具身智能机器人技术

Embodied Intelligent Robotics

邵宏/SHAO Hong，谢大雄/XIE Daxiong
（中兴通讯股份有限公司，中国 深圳 518057）
(ZTE Corporation, Shenzhen 518057, China)

DOI：10.12142/ZTETJ.2024S1006
网络出版地址：http://kns.cnki.net/kcms/detail/34.1228.TN.20240724.1458.018.html
网络出版日期：2024-07-25
收稿日期：2023-11-25

摘要：提出了一种用于智能制造的具身智能机器人技术。提出的具身智能机器人通过主动感知环境、自主学习和自主决策来执行拟人化任务，其“大脑”是以强化学习为核心的智能体，感知部分具有双目视觉、空间六维力感知和本体状态感知等多模态感知能力；通过“感知-行动”反馈环，构建了一个控制周期为1 ms的机器人实时控制系统。在中兴通讯的5G智能制造工厂中部署了具身智能机器人进行实际生产，用来取代工人插拔5G小站产品的RJ45插头和光模块。实践表明，具身智能可打通全自动化生产线的最后断点，是一项在智能制造中有广阔应用前景的机器人技术。

关键词：具身智能；强化学习；机器人；多模态感知

Abstract: An embodied intelligent robot technology in industrial intelligent manufacturing is proposed. The robots can learn the policy, make decisions, and act autonomously to complete anthropomorphic tasks through active perception and environmental interaction. The embodied intelligent robot uses a reinforcement learning agent as its “brain” with a real-time control system of 1 ms control cycle with a multi-modal "perception-action" feedback loop, which includes stereo vision, spatial six-dimensional force sensor, and robot proprioception. The system was subsequently deployed in our 5G intelligent manufacturing factory to replace workers, to plug and unplug network RJ45 crystal heads and optical modules for 5G small station manufacture, which eliminated our last breakpoint of automation. The practice shows that embodied intelligence is a promising robot technology in the future manufacturing industry.

Keywords: embodied AI; reinforcement learning; robotic; multimodal perception

引用格式：邵宏, 谢大雄. 具身智能机器人技术 [J]. 中兴通讯技术, 2024, 30(S1): 40-44. DOI: 10.12142/ZTETJ.2024S1006
Citation：SHAO H, XIE D X. Embodied intelligent robotics [J]. ZTE technology journal, 2024, 30(S1): 40-44. DOI: 10.12142/ZTETJ.2024S1006

在生产制造领域，随着自动化程度的提升，越来越多的人工工位由工业机器人和自动化站点来替代，但仍有一些操作无法采用传统工业机器人那种按预定轨迹执行任务的方式完成，如线缆插拔、光模块安装、气密堵头装配、脆弱物体安装等。因为器件的受力情况和运动轨迹在操作过程中需要实时变化，所以这些操作不能由传统机器人按预定轨迹执行，仍然需要人工操作，是全自动化生产线中的众多“断点”。要实现完全无人化的“黑灯工厂”，我们需要一种全新的具有感知、决策和控制能力的智能机器人技术，即具身智能（Embodied AI）机器人来完成这类柔性装配任务[1]，打通自动化生产线的“断点”。

当前业界对具身智能机器人尚没有统一或确切的定义，一般理解为有身体并支持物理交互的智能体，比如人形机器人。具身智能机器人的核心是在“智”而不在“形”，特别是在工业场景中，机器人可以是各种构形的，其核心是可以像人一样主动感知和决策的“大脑”，使它们能够与环境交互以完成拟人任务。上海交通大学卢策吾教授提出，具身智能不仅仅是具有物理身体，而且是具有与人一样的身体体验的能力。

目前的机器人智能技术广泛采用有监督学习方式，例如用机器视觉引导的智能机器人系统，利用深度神经网络完成目标识别、分割等任务。有监督学习的训练是在带有Ground Truth标签的数据集上进行的，是一种旁观标签学习方式，并非通过主动体验进行学习，因此很难通过这种学习方式发展出未来的具身智能。发展具身智能的关键路径是在与环境的交互中实现主动感知与行动的反馈环。虽然以具身智能机器人第一视角得到的数据不够稳定，但这种反馈环可以帮助机器人解决更多真实问题，特别是完成柔性装配任务。这更类似强化学习（RL）的感知和行动过程[2]。

1 系统模型

为了建立“感知-行动”反馈环模型，本文采用强化学习来构建具身智能机器人的系统。感知部分结合了双目视觉

摄像头获取的视觉信息、机器人末端空间力反馈、机器人本体状态数据等，作为机器人智能体对环境的观察变量。行动部分是由智能体周期性地对机器人输出实时控制指令。

本文将具身智能机器人控制系统简化为一个无限步数的部分可观测马尔可夫决策过程（POMDP）[3-4]，这个过程可参数化描述为元组$(\mathcal{O},\mathcal{A},p,r,\gamma)$。其中$\mathcal{O}$为多模态感知状态空间，$\mathcal{A}$为动作空间，观测状态转移概率$Pr(o_t'|o_{\leqslant t},a_t)$为当前动作$a_t$执行后从包括$t$时刻及以前过往观测感知状态迁移到下一观测感知状态$o_t'$的概率，$r:\mathcal{O}\times\mathcal{A}\rightarrow\mathbb{R}$为当前的观测感知和动作的奖励函数，而$\gamma$为折扣系数。

考虑到机器人动作时，由双目视觉摄像头视觉信息、机器人末端空间力反馈和机器人本体状态数据等组成的观测感知，可看作一系列相关联的连续过程。我们可以将其联合作为t时刻的状态变量s_t构成状态空间$\mathcal{S}$，其中：

$$s_t=\{o_t,o_{t-1},o_{t-2},\cdots\}。\tag{1}$$

这样我们就将POMDP过程转化为标准的马尔可夫决策过程（MDP）[5]，可以用元组$(\mathcal{S},\mathcal{A},p,r,\gamma)$描述。这时其状态转移概率$p$和奖励函数$r_t$分别为：

$$p=Pr(s_t'|s_t,a_t),\tag{2}$$

$$r_t=r(s_t,a_t)。\tag{3}$$

此时控制系统的目标就是找到最佳决策$\pi(a_t|s_t)$使得累计折扣奖励最大化：

$$J(\pi)=\max_\pi[\mathbb{E}[\sum_{t=1}^{\infty}\gamma^t r_t|a_t\sim\pi(\cdot|s_t),\ s_t'\sim p(\cdot|s_t,a_t),s_1\sim p(\cdot)])]。\tag{4}$$

最佳决策可以采用基于神经网络的RL算法通过环境交互来训练。为了得到较好的环境适应能力，通常采用无模型(model-free)的RL算法来优化策略函数π。对于实际智能制造环境中的机器人来说，由于在实际环境中进行强化学习训练的成本比较高，因此需要高效利用样本。比较好的方式是采用Off-policy的RL算法而不是On-policy算法。Off-policy的RL方法，如Double Q-learning[6]深度确定策略梯度（DDPG）[7]和SAC（Soft Actor-Critic）[8]等算法，是用带有随机性的行为策略对环境进行探索获取学习样本，然后使用目标策略在行为策略收集交互样本数据上进行训练，最终找到最优策略。

本文提出的具身智能机器人的系统框架如图1所示。

这里采用在真实机器人控制任务中性能优秀、表现稳定的SAC算法。其在优化策略以获取更高累计收益的同时，也会最大化策略的熵，其价值评判（Critic）函数$Q_\theta(s_t,a_t)$和动作策略（Actor）函数$\pi_\theta(a_t|s_t)$分别由不同参数θ的神经网络模型来拟合，并使用温度控制系数α优化寻找经γ折扣后的最大熵来训练模型参数。

在训练时，s_t是从机器人与环境交互产生的过往数据中抽取出的，而a_t是从当前的策略中采样得来。SAC算法的动作策略函数$\pi_\theta(a_t|s_t)$输出是一个关于动作的分布，用tanh-高斯分布的均值μ_θ和标准差σ_θ来控制，而执行确切的动作时，则对均值和标准差的高斯分布进行一次采样，将采样结果a_t作为策略的决策动作：

$$a_t=\tanh(\mu_\theta(s_t)+\sigma_\theta(s_t)\epsilon),\epsilon\sim\mathcal{N}(0,1)。\tag{5}$$

实际任务的复杂性和控制的鲁棒性对于具身智能机器人仍然是一个很大挑战，所以本文在RL模型之外，增加了预设的动作轨迹规划a_p并和RL算法输出的a_t叠加，从而减少π搜索的空间,叠加动作a_u我们可以用下式来描述:

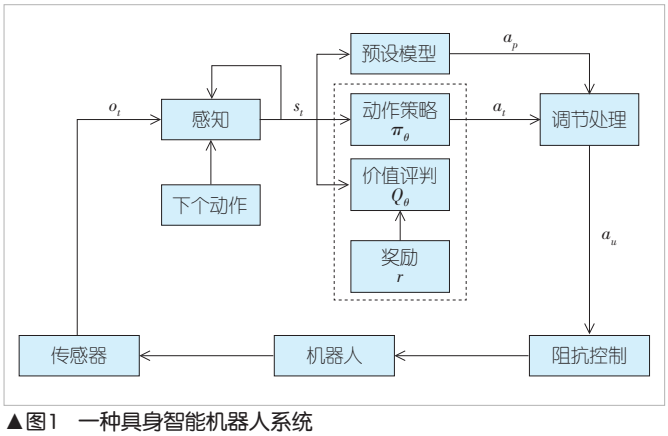
$$a_u=(1-\beta)a_p+\beta a_t,\ \beta\in[0,1],\tag{6}$$

其中，β为模型调节参数。极端情况下，β取值为0时，系统完全工作在基于预先模型的规划动作状态；当取值为1时，系统完全工作在无模型的强化学习状态。

2 多模态感知神经网络结构

通常来说，具身智能机器人对环境的多模态感知会包含视觉、触觉、声音和对自然语言的理解等，但对于制造业中能够完成柔性装配任务的具身智能机器人来说，除了机器人本体姿态感知外，视觉和空间力感知也是最基本的观测变量。这是因为视觉感知可以提供环境和操作工件空间状态信息，空间力感知可以反馈机器人在装配接触过程末端的受力状态。

相较于较低维度的机器人本体状态信息，来自双目视觉摄像头的视觉信息具有很高的维度。因此，在进行多模态感



▲图1　一种具身智能机器人系统

知设计时，我们需要对低维度张量和高维度张量采用不同的处理方法。深度神经网络能够很好地进行高维度的表征学习，但模型训练比较困难，难以在强化学习中直接训练，对此可以在强化学习之前采用自监督学习或自编码器训练一个预训练网络[9]。

我们设计了一种多模态感知神经网络，如图2所示。感知内容主要包括3个部分：双目视觉感知、末端六维力感知、本体状态感知。我们采用带有关节力矩感知器件的7轴机器人作为本体，用装于机器人腕部的双目视觉摄像头作为视觉传感器，结合力反馈和机器人本体状态等低维度信息，以感知更丰富、更深入的周围环境信息。

根据公式（1），将输入传感器感知按照时间序列以8帧$(o_t,o_{t+1},\cdots,o_{t+7})$堆叠后，作为输入的状态变量。

在低维度信息的输入处理部分，本体状态信息自编码器从本体的状态信号中提取出机器人末端工具的位置信息和四元数姿态共7维信息，形成7×8的时间序列帧，然后通过5层因果卷积网络[10]转换为12维特征向量。力反馈自编码器对末端六维力反馈进行处理，读取最近的8帧末端6维F/T力矩，形成8×6时间序列，通过一个4层因果卷积网络[7]转换为12维特征向量。具体实现时，为了训练本体状态信息自编码器和力反馈自编码器，我们用转置反卷积构建了解码器，用采集的数据对两个自编码器进行了预训练。

在高维度的视觉信息处理部分，我们将双目视觉摄像头的图像信息截取为84×84×3的图像块，按8帧进行堆叠，然后输入到4层的卷积神经网络（CNN），再经过1个线性连接层感知机（MLP），最终输出50维的特征向量。为了加快收敛，在训练时参考DrQ算法[11]对图像进行了移动增强处理。

最终上述两部分向量会拼接融合为一个向量作为多模态表征，即动作策略网络和价值评判网络的状态输入维度。动作策略网络输出动作维度为6维的末端笛卡尔坐标变化量，经过控制器生成阻抗控制的动作轨迹，实时控制机器人动作。

3 控制器设计

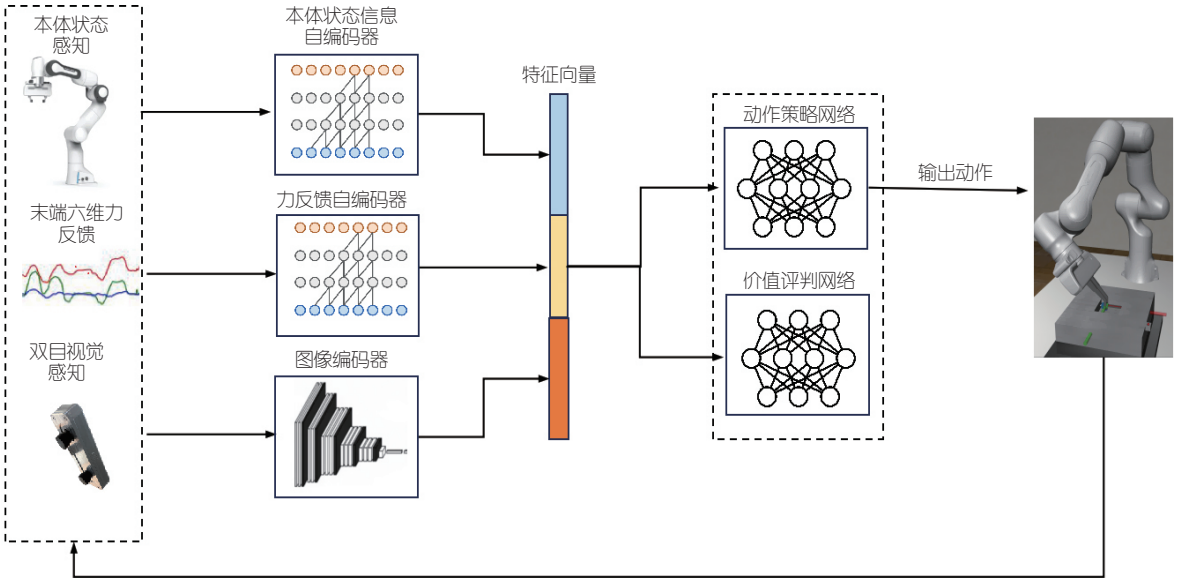
当周期性的多模态感知信息输入到智能体后，智能体向控制器输出20 Hz的末端笛卡尔坐标增量信号a_t，然后由控制器生成1 kHz的机器人关节直接控制力矩τ_u，即控制器将较低带宽的控制策略输出信号转换为高带宽的机器人控制指令，如图3所示。本文设计了阻抗式PID控制器，它以1 kHz频率实时获取机器人当前的末端位置x_t，用x_t+a_t计算出末端期望位置P_d，然后用1 ms的周期做线性插值到X_d，最后阻抗式PID控制器输出控制力矩τ_u：

$$\tau_u = J^T(q)\Lambda\left[P_d - K_p(X - X_d) - K_v(V - V_d)\right], \tag{7}$$

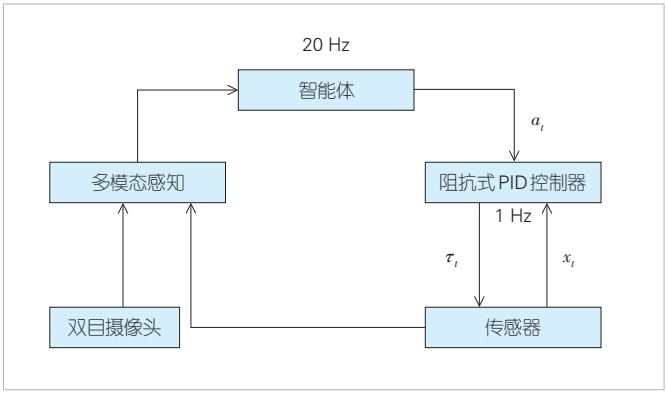
其中，J为实时读取关节角为q的机器人动力学模型和雅可比矩阵，Λ为惯量矩阵，X为机器人末端当前位置，V为空间速度。K_p和K_v分别为可以调节的刚性和阻尼系数。

4 系统仿真与真实环境测试

为了对具身智能机器人系统进行了验证，特别是评估其能否完成5G小站自动插拔线缆等柔性任务，我们搭建了仿



▲图2 多模态感知神经网络设计



▲图3　机器人控制器设计

真环境，使用了Franka七自由度机器人。该机器人具有内置的关节力矩传感器，可以实时给出末端的空间六维力感知数据和机器人的本体姿态数据。

4.1 奖励设计

在插拔5G小站RJ45插头任务时，系统主要依赖根据RJ45插头是否成功插入插口设置稀疏奖励，但为了提高学习效率，也加入密集奖励部分，我们设计了下述奖励函数：

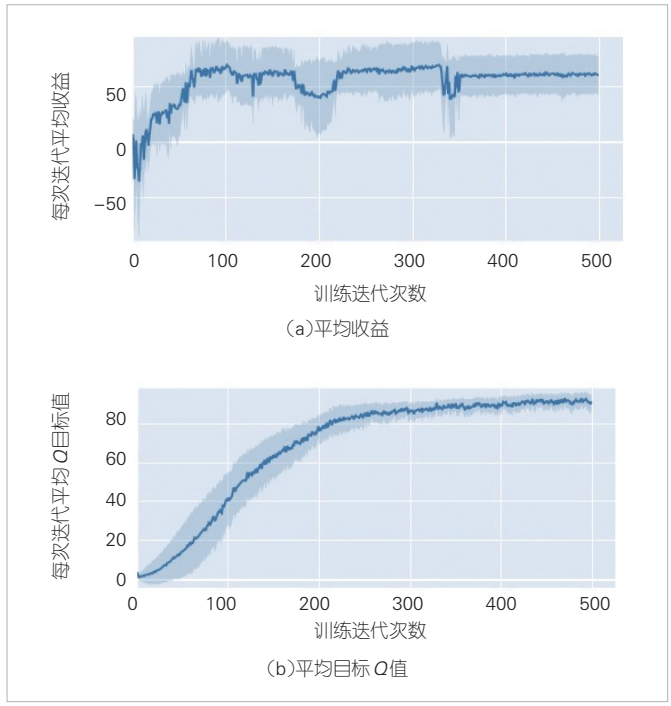
$$r(\mathrm{s})=\begin{cases}\lambda_1\left(c_a-l_x\right), & \text{当 } l_x \geqslant d \text{ 时}\\ \lambda_2\left(c_b-l_x\right), & \text{当 } l_x<d \text{ 且末端六维力超过阈值时}\\ 200, & \text{当Yolo网络判断已经插入成功时}\end{cases},\quad(8)$$

其中，在机器人工具坐标中插头的初始位姿与插入后目标位姿的相对位姿为$(l_x,l_y,l_z,0,0,0)$，d为RJ45插口的深度，式中c_a和c_b为常数，λ_1和λ_2为缩放比例系数。我们根据收到的机器人末端的空间六维力是否超过阈值，来判断操作过程中插头是否成功插入插口。

为了准确判断是否插入成功，我们另外预训练了一个单独的用于判别是否插入成功的目标识别Yolo网络[12]，并采集双目摄像头的视觉信息，做了有监督的预训练。后续在真实的机器人部署环境中，为了增加生产可靠性，我们还对已插入的插头增加了回拉操作，确认插头已经可靠地插入后再给出稀疏奖励信号。

4.2 仿真模拟

仿真模拟中，动作策略网络和价值评判网络的训练采用Adam优化器，批大小设置为128，SAC的软目标更新率设置为0.005。我们训练了500次迭代（epochs）,每次迭代2 500步，采用随机种子进行了10次实验，取平均收益和平均目标Q值，如图4所示。仿真结果表明，系统已经能够很好地完成自动插拔任务。

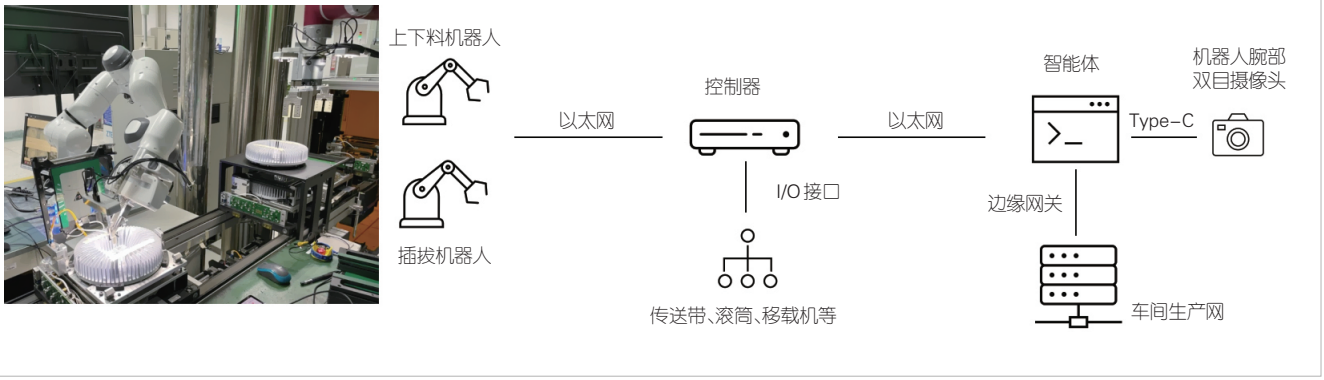


▲图4　仿真强化学习训练结果

4.3 真实生产环境测试

我们将训练后的模型迁移到5G小站生产的器件插拔自动化工位上进行应用和试验。自动插拔工位由两台协作机器人组成，其中一台是与仿真实验相同的Franka机器人，负责网线、光模块的插拔，另外一台机器人则负责自动上下料。自动化插拔工位和后序的环岛测试台的可编程逻辑控制器（PLC）对接以完成整机自动化测试。工位的输入输出（I/O）部分还需要对接控制滚筒和移载机等设备，自动呼叫自动导引车（AGV）接送料等，如图5所示。测试时，为了达到实时性要求，机器人控制器采用打上实时（RT）补丁的Linux操作系统，机器人和控制器之间采用用户数据报协议（UDP）通信，其控制指令的周期为1 ms,并且其要求往返时延（RTT）小于300 μs。超过这个时延的实时数据包将被丢弃。时延连续超过一定阈值，机器人将会停止动作并告警。为了保证机器人控制的稳定性，现场网络必须满足低时延要求，因此测试中采用了低时延的以太网交换机。

在真实生产环境中，RJ45插头插拔和光模块安装任务由对应强化学习模型2个实例化的策略网络来完成，采样和训练分成两个线程并行处理。每次任务操作固定为500步采样，并在每次采样后都进行训练。经过两天的实际训练，机器人能够可靠地自动完成RJ45插头和光模块的插入操作。通过调整上述公式（6）中的参数β可以很好地将有模型和无模型的控制有效结合：在末端工具到达接触范围之前的动



▲图5　具身机器人在实际生产环境中的部署

作中设定β的值为0，到达接触的范围时，将β设定为小于1的一个合适的值，这样就能可靠地批量作业。

5 结束语

具身智能机器人是数字世界融入现实世界的载体，将会成为未来的主流机器人技术之一。通过增加智能体大脑，传统的机器人升级为能够在与环境的互动中进行主动感知和学习的具身智能机器人，并通过获得完成任务后的奖励，在生产制造中不断学习来完成拟人化的任务。本文中的具身智能机器人的设计和应用实践表明，这一技术能够有效打通传统自动化生产中的“断点”，大大提高生产机器人的智能化程度。

参考文献

[1] LIAN W Z, KELCH T, HOLZ D, et al. Benchmarking off-the-shelf solutions to robotic assembly tasks [EB/OL]. [2024-01-15]. http://arxiv.org/abs/2103.05140
[2] SUTTON R S, BARTO A G. Reinforcement learning: an introduction, 2nd edition [M]. London: The MIT Press, 2015
[3] KAELBLING L P, LITTMAN M L, CASSANDRA A R. Planning and acting in partially observable stochastic domains [J]. Artificial intelligence, 1998, 101(1): 99-134
[4] MNIH V, KAVUKCUOGLU K, SILVER D, et al. Playing atari with deep reinforcement learning [EB/OL]. [2024-01-15]. http://arxiv.org/abs/1312.5602
[5] KOSTRIKOV I, YARATS D, FERGUS R. Image augmentation is all you need: regularizing deep reinforcement learning from pixels [EB/OL]. [2024-01-15]. http://arxiv.org/abs/2004.13649
[6] VAN HASSELT H, GUEZ A, SILVER D. Deep reinforcement learning with double Q-learning [EB/OL]. [2024-01-15]. http://arxiv.org/abs/1509.06461
[7] LILLICRAP T P, HUNT J J, PRITZEL A, et al. Continuous control with deep reinforcement learning [EB/OL]. [2024-01-15]. http://arxiv.org/abs/1509.02971
[8] HAARBOJA T, ZHOU A, HARTIKAINEN K, et al. Soft actor-critic algorithms and applications [EB/OL]. [2024-01-15]. https://arxiv.org/pdf/1812.05905
[9] LEE M A, ZHU Y K, ZACHARES P, et al. Making sense of vision and touch: learning multimodal representations for contact-rich tasks [J]. IEEE transactions on robotics, 2020, 36(3): 582-596. DOI: 10.1109/TRO.2019.2959445
[10] OORD A, DIELEMAN S, ZEN H, et al. Wavenet: a generative model for raw audio [EB/OL]. [2024-01-15]. https://arxiv. org/abs/1609.03499
[11] KOSTRIKOV I, YARATS D, FERGUS R. Image augmentation is all you need: regularizing deep reinforcement learning from pixels [EB/OL]. [2024-01-15]. http://arxiv.org/abs/2004.13649
[12] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [EB/OL]. [2024-01-15]. http://arxiv.org/abs/1506.02640

作　者　简　介

邵宏，中兴通讯股份有限公司架构算法专家、移动网络和移动多媒体技术国家重点实验室项目经理，现任中国计算机学会（CCF）智能汽车分会执行常委，是ITU-T、IEEE和IETF资深会员；主要从事具身智能机器人的研究工作；获工业和信息化部ITU-T文稿贡献奖、广东省优秀专利奖、深圳科学技术奖。

谢大雄，中兴通讯股份有限公司监事长、移动网络和移动多媒体技术国家重点实验室主任，教授级高工，中国发明协会会士、国家级领军人才，享受国务院特殊津贴，2023年担任工业和信息化部通信科技术委员会副主任，是《国家中长期科学和技术发展规划纲要（2006-2020年）》“新一代宽带无线移动通信网”重大专项论证委员会委员、国家“973计划”和“863计划”项目带头人；2002年、2010年先后获得国家科技进步奖二等奖2项，2017年获得国家技术发明奖1项，2002年获得首届深圳市市长奖。