计算机科学
COMPUTER SCIENCE

构建场景-行人-行人交互的行人轨迹预测时空图卷积网络

洪铭骏, 纪庆革

引用本文

洪铭骏, 纪庆革. 构建场景-行人-行人交互的行人轨迹预测时空图卷积网络[J]. 计算机科学, 2025, 52(12): 133-140.

HONG Mingjun, JI Qingge. SPP-STGCN:Spatio-Temporal Graph Convolutional Network for Pedestrian Trajectory Predictionwith Scene-Perdestrian-Perdestrain Interactions [J]. Computer Science, 2025, 52(12): 133-140.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.241200212

构建场景-行人-行人交互的行人轨迹预测时空图卷积网络

洪铭骏　纪庆革
中山大学计算机学院　广州 510006
(hongmj6@mail2.sysu.edu.cn)

摘　要　行人轨迹预测是自动驾驶和智能监控系统中的一项基础而关键的任务。场景的限制是影响行人运动轨迹的重要因素之一。尽管现有的研究已经尝试将场景因素融入轨迹预测中,但这些方法在整合场景信息时往往存在不足,尤其是没有考虑到场景的全面融合。对此,提出了一种新的行人轨迹预测模型——构建场景-行人-行人交互的时空图卷积网络 SPP-STGCN(Spatio-Temporal Graph Convolutional Network with Scene-Perdestrian-Perdestrain Interactions)。SPP-STGCN 模型采用两阶段架构来提高预测准确性。第一阶段,模型将轨迹和场景数据整合输入,通过场景邻接矩阵融合块 SAFB 实现两种特征的融合,构建出融合场景特征的时空图邻接矩阵,为预测提供丰富的上下文信息。同时,模型在时间、空间两个维度并行执行,结合轨迹信息分别构造出描述时间相关性的和空间相关性的行人轨迹时空图。第二阶段,场景图卷积网络对时间与空间维度的时空图进行特征提取。提取的特征随后被融合,并通过时间金字塔外推卷积进行处理,以获得行人未来轨迹的二维高斯分布。最后 SPP-STGCN 以该分布作为预测行人轨迹的概率模型,通过采样生成行人未来轨迹。在 ETH 和 UCY 公开数据集上的对比实验结果显示,SPP-STGCN 模型在与 9 种主流模型的对比实验中的表现达到了目前的最佳水平。消融实验与定性分析进一步证实了所提模型的有效性与合理性。SPP-STGCN 模型通过有效整合场景特征,显著提升了行人轨迹预测的性能。

关键词:行人轨迹预测;场景-行人-行人交互;时空图卷积;邻接矩阵生成;注意力机制
中图分类号　TP391

SPP-STGCN:Spatio-Temporal Graph Convolutional Network for Pedestrian Trajectory Prediction with Scene-Perdestrian-Perdestrain Interactions

HONG Mingjun and JI Qingge
School of Computer Science and Engineering,Sun Yat-sen University,Guangzhou 510006,China

Abstract　Pedestrian trajectory prediction is a fundamental and critical task in autonomous driving and intelligent surveillance systems. The constraints of the scene are one of the important factors affecting pedestrian movement trajectories. Despite existing research efforts to incorporate scene factors into trajectory prediction,these methods often fall short in integrating scene information,particularly in terms of comprehensive scene fusion. To overcome these limitations,this study proposes a new pedestrian trajectory prediction model,namely SPP-STGCN. The SPP-STGCN model adopts a two-stage architecture to enhance prediction accuracy. In the first stage,the model integrates trajectory and scene data. Through the Scene Adjacency Fusion Block(SAFB),the model fuses these two types of features to construct a spatio-temporal graph adjacency matrix that incorporates scene features,thereby providing rich contextual information for prediction. Concurrently,the model operates in parallel along the temporal and spatial dimensions,constructing pedestrian trajectory spatio-temporal graphs that describe temporal and spatial correlations based on trajectory information. In the second stage,scene-graph convolutional networks extract features from the temporal and spatial spatio-temporal graphs. The extracted features are then fused and processed through a temporal pyramid extrapolation convolution to obtain the two-dimensional Gaussian distribution of the pedestrian's future trajectory. Finally,SPP-STGCN uses this distribution as a probabilistic model for predicting pedestrian trajectories,generating future trajectories through sampling. Comparative experimental results on the ETH and UCY public datasets show that the SPP-STGCN model has achieved the current state-of-the-art performance in comparison experiments with nine mainstream models. Ablation experiments and qualitative analysis further confirm the effectiveness and rationality of the proposed model. The SPP-STGCN model significantly enhances pedestrian trajectory prediction performance by effectively integrating scene features.

Keywords　Pedestrian trajectory prediction,Scene-pedestrian-pedestrian interaction,Spatio-temporal graph convolution,Adjacency matrix generation,Attention mechanism

到稿日期:2024-12-30　返修日期:2025-05-09
基金项目:广东省自然科学基金(2016A030313288)
This work was supported by the Natural Science Foundation of Guangdong Province,China(2016A030313288).
通信作者:纪庆革(issjqg@mail.sysu.edu.cn)

1 引言

行人轨迹预测在自动驾驶、机器人以及安全监控等领域都是十分重要的任务。无论是在自动驾驶和机器人领域中防止智能体与行人发生碰撞，还是在安全监控中监控行人的异常运动行为，对异常事件的发生做出预警，行人轨迹预测问题都是一个核心且具有挑战性的问题。

本文考虑的是固定视角下，对行人进行未来轨迹预测的问题。在有众多行人的复杂场景下，行人之间会因多种社交关系的存在而调整自己的运动趋势。例如，迎面的两位行人会因为防止碰撞而改变自身的运动方向，造成轨迹偏移原来的趋势。两位结伴而行的行人则会保持着一个紧密的距离和相同的运动趋势。这种交互模式被称为社会交互，如何处理复杂的社会交互是行人轨迹预测中的核心问题之一。

场景因素同样是影响行人轨迹的重要因素。行人在行走的过程中，不仅会尝试躲避其他行人，也会根据场景中的障碍物调整自己的行走轨迹。在固定的视角下，场景信息也是相对固定的，这使得场景特征能够被更好地提取和应用。考虑到这些因素，本文将场景因素融入行人轨迹预测中。

行人轨迹预测早期的工作多使用对运动建模的方法，这部分方法往往更注重行人之间的交互，代表工作为社会力模型 S-F Model[1]。随着神经网络的崛起，采用数据驱动型模型的方法越来越多。社交-长短期记忆网络 Social-LSTM[2] 运用社交池化模块来模拟行人之间的社会交互。社交-时空图卷积网络 Social-STGCN[3] 则通过构建行人轨迹时空图来模拟行人之间的社会交互，而稀疏图卷积神经网络 SGCN[4] 在前者的基础上，采用了神经网络生成轨迹时空图的邻接矩阵。然而，上述神经网络模型仅从行人的轨迹数据出发来构建网络，忽略了行人在运动过程中也会受到场景中各种因素的影响。

事实上，将场景因素融合到行人轨迹预测中的神经网络模型也得到了广泛的研究。符合社会限制和物理限制的注意力生成对抗网络 Sophie[5] 利用场景图像信息作为场景上下文信息，构建物理限制来生成轨迹。然而，该工作将场景图像直接展平，破坏了场景图像的空间性质，导致场景信息与轨迹信息的融合并不充分。场景兼容轨迹预测网络 Y-Net[6] 则通过热力图将轨迹信息融入场景信息，并基于热力图进行行人轨迹预测。该工作取得了优秀的预测结果，但也使问题变得复杂化，计算成本增加。场景限制时空图卷积网络 Scene-STGCN[7] 使用场景语义图构建场景特征向量，实现了场景信息与轨迹信息的融合，但其与社交-时空图卷积网络 Social-STGCNN[3] 一样，对图邻接矩阵的处理过于简单，限制了模型的性能。

为了解决场景信息和轨迹信息难以融合的问题，同时兼顾对行人交互的充分考虑，本文提出了构建场景-行人-行人交互的时空卷积图网络 SPP-STGCN。SPP-STGCN 采用了两阶段的模型架构：第一阶段中，通过行人轨迹构造行人轨迹时空图的邻接矩阵，并利用场景邻接矩阵融合块(Scene Adjacency matrix Fusion Block，SAFB)来融合场景特征；第二阶段中，将构造出的行人轨迹时空图输入主干预测网络预测未来轨迹，并利用场景图卷积网络(Scene-Graph Convolution Network，S-GCN)融合场景特征。通过在邻接矩阵生成模块和主干预测模块都引入场景交互模块，SPP-STGCN 实现了场景特征的充分融合。

本文实验结果显示，SPP-STGCN 在与多个主流的使用图网络模型的行人轨迹预测模型的比较中性能表现最佳。在平均位移误差(ADE)和最终位移误差(FDE)两个关键指标上，SPP-STGCN 分别实现了 0.35 和 0.57 的误差值。

本文的主要贡献如下：

1)提出了一种用于行人轨迹预测的融合场景的时空图卷积网络 SPP-STGCN，构建了场景-行人-行人交互，实现了场景信息与轨迹信息的充分融合；

2)设计了场景融合模块 SAFB，是一种作用于行人轨迹图的邻接矩阵生成模块，使得场景因素能融入行人之间的交互；

3)在公开数据集 ETH 和 UCY 上的实验结果表明，SPP-STGCN 的平均性能优于最新的几种图网络模型。

2 相关工作

2.1 社会交互模型

基于神经网络的数据驱动方法凭借卓越的精度表现，在行人轨迹预测领域广受欢迎。本质上，行人轨迹数据属于序列数据的范畴，但在行人较多的场景下，行人之间不可避免地会产生相互影响，这种影响被称为社会交互。为了使模型能够同时学习行人的运动趋势和行人间的社会交互，融合时间和空间两个维度的模型已成为研究热点。Social-LSTM[2] 采用 LSTM 模型并利用社会-池化层来进行目标行人周围邻居行人特征的聚合。在 State-Refinement LSTM[8] 方法中，进一步引入了加权机制和信息传递机制，以更精细地建模邻居行人的社会影响，并实现对社会交互信息的选择性融合。PTP-STGCN[9] 则运用 Transfomer 来提取时间维度的轨迹特征并使用 GCN 聚合空间维度的轨迹信息。为了更好地学习行人之间的社会交互，相关学者将行人轨迹数据构建成行人轨迹时空图，并使用图神经网络(Graph Neural Networks，GNNs)和其衍生模型，如图卷积网络，实现对行人交互特征的提取。特别地，SocialBiGAT[10] 和 Sophie[5] 使用了图注意力网络，这种网络能够自适应地学习行人之间的相互作用权重。Social-STGCN[3] 和 PTP-STGCN[9] 构建了行人轨迹时空图的图邻接矩阵，并使用了图卷积网络 GCN[11]，通过局部邻域聚合来捕捉行人的时空依赖性。而 SGCN[4] 和 STIGCN[12] 则采用注意力机制，生成时序和空间维度的邻接矩阵，并通过图卷积网络并行提取时序和空间特征。本文借鉴了 SGCN 的方法论，但在生成邻接矩阵和提取特征的过程中引入了场景因素。通过将场景上下文纳入考虑，本文方法能够更全面地理解行人行为的复杂性，为轨迹预测提供了更为丰富和准确的信息。

2.2 多模态轨迹预测

在行人轨迹预测领域，一个核心挑战是如何妥善处理行人运动轨迹所固有的不确定性。为了应对这一挑战，研究者提出了一种方法，即通过综合考虑多种不确定性因素，生成多条符合逻辑的行人轨迹。这种预测模式被称为多模态轨迹预测[13]。例如，Social-GAN[14]，Sophie[5] 和 SocialBiGAT[10] 采用了生成对抗网络(GANs)[15] 来生成行人轨迹。GANs 通过

对抗性训练过程，能够生成符合真实数据分布的轨迹，从而提高了预测的多样性和真实感。另一方面，NSP-SFM[16]，Social-VAE[17]，CSR[18]和 Social-CVAE[19]等则采用了变分自编码器（Variational Autoencoders，VAEs）及其衍生模型来生成未来的轨迹。VAEs 通过学习数据的潜在表示，并从中采样来生成新的轨迹。这种方法能够提供对不确定性的自然量化，并允许生成多种可能的轨迹。然而，GANs 在训练过程中可能存在不稳定性，VAEs 在捕捉复杂数据分布时可能存在局限性。为了克服这些缺点，本文采用了高斯分布来生成行人的未来轨迹，以此模拟行人轨迹的不确定性，旨在实现更加稳定和准确的预测效果。

2.3 场景感知模型

场景因素在行人轨迹预测中扮演着至关重要的角色。行人在移动时会自然地规避障碍物，并受到场景中不同语义元素的社会性约束。随着研究的深入，越来越多的学者开始尝试将场景因素整合到他们提出的模型中。Scene-LSTM[20]和 Sophie[5]将场景图像展平成向量，并将这部分信息融入轨迹预测中，但这种方法可能会损失场景的空间结构信息。Y-Net[6]采用了不同的策略，它将轨迹信息嵌入场景图像中，并基于这些图像进行轨迹预测，尽管这增加了问题的复杂性。Scene-STGCN[7]和 NSP-SFM[16]则是利用场景分割图，构造场景特征向量，并将其有效地整合到轨迹预测过程中。这种方法不仅保留了场景的空间信息，而且通过精确的场景特征提取，提高了预测的准确性。本文采纳了这一策略，以期在考虑场景因素的同时，实现更优的行人轨迹预测效果。

3 本文方法

3.1 问题描述

行人轨迹预测本质上是对坐标序列的预测。本文专注于一个特定的场景，目标是基于已知的行人历史轨迹数据，预测其在未来特定时间段内的移动路径。这一问题的数学表述如下：在一个场景下，给定 N 个行人在 To 个观测帧内的轨迹坐标序列 $L_{1:To}=\{X_1,X_2,\cdots,X_{To}\}$，模型需要预测 T_p 个未来观测帧内行人的轨迹坐标序列 $L_{\widehat{To+1:To+Tp}}=\{X_{\widehat{To+1}},X_{\widehat{To+2}},\cdots,X_{\widehat{To+Tp}}\}$，其中 $X_t=\{x_1^t,x_2^t,\cdots,x_N^t\}\in R^{N\times 2}$ 表示 N 个行人在观测帧第 t 帧的二维世界坐标，对应的 $\hat{X}_t$ 则表示 N 个行人在预测帧第 t 帧的二维世界坐标。同时，为了提高轨迹预测的准确性，本文模型注重场景信息的整合。

总体上，本文的目标是：基于行人在 To 个观测帧内的轨迹数据 L_o 和场景分割图 S，预测未来 T_p 帧内行人轨迹 $\hat{L}_p$。这一预测问题在数学上可以表示为：

$$\hat{L}_p=f(L_o,S) \tag{1}$$

其中，f 表示本文设计的轨迹预测模型。

3.2 模型概况

3.2.1 构建场景输入

行人轨迹的预测不仅受到个体运动模式的影响，还与周围场景环境紧密相关。为了在预测模型中有效融入场景因素，本文采用了一种与 Scene-STGCN[7]一致的场景特征构造方法。通过将场景语义图与行人的运动轨迹相结合，构建出场景特征向量。这些特征向量随后被直接输入到模型中，以实现对行人未来轨迹的准确预测。这种方法使得模型能够综合考虑行人行为和场景环境，提高了预测的准确性和可靠性。

具体而言，本文采用固定场景块内场景语义信息的变化比构造场景特征向量。首先，如果轨迹数据代表的是行人在世界坐标系下的坐标，则需要将行人的世界轨迹坐标转换为语义图像下的坐标，具体如下：

$$\mathcal{P}=\mathcal{K}^{-1}P \tag{2}$$

其中，$\mathcal{P}$代表行人的世界坐标，$\mathcal{K}$代表对应场景下的单应性矩阵，$\bar{\mathcal{P}}$代表行人在语义图像上的坐标。接下来，以行人为中心构造固定大小的场景语义块，得到二值场景语义矩阵 $\mathcal{U}_t^n(i)$，其中 $i(0\leqslant i<C)$代表场景语义种类，C 表示语义种类的总数。并将根据当前帧场景语义像素的占比相对于上一个观测帧场景语义像素占比的变化作为场景特征向量。具体计算式如下：

$$\boldsymbol{f}_{sc_t}^n(i)=\frac{\sum(\mathcal{U}_t^n(i)\oplus\mathcal{U}_{t-1}^n(i))}{S} \tag{3}$$

其中，$\boldsymbol{f}_{sc_t}^n(i)$表示场景特征向量，$\oplus$表示异或操作，$\sum$表示对矩阵的所有元素进行求和，$S$ 表示场景语义块的面积。

3.2.2 模型总览

SPP-STGCN 模型框架如图 1 所示，整体分为两个阶段。第一阶段为邻接矩阵生成模块，它负责接收轨迹特征向量 $\boldsymbol{V}$ 和场景特征向量 $\boldsymbol{f}_{sc}$。该模块分别在时空两个维度对场景特征与轨迹特征进行融合，分别生成行人轨迹与时间维度、空间维度的邻接矩阵 $\mathcal{A}_t$ 和 $\mathcal{A}_s$。随后，邻接矩阵与轨迹特征向量进行组合，得到描述时间交互的行人轨迹时空图 $\mathcal{G}_t=(\mathcal{V},\mathcal{A}_t)$和描述空间交互的行人轨迹时空图 $\mathcal{G}_s=(\mathcal{V},\mathcal{A}_s)$。接下来将两个图和场景特征向量输入第二阶段基于图卷积的主干预测网络中，预测行人未来轨迹的分布。

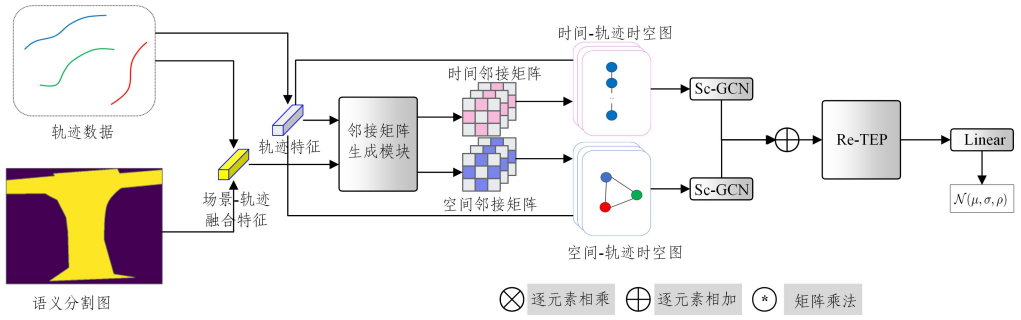

图 1 SPP-STGCN 模型架构

Fig. 1 Framework of SPP-STGCN model

3.3 图邻接矩阵生成模块

图邻接矩阵生成模块的架构如图 2 所示。首先，模型利用简化的自注意力层[21]分别在时间、空间维度上生成邻接矩阵。具体到每个维度上，轨迹特征向量和融合轨迹与场景的融合向量都会经过自注意力模块生成对应的邻接矩阵。随后，在时间、空间维度，轨迹的邻接矩阵和融合特征的邻接矩阵会被输入到场景邻接矩阵融合块 SAFB 中，进行场景维度的特征融合，生成在该维度下的邻接矩阵。最后，时间、空间维度会被输入到进行时空融合的时空交互感知模块 STIL[12]中，并经过规范化生成最后的邻接矩阵。

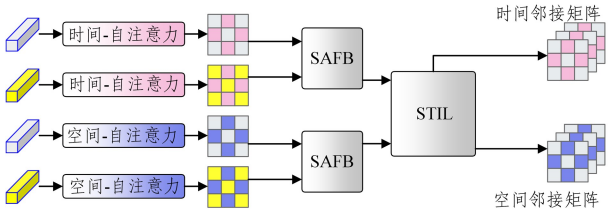

图 2 邻接矩阵生成模块

Fig. 2 Adjacency matrix generating module

自注意力层的架构如图 3 所示。特征向量分别经过两个线性层之后得到 ***Key*** 和 ***Query***，接下来将 ***Key*** 和 ***Query*** 进行矩阵乘法，在经过规范化以保持数值稳定和 softmax 激活后，得到未进行交互的邻接矩阵 $\mathcal{A}$。为了增强模型的表达能力，在这部分中本文采用了多头注意力机制，表示为：

$$\mathcal{A}=\mathrm{softmax}\left(\frac{\boldsymbol{key}(V)\boldsymbol{query}(V)}{\sqrt{d_m}}\right) \tag{4}$$

其中，V 表示输入，***key*** 和 ***query*** 都是线性层，d_m 是隐藏层的深度。

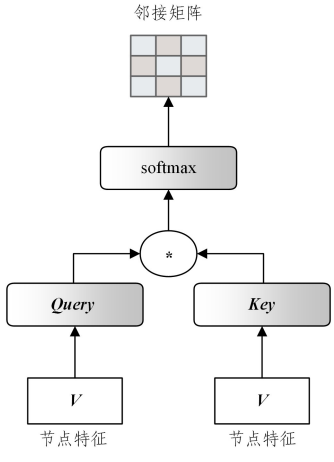

图 3 自注意力生成邻接矩阵

Fig. 3 Self-attention generating adjacency matrix

为了使邻接矩阵能够融合场景维度的特征，本文提出了场景邻接矩阵融合块 SAFB，其架构如图 4 所示。首先，融合邻接矩阵会经过 1×1 的卷积并经过 tanh 函数激活生成权重因子矩阵，该矩阵作为伸缩因子乘以一个可学习的参数 λ 后，与轨迹邻接矩阵逐元素相乘，这一部分将作为残差与原轨迹邻接矩阵相加，得到融合后的轨迹邻接矩阵。表示为：

$$\mathcal{A}_{sc}=\lambda * \tanh(conv(\mathcal{A}_{sc})) \tag{5}$$

$$\mathcal{A}=\mathcal{A}+\mathcal{A}_{sc} * \mathcal{A} \tag{6}$$

其中，$\mathcal{A}$ 和 $\mathcal{A}_{sc}$ 分别表示轨迹邻接矩阵和场景邻接矩阵，$*$ 表示逐元素相乘。

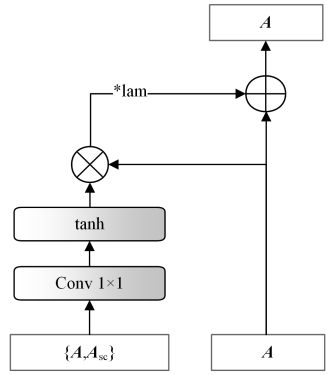

图 4 场景-邻接矩阵融合块

Fig. 4 Scene-adjacency fusion block

STIL 是 STIGCN[12]中提出的用于邻接矩阵时空交互的模块。以时间维度举例，时间维度的基础邻接矩阵会和空间维度的基础邻接矩阵经过一系列的网络层进行融合，生成以时间维度为主、融合空间维度的融合矩阵。随后该融合矩阵会在多头注意力的维度与原基础邻接矩阵进行拼接，得到吸收了空间维度信息的时间维度邻接矩阵。

为了保证图卷积网络的稳定性[11]，每一个生成的邻接矩阵都要加上一个单位矩阵，并再次进行 softmax 激活。

$$\mathcal{A}_{out}=\mathrm{softmax}(\mathcal{A}+\mathbf{I}) \tag{7}$$

其中，$\mathbf{I}$ 是单位矩阵，$\mathcal{A}_{out}$ 表示其作为整个邻接矩阵生成模块的最终输出。

3.4 基于图网络的轨迹预测模块

将生成的时间、空间邻接矩阵与轨迹特征组合后，便可以得到描述时间交互的行人轨迹时空图，以及描述空间交互的行人轨迹时空图。之后，将两种维度的图输入以图卷积为核心的主干网络中，如图 1 所示。

在主干网络中，两个维度的图分别经过场景-图卷积网络(S-GCN)进行图特征提取，随后进行融合，然后输入金字塔时间外推卷积中，最后经过线性层输出对行人未来轨迹分布的预测。

在场景-图卷积网络 S-GCN 中，将场景特征向量与轨迹特征进行融合，如图 5 所示。场景特征向量与对应的轨迹特征向量共享邻接矩阵，同样会经过 GCN 提取特征。随后，二者进行逐元素相乘，并与轨迹特征的卷积残差相加得到此部分的输出。

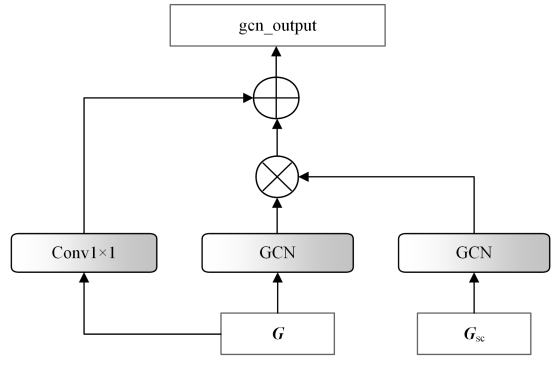

图 5 场景-图卷积

Fig. 5 Scene graph-convolution net

在时间、空间维度都进行图特征提取后，需要对两类特征进行融合。在融合步骤中，两种特征会经过 1×1 的卷积后直接相加得到融合特征。

最后，统一的特征会经过时间外推金字塔卷积以及一个线性层得到最终结果——行人未来轨迹的预测分布 $N(\hat{\mu},\hat{\sigma},\hat{\rho})$。时间外推金字塔卷积 TEP-CNN[12] 是时间外推卷积 TXP-CNN 的变体，文献[12]证实了金字塔架构的有效性。本文方法对 TEP-CNN 进行了改进，在金字塔卷积中加入了规范化模块 Rev-IN[22]，如图 6 所示。该模块首先对特征进行归一化，记录其均值与方差，在此过程中将可学习的参数作为缩放因子，并在特征提取后进行上述操作的逆操作，即逆归一化。在使用多数据集测试模型的情况下，归一化有助于减少不同数据集之间的差异，提升模型的整体性能[22]。本文的消融实验结果证实了这一点。

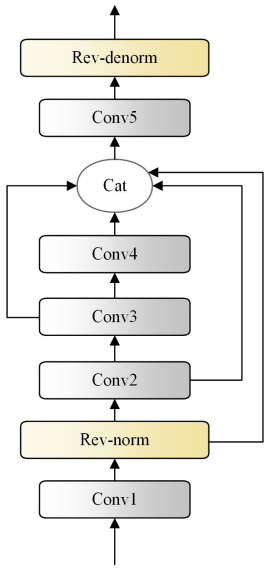

图 6 带有 Rev 正则化的金字塔外推卷积网络

Fig. 6 Pyramid extrapolation convolutional network with Rev-Norm

3.5 损失函数

本文沿用 Social-LSTM[2] 中的假设，假设模型预测未来 t 时刻行人 n 的轨迹 $\bar{p}_n^t=(\bar{x}_n^t,\bar{y}_n^t)$ 遵从二维高斯分布 $N(\bar{\mu}_n^t,\bar{\sigma}_n^t,\bar{\rho}_n^t)$，其中 $\bar{\mu}_n^t$ 为均值，$\bar{\sigma}_n^t$ 为标准差，$\bar{\rho}_n^t$ 为相关系数。本文采用的模型训练方法是最小化预测未来轨迹分布的负对数似然函数。目标函数的形式如下：

$$L^n(W)=-\sum_{t=t_0+1}^{t_p}\log P((x_n^t,y_n^t)\mid\hat{\mu}_n^t,\hat{\sigma}_n^t,\hat{\rho}_n^t) \tag{8}$$

其中，W 代表本文方法中所有可以训练的参数。

4 实验

4.1 数据集与评价指标

本文采用在行人轨迹预测领域被广泛认可的两个标志性且具有挑战性的数据集 ETH[23] 和 UCY[24] 进行大量实验。这两个数据集均采用固定视角下的鸟瞰图方式进行拍摄，提供了丰富的真实世界场景。

ETH 数据集可进一步细分为 ETH 和 HOTEL 两个子数据集。UCY 数据集则包括 UNIV，ZARA1 和 ZARA2 这 3 个子数据集。5 个数据集一共对应 4 个不同的场景。这些数据集汇集了数千条真实的行人轨迹，涵盖了相遇、跟随在内的多种复杂运动模式，符合行人在各种场景下的正常行走行为。

本文采用和之前的工作相同的测试方法和配置：根据观察到的前 3.2 秒(8 帧)内的轨迹数据，预测行人未来 4.8 秒(12 帧)内的轨迹数据。

为了评估模型的性能，本文采用了两个广泛认可的评估指标：平均位移误差(Average Displacement Error，ADE)和最终位移误差(Final Displacement Error，FDE)。假设测试数据中共有 N 个行人，观测帧共有 t_0 帧，预测帧共有 t_p 帧，p 表示行人的真实轨迹，$\hat{p}$ 表示模型的预测轨迹。

平均位移误差(ADE)表示所有行人的预测轨迹与真实轨迹在整个预测帧内位置误差的均值，具体计算式如下：

$$ADE=\frac{\sum_{n=1}^{N}\sum_{t=t_0+1}^{t_0+t_p}\|\hat{p}_t^n-p_t^n\|_2}{N\times t_p} \tag{9}$$

最终位移误差(FDE)表示所有行人的预测轨迹与真实轨迹在预测的最后一帧位置误差的均值，其具体计算式如下：

$$FDE=\frac{\sum_{n=1}^{N}\|\hat{p}_t^n-p_t^n\|_2}{N},t=t_p \tag{10}$$

近期的一些工作[16,18-19]使用了一种新的轨迹生成方式，虽然他们仍然采用 ADE 和 FDE 评价指标，但所提出的方法采用 cvae 逐步生成轨迹，并在每一步生成时运用真实轨迹进行筛选。这些方法在生成轨迹时引入了先验知识，不满足本文提出的问题假设，不适合与本文工作在指标上进行比较。

4.2 实验配置

本文实验均是基于 PyTorch 框架进行的，并使用 Nvidia RTX 2080 GPU 进行训练。使用 Adam 优化器训练 150 个 epoch，批量大小为 64，学习率设置为 0.001。

模型中的部分参数设置如下：多头自注意力中隐藏层的大小为 64，头数设置为 4；SAFB 中的 1×1 卷积隐藏层大小为 16；S-GCN 中的隐藏层大小为 64；TEP-CNN 的卷积层总数设置为 5。

本文实验采用了和之前的工作一致的交叉验证方法。具体来说，对于每个数据集的测试，将其作为测试集，使用其余 4 个数据集作为训练集和验证集。为了构建能进行交互的行人图数据，本文对每个数据集进行了数据清洗，保证每个输入图数据的完整性，以及图的节点数至少为 2。这种清洗的方法为了保证行人之间的交互性，会对数据进行严格筛选。例如，对于 ETH 和 HOTEL 子数据集，从这两个数据集中提取的有效数据量分别约占数据总量的 9.7%和 25.1%。

在测试最终结果时，从模型生成的双变量高斯分布中抽取 $N_{\text{step}}=20$ 个样本，使用其最接近真实轨迹的样本计算平均位移误差和最终位移误差。

4.3 消融实验

为了验证本文提出和采用的模块的有效性，在数据集

ETH 和 UCY 上进行了消融实验，实验结果如表 1 所列。其中，SA 表示场景邻接矩阵融合 SAFB 模块，SG 表示场景图卷积 S-GCN 模块，RN 表示在 TEP 中引入的 Rev-IN 模块。表中数据为各方法在 5 个数据集上的 ADE/FDE。

表 1 消融实验结果

Table 1 Ablation experiment results

SA	SG	RN	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
			0.74/0.95	0.33/**0.42**	0.42/0.74	0.30/0.46	0.26/0.45	0.41/0.60
√			0.60/**0.83**	0.33/0.58	0.40/0.74	0.29/0.47	0.25/0.42	0.37/0.61
√	√		0.58/0.87	0.32/0.45	**0.39**/**0.71**	0.29/0.46	0.25/0.44	0.37/0.59
√	√	√	**0.55**/0.85	**0.28**/**0.42**	0.40/0.74	**0.28**/**0.44**	**0.23**/**0.40**	**0.35**/**0.57**

可以看出，SAFB 模块对于模型性能的提升起到了最大的作用，仅仅加上该模块，便将 ETH 数据集的 ADE 从 0.74 提升到了 0.60，提升率达到了 18.9%，同时将平均 ADE 指标从 0.41 提升到了 0.37，提升率达到了 9.6%。

与 Scene-STGCN[7] 中的场景融合操作类似，本文提出了 S-GCN 模块，其在模型预测步骤中融合了场景因素。该模块在本文模型中的主要效果体现在 UNIV 数据集上，ADE 和 FDE 分别提升了 2.5% 和 4%。Rev-IN 是一个规范化的模块，其主要作用是提升模型的稳定性，同样也能一定程度地提升模型性能。可以看到，该模块在平均 ADE 和平均 FDE 上分别提升了 5.7% 和 3.5%。但该模块在 UNIV 数据集上表现较差，可能是 UNIV 数据集中行人密度远高于其他几个数据集，并且有很多几乎静止的行人作为干扰，导致该数据集的分布较其他 4 个差异较大，从而使模型在该数据集上性能下降。总体而言，SAFB 模块和 S-GCN 模块的引入确实能提升模型的平均性能。

4.4 定量分析

图网络要求对数据进行严苛的过滤，这会导致数据分布的改变。在这种前提下，本文在 ETH 和 UCY 数据集上将 SPP-STGCN 与其他基于图网络的模型进行对比，包括基于图注意力网络的 Sophie[5] 和 STGAT[25]；基于图卷积网络的 STAR[26]，Social-STGCN[3]，SGCN[4]，Causal-STGCN[27]，PTP-STGCN[9]，Scene-STGCN[7]，STIGCN[12]。

对比实验结果如表 2 所列。总体而言，相较于其他 9 种图网络模型，本文方法的平均性能达到了目前的最优水平。本文模型的 ADE 指标与次优基准模型 STIGCN 持平；而在 FDE 方面，本文模型较 STIGCN 提升了 3.4%。具体到每个数据集，相较于基准模型 STIGCN，本文模型在 ETH，HOTEL，ZARA1 和 ZARA2 数据集上均有一定的提升或与其持平，提升幅度在 ETH 上最高，ADE 指标提升了 5.2%，FDE 指标提升了 11.5%。在 UNIV 数据集上没有获得最优结果，原因可能是该数据集的场景是非常宽阔的区域，且行人密度非常大。在这种场景下，场景因素对行人的影响会远小于行人之间的交互影响。在其他几个数据集上的训练会使场景因素的比例过大，导致场景因素的引入在该数据集中反而起到了干扰作用。综合实验结果表明，本文设计的场景融合策略确实能一定程度对行人轨迹预测产生积极的作用，提高行人轨迹预测模型的平均性能。

表 2 ETH 和 UCY 数据集上的对比实验结果

Table 2 Comparative experiment results on ETH and UCY datasets

methods	ETH	HOTEL	UNIV	ZARA1	ZARA2	AVG
Sophie	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.65	0.38/0.78	0.54/1.15
STGAT	0.65/1.12	0.35/0.66	0.52/1.10	0.34/0.69	0.29/0.60	0.43/0.83
STAR	0.56/1.11	**0.26**/0.50	0.52/1.15	0.41/0.90	0.31/0.71	0.41/0.87
Social-STGCN	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
SGCN	0.63/1.03	0.32/0.55	**0.37**/0.70	0.29/0.53	0.25/0.45	0.37/0.65
Causal-STGCN	0.64/1.00	0.38/0.45	0.49/0.81	0.34/0.53	0.32/0.49	0.43/0.66
PTP-STGCN	0.63/1.04	0.34/0.45	0.48/0.87	0.37/0.61	0.30/0.46	0.42/0.68
Scene-STGCN	0.58/**0.77**	0.38/0.57	0.40/0.73	0.29/0.47	0.26/0.44	0.38/0.60
STIGCN	0.58/0.96	0.30/0.44	0.38/**0.67**	**0.28**/0.47	**0.23**/0.42	**0.35**/0.59
SPP-STGCN(Ours)	**0.55**/0.85	0.28/**0.42**	0.40/0.74	**0.28**/**0.44**	**0.23**/**0.40**	**0.35**/**0.57**

4.5 定性分析

本节对 SPP-STGCN 的模型性能进行定性分析。图 7 展示了 SPP-STGCN 在 3 种不同预测模式下的可视化分析结果。

图 7(a)中，两位行人结伴而行，因此二人的轨迹呈现相同的趋势。由可视化结果可以观察到，虽然 SPP-STGCN 的对二人运动趋势的把握不如 Scene-STGCN 准确，但是本文模型对这两位行人的运动模式进行了充分的学习，其预测结果仍然表现出相对合理的未来轨迹预测。同时，本文模型的预测结果有一个从稳定性大到稳定性小的过渡，这表明本文框架中引入时间方面的交互，使得模型对未来轨迹不确定性的把握更强。在图 7(b)中，两位行人迎面而行。SPP-STGCN 充分学习到了两位行人为避免碰撞而产生的“绕开”行为，给出了合理的、准确的未来预测轨迹分布，而 Scene-STGCN 的预测结果虽然合理，但对两位行人运动趋势的预测有些偏差。在图 7(c)中，两位行人即将汇合。可以观察到，下方行人在预测末尾帧发生了小幅度的转向，导致上方行人不得不做出不符合原本趋势的位移。SPP-STGCN 对下方行人的预测比

较准确，对上方行人的预测却不如 Scene-STGCN 准确，导致二者的预测分布图有很大部分重叠。总体而言，相对于 Scene-STGCN，SPP-STGCN 的预测性能在部分运动模式下与其基本持平。但在更多运动模式下，SPP-STGCN 展现出了更合理、更准确的预测。

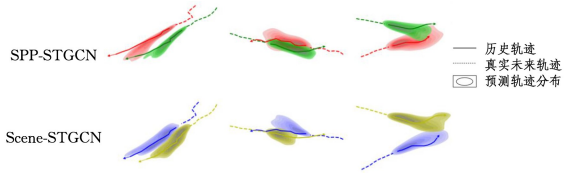

图 7 SPP-STGCN 生成的轨迹在不同运动模式下的结果对比

Fig. 7 Comparison results of trajectory generated by SPP-STGCN under different motion modes

结束语 本文提出了一种用于行人轨迹预测的、构建场景-行人-行人交互的时空图卷积网络 SPP-STGCN。该网络模型从时空两个维度构建行人轨迹时空图，并在构建图邻接矩阵、预测未来轨迹的阶段都融合了场景特征。这样的设计使得 SPP-STGCN 能够充分融合场景信息，实现更精确的行人轨迹预测。相比其他使用图网络的行人轨迹预测模型，SPP-STGCN 的平均性能达到了目前的最优水平，分别在 ADE 和 FDE 上达到了 0.35 和 0.57。未来将会探寻更多场景特征的表现形式，以及一维轨迹和二维场景的多模态融合方法。除此之外，也会从更好的时空特征提取设计、以目的地为导向等多方面持续改进模型。

参 考 文 献

[1] HELBING D, MOLNAR P. Social Force Model for Pedestrian Dynamics[J]. Physical Review E, 1995, 51(5): 4282-4286.

[2] ALAHI A, GOEL K, RAMANATHAN V, et al. Social LSTM: human trajectory prediction in crowded spaces[C]// Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2016: 961-971.

[3] MOHAMED A, QIAN K, ELHOSEINY M, et al. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction[C]// Proceedings of 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2020: 14412-14420.

[4] SHI L, WANG L, LONG C, et al. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2021: 8994-9003.

[5] SADEGHIAN A, KOSARAJU V, SADEGHIAN A, et al. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints[C]// Proceedings of 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 1349-1358.

[6] MANGALAM K, AN Y, GIRASE H, et al. From Goals, Waypoints & Paths to Long Term Human Trajectory Forecasting [C]// Proceedings of 2021 IEEE/CVF International Conference on Computer Vision. IEEE, 2021: 15213-15222.

[7] CHEN H, GI Q. Scene-constrained spatial-temporal graph convolutional network for pedestrian trajectory prediction[J]. Journal of Image and Graphics, 2023, 28(10): 3163-3175.

[8] ZHANG P, OUYANG W, ZHANG P, et al. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 12085-12094.

[9] LIAN J, REN W, LI L, et al. PTP-STGCN: Pedestrian Trajectory Prediction Based on a Spatio-temporal Graph Convolutional Neural Network[J]. Applied Intelligence, 2023, 53: 2862-2878.

[10] KOSARAJU V, SADEGHIAN A, MARTÍN-MARTÍN R, et al. Social-BiGAT: multimodal trajectory forecasting using Bicycle-GAN and graph attention networks[J]. arXiv: 1907.03395, 2019.

[11] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks [J]. arXiv: 1609.02907, 2016.

[12] CHEN W, SANG H, WANG J, et al. STIGCN: spatial-temporal interaction-aware graph convolution network for pedestrian trajectory prediction[J]. The Journal of Supercomputing, 2024, 80(8): 10695-10719.

[13] HUANG R, XUE H, PAGNUCCO M, et al. Multimodal Trajectory Prediction: A Survey[J]. arXiv: 2302.10463, 2023.

[14] GUPTA A, JOHNSON J, FEI-FEI L, et al. Social GAN: Socially Acceptable Trajectories with Generative Adversarial Networks [C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2255-2264.

[15] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative Adversarial Nets [EB/OL]. https://proceedings.neurips.cc/paper_files/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf.

[16] YUE J, MANOCHA D, WANG H. Human Trajectory Prediction via Neural Social Physics[C]// European Conference on Computer Vision. Cham: Springer: 2022: 376-394.

[17] XU P, HAYET J B, KARAMOUZAS I. SocialVAE: Human Trajectory Prediction using Timewise Latents[C]// Proceedings of European Conference on Computer Vision. Cham: Springer, 2022: 511-528.

[18] ZHOU H, YANG X, REN D, et al. CSIR: Cascaded Sliding CVAEs With Iterative Socially-Aware Rethinking for Trajectory Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2023(12): 24.

[19] XIANG W, YIN H, WANG H, et al. SocialCVAE: Predicting Pedestrian Trajectory via Interaction Conditioned Latents[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2024: 6216-6224.

[20] MANH H, ALAGHBAND G. Scene-LSTM: A Model for Human Trajectory Prediction. [J]. arXiv: 1808.04018, 2018.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention Is All You Need[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-

6010.
[22] KIM T, KIM J, TAE Y, et al. Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift [EB/OL]. https://openreview.net/pdf?id=cGDAkQo1C0p.
[23] PELLEGRINI S, ESS A, GOOL L V. Improving Data Association by Joint Modeling of Pedestrian Trajectories and Groupings [C]// Proceedings of the 11th European Conference on Computer Vision. Springer, 2010: 452-465.
[24] LERNER A, CHRYSANTHOU Y, LISCHINSKI D. Crowds by Example[J]. Computer Graphics Forum, 2007, 26(3): 655-664.
[25] HUANG Y, BI H, LI Z, et al. STGAT: Modeling Spatial-Temporal Interactions for Human Trajectory Prediction[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 6272-6281.
[26] YU C, MA X, REN J, et al. Spatio-Temporal Graph Transformer Networks for Pedestrian Trajectory Prediction[C]// Proceedings of Computer Vision-ECCV: 16th European Conference. Cham: Springer, 2020: 507-523.
[27] CHEN G, LI J, LU J, et al. Human Trajectory Prediction via Counterfactual Analysis[C]// Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 9824-9833.

HONG Mingjun, born in 2003, postgraduate. His main research interests include deep learning and computer vision.

JI Qingge, born in 1966, Ph.D, associate professor, is a senior member of CCF (No. 07014S). His main research interests include computer vision, computer graphics and virtual reality.

（责任编辑：何杨）

3 支 CCF 产学合作基金同步申报中

2025 年 CCF 产学合作基金预计发布 20+支基金，目前有 3 支基金同步申报中，现将申报中基金按照已发布时间的先后顺序进行汇总，欢迎 CCF 会员关注并积极申报。

(1)2025 年 CCF-联想蓝海科研基金

2025 年度，CCF-联想蓝海科研基金以 AI、计算与部件两大方向为重点，希望通过该基金搭建产业界与学术界的协作桥梁，以“首创技术”或“成果转化”为共同目标，通过“企业出题、高校揭榜、联合攻关”，达成校企双赢。项目期间，联想也会为学生提供实习机会，让学生能够通过基金项目，接触到真实的应用场景。

本基金会分为单点突破和前沿探索两大类，单点突破类每个课题资助人民币 30 万元，前沿探索类每个课题资助人民币 10 万元。其中，前沿探索鼓励前瞻创新，主要以思想交流和学术报告为主要交付内容。其中 AI 专项共 11 个课题，计算与部件专项共 7 个课题。

截止申请截止时间：2025 年 11 月 15 日 24：00

(2)2025 年 CCF-快手大模型探索者基金

2025 年 CCF-快手大模型探索者基金聚焦于大模型技术基础研究与应用探索，目的加速技术成果产业化转化；同时致力于培养该领域科研与工程人才，通过搭建全球学者产学研合作平台，促进学术界与工业界深度交流合作，助力科技进步与社会发展。

2025 年度共发布 18 项研究课题，围绕“大语言模型”、“视觉理解与生成大模型”、“视频处理大模型”、“生成式推荐/搜索/广告大模型”、“大模型应用”、“大数据方向”这 6 个重点研究方向，为每项课题提供不高于人民币 30 万元的资助，同时快手将提供技术、算力、和脱敏数据等资源支持。

申报截止时间：2025 年 12 月 1 日 24：00。

(3)2025 年 CCF-网易雷火联合基金二期

2025 年 CCF-网易雷火联合基金二期课题聚焦人工智能与游戏、数字人、具身智能等前沿交叉领域，发布了 11 项关键技术课题，本期课题主要是在指定的研究领域主题下，限定课题场景、任务边界和预期指标，与网易雷火及网易伏羲研究团队开展的研究合作，每个项目资助额度为 10-15 万元人民币。

申报截止时间：2025 年 11 月 30 日 24：00。

据 CCF 微信公众号