arXiv:2303.03832v1 [cs.NE] 7 Mar 2023

采用描述符条件化梯度的 MAP-Elites 与将档案库蒸馏为单一策略

Maxence Faldor
m.faldor22
@imperial.ac.uk 伦敦帝国理工学院 伦敦，英国

Félix Chalumeau
f.chalu
meau@instadeep.com
InstaDeep 开普敦，南非

Manon Flageat
manon.flageat18@i
mperial.ac.uk 伦敦帝国理工学院 伦敦，英国

Antoine Cully
a.cully@imperial.ac.uk
伦敦帝国理工学院 伦敦，英国

摘要

质量-多样性算法（如 MAP-Elites）是进化计算的一个分支，旨在生成多样且性能优异的解集，已成功应用于各种领域，尤其是进化机器人学。然而，MAP-Elites 基于源自遗传算法的随机变异进行发散式搜索，因此仅限于进化低维解的种群。PGA-MAP-Elites 通过集成一种受深度强化学习启发、基于梯度的变异算子，克服了这一局限，使大规模神经网络的进化成为可能。尽管 PGA-MAP-Elites 在许多环境中表现出色，但在若干任务中仍会失败，因为基于梯度的变异算子进行收敛式搜索时，无法引导变异朝着能够改进档案库的解发展。在这项工作中，我们提出两项贡献：（1）我们为策略梯度变异算子引入描述符条件评论家，使档案库在整个描述符空间内得到改进；（2）我们利用演员-评论家训练，以零额外成本学习描述符条件策略，将档案库中的知识提炼到单一的多用途策略中，使其能够执行档案库所涵盖的全部行为。我们的算法 DCG-MAP-Elites 在一组具有挑战性的运动任务中，平均将 QD 分数提升了 82%，优于 PGA-MAP-Elites。

1 引言

进化一个引人入胜的方面，是其能够产生大量不同的物种，并且每个物种都适应其生态位。受这一思想启发，质量-多样性（QD）优化是一类进化算法，旨在针对问题生成一组兼具多样性和高性能的解 [3, 7, 32]。不同于返回单个高性能解的传统优化方法，QD 算法的目标是探索一个感兴趣的搜索空间，该空间称为 描述符空间 [27]。生成大量多样且有效的解，可以为解决单个问题提供多个备选方案，这在机器人学中十分有用，有助于提高鲁棒性、从损伤中恢复 [6] 或缩小现实差距 [4]。此外，传统优化方法容易陷入局部最优，而保留针对某一问题的多样化解，有助于找到能够引导我们获得全局更优解的进化阶梯 [27, 28]。多样性的另一项益处在于：当奖励信号稀疏或具有欺骗性时，搜索能够实现高效探索 [2, 8, 31]。

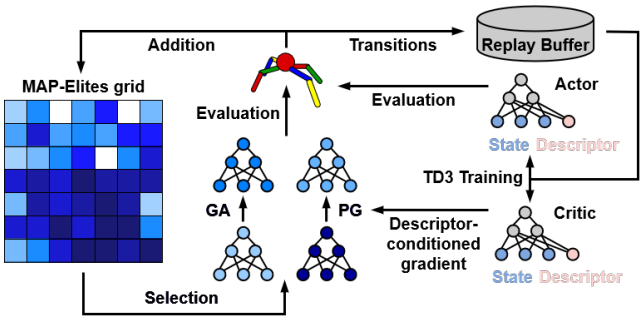

图 1：DCG-MAP-Elites 执行标准的 MAP-Elites 循环，包括选择、变异、评估和添加。它采用两种互补的变异算子：（1）用于探索的标准遗传算法（GA）变异算子；（2）用于提升适应度的描述符条件策略梯度（PG）变异算子。在评论家训练的同时，档案库中的知识也作为副产品蒸馏到描述符条件化演员中。

MAP-Elites [27] 是一种概念简单但有效的 QD 优化算法，在各种应用中均展现出竞争力，可生成大量多样化的技能。然而，MAP-Elites 依赖随机变异，在大型搜索空间中可能导致收敛缓慢 [5, 28, 31],，因此不适合演化具有大量参数的神经网络。

相比之下，深度强化学习（DRL） [25, 26]算法将强化学习与基于梯度的方法的定向搜索能力相结合，以学习单一解。DRL 能在电子游戏中超越人类表现 [40],，在棋盘游戏中击败世界冠军 [35]，并控制连续动作空间中的复杂机器人 [17],；这一直是人工智能领域的一项长期挑战。策略梯度方法已展现出最先进的成果，能够在高维状态空间和连续动作空间中学习包含数千个参数的大型神经网络策略 [18, 24, 36]。

PGA-MAP-Elites [28] 是 MAP-Elites 的扩展，它利用 TD3 算法 [15] 整合了深度强化学习的样本效率。

该算法采用策略梯度（PG）变异算子来高效提升适应度，并结合常用的遗传算法（GA）变异算子。策略梯度变异算子利用深度强化学习推导出的梯度来提升适应度，并引导变异朝全局最优解迈进；同时，遗传算法变异算子的发散式搜索则支持探索和优化 [10]。近期其他研究也提出了将质量多样性算法与强化学习相结合的方法 [31, 38]，以应对复杂的机器人任务。

PGA-MAP-Elites 在迄今文献所考虑的大多数环境中都取得了最先进的表现 [28, 31, 38]。然而，当全局最优解位于搜索空间中某个不太可能产生被纳入档案库的后代的区域时，PG变异算子在相关任务中便会失效。例如，考虑一项运动任务，其中适应度定义为能耗的相反数，描述符则定义为机器人的最终位置。适应度的全局最优解是机器人保持静止的解，以尽可能降低能耗。因此，PG变异算子会促使解保持静止，使其描述符收敛到单个点，即全局最优解的描述符。结果，PG变异算子生成的后代会被丢弃，无法找到有意义的过渡解，进而妨碍多样性。

在本研究中，我们提出描述符条件梯度 MAP-Elites（DCG-MAP-Elites）。该算法以 PGA-MAP-Elites 算法为基础，通过引入描述符条件评论家来增强 PG变异算子；该评论家能够根据目标描述符提供梯度。描述符条件评论家以状态和描述符作为输入，评估动作。借助这种描述符条件评论家，PG变异算子能够变异解，生成适应度更高且以期望描述符为目标的后代，从而避免描述符收敛到单个点。

此外，作为PG变异算子所使用的深度强化学习算法，TD3要求并行训练一个演员和一个评论家。我们利用这种相互交织的演员-评论家训练，同时使演员也具备描述符条件化能力，从而允许其根据当前状态以及我们希望实现的输入描述符来采取动作。因此，演员不再采取全局最大化适应度的动作，而是采取在实现期望描述符的同时最大化适应度的动作。训练结束时，我们可以以期望描述符为条件，使演员执行一种采取能够实现该期望描述符的动作的策略。在一半的任务中，我们观察到，描述符条件化演员能够以相近的QD分数实现档案库中包含的全部描述符范围，从而免除了处理由数千个解组成的集合的负担。

总而言之，我们提出了两项贡献：(1) 使用描述符条件评论家增强策略梯度变异算子；(2) 将档案库中的知识蒸馏到一个单一的多功能策略中，且无需额外成本。我们将该算法与四种最先进的质量多样性算法在四个高维机器人运动任务上进行了比较。结果表明，DCG-MAP-Elites 的质量多样性得分平均比 PGA-MAP-Elites 高 82%。

2 背景

2.1 问题陈述

我们考虑一个智能体在离散时间步 t 上与环境进行顺序交互，回合长度为 T。在每个时间步t，智能体观测一个状态 s_t，执行一个动作 a_t并接收标量奖励 r_t。我们将其建模为一个马尔可夫决策过程（MDP），该过程由一个状态空间 S、一个连续动作空间 A、一个平稳的转移动力学分布 $p(s_{t+1} \mid s_t, a_t)$以及一个奖励函数 r组成：$\mathrm{S} \times \mathrm{A} \to \mathbb{R}$。在本文中，一个策略（也称为解）是一个由 $\phi \in \Phi$参数化的确定性神经网络，记作 $\pi_\phi : \mathrm{S} \to \mathrm{A}$。智能体使用其策略选择动作并与环境交互，从而生成由状态、动作和奖励组成的轨迹。一个解的适应度 由 F给出： $\Phi \to \mathbb{R}$，其定义为期望折扣回报 $\mathbb{E}_{\pi_\phi}\left[\sum_{t=0}^{T-1} \gamma^t r_t\right]$。

在该MDP设定中，质量多样性算法的目标是在描述符空间D的每个点上找到适应度最高的解。描述符函数 D: $\Phi \to$ D通常由用户定义，并以一种有意义的方式刻画解，以满足所需的多样性类型。采用这一记号，我们的目标是演化出一个解的种群，使其相对于 F 具有高性能，并相对于 D具有多样性。

2.2 MAP-Elites

多维表型精英档案（MAP-Elites） [27]是一种简单而有效的 QD 算法，它将描述符空间 D 离散化为由单元组成的多维网格，称为档案库 X，并在每个单元中搜索最佳解，见算法 4。该算法的目标是返回一个尽可能填满高适应度解的档案库。MAP-Elites 首先生成随机解并将其加入档案库。随后，算法重复以下步骤，直至评估完 I 个解为止：(1) 从档案库中均匀选取一批解，通过变异和/或交叉对其进行修改，以生成后代；(2) 评估后代的适应度和描述符，并将每个后代放入对应单元，当且仅当该单元为空，或后代的适应度高于该单元中的当前解时，才执行此操作；在后一种情况下，当前解将被替换。与大多数进化方法一样，MAP-Elites 采用与适应度目标无关的无定向更新。使用遗传算法（GA）变异算子时，MAP-Elites 会进行发散式搜索；由于缺乏定向搜索能力，这可能导致其在高维问题中收敛缓慢，因此它在低维搜索空间中表现最佳 [28]。

2.3 深度强化学习

深度强化学习（DRL） [25, 26] 将强化学习框架与深度神经网络的函数逼近能力相结合，以便在高维状态和动作空间中表示策略和价值函数。与进化算法等黑盒优化方法不同，深度强化学习利用贝尔曼方程所体现的马尔可夫决策过程（MDP）结构，从而实现更高的样本效率。其目标是找到最优策略π_ϕ，使预期回报或适应度 $F(\pi_\phi)$最大化。在强化学习中，许多方法尝试估计动作价值函数 $Q^\pi(s,\ a) = \mathbb{E}_\pi\left[\sum_{t=0}^{T-1} \gamma^t r_t | s, a\right]$

其定义为：从状态 s开始，执行动作 a，之后遵循策略 π所获得的预期折扣回报。

双延迟深度确定性策略梯度（TD3）算法 [15] 是一种演员-评论家式的离策略强化学习方法，能够在动作空间较大且连续的环境中取得最先进的结果。TD3 通过最大化动作价值函数 $Q_\theta(s,a)$，间接学习策略 π_ϕ。这种方法与 Q学习 [15] 密切相关，并尝试近似最优动作价值函数 $Q^*(s,a)$，以找到最优动作 $a^*(s) = \arg\max_a Q^*(s,a)$。然而，在连续动作空间中，计算 $\max_a Q_\theta(s,a)$ 中的动作最大值难以实现，因此通过 $\max_a Q_\theta(s,a) = Q_\theta(s,\pi_\phi(s))$ 进行近似。在 TD3 中，策略 π_ϕ在环境中执行动作，并将状态转移存储在经验回放缓冲区中。随后，使用收集到的经验通过时序差分训练一对评论家网络 Q_{θ_1},Q_{θ_2}，并更新目标网络Q_{θ_1}', Q_{θ_2}'，使其缓慢跟踪主网络。两个评论家都使用单一回归目标 y；该目标通过选取两个评论家中目标值较小者计算得出，并通过采样噪声 $\epsilon \sim \text{clip}(\mathcal{N}(0,\sigma), -c, c)$ 实现目标策略平滑：

$$y = r(s_t,a_t) + \gamma \min_{i=1,2} Q_{\theta_i'}\left(s_{t+1}, \pi_{\phi'}(s_{t+1}) + \epsilon\right) \tag{1}$$

两个评论家网络均通过对该目标进行回归来学习，而策略则延迟更新，每隔 Δ 次迭代才更新一次，只需最大化 Q_{θ_1}，其中 $\max_\phi \mathbb{E}\left[Q_{\theta_1}(s, \pi_\phi(s))\right]$。演员使用确定性策略梯度进行更新：

$$\nabla_\phi J(\phi) = \mathbb{E}\left[\nabla_a Q_{\theta_1}(s,a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)\right] \tag{2}$$

2.4 PGA-MAP-Elites

策略梯度辅助 MAP-Elites（PGA-MAP-Elites） [28] 是 MAP-Elites 的一种扩展，旨在结合深度强化学习方法的定向搜索能力和样本效率，以及遗传算法的探索能力，从而演化深度神经网络，见算法 5。该算法按照常规 MAP-Elites 循环执行选择、变异、评估和添加，迭代预算为 I 次，但采用两个并行的变异算子：一半后代由标准遗传算法（GA）变异算子生成，另一半则由策略梯度（PG）变异算子生成。在循环的每次迭代中，PGA-MAP-Elites 会将后代评估产生的状态转移存入经验回放缓冲区 B，并利用这些数据基于 TD3算法训练一对评论家网络，具体见算法 6。随后，在 PG变异算子中使用训练好的评论家网络，对档案库中选出的解执行 m 个梯度步，以选择能够最大化近似动作价值函数的动作，具体见算法 7。在每次迭代中，评论家网络会进行 n 步梯度下降，朝着式 1 所述的目标进行训练；该目标是在从经验回放缓冲区 B 中均匀采样的 N 个经验状态转移上取平均得到的。演员（也称为贪心演员 [28]）会延迟学习 Δ，通过根据式 2 最大化评论家网络来进行更新。

3 相关工作

3.1 将 Quality-Diversity 扩展到神经进化

在高维搜索空间中进化出多样化解这一挑战，近年来一直是一个活跃的研究课题。ME-ES [5] 将 MAP-Elites 扩展到由大型神经网络参数化的高维解。该算法利用进化策略执行定向搜索，其效率高于遗传算法中使用的随机变异。它通过在父代解的多个扰动版本上进行局部估计，得出适应度梯度，以生成新的解。种群会逐渐趋向参数空间中适应度较高的区域，但这需要对大量解进行采样和评估，因此数据效率尤其低。为了利用时间步级别的信息并提高数据效率，研究者提出了将 MAP-Elites 与强化学习 [28, 30, 31, 38]相结合的方法；实践证明，这些方法能够高效地进化出适用于复杂任务、性能出色且多样化的神经网络种群。PGA-MAP-Elites [28] 在部分变异中使用策略梯度，详情见第 2.4 节。CMA-MEGA [38] 利用进化策略估计描述符梯度，并通过 CMA-ES 机制 [12, 19]结合适应度梯度与描述符梯度。QD-PG [31] 根据访问状态的新颖性引入多样性奖励，并推导出用于最大化这些多样性奖励的策略梯度，从而有助于在奖励稀疏或具有欺骗性的情境中进行探索。PBT-MAP-Elites [30] 将 MAP-Elites 与基于种群的训练流程 [21] 相结合，以优化多样化强化学习智能体的超参数。有意思的是，近期一项研究 [37]使用纯进化策略，将算法 CMA-MAE [13]扩展到机器人任务中的高维策略，并表明其数据效率与 QD-RL 方法相当。该方法展现出竞争力，但表现仍不及 PGA-MAP-Elites。

3.2 对评论家进行条件化

上述方法在推导用于突变解的策略梯度时，都没有考虑描述符。换句话说，它们既不使用描述符条件策略，也不使用像我们的方法 DCG-MAP-Elites 那样的描述符条件评论家。描述符条件评论家的概念与通用价值函数近似器 [33] 的概念相关，而在使用该概念的领域中，与质量-多样性最相关的是技能发现强化学习 [1]。在 VIC、DIAYN、DADS 和 SMERL [9, 16, 23, 34], 中，使用了条件化演员-评论家，但其条件是采样得到的先验，并不像 DCG-MAP-Elites 中那样对应于真实的后验。此外，这些方法在步级别使用多样性，而不是像我们的方法一样显式地在轨迹级别使用多样性。最后，它们不使用档案库来存储种群，因此最终策略集合要小得多。最终，已有研究表明，就其在适应和层次化学习中的应用而言，QD 方法进化出的行为 [1], 与该类方法学习到的技能具有竞争力。

3.3 档案蒸馏

将档案库中的知识蒸馏到单个神经模型中，是一种颇具吸引力的过程：它能减少算法输出的参数数量，并实现泛化以及插值/外推。尽管蒸馏通常指策略蒸馏——即从教师策略中学习观测/动作映射——我们提出将档案蒸馏作为一个通用术语，泛指从档案库向

另一个模型迁移任何类型的知识，无论是策略、在环境中经历的状态转移、完整轨迹，还是发现的描述符。

据我们所知，有两项与 QD 相关的研究采用了档案蒸馏的概念。Go-Explore [8] 会存储到达状态的档案库，并训练一个目标条件策略，以重现到达该状态的策略所执行的轨迹。另一种有趣的档案蒸馏方法是，基于档案库中包含的策略学习一个生成式策略网络 [22]。我们提出的 DCG-MAP-Elites 则将档案库中的经验蒸馏到一个多功能策略中。

4 DCG-MAP-ELITES

我们提出的新方法——描述符条件梯度 MAP-Elites（DCG-MAP-Elites）——通过利用描述符条件评论家来改进 PG 变异算子，从而克服 PGA-MAP-Elites 的局限性；同时，作为演员-评论家训练的副产品，它还能将档案库中的知识蒸馏到一个多功能策略中。伪代码见算法 1。该算法遵循常见的 MAP-Elites 循环，在预算为 I 次迭代的情况下，依次进行选择、变异、评估和添加。算法并行使用两种互补且相互独立的变异算子：1）标准遗传算法算子；2）描述符条件 PG 算子。在每次迭代中，评估步骤产生的状态转移都会存入经验回放缓冲区，并用于训练一个基于 TD3 的演员-评论家对。

与 PGA-MAP-Elites 不同，演员-评论家对以描述符为条件。除了状态 s 和动作 a，评论家 $Q_\theta(s,\ a\ |d)$也取决于描述符 d，并估计从状态 s开始、采取动作 a，随后遵循策略 π并实现描述符 d时的期望折扣回报。实现描述符 d 意味着，由策略 π生成的轨迹的描述符应为 d。除了状态 s，演员 $\pi_\phi(s \mid d)$ 也取决于描述符 d，并在实现描述符 d的条件下最大化期望折扣回报。因此，描述符条件化演员的目标是在最大化适应度的同时，实现输入描述符 d。

4.1 描述符条件评论家

我们不再使用 $Q_\theta(s,\ a)$估计动作价值函数，而是希望使用 $Q_\theta(s,\ a \mid d)$估计描述符条件动作价值函数。当策略 π与环境交互一个长度为 T 的回合时，会生成一条轨迹 τ，它由一系列状态转移组成：

$$(s_0, a_0, r_0, s_1), \ldots, (s_{T-1}, a_{T-1}, r_{T-1}, s_T)$$

具有描述符 $D(\pi) = d$。我们将状态转移（$s,\ a,\ r,\ s'$）的定义扩展为包含策略（$s,\ a,\ r,\ s', d$）的描述符 d。因此，具有描述符 d的轨迹 τ给出一个状态转移序列：

$$(s_0, a_0, r_0, s_1, d), \ldots, (s_{T-1}, a_{T-1}, r_{T-1}, s_T, d)$$

不过，描述符只有在回合结束时才可用，因此只有回合结束后才能用描述符补充状态转移。在我们考虑的所有任务中，奖励函数均为正 $r\colon \mathrm{S} \times \mathrm{A} \to \mathbb{R}^+$，因此适应度函数 F和动作价值函数也均为正。因此，对于任意采样描述符 $d' \in \mathrm{D}$，我们将描述符条件评论家定义为：当策略实现采样描述符 d' 时，其值等于普通动作价值函数；当

算法 1 DCG-MAP-Elites

输入： 批次大小 b，遗传算法变体数量 $g \le b$
使用 b 个随机解初始化档案库 X 和经验回放缓冲区 B
初始化评论家网络Q_{θ_1}、 Q_{θ_2} 和演员网络 π_ϕ
$i \leftarrow 0$
当 $i < I$ **时执行**
训练_演员_评论家(Q_{θ_1}, Q_{θ_2}, π_ϕ, B)
$\pi_{\psi_1}, \ldots, \pi_{\psi_b} \leftarrow$ 选择(X)
$\pi_{\widehat{\psi}_1}, \ldots, \pi_{\widehat{\psi}_g} \leftarrow$ 变异_遗传算法($\pi_{\psi_1}, \ldots, \pi_{\psi_g}$)
$\pi_{\widehat{\psi}_{g+1}}, \ldots, \pi_{\widehat{\psi}_b} \leftarrow$ 变异_PG($\pi_{\psi_{g+1}}, \ldots, \pi_{\psi_b}$, Q_{θ_1}, $\mathcal{B}$)
添加($\pi_{\widehat{\psi}_1} \ldots, \pi_{\widehat{\psi}_b}, \mathcal{X}, \mathcal{B}$)
$i \leftarrow i + b$
函数 添加($\mathcal{X}, \mathcal{B}, \pi_\phi, \pi_{\widehat{\psi}} \ldots$)
对于 $d' \in$ 从 b X 中采样的解 **执行**
(f,状态转移) $\leftarrow F(\pi_\phi(\ .\mid d'))$
插入(B, 状态转移)
对于 $\pi_{\widehat{\psi}} \ldots$ **执行**
(f, 状态转移) $\leftarrow F(\pi_{\widehat{\psi}})$, $d \leftarrow D(\pi_{\widehat{\psi}})$
insert(B, 状态转移)
如果 X (d) $= \emptyset$ 或 $F(\mathrm{X}(d)) < f$ **则**
X (d) $\leftarrow \pi_{\widehat{\psi}}$

策略未能实现采样描述符 d'时，其值等于零。给定一个状态转移 (s, a, r, s', d)，以及 $d' \in \mathrm{D}$，

$$Q_\theta(s, a \mid d') = \begin{cases} Q_\theta(s, a), & \text{if } d = d' \\ 0, & \text{if } d \neq d' \end{cases} \tag{3}$$

然而，采用这种分段定义后，描述符条件动作价值函数不连续，违反了通用逼近定理的连续性假设 [20]。为解决这一问题，我们引入相似度函数 $S\colon \mathrm{D}^2 \to]0,\ 1]$定义为 $S(d,\ d') = e^{-\frac{\|d-d'\|}{l \cdot \mathcal{D}}}$，以平滑描述符条件评论家，并将公式 3 放宽为：

$$\begin{aligned} Q_\theta(s, a \mid d') = S(d, d')\, Q_\theta(s, a) &= S(d, d')\, \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t r_t \middle| s, a \right] \\ &= \mathbb{E}_\pi \left[\sum_{t=0}^{T-1} \gamma^t S(d, d') r_t \middle| s, a \right] \end{aligned} \tag{4}$$

根据公式 4，我们证明，学习描述符条件评论家等价于按照相似度 $S(d,\ d')$ 缩放轨迹 d的描述符与采样描述符 d'之间的奖励。因此，公式 1 中的评论家目标需修改为纳入相似度缩放和描述符条件化演员：

$$y = S(d, d')\, r(s_t, a_t) + \gamma \min_{i=1,2} Q_{\theta_i'} \left(s_{t+1}, \pi_{\phi'}(s_{t+1} \mid d') + \epsilon \mid d' \right) \tag{5}$$

如果采样描述符 d' 大致等于轨迹 d的观测描述符 $d \approx d'$，那么我们有 $S(d, d') \approx 1$，因此奖励保持不变。然而，如果描述符 d'与观测描述符 d差异很大，那么奖励会缩减为 $S(d, d')\ r(s_t,\ a_t) \approx 0$。这种缩放确保奖励的大小不仅取决于动作 a相对于适应度函数 F的质量，还取决于是否实现了描述符 d'。给定一个状态转移（s, a, r, s', d），我们可以通过从 $d' \in \mathrm{D}$ 中采样，生成无限多个评论家更新。这一点被应用于新的

DCG-MAP-Elites 引入的演员-评论家训练，详见算法 2 和第 4.3 节。

算法 2描述符条件化行动者-评论家训练

函数 训练_演员_评论家(Q_{θ_1}, Q_{θ_2}, π_ϕ, B)
  对于 $t = 1 \to n$ **执行**
    从 B 中采样 N状态转移（s, a, r（s, a） s', d, d'）
    采样平滑噪声 ϵ
    $y \leftarrow S$（d, d'）r（s, a）$+\gamma$最小值$_{i=1,\ 2}\ Q_{\theta'_i}$（$s'$, $\pi_{\phi'}$（$s' \mid d'$）$+\epsilon \mid d'$）
    通过回归将两个评论家网络更新为 y
    如果 t 取模 Δ**则**
      使用确定性策略梯度更新演员：
      $\frac{1}{N} \sum \nabla_a Q_{\theta_1}(s,\ a \mid d')|_{a=\pi_\phi(s|d')}\ \nabla_\phi \pi_\phi(s \mid d')$
      软更新目标网络 $Q_{\theta i'}$ 和 $\pi_{\phi'}$

4.2 描述符条件化演员与档案蒸馏

评论家的训练需要训练一个演员 π_ϕ 来近似最优动作 a^*，如第 2.3 节所述。然而，在本文中，评论家估计的动作价值函数以描述符 d为条件。因此，我们不希望估计全局最佳动作，而是希望在策略应于轨迹结束时实现描述符 d 这一条件下，估计最佳动作。因此，演员被扩展为描述符条件策略$\pi_\phi(s \mid d)$，通过$\max_\phi \mathbb{E}\left[Q_\theta(s,\pi_\phi(s \mid d) \mid d)\right]$最大化描述符条件评论家的价值。演员使用确定性策略梯度进行更新，见算法 2：

$$\nabla_\phi J(\phi) = \frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a \mid d')|_{a=\pi_\phi(s|d')} \nabla_\phi \pi_\phi(s \mid d') \quad (6)$$

策略 $\pi_\phi(s \mid d)$ 学习提出动作 a，以优化奖励 $r(s,\ a)$，同时生成一条能够达到描述符 d 的轨迹。因此，描述符条件化演员可以展现广泛的描述符，从而有效地将档案库中的部分能力蒸馏到单个多用途策略中。

4.3 演员-评论家训练

在第 4.1 节中，我们展示了公式 5 中新的描述符条件评论家目标需要一个采样描述符 $d\prime$。在每次迭代中，通过评估后代生成的状态转移 $(s,\ a,\ r(s,\ a)\ s\prime d)$ 会被存储在经验回放缓冲区 B 中。为了学习处于状态 s、采取动作 a并达到描述符 $d\prime$之间的关系，评论家需要在“正”样本（即具有 $d\prime = d$ 的样本）和“负”样本（即具有 $d\prime \neq d$ 的样本）上进行训练。正样本很容易生成，因为我们可以对 $d\prime$ 进行采样，使其等于轨迹的观测描述符 d。然而，负样本需要定义一种任意的采样策略，而该策略可能会影响算法的性能。

此外，不进行环境中的主动交互而仅凭数据学习如何行动颇具难度 [29]。如果演员只被动地学习档案库中所存策略的后代所执行的滚动采样，就可能严重偏离状态-动作分布，从而阻碍收敛。为了解决这一被动学习问题并同时生成负样本，我们在环境中评估演员π_ϕ。每次迭代中，我们采样一批 b描述符 $d\prime$，并评估 $\pi_\phi(\ .\mid d\prime)$。动作器评估步骤会产生 b条轨迹，其中包含状态转移（s_t, a_t, r_t, s_{t+1}, d, $d\prime$），其中d 是轨迹实现的观测描述符，$d\prime$ 是用于评估而采样的输入描述符。通常，观测描述符与采样描述符并不相同，

$$d \neq \prime$$

，从 d而为评论家训练提供负样本。为了获得描述符 $d\prime$，我们从档案库中有放回地均匀采样 b 个解π_ψ，并为其描述符添加高斯噪声。这些解的描述符为 $D(\pi_\psi) = d_\psi$，我们计算一批描述符 $d\prime = d_\psi + \mathrm{N}(0_\mathcal{D}\ \sigma_d I)$。演员评估产生的负样本与后代评估产生的正样本一同存入经验回放缓冲区；对于后者，我们取 $d\prime = d$，详见算法 1 的添加函数。

4.4 描述符条件 PG 变异

学习到评论家 $Q_\theta(s, a \mid d)$ 后，可以按照算法3所述，用它来提升档案库中各解的适应度。从档案库中选取一个描述符为 $D(\pi_\psi)$ 的父代解 $\pi_\psi = d_\psi$。然后，我们应用策略梯度变异算子，使用描述符 d_ψ对评论家进行条件化，从而通过确定性策略梯度执行 m 描述符条件梯度步：

$$\nabla_\psi J(\psi) = \frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a \mid d_\psi)|_{a=\pi_\psi(s)} \nabla_\psi \pi_\psi(s) \quad (7)$$

目标是在保持解与父代处于同一单元的同时，提升该解的适应度。

算法3 描述符条件策略梯度变异

函数 变异_pg($\pi_\psi \ldots$,Q_{θ_1}, B)
  对于 $\pi_\psi \ldots$ **执行**
    $d_\psi \leftarrow D(\pi_\psi)$
    对于 $i = 1 \to m$ **执行**
      从 B 中采样 N 个状态转移（s, a, $r(s,\ a)\ s\prime$, d, $d\prime$）
      使用确定性策略梯度更新演员：
      $\frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a \mid d_\psi)|_{a=\pi_\psi(s)}\ \nabla_\psi \pi_\psi(s)$
  返回 $\pi_{\widehat{\phi}} \ldots$

5 实验

5.1 评估任务

我们在四个 QDGym 任务上评估 DCG-MAP-Elites [28]，这些任务在 Brax 中实现 [14]，源自机器人运动的标准深度强化学习基准。我们将这些任务称为“Ant-Omni”“AntTrap-Omni”“Hexapod-Omni”和“Walker-Uni”。评估涵盖三个全向任务和一个单向任务。Ant-Omni、AntTrap-Omni 和 Hexapod-Omni 是全向任务，目标是让机器人到达平面上的所有可能位置，同时尽量降低能耗。Walker-Uni 是单向任务，目标是让机器人发现所有可能的向前行走方式，同时最大限度地权衡速度与能耗。Walker-Uni 与 PGA-MAP-Elites 论文中使用的任务完全相同 [28],。Flageat 和 Lim 等人使用过 Ant-Omni 和 Hexapod-Omni [11]，而 AntTrap-Omni 则改编自 QD-PG 论文 [31]。两者的区别在于奖励函数中移除了前进奖励项。AntTrap-Omni 的目标是在欺骗性环境中到达平面上的每个位置，

同时尽量降低能耗。这个新任务旨在展示我们的算法在适应度景观的描述符空间中因陷阱而出现不连续性的环境中表现良好。陷阱会造成不连续性，而描述符条件评论家必须学会应对这种不连续性。实际上，陷阱两侧的点在描述符空间中彼此接近，但要通过轨迹实现这些描述符却相距甚远。我们在表 1 中详细介绍这些任务。

PGA-MAP-Elites 此前已在其中一些任务上取得最先进的结果，尤其是在 Walker-Uni 任务上表现突出，但在全向任务中往往表现不佳。在这些任务中，适应度函数的全局最优解是阻止机器人移动的解，这与发现如何到达不同位置的目标直接相悖。因此，PG变异生成的大多数后代都会倾向于减少移动，并缩短行进距离。相比之下，DCG-MAP-Elites 旨在降低能耗，同时保持到达远距离位置的能力。

表1：评估任务

	Ant	AntTrap	Hexapod	Walker
				
状态	身体和关节的位置和速度			
动作	每个关节的扭矩			
Type	Omni	Omni	Omni	Uni
描述符	最终（x，y）位置			足部接触
适应度	存活奖励 + 能耗惩罚			
状态维度	29	29	48	19
动作维度	8	8	18	6
回合长度	250	250	250	1000
参数	21,384	21,384	25,106	19,846

5.2 基线方法

我们将 DCG-MAP-Elites 与四种算法进行比较：CVT-MAP-Elites [39], MAP-Elites-ES [5],PGA-MAP-Elites [28] 和 QD-PG [31]。由于 PBT-MAP-Elites [30] 的实现尚未提供，我们无法将其纳入比较。我们使用 QDax 的实现，并对每项实验使用不同的随机种子重复运行 20 次，评估次数超过一百万次。DCG-MAP-Elites 的源代码可在 github.com/adaptive-intelligent-robotics/DCG-MAP-Elites 获取，其中包含一个可用于复现实验的容器。

5.3 评估指标

我们使用三个主要指标评估档案库，并使用另外两个指标评估描述符条件化演员。对于所有指标，我们基于采用 Bonferroni校正的 Wilcoxon秩和检验计算 p 值。

- **QD分数：** 档案库中存储的所有解的适应度之和。在所考虑的任务中，适应度始终为正，因此不会因算法找到更多解而对其进行惩罚。该分数涵盖了种群的质量和多样性。
- **覆盖率：** 档案库中已填充单元格所占的比例，用于表示描述符空间照亮的成效。
- **最大适应度：** 档案库中最佳解的适应度。

我们还考虑另外两个指标来评估描述符条件策略。这些指标旨在量化档案库与描述符条件策略在 QD分数和描述符空间覆盖率方面的相似程度。

- **描述符条件 QD 分数：** 描述符条件策略在档案库中每个解的描述符上进行评估后所取得的适应度总和。

$$\sum_{\pi_\psi \in \mathcal{X}} F(\pi_\phi(\,.\mid D(\pi_\psi)))$$

- **描述符误差均值：** (i) 对于档案库，所有解的存储的描述符与重新评估的描述符之间的平均差异；(ii) 对于描述符条件策略，所有解的存储的描述符与该策略以存储的描述符为条件时所得到的重新评估的描述符之间的平均差异。(|X| 档案库中的解数量)。

$$\frac{1}{|\mathcal{X}|}\sum_{\pi_\psi \in \mathcal{X}} ||D(\pi_\phi(\,.\mid D(\pi_\psi))) - D(\pi_\psi)||_{\mathcal{D}}$$

5.4 结果

5.4.1 档案库。 图 2 中的实验结果表明，与所有其他基线方法相比，DCG-MAP-Elites 在两个蚂蚁任务上的档案库表现最佳，QD分数（$p < 10^{-6}$）和覆盖率（$p < 10^{-4}$）均为最优。在另外两个任务上，我们的算法的 QD分数与此前最先进的 PGA-MAP-Elites 相近，同时覆盖率显著优于所有基线方法（$p < 10^{-3}$）。覆盖率指标表明，在所有任务中，DCG-MAP-Elites 的探索能力都超过了 QD-PG（$p < 10^{-4}$）。我们还表明，即使在 AntTrap-Omni 这类欺骗性环境中，我们的方法仍能受益于遗传算法算子的探索能力。

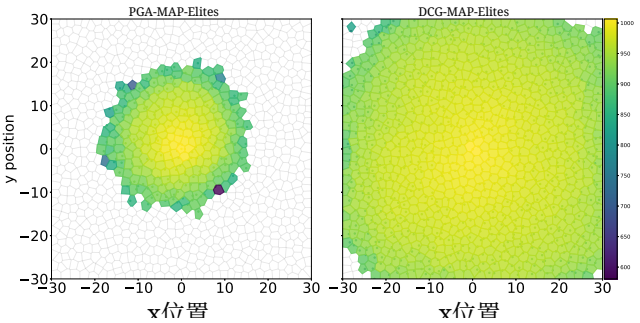

图 3：经过一百万次评估后，PGA-MAP-Elites 和 DCG-MAP-Elites 在 Ant-Omni 上的最终档案库。

实验结果证实，DCG-MAP-Elites 能够克服 PGA-MAP-Elites 在全向任务上的局限，同时在单向任务（Walker-Uni）上的表现仍不相上下；在该任务上，原本就未预期会有显著提升。这进一步证明，采用描述符条件梯度有助于让策略梯度变异算子适用于更广泛的

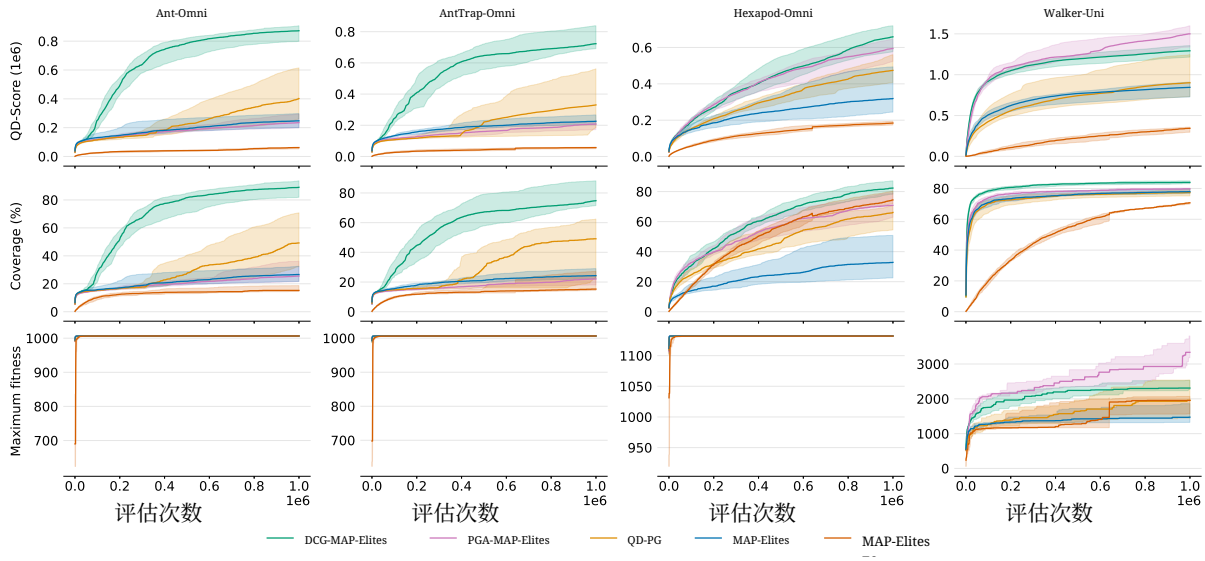

图2：DCG-MAP-Elites 与基线方法在所有任务上的 QD分数、覆盖率和最大适应度。每项实验使用不同的随机种子重复20次。实线表示中位数，阴影区域表示第一和第三四分位数。

任务。总体而言，DCG-MAP-Elites 在所有指标和任务上均表现最佳（或具有竞争力），由此证明它是 QD-RL 文献中首个成功兼顾单向和全向任务的研究成果。此前的研究通常只针对其中一种任务进行调整 [28, 31, 38]。不过，我们的结果也表明仍有改进空间。具体而言，DCG-MAP-Elites 在 Hexapod-Omni 上未能像在其他全向任务上那样，以同等幅度超越基线方法。我们推测，Hexapod-Omni 的环境动力学和高维搜索空间使其挑战性大幅增加。

表2：经过一百万次评估后，档案库和描述符条件策略在20次重复实验中的 QD分数和描述符误差均值（DEM）的中位数。

指标	Type	Ant	AntTrap	Hexapod	Walker
QD分数 (10^5)	档案库	8.7	7.2	6.5	9.8
	策略	8.3	7.0	0.96	5.2
DEM	档案库	4.04	3.44	0.078	0.12
	策略	6.25	5.8	1.39	0.27

5.4.2 描述符条件策略。表 2 给出了档案库和描述符条件策略的最终 QD 分数与描述符误差均值（DEM）；定义请参见第 5.3 节。首先，在 Ant-Omni 和 AntTrap-Omni 上，描述符条件策略取得了与档案库相同的 QD 分数，该分数通过描述符条件 QD 分数衡量（降幅不到 3%）。这表明，尽管单条件策略将信息压缩到了单个网络中，仍能够恢复档案库的质量。其次，值得结合这样一个事实来看待这一观察结果：与此同时，条件策略得到的描述符误差均值也在合理范围内。我们观察到 DEM 为 6.25，不到环境典型尺寸的 8%（即所考虑正方形的对角线）。

读者可参见图 3，直观了解这一数值所代表的含义。此外，档案库的 DEM 为 4.04（占环境尺寸的 4.7%），这表明评估中存在不可避免的随机性，而这种随机性与所用的模拟器 [14] 有关。综合这两项观察结果（QD 分数和 DEM），可以看出，我们的档案库与描述符条件策略在 Ant-Omni 和 AntTrap-Omni 上具有相似的性质，因此，这种多用途策略可以作为档案库的实用概括。第三，我们可以观察到，在 Hexapod-Omni 和 Walker-Uni 任务中，描述符条件策略难以恢复档案库的 QD 分数（两种情况下均下降了 40% 以上）。在 Hexapod-Omni 上，描述符条件策略的 DEM 明显高于档案库的典型误差（1.39 对比 0.078）。这一任务比 Ant-Omni 更具挑战性，因为其维度更多（见表 1）；我们推测所用的优化算法（TD3）可能是限制因素。对于 Walker-Uni，档案库的 DEM（0.12）表明描述符本身也存在固有的变化性，因此，尽管单条件策略得到的 DEM 为 0.27，可能会降低其实用性，但这并不令人那么意外。

总的来说，这些结果表明，我们的单一描述符条件策略已经可以视为对档案库的有前景的概括，在一半的任务上展现出非常相似的特性。其他任务则表明，我们的方法仍有改进空间。

5.4.3 消融实验。 我们进行了额外的消融实验，以展示 DCG-MAP-Elites 三个组成部分的重要性，即：在训练评论家时使用负样本、主动利用演员收集经验，以及根据描述符对演员进行条件化。在**消融实验（1）**中，我们移除了环境中的动作器评估。这样便可以研究被动学习以及缺少负样本的影响，详见第 4.3 节的

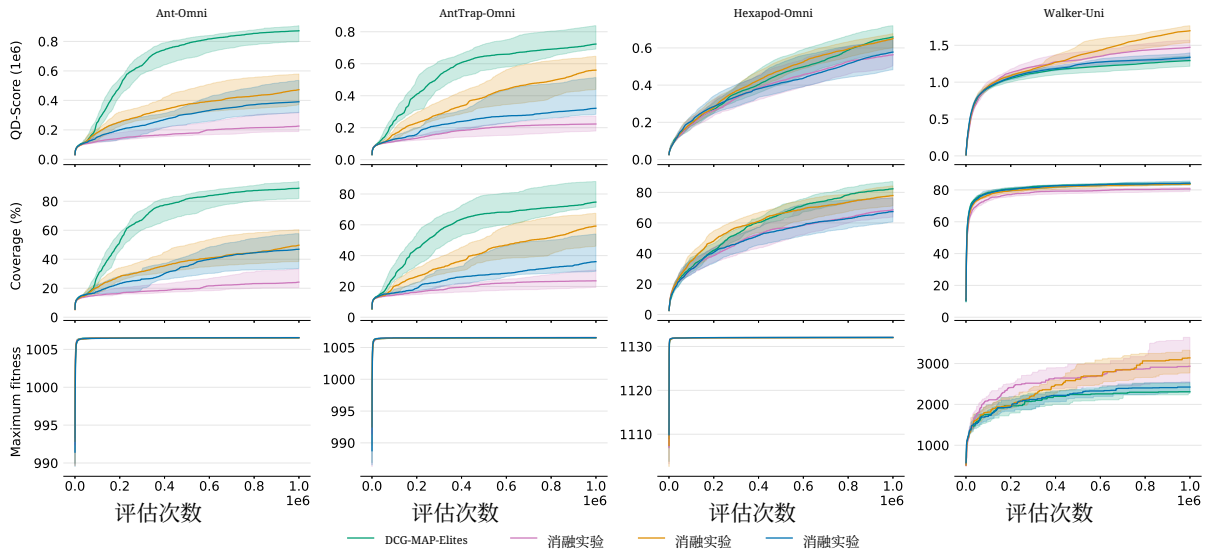

图4：DCG-MAP-Elites 及其消融实验在所有任务上的 QD分数、覆盖率和最大适应度。每项实验使用不同的随机种子重复进行20次。实线表示中位数，阴影区域表示第一和第三四分位数。消融实验（1）是移除了动作器评估和负样本的 DCG-MAP-Elites；消融实验（2）是移除了动作器评估但保留负样本的 DCG-MAP-Elites；消融实验（3）是移除了描述符条件化演员的 DCG-MAP-Elites。

详细说明。缺少主动学习会使学习过程无法获取学习描述符条件化动作价值函数所必需的负样本。我们观察到，在所有全向任务（$p < 0.01$）中，消融实验（1）的表现均逊于 DCG-MAP-Elites。但在 Walker-Uni 上，其表现相近；这可以解释为，非条件化PG发射器已被证明能够有效解决此任务。首次消融实验的结果凸显了主动学习和用于训练评论家的负例的重要性。为补充这一结果，在**消融实验（2）**中，我们移除了动作器评估，但通过在描述符空间中采样来人工生成负样本。结果表明，在蚂蚁任务（$p < 10^{-3}$）中，消融实验（2）的表现显著逊于 DCG-MAP-Elites，但在 Hexapod-Omni 和 Walker-Uni 上表现相近。不过，在蚂蚁任务（$p < 10^{-5}$）中，消融实验（2）的表现显著优于消融实验（1），这表明主动学习和 负样本都至关重要。最后，在**消融实验（3）**中，我们将描述符条件化演员 $\pi_\phi(\,.\,|\,d)$ 替换为普通演员 $\pi_\phi(\,\text{。}\,)$。在所有全向任务（$p < 0.05$）中，消融实验（3）的表现均显著较差，这表明描述符条件化演员对于训练评论家至关重要。消融实验（3）在 Walker-Uni 上的表现与 DCG-MAP-Elites 相当。

6 结论

在这项工作中，我们提出了 DCG-MAP-Elites，并展示了利用描述符条件化梯度来演化大型神经网络群体的优势。作为评论家训练的副产品，我们同步训练一个描述符条件策略，该策略能够实现广泛的描述符。在全向任务中，我们的方法显著优于 PGA-MAP-Elites；而在单向任务中，其表现与 PGA-MAP-Elites 相近。

我们还表明，即使在具有欺骗性的环境中，PG 变体的适应度提升能力与 GA 变体的探索能力之间的协同作用仍得以保留。在某些任务中，描述符条件策略展现出与离散档案相似的特性：它将自身能力汇总到一个神经网络中，并充当连续档案。我们认为，将档案提炼为单一策略是一种很有前景的方法，因为与离散档案相比，它能减少冗余；离散档案中的大多数解可能彼此相似，尤其是相邻单元格中的解。在实际应用中，描述符条件策略还可以免去处理包含数千个解的档案所带来的负担。

将深度强化学习方法与 MAP-Elites 相结合的优势伴随着马尔可夫决策过程设置的局限性。具体而言，我们只能进化可微解，而深度强化学习算法的基础依赖于马尔可夫性质和完全可观测性。尤其在本研究中，由于描述符依赖于完整轨迹，我们面临马尔可夫性质方面的挑战。因此，我们的方法所引入的缩放奖励取决于完整轨迹，而不仅仅取决于当前状态和动作。描述符条件策略的表现也表明，档案库中的知识仍有进一步提炼的空间。

未来的工作中，我们希望研究使用 DCG-MAP-Elites 训练的描述符条件策略的泛化能力，并尝试生成描述符不在档案库中的解，从而有效地进行描述符空间插值。在我们的方法中，评论家会尝试变异解，在保持其描述符不变的同时生成适应度更高的后代。我们认为，可以利用描述符条件评论家变异解，生成朝向不同描述符的后代，从而明确地促进多样性。

参考文献

[1] Felix Chalumeau、Raphael Boige、Bryan Lim、Valentin Macé、Maxime Allard、Arthur Flajolet、Antoine Cully 和 Thomas Pierrot。2022年。《神经进化是技能发现中强化学习的有力替代方案》。https: //doi.org/10.48550/arXiv.2210.03516 arXiv:2210.03516 [cs].[2] Felix Chalumeau、Thomas Pierrot、Valentin Macé、Arthur Flajolet、Karim Beguir、Antoine Cully 和 Nicolas Perrin-Gilbert。2022年。《评估质量-多样性神经进化算法在困难探索问题中的表现》。http: //arxiv.org/abs/2211.13742 arXiv:2211.13742 [cs]. [3] Konstantinos Chatzilygeroudis、Antoine Cully、Vassilis Vassiliades 和 Jean-Baptiste Mouret。2020年。《质量-多样性优化：随机优化的一个新分支》。http://arxiv.org/abs/2012.04322 arXiv:2012.04322 [cs, math, stat]. [4] Konstantinos Chatzilygeroudis、Vassilis Vassiliades 和 Jean-Baptiste Mouret。2018年。《用于机器人损伤恢复的无重置试错学习》。《机器人与自主系统》100（2018年2月），236–250。https://doi.org/10.1016/j. robot.2017.11.010 arXiv:1610.04213 [cs]. [5] Cédric Colas、Joost Huizinga、Vashisht Madhavan 和 Jeff Clune。2020年。《将 MAP-Elites 扩展至深度神经进化》。CoRR abs/2003.01825（2020），67–75。arXiv: 2003 01825 2003 01825 . https://arxiv.org/abs/ 。Antoine [6] Cully、Jeff Clune、Danesh Tarapore 和 Jean-Baptiste Mouret。2015年。《能够像动物一样适应的机器人》。《自然》521，7553（2015年5月），503–507。https: // doi.org/10.1038/nature14422 arXiv:1407.3501 [cs, q-bio]. [7] Antoine Cully 和 Yiannis Demiris。2017年。《质量与多样性优化：一个统一的模块化框架》。http://arxiv.org/abs/ 1708 09251 1708 09251 . arXiv: . cs . [8] [ ] Adrien Ecoffet、Joost Huizinga、Joel Lehman、Kenneth O. Stanley 和 Jeff Clune。2021年。《Go-Explore：困难探索问题的一种新方法》。http: // arxiv.org/abs/1901.10995 arXiv:1901.10995 [cs, stat]. [9] Benjamin Eysenbach、Abhishek Gupta、Julian Ibarz 和 Sergey Levine。2018年。《你所需的不只是一种多样性：无奖励函数的技能学习》。http: //arxiv.org/abs/1802.06070 arXiv:1802.06070 [cs]. [10] Manon Flageat、Felix Chalumeau 和 Antoine Cully。2022年。《不确定领域中用于神经进化的 PGA-MAP-Elites 实证分析》。http://arxiv.org/ abs/2210.13156 arXiv:2210.13156 [cs]. [11] Manon Flageat、Bryan Lim、Luca Grillotti、Maxime Allard、Simón C. S mith 和 Antoine Cully。2022年。《基准评测：用于强化学习神经进化的质量-多样性算法》。http://arxiv.org/abs/2211.02193 arXiv:2211.02193 [cs]. [12] Matthew C. Fontaine 和 Stefanos Nikolaidis。2021年。《可微质量-多样性》。https://doi.org/10.48550/arXiv.2106.03894 arXiv:2106.03894 [cs]. [13] Matthew C. Fontaine 和 Stefanos Nikolaidis。2023年。《协方差矩阵自适应 MAP 退火》。https://doi.org/10.48550/arXiv.2205.10752 arXiv:2205.10752 [cs]. [14] C. Daniel Free man、Erik Frey、Anton Raichuk、Sertan Girgin、Igor Mordatch 和 Olivier Bachem。2021年。《Brax——用于大规模刚体仿真的可微物理引擎》。http://arxiv.org/abs/2106.13281 arXiv:2106.13281 [cs]. [15] Scott Fujimoto、Herke van Hoof 和 David Meger。2018年。《解决行动者-评论家方法中的函数逼近误差》。http://arxiv.org/abs/1802.09477 arXiv:1802.09477 [cs, stat]. [16] Karol Gregor、Danilo Jimenez Rezende 和 Daan Wierstra。2016年。《变分内在控制》。http://arxiv.org/abs/1611.07507 arXiv:1611.07507 [cs]. [17] Shixiang Gu、Ethan Holly、Timothy Lillicrap 和 Sergey Levine。2016年。《通过异步离策略更新实现机器人操作的深度强化学习》。http://arxiv.org/abs/1610.00633 arXiv:1610.00633 [cs]. [18] Tuomas Haarnoja、Aurick Zhou、Pieter Abbeel 和 Sergey Levine。2018年。《软行动者-评论家：采用随机行动者的离策略最大熵深度强化学习》。http://arxiv.org/abs/1801.01290 arXiv:1801.01290 [cs, stat]. [19] Nikolaus Hansen。2016年。《CMA进化策略：教程》。http://arxiv. org/abs/1604.00772 arXiv:1604.00772 [cs, stat]. [20] Kurt Hornik、Maxwell Stinchcombe 和 Halbert White。1989年。《多层前馈网络是通用逼近器》。《神经网络》2，5（1989），359–366。https://doi.org/ 10 1016 0893 6080 89 90020 8 . / - ( ) - Max Jaderberg、Valentin [21] Dalibard、Simon Osindero、Wojciech M. Czarnecki、Jeff Donahue、Ali Razavi、Oriol Vinyals、Tim Green、Iain Dunning、Karen Simonyan、Chrisantha Fernando 和 Koray Kavukcuoglu。2017年。《神经网络的基于种群的训练》。https://doi.org/ 10 48550 1711 09846 1711 09846 . /arXiv. . arXiv: . [ ] cs . [22] Marija Jegorova、Stéphane Doncieux 和 Timothy Hospedales。2019年。《通过生成对抗策略网络构建行为库》。载于《2019年IEEE第九届发展与学习及表观遗传机器人联合国际会议（ICDL-EpiRob）》。https://doi.org/ 10 1109 44920 2019 1811 02945 . /ICDL-EpiRob . [ ] arXiv: . cs, stat . [23] Saurabh Kumar、Aviral Kumar、Sergey Levine 和 Chelsea Finn。2020年。《一种解决方案并非你所需的一切：通过结构化最大熵强化学习进行少样本外推》。http://arxiv.org/abs/2010.14484 arXiv:2010.14484 [cs]。

[24] Timothy P. Lillicrap、Jonathan J. Hunt、Alexander Pritzel、Nicolas Heess、To m Erez、Yuval Tassa、David Silver 和 Daan Wierstra。2019年。《利用深度强化学习实现连续控制》。http://arxiv.org/abs/ 1509 02971 1509 02971 。arXiv：。cs, [ ] stat。[25] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Alex Graves、Ioannis Antonoglou、Daan Wierstra 和 Martin Riedmiller。2013年。《利用深度强化学习玩 Atari》。http://arxiv.org/abs/ 1312 5602 1312 5602 。arXiv：。cs。 [ ] [26] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A. Rusu、Joel Veness、Marc G. Bellemare、Alex Graves、Martin Riedmiller、Andreas K. Fidjeland、Georg Ostrovski、Stig Petersen、Charles Beattie、Amir Sadik、Ioannis Antonoglou、Helen King、Dharshan Kumaran、Daan Wierstra、Shane Legg 和 Demis Hassabis。2015年。通过深度强化学习实现人类水平的控制。《自然》518卷，7540期（2015年2月），529–533。https://doi.org/10.1038/nature14236 期号：7540。出版商：Nature Publishing Group。[27] Jean-Baptiste Mouret 和 Jeff Clune。2015年。通过精英映射照亮搜索空间。http://arxiv.org/abs/1504.04909 arXiv：1504.04909 [cs，q-bio]。[28] Olle Nilsson 和 Antoine Cully。2021年。策略梯度辅助 MAP-Elites。载于《遗传与进化计算会议论文集》。ACM，法国里尔， 866 875 10 1145 3449639 3459304 –。https://doi.org/。/。Georg [29] Ostrovski、Pablo Samuel Castro 和 Will Dabney。2021年。深度强化学习中被动学习的难题。http://arxiv.org/abs/2110. 14020 arXiv：2110.14020 [cs]。[30] Thomas Pierrot 和 Arthur Flajolet。2023年。利用 MAP-Elites 演化多样化的强化学习智能体种群。https://openreview.net/forum?id=CBfYffLqWqb[31] Thomas Pierrot、Valentin Macé、Félix Chalumeau、Arthur Flajolet、Geoffrey Cideron、Karim Beguir、Antoine Cully、Olivier Sigaud 和 Nicolas Perrin-Gilbert。2022年。用于样本高效质量-多样性优化的多样性策略梯度。载于《遗传与进化计算会议论文集》。1075–1083。https://doi.org/10.1145/3512290.3528845 arXiv：2006.08505 [cs]。[32] Justin K. Pugh、Lisa B. Soros 和 Kenneth O. Stanley。2016年。质量多样性：进化计算的新前沿。《机器人学与人工智能前沿》3卷（ 2016 10 3389 2016 00040 ）。 [33] 2015 https://www.frontiersin.org/articles/。/frobt.。Tom Schaul、Daniel Horgan、Karol Gregor 和 David Silver。。通用价值函数近似器。载于第32届国际机器学习会议论文集（《机器学习研究论文集》，第37卷），Francis Bach 和 David Blei（编）。PMLR，法国里尔，1312–1320。https://proceedings.mlr.press/v37/schaul15.html [34] Archit Sharma、Shixiang Gu、Sergey Levine、Vikash Kumar 和 Karol Hausman。2020年。面向动力学的技能无监督发现。http://arxiv.org/abs/1907.01657 arXiv：1907.01657 [cs，stat]。[35] David Silver、Aja Huang、Chris J. Maddison、Arthur Guez、Laurent Sifre、George van den Driessche、Julian Schrittwieser、Ioannis Antonoglou、Veda Panneershel-vam、Marc Lanctot、Sander Dieleman、Dominik Grewe、John Nham、Nal Kalchbrenner、Ilya Sutskever、Timothy Lillicrap、Madeleine Leach、Koray Kavukcuoglu、Thore Graepel 和 Demis Hassabis。2016年。利用深度神经网络和树搜索掌握围棋。《自然》529卷，7587期（2016年1月），484–489。https://doi.org/10.1038/nature16961 期号：7587。出版商：Nature Publishing Group。[36] David Silver、Guy Lever、Nicolas Heess、Thomas Degris、Daan Wierstra 和 Martin Riedmiller。2014年。确定性策略梯度算法。载于第31届国际机器学习会议论文集（《机器学习研究论文集》，第32卷），Eric P. Xing 和 Tony Jebara（编）。PMLR，中国北京，387–395。https://proceedings.mlr.press/v32/silver14.html [37] Bryon Tjanaka、Matthew C. Fontaine、Aniruddha Kalkar 和 Stefanos Nikolaidis。2022年。通过扩展协方差矩阵自适应 MAP 退火训练多样化的高维控制器。https://doi.org/10.48550/arXiv.2210.02622 arXiv：2210.02622 [cs]。[38] Bryon Tjanaka、Matthew C. Fontaine、Julian Togelius 和 Stefanos Nikolaidis。2022年。近似梯度以实现强化学习中的可微质量多样性。https://doi.org/ 10 48550 2202 03666 2202 03666 。/arXiv.。arXiv：。 [ ] cs。[39] Vassilis Vassiliades、Konstantinos Chatzilygeroudis 和 Jean-Baptiste Mouret。2017年。利用质心 Voronoi 镶嵌扩展多维表型精英档案算法。http://arxiv.org/abs/1610.05729 arXiv：1610.05729 [cs]。[40] Oriol Vinyals、Igor Babuschkin、Wojciech M. Czarnecki、Michaël Mathieu、Andrew Dudzik、Junyoung Chung、David H. Choi、Richard Powell、Timo Ewalds、Petko Georgiev、Junhyuk Oh、Dan Horgan、Manuel Kroiss、Ivo Danihelka、Aja Huang、Laurent Sifre、Trevor Cai、John P. Agapiou、Max Jaderberg、Alexander S. Vezhnevets、Rémi Leblond、Tobias Pohlen、Valentin Dalibard、David Budden、Yury Sulsky、James Molloy、Tom L. Paine、Caglar Gulcehre、Ziyu Wang、Tobias Pfaff、Yuhuai Wu、Roman Ring、Dani Yogatama、Dario Wünsch、Katrina McKinney、Oliver Smith、Tom Schaul、Timothy Lillicrap、Koray Kavukcuoglu、Demis Hassabis、Chris Apps 和 David Silver。2019年。利用多智能体强化学习达到《星际争霸 II》宗师级水平。《自然》575卷，7782期（2019年11月），350–354。https://doi.org/10.1038/s41586-019-1724-z 期号：7782。出版商：Nature Publishing Group。

补充材料 — 采用描述符条件化梯度和档案蒸馏为单一策略的 MAP-Elites A 算法 A.1 MAP-Elites

算法 4MAP-Elites

输入： 批次大小 b
用 b 个随机解初始化档案库 X
$i \leftarrow 0$
当 $i < I$ **时执行**
$x_1, \ldots, x_b \leftarrow$ 选择($\mathcal{X}$)
$\widehat{x}_1, \ldots, \widehat{x}_b \leftarrow$ 变异($x_1, \ldots, x_b$)
添加(X $\widehat{x}_1, \ldots, \widehat{x}_b$)
$i \leftarrow i + b$
函数 添加($\mathcal{X}$ $\widehat{x}\ldots$):
对于 $\widehat{x}\ldots$, **执行**
$f \leftarrow F(\widehat{x})$、 $d \leftarrow D(\widehat{x})$
如果 X (d) $= \emptyset$ 或 F(X (d)) $< f$ **则**
X (d) $\leftarrow \widehat{x}$

A.2 PGA-MAP-Elites

算法 5PGA-MAP-Elites

输入： 批次大小 b, GA变体数量 $g \leq b$
用 b 随机解初始化档案库 X 和经验回放缓冲区 B
初始化评论家网络Q_{θ_1}、 Q_{θ_2} 以及演员网络 π_ϕ
$i \leftarrow 0$
当 $i < I$ **时执行**
训练_演员_评论家(Q_{θ_1}, Q_{θ_2}, π_ϕ, B)
$\pi_{\psi_1}, \ldots\ldots, \pi_{\psi_{b-1}} \leftarrow$ 选择(X)
$\pi_{\widehat{\psi}_1}, \ldots\ldots, \pi_{\widehat{\psi}_g} \leftarrow$ 变异_遗传算法($\pi_{\psi_1}, \ldots\ldots, \pi_{\psi_g}$)
$\pi_{\widehat{\psi}_{g+1}}, \ldots\ldots, \pi_{\widehat{\psi}_{b-1}} \leftarrow$ 变异_PG($\pi_{\psi_{g+1}}, \ldots\ldots, \pi_{\psi_{b-1}}, \mathcal{B}$)
添加($\pi_{\widehat{\psi}_1}, \ldots, \pi_{\widehat{\psi}_{b-1}}, \pi_\phi, \mathcal{X}, \mathcal{B}$)
$i \leftarrow i + b$
函数 添加($\mathcal{X}, \mathcal{B}$ $\pi_{\widehat{\psi}}\ldots$)
对于 $\pi_{\widehat{\psi}}\ldots$ **执行**
$(f$, 状态转移$) \leftarrow F(\pi_{\widehat{\psi}})$, $d \leftarrow D(\pi_{\widehat{\psi}})$
插入(B, 状态转移)
如果 X (d) $= \emptyset$ 或 F(X (d)) $< f$ **则**
X (d) $\leftarrow \pi_{\widehat{\psi}}$

算法 6演员-评论家训练

函数 训练_演员_评论家($Q_{\theta_1}, Q_{\theta_2}, \pi_\phi$, B)
对于 $t = 1 \rightarrow n$ **执行**
从 B 中采样 N状态转移（s, a, $(s$, $a)$ s'）
样本平滑噪声 ϵ
$y \leftarrow r$（s, a）$+\gamma$分钟 $_{i=1,2}\, Q_{\theta_i'}\left(s', \pi_{\phi'}(s') + \epsilon\right)$
通过回归更新两个评论家网络，使其达到 y
如果 t 取模 Δ **则**
使用确定性策略梯度更新演员：
$\frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a)|_{a=\pi_\phi(s)} \nabla_\phi \pi_\phi(s)$
软更新目标网络 $Q_{\theta i'}$ 和 $\pi_{\phi'}$

算法 7 PG变异

函数 变异_pg($\pi_\psi \ldots$, Q_{θ_1}, B)
对于 $\pi_\psi \ldots$ **执行**
对于 $i = 1 \rightarrow m$ **执行**
从 B 中采样 N个状态转移（s, a, $r(s$, $a)$ s'）
使用确定性策略梯度更新演员：
$\frac{1}{N} \sum \nabla_a Q_{\theta_1}(s, a)|_{a=\pi_\psi}(s) \nabla_\psi \pi_\psi(s)$
返回 $\pi_{\widehat{\psi}}\ldots$

B 超参数

表3列出了 DCG-MAP-Elites 和各基线方法在所有任务中使用的超参数。

表 3：DCG-MAP-Elites 及所有基线方法的超参数

参数	MAP-Elites	MAP-Elites ES	PGA-MAP-Elites	QD-PG	DCG-MAP-Elites
质心数量	1024	1024	1024	1024	1024
评估批量大小 b	256	1050	256	256	256
策略网络	[128, 128, |A|]	[128, 128, |A|]	[128, 128, |A|]	[128, 128, |A|]	[128, 128, |A|]
GA变体数量 g	256	0	128	86	128
遗传算法变体参数 1 σ_1	0.005		0.005	0.005	0.005
遗传算法变体参数 2 σ_2	0.05		0.05	0.05	0.05
演员网络			[256, 256, |A|]	[256, 256, |A|]	[256, 256, |A|]
评论家网络			[256, 256, 1]	[256, 256, 1]	[256, 256, 1]
TD3批量大小 N			100	100	100
质量评论家训练步数 n			3000	3000	3000
多样性评论家训练步数 n				300	
策略梯度训练步数 m			150	150	150
策略学习率			5×10^{-3}	5×10^{-3}	5×10^{-3}
演员学习率			3×10^{-4}	3×10^{-4}	3×10^{-4}
评论家学习率			3×10^{-4}	3×10^{-4}	3×10^{-4}
经验回放缓冲区大小			10^6	10^6	10^6
折扣因子 γ			0.99	0.99	0.99
演员延迟 Δ			2	2	2
目标更新率			0.005	0.005	0.005
平滑噪声方差 σ			0.2	0.2	0.2
平滑噪声截断值			0.5	0.5	0.5
样本数量		1000			
样本标准差		0.02			
长度尺度 l					0.008
描述符标准差 σ_d					0.0004