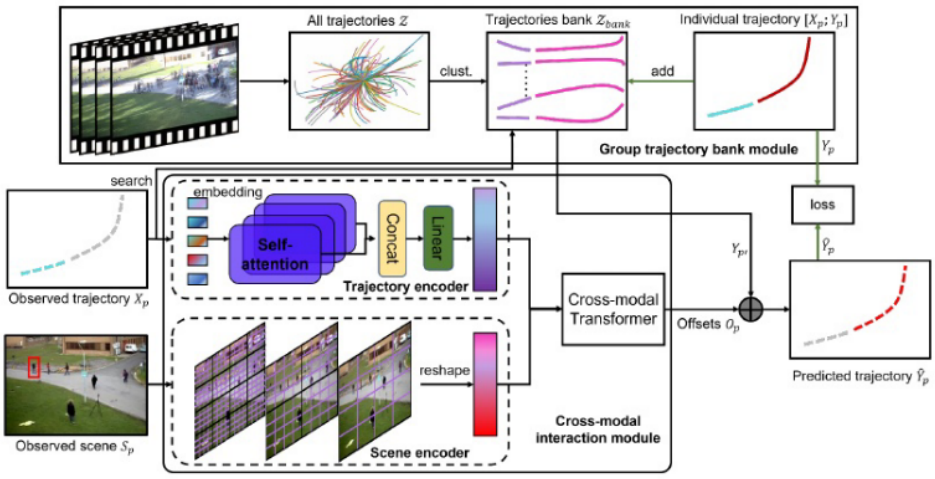
SHENet：从场景历史预测人类轨迹

1. SHENet 概述

人类轨迹预测是从视频片段中预测目标人的未来路径，人类轨迹预测是计算机视觉领域的一个关键任务，对于智能交通、监控系统等应用具有重要意义。然而，由于人类运动的随机性和主观性，准确预测未来轨迹一直是一个挑战。尽管如此，在特定场景中，人类的运动模式往往遵循一定的规律，这些规律受到场景限制（如建筑布局、道路和障碍物）以及人与人、人与物体之间的交互影响。基于此，本文提出利用场景历史中的规律性来预测个体的未来轨迹。

SHENet（Scene History Excavating Network）是一种用于从场景历史中预测人类轨迹的框架。该框架通过学习场景中的隐含规律来预测个体的未来轨迹，这些规律被称为“场景历史”，包括历史群体轨迹和个体与周围环境的交互。SHENet 包含两个主要模块：群体轨迹库模块（GTB）和跨模态交互模块（CMI）。GTB 通过聚类历史轨迹生成代表性群体轨迹，为未来轨迹预测提供候选路径；CMI 则分别对个体轨迹和场景进行编码，并通过跨模态变换器建模它们之间的交互，以优化候选轨迹。此外，为了应对人类运动的随机性和主观性，作者提出了曲线平滑（CS）技术，并将其纳入训练过程和评估指标中，以减少不确定性的影响。实验结果表明，SHENet 在 ETH、UCY 和新提出的 PAV 数据集上均优于现有方法。

2 技术细节



图表 1 SHENet 模型架构图

2.1 群体轨迹库（GTB）的建立与更新

建立: 从视频帧中提取历史人物轨迹Z,分为过去时刻和未来时刻两部分$Z_P = \{X_p|Y_p\}$，划分时间点根据已有轨迹时长确定。计算过去-未来对的欧氏距离，通过 k 聚类法得到代表性历史轨迹库Z_{bank}。

更新: 将通过 CMI 模块得到的偏移量（offsets）O_P与候选的未来轨迹Y_p,相加得到最终的预测轨迹$\widehat{Y_P}$。在训练阶段，如果Y_p到$\widehat{Y_P}$的距离大于距离阈值θ（distance threshold），则人的轨迹（即X_p和Y_p）将被添加到轨迹库Z_{bank}中。如果新加入的轨迹数量大于轨迹阈值β（trajectory threshold），则对轨迹库Z_{bank}重新进行计算聚类。训练结束后，对轨迹库进行固定，用于推理

2.2 轨迹搜索

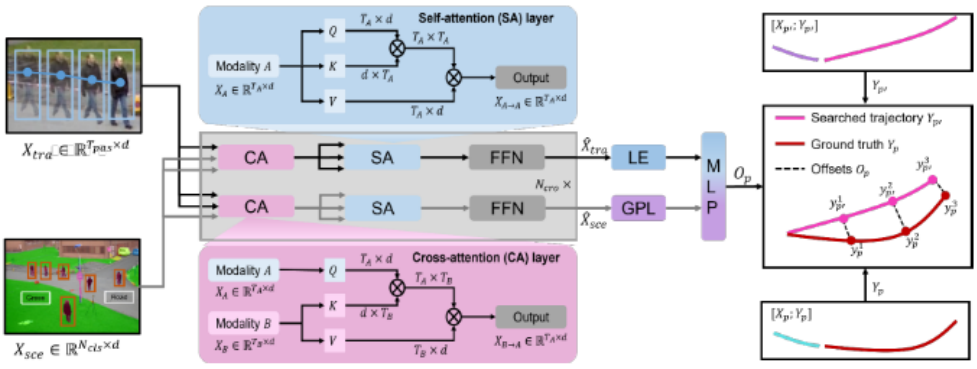
以观察到的X_p作为关键，计算其与Z_{bank}中过去轨迹$\{X_i\}_i^{|Z_{bank}|}$的相似度得分，并根据最大相似度得分找到具有代表性的轨迹$Y_{p'}$,相似度公式如下:

$$s_i = \frac{X_p \cdot X_i}{\|X_P\|\|X_i\|}, \quad 1 \le i \le |Z_{bank}|$$

2.3 跨模态交互（CMI）

CMI 模块由两个单模态编码器和一个跨模态交换器组成，通过轨迹和场景信息的输入，最终得到关于候选轨迹$Y_{p'}$的偏移量O_P

首先，通过轨迹编码器和场景编码器分别对历史轨迹数据X_P和场景信息S_p进行特征提取。轨迹编码器包含嵌入层（embedding）、自注意层（Self-attention）、连接层（Contact）和线性层（Linear），依次完成轨迹数据的高维映射、时间步关系捕捉、向量拼接和归一化处理; 场景编码器则对场景图片进行特征提取并重塑（reshape），使其与轨迹特征维度一致。



图表 2 跨模态交换器

之后，将提取的轨迹特征$X_{tra} \in \mathbb{R}^{\mathrm{Tpas}\times\mathrm{d}}$和场景特征$X_{sce} \in \mathbb{R}^{\mathrm{Ncls}\times\mathrm{d}}$输入交互模块（Cross-modal Transformer），该模块由交互注意层（CA）和自注意层（SA）组成，采用两流结构（two-stream structure），上层 CA 以轨迹特征为查询、场景特征为键值，输出与人体运动相关的场景信息；下层 CA 以场景特征为查询、轨迹特征为键值，输出受环境约束的人体运动信息。自注意层（SA）则对信息进行加权增强，捕捉时间依赖关系，增强特征表示。接着，通过前馈神经网络（FFN）进行非线性变换，增强特征表达能力。

$$\widehat{X_{tra}} = E_{cro}(X_{tra}, X_{sce}) \quad , \quad \widehat{X_{sce}} = E_{cro}(X_{sce}, X_{sce})$$

$$O_P = MLP\left(\left[LE\left(\widehat{X_{tra}}\right); GPL\left(\widehat{X_{sce}}\right)\right]\right)$$

最后，对$\widehat{X_{tra}}$提取最后一个元素$h_{tra} \in \mathbb{R}^{\mathrm{d}}$，代表序列的最终状态或总结，减少计算；对$\widehat{X_{sce}}$进行全局池化操作，进一步整合信息得到$h_{sce} \in \mathbb{R}^{\mathrm{d}}$；最后利用多层感知机（MLP）计算基于场景信息交互的偏移量$O_P \in \mathbb{R}^{\mathrm{Tfut}\times 2}$

3 模型表现

不同于以往的方法，SHENet 将场景信息分为历史群体轨迹（HGT）和个体 - 环境交互（ISI）两类，并通过创新的框架设计充分利用这两类信息进行轨迹预测以往方法多关注个体轨迹或部分场景历史，而 SHENet 通过历史群体轨迹为个体轨迹预测提供更可靠的参考路径，并结合个体 - 环境交互信息进行精细化调整，这种综合考虑场景历史的方式更具创新性和有效性。

在 ETH 和 UCY 数据集上，SHENet 取得了显著优于现有最佳方法的性能。平均 FDE 从 0.39 降低到 0.36，与之前的最佳方法 YNet 相比，实现了 7.7% 的

提升。在 ETH 数据集上，当轨迹运动幅度较大时，SHENet 的优势尤为明显，ADE 和 FDE 分别提升了 12.8% 和 15.3%。

在更具挑战性的 PAV 数据集上，SHENet 同样展现出强大的性能。与 YNet 相比，SHENet 在 CS-ADE 和 CS-FDE 上分别平均提升了 3.3% 和 10.5%。在 PETS 数据集上，CS-FDE 提升了 16.2%，这表明 SHENet 在处理复杂场景和大运动轨迹时具有显著优势。