基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法

裴 炤[1,2],邱文涛[2],王 淼[3],马 苗[2],张艳宁[4,5]
(1. 陕西师范大学现代教学技术教育部重点实验室,陕西西安 710119; 2. 陕西师范大学计算机科学学院,陕西西安710119;
3. 上海交通大学航空航天学院, 上海 200240; 4. 空天地海一体化大数据应用技术国家工程实验室,陕西西安 710129;
5. 西北工业大学计算机学院, 陕西西安 710129)

摘 要: 行人轨迹预测是视频监控的重要组成部分,因现有方法未充分利用场景特征信息造成其预测轨迹不符合生活常识,导致行人轨迹预测精度较低出现明显偏离真实轨迹的情况. 针对上述不足本文提出一种基于Transformer动态场景信息生成对抗网络(Generative Adversarial Network,GAN)的行人轨迹预测方法. 该方法利用动态场景特征提取模块的卷积神经网络(Convolutional Neural Networks,CNN)模型对目标行人的动态场景信息进行特征提取,同时生成器网络中的编码器利用Transformer对行人的社会交互信息特征以及轨迹信息特征进行建模. 在ETH和UCY数据集上的实验结果表明,与Social GAN模型相比,本文方法在多个场景下的平均位移误差准确率提高了25.61%,最终位移误差准确率提高了38.44%.

关键词: 行人轨迹预测; 生成对抗网络; 转换器; 深度学习; 长短期记忆网络
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112(2022)07-1537-11
电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20210762

Pedestrian Trajectory Prediction Method Using Dynamic Scene Information Based Transformer Generative Adversarial Network

PEI Zhao[1,2], QIU Wen-tao[2], WANG Miao[3], MA Miao[2], ZHANG Yan-ning[4,5]
(1. *Key Laboratory of Modern Teaching Technology (Ministry of Education), Shaanxi Normal University, Xi'an, Shaanxi* 710119, *China*;
2. *School of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi* 710119, *China*;
3. *School of Aeronautics and Astronautics, Shanghai Jiao Tong University, Shanghai* 200240, *China*;
4. *National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application Technology, Xi'an, Shaanxi* 710129, *China*;
5. *School of Computer Science, Northwestern Polytechnical University, Xi'an, Shaanxi* 710129, *China*)

Abstract: Pedestrian trajectory prediction is an important part of video surveillance. The current methods are not accurate and sometimes violate common senses because scene information is not fully used. To eliminate the above shortcomings, this paper proposes a transformer generated adversarial network(GAN) algorithm which combines dynamic scene information with pedestrian social interaction information. The convolution neural network model of the dynamic scene extraction module is utilized to extract the dynamic scene information features of the target pedestrian, and the encoder in the generator network uses transformer to model the features of social interaction information and trajectory information of pedestrians. Experimental results on ETH and UCY datasets show that, compared with social GAN model, our method improves the accuracy of average displacement error by 25.61% and the accuracy of average final displacement error by 38.44% in multiple scenarios.

Key words: pedestrian trajectory prediction; generative adversarial networks; transformer; deep learning; long short-term memory

收稿日期:2021-06-16;修回日期:2021-11-07;责任编辑:李勇锋
基金项目:国家自然科学基金(No.61971273,No.61877038);陕西省重点研发计划(No.2021GY-032);中央高校基本科研业务(No.GK202003077);上海市自然科学基金(No.20ZR1427800)

1 引言

基于深度学习的行人轨迹预测[1]是近年来人工智能和计算机视觉领域的热点研究问题,应用在视频安防监控、目标跟踪等方面.行人轨迹预测是根据目标行人的历史轨迹以及行为特征综合分析后,推测出目标行人在未来的位置坐标[2].在行人密集的公共场所,监测场所内行人的活动轨迹,并分析人群的运动、检测异常的行人轨迹,对犯罪预防、防恐防暴等公共安全领域有着积极的作用[3,4].在目标跟踪[5,6]领域,在跟踪过程中因目标行人被短暂遮挡而导致跟踪失败时,可以使用行人轨迹预测技术预测目标行人的未来轨迹,实现对目标行人的继续跟踪.

行人间的社交关系与所处的场景都会影响行人对未来路径的规划.例如当目标行人前方有结伴而行的路人时,根据社交惯例,其不会从路人之间径直穿越,而是选择绕行.在道路上遇到不同障碍物时会选择不同的策略改变其行进方向,其可以分为静态障碍物和动态障碍物两类:当目标行人遇见静态障碍物,如道路旁停放的汽车、树木以及建筑物,这时行人会选择绕行,而当其遇见动态障碍物,如行驶的汽车,行人首先会预估汽车的行进速度及其对自身前进路径的影响,进而会选择减速慢行或者驻足等候汽车通过.

行人轨迹预测本质上是基于时间序列的预测问题,该问题更关注近距离范围内的邻居行人及环境对目标行人的影响,较远距离的邻居行人及环境对目标行人的影响相对较弱,LSTM在处理长距离依赖的时序问题上有较好的效果,但在短距离预测方面稍显不足,此外,静态场景信息对行人路径规划的影响体现在当前短时间内,而动态场景信息会影响行人对未来长远的路径规划.

因此,有效利用物理环境以及行人间的社交关系对解决行人轨迹问题至关重要,为解决上述问题,本文提出一种基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法,该方法首先构造动态场景信息提取模块,提取动态场景信息特征,同时利用Transformer在解决短距离依赖的时序问题上的优势,以此构造基于Transformer的生成对抗网络对行人轨迹进行特征提取,同时利用池化模块将动态场景信息和行人社会交互信息进行特征融合,增强模型对物理场景信息以及社交信息的学习,进而提高模型预测的精准率.

主要贡献如下:

1. 首先为了解决LSTM在短距离依赖的时序预测问题上的不足,本文使用在短距离依赖的时序预测问题表现更好的Transformer网络取代LSTM,Transformer网络的自注意力机制使网络在提取目标行人的社会交互信息特征与历史轨迹特征时更加关注近距离的邻居行人.

2. 其次通过构造动态场景信息提取模块,使用卷积神经网络[7]提取动态场景信息特征,并利用池化模块将动态场景信息特征、历史轨迹特征、行人社会交互信息进行特征融合.池化模块利用社交边界模型对其交互信息进行池化操作,选取对行人轨迹产生最大影响的特征信息,将其与动态场景信息特征进行特征融合后反馈至解码器进行预测,从而实现将动态场景信息和行人社会交互信息结合,提升模型合理预测的精度.

3. 最后构建基于Transformer的生成对抗网络,生成器以池化层和随机高斯噪声为输入,将生成的符合日常生活规范的行人轨迹信息持续输入到鉴别器网络,生成器和鉴别器进行博弈,不断优化双方网络参数,最终使生成器可以生成高质量的行人轨迹信息扩充训练集,从而提高模型预测的准确率.

在ETH[8]和UCY[9]数据集上的实验结果和相关实验分析表明,本文提出的行人轨迹预测方法相较于以往基于传统循环神经网络模型的行人轨迹预测算法具有更高的准确率,验证了本文提出的行人轨迹预测方法的有效性.

2 相关工作

传统的行人轨迹预测研究[10~14]通常使用相对复杂的数学统计模型如:本领域的开创工作是Helbing[10]提出的基于社会力的线性模型Social Force,它将行人和障碍物对目标的影响简单抽象为引力与斥力,行人与目标相互靠近称之为引力,反之行人与目标相互排斥从而避免碰撞称之为斥力,以此进行建模.Kitani[11]等人使用基于隐含马尔科夫模型和逆最优控制的方式通过对行人的动作理解进行强化学习建模,从而更好地学习静态环境对行人轨迹的影响.但此类模型需要对场景进行语义标注,模型对复杂场景的泛化能力较低,在面对动态场景无法取得很好的预测效果.

此后基于数据驱动的深度学习模型[15~22]成为行人轨迹预测的主要方法,如基于循环神经网络模型(Recurrent Neural Network,RNN)以及长短期记忆网络模型(Long Short-Term Memory,LSTM)的方法[23~26]逐渐用在解决此类时间序列问题上,此类模型相较于社会力等数学统计类的模型可以处理复杂的场景,且预测准确率有较大提升,逐步成为行人轨迹预测的主流模型.现阶段基于LSTM的社交网络模型有SR-LSTM[25]、Social-LSTM[18]等模型,此类模型引入了行人社交机制,利用行人之间的欧式距离和LSTM的隐藏特征信息进行社会化建模,通过社会池化层对其进行池化后根据隐藏状态信息进行预测.Pei[1]提出了一种在行人密集场景下的基于Social-affinity LSTM的行人

轨迹预测方法,其根据邻居行人的相对位置构造了一种社会亲和力图用于记录邻居行人的社交影响权重,Social-affinity LSTM根据目标行人的个人轨迹特征和邻居行人的影响进行轨迹预测.上述方法的缺点在于并未考虑行人的轨迹是多模态的,在许多情况下对于行人而言可供选择的路径是多样的,并非单一路径.

生成对抗网络(Generative Adversarial Network,GAN)的出现为多模态的行人轨迹预测提供了技术途径.Gupta[27]等人提出了一种基于生成对抗网络(Social-GAN,SGAN)的行人轨迹预测方法,其通过LSTM构造生成对抗网络,利用生成对抗网络的生成器网络和鉴别器网络不断博弈,从而强迫网络不断优化模型参数、生成符合社会规范的轨迹,以此扩充数据集,提高预测精度,但它未利用任何场景信息,仅利用行人之间的社会交互信息,未考虑场景对行人的影响,因此可能会出现违背生活常识的预测轨迹.

此后Sadeghian[28]等人将场景信息与注意力机制[29,30]结合,同时利用生成对抗网络生成多模态的轨迹.Vineet[31]等将图注意力(Graph ATtention network,GAT)网络和生成对抗网络相结合,其利用图注意力网络对静态场景中所有行人之间的社会交互进行建模,通过生成对抗网络构造预测轨迹与目标行人的行为特征之间的可逆映射来生成符合社会规范的轨迹.上述方法仅考虑当前时刻静态场景对行人的影响,未考虑动态场景的影响.

3　问题定义

行人轨迹预测问题可以看作是在固定场景中根据给定n个目标行人的历史轨迹以及状态特征,预测目标行人的未来轨迹坐标的问题,其本质上是基于时间序列的预测问题.在本文中,给定目标行人的轨迹$\mathbf{X}=(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_n)$,其中$\boldsymbol{X}_i=\left\{(x_i^t,y_i^t)|t\in(1,\cdots,t_{\text{obs}})\right\}$,$n$为场景中所有目标行人的个数,$(x_i^t,y_i^t)$为目标行人$i$在$t$时刻的坐标,$t_{\text{obs}}$为观测的时序时长.将行人的真实轨迹表示如下:

$$\mathbf{Y}=(\boldsymbol{Y}_1,\boldsymbol{Y}_2,\cdots,\boldsymbol{Y}_n) \tag{1}$$

$$\boldsymbol{Y}_i=\left\{(x_i^t,y_i^t)|t\in(t_{\text{obs}}+1,t_{\text{obs}}+2,\cdots,t_{\text{obs}}+t_{\text{pred}})\right\} \tag{2}$$

其中t_{pred}为预测的时序长度,相似的,本文方法预测的行人轨迹表示如下:

$$\hat{\boldsymbol{Y}}=(\hat{\boldsymbol{Y}}_1,\hat{\boldsymbol{Y}}_2,\cdots,\hat{\boldsymbol{Y}}_n) \tag{3}$$

$$\hat{\boldsymbol{Y}}_i=\left\{(x_i^t,y_i^t)|t\in(t_{\text{obs}}+1,t_{\text{obs}}+2,\cdots,t_{\text{obs}}+t_{\text{pred}})\right\} \tag{4}$$

4　基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法

本文提出的基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法总体网络结构如图1所示,模型整体由动态场景信息提取模块、生成器网络、池化模块、鉴别器网络和损失函数组成,其中动态场景信息提取模块由卷积神经网络构成,生成器网络包含编码器和解码器,池化模块包含行人社会交互计算模块,鉴别器网络包含解码器、全连接层和多层感知机.由于本文中的生成对抗网络与Transformer都由编码器与解码器组成,作为区分,本文将生成对抗网络中的生成器网络与鉴别器网络中的编码器分别表示为G-Encoder、D-Encoder,将生成器的解码器表示为G-Decoder,将Transformer的编码器与解码器表示为T-Encoder、T-Decoder.

本模型的预测过程如图1所示,首先由场景提取模块进行动态场景信息特征提取,G-Encoder将场景内所有行人的轨迹作为Transformer的输入,学习行人的历史轨迹特征.池化模块根据G-Encoder传入的行人轨迹特征信息计算出目标行人的社会交互信息,之后将社会交互信息与动态场景信息进行特征融合获得行人状态信息.G-Decoder将行人状态信息加入随机高斯噪声进行解码后生成相应的预测路径.生成器网络产生的预测路径与真实的行人数据作为鉴别器的输入,D-Encoder将路径信息进行编码之后由多层感知机对其进行分类鉴别.损失函数模块负责计算行人轨迹预测模型的误差,并将误差进行反向传播,从而增强生成器网络生成轨迹的能力.生成器网络和鉴别器网络会持续进行对抗训练,鉴别器网络对真假轨迹信息的鉴别能力也在对抗过程中不断提高,整个网络的参数也不断优化,最终生成器网络将产生可以媲美真实轨迹的高质量轨迹序列信息,模型的预测能也随之提升.

4.1　动态场景信息提取模块

行人当前时刻所处的静态场景会影响行人短时间内的行进方向,而动态场景会对其未来长远的路径规划产生重要影响,因此将动态场景信息引入行人轨迹预测方法显得尤为必要.为了获取行人所处的场景并加以利用,本文设计了动态场景提取模块,如图2所示.

本模块由两个关键部分组成,一个是场景关键帧提取模块,用于在视频中获取行人所处的场景.场景提取模块首先将目标行人的编号视为键,将其出现的时刻视为值,由此构造哈希表.在哈希表中检索出目标行人出现的起止时间,根据起止时间获得视频对应的场景关键帧P_t,将当前时刻到t_{obs}时刻的帧集合设为场景集合I_i^t.另是卷积神经网络模块,其首先对I_{i}^{t}中的场景关键帧进行特征提取,对其进行最大池化计算得到动态场景信息张量$\boldsymbol{V}_p^t$.动态场景信息提取模块工作的相关过程如下所示:

$$I_i^t=(P_t,\cdots,P_{t_{\text{obs}}}) \tag{5}$$

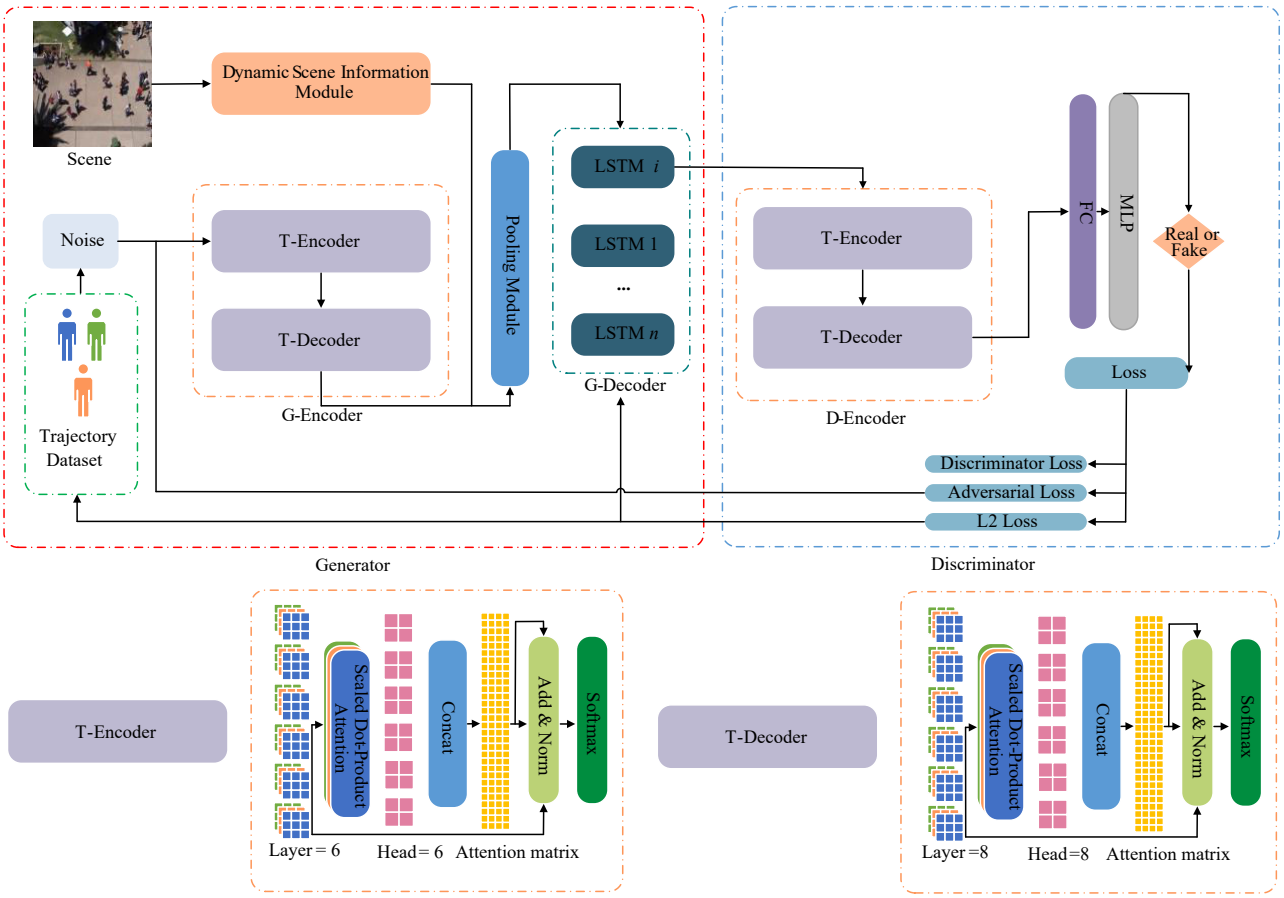


图1 基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法总体网络结构

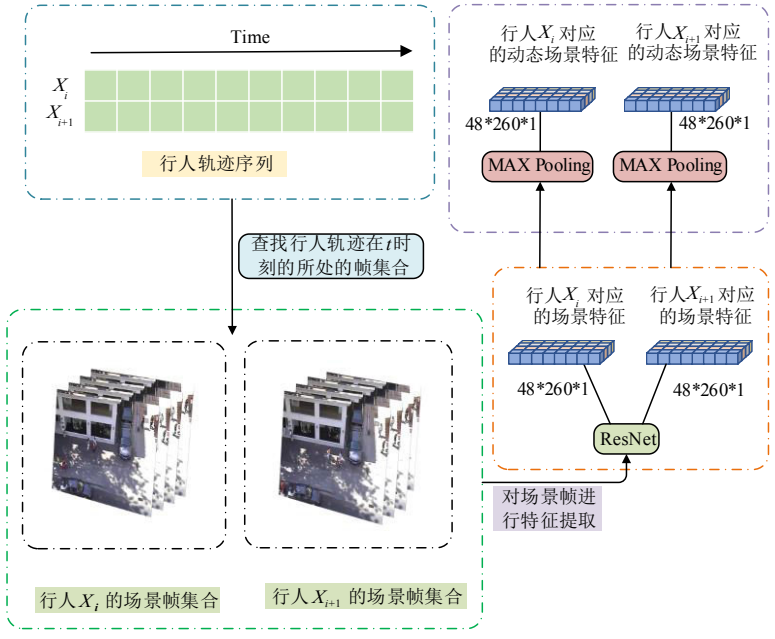


图2 动态场景提取模块的工作流程

$$V_p^t = \mathrm{MAX}\left\{\mathrm{CNN}\left(I_i^t, W_{\mathrm{CNN}}\right)\right\} \tag{6}$$

在本文中使用的卷积神经网络CNN(·)为ResNet，其网络初始化参数为使用ImageNet预训练之后得到的参数，MAX(·)代表最大池化运算.

4.2　生成器网络

在处理时序问题上通常采用以长短期记忆网络(LSTM)为代表的循环神经网络(RNN),最近研究[32]表明LSTM在解决长距离依赖的问题上表现较好,但在解决短距离依赖的问题上Transformer网络表现较好,因此本文选择使用Transformer网络与LSTM共同构造生成对抗网络. 与一般的生成对抗网络相似,本文方法也由生成器网络和鉴别器网络组成,在本文中生成器网络用于学习行人真实轨迹的数据分布、生成预测轨迹序列,其中G-Encoder编码器由Transformer网络构成,G-Decoder解码器由LSTM构成.

4.2.1　G-Encoder编码器

本文将所有行人的轨迹看作是二维坐标序列,G-Encoder编码器首先使用多层感知机将每个行人的轨迹序列由二维坐标序列转换为时空位置张量$\boldsymbol{g}_i^t$,将其作为Transformer网络的输入,Transformer网络将学习并得到每位行人时空位置特征信息$\boldsymbol{T}_i^t$. 具体过程如下:

$$\boldsymbol{g}_i^t=\phi(x_i^t,y_i^t;\ w_{\mathrm{ee}}) \tag{7}$$

$$\boldsymbol{T}_i^t=\mathrm{Trans}\left[\boldsymbol{g}_i^{t-1};\ \boldsymbol{T}_i^{t-1}\right] \tag{8}$$

其中,$\phi(\cdot)$为含有非线性激活函数ReLU嵌入层(Embedding Layer)网络,w_{ee}为嵌入层网络的权重参数. 式(8)中Trans(·)为G-Encoder编码器中的Transformer网络.

4.2.2　G-Decoder解码器

G-Decoder解码器由LSTM序列模型构成,用于生成预测的轨迹序列. 经过池化模块(见下文4.3节)将轨迹特征、动态信息场景特征、社会交互特征进行特征融合后得到融合特征$\boldsymbol{h}_i^t$(式(19)),为了获得多模态轨迹信息,在此使用多层感知机为$\boldsymbol{h}_i^t$加入随机高斯噪声z得到隐藏状态张量$\boldsymbol{h}_{di}^t$,之后使用全连接网络将$t-1$时刻的二维坐标转换为空间张量$\boldsymbol{LP}_i^{t-1}$,LSTM网络对其计算得到当前状态张量$\boldsymbol{h}_{ci}^t$,最后通过多层感知机将其转化为坐标序列,即预测得到的坐标序列$\hat{\boldsymbol{Y}}_i^t$. 如下所示:

$$\boldsymbol{h}_{di}^t=\mathrm{MLP}(\boldsymbol{h}_i^t,z;\ w_{\mathrm{dp1}}) \tag{9}$$

$$\boldsymbol{LP}_i^{t-1}=\mathrm{FC}(\boldsymbol{Y}_i^{t-1};\ w_{\mathrm{dfc}}) \tag{10}$$

$$\boldsymbol{h}_{ci}^t=\mathrm{LSTM}(\boldsymbol{h}_{di}^{t-1},\boldsymbol{LP}_i^{t-1};\ w_{\mathrm{de}}) \tag{11}$$

$$\hat{\boldsymbol{Y}}_i^t=\mathrm{MLP}(\boldsymbol{h}_{ci}^t;\ w_{\mathrm{dp2}}) \tag{12}$$

其中,w_{de}为G-Decoder解码器中LSTM网络的权重参数,w_{dfc}为全连接网络权重参数,w_{dp1}与w_{dp2}为多层感知机MLP(·)的不同权重参数.

4.3　池化模块

本文方法分别使用动态场景信息池化模块和行人社会交互信息池化模块来处理动态场景信息和行人社会交互信息.

4.3.1　动态场景信息池化模块

在本文方法中,动态场景信息池化模块首先使用多层感知机将随机高斯噪声z与动态场景信息特征张量V_p^t进行特征融合得到$\boldsymbol{h}_I^t$,其次是根据行人社会交互信息池化模块确定在$t=t_{\mathrm{obs}}$时社会影响范围内的行人(本文将其视为邻居行人),将行人i的所有邻居行人的轨迹坐标$\boldsymbol{X}_i^t,\boldsymbol{X}_{i+1}^t,\ldots,\boldsymbol{X}_n^t$编码为邻居行人张量$\mathbf{X}_{i,\mathrm{ngb}}^t$,最后是通过使用多层感知机将$\boldsymbol{h}_I^t$、$\mathbf{X}_{i,\mathrm{ngb}}^t$进行前向传播对目标行人及其邻居的动态场景信息特征进行更新. 具体过程如下:

$$\boldsymbol{h}_I^t=\varnothing(\boldsymbol{V}_p^t,z,w_{\mathrm{eh}}) \tag{13}$$

$$\boldsymbol{X}_{i,\mathrm{ngb}}^t=\left[\boldsymbol{X}_i^t,\boldsymbol{X}_{i+1}^t,\ldots,\boldsymbol{X}_n^t\right] \tag{14}$$

$$\boldsymbol{P}_i^t=\gamma(\boldsymbol{h}_I^t,\boldsymbol{X}_{i,\mathrm{ngb}}^t,W_{\mathrm{ep}}) \tag{15}$$

其中$\varnothing$是含有ReLU非线性激活函数的多层感知器,w_{eh}是$\varnothing$的权重参数. $\boldsymbol{X}_{i,\mathrm{ngb}}^t$行人$i$的所有邻居行人在$t=t_{\mathrm{obs}}$时的轨迹坐标张量. γ为多层感知机,W_{ep}为其权重参数.

4.3.2　行人社会交互信息池化模块

社交信息池化社交信息池化模块首先确定影响行人的社交边界. 例如,当目标行人行走时,离其最近的人对其规划路径时的决策影响最大,为此本文设计了社交边界模型来衡量行人间的社会交互影响,利用邻里之间的相对距离和行人的当前坐标去构造边界模型,得到社交边界特征张量$\boldsymbol{H}_i^t$,将其与动态场景信息张量、轨迹特征张量进行特征融合后得到行人状态信息特征$\boldsymbol{h}_i^t$,具体过程如下:

$$\boldsymbol{H}_i^t=\sum_{j\in N_i}R_{mn}\left[x_j^t-x_i^t,y_j^t-y_i^t\right] \tag{16}$$

$$\boldsymbol{e}_i^t=\partial(x_i^t,y_i^t;\ w_e) \tag{17}$$

$$\boldsymbol{a}_i^t=\partial(\boldsymbol{H}_i^t;\ w_a) \tag{18}$$

$$\boldsymbol{h}_i^t=\mathrm{MAX}\{\boldsymbol{h}_i^{t-1},\boldsymbol{e}_i^t,\boldsymbol{a}_i^t,\boldsymbol{P}_i^t,\boldsymbol{T}_i^t\} \tag{19}$$

其中,式(16)中$R_{mn}(\cdot)$为指示函数,用于检查坐标(x,y)是否在$m\bullet n$表示的方格内部(在则返回1,否则返回0),N_i表示第i个行人社会边界区域内的所有邻居集合. $\boldsymbol{h}_i^t$表示第i个人在$t-1$时刻的状态特征信息,$\partial(\cdot)$是含有ReLU非线性激活函数的映射函数,w_e和w_a是映射函数$\partial(\cdot)$的权重系数.

4.4　鉴别器网络

鉴别器网络用于鉴别输入轨迹是来自生成器的生成轨迹$\hat{\boldsymbol{Y}}_i=(x_i^t,y_i^t)$,还是来自数据集中用于训练的真实轨迹数据$\boldsymbol{Y}_{\mathrm{i}}=(x_i^t,y_i^t)$,本文将输入轨迹表示为$\boldsymbol{D}_i$,首先使用全连接层将$\boldsymbol{D}_i$由坐标序列转换为时空特征张量$\boldsymbol{S}_i^t$,然后使用Transformer对其编码得到$\boldsymbol{d}_i^t$,最后由多层感知机对其进行计算,输出计算后的结果$\boldsymbol{Y}_{\mathrm{disc}}$,其值越高则表示输入轨迹为真实轨迹的概率越大. 具体如下:

$$\boldsymbol{D}_i=\left\{\mathbf{Y}_i,\hat{\boldsymbol{Y}}_i\right\} \tag{20}$$

$$\boldsymbol{S}_i^t = \mathrm{FC}(\boldsymbol{T}_i;\ w_p) \tag{21}$$

$$\boldsymbol{d}_i^t = \mathrm{Trans}(\boldsymbol{S}_i^t) \tag{22}$$

$$\boldsymbol{Y}_{\mathrm{disc}} = \mathrm{MLP}(\boldsymbol{d}_i^t;\ w_y) \tag{23}$$

其中，w_p为全连接层FC的权重参数，w_y为多层感知机MLP的权重参数.

4.5 损失函数

本文采用的损失函数由$L_{\mathrm{GAN}}(\mathrm{G},\mathrm{D})$和$L_{\mathrm{L2}}(\mathrm{G})$两部分组成，其中$L_{\mathrm{GAN}}(\mathrm{G},\mathrm{D})$是生成对抗网络的损失函数，$L_{\mathrm{L2}}(\mathrm{G})$是L2坐标偏移的损失函数，其本质是基于最大似然定理的概率分布函数，用于计算真实坐标位移与预测得到的K个位移$\mathrm{G}(z)$之间的最小差值以便提升预测轨迹的质量. 通过对各个损失函数进行反向传播，不断地优化生成对抗网络各层的权重参数. 其表达式如下：

$$L = \arg\ \min_{\mathrm{G}} \max_{\mathrm{D}} (\mathrm{G},\mathrm{D}) + L_{\mathrm{L2}}(\mathrm{G}) \tag{24}$$

$$L_{\mathrm{GAN}}(\mathrm{G},\mathrm{D}) = \mathrm{E}_{\mathrm{x}\sim\mathbf{Y}_i(\mathrm{x})}\left[\log \mathrm{D}(\mathbf{Y}_i)\right] \tag{25}$$

$$L_{12}(\mathrm{G}) = \min_{\mathrm{G}} \mathrm{E}_{\mathrm{x}\sim\hat{\mathbf{Y}}_i(\mathrm{x})}\left[\|\mathbf{Y}_i - \mathrm{G}(\mathbf{Y}_i, z)\|_2\right] \tag{26}$$

其中，γ为超参数，用于平衡$L_{\mathrm{GAN}}(\mathrm{G},\mathrm{D})$与$L_{\mathrm{L2}}(\mathrm{G})$，E为期望.

5 实验与分析

本文实验环境为Ubuntu 16.04，GPU为NVIDIATITAN XP，CPU为Intel(R)Core(TM)i7-7700K CPU @ 4.20 GHz×8，使用的深度学习框架为PyTorch 1.7.0.

本文实验首先在ETH和UCY两个公共数据集上评估我们提出的方法的可行性，这两个数据集包含真实的行人轨迹和社会活动，包括对物理障碍物的躲避、行人之间行走. 其中ETH数据集包含ETH和Hotel两个场景，UCY数据集包含Zara1、Zara2和Univ三个场景.

5.1 实验数设置及评价指标

在本文实验中Transformer网络的参数如下：T-Encoder的层数为6，head个数为6，T-Decoder的层数为8，head个数为8. G-Encoder中嵌入层单元数为64，隐藏层单元数为64，多层感知机单元数为1 024，G-Decoder的嵌入层单元数为64，隐藏层单元数为128，多层感知机单元数为1 024，瓶颈层单元数为1 024，使用ReLU作为激活函数，生成器网络的学习率设置为0.001. 鉴别器中编码器的嵌入层单元数设置为64，隐藏层单元个数设置为64，多层感知机单元数为1 024，学习率设置为0.001. 池化模块中的嵌入层单元数为64，隐藏层单元数为64，多层感知机单元数为1 024，使用ReLU作为激活函数. 场景提取模块使用在ImageNet数据集上预训练的ResNet模型，整个网络中噪声为8个维度的高斯噪声，训练时的数据的批次大小为32，epochs大小设置为500，训练迭代次数设置为15 000次，观察轨迹的长度设置为8步，预测轨迹长度为12步.

与之前的研究方法[3,4]类似，在此本文选用ADE(平均偏移误差)和FDE(最终偏移误差)作为评价指标来刻画预测轨迹的准确性. ADE是通过计算每个时刻的预测轨迹与真实轨迹的平均欧氏距离来评估预测序列的准确性. FDE是通过计算最终时刻的预测轨迹位置与真实轨迹位置的平均欧氏距离来评估预测序列的准确性.

5.2 实验结果与分析

本文将文中方法和LSTM、Social-LSTM、Social-GAN、Sophie、Social-BiGAT在ETH和UCY数据集上进行对比实验.

5.2.1 定量分析

本文将文中方法和LSTM、Social-LSTM、Social-GAN、Sophie、Social-BiGAT在ETH和UCY数据集上进行对比实验. 各种轨迹预测方法的ADE和FDE的对比结果如表1所示. 其中ADE和FDE的数值表示预测轨迹与真实轨迹误差，数值越小表示预测误差越小、准确率越高，各种场景下的最优结果已在表中标记. 从表1中可以看出，本文方法的ADE和FDE表现在ETH和UCY两大数据集中的多个场景取得了较好的效果. 本文方法的行人社会交互信息池化模块将来自于Transformer的自注意力机制提取的社交特征与社交边界特征进行融合，从而更准确的刻画行人之间的社交影响. 不同于上述模型仅考虑了社交因素而忽略了动态场景信息对目标行人的影响，本文方法中同时引入了动态场景信息池化模块，将其与行人社会交互信息池化模块相结合后产生社会交互约束，在对轨迹进行预测时会迫使模型生成符合日常生活规范的轨迹，使得模型对真实场景的拟合效果更好，模型的预测能力也随之提

表1 不同模型的ADE和FDE结果对比

Metric	Dataset	LSTM	Social-LSTM	Social-GAN	Sophie	Social-BiGAT	本文方法
ADE	ETH	1.09	1.09	0.81	0.86	0.69	**0.65**
	Hotel	0.86	0.79	0.72	0.76	0.49	**0.34**
	Univ	0.61	0.67	0.60	0.54	0.55	**0.53**
	Zara1	0.41	0.47	0.34	**0.30**	**0.30**	0.32
	Zara2	0.52	0.56	0.42	0.38	0.36	**0.31**
Average		0.70	0.72	0.58	0.61	0.48	**0.44**
FDE	ETH	2.41	2.35	1.52	1.65	1.29	**1.18**
	Hotel	1.91	1.76	1.61	1.67	1.01	**0.64**
	Univ	1.31	1.40	1.84	1.24	1.32	**1.15**
	Zara1	1.88	1.00	1.26	0.63	**0.62**	0.66
	Zara2	1.11	1.17	0.69	0.78	0.75	**0.63**
Average		1.52	1.54	1.38	1.24	1.00	**0.89**

升. 因此本文方法在大多数场景下的ADE和FDE优于LSTM、Social-LSTM、Social-GAN、Sophie、Social-BiGAT等模型.

5.2.2　消融实验

为了进一步验证本文提出方法的有效性,本小节中使用定量分析方法进行验证. 首先,本文选择Social GAN作为基线方法,测试其在各个数据集场景中的实验结果. 在此基础上,保持相同的试验参数设置,本文分别设计为其加入动态场景信息提取模块、Transformer网络以及两者结合方法的试验,具体对比结果如表2所示.

表2表明:在单独使用动态场景信息提取模块或Transformer网络的情况下,本文方法在大多数场景中的ADE和FDE优于基线方法,在使用两者结合的方法时,本文方法在全部场景中的ADE和FDE均优于基线方法.

在ETH数据集中,受数据集中场景的制约,行人行进路线基本固定,故动态场景信息对行人的路径规划有一定影响,本文方法相较于基线方法ADE提高了19.75%,FDE提高了22.37%,但略低于单独使用动态场景信息的方法,推测是因为Transformer网络自注意力机制中的位置编码器,使得本文方法更关注行人自身的轨迹,从而弱化了动态场景信息的影响权重.

Hotel数据集中场景较为复杂,对行人的路径规划影响较大,因此本文方法相较于基线方法ADE提高了52.78%,FDE提高了60.25%,和ETH数据集中的情况相反,单独使用Transformer方法的准确率略高于本文方法,推测和ETH数据集中情况相似,动态场景信息对模型的影响权重略大,使得模型侧重于学习动态场景信息.

Univ数据集中行人较为密集,障碍物处于道路边缘,因此对目标行人影响最大的是周围行人,得益于Transformer网络的自注意力机制,本文方法相较于基线方法ADE提高了11.67%,FDE提高了37.5%.

Zara1与Zara2数据集场景相同,场景中的车辆、建筑物会影响行人对未来路径的规划,本文方法相较于基线方法ADE分别提高了5.88%、26.19%,FDE分别提高了47.62%、8.7%.

表2　消融实验结果对比

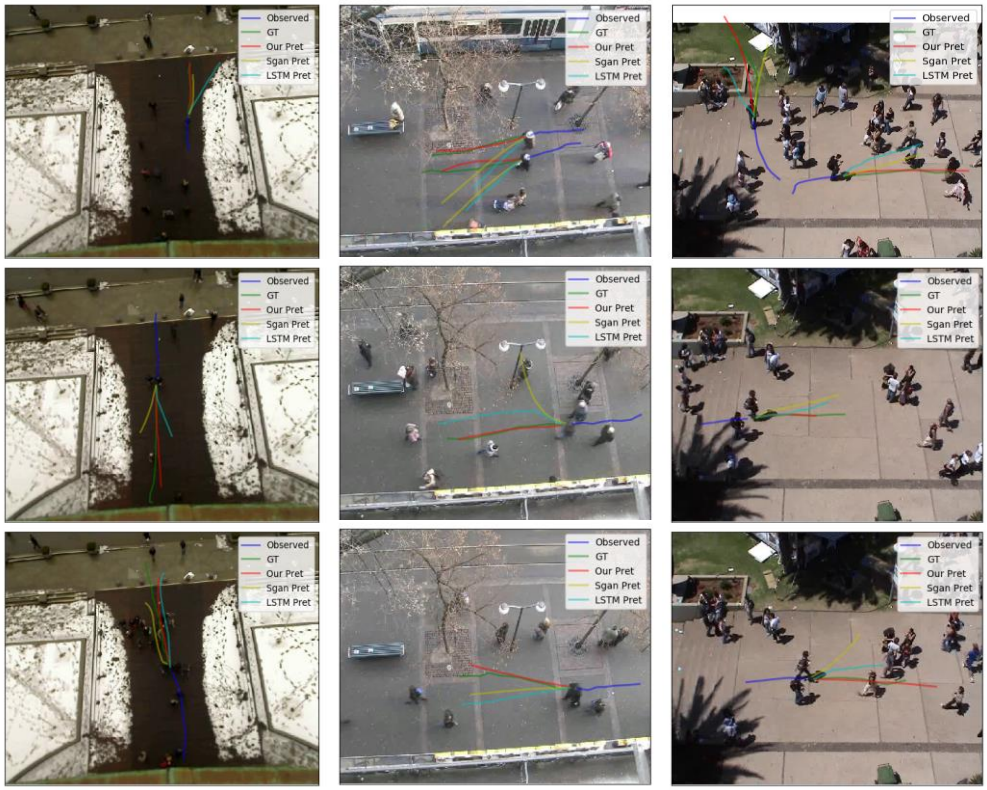
Metric	Dataset	Baseline	Dynamic scene information	Transformer	Dynamic scene information+ Transformer
ADE	ETH	0.81	**0.64**	0.68	0.65
	Hotel	0.72	0.38	**0.32**	0.34
	Univ	0.60	0.7	0.64	**0.53**
	Zara1	0.34	0.35	0.37	**0.32**
	Zara2	0.42	0.34	0.35	**0.31**
Average		0.58	0.48	0.47	**0.43**
FDE	ETH	1.52	**1.1**	1.21	1.18
	Hotel	1.61	0.72	**0.59**	0.64
	Univ	1.84	1.35	1.24	**1.15**
	Zara1	1.26	0.77	0.76	**0.66**
	Zara2	0.69	0.65	0.73	**0.63**
Average		1.39	0.91	0.90	**0.85**

5.2.3　定性分析

图3展示了各模型在ETH和UCY数据集中各个场景中的轨迹预测可视化对比图. 其中图3(a)为ETH数据集场景下的轨迹预测对比图,该场景两侧是积雪与围墙,场景前方有路障球. 从图3(a)中可以看出,仅有本文方法预测的轨迹接近真实轨迹,LSTM、Social-GAN模型预测得到的轨迹与真实轨迹偏差较大.

图3(b)为Hotel数据集场景下的预测对比图,该场景是位于车站的一个旅馆前,行人的轨迹主要是进出车站或者直行经过旅馆,场景中行人轨迹比较复杂. 从图3(b)第一张图像中可以看出行人真实轨迹是直行,但Social-GAN、LSTM预测行人将会转向. 图3(b)第二张图像中可以看出目标行人的真实意图是直行路过,本文方法预测得到行人的轨迹与真实轨迹十分贴合,但Social-GAN预测行人将会转向进入车站,LSTM预测的行人行进方向基本正确,但与真实轨迹相差太大. 图3(b)第三张图像场景内行人行进方向与图3(b)第一张图像刚好相反,目标行人的真实轨迹是转向,Social-GAN、LSTM均对行人未来的行进方向判断失误,只有本文方法预测得到的轨迹与真实轨迹最相符.

图3(c)为Univ数据集场景下的预测对比图,该场景是大学校园的一个交叉路口,该场景中人群密度大,可以看作是典型的拥挤社交场景. 人群密度大带来的问题就是行人轨迹无序,社交信息对目标行人的路径规划产生决定性的影响,这体现在目标行人随时会调整前进方向,同时还会因为与其他行人交谈而产生中



(a) ETH (b) Hotel (c) Univ

图3 各模型在不同场景的预测轨迹可视化对

途长时间逗留的现象. 从图3(c)中可以看出,本文方法在该拥挤社交场景中的预测表现显著优于其他的模型,这得益于本文使用的Transformer网络的自注意力机制与其位置编码器在处理时序问题上的优异表现.

图4(a)为Zara1数据集场景下的预测对比图,图4(b)为Zara2数据集场景下的预测对比图. 两个场景均为商场前的道路,行人的运动轨迹主要为进出商场或者路过. 从图4(a)中可以看出在行人稀疏时,各个模型的预测结果大致相似,本文方法预测的轨迹与真实轨迹几乎重合,在各个模型中表现最优. 图4(a)中第一张图像展示了行人转向时各种模型的轨迹预测对比图,从图中可以看出LSTM、Social-GAN模型均未预测到目标的转向,另外从图4(b)中第二张图片可以看出其他模型的预测轨迹会与汽车障碍物发生接触,这显然违背了生活常识,而本文方法预测得到的轨迹明显优于其他模型,这是因为本文方法引入的动态场景信息可以综合考虑到目标旁边的汽车障碍物,从而选择绕过汽车调整行进方向.

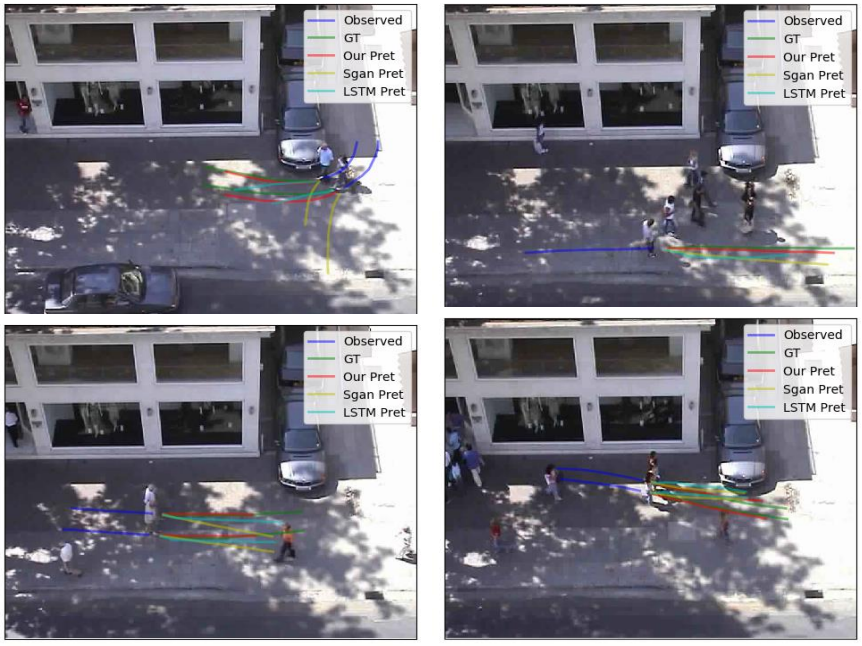
5.2.4 预测时效分析

表3中LSTM模型最为简单,预测的精准度也最低,其预测所耗费的时间为2.7 ms. Social-LSTM在LSTM的基础上加入了社会池化模块,计算量大幅增加导致时间开销增加,其预测所耗费的时间为4.2 ms. Social-GAN与本文方法都基于生成对抗网络,需要进行大量前向传播以及通过优化鉴别器进行反向传播更新生成器参数,其中Social-GAN预测所耗费的时间为29.4 ms,本文方法引入的动态场景信息提取模块会进行多次卷积、池化,所以耗时比Social-GAN稍长,其预测所耗费时间为34.3 ms. 对比结果如表3所示,虽相对于其它对比方法预测耗时略长,但本文方法在34.3 ms仍然能够预测未来120帧的轨迹,完全满足视频处理实时性的要求(该数据集视频帧率为25 FPS). 考虑到本文方法预测精度在对比方法中最高,因此该方法综合表现优异.

5.2.5 合理性分析

为了进一步验证本文方法的预测结果是否符合日常规范,如图5所示,本小节分别展示了本文方法在面对静态遮挡物和场景中移动目标时的预测结果(包含场景ETH、Hotel、Univ和Zara1). 为了将可视化的结果更好的展示,在此对每组目标生成10次轨迹预测结果(多模态轨迹预测). 其中(a)、(b)、(c)展示了本文方法面对静态障碍物时的预测结果,(d)展示了本文方法在面对动态障碍物时的预测结果.

从图5(a)中可以看出本文方法在面对路障球进行

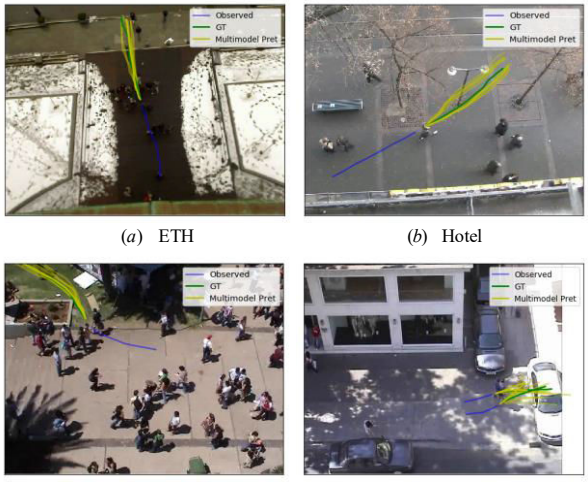


(*a*) Zara1　　(*b*) Zara2

图4　各模型在zara1和zara2场景的预测轨迹可视化对比

表3　各模型预测时效分析

预测模型	预测耗费时间/ms
LSTM	2.7
Social-LSTM	4.2
Social-GAN	29.4
本文方法	34.3



(*a*) ETH　　(*b*) Hotel

(*c*) Univ　　(*d*) Zara1

图5　不同场景的多模态轨迹可视化预测结果

预测时,其预测的轨迹分布在路障球的左右两侧,从而避开路障球. 图5(*b*)中展示了本文方法预测的轨迹会绕过路灯. 图5(*c*)中展示了本文方法预测的轨迹分布在花坛旁边的空地上. 在日常生活中,行人在避让车辆时会让车辆先行通过,本文方法在图5(*d*)Zara1场景中生成的轨迹均未与行进中的汽车车头部分接触(图中轨迹与车的其他部分也并未接触,在第4帧之后汽车已经驶离场景). 以上场景的预测轨迹符合日常规范,也证明本文方法提出的动态场景信息提取模块是合理有效的,所预测的结果是符合日常规范的.

6　结论

针对目前行人轨迹预测方法对物理环境以及行人间的社交关系利用不充分问题,本文提出了一种基于Transformer动态场景信息生成对抗网络的行人轨迹预测方法. 与其他行人轨迹预测方法相比,本文方法在ETH和UCY数据集的多数场景中ADE和FDE的表现优于其他方法,在复杂场景中可以较为准确的预测目标行人的轨迹,证明本文方法提出的动态场景信息提取模块与引入的Transformer网络对模型的预测效果有显著提升作用. 但是在拥挤场景中,本文方法的预测效果距离预期还有提升空间. 在接下来的工作中,将引入图注意力神经网络对行人之间的社会交互建模,以此提高本文方法在各场景中的预测精度与预测效率.

参考文献

[1] PEI Z, QI X, ZHANG Y, et al. Human trajectory prediction in crowded scene using social-affinity long short-term memory[J]. Pattern Recognition, 2019, 93: 273-282.

[2] YAMAGUCHI K, BERG A C, ORTIZ L E, et al. Who are

you with and where are you going?[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs: IEEE, 2011: 1345-1352.

[3] DESOUZA G N, KAK A C. Vision for mobile robot navigation: A survey[J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2002, 24(2): 237-267.

[4] RUDENKO A, PALMIERI L, HERMAN M, et al. Human motion trajectory prediction: A survey [J]. The International Journal of Robotics Research, 2020. 39(8): 895-935.

[5] 李康, 李亚敏, 胡学敏, 等. 基于卷积神经网络的鲁棒高精度目标跟踪算法[J]. 电子学报, 2018, 46(9): 2087-2093.
LI K, LI Y M, HU X M, et al. A robust and accurate object tracking algorithm based on convolutional neural network [J]. Acta Electronica Sinica, 2018, 46(9): 2087-2093. (in Chinese)

[6] 马少雄, 邱实, 唐颖, 等. 基于工地场景的深度学习目标跟踪算法 [J]. 电子学报, 2020, 48(9): 1665-1671.
MA S X, QIU S, TANG Y, et al. Deep learning target tracking algorithm based on construction site scene[J]. Acta Electronica Sinica, 2020, 48(9): 1665-1671. (in Chinese)

[7] S-C B LO, H-P CHAN, LIN J-S, et al. Artificial convolution neural network for medical image pattern recognition [J]. Neural Networks, 1995, 8(7-8): 1201-1214.

[8] PELLEGRINI S, ESS A, VAN G L. Improving data association by joint modeling of pedestrian trajectories and groupings[C]//Proceedings of the European Conference on Computer Vision. Crete: Springer, 2010: 452-465.

[9] LERNER A, CHRYSANTHOU Y, LISCHINSKI D. Crowds by example[C]//Proceedings of the Computer Graphics Forum. Oxford: Blackwell Publishing Ltd, 2007: 26(3): 655-664.

[10] HELBING D, MOLNAR P. Social force model for pedestrian dynamics[J]. Physical Review E, 1995, 51(5): 4282-4286.

[11] KITANI K M, ZIEBART B D, BAGNELL J A, et al. Activity forecasting[C]//Proceedings of the European Conference on Computer Vision. Florence, Italy: Springer, 2012: 201-214.

[12] LEE N, CHOI W, VERNAZA P, et al. Desire: Distant future prediction in dynamic scenes with interacting agents [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Hawaii, USA: IEEE, 2017: 336-345.

[13] PELLEGRINI S, ESS A, SCHINDLER K, et al. You' ll never walk alone: Modeling social behavior for multi-target tracking[C]//Proceedings of the 2009 IEEE 12th International Conference on Computer Vision. Kyoto, Japan: IEEE, 2009: 261-268.

[14] MOUSSAID M, PEROZO N, GARNIER S, et al. The walking behaviour of pedestrian social groups and its impact on crowd dynamics[J]. Plos One, 2010, 5(3): e10047.

[15] XU Y, PIAO Z, GAO S. Encoding crowd interaction with deep neural network for pedestrian trajectory prediction [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 5275-5284.

[16] ZHAO T, XU Y, MONFORT M, et al. Multi-agent tensor fusion for contextual trajectory prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 12126-12134.

[17] ALAHI A, RAMANATHAN V, FEI F L. Socially-aware large-scale crowd forecasting[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Columbus, USA: IEEE, 2014: 2203-2210.

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

[19] BALLAN L, CASTALDO F, ALAHI A, et al. Knowledge transfer for scene-specific motion prediction[C]// Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 697-713.

[20] LIU J, SHAHROUDY A, XU D, et al. Spatio-temporal LSTM with trust gates for 3d human action recognition [C]//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 816-833.

[21] ROBICQUET A, SADEGHIAN A, ALAHI A, et al. Learning social etiquette: Human trajectory understanding in crowded scenes[C]//Proceedings of the European Conference on Computer Vision. Amsterdam, Netherlands: Springer, 2016: 549-565.

[22] ALTCHÉ F, DEL A. An LSTM network for highway trajectory prediction[C]//Proceedings of the 2017 IEEE 20th International Conference on Intelligent Transportation Systems. Yokohama, Japan: IEEE, 2017: 353-359.

[23] CHENG B, XU X, ZENG Y J, et al. Pedestrian trajectory prediction via the social-grid LSTM model[J]. Journal of Engineering-Joe, 2018, 2018(16): 1468-1474.

[24] FERNANDO T, DENMAN S, SRIDHARAN S, et al.

Soft+hardwired attention: An LSTM framework for human trajectory prediction and abnormal event detection [J]. Neural Network, 2018, 108(1): 466-478.

[25] ZHANG P, OUYANG W, ZHANG P, et al. SR-LSTM: State refinement for LSTM towards pedestrian trajectory prediction[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 12085-12094.

[26] 金苍宏, 董腾然, 陈天翼, 等. 融合序列分解与时空卷积的时序预测算法[J]. 电子学报, 2021, 49(2): 233-238.
JIN C H, DONG T R, CHEN T Y, et al. Spatio-temporal convolutional forecasting based on time-series decomposition strategy [J]. Acta Electronica Sinica, 2021, 49(2): 233-238. (in Chinese)

[27] GUPTA A, JOHNSON J, FEI F L, et al. Social GAN: Socially acceptable trajectories with generative adversarial networks[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Salt Lake City, USA: IEEE, 2018: 2255-2264.

[28] SADEGHIAN A, KOSARAJUV. et al. Sophie: An attentive GAN for predicting paths compliant to social and physical constraints[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. Long Beach, USA: IEEE, 2019: 1349-1358.

[29] 李志欣, 孙亚茹, 唐素勤, 等. 双路注意力引导图卷积网络的关系抽取[J]. 电子学报, 2021, 49(2): 315-323.
Li Z X, Sun Y R, Tang S Q, et al. Dual attention guided graph convolutional networks for relation extraction[J]. Acta Electronica Sinica, 2021, 49(2): 315-323. (in Chinese)

[30] VEMULA A, MUELLING K, OH J. Social attention: Modeling attention in human crowds[C]//Proceedings of the 2018 IEEE International Conference on Robotics and Automation. China: IEEE, 2018: 4601-4607.

[31] KOSARAJU V, SADEGHIAN A, MARTÍN M R, et al. Social-BIGAT: Multimodal trajectory forecasting using bicycle-GAN and graph attention networks[C]//Proceedings of the Advances in Neural Information Processing Systems. Vancouver, Canada: NIPS, 2019: 137-146.

[32] DAI Z, YANG Z, YANG Y, et al. Transformer-XL: Attentive language models beyond a fixed-length context [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: ACL,2019:2978-2988.

作者简介

裴　炤　男,1983年2月生,陕西西安人,博士、教授、博士生导师. 主要从事计算机视觉与人工智能、图像处理与模式识别、机器学习的相关研究.
E-mail: zpei@snnu.edu.cn

邱文涛　男,1996年8月生,山东枣庄人. 陕西师范大学计算机科学学院研究生,主要研究方向为计算机视觉和行人轨迹预测.
E-mail: qiuwentao@snnu.edu.cn

王　淼(通讯作者)　男,1981年7月生,河南义马人,博士,上海交通大学航空航天学院助理研究员,主要研究方向为智能信息处理、数据挖掘、计算机视觉.
E-mail: miaowang@sjtu.edu.cn