 陈兴国，俞扬．强化学习及其在电脑围棋中的应用[J]．自动化学报，2016，42(5)：685-695.
Chen X G, Yu Y. Reinforcement learning and its application to the game of Go[J]. Acta Automatica Sinica, 2016, 42(5): 685-695.

[32] Dosovitskiy A, Koltun V. Learning to act by predicting the future[DB/OL]. (2017-02-14) [2021-04-02]. https://arxiv.org/abs/1611.01779.

[33] Ha D, Schmidhuber J. World models[DB/OL]. (2017-05-09) [2021-04-02]. https://arxiv.org/abs/1803.10122.

[34] Oh J, Guo X, Lee H, et al. Action-conditional video prediction using deep networks in Atari games[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2015: 2863-2871.

[35] Oh J, Chockalingam V, Singh S, et al. Control of memory, active perception, and action in minecraft[DB/OL]. (2016-05-30) [2021-04-02]. https://arxiv.org/abs/1605.09128.

[36] Lample G, Chaplot D S. Playing FPS games with deep reinforcement learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2017, 31(1): 2140-2146.

[37] Kempka M, Wydmuch M, Runc G, et al. ViZDoom: A Doom-based AI research platform for visual reinforcement learning [C]//IEEE Conference on Computational Intelligence and Games. Piscataway, USA: IEEE, 2016: 1-8.

[38] Vinyals O, Ewalds T, Bartunov S, et al. StarCraft II: A new challenge for reinforcement learning[DB/OL]. (2017-08-16) [2021-04-02]. https://arxiv.org/abs/1708.04782.

[39] 孙长银，穆朝絮．多智能体深度强化学习的若干关键科学问题[J]．自动化学报，2020，46(7)：1301-1312.
Sun C Y, Mu C X. Important scientific problems of multi-agent deep reinforcement learning[J]. Acta Automatica Sinica, 2020, 46(7): 1301-1312.

[40] Lillicrap T P, Hunt J J, Pritzel A, et al. Continuous control with deep reinforcement learning[DB/OL]. (2017-08-16) [2021-04-02]. https://arxiv.org/abs/1509.02971.

[41] Heess N, TB D, Sriram S, et al. Emergence of locomotion behaviours in rich environments[DB/OL]. (2019-07-05) [2021-04-02]. https://arxiv.org/abs/1707.02286.

[42] Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[DB/OL]. (2017-08-28) [2021-04-02]. https://arxiv.org/abs/1707.06347.

[43] Al-Shedivat M, Bansal T, Burda Y, et al. Continuous adaptation via meta-learning in nonstationary and competitive environments[DB/OL]. (2018-02-23) [2021-04-02]. https://arxiv.org/abs/1710.03641.

[44] Levine S, Abbeel P. Learning neural network policies with guided policy search under unknown dynamics[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2014: 1071-1079.

[45] Levine S, Finn C, Darrell T, et al. End-to-end training of deep visuomotor policies[J]. The Journal of Machine Learning Research, 2016, 17(1): 1334-1373.

[46] Mnih V, Kavukcuoglu K, Silver D, et al. Playing Atari with deep reinforcement learning[DB/OL]. (2013-12-19) [2021-04-02]. https://arxiv.org/abs/1312.5602.

[47] van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double Q-learning[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2016, 30(1): 2094-2100.

[48] Wang Z, Schaul T, Hessel M, et al. Dueling network architectures for deep reinforcement learning[J]. Proceedings of Machine Learning Research, 2016, 48: 1995-2003.

[49] Bellemare M G, Dabney W, Munos R. A distributional perspective on reinforcement learning[J]. Proceedings of Machine Learning Research, 2017, 70: 449-458.

[50] Dabney W, Rowland M, Bellemare M G, et al. Distributional reinforcement learning with quantile regression[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2018, 32(1): 2892-2901.

[51] Bellman R E. Adaptive control processes: A guided tour[M]. New Jersey, USA: Princeton University Press, 2015.

[52] Mnih V, Badia A P, Mirza M, et al. Asynchronous methods for deep reinforcement learning[J]. Proceedings of Machine Learning Research, 2016, 48: 1928-1937.

[53] TORCS. TORCS: The open racing car simulator[EB/OL]. (2020-02-06) [2021-04-02]. http://torcs.sourceforge.net.

[54] Todorov E, Erez T, Tassa Y. MuJoCo: A physics engine for model-based control[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2012: 5026-5033.

[55] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[J]. (2018-08-08)[2021-04-02]. https://arxiv.org/abs/1801.01290.

[56] Brockman G, Cheung V, Pettersson L, et al. OpenAI Gym[J]. (2016-06-05) [2021-04-02]. https://arxiv.org/abs/1606.01540.

[57] Fujimoto S, van Hoof H, Meger D. Addressing function approximation error in actor-critic methods[DB/OL]. (2018-10-22) [2021-04-02]. https://arxiv.org/abs/1802.09477.

[58] Kalweit G, Boedecker J. Uncertainty-driven imagination for continuous deep reinforcement learning[J]. Proceedings of Machine Learning Research, 2017, 78: 195-206.

[59] Schulman J, Levine S, Abbeel P, et al. Trust region policy optimization[J]. Proceedings of Machine Learning Research, 2015, 37: 1889-1897.

[60] Kurutach T, Clavera I, Duan Y, et al. Model-ensemble trust-region policy optimization[DB/OL]. (2018-10-05) [2021-04-02]. https://arxiv.org/abs/1802.10592.

[61] Luo Y, Xu H, Li Y, et al. Algorithmic framework for model-based deep reinforcement learning with theoretical guarantees[DB/OL]. (2021-02-15) [2021-04-02]. https://arxiv.org/abs/1807.03858.

[62] Andrychowicz M, Wolski F, Ray A, et al. Hindsight experience replay[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2017: 5048-5058.

[63] Weber T, Racanière S, Reichert D P, et al. Imagination-augmented agents for deep reinforcement learning[DB/OL]. (2018-02-14) [2021-04-02]. https://arxiv.org/abs/1707.06203.

[64] Nagabandi A, Kahn G, Fearing R S, et al. Neural network dynamics for model-based deep reinforcement learning with model-freefine-tuning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 7559-7566.

[65] Feinberg V, Wan A, Stoica I, et al. Model-based value estimation for efficient model-free reinforcement learning[DB/OL]. (2018-02-28) [2021-04-02]. https://arxiv.org/abs/1803.00101.

[66] Buckman J, Hafner D, Tucker G, et al. Sample-efficient reinforcement learning with stochastic ensemble value expansion[DB/OL]. (2019-06-07) [2021-04-02]. https://arxiv.org/abs/1807.01675.

[67] Kaiser L, Babaeizadeh M, Milos P, et al. Model-based reinforcement learning for Atari[DB/OL]. (2020-02-19) [2021-04-02]. https://arxiv.org/abs/1903.00374.

[68] Silver D, Lever G, Heess N, et al. Deterministic policy gradient algorithms[J]. Proceedings of Machine Learning Research, 2014, 32(1): 387-395.

[69] Sutton R S, McAllester D A, Singh S P, et al. Policy gradient methods for reinforcement learning with function approximation[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2000: 1057-1063.

[70] 多南讯，吕强，林辉灿，等. 迈进高维连续空间：深度强化学习在机器人领域中的应用[J]. 机器人，2019，41(2)：276-288.

Duo N X, Lü Q, Lin H C, et al. Step into high-dimensional and continuous action space: A survey on applications of deep reinforcement learning to robotics[J]. Robot, 2019, 41(2): 276-288.

[71] Finn C, Tan X Y, Duan Y, et al. Deep spatial autoencoders for visuomotor learning[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2016: 512-519.

[72] Kalashnikov D, Irpan A, Pastor P, et al. Scalable deep reinforcement learning for vision-based robotic manipulation[C]//Conference on Robot Learning. Cambridge, USA: JMLR, 2018: 651-673.

[73] Zeng A, Song S, Welker S, et al. Learning synergies between pushing and grasping with self-supervised deep reinforcement learning[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 4238-4245.

[74] Berscheid L, Meißner P, Kröger T. Robot learning of shifting objects for grasping in cluttered environments[DB/OL]. (2019-07-25) [2021-04-02]. https://arxiv.org/abs/1907.11035.

[75] Yu K T, Rodriguez A. Realtime state estimation with tactile and visual sensing for inserting a suction-held object[C]//IEEE/RSJ Internaitonal Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 1628-1635.

[76] Viereck U, Pas A, Saenko K, et al. Learning a visuomotor controller for real world robotic grasping using simulated depth images[J]. (2017-11-17) [2021-04-02]. https://arxiv.org/ abs/1706.04652.

[77] Calandra R, Owens A, Jayaraman D, et al. More than a feeling: Learning to grasp and regrasp using vision and touch[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3300-3307.

[78] Zhang F Y, Leitner J, Milford M, et al. Towards vision-based deep reinforcement learning for robotic motion control [C/OL] //Australasian Conference on Robotics and Automation. 2015. [2021-04-02]. https://www.araa.asn.au/acra/acra2015/papers/pap168.pdf.

[79] Finn C, Tan X Y, Duan Y, et al. Learning visual feature spaces for robotic manipulation with deep spatial autoencoders [DB/OL]. (2016-03-01) [2021-04-02]. https://arxiv.org/abs/1509.06113.

[80] Mahler J, Matl M, Satish V, et al. Learning ambidextrous robot grasping policies[J]. Science Robotics, 2019, 4(26). DOI: 10.1126/scirobotics.aau4984.

[81] Jiang R, Wang Z, He B, et al. Vision-based deep reinforcement learning for UR5 robot motion control[C]//IEEE International Conference on Consumer Electronics and Computer Engineering. Piscataway, USA: IEEE, 2021: 246-250.

[82] Popov I, Heess N, Lillicrap T, et al. Data-efficient deep reinforcement learning for dexterous manipulation[DB/OL]. (2017-04-10) [2021-04-02]. https://arxiv.org/abs/1704.03073.

[83] Jang E, Devin C, Vanhoucke V, et al. Grasp2Vec: Learning object representations from self-supervised grasping[DB/ OL]. (2018-11-19) [2021-04-02]. https://arxiv.org/abs/1811. 06964.

[84] Li Z J, Zhao T, Chen F, et al. Reinforcement learning of manipulation and grasping using dynamical movement primitives for a humanoidlike mobile manipulator[J]. IEEE/ASME Transactions on Mechatronics, 2017, 23(1): 121-131.

[85] Fang K, Zhu Y K, Garg A, et al. Learning task-oriented grasping for tool manipulation from simulated self-supervision[J]. International Journal of Robotics Research, 2020, 39(2/3): 202-216.

[86] Andrychowicz O M, Baker B, Chociej M, et al. Learning dexterous in-hand manipulation[J]. International Journal of Robotics Research, 2020, 39(1): 3-20.

[87] Akkaya I, Andrychowicz M, Chociej M, et al. Solving rubik's cube with a robot hand[DB/OL]. (2019-10-16) [2021-04-02]. https://arxiv.org/abs/1910.07113.

[88] Gupta A, Eppner C, Levine S, et al. Learning dexterous manipulation for a soft robotic hand from human demonstrations[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2016: 3786-3793.

[89] Nagabandi A, Konoglie K, Levine S, et al. Deep dynamics models for learning dexterous manipulation[C]//Conference on Robot Learning. Cambridge, USA: JMLR, 2020: 1101-1112.

[90] Yahya A, Li A, Kalakrishnan M, et al. Collective robot reinforcement learning with distributed asynchronous guided policy search[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 79-86.

[91] Chebotar Y, Kalakrishnan M, Yahya A, et al. Path integral guided policy search[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 3381-3388.

[92] Urakami Y, Hodgkinson A, Carlin C, et al. DoorGym: A scalable door opening environment and baseline agent[DB/OL]. (2020-05-13) [2021-04-02]. https://arxiv.org/abs/1908.01887.

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

[94] Zakka K, Zeng A, Lee J, et al. Form2Fit: Learning shape priors for generalizable assembly from disassembly[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 9404-9410.

[95] Li F M, Jiang Q, Zhang S S, et al. Robot skill acquisition in assembly process using deep reinforcement learning[J]. Neurocomputing, 2019, 345: 92-102.

[96] Khansari M, Kappler D, Luo J L, et al. Action image representation: Learning scalable deep grasping policies with zero real world data[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 3597-3603.

[97] Murali A, Mousavian A, Eppner C, et al. 6-DOF grasping for target-driven object manipulation in clutter[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 6232-6238.

[98] Matas J, James S, Davison A J. Sim-to-real reinforcement learning for deformable object manipulation[DB/OL]. (2018-10-08) [2021-04-02]. https://arxiv.org/abs/1806.07851.

[99] Li Y X, Yue Y H, Xu D F, et al. Folding deformable objects using predictive simulation and trajectory optimization[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2015: 6000-6006.

[100] Cusumano-Towner M, Singh A, Miller S, et al. Bringing clothing into desired configurations with limited perception [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2011: 3893-3900.

[101] Yamakawa Y, Namiki A, Ishikawa M. Motion planning for dynamic folding of a cloth with two high-speed robot hands and two high-speed sliders[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2011: 5486-5491.

[102] Maitin-Shepard J, Cusumano-Towner M, Lei J N, et al. Cloth grasp point detection based on multiple-view geometric cues with application to robotic towel folding[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2010: 2308-2315.

[103] Osawa F, Seki H, Kamiya Y. Unfolding of massive laundry and classification types by dual manipulator[J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2007, 11(5): 457-463.

[104] Bersch C, Pitzer B, Kammel S. Bimanual robotic cloth manipulation for laundry folding[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2011: 1413-1419.

[105] Tsurumine Y, Cui Y, Uchibe E, et al. Deep reinforcement learning with smooth policy update: Application to robotic cloth manipulation[J]. Robotics and Autonomous Systems, 2019, 112: 72-83.

[106] Wu Y L, Yan W, Kurutach T, et al. Learning to manipulate deformable objects without demonstrations[C]//Robotics: Science and Systems. Cambridge, USA: MIT Press, 2020. DOI: 10.15607/RSS.2020.XVI.065.

[107] Seita D, Florence P, Tompson J, et al. Learning to Rearrange deformable cables, fabrics, and bags with goal-conditioned transporter networks[DB/OL]. (2021-03-26) [2021-04-02]. https://arxiv.org/abs/2012.03385.

[108] Lin X, Wang Y, Olkin J, et al. SoftGym: Benchmarking deep reinforcement learning for deformable object manipulation[DB/OL]. (2021-03-08) [2021-04-02]. https://arxiv.org/abs/2011.07215.

[109] Sajjan S, Moore M, Pan M, et al. Clear grasp: 3D shape estimation of transparent objects for manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 3634-3642.

[110] Pan S J, Yang Q. A survey on transfer learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2009, 22(10): 1345-1359.

[111] Lin Y C, Zeng A, Song S R, et al. Learning to see before learning to act: Visual pre-training for manipulation[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 7286-7293.

[112] Rusu A A, Večerík M, Rothörl T, et al. Sim-to-real robot learning from pixels with progressive nets[J]. Proceedings of Machine Learning Research, 2017, 78: 262-270.

[113] Rusu A A, Rabinowitz N C, Desjardins G, et al. Progressive neural networks[DB/OL]. (2016-09-07) [2021-04-02]. https://arxiv.org/abs/1606.04671.

[114] Gupta A, Devin C, Liu Y X, et al. Learning invariant feature spaces to transfer skills with reinforcement learning[DB/OL]. (2017-03-08) [2021-04-02]. https://arxiv.org/abs/1703.02949.

[115] Wang M, Deng W H. Deep visual domain adaptation: A survey[J]. Neurocomputing, 2018, 312: 135-153.

[116] Peng X B, Andrychowicz M, Zaremba W, et al. Sim-to-real transfer of robotic control with dynamics randomization [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 3803-3810.

[117] Tobin J, Biewald L, Duan R, et al. Domain randomization and generative models for robotic grasping[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 3482-3489.

[118] James S, Wohlhart P, Kalakrishnan M, et al. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 12627-12637.

[119] Hundt A, Killeen B, Greene N, et al. “Good robot!”: Efficient reinforcement learning for multi-step visual tasks with sim to real transfer[J]. IEEE Robotics and Automation Letters, 2020, 5(4): 6724-6731.

[120] Pedersen O M, Misimi E, Chaumette F. Grasping unknown objects by coupling deep reinforcement learning, generative adversarial networks, and visual servoing[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 5655-5662.

[121] Zhu J Y, Park T, Isola P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//IEEE International Conference on Computer Vision. Piscataway, USA: IEEE, 2017: 2223-2232.

[122] Mahmood A R, Korenkevych D, Komer B J, et al. Setting up a reinforcement learning task with a real-world robot[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2018: 4635-4640.

[123] Finn C, Levine S. Deep visual foresight for planning robot motion[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2017: 2786-2793.

[124] Yahya A, Li A, Kalakrishnan M, et al. Collective robot reinforcement learning with distributed asynchronous guided policy search[C]//IEEE/RSJ International Conference on Intelligent Robots and Systems. Piscataway, USA: IEEE, 2017: 79-86.

[125] Levine S, Koltun V. Guided policy search[J]. Proceedings of Machine Learning Research, 2013, 28(3): 1-9.

[126] Riedmiller M, Hafner R, Lampe T, et al. Learning by playing-solving sparse reward tasks from scratch[DB/OL]. (2018-02-28) [2021-04-02]. https://arxiv.org/abs/1802.10567.

[127] Zhang M, Vikram S, Smith L, et al. Solar: Deep structured representations for model-based reinforcement learning [J]. Proceedings of Machine Learning Research, 2019, 97: 7444-7453.

[128] Pathak D, Agrawal P, Efros A A, et al. Curiosity-driven exploration by self-supervised prediction[C]//IEEE Conference on Computer Vision and Pattern Recognition Workshops. Piscataway, USA: IEEE, 2017: 16-17.

[129] Savinov N, Raichuk A, Marinier R, et al. Episodic curiosity through reachability[DB/OL]. (2019-08-06) [2021-04-02]. https://arxiv.org/abs/1810.02274.

[130] Li R, Jabri A, Darrell T, et al. Towards practical multi-object manipulation using relational reinforcement learning [C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 4051-4058.

[131] Liu D, Wang Z, Lu B, et al. A reinforcement learning-based framework for robot manipulation skill acquisition[J]. IEEE Access, 2020, 8: 108429-108437.

[132] Zuo Y M, Qiu W C, Xie L X, et al. CRAVES: Controlling robotic arm with a vision-based economic system[C]//IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, USA: IEEE, 2019: 4214-4223.

[133] Singh A, Yang L, Finn C, et al. End-to-end robotic reinforcement learning without reward engineering[C]//Robotics: Science and Systems. Cambridge, USA: MIT Press, 2019. DOI: 10.15607/RSS.2019.XV.073.

[134] Zhu H, Yu J, Gupta A, et al. The ingredients of real-world robotic reinforcement learning[DB/OL]. (2020-04-27) [2021-04-02]. https://arxiv.org/abs/2004.12570.

[135] Cabi S, Colmenarejo S G, Novikov A, et al. Scaling data-driven robotics with reward sketching and batch reinforcement learning[C]//Robotics: Science and Systems. Cambridge, USA: MIT Press, 2020. DOI: 10.15607/RSS.2020.XVI.076.

[136] Ibarz B, Leike J, Pohlen T, et al. Reward learning from human preferences and demonstrations in Atari[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2018: 8011-8023.

[137] Xu D, Denil M. Positive-unlabeled reward learning[DB/OL]. (2019-11-01) [2021-04-02]. https://arxiv.org/abs/1911.00459.

[138] Wu Z, Lian W Z, Unhelkar V, et al. Learning dense rewards for contact-rich manipulation tasks[DB/OL]. (2020-11-17) [2021-04-02]. https://arxiv.org/abs/2011.08458.

[139] Rajeswaran A, Kumar V, Gupta A, et al. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations[C]//Robotics: Science and Systems. Cambridge, USA: MIT Press, 2018. DOI: 10.15607/RSS.2018.XIV.049.

[140] Zhu H, Gupta A, Rajeswaran A, et al. Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost[C]//International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2019: 3651-3657.

[141] Zhu Y, Wang Z Y, Merel J, et al. Reinforcement and imitation learning for diverse visuomotor skills[C]//Robotics: Science and Systems. Cambridge, USA: MIT Press, 2018. DOI: 10.15607/RSS.2018.XIV.009.

[142] Chen X Y, Zhou Z J, Wang Z, et al. BAIL: Best-action imitation learning for batch deep reinforcement learning[DB/OL]. (2020-11-02) [2021-04-02]. https://arxiv.org/abs/1910.12179.

[143] Gupta A, Kumar V, Lynch C, et al. Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning[DB/OL]. (2019-10-25) [2021-04-02]. https://arxiv.org/abs/1910.11956.

[144] Berscheid L, Meißner P, Kröger T. Self-supervised learning for precise pick-and-place without object model[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4828-4835.

[145] Vecerik M, Hester T, Scholz J, et al. Leveraging demonstrations for deep reinforcement learning on robotics problems with sparse rewards[DB/OL]. (2018-10-08) [2021-04-02]. https://arxiv.org/abs/1707.08817.

[146] Chen S A, Tangkaratt V, Lin H T, et al. Active deep Q-learning with demonstration[J]. Machine Learning, 2020, 109: 1699-1725.

[147] Nair A, McGrew B, Andrychowicz M, et al. Overcoming exploration in reinforcement learning with demonstrations[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2018: 6292-6299.

[148] Xiang G F, Su J B. Task-oriented deep reinforcement learning for robotic skill acquisition and control[J]. IEEE Transactions on Cybernetics, 2021, 51(2): 1056-1069.

[149] Mandlekar A, Ramos F, Boots B, et al. IRIS: Implicit reinforcement without interaction at scale for learning control from offline robot manipulation data[C]//IEEE International Conference on Robotics and Automation. Piscataway, USA: IEEE, 2020: 4414-4420.

[150] Mandlekar A, Zhu Y, Garg A, et al. Roboturk: A crowdsourcing platform for robotic skill learning through imitation [DB/OL]. (2018-11-07) [2021-04-02]. https://arxiv.org/abs/1811.02790.

[151] Gao Y, Xu H, Lin J, et al. Reinforcement learning from imperfect demonstrations[DB/OL]. (2019-05-30) [2021-04-02]. https://arxiv.org/abs/1802.05313.

[152] Brown D S, Goo W, Nagarajan P, et al. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations[J]. Proceedings of Machine Learning Research, 2019, 97: 783-792.

[153] Choi S, Lee K, Oh S. Robust learning from demonstrations with mixed qualities using leveraged gaussian processes [J]. IEEE Transactions on Robotics, 2019, 35(3): 564-576.

[154] Hadfield-Menell D, Dragan A, Abbeel P, et al. Cooperative inverse reinforcement learning[DB/OL]. (2016-11-12) [2021-04-02]. https://arxiv.org/abs/1606.03137.

[155] Zheng J, Liu S, Ni L M. Robust bayesian inverse reinforcement learning with sparse behavior noise[C]//AAAI Conference on Artificial Intelligence. Pola Alto, USA: AAAI, 2014: 2198-2205.

[156] Kilinc O, Hu Y, Montana G. Reinforcement learning for robotic manipulation using simulated locomotion demonstrations[DB/OL]. (2020-06-29) [2021-04-02]. https://arxiv.org/abs/1910.07294.

[157] Song S, Zeng A, Lee J, et al. Grasping in the wild: Learning 6DOF closed-loop grasping from low-cost demonstrations[J]. IEEE Robotics and Automation Letters, 2020, 5(3): 4978-4985.

[158] Young S, Gandhi D, Tulsiani S, et al. Visual imitation made easy[DB/OL]. (2020-08-11) [2021-04-02]. https://arxiv.org/abs/2008.04899.

[159] Wang J X, Kurth-Nelson Z, Kumaran D, et al. Prefrontal cortex as a meta-reinforcement learning system[J]. Nature Neuroscience, 2018, 21(6): 860-868.

[160] Botvinick M, Ritter S, Wang J X, et al. Reinforcement learning, fast and slow[J]. Trends in Cognitive Sciences, 2019, 23(5): 408-422.

[161] Kirsch L, van Steenkiste S, Schmidhuber J. Improving generalization in meta reinforcement learning using learned objectives[DB/OL]. (2020-02-14) [2021-04-02]. https://arxiv.org/abs/1910.04098.

[162] Oh J, Hessel M, Czarnecki W M, et al. Discovering reinforcement learning algorithms[DB/OL]. (2021-01-05) [2021-04-02]. https://arxiv.org/abs/2007.08794.

[163] Rakelly K, Zhou A, Finn C, et al. Efficient off-policy metareinforcement learning via probabilistic context variables[J]. Proceedings of Machine Learning Research, 2019, 97: 5331-5340.

[164] Bechtle S, Molchanov A, Chebotar Y, et al. Meta-learning via learned loss[C]//International Conference on Pattern Recognition (ICPR). Piscataway, USA: IEEE, 2021: 4161-4168.

[165] Fakoor R, Chaudhari P, Soatto S, et al. Meta-Q-learning[DB/OL]. (2020-04-04) [2021-04-02]. https://arxiv.org/abs/1910.00125.

[166] Goyal P, Niekum S, Mooney R J. PixL2R: Guiding reinforcement learning using natural language by mapping pixels to rewards[DB/OL]. (2017-08-16) [2021-04-02]. https://arxiv.org/abs/1708.04782.

[167] Wu B H, Xu F, He Z P, et al. SQUIRL: Robust and efficient learning from video demonstration of long-horizon robotic manipulation tasks[DB/OL]. (2020-03-10) [2021-04-02]. https://arxiv.org/abs/2003.04956.

[168] Yu L T, Yu T H, Finn C, et al. Meta-inverse reinforcement learning with probabilistic context variables[DB/OL]. (2019-10-26) [2021-04-02]. https://arxiv.org/abs/1909.09314.

[169] Ghasemipour S K S, Gu S, Zemel R. SMILe: Scalable meta inverse reinforcement learning through context-conditional policies[C]//Advances in Neural Information Processing Systems. La Jolla, USA: Neural Information Processing Systems Foundation, 2019.

[170] Yu T, Quillen D, He Z, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning[J]. Proceedings of Machine Learning Research, 2020, 100: 1094-1100.

[171] Xu Z J, Qi B C, Agrawal S, et al. AdaGrasp: Learning an adaptive gripper-aware grasping policy[DB/OL]. (2021-03-14) [2021-04-02]. https://arxiv.org/abs/2011.14206.

[172] Yu T H, Kumar S, Gupta A, et al. Gradient surgery for multi-task learning[DB/OL]. (2020-12-22) [2021-04-02]. https://arxiv.org/abs/2001.06782.

作者简介：

陈佳盼（1993 –），男，硕士生。研究领域：深度强化学习，机器人操作。

郑敏华（1988 –），女，博士，讲师。研究领域：人-机器人交互，机器学习，虚拟现实，智能感知。