中图法分类号:TP391　　**文献标识码**: A　　**文章编号**: 1006-8961(XXXX)XX-0001-14

论文引用格式:Wang Yunke, Tao Linwei, Lin Yutian, Du Bo, Xu Chang. Visual Adversarial Imitation Learning with Calibrated Contrastive Representation[J/OL]. Journal of Image and Graphics, XXXX:1-14. DOI: 10.11834/jig.260149.(王云柯, 陶林伟, 林雨恬, 杜博, 徐畅. 校准对比学习表征驱动的视觉对抗模仿学习[J/OL]. 中国图象图形学报,XXXX:1-14. DOI: 10.11834/jig.260149.)[DOI:10.11834/jig.260149]

校准对比学习表征驱动的视觉对抗模仿学习

王云柯[1],陶林伟[2],林雨恬[1],杜博[1*],徐畅[2]

1. 武汉大学　计算机学院,湖北　武汉　430072; 2. 悉尼大学　计算机学院,澳大利亚　悉尼　2006

摘　要:**目的**　视觉模仿学习旨在从高维图像观测中学习智能体控制策略,但相比基于低维本体状态的方法,其性能仍存在明显差距。主要原因在于,像素观测中的关键行为差异较为细微,视觉编码器难以学习具有充分判别性的状态表征。已有视觉对抗模仿学习方法主要关注专家样本与智能体样本之间的区分,未充分利用智能体回放样本的内部结构信息,也未显式建模智能体策略在训练过程中逐步接近专家策略的动态演化特征。为提升高维视觉观测下的表征判别能力与训练稳定性,本文提出一种基于校准对比学习的视觉对抗模仿学习方法。**方法**　本文在生成对抗模仿学习框架中引入校准对比表示学习机制,通过"拉近相似状态、分离差异状态"的方式增强视觉编码器的判别能力。不同于已有方法主要关注专家样本与智能体样本之间的静态区分,本文进一步挖掘回放缓冲区的内部样本结构,并建模智能体样本质量随训练过程逐步提升的动态特征。具体而言,本文将智能体样本视为高质量样本与低质量样本的混合分布,并利用校准监督对比损失自适应调整其与专家样本之间的对比关系,从而提升视觉表征质量和对抗训练稳定性。**结果**　在DMControl Suite的9个连续控制任务上进行了实验验证。实验结果表明,所提出方法CAIL(contrastive adversarial imitation learning)在多个任务上取得了更高的累计回报,并在训练早期表现出更好的样本效率。与代表性方法PCIL(policy contrastive imitation learning)相比,CAIL在1M时间步的平均性能提升了22.6%。消融实验进一步验证了智能体无监督对比损失和校准监督对比损失的有效性,可视化结果表明CAIL能够更加准确地关注智能体关节等行为相关区域。**结论**　本文提出的校准对比视觉对抗模仿学习方法能够更充分地利用智能体回放样本,并动态刻画智能体样本质量随训练过程变化的特征,从而提升视觉状态表征的判别能力和对抗训练稳定性。该方法为高维视觉观测条件下的模仿学习提供了一种有效的表征学习思路。

关键词:强化学习;模仿学习;对抗模仿学习;对比学习;表征学习

Visual Adversarial Imitation Learning with Calibrated Contrastive Representation

Wang Yunke[1], Tao Linwei[2], Lin Yutian[1], Du Bo[1*], Xu Chang[2]

1. School of Computer Science, Wuhan University, Hubei Wuhan 430072, China; 2. School of Computer Science, The University of Sydney, Sydney NSW 2006, Australia

Abstract: **Objective** Visual imitation learning aims to learn control policies directly from high-dimensional image observations. It is important for robotics tasks, where agents often need to make decisions from visual inputs. However, compared with imitation learning based on low-dimensional proprioceptive states, visual imitation learning still suffers from a clear

收稿日期:2026-03-25;**修回日期**:2026-06-28

*通信作者:杜博 dubo@whu.edu.cn

基金项目:国家自然科学基金项目(62225113);湖北省自然科学基金创新研究群体项目(2024AFA017)

Supported by: National Natural Science Foundation of China (62225113), Innovative Research Group Project of Hubei Province (2024AFA017)

performance gap. One major reason is that behavior-related differences in pixel observations are often subtle. As a result, the visual encoder may fail to learn sufficiently discriminative state representations. This problem becomes more serious in visual adversarial imitation learning. In this framework, the discriminator needs to distinguish expert samples from agent samples and provide reward signals for policy learning. If the learned visual representation is weak, the discriminator may produce unstable or inaccurate rewards. This further affects policy optimization. Existing visual adversarial imitation learning methods mainly focus on the distinction between expert samples and agent samples. They usually treat this distinction as a static discrimination problem. However, they do not fully exploit the internal structure of agent samples stored in the replay buffer. They also do not explicitly model the dynamic evolution of agent samples during training. In practice, the quality of agent samples gradually improves as the learned policy approaches the expert policy. Ignoring this process may limit the quality of visual representation learning. To address these issues, this paper proposes a visual adversarial imitation learning method based on calibrated contrastive learning. **Method** This paper introduces a calibrated contrastive representation learning method in the visual adversarial imitation learning framework. The main idea is to improve the discriminative ability of the visual encoder by contrastive learning. Specifically, similar states are pulled closer in the feature space, while different states are pushed apart. In this way, the encoder can learn more effective visual representations for adversarial imitation learning. Different from existing methods that mainly focus on the static distinction between expert samples and agent samples, the proposed method further explores the internal sample structure in the replay buffer. It also models the dynamic improvement of agent sample quality during training. In the proposed framework, agent samples are regarded as a mixture of high-quality samples and low-quality samples. High-quality agent samples may be close to expert samples, especially in the later stage of training. Low-quality agent samples are still far from expert behavior and should not be treated as expert-like samples. Based on this observation, this paper introduces a calibrated supervised contrastive loss. This loss adaptively adjusts the contrastive relationship between agent samples and expert samples. It reduces the risk of incorrectly pushing high-quality agent samples away from expert samples. At the same time, it preserves the discrimination between expert behavior and low-quality agent behavior. The proposed method therefore improves both visual representation quality and adversarial training stability. The final objective combines the adversarial imitation learning loss with the proposed contrastive representation learning objective. The whole framework can be optimized in an end-to-end manner. **Result** Experiments are conducted on 9 continuous control tasks from the DMControl Suite. These tasks cover different types of control behaviors, including balancing, locomotion, and complex body movement. The experimental results show that the proposed method achieves higher cumulative rewards on multiple tasks. It also shows better sample efficiency in the early training stage. Compared with the representative method PCIL, CAIL improves the average performance by 22.6% at 1M training steps. This result indicates that calibrated contrastive representation learning can effectively improve visual adversarial imitation learning. The ablation study further verifies the effectiveness of the proposed contrastive losses. The unsupervised contrastive loss on agent samples improves the use of the replay buffer. It helps the encoder learn from a large number of dynamically collected visual states. The calibrated supervised contrastive loss further improves performance by modeling the changing quality of agent samples. The results show that these components are complementary. The full CAIL model achieves better performance than the version without calibration. This demonstrates the importance of dynamically modeling agent samples during training. In addition, visualization results show that the discriminator learned by CAIL can focus more accurately on behavior-related regions, such as the joints and body parts of the agent. These regions are important for distinguishing different motion states. The visualization results further support that CAIL learns more meaningful and discriminative visual representations. **Conclusion** This paper proposes a calibrated contrastive visual adversarial imitation learning method for high-dimensional visual control tasks. The proposed method makes better use of agent samples in the replay buffer. It also explicitly models the change of agent sample quality during training. By introducing calibrated contrastive representation learning, the method improves the discriminative ability of visual state representations and stabilizes adversarial training. Experimental results on DMControl tasks show that the proposed method achieves better average performance and sample efficiency than representative visual imitation learning methods. The proposed framework provides an effective representation learning strategy for imitation learning from high-dimensional image observations.

Key words: reinforcement learning; imitation learning; adversarial imitation learning; contrastive learning; representation learning

论文引用格式：Wang Yunke, Tao Linwei, Lin Yutian, Du Bo, Xu Chang. Visual adversarial imitation learning with calibrated contrastive representation [J/OL]. Journal of Image and Graphics, xxxx: 1-15. DOI: 10.11834/jig.260149. (王云柯,陶林伟,林雨恬,杜博,徐畅. 校准对比学习表征驱动的视觉对抗模仿学习[J/OL]. 中国图像图形学报,xxxx: 1-15. DOI: 10.11834/jig.260149.) [DOI: 10.11834/jig.260149]

0 引 言

模仿学习(imitation learning,IL)是一类用于解决基于马尔可夫模型的序列决策问题的重要方法(Liu等,2021)。其核心目标是通过模仿专家行为,使智能体学习在给定状态下的决策策略。模仿学习方法中,行为克隆(behavioral cloning,BC)是最基础且应用最广泛的算法之一(Pomerleau,1988)。该方法通过对专家演示数据进行监督学习,直接学习从状态到动作的映射关系。然而,由于行为克隆依赖离线监督学习进行训练,在策略执行过程中容易出现误差累积现象(Brantley等,2019),导致智能体进入训练数据未覆盖的状态区域,进而带来潜在的决策风险(Tu等,2022)。为缓解上述问题,对抗模仿学习(adversarial imitation learning,AIL)通过分布匹配机制,使智能体策略的分布逐渐逼近专家策略分布(Ho和Ermon,2016),从而提升策略的泛化能力。在标准基准任务上,AIL已取得显著效果(Todorov等,2012),其中专家演示通常由低维本体状态(如位置、速度等)构成。相比之下,面向视觉输入的模仿学习任务更具挑战性。已有研究表明,视觉状态输入会显著增加策略学习的难度(Tucker等,2018),而在复杂的三维交互环境中,上述问题会进一步加剧(Jaderberg等,2018)。

在高维视觉输入条件下,状态表征学习是影响策略模型性能的关键因素。强化学习(reinforcement learning,RL)通过与环境交互并依回报信号进行策略优化(Sutton和Barto,2018)。当状态由原始像素表示时,通常需要借助图像编码器获得紧凑且判别性强的潜在表征。在视觉强化学习中,策略网络与价值网络通常共享编码器表征,编码器主要通过价值函数损失进行更新。已有研究表明,简单的数据增强策略即可显著提升视觉强化学习的泛化能力(Cobbe等,2019)。CURL(contrastive unsupervised representations for reinforcement learning)方法(Laskin等,2020)进一步将对比学习引入强化学习框架,通过数据增强构建对比目标,从而提升视觉强化学习的数据效率。此外,已有研究将表征学习作为辅助任务与强化学习过程联合优化,以提升视觉表征质量(Yarats等,2021)。

然而,视觉强化学习通常依赖明确的回报函数,为编码器学习提供稳定的监督信号。在视觉模仿学习(visual imitation learning,VIL)中,智能体往往仅依赖专家演示数据,缺乏可直接用于表征优化的稠密回报信号。在对抗式训练框架下,判别器与策略网络的相互博弈进一步加剧了训练的不稳定性,使得编码器难以获得稳定有效的梯度指导。因此,相较于视觉强化学习,视觉模仿学习对表征判别能力与训练稳定性提出了更高要求。由于在自动驾驶和机器人学习等真实场景中的应用潜力,视觉模仿学习受到广泛关注(Ross等,2011)。部分研究通过分阶段方式先学习回报函数或视觉编码器,再进行策略优化(Brown等,2019)。也有方法尝试直接改造对抗模仿学习框架,使其适配视觉输入(Rafailov等,2021)。例如,PatchAIL(patch-based adversarial imitation learning)算法(Liu等,2023)通过局部图像块判别机制评估不同图像patch的"专家性",在一定程度上增强了对细粒度视觉差异的敏感性。此外,PCIL(policy contrastive imitation learning)算法(Huang等,2023)在判别器中引入对比约束以拉近专家样本并分离智能体样本,从而增强判别能力。然而,该类方法未充分利用回放缓冲区中大量智能体样本,也未显式刻画智能体在对抗训练过程中逐步接近专家策略的演化特性。

尽管上述方法取得一定进展,但视觉对抗模仿学习整体性能仍落后于基于低维本体状态的对抗模仿学习。其根本原因在于,低维本体状态在特征空间中具有更强的可分性,而图像空间中的关键行为

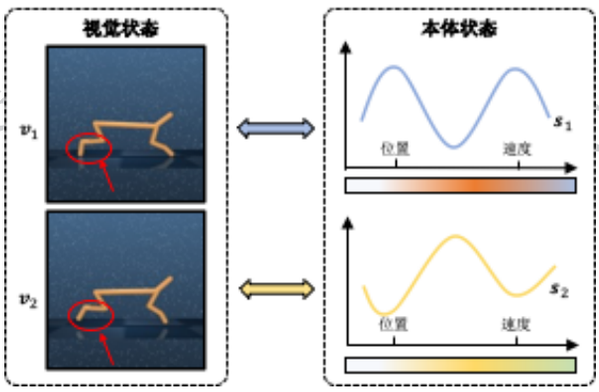


图1　视觉状态以及它们对应的本体状态示意图

Fig. 1　Figure of visual states and their corresponding physical states.

差异往往难以直接辨识(Van Hasselt 等,2016)。如图1所示,物理状态的显著变化可能仅对应视觉观测中的微小差异,从而导致表征判别难度增加。因此,构建能够准确编码视觉状态的判别性表征,是缩小视觉 AIL 与本体 AIL 性能差距的关键问题。对比表征学习通过"拉近相似样本、分离不相似样本"构建判别性特征空间,在视觉表征学习领域已得到系统验证(Chen 等,2020)。然而,现有视觉对抗模仿学习方法多侧重于专家与智能体样本之间的区分,而忽视了智能体样本内部结构信息及其随训练逐步演化的动态特征(Huang 等,2023)。

为了解决这个问题,本文提出了对比对抗模仿学习(Contrastive Adversarial Imitation Learning,简称 CAIL)。本文的核心思想是通过对比表示学习提升视觉编码器的判别能力,使模型能够在特征空间中拉近相似视觉状态,并分离行为差异较大的视觉状态。具体而言,CAIL 在视觉对抗模仿学习框架中引入无监督对比学习约束,以充分利用智能体与环境交互过程中存储在回放缓冲区中的大量视觉状态,从而增强编码器对不同智能体状态的实例级判别能力。同时,为了提升专家样本与智能体样本之间的可分性,本文引入监督对比学习思想,对两类样本在特征空间中的关系进行约束。进一步地,考虑到智能体策略在训练过程中会逐渐接近专家策略,本文将智能体样本建模为高质量样本与低质量样本组成的混合分布,并提出校准监督对比损失,自适应调整智能体样本与专家样本之间的对比关系。该设计可以缓解普通监督对比学习中将高质量智能体样本错误视为专家负样本的问题,从而提升视觉表征质量和对抗训练稳定性。本文的主要贡献总结如下:

(1)提出一种基于校准对比学习的视觉对抗模仿学习方法。该方法在对抗模仿学习框架中引入对比表示学习机制,通过拉近相似状态、分离差异状态来增强视觉编码器的判别能力,从而缓解高维视觉输入下表征学习不足和对抗训练不稳定的问题。

(2)提出校准监督对比学习机制,以建模智能体样本质量随训练过程逐步提升的动态特征。该机制将智能体样本视为高质量样本与低质量样本的混合分布,并自适应调整其与专家样本之间的对比关系,从而减少高质量智能体样本被错误排斥的问题,提升表征学习的稳定性。

(3)在 DMControl Suite 的9个视觉连续控制任务上验证了本文方法的有效性。实验结果表明,CAIL 在平均性能、样本效率和训练稳定性方面均优于多种代表性视觉模仿学习方法。消融实验进一步验证了提出的无监督对比损失和校准监督对比损失的有效性。可视化结果表明 CAIL 能够更加准确地关注智能体关节等行为相关区域。

1　校准对比对抗模仿学习

1.1　预备知识

1.1.1　马尔可夫决策过程

马尔可夫决策过程(Markov Decision Process,简称 MDP)是一种数学模型(Puterman,1994),用于在部分随机且部分受控的环境中进行决策。它为强化学习、模仿学习算法提供了理论基础。MDP 通常用于刻画决策者在时间序列中的决策过程。MDP 由以下几个核心要素构成:

$$M=(S,A,P,R,\gamma,\mu_0),\#(1)$$

其中 S 为状态空间,A 为动作空间,$P(s'|s,a)$ 表示在状态 s 下执行动作 a 中转移至 s' 的概率,$R(s,a)$ 表示在状态 s 下执行动作 a 所获得的回报,$\gamma\in[0,1]$ 为折扣因子,用于计算未来回报的当前价值,μ_0 为初始状态分布。马尔可夫决策过程的目标是为决策者求解一个最优策略模型 $\pi(a|s)$,使得决策者能够在该策略的支持下,最大化其在 MDP 模型中获得的预期累积回报。一般来说,策略的分布可以由占有率度量 ρ_π 来表示,用以衡量该策略在状态-动作空间中的分布。

1.1.2 强化学习

强化学习是一种通过与环境交互学习最优决策策略的方法。在强化学习中，智能体通过在环境中“探索试错”，不断优化其自身策略。大部分强化学习算法都是基于马尔可夫决策过程的，其目标是为马尔可夫决策过程中的决策者学得一个策略π，使其能够在决策过程中获得尽可能大的预期累积回报。在强化学习中，最优策略可以通过以下目标函数求解：

$$\pi^{*}=\arg\max_{\pi}\mathbb{E}_{\pi}\left[\sum_{t=0}^{T}\gamma^{t}r\left(s_{t},a_{t}\right)\right].\#(2)$$

强化学习通常依赖精心设计的回报函数，然而在实际场景中，该设计往往较为困难。相比之下，模仿学习能够让决策者通过直接模仿专家演示来学习策略，省略了回报函数设计的步骤，是一个更加实际的选择。

1.1.3 对抗模仿学习

生成对抗模仿学习（Generative Adversarial Imitation Learning，GAIL）（Ho等，2016）是模仿学习中最具代表性的框架。GAIL引入了一个判别器来区分专家演示（即专家轨迹中的“状态-动作”对）与智能体演示（即智能体轨迹中的“状态-动作”对）。智能体的目标是“欺骗”判别器的判断，使其将智能体演示分类为和专家演示一致的类别，从而在分布意义上逼近专家策略。

从目标函数的角度看，GAIL通过一个极小极大问题实现分布匹配。判别器D负责区分样本来源，策略π_{θ}则不断调整自身，使得判别器难以区分两者。该过程本质上对应于最小化专家与智能体占用测度之间的散度（Ke等，2020）。其目标函数形式为：

$$\min_{\pi_{\theta}}\max_{D}\ \mathrm{E}_{(s,a)\sim\rho_{\pi}}\left[\log D\left(s,a\right)\right]+\mathrm{E}_{(s,a)\sim\rho_{\pi_{\theta}}}\left[\log\left(1-D\left(s,a\right)\right)\right],\#(3)$$

在上述目标函数中，智能体策略π_{θ}通过最小化$\mathrm{E}_{(s,a)\sim\rho_{\pi_{\theta}}}\log\left(1-D\left(s,a\right)\right)$来训练。因此，判别器的输出$-\log\left(1-D\left(s,a\right)\right)$可以被视作智能体策略$\pi_{\theta}$学习的回报值。随后，即可使用常见的强化学习算法更新策略模型，例如TRPO（Schulman等，2015）、PPO（Schulman等，2017）。

1.1.4 从低维状态到高维视觉状态

在上述理论框架基础上，进一步考虑视觉输入条件下的模仿学习问题。在标准控制任务中，状态通常为低维本体信息（例如关节位置、速度等），且专家演示数据往往包含完整的状态-动作对。然而，在视觉模仿学习任务中，专家演示数据通常仅提供第三人称视觉观测，缺乏显式动作信息。例如，在DMControl环境中，视觉状态被表示为智能体第三人称视角3个连续帧的84×84分辨率的RGB渲染图像的组合，以在视觉状态空间中融合时序信息和空间信息。因此，需要对原始生成对抗模仿学习的框架进行改写。具体而言，判别对象由状态-动作对(s,a)替换为视觉状态v，图像编码器$f(v)$输出表征r，随后通过投影头$h_{d}(r)$输出判别得分，判别器可写为$D(v)=h_{d}\left(f(v)\right)$。对应的判别损失为：

$$L_{dis}\left(h_{d}f\right)=-\mathrm{E}_{v\sim\rho_{\pi_{\theta}}}\left[\log h_{d}\left(f(v)\right)\right]-\mathrm{E}_{v\sim\rho_{\pi_{\theta}}}\left[\log\left(1-h_{d}\left(f(v)\right)\right)\right],\#(4)$$

图2 对比对抗模仿学习算法框架图

Fig. 2 Figure of the framework of Contrastive Adversarial Imitation Learning.

在该框架下，判别器仍负责区分“专家视觉轨迹”与“智能体视觉轨迹”，智能体策略模型π_{θ}则通过强化学习算法更新。

一般来说，将公式4直接应用于视觉模仿学习难以取得好的效果，主要原因在于编码器f对于视觉相似但语义不同的智能体状态判别能力有限。为了让视觉模仿学习取得更好的效果，对编码器的表征能力提出了更高的要求。然而，由于算法中对抗训练的不稳定性和强化学习步骤中随机采样的不确定性，导致视觉状态编码器难以获得较好的表征。

1.2 算法表述

1.2.1 算法描述

近年来，对比学习已经成为学习判别性表征的重要方法。将对比学习引入视觉对抗模仿学习，可以为视觉编码器提供额外的表征约束，从而缓解高维视觉输入下判别器表征能力不足的问题。其核心思想是在特征空间中拉近相似视觉状态的表征，并分离行为差异较大的视觉状态表征。具体而言，给定视觉状态v，视觉编码器f将其映射为潜在的表征$r=f(v)$。通过构造正样本和负样本，对比学习约束能够促使编码器学习更加稳定且具有判别性的视觉状态表征。基于上述思想，本文提出对比对抗模仿学习算法（Contrastive Adversarial Imitation Learning，

CAIL)。CAIL在原始对抗模仿学习框架的基础上引入对比表示学习机制,包括无监督对比学习约束L_{UnSupCon}和有监督对比学习约束L_{SupCon}。无监督对比损失函数的目的是充分利用缓冲区中大量的智能体视觉状态,有监督对比损失函数的目标是增强编码器的判别能力。进一步地,考虑到智能体策略会随着模仿学习训练逐步接近专家策略,训练后期的部分智能体视觉状态可能已经具有较高质量。若仍将所有智能体样本简单视为专家样本的负例,可能会影响表征学习的稳定性。为此,本文提出校准有监督对比损失$L_{\text{C-SupCon}}$,将智能体样本建模为高质量样本与低质量样本的混合分布,并自适应调整智能体样本与专家样本之间的对比关系。在完整CAIL中,$L_{\text{C-SupCon}}$用于替代普通有监督对比损失,而L_{SupCon}主要用于构造无校准版本CAIL(w/o cal),以验证校准机制的有效性。CAIL算法的总体框架如图2所示。

1.2.2 无监督对比学习损失函数

在视觉对抗模仿学习过程中,智能体与环境交互产生的大量视觉演示被存储于回放缓冲区中,用于后续策略模型的训练。从另一个方面来看,这些数据亦可被视为大规模无标签数据,可被用于基于对比学习的表征训练。CAIL算法中加入的无监督对比学习损失函数充分利用了回放缓冲区中大量的智能体演示来学习视觉演示表征。在训练中,CAIL算法对一个批次的N个智能体视觉状态进行数据增广,最终得到$2N$个增强的智能体视觉状态$v^a=\{v_i^a\}_{i=1}^{2N}$。同一个智能体视觉状态的两个增强状态分别表示为$\left(v_i^a,p\left(v_i^a\right)\right)$,其中$p\left(v_i^a\right)$是$v_i^a$以外的两个增强状态之一。由于缺乏标签信息,将$p\left(v_i^a\right)$视为$v_i^a$的唯一正例,并期望最小化$v_i^a$和$p\left(v_i^a\right)$在特征空间上的距离。同一批次中剩余的$2N-2$个增强状态被视为$v_i^a$的负例,因此它们在特征空间上的表征应和$v_i^a$在特征空间上的表征距离更远。基于这个思想,CAIL采用在对比学习中广泛使用的InfoNCE损失来定义基于单个智能体视觉状态v_i^a的无监督对比学习损失:

$$L_{\text{InfoNCE}}\left(v_i^a,v_i^{a+},v^{a-},f,h\right)=-\log\frac{\exp\left(\text{sim}\left(h\left(f\left(v_i^a\right)\right),h\left(f\left(v_i^{a+}\right)\right)\right)/\tau\right)}{\sum_j\exp\left(\text{sim}\left(h\left(f\left(v_i^a\right)\right),h\left(f\left(v_j^{a-}\right)\right)\right)/\tau\right)},\#(5)$$

其中sim(·)表示余弦相似度函数,v_i^{a+}是与v_i^a对应的正例,v^{a-}是负例的集合,h是MLP投影头,τ是温度参数。最小化L^{InfoNCE}损失会导致v_i^{a+}和v_i^a的表征距离减小,而不相似状态(v^{a-})的表征距离则会增加。进一步地,将基于单个智能体视觉状态v_i^a的无监督对比学习损失扩展到同一批次的全部智能体视觉演示,即将无监督对比损失L_{UnSupCon}定义为批次中所有$2N$个状态增强的L_{InfoNCE}损失的平均值,损失函数表示为:

$$L_{\text{UnSupCon}}=\frac{1}{2N}\sum_{i=1}^{2N}L_{\text{InfoNCE}}\left(v_i^a,p\left(v_i^a\right),v^a\setminus v_i^a,f,h_{\text{unsup}}\right)\#(6)$$

其中h_{unsup}是无监督对比学习中使用的专用投影头。

1.2.3 有监督对比学习损失函数

无监督对比学习损失函数充分利用了缓冲回放区大量的智能体视觉状态来为编码器f学习一个有判别力的表征,提升了模型对不同的视觉演示的判别力。除此之外,模型还需要具备区分智能体状态与专家状态的能力。通常情况下,专家视觉状态往往有着规律的模式和明显的集群效应,然而智能体视觉状态的不规律噪声和次优性往往使其在特征空间上较为分散。因此,一个好的模型需要将专家视觉状态在特征子空间的距离拉近,并同时将专家视觉状态和智能体视觉状态的距离推远。为了达到这个目标,可以使用有监督对比学习损失函数(Khosla等,2020)。有监督对比学习损失函数需要样本的类别信息,因此可以将专家视觉状态和智能体视觉状态视作两个类别,在这种情况下进行有监督的对比学习训练。

在无监督对比学习损失中,每个样本仅仅只有一个对应的正样本。然而,在有监督对比学习中,其将无监督对比学习损失L_{UnSupCon}扩展到了任意数量正样本的形式。具体来说,一个批次中所有增广的专家视觉状态$v^e=\{v_i^e\}_{i=1}^N$被看作是一个类别,而所有增广的智能体视觉状态v^a被视作另一类别。如果将一个批次中所有增广的视觉状态表示为$v=v^a\cup v^e$,那么对于其中每个视觉状态$v_i^e\in v^e$的有监督对比学习损失可以被定义为:

$$L_{\text{Sup}}\left(v_i^e,v^e\right)=-\frac{1}{\left|v^e\right|-1}\sum_{v_j^+\in v^e\setminus v_i^e}L_{\text{InfoNCE}}\left(v_i^e,v_j^+,v\setminus v_i^e,f,h_{\text{sup}}\right),\#(7)$$

其中h_{sup}是有监督对比损失中使用的投影头,$\left|v^e\right|$是

v^e中的样本数量。将单个样本的对比损失扩展至整个批次的数据，得到有监督对比学习损失，其可以被定义如下：

$$L_{\mathrm{SupCon}}=\frac{1}{N}\sum_{i=1}^{N}L_{\mathrm{Sup}}\left(v_i^e,v^e\right)\#(8)$$

利用有监督对比学习损失L_{SupCon}，模型不仅能够学习到与无监督对比学习损失L_{UnSupCon}一致的表征，还增强了区分专家和智能体视觉状态的能力。

1.2.4　校准有监督对比学习损失函数

在普通有监督对比学习损失中，类别标签是固定的，即专家视觉状态被视为一类，智能体视觉状态被视为另一类。然而，在对抗模仿学习过程中，智能体策略会随着训练逐步提升，其生成的视觉状态也会逐渐接近专家视觉状态。尤其在训练后期，回放缓冲区中的部分智能体视觉状态可能已经具有较高质量。此时，如果仍将所有智能体视觉状态简单视为专家视觉状态的负例，可能会错误地排斥高质量智能体样本，从而影响表征学习的稳定性。因此，在视觉对抗模仿学习中，将智能体视觉状态建模为高质量样本与低质量样本的混合分布更为合理。

然而，直接判断某个智能体视觉状态是否为高质量样本并不容易。为此，本文引入校准参数α，将智能体样本v_i^a视为一个混合样本：其有α的概率是高质量样本，有$1-\alpha$的概率是低质量样本。α用于刻画智能体样本质量随训练过程逐步提升的趋势。当把v_i^a视作高质量样本时，其行为表现被认为更接近专家样本。在这种情况下，所有的专家状态v_i^e和其对应的增广状态$p\left(v_i^a\right)$被视作v_i^a的正例。当将智能体状态视为低质量样本时，本文不将其与专家样本拉近，而是退化为无监督对比学习形式，仅有增广的状态$p\left(v_i^a\right)$被视作v_i^a的正例，其他所有状态都被视作v_i^a的负例。基于上述混合建模的思想，校准有监督对比学习损失可以被定义如下：

$$\begin{aligned}L_{\mathrm{C-SupCon}}&=\alpha E_{v_i^a\sim\rho_{\pi_\theta}}\left[L_{\mathrm{Sup}}\left(v_i^a,v^e\cup\{v_i^a,p\left(v_i^a\right)\}\right)\right]\\&+(1-\alpha)E_{v_i^a\sim\rho_{\pi_\theta}}\left[L_{\mathrm{InfoNCE}}\left(v_i^a,p\left(v_i^a\right),v\setminus v_i^a,f,h_{\mathrm{sup}}\right)\right],\#(9)\end{aligned}$$

其中，第一项表示当智能体样本被视为高质量样本时，将其与专家样本及自身增强视图拉近；第二项表示当智能体样本被视为低质量样本时，仅保持其不同增强视图之间的一致性。结合公式6和公式9，CAIL的总体目标函数定义为：

$$L_{\mathrm{CAIL}}=L_{\mathrm{dis}}+\lambda_1L_{\mathrm{UnSupCon}}+\lambda_2L_{\mathrm{C-SupCon}},\#(10)$$

其中λ_1和λ_2是控制不同对比学习损失项的超参数，判别器损失定义为$L_{\mathrm{dis}}=\frac{1}{N}\sum_{i=1}^{N}-\log\left(h_d\left(r_i^e\right)\right)-\log\left(1-h_d\left(r_i^a\right)\right)$。在本文实现中，超参数$\lambda_1$和$\lambda_2$均设置为定量1。

1.2.5　算法综述

在上一节中已经给出了CAIL算法中判别器的训练目标函数，即公式10。除了通过对抗损失和对比损失训练视觉表征外，CAIL还需要利用判别器输出的奖励信号更新策略模型。本文采用基于DDPG（Silver等，2014）的强化学习算法训练策略网络，并引入Double Q-learning（Van等，2016）的思想以缓解评论网络的过估计问题。具体而言，本文使用策略网络π_θ和两个评论网络Q_{ϕ_1}、Q_{ϕ_2}进行交替优化。给定回放缓冲区$\mathcal{B}$中的转移样本$\mathcal{B}=(v,a,r,v',d)$，其是一个包含视觉状态v，动作a，回报r，下一视觉状态v'和终止信号d的五元组。回报r通过对抗训练中判别器的输出定义。评论网络通过最小化差分损失进行更新：

$$\begin{aligned}&L_Q(\phi,\mathcal{B})=E_{t\sim\mathcal{B}}\left[\left(Q_{\phi_i}(v,a)-\left(r+\gamma(1-d)\Gamma\right)\right)^2\right]\\&\forall i\in\{1,2\},\#(11)\end{aligned}$$

其中下一状态价值Γ的估算值可被定义为$\Gamma=\left(\min_{i=1,2}Q_{\phi_i}\left(v',\pi_\theta\left(a'|v'\right)\right)\right)$。

在评论网络能够评估给定视觉状态和动作的价值后，策略网络通过最大化评论网络估计的动作价值进行更新。对应地，策略网络的优化目标可以写为：

$$L_\pi(\theta)=-E_{(v,a)\sim\pi_\theta}\left[\min_{i=1,2}Q_{\phi_i}(v,a)\right],\#(12)$$

在策略更新过程中，本文不通过策略损失更新视觉编码器的参数，以减少强化学习梯度对视觉表征学习带来的不稳定影响。视觉编码器主要由判别器损失和对比学习损失进行更新。

2　实　验

本节通过多组实验评估所提出CAIL算法的有效性。实验主要在基于像素输入的DMControl连续控制任务上开展，并与多种视觉模仿学习方法进行比较。本文从性能、样本效率、训练开销、消融分析以及可视化表征等多个方面分析CAIL的有效性。

2.1 实验设置

2.1.1 实验环境

实验环境选取DMControl中九个最具代表性的连续控制任务进行评估，包括Cartpole Swingup、Finger Spin、Cheetah Run、Hopper Hop、Hopper Stand、Walker Stand、Walker Walk、Walker Run以及Quadruped Run。这些任务覆盖了平衡控制、运动控制以及复杂动力学交互等不同难度的控制问题，能够较全面地检验视觉模仿学习算法的泛化能力。在专家演示数据方面，本文采用ROT（Regularized Optimal Transport）公开数据集（Haldar等，2022）中的专家演示作为训练样本。在DMControl环境中，视觉状态被表示为智能体3个连续帧的84×84大小的RGB渲染图像的组合，以在视觉状态空间中融合时序信息和空间信息。

2.1.2 网络架构

关于图像编码器结构，本文遵循文献（Yarats等，2021）中的默认视觉编码器结构，其包含一个

4层的卷积神经网络，每层的卷积核大小，通道数，步长，padding值分别为[3×3，32，2，0]，[3×3，32，1，0]，[3×3，32，1，0]，[3×3，32，1，0]。经过视觉编码器的处理，视觉状态最终被转换为特征图。在图像编码器f对视觉状态v进行编码后，生成的特征图会被展平为长度为39200的向量。接着，该向量被传递到后续的多层感知机策略网络和评论网络。判别器的结构与文献（Yarats等，2021）中的默认结构一致，其为一个简单的3层MLP结构。CAIL算法的整体结构如图3所示。

2.1.3 对比算法

为了客观评估CAIL的性能，本文将其与五种代表性方法进行比较，即BC，GAIL，共享编码器的GAIL（GAIL-SE）（Cohen等，2021），PCIL（Huang等，2023）和PatchAIL（Liu等，2023）。行为克隆方法通过监督学习直接学习得到视觉状态到动作之间的映射关系，是最基础的视觉模仿学习方法。GAIL采用生成对抗机制，通过判别器区分专家轨迹与智能体轨迹，并利用对抗回报更新策略网络。GAIL-SE在GAIL基础上引入共享视觉编码器，使策略网络与判别器共享同一特征提取模块，从而提高视觉特征利用效率。PCIL在策略编码器中加入对比约束，使专家轨迹在特征空间中更加集中，同时将智能体轨迹推离专家轨迹。PatchAIL则在对抗训练过程中引入图像块级判别器，对图像不同区域分别计算回报信号。

Tab. 1 Table of performance of CAIL and compared methods at 500K and 1M timesteps.

图3 CAIL算法编码器、策略模型更新框架图

Fig. 3 Framework of CAIL's encoder and policy's update.

2.1.4 超参数的选择

在CAIL算法中，学习率设置为1×10^{-4}，三个神经网络的训练批次大小为256。判别器网络在每个时间步都进行更新，而策略网络每两个时间步进行一次优化和更新。模仿学习中默认的回放缓冲区大小设置为150000，采样轨迹的折扣率γ设置为0.99，神经网络参数更新的优化器为Adam优化器。

2.1.5 评价指标

在评价指标方面，从智能体在不同任务中的性能通过轨迹累计回报进行衡量，即在一条完整交互轨迹中获得的回报总和，平均回报越高，表示策略性能越优。实验分别在500K训练步和1M训练步两个阶段报告主要结果。所有实验均使用5个不同随机种子重复运行，并报告平均值与标准差，以减少随机因素对实验结果的影响。

2.2 实验结果

CAIL及其对比算法的实验结果如表1所示，其中CAIL（w/o cal）表示将CAIL算法中校准的有监督对比学习损失替换为普通的有监督对比学习损失。从表格中可以观察到，在视觉模仿学习任务中，即使在最简单的任务Cartpole Swingup中，原始的GAIL算法也难以取得较好的结果。此外，在Hopper Stand任务中，基于GAIL算法训练得到的智能体甚至无法学到比随机策略更好的策略。这些结果与之前在文献（Tucker等，2018）中得到的结论一致。相比之下，通过为策略和判别器采用共享图像编码器，基于GAIL-SE算法训练得到的智能体在所有九个环境中获得的回报至少是原始GAIL算法的两倍。然而，GAIL-SE的性能与专家性能相比仍存在明显差距。具体而言，原始GAIL在高维视觉输入下表现较弱，说明视觉表征学习是视觉对抗模仿学习中的关键瓶颈。GAIL-SE和PCIL的性能提升表明共享视觉编码器和对比约束能够改善视觉模仿学习效果，而CAIL在多个任务上取得更高累计回报，并在1M时间步下相较PCIL平均性能提升22.6%，说明校准对

500K时间步	GAIL	GAIL-SE	PCIL	CAIL (w/o cal)	CAIL	BC	专家
Cartpole Swingup	186±20	734±160	296±52	790±61	**838 4**	521±120	859±0
Finger Spin	0±0	303±191	408±297	460± 193	**642 186**	284±120	976±9
Cheetah Run	69±27	514±33	461±51	497±44	**538 34**	185±49	890±9
Hopper Hop	10±7	8±8	38±18	51±23	**73 12**	109±18	318±7
Hopper Stand	5±3	270±343	451±199	290± 306	**545 305**	386±72	976±9
Walker Stand	272±95	513±286	960±18	910±44	**961 9**	496±70	939±9
Walker Walk	69±49	268±61	203±19	350± 136	**463 126**	556±110	970±20
Walker Run	24±6	57±17	146±10	148±28	**157 9**	378±82	778±10
Quadruped Run	151±57	212±5	228±119	238±50	**306 10**	277±58	547±136
1M时间步	GAIL	GAIL-SE	PCIL	CAIL (w/o cal)	CAIL	BC	专家
Cartpole Swingup	199±17	801±91	67±36	687± 312	**837 23**	521±120	859±0
Finger Spin	0±0	407±250	534±233	704± 122	**785 99**	284±120	976±9
Cheetah Run	84±30	624±35	662±27	689±37	**725 31**	185±49	890±9
Hopper Hop	0±0	121±4	158±8	**184 25**	182±20	109±8	318±7
Hopper Stand	5±3	747±63	733±104	754± 106	**777 42**	386±72	976±9
Walker Stand	275±100	764±251	827±38	**859 101**	831±154	496±70	939±9
Walker Walk	63±34	**953 3**	952±0	940±13	938±6	556±110	970±20
Walker Run	28±8	133±28	519±59	468±66	**526 8**	378±82	778±10
Quadruped Run	115±60	296±82	**427 35**	322±38	382±27	277±58	547±136

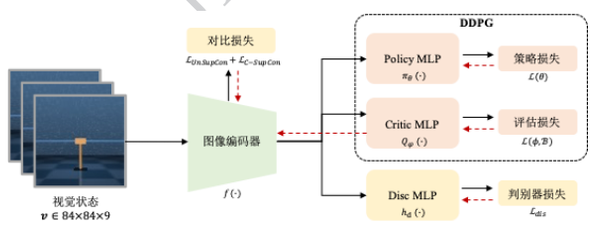
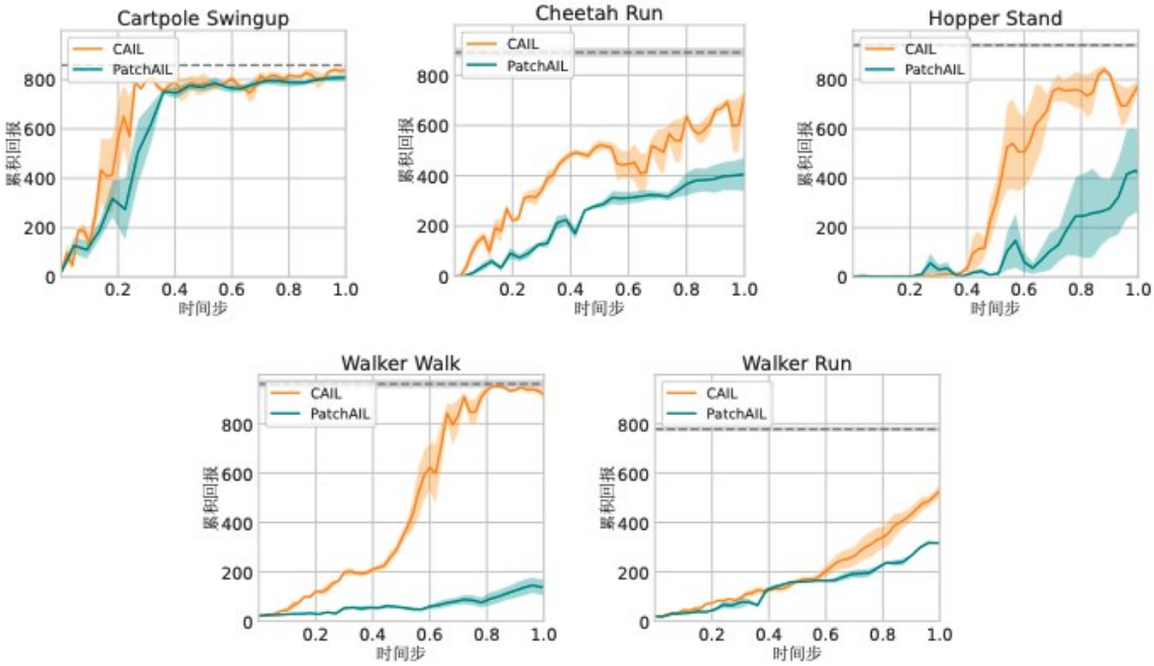


表1 CAIL和对比算法在500K时间步和1M时间步的实验结果对比表

比学习能够进一步提升视觉表征质量和策略性能。

PatchAIL算法是一种基于图像块判别器的最先进的视觉对抗模仿学习方法。虽然PatchAIL算法在收敛时间步(1M)时性能最佳,但是与其他算法相比,PatchAIL的训练成本显著增加了。在图5中,展示了PatchAIL在Cartpole Swingup任务中的训练时间和GPU内存使用情况。所有训练成本均在单张NVIDIA RTX 4090 GPU上统计,训练时间以小时为



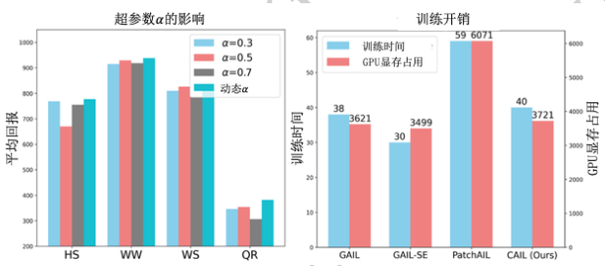


图5　参数α的影响(左图)和训练成本对比(右图)

Fig. 5　Figure of impact of α and comparison on training cost.

单位,GPU 显存占用以 MB 为单位。实验结果显示相比于其他对抗模仿学习算法,PatchAIL 的训练开销十分显著。例如,与 GAIL 算法相比,PatchAIL 算法的训练时间和 GPU 内存使用量几乎翻倍。考虑到训练效率的问题,图 4 中给出了 CAIL 算法和 PatchAIL 算法在相同训练时间预算下的训练曲线。

该训练曲线能够反映给定不同算法相同的训练时间,模型分别能够达到的性能。在相同训练时间的设置下,实验结果表明 CAIL 算法比 PatchAIL 算法的表现更好。

此外,复杂度分析结果表明,CAIL 的参数量和 FLOPs 分别为 4.25M 和 93.0M,而 PatchAIL 的参数量和 FLOPs 分别为 3.27M 和 659.0M。可以看出,PatchAIL 的参数量并未显著高于 CAIL,但其 FLOPs 明显更高。这是因为 PatchAIL 的 patch discriminator 采用卷积结构,参数共享使其参数量相对有限,但其需要在高维视觉状态上进行密集卷积计算以生成基于 patch 的回报值,因此带来了更高的计算开销。

在 CAIL 算法中,参数α是一个预先定义的先验

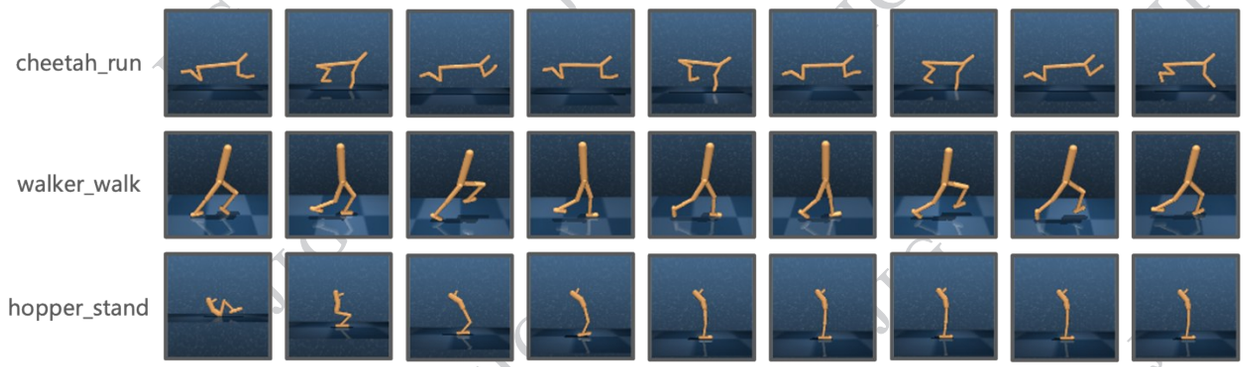


图6　智能体可视化图

Fig. 6　Visualized agent's trajectories.

参数，代表了在模仿学习训练期间智能体演示被视为“正样本”的概率。为了研究不同α值对实验中校准对比损失的性能的影响，实验中设置了不同的α值并进行了CAIL算法的评估。一般来说，在模仿学习训练的初始阶段，智能体演示更有可能是“负样本”，因为其质量较差。随着模仿学习训练的进行，智能体演示被视为“正样本”的概率应该相应地增加。因此，继续将智能体演示视为含有大量的“负样本”是不合适的。为解决这个问题，CAIL算法中设计了一个动态的α更新方法。α的值在训练过程中会从0.3线性增加至0.5。α_t表示智能体样本在校准监督对比学习中被视为高质量、专家相似样本的程度。在训练早期，智能体策略与专家策略仍有明显差距，因此不应该赋予过高的α_t。随着训练进行，智能体样本质量逐渐提升，因此逐步增大α_t可以更好地反映智能体策略逐渐接近专家策略的过程。我们将上限设置为0.5是出于保守建模的考虑：即使在训练后期，也不能假设大多数智能体样本都已经达到专家质量，0.5表示高质量样本和低质量样本处于相对平衡的混合状态，能够避免过度乐观地将次优智能体样本作为专家正样本。图5中展示了基于动态超参数α的实验结果，可以观察到设置不同α值对CAIL算法的影响不大，且在大多数情况下使用动态α可以导致更好的性能。

2.3 消融实验

消融实验主要验证了算法中不同模块的功能和有效性。在CAIL算法的消融实验中，主要研究了不同数据增广方式对算法结果的影响，以及对比学习和直接进行数据增强的比较。

表2 数据增广和对比学习的实验结果对比

Tab. 2 Table of results between using data augmentation and contrastive learning

500K时间步	GAIL-SE（无数据增广）	GAIL-SE（有数据增广）	CAIL
Cartpole Swing	734	729±40	**838 4**
Cheetah Run	514	533±34	**538 34**
Hopper Stand	270	206±89	**541 305**
Walker Stand	513	881±15	**961 9**
Walker Walk	268	200±5	**463 126**
Quadruped Run	212	222±74	**306 110**

2.3.1 对比学习算法分析

根据文献（Jeong等，2021）中的结论，在模型的对抗训练中直接使用数据增强可能会导致不稳定性，进而导致模型训练的失败。为了在对抗模仿学习中验证这一观点，实验中对GAIL-SE算法直接应用不同的数据增强方式，实验结果如表2中所示。除了Walker Stand环境外，可以观察到在GAIL-SE环境中使用数据增强对提高性能的作用不明显，和不使用数据增强相比，其回报平均下降了40.4%。相比之下，CAIL算法在这5个环境中表现均优于GAIL-SE算法，回报平均提高了47.2%。因此，可以得出结论，使用基于校准的对比损失结合数据增强能够增强对抗模仿学习算法的表现，而直接在对抗模仿学习中应用数据增强难以使得模型提升效果。

2.3.2 数据增强策略分析

在实验中，使用Random-Shift数据增广方式来生成智能体专家视觉状态的增广数据，该数据增广方式已经被广泛应用于视觉强化学习中，并被证明是最有效的数据增强技术（Laskin等，2020）。为了验证其他数据增强方式（例如Random Crop、Random Cutout、Random Aug）对模仿学习性能产生的影响，在Hopper Stand环境中进行了实验验证，实验结果见表3。从表格中可以观察到CAIL算法在所有4种不同的增强方式下都取得了比GAIL-SE算法更好的表现，并且使用Random-Shift增强能够达到最佳性能。同时，表格中的结果也说明了使用部分数据增强方式可能会降低对抗模仿学习的性能（例如Random Cutout和Random Aug）。该结果的原因在于直接将数据增强应用于GAIL-SE时，增强样本会直接

参与对抗判别和奖励估计，可能增加训练不稳定性，而CAIL使用数据增强构造对比学习中的正样本视图，将其转化为表征一致性约束，因此能够更有效地利用数据增强。

表3　使用不同数据增广方式的实验结果对比

Tab. 3　Table of results using different data augmentation ways

增广方式	GAIL-SE	CAIL
无数据增广方式	602±163	N/A
+ Random Shift	676±130	**777　42**
+ Random Crop	649±208	**755　71**
+ Random Cutout	530±260	**691　92**
+ Random Aug	524±165	**549　115**

2.3.3　不同损失函数的分析

本文针对不同对比损失函数的消融实验，以分别验证无监督对比损失、有监督对比损失和校准监督对比损失的独立贡献。本文新增在Quadruped Run环境进行消融实验结果如表4所示。

表4　关于不同对比损失函数的消融实验结果表

Tab. 4　Table of results using different contrastive losses

损失函数	平均回报
GAIL-SE	259±78
$+L_{\mathrm{UnSupCon}}$（作用于智能体样本）	329±27
$+L_{\mathrm{UnSupCon}}$（作用于专家演示样本）	253±36
$+L_{\mathrm{SupCon}}$	302±90
$+L_{\mathrm{C-SupCon}}$	319±56
CAIL (w/o cal)	332±62
CAIL	**346　73**

从结果可以看出，GAIL-SE加入作用于智能体样本的无监督对比损失后，性能显著提升，说明智能体回放缓冲区中大量且持续更新的样本能够为视觉编码器提供有效的表征学习信号。相比之下，作用于专家样本的无监督对比损失的结果未带来明显提升，主要原因是专家演示数据规模较小且固定，难以提供足够丰富的样本变化。进一步地，加入普通有监督对比损失L_{SupCon}后，平均回报亦大幅提升，说明区分专家样本和智能体样本有助于增强表征判别性。替换为校准监督对比损失$L_{\mathrm{C-SupCon}}$后，性能进一步提升，表明动态建模智能体样本质量能够带来更稳定的表征学习效果。最终完整CAIL，取得最高的平均回报。该结果说明，智能体样本上的无监督对比损失和校准监督对比损失具有互补作用，完整CAIL能够取得最佳性能。

2.4　可视化实验

图6中展示了DMControl软件中智能体的可视化运动轨迹，智能体以CAIL算法在1M时间步的训练出的智能体策略为支撑，并与环境交互。在交互过程中，记录智能体的运动状态。图中的结果表明，通过CAIL算法学习的策略可以使智能体成功地完成快速奔跑的任务。例如，在Cheetah智能体的跑步任务中，经过训练的智能体始终保持快速步伐。此外，在Walker智能体的行走任务中，经过训练的智能体表现出平稳的行走动作。在Hopper智能体站立任务中，经过训练的Hopper智能体能够从平躺的随机状态出发，最终完成站立的姿势并持续保持该动作。

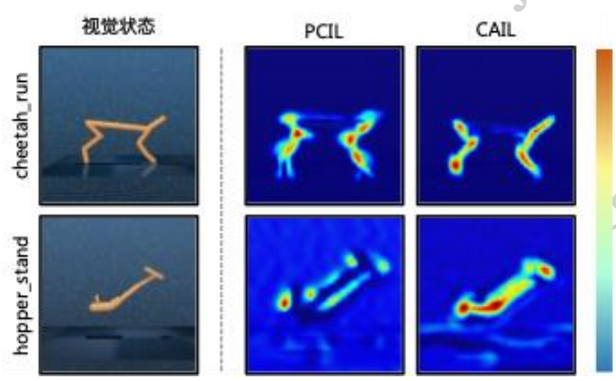


图7　模型的空间注意力图

Fig. 7　Spatial attention of trained model.

如图7所示，通过Grad-CAM（Selvaraju等，2017）对Cheetah Run和Hopper Stand两个环境中训练得到的判别器的空间注意力图进行了可视化。图中展示了CAIL和与其最接近的对比算法PCIL的结果。通常，空间注意力图能够直观地突出显示判别器在决策中优先考虑的区域，红色区域表示对决策的影响更大，蓝色区域表示对决策的影响更小。从图中可以明显观察出，两种算法的注意图都能够聚焦在智能体的身体关节上，这说明智能体的关节部分的运动情况对区分不同的视觉演示很重要。与PCIL算法相比，CAIL算法的注意图在智能体身体关节上的覆盖面更广且定位更精确。该实验结果说明了

CAIL算法训练得到的判别器在捕捉不同视觉状态之间的差异方面具有更好的能力，进而为后续策略模型的训练提供了精准的反馈。因此，CAIL算法相比PCIL算法能够取得更好的性能。可视化样本均由训练至1M步后收敛的智能体策略模型生成。具体而言，我们使用训练完成后的策略与环境交互，并从得到的测试轨迹中随机采样视觉状态进行Grad-CAM可视化。

3 结 论

本文针对高维视觉观测条件下对抗模仿学习性能下降的问题，从表示学习角度对视觉对抗模仿学习方法进行了研究，为视觉模仿学习中的表征学习问题提供了新的思路。分析表明，当输入由低维本体状态转变为像素级视觉观测时，判别器难以稳定提供具有判别性的回报信号，进而影响视觉编码器的表示学习能力，最终限制策略优化效果。为此，本文在对抗模仿学习框架中引入对比表示学习机制，提出校准对比对抗模仿学习方法，通过无监督对比学习挖掘智能体视觉样本之间的结构信息，并结合监督对比约束增强专家样本与智能体样本之间的可分性，同时利用校准机制动态刻画智能体策略逐步接近专家策略的演化过程，从而提升视觉表示的判别能力与训练稳定性。在DMControl视觉控制任务上的实验结果表明，所提出方法在模仿性能和样本效率方面均优于多种现有视觉模仿学习方法，并能够更有效地利用回放缓冲区中的视觉数据。

尽管如此，本文方法仍存在一定局限性。首先，算法依赖数据增强策略来构造对比学习中的正样本视图，不同增强方式可能会影响最终性能，如何设计更加稳健、任务自适应的数据增强策略仍值得进一步研究。其次，本文实验主要在仿真环境中进行。真实机器人视觉场景通常包含更复杂的光照变化、相机视角变化以及仿真到真实环境之间的动力学差异，因此所提出方法在真实机器人任务中的泛化能力仍需要进一步验证。未来工作将探索更加稳健的自监督视觉表征学习机制，并在真实机器人视觉模仿学习任务中进一步评估本文方法的实际应用潜力。

参考文献(References)

Brantley K, Sun W and Henaff M. 2020. Disagreement-regularized imitation learning//Proceedings of 2020 International Conference on Learning Representations [EB/OL: https://openreview.net/pdf? id=rkgbYyHtwB]

Brown D, Goo W, Nagarajan P and Niekum S. 2019. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR: 783-792 [EB/OL: https://proceedings.mlr.press/v97/brown19a/brown19a.pdf]

Chen T, Kornblith S, Norouzi M and Hinton G. 2020. A simple framework for contrastive learning of visual representations//Proceedings of the 37th International Conference on Machine Learning. Virtual Event: PMLR: 1597-1607 [DOI: 10.5555/3524938.3525087]

Cobbe K, Klimov O, Hesse C, Kim T and Schulman J. 2019. Quantifying generalization in reinforcement learning//Proceedings of the 36th International Conference on Machine Learning. Long Beach: PMLR: 1282-1289 [EB/OL: https://proceedings. mlr. press/v97/cobbe19a.pdf]

Haldar S, Mathur V, Yarats D and Pinto L. 2022. Watch and match: Supercharging imitation with regularized optimal transport [EB/OL]. [2026-03-15]. https://arxiv.org/abs/2206.15469

Ho J and Ermon S. 2016. Generative adversarial imitation learning// Advances in Neural Information Processing Systems. Barcelona: Curran Associates: 4565-4573 [DOI: 10.5555/3157382.3157608]

Huang J, Yin Z H, Hu Y and Gao Y. 2023. Policy contrastive imitation learning//Proceedings of the 40th International Conference on Machine Learning. Honolulu: PMLR: 14007-14022 [DOI: 10.5555/3618408.3618978]

Jaderberg M, Czarnecki W M, Dunning I, Marris L, Lever G and Castaneda A G, et al. 2018. Human-level performance in 3D multiplayer games with population-based reinforcement learning [EB/OL]. [2018-07-03]. https://arxiv.org/abs/1807.01281

Ke L, Choudhury S, Barnes M, Sun W, Lee G and Srinivasa S. 2020. Imitation learning as f-divergence minimization//Proceedings of the International Workshop on the Algorithmic Foundations of Robotics. Cham: Springer: 313-329 [DOI: 10.1007/978-3-030-66723-8_19]

Khosla P, Teterwak P, Wang C, Sarna A, Tian Y and Isola P, et al. 2020. Supervised contrastive learning// Proceedings of the 34th International Conference on Neural Information Processing Systems [DOI: 10.5555/3495724.3497291]

Laskin M, Srinivas A and Abbeel P. 2020. CURL: Contrastive unsupervised representations for reinforcement learning//Proceedings of the

37th International Conference on Machine Learning. Virtual Event: PMLR: 5639-5650 [DOI: 10.5555/3524938.3525461]

Liu M H, He T R, Zhang W N, Yan S C and Xu Z W. 2023. Visual imitation learning with patch rewards//Proceedings of the 2023 International Conference on Learning Representation [EB/OL: https://openreview.net/forum? id=OnM3R47KIiU]

Pomerleau D A. 1988. ALVINN: An autonomous land vehicle in a neural network//Proceedings of the 2nd International Conference on Merual Information Processing Systems. Denver: Morgan Kaufmann [DOI: 10.5555/2969735.2969771]

Puterman M L. 1994. Markov Decision Processes: Discrete Stochastic Dynamic Programming. New York: Wiley [DOI: 10.1002/9780470316887]

Rafailov R, Yu T, Rajeswaran A and Finn C. 2021. Visual adversarial imitation learning using variational models//Proceedings of the 35th International Conference on Neural Information Processing Systems 3016-3028 [DOI: 10.5555/3540261.3540492]

Ross S, Gordon G and Bagnell D. 2011. A reduction of imitation learning and structured prediction to no-regret online learning//Proceedings of the 14th International Conference on Artificial Intelligence and Statistics. Fort Lauderdale: JMLR Workshop and Conference Proceedings: 627-635 [EB/OL: https://proceedings.mlr.press/v15/ross11a.html]

Schulman J, Levine S, Abbeel P, Jordan M and Moritz P. 2015. Trust region policy optimization//Proceedings of the 32nd International Conference on Machine Learning. Lille: PMLR: 1889-1897 [DOI: 10.5555/3045118.3045319]

Schulman J, Wolski F, Dhariwal P, Radford A and Klimov O. 2017. Proximal policy optimization algorithms [EB/OL]. [2026-03-15]. https://arxiv.org/abs/1707.06347

Selvaraju R R, Cogswell M, Das A, Vedantam R, Parikh D and Batra D. 2017. Grad-CAM: Visual explanations from deep networks via gradient-based localization//Proceedings of the IEEE International Conference on Computer Vision. Venice: IEEE: 618-626 [DOI: 10.1109/ICCV.2017.74]

Silver D, Huang A, Maddison C J, Guez A, Sifre L and Van Den Driessche G, et al. 2016. Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587): 484-489 [DOI: 10.1038/nature16961]

Silver D, Lever G, Heess N, Degris T, Wierstra D and Riedmiller M. 2014. Deterministic policy gradient algorithms//Proceedings of the 31st International Conference on Machine Learning. Beijing: PMLR [DOI: 10.5555/3044805.3044850]

Sutton R S and Barto A G. 2018. Reinforcement Learning: An Introduction. 2nd ed. Cambridge: MIT Press [DOI: 10.5555/3312046]

Todorov E, Erez T and Tassa Y. 2012. MuJoCo: A physics engine for model-based control//Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems. Vilamoura: IEEE: 5026-5033 [DOI: 10.1109/IROS.2012.6386109]

Tu S, Robey A, Zhang T and Matni N. 2022. On the sample complexity of stability constrained imitation learning//Proceedings of the 4th Annual Learning for Dynamics and Control Conference. Palo Alto: PMLR: 180-191 [EB/OL: https://proceedings.mlr.press/v168/tu22a.pdf]

Tucker A, Gleave A and Russell S. 2018. Inverse reinforcement learning for video games [EB/OL]. [2026-03-15]. https://arxiv.org/abs/1810.10593

Tunyasuvunakool S, Muldal A, Doron Y, Liu S, Bohez S and Merel J, et al. 2020. dm_control: Software and tasks for continuous control [EB/OL]. [2020-09-07]. https://arxiv.org/abs/2006.12983

Van Hasselt H, Guez A and Silver D. 2016. Deep reinforcement learning with double Q-learning//Proceedings of the 30th AAAI Conference on Artificial Intelligence. Phoenix: AAAI Press [DOI: 10.5555/3016100.3016191]

Yarats D, Fergus R, Lazaric A and Pinto L. 2021a. Reinforcement learning with prototypical representations//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: PMLR: 11920-11931 [EB/OL: https://proceedings.mlr.press/v139/yarats21a/yarats21a.pdf]

Yarats D, Zhang A, Kostrikov I, Amos B, Pineau J and Fergus R. 2021b. Improving sample efficiency in model-free reinforcement learning from images//Proceedings of the 35th AAAI Conference on Artificial Intelligence. Virtual Event: AAAI Press: 10674-10681 [DOI: 10.1609/aaai.v35i12.17276]

Yifei L, Xuemin H, Guowen C, Shihao L, Long C. Review of end-to-end motion planning for autonomous driving with visual perception [J]. Journal of Image and Graphics, 2021, 26(1): 49-66. (刘旖菲, 胡学敏, 陈国文, 刘士豪, 陈龙. 视觉感知的端到端自动驾驶运动规划综述[J]. 中国图象图形学报, 2021, 26(1): 49-66) DOI: 10.11834/jig.200276.

作者简介

王云柯,男,博士后研究员,主要研究方向为强化学习、具身智能。E-mail:yunke.wang@whu.edu.cn

陶林伟,男,博士研究生,主要研究方向为可信人工智能。E-mail: linwei.tao@sydney.edu.au

林雨恬,女,副教授,主要研究方向为计算机视觉。E-mail: yutian.lin@whu.edu.cn

杜博,通信作者,男,教授,主要研究方向为计算机视觉。E-mail:dubo@whu.edu.cn

徐畅,男,副教授,主要研究方向为计算机视觉。E-mail: c.xu@sydney.edu.au