基于双向时间建模与时空自监督学习的多行人轨迹预测

赵帅[1]，李琳[2]

（1.上海理工大学光电信息与计算机工程学院，上海 200093；2. 上海理工大学光电信息与计算机工程学院，上海 200093）

摘　要：为了捕捉行人轨迹中的复杂时空依赖关系，本文提出了一种结合双向时间学习模块与时空交互学习模块的行人轨迹预测模型。模型通过双向时间特征建模和自监督学习挖掘时空交互特征。在双向时间学习模块中，利用双向时间卷积网络同时建模历史与未来的轨迹信息，以捕捉轨迹动态变化特征。在时空交互学习模块中，通过 TTT (Test-Time Training) 层的自监督学习机制，在推理阶段动态调整特征表示，从而建模时空关联性。最终，通过自适应融合策略对两模块提取的特征进行加权组合，以实现对关键特征的聚焦和无关信息的抑制。实验结果表明，该模型在 ETH 和 UCY 数据集上具有良好的预测性能。

关键词：行人轨迹预测；时空模型；**自监督学习**；BiTCN；TTT

中图分类号：TP391　　**文献标志码：**A

Multi-Pedestrian Trajectory Prediction based on Bidirectional Temporal Modeling and Spatiotemporal Self-supervised Learning

ZHAO Shuai[1], LI Lin[2]

（1. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China; 2. School of Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China）

Abstract: To capture the complex spatiotemporal dependencies in pedestrian trajectories, this paper proposes a trajectory prediction model that combines a bidirectional temporal learning module and a spatiotemporal interaction learning module. The model leverages bidirectional temporal feature modeling and self-supervised learning to extract spatiotemporal interaction features. In the bidirectional temporal learning module, a bidirectional temporal convolutional network is utilized to simultaneously model both historical and future trajectory information, enabling the capture of dynamic trajectory changes. In the spatiotemporal interaction learning module, the Test-Time Training (TTT) layer is employed with a self-supervised learning mechanism to dynamically adjust feature representations during the inference stage, thereby modeling spatiotemporal correlations. Finally, an adaptive fusion strategy is used to combine the features extracted by the two modules, focusing on key features while suppressing irrelevant information. Experimental results demonstrate that the proposed model achieves competitive prediction performance on the ETH and UCY datasets.

Key words: pedestrian trajectory prediction; spatiotemporal model; self-supervised learning; BiTCN; TTT

引　言

行人轨迹预测在诸如视频监控[1]、自动驾驶[2]和视觉感知[3]等多个领域极为重要，能帮系统提前预判风险，保障安全与效率。但因行人行为的不确定性和复杂场景的影响，轨迹预测面临诸多挑战。

基金项目：国家自然科学基金资助项目（61673277）。

传统的预测方法多结合运动学模型与贝叶斯滤波器，通过状态传播实现轨迹外推[4]进行预测。但这类方法假设行人运动恒定，难以应对复杂动态行为。像 Schneider 等人[5]对比发现多模型交互方法虽有一定优势，但对速度恒定的假设仍使其局限。Pavlovic V 等人[6]提出切换线性动力学系统（SLDS）通过马尔可夫链实现线性模型间的概率切换，能够处理非线性运动模式，却需要大量数据进行先验和转移概率的优化，实际应用中存在数据依赖性强的问题。

近年来，基于深度学习的预测方法兴起，不依赖固定数学模型，靠大规模数据学习映射关系。例如，Alahi 等人[7]提出了社会长短期记忆网络（Social-LSTM），通过社交池化层建模多行人的交互约束，生成无冲突轨迹。然而，Social-LSTM 存在计算复杂度高、实时性差的问题，尤其在实时预测场景中。SR-LSTM[8]在 Social-LSTM 基础上扩展了视觉特征和新的池化机制，并通过加权机制来衡量每个行人对其他行人的贡献，但在时间维度难捕捉序列之间的长期依赖关系。

生成对抗网络（GAN）[9]和变分自编码器（VAE）[10]等生成模型通过引入噪声和多样性损失，能够输出多模态预测结果。 例如，Gupta 等人[11]的 SGAN 模型通过生成器 - 判别器对抗训练，在轨迹多样性和预测速度上优于传统 LSTM 方法，但存在长时社交关系建模不足与训练稳定性差的缺陷。Fang 等人[12]提出的 Atten-GAN 引入注意力池化模块来充分提取行人间的交互信息，并通过在损失函数中加入随时间减少的噪声来解决 GAN 训练中的梯度消失问题。Kothari 等人[13]提出一种改进的 SGAN 架构(SGANv2)，通过协同采样策略，在测试时也利用了学习到的鉴别器，不仅细化了碰撞轨迹，而且防止发生模式崩溃问题。Yang 等人[14]提出一种结合 GAN 和社会自注意机制的行人轨迹预测模型，该模型通过生成器预测未来轨迹，判别器判断其真实性，同时利用社会自注意机制提取重要的交互信息，从而帮助模型聚焦于关键的社会互动特征，提升预测精度。通过结合 MLP 和 LSTM，模型能够在时间和空间维度上提取深层特征，实现多模态预测。

图神经网络（GCN[15]和 GAT[16]）的发展为空间交互建模提供了新思路。Mohamed 等人[17]提出的 Social-STGCNN 采用图卷积和 TCN[18]提取时空特征，有效避免了循环结构的误差积累问题。Zhu 等人[19]提出了 Tri-HGNN 模型，通过分层图网络融合外部交互与内在意图，提升了复杂场景下的建模能力，但较高的计算复杂度限制了其实时应用。杨永鹏等人[20]提出了一种基于时序分解和注意力图神经网络（TDAGNN）的交通预测模型，通过双分支时序分解卷积神经网络挖掘时间依赖，结合多头交互注意力网络和自缩放动态扩散图神经网络捕捉动态异质信息与空间依赖，为复杂时空依赖建模提供了新思路。

现有方法普遍存在两方面局限：其一，单向时间建模难以利用未来轨迹的潜在影响，导致时序特征表征不完整；其二，动态场景下行人运动模式的突变性，使得固定参数的模型适应性不足。针对上述问题，本文提出一种融合双向时间建模与时空自监督学习的轨迹预测模型：通过双向时间卷积网络（BiTCN）同时捕捉历史与未来的时序依赖，借助测试时训练（TTT）层[21]的自监督机制动态优化时空交互特征，最终通过自适应融合策略实现关键特征聚焦。实验结果表明，该模型在 ETH 和 UCY 数据集上显著提升了复杂场景下的预测精度。

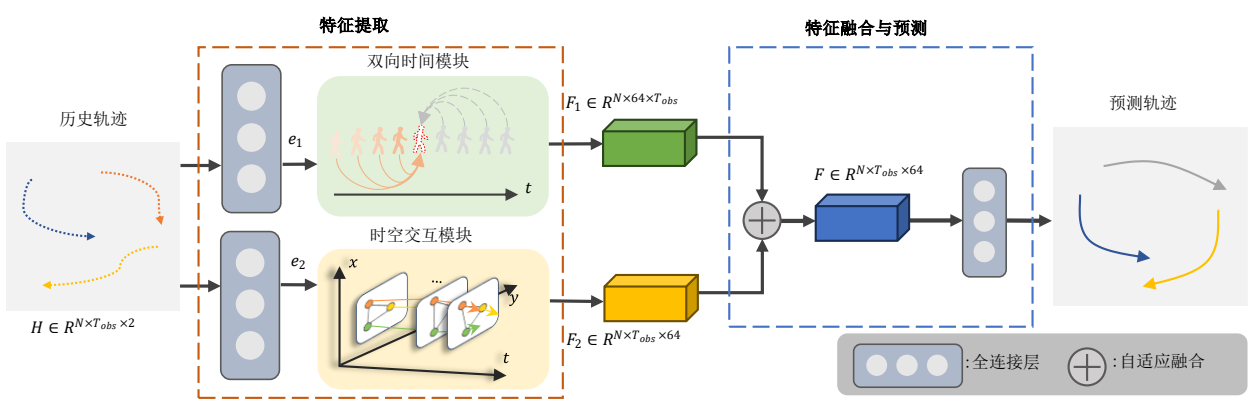


图 1　模型整体结构

Fig. 1　Overall model architecture

1 模型方法

1.1 问题定义

在行人轨迹预测中，首先从固定间隔的视频帧中提取每个行人的空间坐标。设第 i 个行人在时间 t 的二维坐标为 (x_t^i, y_t^i)，表示为集合 $\{\vec{p}_t^i | i = 1, \dots, N, t = 1, \dots T\}$，其中 N 为需要预测的行人总数，每个行人的轨迹总时间长度定义为 $T = T_{obs} + T_{pred}$，其中 T_{obs}表示历史轨迹的时间长度，而 T_{pred}为预测轨迹的时间长度。行人轨迹预测的目标是基于行人 i 的历史轨迹 $H^i = (\vec{p}_1^i, \dots, \vec{p}_{T_{obs}}^i)$，预测行人 i 在未来时间段 $[T_{obs} + 1, T_{\text{pred}}]$ 内的路径 $\Gamma^i = (\vec{p}_{T_{obs}+1}^i, \dots, \vec{p}_{T_{pred}}^i)$。

该任务可以形式化为通过学习模型的参数W^*，使模型能够在未来特定时间段内对每个行人的位置进行准确预测。其数学表达式如下：

$$\Gamma = f(H^1, H^2, \dots, H^N; W^*) \tag{1}$$

其中，$\Gamma = \{\Gamma^i : i \in 1,2, \dots, N\}$ 表示所有行人的预测轨迹集合。在本文后续表述中，除非特别说明，所有符号均默认针对第 i 个行人。例如(x_t, y_t)表示第 i 个行人在 t 时刻的坐标位置，轨迹集合$H = (\vec{p}_1, \dots, \vec{p}_{T_{obs}})$ 表示其历史轨迹。

1.2 模型整体概述

如图 1 所示，模型主要由特征提取和特征融合与预测两部分组成。首先，将历史轨迹通过全连接层分别映射到两个高维空间，得到行人的轨迹特征矩阵。随后，这些特征分别被传递到时间依赖模块和时空交互模块。时间依赖模块用于提取轨迹特征的前向和后向时间依赖特征，而时空交互模块则用于捕捉行人之间的动态交互关系。最后，通过自适应融合方式整合从两个模块提取到的特征，利用全连接层输出最终的预测轨迹特征。接下来将对这两部分详细描述。

1.2.1 双向时间学习模块

由于 TCN 提取时序特征时只能利用过去的信息，无法捕捉未来的上下文信息，在处理长期依赖关系时会导致部分特征信息的丢失。因此，本模型选择双向建模的方式，采用 BiTCN 网络同时利用过去和未来的轨迹特征，通过结合正向卷积和反向卷积对历史轨迹与未来轨迹双向建模，扩展了模型对时间依赖关系的理解，提升了特征的全局性与丰富性。图 2 展示了时间依赖模块利用 BiTCN 网络学习行人轨迹特征的过程。

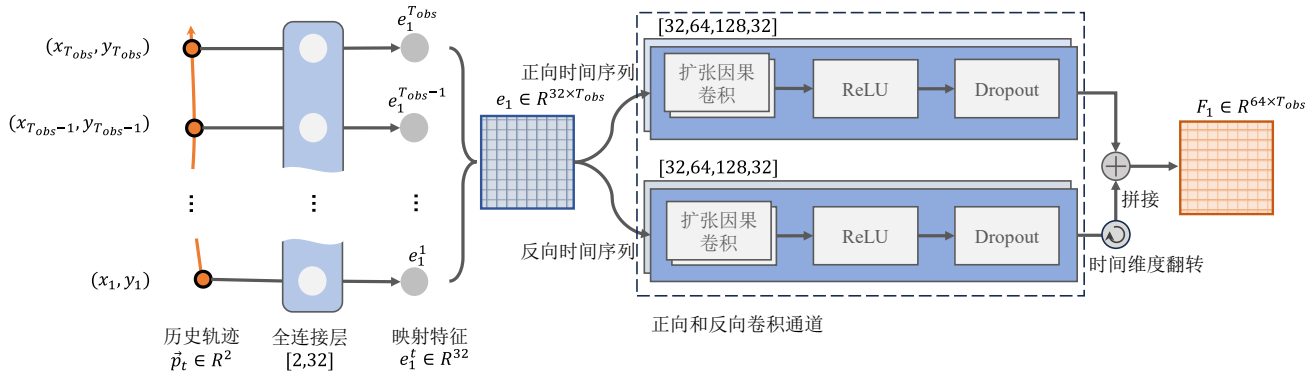


图 2 时间依赖模块

Fig. 2 Time-dependent module

首先通过一个全连接层将历史轨迹 $\boldsymbol{H} \in R^{2 \times T_{obs}}$从原始空间映射到一个更高维空间，其中，2 表示在每个时刻的 x 和 y 坐标特征。具体映射为：

$$\boldsymbol{e_1} = \phi_1(\boldsymbol{H}, \boldsymbol{W_{e1}}) \tag{2}$$

其中，$\boldsymbol{e_1} \in R^{32\times \mathrm{T}_{obs}}$表示经过映射后的 32 维特征，$\phi_1(\cdot)$表示全连接层操作，$\boldsymbol{W_{e1}}$ 是可学习的参数矩阵。

接着，为了学习轨迹特征中的后向时间信息，沿时间维度翻转轨迹序列生成反向轨迹，此时模型有正向序列和反向序列两个输入。分别将正向和反向序列传递到正向卷积和反向卷积通道中，学习轨迹特征中的正向和后向特征。以下以正向序列输入为例说明正向卷积的操作细节。

对于输入的轨迹序列 $\boldsymbol{e_1} \in R^{32\times T_{obs}}$,首先需要构建一个卷积核 $\boldsymbol{w} \in R^{K\times 32\times \tau}$,其中$K = 32$ 是卷积核的数量，τ 是卷积核的时间跨度，即每次卷积操作的范围为 τ 个时间步。TCN 采用一维卷积，卷积核沿着时间维度滑动提取特征。每个卷积核 $k \in K$ 在时刻 t 的因果卷积输出 $f_k(t)$ 可以表示为：

$$f_k(t) = \sum_{i=1}^{F}\sum_{j=1}^{\tau} w_{i,t-j} \cdot e_{1_{i,t-j}}, \qquad k \in 1,\dots,K \tag{3}$$

其中，$w_{i,j}$ 表示卷积核在第 i 个特征通道上第 j 个时间步的卷积核权重。为在不增加网络深度的情况下扩大感受野，TCN 引入扩张卷积，在卷积核中引入扩张率d，扩展了感受野范围。扩张因果卷积的计算公式为：

$$f_k(t) = \sum_{i=1}^{F}\sum_{j=1}^{\tau} w_{i,t-d\cdot j} \cdot e_{1_{i,t-d\cdot j}} \tag{4}$$

其中，d 表示扩张率。如图 3 所示，令 $d = [1,2,4]$，则堆叠三层卷积层即可在 $t = 8$ 时覆盖 8 个时间步的历史信息。最终经过 K 个卷积核提取到正向特征 $\boldsymbol{f}_{\mathbf{forward}} \in R^{K\times T_{obs}}$。

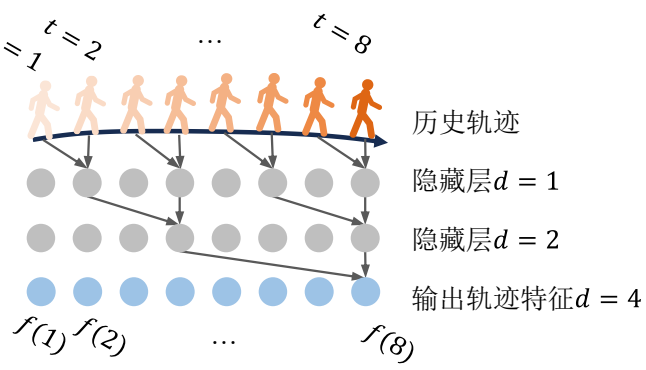


图 3　扩张因果卷积示意图

Fig. 3　Illustration of the Expanded Causal Convolution

从反向轨迹序列提取特征的过程与正向类似，将反向序列输入到反向卷积通路中得到反向特征 $\boldsymbol{f}_{\mathbf{backward}} \in R^{K\times T_{obs}}$，其中 K=32。由于反向特征表示的是从未来时间到过去时间的依赖关系，而正向特征表示的是从过去时间到未来时间的依赖关系，为确保时间一致性，需要将 $\boldsymbol{f_{backward}}$ 沿时间维度翻转以对齐时间特征关系。翻转后的反向特征与正向特征在特征维度上进行拼接，得到双向特征：

$$\boldsymbol{f_1} = [\boldsymbol{f_{forward}}; \boldsymbol{f_{backward}}] \tag{5}$$

其中，符号 $[\cdot;\cdot]$ 表示在特征维度上进行拼接，$\boldsymbol{f_1} \in R^{2K\times T_{obs}}$表示整合的双向特征。对于场景中的 N 个行人，可以得到 N 个双向特征，即$\boldsymbol{F_1} \in R^{N\times 2K\times T_{obs}}$，包含了 N 个行人的正向与反向轨迹的双向时序信息。

1.2.2　时空交互学习模块

时空交互模块以 TTT 层作为网络的核心，其通过多视图特征变换和自监督学习机制，动态捕捉行人轨迹的时空依赖特征。TTT 层通过外循环优化投影矩阵等全局参数，内循环逐时间步更新隐藏状态，实现对动态交互模式的自适应学习。如图 4 所示，输入历史轨迹 $\boldsymbol{H} \in R^{T_{obs}\times N\times 2}$ 首先通过一个全连接层进行特征映射，将每个行人的坐标转换为高维特征表示，生成初始特征表示 $\boldsymbol{e_2} \in R^{T_{obs}\times N\times F}$，其中 $F = 32$ 为映射后的特征维度。

之后，将映射后的高维特征输入到特征聚合模块 1 中。在特征聚合模块 1 中，首先通过层归一化对输入特征进行标准化处理后输入至 TTT 层。在 TTT 层中，每个时间步 t 通过可学习的投影矩阵将**输入特征** $\boldsymbol{e_2}$ 转换为多视图表示以构建自监督学习任务。具体而言，通过投影矩阵 $\boldsymbol{\theta_K} \in \mathbb{R}^{d\times F}$ 对特征进行映射，生成包含噪声的查询视图 $\boldsymbol{q_t^{train}} = \boldsymbol{\theta_K} \cdot \boldsymbol{e_2}$ ，该视图保留输入主要结构并引入噪声作为动态学习的输入信号；同时，利用投影矩阵 $\boldsymbol{\theta_V} \in \mathbb{R}^{d\times F}$ 生成目标视图 $\boldsymbol{v_t} = \boldsymbol{\theta_V} \cdot \boldsymbol{e_2}$,作为自监督学习的参考信号以约束特征重构的准确性。

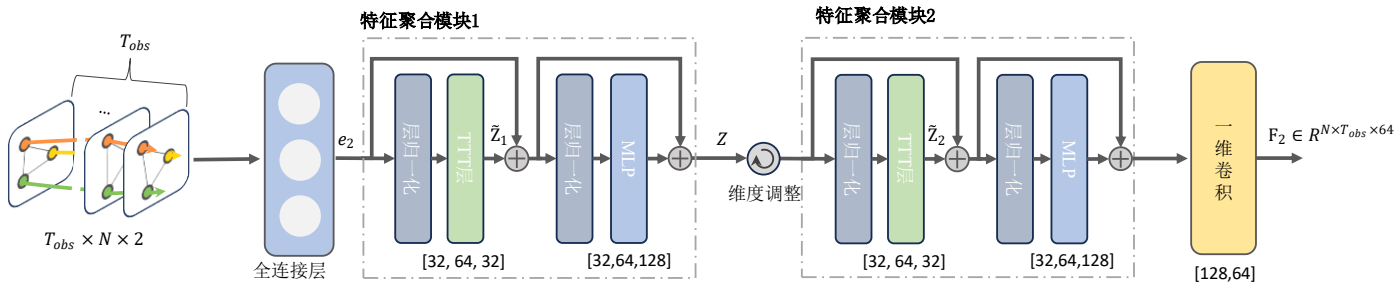


图 4　时空交互模块

Fig. 4　Spatiotemporal interaction module

内循环网络$f(\cdot)$对查询视图 $\boldsymbol{q_t^{train}}$ 进行映射，通过最小化与目标视图 v_t 的均方误差损失优化隐藏状态 $\boldsymbol{W_t}$ 。损失函数定义如下：

$$\ell(\boldsymbol{W_t}; \boldsymbol{e_2}) = \left\| f\left(\boldsymbol{q_t^{train}}; \boldsymbol{W_t}\right) - \boldsymbol{v_t} \right\|^2 \tag{6}$$

其中，$\boldsymbol{W_t}$ 为第 t 步的可学习参数，$\|\cdot\|_2$表示 L2 范数。通过梯度下降更新，更新规则如下：

$$\boldsymbol{W_{t+1}} = \boldsymbol{W_t} - \eta \nabla_{w_t} \ell(\boldsymbol{W_t}; \boldsymbol{e_2}) \tag{7}$$

其中，η 为学习率，逐时间调整 $\boldsymbol{W_t}$ 使得模型适应轨迹数据的动态交互模式。基于更新后的隐藏状态 $\boldsymbol{W_t}$ 和测试投影矩阵 $\boldsymbol{\theta_Q} \in \mathbb{R}^{d\times F}$生成的测试视图 $\boldsymbol{q_t^{test}} = \boldsymbol{\theta_Q} \cdot \boldsymbol{e_2}$提取时空特征:

$$\boldsymbol{z_t} = f(\boldsymbol{q_t^{test}}; \boldsymbol{W_t}) \tag{8}$$

累积所有时间步输出得到 $\boldsymbol{TTT_{out}} = \left[z_1, z_2, \ldots, z_{T_{obs}}\right] \in \mathbb{R}^{T_{obs}\times N\times d}$，编码行人间的动态交互模式。

为增强模块表达能力并缓解梯度消失问题，特征聚合模块采用残差连接结构。首先，将层归一化后的输入特征与 TTT 层的输出特征相加:

$$\boldsymbol{Z_1} = \boldsymbol{e_2} + \boldsymbol{TTT_{out}} \tag{9}$$

随后，通过多层感知机对层归一化后的Z_1 进行处理，提取复杂非线性模型，并经残差连接得到最终的时空交互特征：

$$\boldsymbol{Z} = \boldsymbol{Z_1} + MLP\left(LayerNorm(\boldsymbol{Z_1})\right) \tag{10}$$

其中，特征 $\boldsymbol{Z} \in \mathbb{R}^{T_{obs}\times N\times 32}$ 已有效编码行人间的动态空间交互关系。进一步通过结构相同的特征聚合模块 2 处理，调整特征维度至 $\mathbb{R}^{T_{obs}\times N\times 32}$ ，并通过一维卷积层调整通道数至 64 维，最终得到时空交互特征$\boldsymbol{F_2} \in \mathbb{R}^{N\times T_{obs}\times 64}$ 。该表示能够捕捉每个行人在不同时刻对周围环境的依赖关系，为后续的自适应特征融合提供关键的时空上下文信息。

1.4　自适应融合

在前面我们已经分别构建了时间依赖模块和时空交互模块，两个模块分别注重于构建行人轨迹序列在时间上的依赖关系以及构建行人在每个时刻的动态交互关系。为了更好的将两个模块的特征融合在一起，这里借鉴了 MotionBERT[22]的方法， 即采用自适应融合策略来融合时间依赖特征和时空交互特征，具体方法如下：

$$\boldsymbol{F^i} = \alpha_1^i \odot F_1^i + \alpha_2^i \odot F_2^i, \qquad i \in 1, \ldots, 64 \tag{11}$$

$$\alpha_1^i, \alpha_2^i = softmax\left(W_\alpha^i \cdot [F_1^i; F_2^i]\right) \tag{12}$$

其中，$\boldsymbol{F} \in R^{N\times T_{obs}\times 128}$，⊙表示元素间乘法操作，$\boldsymbol{F_1}$和$\boldsymbol{F_2}$分别是时间依赖特征和时空交互特征，$\boldsymbol{\alpha_1}$ 和$\boldsymbol{\alpha_2}$表示自适应融合的权重，$[\cdot;\cdot]$表示沿着特征维度拼接，$\boldsymbol{W_\alpha}$ 是一个可学习的线性变换矩阵，通过多轮训练后，可以优化出一个最大化预测精度的权重值。自适应融合方式在此基础上，能够增强对提高预测精度更具价值的特征的利用，同时削弱那些对预测精度贡献较低甚至可能降低预测精度的特征的影响，从而实现特征的最优融合。

1.5 轨迹预测和损失函数

遵循 Social-LSTM[7]的先例，本文假设某行人在时间步 t 的位置坐标 (x_n^t, y_n^t) 服从双变量高斯分布函数，即 (x_n^t, y_n^t)~F$\left(\mu_n^{x_t}, \mu_n^{y_t}, \sigma_n^{x_t}, \sigma_n^{y_t}, \rho_n^t\right)$，其中，$\mu$ 代表均值，σ 代表标准差，而ρ 是相关性系数，双高斯分布函数的概率密度函数如下所示：

$$N(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} e^{\left(\frac{-(x-\mu_x)^2}{2\sigma_x^2} + \frac{-(y-\mu_y)^2}{2\sigma_y^2}\right)} \tag{13}$$

在预测过程中，模型通过一个全连接层输出第 n 个行人在 t 时刻预测的轨迹的 5 维特征，分别对应于双变量高斯分布中的 $\mu_n^{x_t}, \mu_n^{y_t}, \sigma_n^{x_t}, \sigma_n^{y_t}, \rho_n^t$。预测轨迹更精确的匹配真实轨迹的双变量高斯分布，则意味着损失函数值最小，即双变量高斯分布趋近于最大似然估计值，因此这里采用负对数似然损失函数作为损失函数，见公式（14）。这一计算过程依靠一个全连接层进行预测。此模型通过最小化下面的负对数似然损失来训练：

$$L_n(\boldsymbol{W}^*) = -\sum_{t=T_{obs}+1}^{T_{pred}} ln\, P\left((x_n^t, y_n^t) \,\middle|\, \mu_n^{x_t}, \mu_n^{y_t}, \sigma_n^{x_t}, \sigma_n^{y_t}, \rho_n^t\right) \tag{14}$$

其中，$\boldsymbol{W}^*$代表模型中当前学习到的模型参数，T_{obs}表示预测的时间步长，μ、σ、ρ均为每个时间步下每个行人的x、y 坐标对应于双变量高斯分布函数中的必要参数，通过最小化损失函数得到最优的参数值。

2 实 验

2.1 数据集和评估指标

提出的轨迹预测模型在两个公开的轨迹预测数据集 ETH 和 UCY 上进行了评估，这两个数据集涵盖了丰富的社交互动场景。这些数据集包括 ETH-eth、ETH-hotel、UCY-zara1、UCY-zara2 及 UCY-univ 共五个场景。ETH 数据集中平均行人数量为 5，UCY 数据集中平均行人数据量为 18。行人在数据集中的行为多样，包括走非直线路径、从多方向移动、群体行走、避免碰撞及停留等。此外，还开展了一系列消融实验和对比实验，来探讨每一个新提出组件的效果。有两种类型的指标用于评估轨迹预测的性能，包括平均位移误差（ADE）和最终位移误差（FDE），单位为米。

1)平均位移误差（ADE）：ADE 指标用于测量提出的方法生成的预测轨迹与真实轨迹之间的欧几里得距离的平均值，覆盖所有预测时间步。较小的值表示更好的结果。ADE 的定义如下：

$$ADE\left(\hat{Y}, Y\right) = \frac{\sum_{i=1}^{N}\sum_{t=T_{\text{obs}}+1}^{T_{\text{pred}}} ||\hat{y}_i^t - y_i^t||_2}{N\left(T_{\text{pred}} - T_{\text{obs}}\right)} \tag{15}$$

2)最终位移误差（FDE）：FDE 通过计算最终预测位置与每个行人的真实目的地之间的平均欧几里得距离来计算，较小的值表明性能更佳。FDE 的定义如下：

$$FDE\left(\hat{Y}, Y\right) = \frac{\sum_{i=1}^{N} ||\hat{y}_i^{T_{pred}} - y_i^{T_{\text{pred}}}||_2}{N} \tag{16}$$

2.2 实现细节

本文实验环境配置如下，硬件方面使用了 RTX 3080 Ti（12GB）显卡以及 12 核 Intel(R) Xeon(R) Silver 4214R CPU（主频 2.40GHz）。软件环境采用 PyTorch 2.0、Python 3.8（运行于 Ubuntu 20.04 系统）和 CUDA 11.8。设置观察期内的步数T_{obs}为 8 步（等于 3.2 秒），而预测期内的步数T_{pred}设为 12 步（等于 4.8 秒）。TTT 层的内循环网络模型 MLP 设置参考了 TTT 原文中的 TTT-MLP 模型配置，BiTCN 中卷积核卷积步长 τ 设置为 3，每个 TCN 通道的层数为 3。该算法选择了 Adam 优化器进行模型训练，设置学习率为 0.01。

2.3 与现有方法的比较

提出的方法与 15 种基线方法在公共基准数据集 ETH 和 UCY 进行了比较。包括 Sophie[23]、SR-LSTM[8]、DSCMP[24]、Social-STGCNN[17]、AST-GNN[25]、SGCN[26]、Atten-GAN[12]、Conv2D-tobs-NR-Ks5[27]、Social-Implicit[28]、Social-SAGAN[14]、Tri-HGNN[19]、Social TAG[29]、RDGCN[30]、MSTCNN[31]以及 DSTCNN[32]等。所有的方法都输入 8 帧，输出 12 帧。与这些基线方法的对比实验结果展示在表 1 中，ADE/FDE 指标越低越好。

表 1 显示，本文模型与除 DSTCNN [37] 之外的对比模型相比，表现出明显的性能优势。但是，DSTCNN 在 hotel 场景下的 ADE 和 FDE 指标均优于本文模型，这可能是因为 hotel 场景通常表现出较低的人群密度，相比于其他场景，行人之间的交互较少，这使得行人轨迹呈现出更独立、更可预测的特点。在这种情况下，可能是因为单一行人轨迹中的时空依赖关系较为简单，预测轨迹时不需要大量的复杂交互信息。因此，时间依赖模块和时空学习交互模块的特征融合在这种场景中相对冗余，甚至可能引入噪声，削弱了模型的性能表现。但本模型在 eth、univ、zara1、zara2 上相比于现有的模型均具有良好表现。

表 1 不同算法的 ADE/FDE 指标对比

Table 1 Comparison of ADE/FDE indicators of different algorithms

算法模型	Years	eth	hotel	univ	zara1	zara2	平均值
Sophie[23]	2019	0.70/1.43	0.76/1.67	0.54/1.24	0.30/0.63	0.38/0.78	0.54/1.15
SR-LSTM[8]	2019	0.69/1.29	0.49/1.01	0.55/1.32	0.30/0.62	0.36/0.75	0.48/1.00
DSCMP[24]	2020	0.66/1.21	0.27/0.46	0.50/1.07	0.33/0.68	0.28/0.60	0.41/0.80
Social-STGCNN[17]	2020	0.64/1.11	0.49/0.85	0.44/0.79	0.34/0.53	0.30/0.48	0.44/0.75
AST-GNN[25]	2021	0.66/1.02	0.37/0.61	0.46/0.83	0.32/0.52	0.28/0.45	0.42/0.69
SGCN[26]	2021	0.63/1.03	0.32/0.55	0.37/0.70	0.29/0.53	0.25/0.45	0.37/0.65
Atten-GAN[12]	2022	0.64/1.12	0.36/0.72	0.51/1.13	0.36/0.61	0.34/0.66	0.44/0.84
Conv2D-tobs-NR-Ks5[27]	2022	0.56/1.11	0.24/0.46	0.58/1.23	0.46/0.99	0.35/0.75	0.44/0.91
Social-Implicit[28]	2022	0.66/1.44	0.20/0.36	0.31/0.60	0.25/0.50	0.22/0.43	0.33/0.67
Social-SAGAN[14]	2023	0.65/1.19	0.36/0.70	0.54/1.14	0.33/0.66	0.29/0.61	0.43/0.86
Tri-HGNN[19]	2023	0.62/0.86	0.38/0.65	0.49/0.88	0.27/0.44	0.25/0.40	0.40/0.65
Social TAG[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
RDGCN[30]	2023	0.58/0.94	0.30/0.45	0.35/0.65	0.28/0.48	0.25/0.44	0.35/0.59
MSTCNN[31]	2024	0.63/0.98	0.32/0.49	0.42/0.72	0.32/0.50	0.28/0.44	0.39/0.63
DSTCNN[32]	2024	0.53/1.08	**0.19/0.34**	0.29/0.53	0.23/0.43	0.23/0.43	0.29/0.53
本模型	-	**0.27/0.58**	0.28/0.55	**0.21/0.38**	**0.15/0.24**	**0.14/0.21**	**0.21/0.39**

总的来说，与表 1 中的其他模型相比，本模型获得了最好的 ADE 均值和最好的 FDE 均值，并且分别在 eth、univ、zara1、zara2 数据集上得到的 ADE 和 FDE 值均为表 1 中所有模型方法的最小值。因此，本模型具有一定的竞争力，在行人轨迹预测问题中具备良好的预测效果。

2.4 对比实验

2.4.1 特征融合方法比较

在特征融合阶段，如图 5 所示，利用四种融合方法聚合两个模块提取到的特征并进行了对比。具体来说，特征拼接方法是将两个模块提取的特征沿特征维度进行拼接；元素平均是将两个模块提取的特征取平均值；序列池化是在特征拼接的基础上对特征的时间维度作平均池化处理；自适应融合是根据两个模块提取到的特征对于预测准确率的贡献占比进行加权求和。从实验结果上来看，自适应融合的方式在平均 ADE 和平均 FDE 上都是最优的，因此本模型采用了自适应融合的方式作为最后的特征融合方案。

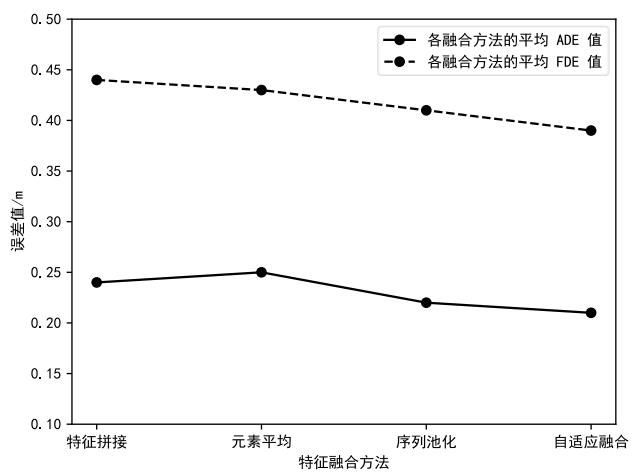


图 5　不同融合方法对比

Fig. 5　Comparison of Different Fusion Methods

2.4.2 序列建模方法比较

本模型在时空交互模块中引入了 TTT 层用于建模轨迹特征。TTT 层是最近提出的一种用于序列建模的网络层，为了评估 TTT 层与经典序列建模网络（如 LSTM 和 Transformer）的性能差异，本文设计了对比实验。具体而言，在保持时间依赖模块不变的情况下，将时空交互模块中的 TTT 层分别替换为 LSTM 和 Transformer 网络，并比较模型在不同场景下的预测性能。实验结果如表 2 所示。

从实验结果可以看出，使用 TTT 层作为时空交互模块的核心网络，在 eth、hotel、zara1 和 zara2 场景下，其预测性能显著优于 LSTM 和 Transformer 网络。此外，从平均指标来看，TTT 层的 ADE 和 FDE 相较于 Transformer 分别降低了 8.7% 和 7.1%，展现出更强的整体性能和泛化能力。然而，在 univ 场景中，TTT 层的预测性能略差于 Transformer，其 ADE 和 FDE 分别高出 5% 和 8.6%。这一现象可能归因于 univ 场景中轨迹间交互关系的复杂性较高，而 Transformer 的全局注意力机制更适合建模这种场景下的长距离依赖关系；相较之下，TTT 层更擅长于捕捉局部特征的动态变化，因此在该场景中表现相对较弱。

表 2　与经典序列建模网络的 ADE/FDE 对比

Table 2　Comparison of ADE/FDE with Classic Sequence Modeling Networks

序列建模网络	eth	hotel	univ	zara1	zara2	平均
LSTM	0.33/0.61	0.28/0.57	0.23/0.39	0.16/0.24	0.14/0.22	0.23/0.41
Transformer	0.27/0.60	0.34/0.70	0.20/0.35	0.16/0.24	0.16/0.21	0.23/0.42
TTT(本模型)	0.27/0.58	0.28/0.55	0.21/0.38	0.15/0.24	0.14/0.21	0.21/0.39

2.4.3 TCN与BiTCN比较

本文在时间依赖模块中采用 BiTCN 网络用于学习轨迹序列中的双向时间依赖关系，为了验证该通路的有效性，采用 TCN 网络替换模型中的 BiTCN 网络进行对比实验。如表 3 所示，BiTCN 具有更高的预测精度，因为双向结

构相比于单向结构能够更好的建模时间序列的长期依赖性，可以利用过去和未来的上下文信息以捕捉全局特征，从表中数据可见，相比于 TCN，BiTCN 的 ADE 和 FDE 的平均值分别降低了 4.5%和 7.1%，证明双向建模能够提高预测精度。因此本模型采用 BiTCN 网络。

表 3　TCN 与 BiTCN 的 ADE/FDE 比较

Table 3　Comparison of ADE/FDE Between TCN and BiTCN

方法	eth	hotel	univ	zara1	zara2	平均
TCN	**0.26**/0.59	0.30/0.60	0.22/0.39	0.17/**0.22**	0.17/0.32	0.22/0.42
BiTCN	0.27/**0.58**	**0.28/0.55**	**0.21/0.38**	**0.15**/0.24	**0.14/0.21**	**0.21/0.39**

2.4.4　模型参数量与推理时间的比较

从表 4 中数据可以看出，本模型通过结合 RNN 和 TCN 的优势，在参数量、推理时间以及建模能力之间实现了良好的平衡。与传统 RNN 模型（如 Social-LSTM 和 SR-LSTM）相比，本模型的参数量显著减少，仅为 Social-LSTM 的 22.4%，推理时间也减少至其 9%，表明模型的轻量化设计有效降低了计算资源的消耗。相比参数量较大的 Transformer 模型（如 STAR），本模型显著减少了计算复杂度，仅为其 6.1%，同时推理效率接近 CNN 模型，但保留了更强的时间依赖建模能力，适合复杂场景下的轨迹预测任务。此外，与轻量化的 CNN 模型（如 MSTCNN 和 Social-Implicit）相比，本模型虽然推理时间稍长，但与之相比本模型在时空特征建模的精度和泛化性上具备显著优势。因此，本模型不仅兼顾了高效性与轻量化设计，还在时间序列建模的表达能力上超越了现有方法，特别适合实际应用中对实时性和准确性要求较高的场景。整体而言，该模型在性能和效率上的平衡为轨迹预测研究提供了新的方向。

表 4　模型参数量和推理时间对比

Table 4　Comparison of Model Parameters and Inference Time

模型方法	类型	参数量/K	推理时间/s
Social-LSTM	RNN	264	0.1541
SR-LSTM	RNN	64.9	0.0708
STAR	Transformer	964.9	0.0214
Social-Implicit	CNN	5.8	0.0010
MSTCNN	CNN	3.1	0.0007
本模型	RNN+CNN	59.14	0.014

2.5　消融实验

表 5　模块消融实验

Table 5　Ablation Study of Modules

版本	方案			ADE/FDE					
	时间依赖模块	时空交互模块	自适应融合	eth	hotel	univ	zara1	zara2	平均
配置 1	√	×	×	0.30/0.64	0.33/0.58	0.27/0.45	0.18/0.25	0.18/0.22	0.25/0.43
配置 2	×	√	×	0.30/0.64	0.31/0.60	0.27/0.43	0.17/0.26	0.16/0.23	0.24/0.43
本模型	√	√	√	**0.27/0.58**	**0.28/0.55**	**0.21/0.38**	**0.15/0.24**	**0.14/0.21**	**0.21/0.39**

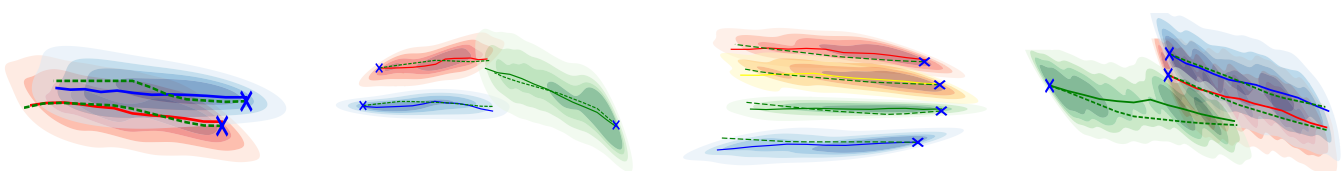
为了研究提出的方法的有效性，本文算法在 ETH 和 UCY 数据集上进行了通路消融实验。为了分析模型中两个模块结合的效果，将两个模块分别单独作为预测网络进行测试。由实验结果表 5 可知，当 TTT 网络单独作为主干网络进行预测时，其在 ETH 和 UNIV 场景下的 ADE 指标和 BiTCN 单独作为主干网络时的 ADE 指标相等，而在

HOTEL、ZARA1、ZARA2 上则拥有更低的 ADE；当 BiTCN 单独作为主干网络时，其在 ETH、HOTEL、ZARA1、ZARA2 场景下的 FDE 指标等于或优于单独的 TTT 网络预测时的 FDE 指标。因此通过该实验可知，TTT 网络对于全局预测的能力更强，而 BiTCN 对于轨迹端点的预测更优秀，通过结合以上两个模块，最终得到本文的算法模型，其具有比以上单个模块独立作用时更优秀的预测能力。

2.6 定性分析

2.6.1 多模态轨迹预测

图 6 展示了本模型在不同场景下的多模态轨迹预测结果，其中虚线表示真实的未来轨迹，实线表示预测的轨迹分布的均值。在双人并行场景 (a) 中，模型成功预测出两名行人平行前进的轨迹，且预测结果与真实轨迹高度一致，说明其对简单线性运动有较强的建模能力。在双人相遇场景 (b) 中，模型捕捉到了有两名行人交叉行走的趋势，并给出了合理的多模态轨迹分布，反映了对轨迹不确定性的良好建模能力。在多人并行场景 (c) 中，模型能够准确区分不同个体的轨迹，避免了预测结果的轨迹重叠，展现出对多人场景下轨迹独立性的良好建模表现。而在多人汇合场景 (d) 中，模型对多名行人轨迹汇聚的趋势进行了准确预测，预测结果具有较高的集中性和合理性。总体来看，本模型不仅能够捕捉行人的主要轨迹趋势，还可以通过多模态分布预测反映潜在的多种运动模式，验证了其在复杂和不确定性场景下的强大适应性和泛化能力。



(a)双人并行　(b)行人相遇　(c)多人并行　(d)多人汇合

图 6　多模态轨迹预测分布可视化

Fig.6　Visualization of multimodal trajectory prediction distributions

2.6.2 真实场景下的轨迹生成

在 2.1 节描述了 ETH 和 UCY 两个数据集的特点，在此从 ETH 和 UCY 数据集中分别选择了 eth 和 zara1 场景作为实例进行研究，如图 7 所示，其中间隔画线表示历史轨迹，实线表示真实的未来轨迹，点线表示预测的未来轨迹。从图中可以看出，该模型在行人轨迹预测中展现了多方面的优势。对于简单场景（如 a 和 b），预测轨迹与真实轨迹高度吻合，准确捕捉了行人的移动方向和路径，展现出较强的拟合能力。在复杂交互场景（如 c 和 d），模型能够清晰反映出交汇点、分离趋势以及转弯方向，即便存在轻微偏差，也能合理预测动态交互行为的趋势。在高密度场景（如 e 和 f），尽管行人轨迹密集，模型依然能够保持良好的分布特性，避免大规模重叠，表现出较高的鲁棒性和适应性。这些特点表明模型在不同场景中均具有较高的预测精度，尤其在常规场景中的准确性和复杂场景中的适应性，体现了本模型在行人轨迹预测任务中的优越性能。

(a) zara1 中双人并行　(b) eth 中双人并行　(c) zara1 中行人相遇　(d) eth 中行人相遇　(e) zara1 多人并行　(f) eth 中多人并行

图 7　真实场景下预测轨迹可视化

Fig 7　Visualization of Predicted Trajectories in Real-World Scenarios

3 结　论

本文提出一种融合双向时间建模与时空自监督学习的行人轨迹预测模型，通过 **BiTCN** 网络捕捉轨迹的双向时间依赖特征，借助 **TTT** 层的自监督机制动态优化时空交互特征，并通过自适应策略实现特征融合。实验表明，模型在 **ETH** 和 **UCY** 数据集上显著优于现有方法，尤其在密集动态场景中展现出对复杂交互模式的建模能力。

然而，模型在行人关联性较弱或人数较少的场景中，时空交互模块的自监督信号可能因交互信息不足而贡献受限，单轨迹长期依赖的建模精度仍有提升空间。未来研究可聚焦于强化时间维度的分层注意力机制，探索个体先验特征与物理运动学约束的融合，为该领域的进一步研究提供参考。

参 考 文 献：

[1] Yasuno M, Yasuda N, Aoki M. Pedestrian Detection and Tracking in Far Infrared Images[C/OL]//2004 Conference on Computer Vision and Pattern Recognition Workshop. 2004: 125-125[2025-05-07].

[2] Luo Y, Cai P, Bera A, et al. Porca: Modeling and planning for autonomous driving among many pedestrians[J]. IEEE Robotics and Automation Letters, 2018, 3(4): 3418-3425.

[3] Hu T, Long C, Xiao C. A Novel Visual Representation on Text Using Diverse Conditional GAN for Visual Recognition[J]. IEEE Transactions on Image Processing, 2021, 30: 3499-3512.

[4] Keller C G, Gavrila D M. Will the Pedestrian Cross? A Study on Pedestrian Path Prediction[J]. IEEE Transactions on Intelligent Transportation Systems, 2014, 15(2): 494-506.

[5] Schneider N, Gavrila D M. Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study[C]//WEICKERT J, HEIN M, SCHIELE B. Pattern Recognition. Berlin, Heidelberg: Springer, 2013: 174-183.

[6] Pavlovic V, Rehg J M, MacCormick J. Learning Switching Linear Models of Human Motion [C/OL]//Advances in Neural Information Processing Systems 13. MIT Press, 2000 [2025-05-07].

[7] Alahi A, Goel K, Ramanathan V, et al. Social LSTM: Human Trajectory Prediction in Crowded Spaces[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016: 961-971[2025-05-07].

[8] Zhang P, Ouyang W, Zhang P, et al. SR-LSTM: State Refinement for LSTM Towards Pedestrian Trajectory Prediction[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 12085-12094[2025-05-07].

[9] Krichen M. Generative Adversarial Networks[C/OL]//2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). 2023: 1-7[2025-05-07].

[10] Kipf T N, Welling M. Variational Graph Auto-Encoders[EB/OL]. arXiv, 2016[2025-05-07].

[11] Gupta A, Johnson J, Fei-Fei L, et al. Social GAN: Socially Acceptable Trajectories With Generative Adversarial Networks[C/OL]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 2255-2264[2025-05-07].

[12] Fang F, Zhang P, Zhou B, et al. Atten-GAN: Pedestrian Trajectory Prediction with GAN Based on Attention Mechanism[J]. Cognitive Computation, 2022, 14(6): 2296-2305.

[13] Kothari P, Alahi A. Safety-Compliant Generative Adversarial Networks for Human Trajectory Forecasting[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(4): 4251-4261.

[14] Yang C, Pan H, Sun W, et al. Social Self-Attention Generative Adversarial Networks for Human Trajectory Prediction[J]. IEEE Transactions on Artificial Intelligence, 2024, 5(4): 1805-1815.

[15] Kipf T N, Welling M. Semi-Supervised Classification with Graph Convolutional Networks[EB/OL]. arXiv, 2017[2025-05-07].

[16] Veličković P, Cucurull G, Casanova A, et al. Graph Attention Networks[EB/OL]. arXiv, 2018[2025-05-07].

[17] Mohamed A, Qian K, Elhoseiny M, et al. Social-STGCNN: A Social Spatio-Temporal Graph Convolutional Neural Network for Human Trajectory Prediction[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 14424-14432[2025-05-07].

[18] Lea C, Vidal R, Reiter A, et al. Temporal Convolutional Networks: A Unified Approach to Action Segmentation[C]//HUA G, JÉGOU H. Computer Vision – ECCV 2016 Workshops. Cham: Springer International Publishing, 2016: 47-54.

[19] Zhu W, Liu Y, Wang P, et al. Tri-HGNN: Learning triple policies fused hierarchical graph neural networks for pedestrian trajectory prediction[J]. Pattern Recognition, 2023, 143: 109772.

[20] 杨永鹏,杨震,杨真真.基于时序分解和注意力图神经网络的交通预测[J].数据采集与处理,2025,40(02):417-430.

[21] Sun Y, Li X, Dalal K, et al. Learning to (Learn at Test Time): RNNs with Expressive Hidden States[EB/OL]. arXiv, 2025[2025-05-07].

[22] Zhu W, Ma X, Liu Z, et al. MotionBERT: A Unified Perspective on Learning Human Motion Representations[C/OL]//2023 IEEE/CVF International Conference on Computer Vision (ICCV). Paris, France: IEEE, 2023: 15039-15053[2025-05-07].

[23] Sadeghian A, Kosaraju V, Sadeghian A, et al. SoPhie: An Attentive GAN for Predicting Paths Compliant to Social and Physical Constraints[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 1349-1358[2025-05-07].

[24] Tao C, Jiang Q, Duan L, et al. Dynamic and Static Context-Aware LSTM for Multi-agent Motion Prediction[C]//VEDALDI A, BISCHOF H, BROX T, et al. Computer Vision – ECCV 2020. Cham: Springer International Publishing, 2020: 547-563.

[25] Zhou H, Ren D, Xia H, et al. AST-GNN: An attention-based spatio-temporal graph neural network for Interaction-aware pedestrian trajectory prediction[J]. Neurocomputing, 2021, 445: 298-308.

[26] Shi L, Wang L, Long C, et al. SGCN: Sparse Graph Convolution Network for Pedestrian Trajectory Prediction[C/OL]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 8994-9003[2025-05-07].

[27] Zamboni S, Kefato Z T, Girdzijauskas S, et al. Pedestrian trajectory prediction with convolutional neural networks[J]. Pattern Recognition, 2022, 121: 108252.

[28] Mohamed A, Zhu D, Vu W, et al. Social-Implicit: Rethinking Trajectory Prediction Evaluation and The Effectiveness of Implicit Maximum Likelihood Estimation[C]//AVIDAN S, BROSTOW G, CISSÉ M, et al. Computer Vision – ECCV 2022. Cham: Springer Nature Switzerland, 2022: 463-479.

[29] Zhang X, Angeloudis P, Demiris Y. Dual-branch spatio-temporal graph neural networks for pedestrian trajectory prediction[J]. Pattern Recognition, 2023, 142: 109633.

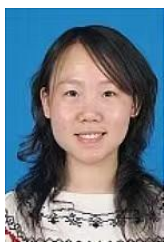
[30] Sang H, Chen W, Wang J, et al. RDGCN: Reasonably dense graph convolution network for pedestrian trajectory prediction[J]. Measurement, 2023, 213: 112675.

[31] Sang H, Chen W, Wang H, et al. MSTCNN: multi-modal spatio-temporal convolutional neural network for pedestrian trajectory prediction[J]. Multimedia Tools and Applications, 2024, 83(3): 8533-8550.

[32] Chen W, Sang H, Wang J, et al. DSTCNN: Deformable spatial-temporal convolutional neural network for pedestrian trajectory prediction[J]. Information Sciences, 2024, 666: 120455.

作者简介：

赵帅(1999-)，男，硕士研究生，研究方向：行人轨迹预测，E-mail:978317890@qq.com。



李琳(1983-)，**通信作者**，女，副教授，研究方向：多智能体系统控制，图像处理等，E-mail:lilin0211@163.com。