No.29

基于空间编码-注意力网络的多智能体社交行为理解

安琪，萧德雄
经管学院，计算机系

Background and Aims

Background

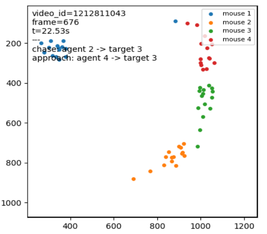
在行为神经科学与动物模型研究中，对小鼠社交行为进行**自动化、精细化**识别是理解神经回路功能、疾病模型与药物干预机制的重要基础。然而，现有行为分析流程仍高度依赖人工标注，不仅成本高、主观性强，而且在实验条件或个体差异发生变化时，标注一致性与模型泛化能力均受到严重限制。MABe（Mouse Action Behavior estimation）任务正是针对这一瓶颈提出的标准化挑战，其核心目标是将多小鼠社交行为识别问题表述为一个**基于骨架关键点序列的时序学习问题**。不同于直接处理原始视频，MABe 使用姿态估计方法提取多只小鼠的身体关键点轨迹作为输入，要求模型在仅依赖骨架信息的条件下，识别细粒度的个体行为与社交互动行为，其中许多类别具有明确的施动者—受动者（agent–target）结构。

Challenges

高维交互性：社交行为并非单体姿态的简单函数，而高度依赖个体之间的相对位置、朝向、接触关系及其随时间变化的拓扑结构，因此模型必须显式刻画跨个体的空间关系。

长时序依赖：诸如追逐、嗅探等行为通常跨越数百帧，要求模型在高频、长序列输入下有效整合远期时间信息。

极端类别不平衡：在 70 余种行为类别中，大量帧属于背景或非社交状态，而具有生物学意义的社交行为样本极为稀缺，这对模型训练与评估提出了严峻挑战。



图一：小鼠骨架关键点一例

Aims

针对上述问题，我们提出了一种**端到端的混合时空建模框架**。该框架通过显式的拓扑与关系建模刻画多小鼠之间的交互结构，结合高效的时间建模机制以捕捉长距离时序依赖，并在训练过程中引入针对类别不平衡的策略，以提升对稀有社交行为的识别能力。整体目标是在保持计算效率的同时，实现对多主体社交行为的稳定、可扩展识别。

Related works

传统序列建模（Traditional Sequence Modeling）

早期的行为识别方法常使用 RNN、LSTM 等循环神经网络对骨架坐标序列建模，但这些方法通常将每帧关键点简单拼接成向量输入序列模型，忽略了骨架本身丰富的空间结构与个体间的交互关系，使得空间信息难以被有效利用。

图神经网络与注意力机制（GNN & Attention）

为解决结构性弱建模的问题，图神经网络（GNN）被广泛用于骨架数据的空间关系建模，通过定义关节节点及其连边来捕获非欧几里得几何结构。在时序动作识别领域，引入注意力机制的 Transformer 结构能够灵活建模长距离依赖与跨时间上下文，从人类动作识别扩展到动物行为建模中表现出良好潜力。

MABe/CalMS21 等 Benchmark 及相关方法

MABe Challenge 源于 CalMS21 数据集与 Multi-Agent Behavior Benchmark，该基准由多个实验室采集的大规模小鼠社交行为轨迹构成，是行为识别研究的重要推动力量。基于这些数据集的顶会论文表明，在动物行为识别中不仅需要处理关键点时序，还须考虑交互代理之间的关系结构，并能在不同实验设置下保持泛化。相关研究提出了包括自监督表示学习、多任务学习及跨数据集迁移学习的方法，以评估和提升对复杂行为的判别性能。

最新进展

近期的基准（例如 MABe22）进一步扩展了多物种、多任务行为分析框架，并对比了多种视频与轨迹表示学习方法，指出从人类动作识别直接迁移的模型在动物行为建模上仍存在性能差异。

Results

在实现相对复杂的模型之前，我们先使用基础空间编码与课上所学的基础模型进行了baseline测试：

	Lab-averaged Macro F1-score
LSTM	0.21
CNN	0.20
XGBoost	0.17

通过引入更复杂的空间编码与特征信息，对不同行为分别训练一个XGBoost，性能得到了显著提升：

XGBoost	0.42

进一步的，采用我们提出的架构：

HHSTF	0.50

Key References

Yan, S., Xiong, Y., & Lin, D. (2018). Spatial temporal graph convolutional networks for skeleton-based action recognition. In Proceedings of the AAAI conference on artificial intelligence (Vol. 32, No. 1).

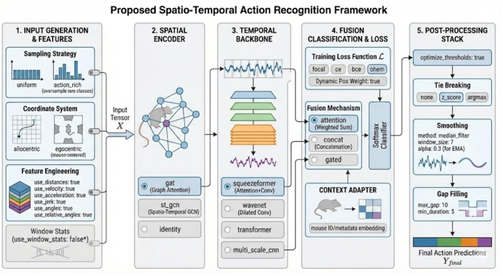
Kim, S., Gholami, A., Shaw, A., et al. (2022). Squeezeformer: An Efficient Transformer for Automatic Speech Recognition. Advances in Neural Information Processing Systems (NeurIPS), 35, 9361-9373.

Sun J J, Karigo T, Chakraborty D, et al. The multi-agent behavior dataset: Mouse dyadic social interactions[J]. Advances in neural information processing systems, 2021, 2021(DB1): 1.

Sun J J, Marks M, Ulmer A W, et al. Mabe22: A multi-species multi-task benchmark for learned representations of behavior[C]//International Conference on Machine Learning. PMLR, 2023: 32936-32990.

Model and Methods

针对 MABe 的 HHSTF Network



图二：HHSTF网络结构

我们提出的 **HHSTF（Hybrid Hierarchical Spatio-Temporal Framework）Network** 面向 MABe 的多小鼠骨架序列识别任务，整体流程为：**输入特征构建 → 单帧空间拓扑编码 → 长时序骨干建模 → 分类与损失设计（不平衡友好）→ 推理与后处理**。模型目标是在保持计算效率的同时，对"多主体交互 + 长跨度行为 + 稀有类别"三类困难同时有针对性处理。

评估指标（Evaluation Metric）

$$F_\beta = (1+\beta^2)\cdot\frac{\text{Precision}\cdot\text{Recall}}{(\beta^2\cdot\text{Precision})+\text{Recall}}$$

本研究采用分**实验室宏平均 F-beta 分数**，实验中取 β=1。先对实验室内所有类别的 F-beta 取平均（Macro-average），再对所有实验室的得分取算术平均，得到最终 Score。该指标对类别不平衡具有鲁棒性，且能有效评估模型在不同实验环境下的泛化能力。

输入处理（Input Representation）

为减少场景与绝对位置干扰，所有关键点坐标均转换为**以小鼠为中心的相对坐标系**。在此基础上，引入速度、加速度及三阶差分（Jerk）等运动学特征，以增强对快速突变动作与细微震颤的刻画能力。最终输入以固定长度时间窗口（最长 512 帧）组织。

空间拓扑编码（Spatial Graph Encoding）

在单帧层面，我们采用**基于图的空间编码器**建模骨架结构与个体间交互关系。节点对应小鼠身体关键点，边同时包含：

- 体内连接（解剖结构先验），
- 体间连接（跨小鼠的接近与互动关系）。

通过图注意力机制，自适应聚合关键点及交互信息，得到结构感知的帧级空间表示。

长时序建模（Temporal Backbone）

为捕捉跨数百帧的社交行为动态，模型使用**高效 Transformer 变体**对时间维进行全局建模。同时引入卷积模块增强局部时序特征，有助于识别动作的起始与终止边界，从而兼顾长程依赖与短时动态。

训练与不平衡处理（Loss & Training Strategy）

针对 MABe 中严重的类别不平衡问题，训练阶段采用**动态加权损失策略**，提高稀有社交行为与困难样本的权重，抑制背景帧主导效应，从而提升对低频行为的识别性能。