



## 视觉强化学习方法研究综述

王荣荣 程玉虎 王雪松

### Overview of Visual Reinforcement Learning Methods

WANG Rong-Rong, CHENG Yu-Hu, WANG Xue-Song

在线阅读 View online: <https://doi.org/10.16383/j.aas.c250422>

---

## 您可能感兴趣的其他文章

### 多智能体强化学习控制与决策研究综述

Survey on Multi-agent Reinforcement Learning for Control and Decision-making

自动化学报. 2025, 51(3): 510–539 <https://doi.org/10.16383/j.aas.c240392>

### 基于表征学习的离线强化学习方法研究综述

A Review of Offline Reinforcement Learning Based on Representation Learning

自动化学报. 2024, 50(6): 1104–1128 <https://doi.org/10.16383/j.aas.c230546>

### 基于梯度损失的离线强化学习算法

Gradient Loss for Offline Reinforcement Learning

自动化学报. 2025, 51(6): 1218–1232 <https://doi.org/10.16383/j.aas.c240481>

### 基于强化学习的流程工业智能决策研究与展望

A Review and Perspective on Reinforcement Learning for Intelligent Decision-making in Process Industries

自动化学报. 2025, 51(10): 2163–2177 <https://doi.org/10.16383/j.aas.c250272>

### 基于滚动时域强化学习的智能车辆侧向控制算法

Receding Horizon Reinforcement Learning Algorithm for Lateral Control of Intelligent Vehicles

自动化学报. 2023, 49(12): 2481–2492 <https://doi.org/10.16383/j.aas.c210555>

### 基于多智能体强化学习的博弈综述

Multi-agent Reinforcement Learning Based Game: A Survey

自动化学报. 2025, 51(3): 540–558 <https://doi.org/10.16383/j.aas.c240478>

# 视觉强化学习方法研究综述

王荣荣<sup>1</sup> 程玉虎<sup>1</sup> 王雪松<sup>1</sup>

**摘要** 视觉作为强化学习智能体感知环境的主要途径,能够提供丰富的细节信息,从而支持智能体实现更复杂、精准的决策.然而,视觉数据的高维特性易导致信息冗余与样本效率低下,成为强化学习应用中的关键挑战.如何在有限交互数据中高效提取关键视觉表征,提升智能体决策能力,已成为当前研究热点.为此,系统梳理视觉强化学习方法,依据核心思想与实现机制,将其归纳为五类:图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习,深入分析各类方法的研究进展及代表性工作的优势与局限.同时,综述 DMControl、DMControl-GB、DCS 和 RL-ViGen 四大主流基准平台,总结视觉强化学习在机器人控制、自动驾驶以及多模态大模型等典型场景中的应用实践.最后,结合当前研究瓶颈,探讨未来发展趋势与潜在研究方向,以期为该领域提供清晰的技术脉络与研究参考.

**关键词** 强化学习;视觉表征;视觉强化学习;智能体

**引用格式** 王荣荣,程玉虎,王雪松.视觉强化学习方法研究综述.自动化学报,2026,52(3):381-410

**DOI** 10.16383/j.aas.c250422 **CSTR** 32138.14.j.aas.c250422

## Overview of Visual Reinforcement Learning Methods

WANG Rong-Rong<sup>1</sup> CHENG Yu-Hu<sup>1</sup> WANG Xue-Song<sup>1</sup>

**Abstract** Vision, as the primary means for reinforcement learning agents to perceive their environment, provides rich and detailed information that supports agents in making more complex and precise decisions. However, the high-dimensional nature of visual data often leads to information redundancy and low sample efficiency, posing a key challenge in the application of reinforcement learning. How to efficiently extract key visual representations from limited interaction data to enhance agents' decision-making capabilities has become a current research focus. To address this, this paper systematically reviews visual reinforcement learning methods, categorizing them into five categories based on their core ideas and implementation mechanisms: Image-enhanced, model-enhanced, task-assisted, knowledge-transferred, and offline visual reinforcement learning approaches. It provides an in-depth analysis of the research progress in each category, as well as the strengths and limitations of representative works. Meanwhile, this paper reviews four major benchmark platforms: DMControl, DMControl-GB, DCS, and RL-ViGen, and summarizes the applications of visual reinforcement learning in typical scenarios such as robotic control, autonomous driving, and multimodal large models. Finally, based on current research bottlenecks, this paper discusses future development trends and potential research directions, aiming to offer a clear technical framework and research reference for this field.

**Keywords** reinforcement learning; visual representation; visual reinforcement learning; agent

**Citation** Wang Rong-Rong, Cheng Yu-Hu, Wang Xue-Song. Overview of visual reinforcement learning methods. *Acta Automatica Sinica*, 2026, 52(3): 381-410

在人工智能领域,强化学习(reinforcement learning, RL)<sup>[1]</sup>关注的是智能体在环境中采取动作以获得最大化累积奖励的过程.近年来,随着计算机技术的飞速发展,强化学习在游戏<sup>[2]</sup>、机器人自主

决策<sup>[3]</sup>、视觉导航<sup>[4]</sup>、智能交通系统<sup>[5]</sup>等领域取得突破性进展.我国高度重视强化学习的发展:2017年《国务院关于印发新一代人工智能发展规划的通知》<sup>[6]</sup>首次将强化学习列为算法创新的重点方向;2024年颁布的《国家人工智能产业综合标准化体系建设指南(2024版)》<sup>[7]</sup>进一步明确提出,要规范机器学习中的强化学习技术标准,以及以通用大模型为核心的智能体实例和智能体基本功能、应用架构等技术要求,具体涵盖智能体强化学习、多任务分解、推理、提示词工程等多个标准.这些政策部署不仅印证了强化学习在技术创新中的战略地位,更凸显了其作为国家科技发展重要基石的现实意义.

收稿日期 2025-08-29 录用日期 2025-11-21

Manuscript received August 29, 2025; accepted November 21, 2025

国家自然科学基金(62373364, 62573416),江苏省重点研发计划(BE2022095)资助

Supported by National Natural Science Foundation of China (62373364, 62573416) and Key Research and Development Program of Jiangsu Province (BE2022095)

本文责任编辑 穆朝黎

Recommended by Associate Editor MU Chao-Xu

1. 中国矿业大学信息与控制工程学院 徐州 221116

1. School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116

传统的强化学习方法,如深度 Q 学习<sup>[8]</sup>、柔性 Actor-Critic 学习<sup>[9]</sup>以及近端策略优化<sup>[10]</sup>等,通常依赖人工精心设计的状态输入来引导智能体的学习过程.然而,在现实世界中,大约 80% 的信息是通过视觉途径获取的<sup>[11]</sup>.视觉作为感知外部环境的主要方式之一,不仅提供海量且高维度的输入数据,还蕴含环境的丰富细节.因此,将视觉信息融入到强化学习框架中,能够使智能体基于视觉感知做出更为复杂和精准的决策.基于这一理念,视觉强化学习<sup>[12]</sup>(visual reinforcement learning, VRL)逐渐成为人工智能领域的研究热点.该技术强调利用视觉输入来指导智能体的决策过程,其核心优势在于能够直接从高维、丰富的图像数据中提取环境特征,使智能体在面临视觉复杂的任务时能够有效地学习和执行相应的策略.

然而,从高维视觉输入中学习有效策略仍面临诸多挑战.高维视觉数据通常包含大量与任务无关的冗余信息或噪声,如何从中准确提取与决策密切相关的关键特征,构建低维且具有表征力的状态信息,成为亟待突破的核心难题.与此同时,样本效率低下的问题也严重制约着实际应用的发展:在高维输入下进行策略学习往往依赖海量的环境交互数据,而现实场景中数据采集成本高昂、交互过程缓慢,这一矛盾显著限制了模型的训练效率和应用可行性.因此,在交互数据有限的条件下,如何提升智能体从高维视觉输入中提取关键信息的能力,以实现其在复杂视觉环境中的高效与稳健决策,已成为当前研究的热点与难点.

本文旨在对视觉强化学习方法进行全面而深入的综述.首先,通过阐述该领域的基本概念与问题描述,为后续内容的展开奠定理论基础.在此基础上,系统梳理视觉强化学习方法的发展脉络,重点归纳并分析五类主流技术路线:图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习方法,深入剖析各类方法的技术原理、代表性成果、适用场景以及各自的优缺点.随后,介绍当

前广泛使用的四种基准测试平台,为方法的评估提供支撑.此外,本文还探讨视觉强化学习在机器人控制、自动驾驶、多模态大模型等前沿领域的实际应用案例,展示其应用价值与发展潜力.最后,总结当前研究中存在的问题和挑战,并对未来的研究方向进行展望,以期为该领域的研究者提供有益参考.

本文与文献 [12]、文献 [13] 均为视觉强化学习领域的综述性论文.为更清晰地呈现本文的独特价值,表 1 从覆盖范围、分类框架、基准平台和应用场景四个维度,将本文与上述两篇文献进行系统比较.文献 [12] 主要聚焦于图像增强型视觉强化学习,重点关注无模型在线场景,其分类框架围绕增强数据的生成方式以及对增强数据的利用两个角度展开,并对 Atari 游戏、DMControl 及其变体等常用视觉强化学习基准平台进行介绍.文献 [13] 则侧重于状态表征的生成方式,将现有方法划分为基于度量、辅助任务、数据增强、对比学习、非对比学习以及基于注意力等类别,但其涵盖范围仍局限于无模型在线场景,且未对相关基准平台展开系统评述.相比之下,本文在覆盖广度和内容深度上均有显著拓展:不仅全面涵盖无模型与基于模型两大类方法,还系统梳理了在线与离线两种学习范式下的视觉强化学习进展.在基准平台方面,本文进一步讨论一个覆盖多样化视觉应用场景的综合基准平台 RL-ViGen,为视觉强化学习的系统评估与比较提供重要参考.

本文结构安排如下:第 1 节介绍视觉强化学习的背景知识与问题定义;第 2 节系统梳理图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习五类方法的研究进展;第 3 节概述四种常用基准测试平台及其特点;第 4 节阐述视觉强化学习在机器人控制、自动驾驶和多模态大模型等典型场景中的应用;第 5 节探讨该领域面临的挑战与未来发展趋势;第 6 节对全文进行总结.

## 1 问题陈述

视觉强化学习问题通常被建模为一个马尔科夫决策过程 (Markov decision process, MDP)<sup>[1]</sup>,形式

表 1 与已有综述对比  
Table 1 Comparison with existing reviews

综述	覆盖范围	分类框架	基准平台	应用场景
文献 [12]	图像增强型与任务辅助型 VRL	基于数据增强在 VRL 中的使用方式:生成增强数据与利用增强数据	主要涵盖: Atari 游戏、DMControl 及其变体;简要提及: OpenAI Procgen、DeepMind Lab、CARLA	涵盖无模型方法;适用于在线 VRL 环境
文献 [13]	图像增强型与任务辅助型 VRL	基于状态表征学习方法:基于度量、辅助任务、数据增强、对比学习、非对比学习和基于注意力的方法	未明确讨论或比较基准平台	涵盖无模型方法;适用于在线 VRL 环境
本文	五种类型全覆盖	基于方法的核心思想与实现机制:图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习	DMControl、DMControl-GB、DCS 和 RL-ViGen	涵盖无模型与基于模型方法;适用于在线与离线 VRL 环境

化表示为六元组  $M = \langle O, S, A, P, r, \gamma \rangle$ . 在该框架中,  $O$  表示观测空间;  $S$  表示状态空间;  $A$  表示动作空间;  $P(s_{t+1}|s_t, a_t)$  为状态转移函数, 描述在状态  $s_t \in S$  下执行动作  $a_t \in A$  后转移到下一个状态  $s_{t+1} \in S$  的概率;  $r$  表示奖励函数;  $\gamma \in [0, 1)$  为折扣因子, 用于权衡当前和未来奖励的重要性. 智能体通过策略  $\pi$  从动作空间  $A$  中采样一个动作  $a$  并与环境交互, 其核心目标是学习一个最优策略  $\pi^*$ , 以最大化长期累积折扣奖励的期望值, 即  $E_{\pi} [\sum_{t=0}^{\infty} \gamma^t \times r_t]$ .

根据环境感知信息的可获取性, 视觉强化学习可划分为以下两类典型场景:

1) 全观测场景: 智能体在获取环境内部状态信息的同时, 也能够接收视觉观测数据. 在此类场景中, 视觉信息通常作为状态的有益补充, 用于提升策略学习的效果. 例如通过设计内在奖励等方式进一步优化策略.

2) 部分观测场景: 智能体仅能通过视觉传感器获取图像观测, 无法直接访问环境的完整状态. 由于单帧图像往往难以充分捕捉系统动态, 为缓解部分观测性带来的挑战, 通常参考文献 [14] 的做法, 用连续  $k$  帧观测堆叠来近似当前环境状态, 即  $s_t = (o_{t-(k-1)}, \dots, o_{t-1}, o_t)$ , 其中  $o_t \in O$  表示时刻  $t$  所获取的视觉观测. 该方式能够在一定程度上保留环

境的时间动态信息, 为策略学习提供更丰富的感知上下文.

## 2 方法分类

为系统梳理视觉强化学习方法的研究进展, 本文依据方法的核心思想与实现机制, 将其划分为五类: 图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习. 这五类方法在总体目标、核心思想、适用场景等方面各具优势与局限, 其具体特性与对比详见表 2. 基于这一分类体系, 本文将系统综述各类方法的研究现状, 详细分析代表性工作的技术特点, 并客观评述现有方法面临的挑战与发展方向.

### 2.1 图像增强型视觉强化学习

图像增强型视觉强化学习方法的核心思想在于: 通过随机裁剪、缩放、掩码等图像增强技术生成的增强样本, 应保持与原始样本相似的特征分布, 从而确保模型能够有效学习数据中的关键信息 (如图 1 所示). 这类方法旨在提升模型的样本效率和泛化能力, 使其在面对真实视觉决策任务中的多种变化时, 仍能保持稳定且高效的性能表现. 根据增强方式的不同, 图像增强型视觉强化学习方法可分为两类: 基础特征增强与高级语义增强. 图 2 展示了

表 2 视觉强化学习方法对比  
Table 2 Comparison of visual reinforcement learning methods

方法类别	子类	总体目标	核心思想	优势	劣势	适用场景
图像增强型 VRL	基础特征增强	提升数据利用率	直接修改像素值或频谱信息	实现简单、计算开销低、通用性强	易破坏关键语义信息, 误导策略学习	通用视觉任务
	高级语义增强	优化关键区域学习	基于语义显著性进行针对性增强	保留并强化关键区域, 提升语义理解能力	依赖显著性检测精度, 实现复杂	需精细感知的任务
模型增强型 VRL	基于世界模型的 VRL	减少环境交互, 提升推理能力	构建内部环境动力学模型, 预测状态转移	可在模拟中规划, 减少真实交互	模型偏差易导致策略退化	复杂动力学环境、长视野任务
	视觉表征学习	提升视觉理解能力	利用预训练大模型提取低维语义特征	特征提取能力强, 泛化性好	存在领域差异, 计算资源消耗大	复杂视觉理解任务
任务辅助型 VRL	奖励生成	解决奖励稀疏问题	利用大模型生成密集、语义合理的奖励信号	减少人工设计成本, 缓解稀疏奖励	存在幻觉风险, 奖励可能不准确	奖励设计困难的任務
	自监督学习	提升表征学习效率	设计自监督任务促进特征学习			
	未来帧预测相似性度量	学习环境动态	预测未来图像帧来学习状态表征	提升智能体提取特征能力, 增强对环境理解	可能与主任务冲突, 增加训练复杂度	需丰富表征的任务
知识迁移型 VRL	信息论	提升状态一致性	通过相似性度量增强状态表征一致性			
	多视角 VRL	学习稳健表征, 提升泛化力	压缩无关信息, 保留任务相关信息			
离线 VRL	多视角 VRL	克服遮挡与盲区	融合多个视角图像信息	提升环境感知完整性	视角缺失或质量差时性能下降	多摄像头环境
	—	提升跨域泛化能力	迁移已有知识至新环境	加速适应, 提升泛化	领域差异大时可能负迁移	跨环境、跨任务迁移
离线 VRL	—	从静态图像数据集中学习策略	利用历史数据集训练, 无需环境交互	安全性高, 避免在线试错风险	易受数据分布偏移影响, 泛化能力受限	数据集丰富但交互受限场景

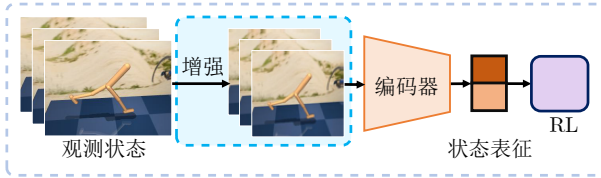


图 1 图像增强型视觉强化学习示意图

Fig. 1 The diagram of image-enhanced VRL

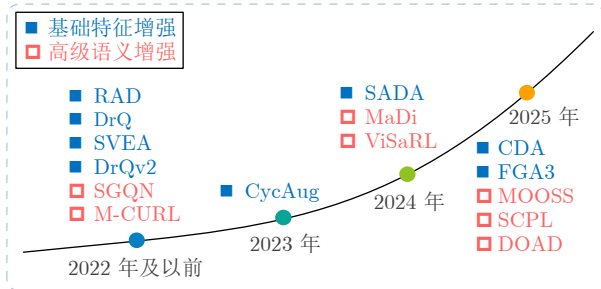


图 2 图像增强型视觉强化学习时间线

Fig. 2 The timeline of image-enhanced VRL

该类方法的时间线。

下面介绍一种经典的图像增强型视觉强化学习算法。2020年, Laskin等<sup>[15]</sup>首次系统探讨了利用基本图像增强技术(如裁剪、平移、翻转、旋转等)提升视觉强化学习性能的可行性,并提出基于增强数据的强化学习方法(reinforcement learning with augmented data, RAD)<sup>[15]</sup>。本文以在Actor-Critic类算法中引入RAD为例。假设当前状态 $s$ 由多帧观测图像构成,通过引入图像变换算子 $f(\cdot)$ ,可将原始状态 $s$ 转换为增强后的状态 $f(s)$ ,并将其作为Critic网络的输入。此时,Critic网络的优化目标为最小化贝尔曼误差,其损失函数定义如下:

$$L_Q(\zeta) = \mathbb{E}_{(s, a, r, s') \sim D} [(Q_\zeta(f(s), a) - (r + \gamma V_\psi(f(s'))))^2] \quad (1)$$

其中, $D$ 表示经验回放池,用于存储与环境交互产生的样本; $Q$ 为动作值函数; $\zeta$ 为Critic网络的参数; $V_\psi(\cdot)$ 是由参数 $\psi$ 参数化的状态值函数网络; $s'$ 为下一时刻状态。

可以看到,RAD的核心思路为:在强化学习框架(如柔性Actor-Critic<sup>[9]</sup>)中,直接对输入的观测图像施加数据增强,然后利用编码器对增强后的样本进行特征编码,从而得到状态表征。这种方法使得智能体能够更有效地从增强后的观测数据中学习策略。该研究为深入探索数据增强技术与强化学习的结合提供了重要基础。

### 2.1.1 基础特征增强

基础特征增强侧重于图像的底层物理属性,如

亮度、对比度、几何形变、频域成分及像素级噪声等。该方法通过几何变换、光度调整、频域滤波或添加噪声等数学操作,直接修改像素值或频谱信息,无需理解图像的具体内容,具有较强的通用性和计算效率。

上述RAD并未对强化学习的核心架构进行本质上的改进。相比之下,在RAD<sup>[15]</sup>的基础上进一步发展而来的数据正则化Q学习(data-regularized Q, DrQ)<sup>[16]</sup>不仅允许灵活调整状态增强,而且还在模型训练过程中引入两种不同的Q函数正则项(对目标Q值进行平均化和对在线Q函数进行平均化),以使得相同输入图像的不同转换结构都能有相似的Q值。

RAD<sup>[15]</sup>和DrQ<sup>[16]</sup>均聚焦于如何将图像增强技术引入到柔性Actor-Critic的状态观测图像中,DrQ的改进版本DrQv2<sup>[17]</sup>则是采用深度确定性策略梯度(deep deterministic policy gradient, DDPG)<sup>[18]</sup>作为基础强化学习算法。此外,为更好地应对基于视觉的控制挑战,DrQv2还在多个实现细节上进行精细而重要的调整,具体包括:应用双线性插值方法对随机平移图像进行平滑处理,从而减少图像平移时所产生的锯齿现象;通过整合多步回报与目标Critic技术实现更快的收敛速度,这表明DrQv2能够更加高效地从历史经验中学习并逐步优化策略;使用一种逐渐衰减的探索噪声调度机制,旨在降低过拟合风险并增强智能体的泛化能力。

然而,Hansen等<sup>[19]</sup>在视觉强化学习研究中发现,RAD<sup>[15]</sup>和DrQ<sup>[16]</sup>采用的裁剪、平移、翻转等图像增强技术因两大问题易导致模型训练不稳定:一是无差别应用图像增强导致Q目标高方差;二是严格从增强数据中估计Q值引发过度正则化。为此,Hansen等<sup>[19]</sup>提出一个用于稳定Q值估计的数据增强框架(stabilized Q-value estimation under augmentation, SVEA),主要包括三个关键点:仅在当前状态的Q值估计中使用图像增强技术,避免错误自举;优化一个结合增强和未增强观测状态的Q值目标;在未增强数据上优化Actor网络,并通过参数共享间接学习通用策略。在这些改进措施的共同作用下,SVEA无需额外的前向传递或引入额外的可学习参数即可提高Q值估计的稳定性。但是,SVEA假设编码器的输出对输入增强完全不变,这种假设会导致两个主要问题:卷积神经网络编码器的输出无法对几何变换(如旋转和平移)保持不变;由于编码器与Critic网络一同实施端到端训练,部分不变性可能会被转移到Critic网络上。

为克服这些问题,Almuzairee等<sup>[20]</sup>提出一种基于数据增强的稳定Actor-Critic学习方法(stabil-

ized Actor-Critic under data augmentation, SADA). SADA 不仅增强了 Critic 网络的输入, 而且增强了 Actor 网络的输入, 并进行微小的调整以避免训练不稳定, 主要包括: 在更新 Actor 网络时, 仅增强策略输入而不改变 Q 函数输入; 在更新 Critic 网络时, 仅增强在线 Q 函数输入而不改变目标 Q 函数输入; 在增强和未增强的数据上进行联合优化.

上述方法依赖手动选择图像增强算子, 难以适应不同任务需求, 且单一算子无法充分利用多样化的先验知识. 针对此问题, Xiong 等<sup>[21]</sup>提出的组合数据增强 (combined data augmentation, CDA) 框架将增强方法分为三个数据流: 原始图像、像素级增强和空间级增强. 该框架通过原图与像素级增强图的凸组合减少分布偏移, 并将空间级增强作为正则项, 由超参数控制其影响以避免误导模型学习. CDA 无需修改网络结构即可灵活集成多种增强算子. 在 DMControl 及其泛化任务 DMControl-GB 上的大量实验表明, CDA 在所有训练环境中均达到相当或更优的渐近性能. 此外, 在 DMControl-GB 的“color hard”与“video easy”设置中, CDA 在 10 项任务中有 6 项的泛化性能超越基线方法. 但该方法仍需手动调整凸组合系数和正则化超参数, 难以实现自适应优化.

尽管 CDA 在一定程度上提升了增强策略的灵活性, 但其对人工调参的依赖限制了泛化能力. Ma 等<sup>[22]</sup>则进一步探讨图像增强在提升视觉强化学习样本效率方面的关键属性, 并提出两种策略来增强其有效性. 研究指出, 空间多样性和适度的硬度是单个数据增强操作中的重要属性. 基于此, Ma 等<sup>[22]</sup>引入一种新的增强算子 Rand PR (random pad-size). 此外, 考虑到强化学习的非平稳特性, 还设计一种称为 CycAug (cycling augmentation) 的多类型数据增强融合方案. 该方案通过周期性地变换多种数据增强技术的应用, 既增加了增强类型的多样性, 又保持了数据分布的一致性.

上述方法虽然从空间域的角度提升了增强的多样性和一致性, 然而, 文献 [23] 指出, 直接基于原始图像进行学习容易因训练与测试环境之间存在域间隙, 导致在未见环境中的性能下降. 为此, 该研究转向频域进行图像增强, 提出傅里叶引导的自适应对抗增强方法 (Fourier guided adaptive adversarial augmentation, FGA3)<sup>[23]</sup>. 该方法首先对图像进行傅里叶变换以分离幅度和相位信息, 随后在傅里叶幅度域生成对抗性样本, 最终联合原始样本和对抗性样本训练并优化值网络. 这种频域处理方法能有效保持语义一致性. 在 DMControl-GB 的随机颜色基准测试中, SAC 结合 FGA3 在颜色变化环境下

展现出最优泛化性能, 提升达 52.6%. 在自然视频基准测试中, 该方法在背景变化环境下也表现出最佳适应能力, 性能提升达 69.6%.

总之, 基础特征增强的优点是实现简单、计算成本低且通用性强, 能通过增加数据多样性有效提升模型的泛化能力; 然而, 其缺点在于增强过程是盲目和均匀的, 可能破坏图像中的关键决策信息 (如裁剪掉重要物体或扭曲关键颜色), 从而引入噪声甚至误导学习过程, 导致策略性能下降.

### 2.1.2 高级语义增强

高级语义增强则基于图像的高层语义信息, 例如通过显著性检测、语义分割生成的掩码进行有针对性的增强. 这类方法通常依赖深度学习模型来理解图像内容, 并结合大规模标注数据实现对关键区域的优化, 从而提升模型对语义信息的学习能力.

尽管上述基础特征增强方法充分阐释了图像增强技术在视觉强化学习上的有效性, 但是直接对视觉观测进行图像增强会受到任务无关因素的干扰, 难以有效应对智能体在未知环境中的泛化挑战. 针对高维图像或视频帧输入中存在的任务无关视觉干扰问题, Grooten 等<sup>[24]</sup>通过将掩码器模块集成到标准的 Actor-Critic 框架中, 提出一种基于扰动掩码的视觉强化学习方法 (masking distractions, MaDi), 主要思路为: 运用掩码算子动态调整状态观测图像每个像素的亮度水平, 从而有效过滤与任务无关的视觉干扰. 但是, 由于 MaDi 使用 11 层深度的卷积神经网络作为状态编码器以及端到端的训练方式, 这不仅在一定程度上增加了计算成本, 而且导致其难以处理非结构化的高维场景.

在提升模型对关键视觉区域感知能力的探索中, 为使策略学习更专注于重要的任务相关区域而忽略模糊或干扰像素, Bertoin 等<sup>[25]</sup>提出显著性引导 Q 网络 (saliency-guided Q-networks, SGQN), 将二值化的归因图作为输入状态上的掩码, 并利用值函数的一致性来规范策略学习目标. 同时, SGQN 还定义一个自监督学习辅助任务, 旨在匹配增强图像与原始图像的归因图, 引导模型在训练过程中聚焦于状态间共享的特征. 然而, SGQN<sup>[25]</sup>引入的额外网络参数数量相当庞大, 给强化学习基础架构增加了约 1.6 M 的参数 (相当于增加了 25%), 这无疑加大了内存需求.

为更高效地利用视觉序列中的全局依赖关系, Zhu 等<sup>[26]</sup>则通过引入基于 Transformer 的掩码重构机制来充分利用视觉数据之间的高相关性, 提出一种掩码对比表征学习 (masked contrastive representation learning for RL, M-CURL)<sup>[26]</sup>. M-CURL 采用两个编码器: 一是用于提取图像特征的在线编

码器,使用梯度更新的方式端到端地进行训练;二是势能编码器,其编码结果被视作重构目标,通过在线编码器的指数滑动平均更新. M-CURL 中的 Transformer 解码器主要用于根据动作表征、位置表征以及在线编码器得到的表征来预测重构后的状态表征. 为提升重构效率,该研究还额外引入 BYOL (bootstrap your own latent)<sup>[27]</sup> 方法中的投影模块和预测模块来对 Transformer 解码器的输出做进一步处理. M-CURL 能够提高智能体在学习状态表征时对全局范围动态的感知能力,但是在建模状态的细微演化方面仍存在不足. 为更精细地刻画状态演变过程, Sun 等<sup>[28]</sup> 通过结合时间对比目标和基于图的时空掩码技术,设计一种基于未来帧预测辅助任务的掩码增强时间对比学习方法 (mask-enhanced temporal contrastive learning for smooth state evolution, MOOSS),能够清晰地建模视觉强化学习中的状态演化过程.

除了纯算法层面的改进,也有研究尝试引入人类视觉认知机制. 进一步, Liang 等<sup>[29]</sup> 提出视觉显著性引导强化学习 (visual saliency-guided RL, ViSaRL) 算法,利用人类视觉注意力提升机器人控制任务的效果. 该方法首先基于人工标注显著性图训练预测模型,识别图像关键区域;其次,利用该模型对离线数据集进行伪标注,并通过多模态自编码器预训练图像编码器,使其聚焦显著特征;最后固定编码器,利用其潜在表征进行策略训练. ViSaRL 在仿真与真实机器人任务中均表现出色. 在 Meta-World 任务中,基于 CNN 与 Transformer 编码器的任务成功率分别提升 13% 和 18%;在 DMControl 任务中,采用显著性输入的平均回报相对提高 256%. 此外,在 DMControl-GB 基准测试中, ViSaRL 在颜色与视频背景扰动下的性能分别相对提升 19% 和 35%.

延续显著性引导的思路并着眼于提升策略的泛化鲁棒性,为有效提升视觉强化学习的策略泛化能力, Sun 等<sup>[30]</sup> 提出显著性不变的一致性策略学习 (saliency-invariant consistent policy learning, SCPL) 算法. 该算法通过三个关键设计实现这一目标:首先,引入显著性属性掩码图,使编码器和值函数在原始与扰动观测中均能聚焦任务相关区域;其次,利用增强数据生成动力学相关表征,帮助编码器捕获任务及奖励特征;最后,通过 KL 散度约束保持原始与扰动观测间的策略一致性. SCPL 在 15 个 DMControl-GB 视觉扰动任务中的 13 个上表现优异,尤其在“video hard”设置中,平均性能领先其他基线 14%. 然而, SCPL 采用固定的显著性分位数阈值来生成注意力掩码,这一设计可能限制其在不同应用场景中的适应性.

值得注意的是,在数据稀缺的场景下,上述基于高级语义增强的方法面临过拟合的风险. 在数据样本有限的情况下(如 Atari 100K),单纯引入视觉注意力机制会加剧强化学习的过拟合问题. 针对这一挑战, Ma 等<sup>[31]</sup> 提出“不忽视任何细节”(don't overlook any detail, DOAD) 算法. 首先,通过神经网络提取动力学任务驱动的显著性图,并将其权重融入原始观测图像,模拟人类在视觉决策中对任务相关区域的专注能力;其次,采用两阶段训练策略,优化视觉表征学习和 Q 值估计;最后,引入条件网络重置方法,模拟人类学习的灵活性以促进模型适应新信息. 仿真结果表明, DOAD 在 Atari 100K 基准测试中有效提升了有限数据环境下的学习性能.

上述高级语义增强方法的优点是能够聚焦于任务关键区域,通过抑制无关背景干扰来提升学习效率、最终性能以及策略的可解释性;但其缺点是实现复杂、计算开销大,并且高度依赖显著性图生成的准确性,一旦显著性识别错误,会系统性地将智能体的注意力引向歧途,造成灾难性的误导.

综上所述,尽管图像增强型视觉强化学习方法在一定程度上能够促使模型学习到更加泛化的状态表征,然而,这类方法也存在一些局限性:其一,增强效果依赖于具体方式的选择,不当的增强可能损害模型性能,如过度增强会掩盖原始数据的关键信息;其二,在端到端训练中,图像增强并未直接促进对状态表征的深度理解,缺乏专门的编码器优化机制,难以应对复杂视觉任务;其三,静态的图像增强难以适应任务的动态变化,无法准确捕捉特定任务所需的多样化特征. 这些局限制约了模型的表征能力和泛化性能.

## 2.2 模型增强型视觉强化学习

模型增强型视觉强化学习聚焦于通过引入或构建特定模型来提升算法性能. 主要包含两个分支:一是基于世界模型的视觉强化学习,通过构建环境动力学模型进行规划,降低对真实环境交互的依赖(如图 3 所示);二是大模型增强的视觉强化学习,利用预训练的大模型提取特征或提供先验知识,加速策略学习过程. 图 4 给出了模型增强型视觉强化学习方法的时间线.

接下来介绍一种经典的模型增强型视觉强化学习方法 Dreamer<sup>[32]</sup>. 其核心思想是通过在学习到的紧凑潜在空间中想象轨迹来高效训练策略. 具体流程为:首先利用历史经验学习一个包含表征模型、状态转移模型和奖励模型的潜在动力学世界模型;然后在该模型中进行轨迹想象,并联合训练动作策略网络和状态值网络. 其中,策略通过反向传播值

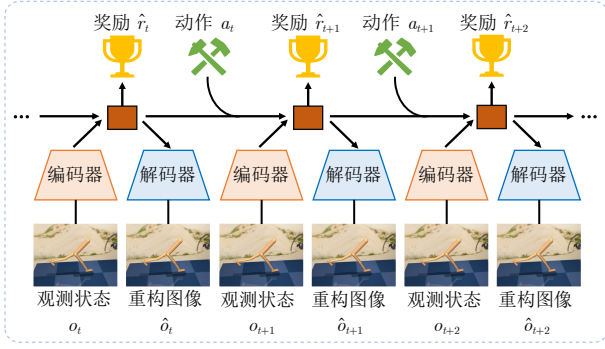


图 3 模型增强型视觉强化学习示意图 (以基于世界模型的视觉强化学习为例)

Fig. 3 The diagram of model-enhanced VRL (Taking world model-based VRL as an example)

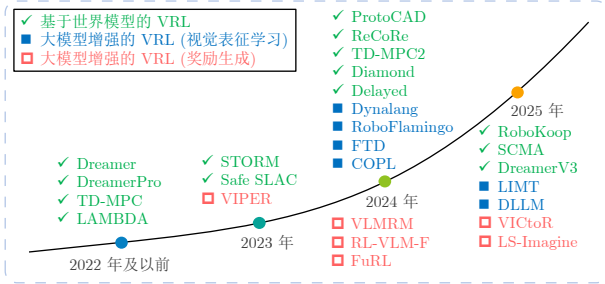


图 4 模型增强型视觉强化学习时间线

Fig. 4 The timeline of model-enhanced VRL

估计的解析梯度进行优化, 从而兼顾想象范围之外的长期回报. Dreamer 的模型结构如下<sup>[32]</sup>:

- 1) 表征模型:  $z_t \sim p_\theta(z_t | z_{t-1}, a_{t-1}, o_t)$
- 2) 状态转移模型:  $\hat{z}_t \sim q_\theta(\hat{z}_t | z_{t-1}, a_{t-1})$
- 3) 观测模型:  $\hat{o}_t \sim q_\theta(\hat{o}_t | z_t)$
- 4) 奖励模型:  $\hat{r}_t \sim q_\theta(\hat{r}_t | z_t)$

其中,  $p$  表示真实环境中的样本分布;  $q$  为潜在想象过程中的近似分布;  $z$  表示经过编码器得到的潜在状态表征;  $\hat{z}$  为状态转移模型得到的预测状态表征;  $\hat{o}$  为观测模型得到的预测观测表征;  $\hat{r}$  为奖励模型得到的预测奖励;  $\theta$  为世界模型参数.

通过变分信息瓶颈联合优化上述组件, 得到如下损失函数<sup>[32]</sup>:

$$L(\theta) = \mathbb{E}_{p_\theta(z_{1:T} | a_{1:T}, o_{1:T})} \left[ \sum_{t=1}^T (-\log q_\theta(o_t | z_t) - \log q_\theta(r_t | z_t) + \beta \text{KL}(p_\theta(z_t | z_{t-1}, a_{t-1}, o_t) \| q_\theta(\hat{z}_t | z_{t-1}, a_{t-1}))) \right] \quad (2)$$

其中,  $\beta$  为权衡因子;  $T$  为一个序列批次的长度; KL

表示 Kullback-Leibler 散度. 该损失函数包含三项: 观测重构损失、奖励重构损失以及表征先验与后验分布之间的动力学正则项.

状态转移模型能够在紧凑的潜在空间中进行前向预测, 无需观察或想象对应的图像, 从而实现成千上万条想象轨迹的低内存、快速并行预测. 在有限时间范围  $H$  的想象轨迹中, Dreamer 在世界模型的潜在空间中学习一个动作模型和一个值模型: 动作模型执行策略并预测动作; 值模型估计从状态  $s_\tau$  出发所能获得的期望累积奖励.

1) 动作模型:  $a_\tau \sim q_\phi(a_\tau | s_\tau)$

2) 值模型:  $V_\psi(s_\tau) \approx \mathbb{E}_{q_\phi(a_\tau | s_\tau)} \left[ \sum_{\tau=t}^{t+H} \gamma^{\tau-t} r_\tau \right]$

其中,  $\phi$  和  $\psi$  分别为动作模型和值模型的网络参数. 两者以策略迭代方式协同训练: 动作模型旨在最大化值估计; 值模型则拟合动作模型变化后的值估计. 值估计采用  $V_\lambda(s_\tau)$  来实现偏差与方差之间的权衡,  $\lambda$  为权衡因子. 动作通过 tanh 变换的高斯分布重参数化采样得到:  $a_\tau = \tanh(\mu_\phi(s_\tau) + \sigma_\phi(s_\tau)\varepsilon)$ . 其中,  $\mu$  和  $\sigma$  分别为动作模型输出的均值和标准差;  $\varepsilon$  从标准正态分布中采样.

动作模型  $q_\phi(a_\tau | s_\tau)$  的目标是预测能够产生高价值状态轨迹的动作, 值模型  $V_\psi(s_\tau)$  的目标是回归值估计, 其目标函数分别为<sup>[32]</sup>:

$$\max_{\phi} \mathbb{E}_{q_\theta, q_\phi} \left[ \sum_{\tau=t}^{t+H} V_\lambda(s_\tau) \right] \quad (3)$$

$$\min_{\psi} \mathbb{E}_{q_\theta, q_\phi} \left[ \sum_{\tau=t}^{t+H} \frac{1}{2} \|V_\psi(s_\tau) - V_\lambda(s_\tau)\|_2^2 \right] \quad (4)$$

Dreamer 通过在学习的紧凑潜在空间中传播解析梯度, 高效地学习行为策略, 在数据效率、计算时间与最终性能上均超越了现有基于模型和无模型的方法.

### 2.2.1 基于世界模型的视觉强化学习

世界模型通常是指智能体通过内部学习和推理构建的环境模型. 该模型通常在智能体与环境交互的过程中逐步形成, 使智能体能够在不直接观察环境变化的情况下推断动作对未来状态的影响. 在视觉强化学习中, 世界模型可以在基于学习到的像素级信息或者更高层次的潜在表征上进行未来状态的预测. 先前在图像编码领域的成功案例表明<sup>[33]</sup>: 生成式模型能够有效融合合理随机噪声到潜在空间中, 同时自动学习高维数据的低维表征. 因此, 可以考虑在视觉强化学习中利用生成式模型将高维视觉观测映射到一个潜在状态空间, 并构建递归潜在动力学模型.

Zhang 等<sup>[34]</sup>将 Transformer 强大的序列建模和生成能力与变分自编码器 (variational autoencoder, VAE) 的随机性相结合, 为视觉强化学习智能体构建一种基于高效随机 Transformer 的世界模型 (stochastic Transformer-based world model, STORM). 尽管 STORM 能够捕捉环境中的局部动力学特性, 但由于其仅对状态转移进行建模而不显式地对环境情境进行感知, 因而限制了世界模型的动力学泛化性能.

针对 STORM 在情境感知方面的不足, Wang 等<sup>[35]</sup>提出基于原型情境感知动力学模型的视觉强化学习方法 (prototypical context-aware dynamics, ProtoCAD). 主要思路为: 利用 VAE 建模递归状态空间模型并引入时序一致的原型正则器, 激励相同潜在空间轨迹的不同时间序列所对应的原型在时间维度上相匹配. 通过将潜在状态的投影与聚类得到的原型相结合, ProtoCAD 显著增强了世界模型的泛化能力.

除了情境感知的缺失, 另一个影响泛化能力的关键因素是状态表征对任务相关信息的区分度. Poudel 等<sup>[36]</sup>指出: VAE 学到的状态表征无法有效区分任务相关状态和任务无关状态, 因此所学策略仍有可能过拟合于特定环境的特征. 基于这一认知, Poudel 等<sup>[36]</sup>提出一种用于生成世界模型的正则化对比表征学习方法 (regularized contrastive representation learning, ReCoRe). 主要包括四个部分: 不变表征学习模块, 利用数据增强和对比学习提取具有不变性的状态表征; 干预不变正则器, 用于防止对比学习过程中的特征坍塌; 潜在动力学模型, 用于预测随机潜在先验状态及其奖励; Actor-Critic 机制, 致力于最大化期望回报. iGibson 基准测试环境上的实验结果表明: ReCoRe 在处理分布外数据和实现模拟到现实世界的泛化方面展现出较强的稳健性. 但是, 上述 STORM<sup>[34]</sup>、ProtoCAD<sup>[35]</sup>和 ReCoRe<sup>[36]</sup>也存在一些局限: 对噪声和虚假信息较为敏感; 解码器的使用增大了计算复杂性, 需要更多的资源且延长了训练时间; 过于专注于复制输入而忽视了高层次的理解, 限制了模型的表达能力.

为克服生成式模型在计算效率与表征抽象性方面的不足, 研究者开始转向直接状态编码方法. 这类模型无需解码器重构输入, 从而更高效地提取有用信息、降低计算成本并缓解过拟合. 作为典型代表, 基于原型的梦想家 (Dreamer with prototypes, DreamerPro)<sup>[37]</sup>融合了原型表征学习与时序动力学学习, 一方面从世界模型想象推理得到的模拟轨迹

中提取原型, 与此同时还整合了历史观测状态与动作的时序信息. DMControl 任务上的实验结果表明: DreamerPro 能够有效减少视觉干扰对潜在动力学模型构建的影响, 进而提升强化学习智能体的稳健性.

近年来, 来自动态系统理论的 Koopman 算子<sup>[38]</sup>为非线性系统建模提供了新范式, 其能够从数据中学习出全局线性的状态空间表示, 从而实现对非线性动力系统的分析、预测与控制. 受此启发, Kumawat 等<sup>[39]</sup>开发一种对比谱 Koopman 编码器, 将高维视觉输入映射为复数形式的任务嵌入空间, 实现视觉表征的线性化. 进一步地, 他们将世界模型预测作为辅助任务, 结合强化学习联合优化嵌入空间、Koopman 算子及线性控制器, 提出适用于下游任务自适应调优的 RoboKoop 方法<sup>[39]</sup>.

为进一步整合基于模型的规划与无模型学习的优势, Hansen 等<sup>[40]</sup>提出一种结合时序差分学习和模型预测控制的强化学习方法 (temporal difference learning for model predictive control, TD-MPC). 主要思路为: 首先, 通过直接编码状态表征损失来学习任务导向的潜在动力学模型; 然后, 利用模型预测控制在已学到的环境模型中进行多步预测和优化; 最后, 结合时序差分学习来更新模型和策略.

尽管 TD-MPC 在单一任务中表现良好, 但其泛化至多任务的能力仍显不足. 为此, Hansen 等<sup>[41]</sup>进一步提出 TD-MPC2, 通过引入可学习的固定维度任务嵌入空间, 构建能够处理大规模多任务的世界模型. 该方法结合联合嵌入预测、奖励预测和时序差分学习, 通过交互数据隐式学习以控制为中心的世界模型. TD-MPC2 能够直接从混合质量的多任务数据中学习通用表征, 对任务差异具有强稳健性, 且无需针对不同任务调整超参数, 在机器人控制、自动驾驶和游戏等多个领域均取得优异表现.

当前主流世界模型多依赖于离散潜在变量, 可能导致视觉细节丢失. 扩散模型在图像生成中的成功为高保真环境建模提供了新思路. 文献 [42] 提出基于扩散模型的环境建模方法 Diamond (diffusion as a model of environment dreams), 用反向去噪过程刻画环境动力学, 在提升样本效率的同时保留丰富的视觉细节. 该方法采用 EDM (elucidating diffusion models) 框架, 通过特定的噪声调度与网络预处理, 仅需少量采样步骤即可稳定生成高质量的状态预测. 实验表明, Diamond 在 26 个 Atari 100k 离散控制任务中的平均人类归一化得分达 1.46, 优于所有其他在世界模型中训练的智能体, 但

其在连续控制任务中的应用仍需进一步探索。

视觉干扰是影响现实世界强化学习应用的另一关键挑战。Zhou 等<sup>[43]</sup>提出自洽模型适应 (self-consistent model-based adaptation, SCMA) 算法, 通过无监督分布匹配优化去噪模型, 将含噪观测状态转换为干净观测样本, 从而在不改变策略的情况下提升智能体的稳健性。实验表明, SCMA 在多种干扰环境和真实机器人数据上的表现均优于基线方法, 但其需要同时维护世界模型和去噪模型的特点也带来了较高的内存开销。

尽管上述方法在不同方面取得了进展, 它们通常仍需针对具体任务进行精细调参。2025 年, Hafner 等<sup>[44]</sup>在 *Nature* 发表的 DreamerV3 算法通过多项技术创新实现了广泛的任务泛化能力。该算法采用自编码器学习状态表征, 结合递归状态空间模型进行状态与奖励预测, 并整合 Symlog-Symexp 变换、KL 平衡、百分位回报归一化和双热损失四项技术以确保训练稳定性。作为 Dreamer 系列的第三代, DreamerV3 在 8 个领域的 150 多项任务中全面领先, 更在 Minecraft 中仅凭稀疏奖励实现从零获取钻石的里程碑突破, 显著降低了强化学习在新任务中的应用门槛。

此外, 现实系统中的反馈延迟也对传统强化学习构成严峻挑战。文献<sup>[45]</sup>借助世界模型提出三种应对策略: Extended Actor 通过扩展状态空间 (包含延迟状态与历史动作序列) 直接处理延迟; Memoryless Actor 仅基于当前延迟状态进行决策, 依赖网络隐式记忆动作历史; Latent Actor 则通过世界模型预测当前潜在状态再做出决策。该研究基于 DreamerV3<sup>[44]</sup> 框架实现上述方法, 并在延迟部分可观测马尔科夫决策过程环境中进行理论分析与实验验证。该工作首次在视觉延迟环境中系统验证了方法的有效性, 填补了相关领域的研究空白, 为实际应用提供了可行的技术路径。

在安全无关的任务中, 上述方法通过最大化累积奖励已取得显著成功。然而在自动驾驶、医疗等高风险领域中, 安全性至关重要<sup>[46]</sup>。As 等<sup>[47]</sup>探索从高度部分可观测环境 (即像素级) 中学习安全策略。为此, 他们提出拉格朗日引导的基于模型的智能体 (LAMBDA), 利用由贝叶斯世界模型生成的虚拟轨迹来近似真实动力学模型。通过引入不确定性, LAMBDA 能够估计奖励函数的乐观上限和安全约束的悲观上限。此外, 还采用增广拉格朗日乘子和近端松弛机制来解决约束马尔科夫决策过程问题。实验结果表明, LAMBDA 在训练过程中能够有效保持安全约束。

类似地, Hogewind 等<sup>[48]</sup>在 SLAC (一种基于表征学习的强化学习方法)<sup>[49]</sup>基础上施加安全约束, 以促进从像素中学习安全策略。所提方法确保了状态表征能够捕捉与安全相关的信息。为优化这个带有安全约束的目标函数, 采用拉格朗日松弛技术来识别与最大化问题可行解对应的鞍点。与 SLAC 相似, 该变分推断模型仅能捕捉输入变量与潜在变量之间的依赖性。然而, 它缺乏捕捉状态之间相互依赖关系的能力, 这一限制可能阻碍其理解数据内部复杂关系。

综上所述, 基于世界模型的视觉强化学习方法能够在较少的交互样本下学到有效的决策策略, 并允许智能体在安全的训练环境中进行学习和优化, 从而减少潜在的风险和不安全因素。然而, 这类方法也面临世界模型预测偏差可能导致的策略退化问题, 其训练过程往往复杂且难以稳定, 对建模精度与计算资源的要求较高, 同时潜在误差在长期迭代中可能累积扩散, 进而影响策略的真实环境适应性。

## 2.2.2 大模型增强的视觉强化学习

大模型是指使用大规模数据和强大的计算能力训练出来的规模庞大的深度学习模型。大模型通过其庞大的参数量、深层次的网络结构和广泛的预训练能力, 能够捕捉复杂的数据模式, 具有强大的表征学习能力和高度的通用性。随着计算机硬件性能的不不断提升以及深度学习算法的快速优化, 大模型的发展日新月异, 已经在自然语言处理、图像识别、语音识别等领域得到成功应用, 可分为大语言模型、视觉大模型、多模态大模型 (如视觉-语言模型) 等多种类型。将大模型技术应用于视觉强化学习中, 能够提升智能体对环境的理解能力, 进而增强其在复杂任务中的适应性和决策能力。

### 1) 视觉表征学习

视觉表征学习是指利用大规模预训练模型 (如对比语言-图像预训练 (contrastive language-image pre-training, CLIP)<sup>[50-51]</sup> 和 Flamingo<sup>[52]</sup>) 从高维的原始像素观测中提取具有高层语义和任务相关性的低维特征表示, 以替代传统的手工设计特征或端到端训练的编码器, 从而帮助智能体更好地理解环境状态、提升样本效率并缓解维度灾难问题。

大语言模型 (large language model, LLM) 可以用于改进视觉强化学习中的语义理解和生成, 为智能体提供更准确的任务指导或生成自然语言描述。为将语言理解与未来预测统一为一个自监督的学习框架, Lin 等<sup>[53]</sup>设计一个学习多模态世界模型的智能体 Dynalang。Dynalang 整合视觉环境中的语言输入, 通过编码多模态数据构建状态表征, 并

从模拟轨迹中学习动作,以预测未来的文本和图像表征序列。HomeGrid、Messenger 等环境上的实验结果表明, Dynalang 通过理解环境描述、游戏规则和指令,显著提高了语言处理能力,能够有效完成指定任务。

继 Dynalang 之后,研究者进一步探索语言模型在多任务场景下的潜力。为有效处理多任务问题, Aljalbout 等<sup>[54]</sup>提出基于语言的多任务视觉世界模型训练方法 (language-informed multi-task visual world models, LIMT): 利用预训练的语言模型将任务指令映射到一个语义丰富的隐空间中,形成任务表征;随后,这些表征被世界模型和策略所利用,以推断任务间的动力学与行为相似性。通过使用语言驱动的任务表征, LIMT 可以显著提升世界模型和基于模型的多任务学习的效果。然而, LIMT 依赖于预先计算的语言嵌入作为任务表征,并将其作为策略的条件。对于具有相似文本描述但动力学特性不同的任务而言,这些语言嵌入可能会产生混淆,导致策略判断上的冲突,从而影响整体性能。

除了多任务学习,大语言模型也被用于缓解强化学习中长视界与稀疏奖励带来的挑战。在上述研究基础上, Liu 等<sup>[55]</sup>提出大语言模型增强的梦想家 (dreaming with LLM, DLLM)。DLLM 利用自然语言描述环境动力学,并将大语言模型的指导融入模拟轨迹的构建中,以提高智能体的探索效率和目标达成能力。同时, DLLM 还利用自动衰减机制生成内在奖励,从而促进策略优化。值得注意的是,由于 Dynalang<sup>[53]</sup>、LIMT<sup>[54]</sup>和 DLLM<sup>[55]</sup>均依赖于大语言模型提供的指导,因而这些方法的性能容易受到大语言模型输出的固有不稳定性的影响。

与大语言模型相辅相成,视觉大模型 (large visual model, LVM) 能够从原始图像中提取具有语义意义的对象和区域信息,帮助智能体更精准地识别和理解环境中的关键元素,从而提升决策质量。在这一方向上, Chen 等<sup>[56]</sup>通过使用基础分割模型来区分任务相关与任务无关的视觉状态,提出一种分段辅助视觉强化学习方法: 先聚焦再决策 (focus-then-decide, FTD)。FTD 的主要思路为: 首先使用基础分割模型将观测状态划分为不同对象集,然后引入注意力选择模块来评估每个对象的重要性,并将相关信息整合进决策模型训练中,从而有效识别任务相关对象。然而, FTD 的性能在很大程度上取决于基础分割模型的分割准确性。由于分割的速度和准确性往往难以兼顾,因此探索更适合强化学习下游任务的分割模型显得尤为重要。其次, FTD 假设奖励信号是事先已知的,但在实际应用中,奖励函数的设计往往受到许多因素的影响。

尽管上述方法在各自任务中取得了一定成效,但是它们在训练数据有限时仍难以有效处理未见指令。因此, Jiang 等<sup>[57]</sup>提出 CLIP 引导的目标定位策略学习 (CLIP-guided object-grounded policy learning, COPL), 利用视觉-语言模型 (VLM) 的视觉定位能力,将视觉-语言知识迁移到强化学习中,以提升智能体的泛化能力。具体而言,通过 Mine-CLIP 对指令中的目标对象进行视觉定位,生成置信度图,并设计基于该图的内在奖励函数,引导智能体瞄准目标。同时,将置信度图作为策略网络输入,使智能体通过可视化表征理解新指令,实现零样本泛化。在 Minecraft 实验中, COPL 在狩猎和采集任务上显著优于语言条件强化学习方法,但对挖洞、建造房屋等复杂规划任务泛化能力有限。

除了泛化能力,现有视觉-语言模型还存在另一明显局限: 它们大多基于静态图像-文本对训练,难以满足机器人任务对视频理解与闭环控制的需求,且语言输出与动作空间之间存在表征差异。针对这一问题, Li 等<sup>[58]</sup>提出 RoboFlamingo 框架,将视觉语言理解与决策过程解耦: 利用预训练 VLM 处理单步观测图像与语言指令,通过显式策略头建模历史信息,仅需少量真实机器人演示即可微调适配下游任务。实验表明,该方法在 CALVIN 基准上显著优于现有技术,尤其在零样本泛化方面表现突出,验证了预训练模型在数据效率与泛化能力上的优势。但其仍局限于仿真环境,缺乏真实机器人验证,且对开环控制、端到端微调敏感,对语言同义词的稳健性也较弱。

总之,视觉表征学习类方法能够克服传统卷积编码器表征能力有限的问题,通过预训练大模型所提供的强大先验知识,显著提升智能体对高维视觉输入的理解深度与泛化能力。然而,预训练模型往往依赖于静态数据集,与强化学习动态环境存在领域差异,可能导致表征与当前任务无关的冗余信息引入,且模型计算开销较大,对硬件资源要求较高,在实时控制场景中可能带来延迟问题。

## 2) 奖励生成

奖励生成是指利用大模型的先验知识,通过理解指令、分析环境或预测目标来合成奖励函数,以解决传统强化学习中奖励设计困难、稀疏或难以对齐人类意图的问题,从而更高效地引导智能体学习目标策略。在视觉强化学习中,奖励设计对智能体行为至关重要<sup>[59]</sup>,但手动设计奖励函数往往不现实,从人类反馈中学习又成本高昂。因此,借助预训练基础模型,通过自然语言提供辅助奖励信号,成为一种高效且自然的解决方案,可显著降低奖励设计成本并提升其实用性。

如何设计合理的奖励信号以有效引导智能体学习复杂行为一直是强化学习面临的一大挑战, 一个比较有前景的解决方案是利用互联网上丰富的未标记视频资源来提取行为偏好. 基于此, Escontrela 等<sup>[60]</sup> 提出一种视频预测奖励方法 (video prediction rewards, VIPER), 通过训练自回归 Transformer 基础模型来预测视频中的行为, 并将视频预测的似然作为奖励信号; 与此同时采用熵最大化目标, 鼓励智能体匹配视频模型的轨迹分布. VIPER 不需要任何真实值奖励或动作标注, 只需提供智能体行为的相关视频即可. Atari 游戏和 MuJoCo 机器人控制任务上的实验结果表明, VIPER 能够在没有人工手动设计任务奖励的情况下实现专家水平的控制. 然而, VIPER 仍需使用专家领域内数据进行模型训练, 这在现实世界中并不容易获得.

尽管 VIPER<sup>[60]</sup> 在视频驱动的奖励生成方面取得进展, 但在处理复杂长时程任务时仍面临挑战. 现有视觉指令校正方法在长时程操作任务中存在三方面局限: 缺乏阶段划分意识、难以适应任务复杂度变化、缺少目标状态显式估计, 易导致奖励信号偏差. 为此, Hung 等<sup>[61]</sup> 提出视觉指令校正奖励 (vision-instruction correlation reward, VICtoR) 模型, 通过任务知识生成器、阶段检测器和运动进度评估器, 将复杂任务分解为阶段、运动和进度三个层次, 实现细粒度奖励信号生成. 实验表明, VICtoR 能够为长时程任务生成有效的密集奖励信号, 并在真实视频中准确识别任务进度, 但其对未见运动类型缺乏零样本泛化能力.

与 VICtoR 这类需要任务分解的方法不同, 另一种思路是直接利用大规模预训练模型作为奖励来源. 为避免手动设计奖励函数或额外收集昂贵的数据来学习奖励模型, Rocamonde 等<sup>[62]</sup> 提出一种用预训练的视觉-语言模型 (如 CLIP) 作为强化学习任务的零样本奖励模型 (using VLMs as reward models, VLM-RM), 通过计算状态表征与自然语言指令之间的余弦相似度来定义奖励函数. 基于 VLM-RM, Rocamonde 等<sup>[62]</sup> 训练一个人形机器人学习复杂的任务, 如举起手臂、做劈叉和跪下等. 但是, 不同 (简单或详尽) 的语言提示可能会使 VLM-RM 产生不同的奖励值, 进而影响强化学习方法的性能.

尽管 VLM-RM 能够实现零样本奖励生成, 但现有方法中, 基于大型语言模型的奖励生成依赖环境源代码或底层状态信息, 而基于视觉-语言模型直接输出奖励分数的方法则存在噪声和不一致性问题. 为此, RL-VLM-F<sup>[63]</sup> 利用视觉-语言模型对智能

体的视觉观测图像进行偏好比较, 并基于比较结果自动生成奖励函数. 这种方法既避免了直接输出带来的噪声问题, 又无需依赖底层状态信息, 适用于复杂操作任务. 实验显示, 该方法在物体操作任务中表现优于基线方法, 在部分任务上甚至超越人工设计的奖励函数. 但需注意的是, 该方法依赖于预训练的视觉-语言模型, 可能继承模型偏见, 且在长时程任务中的性能仍有待提升.

针对上述 VLM 奖励中普遍存在的模糊性和不一致性问题, Fu 等<sup>[64]</sup> 重点探讨 VLM 中存在的模糊奖励问题, 即, 在某些情况下基于 VLM 构建的奖励函数虽然有意义, 但因自身的不精确性也可能会产生误导性. 为此, Fu 等<sup>[64]</sup> 提出一种轻量级微调方法——模糊 VLM 奖励辅助强化学习 (fuzzy VLM reward-aided RL, FuRL): 通过对 VLM 生成的嵌入进行微调以实现奖励信号的对齐, 进而促进环境探索与策略学习的协同; 与此同时, 引入中继强化学习机制来帮助智能体在探索过程中避免陷入局部最优解. Meta-World 环境<sup>[65]</sup> 中的实验结果显示: 在稀疏奖励任务上, FuRL 能够显著提升柔性 Actor-Critic 和 DrQ 这两种基准强化学习方法的表现.

除了在奖励函数设计上的创新, 另一类方法从模型结构入手提升视觉强化学习性能. 基于模型的视觉强化学习方法 (如 DreamerV3) 利用世界模型提高样本效率, 但其依赖短时想象的特性导致探索能力受限. 为此, Li 等<sup>[66]</sup> 提出一种名为 LS-Imagine 的长短期世界模型解决方案: 首先通过虚拟探索和 U-Net 模块生成可操作性地图, 并设计相应内在奖励机制引导智能体关注任务相关目标; 其次开发了支持短时单步与长时跳跃想象模式的自适应世界模型, 实现潜在空间中的多步收益预测. MineDojo 平台实验表明, 该方法在成功率和任务效率上均超越现有技术, 但仍面临模型训练复杂度高和跨任务泛化性不足的挑战.

上述奖励生成方法借助大模型的外部先验知识, 自动生成或辅助设计奖励函数, 缓解了手工设计奖励的繁琐性和主观性问题. 然而, 生成奖励的可靠性受模型幻觉问题影响, 可能产生错误或模糊的奖励信号, 进而误导策略优化.

### 2.3 任务辅助型视觉强化学习

任务辅助型视觉强化学习则通过引入额外的辅助任务或多视角信息, 引导智能体更有效地利用视觉信息 (如图 5 所示). 该方法主要分为两种: 一是辅助任务引导的视觉强化学习, 设计与强化学习主任务相关的辅助视觉任务, 促进表征学习; 二是

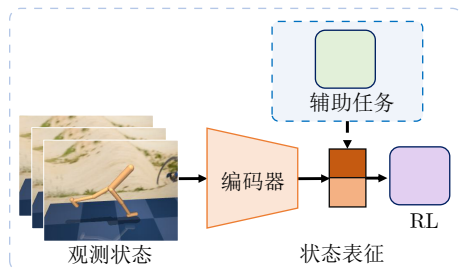


图 5 任务辅助型视觉强化学习示意图

Fig.5 The diagram of task-assisted VRL

多视角视觉强化学习, 通过整合不同视角的信息, 增强智能体对复杂环境的理解能力. 图 6 展示了任务辅助型视觉强化学习方法的时间线.

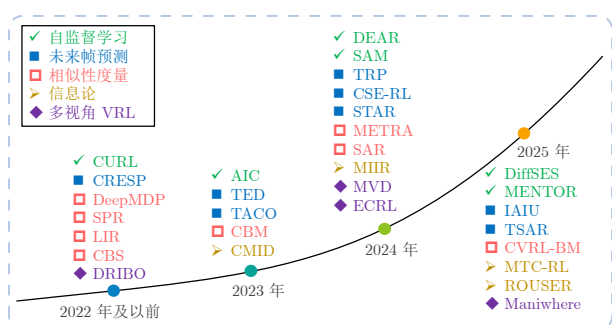


图 6 任务辅助型视觉强化学习时间线

Fig.6 The timeline of task-assisted VRL

以下是一种代表性的任务辅助型视觉强化学习方法. 为提升视觉强化学习方法在下游决策任务中的计算效率和泛化性能, Laskin 等<sup>[67]</sup> 创新性地引入对比表征学习作为强化学习的辅助任务, 提出基于对比无监督表征的强化学习方法 (contrastive unsupervised representations for RL, CURL)<sup>[67]</sup>. 其主要思路为: 通过提升同一观测不同增强样本间的互信息并降低不同观测增强样本间的互信息, 从而促使模型捕捉状态间的关键差异. 该方法通常涉及到两个核心网络组件: 一是用于处理锚点样本的在线编码器; 二是针对目标样本的目标编码器. 具体流程为: 从经验池中随机采样一批观测样本, 选择其中一个样本作为锚点样本  $o_q$ , 并以其增强样本作为正样本  $o_p$ , 而负样本  $o_n$  则来自同一批次中的其他样本, 将正样本与负样本统称为目标样本. 锚点样本与目标样本之间的嵌入相似性可通过双线性映射  $q^T W k$  来度量, 其中  $q$  代表锚点样本经过在线编码器得到的输出;  $k$  表示每个目标样本通过目标编码器得到的输出;  $W$  为随机初始化的权重矩阵. 为建模上述相似性关系, CURL<sup>[67]</sup> 的损失函数定义如下:

$$L_{\text{CURL}} = -\mathbb{E}_{(q, k_+, \{k_i\})} \left[ \log \frac{\exp(q^T W k_+)}{\exp(q^T W k_+) + \sum_{i=0}^{K-1} \exp(q^T W k_i)} \right] \quad (5)$$

其中,  $k_+$  和  $k_i$  分别表示正样本  $o_p$  和第  $i$  个负样本  $o_n$  经过目标编码器处理后的输出;  $K$  表示负样本的数量.

尽管 CURL<sup>[67]</sup> 在 DMControl 和 Atari 上的连续和离散视觉控制任务中展示了良好的结果, 但随后的研究表明, CURL 的优势大多数可以归因于状态增强, 而非辅助对比损失<sup>[68]</sup>.

### 2.3.1 辅助任务引导的视觉强化学习

辅助任务引导的视觉强化学习方法的主要思路为: 通过设置表征学习辅助任务为编码器提供额外的监督信息, 使其能够学习到更为丰富有效的状态表征, 进而提升模型的适应性和稳健性. 常见的辅助任务构造方式包括自监督学习、未来帧预测以及相似性度量.

#### 1) 自监督学习

除 CURL 外, 在表征学习的另一类思路, Pore 等<sup>[69]</sup> 着眼于环境和智能体的解纠缠表征学习. 他们利用特征分离约束, 分别学习环境与智能体的解耦表征, 并将这些表征作为辅助自监督损失与强化学习目标相结合, 提出一种基于解纠缠表征的视觉强化学习方法, 为后续结构化表征研究提供了思路.

然而, 传统深度强化学习方法仍面临策略可解释性差、计算复杂度高问题, 而符号强化学习方法在高维视觉场景中又存在可扩展性不足和依赖人工先验的局限. 针对这一挑战, Zheng 等<sup>[70]</sup> 提出可微符号表达式搜索 (differentiable symbolic expression search, DiffSES) 方法. DiffSES 通过自监督对象检测模块将高维图像降维为对象级特征 (如位置、速度), 并结合部分可微的遗传编程与神经网络引导机制, 搜索简洁的离散符号策略, 提升学习效率与可扩展性. 实验结果显示, DiffSES 生成的符号策略较现有最优方法更简洁、可扩展性更强, 且降低了对符号先验的依赖. 但该方法仍存在一定局限: 复杂场景下对象检测的分割误差可能影响性能, 且符号策略的完全可解释性有待进一步提升.

除了可解释性, 训练稳定性也是视觉强化学习中的重要问题. Zhai 等<sup>[71]</sup> 指出, 大型编码器在拟合敏感的时序差分目标时容易出现参数振荡和表征漂移 (即“振荡自过拟合”现象). 为此, 他们提出非对

称交互合作 (asymmetric interactive cooperation, AIC) 框架, 通过一个大型可优化编码器与一个辅助轻型编码器之间的协同交互, 在保持表征判别力的同时显著提高训练稳定性. AIC 还采用贪婪自举优化策略, 交替优化表征学习和策略学习以强化整体训练过程. 在 CARLA 自动驾驶和 Vizdoom 环境上的实验表明, AIC 性能显著优于 CURL、SVEA 等基线方法, 并能稳定训练如 ResNet 等大型模型. 但该方法的双编码器结构也带来了额外的计算负担, 其拓扑蒸馏所采用的线性变换也可能限制知识迁移的灵活性.

值得注意的是, 上述方法大多集中于状态表征的优化, 而忽视了动作空间中所蕴含的丰富语义信息. 针对这一局限, 刘宇昕等<sup>[72]</sup>通过构建基于负余弦相似性的状态序列和动作序列来预测辅助损失并对其进行联合掩码重构, 提出一种基于状态-动作联合掩码的自监督强化学习方法 (state-action joint mask-based self-supervised learning, SAM). 尽管 SAM 能够挖掘出与任务密切相关的表征信息, 但在处理具有视觉干扰的任务时其性能仍受限.

另一方面, 网络结构的设计也直接影响强化学习的效率与性能. 现有方法通常采用多层感知机作为策略主干, 但容易引发梯度冲突, 导致训练低效, 且随机扰动难以帮助模型逃离局部最优. 为此, MENTOR 方法<sup>[73]</sup>引入两项关键改进: 首先, 采用混合专家 (mixture-of-experts, MoE) 架构替代传统的多层感知机作为策略主干, 通过动态路由机制将不同任务的梯度分配给专用专家, 以缓解梯度冲突; 其次, 引入任务导向的扰动机制, 从历史高性能智能体中采样扰动候选, 取代随机噪声, 提升优化的针对性. 在 DMControl 任务中, MENTOR 在 Dog Stand 和 Dog Walk 任务上的回合奖励分别提升约 17% 和 10%; 在 Meta-World 的 Hammer 任务中, MENTOR 仅用 30% 的训练帧数就取得显著性能提升; 在 Adroit 任务中, MENTOR 几乎达到 100% 成功率, 且训练时间大幅缩短. 然而受限于 MoE 的基础实现方式, 其计算开销仍然较大.

但是, 这类基于自监督学习辅助任务的视觉强化学习方法忽略了强化学习场景中的一个重要特性: 状态之间存在的时间连续性和关联性<sup>[74]</sup>. 这意味着在实际应用中, 仅仅依靠空间层面的数据增强与状态表征可能不足以全面反映真实世界的动态变化规律. 因此, 如何有效结合时间维度的信息成为当前视觉强化学习研究的一个关键挑战.

## 2) 未来帧预测

为应对上述挑战, 一些研究者通过构造未来帧

预测辅助损失来学习未来的状态表征, 从而捕捉数据的内在结构特征.

给定起始状态观测和预定义的后续动作序列, CRESP (characteristic reward sequence prediction)<sup>[75]</sup>通过预测奖励序列分布的特征函数来提取与任务相关的表示. 实验证明了特征函数在确定奖励序列分布方面的有效性, 这使得 CRESP 能够从奖励和状态观测转移中提取与长期任务相关的信息, 而不会受到无关视觉干扰的影响.

在理论分析方面, Wang 等<sup>[76]</sup>首次对视觉强化学习的泛化性能进行系统研究, 并确立泛化目标的上界. 该上界包含两个关键组成部分: 一是策略散度, 它衡量在不同环境中执行策略时状态-动作分布的差异性; 二是贝尔曼误差, 它反映训练策略与最优策略之间的差异. 基于这一理论框架, 研究中引入一个名为 TRP (truncated return prediction) 的辅助任务, 通过预测历史轨迹的截断回报来确保不同领域之间策略的一致性. 实验结果显示, 在 DMControl-GB 连续控制视觉任务及机器人操作任务中, TRP 有效地提升了训练过程中的样本效率.

为进一步增强强化学习的泛化能力, Dunion 等<sup>[77]</sup>提出一种名为时序解耦 (temporal disentanglement, TED) 的未来帧预测辅助任务. 该算法利用强化学习中观测状态的时序连续性, 从连续时间步中提取时序数据, 从而在线学习出一种具备稳健性的图像编码解耦表征. 由于 TED 不依赖于解码器, 因此其计算成本相对较低. 此外, 通过使用解耦表征, TED 能够优化学习策略, 使其在面对未曾见过的任务时表现更加出色.

针对传统强化学习算法在机器人连续控制任务中易受视觉干扰的问题, 文献<sup>[78]</sup>提出一种基于聚类驱动的状态嵌入强化学习方法 (clustering-driven state embedding for RL, CSE-RL)<sup>[78]</sup>. 主要思路为: 首先, 引入基于对比学习的未来帧预测辅助损失, 根据历史观测状态与动作来预测未来状态表征, 促进学习过程; 然后, 利用学生 t 分布来建模预测的聚类分配, 并通过强化目标分布来定义奖励中心, 进而将状态表征、奖励及其对应中心输入到预测的聚类分配函数中进行聚类. CSE-RL 不仅能够加深智能体对于不同状态之间内在关系的理解, 而且能够有效过滤掉无关或冗余的信息, 从而构建出更加高效且具有更强抗视觉干扰能力的状态表征.

上述方法在从复杂高维视觉输入中直接学习行为方面已展现出显著能力, 但仍面临表征坍塌 (包括完全坍塌与维度坍塌) 以及维度冗余等挑战. 尽管对比学习在一定程度上缓解了完全坍塌问题, 但

其容易陷入类冲突困境. 针对上述问题, Wang 等<sup>[79]</sup>提出基于实例重加权对齐性与实例维度均匀性的视觉强化学习算法 (instance-reweighted alignment and instance-dimension uniformity, IAIU). 该算法引入实例重加权对齐表征学习机制, 通过最小化预测下一状态表征分布与其加权实际对应项之间的 KL 散度, 有效实现了相同语义类中状态表征的对齐, 从而显著缓解了类冲突问题. 同时, 采用实例维度均匀正则化机制, 通过在实例和维度层面分别运用 Hilbert-Schmidt 独立性准则和标准正交约束, 有效抑制了表征坍塌现象, 确保了任务相关状态表征的有效提取. IAIU 算法通过对齐性和均匀性双重机制, 不仅成功解决了类冲突这一关键问题, 还保证了实例和维度层面的均匀性, 为视觉强化学习处理表征坍塌问题提供了新的解决方案.

尽管上述研究在状态表征方面取得了显著进展, 但它们主要聚焦于视觉强化学习的状态表征, 却较少关注连续控制任务中动作表征的关键作用. 通过在潜在动作空间内将语义上相似的动作进行分组, 智能体能够更好地在各种状态-动作对之间实现知识泛化, 进而提升强化学习算法的样本效率. 为此, Zheng 等<sup>[80]</sup>提出时间动作驱动的对比学习 (temporal action-driven contrastive learning, TACO), 旨在帮助智能体同时获得潜在状态和动作表征. TACO 通过优化当前状态及其对应动作序列的表征与相应未来状态表征之间的互信息, 实现状态与动作表征的同步学习. 理论上, TACO 学到的状态-动作表征包含足够的控制信息, 适合处理高维连续控制任务. 然而, 由于所采用的对比学习框架的固有特性, 当面对大批量数据时, 计算效率可能会受到影响.

除了动作表征的引入, 已有方法多聚焦于从观测图像中提取静态状态表征, 忽视状态-动作交互中蕴含的环境动力学信息, 这也制约了样本效率的进一步提升. 为此, Yan 等<sup>[81]</sup>提出状态-动作表征学习 (state-action representation learning, STAR) 框架. STAR 在 DrQv2<sup>[17]</sup> 基础上引入状态-动作联合编码器, 通过潜在动力学预测、对比学习和奖励预测等多任务损失进行联合训练, 并将学习到的联合表征与状态表征、动作一同输入 Q 网络, 以增强值函数估计. 他们还从理论上证明了该方法不影响值函数的收敛性. 在 DMControl 上的实验表明, STAR 在多数任务中均展现出显著的样本效率优势. 以具有挑战性的“Finger Turn Hard”任务为例, STAR 仅使用 100 万训练步数即达到最优回报的 80%, 效率为 DrQv2 的 3.5 倍. 然而, 其多任务损失

的超参数仍需人工调优, 限制了在特定任务上的泛化能力.

为更全面地融合状态、动作与奖励信息, 刘民颂等<sup>[82]</sup>进一步提出基于 Transformer 的状态-动作-奖赏预测表征学习框架 (Transformer-based state-action-reward prediction representation learning, TSAR), 通过基于对比学习的掩码序列状态预测学习、基于均方误差的动作预测学习和奖赏预测学习三个辅助任务共同作用以学习状态与动作表征. 通过同时将状态表征和动作表征显式地纳入强化学习策略的优化中, TSAR 显著提高了表征对策略学习的促进作用. 然而, 上述方法大多采用对比学习或二值交叉熵来构建未来帧预测辅助任务, 因此这类方法可能因过于专注于区分性特征而忽略了一些细微但重要的细节, 且存在表征坍塌的风险.

### 3) 相似性度量

另一种简单有效的方法是借助于相似性度量 (如 L1 范数、L2 范数、负余弦相似性、双模拟度量等) 来为视觉强化学习构造辅助任务<sup>[83]</sup>, 以确保实际观测到的未来状态表征与基于潜在动力学模型预测得到的状态表征之间的高度一致性.

这类方法的典型代表是自预测表征 (self-predictive representations, SPR)<sup>[84]</sup>. Schwarzer 等<sup>[84]</sup>利用数据增强和基于负余弦相似性的距离度量辅助任务来预测未来潜在状态表征, 从而在不同环境观测视图下能够学习到既具有时序预测能力又保持增强一致性的表征.

与 SPR 不同, Park 等<sup>[85]</sup>将关注点转向状态之间的时间距离, 提出一种基于度量感知抽象的强化学习方法 (metric-aware abstraction, METRA). 该方法以两个状态之间的最小环境步数作为潜在空间的度量标准, 通过在紧凑的潜在状态空间中最大化覆盖范围, 促使智能体获得多样化的行为, 从而有效逼近整个状态空间的覆盖. 然而, 由于仅考虑两个状态之间的最小时间距离, METRA 在处理高度不对称的环境时可能显得过于保守. 此外, 该方法仅聚焦于状态层面的表征, 忽略了动作空间中蕴含的丰富语义信息.

为从具有视觉干扰的高维观测中精准提炼出与任务相关的状态表征, Liang 等<sup>[86]</sup>提出序列动作诱导不变表征 (sequential action-induced invariant representation, SAR). 该方法通过最小化预测动作序列与真实动作序列之间的潜在距离, 确保状态编码器的输出与实际控制信号保持一致, 从而过滤任务无关信息, 提升表征的任务相关性. 然而, 综合分析 SPR、METRA 和 SAR 等方法可以发现, 它们

所采用的传统距离度量方式往往过于强调状态或动作空间中点的不重合性,在一定程度上限制了其应用范围和灵活性。

在这一背景下,伪度量<sup>[87]</sup>提供了一种更为灵活的距离定义方式,它允许不同状态之间存在零距离,即将某些不同状态视为等价或高度相似。这一特性在特定任务中有助于提升模型的泛化能力和适应性。作为伪度量的一种,双模拟度量<sup>[88]</sup>通过比较两个状态的系统参数来评估它们在行为层面的相似性,为状态表征的构建提供新的理论依据。

基于双模拟度量的思想,Gelada等<sup>[89]</sup>提出DeepMDP模型,通过预测下一时刻的潜在状态和奖励来优化状态表征,并借助双模拟度量建立理论保障。为缓解潜在空间中的表征坍塌问题,他们引入辅助的重构损失以提升学习性能。然而,重构损失容易受到任务无关信息的干扰,从而影响表征的纯净性。

针对上述问题,Zhang等<sup>[90]</sup>对双模拟度量进行进一步拓展,提出一种无需状态重构的辅助任务,直接学习具备双模拟行为的表征空间。在视觉MuJoCo基准测试中的实验结果表明,该方法在面对任务无关干扰时表现出更强的稳健性,显著优于基于重构或对比损失的现有辅助任务视觉强化学习方法。

然而,Agarwal等<sup>[91]</sup>指出,在双模拟度量中引入奖励机制可能导致约束条件过紧或过松的问题。为此,他们转而从策略相似性的角度提出一种基于行为接近度的策略相似性度量辅助任务。在该框架下,若两个状态对应的最优策略相似,则状态本身也被视为相似。通过将这一度量整合至对比学习框架,能够有效学习对观测变化具有不变性的策略表征,从而增强泛化能力。

在此基础上,Liu等<sup>[92]</sup>提出基于双模拟度量的聚类辅助方法(clustering with bisimulation metrics, CBM),通过评估状态观测值与聚类中心之间的双模拟距离预测聚类分配。CBM通过最小化预测簇分配与目标分布之间的交叉熵损失,联合优化编码器与聚类中心参数,最终学到任务相关的紧凑状态表征。

上述方法并未考虑安全因素。在现实世界的应用中,尤其是在依赖视觉输入的场景下,核心挑战在于如何高效提取安全决策所需的关键特征,同时保持较高的样本效率。针对这一难题,文献<sup>[93]</sup>提出一种基于安全双模拟度量的视觉强化学习方法(constrained visual representation learning with bisimulation metrics, CVRL-BM)。通过构建序列

条件变分推断模型,将高维视觉观测样本压缩为低维状态表征。在此基础上,引入安全双模拟度量机制,用以量化状态间的行为相似性,其核心目标是通过优化使潜在状态表征之间的距离与其对应状态间的安全双模拟度量尽可能接近。通过有机整合这两个关键组件,CVRL-BM不仅能够学习到紧凑且信息丰富的视觉状态表征,还能最大限度降低安全风险。

#### 4) 信息论

基于信息论的视觉强化学习方法通常将互信息、信息瓶颈等信息论工具作为辅助目标或正则项,融入强化学习智能体的训练流程,从而学习具备稳健性和强泛化能力的视觉表征。这类方法能够有效压缩或滤除任务无关的视觉干扰,同时保留并增强与任务相关的信息,进而提升智能体在复杂视觉环境中的决策性能。

在强化学习中,训练数据有限或特征覆盖不足易导致特征间的虚假相关性,降低智能体在环境变化时的泛化能力。现有解耦方法多依赖最小化特征间互信息,但难以处理相关特征。为解决这一问题,文献<sup>[94]</sup>提出基于条件互信息的解耦表征(conditional mutual information for disentanglement, CMID),通过最小化潜在特征间的条件互信息,结合MDP因果图确定通用条件集,以实现给定条件下特征间的条件独立。在DMControl连续控制任务上的实验表明,CMID能有效提升泛化性能,其零样本泛化回报平均提高77%,优于DrQ<sup>[16]</sup>、CURL<sup>[67]</sup>和TED<sup>[77]</sup>等基线方法。但其计算复杂度较高,平均运行时间增加67%,且在复杂因果场景中可能需要更长历史信息作为条件集。

除了解决特征间虚假相关性的挑战,如何系统性地构建具有理论保障的跨域泛化方法也成为研究重点。当前主流方法通常依赖数据增强、注意力机制等经验性技巧,缺乏理论支撑,难以系统提升跨域泛化性能。为应对这一挑战,Wang等<sup>[95]</sup>从信息论角度出发,提出三个互信息项作为学习域不变表征的理论基础,并据此设计基于互信息的不变表征算法(mutual information-based invariant representation, MIIR)。该算法包含三个组件:Q值预测、动力学预测和偏差扰动正则化,分别用于保留回报信息、维持动力学转移一致性以及通过扰动动作无关像素来增强不变性。在DMControl-GB、机器人操纵和CARLA自动驾驶三大基准测试上的实验结果表明,MIIR在未见环境中均取得最优泛化性能和样本效率(例如在DMControl的“video hard”场景中,平均回报提升22%)。但其动作相关像素的

判定依赖频率假设,在复杂场景(如 CARLA 极端天气)下可能因纹理干扰而降低扰动效果。

在关注单一状态表征的同时,研究者也意识到需要在更宏观的轨迹层面提升策略的稳健性和一致性。现有视觉强化学习方法通常仅关注状态或动作的单方面一致性,未能在整体轨迹层面推动智能体学习简洁、一致且易于压缩的行为模式,导致所学策略在状态噪声、动作干扰或系统动力学变化时表现不稳。为此,文献 [96] 提出最大全相关强化学习方法(maximum total correlation reinforcement learning, MTC-RL),通过最大化轨迹中潜在状态表征与动作之间的全相关,引导智能体学习更具一致性和压缩性的轨迹,以增强策略稳健性。该方法基于 SAC 算法框架,联合优化策略、状态表征、动力学模型及动作预测模型,并引入变分下界近似求解全相关优化问题。在 DMControl 基准测试中, MTC-RL 在原始任务上性能提升 10%,且在观测噪声、动作干扰和质量扰动等稳健性测试中均显著优于基线。然而,该方法的全相关变分下界始终为负,无法直接估计真实全相关值。

尽管上述方法在提升表征稳健性方面取得了进展,但它们往往忽略了表征与长期决策任务之间的紧密联系。上述方法虽能抵抗视觉干扰,但未考虑下游的序列决策任务,导致所学表征缺乏长期决策信息,从而影响泛化能力。针对这一问题, Yang 等 [97] 提出基于信息瓶颈的稳健动作值表征学习方法(robust action-value representation learning, ROUSER)。该方法通过最大化表征与动作值之间的互信息来捕捉长期信息,同时最小化其与状态-动作对的互信息以过滤无关特征。由于真实动作值未知, ROUSER 将其拆解为单步奖励与后续稳健表征的递归组合,从而借助已知奖励计算损失。实验显示,在多种存在颜色或视频背景干扰的 DMControl 任务中, ROUSER 显著优于 DrQv2 [17]、CURL [67]、TACO [80] 等方法,其表征能更有效预测长达 300 步的未来奖励序列。然而,该方法仍依赖人工设计的奖励信号,在奖励未知的场景中适用性有限。

辅助任务引导的视觉强化学习方法通过引入额外的表征学习目标,有效提升了智能体从高维像素中提取特征的能力,增强智能体对环境的理解。然而,这类方法也显著增加了计算复杂性和训练难度,辅助任务与强化学习主任务目标可能存在冲突或冗余,若设计不当反而会导致训练不稳定或分散智能体对核心奖励信号的注意力,从而影响最终策略的性能。

### 2.3.2 多视角视觉强化学习

作为任务辅助型视觉强化学习的一个分支,多视角视觉强化学习方法通过从多个不同视角或摄像头获取同一环境的视觉观测,并对这些异构或同构的视觉输入进行融合与对齐,学习出更为全面且具备视角不变性的环境状态表征,从而辅助强化学习策略的训练。该方法的核心优势在于:能够提取更丰富、更稳健的环境状态表征,通过多源信息互补有效降低感知不确定性;借助辅助任务挖掘数据中的隐含结构,显著提升样本效率与学习稳定性;同时增强策略在面对视角切换或部分观测缺失时的泛化能力。

前述视觉强化学习方法通常仅能应对单一类型的视觉干扰(如相机视角、外观或光照变化),且难以有效区分任务相关与任务无关的视觉信息,导致智能体在视觉变化显著的环境中泛化能力较弱。基于多视角信息瓶颈的稳健深度强化学习(robust deep reinforcement learning via multi-view information bottleneck, DRIBO) [98] 通过最大化多视角观测图像序列与表征序列之间的互信息,同时压缩从多视角图像中提取的任务无关信息,从而学习到更具稳健性与预测能力的表征。在 DMControl 环境中, DRIBO 在面对高维视觉干扰时表现出良好效果,其平均回报较现有最优方法提升 25%。然而, DRIBO 仅聚焦于全局互信息,可能在某些环境中忽略局部关键特征,导致其在部分特定任务上的表现不如其他基于数据增强的方法。

然而,基于数据增强的方法易因随机化导致训练不稳定,难以实现零样本的仿真到现实迁移。为此, Yuan 等 [99] 提出 Maniwhere 框架:通过结合对比损失与特征图对齐损失的多视角表征学习,提取共享语义信息;引入空间变换网络增强空间感知能力;采用课程式域随机化策略提升训练稳定性。该框架旨在提升视觉运动机器人在多种视觉干扰下的操作能力,实现多样化真实环境中的良好泛化。实验结果表明, Maniwhere 在仿真与真实机器人平台上均表现出优越性能,可实现对真实世界复杂视觉条件的零样本适应。但该方法对目标物体的颜色变化较为敏感,且在长时程复杂任务中仍存在局限。

尽管 Maniwhere 在视觉稳健性方面取得了显著进展,但其对多视角信息的整合能力仍显不足,尤其在对视角间共享与私有特征的解耦方面存在改进空间。为更有效地整合多相机视角的信息, Dunion 等 [100] 提出多视角解纠缠方法(multi-view disentanglement, MVD)。MVD 通过引入两种自监督学习损失(共享辅助损失和私有辅助损失)来分

别增强视角间共享特征的一致性和减小私有特征间的相似度, 从而为多视角强化学习策略提供辅助任务, 提高模型泛化能力。

在涉及多物体操作的复杂任务中, 现有方法仍难以有效建模物体间的依赖关系, 且对环境变化的泛化能力有限。为此, Haramati 等<sup>[101]</sup>提出一种以实体为中心的强化学习方法。首先, 通过预训练的深度潜在粒子模型从图像数据中提取实体及其属性, 随后采用实体交互 Transformer 来建模物体间交互关系和跨视角关联, 并采用广义密度感知 Chamfer 距离作为基于图像的奖励机制, 引导机器人将物体移至目标位置。实验表明, 该方法在三物体训练任务中表现优异, 并能有效扩展到十物体以上的复杂场景, 展现了良好的泛化能力。然而, 该方法依赖预训练模型, 且基于图像的奖励设计在真实机器人应用中可能面临挑战, 性能也易受遮挡和噪声影响。

多视角视觉强化学习方法能有效克服单一视角下的视觉遮挡和视野盲区问题, 并通过视角间的一致性约束增强模型对视角变化的稳健性, 从而提升策略在部署时对摄像机位姿变化的适应能力。然而, 此类方法也面临一些局限: 首先, 它高度依赖多视角数据的同步采集, 对硬件配置与数据存储要求较高; 其次, 不同视角间可能存在明显的外观差异, 为特征对齐与融合模块的设计带来挑战, 若处理不当易引入噪声; 此外, 在视角数量有限或某些视角质量较差时, 其性能提升可能较为有限。

## 2.4 知识迁移型视觉强化学习

知识迁移型视觉强化学习旨在解决视觉任务在不同领域或环境间的泛化问题, 尤其关注源域与目标域之间存在显著视觉差异时的跨域适应能力。此类方法通过迁移已有知识(如策略、表征或模型参数), 有效提升智能体在新环境中的学习效率与适应性(如图 7 所示)。

接下来介绍一种知识迁移型视觉强化学习的代表性方法。在视觉强化学习领域, 智能体在有限视角下训练后, 往往难以将其学习到的技能有效迁移到新的、未见过的视角, 这一挑战被称为视角泛化问题。为解决该问题, Yang 等<sup>[102]</sup>提出一个重要观点: 在测试阶段对新视角的适应性是关键, 而不是一味追求视角不变性。为此, 他们提出 MoVie (visual model-based policy adaptation for view generalization), 旨在使基于模型的视觉强化学习策略能够适应并泛化到新的视角。MoVie 的核心在于它能够利用从交互中收集的状态转移元组, 并通过动力学模型的学习目标, 将 STN (spatial Transformer

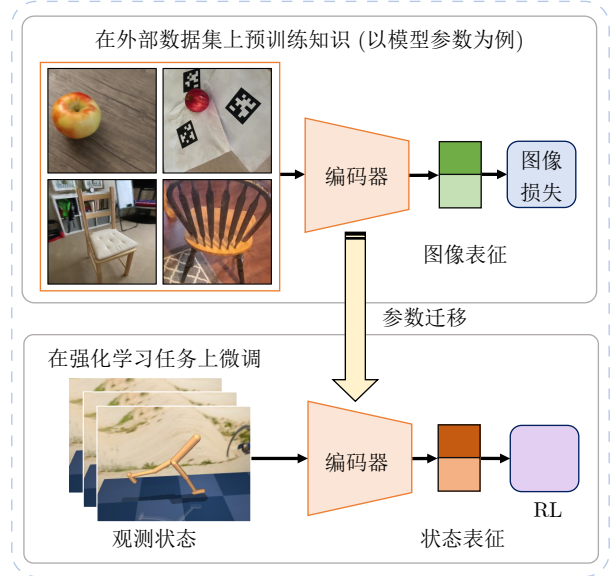


图 7 知识迁移型视觉强化学习示意图

Fig. 7 The diagram of knowledge-transferred VRL

networks) 融入到浅层架构中。这种方法的巧妙之处在于, 它在训练过程中不需要任何额外的调整, 并且能够与各种基于模型的视觉强化学习算法无缝协作, 展现出较高的通用性和适应性。

在测试阶段, 原始潜在状态动力学预测目标可表示为:

$$L_{\text{dynamics}} = \|d(h(o_t), a_t) - h(o_{t+1})\|_2 \quad (6)$$

其中,  $h$  为图像编码器, 将高维观测映射到潜在状态空间;  $d$  表示潜在动力学模型。

测试时, 来自未见视图的观测位于不同空间  $O'$ , 其对应的潜在空间为  $Z'$ 。然而, 训练阶段学习的投影  $h$  仅实现从  $O$  到  $Z$  的映射, 策略  $\pi$  也只学习从  $Z$  到  $A$  的映射, 因此难以直接从  $Z'$  泛化到  $A$ 。为解决这一问题, MoVie 将投影  $h: O \mapsto Z$  适配为  $h': O' \mapsto Z$ , 使策略无需训练即可执行正确映射  $Z \mapsto A$ 。同时, 固定潜在动力学模型  $d$  为  $d^*$ , 使其成为监督信号而非训练目标。在  $h$  的浅层结构中插入 STN 模块以优化投影适配, 因此将  $h$  表示为  $h^{SAE}$  (空间自适应编码器)。尽管目标仍是潜在状态动力学预测损失, 但此监督方式与训练阶段存在本质区别。具体目标可写为:

$$L_{\text{view}} = \|d^*(h^{SAE}(o_t), a_t) - h^{SAE}(o_{t+1})\|_2 \quad (7)$$

下面进一步介绍空间自适应编码器的结构。为保持方法简洁并实现快速适应, 只在原始编码器的浅层插入两个不同的 STN 模块。每个 STN 模块由两部分组成: 定位网络, 用于预测包含 6 个参数的仿射变换; 网格采样器, 用于生成仿射网格并从原

始特征图中采样特征. 仿射变换的点对点映射公式为:  $(x^s, y^s)^T = A_\phi(x^t, y^t, 1)^T$ . 其中,  $(x^s, y^s)$  是输入特征图中定义样本点的源坐标;  $(x^t, y^t)$  为输出特征图中规则网格的目标坐标;  $A_\phi$  表示仿射变换矩阵.

综上所述, MoVie 方法通过在测试时微调空间自适应图像编码器来实现视角泛化. 该方法无需显式奖励信号, 也无需在训练阶段进行结构修改, 即可成功将视觉模型驱动的策略泛化至未见过的视角.

尽管视角适应方法(如 MoVie<sup>[102]</sup>)提升了跨视角泛化能力, 但视觉强化学习方法仍普遍面临另一类挑战: 当前方法虽能直接从高维视觉输入中学习策略, 但在同时学习特征提取与高奖励动作时, 常面临样本效率低和计算成本高的问题. 针对这一挑战, Mu 等<sup>[103]</sup>提出两阶段学习框架 State-to-Visual DAgger: 首先基于低维状态观测训练状态策略, 再通过在线模仿学习将其迁移至视觉空间. 实验表明, 该方法在复杂任务中展现出显著优势, 不仅收敛更稳定, 计算效率也优于传统视觉强化学习方法. 然而在简单任务场景下, 其性能优势及样本效率提升则相对有限.

然而, 上述两阶段学习方法面临环境迁移带来的领域偏移挑战. 当训练与部署环境之间存在显著差异时, 视觉强化学习算法所学到的策略可能表现不佳, 且其泛化能力较弱. 对此, Wang 等<sup>[104]</sup>指出, 在面对明显的领域偏移时, 即使预测的奖励不够准确, 但仍然可以作为一种有效的信号用于指导策略的微调. 为此, 针对领域适应问题, 他们提出 PRFT (predicted reward fine-tuning)<sup>[104]</sup>. 具体而言, PRFT 在训练时联合学习一个策略和一个奖励预测模型. 然后, 在测试和部署之前, 利用预测的奖励对策略进行微调, 以适应目标测试环境. 实验结果表明, 该奖励预测模型在应对视觉域变化时表现出较强的泛化能力, 并且通过微调显著提升了策略性能. 然而, 当领域变化极为显著时, 错误的奖励预测可能会误导策略的微调过程, 导致其性能相对于零样本测试有所下降.

除了应对领域偏移, 跨域视觉控制任务还面临预训练阶段的效率与泛化能力之间的权衡问题. 现有方法在跨域视觉控制任务中面临两大挑战: 一是主动探索预训练方法因探索策略与编码器训练间的耦合关系导致跨域性能下降; 二是额外探索智能体的引入造成预训练效率低下. 为此, CRPTpro (cross-domain random pretraining with prototypes for RL)<sup>[105]</sup>通过随机策略解耦数据采样与编码器训练, 实现高效跨域数据收集, 并结合原型自监督算法预训练通用视觉编码器. 实验表明, CRPT-

pro 在多个复杂任务中显著优于现有方法, 尤其在跨域下游任务中表现突出, 且大幅提升了预训练效率. 但该方法依赖随机策略, 可能在复杂探索场景中受限, 且原型扩散损失的权重需精细调参, 增加了实现难度.

尽管以 CRPTpro<sup>[105]</sup>为代表的预训练方法提升了跨域泛化能力, 但是当前视觉强化学习方法仍主要依赖二维视觉技术, 难以有效处理真实世界中的三维场景. 针对这一局限, Ze 等<sup>[106]</sup>提出一种基于自监督三维表征学习的框架, 旨在提升强化学习在机器人控制任务中的样本效率和泛化能力. 该框架首先在大规模三维数据集上预训练几何感知的三维自编码器; 随后在强化学习过程中联合微调三维重建与策略学习目标, 并通过静态-动态视角合成任务强化空间理解能力. 实验表明, 该方法可实现对真实机器人环境的零样本迁移, 且对相机视角和光照变化具有较强的稳健性. 然而, 该方法依赖多视角输入数据, 而真实场景中多视角采集往往受限, 且三维预训练过程计算开销较大.

### 2.5 离线视觉强化学习

前述类别均属于在线视觉强化学习方法, 其特点是智能体通过与环境的持续交互来学习策略. 相比之下, 离线视觉强化学习则旨在使智能体仅通过预先收集的、包含高维图像数据的静态数据集来学习决策策略, 而无需与真实环境进行在线交互<sup>[107]</sup>(如图 8 所示). 图 9 展示了知识迁移型与离线视觉强化学习方法的时间线.

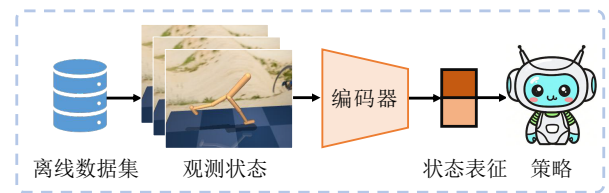


图 8 离线视觉强化学习示意图  
Fig.8 The diagram of offline VRL

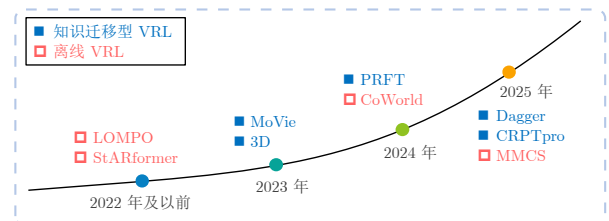


图 9 知识迁移型与离线视觉强化学习时间线  
Fig.9 The timeline of knowledge-transferred and offline VRL

下面介绍一种典型的离线视觉强化学习方法. 现有基于模型的视觉强化学习在低维状态空间中表现良好, 但处理图像输入时面临不确定性量化难、计算成本高的问题, 且离线设定下无法通过交互修正模型误差. 为此, Rafailov 等<sup>[108]</sup>提出一种基于潜在空间建模的方法 LOMPO (latent offline model-based policy optimization), 利用变分自编码器将高维图像映射到潜在空间, 并结合集成潜在动力学模型进行不确定性估计 (以预测差异为度量). 该方法构建带不确定性惩罚的潜在马尔科夫决策过程, 通过悲观正则化缓解分布偏移.

具体而言, LOMPO 利用可获得的离线数据学习变分模型, 该模型包括图像编码器、解码器以及一组潜在动力学模型. 该变分模型的优化目标如下:

$$L(\theta) = \mathbb{E}_{p_\theta} \left[ \sum_{t=0}^{H-1} \left( -\log q_\theta(o_{t+1}|s_{t+1}) + \text{KL} \left( p_\theta(s_{t+1}|o_{t+1}, s_t, a_t) \parallel \hat{T}_t(s_{t+1}|s_t, a_t) \right) \right) \right] \quad (8)$$

其中, 每一步从前向转移模型集合  $\{\hat{T}_1, \dots, \hat{T}_K\}$  中随机抽取一个模型  $\hat{T}_t$  用于训练,  $K$  指的是学到的前向转移模型集中模型的数量. 真实样本分布  $p$  通过标准的均值场近似建模为单模态高斯分布, 并在所有时间步中共享.

在潜在空间表征的基础上, 分别训练 Actor 网络  $\pi_\phi(a_t|s_t)$  和 Critic 网络  $Q_\psi(s_t, a_t)$ , 并同时对二者进行模型推理. 为此, 维护两个经验缓冲池: 真实数据经验缓冲池  $B_{\text{real}}$  和潜在数据经验缓冲池  $B_{\text{sample}}$ . 真实数据经验缓冲池包含来自潜在马尔科夫决策过程的转移元组  $s_t, a_t, r_t, s_{t+1}$ , 其中状态是从真实样本分布  $s_{1:H} \sim p_\theta(s_{1:H}|s_{1:H}, a_{1:H-1})$  中采样的, 该分布覆盖了来自真实数据集  $D_{\text{env}}$  的轨迹  $\{o_{1:H}, r_{1:H}, a_{1:H-1}\}$ . 潜在数据经验缓冲池则包含在模型潜在空间中通过策略展开得到的状态转移, 其中每一步的状态转移通过从集合中随机选择一个前向模型执行. 展开过程中的奖励  $\tilde{r}_t$  使用集合估计与不确定性惩罚项计算, 具体形式如下:

$$\tilde{r}_t(s_t, a_t) = \frac{1}{K} \sum_{i=1}^K r_\theta(s_t^{(i)}, a_t) - \lambda u(s_t, a_t) \quad (9)$$

其中,  $s_t^{(i)} \sim \hat{T}_{\theta_i}(s_{t-1}, a_{i-1})$  表示从每个前向模型中采样的状态, 而  $s_t$  是从集合  $\{s_t^{(i)}, i = 1, \dots, K\}$  中采样得到的. 这里  $u(s_t, a_t)$  是模型不确定性的估计,  $\lambda$  为惩罚参数. 特别地, 将  $u(s_t, a_t)$  定义为集合中各潜在模型预测之间的一致性, 即对数似然的方

差  $u(s_t, a_t) = \text{Var}(\{\log \hat{T}_{\theta_i}(s_t|s_{t-1}, a_{t-1}), i = 1, \dots, K\})$ . 使用这个启发式方法来估计不确定性, 旨在反映前向模型在均值与方差层面的不一致性. 最后, Actor 网络  $\pi_\phi(a_t|s_t)$  与 Critic 网络  $Q_\psi(s_t, a_t)$  采用标准的异策略训练算法, 使用从真实与潜在数据经验缓冲池中按相等比例抽取的批次进行训练.

实验显示, LOMPO 在视觉运动控制仿真和真实机器人任务 (如 drawer-closing) 中的表现优于离线无模型及在线视觉强化学习方法, 在数据有限时仍表现稳健. 但它在专家数据分布极窄 (如 Door Open) 时表现受限, 且潜在空间假设可能影响其在复杂动态环境中的泛化能力.

除基于模型的方法外, 序列建模方法通过将强化学习转化为序列建模问题也引起了广泛关注, 但仍存在明显局限: 一方面, 全局自注意力机制忽视了状态-动作-奖励三元组在相邻时间步的强因果关系, 导致长序列建模效率低下且局部关系捕捉不足; 另一方面, 传统卷积编码会损失图像中的细粒度空间信息. 针对这些问题, Shang 等<sup>[109]</sup>提出双层架构 StARformer (state-action-reward Transformer) 模型, 其中 Step Transformer 通过 ViT 式分块编码建立单步内状态-动作-奖励的细粒度关联, Sequence Transformer 则结合全局卷积特征进行长序列建模. 实验表明, StARformer 在离线视觉强化学习与模仿学习任务中优于 Decision Transformer, 尤其在长序列场景下优势显著. 然而, 其双 Transformer 结构增加了模型复杂性, 且对超参数选择较为敏感, 可能限制其在更复杂任务中的扩展性.

进一步地, 在离线视觉强化学习中, 高维观测易受混杂因素干扰而产生虚假关联, 且因无法通过交互探索加剧了无关信息的影响. 为此, Zhang 等<sup>[110]</sup>提出基于掩码的最小因果状态表征学习方法 (mask-based minimal causal state, MMCS). 该方法通过掩码网络划分潜在空间, 结合条件独立性约束与因果充分性目标, 有效分离出任务相关的最小因果变量, 消除变量间冗余依赖. 作为可插拔模块, MMCS 能兼容多种离线强化学习算法. 在 VisualD4RL 基准测试中, MMCS 显著提升了策略性能与样本效率, 并在多种视觉干扰环境中展现出优越的稳健性. 然而, 该方法在处理高复杂度任务 (如 Humanoid-walk) 时仍存在因探索-利用权衡导致的性能波动问题, 其任务适应能力及损失项权重的动态平衡机制仍需进一步研究.

除视觉干扰和因果混淆问题外, 离线视觉强化学习还常面临过拟合和值函数高估问题, 影响策略的稳健性. 为此, Wang 等<sup>[111]</sup>提出协同式世界模型框架 (collaborative world models, CoWorld), 旨在

通过“在线化”提升离线视觉强化学习性能。该方法利用在线模拟器作为离线策略测试平台,通过在线到离线的知识迁移优化离线策略。具体包括:通过离线-在线状态对齐减小域间差异,利用在线-离线奖励对齐使源域 Critic 可评估目标策略,并引入最小-最大值约束调节值估计,避免对分布外数据过度惩罚。实验表明,CoWorld 在多种视觉控制任务尤其是跨任务与跨环境设置下表现优异,有效提升了策略性能。然而,该方法依赖训练良好的源域在线模型,性能受限于源域与目标域的相关性,且因引入额外模型组件而增加了复杂性和计算开销,可能影响其在大规模或复杂场景中的应用效率。

总而言之,离线视觉强化学习的优势在于其训练过程具有较高的安全性,无需与环境实时交互,仅依靠静态数据集即可学习策略,从而有效规避了在高风险场景中的试错代价,并能够充分利用已有的历史数据。然而,该方法的核心局限在于,智能体不仅需要从高维且冗余的像素信息中提取有效特征,还严重受限于离线数据集的质量和覆盖范围,容易遭遇分布偏移问题。当面对数据分布以外的状态时,策略可能因外推错误而产生严重失效,甚至导致整体性能崩溃。

### 3 基准平台

本节系统性地介绍四种广泛应用于视觉强化学习研究的基准平台,包括 DeepMind Control Suite (DMControl)<sup>[112]</sup>、基于其改进的 DeepMind Control Generalization Benchmark (DMControl-GB)<sup>[113]</sup>、专注于视觉干扰场景的 Distracting Control Suite (DCS)<sup>[114]</sup> 以及综合性更强的 RL-ViGen<sup>[115]</sup> 平台。这些平台各具特色,在核心目标、视觉观测、环境复杂度等方面展现出明显的差异性特征,其详细对比如表 3 所示。接下来,本文将从平台架构设计、任务设置、视觉观测与奖励设计等多个维

度,深入剖析这四种基准平台的技术细节与核心特性。

#### 3.1 DMControl

DMControl<sup>[112]</sup> 是专为强化学习研究设计的连续控制任务基准平台,基于 MuJoCo 物理引擎构建。该平台提供标准化的测试环境,任务结构清晰,奖励函数设计合理且易于理解,便于算法评估与对比。所有任务均采用 Python 实现,代码可读性强,并支持灵活扩展。项目开源地址: [https://github.com/google-deepmind/dm\\_control](https://github.com/google-deepmind/dm_control)。

在环境设置方面,DMControl 严格遵循连续马尔科夫决策过程框架,构建一个完整的连续控制任务体系。该平台定义连续的状态空间、动作空间和观测空间。其中,状态空间由位置、速度等物理量构成。动作空间经过规范化处理,统一映射到  $[-1, 1]$  区间,以确保跨任务一致性。观测空间则提供两种模式:一种是低维特征观测,包括关节角度、运动速度和传感器读数等结构化数据,这些数据以有序字典的形式组织,便于智能体处理;另一种是像素级视觉观测,通过封装器实现,为视觉强化学习研究提供支持。这种双重观测模式的设计,使得研究者能够在同一任务框架下,对比基于状态和基于视觉的不同算法表现。

在奖励机制方面,DMControl 采用精心设计的奖励函数体系。多数任务的奖励值被归一化至  $[0, 1]$  区间,既包含稀疏奖励设置,也提供基于距离的连续奖励形式。此外,平台还通过实时颜色可视化来增强反馈的直观性,帮助研究人员更清晰地理解算法的表现。虽然任务本身采用无限时域设定,但在评估阶段,DMControl 采用固定 1000 步的标准化流程,以累计奖励作为主要评价指标。这种设计使得不同算法的学习曲线能够在统一的  $[0, 1000]$  量纲范围内进行直观的比较和分析,为算法的评估

表 3 基准平台对比  
Table 3 Comparison of benchmark platforms

	基准平台			
	DMControl	DMControl-GB	DCS	RL-ViGen
核心目标	基础连续控制任务评估	视觉泛化能力评估	抗视觉干扰能力评估	全面视觉泛化评估
基础环境	MuJoCo	基于 DMControl	基于 DMControl	融合多个仿真平台
任务类型	连续控制	同 DMControl, 但增加泛化测试	同 DMControl, 但增加视觉干扰	多样化 (运动控制、自动驾驶、灵巧操作、桌面操作、室内导航)
视觉观测	固定视角、简单背景	支持视觉变化 (颜色、背景)	动态干扰 (相机位姿、颜色、背景)	高保真、多视角、动态光照、复杂场景、跨形态
泛化能力评估	不支持	支持背景变化下的泛化	支持干扰下的零样本泛化	支持多种泛化类型, 强调跨任务、跨形态、跨视角的综合泛化能力
环境复杂度	低 (简单物理仿真)	中等 (视觉变化但任务单一)	中等 (动态干扰但任务单一)	高 (真实仿真、多样化任务)

提供明确且一致的标准。

DMControl 构建一个全面且层次分明的连续控制任务体系, 涵盖从基础物理系统到复杂机器人控制的广泛挑战。该平台的任务设计按照难度和应用场景可分为以下几类: 在平衡控制领域, DMControl 设置 Pendulum (倒立摆)、Acrobot (双摆) 等经典任务, 用于评估算法的基本平衡能力。其中, Cartpole (小车倒立摆) 系列任务通过设置不同的初始条件 (如摆起和平衡) 来考查算法的适应能力, 而多连杆的 Cart-k-pole 任务则进一步提升控制难度; 在物体操控方面, Ball in cup 测试动态物体捕捉能力, Finger 模拟精细物体操控, Manipulator 和 Stacker 则涉及抓取、搬运和堆叠等复杂操作技能。运动控制任务则包含 Hopper (平面跳跃)、Cheetah (猎豹奔跑)、Walker (双足机器人行走) 以及 Swimmer (多节游泳) 等多样化场景。平台还包含 Point-mass (点质量导航)、Reacher (机械臂抓取) 和 Fish (鱼类游泳) 等特殊场景任务。最具挑战性的是 Humanoid 系列任务, 包括简化版人形机器人和基于真实动作捕捉数据的 Humanoid\_CMU, 这些任务支持站立、行走和奔跑等高难度动作, 为研究复杂机器人控制算法提供极具挑战性的测试场景。此外, LQR 领域为线性系统控制提供理论验证环境, 为控制理论的研究和应用提供有力支持。这些任务共同构成一个标准化、可扩展的强化学习基准测试平台, 已成为评估 A3C、DDPG、D4PG 等先进算法性能的重要基准。

### 3.2 DMControl-GB

DMControl-GB<sup>[13]</sup> 是基于 DMControl 构建的强化学习基准平台, 其核心目标是评估智能体在视觉输入条件下的泛化性能。该平台通过将训练环境与测试环境进行解耦, 模拟智能体在真实场景中可能面临的泛化挑战。项目开源地址: <https://github.com/nicklashansen/dmcontrol-generalization-benchmark>。

DMControl-GB 的基准测试涵盖平衡控制、物体操控和运动控制三大核心领域, 包含多种典型任务, 如 cartpole\_swingup (倒立摆摆动)、ball\_in\_cup\_catch (球杯捕捉)、finger\_spin (指尖旋转)、walker\_walk (双足机器人行走) 与 walker\_stand (站立) 等典型任务。这些任务从基础运动到精细操作, 构成一个多层次的复杂控制挑战体系。DMControl-GB 采用纯视觉输入模式, 智能体仅能获取经过裁剪的  $84 \times 84$  像素 RGB 图像观测, 而无法直接访问系统底层的状态信息。训练阶段在固定的环境中进行, 而测试阶段则引入显著的视觉干扰因素,

比如通过随机颜色扰动改变背景、物体和光照条件, 或者采用动态视频背景替换来增加视觉复杂性。这种设计迫使智能体必须建立对任务关键特征的稳健表征能力, 同时过滤无关视觉干扰, 从而系统地评估其在未见视觉条件下的适应性。

在奖励机制方面, 各个任务都采用原生定义的奖励函数。这些奖励函数基于运动速度、平衡稳定性或物体间距等指标, 引导智能体在避免依赖易变视觉特征的前提下, 学习最大化累积奖励的最优策略。通过这种多维度的任务设计和训练-测试环境解耦的评估机制, DMControl-GB 为视觉强化学习领域的泛化性能研究建立了全面且严谨的标准化测试平台。

### 3.3 DCS

DCS<sup>[14]</sup> 是基于 DMControl 开发的一个强化学习基准平台, 其核心目标是通过引入视觉干扰来模拟真实世界中的感知挑战, 从而促进算法在机器人控制等实际应用中的稳健性提升。DCS 在完整保留 DMControl 原有物理引擎、任务逻辑和奖励函数的基础上, 巧妙地引入三种视觉干扰机制: 相机位姿干扰通过随机调整相机的位置和方向来模拟真实场景中的视角变化; 物体颜色干扰则随机改变场景中物体的颜色特征; 动态背景干扰则利用 DAVIS 2017 数据集中的随机视频片段替代原始静态背景。这些干扰的强度可以通过参数调整, 支持静态或动态两种模式运行, 为算法测试提供多层次的难度梯度。这些设计有效模拟了真实场景中的视角偏移、光照变化和环境动态, 大幅增加了从像素观测中提取状态信息的难度, 从而更精准地评估算法在复杂视觉干扰下的性能。项目开源地址: [https://github.com/google-research/google-research/tree/master/distracting\\_control](https://github.com/google-research/google-research/tree/master/distracting_control)。

在观测设置方面, 智能体接收的观测数据是包含多种干扰因素的 RGB 图像。与原始 DMControl 环境不同, 这些图像的变化不仅由任务状态本身驱动, 还包含大量与任务无关的视觉干扰信息。面对这样的输入, 智能体需要具备从复杂像素流中过滤干扰、识别关键状态特征的能力。在奖励机制方面, DCS 完全继承 DMControl 的奖励函数设计, 其反馈仅基于机器人的物理状态参数 (如位置、速度、目标达成度等), 确保智能体的学习目标始终聚焦于解决底层控制问题, 而非单纯适应视觉层面的变化。

DCS 整合了 Cheetah、Walker、Reacher 等一系列经典的连续控制任务。该套件通过设定 Easy

和 Medium 两个难度级别来建立标准化的评估体系, 其中难度调节主要通过控制干扰强度实现, 包括相机扰动幅度、颜色变化范围以及背景视频数量等参数. 实验结果显示, 视觉强化学习算法在 DCS 环境中的表现显著低于其在原始 DMControl 环境中的性能, 特别是在面对多种干扰因素叠加的复杂场景时, 算法性能出现急剧下滑. 这一现象清晰地揭示了现有方法在处理真实世界视觉复杂性方面存在的局限性, 同时也为后续研究指明了需要突破的技术难点和发展方向.

### 3.4 RL-ViGen

RL-ViGen<sup>[115]</sup> 是一个专为评估视觉强化学习智能体泛化性能而设计的综合性基准平台. 该平台集成 Adroit (灵巧操作)、CARLA (自动驾驶)、Habitat (室内导航)、Robosuite (桌面操作) 和 DMControl (运动控制) 等多个高保真仿真环境, 构建一个覆盖机器人操作到复杂导航等多样化任务的测试体系. 通过对这些环境进行功能扩展, RL-ViGen 支持多种视觉泛化类型测试, 能够全面评估智能体在未知场景中的适应能力. 平台设计的任务不仅考查智能体的策略学习能力, 更强调其在真实像素输入条件下应对各类动态变化的稳健性, 从而有效提升测试结果与实际部署需求的一致性. 项目开源地址: <https://github.com/gemcollector/RL-ViGen>.

在观测设置方面, RL-ViGen 采用基于摄像头的像素输入作为状态表示, 模拟真实机器人系统的感知方式. 为全面评估智能体的稳健性与泛化能力, 平台引入五种视觉干扰因素: 视觉外观差异、相机视角变化、动态光照变化、多样化场景结构和跨形态变化. 这些因素完整覆盖了真实环境中智能体可能面临的主要视觉挑战. 针对不同任务特性, 平台设计特定的泛化评估方案. 例如, 桌面操作任务重点考查智能体对前四类干扰因素的适应能力, 而跨形态评估则专门测试智能体在不同机器人形态间的知识迁移能力. 这些设定显著提升了任务难度, 促使算法学习更具不变性的特征表示. 在奖励机制方面, 平台保留各个任务原有的设计, 通过稀疏或密集的奖励信号, 结合任务完成度、距离误差或稳定性等多元指标来驱动学习过程. 实验采用训练-测试解耦模式, 智能体首先在固定环境中进行充分训练, 随后在包含各类干扰因素的零样本测试场景中接受性能评估. 主要评价指标包括任务成功率和累计奖励值等关键性能参数.

RL-ViGen 围绕五大核心领域构建, 涵盖多种任务类型与复杂的视觉泛化挑战, 全面评估视觉强

化学习模型在多样化环境下的适应能力. 在灵巧操作方面, 平台基于 Adroit 环境设计开门、使用锤子以及笔操作等需要精细手部控制的任务, 重点考查智能体在物体颜色、手部形态及视角发生显著变化时的泛化表现; 自动驾驶领域则依托 CARLA 仿真器, 构建涵盖动态天气变化、复杂交通状况以及不同道路结构的驾驶场景, 模拟真实驾驶中的多重不确定性; 室内导航任务利用 Habitat 平台的高保真 3D 环境, 要求智能体在光照条件波动和场景布局调整的情况下完成目标定位, 以测试其在真实室内空间中的导航鲁棒性; 桌面操作部分基于 Robosuite 框架设置机械臂抓取与装配任务, 同时引入动态背景干扰和跨形态设备 (如不同构型的机械臂) 的泛化测试, 增强任务的现实挑战性; 运动控制方面对 DMControl 进行扩展, 包含四足机器人行走、平衡等典型任务, 并通过改变物体外观、光照条件及背景变化来系统评估运动策略的稳定性与适应能力. 这五个领域共同构成一个高度贴近真实世界复杂视觉条件的综合性测试平台, 为视觉强化学习算法的泛化能力研究提供丰富、多样且极具挑战性的实验环境.

该平台在统一的优化框架下集成多种前沿的视觉强化学习算法, 如 DrQv2<sup>[117]</sup>、CURL<sup>[67]</sup> 以及 SGQN<sup>[25]</sup> 等, 确保算法间比较的公平性. 实验结果表明, 尽管部分方法在特定泛化场景中展现出一定优势 (例如 SGQN 在视角变化任务中表现优异), 但在面对跨形态迁移或复杂场景结构发生显著变化等更具挑战性的条件下, 现有算法整体表现仍不尽如人意, 这揭示了当前视觉强化学习在实现真正稳健泛化能力上的局限性. RL-ViGen 的构建旨在推动研究者探索更具普适性与环境适应性的智能体, 进而加速强化学习技术在真实世界场景中的广泛应用与落地进程.

综上所述, DMControl 是一种基础的连续控制视觉强化学习基准, 其环境无视觉干扰. DMControl-GB 和 DCS 在 DMControl 的基础上分别引入背景变化和强干扰, 主要用于测试策略在视觉扰动下的稳健性, 但它们的任务类型较为单一. 相比之下, RL-ViGen 提供了更为全面的视觉泛化测试, 涵盖多种任务类别和泛化类型, 非常适合用于开发和测试能够在复杂现实世界场景中实现泛化的强化学习算法.

## 4 应用

视觉强化学习技术已在多个前沿领域展现出广泛应用潜力. 例如, 在机器人控制的布料整理任务

中, 机器人基于视觉观测图像对布料状态进行估计, 并执行相应的整理动作; 在自动驾驶场景中, 系统通过分析连续图像帧提取时序运动特征, 进而优化车辆行驶轨迹; 在多模态大模型的视觉推理任务中, 智能体能够根据语言描述对图像内容进行分析与推理, 并生成对应的响应结果. 除上述场景外, 该技术还在图像生成、多模态情感识别及芯片布局等领域发挥着重要作用. 下面将系统介绍这几类典型应用及其代表性方法.

#### 4.1 机器人控制

视觉强化学习技术近年来在机器人控制领域得到广泛应用, 其通过融合高维视觉输入与强化学习的决策机制, 显著提升了机器人在复杂环境中的感知与响应能力. 典型应用包括移动机器人的自主导航、机器人对柔性物体如布料的精细操作以及无人机在动态环境中的自主避障等. 这些场景共同体现出视觉感知与强化学习结合在应对环境不确定性、高维状态空间等问题方面的优势. 下面将分别介绍三类具有代表性的应用案例.

在无地图环境中, 目标驱动的移动机器人的导航依赖于有效的状态表征来确保可靠的决策. 受到点云中鸟瞰图 (bird's-eye view, BEV) 视觉感知优势的启发, Jiang 等<sup>[116]</sup> 提出一种新的导航策略——BEVNav. 该方法通过设计一种稀疏-密集 BEV 网络来高效地将三维点云转换为 BEV 特征. 这种转换不仅创建了有效的场景表征, 还借助时空对比学习机制促进了有效状态表征的学习. 具体而言, 在空间层面上, 来自点云的两个随机增强视图相互预测, 以增强空间表征; 在时间层面上, 将当前观测状态与连续帧的动作相结合来预测未来的状态表征, 以此建立观测状态转移与动作之间的关系, 捕捉随时间演变的信息. 实验结果表明, BEVNav 在多个公开基准测试中展现出优越的导航性能, 显著提升了在复杂环境下的导航能力. 然而, BEVNav 仅采用当前帧的点云作为观测状态, 而不依赖于之前帧中的时间信息. 在实际应用中, 结合当前帧与之前帧的数据来有效预测行人的运动轨迹更具潜力.

除了在结构化环境中的导航任务, 视觉强化学习在非刚性物体操作中同样展现出广泛潜力, 例如机器人布料整理. 现有方法在处理高动态、易变形的布料时, 通常依赖难以在真实环境中获取的精确状态信息 (如粒子位置); 若仅使用 RGB 图像, 则面临输入维度高、特征跟踪困难等问题. 为此, Chen 等<sup>[117]</sup> 提出基于 Transformer 的知识蒸馏方法 TraKDis, 通过视觉强化学习解决机器人布料整理

任务. 该方法首先利用特权状态信息 (布料粒子位置) 训练教师模型, 再通过知识蒸馏将其知识迁移至仅以 RGB 图像为输入的学生模型, 并结合预训练 CNN 编码器估计布料状态以缩小域差距, 同时采用权重初始化策略提升训练效率. 实验表明, TraKDis 在三种布料折叠任务中显著优于现有强化学习方法. 然而, 该方法仍需大量训练数据且模型规模较大, 可能影响推理速度与内存效率.

进一步地, 视觉强化学习也被广泛应用于无人机等空中机器人的自主控制任务, 尤其是在复杂环境中的避障问题. 针对无人机在虚拟管道环境中的自主避障问题, 现有激光雷达方案存在重量大、成本高及依赖定位信号等局限. 虽然视觉传感器能提供丰富环境信息, 但传统方法在复杂动态场景中避障效率不足. 为此, 赵静等<sup>[118]</sup> 提出基于视觉传感器的深度强化学习架构: 通过轻量化 CNN-LSTM 双网络处理深度相机数据, 设计新型奖励函数解决稀疏奖励问题, 并利用广义优势估计优化连续动作空间的轨迹平滑性. AirSim 仿真实验表明, 该方法在动/静态障碍环境中具有更快的收敛速度和更强的稳健性. 但需指出的是, 当前研究仅完成单机虚拟管道仿真, 尚未涉及非结构化环境、多机协同等复杂场景, 且实际平台的迁移性能和实时性仍需进一步验证.

#### 4.2 自动驾驶

视觉强化学习方法在自动驾驶领域应用广泛. 该类方法以原始图像或视频序列作为输入, 通过端到端的方式学习驾驶策略, 具备感知与决策一体化的优势. 下面以两种典型应用为例, 分别介绍其在特征提取与推理机制方面的代表性进展.

在基于视觉的自动驾驶任务中, 现有方法常因环境信息维度高而难以有效提取时空特征, 且受制于训练数据的分布偏差, 导致模型泛化能力不足. 针对这些问题, 杨蕾等<sup>[119]</sup> 提出一种融合时空特征的视觉自动驾驶强化学习算法 (STRLAD). 该方法采用双流网络结构, 分别从 RGB 图像中提取空间语义特征, 并从灰度图中提取时序运动特征, 同时引入注意力机制实现多层次特征融合, 从而更全面地感知环境. 随后, 利用柔性 Actor-Critic 算法学习驾驶策略, 以减轻数据偏差的影响并提升泛化性能. 在 CARLA 仿真平台上的实验结果表明, STRLAD 在复杂城市环境 (尤其是存在多动态物体的场景) 中可实现 89% 的自动驾驶完成率, 显示出良好的性能. 然而, 该方法仍存在单帧推理耗时较长、收敛速度较慢以及超时率偏高等问题.

尽管 STRLAD 等方法在特征提取和策略学习方面取得了一定进展, 但当前自动驾驶领域的端到端模型在应对长尾场景时, 仍常因缺乏常识与推理能力而表现不佳. 特别是现有视觉-语言模型方法多依赖监督微调, 未能充分结合强化学习与推理机制以提升规划性能. 为此, AlphaDrive 方法<sup>[120]</sup>构建了基于群组相对策略优化的强化学习框架, 设计准确性、动作权重、多样性和输出格式四类规划奖励, 并采用结合知识蒸馏的两阶段训练策略: 先通过大模型生成高质量推理数据进行监督微调预训练, 再进行强化学习微调, 显著提升了规划性能与训练效率. 实验表明, AlphaDrive 在 MetaAD 数据集上的表现优于现有监督微调方法, 具备多模态规划能力. 但该方法尚未支持变道等复杂行为, 且依赖大模型生成的伪标签, 存在感知误差风险.

### 4.3 多模态大模型

视觉强化学习方法在多模态大模型尤其是视觉-语言模型中应用广泛. 为更具体地展示其应用价值, 以下将分别从推理机制优化和图像质量评估两个典型方向, 介绍相关研究案例.

首先, 在视觉推理机制方面, 现有大型视觉-语言模型在多任务视觉感知中常面临缺乏统一框架、推理冗余以及多目标匹配效率低等问题. 针对这些挑战, Liu 等<sup>[121]</sup>提出一种名为 VisionReasoner 的统一视觉感知与推理框架. 该框架通过多对象认知学习和系统级任务重构, 将检测、分割、计数等任务整合于单一共享模型, 并设计包含思维奖励与非重复奖励的结构化推理机制, 以及 IoU 和 L1 等精准定位奖励, 以提升性能. 实验结果显示, VisionReasoner 在涵盖上述任务的十个基准上的表现均显著优于基线模型. 然而, 该框架在处理复杂任务时仍可能存在推理效率瓶颈, 且受限于较小的训练数据规模 (仅 7000 个样本), 其在更广泛场景下的泛化能力仍有待验证.

除了推理机制, 图像质量评估是多模态大模型实际应用中另一个关键问题. 当前该类方法普遍面临两大挑战: 一是依赖大量标注数据进行有监督微调, 导致灵活性差、成本高; 二是评估结果往往难以兼顾准确性与可解释性, 通常仅输出缺乏解释的分数或仅提供描述性判断而无法量化评分. 针对这些问题, Li 等<sup>[122]</sup>提出一种基于群组相对策略优化的强化学习框架 Q-Insight. 该方法联合优化图像质量评分与退化感知任务, 通过设计可验证评分、退化分类和强度感知三类奖励函数, 仅需少量标注数据即可实现高效的视觉推理. 实验表明, Q-Insight 在

评分回归与退化感知任务上均显著优于现有最先进方法, 并在零样本图像比较中展现出优异的泛化能力. 然而, 该模型在“无退化”类别的识别准确率略低于基线方法, 反映出其对潜在退化可能过于敏感, 存在一定误判风险; 此外, 其训练需消耗 16 块 A100 级别 GPU, 计算资源成本较高, 限制了实际部署的可行性.

综上所述, 尽管视觉强化学习方法在多个视觉任务中展现出显著优势, 其在效率、泛化能力、资源消耗等方面仍面临诸多挑战, 有待后续研究进一步优化.

### 4.4 其他应用

除了上述三个典型应用领域, 视觉强化学习在其他相关领域也得到广泛应用, 例如图像生成、多模态情感识别和芯片布局等方向. 以下将分别介绍其在这些具体任务中的应用案例及代表性方法.

首先, 在图像生成方面, 扩散模型虽已表现出卓越性能, 但仍存在放大训练数据固有偏差的倾向, 导致生成结果缺乏多样性. 例如, 输入与特定职业相关的文本提示时, 生成图像可能呈现明显的性别偏见. 针对这一问题, 文献<sup>[123]</sup>采用强化学习方法对扩散模型进行微调, 以提升生成图像的多样性. 该方法设计一种基于图像集的多样性奖励函数, 通过评估当前生成图像集与一个无偏参考图像集之间的分布差异, 来衡量生成结果的多样性水平. 进一步地, 引入边际效用概念为每张生成图像分配个体奖励, 从而将扩散过程建模为一个多步决策问题, 并借助策略梯度方法对模型进行微调, 以最大化多样性奖励. 实验结果表明, 该方法在多个图像分类数据集上有效提高了生成多样性, 但其性能仍依赖于外部参考图像集的选择质量.

除了视觉内容生成, 视觉感知与强化学习的结合也在多模态情感识别中展现出潜力. 现有音频-视觉情感分类方法在特征融合、模型优化与集成策略上存在局限, 如 3D CNN 缺乏光流信息, Wav2Vec2 忽略局部上下文, 传统粒子群优化易陷入局部最优. 为此, Kondrotas 等<sup>[124]</sup>提出一种基于强化学习增强粒子群优化的多模态集成模型. 该方法采用双流 3D CNN 处理 RGB 与光流视频, 改进 Wav2Vec2 并融合长短期记忆网络与注意力机制以增强音频特征提取, 同时利用强化学习优化粒子群的参数与动作选择, 协同搜索最优超参数、集成权重, 并构建具有互补性的动态集成模型. 实验表明, 该方法在音频-视觉情感数据集上的表现优于主流模型与搜索策略, 在多模态融合与优化方面优势显著. 但其计算复杂度高且依赖高质量多模态数据, 在数据质量

较差时性能会出现下降。

进一步地, 视觉感知的强化学习还在芯片布局这类复杂优化任务中取得显著进展. 文献 [125] 研究数字芯片后端设计中的全局布局优化问题, 针对传统优化方法 (如模拟退火、遗传算法) 计算成本高、易陷入局部最优, 以及现有强化学习方法忽略布局视觉信息、宏块尺寸导致重叠率高等不足, 提出一种基于视觉强化学习的解决方案. 该方法将电路网表转化为多通道特征图像, 通过融合卷积神经网络与图卷积网络构建决策模型, 利用视觉热力图指导布局, 并设计基于引脚数量的宏块排序机制和密集奖励函数, 在确保零重叠的同时实现线长最小化. ISPD2005 基准测试验证了该方法的有效性, 但其在更大规模或先进工艺节点电路中的适用性仍需进一步研究.

视觉强化学习在图像生成、多模态情感识别和芯片布局等领域的应用表明, 该方法通过将感知与决策相结合, 能够有效处理高维状态空间和复杂优化问题. 尽管这些方法在各自领域取得了显著进展, 但仍普遍面临计算成本高、对数据质量依赖性强以及泛化能力有限等共同挑战.

## 5 展望

尽管视觉强化学习在机器人控制、自动驾驶和多模态大模型等领域取得了显著进展, 但在走向实际应用过程中仍面临诸多挑战. 本节从视觉域差距、模型推理效率、可解释性与安全性四个维度, 梳理当前研究的主要问题, 并探讨未来发展方向.

1) 视觉域差距. 当前方法在从仿真迁移至实时面临显著视觉差异, 包括光照变化、传感器噪声、动态模糊及复杂物理交互 (如柔性物体变形与非线性摩擦), 这些因素常被简化处理, 导致策略在实际场景中性能下降. 此外, 现实系统中的信号延迟也进一步增加了策略迁移的不确定性. 未来研究应聚焦于构建高真实感的“神经仿真器”, 结合生成模型与神经渲染技术还原真实视觉与动力学特性; 同时发展域随机化、元学习和在线自适应等方法, 提升智能体的跨域迁移与自主调整能力.

2) 多模态大模型推理效率瓶颈. 多模态大模型虽具备出色的视觉理解能力, 但参数规模大导致推理延迟高、资源消耗大, 难以满足实时系统 (如自动驾驶) 的毫秒级响应要求. 未来应致力于提升推理效率, 包括模型压缩、知识蒸馏、稀疏化与专用硬件加速; 也可探索模块化架构, 以大模型为指导驱动轻量化策略网络, 实现性能与效率的平衡.

3) 可解释性不足. 策略因直接从像素学习而缺

乏语义透明性, 在安全敏感场景中调试困难、风险突出. 现有事后解释方法 (如显著性图) 提供信息有限, 理论层面也缺乏对策略泛化与收敛性的系统分析. 未来应融合表征学习、信息论与因果推断等方法, 构建视觉表征质量评估体系, 并加强策略中的因果建模以提升可解释性.

4) 安全性问题. 视觉系统易受对抗攻击, 微小扰动可能导致严重误判; 在未知环境中可能因缺乏先验而产生危险行为, 甚至因视觉错觉误导奖励函数. 未来需构建多层次防御机制: 在感知层通过对抗训练、输入净化与多传感器融合提升鲁棒性; 在决策层引入不确定性量化, 实现高风险状态下的安全降级; 在学习过程中嵌入物理安全约束或通过形式化方法验证策略可靠性; 还应建立“人类在环”监督机制, 平衡性能与安全.

总之, 视觉强化学习在迈向更广泛应用的道路上, 需要在视觉域适应、推理效率、可解释性和安全性等方面持续突破, 以支持智能体在复杂现实环境中的安全、高效与可持续学习.

## 6 总结

本文系统综述视觉强化学习的研究现状与关键技术. 首先围绕不同方法的核心思想与实现机制, 深入分析图像增强型、模型增强型、任务辅助型、知识迁移型以及离线视觉强化学习方法的研究进展与演变脉络, 比较各类方法的特点与优劣. 随后, 介绍四种常用的基准平台: DMControl、DMControl-GB、DCS 和 RL-ViGen, 并辨析它们之间的区别与联系. 此外, 还结合机器人控制、自动驾驶和多模态大模型等多个应用场景, 开展典型案例分析. 最后, 在总结现有研究基础上, 指出当前视觉强化学习面临的主要挑战, 并对未来发展方向进行展望.

### 参考文献

- 1 Sutton R S, Barto A G. *Reinforcement Learning: An Introduction* (Second edition). Cambridge: MIT Press, 2018.
- 2 Li B Y, Zhang Z N, Zheng G Z, Cai C R, Zhang J Q, Chen L. Cooperation in public goods games: Leveraging other-regarding reinforcement learning on hypergraphs. *Physical Review E*, 2025, **111**(1): Article No. 014304
- 3 Haarnoja T, Moran B, Lever G, Huang S H, Tirumala D, Humplik J, et al. Learning agile soccer skills for a bipedal robot with deep reinforcement learning. *Science Robotics*, 2024, **9**(89): Article No. eadi8022
- 4 Gao Yu-Ning, Wang An-Cheng, Zhao Hua-Kai, Luo Hao-Long, Yang Zi-Di, Li Jian-Sheng. Review on visual navigation methods based on deep reinforcement learning. *Computer Engineering and Applications*, 2025, **61**(10): 66-78  
(高宇宁, 王安成, 赵华凯, 罗豪龙, 杨子迪, 李建胜. 基于深度强化学习的视觉导航方法综述. *计算机工程与应用*, 2025, **61**(10): 66-78)
- 5 He X K, Huang W H, Lv C. Trustworthy autonomous driving

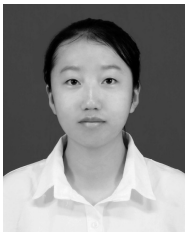
- via defense-aware robust reinforcement learning against worst-case observational perturbations. *Transportation Research Part C: Emerging Technologies*, 2024, **163**: Article No. 104632
- 6 Notice of the State Council on issuing the new generation of artificial intelligence development plan [Online], available: [http://www.gov.cn/zhengce/content/2017-07/20/content\\_5211996.htm](http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm), February 4, 2026  
(国务院关于印发新一代人工智能发展规划的通知 [Online], available: [http://www.gov.cn/zhengce/content/2017-07/20/content\\_5211996.htm](http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm), 2026-02-04)
- 7 Guidelines for the construction of a comprehensive standardization system for the national artificial intelligence industry (2024 Edition) [Online], available: <https://www.gov.cn/zhengce/zhengceku/202407/P020240702716282797987.pdf>, February 4, 2026  
(国家人工智能产业综合标准化体系建设指南 (2024 版) [Online], available: <https://www.gov.cn/zhengce/zhengceku/202407/P020240702716282797987.pdf>, 2026-02-04)
- 8 Mnih V, Kavukcuoglu K, Silver D, Rusu A A, Veness J, Bellemare M G, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, **518**(7540): 529–533
- 9 Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning. Stockholm, Sweden: PMLR, 2018. 1861–1870
- 10 Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv: 1707.06347, 2017.
- 11 Li Jia-Ning, Tian Yong-Hong. Recent advances in neuromorphic vision sensors: A survey. *Chinese Journal of Computers*, 2021, **44**(6): 1258–1286  
(李家宁, 田永鸿. 神经形态视觉传感器的研究进展及应用综述. 计算机学报, 2021, **44**(6): 1258–1286)
- 12 Ma G Z, Wang Z, Yuan Z C, Wang X Q, Yuan B, Tao D C. A comprehensive survey of data augmentation in visual reinforcement learning. *International Journal of Computer Vision*, 2025, **133**(10): 7368–7405
- 13 Eehchahed A, Castro P S. A survey of state representation learning for deep reinforcement learning. arXiv preprint arXiv: 2506.17518, 2025.
- 14 Yarats D, Fergus R, Lazaric A, Pinto L. Reinforcement learning with prototypical representations. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 11920–11931
- 15 Laskin M, Lee K, Stooke A, Pinto L, Abbeel P, Srinivas A. Reinforcement learning with augmented data. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1669
- 16 Yarats D, Kostrikov I, Fergus R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. In: Proceedings of the International Conference on Learning Representations. Vienna, Austria: OpenReview.net, 2021. Article No. 121
- 17 Yarats D, Fergus R, Lazaric A, Pinto L. Mastering visual continuous control: Improved data-augmented reinforcement learning. In: Proceedings of the 10th International Conference on Learning Representations. Virtual Event: OpenReview.net, 2022. Article No. 113
- 18 Lillicrap T P, Hunt J J, Pritzel A, Heess N, Erez T, Tassa Y, et al. Continuous control with deep reinforcement learning. In: Proceedings of the 4th International Conference on Learning Representations (ICLR). San Juan, Puerto Rico: ICLR, 2016.
- 19 Hansen N, Su H, Wang X L. Stabilizing deep Q-learning with ConvNets and vision Transformers under data augmentation. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. Virtual Event: Curran Associates Inc., 2021. Article No. 281
- 20 Almuzaire A, Hansen N, Christensen H I. A recipe for unbounded data augmentation in visual reinforcement learning. arXiv preprint arXiv: 2405.17416, 2024.
- 21 Xiong X, Shen C, Wu J H, Lv S, Zhang X D. Combined data augmentation framework for generalizing deep reinforcement learning from pixels. *Expert Systems With Applications*, 2025, **264**: Article No. 125810
- 22 Ma G Z, Zhang L R, Wang H Y, Li L, Wang Z L, Wang Z, et al. Learning better with less: Effective augmentation for sample-efficient visual reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2614
- 23 Lee J W, Hwang H. Fourier guided adaptive adversarial augmentation for generalization in visual reinforcement learning. In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, USA: AAAI Press, 2025. 18110–18118
- 24 Grooten B, Tomilin T, Vasan G, Taylor M E, Mahmood A R, Fang M, et al. MaDi: Learning to mask distractions for generalization in visual deep reinforcement learning. In: Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, 2024. 733–742
- 25 Bertoin D, Zouitine A, Zouitine M, Rachelson E. Look where you look! Saliency-guided Q-networks for generalization in visual reinforcement learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 2225
- 26 Zhu J H, Xia Y C, Wu L J, Deng J J, Zhou W G, Qin T, et al. Masked contrastive representation learning for reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023, **45**(3): 3421–3433
- 27 Grill J B, Strub F, Altché F, Tallec C, Richemond P H, Buchatskaya E, et al. Bootstrap your own latent a new approach to self-supervised learning. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 1786
- 28 Sun J R, Akcal M U, Chowdhary G, Zhang W. MOOSS: Mask-enhanced temporal contrastive learning for smooth state evolution in visual reinforcement learning. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). Tucson, USA: IEEE, 2025. 6719–6729
- 29 Liang A, Thomason J, Biyik E. ViSaRL: Visual reinforcement learning guided by human saliency. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 2907–2912
- 30 Sun J B, Tu S J, Zhang Q C, Chen K, Zhao D B. Saliency-invariant consistent policy learning for generalization in visual reinforcement learning. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems. Detroit, USA: International Foundation for Autonomous Agents and Multiagent Systems, 2025. 1987–1995
- 31 Ma J L, Li C, Feng Z Q, Xiao L M, He C D, Zhang Y. Don't overlook any detail: Data-efficient reinforcement learning with visual attention. *Knowledge-Based Systems*, 2025, **310**: Article No. 112869
- 32 Hafner D, Lillicrap T, Ba J, Norouzi M. Dream to control: Learning behaviors by latent imagination. In: Proceedings of the International Conference on Learning Representations (ICLR). Addis Ababa, Ethiopia: OpenReview.net, 2020. Article No. 120
- 33 Hu Ming-Fei, Zuo Xin, Liu Jian-Wei. Survey on deep generative model. *Acta Automatica Sinica*, 2022, **48**(1): 40–74

- (胡铭菲, 左信, 刘建伟. 深度生成模型综述. 自动化学报, 2022, 48(1): 40–74)
- 34 Zhang W P, Wang G, Sun J, Yuan Y T, Huang G. STORM: Efficient stochastic transformer based world models for reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 1182
- 35 Wang J J, Zhang Q C, Mu Y, Li D, Zhao D B, Zhuang Y Z, et al. Prototypical context-aware dynamics for generalization in visual control with model-based reinforcement learning. *IEEE Transactions on Industrial Informatics*, 2024, 20(9): 10717–10727
- 36 Poudel R P K, Pandya H, Liwicki S, Cipolla R. ReCoRe: Regularized contrastive representation learning of world model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 22904–22913
- 37 Deng F, Jang I, Ahn S. DreamerPro: Reconstruction-free model-based reinforcement learning with prototypical representations. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 4956–4975
- 38 Eyüboğlu M, Powell N, Karimi A. Data-driven control synthesis using Koopman operator: A robust approach. In: Proceedings of the American Control Conference (ACC). Toronto, Canada: IEEE, 2024. 1879–1884
- 39 Kumawat H, Chakraborty B, Mukhopadhyay S. RoboKoop: Efficient control conditioned representations from visual input in robotics using Koopman operator. In: Proceedings of the 8th Conference on Robot Learning. Seoul, South Korea: PMLR, 2025. 3474–3499
- 40 Hansen N, Su H, Wang X L. Temporal difference learning for model predictive control. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 8387–8406
- 41 Hansen N, Su H, Wang X L. TD-MPC2: Scalable, robust world models for continuous control. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. Article No. 130
- 42 Alonso E, Jelley A, Micheli V, Kanervisto A, Storkey A, Pearce T, et al. Diffusion for world modeling: Visual details matter in Atari. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2024. Article No. 1873
- 43 Zhou X X, Ying C Y, Feng Y, Su H, Zhu J. Self-consistent model-based adaptation for visual reinforcement learning. In: Proceedings of the 34th International Joint Conference on Artificial Intelligence (IJCAI). Montreal, Canada: IJCAI, 2025. 7191–7199
- 44 Hafner D, Pasukonis J, Ba J, Lillicrap T. Mastering diverse control tasks through world models. *Nature*, 2025, 640(8059): 647–653
- 45 Karamzade A, Kim K, Kalsi M, Fox R. Reinforcement learning from delayed observations via world models. In: Proceedings of the Reinforcement Learning Conference (RLC). Amherst, USA: RLJ, 2024. 2123–2139
- 46 Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. Safe reinforcement learning: A survey. *Acta Automatica Sinica*, 2023, 49(9): 1813–1835  
(王雪松, 王荣荣, 程玉虎. 安全强化学习综述. 自动化学报, 2023, 49(9): 1813–1835)
- 47 As Y, Usmanova I, Curi S, Krause A. Constrained policy optimization via Bayesian world models. In: Proceedings of the 10th International Conference on Learning Representations (ICLR). Virtual Event: OpenReview.net, 2022. Article No. 124
- 48 Hogewind Y, Simão T D, Kachman T, Jansen N. Safe reinforcement learning from pixels using a stochastic latent representation. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023. Article No. 114
- 49 Lee A X, Nagabandi A, Abbeel P, Levine S. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. Vancouver, Canada: Curran Associates Inc., 2020. Article No. 63
- 50 Radford A, Kim J W, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning transferable visual models from natural language supervision. In: Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR, 2021. 8748–8763
- 51 Doroudian E, Taghavifar H. CLIP-RLDrive: Human-aligned autonomous driving via CLIP-based reward shaping in reinforcement learning. arXiv preprint arXiv: 2412.16201, 2024.
- 52 Alayrac J B, Donahue J, Luc P, Miech A, Barr I, Hasson Y, et al. Flamingo: A visual language model for few-shot learning. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2022. Article No. 1723
- 53 Lin J, Du Y Q, Watkins O, Hafner D, Abbeel P, Klein D, et al. Learning to model the world with language. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 29992–30017
- 54 Aljalbout E, Sotirakis N, van der Smagt P, Karl M, Chen N. LIMT: Language-informed multi-task visual world models. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Atlanta, USA: IEEE, 2025. 8226–8233
- 55 Liu Z Y, Huan Z Y, Wang X Y, Lv J F, Tao J, Li X, et al. World models with hints of large language models for goal achieving. In: Proceedings of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies. Albuquerque, New Mexico: Association for Computational Linguistics, 2025. 50–72
- 56 Chen C, Xu J C, Liao W J, Ding H, Zhang Z Z, Yu Y, et al. Focus-then-decide: Segmentation-assisted reinforcement learning. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 11240–11248
- 57 Jiang H B, Lu Z Q. Visual grounding for object-level generalization in reinforcement learning. In: Proceedings of the 18th European Conference on Computer Vision (ECCV). Milan, Italy: Springer, 2024. 55–72
- 58 Li X H, Liu M H, Zhang H B, Yu C J, Xu J, Wu H T, et al. Vision-language foundation models as effective robot imitators. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. Article No. 119
- 59 Ma H Z, Sima K K, Vo T V, Fu D, Leong T Y. Reward shaping for reinforcement learning with an assistant reward agent. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 33925–33939
- 60 Escontrela A, Adeniji A, Yan W, Jain A, Peng X B, Goldberg K, et al. Video prediction models as rewards for reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 3009
- 61 Hung K H, Lo P C, Yeh J F, Hsu H Y, Chen Y T, Hsu W H. VICtoR: Learning hierarchical vision-instruction correlation rewards for long-horizon manipulation. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. Article No. 128
- 62 Rocamonde J, Montesinos V, Nava E, Perez E, Lindner D. Vision-language models are zero-shot reward models for reinforcement learning. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. Article No. 118
- 63 Wang Y F, Sun Z Y, Zhang J, Xian Z, Biyik E, Held D, et al.

- RL-VLM-F: Reinforcement learning from vision language foundation model feedback. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 51484–51501
- 64 Fu Y W, Zhang H C, Wu D, Xu W, Boulet B. FuRL: Visual-language models as fuzzy rewards for reinforcement learning. In: Proceedings of the 41st International Conference on Machine Learning. Vienna, Austria: PMLR, 2024. 14256–14274
- 65 Yu T H, Quillen D, He Z P, Julian R, Hausman K, Finn C, et al. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In: Proceedings of the Conference on Robot Learning. Osaka, Japan: PMLR, 2020. 1094–1100
- 66 Li J J, Wang Q, Wang Y B, Jin X, Li Y, Zeng W J, et al. Open-world reinforcement learning over long short-term imagination. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. Article No. 123
- 67 Laskin M, Srinivas A, Abbeel P. CURL: Contrastive unsupervised representations for reinforcement learning. In: Proceedings of the 37th International Conference on Machine Learning. Vienna, Austria: PMLR, 2020. 5639–5650
- 68 Zhang Chong-Sheng, Chen Jie, Li Qi-Long, Deng Bin-Quan, Wang Jie, Chen Cheng-Gong. Deep contrastive learning: A survey. *Acta Automatica Sinica*, 2023, **49**(1): 15–39 (张重生, 陈杰, 李岐龙, 邓斌权, 王杰, 陈承功. 深度对比学习综述. *自动化学报*, 2023, **49**(1): 15–39)
- 69 Pore A, Muradore R, Dall'Alba D. DEAR: Disentangled environment and agent representations for reinforcement learning without reconstruction. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Abu Dhabi, United Arab Emirates: IEEE, 2024. 650–655
- 70 Zheng W Q, Sharan S P, Fan Z W, Wang K, Xi Y H, Wang Z Y. Symbolic visual reinforcement learning: A scalable framework with object-level abstraction and differentiable expression search. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025, **47**(1): 400–412
- 71 Zhai Y P, Peng P X, Zhao Y F, Huang Y R, Tian Y H. Stabilizing visual reinforcement learning via asymmetric interactive cooperation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023. 207–216
- 72 Liu Yu-Xin, Xiang Liu-Yu, He Zhao-Feng, Wei Yun, Wu Hui-Jia, Wang Yong-Gang. State-action joint mask-based self-supervised learning algorithm. *Computer Technology and Development*, 2024, **34**(11): 125–132 (刘宇昕, 项刘宇, 何召锋, 魏运, 吴惠甲, 王永钢. 基于状态-动作联合掩码的自监督学习算法. *计算机技术与发展*, 2024, **34**(11): 125–132)
- 73 Huang S N, Zhang Z A, Liang T H, Xu Y H, Kou Z H, Lu C H, et al. MENTOR: Mixture-of-experts network with task-oriented perturbation for visual reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 26143–26161
- 74 Shi W J, Huang G, Song S J, Wu C. Temporal-spatial causal interpretations for vision-based reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022, **44**(12): 10222–10235
- 75 Yang R, Wang J, Geng Z J, Ye M X, Ji S W, Li B, et al. Learning task-relevant representations for generalization via characteristic functions of reward sequence distributions. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Washington, USA: Association for Computing Machinery, 2022. 2242–2252
- 76 Wang S, Wu Z H, Hu X B, Wang J W, Lin Y F, Lv K. What effects the generalization in visual reinforcement learning: Policy consistency with truncated return prediction. In: Proceedings of the 38th AAAI Conference on Artificial Intelligence. Vancouver, Canada: AAAI Press, 2024. 5590–5598
- 77 Dunion M, McInroe T, Luck K S, Hanna J P, Albrecht S V. Temporal disentanglement of representations for improved generalisation in reinforcement learning. In: Proceedings of the 11th International Conference on Learning Representations (ICLR). Kigali, Rwanda: OpenReview.net, 2023. Article No. 116
- 78 Wang R R, Cheng Y H, Wang X S. Clustering-driven state embedding for reinforcement learning under visual distractions. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2024, **54**(12): 7382–7395
- 79 Wang R R, Cheng Y H, Wang X S. Visual reinforcement learning control with instance-reweighted alignment and instance-dimension uniformity. *IEEE Transactions on Neural Networks and Learning Systems*, 2025, **36**(6): 9905–9918
- 80 Zheng R J, Wang X Y, Sun Y C, Ma S, Zhao J Y, Xu H Z, et al. TACO: Temporal latent action-driven contrastive loss for visual reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 2092
- 81 Yan M B, Lv J F, Li X. Enhancing visual reinforcement learning with state-action representation. *Knowledge-Based Systems*, 2024, **304**: Article No. 112487
- 82 Liu Min-Song, Zhu Yuan-Heng, Zhao Dong-Bin. State-action-reward prediction representation learning based on Transformer. *Acta Automatica Sinica*, 2025, **51**(1): 117–132 (刘民颂, 朱圆恒, 赵冬斌. 基于 Transformer 的状态-动作-奖赏预测表征学习. *自动化学报*, 2025, **51**(1): 117–132)
- 83 Choi H, Lee H, Song W, Jeon S, Sohn K, Min D B. Local-guided global: Paired similarity representation for visual reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Vancouver, Canada: IEEE, 2023. 15072–15082
- 84 Schwarzer M, Anand A, Goel R, Hjelm R D, Courville A, Bachman P. Data-efficient reinforcement learning with self-predictive representations. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2021. Article No. 118
- 85 Park S, Rybkin O, Levine S. METRA: Scalable unsupervised RL with metric-aware abstraction. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. Article No. 125
- 86 Liang D Y, Chen Q H, Liu Y L. Sequential action-induced invariant representation for reinforcement learning. *Neural Networks*, 2024, **179**: Article No. 106579
- 87 Cai F Z, Kwietniak D, Li J, Pourmand H. On the properties of the mean orbital pseudo-metric. *Journal of Differential Equations*, 2022, **318**: 1–19
- 88 Ferns N, Panangaden P, Precup D. Metrics for finite Markov decision processes. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. Banff, Canada: AUAI Press, 2004. 162–169
- 89 Gelada C, Kumar S, Buckman J, Nachum O, Bellemare M G. DeepMDP: Learning continuous latent space models for representation learning. In: Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA: PMLR, 2019. 2170–2179
- 90 Zhang A, McAllister R T, Calandra R, Gal Y, Levine S. Learning invariant representations for reinforcement learning without reconstruction. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2021. Article No. 117
- 91 Agarwal R, Machado M C, Castro P S, Bellemare M G. Contrastive behavioral similarity embeddings for generalization in reinforcement learning. In: Proceedings of the International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2021. Article No. 130
- 92 Liu Q Y, Zhou Q, Yang R, Wang J. Robust representation

- learning by clustering with bisimulation metrics for visual reinforcement learning with distractions. In: Proceedings of the 37th AAAI Conference on Artificial Intelligence. Washington, USA: AAAI Press, 2023. 8843–8851
- 93 Wang R R, Cheng Y H, Wang X S. Constrained visual representation learning with bisimulation metrics for safe reinforcement learning. *IEEE Transactions on Image Processing*, 2025, **34**: 379–393
- 94 Dumion M, McInroe T, Luck K S, Hanna J P, Albrecht S V. Conditional mutual information for disentangled representations in reinforcement learning. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 3509
- 95 Wang S, Wu Z H, Wang J W, Hu X B, Lin Y F, Lv K. How to learn domain-invariant representations for visual reinforcement learning: An information-theoretical perspective. In: Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI). Jeju, South Korea: IJCAI, 2024. 1389–1397
- 96 You B, Liu P Z, Liu H P, Peters J, Arenz O. Maximum total correlation reinforcement learning. In: Proceedings of the 42nd International Conference on Machine Learning. Vancouver, Canada: PMLR, 2025. 72677–72699
- 97 Yang R, Wang J, Peng Q J, Guo R B, Wu G P, Li B. Learning robust representations with long-term information for generalization in visual reinforcement learning. In: Proceedings of the 13th International Conference on Learning Representations (ICLR). Singapore: OpenReview.net, 2025. Article No. 129
- 98 Fan J M, Li W C. DRIBO: Robust deep reinforcement learning via multi-view information bottleneck. In: Proceedings of the 39th International Conference on Machine Learning. Baltimore, USA: PMLR, 2022. 6074–6102
- 99 Yuan Z C, Wei T M, Cheng S Q, Zhang G, Chen Y P, Xu H Z. Learning to manipulate anywhere: A visual generalizable framework for reinforcement learning. In: Proceedings of the 8th Conference on Robot Learning. Munich, Germany: PMLR, 2025. 1815–1833
- 100 Dumion M, Albrecht S V. Multi-view disentanglement for reinforcement learning with multiple cameras. In: Proceedings of the Reinforcement Learning Conference (RLC). Amherst, USA: RLJ, 2024. 498–515
- 101 Haramati D, Daniel T, Tamar A. Entity-centric reinforcement learning for object manipulation from pixels. In: Proceedings of the 12th International Conference on Learning Representations (ICLR). Vienna, Austria: OpenReview.net, 2024. Article No. 135
- 102 Yang S Z, Ze Y J, Xu H Z. MoVie: Visual model-based policy adaptation for view generalization. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 940
- 103 Mu T Z, Li Z Y, Strzelecki S W, Yuan X, Yao Y C, Liang L T, et al. When should we prefer state-to-visual dagger over visual reinforcement learning? In: Proceedings of the 39th AAAI Conference on Artificial Intelligence. Philadelphia, Pennsylvania: AAAI Press, 2025. 14637–14645
- 104 Wang W Y, Fang X Y, Hager G. Adapting image-based RL policies via predicted rewards. In: Proceedings of the 6th Annual Learning for Dynamics & Control Conference. Oxford, UK: PMLR, 2024. 324–336
- 105 Liu X, Chen Y R, Li H R, Li B Y, Zhao D B. Cross-domain random pretraining with prototypes for reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2025, **55**(5): 3601–3613
- 106 Ze Y J, Hansen N, Chen Y B, Jain M, Wang X L. Visual reinforcement learning with self-supervised 3D representations. *IEEE Robotics and Automation Letters*, 2023, **8**(5): 2890–2897
- 107 Wang Xue-Song, Wang Rong-Rong, Cheng Yu-Hu. A review of offline reinforcement learning based on representation learning. *Acta Automatica Sinica*, 2024, **50**(6): 1104–1128 (王雪松, 王荣荣, 程玉虎. 基于表征学习的离线强化学习方法研究综述. *自动化学报*, 2024, **50**(6): 1104–1128)
- 108 Rafailov R, Yu T H, Rajeswaran A, Finn C. Offline reinforcement learning from images with latent space models. In: Proceedings of the 3rd Conference on Learning for Dynamics and Control. Virtual Event: PMLR, 2021. 1154–1168
- 109 Shang J H, Kahatapitiya K, Li X, Ryoo M S. StARformer: Transformer with state-action-reward representations for visual reinforcement learning. In: Proceedings of the 17th European Conference on Computer Vision (ECCV). Tel Aviv, Israel: Springer, 2022. 462–479
- 110 Zhang Y R, Chen K Z, Liu Y L. Causal representation learning in offline visual reinforcement learning. *Knowledge-Based Systems*, 2025, **320**: Article No. 113565
- 111 Wang Q, Yang J M, Wang Y B, Jin X, Zeng W J, Yang X K. Making offline RL online: Collaborative world models for offline visual reinforcement learning. In: Proceedings of the 38th International Conference on Neural Information Processing Systems (NeurIPS). Vancouver, Canada: Curran Associates Inc., 2024. 97203–97230
- 112 Tassa Y, Doron Y, Muldal A, Erez T, Li Y Z, de Las Casas D, et al. DeepMind control suite. arXiv preprint arXiv: 1801.00690, 2018.
- 113 Hansen N, Wang X L. Generalization in reinforcement learning by soft data augmentation. In: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA). Xi'an, China: IEEE, 2021. 13611–13617
- 114 Stone A, Ramirez O, Konolige K, Jonschkowski R. The distracting control suite—A challenging benchmark for reinforcement learning from pixels. arXiv preprint arXiv: 2101.02722, 2021.
- 115 Yuan Z C, Yang S Z, Hua P, Chang C, Hu K Z, Xu H Z. RL-ViGen: A reinforcement learning benchmark for visual generalization. In: Proceedings of the 37th International Conference on Neural Information Processing Systems. New Orleans, USA: Curran Associates Inc., 2023. Article No. 295
- 116 Jiang J H, Yang Y X, Deng Y Q, Ma C L, Zhang J. BEVNav: Robot autonomous navigation via spatial-temporal contrastive learning in bird's-eye view. *IEEE Robotics and Automation Letters*, 2024, **9**(12): 10796–10802
- 117 Chen W, Rojas N. TraKDis: A Transformer-based knowledge distillation approach for visual reinforcement learning with application to cloth manipulation. *IEEE Robotics and Automation Letters*, 2024, **9**(3): 2455–2462
- 118 Zhao Jing, Pei Zi-Nan, Jiang Bin, Lu Ning-Yun, Zhao Fei, Chen Shu-Feng. Virtual tube visual obstacle avoidance for UAV based on deep reinforcement learning. *Acta Automatica Sinica*, 2024, **50**(11): 2245–2258 (赵静, 裴子楠, 姜斌, 陆宁云, 赵斐, 陈树峰. 基于深度强化学习的无人机虚拟管道视觉避障. *自动化学报*, 2024, **50**(11): 2245–2258)
- 119 Yang Lei, Lei Wei-Min, Zhang Wei. Reinforcement learning algorithm for visual auto-driving based space-time features. *Journal of Chinese Computer Systems*, 2023, **44**(2): 356–362 (杨蕾, 雷为民, 张伟. 融合时空特征的视觉自动驾驶强化学习方法. *小型微型计算机系统*, 2023, **44**(2): 356–362)
- 120 Jiang B, Chen S Y, Zhang Q, Liu W Y, Wang X G. AlphaDrive: Unleashing the power of VLMs in autonomous driving via reinforcement learning and reasoning. arXiv preprint arXiv: 2503.07608, 2025.
- 121 Liu Y Q, Qu T Y, Zhong Z S, Peng B H, Liu S, Yu B, et al. VisionReasoner: Unified reasoning-integrated visual perception via reinforcement learning. arXiv preprint arXiv: 2505.12081, 2025.
- 122 Li W Q, Zhang X Y, Zhao S J, Zhang Y B, Li J L, Zhang L, et al. Q-Insight: Understanding image quality via visual reinforcement learning. arXiv preprint arXiv: 2503.22679, 2025.

- 123 Miao Z C, Wang J, Wang Z, Yang Z Y, Wang L J, Qiu Q, et al. Training diffusion models towards diverse image generation with reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, USA: IEEE, 2024. 10844–10853
- 124 Kondrotas K, Zhang L, Lim C P, Asadi H, Yu Y H. Audio-visual emotion classification using reinforcement learning-enhanced particle swarm optimisation. *IEEE Transactions on Affective Computing*, 2025, **16**(4): 3434–3451
- 125 Xu Fan-Feng, Tong Ming-Lei. Visual-based reinforcement learning for digital chip global placement. *Application Research of Computers*, 2024, **41**(4): 1270–1274  
(徐樊丰, 仝明磊. 基于视觉强化学习的数字芯片全局布局方法. 计算机应用研究, 2024, **41**(4): 1270–1274)



**王荣荣** 中国矿业大学博士研究生. 2021 年获得济南大学硕士学位. 主要研究方向为深度强化学习.

E-mail: [wangrongrong1996@126.com](mailto:wangrongrong1996@126.com)

**(WANG Rong-Rong** Ph.D. candidate at China University of Mining and Technology. She received

her master degree from University of Jinan in 2021. Her main research interest is deep reinforcement learning.)

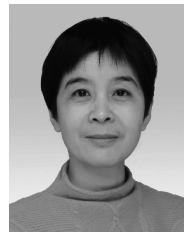


**程玉虎** 中国矿业大学教授. 2005 年获得中国科学院自动化研究所博士学位. 主要研究方向为机器学习, 智能系统.

E-mail: [chengyuhu@163.com](mailto:chengyuhu@163.com)

**(CHENG Yu-Hu** Professor at China University of Mining and

Technology. He received his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences in 2005. His research interests include machine learning and intelligent systems.)



**王雪松** 中国矿业大学教授. 2002 年获得中国矿业大学博士学位. 主要研究方向为机器学习, 模式识别. 本文通信作者.

E-mail: [wangxuesongcumt@163.com](mailto:wangxuesongcumt@163.com)

**(WANG Xue-Song** Professor at China University of Mining and

Technology. She received her Ph.D. degree from China University of Mining and Technology in 2002. Her research interests include machine learning and pattern recognition. Corresponding author of this paper.)