计 算 机 研 究 与 发 展
Journal of Computer Research and Development
https://crad.ict.ac.cn
63(5): 1273−1289, 2026

# 基于记忆增强和环境解耦的视觉强化学习表征优化

徐东红　武世龙
（中国矿业大学计算机科学与技术学院/人工智能学院　江苏徐州　221116）
（矿山数字化教育部工程研究中心（中国矿业大学）　江苏徐州　221116）
（地下空间智能感知与应急物联江苏省产业技术工程化中心（中国矿业大学）　江苏徐州　221116）
（xudh123@cumt.edu.cn）

# Visual Reinforcement Learning Representation Optimization Based on Memory Augmentation and Environment Disentanglement

Xu Donghong and Wu Shilong
（*School of Computer Science and Technology/ School of Artificial Intelligence*, *China University of Mining and Technology*, *Xuzhou*, *Jiangsu* 221116）
（*Mine Digitization Engineering Research Center of the Ministry of Education* (*China University of Mining and Technology*), *Xuzhou*, *Jiangsu* 221116）
（*Jiangsu Provincial Industrial Technology Engineering Center for Intelligent Sensing and Emergency IoT in Underground Space*（*China University of Mining and Technology*), *Xuzhou*, *Jiangsu* 221116）

**Abstract**　Visual reinforcement learning has demonstrated enormous potential across a wide range of domains. However, current algorithms still face two core challenges: insufficient generalization capability and low sample efficiency. Due to the limited availability of training data and the diversity and complexity of environments, reinforcement learning agents often become overly reliant on specific features present in the training settings, which hampers their ability to adapt to novel environments. Even minor environmental changes can induce significant alterations in image pixels, thereby affecting the latent representations learned by the agents and ultimately causing their policies to fail. To develop robust representations, we introduce the environment disentangled representation learning (EDRL) algorithm for visual reinforcement learning. It employs a self-supervised framework to extract environment-invariant features, thereby significantly enhancing the agent's generalization and sample efficiency. Firstly, by employing periodic data augmentation and integrating historical observations, the method simulates complex environmental changes and expands the effective observation range, thus reducing policy biases caused by training instability. Secondly, robust, environment-invariant features are obtained by disentangling latent features and isolating them through reconstruction. Finally, by predicting the changes in representations across time steps and incorporating a dynamic consistency loss, the approach ensures the consistency and robustness of the representations. Experimental results demonstrate the effectiveness of EDRL, as validated on the DMControl generalization benchmark (DMControl-GB) and the distracting control suite (DistractingCS). The results indicate that, compared with state-of-the-art methods, EDRL achieves an average performance improvement of over 15% in complex scenarios on DMControl-GB and over 20% in highly distracting environments on DistractingCS.

**Key words**　visual reinforcement learning; representation disentanglement; data augmentation; generalization; self-supervised learning

**收稿日期**: 2025-05-15; **修回日期**: 2025-12-15
**基金项目**: 国家自然科学基金项目(52374242); OPPO 研究资金项目(H7D220060)
This work was supported by the National Natural Science Foundation of China (52374242) and the OPPO Research Fund (H7D220060).
**通信作者**: 武世龙(ts23170090p31@cumt.edu.cn)

**摘 要** 视觉强化学习已在诸多领域中展现出了巨大的潜力。然而，现有的算法仍面临着两大核心挑战：泛化能力不足和样本效率低下。由于训练数据的有限性以及环境的多样性和复杂性，强化学习智能体常因过度依赖训练环境中的特定特征，从而难以适应新的未见环境。即便是微小的环境变化，也可能导致图像中像素的显著变化，影响智能体已学习到的潜在表征，进而使其学到的策略失效。为了学习更具鲁棒性的表征，提出一种基于记忆增强与环境解耦的视觉强化学习表征算法 EDRL（environment disentangled representation learning），旨在通过自监督学习框架提取环境无关的鲁棒表征，提升智能体的泛化能力和样本效率。首先，通过周期性数据增强和历史观测融合，模拟复杂环境变化并扩大有效观测范围，降低因训练不稳定性导致的策略偏差。其次，通过表征解耦与重构分离出环境无关的鲁棒表征。最后，通过预测时间步间的表征变化并引入动态一致性损失，确保表征的一致性与鲁棒性。在 DMControl-GB（deepmind control generalization benchmark）和 DistractingCS（distracting control suite）基准测试上的实验验证了 EDRL 的有效性。结果表明，相较于当前最优方法，EDRL 在 DMControl-GB 的复杂场景下平均性能提升 15% 以上，在 DistractingCS 的高干扰环境下平均性能提升 20% 以上。

**关键词** 视觉强化学习；表征解耦；数据增强；泛化性；自监督学习

**中图法分类号** TP391

**DOI:** 10.7544/issn1000-1239.202550320 **CSTR:** 32373.14.issn1000-1239.202550320

通过视觉信号进行强化学习是当前研究领域的一个重要方向，已在多个应用领域取得显著进展，包括视频游戏[1-3]、机器人控制[4]以及自主导航[5-6]等。视觉强化学习专注于从高维图像输入中学习最优控制策略[7-8]，然而，现有方法在应对未见环境时，其泛化能力表现出明显的不足。例如，模型通常会过度依赖训练环境中的特定特征（如位置分布、光照强度、物体颜色等）[9-10]，而无法适应具有干扰和变化的新环境。这种现象不仅限制了视觉强化学习在真实场景中的应用，也暴露了其在表征学习方面的局限性。如何使强化学习智能体学习到不受视觉变化影响且能够有效表达关键动态信息的图像表征，提高视觉强化学习的泛化能力，仍是该领域面临的一项核心挑战[11-13]。为了应对这一挑战，本文提出基于记忆增强与环境解耦的视觉强化学习表征算法 EDRL（environment disentangled representation learning）。通过周期性地应用多类型数据增强，并融合多步观测数据，最后结合表征解耦算法提取环境无关的鲁棒表征，有效提升了智能体在未见环境中的适应能力与泛化性能。

强化学习智能体难以适应未见环境，其主要原因之一是智能体在训练过程中容易过度依赖环境中的特定元素[7]。这一问题在视觉强化学习任务中尤为突出，因为环境变量的细微变化可能导致智能体接收到完全不同的像素分布，致使此前训练出的强化学习策略失去最优性。例如，在机器人控制任务中，训练环境可能包含一些对实际任务无关的视觉信息，如墙面颜色、摄像头本身的镜头污渍或其他固定背景元素。如果智能体在训练过程中误将这些元素作为判断自身位置或运动策略的重要依据，那么当部署到新环境中时，若这些视觉信息发生变化，智能体便无法正确判断，导致策略失效[14]。类似地，在视频游戏中，智能体可能会过度依赖环境中特定的光影效果、背景纹理或其他不具备任务相关性的装饰性元素，这些因素在游戏胜负中并不起决定性作用，但会被模型误认为是解决任务的关键线索[15]。一旦这些元素在新环境中改变，过拟合于这些无关元素的智能体就无法有效地调整策略，进而表现出较差的泛化能力。

为了解决这一问题，一种方法是在训练期间使用域随机化以最大化观察到的环境变化[7,16]。然而，在实际应用中，智能体在未来可能遇到的所有变化类型通常是未知的。即使能够提前知晓所有可能的变化，使用全部变化集合进行训练往往效率不高，且可能导致智能体学习到一种对多个环境的折中策略，从而无法实现真正的泛化。此外，由于实验或模拟环境的局限性，在训练过程中生成所有可能的环境变化是不可行的。领域随机化的另一种替代方法是学习一种能够泛化到未见变化的鲁棒表示[17]，但如何实现这一目标仍是一个开放问题[18]。已有研究尝试提取对图像中干扰因素不敏感的鲁棒表征[19-20]，用于改善智能体的泛化能力。然而，这些方法并未对表征中任务相关因素进行强制约束，因此在面对相关因素的未见值时泛化效果有限。本文所提方法通

过 3 个关键损失函数(表征解耦损失 $\mathcal{L}_{\text{fd}}$、特征重构损失 $\mathcal{L}_{\text{rec}}$、动态一致性损失 $\mathcal{L}_{\text{dyn}}$)，确保模型能够在不同环境下保持特征的准确性和一致性。

与域随机化类似，近期的研究表明，选择合适的数据增强方法同样可以显著提升强化学习策略的泛化能力[21-23]。数据增强通过对训练数据应用一系列变换，如旋转、缩放、裁剪等，来模拟环境变化，提高模型的鲁棒性。然而，现有的数据增强方法主要关注于单一类型的变换，这些变换无法全面反映现实环境中的复杂性和多样性。此外，多类型数据增强的融合方案往往会导致训练数据分布的不稳定，影响智能体的学习效果[24-25]。

另一类有效的方法是通过表征解耦来增强模型的鲁棒性。表征解耦的目标是寻找一个低维表征空间，从输入数据中分离出可解释的潜在因素。其中，特征向量的某些维度仅会随着对应潜在因素的变化而发生变化，而不受其他潜在因素变化的影响[26]。这使得强化学习智能体可以依赖未变部分快速适应环境变化，实现更强的泛化能力和性能恢复，这类似于基于状态的强化学习方法。Qian 等人[27] 和 Liu 等人[28] 的研究表明，解耦表征能够显著提升强化学习的泛化能力。然而，大多数解耦方法在处理输入数据时，未能充分利用多时间步的观测信息，仅依赖单帧或短时间窗口的输入进行表征建模，导致解耦出的特征只能反映当前时间步的状态信息，这使得智能体难以捕捉长期动态特征。在实际应用中，动态环境中的状态变化常常涉及较长时间跨度的关联信息，例如机器人在导航过程中，需要参考历史路径信息来正确规划下一步操作。观测范围的受限直接导致解耦模型无法有效分离出任务相关因素和环境相关干扰，从而在应对复杂动态环境时表现出泛化能力不足。此外，现有的表征解耦方法通常需要在弱监督或有监督的条件下进行，不仅增加了训练的复杂性，而且限制了其在实际应用中的适用范围。

为了应对上述问题，本文提出了 EDRL，如图 1 所示。这是一个用于提高视觉强化学习泛化能力的自监督框架，旨在从高维图像输入中提取环境无关的鲁棒特征，显著提升智能体在未见环境中的泛化能力和样本效率。

本文方法主要包括 3 个部分：

1)为了解决现有研究的多类型数据增强方案所存在的不稳定问题，以及大多数解耦方法在处理输入数据时没有充分利用多时间步的观测信息，导致解耦出的特征只能反映当前时间步的状态信息这一问题，提出一种记忆增强与融合算法，通过周期性地应用多类型数据增强操作，有效模拟复杂多变的真实环境，同时保持了训练过程的稳定性。其次，通过存储和融合多步前的增强观测，扩展了观测区域，充分捕捉长期动态特征。

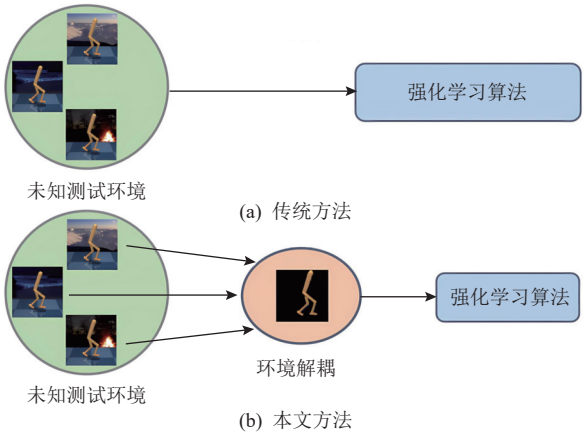



Fig. 1 Overall comparison of traditional methods and proposed method

图 1 传统方法与本文方法的总体比较

2)提出自适应表征解耦算法，通过自适应稀疏自注意力机制，有效过滤掉冗余信息和环境噪声，并通过表征解耦损失和表征重构损失优化解耦表示的独立性和表达能力。

3)为了解决连续动态场景中特征提取的不稳定性问题，本文提出动态一致性迁移算法，通过预测和对比连续时间步间的特征变化，约束特征的动态迁移过程，提升了模型在动态场景中的稳定性和鲁棒性。

通过在 DMControl-GB(DeepMind control generalization benchmark)[29] 和 DistractingCS(distracting control suite)[30] 基准测试上的实验，验证了 EDRL 的有效性。结果表明，在 DMControl-GB 测试环境下，EDRL 在不同的任务中均表现出更高的性能，尤其在任务复杂度较高的场景下，平均性能提升超过 15%，最高接近 30%。在 DistractingCS 测试环境下，EDRL 展现出对多类型视觉干扰(如背景变化、颜色扰动和摄像头角度调整)的强适应性。在干扰强度增加的条件下，EDRL 的性能下降幅度明显低于对比方法，表现出更稳定的泛化能力。与主流方法相比，EDRL 在复杂环境中的泛化能力和样本效率均有显著提升，尤其是在高噪声和强干扰的视觉场景中表现优异。

本文的主要贡献有 3 个方面：

1)为提高视觉强化学习的泛化能力设计了一种新的算法 EDRL，能够有效从高维视觉输入中提取鲁棒特征。此外，EDRL 的训练无需依赖手动标注的数

据，也不需要环境的先验知识，因此具有高度的通用性，能够无缝集成到多种主流的强化学习算法中。

2)为解决多类型数据增强方案的不稳定问题，提出一种记忆增强和融合算法，通过周期性应用多类型数据增强，模拟复杂的环境变化，同时减少因频繁改变数据分布而导致的训练不稳定性。此外，为了充分利用多时间步的观测信息，有效捕捉长期动态特征，本文通过存储和融合多步前的观测数据，扩展了智能体的有效观测区域，提高样本效率并增强模型的泛化能力。

3)通过广泛的实验，与主流方法相比较，EDRL展现了更强的泛化能力，特别是在高干扰和复杂任务场景中取得了显著的性能提升。

## 1 相关工作

### 1.1 视觉强化学习中的泛化问题

视觉强化学习中的泛化能力指的是，通过视觉信号训练的智能体在未知的环境或任务中仍能保持良好表现的能力[31]。尽管视觉强化学习算法在解决复杂任务方面取得了显著进展，但许多研究仍指出其泛化能力不足的问题[7-8,32]。Soualhi 等人[9]通过 Grid-World 迷宫环境的实验，发现视觉强化学习模型在某些情况下倾向于记住训练中的特定状态或任务，而不是学会泛化，这种现象在带有随机奖励的环境中尤其明显。Han 等人[8]使用归因方法揭示了“观察过拟合”现象，即智能体倾向于在训练过程中基于观察到的无关背景元素(例如游戏中的计时器或背景物体)做出决策，而非基于与任务相关的语义特征(如物体的位置和关系)。这会导致模型在测试环境中表现较差，因为这些无关特征无法有效泛化到新环境中。Weissenbacher 等人[32]将强化学习中的泛化问题定义为智能体在不同环境分布下的鲁棒性，并通过从这些分布中随机采样环境参数进行多域训练，促使智能体学习能够适应环境变化的鲁棒策略。为了应对上述研究所指出的视觉强化学习在未知环境下表现出的泛化能力不足的问题，本文的方法通过结合记忆增强与表征解耦技术，使得模型可以专注于与任务相关的语义特征，最终实现强化学习算法在未知环境下仍能保持较高的性能，提高模型的泛化能力。

### 1.2 数据增强

数据增强是一种通过对原始数据进行变换，生成新的数据样本以提高模型泛化能力的技术。它广泛应用于机器学习和深度学习，尤其是在计算机视觉领域。通过增加训练数据的多样性，数据增强可以有效防止模型过拟合，提高表示的鲁棒性。广泛的研究表明，数据增强同样可以显著提升视觉强化学习的泛化性能。Laskin 等人[21]设计了一种即插即用的 RAD 模块，旨在将数据增强推广至任何强化学习算法，他们通过广泛的实验对比分析了多种数据增强策略，并验证了这些方法在视觉强化学习中的优势。Yarats 等人[33]在 DrQ 方法中结合了多种图像增强技术，通过对每个图像进行多次随机变换，分别计算 Q 值和目标值，并对这些结果进行平均，以降低估计的方差并提高模型的稳定性。Hansen 等人[34]仅在 Q 值估计中应用数据增强，而不增强用于引导的目标值。Gorinski 等人[23]将以往的方法与 UCB[35]相结合，并设计了 3 种自动寻找有效增强的方法。Yuan 等人[36]提出了一种基于任务的 Lipschitz 数据增强方法 TLDA，区分并增强与任务无关的像素。Fan 等人[37]设计了 SECANT 算法，通过分阶段的自我专家克隆技术，结合弱增强和强增强策略，显著提升了视觉强化学习在未见过环境中的零样本泛化能力。

除了对原始输入进行数据增强以外，有些研究方法将增强操作直接在网络内部进行。Liu 等人[22]提出随机卷积算法，通过引入随机化神经网络来扰动视觉观测，增强智能体的泛化能力。Choi 等人[38]通过在卷积神经网络的深层特征图上进行局部像素位置交换，同时保持通道一致性，减少模型对无关视觉特征(如背景元素)的过度拟合，增强模型的泛化能力。Li 等人[39]基于 CrossNorm 和 SelfNorm 两种归一化技术，通过交换不同输入图像通道的均值和标准差，增加训练数据的多样性，从而扩展了模型的训练分布，同时学习标度函数和偏移函数，动态地调整特征的统计值，使模型在训练阶段和测试阶段都能更好地表现。

然而，现有的研究中通过单一类型的增强操作只能对数据进行特定维度的扰动，其变化范围和多样性有限，难以模拟现实环境中更复杂的视觉变化。而现有的多类型增强融合方案虽然能够提供更丰富的增强类型，但容易导致训练过程中的数据分布不稳定。强化学习的训练过程本身具有非稳态性，频繁的增强类型切换会加剧这种不稳定性，反而降低了样本效率。

### 1.3 自监督表征学习

自监督表征学习旨在从大规模的未标注图像或视频中学习通用特征，而无需进行繁琐昂贵的数据

标注。近年来，受到自监督学习在 NLP 和 CV 领域取得巨大成功的启发，研究人员尝试以自监督的方式学习视觉强化学习的表示。Laskin 等人[40]证明了对比学习能够有效增强基于像素输入的强化学习算法的数据效率，特别是在没有世界模型的情况下，也能够在多种复杂任务中达到最先进的性能，为强化学习中的自监督表示学习提供了新的方向。Ye 等人[41]提出通过时间对比学习来提高基于模型的强化学习的采样效率。

除了通过自监督的对比学习外，Stooke 等人[42]引入了一项新的无监督学习任务，将表示学习与策略学习分离。Hansen 等人[43]在训练阶段同时优化强化学习目标和自监督任务，并在部署阶段没有奖励信号的情况下，只通过自监督任务持续更新策略，使模型能够适应新环境的变化。Wang 等人[44]利用无监督的关键点检测和视觉注意力机制，生成与视觉干扰无关的前景掩码，为强化学习策略提供稳定、清晰的视觉输入。相比于通过自监督任务更新策略、关键点检测或视觉前景掩码生成的方法，本文的方法更加专注于在自监督的条件下，从原始图像中提取与环境无关的稳定特征。如图 1 所示，传统方法直接依赖原始特征输入，其中大量因素与任务无关，极易导致智能体在训练过程中出现过拟合问题。当部署到新环境时，这类无关因素的变化会引发策略失效。本文通过显式分离环境特征，保留与任务强相关且对环境变化不敏感的特征，从而降低智能体对外部噪声的依赖，提升模型的泛化能力和在新环境中的适应性。

### 1.4 表征解耦

表征解耦旨在将复杂问题中的不同因素分离出来，以增强模型的泛化能力和可解释性。通过这种方式，模型能够更准确地捕捉数据的内在结构，减少对噪声或偶然关联的依赖，提高在未见过的数据上的表现。在近几年的研究中，变分自编码器（VAE）是最常用的框架之一。例如：Shen 等人[45]和 Nademi 等人[46]设计了$\beta$-VAE，通过放大独立性约束，增强表征的解耦能力；Xie 等人[47]提出的 Factor-VAE 则通过增加一个惩罚项，使潜在表示的边缘分布更加因子化，以此促进解耦；Wang 等人[26]在训练期间增加了监督，以防止从独立同分布数据中学习解耦表征。然而，这些基于 VAE 的方法通常建立在假设变差因素之间相互独立的基础上，因此在处理相关因素的场景中存在显著局限性。

针对 VAE 在处理相关数据时的局限性，研究者们提出了一些改进算法[48-51]。Träuble 等人[48]对基于 VAE 的解耦技术在相关数据上的表现进行了分析，发现特征之间的相关性被编码到表示中。为此，他们提出了一种在训练中使用弱监督的方法，以在相关数据上学习解耦表征，同时通过适应性技术对潜在的纠缠特征进行“修正”。此外，Funke 等人[49]提出了基于条件互信息 CMI 的解耦方法，用于在有监督的环境中，通过学习解耦表征来改进相关性偏移下的泛化能力。其他近期研究利用行为相似性结合对比学习[50]或聚类[51]将行为相似的观察结果映射到相似的表示中。然而，这些方法在重建损失阶段需要额外的解码器网络，且表征解耦需要在弱监督或有监督的条件下进行。本文的方法在自监督的环境下进行表征解耦，同时无需在重建损失阶段引入额外的解码器网络。

## 2 方　　法

为了解决视觉强化学习在面对未知环境时泛化能力不足的问题，本文提出 EDRL，这是一个高效的自监督算法，旨在通过完全自监督的方式提升视觉强化学习的泛化能力。EDRL 通过 3 个关键算法提升视觉强化学习的泛化能力：记忆增强与融合算法（memory augmentation and fusion module, MAFM）、自适应表征解耦算法（adaptive feature decoupling module, AFDM）和动态一致性迁移算法（dynamic consistency transfer module, DCTM）。如图 2 所示，MAFM 首先对输入帧序列进行循环数据增强。为模拟人眼和记忆的工作方式，本文将 $K$ 步（本文预设的步数为 3，保证跨帧信息融合的效果，同时避免更大窗口带来的计算开销与响应滞后[21]）前的增强观测存储在缓冲池中，通过 CNN-LSTM 将增强后的部分观测组合成一个单一的观测，以此扩展有效观测区域，随后利用交叉注意力机制进一步提升特征的表达能力。在通过 MAFM 获取了丰富的时空特征表示后，AFDM 通过自适应稀疏自注意力和特征解码器对特征进行解耦，分离出环境无关的特征。最后，DCTM 通过动作感知迁移单元预测未来帧的环境无关特征，并通过动态一致性损失确保时间步间的特征一致性。

### 2.1 记忆增强与上下文感知

如图 2 所示，EDRL 从 2 组视觉输入 $O_t$ 和 $O_{t+1}$ 开始。为了提高样本效率并增强模型的泛化能力，现有的方法通常会对原始输入进行数据增强操作。单一类型的增强操作只能对数据进行特定维度的扰动，

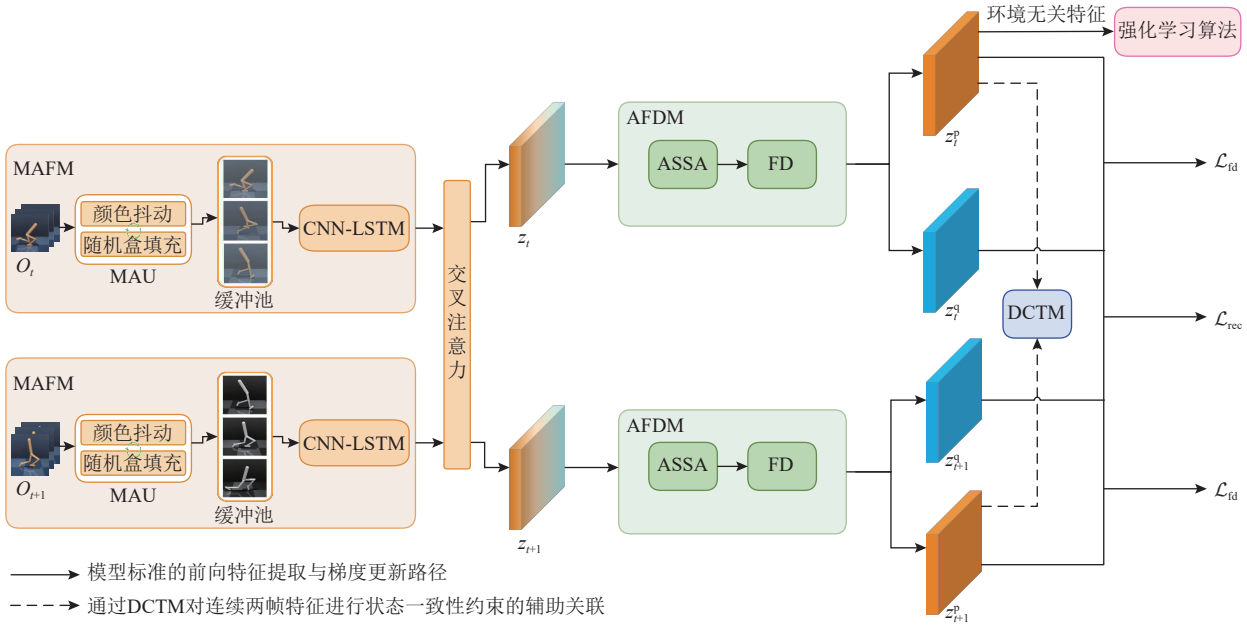



Fig. 2 Overall architecture of EDRL

图 2 EDRL 整体架构

其变化范围和多样性有限，难以模拟现实环境中更复杂的视觉变化。相比于单一数据增强，多类型数据增强不仅能够模拟更广泛的变化场景，还能在有限的数据下提供更丰富的信息，减少过拟合的风险，提高模型的泛化性。然而，现有的多类型数据增强融合方法存在明显的缺陷。组合式和混合式增强方法容易导致数据难度过高，给模型训练带来不稳定性[24]；而采样式增强方法虽然能够增加类型多样性，但频繁切换数据增强类型会导致数据分布的不稳定，进一步影响强化学习的收敛效果[25,52]。针对这些问题，本文提出记忆增强与融合算法，如算法 1 所示。MAFM 会对输入的帧序列进行数据增强处理，以此来假设一个新的环境，从而有效提高模型的泛化能力。与传统的随机或同时应用所有增强操作的方法不同，本文设计了一种新的记忆增强单元(memory augmentation unit, MAU)，它采用周期性应用策略。首先，每个 MAU 预设了一组有效的数据增强操作，这些操作可能包括颜色抖动、对比度调整、模糊处理等，每种操作都会对输入帧产生不同程度的变换，旨在模拟真实环境中可能出现的多种视觉变化。本文采用颜色抖动和随机盒填充作为 MAU 的预设增强操作。在每次训练更新中，每个 MAU 仅应用 1 种数据增强操作，然后按照预定的周期顺序切换到下一种操作(本文预设的循环间隔为 $10^5$ 步)。MAU 有助于在增加样本多样性的同时，保持数据分布的一致性，降低训练难度，减少因频繁改变数据分布而导致的训练不稳定性。

**算法 1.** 记忆增强与融合算法。

输入：帧序列 $O_t, O_{t+1}$；

输出：特征 $z_t, z_{t+1}$。

① 初始化数据增强操作集合 $A$ 和缓冲池 $B_1(K), B_2(K)$；

② 初始化 CNN-LSTM 和交叉注意力模型参数；

③ while 训练未完成 do

④ for 序列中的每一帧 do

⑤ 从 $A$ 中采样增强操作 $A_1, A_2$；

⑥ 对图像帧应用增强操作：$s_t^{aug_1} \leftarrow A_1(O_t), s_{t+1}^{aug_2} \leftarrow A_2(O_{t+1})$；

⑦ 将 $s_t^{aug_1}, s_{t+1}^{aug_2}$ 分别存入缓存 $B_1, B_2$；

⑧ if $B_1.size > K$ 或 $B_2.size > K$ then

⑨ 弹出缓存中最旧的帧；

⑩ end if

⑪ end for

⑫ 从缓存中取出最近 $K$ 帧并拼接：

⑬ $f_t \leftarrow concat(s_{t-K+1}^{aug_1}, \cdots, s_t^{aug_1})$；

⑭ $f_{t+1} \leftarrow concat(s_{t-K+2}^{aug_2}, \cdots, s_{t+1}^{aug_2})$；

⑮ 使用 CNN-LSTM 提取时空特征：

⑯ $f_t' \leftarrow CNN\text{-}LSTM(f_t)$；

⑰ $f_{t+1}' \leftarrow CNN\text{-}LSTM(f_{t+1})$；

⑱ 使用交叉注意力机制获得最终特征：

⑲ $z_t, z_{t+1} \leftarrow CrossAttention(f_t', f_{t+1}')$；

⑳ end while

受人类视觉系统记忆特性的启发，本文通过将缓冲池与数据增强相结合，实现记忆增强的效果。如算法 1 所示，MAFM 将 $K$ 步前的增强观测存储在缓冲池中，并通过模拟人眼和记忆的工作方式，将增强后的部分观测组合成一个单一的观测，以此扩展有效观测区域，并防止随着网络的加深和输入图像数量的增加而逐渐忘记早期输入的图像。与 Quan 等人[53]通过联合建模空间邻域和时间依赖性来提高复杂行为序列预测能力的方法类似，本文采用卷积神经网络（convolutional neural networks, CNN）和长短期记忆网络（long short-term memory, LSTM）的组合 CNN-LSTM[54]来处理增强后的帧序列。CNN 用于提取帧中的局部空间特征，它能够有效捕捉场景中的关键视觉信息，并将低级特征经过多个卷积层逐步抽象，转化为更加高级的表示形式。随后，经过 CNN 提取的特征会传递到 LSTM 中，LSTM 是一种适用于序列数据的网络结构，它能够处理时间依赖性并捕获帧序列之间的长期关联性。LSTM 的记忆单元可以存储过去帧的信息，并在处理当前帧时将这些历史信息纳入考虑。因此，LSTM 能够捕捉帧序列的时间维度特征，从而更好地理解视频中的动态变化。CNN-LSTM 不仅能够从图像帧中提取空间特征，还能够有效捕捉帧序列中的时间特性，为后续算法提供更加丰富的输入特征表示。

为了进一步提升特征的表达能力，MAFM 在利用 CNN-LSTM 提取空间和时间特征的基础上，引入了交叉注意力机制，交叉注意力机制的数学模型如式（1）所示，其中查询（$\boldsymbol{Q}$）来自其中一个输入序列，键（$\boldsymbol{K}$）和值（$\boldsymbol{V}$）来自另一个输入序列，$d_k$ 是 $\boldsymbol{K}$ 的维度。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = softmax\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}。 \tag{1}$$

这一机制使得模型能够学习不同帧之间的相互关系，加强它们之间的信息互动。通过动态调整每个时间步的特征权重，模型能够集中注意力在当前任务决策中最为关键的帧特征上，同时忽略那些不相关或冗余的信息。综上，MAFM 为后续的任务提供了更加丰富和鲁棒的特征表示 $z_t$ 和 $z_{t+1}$：

$$z_t = MAFM(O_{t-k+1},\cdots,O_t)。 \tag{2}$$

## 2.2 环境无关表征解耦

在 2.1 节，本文通过 MAFM 获取了丰富的时空特征表示 $z_t$ 和 $z_{t+1}$。然而，特征中仍可能包含与特定环境相关的冗余信息，影响模型在不同环境中的泛化能力。为此，本文将特征 $z_t$ 形式化地分解为

$$z_t = z_t^{\mathrm{p}} + z_t^{\mathrm{q}}, \tag{3}$$

其中：$z_t^{\mathrm{p}}$ 定义为环境无关特征，仅与任务动态相关，不随背景或外部扰动而改变；$z_t^{\mathrm{q}}$ 定义为环境相关特征，包含光照、背景纹理等外部噪声因素。为了从 $z_t$ 和 $z_{t+1}$ 中分离出环境无关的特征，本文提出算法 AFDM。AFDM 由自适应稀疏自注意力（adaptive sparse self-attention, ASSA）和特征解码器（feature decoder, FD）组成，最终由 FD 输出环境无关特征 $z_t^{\mathrm{p}}$。

### 2.2.1 自适应稀疏自注意力

传统 Transformer 中的自注意力机制采用密集的注意力关系，会计算所有查询和键的相关性，但并非所有查询和键之间都有显著的相关性，因此强制计算所有注意力分数会引入不必要的噪声和无关信息，影响环境无关表征解耦的效果。同时，密集的注意力关系还会导致冗余信息的传播，尤其是在图像的无关区域，这进一步削弱了网络捕捉重要特征的能力。为了解决上述问题，AFDM 引入机制 ASSA[55]。如图 3 所示，ASSA 通过自适应调整注意力权重，在全局特征空间中选择性地关注那些与任务目标高度相关的特征，而弱化或忽略环境噪声带来的影响。

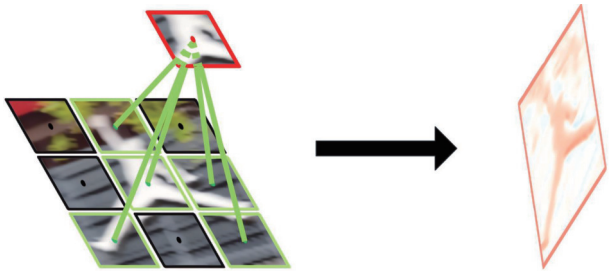


Fig. 3 ASSA selectively attends to task highly relevant features in the global feature space

图 3 ASSA 在全局特征空间中选择性地关注那些与任务目标高度相关的特征

与 Quan 等人[56]的双分支方法类似，ASSA 包括 2 个主要分支：稀疏自注意力（sparse self-attention, SSA）和密集自注意力（dense self-attention, DSA），如图 4 所示。SSA 使用了基于平方 ReLU（squared ReLU）的注意力机制，这种机制过滤掉了那些匹配分数较低的键-值对，仅保留有用的特征。这种稀疏机制有助于减少冗余信息和噪声干扰，使得模型能够更专注于有用的特征聚合。然而，完全依赖稀疏机制会导致信息丢失，尤其是在需要保留大量上下文信息的场景下。为了解决稀疏机制可能导致的信息丢失问题，ASSA 引入了 DSA 分支。该分支采用了传统的 Softmax 激活函数，确保在特征聚合过程中信息流

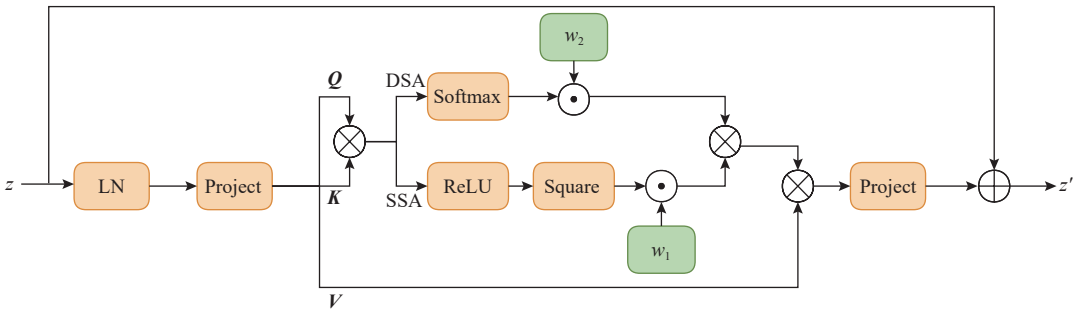



Fig. 4 Overall architecture of ASSA

图 4 ASSA 整体结构

的连续性，避免过度稀疏带来的信息丢失问题。为了在稀疏和密集自注意力之间找到最佳平衡，ASSA 通过自适应融合机制，动态调整 SSA 和 DSA 的贡献比例（$w_1$和$w_2$），以便根据输入特征的需求，灵活选择不同的注意力模式。ASSA 为每个分支分配了可学习的权重参数，通过这些权重实现对 2 个分支的自适应调节：

$$\boldsymbol{A}=(w_1\times SSA+w_2\times DSA)\boldsymbol{V}, \tag{4}$$

其中$w_1$和$w_2$是自适应学习的权重，用于控制 SSA 和 DSA 的相对重要性，$\boldsymbol{V}$是值矩阵。$w_1$和$w_2$的具体计算方法为：

$$w_n=\frac{\mathrm{e}^{a_n}}{\sum_{i=1}^{n}\mathrm{e}^{a_i}},n\in\{1,2\}, \tag{5}$$

其中$a_1,a_2$是可学习的参数，当模型检测到需要保留更多上下文信息时，DSA 的权重会增加，确保特征的全面性；而当冗余信息和噪声较多时，SSA 的权重则会增大，以过滤掉无关特征，实现更加精准的特征提取。通过自适应融合机制，ASSA 既能保证任务相关特征的有效提取，又能避免因过度稀疏化而导致的信息不足，使得模型在处理多样化的环境和复杂任务时具备更好的表征解耦能力。

2.2.2 表征解耦与重构

ASSA 将特征$z_t$和$z_{t+1}$中的冗余信息进行稀疏化处理，并初步保留了与任务相关的关键信息。在此基础上，FD 通过卷积操作解码并输出环境无关特征$z_t^{\mathrm{p}}$和$z_{t+1}^{\mathrm{p}}$，如图 2 所示。FD 采用轻量化的多层卷积结构，能够有效地剔除残留的环境相关信息，确保最终提取出专注于任务的环境无关信息。FD 的具体实现细节如表 1 所示，其中$z_t'$是 ASSA 的输出。

$$z_t^{\mathrm{p}}=FD(z_t), \tag{6}$$

$$z_t^{\mathrm{q}}=z_t-z_t^{\mathrm{p}}。 \tag{7}$$

基于环境无关特征的独立性，本文引入了表征解耦损失$\mathcal{L}_{\mathrm{fd}}$。通过最大化环境无关特征和环境特征之间的余弦距离，确保这 2 类特征的相互独立性和不相关性。

$$\mathcal{L}_{\mathrm{fd}}=\left|\left\|g(z_t^{\mathrm{p}})^{\mathrm{T}}g(z_t^{\mathrm{q}})\right\|_2\right|, \tag{8}$$

其中$g(\cdot)$代表全局平均池化操作，将特征图转换为向量。这种池化操作不仅减少了局部噪声的影响，还增强了特征在全局范围内的可分离性，使环境无关特征与环境相关特征能够在空间中保持清晰的界限。为了确保分离出的环境无关特征不仅具有独立性，还能准确保留输入信息的核心内容，本文设计了特征重构损失$\mathcal{L}_{\mathrm{rec}}$。通过将环境无关特征$z_t^{\mathrm{p}}$与环境相关特征$z_{t+1}^{\mathrm{q}}$重新组合，并使用均方误差（mean-square error，MSE）进行重构损失计算，来监督模型的重构精度，保证解耦出的特征组合能够还原原始特征。$\mathcal{L}_{\mathrm{rec}}$的具体公式为：

$$\mathcal{L}_{\mathrm{rec}}=\left\|(z_t^{\mathrm{p}}+z_{t+1}^{\mathrm{q}})-z_t'\right\|_2^2+\left\|(z_{t+1}^{\mathrm{p}}+z_t^{\mathrm{q}})-z_{t+1}'\right\|_2^2。 \tag{9}$$

其中$z_t'$和$z_{t+1}'$是$o_t$和$o_{t+1}$分别应用不同增强$A_2$和$A_1$后的特征。

特征重构损失函数能够确保模型在分离环境无关特征的同时，保留输入信息的主要结构和特征，使得模型能够在复杂环境下保持特征的准确性和一致性。

**Table 1 Implementation Details of FD**

**表 1 FD 的实现细节**

| 层数 | 配置 | 输入 | 输出 |
|---|---|---|---|
| 1 | Conv（$N16$，$K1$，$S1$）- BN - ReLU-DropOut | $z_t'$ | $z_{1t}$ |
| 2 | Conv（$N16$，$K1$，$S1$）- BN - ReLU-DropOut | $z_{1t}$ | $z_{2t}$ |
| 3 | Conv（$N32$，$K1$，$S1$）- BN | $z_{2t}$ | $z_t^{\mathrm{p}}$ |

### 2.3 动态一致性迁移

通过上述的 MAFM 与 AFDM，模型已经能够提取丰富的时空特征并有效分离环境无关特征。然而，在处理连续帧时，缺乏对特征变化的约束可能会导致分离的不稳定。为了解决这一问题，本文提出了

动态一致性迁移算法，通过预测和对比时间步间的特征变化，确保模型在动态场景中保持高效的特征迁移与一致性。

DCTM 的核心是利用当前的环境无关特征和动作信息，预测下一时刻的环境无关特征，在时间维度上对特征变化进行约束。如图 5 和算法 2 所示，给定当前时刻的环境无关特征 $z_t^{\mathrm{p}}$，DCTM 基于 $z_t^{\mathrm{p}}$ 和动作 $a_t$ 预测下一个时刻 $t$+1 的环境无关特征 $\hat{z}_{t+1}^{\mathrm{p}}$，这一预测由动作感知迁移单元（action aware transfer unit, AATU）完成，确保模型能够模拟状态的时间演化过程，帮助捕捉不同时间步间的特征迁移。AATU 的具体实现如表 2 所示，其核心流程是将当前时刻的环境无关特征与动作信息拼接后，经由卷积网络映射，得到对下一时刻环境无关特征的预测结果，在潜在空间中模拟动态转移过程。AATU 相当于一个轻量级的隐空间动力学预测器，通过建模动作对特征转移的调控作用，使得特征能够随动作变化而平滑演进，其目标是学习特征在时间维度上的动态变化规律，以及动作对特征转移的影响。

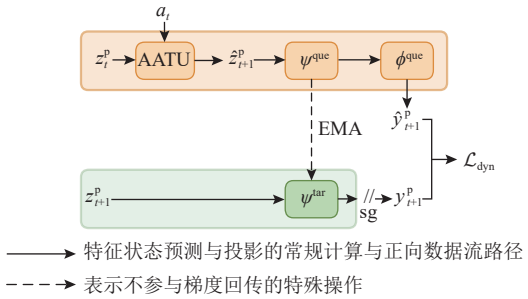



Fig. 5 Overall architecture of DCTM

图 5 DCTM 整体架构

**Table 2 Implementation Details of AATU**

**表 2 AATU 的实现细节**

| 层数 | 配置 | 输入 | 输出 |
|---|---|---|---|
| 1 | Concatenate | $z_t^{\mathrm{p}}, a_t$ | *conc* |
| 2 | Conv（*N*96，*K*3，*S*1）- ReLU - BN | *conc* | *conv*1*t* |
| 3 | Conv（*N*32，*K*3，*S*1）- ReLU | *conv*1*t* | $\hat{z}_{t+1}^{\mathrm{p}}$ |

**算法 2.** 动态一致性迁移算法。

输入：环境无关特征 $z_t^{\mathrm{p}}$，$z_{t+1}^{\mathrm{p}}$，动作 $a_t$。

① 初始化网络参数：AATU、查询编码器 $\psi^{\mathrm{que}}$、目标编码器 $\psi^{\mathrm{tar}}$、查询预测器 $\phi^{\mathrm{que}}$；

② 初始化 EMA 动量系数 $\tau$；

③ while 训练未完成 do

④ 通过 AATU 预测下一时刻的环境无关特征：

$$\hat{z}_{t+1}^{\mathrm{p}} \leftarrow AATU(z_t^{\mathrm{p}}, a_t);$$

⑤ 通过查询编码器与查询预测器得到预测特征：

$$\hat{y}_{t+1}^{\mathrm{p}} \leftarrow \phi^{\mathrm{que}}(\psi^{\mathrm{que}}(\hat{z}_{t+1}^{\mathrm{p}}));$$

⑥ 通过目标编码器得到目标特征：

$$y_{t+1}^{\mathrm{p}} \leftarrow \psi^{\mathrm{tar}}(z_{t+1}^{\mathrm{p}});$$

⑦ 计算动态一致性损失：

$$\mathcal{L}_{\mathrm{dyn}} \leftarrow -\frac{\langle \hat{y}_{t+1}^{\mathrm{p}}, y_{t+1}^{\mathrm{p}} \rangle}{\|\hat{y}_{t+1}^{\mathrm{p}}\|_2 \|y_{t+1}^{\mathrm{p}}\|_2};$$

⑧ 使用指数移动平均（EMA）更新目标编码器参数：

$$\xi_{\mathrm{tar}} \leftarrow \tau\xi_{\mathrm{tar}} + (1-\tau)\xi_{\mathrm{que}};$$

⑨ end while

在 AATU 完成预测后，查询编码器 $\psi^{\mathrm{que}}$ 和目标编码器 $\psi^{\mathrm{tar}}$ 分别负责对预测特征和目标特征进行编码，并将它们映射到一个低维潜在空间，从而提高计算效率并确保特征稳定性。编码后的预测特征通过查询预测器 $\phi^{\mathrm{que}}$ 进一步处理，使其能够更好地与目标特征对齐。目标编码器中的参数不会直接参与反向传播，而是采用 EMA 更新，即 $\xi_{\mathrm{tar}} \leftarrow \tau\xi_{\mathrm{tar}} + (1-\tau)\xi_{\mathrm{que}}$，其中 $\tau$ 为动量系数。EMA 将当前时刻参数与历史参数进行加权平均，实现训练过程中参数的平滑更新。与传统的梯度更新相比，EMA 能够有效抑制因单步梯度波动导致的剧烈变化，保证模型在训练过程中的连续性和稳定性。

最后，通过动态一致性损失确保特征一致性：

$$\mathcal{L}_{\mathrm{dyn}}(\hat{y}_{t+1}^{\mathrm{p}}, y_{t+1}^{\mathrm{p}}) = -\frac{\langle \hat{y}_{t+1}^{\mathrm{p}}, y_{t+1}^{\mathrm{p}} \rangle}{\|\hat{y}_{t+1}^{\mathrm{p}}\|_2 \|y_{t+1}^{\mathrm{p}}\|_2}, \tag{10}$$

其中

$$\hat{y}_{t+1}^{\mathrm{p}} = \phi^{\mathrm{que}}\left(\psi^{\mathrm{que}}\left(\hat{z}_{t+1}^{\mathrm{p}}\right)\right), \tag{11}$$

$$y_{t+1}^{\mathrm{p}} = \psi^{\mathrm{tar}}\left(z_{t+1}^{\mathrm{p}}\right)。 \tag{12}$$

$\mathcal{L}_{\mathrm{dyn}}$ 采用负余弦相似度，通过最大化 $\hat{y}_{t+1}^{\mathrm{p}}$ 和 $y_{t+1}^{\mathrm{p}}$ 之间的相似性，使得预测的环境无关特征 $\hat{z}_{t+1}^{\mathrm{p}}$ 和真实的环境无关特征 $z_t^{\mathrm{p}}$ 尽可能接近，从而增强 $z_t^{\mathrm{p}}$ 的时间一致性，确保其在动态场景下的稳定性。与 MSE 和 KL 散度相比，负余弦在 L2 归一化后具有尺度不变性，专注方向对齐；而 MSE 仅在单位范数时与余弦最大化等价，实际训练会额外惩罚模长差异并对幅度噪声更敏感；KL 散度则需要将特征概率化并引入温度等校准超参数，且当预测分布与真实分布的有效范围不重合时，KL 散度的损失值容易趋于无穷大，这会迫使模型为了强行拟合而输出极端且过度自信的预测[40]。

最终，对上述的损失函数进行加权求和，构建了最终的目标函数用于 EDRL 的训练。

$$\mathcal{L}_{\text{EDRL}} = \mathcal{L}_{\text{RL}} + \alpha\mathcal{L}_{\text{fd}} + \beta\mathcal{L}_{\text{rec}} + \gamma\mathcal{L}_{\text{dyn}}, \tag{13}$$

其中$\alpha$, $\beta$, $\gamma$是权重系数，用于在训练过程中控制不同损失函数的权重。

综上所述，EDRL 的伪代码如算法 3 所示。

**算法 3.** 基于记忆增强与环境解耦的视觉强化学习表征算法。

输入：帧序列$O_t, O_{t+1}$；

输出：环境无关特征$z_t^{\text{p}}$。

① 初始化数据增强操作集合$A$和缓冲池$B_1(K)$，$B_2(K)$；

② 初始化网络参数：*MAFM*, *ASSA*, *FD*, *DCTM*；

③ while 训练未完成 do

④ 通过算法 1 获取时空特征$z_t, z_{t+1}$：

$z_t \leftarrow MAFM(O_{t-K+1}, \cdots, O_t)$，

$z_{t+1} \leftarrow MAFM(O_{t-K+2}, \cdots, O_{t+1})$；

⑤ 通过 *ASSA* 和 *FD* 获得环境无关特征和环境特征：

$z_t^{\text{p}}, z_t^{\text{q}} \leftarrow FD(ASSA(z_t))$，

$z_{t+1}^{\text{p}}, z_{t+1}^{\text{q}} \leftarrow FD(ASSA(z_{t+1}))$；

⑥ 通过算法 2 预测下一时刻环境无关特征：

$\hat{y}_{t+1}^{\text{p}} \leftarrow DCTM(z_t^{\text{p}}, z_{t+1}^{\text{p}}, a_t)$；

⑦ 计算动态一致性损失$\mathcal{L}_{\text{dyn}}$；

⑧ 计算表征解耦损失$\mathcal{L}_{\text{fd}}$；

⑨ 计算特征重构损失$\mathcal{L}_{\text{rec}}$；

⑩ 最小化总损失$\mathcal{L}_{\text{EDRL}} = \mathcal{L}_{\text{dyn}} + \mathcal{L}_{\text{fd}} + \mathcal{L}_{\text{rec}}$；

⑪ 更新所有模块参数；

⑫ end while

## 3 实　　验

### 3.1 实验环境

本文的核心目标是提高智能体面临多变环境及视觉干扰时的泛化性能。与现有研究的实验方法类似[29,34,43-44]，本文先在初始环境中对智能体进行训练，然后在缺乏奖励信号和测试环境先验知识的条件下评估其性能。本文在连续控制基准 DMControl 上进行测试，包括 DMControl-GB[29] 和 DistractingCS[30]。这些基准测试由一系列具有不同复杂性和噪声水平的环境组成，能够全面评估智能体泛化到未知环境的能力。

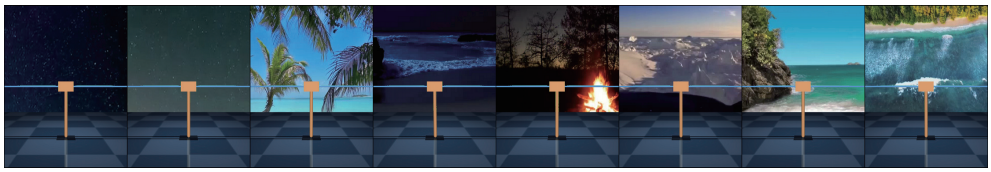
1）DMControl-GB。DMControl 是一套由 DeepMind 开发，并被广泛用于视觉强化学习研究的连续控制基准套件。它提供多个连续物理控制任务环境，每个任务有各自的特征和难度，但得分范围统一为 0~1000。如图 6(a)和图 6(b)所示，DMControl-GB 在此基础上通过引入额外的视觉扰动，如背景颜色随机化、视频背景替换等，模拟了更具挑战性的视觉场景。这些扰动帮助本文在标准控制任务中测试模型的鲁棒性，进一步揭示模型在应对视觉域转换时的表现。与原始的 DMControl 不同，DMControl-GB 侧重于测试智能体在训练环境与测试环境存在显著视觉差异情况下的表现，从而评估模型的泛化能力。本文通过 DMControl-GB 的 2 个不同级别的视觉干扰来测试模型的泛化能力，具体包括 video_easy 和 video_hard。在 video_easy 模式下，系统将从 1 组预设的视频库中随机选取背景视频，训练环境的背景被替换为相对简单且动态性较低的视频。这种设置旨在测试智能体在面对环境变化时，是否能够保持对任务的持续专注。虽然背景会变化，但地面和阴影保持不变，使得智能体可以更容易地专注于与任务相关的特征。而在 video_hard 模式下，系统将从一组更大的视频库中随机选取背景视频，不仅背景被替换为更复杂、动态性更强的视频，而且地面和阴影也会发生变化。这种测试环境增加了视觉复杂性，迫使智能体必须在复杂的背景噪声中区分出与任务相关的重要特征，它测试的是智能体在面对严重干扰时的泛化能力。

2）DistractingCS。DistractingCS 在 DMControl 的基础上进一步增强了视觉干扰的复杂性，不仅考虑了单一干扰类型的影响，还考察了多种干扰类型组合时智能体的表现，如图 6(c)和图 6(d)所示。DistractingCS 通过综合运用背景视频、颜色变化和相机姿态调整，创造出一个更具挑战性的视觉环境。DistractingCS 的测试难度由其强度设定决定，强度越高，环境变化越大，具体表现为：更快地调整摄像头角度，且角度变化更为剧烈；更快地改变智能体颜色，并且颜色变化的幅度更大；从更大的视频集合中随机选择背景。通过引入多种类型的视觉干扰，使得测试环境更加贴近现实，从而更加有效地评估算法在复杂任务中的鲁棒性和泛化能力。

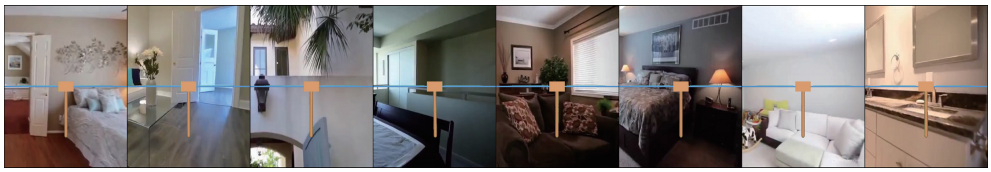
### 3.2 基线方法

本文将 EDRL 与以下基线方法进行比较，包括 SAC, DrQ, RAD, SODA, SVEA, SGQN, MaDi。

SAC[57] 结合了软策略和双 Q 网络的优势，通过最大化预期回报和最小化策略的熵来鼓励探索，从而提高学习效率。同时它还引入自适应温度参数，动态平衡奖励和熵，提升学习的稳定性和样本效率。



(a) DMControl-GB video_easy



(b) DMControl-GB video_hard



(c) DistractingCS（强度为0.1）



(d) DistractingCS（强度为0.5）

Fig. 6 Experimental environment

图 6 实验环境

该方法在复杂控制任务中表现优异，成为强化学习领域的重要方法之一。

DrQ[33] 是一种增强无模型强化学习性能的方法，通过应用数据增强技术（如随机图像平移）来减少过拟合，并提高样本效率。DrQ 将图像增强应用于输入观察值，同时对 Q 函数进行正则化，使得相同状态的不同变换具有相似的 Q 值。该方法可以与多种无模型强化学习算法结合使用，显著提升它们在像素级任务上的表现。

RAD[21] 通过在训练过程中对输入数据进行随机变换，迫使学习模型关注环境中的关键特征，在像素级控制任务和状态级控制任务中均实现了性能的显著提升。它作为一个即插即用的模块，能够与多种现有的强化学习算法无缝集成，且无需更改底层算法结构。

SODA[29] 旨在通过将数据增强与策略学习解耦来提升算法的泛化能力。SODA 使用未增强的数据进行策略优化，而将增强数据用于辅助的表示学习任务，通过最大化增强与非增强数据的潜在表征之间的互信息，促进编码器学习到在不同视图间共享的特征。这种方法不仅减少了对强化学习优化过程的干扰，而且显著提高了样本效率和训练稳定性，使得 SODA 在多个基准测试中实现了优异的泛化性能。

SVEA[34] 通过稳定 Q 值估计来缓解数据增强引发的不稳定性。该方法通过增强当前状态的 Q 值估计中应用数据，而不增强目标 Q 值，并通过同时优化增强与未增强数据，有效解决了因数据增强带来的 Q 值估计高方差和过度正则化问题。SVEA 显著提高了算法的稳定性和效率，特别是在视觉强化学习任务中。

SGQN[58] 通过显著性图引导模型关注输入图像中的关键像素，以提高泛化能力。SGQN 结合了一致性正则化和自监督学习，使得 Q 网络在遮罩图像和原始图像上保持一致性，减少对背景噪声的依赖，提升稳定性，确保模型能够在不同环境中泛化。SGQN 显著提升了在视觉扰动任务中的泛化能力，同时保持了较高的训练效率和模型的可解释性。

MaDi[59] 通过屏蔽视觉干扰来提升深度强化学习的泛化能力。它在传统的 Actor-Critic 架构中引入了一个轻量级的 Masker 网络，动态生成遮罩来过滤任务中无关的像素信息。MaDi 不需要额外的损失函数，仅依靠奖励信号来训练，显著减少了模型的复杂度。MaDi 在多个具有视觉干扰的基准任务和现实

机器人任务中表现优异，极大提升了智能体在未见过环境中的泛化能力。

为了确保比较的公平性和有效性，本文统一采用 Actor-Critic 架构实现包括 EDRL 在内的所有方法。所有方法均在无干扰的原始训练环境中进行 500 000 步的训练(DMControl-500k 通常被用于测量整体性能)，并且遵循 DMControl-GB 中为所有基线算法设定的默认超参数[29]。

### 3.3 实验结果与分析

#### 3.3.1 DMControl-GB

为了减少随机性对实验结果的影响，本文在 5 个不同的随机种子下进行训练，并最终计算结果的平均值及其标准差。表 3 和表 4 分别展示了 EDRL 及基线算法在 DMControl-GB 基准测试中，video_easy 和 video_hard 条件下的性能测试结果。实验结果显示，在 video_easy 条件下的 6 个测试环境中，EDRL 在 Walker stand, Walker walk, Ball-in-cup catch, Finger spin 四个环境中表现最佳，经过 500 000 个训练步骤后的平均回报相比基线算法 MaDi 提升了 7.3%。其中在相对简单的 Walker stand, Walker walk, Cartpole balance 任务中，EDRL 的性能提升并不明显或取得次优水平，但在任务相对困难的 Ball-in-cup catch 和 Finger spin 上，EDRL 的分数相比基线算法 MaDi 平均提升了约 15% 以上，可以看出 EDRL 在复杂任务上的性能提升较为明显。通过移除地面和阴影，video_hard 会造成更大、更混乱的分布偏移，从表 3 可以看出 EDRL 所有竞争对手的泛化能力都大幅下降，EDRL 受到的影响相对较小。在 video_hard 条件下的 6 个测试环境中，EDRL 在 Walker walk, Cartpole swingup, Ball in cup catch 和 Finger spin 四个环境中表现最佳，平均回报相比基线算法 MaDi 提升了 15.9%，取得了较大幅度的领先。与 video_easy 条件下实验结果类似，在任务相对困难的 Cartpole swingup, Ball in cup catch 和 Finger spin 上性能提升显著，提升范围在 5%~26%。由于 EDRL 将解耦后的特征输入到强化学习算法中，因此经过训练的智能体可以有效抵挡来自各种复杂环境的干扰。

#### 3.3.2 DistractingCS

图 7 展示了经过 500 000 步训练后，EDRL 和基线算法在 DistractingCS 基准测试下的 6 种不同环境中的泛化性能，结果取自 5 个不同随机种子的平均值。其中，强度表示变化的程度，强度越高，环境变

**Table 3 Experimental Scores of Different Algorithms Across Various Test Environments Under the Condition of video_easy**

**表 3 video_easy 条件下不同算法在各个测试环境下的实验得分**

| 测试环境 | SAC | DrQ | RAD | SODA | SVEA | SGQN | MaDi | EDRL（本文） | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|
| Walker，stand | 501±113 | 954±10 | 961±1 | 147±17 | 977±3 | 672±153 | 967±3 | **979±17** | +2 |
| Walker，walk | 334±37 | 821±38 | 726±42 | 479±168 | 936±14 | 882±26 | 895±24 | **939±29** | +3 |
| Cartpole，balance | 924±19 | 932±33 | 801±95 | 961±10 | 967±2 | 965±5 | **982±4** | 967±8 | −15 |
| Cartpole，swingup | 782±21 | 613±74 | 658±17 | 215±125 | 786±15 | 798±13 | **848±6** | 821±43 | −27 |
| Ball-in-cup，catch | 602±91 | 714±131 | 561±147 | 750±98 | 757±138 | 761±171 | 807±144 | **930±31** | +123 |
| Finger，spin | 227±26 | 543±50 | 479±65 | 429±100 | 645±39 | 592±11 | 679±17 | **924±13** | +245 |
| 平均值 | 563 | 763 | 698 | 497 | 845 | 778 | 863 | 926 | +331 |

注：黑体数值表示每个测试环境下的最优得分；$\Delta$ 表示每个测试环境下的最优得分与次优得分的分数差；“±”前后数值分别表示均值和标准差。

**Table 4 Experimental Scores of Different Algorithms Across Various Test Environments Under the Condition of video_hard**

**表 4 video_hard 条件下不同算法在各个测试环境下的实验得分**

| 测试环境 | SAC | DrQ | RAD | SODA | SVEA | SGQN | MaDi | EDRL（本文） | $\Delta$ |
|---|---|---|---|---|---|---|---|---|---|
| Walker，stand | 233±28 | 361±44 | 343±31 | 143±12 | 847±18 | 606±92 | **920±14** | 881±19 | −39 |
| Walker，walk | 168±19 | 111±16 | 116±18 | 214±68 | 526±55 | 718±37 | 504±33 | **753±27** | +35 |
| Cartpole，balance | 314±12 | 250±6 | 263±18 | 449±46 | 579±26 | 703±17 | **827±25** | 809±34 | −18 |
| Cartpole，swingup | 140±10 | 155±10 | 133±19 | 172±46 | 453±26 | 488±18 | 619±24 | **673±36** | +54 |
| Ball-in-cup，catch | 176±38 | 180±51 | 261±58 | 194±39 | 327±59 | 687±155 | 758±135 | **796±81** | +38 |
| Finger，spin | 21±4 | 16±3 | 8±2 | 107±30 | 154±31 | 554±8 | 358±25 | **701±18** | +147 |
| 平均值 | 175 | 179 | 187 | 213 | 481 | 626 | 664 | 770 | +217 |

注：黑体数值表示每个测试环境下的最优得分；$\Delta$ 表示每个测试环境下的最优得分与次优得分的分数差；“±”前后数值分别表示均值和标准差。

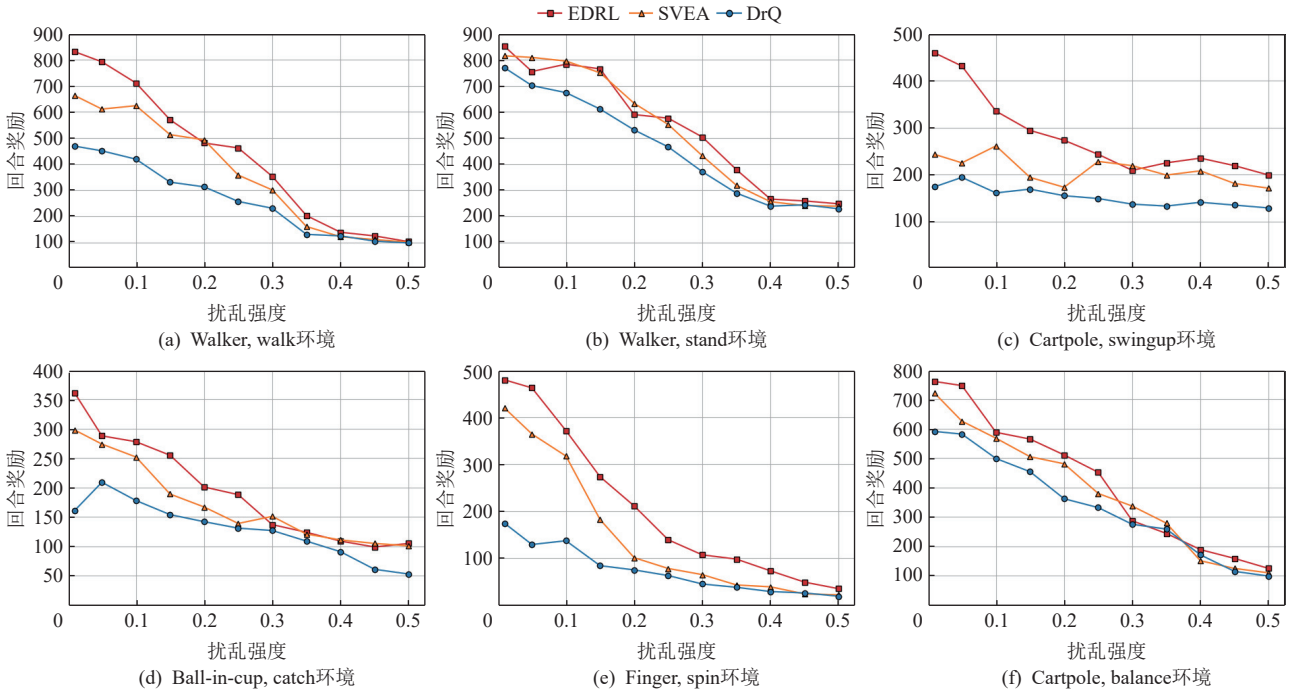



Fig. 7 Experimental results of DistractingCS testing

图 7 DistractingCS 测试下的实验结果

化越大。

如图 7 所示，当强度为 0.5 时，相机姿态会发生显著变化，在这种强烈的强度变化下，大多数方法都无法正常工作。实验结果表明，DistractingCS 各任务难度存在显著差异，而 EDRL 在所有任务上都与强基线算法 SVEA 相媲美甚至领先，对比基线算法 DrQ 则在所有任务上都取得较大幅度领先。在低强度变化的条件下，EDRL 的平均提升超过了 25%，在 Cartpole swingup 任务中提升甚至超过了 50%，并且随着强度的增加，其性能下降的速度相较于基线算法更为缓慢，这确保了 EDRL 在不同条件下的高鲁棒性。

### 3.3.3 消融实验

为了研究本文提出的 3 个关键损失函数（特征分解 $\mathcal{L}_{\text{fd}}$、重构 $\mathcal{L}_{\text{rec}}$ 和状态转移 $\mathcal{L}_{\text{dyn}}$）和循环数据增强对模型泛化性能的影响以及 MAFM 模块中引入的 CNN-LSTM 结构的有效性，本文在 DMControl-GB 基准测试中的 video_easy 条件下基于 5 个不同的随机种子进行了一系列消融实验，实验结果汇总于图 8、图 9 和图 10 中。其中纵轴为平均奖励，与智能体的任务完成质量呈正相关，质量越高，数值越大，理论范围为 [0, 1 000]。

如图 8 所示，仅采用单一损失函数进行训练的智能体相较于采用 2 个或更多损失函数训练的智能体，在性能上存在明显差距，但其性能仍可与基线算法 DrQ，RAD 相媲美甚至领先基线算法 SAC。特别是仅使用表征解耦损失 $\mathcal{L}_{\text{fd}}$ 的智能体表现出了较高的

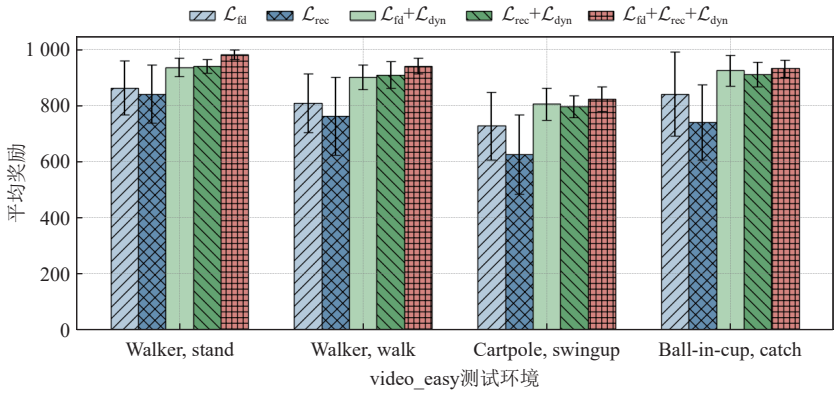



Fig. 8 Ablation experimental results for the three loss functions

图 8 针对 3 个损失函数的消融实验结果

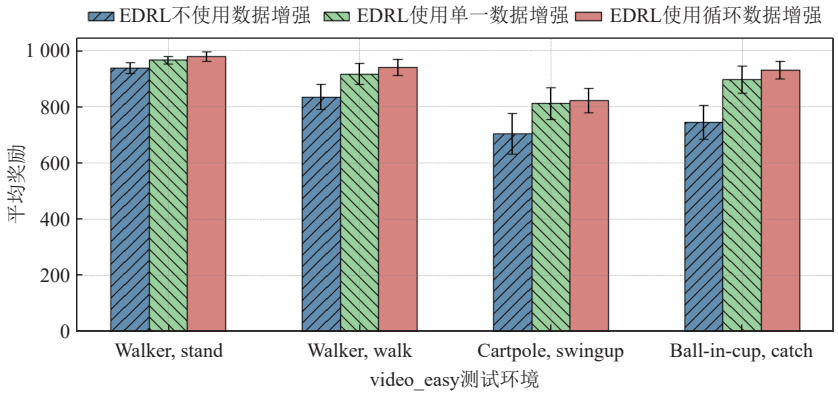



Fig. 9 Ablation experimental results for cyclic data augmentation

图 9 针对循环数据增强的消融实验结果

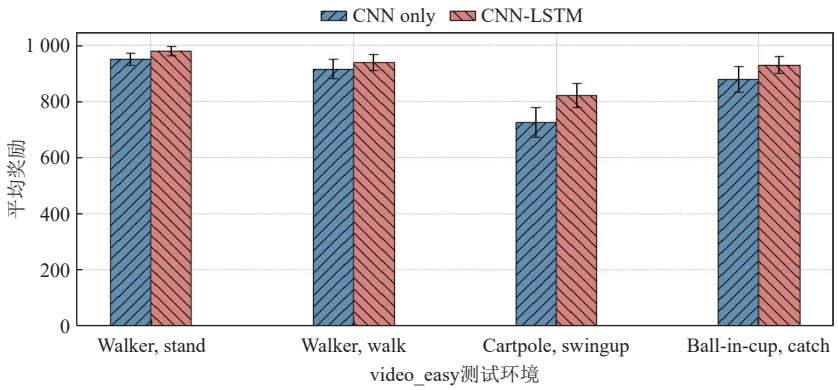



Fig. 10 Ablation experimental results for CNN-LSTM

图 10 针对 CNN-LSTM 的消融实验结果

性能提升。这表明，即便没有其他约束条件，EDRL仅通过$\mathcal{L}_{\mathrm{fd}}$也能够在一定程度上学习到环境无关特征。其次，当不使用状态转移损失$\mathcal{L}_{\mathrm{dyn}}$时，智能体的性能波动较大，这说明$\mathcal{L}_{\mathrm{dyn}}$有助于提升智能体的泛化稳定性。此外，$\mathcal{L}_{\mathrm{fd}}$和$\mathcal{L}_{\mathrm{rec}}$两个损失函数的性能提升效果相当，这可能是因为它们均旨在实现特征的有效解耦和分离。最终，当 3 个损失函数联合使用时，智能体达到了最佳性能。

图 9 展示了 EDRL 在不使用数据增强、使用单一数据增强和使用循环数据增强 3 种情况下的性能表现。结果表明，在未进行数据增强的情况下，模型的性能和稳定性在大多数任务中均有较大的损失。在使用单一数据增强的情况下，性能相比不使用数据增强有较大的提升，但仍与使用循环数据增强时的性能有一定差距。使用循环数据增强的模型在性能和稳定性上均有不同程度的领先。这一领先在复杂程度较高的任务中更为明显。这一结果进一步证明了本文提出的循环数据增强方法在减少过拟合、改善训练的稳定性以及提高模型泛化能力方面的有效性。

图 10 展示了 MAFM 模块分别采用 CNN 和 CNN-LSTM 两种结构处理增强帧序列后 EDRL 的性能对比结果。结果表明，CNN-LSTM 在所有任务中的分数均优于仅使用 CNN 的结构，说明在跨时间步建模中引入 LSTM 对性能具有积极作用。特别是在 Cartpole swingup 等动态变化较为剧烈的任务中，CNN-LSTM 相较于 CNN 显著提升了平均奖励，并有效降低了结果的波动性，表明其在建模时序依赖与提升训练稳定性方面具有明显优势。在其他任务中，CNN-LSTM 始终保持优于 CNN 的表现，进一步验证了在 MAFM 模块中引入 LSTM 的合理性和必要性。

## 4 结 论

本文针对视觉强化学习智能体在面对未知环境时泛化能力不足的问题，提出了一种基于记忆增强和环境解耦的表征算法 EDRL，从高维视觉输入中提取鲁棒的环境无关特征，显著提升了智能体在复杂和多变环境下的泛化能力和样本效率。所提出的算法首先通过多类型的数据增强和多步前的观测融合，模拟了真实环境中的复杂变化，并扩展了智能体的

观测能力，增强了模型应对环境波动和复杂性变化的鲁棒性。其次，通过自适应表征解耦算法将环境无关特征与环境相关表征解耦，增强了特征的独立性和表达能力。最后，动态一致性迁移算法通过预测连续时间步的特征变化，优化了特征迁移过程，增强了模型在动态环境中的稳定性。本文通过 DMControl-GB 和 DistractingCS 基准测试验证了 EDRL 的有效性。在 DMControl-GB 场景下平均性能提升 15% 以上，在 DistractingCS 的高干扰环境下平均性能提升 20% 以上，验证了 EDRL 在复杂多变与强干扰的视觉任务中具备更强的泛化能力与鲁棒性。

尽管本文所提算法在提高视觉强化学习的泛化能力方面取得了显著成果，但由于自适应稀疏自注意力机制和动态一致性迁移算法的引入不可避免地增加了计算开销。因此，如何优化这些算法，减少计算成本，提升模型的运行效率，是未来研究内容的重要方向。

**作者贡献声明**：徐东红提出了系统设计、解决方案、算法思路和实验方案，进行了论文的多次审阅与修订；武世龙进行了论文的撰写，完成了实验验证。

## 参 考 文 献


[1] Zhao Jing, Pei Zinan, Jiang Bin, et al. Virtual tube visual obstacle avoidance for UAV based on deep reinforcement learning[J]. Acta Automatica Sinica, 2024, 50(11): 2245−2258 (in Chinese)
(赵静, 裴子楠, 姜斌, 等. 基于深度强化学习的无人机虚拟管道视觉避障[J]. 自动化学报, 2024, 50(11): 2245−2258)

[2] Yao Hongge, Zhang Wei, Yang Haoqi, et al. Union regression object localization based on deep reinforcement learning[J]. Acta Automatica Sinica, 2023, 49(5): 1089−1098 (in Chinese)
(姚红革, 张玮, 杨浩琪, 等. 深度强化学习联合回归目标定位[J]. 自动化学报, 2023, 49(5): 1089−1098)

[3] Noh S, Kim S, Jang I. Exploring visual reinforcement learning for sample-efficient robotic manipulation[C] //Proc of the 15th Int Conf on Information and Communication Technology Convergence. Piscataway, NJ: IEEE, 2024: 207−210

[4] Zhang Yi, Feng Wei, Yang Zhile, et al. Visual navigation of mobile robots in complex environments based on distributed deep reinforcement learning[C] //Proc of the 6th Asian Conf on Artificial Intelligence Technology. Piscataway, NJ: IEEE, 2022: 1−5

[5] Wang Shuo, Wu Zhihao, Hu Xiaobo, et al. Skill-based hierarchical reinforcement learning for target visual navigation[J]. IEEE Transactions on Multimedia, 2023, 25: 8920−8932

[6] Xiao Wendong, Yuan Liang, He Li, et al. Multigoal visual navigation with collision avoidance via deep reinforcement learning[J]. IEEE Transactions on Instrumentation and Measurement, 2022, 71: 1−9

[7] Zeng Junjie, Qin Long, Xu Haotian, et al. Exploration approaches in deep reinforcement learning based on intrinsic motivation: A review[J]. Journal of Computer Research and Development, 2023, 60(10): 2359−2382 (in Chinese)
(曾俊杰, 秦龙, 徐浩添, 等. 基于内在动机的深度强化学习探索方法综述[J]. 计算机研究与发展, 2023, 60(10): 2359−2382)

[8] Han Changlin, Peng Zhiyong, Liu Yadong, et al. Overfitting-avoiding goal-guided exploration for hard-exploration multi-goal reinforcement learning[J]. Neurocomputing, 2023, 525: 76−87

[9] Soualhi T, Crombez N, Lombard A, et al. Leveraging motion perceptibility and deep reinforcement learning for visual control of nonholonomic mobile robots[J]. Robotics and Autonomous Systems, 2025, 189: 104920

[10] Tian Yantao, Cao Xuanhao, Huang Kai, et al. Learning to drive like human beings: A method based on deep reinforcement learning[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(7): 6357−6367

[11] Sun Jiarui, Akcal M U, Chowdhary G, et al. MOOSS: Mask-enhanced temporal contrastive learning for smooth state evolution in visual reinforcement learning[C] //Proc of the 2025 IEEE/CVF Winter Conf on Applications of Computer Vision. Piscataway, NJ: IEEE, 2025: 6719−6729

[12] Schneider M, Krug R, Vaskevicius N, et al. The surprising ineffectiveness of pre-trained visual representations for model-based reinforcement learning[J]. Advances in Neural Information Processing Systems, 2024, 37: 32916−32946

[13] Mu Tongzhou, Li Zhaoyang, Strzelecki S W, et al. When should we prefer state-to-visual dagger over visual reinforcement learning?[C] //Proc of the 39th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2025: 14637−14645

[14] Fu Yuwei, Zhang Haichao, Wu Di, et al. FuRL: Visual-language models as fuzzy rewards for reinforcement learning[C] //Proc of the 42nd Int Conf on Machine Learning. New York: PMLR, 2024: 14256−14274

[15] Sarkar S, Babu A R, Mousavi S, et al. Robustness and visual explanation for black box image, video, and ecg signal classification with reinforcement learning[C] //Proc of the 38th AAAI Conf on Artificial Intelligence. Palo Alto, CA: AAAI, 2024: 23829−23831

[16] Huang Jiaxing, Guan Dayan, Xiao Aoran, et al. FSDR: Frequency space domain randomization for domain generalization[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6891−6902

[17] Beigomi B, Zhu Z H. Towards real-world efficiency: Domain randomization in reinforcement learning for pre-capture of free-floating moving targets by autonomous robots[C] //Proc of the 41st IEEE Int Conf on Robotics and Automation. Piscataway, NJ: IEEE, 2024: 11753−11759

[18] Pore A, Muradore R, Dall'Alba D. DEAR: Disentangled environment and agent representations for reinforcement learning without reconstruction[C] //Proc of the 2024 IEEE/RSJ Int Conf on Intelligent Robots and Systems. Piscataway, NJ: IEEE, 2024: 650−655

[19] Zhang A, Lyle C, Sodhani S, et al. Invariant causal prediction for block MDPs[C] //Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2020: 11214−11224

[20] Zhang A, McAllister R, Calandra R, et al. Learning invariant representations for reinforcement learning without reconstruction[J]. arXiv preprint, arXiv: 2006.10742, 2020

[21] Laskin M, Lee K, Stooke A, et al. Reinforcement learning with augmented data[J]. Advances in Neural Information Processing Systems, 2020, 33: 19884−19895

[22] Liu Sicong, Zhang X S, Li Yushuo, et al. On the data-efficiency with contrastive image transformation in reinforcement learning[C] //Proc of the 11th Int Conf on Learning Representations. Vancouver, CA: ICLR, 2023: 17307−17319

[23] Gorinski P J, Zimmer M, Lampouras G, et al. Automatic unit test data generation and actor-critic reinforcement learning for code synthesis[J]. arXiv preprint, arXiv: 2310.13669, 2023

[24] Lee Y J, Kim J, Park Y J, et al. Masked and inverse dynamics modeling for data-efficient reinforcement learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5): 8814−8827

[25] Müller S G, Hutter F. Trivialaugment: Tuning-free yet state-of-the-art data augmentation[C] //Proc of the 34th IEEE/CVF Int Conf on Computer Vision. Piscataway, NJ: IEEE, 2021: 774−782

[26] Wang Yifan, Luo Xiao, Chen Chong, et al. DisenSemi: Semi-supervised graph classification via disentangled representation learning[J]. IEEE Transactions on Neural Networks and Learning Systems, 2025, 36(5): 8192−8204

[27] Qian Zhifeng, You Mingyu, Zhou Hongjun, et al. Weakly supervised disentangled representation for goal-conditioned reinforcement learning[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2202−2209

[28] Liu Shunyu, Song Jie, Zhou Yihe, et al. Interaction pattern disentangling for multi-agent reinforcement learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(12): 8157−8172

[29] Hansen N, Wang Xiaolong. Generalization in reinforcement learning by soft data augmentation[C] //Proc of the 41st IEEE Int Conf on Robotics and Automation. Piscataway, NJ: IEEE, 2021: 13611−13617

[30] Stone A, Ramirez O, Konolige K, et al. The distracting control suite —A challenging benchmark for reinforcement learning from pixels[J]. arXiv preprint, arXiv: 2101.02722, 2021

[31] Kirk R, Zhang A, Grefenstette E, et al. A survey of zero-shot generalisation in deep reinforcement learning[J]. Journal of Artificial Intelligence Research, 2023, 76: 201−264

[32] Weissenbacher M, Agarwal R, Kawahara Y. SiT: Symmetry-invariant transformers for generalisation in reinforcement learning[J]. arXiv preprint, arXiv: 2406.15025, 2024

[33] Yarats D, Kostrikov I, Fergus R. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels[C] //Proc of the 9th Int Conf on Learning Representations. Vancouver, CA: ICLR, 2021: 2517−2537

[34] Hansen N, Su Hao, Wang Xiaolong. Stabilizing deep q-learning with convnets and vision transformers under data augmentation[J]. Advances in Neural Information Processing Systems, 2021, 34: 3680−3693

[35] Kolarijani M A S, Proskurnikov A V, Esfahani P M. Macroscopic noisy bounded confidence models with distributed radical opinions [J]. IEEE Transactions on Automatic Control, 2020, 66(3): 1174−1189

[36] Yuan Zhecheng, Ma Guozheng, Mu Yao, et al. Don't touch what matters: Task-aware Lipschitz data augmentation for visual reinforcement learning[C] //Proc of the 31st Int Joint Conf on Artificial Intelligence. Palo Alto, CA: IJCAI, 2022: 3702–3708

[37] Fan Linxi, Wang Guanzhi, Huang Dian, et al. SECANT: Self-expert cloning for zero-shot generalization of visual policies[C] //Proc of the 39th Int Conf on Machine Learning (ICML 2021). New York: PMLR, 2021: 3088–3099

[38] Choi H, Lee H, Song W, et al. Local-guided global: Paired similarity representation for visual reinforcement learning[C] //Proc of the 32nd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2023: 15072−15082

[39] Li Lu, Lyu Jiafei, Ma Guozheng, et al. Normalization enhances generalization in visual reinforcement learning[C] //Proc of the 23rd Int Conf on Autonomous Agents and Multiagent Systems. Richland, SC: IFAAMAS, 2024: 1137–1146

[40] Laskin M, Srinivas A, Abbeel P. Curl: Contrastive unsupervised representations for reinforcement learning[C] //Proc of the 38th Int Conf on Machine Learning. New York: PMLR, 2020: 5639−5650

[41] Ye Weirui, Liu Shaohua, Kurutach T, et al. Mastering atari games with limited data[J]. Advances in Neural Information Processing Systems, 2021, 34: 25476−25488

[42] Stooke A, Lee K, Abbeel P, et al. Decoupling representation learning from reinforcement learning[C] //Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2021: 9870−9879

[43] Hansen N, Jangir R, Sun Yu, et al. Self-supervised policy adaptation during deployment[C] //Proc of the 9th Int Conf on Learning Representations. Vancouver, CA: ICLR, 2021: 2415−2432

[44] Wang Xudong, Lian Long, Yu S X. Unsupervised visual attention and invariance for reinforcement learning[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 6677−6687

[45] Shen Bingbing, Yao Le, Yang Zeyu, et al. Mode information separated β-VAE regression for multimode industrial process soft sensing[J]. IEEE Sensors Journal, 2023, 23(9): 10231−10240

[46] Nademi Y, Kalmady S V, Sun Weijie, et al. Generative data by β-variational autoencoders help build stronger classifiers: ECG use case[C] //Proc of the 19th Int Symp on Medical Information Processing and Analysis. Piscataway, NJ: IEEE, 2023: 1−7

[47] Xie Yuying, Arildsen T, Tan Zhenghua. Improved disentangled speech representations using contrastive learning in factorized hierarchical variational autoencoder[C] //Proc of the 31st European Signal Processing Conf. Piscataway, NJ: IEEE, 2023: 1330−1334

[48] Träuble F, Creager E, Kilbertus N, et al. On disentangled representations learned from correlated data[C] //Proc of the 39th Int Conf on Machine Learning. New York: PMLR, 2021: 10401−10412

[49] Funke C M, Vicol P, Wang K C, et al. Disentanglement and generalization under correlation shifts[C] //Proc of the 1st Conf on Lifelong Learning Agents. New York: PMLR, 2022: 116−141

[50] Agarwal R, Machado M C, Castro P S, et al. Contrastive behavioral similarity embeddings for generalization in reinforcement learning[C] // Proc of the 9th Int Conf on Learning Representations. Vancouver, CA:

ICLR, 2021: 2991−3020

[51] Mazoure B, Ahmed A M, MacAlpine P, et al. Cross-trajectory representation learning for zero-shot generalization in RL[C] //Proc of the 10th Int Conf on Learning Representations. Vancouver, CA: ICLR, 2022: 5608−5631

[52] Cubuk E D, Zoph B, Shlens J, et al. Randaugment: Practical automated data augmentation with a reduced search space[C] //Proc of the 33rd IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2020: 702−703

[53] Quan Ruijie, Zhu Linchao, Wu Yu, et al. Holistic LSTM for pedestrian trajectory prediction[J]. IEEE Transactions on Image Processing, 2021, 30: 3229−3239

[54] Yu Rui, Kershaw J, Wang Peng, et al. How to accurately monitor the weld penetration from dynamic weld pool serial images using CNN-LSTM deep learning model?[J]. IEEE Robotics and Automation Letters, 2022, 7(3): 6519−6525

[55] Zhou Shihao, Chen Duosheng, Pan Jinshan, et al. Adapt or perish: Adaptive sparse transformer with attentive feature refinement for image restoration[C] //Proc of the 37th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2024: 2952−2963

[56] Quan Ruijie, Yu Xin, Liang Yuanzhi, et al. Removing raindrops and rain streaks in one go[C] //Proc of the 34th IEEE/CVF Conf on Computer Vision and Pattern Recognition. Piscataway, NJ: IEEE, 2021: 9147−9156

[57] Haarnoja T, Zhou A, Abbeel P, et al. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor[C] //Proc of the 36th Int Conf on Machine Learning. New York: PMLR, 2018: 1861−1870

[58] Bertoin D, Zouitine A, Zouitine M, et al. Look where you look! Saliency-guided Q-networks for generalization in visual Reinforcement Learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 30693−30706

[59] Grooten B, Tomilin T, Vasan G, et al. MaDi: Learning to mask distractions for generalization in visual deep reinforcement learning[C] //Proc of the 23rd Int Conf on Autonomous Agents and Multiagent Systems. Richland, SA: IFAAMAS, 2024: 733−742



**Xu Donghong**, born in 1977. PhD, associate professor, master supervisor. His main research interests include deep reinforcement learning, visual reinforcement learning, cloud computing, and large model.

**徐东红**, 1977 年生。博士，副教授，硕士生导师。主要研究方向为深度强化学习、视觉强化学习、云计算、大模型。



**Wu Shilong**, born in 2000. Master candidate. His main research interests include deep reinforcement learning and visual reinforcement learning.

**武世龙**, 2000 年生。硕士研究生。主要研究方向为深度强化学习、视觉强化学习。