北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

论文分享

MAXIMUM ENTROPY HETEROGENEOUS-AGENT REINFORCEMENT LEARNING

Published as a conference paper at ICLR 2024

https://sites.google.com/view/meharl

ICLR2024 强化学习相关文章汇总

Dragon
强化学习

+ 关注

350 人赞同了该文章

目录 收起

最近花时间过了一遍ICLR2024中稿的文章，我筛选并整理了与强化学习相关的论文，将其归类成17个领域，其中每篇文章只会被归到一个类别内；有些文章肯定同时横跨多个领域，我选择了一个最主要的类别。如果有遗漏的强化学习文章或者其他宝贵意见，欢迎大家评论区补充交流。

值得一提的是17个类别中的最后两个类，RL Theory专注于理论推导，涉及边界推导和复杂度分析✦，通常不包含实验部分，或仅包含有限的数值模拟✦；Interdisciplinary of RL则涵盖了强化学习的应用和其他领域知识方法在强化学习中应用。这些论文在标题前都标注了关键词，以突出其核心问题、领域或方法。

接下来，我计划精读**Oral**和**Spotlight**的文章，然后按照大类逐一阅读，其中优先关注offline RL和RLHF这两个领域，并在知乎上分享我的阅读笔记和见解，欢迎大家关注催更✦。对于multi-agnet RL的深入解读解读，推荐阅读这位博主的精彩总结ICLR 2024 多智能体强化学习论文汇总 - 知乎 (zhihu.com)。

---

1. Multi-Agent RL

（**Oral**）Efficient Episodic Memory Utilization of Cooperative Multi-Agent Reinforcement Learning

（**Spotlight**）Greedy Sequential Execution: Solving Homogeneous and Heterogeneous

▲ 赞同 350 ▼ 24 条评论 分享 喜欢 收藏 申请转载 ···

会议介绍——International Conference on Learning Representations

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

人工智能三大会：NeurIPS、ICLR、ICML

中国计算机学会推荐国际学术会议

（人工智能）

一、A类

序号	会议简称	会议全称	出版社	网址
1	AAAI	AAAI Conference on Artificial Intelligence	AAAI	http://dblp.uni-trier.de/db/conf/aaai/
2	NeurIPS	Conference on Neural Information Processing Systems	MIT Press	http://dblp.uni-trier.de/db/conf/nips/
3	ACL	Annual Meeting of the Association for Computational Linguistics	ACL	http://dblp.uni-trier.de/db/conf/acl/
4	CVPR	IEEE/CVF Computer Vision and Pattern Recognition Conference	IEEE	http://dblp.uni-trier.de/db/conf/cvpr/
5	ICCV	International Conference on Computer Vision	IEEE	http://dblp.uni-trier.de/db/conf/iccv/
6	ICML	International Conference on Machine Learning	ACM	http://dblp.uni-trier.de/db/conf/icml/
7	IJCAI	International Joint Conference on Artificial Intelligence	Morgan Kaufmann	http://dblp.uni-trier.de/db/conf/ijcai/

	Publication	h5-index	h5-median
1.	Nature	414	607
2.	The New England Journal of Medicine	410	704
3.	Science	391	564
4.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	356	583
5.	The Lancet	345	600
6.	Advanced Materials	294	406
7.	Cell	288	459
8.	Nature Communications	287	389
9.	Chemical Reviews	270	434
10.	International Conference on Learning Representations	253	470
11.	JAMA	253	446
12.	Neural Information Processing Systems	245	422
13.	Proceedings of the National Academy of Sciences	245	337
14.	Journal of the American Chemical Society	245	330
15.	Angewandte Chemie	235	314
16.	Chemical Society Reviews	234	339

Single-blind review | double-blind review | **open review**

会议介绍——International Conference on Learning Representations

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

Paper Decision

Decision | Program Chairs | 16 Jan 2024, 19:53 (modified: 17 Feb 2024, 04:40) | Everyone | Revisions

Decision: Accept (spotlight)

Add: Public Comment

Public Comment

Public Comment | 03 Aug 2024, 11:28 (modified: 03 Aug 2024, 22:09) | Everyone | Revisions

[Deleted]

Meta Review of Submission5524 by Area Chair F59c

Meta Review | Area Chair F59c | 07 Dec 2023, 22:43 (modified: 17 Feb 2024, 04:29) | Everyone | Revisions

Metareview:

The paper introduces HASAC, a novel algorithm in the Maximum Entropy (MaxEnt) framework, addressing challenges in cooperative Multi-Agent Reinforcement Learning (MARL). Theoretical proofs validate HASAC's effectiveness, demonstrating monotonic improvement and convergence to quantal response equilibrium (QRE). The paper also presents MEHAML, a generalized template. Empirical evaluations on six benchmarks consistently show HASAC outperforming baselines, emphasizing enhanced sample efficiency, robustness, and exploration. The work contributes a comprehensive solution, bridging theory and practice in cooperative MARL.

The paper stands out for its excellent presentation, well-motivated methodology, and clear discussion of related work. The technical contributions, specifically the novel algorithmic framework MEHAML and the practical algorithm HASAC, are noteworthy. Unlike a simple combination of MaxEnt RL and MARL, the paper's approach is uniquely derived from the Probabilistic Graphical Models (PGM) formulation. Empirically, HASAC consistently outperforms baselines across various benchmarks, achieving the best performance in 31 out of 35 tasks. Overall, the paper contributes significantly to cooperative Multi-Agent Reinforcement Learning (MARL) with its clarity, innovation, and empirical success.

While the paper has several strengths, a few weaknesses have been identified. The identified weaknesses in the paper include concerns about limited novelty due to applying the Soft Actor-Critic (SAC) algorithm, the absence of testing in crucial sample efficiency scenarios, and doubts about the validity of experimental results due to significant fluctuations in training curves and selective result presentation. Additionally, there are calls for a clearer explanation of the effect of sequential updates in the MaxEnt MARL objective and a more robust method to tune entropy terms, given the algorithm's sensitivity to entropy temperature in the ablation study. Addressing these concerns would enhance the overall quality and credibility of the paper.

Given the consensus among reviewers, we propose accepting the paper for publication, with the understanding that the authors will address the identified weaknesses and incorporate any necessary clarifications or improvements in the final revision. We are confident that the suggested changes will further enhance the paper's quality and contribution to the field.

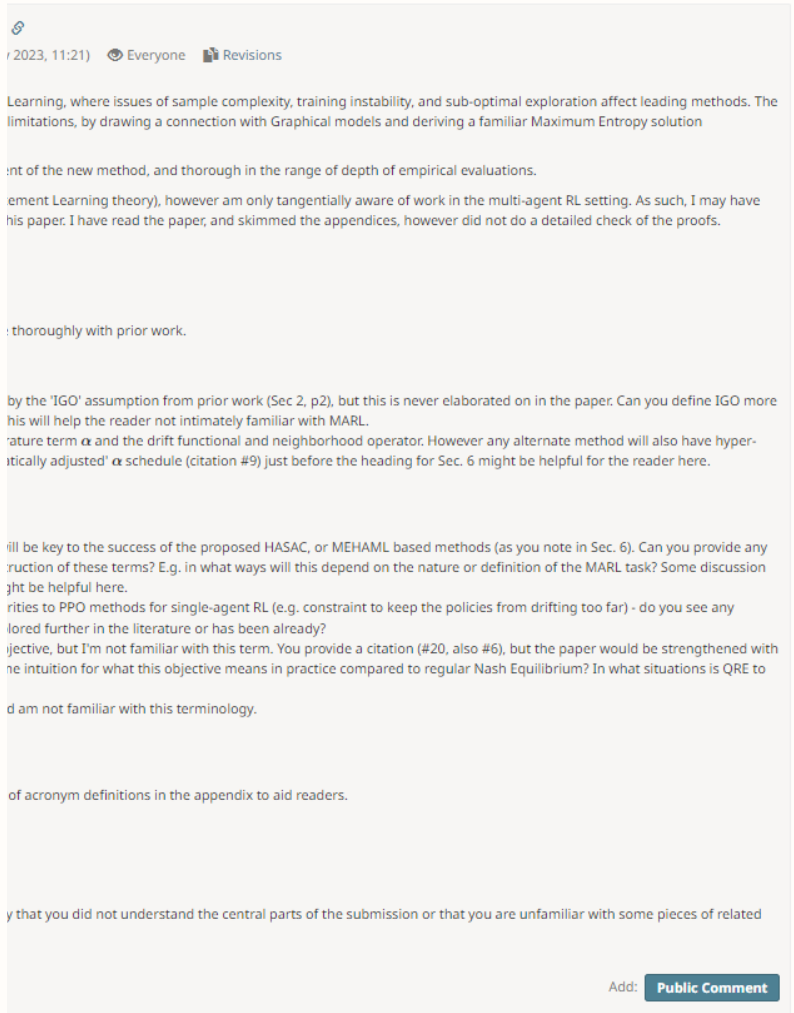
Justification For Why Not Higher Score:

While the paper demonstrates notable strengths, there are aspects (like the novelty) that should be stronger in a paper accepted for oral presentation.

Justification For Why Not Lower Score:

The paper's substantial theoretical contribution, superior empirical performance, clear presentation, and potential for broader impact collectively make it a strong candidate for a spotlight presentation rather than being limited to a poster.

Add: Public Comment



https://openreview.net/forum?id=tmqOhBC4a5

作者简介

Jiarong Liu[1♯,*], Yifan Zhong[1,2*], Siyi Hu[3],
Haobo Fu[4], Qiang Fu[4], Xiaojun Chang[3], Yaodong Yang[1†]
[1]Institute for AI, Peking University, [2]National Key Laboratory of General AI, BIGAI,
[3]University of Technology Sydney, [4]Tencent AI Lab

Jiarong Liu
guazimao

https://github.com/guazimao

https://ivan-zhong.github.io/

Dr. Yaodong Yang (杨耀东)
PKU Alignment & Interaction Lab

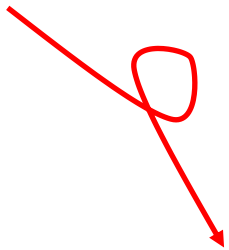
https://www.yangyaodong.com/

课题背景

Multi-agent reinforcement learning (MARL) 类型的任务其特点是难以协调单个智能体的策略改进以提高团队的整体表现，因此传统的独立学习方法往往具有较差的收敛性。

Centralized training decentralized execution (CTDE) 范式假设在训练阶段可以访问全局状态和团队成员的操作和策略，引出了multi-agent policy gradient algorithms以及value decomposition algorithms方法的发展。

Heterogeneous-agent mirror learning (HAML)为算法设计提供了模板，保证任何算法都能单调地改进联合目标并收敛到纳什均衡。



作者2022年的工作，也是本文的工作基础

Heterogeneous-agent mirror learning (HAML) 是一种针对MARL算法设计的通用模板，经过理论推导，基于此模板派生出的算法满足联合奖励的单调改进以及收敛到纳什均衡的理想属性。

文章证明了两种SOTA的MARL算法，HATRPO和HAPPO符合HAML模板范式，即是HAML的实例，并且作为理论的自然结果，提出了两种著名RL算法的HAML扩展，HAA2C（针对A2C）和HADDPG（针对DDPG），并在StarCraftII和Multi-Agent MuJoCo任务上展示了它们对强baseline的有效性。

作者2022年的工作，发表在ICLR 2022

主要挑战

HAML框架下的方法面临样本复杂性和训练不稳定的挑战。具体表现为on-policy的方法每个梯度步骤需要新的样本数据，随着任务复杂性和智能体数量的增加，这一过程的开销将变得非常昂贵。off-policy的方法则面临训练不稳定以及超参数敏感等问题。

HAML衍生的方法可能由于探索不足导致的次优纳什均衡收敛。标准的MARL总是会最大化目标奖励以获取理论上存在的确定性收敛解，因此随机性往往是not inherently encouraged。多个纳什均衡的存在是在多智能体博弈中经常观察到的现象， HAML衍生的方法可能无法充分探索并过早地收敛于次优纳什均衡。

虽然很多随即策略方法在单智能体RL中取得了巨大的成功，在需要合作的MARL中解决随机策略学习问题仍具有挑战性。一般情况下，现有的CTDE方法无法保证随机策略学习的收敛性。

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

论文分享

GREIL-Crowds: Crowd Simulation with Deep Reinforcement Learning and Examples

发表于SIGGRAPH 2023 / TOG 2023

作者简介

PANAYIOTIS CHARALAMBOUS, CYENS - Centre of Excellence, Cyprus
JULIEN PETTRÉ, Univ Rennes, Inria, CNRS, IRISA, France
VASSILIS VASSILIADES, CYENS - Centre of Excellence, Cyprus
YIORGOS CHRYSANTHOU, CYENS - Centre of Excellence, Cyprus
NURIA PELECHANO, Universitat Politecnica de Catalunya (UPC), Spain

https://totis77.github.io/

https://www.cs.upc.edu/~npelechano/

课题背景

- 微观人群建模的研究重点是找到一个人在人群中移动的主要准则，逐步去除死锁、不稳定的行为。上述工作成功的一个副作用是会产生过于流畅和完美的避障和目标追踪，失去真实人群中存在的突然变向、犹豫等细节。这些细节本质也不是基于避障或目标追踪而产生的。

- 一种有效的解决方案是数据驱动，基于包含上述细节的真实人群轨迹，无需对其进行解释建模。

主要挑战

- 基于数据驱动的方法需要大量的训练资源，并且只能模拟原始数据中的情况。
- 没有考虑历史和未来结果，总体轨迹往往表现出不一致性。
- 仿真数据和原始数据存在偏差导致预测错误。在多代理的环境中错误会随时间推移累计，导致整体效果与预期的巨大偏差。

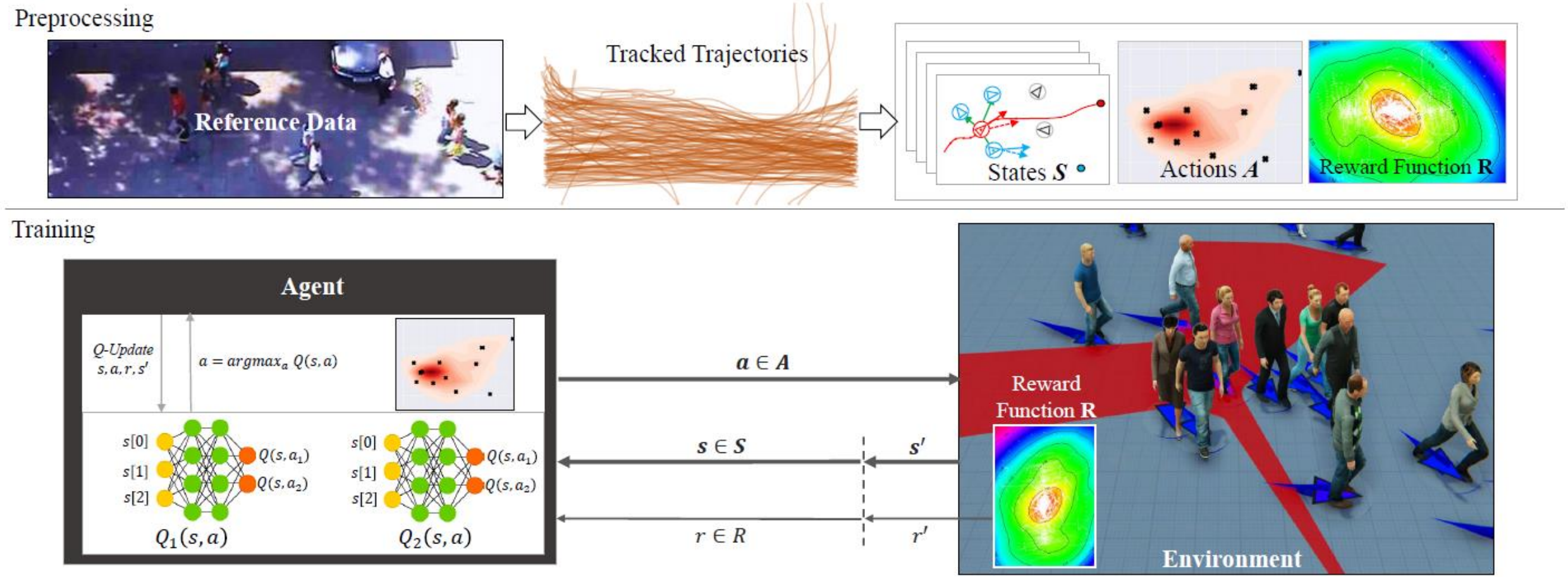
本文贡献

北京理工大学 BEIJING INSTITUTE OF TECHNOLOGY

- 实现了一个完整的学习框架可以从相对小的输入人群数据集中产生连续、合理的人群行为。

- 提出了一种数据驱动的基于novelty detection的奖励函数，其中考虑了给定状态的合理性。

Guided REinforcement Learning Crowds (GREIL Crowds)



状态空间与动作空间设计

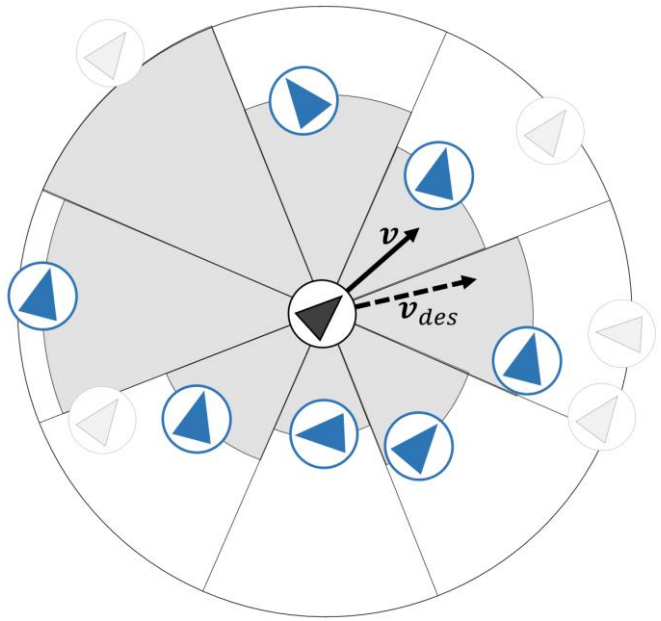
$$s = s_a \cup s_g \cup s_n$$

$$s_a = \{|\mathbf{v}|\} \in \mathbb{R}$$

$$s_g = \{\mathbf{v} - \mathbf{v_d}\} \in \mathbb{R}^2$$

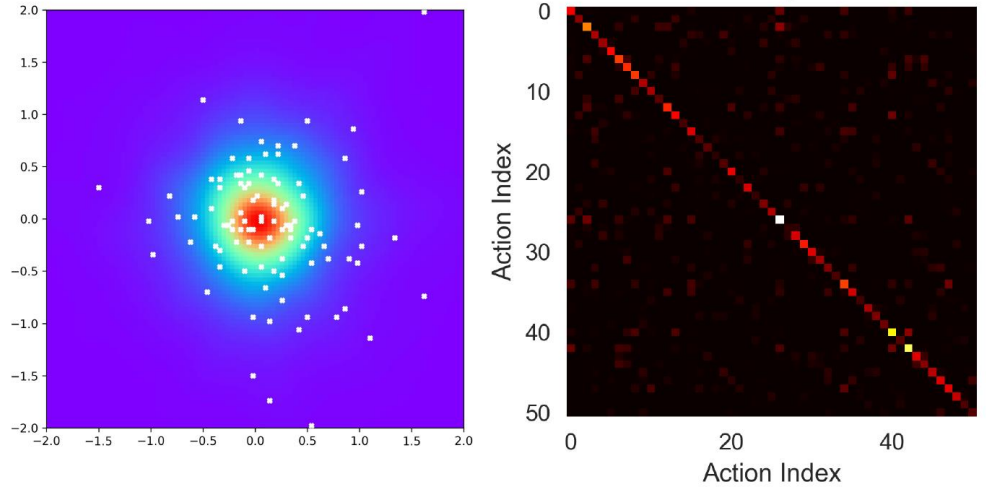
$$s_n = \{(d_i, \mathbf{v}_i - \mathbf{v})\} \in \mathbb{R}^{33}$$

$$\{\mathbf{a}_i\} \in \mathbb{R}^2$$



Goal

Kernel Density Estimation



(a) Action Space

(b) Action Sequences

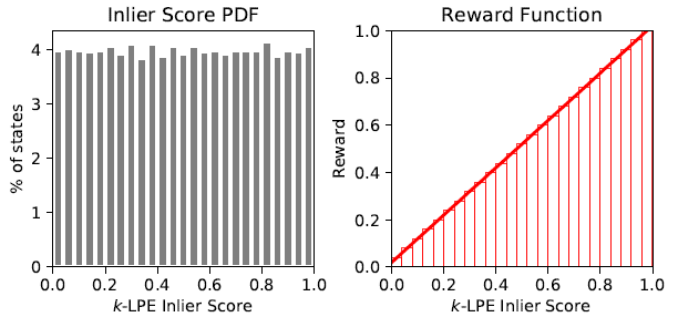
基于novelty detection的奖励函数设计

思路：人们倾向于在感到舒适的状态空间S中移动，有时倾向于高效直接（例如直行），有时有可能倾向于一些不常见的情况（例如停下来接电话），手工定义定理函数R无法解决这一问题。

首先计算基于k-LPE的分数R_S：

$$kNN(\mathbf{s}_i) = \{\mathbf{s}_{i1}, \mathbf{s}_{i2}, \ldots, \mathbf{s}_{ik}\}$$

$$R_S(\mathbf{s}_i) = d(\mathbf{s}_i, \mathbf{s}_{ik})$$



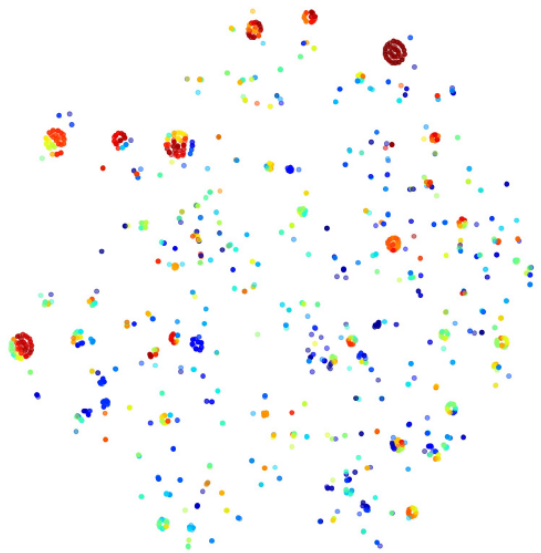
然后计算异常值pK：

$$p_K(\mathbf{s}_t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{\{R_S(\mathbf{s}_t) \leq R_S(\mathbf{s}_i)\}}$$

奖励函数：

$$\mathcal{I} = \{1 - p_K(s) : \forall s \in S\}$$

$$R(s, a, s') = P(\mathcal{I} \leq 1 - p_K(s'))$$



在奖励中引入了数据集中包含的隐式信息（包括隐式的避障），且agent访问到不常见情况的比例也与原数据集一致。

训练细节

三个选项设置起始点、终点、预期速度以及初始速度：

a) we leave the reference data values

b) we jitter them using Gaussian noise

c) we select random values in acceptable ranges

三种方法为agent选取动作：

a) by following the currently learned policy

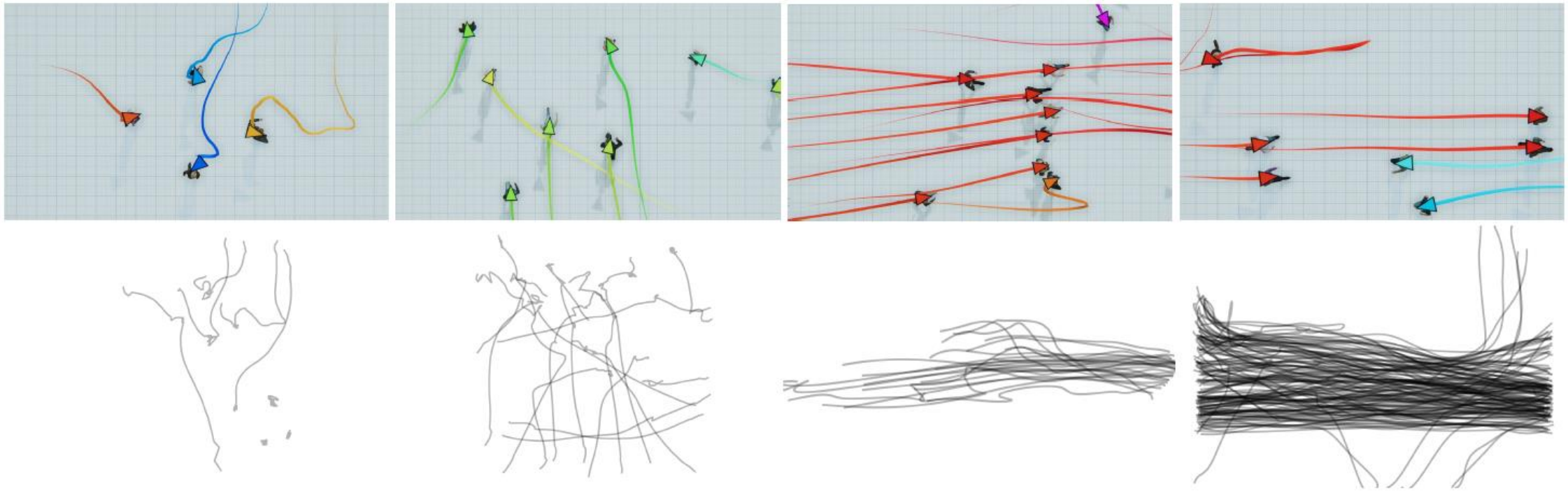
b) by choosing a completely random action

c) by selecting an action based on the statistics of the action sequences from the reference data

Table 1. Default values for hyperparameters.

Parameter	Value	Description
r	$.3m$	Agent Radius
R_s	$5m$	Maximum search distance for neighbours
H	$10s$	Maximum time for agent training
γ	.90	Discount factor
M	64000	Experience replay memory size
M_{min}	6400	Minimum experiences before training starts
E_e	1000	Episodes of that E_e is linearly changed
e_r	$1.0 - .1$	Random exploration linearly changed in E_e
e_g	.4	Guided exploration
batch size	64	Number of experiences used per update
k	100	Number of nearest neighbours used by k-LPE
freq	$5Hz$	Agent update frequency (training+simulation)
α	$1e^{-4}$	Initial learning rate
α_e	10000	Episode period to drop α
α_d	.5	Drop factor of α every α_e

数据源



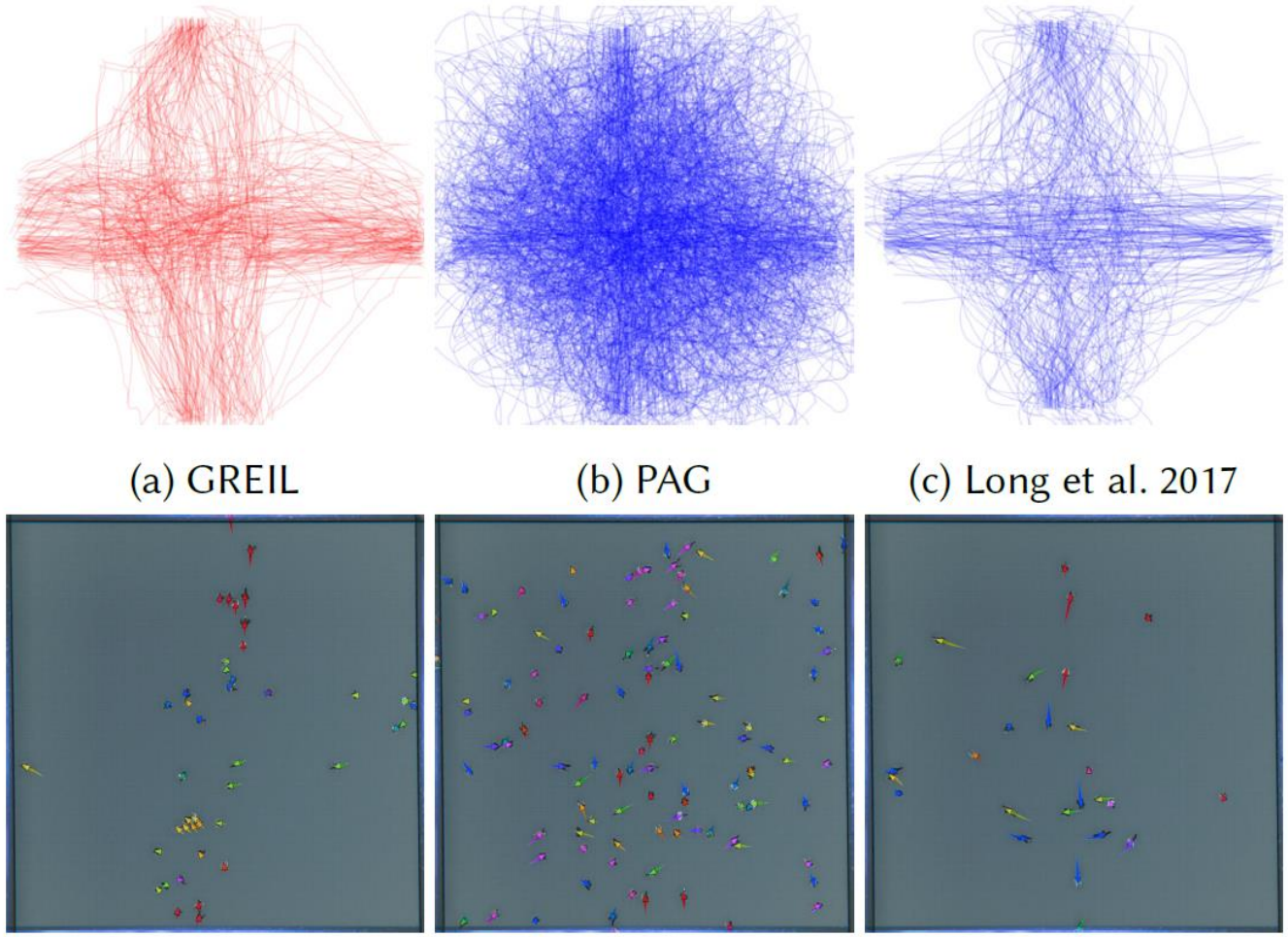
(a) Chat (b) Flock-1 (c) Flock-2 (d) Pedestrians

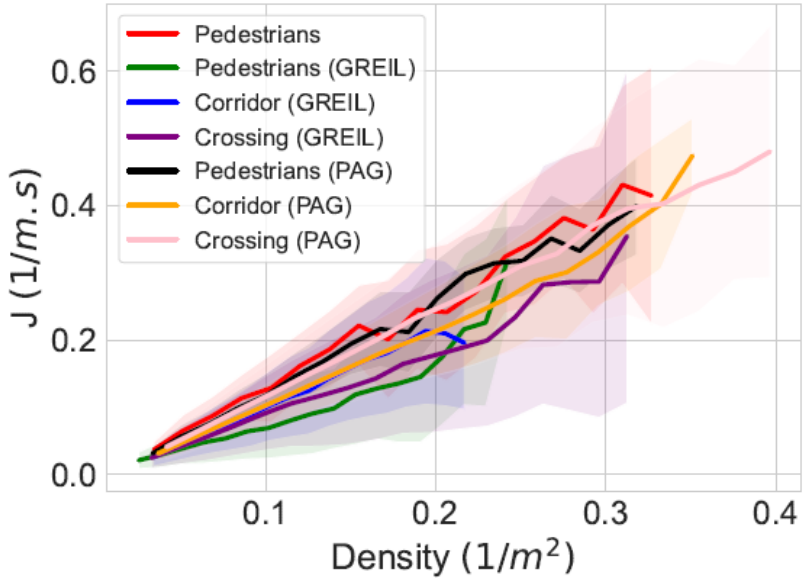
Chat is a dataset of 8 agents moving around, stopping and talking to each other.

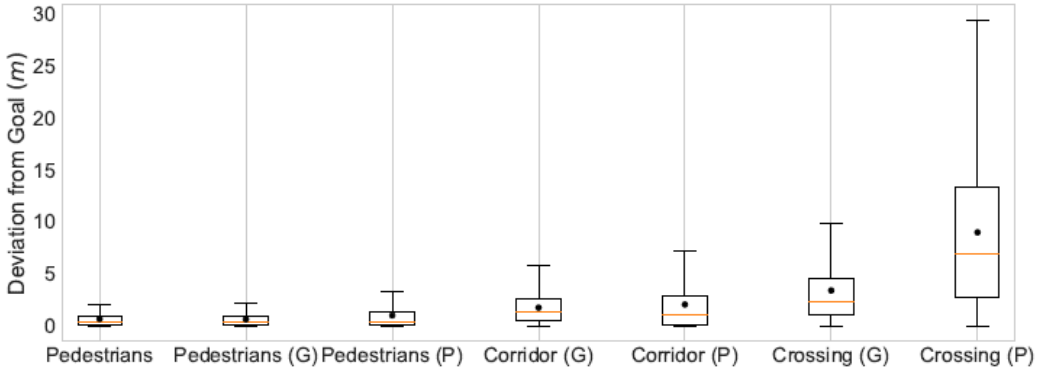
The Flocking datasets (Flock-1, Flock-2) consist of 16 and 24 agents respectively and demonstrate crowds moving together in roughly the same direction

The pedestrian dataset is the more complex one since it has 148 people with mixed behaviors such as walking by themselves and/or in small groups of 2-4 people entering and/or leaving groups, etc.

对比实验







相比数据源双倍的人群密度，强化学习效果优于监督学习

Discussion

- 可以用actor-critic或者policy-based的方法来解决连续动作空间的问题
- 当前奖励函数需要执行k近邻搜索，当状态空间维度较高时效率较低
- 仿真效果和风格依赖数据集，如果想要学习具有个性化行为的细节需要实现基于“profile”的策略
- 可以开发一个controller控制代理的生成、初始化（基于密度、流量等）（我也想要）
- 在特定场景中（游戏、电影、疏散、自动驾驶等）的实际应用

北京理工大學
BEIJING INSTITUTE OF TECHNOLOGY

Thanks

Text-Guided Synthesis of Crowd Animation (SIGGRAPH 2024)

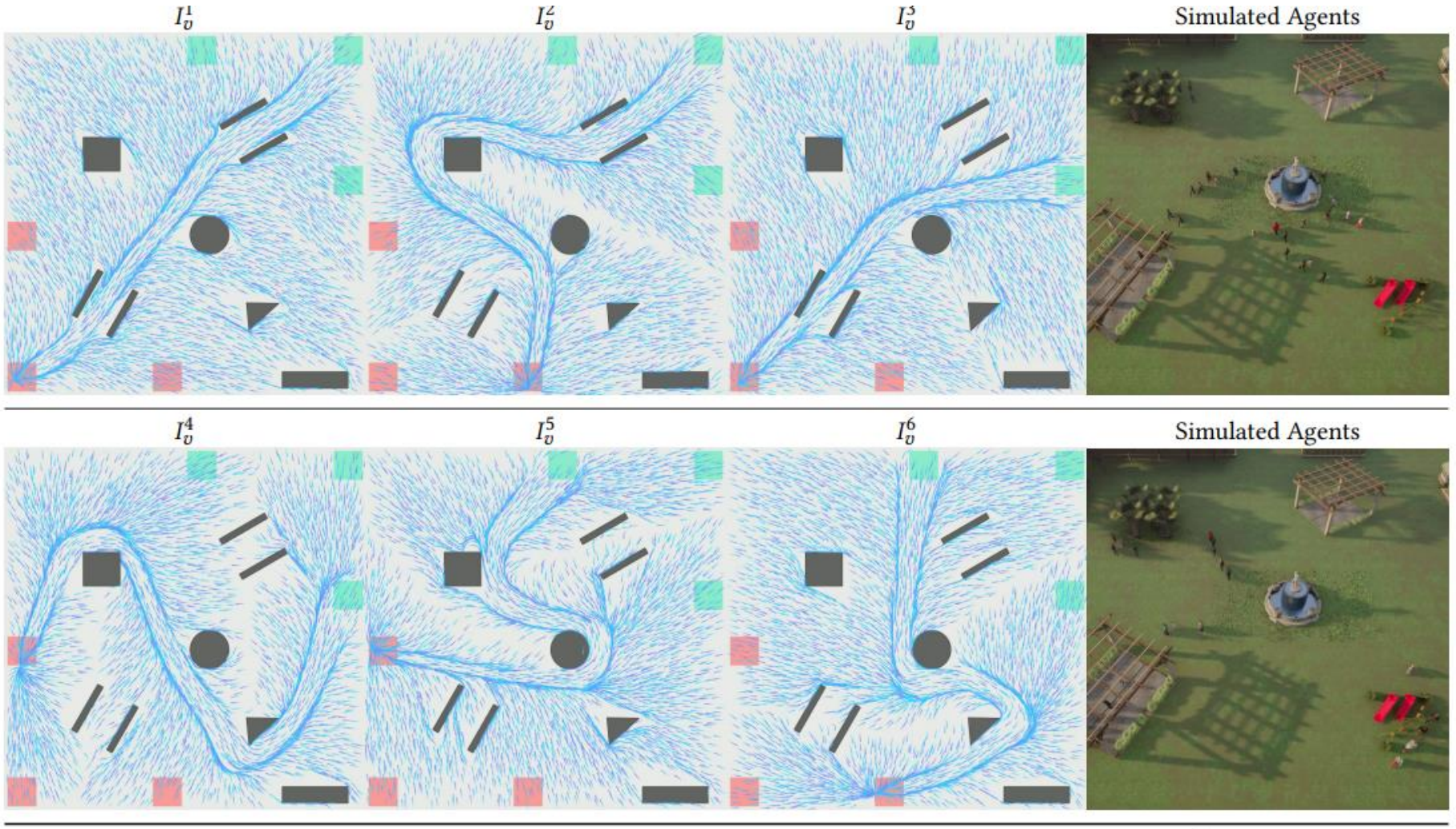


Figure 6: *A complex scenario with 6 agent groups. We show the input T, the canonicalized $\{s^j\}$, and the predicted $\{I_v^j\}$. The start/goal regions are marked in green and red, respectively.*