基于图卷积与自适应Transformer的行人轨迹预测

王芳芳,刘明华*,渠连恩,王 贺,李丹宁
(青岛科技大学信息科学技术学院,山东青岛 266061)

摘 要: 行人轨迹预测是自动驾驶和机器人导航等领域的核心挑战之一,其关键在于如何有效建模行人间的复杂交互关系并提取多尺度时空特征. 本文提出一种基于图卷积与自适应Transformer的行人轨迹预测方法(pedestrian trajectory prediction method based on Graph Convolution and Adaptive Transformer,GCAT),通过层次化的特征提取与自适应交互建模实现高精度的轨迹预测. 模型以历史观测时间窗口内所有行人的位置与速度信息作为输入,首先通过线性投影与正弦-余弦位置编码将原始观测映射至高维特征空间,以显式保留时序顺序信息. 随后,引入关系图卷积网络捕获行人之间的局部拓扑结构及空间交互强度,通过基于特征余弦相似度的自适应邻接矩阵实时构建交互图,使图结构能够根据场景特征自适应调整. 同时,引入增强型多层卷积结构,通过可学习的残差权重自适应平衡不同层级特征的贡献,有效缓解深层网络的梯度消失问题,增强局部交互特征的表达能力. 此外,模型进一步引入空间自适应Transformer建模全局时空依赖关系,该模块通过可学习的空间偏移量实现特征图上的连续采样. 具体实现中,模型通过线性层从输入特征中生成空间偏移量和注意力权重,偏移量与参考点坐标相加后经归一化得到实际采样位置,利用双线性插值从特征图中提取对应位置的特征值,再通过注意力权重进行加权聚合,获得对局部几何变化与全局时序依赖的增强表达. 这种连续采样策略使模型能够聚焦于对轨迹预测最相关的空间区域,自适应地应对不同场景的几何布局变化. 同时,模型融合多粒度时序特征,逐步提取从局部交互到全局依赖的多层次时空表达,有效解决了现有方法在长程依赖建模、环境适应性以及多尺度特征融合等关键方面存在的问题. 在实验验证方面,本文在两个广泛使用的公共行人轨迹预测数据集ETH和UCY上对所提出的方法进行了系统评估. 相比现有基线模型,所提出模型在平均位移误差(Average Displacement Error,ADE)和最终位移误差(Final Displacement Error,FDE)指标上分别取得了5.1%和13.2%的性能提升,验证了模型在复杂交互关系建模和多尺度时空特征提取方面的有效性与先进性.

关键词: 轨迹预测;局部拓扑结构;全局时序依赖;多尺度特征融合;预测性能
中图分类号: TP391.4 **文献标识码:** A **文章编号:** 0372-2112(2025)12-4507-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20250855

Pedestrian Trajectory Prediction Based on Graph Convolution and Adaptive Transformer

WANG Fang-fang, LIU Ming-hua*, QU Lian-en, WANG He, LI Dan-ning
(*School of Information Science and Technology*, *Qingdao University of Science and Technology*, *Qingdao*, *Shandong* 266061, *China*)

Abstract: Pedestrian trajectory prediction is one of the core challenges in fields such as autonomous driving and robotic navigation. Its key difficulty lies in effectively modeling complex interactions among pedestrians and extracting multi-scale spatiotemporal features. This paper proposes a pedestrian trajectory prediction method based on graph convolution and adaptive transformer (GCAT), which achieves high-precision trajectory prediction through hierarchical feature extraction and adaptive interaction modeling.The model takes the position and velocity information of all pedestrians within a historical observation window as input. First, linear projection and sinusoidal positional encoding are applied to map the raw observations into a high-dimensional feature space, explicitly preserving temporal order information. Subsequently, a relational graph convolutional network is introduced to capture local topological structures and spatial interaction strengths among pedestrians. An adaptive adjacency matrix based on feature cosine similarity is constructed in real time to model pedestrian interactions, enabling the graph structure to dynamically adjust according to scene characteristics. In addition, an enhanced multi-layer convolutional structure is employed, where learnable residual weights are used to adaptively balance the contri-

收稿日期:2025-09-27;录用日期:2025-12-15;责任编辑:覃怀银
*通讯作者:刘明华

butions of features at different layers. This design effectively alleviates the gradient vanishing problem in deep networks and strengthens the representation capability of local interaction features. Furthermore, the model incorporates a spatially adaptive Transformer to model global spatiotemporal dependencies. This module achieves continuous sampling over feature maps through learnable spatial offsets. Specifically, spatial offsets and attention weights are generated from the input features via linear layers. The offsets are added to reference point coordinates and normalized to obtain actual sampling locations. Bilinear interpolation is then used to extract feature values at these locations from the feature maps, which are subsequently aggregated using the attention weights. This process yields enhanced representations that capture both local geometric variations and global temporal dependencies. The continuous sampling strategy enables the model to focus on spatial regions most relevant to trajectory prediction and to adaptively handle geometric layout variations across different scenes. Meanwhile, the model further integrates multi-granularity temporal features, progressively extracting multi-level spatiotemporal representations ranging from local interactions to global dependencies. This design effectively addresses key limitations of existing methods in modeling long-range dependencies, environmental adaptability, and multi-scale feature fusion. For experimental validation, the proposed method is systematically evaluated on two widely used public pedestrian trajectory prediction datasets, ETH and UCY. Compared with existing baseline models, the proposed approach achieves improvements of 5.1% and 13.2% in terms of average displacement error (ADE) and final displacement error (FDE), respectively, demonstrating its effectiveness and superiority in complex interaction modeling and multi-scale spatiotemporal feature extraction.

Key words: trajectory prediction; local topology structure; global temporal dependencies; multi-scale feature fusion; prediction performance

1 引言

行人轨迹预测是人工智能领域研究的核心问题之一,在自动驾驶[1]、智能交通[2]以及社交机器人[3]等众多实际应用中具有关键作用. 现有的行人轨迹预测方法主要可分为基于物理模型的方法、基于统计学习的方法以及基于深度学习的方法. 传统的基于物理模型的方法虽然具有较强的可解释性,但往往难以处理复杂的交互关系和不确定性因素[4]. 基于统计学习的方法通过概率建模,在一定程度上刻画不确定性和多模态分布,但由于模型表达能力有限,这类方法在处理大规模场景和复杂群体交互时往往表现不足[5].

近年来,随着深度学习技术的发展,图卷积神经网络(Graph Convolutional Networks, GCN)和 Transformer 架构在行人间建模领域展现出显著优势. GCN 能够通过显式构建图结构,有效刻画交通参与者或行人之间的空间拓扑关系与局部交互模式,在交通流预测和多主体建模任务中取得了良好效果[6]. 而 Transformer 通过自注意力机制在全局范围内建模时序依赖关系,能够高效融合多模态时空信息,在自动驾驶运动预测与规划等复杂场景中展现出强大的建模能力[7]. 然而,单独使用 GCN 或 Transformer 均存在局限性:GCN 难以建模复杂的长距离依赖关系,而标准 Transformer 则缺乏对局部空间结构的敏感性,且计算复杂度随行人数量增长而上升[8].

为了解决上述挑战,本文提出了基于图卷积与自适应 Transformer 的行人轨迹预测方法(pedestrian trajectory prediction method based on Graph Convolution and Adaptive Transformer, GCAT). 在 ETH、UCY 数据集上的大量实验表明,本文的方法优于现有的基线方法,模型计算量及性能均有了大幅度提升. 消融实验和定性分析进一步验证了本文的方法可以准确地模拟行人复杂的时空相互作用,并描述行人未来运动的多样性. 主要贡献总结如下:

(1)提出层级关系增强的 GCN-Transformer 融合架构,特别引入双层 GCN 模块,有效结合局部结构感知与全局依赖建模,能够多层次地捕捉行人间复杂的时空依赖关系.

(2)构建自适应空间感知机制,使模型能够根据场景动态调整对空间位置信息的关注度,从而在不同交互环境下实现更灵活的注意力分配与更强的空间建模能力.

(3)引入多粒度时序特征融合策略,通过不同时间粒度的特征提取和交互建模,实现对行人行为模式的多层次理解,进一步强化了模型在复杂场景下的预测能力.

2 相关工作

2.1 基于图卷积神经网络的行人轨迹预测

GCN 因其能有效建模行人间的复杂交互关系,在行人轨迹预测领域展现出显著优势. 早期研究方面,Social-GAN[9](Generative Adversarial Networks)首次将图卷积网络引入轨迹预测任务,通过构建动态图结构捕获行人间的社会交互模式,奠定了 GCN 在该领域的基础;随后,Huang 等人[10]提出时空图注意力网络,通过结合时间和空间维度的图卷积操作,显著增强了对长期依赖关系的建模能力;文献[11]则采用时空图卷积

架构,通过交替堆叠图卷积层和时间卷积层来处理动态图序列;文献[12]提出自适应图卷积机制,能够动态调整邻接矩阵以适应不同的交互模式. 然而,现有方法大多采用固定的图构建策略,难以自适应学习复杂场景中的动态关系拓扑,这一局限性制约了模型的泛化能力. 针对上述问题,本文提出GCAT模型,能够基于行人特征的余弦相似度动态构建邻接矩阵,自适应地学习群体层面的交互拓扑结构,并在多尺度变换框架中与局部成对关系建模相结合,实现了从细粒度个体交互到宏观群体动态的全方位图结构建模,显著提升了复杂场景下轨迹预测的准确性,特别是在高密度人群和复杂交互场景中表现出优越性能.

2.2　基于Transformer的行人轨迹预测

近年来,Transformer模型[13~16]凭借其在长距离依赖建模方面的卓越能力,开始被研究者应用于行人轨迹预测领域. 早期工作如Yao等人[17]主要利用其强大的时序建模能力,通过自注意力机制捕捉历史轨迹中的时间依赖性. 然而,复杂交互场景中仅依赖时序信息的模型预测能力不足. 为同时建模时空依赖性,后续研究引入了空间交互信息. Yu等人[18]利用图结构来描述行人间复杂的空间交互,并借助Transformer来捕捉这些交互在时间维度上的动态演化;Mangalam等人[19]则通过强化位置编码,显著提升了模型对空间关系的理解能力;而Yuan等人[20]创新性地提出代理级注意力机制,专门用于建模行人间的复杂交互关系. 但这些方法往往缺乏对多尺度时空依赖性的统一建模能力,难以同时捕获不同时间跨度和空间尺度下的复杂交互模式. 不同于上述方法,本文提出的GCAT模型在整体架构中引入多尺度特征建模机制,并通过结合自适应空间感知Transformer,并行处理表征行人个体的节点特征与表征交互的边特征. 该设计不仅保留了捕捉个体时序动态的能力,更能以多尺度方式显式建模行人间的社会影响,深度编码复杂时空依赖关系,从而提升轨迹预测精度.

3　用于轨迹预测的GCAT模型

GCAT模型的核心在于构建一条“由近及远”的信息传递链路. 首先,利用关系型GCN层提取行人邻域的局部拓扑特征;随后,通过自适应空间感知Transformer层,将这些局部线索整合为全局关系表征,从而实现局部结构感知与全局依赖建模的有机衔接. 如图1所示,GCAT模型包含4个部分:特征初始化(左下角)、关系型GCN(左上角虚线框内)、自适应空间感知Transformer(中间虚线框内)以及未来轨迹解码器. 其中,右上角虚线框内是GCN层核心原理.

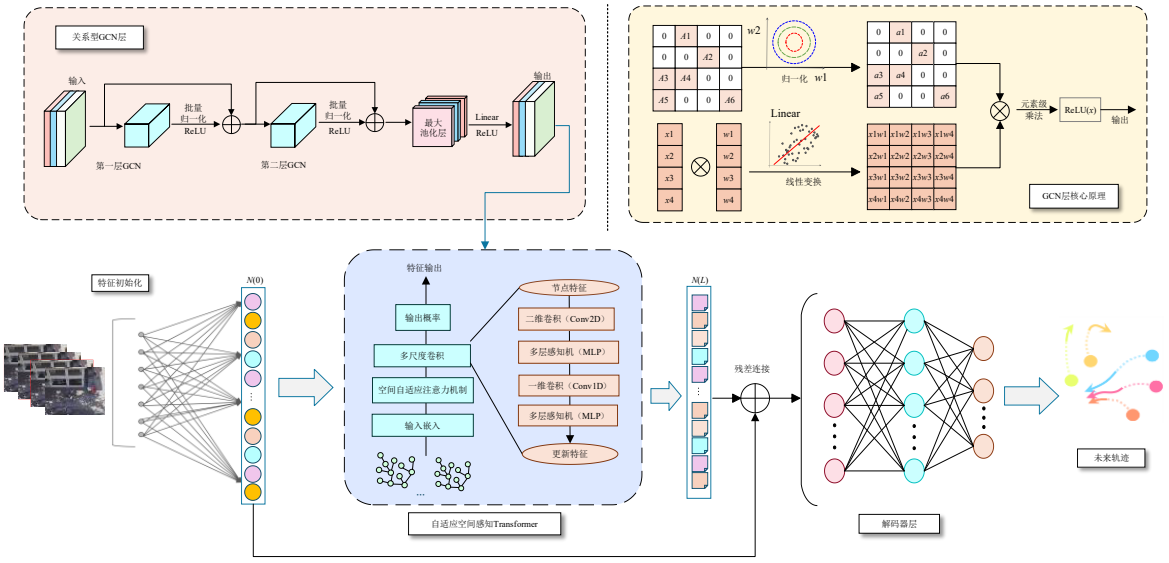


图1　GCAT模型架构图

3.1　关系型GCN

本文提出一种基于GCN的关系建模框架,采用“动态构建-分层提炼”的层级化策略,有效提升了在复杂交互场景下的建模能力. 如图2所示,GCN的核心在于节点特征的线性变换与图结构信息的深度聚合. 节点原始特征向量经可学习权重矩阵映射至高维空间后,与归一化邻接矩阵进行聚合,并通过激活函数引入非线性,最终生成更具判别性的节点表示. 其核心操作如式(1)所示:

$$\boldsymbol{X}'=\sigma\left(\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}}\boldsymbol{X}\boldsymbol{W}\right) \tag{1}$$

其中,$\boldsymbol{A}$表示邻接矩阵,$\boldsymbol{D}$为度矩阵,$\boldsymbol{X}$为节点特征,$\boldsymbol{W}$

为可学习权重矩阵.

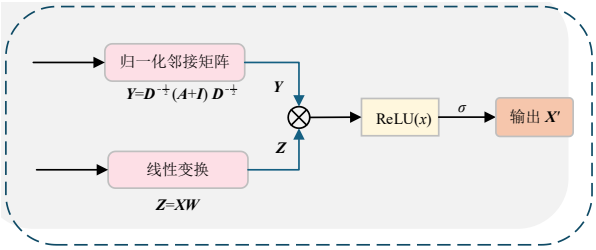


图2 GCN核心原理

所提框架进一步采用双层GCN的架构,如图1左上角所示,其核心原理在于通过堆叠GCN层逐步聚合多跳邻居信息,以实现由浅入深的特征提炼,并利用残差连接增强信息流和模型深度. 第一层GCN进行初步的节点特征转换和一跳邻居信息聚合,如式(2)所示:

$$\boldsymbol{H}^{(1)\mathrm{raw}}=\check{\boldsymbol{A}}\boldsymbol{X}^{(0)}\boldsymbol{W}^{(1)} \tag{2}$$

其中,$\boldsymbol{W}^{(1)}\in\boldsymbol{R}^{[d_1\times d_0]}$是第一层的可学习权重矩阵.

第二层在此基础上捕捉更复杂的多跳关系. 它继承第一层生成的图结构G及已更新的特征,并执行第二轮的特征传播与提炼,有效推断高阶社会动态,扩大模型感受野,如式(3)所示. 最终,聚合后的特征经筛选后用于生成多样化未来轨迹.

$$\boldsymbol{H}^{(2)\mathrm{raw}}=\check{\boldsymbol{A}}\boldsymbol{H}^{(1)}\boldsymbol{W}^{(2)} \tag{3}$$

在复杂关系建模中,本文提出一种基于特征相似度的自适应邻接矩阵生成机制,以动态捕捉场景中隐含的交互结构. 具体而言,给定当前时刻的节点特征矩阵,模型首先对特征向量进行L_2归一化,将其约束在高维特征空间的单位范数下. 随后,通过计算节点间的成对余弦相似度构建原始邻接矩阵$\boldsymbol{A}$. 为保证图卷积的数值稳定性,该矩阵进一步经过对称归一化拉普拉斯变换处理,并最终应用于图卷积操作中,如式(4)所示:

$$\boldsymbol{Y}=\boldsymbol{D}^{-\frac{1}{2}}(\boldsymbol{A}+\boldsymbol{I})\boldsymbol{D}^{-\frac{1}{2}} \tag{4}$$

这种生成机制不再依赖预定义的固定拓扑,而是允许图结构随着行人特征的演变而自适应更新,从而精准地聚合语义相似的邻居信息.

实验结果表明,本文的关系型GCN框架能够有效捕捉行人间的复杂交互模式,显著提升轨迹预测的准确性,特别是在拥挤场景和复杂交互情境下表现出色.

3.2 自适应空间感知Transformer

该模块通过融合自适应注意力机制与多尺度卷积,有效提取输入中的空间上下文信息,如图3所示.

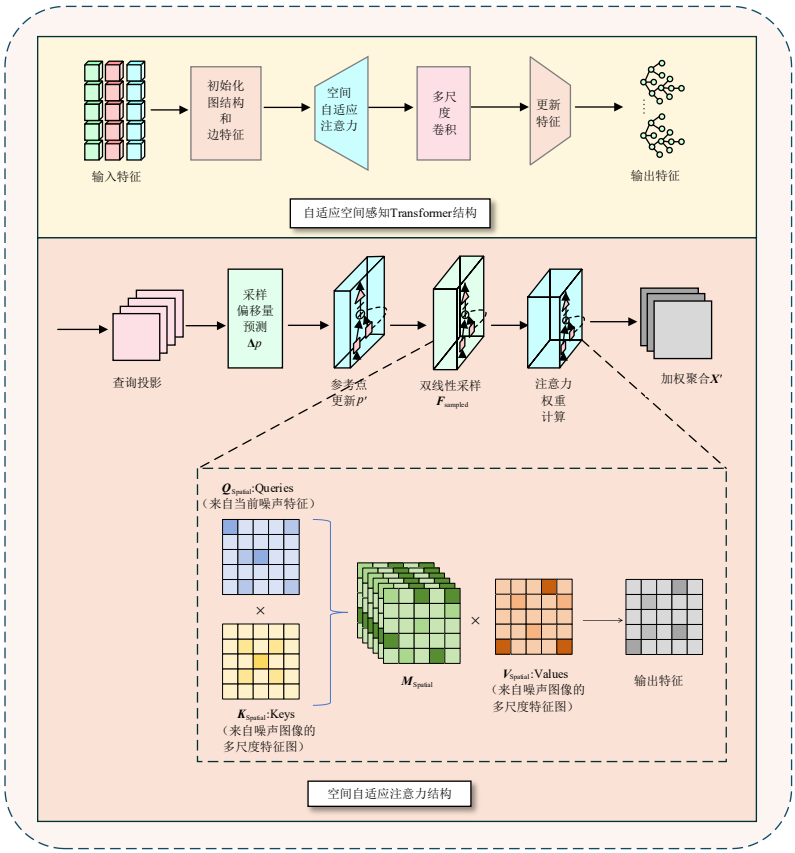


图3 自适应空间感知Transformer模型图

本文采用的空间自适应注意力机制是自适应空间感知Transformer层的重要组件,它能够根据场景中动态变化的空间信息自适应地调整注意力权重. 给定输入节点特征$\boldsymbol{X}\in\mathbf{R}^{N\times d}$,$N$为节点数量,$d$为特征维度. 空间感知Transformer首先通过空间自适应注意力机制计算查询特征、键特征和值特征,如式(5)所示:

$$\begin{cases}\boldsymbol{Q}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{Q}}\\ \boldsymbol{K}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{K}}\\ \boldsymbol{V}=\boldsymbol{X}\boldsymbol{W}_{\boldsymbol{V}}\end{cases}\tag{5}$$

其中,$\boldsymbol{W}_{\boldsymbol{Q}}$、$\boldsymbol{W}_{\boldsymbol{K}}$、$\boldsymbol{W}_{\boldsymbol{V}}\in\mathbf{R}^{d\times d}$为可学习的权重矩阵.

随后,基于查询特征和参考点p学习采样偏移量Δp,模型根据Δp计算实际的采样坐标p',如式(6)所示,p为初始规则网格参考点. 从而使得模型能够根据内容动态调整采样位置,更精确地关注重要的空间区域,即

$$p'=p+\Delta p\tag{6}$$

在采样阶段,该模块采用双线性插值在连续空间中精确提取局部特征,以解决离散网格难以对齐不规则运动模式的问题. 多尺度特征图由输入节点特征经过不同卷积核尺寸的多尺度卷积映射而来,从而表征不同感受野范围内的空间上下文信息. 由于采样坐标p'通常为非整数坐标,无法直接索引离散特征图,模型在多尺度特征图上以p'为中心,在其邻域四个最近网格点处执行基于距离权重的双线性插值,从而获得跨尺度的连续采样特征$\boldsymbol{F}_{\text{sampled}}$. 随后,将来自不同尺度的插值特征融合,并结合注意力权重α进行加权聚合,得到更新表示$\boldsymbol{X}'$,如式(7)所示,其中$\boldsymbol{W}_{\text{O}}\in\mathbf{R}^{d\times d}$为输出投影矩阵. 该机制能够动态反映不同采样位置特征的重要性,从而更精确地捕获图结构中的复杂空间依赖关系.

$$\boldsymbol{X}'=\alpha\boldsymbol{F}_{\text{sampled}}\boldsymbol{W}_{\text{O}}\tag{7}$$

为增强特征表达能力,该层进一步引入卷积增强模块$\boldsymbol{X}_{\text{conv}}$和空间卷积模块$\boldsymbol{X}_{\text{spatial}}$,节点特征卷积模块分别捕获节点特征交互与局部空间模式. 最终,模型通过残差连接和归一化得到输出$\boldsymbol{X}_{\text{out}}$,如式(8)所示:

$$\boldsymbol{X}_{\text{out}}=\text{LayerNorm}\left(\boldsymbol{X}+\boldsymbol{X}_{\text{conv}}+\boldsymbol{X}_{\text{spatial}}+\boldsymbol{X}'\right)\tag{8}$$

因此,自适应空间感知Transformer能够自适应整合跨尺度空间依赖,实现多粒度信息捕获并兼顾计算效率. 同时,该模块包含边特征更新机制,用于动态建模节点间关系. 基于此,结合全局自适应注意力与空间卷积操作,模型能够有效整合局部拓扑信息与全局空间依赖,从而赋予GCAT在复杂图结构数据上更强的特征学习与预测能力.

3.3　多粒度时序特征融合框架

本文提出的多粒度时序特征融合框架如图4所示. 该框架采用分层多尺度结构,能够同时捕捉局部粒度交互与全局群体动态,从而突破传统方法在非线性、时变与不确定交互关系建模中的局限.

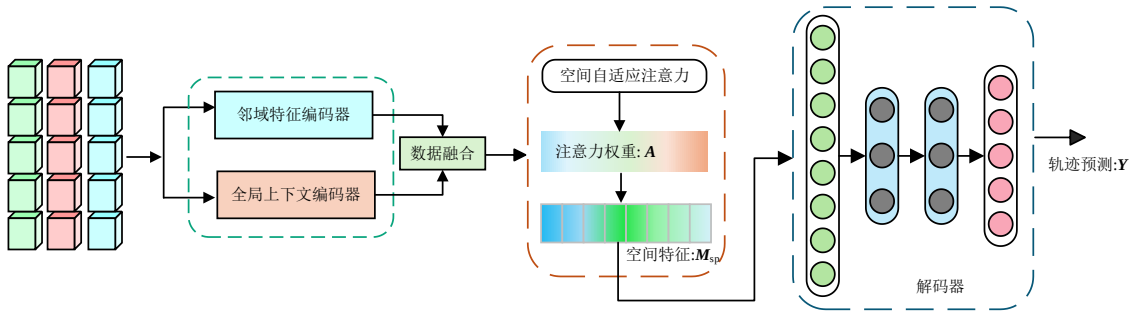


图4　多粒度时序特征融合框架图

该框架首先将输入轨迹序列$\boldsymbol{X}=\left\{x_t^i\right\}_{t=1}^{T}$映射到高维特征空间,通过特征嵌入与位置编码获得初始表示,如式(9)所示:

$$\boldsymbol{H}_0=\text{Embed}(\boldsymbol{X})+\text{PE}(\boldsymbol{X})\tag{9}$$

其中,特征嵌入函数Embed(·)提取速度、加速度等运动语义信息,位置编码函数PE(·)注入时间顺序依赖.

随后,采用双尺度并行处理策略全面捕捉行人间的复杂交互. 局部特征编码器捕捉细粒度交互关系$\boldsymbol{H}_{\text{local}}$;全局特征编码器建模群体层面的动态模式$\boldsymbol{H}_{\text{global}}$. 两个尺度的特征表示进行融合,如式(10)所示. 然后通过双线性插值进一步增强特征表达能力,动态调控感受野范围.

$$\boldsymbol{H}_{\text{fused}}=\text{Concat}\left(\left[\boldsymbol{H}_{\text{local}}+\boldsymbol{H}_{\text{global}}\right]\right)\tag{10}$$

在预测阶段,框架利用多头解码器并行生成多模态轨迹分布,从而有效应对未来轨迹的不确定性与多样性. 通过级联式特征融合策略,框架实现了从粗粒度到细粒度的渐进式特征提取,最终生成包含丰富时空信息的轨迹表示.

3.4　损失函数

假设对于每个样本,模型输入为历史轨迹$\boldsymbol{X}$,输出为未来轨迹的预测值,真实未来轨迹为$\boldsymbol{Y}$. 对于每一帧的每个行人,损失函数定义如式(11)所示:

$$L=\frac{1}{B\times N}\sum_{b=1}^{B}\sum_{n=1}^{N}\min_{k}\left(\frac{1}{T}\sum_{t=1}^{T}\left(\hat{\boldsymbol{y}}_{b,n,k,t}-\boldsymbol{y}_{b,n,t}\right)_2\right) \quad (11)$$

其中，$\hat{\boldsymbol{y}}_{b,n,k,t}$为第$b$个样本中第$n$个行人在第$k$条预测轨迹的第$t$帧的预测位置，$\boldsymbol{y}_{b,n,t}$为对应的真实位置. 这种损失函数能够有效地反映模型在轨迹预测任务中的性能，尤其适用于多模态预测场景，有助于模型学习到更为准确和多样化的轨迹分布.

4 结果与分析

4.1 数据集及实验设置细节

(1)数据集. 为了评估模型，实验在2个开放数据集ETH和UCY上进行了验证. 如表1所示，这2个数据集包括5个室外拍摄的鸟瞰场景，共2 206条行人轨迹.

(2)实验设置. 对于所有数据集的训练过程，本文均采用Adam优化器[21]，其初始学习率设置为0.001. 为防止模型过拟合并加速收敛，本文在每训练100个epoch后，将学习率乘以0.5进行衰减. 该网络使用NVIDIA GeForce RTX 5000进行训练，所有实现都是基于PyTorch 2.1.0[22].

表1 ETH/UCY数据集

数据集	场景	帧数	人数	分组数	障碍物数
ETH	ETH	1 448	360	243	44
	HOTEL	1 168	390	623	25
UCY	UNIV	541	434	297	16
	ZARA1	866	148	91	34
	ZARA2	1 052	204	140	34

(3)评估指标. 为评估轨迹预测性能，本文采用2个指标——平均位移误差(Average Displacement Error，ADE)和最终位移误差(Final Displacement Error，FDE). ADE是指预测轨迹与真实轨迹之间每个时间步的平均欧几里得距离，它反映了预测轨迹在整个预测时间段内的平均误差. FDE是指预测轨迹的最终位置与真实轨迹的最终位置之间的欧几里得距离，它反映了预测轨迹在最终时刻的误差. 具体计算如式(12)和式(13)所示：

$$\mathrm{ADE}=\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\left(\boldsymbol{p}_t^{(i)}-\hat{\boldsymbol{p}}_t^{(i)}\right) \quad (12)$$

$$\mathrm{FDE}=\frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{p}_T^{(i)}-\hat{\boldsymbol{p}}_T^{(i)}\right) \quad (13)$$

4.2 实验结果比较

为了综合评估所提GCAT模型的性能，本文在2个公开的基准数据集ETH和UCY上进行了一系列详尽的对比实验，如表2所示.

表2 在ETH和UCY数据集上的实验结果比较 单位：m

方法	年份	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Social-STGCNN	2020	0.64/1.11	0.64/1.11	0.64/1.11	0.64/1.11	0.64/1.11	0.64/1.11
STAR	2020	0.36/0.65	0.17/0.36	0.31/0.62	0.26/0.55	0.22/0.46	0.26/0.53
AST-GNN[23]	2021	0.69/1.27	0.36/0.62	0.46/0.83	0.32/0.53	0.28/0.44	0.40/0.66
GCH-GAT[24]	2022	0.63/1.10	0.38/0.73	0.55/1.16	0.33/0.66	0.30/0.64	0.44/0.86
PTP-STGCN[25]	2022	0.63/1.04	0.34/0.45	0.48/0.87	0.37/0.61	0.30/0.46	0.42/0.68
EvoSTGAT[26]	2022	0.64/1.19	0.35/0.51	0.44/0.82	0.31/0.50	0.28/0.47	0.41/0.70
SRAI-LSTM[27]	2022	0.32/0.59	0.18/0.34	0.35/0.72	0.24/0.51	0.23/0.50	0.26/0.53
SKGACN[28]	2023	0.55/0.83	0.30/0.50	0.39/0.75	0.30/0.51	0.26/0.45	0.36/0.61
SocialTAG[29]	2023	0.61/1.00	0.37/0.56	0.51/0.87	0.33/0.50	0.30/0.49	0.42/0.68
MRGTraj[30]	2024	**0.28**/0.47	0.21/0.39	0.33/0.60	0.24/0.44	0.22/0.41	0.26/0.46
SMEMO[31]	2024	0.39/0.59	**0.14/0.20**	**0.23/0.41**	0.19/0.32	0.15/0.26	0.22/0.35
WTGCN[32]	2024	0.60/0.95	0.25/0.37	0.36/0.65	0.27/0.46	0.23/0.39	0.34/0.56
DSTIGCN[33]	2025	0.43/0.70	0.22/0.41	0.25/0.45	0.20/0.37	0.17/0.32	0.25/0.45
GCAT	—	0.38/**0.50**	**0.14**/0.23	0.25/0.45	**0.16/0.29**	**0.14/0.23**	**0.21/0.33**

注：加粗表示最优结果，下划线表示次优结果，数值表示最小ADE/最小FDE.

从平均预测误差(Average Prediction Error，AVG)指标可以看出，与近两年提出的强基线模型相比，GCAT展现出了更为全面的性能优势. 以SMEMO模型为例，虽然该模型通过记忆模块增强了预测的稳定性，但在ETH等稀疏场景中，其基于历史记忆的检索机制可能难以即时响应突发的运动变化. 相较之下，GCAT通过基于特征相似度的自适应邻接矩阵，在每个时间步动态重构交互拓扑，不依赖历史记忆即可即时反映当前行为模式. 在ETH数据集中，GCAT在ADE/FDE上分别较SMEMO提升了2.5%与15.2%，验证了动态拓扑建模在非平稳交互场景中的有效性.

同时，相较于MRGTraj模型，尽管其多模态回归与残差建模在复杂交互中表现不俗，但在UNIV、ZARA1等高密度拥挤场景中仍显现出一定的累积误差. GCAT

在UNIV和ZARA1场景中实现了0.08 m/0.15 m的ADE/FDE改进. 这一优势主要归因于模型的自适应邻接矩阵与跨尺度空间感知机制,前者能够动态重构拥挤人群中的关键交互拓扑,后者则通过细粒度的连续特征采样,修正了密集交互下的微小轨迹偏差,从而在全局趋势跟踪和局部避让预测上均实现了更高的精度.

综上所述,GCAT不仅在统计指标上优于SMEMO和MRGTraj等最新基线,更重要的是,它通过解决动态交互的实时性与空间感知的连续性这两个关键难题,证明了其核心设计在应对多样化复杂场景时的有效性与泛化能力.

此外,GCAT在模型轻量化和计算效率方面也具有突出优势,如表3所示. 与参数量动辄数百万甚至上千万的复杂模型相比,GCAT的参数量仅为0.91×10^6,实现了超过90%的参数削减. 在计算复杂度方面,其19.0×10^6的MAC远低于PECNet和GroupNet等方法,甚至与MID*相比,在计算效率上提升了3个数量级. 这种卓越的效率并非以牺牲性能为代价,与参数量相近的STAR相比,GCAT在平均ADE和FDE上分别提升了19.2%和37.7%.

表3　所提出方法与最新方法的计算效率对比

方法	参数量	计算量
PECNet	2.1×10^6	259.2×10^6
STAR	1.0×10^6	12.0×10^9
MemoNet[34]	10.7×10^6	6.0×10^9
GroupNet[35]	2.2×10^6	411.5×10^6
MID*[36]	9.0×10^6	40.3×10^9
EqMotion[37]	3.0×10^6	147.1×10^6
GCAT	$\mathbf{0.9\times10^6}$	$\mathbf{19.0\times10^6}$

注:加粗表示最优结果.

从表4可以看出,GCAT的推理时间仅为0.009 42 s,在所有对比方法中实现了最优的推理速度. 尽管WTGCN和GADG等图卷积方法同样具备较高的推理效率,GCAT仍保持性能领先优势. 这一结果充分体现了GCAT所采用的轻量级多尺度时序建模架构的设计优势,通过高效的空间自适应注意力机制和优化的网络结构,成功实现了预测精度与计算效率的最佳平衡. 相较于传统的基于长短期记忆网络(Long Short-Term Memory,LSTM)的方法,如SRAI-LSTM和基于注意力机制的复杂模型,GCAT在保证高精度轨迹预测的同时,显著降低了计算开销,这对于实时应用场景,特别是在自动驾驶和智能监控等对响应时间敏感的领域中具有较高的应用价值.

这些结果充分证明了本文所提出的图卷积与自适应Transformer融合架构的有效性,该架构能够自适应建模复杂的时空依赖关系,兼顾局部交互与全局动态,从而实现精准的轨迹预测.

表4　推断时间比较

方法	推断时间/s
NMMP[38]	0.015 26
STAR	0.027 12
Introver[39]	0.120 00
GADG[40]	0.012 70
SRAI-LSTM	0.019 00
WTGCN	0.011 60
GCAT	**0.009 42**

注:加粗表示最优结果.

4.3　模块间的消融实验

为深入验证模型中各核心组件的有效性,本文进行了一系列的消融研究,旨在定量分析不同模块对模型整体性能的贡献. 为了确定模型的最佳超参数配置,本文评估了GCN的堆叠层数(layers)以及训练轮数(epochs)对预测性能的影响,实验结果如表5所示. 首先,从GCN层数的对比可以看出,当层数设置为2时,模型在ADE和FDE指标上均达到最佳表现;当层数较浅时,模型对高阶邻域信息的提取能力受限,预测精度有所下降;而当层数增加至3层或5层时,性能不升反降,这可能由于深层GCN易出现的过平滑现象,使节点特征趋于同质化,从而削弱不同行人之间的判别性. 其次,在训练轮数方面,实验结果显示随着epochs从200增加至300,模型性能稳步提升并在300轮附近收敛;继续训练至350轮并未带来进一步的性能改善,反而增加了额外的训练开销. 因此,本文最终采用2层GCN结构并将训练轮数设为300,以在预测精度、稳定性与计算成本之间实现最优平衡.

表5　GCN层数(layers)、训练轮数(epochs)的消融研究

评估指标	GCN层数(epochs=300)				训练轮数(layers=2)			
	1	2	3	5	200	250	300	350
ADE	0.22	**0.21**	0.23	0.22	0.23	0.22	**0.21**	0.21
FDE	0.34	**0.33**	0.33	0.35	0.36	0.35	**0.33**	0.33

注:加粗表示最优结果.

在表6的模块消融实验中,首先评估了单独使用自适应空间感知Transformer模块的性能. 该变体虽能捕捉时序依赖,但由于缺乏对行人间空间关系的显式建模,其平均性能显著低于完整模型,证明了GCN在复杂场景中捕捉局部交互结构的重要性. 随后,仅使用关系型GCN模块的变体性能出现明显下降,表明准确建模长期时序动态对轨迹预测结果具有重要影响. 而本文提出的模型通过有效融合GCN的空间感知能力与Transformer的时序建模优势,取得了最佳平均ADE/FDE性能,证实二者对于预测任务是互补且不可或缺的.

表6 各个模块的消融实验 单位:m

关系型GCN	自适应空间感知Transformer	GCAT	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
—	√	—	0.38/0.52	0.16/0.26	0.27/0.50	0.18/0.31	0.13/0.22	0.22/0.36
√	—	—	0.41/0.55	0.32/0.59	0.34/0.61	0.17/0.29	0.25/0.41	0.29/0.49
√	√	√	0.38/0.50	0.15/0.23	0.25/0.45	0.17/0.29	0.14/0.23	0.21/0.33

注:数值表示最小ADE/最小FDE.

为验证所提出模型框架的计算高效性,对其中2个核心模块的计算复杂度进行了量化分析. 如表7所示,通过参数量和计算量这2个关键指标来评估其计算成本.

表7 关系型GCN与自适应空间感知Transformer的计算效率消融实验

模块	参数量	计算量
关系型GCN	0.1×10^6	0.1×10^6
自适应空间感知Transformer	0.3×10^6	6.1×10^6

实验结果充分证明了GCAT模型设计的轻量化与高效性,使其能够在有效捕捉复杂空间交互关系的同时,仅产生极低的计算开销. 这对于模型的整体效率和在资源受限环境下的部署尤为重要,也验证了所提模型架构设计的合理性.

4.4 实验过程对比

图5展示了GCAT模型在ETH、HOTEL、UNIV、ZARA1和ZARA2这5个基准数据集上预测未来12时间步(4.8 s)的平均位移误差变化趋势. 可以看出,所有场景中预测误差均随时间步增加而稳定上升,符合预测时长增加导致不确定性累积的普遍规律. 在预测的初始阶段,不同数据集上的误差反映了不同场景行人运动的复杂性,而后期阶段误差曲线逐渐收敛,表明模型在长时预测任务中具有良好的泛化能力. 图6对比了相同实验环境下GCAT与MRGTraj的训练过程. 结果显示,GCAT在ADE和FDE两项指标上均具有明显优势,训练曲线也在70%进度后趋于稳定,收敛性良好;而MRGTraj虽持续下降,但其绝对性能始终低于GCAT.

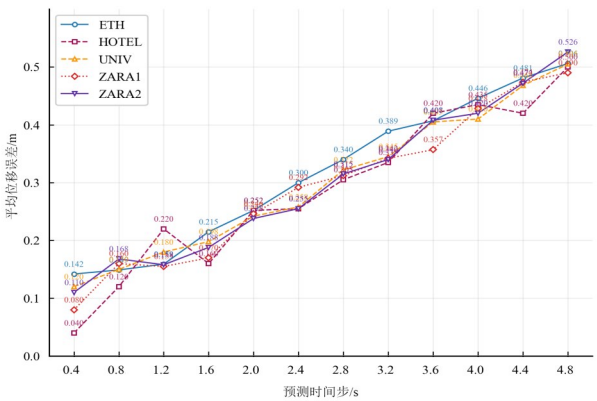


图5 GCAT模型在不同时间步的位移误差变化

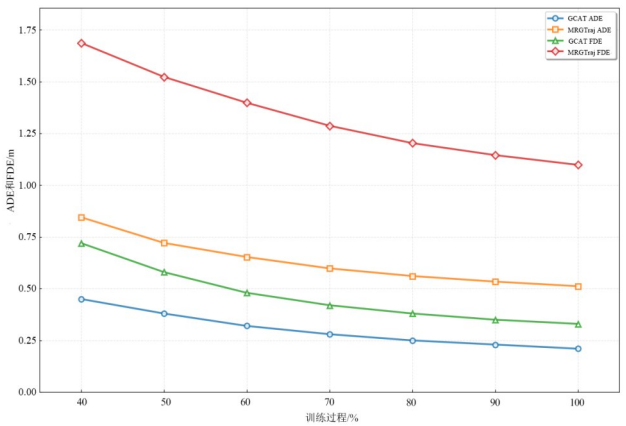


图6 训练过程分析

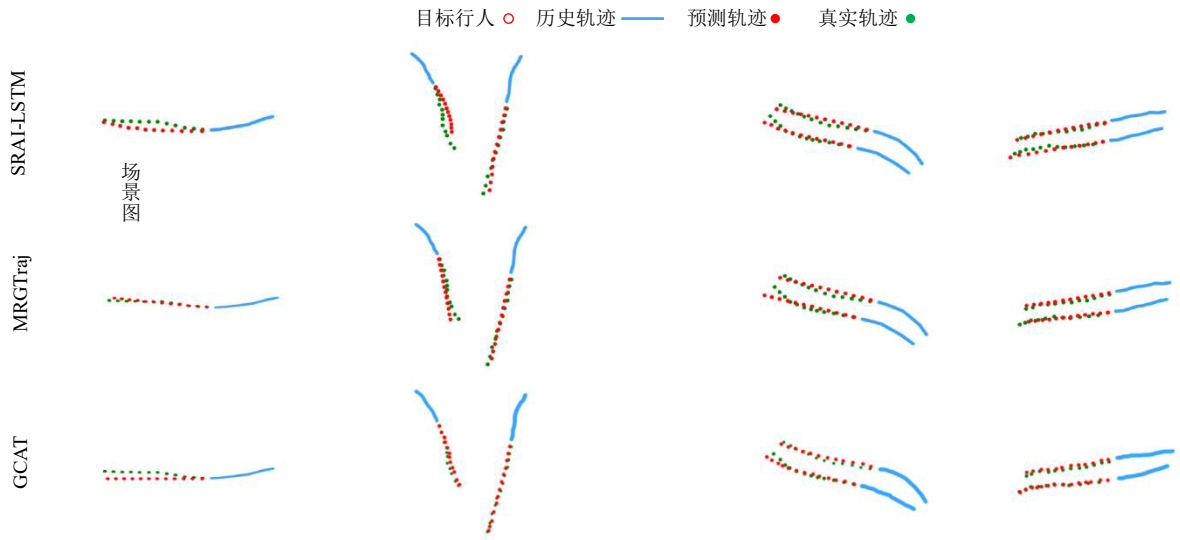
4.5 轨迹可视化

为避免仅凭视觉判断带来的主观性,结合图7所示样例对应场景的平均误差(ADE/FDE)进行了量化分析. 如表2所示,在ETH场景中,GCAT的ADE/FDE为0.38 m/0.50 m,其误差水平优于SRAI-LSTM(0.32 m/0.59 m)和MRGTraj(0.28 m/0.47 m);在UNIV场景中,GCAT的ADE/FDE为0.25 m/0.45 m,同样低于SRAI-LSTM(0.35 m/0.72 m)和MRGTraj(0.33 m/0.60 m). 量化结果与图7中的可视化趋势一致,表明GCAT不仅在预测轨迹特征更接近真实分布,其数值误差也更低,从而验证了模型预测精度的提升.

在ETH、HOTEL、UNIV和ZARA1这4个数据集的可视化样例中(图7),可进一步观察到各模型在不同场景下的预测差异. 在ETH场景(第2列)中,MRGTraj和SRAI-LSTM的预测轨迹与真实轨迹在整体走向和终点位置上存在明显偏移;而在UNIV场景(第4列)中,2种基线方法的轨迹随时间累积偏差更为突出. 相比之下,GCAT在运动轨迹的整体趋势、运动方向变化及目标位置等关键轨迹特征上更贴近真实轨迹,体现了模型在复杂场景中的预测稳定性.

因此,与MRGTraj和SRAI-LSTM等依赖固定图结构或循环网络的方法相比,GCAT的多尺度设计和动态注意力机制使其在处理人群密集、交互复杂的场景时表现出更高的预测准确率,从而生成更贴近真实的轨迹.

(*a*) 展示场景原图,场景来自ZARA1和ETH数据集



(*b*) 分别对应SRAI-LSTM、MRGTraj与GCAT模型的预测结果可视化

图7　SRAI-LSTM、MRGTraj与GCAT这3种模型预测结果的可视化对比

5　结论

本文提出了一种基于图卷积与自适应Transformer的行人轨迹预测方法,通过关系图卷积网络建模局部交互,并利用多尺度时空特征提取机制增强Transformer的全局动态建模能力,从而在局部与全局2个层面有效提升对复杂场景中行人运动模式的建模能力.通过在ETH、UCY行人轨迹数据集上的评估,所提出的方法在多个场景中取得了较好结果,平均ADE/FDE为0.21 m/0.33 m,验证了模型在行人轨迹预测任务中的有效性与先进性.

本文方法主要针对模型在行人场景下的性能,未来工作将致力于将该方法扩展至更广泛的智能体类型,包括无人车辆、机器人集群等动态系统,以验证模型的跨主体泛化能力.同时,我们计划引入更丰富的环境语义信息与上下文约束,以进一步提升模型在复杂真实场景中的预测稳定性与实用性,为自动驾驶、智能监控和人机交互等应用提供更加可靠的轨迹预测支持.

参考文献

[1] DEO N, TRIVEDI M M. Convolutional social pooling for vehicle trajectory prediction[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Piscataway: IEEE, 2018: 1549-15498.

[2] IVANOVIC B, PAVONE M. The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2019: 2375-2384.

[3] ROBICQUET A, SADEGHIAN A, ALAHI A, et al. Learning social etiquette: Human trajectory understanding in crowded scenes[C]//Computer Vision - ECCV 2016. Cham: Springer, 2016: 549-565.

[4] LEFÈVRE S, VASQUEZ D, LAUGIER C. A survey on motion prediction and risk assessment for intelligent vehicles[J]. ROBOMECH Journal, 2014, 1(1): 1.

[5] SHI L S, WANG L, LONG C J, et al. SGCN: Sparse graph convolution network for pedestrian trajectory prediction[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 8990-8999.

[6] 胡春华, 曾萼岚, 荣辉桂. 基于双图卷积机制的数字孪生交通流预测[J]. 电子学报, 2025, 53(1): 141-150.
HU C H, ZENG E L, RONG H G. Traffic flow prediction of digital twin based on two-graph convolution mechanism[J]. Acta Electronica Sinica, 2025, 53(1): 141-150. (in Chinese)

[7] 袁丁, 李源, 孟羽倩, 等. 基于时空注意力Transformer的自动驾驶运动规划方法[J]. 电子学报, 2025, 53(7): 2418-2427.
YUAN D, LI Y, MENG Y Q, et al. A motion planning method for autonomous driving based on spatiotemporal attention transformer[J]. Acta Electronica Sinica, 2025, 53(7): 2418-2427. (in Chinese)

[8] GIULIARI F, HASAN I, CRISTANI M, et al. Transformer networks for trajectory forecasting[C]//2020 25th International Conference on Pattern Recognition. Piscataway: IEEE, 2021: 10335-10342.

[9] GUPTA A, JOHNSON J, LI F F, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2018: 2255-2264.

[10] HUANG Y F, BI H K, LI Z X, et al. STGAT: Modeling spatial-temporal interactions for human trajectory prediction[C]//2019 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2020: 6271-6280.

[11] MOHAMED A, QIAN K, ELHOSEINY M, et al. Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 14412-14420.

[12] LI J C, YANG F, TOMIZUKA M, et al. EvolveGraph: Multi-agent trajectory prediction with dynamic relational reasoning[EB/OL]. (2020-10-22) [2025-10-10]. https://arXiv.org/abs/2003.13924.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Kerrville: Association for Computational Linguistics, 2019: 4171-4186.

[14] SU Z X, HUANG G, ZHANG S Y, et al. Crossmodal transformer based generative framework for pedestrian trajectory prediction[C]//2022 International Conference on Robotics and Automation. Piscataway: IEEE, 2022: 2337-2343.

[15] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[EB/OL]. (2018)[2025-10-10]. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf.

[16] WANG A, SINGH A, MICHAEL J, et al. GLUE: A multi-task benchmark and analysis platform for natural language understanding[EB/OL]. (2019-02-22) [2025-10-10]. https://arXiv.org/abs/1804.07461.

[17] YAO H Y, WAN W G, LI X. End-to-end pedestrian trajectory forecasting with transformer network[J]. ISPRS International Journal of Geo-Information, 2022, 11(1): 44.

[18] YU C J, MA X, REN J W, et al. Spatio-temporal graph transformer networks for pedestrian trajectory prediction[C]// Computer Vision - ECCV 2020. Cham: Springer, 2020: 507-523.

[19] MANGALAM K, GIRASE H, AGARWAL S, et al. It is not the journey but the destination: Endpoint conditioned trajectory prediction[C]//Computer Vision - ECCV 2020. Cham: Springer, 2020: 759-776.

[20] YUAN Y, WENG X S, OU Y L, et al. AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting[C]//2021 IEEE/CVF International Conference on Computer Vision. Piscataway: IEEE, 2022: 9793-9803.

[21] KINGA D, ADAM J B. A method for stochastic optimization[C]//International Conference on Learning Representations(ICLR). Appleton: ICLR, 2015: 50478691.

[22] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An imperative style, high-performance deep learning library[EB/OL]. (2019-12-03)[2025-10-10]. https://arXiv.org/abs/1912.01703.

[23] ZHOU H, REN D C, XIA H X, et al. AST-GNN: An attention-based spatio-temporal graph neural network for Interaction-aware pedestrian trajectory prediction[J]. Neurocomputing, 2021, 445: 298-308.

[24] ZHOU L, ZHAO Y L, YANG D Y, et al. GCHGAT: Pedestrian trajectory prediction using group constrained hierarchical graph attention networks[J]. Applied Intelligence, 2022, 52(10): 11434-11447.

[25] LIAN J, REN W W, LI L H, et al. PTP-STGCN: Pedestrian trajectory prediction based on a spatio-temporal graph convolutional neural network[J]. Applied Intelligence, 2023, 53(3): 2862-2878.

[26] TANG H W, WEI P, LI J P, et al. EvoSTGAT: Evolving spatiotemporal graph attention networks for pedestrian trajectory prediction[J]. Neurocomputing, 2022, 491: 333-342.

[27] PENG Y S, ZHANG G F, SHI J, et al. SRAI-LSTM: A social relation attention-based interaction-aware LSTM for human trajectory prediction[J]. Neurocomputing, 2022, 490: 258-268.

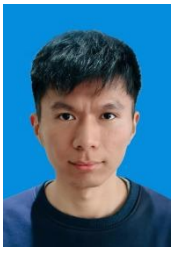
[28] LV K, YUAN L. SKGACN: Social knowledge-guided graph attention convolutional network for human trajectory prediction[J]. IEEE Transactions on Instrumentation and Measurement, 2023, 72: 2517111.

[29] ZHANG X C, ANGELOUDIS P, DEMIRIS Y. Dual-branch spatio-temporal graph neural networks for pedestrian trajectory prediction[J]. Pattern Recognition, 2023,

142: 109633.
[30] PENG Y S, ZHANG G F, SHI J, et al. MRGTraj: A novel non-autoregressive approach for human trajectory prediction[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(4): 2318-2331.
[31] MARCHETTI F, BECATTINI F, SEIDENARI L, et al. SMEMO: Social memory for trajectory forecasting[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, 46(6): 4410-4425.
[32] CHEN W X, SANG H F, WANG J Y, et al. DSTIGCN: Deformable spatial-temporal interaction graph convolution network for pedestrian trajectory prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(5): 6923-6935.
[33] CHEN W X, SANG H F, WANG J Y, et al. WTGCN: Wavelet transform graph convolution network for pedestrian trajectory prediction[J]. International Journal of Machine Learning and Cybernetics, 2024, 15(12): 5531-5548.
[34] XU C X, MAO W B, ZHANG W J, et al. Remember intentions: Retrospective-memory-based trajectory prediction[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 6478-6487.
[35] XU C X, LI M S, NI Z Y, et al. GroupNet: Multiscale hypergraph neural networks for trajectory prediction with relational reasoning[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 6488-6497.
[36] GU T P, CHEN G Y, LI J L, et al. Stochastic trajectory prediction via motion indeterminacy diffusion[C]//2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2022: 17092-17101.
[37] XU C X, TAN R T, TAN Y H, et al. EqMotion: Equivariant multi-agent motion prediction with invariant interaction reasoning[C]//2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2023: 1410-1420.
[38] HU Y, CHEN S H, ZHANG Y, et al. Collaborative motion prediction via neural motion message passing[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2020: 6318-6327.
[39] SHAFIEE N, PADIR T, ELHAMIFAR E. Introvert: Human trajectory prediction via conditional 3D attention[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE, 2021: 16810-16820.
[40] 孔玮, 刘云, 李辉, 等. 基于全局自适应有向图的行人轨迹预测[J]. 电子学报, 2022, 50(8): 1905-1916.
KONG W, LIU Y, LI H, et al. Pedestrian trajectory prediction based on global adaptive directed graph[J]. Acta Electronica Sinica, 2022, 50(8): 1905-1916. (in Chinese)

作者简介

王芳芳　女,1997年11月出生于山东省潍坊市. 现为青岛科技大学信息科学技术学院硕士研究生. 主要研究方向为计算机视觉、轨迹预测.
E-mail: 2024111009@mails.qust.edu.cn



王　贺　男,1999年10月出生于山东省菏泽市. 现为青岛科技大学信息科学技术学院硕士研究生. 主要研究方向为计算机视觉、语音识别.
E-mail: 19854299311@163.com

刘明华　男,1980年2月出生于山东省聊城市. 现为青岛科技大学信息科学技术学院教授. 主要研究方向为计算机视觉、目标识别与跟踪、视频图像理解与分割等.
E-mail: qustlmh@qust.edu.cn

李丹宁　女,2000年9月出生于山东省泰安市. 现为青岛科技大学信息科学技术学院硕士研究生. 主要研究方向为计算机视觉、时间序列.
E-mail: 15666254825@163.com

渠连恩　男,1980年12月出生于山东省聊城市. 现为青岛科技大学信息科学技术学院教授. 主要研究方向为计算机视觉、智能交通、天气预测等.
E-mail: lianen.qu@qust.edu.cn