arXiv:2601.03413v1 [cs.LG] 6 Jan 2026

Sensor toPixels：通过基于图像的强化学习实现去中心化群体聚集

Yigal Koifman †1[0009−0004−2874−6928]、Eran Iceland †1[0000−0002−2126−3526]、Erez Koifman1[0009−0001−4606−3221]、Ariel Barel1[0000−0003−3275−4264]和AlfredM. Bruckstein1[0000−0001−5669−0037]

[1] 以色列理工学院，海法，以色列 igal.k@cs.technion.ac.il[2] 以色列理工学院，海法，以色列 eran.iceland@gmail.com[3] 以色列理工学院，海法，以色列 erez.koifman@cs.technion.ac.il[4] 以色列理工学院，海法，以色列 arielba@technion.ac.il[5] 以色列理工学院，海法，以色列 freddy@cs.technion.ac.il

†这些作者对本研究的贡献相同

摘要。 本研究展现了基于图像的强化学习方法在解决群体相关任务方面的潜力。在多智能体强化学习中，有效的策略学习取决于智能体如何感知、解读和处理输入。传统方法通常依赖手工设计的特征提取，或基于原始向量的表示，这会限制学习到的策略在输入顺序和规模方面的可扩展性与效率。在本研究中，我们提出了一种用于多智能体系统去中心化控制的基于图像的强化学习方法：将观测编码为结构化视觉输入，以便由神经网络处理，从中提取空间特征，并生成新颖的去中心化运动控制规则。我们在一项多智能体收敛任务中评估了该方法。该任务中的智能体具有有限感知范围和仅方位感知能力，目标是在聚合过程中保持群体的凝聚力。我们将算法的性能与两个基准进行比较：Bellaiche 和 Bruckstein 提出的解析解，该方法能够确保收敛，但进展缓慢；以及基于神经网络的框架 $VariAntNet$，该框架收敛速度快得多，但在困难编队中的成功率中等。我们的方法实现了较高的收敛率，收敛速度几乎与 $VariAntNet$ 相当。在某些场景中，它是唯一切实可行的替代方案。

关键词： 多智能体强化学习 · 多智能体系统 · 去中心化控制

1 引言

群体机器人学因具备在动态环境中解决复杂任务的潜力而受到广泛关注。群体机器人学的协同通常

采用集中式或去中心化控制范式。集中式方法依赖掌握完整系统信息的全局控制器，但往往受限于可扩展性，且容易受到单点故障的影响。相比之下，群体机器人学控制可采用去中心化方法，其特点是局部决策、自组织，以及能够抵御单个智能体故障。

群体机器人学的一个分支——蚁群机器人学，研究在有限感知、计算资源极少且通信能力有限或完全缺失的条件下运行的智能体。这使得搜索与救援、环境监测等真实世界应用能够以可扩展且经济高效的方式实现。与此同时，由于智能体仅依据本地感官输入采取行动，这也给协同与集体行为带来了重大挑战。

连通性与收敛在群体任务中发挥着关键作用。连通性确保智能体在可见性图中保持连接，从而实现合作与协同。收敛则使所有智能体聚集到一个小范围内，以支持会合或资源共享等任务，并体现出局部规则如何产生连贯的全局行为。

本研究的主要目标是探索一种基于多智能体强化学习（MARL）的算法，以生成去中心化控制策略，使智能体能够实现全局目标。我们通过一项基础的群体收敛任务对该方法进行评估，同时避免群体碎片化。在多智能体强化学习中，有效的策略学习取决于智能体如何处理和解读感官输入。传统方法通常依赖手工设计的特征提取，或原始的基于向量的表示；当观测的规模和顺序发生变化时，这可能会限制所学策略的可扩展性。

所提出的框架采用基于图像的感知表示，实现集中训练、分散执行（CTDE）：将智能体的观测编码为结构化视觉输入，并由卷积神经网络（CNN）处理，从而有效提取空间特征，促进控制规则的形成，如图 1 所示。

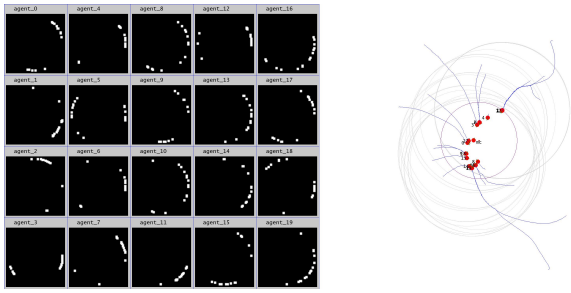


图 1。 20 个智能体的收敛过程。左：典型智能体的局部观测，其中白色像素表示在有限感知范围内检测到的邻居。右：群体收敛轨迹的全局视图。

Bellaiche 和 Bruckstein [4], 研究了有限可见范围和仅方位感知条件下聚集问题的解析解，也就是说，智能体只能感知邻居的相对方向。他们提出了一种局部运动规则，以保证智能体在有限时间内完成聚集。

我们将强化学习（RL）策略的性能与解析解以及基于神经网络的框架 VariAntNet [10], 进行比较。结果表明，即使在极具挑战性的编队中，强化学习策略也能保持群集连通性；与解析方法相比，它的收敛速度显著更快，但在整个过程中无法保证连通性。我们的主要贡献如下：

1. **基于图像的方位表示**：将仅包含方位信息的智能体观测编码为图像，从而能够使用卷积神经网络。这种方法在去中心化群体协调领域似乎尚属首创。2. **面向连通性的强化学习控制**：一种能够大力促进群集连通性、并相较于解析解加快收敛速度的强化学习策略，在复杂编队下优于以往方法。

2 相关工作

多智能体强化学习已成为分布式决策研究的基石，使智能体能够在共享环境中学习协作或竞争行为。Busoniu 等人 [6] 和 Zhang 等人 [19]的奠基性研究为多智能体强化学习奠定了理论基础，凸显了其与博弈论、控制和分布式优化的联系。近期的综述 [7], [1] 进一步拓展了这些基础，着重探讨了多智能体场景中的非平稳性、信用分配和可扩展性等挑战。

多智能体强化学习（MARL）中广泛采用的一种方法是集中训练、分散执行（CTDE），该方法允许智能体在训练期间利用全局信息，并在执行时独立行动。这种方法能够解决非平稳环境导致的不稳定问题，并由 Lowe 等人在多智能体深度确定性策略梯度（MADDPG）中通过 [12] 进行了形式化，随后又在 Amato 对 CTDE 的理论性论述（如 [2]）中得到进一步探讨。事实证明，这种方法在合作任务中尤其有效，因为智能体在部署期间必须在没有直接通信的情况下进行协调。

在群体机器人学领域，多智能体强化学习已应用于队形控制、会合和追逃等任务。Hüttenrauch 等人 [8] 提出了一种均值嵌入表示，用于可扩展的群体学习，以应对如何表示规模可变的智能体群组这一挑战。这些方法通常假设完全可观测性或通信，而在仅方位模型中，智能体只能感知邻居的相对方向，因此有限范围可见性构成了更受限的感知范式。近期研究开始探讨有限感知条件下的多智能体强化学习，例如仅方位观测。Li 等人 [11] 提出了一种用于目标追踪的合作式多智能体强化学习框架，仅使用方位测量来应对在有限感知条件下进行状态估计和控制所面临的挑战。同样，Kang 等人 [9] 提出了

多智能体掩码自编码器（MA²E），使智能体无需显式通信便能根据部分观测推断全局信息，从而增强部分可观测环境中的去中心化协同。

Bellaiche 和 Bruckstein [4] 提出了一种解析方法，用于解决受有限可见范围和仅方位感知约束的聚集问题。该方法要求每个智能体识别其可见范围内能够涵盖所有邻居的最小圆形扇区。如果该扇区的角度小于 π，智能体便将速度向量设为指向最外侧两个邻居的单位向量之和，并据此移动。否则，智能体被视为受到包围，并保持静止。已有证明表明，无论初始配置和群体规模如何，该方法都能保证收敛。

Koifman等人 [10] 提出了 VariAntNet，这是一种基于神经网络的方法，旨在应对同一挑战。该方法采用监督式集中训练、分散执行的学习方法，利用几何特征和基于可视图拉普拉斯算子的损失来训练无状态神经网络。

VariAntNet 通过集中式访问对控制策略进行时间优化，并训练其模仿预定义行为；而我们基于多智能体强化学习的方法则使智能体能够通过局部交互学习去中心化策略，并优化长期累积奖励。这种方法支持更丰富的时间行为和更高的灵活性，同时保留去中心化执行。感知能力受限迫使解析方法采用谨慎的规则，因此其收敛速度远慢于统计训练得到的解。

3 问题设定

本工作的主要目标是研究一种用于表示多智能体问题的多智能体强化学习框架。作为测试案例，我们考察一项基础的多智能体聚集任务，其中智能体必须聚集到一个小区域内。智能体被设定为仅具备有限范围的方位感知能力，因此只能探测到位于其可见范围内的邻居的方向。一项关键约束是，群体必须保持凝聚，防止碎片化；这可通过遵循“永不失去邻居”的保守原则来实现， [3],这意味着智能体应避免失去任何已感知到的邻居。

3.1 符号与定义

智能体在时间 t 的位置表示为 $P(t) \triangleq \{p_i(t)\}_{i=1,2,3,...,N}$，其中每个智能体的位置表示为 $p_i(t) \triangleq \{(x_i, y_i)^T\}_{i=1,2,3,...,N}$。智能体的可见范围表示为 V， p_i 与 p_j 之间的距离定义为欧几里得距离： $d(p_i, p_j) = \| p_j - p_i \| \triangleq \left((p_j - p_i)^T (p_j - p_i)\right)^2$。智能体 $i's$ 的邻居集合 $\mathcal{N}_i$ 定义为 $\mathcal{N}_i = \{p_j \in P, d(p_i, p_j) \leq V\}$。在仅方位感知中，智能体 i 在时间 t 对智能体 j 的观测定义为单位向量 $\hat{u}_{ij}$，其方向从 $p_i(t)$ 指向 $p_j(t)$: $\hat{u}_{ij}(t) = \frac{p_j(t) - p_i(t)}{\|p_j(t) - p_i(t)\|}$

4 方法论

从传感器到像素框架在多智能体强化学习环境中采用集中训练、分散执行范式，并使用演员-评论家架构。在训练期间，利用集中式数据获取全局环境信息，并根据群体的动作计算基于全局信息的奖励。在执行期间，每个智能体以分散方式独立运行，仅依赖其部分局部观测。

问题设定可以建模为去中心化部分可观测马尔可夫决策过程（DEC-POMDP）[5]，该模型用于描述不确定环境下的多智能体决策：智能体必须协作以实现共同目标，同时只能部分观测环境：

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{N}, \{\mathcal{O}_i\}, \{\mathcal{A}_i\}, \mathcal{P}(s' \mid s, \mathbf{a}), \{\mathcal{R}_i\}, \gamma \rangle, \tag{1}$$

其中，一组 $\mathcal{N}$ 智能体在共享环境中运行，只能获取实际全局状态 $\mathcal{S}_t$的部分信息，因此每个智能体 i 只能观测到环境的有限局部视图，定义为 $\mathcal{O}_i$（t）$= f$（$\mathcal{S}_t$）。因此，$\mathcal{S}$ 表示全局状态，$\mathcal{N} = \{1, 2, \cdots, N\}$ 是智能体集合，$\mathcal{O}_i$ 是智能体 i的观测，$\mathcal{A}_i$是智能体 i 的动作空间，$\mathcal{P}$（$s' \mid s$，$\mathbf{a}$）表示在给定动作 a 的情况下，环境从 s 转移到 s' 的概率，$\mathcal{R}_i$是智能体 i的奖励函数，γ 是折扣因子。

目标是在训练期间优化策略函数 π，以最大化累积折扣奖励的期望值：$J(\theta) = \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R_t\right]$，其中 θ 表示策略网络的参数。

在每个时间步，每个智能体感知其局部环境，获得部分观测。该观测经过预处理并转换为基于像素的表示。随后，所得数据通过基于卷积神经网络的特征提取器，在用于策略学习之前对相关信息进行编码，如图 2 所示。策略正在更新过程中

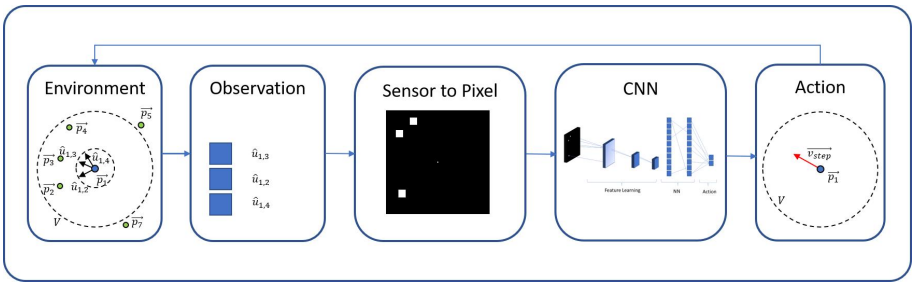


图 2. 从传感器到像素框架。

采用由局部奖励和全局奖励组成的奖励函数进行训练

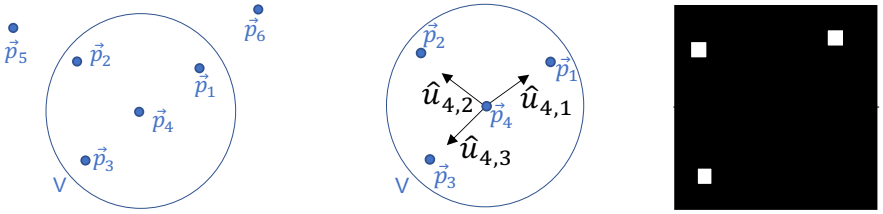
函数。局部奖励函数评估智能体的个体动作和状态，而全局奖励函数则促进整个群体的行为，鼓励协同行动。这种设计使智能体能够学习有效的个体行为，同时推动整个群体实现聚集到一个小区域的目标。

4.1 观测空间

每个智能体都使用其仅测方位的有限量程传感器观测邻近智能体。因此，智能体 i 在时间 t 的观测记为 $\mathcal{O}_i(t)$，定义为一个指向其可见邻居的单位向量多重集：

$$O_i(t) = \{\hat{u}_{ij}(t) : \forall p_j \in \mathcal{N}_i(t)\} \tag{2}$$

该观测被投影到一个大小为 75×75 像素的二值矩阵上，其中智能体 i 位于中心像素处，观测到的每个智能体均以 3×3 矩阵块表示，并根据其相对于智能体局部参考系的方向进行定位，如图3所示。



(a)环境视图围绕 p_4，其中 p_4 感知到 p_1，p_2，p_3的方向。

(b) p_4观测为$\hat{u}_{4,2}$，$\hat{u}_{4,1}$，$\hat{u}_{4,3}$。

(c)所有$\hat{u}_{ij}$ 均为投影到一个二值矩阵。

图3. 从局部观测到像素网格表示的预处理。

4.2 动作空间

在每一步中，每个智能体从连续动作空间中选择一个动作，$\mathcal{A}$。此动作由方向角 α 和步长 σ定义：

$$\mathcal{A} = \{(\alpha, \sigma) \mid \alpha, \sigma \in \mathbb{R},\ -\pi < \alpha \leq \pi,\ 0 \leq \sigma \leq 1\} \tag{3}$$

智能体的策略 $\pi(a \mid o)$ 将局部观测映射为动作 $a = (\alpha,\ \sigma)$，从而实现精细方向控制和自适应运动规划。角度 α 决定航向，而标量 σ 则调节运动幅度。这种连续控制设置能够实现平稳、精确的协调，并有助于在去中心化群体行为中维持凝聚力、加快收敛速度。

4.3 网络架构

特征提取 框架的下一阶段是特征提取，由卷积神经网络完成。该网络处理投影观测，学习提取空间模式并识别嵌入特征。使用卷积神经网络的一项关键优势在于，无论观测的顺序、数量或方向如何，它都能捕捉观测之间的空间和几何关系。

该架构由多个卷积层组成，其设计灵感源自 Mnih 等人的开创性工作。他们 [14], 证明，基于卷积神经网络的特征提取能够帮助智能体以人类水平的表现学会玩雅达利游戏。这些层及其架构定义见表1。

表1。卷积神经网络流水线。

层	输入通道	输出通道	卷积核大小	步幅	激活
第1层	1	32	8,8	4	ReLU
第2层	32	64	4,4	2	ReLU
第3层	64	64	3,3	1	ReLU

最后一个卷积层的输出首先被展平为一个包含 1600 个神经元的向量，然后由多层感知机（MLP）处理，将其维度降至 512 维特征表示。

演员-评论家网络尽管演员网络和评论家网络共用一个特征提取模块，但后续架构彼此独立。演员网络以确定性方式输出当前状态下应采取的具体动作，直接将学习策略映射为动作向量。评论家网络则输出一个标量值，用于估计当前（状态、动作）对所对应的预期价值函数，并据此评估和改进策略，如图4所示。

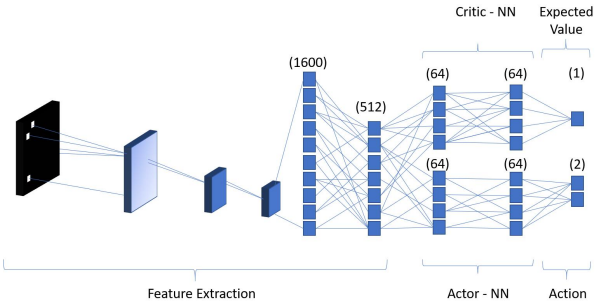


图 4. 演员-评论家神经网络架构。

4.4 奖励函数

目标是使群体汇聚到一片紧凑区域，同时保持凝聚力并防止碎片化。这种集体行为源于学习策略，而学习策略由奖励函数塑造。智能体的奖励函数包含两个组成部分：

– **局部奖励函数**： R^i_{local}旨在促进群体凝聚，类似于 [13] 中所述的“永不失去邻居”原则：从连通状态开始的智能体应与其已累积的邻居集合保持可见。因此，对于每个丢失的单个邻居，无论丢失发生在何处，都会对局部奖励施加邻居丢失惩罚 P_{ln}。 $R^i_{ln}(t)$ 表示在时间步 t 因邻居丢失而累积的总惩罚。

$$R^i_{ln}(t) = (|\mathcal{N}_i(t)| - |\mathcal{N}_i(t+1)|) \cdot P_{ln}, \quad for \ \ |\mathcal{N}_i(t+1)| < |\mathcal{N}_i(t)|$$

为加快群体收敛并减少回合步数，每个智能体执行的每一步都会施加额外惩罚 P_{acc}。因此，局部奖励表示为 $R^i_{local}(t) = R^i_{ln}(t) + P_{acc}$。

– **全局奖励**： R_{global}旨在通过最小化以智能体平均位置为圆心的圆的边界半径 $D_{\text{global}}(t)$，促进群体收敛。

边界半径定义为： $D_{\text{global}}(t) = \max_{i \in \mathcal{N}} ||p^i(t) - C_{\text{swarm}}(t)||$，其中平均位置表示为 $C_{\text{swarm}}(t) = \frac{1}{N} \sum_{i=1}^{N} p_i(t)$，全局奖励计算为 $R_{\text{global}}(t) = (D_{\text{global}}(t) - D_{\text{global}}(t+1)) \cdot C_g$，其中$C_g$ 为归一化常数。

智能体 i 的奖励 $\mathcal{R}^i$ 定义为 $R^i(s, a^i, \mathbf{a}) = R^i_{\text{local}}(s, a^i) + R_{\text{global}}(s, \mathbf{a})$。

5 结果

该框架通过基于仿真的方法进行评估。首先，我们采用结构化学习过程训练模型，以优化其性能。其次，我们在一系列多样化的场景中评估训练后的模型，以考察其适应性和有效性。最后，通过测量两个关键指标，将其性能与解析算法 [4] 和 VariAntNet 进行比较：成功完成聚集的场景百分比，以及收敛时间。我们使用 PettingZoo 库 [17] 进行多智能体环境仿真，并结合 Stable-Baselines3 [15]来训练和评估策略。

我们的数据集包含由随机定位的智能体组成的连通编队，这些智能体具有不同的规模、可见范围和难度等级。难度由智能体的初始空间分布决定，并通过有效可见范围量化，$V_{\text{eff}} = V \times VR$，其中 $0 < VR \leq 1$ 为可见率。生成过程中，智能体从一个随机初始位置开始依次放置。每个新智能体都会反复调整位置，直至其位于至少一个已放置智能体的 V_{eff} 范围内，从而确保连通性。增大 VR 会提高难度，因为断连风险随之增加。我们将$V_{\text{eff}} = 0.75$ 的编队定义为“有挑战性”，将 $V_{\text{eff}} = 1$ 的编队定义为“临界”。

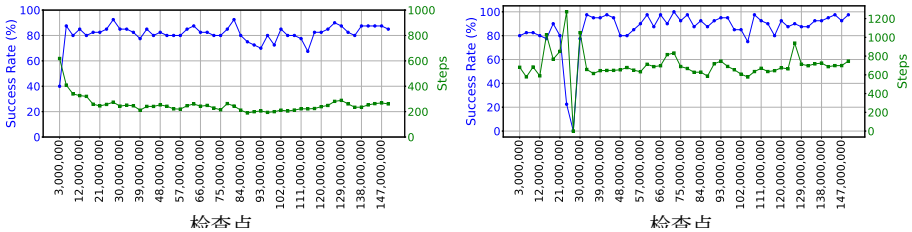
5.1 训练参数

我们采用近端策略优化（PPO）算法的多智能体变体 [16], ；该算法在单智能体强化学习任务中表现出色。近期研究 [18] 表明，PPO 可有效扩展至协作式多智能体环境，因此适用于我们的任务。

可见范围设置为 50，步长设置为 0.5。该模型的训练课程包括：先在由 10 个智能体组成的编队上进行 1.5 亿步的学习阶段，随后在由 20 个智能体组成的编队上再进行 1.5 亿步训练。每个智能体的回合截断步数分别为 1500 步和 3000 步。我们的算法采用 0.00002 的学习率、 $\gamma = 0.95$ 的折扣因子、2048 的批次大小以及 8 个并行环境进行训练。奖励函数参数为：$P_{ln} = -0.5$、 $P_{acc} = -0.01$ 和 $C_g = 0.1$。

5.2 训练评估

为评估训练过程，我们每隔 3M 步保存一次检查点。对于 10 个和 20 个智能体这两种设置，我们在每个检查点上都使用一组固定的 40 个编队进行评估。结果汇总于图5，表明策略在训练初期迅速收敛至较优解。此后，在剩余的训练过程中，性能出现了明显波动。选定的检查点兼具较高的收敛速度和较快的收敛过程。具体而言，RL10 选用 81M 的检查点模型，RL20 选用 42M 的检查点模型。



(a)RL10 训练过程评估。 (b) RL20 训练过程评估。

图5. 10 个和 20 个智能体收敛至稳定策略后的训练过程评估。从每次训练会话中选取一个检查点模型进行进一步评估。

5.3 执行结果

模型在“有挑战性”和“临界”初始构型VR = (0.75,1)上进行评估。结果与解析解及 VariAntNet 方法进行比较，汇总如表2所示。结果表明，采用基于图像的感知表示训练的智能体，在.....方面更快且更

表2。 比较 Bellaiche 的解析模型 [4], VariAntNet（最大值）和 VariAntNet（最大值、加权 1,10），以及多智能体强化学习方法 RL20、RL10 在严峻条件和边缘初始配置下的 1000 个场景中的表现。报告了平均收敛时间，以及在保持连通性的同时实现收敛的智能体百分比。高亮结果表示在可接受标准下表现最佳。

模型	严峻条件						边缘型					
	10 个智能体		20 个智能体		30 个智能体		10 个智能体		20 个智能体		30 个智能体	
	步数	连通性 (%)	步数	连通性 (%)	步数	连通性 (%)	步数	连通性 (%)	步数	连通性 (%)	步数	连通性 (%)
解析基线	271	100.0	702	100.0	1235	100.0	362	100.0	950	100.0	1629	100.0
VariAntNet(Max)	165	99.5	346	95.2	554	91.6	217	84.5	451	58.4	711	45.8
VariAntNet(Max, W)	**159**	**98.8**	**322**	**94.9**	**506**	**90.6**	209	82.6	418	53.8	646	41.3
RL20	241	100.0	506	100.0	776	99.8	**354**	**97.0**	**753**	**93.6**	**1173**	**91.6**
RL10	186	99.7	383	95.4	565	89.2	260	82.9	543	61.7	797	40.6

能可靠地收敛到“临界”初始构型。在“有挑战性”条件下，对于由20个智能体组成的群体，RL20经过506步实现了100%的连通性，在连通性保持方面优于 VariAntNet（94.9%，322步），但收敛时间更长；不过，它仍比解析基线（100%，702步）收敛得更快。我们将连通性超过90%的结果定义为可接受。在有挑战性的条件下，对于由30个智能体组成的群体，VariAntNet的性能明显下降，而解析模型的收敛速度最慢；RL20则保持了更优的性能，性能损失仅属中等。RL20是在临界条件下唯一可行的强化学习变体，表现比解析基线高出2–28%。这归因于强化学习框架能够为智能体之间的任何断连给予负奖励，并为群体收敛给予正奖励。由于强化学习不仅考虑下一步奖励，也会考虑未来奖励，因此模型能够随着时间推移学会最优策略。在有挑战性的条件下，加权 VariAntNet 的表现比 RL10 高出15%。

训练从 10 个智能体增加到 20 个智能体时，会出现一个有趣的现象：模型变得更加保守，以降低收敛速度为代价减少断连次数。这是因为随着智能体数量增加，初始构型中出现“叶节点”智能体（仅有一个邻居的智能体）的概率也会增加。断开连接这些智能体的唯一边会导致图断连。此外，桥（割边）在断开后将群体分割成多个片段，其出现概率也会增加。因此，训练阶段的智能体数量越多，模型学到的行为就越保守。

6 消融研究

这项消融研究考察了局部奖励（即邻居损失惩罚因子 $P_{\ln}$）对群体行为的影响。除基线值 $P_{\ln} = -0.5$ 外，我们还评估了另外两个取值： $P_{\ln} = 0$ 和 $P_{\ln} = -1$。每种配置下，模型均训练了 150M 步。

对于 $P_{\mathbf{ln}} = -1$，学习策略导致所有智能体都保持静止。失去邻居的惩罚高得不成比例，使智能体变得非常保守，避免任何可能导致断连的移动。

对于 $P_{\mathbf{ln}} = 0$，这里观察到了相反的行为。智能体急于收敛，却忽视了失去连接所带来的后果。在许多情况下，尽管个别边丢失，群体仍保持连通。这使收敛速度加快，但代价是整体收敛质量下降。

表3将此配置的结果与基线惩罚 $P_{\mathrm{ln}} = -0.5$ （表2中的 RL10）进行了比较。对于具有 $V_{\mathrm{eff}} = 0.75$ 的 10 个智能体编队，系统实现了合理的收敛速度，且收敛较快。然而，对于 20 个智能体编队，或在 $V_{\mathrm{eff}} = 1$ 的情况下，收敛速度显著下降。

表3. 不同邻居丢失惩罚值 P_{ln} 的消融研究结果。

	10 个智能体				**20 个智能体**			
V_{ref}	$P_{\mathrm{ln}} = 0$		$P_{\mathrm{ln}} = -0.5$		$P_{\mathrm{ln}} = 0$		$P_{\mathrm{ln}} = -0.5$	
	步数	收敛率（%）	步数	收敛率（%）	步数	收敛率（%）	步数	收敛率（%）
0.75	147	92.1	186	99.7	245	59.8	383	95.4
1.00	193	55.2	260	82.9	313	11.6	543	61.7

7 局限性

基于强化学习的方法的一个显著局限是训练时间长，且需要经过挑选过程来确定更好的学习参数，这通常需要大量的计算资源和谨慎的超参数调优。即使经过长时间训练，这些方法也无法保证收敛到稳定或最优策略，因为其性能可能在不同运行过程中波动，或无法泛化到未见过的场景。

8 结论

在这项工作中，我们提出了从传感器到像素框架，这是一种基于多智能体强化学习的方法，结合了集中训练、分散执行与基于卷积神经网络的感知，用于去中心化群体协调。结果表明，该框架显著优于已知模型，在不同群体规模下实现了更快的收敛和更强的凝聚力。这些发现凸显了基于强化学习的方法在解决群体机器人学基本挑战方面的潜力，尤其是在灾害应急响应等时间关键型场景中。展望未来，后续研究将探索更复杂的多智能体行为和环境。另一个研究方向是开发一种自适应决策框架，在推理过程中动态选择最合适的算法，在解析保证与基于学习的效率之间取得平衡。

参考文献

1. Albrecht, S., et al.：《多智能体强化学习：基础与现代方法》。剑桥大学出版社（2023），https://www.marl-book.com 2. Amato, C.：合作式多智能体强化学习中去中心化执行的集中式训练简介。arXiv预印本 arXiv:2409.03052（2024） 3. Barel, A., Manor, R., Bruckstein, A.M.：聚在一起：多智能体几何共识（聚集、会合、聚类、聚合）。arXiv预印本 arXiv:1902.01455（2019） 4. Bellaiche, L.I., Bruckstein, A.：有限可视范围与仅方位感知条件下智能体的连续时间聚集。《群体智能》**11**(3)，271–293（2017） 5. Bellman, R.：一个马尔可夫决策过程。《数学与力学杂志》页码 679–684（1957） 6. Busoniu, L., Babuska, R., De Schutter, B.：多智能体强化学习综合综述。IEEE《系统、人类与控制论汇刊C辑（应用与综述）》**38**(2)，156–172（2008） 7. Hernandez-Leal, P., Kartal, B., Taylor, M.E.：多智能体深度强化学习综述与评析。《自主智能体与多智能体系统》**33**，750–797（2019） 8. Hüttenrauch, M., Sosic, A., Neumann, G.：面向群体系统的深度强化学习。《机器学习研究杂志》**20**(54)，1–31（2019） 9. Kang, S., Lee, Y., Kim, G., Chong, S., Yun, S.Y.：MA^{2}E：利用掩码自编码器解决多智能体强化学习中的部分可观测性问题。载于：《第十三届国际学习表征会议》 10. Koifman, Y., Koifman, E., Iceland, E., Barel, A., Bruckstein, A.M.：Vari-Ant Net：学习多智能体系统的去中心化控制（2025），https://arxiv.org/abs/2509 02271 11 . . Li, J., et al.：通过多智能体强化学习进行合作式仅方位目标追踪。arXiv预印本 arXiv:2503.08740（2025） 12. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., Mordatch, I.：面向混合合作竞争环境的多智能体演员-评论家。载于：《神经信息处理系统进展》。页码 6379–6390（2017） 13. Manor, R., Barel, A., Bruckstein, A.M.：凝聚型灵活群体的局部交互。arXiv预印本 arXiv:1903.09259（2019） 14. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., et al.：通过深度强化学习实现人类水平的控制。《自然》**518**(7540)，529–533（2015） 15. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., Dormann, N.：Stable-Baselines3：可靠的强化学习实现。《机器学习研究杂志》**22**(268)，1–8（2021），http://jmlr.org/papers/v22/20-1364.html 16. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O.：近端策略优化算法。载于：arXiv预印本 arXiv:1707.06347（2017） 17. Terry, J., Black, B., Grammel, N., Jayakumar, M., Hari, A., Sullivan, R., Santos, L.S., Dieffendahl, C., Horsch, C., Perez-Vicente, R., et al.：PettingZoo：面向多智能体强化学习的Gym。《神经信息处理系统进展》**34**，15032–15043（2021） 18. Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., Wu, Y.：近端策略优化在合作式多智能体游戏中的惊人效果。arXiv预印本 arXiv:2103.01955（2021） 19. Zhang, K., Yang, Z., Basar, T.：多智能体强化学习：理论与算法选择性综述。arXiv预印本 arXiv:1911.10635（2019）