doi:10.19562/j.chinasae.qcgc.2024.11.004

基于多信息融合网络的行人轨迹预测方法

高 嵩[1,2],周江邻[2],高博麟[1],芦 健[2],王 鹤[2],徐月云[2]
(1. 清华大学车辆与运载学院,北京 100000; 2. 国家智能网联汽车创新中心,北京 102600)

[摘要] 随着自动驾驶技术的不断发展,准确预测行人的未来轨迹已经成为确保系统安全和可靠的关键要素。然而,现有行人轨迹预测研究多数依赖于固定摄像头视角,进而限制了对行人运动的全面观测,因此难以直接应用于自动驾驶车辆自车视角(ego-vehicle)下的行人轨迹预测。针对该问题,本文提出了一种基于多行人信息融合网络(MPIFN)的自车视角行人轨迹预测方法。该方法通过融合社会信息、局部环境信息和行人时间信息,实现了对行人未来轨迹的准确预测。本文构建了一个局部环境信息提取模块,结合了可形变卷积与传统卷积和池化操作,旨在更有效地提取复杂环境中的局部信息。该模块通过动态调整卷积核的位置,增强了模型对不规则和复杂形状的适应能力。同时,构建了行人时空信息提取模块和多模态特征融合模块,以实现对社会信息与环境信息的充分融合。实验结果表明,该方法在JAAD和PSI两个自车视角下驾驶数据集上均取得了先进的性能。在JAAD数据集上,累积均方误差(CF_MSE)为4 063,累积平均均方误差(C_MSE)为829。在PSI数据集上平均相对偏差(ARB)和最终相对偏差(FRB)也分别在预测时间为0.5、1.0、1.5 s时取得了18.08、29.21、44.98和25.27、54.62、93.09的优异表现。

关键词:自动驾驶;行人轨迹预测;多行人信息融合网络;自车视角

Pedestrian Trajectory Prediction Method Based on Multi-information Fusion Network

Gao Song[1,2], Zhou Jianglin[2], Gao Bolin[1], Lu Jian[2], Wang He[2] & Xu Yueyun[2]
1. *School of Vehicles and Mobility, Tsinghua University, Beijing* 100000;
2. *National Innovation Center of Intelligent and Connected Vehicles, Beijing* 102600

[**Abstract**] With the continuous development of autonomous driving technology, accurately predicting the future trajectories of pedestrians has become a critical element in ensuring system safety and reliability. However, most existing studies on pedestrian trajectory prediction rely on fixed camera perspectives, which limits the comprehensive observation of pedestrian movement and thus makes them unsuitable for direct application to pedestrian trajectory prediction under the ego-vehicle perspective in autonomous vehicles. To solve the problem, in this paper a pedestrian trajectory prediction method under the ego-vehicle perspective based on the Multi-Pedestrian Information Fusion Network (MPIFN) is proposed, which achieves accurate prediction of pedestrians' future trajectories by integrating social information, local environmental information, and temporal information of pedestrians. In this paper, a Local Environmental Information Extraction Module that combines deformable convolution with traditional convolutional and pooling operations is constructed, aiming to more effectively extract local information from complex environment. By dynamically adjusting the position of convolutional kernels, this module enhances the model's adaptability to irregular and complex shapes. Meanwhile, the pedestrian spatiotemporal information extraction module and multimodal feature fusion module are developed to facilitate comprehensive integration of social and environmental information. The experimental results show that the proposed method achieves advanced performance on two ego-vehicle driving datasets, JAAD and PSI. Specifically, on the JAAD dataset, the Center Final Mean Squared Er-

原稿收到日期为 2024 年 05 月 28 日,修改稿收到日期为 2024 年 07 月 01 日。
通信作者:高博麟,副研究员,博士,E-mail:gaobolin@tsinghua.edu.cn。

ror (CF_MSE) is 4 063, and the Center Mean Squared Error (C_MSE) is 829. On the PSI dataset, the Average Root Mean Square Error (ARB) and Final Root Mean Square Error (FRB) also achieve outstanding performance with values of 18.08/29.21/44.98 and 25.27/54.62/93.09 for prediction horizons of 0.5 s, 1.0 s, and 1.5 s, respectively.

Keywords: autonomous driving; pedestrian trajectory prediction; multiple pedestrian information fusion network; ego-vehicle

前言

预测行人的未来轨迹有助于提高自动驾驶汽车对驾驶风险的认知,并预估潜在的危险。例如,轨迹预测[1-3]系统能够在行人突然从车辆侧面或视野盲区闯入时,提前规划自动驾驶车辆行驶路径,并帮助车辆采取紧急制动或转向等应对措施。现有行人轨迹预测研究多数依赖于固定摄像头视角,限制了对行人运动的全面观察,因此难以直接应用于自动驾驶车辆的自车视角(ego-vehicle)下的行人轨迹预测。如表1所示,固定摄像头视角通常捕捉到的行人移动和环境信息是静态的,即环境布局和背景在一段时间内保持不变。而自动驾驶汽车自车视角(如车载摄像头)捕捉到的环境信息是动态的,行人、车辆、道路状况等因素都在不断变化。在自车视角下,行人的移动不仅受到其他行人和环境的影响,还受到车辆行驶状态、速度、转向等因素的直接影响。因此,基于自车视角的行人轨迹预测方法受到了广泛的关注[4-6]。

表1 固定视角与自车视角的差异对比

对比项	固定视角	自车视角
摄像头位置	固定于某一位置(如路口、建筑顶部)	安装于车辆上,随车辆移动
视野范围	有限的、特定的视野范围	宽广的、移动的视野范围
环境稳定性	环境相对固定,不随摄像头移动	环境随车辆移动而持续变化

在基于自车视角的行人轨迹预测领域,观察对象和车载摄像头同时运动,将车载摄像头作为参考点,须同时考虑观察对象的运动特征和环境信息,从而建立物体运动的数学模型[7-8]。通过将自车视角纳入轨迹预测过程,能够更好地理解和预测行人的运动行为,特别是在复杂的城市交通环境中。然而,尽管自车视角下的轨迹预测在推动自动驾驶和智慧交通建设方面取得了显著的进展,但其仍面临着诸多方面的挑战,如行人运动的随机性以及场景的复杂性。为应对挑战,研究人员提出了多种不同的方法和技术。

Bhattacharyya等[9]提出首个基于自车视角观测数据的行人轨迹预测模型,即Bayesian LSTM。该模型通过对观测轨迹的不确定性进行建模,从而推测目标行人未来位置的概率分布,进而达到长期行人轨迹预测的目的。Yao等[10]提出了一种多编码流RNN (recurrent neural network, RNN) 编码器-解码器模型,该模型能够预测自车视角视频中目标物体的相对位置,并捕捉邻近物体的外观。随着研究的深入,基于人体姿态特征和行人意图的方法开始受到了研究者们的广泛关注。Yagi等[11]在卷积-反卷积框架中利用人体姿态、行人尺度和相机运动等数据作为线索,预测行人未来的轨迹。Chen等[12]提出一种新颖的可解释性行人轨迹预测(eP2P)模型,该模型通过融合行人意图和驾驶行为来指导未来轨迹的生成。Rasouli等[13]提出行人意图、车辆速度和行人轨迹联合预测框架。该框架基于LSTM (long short-term memory, LSTM) 实现。然而,人体姿态关键点的检测精度受相机视角、人车距离、遮挡等因素的影响较大,且行人意图具有较高的可变性,因此基于人体姿态和行为意图的行人轨迹预测方法在复杂的城市交通环境中受限。

在行人轨迹预测领域,社会信息(social information)通常是指目标行人与其周围行人之间的历史交互或动态关系[1-2, 14]。这种信息包括但不限于行人之间的距离、速度、方向、避让行为、群体动态等。它们共同构成了行人在特定环境中的社会行为模式。环境信息(environmental information)定义为除行人自身运动状态之外的所有影响行人行为的外部因素[15-17]。这些因素可能包括道路类型、交通信号、周围行人和车辆的行为、天气状况等。Wang等[14]提出SGNet网络,该网络首先对行人间的社会信息进行建模,随后对每个行人在每一时刻的目标位置进行逐一估计,从而获得行人在所有时刻下的完整轨迹。Neumann等[15]使用自监督训练范式,通

过融合环境信息与行人轨迹从而推断出行人的未来轨迹。然而,上述方法仅将社会信息或环境信息单独地纳入到网络中,缺少对社会信息和环境信息的充分融合,导致网络信息损失,从而影响轨迹预测精度。

为了解决上述问题,本文提出一种基于多行人信息融合网络(multiple pedestrian information fusion network, MPIFN)的自车视角行人轨迹方法,实现了对社会信息和局部环境信息,以及行人时间信息的充分融合,并在学习过程中考虑了长距离依赖。首先,建立了一种局部环境信息提取模块,该模块包含一个空间编码器和一个时间编码器,依赖可形变卷积和传统卷积、池化操作对局部环境信息进行提取。其次,构建了行人时空信息提取模块,该模块结合时间注意力机制和空间注意力机制提升了模型捕获复杂场景行人社会信息的能力,并在学习的过程中充分考虑了长距离依赖;最后,设计了一种基于交叉注意力的多模态特征融合模块,实现对社会信息和局部环境信息的充分融合,为解码器预测模块提供了可靠信息。综上所述,本文的主要贡献可以归纳为:

(1)现有自车视角下的轨迹预测方法仅将社会信息或环境信息单独地融入到深度学习的网络中,缺少对社会信息和环境信息的融合。鉴于此,本文提出了一种基于多信息融合网络的自车视角行人轨迹预测方法。该方法通过联合学习社会信息和局部环境信息,旨在实现准确的行人轨迹预测。

(2)现有方法在处理自车视角下的环境信息时往往简单使用卷积神经网络提取图像特征,忽略了自车视角下环境信息的时空特性。为此,本文建立了一种局部环境信息提取模块,该模块包含一个空间编码器和一个时间编码器,通过可形变卷积、池化操作等捕捉更精确的局部环境时空特征。

(3)在自车视角下行人轨迹预测公开数据集JAAD和PSI上对提出的方法进行了验证,实验结果表明,与现有同类型方法相比,本文提出的方法具有更好的性能。

1　方法

1.1　问题描述

轨迹预测问题可以被视为时间序列预测任务[17-19]。在行人轨迹预测中,通常将目标行人及其周围行人的过去轨迹以及行人所处场景图像作为模型的输入,以此预测目标行人的未来轨迹。在本文中,对于目标行人i,给定自车视角下从观测时间步$t=1$到$t=t_{\text{obs}}$的二维边界框信息$\boldsymbol{X}=[\boldsymbol{X}_i^1,\boldsymbol{X}_i^2,\ldots,\boldsymbol{X}_i^{t_{\text{obs}}}]$及其场景图像序列$\boldsymbol{S}=[S_i^1,S_i^2,\ldots,S_i^{t_{\text{obs}}}]$,预测未来时间步$t=t_{\text{obs}}+1$到$t=t_{\text{pred}}$下的二维边界框信息$\widehat{\boldsymbol{Y}}=[\widehat{\boldsymbol{Y}}_i^{t_{\text{obs}}+1},\widehat{\boldsymbol{Y}}_i^{t_{\text{obs}}+2},\ldots,\widehat{\boldsymbol{Y}}_i^{t_{\text{pred}}}]$。其中$\boldsymbol{X}_i^t=(u_i^t,v_i^t,r_i^t,o_i^t)\in\mathrm{R}^4$、$\widehat{\boldsymbol{Y}}_i^t=\left(\widehat{u}_i^t,\widehat{v}_i^t,\widehat{r}_i^t,\widehat{o}_i^t\right)\in\mathrm{R}^4$表示行人$i$在$t$时刻边界框左上角和右下角的二维像素坐标,$S_i^t\in\mathrm{R}^{t\times h\times w\times c}$表示行人$i$在$t$时刻的场景图像。

1.2　框架概览

所提出的MPIFN模型架构如图1所示。该模型包含3个主要部分:(1)由空间编码器和时间编码器组成的局部环境信息提取模块。在空间编码器部分,与传统使用2D卷积网络(如ResNet34)获取帧级(frame-level)特征$f_{\text{s}}\in\mathbb{R}^{t_{\text{obs}}\times h\times w\times c}$(其中,$h$、$w$、$c$分别表示特征$f_{\text{s}}$的长、宽及通道数)方法不同,考虑到自车视角下行人的位置和姿态不断变化,传统固定形状的卷积核无法有效地提取局部环境信息,因此在空间编码器部分采用可形变卷积(deformable convolution)[20]的策略缓解自车视角下行人存在的"近大远小"的问题。对于时间编码器,利用卷积神经网络(CNN)生成局部环境信息$f_{\text{st}}\in\mathbb{R}^{1\times h\times w\times c}$。(2)结合时间注意力和空间注意力的时空信息提取模块。在该模块中,行人二维边界框信息$\boldsymbol{X}=[\boldsymbol{X}_i^1,\boldsymbol{X}_i^2,\ldots,\boldsymbol{X}_i^{t_{\text{obs}}}]$首先被映射到高维空间$\boldsymbol{X}_{\text{emb}}$,然后通过时间注意力和空间注意力学习时间信息$l_i^t$和社会信息$l_{\text{st}}$。(3)基于交叉注意力的多模态特征融合模块。在该模块中,局部环境信息f_{st}与社会信息l_{st}之间的权重通过交叉注意力机制计算得到,并通过交叉注意力机制对局部环境信息与社会信息进行加权求和,最终得到融合后的特征。下面对该模型中各个模块进行详细介绍。

1.3　局部环境信息提取模块

局部环境信息是指目标行人在静态或动态场景中所处环境的语义信息。传统方法在提取局部环境信息时通常基于静态场景假设。然而,由于自车视角下的行人场景是动态变化的,这种假设并不适用于实际交通场景。因此,本文提出了一种局部环境信息提取模块,该模块包含一个空间编码器和一个时间编码器,通过可形变卷积和传统卷积、池化操作对局部环境信息进行提取。局部环境信息提取模块的网络结构如图2所示。

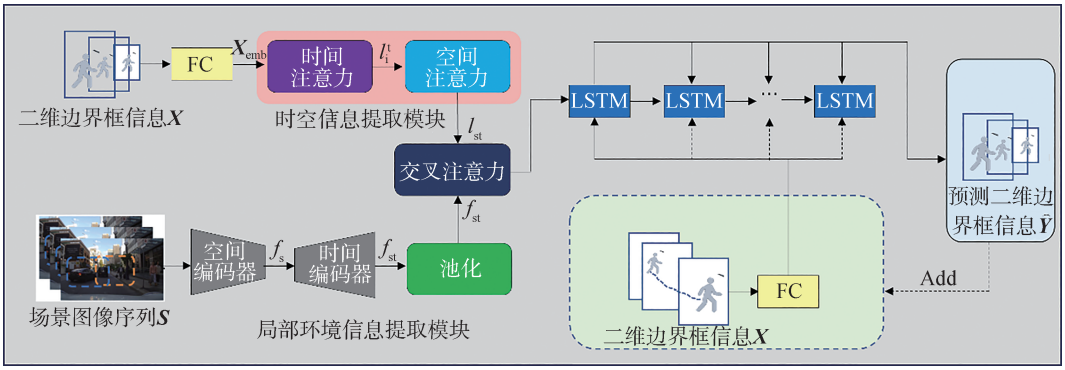

图 1 本文所提 MPIFN 框架图

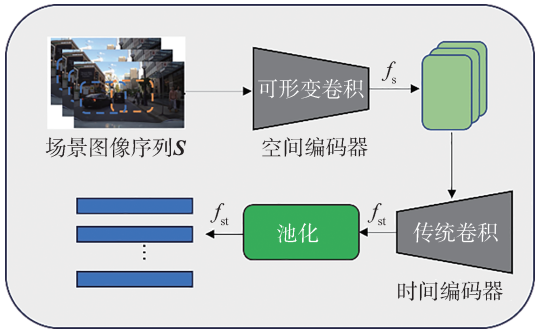

图 2 局部环境信息提取模块

首先,通过对抖动后的二维边界框进行剪裁,获取行人的局部环境图像序列 $\boldsymbol{S}=[S_i^1,S_i^2,\ldots,S_i^{t_{obs}}]$,并调整边界框尺寸使其高度与宽度相匹配,从而作为空间编码器的输入。具体而言,如果剪裁后的图像的高宽比大于或小于预设的比例(例如 1∶1),本文根据需要进行缩放操作。如果高度大于宽度,将按照宽度等比例缩小图像的高度;反之,如果宽度大于高度,则按照高度等比例缩小图像的宽度。这样处理后,图像的高度和宽度将相等,从而满足空间编码器的输入要求。与此同时,空间编码器通过一个可学习的偏移量调整卷积核上局部感受野的位置和形状,以匹配目标物体的形状和大小,从而得到图片序列的帧级特征 $f_s\in\mathbb{R}^{t_{obs}\times h\times w\times c}$。该过程可以表示为

$$f_s=\sum_{p_n\in R}\left(w(p_n)*x(p_0+p_n+\Delta p_n)\right) \tag{1}$$

式中:p_0 表示卷积窗口的中心位置;p_n 是距离 p_0 不超过 1 个像素的相对位置;Δp_n 为可学习的偏移量;R 表示卷积窗口的大小;$w(\cdot)$ 表示取出对应位置的权重值;$x(\cdot)$ 表示取出对应位置的像素值。

随后,将帧级特征 f_s 输入到时间编码器,得到局部环境信息 f_{st}。具体而言,使用通道数 $c=64$ 的 3D 卷积滤波器 $t_{obs}\times 1\times 1$,沿着 f_s 的时间维度方向对其进行卷积操作,从而得到局部环境信息 $f_{st}\in\mathbb{R}^{1\times h\times w\times c}$。这里,采用具有较长时序记忆的 3D 卷积滤波器 $t_{obs}\times 1\times 1$,可以更好地捕捉时间序列中的长期依赖关系。该过程可以表示为

$$f_{st}=\text{Conv3D}\left(f_s,t_{obs}\times 1\times 1\right) \tag{2}$$

式中:Conv3D 表示 3D 卷积操作;f_s 是输入的帧级特征;$t_{obs}\times 1\times 1$ 是具有 64 个通道的 3D 卷积滤波器。

为了调节数据维度并降低数据冗余,环境信息 f_{st} 进一步通过池化层进行处理。传统的池化操作是在每个通道上独立进行,没有考虑不同通道之间的关联。通道池化[21]则是对每个通道的特征进行整体池化,使特征图变得更加平滑从而减少噪声和冗余信息,但通道池化会将每个通道内的信息压缩成一个标量,进而导致细节信息的丢失。为了继承上述方法的优点,同时避开其缺点,局部环境信息 f_{st} 首先由通道池化进行处理,保留有用特征并压缩冗余信息;随后,利用最大池化降低数据维度,得到池化后的环境信息 f_{st}^p。该方法有效地平衡了特征保留和维度降低之间的矛盾,为后续预测提供了更为准确的环境信息表示。

1.4 时空信息提取模块

在现有的自车视角下的行人轨迹预测方法中,通常采用 RNN[22]或 LSTM[23]来提取行人的时空信息并对其轨迹进行建模。尽管 LSTM 在处理长序列时有所改进,但它们仍然很难建模长距离依赖关系。为了缓解该问题并增强模型捕获复杂场景下行人社会信息的能力,本文提出了一种结合时间注意力和空间注意力的行人时空信息提取模块。该模块在学习过程中,充分考虑了长距离依赖,能够有效地从复杂的行人轨迹数据中提取有用的信息。

图3展示了时空信息提取模块的网络架构。在图3中,输入向量首先通过一个线性变换层得到所需的查询矩阵(Query)和键值对矩阵(Key-Value)。然后通过点积计算查询矩阵和键值对矩阵之间的相似度,并计算每个键值对对应的权重。最后,这些权重被用来加权求和对应的值矩阵(Value),从而得到最终的输出向量:

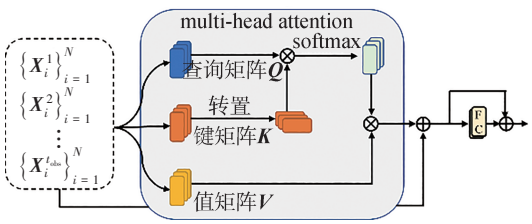

图3　时空信息提取模块

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{3}$$

式中:$\sqrt{d}$用于归一化以确保数值稳定性;$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$分别表示查询矩阵、键矩阵和值矩阵。此外,为了捕捉输入序列中不同子空间复杂的交互关系,多头注意力机制(multi-head attention)在上述机制的基础上进行了扩展和优化。在多头注意力机制中,每个头都会独立地计算Query和每个Key的点积,从而得到一个权重分布。这些权重分布会被用来加权求和对应的Value向量,得到输出向量,该输出向量是多个头计算结果的拼接:

$$\begin{cases}\boldsymbol{Q}_j=\boldsymbol{Q}\boldsymbol{W}_j^{Q},\boldsymbol{K}_j=\boldsymbol{K}\boldsymbol{W}_j^{K},\boldsymbol{V}_j=\boldsymbol{V}\boldsymbol{W}_j^{V}\\ head_j=Att(\boldsymbol{Q}_j,\boldsymbol{K}_j,\boldsymbol{V}_j),j=1,\ldots,n\\ MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\\ \quad Concat(head_1,head_2,\ldots,head_n)\boldsymbol{W}^{o}\end{cases} \tag{4}$$

式中:n表示多头数;$\boldsymbol{W}_j^{Q}$、$\boldsymbol{W}_j^{K}$、$\boldsymbol{W}_j^{V}$以及$\boldsymbol{W}^{o}$为对应的参数矩阵。

在上述理论的指导下,时空信息提取模块首先对第i个行人在观测时间步$t\in[1,t_{\text{obs}}]$的二维边界框信息$\boldsymbol{X}=[\boldsymbol{X}_i^1,\boldsymbol{X}_i^2,\ldots,\boldsymbol{X}_i^{t_{\text{obs}}}]$通过全连接层映射到高维空间:

$$\boldsymbol{X}_{\text{emb}}=\phi(\boldsymbol{X};\ \boldsymbol{W}) \tag{5}$$

式中:$\phi(\cdot)$具有ReLU激活的线性嵌入函数;$\boldsymbol{W}$表示嵌入权重。

由于行人的移动轨迹具有时序性,每个时间步的轨迹点都可能对预测未来的轨迹产生影响,但不同时间步的影响程度可能不同。时间注意力机制能够为每个时间步的轨迹点分配不同的权重,使得模型能够更加关注那些对预测未来轨迹更为关键的时间步。因此,利用时间注意力机制挖掘行人的时间信息l_i^t,学习序列长距离依赖关系,以强化模型对数据的处理能力。这一过程可表示为

$$l_i^t=Att(\boldsymbol{X}_{\text{emb}})=\text{softmax}\left(\frac{(\boldsymbol{X}_{\text{emb}}\boldsymbol{W}_{\text{q}})(\boldsymbol{X}_{\text{emb}}\boldsymbol{W}_{\text{k}})^{\mathrm{T}}}{\sqrt{d}}\right)(\boldsymbol{X}_{\text{emb}}\boldsymbol{W}_{\text{v}}) \tag{6}$$

式中:$\boldsymbol{X}_{\text{emb}}=[\boldsymbol{X}_i^1,\boldsymbol{X}_i^2,\cdots,\boldsymbol{X}_i^t]$;$\boldsymbol{W}_q$、$\boldsymbol{W}_k$和$\boldsymbol{W}_v$表示参数矩阵。

与此同时,空间注意力关注于空间位置的重要性。由于行人在移动过程中会受到周围环境、其他行人等多种因素的影响,因此不同空间位置的信息对预测行人的未来轨迹具有不同的重要性,而空间注意力机制能够为不同的空间位置分配不同的权重,使得模型能够更加关注那些对预测未来轨迹更为关键的空间位置。因此,利用空间注意力机制捕捉行人间的空间信息$l_{\text{st}}=Att(l_i^t)$,从而完善模型对社会信息的提取。其中$\boldsymbol{l}_t=[l_1^t,l_2^t,\cdots,l_N^t]$。值得注意的是,时空信息提取模块在挖掘行人的时间信息和行人间的空间信息时,充分考虑了长距离依赖关系,有助于模型更好地理解行人在真实世界中的交互行为,从而更准确地捕捉社会信息。

1.5　多模态特征融合模块

在多人场景下的轨迹预测中,行人与行人之间、行人与环境之间的交互信息对预测行人未来轨迹至关重要。为了充分融合局部环境信息和社会信息,本文利用交叉注意力机制设计了一个多模态特征融合模块。具体而言,局部环境信息f_{st}与社会信息l_{st}之间的权重通过交叉注意力机制计算得到,并通过该机制对这两种信息进行加权求和,从而得到融合后的特征:

$$CATT(f_{\text{st}},l_{\text{st}})=\text{softmax}\left(\frac{(f_{\text{st}}\boldsymbol{W}_q')(l_{\text{st}}\boldsymbol{W}_k')}{\sqrt{d}}\right)l_{\text{st}}\boldsymbol{W}_v' \tag{7}$$

式中$\boldsymbol{W}_q'$、$\boldsymbol{W}_k'$和$\boldsymbol{W}_v'$表示参数矩阵。

需要说明的是,本文在空间和时间注意力以及交叉注意力中均采用了如式(4)所述的多头注意机制,以捕捉来自不同子空间的信息。

为了对交叉注意力的信息进行解码,融合后的特征将作为解码器LSTM的初始隐藏状态。二维边界框信息$\boldsymbol{X}=[\boldsymbol{X}_i^1,\boldsymbol{X}_i^2,\ldots,\boldsymbol{X}_i^{t_{\text{obs}}}]$经过全连接层FC处理后将作为LSTM的输入序列。通过LSTM的解码过程,模型能够捕捉序列中的时间依赖性和上下文信息,从而更新隐藏状态。更新后的隐藏状态将进

一步通过全连接层进行解码,从而生成预测的二维边界框坐标 $\widehat{\boldsymbol{Y}} = [\widehat{\boldsymbol{Y}}_i^{t_{\text{obs}}+1}, \widehat{\boldsymbol{Y}}_i^{t_{\text{obs}}+2}, \ldots, \widehat{\boldsymbol{Y}}_i^{t_{\text{pred}}}]$。

本文采用均方值误差(mean square error, MSE)损失函数优化深度网络。均方误差是一种常用的回归问题损失函数,它通过计算预测值与实际值之间差的平方均值来衡量预测的准确性。本文采用均方值误差衡量网络预测的行人轨迹与实际轨迹之间的差异:

$$L_{\text{traj}} = \frac{1}{N \cdot t_{\text{pred}}} \sum_{i=1}^{N} \sum_{t=t_{\text{obs}}+1}^{t_{\text{pred}}} \left\| \boldsymbol{Y}_i^t - \widehat{\boldsymbol{Y}}_i^t \right\|_2 \tag{8}$$

式中:N表示场景中行人的数量;$\boldsymbol{Y}_i^t$表示行人i在t时刻真实边界框左上角和右下角的二维像素坐标。

2 实验设计

本节将介绍所使用的数据集和评估指标,并详细说明具体的实施细节。随后,通过与其他方法进行比较,证明本文所提出方法的优越性能。除此之外,为验证本文所提出方法关键组件的有效性,还进行了消融实验。最后,通过可视化结果直观地验证了本文所提出方法的效果。

2.1 公开数据集介绍

自动驾驶联合注意力(joint attention for autonomous driving, JAAD[24])数据集专注于从车辆视角捕捉驾驶场景,它由一系列不连续的视频片段构成。该数据集详细记录了4 365段行人的运动轨迹,并为每条轨迹提供了精确的二维边界框标注。此外,它还包含了各种情境下行人穿越行为的意图分类,为研究提供了丰富的信息资源。根据文献[4],将数据集分为训练集(50%)、验证集(10%)和测试集(40%)。

IUPUI-CSRC行人意图(PSI)[12]数据集是另一个自车视角下的驾驶数据集,并引入了两个独特的任务标签。首先,它记录了行人在车辆视角下的动态意图变化;其次,它提供了行人意图的详细解读。这些创新的标签使得数据集适用于多种计算机视觉任务,包括行人轨迹预测、行人意图识别、车辆与行人交互分割以及视频到语言的转换等。本文对PSI数据集中的110个视频进行了划分:前75个视频用于训练,第76至80个视频用于验证模型的性能,而剩余的视频则用于最终的测试。

本文采用3种常用的性能评估指标,即平均位移误差(average displacement error, ADE)、最终位移误差(final displacement error, FDE)和均方误差(mean squared error, MSE)来评估预测结果的准确性。其中,ADE是指预测值与真实值在所有观测时间步下的平均差距,FDE是指预测值与真实值在时间步t_{pred}下的误差,MSE则是预测值与真实值之间的平均平方误差:

$$\begin{cases} ADE = \dfrac{\sum_{i=1}^{N} \sum_{t=t_{\text{obs}}+1}^{t_{\text{pred}}} \left\| \boldsymbol{Y}_i^t - \widehat{\boldsymbol{Y}}_i^t \right\|_2}{N \times \left[t_{\text{pred}} - t_{\text{obs}} \right]} \\ FDE = \dfrac{\sum_{i=1}^{N} \left\| \boldsymbol{Y}_i^{t_{\text{pred}}} - \widehat{\boldsymbol{Y}}_i^{t_{\text{pred}}} \right\|_2}{N} \end{cases} \tag{9}$$

式中:Y_i^t表示真实的二维边界框坐标;$\widehat{Y}_i^t$表示预测的二维边界框坐标;$\| \cdot \|_2$表示l_2范数。

在实验中,使用MSE、中心均方误差(center mean squared error, C_{MSE})和中心最终均方误差(center final mean squared error ,CF_{MSE})评估模型在JAAD数据集上的性能,其中C_{MSE}和CF_{MSE}的计算与ADE/FDE相似[14-16];使用ADE、FDE、ARB和FRB评估模型在PSI数据集上的性能,其中ARB和FRB分别为边界框坐标的平均RMSE和最终RMSE[12]。

对于所有注释帧,本文利用观察到的前15帧序列作为输入,预测目标行人在随后0.5 s(后15帧)、1.0 s(后30帧)和1.5 s(后45帧)的边界框位置。

2.2 实施细节

在本文中,局部环境图像序列S的维度为15 × 3 × 224 × 224,使用64个大小为3 × 3的可形变卷积窗口对其处理,因此式(1)中f_s的维度为64 × 15 × 224 × 224。f_s通过c = 64通道的3D卷积滤波器,得到维度大小为64 × 1 × 224 × 224的局部环境信息f_{st}。此外,通道池化与最大池化的池化核大小分别设置为1 × 64和14 × 14。对于时空信息提取模块,式(5)中$\boldsymbol{X}_{\text{emb}}$的维度设置为256- d。时间注意力、空间注意力以及多模态特征融合模块中的交叉注意力的多头数都设置为8。

网络使用随机梯度下降(stochastic gradient descent, SGD[25])优化器进行优化,学习率初始化为0.001。batchsize设置为64 ,epoch设置为300。

2.3 实验结果

JAAD实验结果见表2,表中数值越小,性能越好。最佳和次佳的结果分别以粗体和下划线标出。Soc、Env和Soc-Env分别代表社会信息、局部环境信息和两者的充分融合。从表2中可以看到,所提出

的MPIFN充分融合了社会信息和环境信息,因而在JAAD数据集上与FOL-X[22]、PIE_{traj}[13]、BiTrap[4]相比取得了较好的效果。具体而言,MPIFN在C_{MSE}和CF_{MSE}上优于SGNet 16.7%和0.03%,这得益于MPIFN实现了对社会信息和局部环境信息以及行人时间信息的充分融合。MPIFN与SGNet[15]相比较在MSE(0.5 s、1.0 s)、C_{MSE}和CF_{MSE}上均取得了较好结果。

表2　在JAAD上实验结果

方法	Soc	Env	Soc-Env	PSI		
				MSE↓(0.5 s/1.0 s/1.5 s)	CF_{MSE}↓(1.5 s)	C_{MSE}↓(1.5 s)
Bayesian-LSTM[9]	√			159/539/1 535	5 615	1 447
FOL-X[26]		√		147/484/1 374	4 924	1 290
PIE_{traj}[13]		√		110/399/1 280	4 780	1 183
BiTraP[4]	√			93/378/1 206	4 565	1 105
eP2P[12]	√	√				
SGNet[15]	√			82/328/**1 049**	4 076	996
CVTF[27]		√		98/314/1 190	4 520	1 022
Pedformer[28]	√	√	√	93/364/1 134	4 364	1 080
VOSTN[29]	√			94/364/1 134	**3 980**	947
MPIFN(Ours)	√	√	√	**81/307**/1 106	4 063	**829**

PSI实验结果见表3。表中数值越小,性能越好。最佳和次佳的结果分别以粗体和下划线标出。由表3可见,MPIFN同样优于现有方法,并且优势较大。主要原因为:(1)PSI数据集包含的序列较长,本文提出方法学习长距离依赖关系,因此预测网络能够更好地捕捉到自车视角下行人轨迹的变化;(2)MPIFN能够更有效地对社会信息和局部环境信息进行挖掘。

表3　在PSI上实验结果

方法	Soc	Env	Soc-Env	PSI			
				ADE↓(0.5 s/1.0 s/1.5 s)	FDE↓(0.5 s/1.0 s/1.5 s)	ARB↓(0.5 s/1.0 s/1.5 s)	FRB↓(0.5 s/1.0 s/1.5 s)
Bayesian-LSTM[9]	√						
FOL-X[26]		√					
PIE_{traj}[13]		√					
BiTraP[4]	√						
eP2P[12]	√	√		22.67/31.07/44.90	27.76/52.03/93.14	27.12/35.03/48.51	31.59/55.08/95.97
SGNet[15]	√						
本文	√	√	√	**10.00/17.13/27.67**	**14.66/34.56/62.21**	**18.08/29.21/44.98**	**25.27/54.62/93.99**

2.4　消融实验

在本节中,将详细阐述在JAAD数据集上进行的消融实验结果,以验证所提方法中关键模块的有效性。表4列出了消融实验的结果,表中数值越小,性能越好。其中Bbox表示行人二维检测框,Image表示局部环境图像序列,Concat表示多模态特征融合模块由拼接操作构成,FC表示全连接层,符号'-'意味着不使用该模块。

从表4中可以得出如下结论:(1)Variant 2在性能上比Variant 1高14.4%,这表明局部环境图像序列的加入,有利于模型学习到更多可用信息;(2)在将时空信息提取模块由LSTM替换成时空注意力后,Variant 3的结果大幅领先于Variant 2,这反映出时空注意力在学习长序列依赖关系的有效性;(3)Variant 4的结果优于Variant 3,表明可形变卷积在提取动态场景下的序列的帧级特征具有独特优势;(4)Variant 5的结果在MSE(1.5 s)上领先Variant 48.9%,证明使用基于交叉注意力的多模态特征融合模块,对社会信息和局部环境信息的充分融合有积极影响;最后,在将解码器换成LSTM后,性能得到进一步提高。

表4 JAAD上消融实验结果

项目	输入	时空信息提取模块 + 局部环境信息提取模块 + 多模态特征融合模块	解码器	MSE(1.5 s)
Variant1	Bbox	LSTM + − + −	FC	2 458
Variant2	Bbox+Image	LSTM + 传统卷积 + Concat	FC	2 103
Variant3	Bbox+Image	时空注意力 + 传统卷积 + Concat	FC	1 502
Variant4	Bbox+Image	时空注意力 + 可形变卷积 + Concat	FC	1 435
Variant5	Bbox+Image	时空注意力 + 可形变卷积 + 交叉注意力	FC	1 307
本文	Bbox+Image	时空注意力 + 可形变卷积 + 交叉注意力	LSTM	**1 106**

2.5 可视化

图4和图5展示了MPIFN在JAAD和PSI数据集上可视化结果,图中红色为地面真实位置,绿色为预测结果,最佳观看效果为放大后的彩色框。由于MPIFN充分融合了社会信息和局部环境信息,并在学习过程中考虑了长距离依赖,因此MPIFN的预测结果接近真实实况。图6展示了所提出方法与其它方法的对比效果,图中黄色框/虚线表示真实位置/真实轨迹,其余彩色框/虚线表示预测位置/预测轨迹。

在JAAD数据集的可视化对比结果中(图6上半部分),左侧图像展示了一个行人走在街道上的场景,黄色虚线表示行人的实际轨迹。可以看到,本文提出的方法(绿色虚线)与实际轨迹的吻合度较高,而ePP2P[12](红色虚线)和VOSTN[29](蓝色虚线)存在一定偏差。右侧图像展示了一个行人过马路的场景,所提出方法的预测轨迹与实际轨迹比较接近,相比之下,ePP2P和VOSTN的预测轨迹与实际轨迹存在一定的偏差。在PSI数据集的可视化对比结果中(图6下半部分),左侧图像展示了一个行人从右侧走向左侧的场景,右侧图像展示了一个行人从中间向右侧移动的场景,可以看到,所提出方法在复杂场景中的良好表现,与真实轨迹大体一致。

图4 在JAAD数据集上的可视化结果

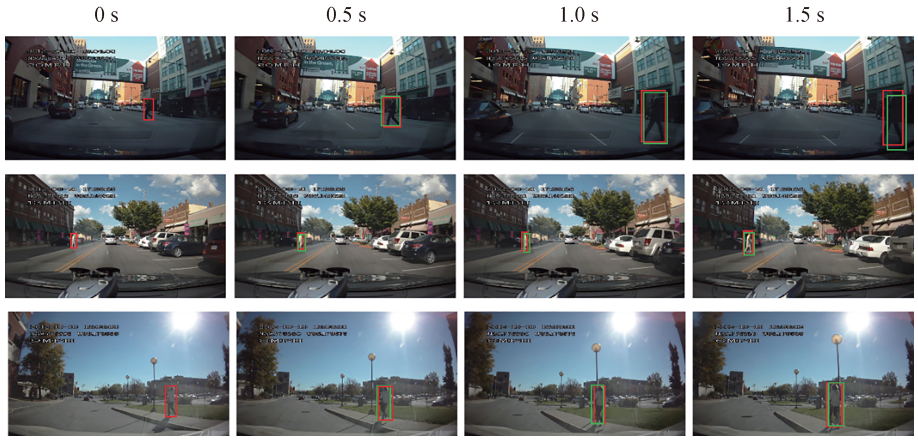

图5 在PSI数据集上的可视化结果

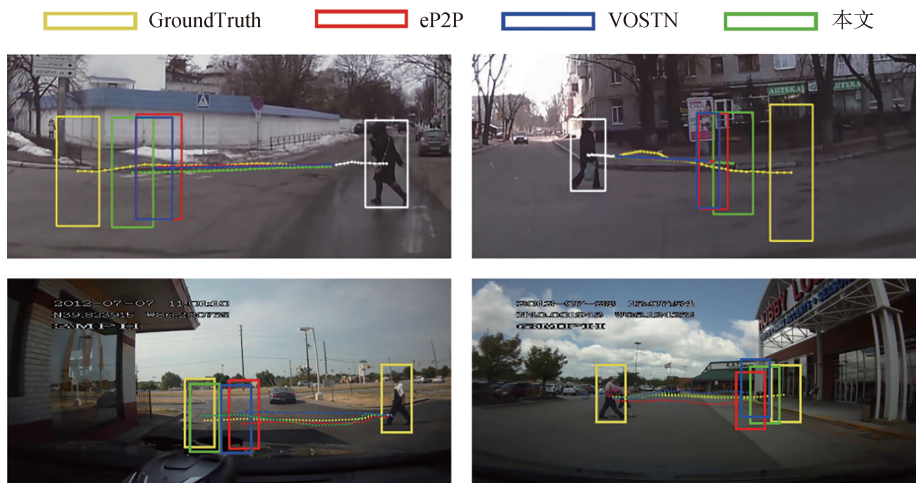

图6　在JAAD数据集(上)和PSI数据集(下)的可视化对比结果

3　结论

本文提出了一种基于多行人信息融合网络(MPIFN)的自车视角行人轨迹预测方法。该方法实现了对社会信息、局部环境信息以及行人时间信息的充分融合,并在学习过程中考虑了长距离依赖。具体来说,局部环境信息提取模块利用可形变卷积和传统卷积、池化操作来提取局部环境信息;时空信息提取模块结合时间注意力和空间注意力机制,提高了模型捕获复杂场景下行人社会信息的能力。基于交叉注意力的多模态特征融合模块,实现了对社会信息和局部环境信息的充分融合,为解码器预测模块提供了可靠的输入信息。在两个公开数据集上的实验结果证明了所提出方法的有效性。

参考文献

[1] 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.

[2] SUN J, LI Y, CHAI L, et al. Modality exploration, retrieval and adaptation for trajectory prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(12): 15051-15064.

[3] SHI L, WANG L, LONG C, et al. Representing multimodal behaviors with mean location for pedestrian trajectory prediction[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(9): 11184-11202.

[4] YAO Y, ATKINS E, JOHNSON-ROBERSON M, et al. BiTraP: bi-directional pedestrian trajectory prediction with multi-modal goal estimation[J]. IEEE Robotics and Automation Letters, 2021, 6(2): 1463-1470.

[5] 郭景华, 何智飞, 罗禹贡, 等. 人机混驾环境下基于深度学习的车辆切入轨迹预测[J]. 汽车工程, 2022, 44(2): 153-160.
GUO J H, HEI Z F, LUO Y G, et al. Vehicle cut-in trajectory prediction based on deep learning in a human-machine mixed driving environment [J]. Automotive Engineering, 2022, 44(2): 153-160.

[6] 郭景华, 肖宝平, 王靖瑶, 等. 基于 Residual BiLSTM 网络的车辆切入意图预测研究[J]. 汽车工程, 2021, 43(7): 971-977.
GUO J H, XIAO B P, WANG J Y, et al. Study on vehicle cut-in intention prediction based on residual BiLSTM network [J]. Automotive Engineering, 2021, 43(7): 971-977.

[7] GODARD C, MAC AODHA O, BROSTOW G J. Unsupervised monocular depth estimation with left-right consistency[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 270-279.

[8] CHEN Y, SCHMID C, SMINCHISESCU C. Self-supervised learning with geometric constraints in monocular video: connecting flow, depth, and camera[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 7063-7072.

[9] BHATTACHARYYA A, FRITZ M, SCHIELE B. Long-term on-board prediction of people in traffic scenes under uncertainty[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4194-4202.

[10] YAO Y, XU M, CHOI C, et al. Egocentric vision-based future vehicle localization for intelligent driving assistance systems[C]. 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 9711-9717.

[11] YAGI T, MANGALAM K, YONETANI R, et al. Future person localization in first-person videos[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 7593-7602.

[12] CHEN T, JING T, TIAN R, et al. Psi: a pedestrian behavior dataset for socially intelligent autonomous car [J]. arXiv preprint

arXiv: 2112.02604, 2021.

[13] RASOULI A, KOTSERUBA I, KUNIC T, et al. Pie: a large-scale dataset and models for pedestrian intention estimation and trajectory prediction[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6262-6271.

[14] WANG C, WANG Y, XU M, et al. Stepwise goal-driven networks for trajectory prediction[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 2716-2723.

[15] NEUMANN L, VEDALDI A. Pedestrian and ego-vehicle trajectory prediction from monocular camera[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 10204-10212.

[16] SU Y, LI Y, WANG W, et al. A unified environmental network for pedestrian trajectory prediction[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2024, 38(5): 4970-4978.

[17] FU Z, JIANG K, XIE C, et al. Summary and reflections on pedestrian trajectory prediction in the field of autonomous driving [J]. IEEE Transactions on Intelligent Vehicles, 2024.

[18] HASAN F, HUANG H. MALS-Net: a multi-head attention-based LSTM sequence-to-sequence network for socio-temporal interaction modelling and trajectory prediction [J]. Sensors, 2023, 23(1): 530.

[19] 桑海峰, 赵梓杉, 王金玉, 等. 基于车辆轨迹预测对抗性攻击与鲁棒性研究[J]. 汽车工程, 2024, 46(3): 407-417.
SANG H F, ZHAO Z S, WANG J Y, et al. Research on adversarial attacks and robustness in vehicle trajectory prediction [J]. Automotive Engineering, 2024, 46(3): 407-417.

[20] DAI J, QI H, XIONG Y, et al. Deformable convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2017: 764-773.

[21] DU J, WANG S, MIAO H, et al. Multi-channel pooling graph neural networks[C]. IJCAI. 2021: 1442-1448.

[22] GROSSBERG S. Recurrent neural networks [J]. Scholarpedia, 2013, 8(2): 1888.

[23] GRAVES A, GRAVES A. Long short-term memory[J]. Supervised Sequence Labelling with Recurrent Neural Networks, 2012: 37-45.

[24] RASOULI A, KOTSERUBA I, TSOTSOS J K. Are they going to cross? a benchmark dataset and baseline for pedestrian crosswalk behavior[C]. Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017: 206-213.

[25] KOSARAJU V, SADEGHIAN A, MARTÍN-MARTÍN R, et al. Social-bigat: multimodal trajectory forecasting using bicycle-gan and graph attention networks[J]. Advances in Neural Information Processing Systems, 2019, 32.

[26] YAO Y, XU M, WANG Y, et al. Unsupervised traffic accident detection in first-person videos[C]. 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2019: 273-280.

[27] HE Y, YANG Y, CAI Y, et al. Predicting pedestrian tracks around moving vehicles based on conditional variational transformer [J]. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 2023: 09544070231175536.

[28] RASOULI A, KOTSERUBA I. PedFormer: pedestrian behavior prediction via cross-modal attention modulation and gated multitask learning[C]. 2023 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2023: 9844-9851.

[29] WANG J, SANG H, CHEN W, et al. VOSTN: variational one-shot transformer network for pedestrian trajectory prediction[J]. Physica Scripta, 2024, 99(2): 026002.

(上接第1961页)

[10] LABRíN C, URDINEZ F. Principal component analysis[M]. R for Political Data Science. Chapman and Hall/CRC, 2020: 375-393.

[11] SANAGA K P, YANG M S. Unsupervised K-means clustering algorithm [J]. IEEE Access, 2020, 8: 80716-80727.

[12] HAN K, XIAO A, WU E, et al. Transformer in transformer[J]. Advances in Neural Information Processing Systems, 2021, 34: 15908-15919.

[13] OKK AN U, SERBES Z A. Rainfall-runoff modeling using least squares support vector machines[J]. Environmetrics, 2012, 23(6): 549-564.

[14] TREIBER M, HENNECKE A, HELBING D. Congested traffic states in empirical observations and microscopic simulations[J]. Physical Review E, 2000, 62(2): 1805.

[15] LIU L, FENG S, FENG Y, et al. Learning-based stochastic driving model for autonomous vehicle testing[J]. Transportation Research Record, 2022, 2676(1): 54-64.