doi:10.19562/j.chinasae.qcgc.2024.01.004

基于动作条件交互的高效行人过街意图预测*

杨 彪[1],韦智文[1],倪蓉蓉[1],王 海[2],蔡英凤[3],杨长春[1]
(1. 常州大学微电子与控制工程学院,常州 213159; 2. 江苏大学汽车与交通工程学院,镇江 212013;
3. 江苏大学汽车工程研究院,镇江 212013)

[摘要] 城市化的进程不断加速,人车冲突问题已成为现代社会亟待解决的重大难题。复杂交通场景下,行人横穿马路行为导致交通事故频发,准确、实时地预测行人过街意图对避免人车冲突、提高驾驶安全系数和保障行人安全至关重要。本文提出基于动作条件交互的高效行人过街意图预测框架(efficient action-conditioned interaction pedestrian crossing intention anticipation framework,EAIPF)来预测行人过街意图。EAIPF引入行人动作编码模块增强多模态动作模式下的表征能力,挖掘深层骨架上下文信息。同时,引入场景对象交互模块挖掘与对象交互信息,理解交通场景中高级语义线索。最后,意图预测模块融合行人动作特征和对象交互特征,实现行人过街意图的鲁棒预测。所提出的方法在两个公共数据集JAAD和PIE上验证算法性能,准确率分别达到了89%和90%,表明本文方法可以在复杂交通场景下准确预测行人穿越意图。

关键词:人车冲突;行人过街意图预测;图卷积网络;行人动作编码;场景理解

Efficient Pedestrian Crossing Intention Anticipation Based on Action-Conditioned Interaction

Yang Biao[1], Wei Zhiwen[1], Ni Rongrong[1], Wang Hai[2], Cai Yingfeng[3] & Yang Changchun[1]
1. *School of Microelectronics and Control Engineering, Changzhou University, Changzhou* 213159;
2. *School of Automotive and Traffic Engineering, Jiangsu University, Zhenjiang* 212013;
3. *Institute of Automotive Engineering, Jiangsu University, Zhenjiang* 212013

[**Abstract**] With acceleration of the urbanization process, pedestrian-vehicle conflicts have become a significant issue that modern society urgently needs to solve. In complex traffic scenarios, pedestrian crossing behavior leads to frequent traffic accidents. Accurately and timely anticipating pedestrian crossing intentions is crucial for avoiding pedestrian-vehicle conflicts, improving driving safety, and ensuring pedestrian safety. An Efficient Action-Conditioned Interaction Pedestrian Crossing Intention Anticipation Framework (EAIPF) is proposed in this paper to anticipate pedestrian crossing intention. EAIPF introduces in a pedestrian action encoding module to enhance the representation ability of multimodal action patterns and discover deep skeletal context information. At the same time, the scene object interaction module is introduced to explore interaction information with objects and understand advanced semantic clues in traffic scenes. Finally, the intention anticipation module fuses pedestrian action and object interaction features to achieve robust anticipation of pedestrian crossing intentions. The proposed method is verified on two public datasets, JAAD and PIE, achieving the accuracy of 89% and 90%, respectively, indicating that the proposed method can accurately anticipate pedestrian crossing intentions in complex traffic scenarios.

Keywords: pedestrian-vehicle conflict; crossing intention anticipation; graph convolution network; pedestrian action encoding; scene understanding

*江苏省博士后基金(2021K187B)、国家博士后基金(2021M701042)和江苏省科技厅面上项目(BK20221380)资助。
原稿收到日期为 2023 年 06 月 04 日,修改稿收到日期为 2023 年 07 月 03 日。
通信作者:蔡英凤,教授,博士,E-mail: caicaixiao0304@126.com。

前言

智能交通系统[1]可以为车辆提供自动驾驶、紧急制动以及行人横穿预警等辅助功能。在物流配送、智能出行、公共交通等领域,人车交互技术[2]已成为热点话题。在复杂城市道路中,特别是在没有十字路口的道路两侧,车辆在行驶中面对行人突发性行为无法及时做到科学决策。同时,行人在过街过程中遇到突发状况,会在极短的时间内改变原来的行为动作和运动方向,导致过街行为复杂和难以预测。如果驾驶员忽略行人过街,将会导致交通拥堵、财产损失,甚至威胁行人生命。因此,准确预测行人的过街意图,并辅助驾驶员进行减速和制动,可以保障行人安全,并提供更舒适的驾乘体验。

为预测行人过街意图,估计行人姿态是过街意图预测的基础。近年来,姿态估计[3-7]已经广泛开展。Li 等[3]引入动态阈值策略和向量姿态表征建模不同身体部位之间的关系。Luo 等[4]引入关节级回归方法部分关联场,解决关节点的重叠问题和提高低分辨率图像上姿态估计的效果。Liu 等[5]在图神经网络中嵌入图注意力机制以帮助模型准确地聚焦关键关节点信息,提高模型的判别能力和准确率。在复杂交通场景中,时间、天气、光线和距离等外部环境因素会影响姿态估计结果。特别是低分辨率图像中远距离目标行人的检测,传统的上采样计算效率有限。因此,有必要对上述场景下的行人姿态准确估计展开深入研究,以提升复杂交通场景中的姿态估计性能。

除准确估计姿态,行人动作编码方法也至关重要。基于骨架的动作识别方法对复杂交通场景和动态行人特征有很强的学习能力。ST-GCN[8]首次用图卷积网络学习行人骨架时空信息。但是,单一的图卷积网络很难灵活学习不同运动模式的骨架信息。2s-agcn[9]提出双流自适应图卷积网络,解决了ST-GCN处理骨架灵活性的问题。Ye等[10]提出动态图卷积网络,通过融合所有关节的上下文特征学习关节对之间的相关性。但是,由于行人的动作模式复杂多样,共享拓扑难以学习不同动作模式下关节之间的多样关系。通过参数化的多通道图卷积网络[9]可以独立建模不同动作的骨架信息,但是存在参数开销大、推理时间长的不足。因此,实时、动态、有效的动作编码仍然是具有挑战性的任务。

此外,场景对象交互模块是理解交通场景的关键,当前研究利用语义分割[11-14]得到交通对象分类和可行驶区域。杨彪等[11]提出轻量级E-Net网络编码行人局部场景。Pedestrian Graph+[12] 提出用2D卷积提取行人局部上下文信息。然而,局部上下文感受野小,容易捕捉与行人意图无关的细粒度特征。全局语义图感受野变大,可以补充局部上下文缺失的交通对象信息。Yang 等[13]引用DeepLabV3模型提取语义掩码,对场景中交通对象进行分类和定位。Ni 等[14]在局部交通场景的语义级感知基础上,引入HGAN建模动静态交通对象,提供全局交通场景的对象级感知。但是,全局语义图需要深层网络,与行人无关的语义信息冗余、计算开销大、推理时间长。

行人在交通场景中占比最大、分布最广且极易在过街过程中受到伤害。因此,准确预测行人过街意图对于防治人车冲突至关重要。Rasouli 等[15]提出了一个注释丰富的数据集JAAD来研究交通场景对行人行为的影响。研究发现,融合多模式特征可以帮助推断行人过马路的意图。SF-GRU[16]提出堆叠循环网络逐步融合车速、行人边框、人体姿态、局部上下文和全局语义图,实现行人过街意图预测。上述研究通过融合多特征提升过街意图预测准确率,却忽略了特征融合策略设计。Yang 等[13]提出基于注意力机制的循环神经网络,解决了不同时空特征的融合策略问题。上述研究揭示了行人动作信息和场景语义信息融合的可行性,能够实现行人过街意图预测。

针对上述不足,本文中提出基于动作条件交互的高效行人过街意图预测框架EAIPF,通过融合行人动作特征和对象交互特征,实现行人过街意图预测。针对低分辨率图像骨架表征信息丢失导致动作编码信息不足的问题,利用人体关节点的向量表征方式[6]和基于自注意力图卷积网络[17],增强多模态动作表征能力和挖掘骨架上下文信息。针对交通对象交互信息瓶颈,利用场景对象交互模块建模目标行人和相关交通对象的交互信息,实现深入理解交通场景中高级语义线索。最后,引入自适应平均池化层[18]融合行人动作特征和对象交互特征,准确预测行人过街意图。

1 行人过街意图预测算法

1.1 算法概述

行人过街意图预测是在复杂交通场景中预测出行人未来短时间内是否有过街的意图,从而辅助驾

驶员提前做出决策,避免发生人车冲突。行人意图被定义为有过街意图(C)和无过街意图(NC)的二元分类任务。图1为本文提出的基于动作条件交互的高效行人过街意图预测框架EAIPF。EAIPF包含行人动作编码模块、场景对象交互模块和意图预测模块。首先,动作编码模块引入SimCC[16]提取行人骨架序列$\boldsymbol{X}=[P_1,P_2,\ldots,P_t]$,并通过基于自注意力图卷积网络编码行人动作信息$S_A$。其次,场景对象交互模块引用目标检测网络检测观测帧最后一帧图像中交通对象,并通过对象特征网络和对象交互网络对m个交通对象的对象交互特征S_I进行建模。最后,意图预测模块通过池化层和注意力层融合行人动作特征S_A和对象交互特征S_I,实现行人过街意图的鲁棒预测。

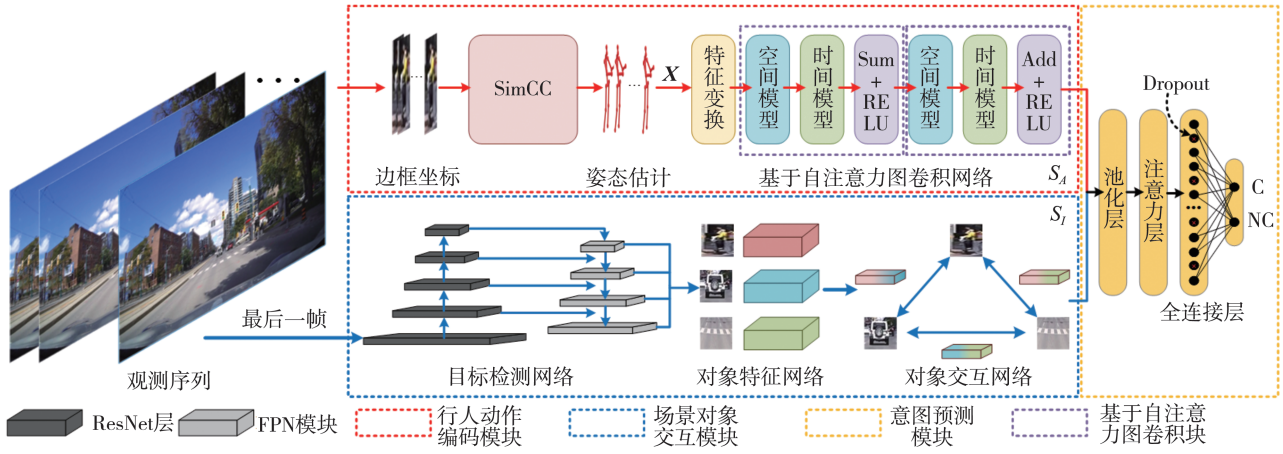

图1　基于动作条件交互的高效行人过街意图预测框架

1.2　行人动作编码模块

行人产生过街意图的同时会呈现相应动作,如转头、转身、抬脚和注视等。准确捕捉行人的动作信息,能够提升过街意图预测的准确性。行人骨架作为姿态的紧凑表示,可以将行人骨架视为图结构数据$G=(\mathrm{V},\mathrm{E})$。其中,$\mathrm{V}=\{v_1,v_2,\ldots,v_N\}$为$N$个关节点集合,$\mathrm{E}=\{e_{mn}|m=1,2,\ldots,N,n=1,2,\ldots,N,m\neq n\}$为关节点之间的躯干集合。本文利用骨架提取网络和两块基于自注意力图卷积网络提高行人动作编码能力。

1.2.1　骨架提取网络

行人过街意图预测中最重要的任务就是行人的姿态估计,行人骨架是对行人动作最确切的描述,能否在动态环境和复杂背景中准确识别行人骨架,不仅影响行人动作识别结果,还影响后续行人过街意图的预测任务。如图2所示,本文引入SimCC作为行人骨架提取方法。首先,引入HRNet[7]网络作为编码器提取图片中目标行人n个关节点表示。其次,提出姿态估计坐标解耦表征的方法,通过向量表征模块将第i关节点坐标(x_i,y_i)表征为垂直和水平的一维向量。公式被定义如下:

$$(\boldsymbol{x}_i',\boldsymbol{y}_i')=(P(x_i)\times\alpha,P(y_i)\times\alpha)\tag{1}$$

式中:$P(\cdot)$为线性投影变换函数;$\alpha(\alpha\geqslant 1)$为缩放因子,用于增强关节点的精确定位。

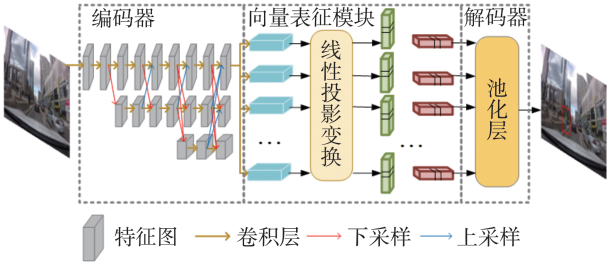

图2　骨架提取网络结构图

最后,引入池化层作为解码器,将每个连续坐标值离散化生成第i个关节点坐标$(\hat{x}_i,\hat{y}_j)$。该公式定义如下:

$$(\hat{x}_i,\hat{y}_i)=\left(\frac{\arg\max(x_i')}{k},\frac{\arg\max(y_i')}{k}\right)\tag{2}$$

1.2.2　特征变换模型

特征变换模型将行人姿态序列$\boldsymbol{X}\in\mathbb{R}^{T\times N\times C}$变换为高级特征表示。其中,$T$为时间维度,$N$为关节点数量,$C$为特征维度。首先,姿态序列$\boldsymbol{X}$通过初始共享拓扑学习一般关节空间属性。然后,线性变换后融合关节位置信息。特征变换模型定义如下:

$$\widehat{\boldsymbol{X}}=Linear(\boldsymbol{X}\boldsymbol{W}_o)+\boldsymbol{Z}\tag{3}$$

式中：$\boldsymbol{W}_0 \in \mathbb{R}^{N\times N\times C'}$ 为参数化初始共享邻接矩阵；$\boldsymbol{Z} \in \mathbb{R}^{N\times C'}$ 为可学习位置编码矩阵；$\widehat{\boldsymbol{X}} \in \mathbb{R}^{N\times C'}$ 为特征变化后的行人骨架特征。

1.2.3 空间模型

空间模型用于提取特征变化后的行人骨架特征 $\widehat{\boldsymbol{X}}$ 的空间特征。如图 3 所示，上部分为空间模型结构图，包含 3 个基于自注意力图卷积模型和残差链接，下部分为基于自注意力图卷积模型的结构图。首先，引用多头注意力机制挖掘骨架上下文信息，得到骨架自注意力图。公式表示为

$$S(\widehat{\boldsymbol{X}}) = \text{soft max}\left(\frac{\widehat{\boldsymbol{X}}\boldsymbol{W}_K(\widehat{\boldsymbol{X}}\boldsymbol{W}_Q)^{\mathrm{T}}}{\sqrt{C'}}\right) \tag{4}$$

式中：$\boldsymbol{W}_K$、$\boldsymbol{W}_Q \in \mathbb{R}^{C\times C'}$ 分别为多头注意力机制的键向量和查询向量线性变换的权重矩阵；$S(\widehat{\boldsymbol{X}}) \in \mathbb{R}^{T\times N\times N}$ 为骨架自注意力图。接着，骨架自注意力图与可学习自适应图卷积[9]逐元素乘积，学习多模态动作特征的表征能力。最后，引用残差连接[19]更新动作特征的表示。公式如下：

$$f_{SA} = \sigma\left(\sum_{k=1}^{3}(\boldsymbol{W}_A^k \odot S^k(\widehat{\boldsymbol{X}})\boldsymbol{W}_{\mathrm{GCL}}^k) + \widehat{\boldsymbol{X}}\right) \tag{5}$$

式中：$\boldsymbol{W}_A$ 为可学习共享邻接矩阵；$\boldsymbol{W}_{\mathrm{GCL}} \in \mathbb{R}^{N\times N\times C'}$ 为图卷积层权重；σ 为 ReLU 激活函数。空间模型的整体表达式为

$$f_{\text{spacial}} = \sigma\left(\sum_{k=1}^{3}(f_{SA}^k) + \widehat{\boldsymbol{X}}\right) \tag{6}$$

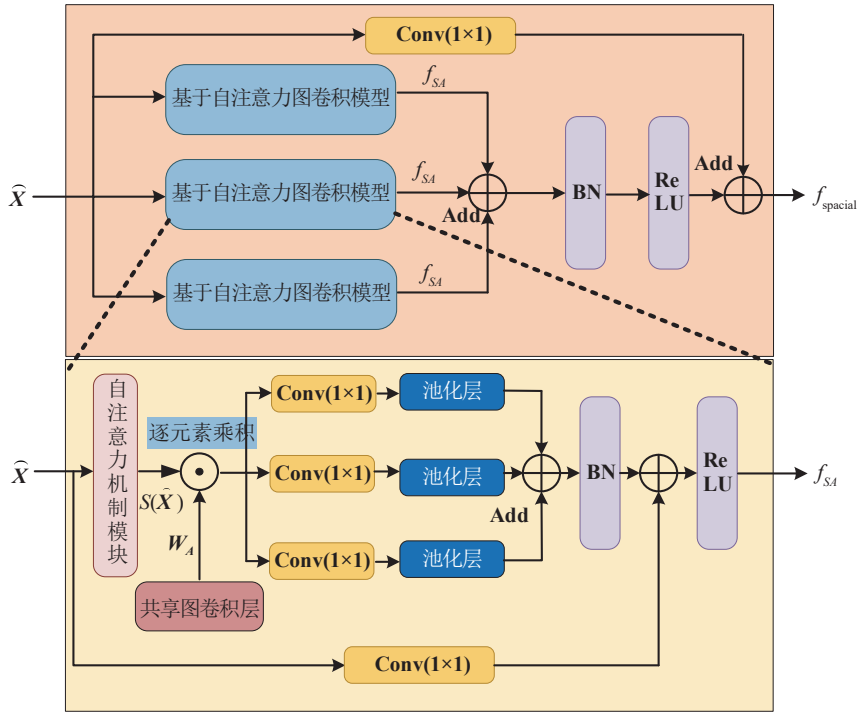

图 3 空间模型结构图

1.2.4 基于自注意力图卷积网络时间模型

如图 4 所示，时间模型以空间模型输出 f_{spacial} 作为输入，引入多尺度时间模块[20]提取行人动作特征的时间信息。时间模型由 4 个分支组成，每个分支包含一个 1×1 的卷积块来减少特征维度。前两个分支还包含一个 1×1 的卷积块和不同膨胀率的时间卷积。第 3 个分支还包含一个 MaxPool。第 4 个分支为 1×1 卷积的残差连接。时间模型的表达式为

$$S_A = \sigma\left(\left(\sum_{k=1}^{3} f_{\text{spacial}} \cdot \boldsymbol{W}_{\mathrm{TCL}}^k\right) + \widehat{\boldsymbol{X}}\right) \tag{7}$$

式中 $\boldsymbol{W}_{\mathrm{TCL}} \in \mathbb{R}^{T\times N\times C'}$ 为多尺度时间卷积权重。

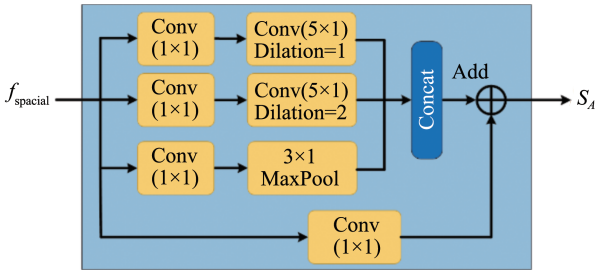

图 4 时间模型结构图

1.3　场景对象交互模块

探索交通场景对象交互关系对于理解交通场景意义重大,可以辅助行人动作特征更好地预测过街意图。为此,本文中利用场景对象交互模块对交通对象交互信息进行建模。首先,目标检测网络引用Faster R-CNN[21]提取输入图像的视觉特征。然后,利用对象特征网络提取每个对象边框中的属性特征。接着,通过对象交互网络根据成对边框对应的属性特征融合生成对象交互特征。最后,拼接输出所有对象交互特征。

场景对象交互模块结构如图1中蓝色虚线框所示。首先,目标检测网络引用ResNet层和FPN模块[22]提取观测帧中最后一帧图像的视觉特征。然后,利用对象特征网络根据每个对象边框提取属性特征$A \in \mathbb{R}^d$。接着,对象交互网络以对象O_p和O_q的属性特征A_p和A_q为输入,融合建模对象间的交互关系$I_{p:q} \in \mathbb{R}^d$,最后聚合交互关系,输出对象交互关系$S_I \in \mathbb{R}^{m(m-1)\times d}$。表达式如下:

$$I_{p:q} = \sigma(\boldsymbol{W}_p A_p + \boldsymbol{W}_q A_q) - (\boldsymbol{W}_p A_p - \boldsymbol{W}_q A_q) \odot (\boldsymbol{W}_p A_p - \boldsymbol{W}_q A_q) \tag{8}$$

$$S_I = \sum_{p=1}^{m}\sum_{q=1}^{m} \mathrm{Soft\,max}(\boldsymbol{W}_I I_{p:q} + \boldsymbol{f}_{pq}), p \neq q \tag{9}$$

式中:$\boldsymbol{W}_p$、$\boldsymbol{W}_q$为对象O_p、O_q特征空间映射;$\boldsymbol{W}_I \in \mathbb{R}^d$为对象交互特征融合器;$\boldsymbol{f}_{pq}$为对象$O_p$、$O_q$类别之间的关系分布向量。

1.4　意图预测模块

意图预测模块结构如图1中黄色虚线边框所示。行人观测样本S通过1.2节中行人动作编码模块得到行人动作特征S_A和1.3节中场景对象交互模块得到对象交互特征S_I。为充分融合这两种不同类别的特征数据,本文利用自适应平均池化层压缩对象交互特征S_I,输出与行人动作特征S_A通道数相同的向量。最后,$sigmoid(\cdot)$函数对S_A和池化后的S_I归一化。融合后的特征为$M \in \mathbb{R}^{T\times N\times C'}$,融合过程的表达式如下:

$$M = S_A \otimes \sigma(AAP(S_I)) \tag{10}$$

为平衡基于动作条件的交互关系,本文利用注意力机制,突出行人动作特征S_A对行人过街意图的影响。防止数据过拟合将dropout设置为0.5,并用全连接层预测每个行人观测样本S的过街意图,其公式如下:

$$Y_P = Linear(\sigma(Att(SiLU(M) \otimes \boldsymbol{W}_M)) + M) \tag{11}$$

式中:$SiLU(\cdot)$为激活函数;$Att(\cdot)$为基于注意力的变换函数;$\boldsymbol{W}_M$为可学习的注意力权重。最后,通过$Linear(\cdot)$分类器输出C/NC标签。

1.5　损失函数

行人过街意图预测是一种有过街意图和无过街意图的二元分类任务。为评估两个概率之间的分布差异,本文引用二元交叉熵损失函数评估模型的性能。损失函数表达式如下:

$$L = -Y_T \cdot \log(Y_P) - (1 - Y_T) \cdot \log(1 - Y_P) \tag{12}$$

式中:Y_T为行人样本S过街意图的真实标签;Y_P为EAIPF的行人过街意图预测概率。

2　实验结果与分析

2.1　数据集

(1)JAAD[15]:JAAD数据集研究日常城市环境中自动驾驶背景下行人和驾驶员的行为和其他因素对他们的影响。为此,JAAD数据集提供了从240 h多的驾驶镜头中裁剪的346个视频剪辑,每个视频时长5~10 s且包含丰富的行人属性、外观和行为标签。

(2)PIE[23]:PIE数据集与JAAD数据集类似,同样研究交通场景中行人的行为。PIE数据集提供了加拿大多伦多晴朗天气连续6 h多的驾驶镜头。与JAAD数据集不同的是,PIE数据集提供了OBD传感器的自运动车辆信息标签,提供了交通对象属性信息的同时,补充了交通对象的位置边框信息。

2.2　数据预处理

如图5所示,事件(Event)定义为区分样本行人有过街意图和行人无过街意图标志。行人有过街意图样本以过街行为起始帧为Event,行人无过街意图样本以行人可观测帧最后一帧为Event。为短时预警路边行人突发性过街行为和学习行人行为动作,TTE(time to event)设置为1~2 s(30-60帧)。每个行人观测样本长度为16帧,在TTE内重叠采样,JAAD和PIE的样本重叠采样率分别为0.8和0.6。

行人观测样本数据格式为(T, N, C)。其中,T表示行人观测帧长度,N表示行人关节点数量,C表示(x, y, d, sc)的关节点维度。其中,x和y表示关节点的2D坐标,d表示关节点深度,sc表示关节点置信度。为增强多模态动作表征能力,本文利用SimCC提取行人骨架数据,并引入在KITTI数据集上预训练的R-MSFM单目深度估计器[24],提取观测序列每一帧行人骨架的深度信息。此外,添加了基于MLP

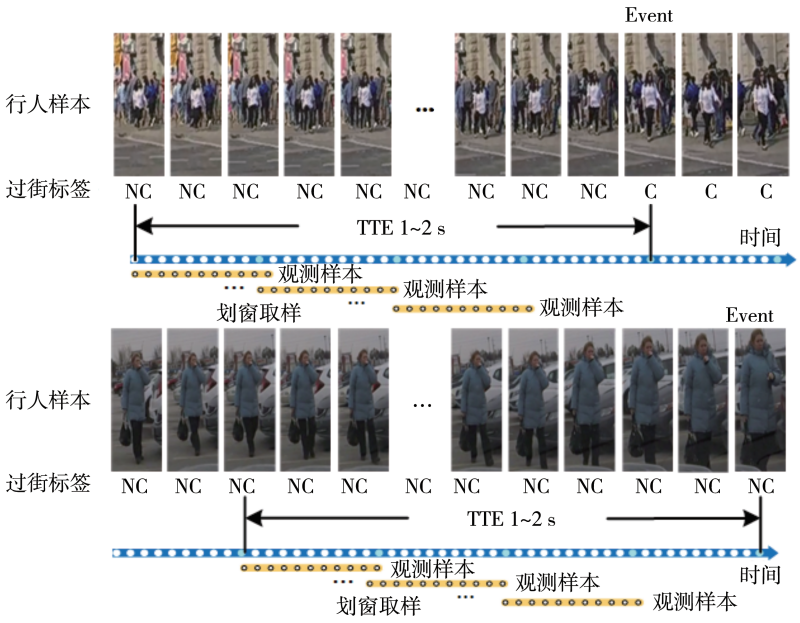

图 5 数据集行人样本采样过程

的人体姿态预测模块[25]。通过学习行人观测序列,预测行人未来 30 帧姿态,提升行人过街意图预测模型的性能。使用观测帧最后一帧作为交通场景,提取交通对象交互特征。PIE 数据集提供了每一帧中交通对象的边框坐标。与 PIE 数据集不同,JAAD 数据集只提供场景中是否有交通对象的数量标签,并没有交通对象的位置信息。为此,本文引入 Fast-RCNN 检测交通对象目标,并根据行人边框中心选取距离最近、交互频繁的 4 个交通对象进行交互关系建模。

2.3 评价指标

行人过街意图预测为二元分类任务。EAIPF 的性能可以用 Precision(Pre)、Recall(Rec)、Accuracy(Acc)、F1-score(F1)和 ROC_AUC(AUC)评价指标评估。这些评价指标可以用真阳性(TP)、伪阳性(FP)、真阴性(TN)和伪阴性(FN)表示:

$$\mathrm{Pre} = \frac{TP}{TP + FP} \tag{13}$$

$$\mathrm{Rec} = \frac{TP}{TP + FN} \tag{14}$$

$$\mathrm{Acc} = \frac{TP + TN}{TP + FP + TN + FN} \tag{15}$$

$$\mathrm{F1} = 2 \cdot \frac{\mathrm{Precision} \cdot \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{16}$$

2.4 实验细节设置

所采用的实验平台搭载了一块英伟达 Nvidia 3090 显卡和一块英特尔 I7 CPU,实验环境为 Ubuntu 18.04 系统、Pytorch 框架。本文使用 Adamw 优化器训练网络,每次训练批次大小为 128,训练 epoch 为 30。JAAD 的初始学习率为 0.005,PIE 的初始学习率为 0.003,使用 CosineAnnealingLR 学习策略,学习率以正弦规律变化进行训练。

2.5 销蚀实验

本文主要贡献是提出基于动作条件交互的高效行人过街意图预测框架 EAIPF,除从行人动作中预测其过街意图,还能够融合交通对象交互信息,做出综合决策。分别在 JAAD 与 PIE 数据集上进行销蚀实验,评价动作编码模块、预测模块和场景对象交互模块不同融合策略的有效性。结果如表 1~表 3 所示,预测模型在 5 个指标上都优于场景对象交互模块,说明行人动作中包含其是否准备过街的主要信息,交通对象交互信息可以作为行人动作的重要补充。EAIPF 将行人动作信息和交通对象交互信息深度融合,与单通道预测模型相比,在 3 个数据集中 Acc、AUC、F1、Pre 和 Rec 值都有显著提升。

2.6 对比实验

使用行人过街意图预测基准[26]对不同过街意图预测模型进行对比实验。表 4 展示了 JAAD 和 PIE 数据集上的对比结果。行人过街意图预测的主流方法主要集中在 3 种趋势上:卷积模型、循环模型和图卷积模型。ATGC[16]和 ConvLSTM[27]是较早提出的用卷积模型实现过街意图预测任务,使用卷积神经网络分析场景和预测行人的过街行为。循环模型对连

表1　EAIPF在JAAD_all数据集上销蚀实验　　%

2D动作编码	3D动作编码	预测模块	场景对象交互模块	Acc	AUC	F1	Pre	Rec
√				84.15	72.66	55.04	55.22	54.87
	√			84.04	74.59	57.07	54.43	59.98
√		√		86.27	82.88	66.43	58.03	77.66
	√	√		86.01	79.95	63.85	54.27	77.57
√			√	85.47	76.63	60.27	57.75	63.02
	√		√	84.61	77.23	59.96	55.01	65.88
√		√	√	88.96	84.10	78.81	85.66	72.98
	√	√	√	88.01	79.23	65.70	65.67	65.73

表2　EAIPF在JAAD_beh数据集上销蚀实验　　%

2D动作编码	3D动作编码	预测模块	场景对象交互模块	Acc	AUC	F1	Pre	Rec
√				63.76	55.48	75.12	86.76	66.18
	√			63.85	55.72	75.62	86.79	63.85
√		√		67.86	67.89	72.51	77.98	67.76
	√	√		68.05	67.74	72.93	77.49	68.97
√			√	64.15	61.77	71.31	71.41	71.21
	√		√	65.69	63.53	72.44	72.81	72.08
√		√	√	68.64	67.72	74.01	76.86	71.37
	√	√	√	69.10	65.03	76.69	72.64	81.21

表3　EAIPF在PIE数据集上销蚀实验　　%

2D动作编码	3D动作编码	预测模块	场景对象交互模块	Acc	AUC	F1	Pre	Rec
√				78.57	66.69	50.92	71.81	39.45
	√			79.80	69.60	56.34	72.07	46.24
√		√		87.75	81.26	75.31	86.98	66.41
	√	√		88.86	82.92	77.79	88.64	69.31
√			√	82.62	74.06	63.82	77.05	54.47
	√		√	84.10	78.98	70.43	73.91	67.26
√		√	√	88.93	82.59	77.58	90.16	68.08
	√	√	√	90.33	86.57	81.95	86.35	77.97

续视频帧的视觉特征之间的依赖关系进行建模,如SingleRNN[28]、MultiRNN[29]、StackedRNN[30]、HierarchicalRNN[31]和SFRNN[15]将图像堆栈作为输入,并使用堆叠RNN逐层融合视觉特征,提升行人过街意图预测性能。3D卷积模型能够捕获时序上的信息,如C3D[32]、I3D[33]和PCPA[34]将视频作为输入,使用3D卷积分支编码视觉信息,并使用RNN分支并行处理数据集提供的外观特征。图卷积模型建模图的结构属性和节点特征信息,如Pedestrian Graph+[11]引入图卷积网络建模行人姿态数据,并使用卷积模块处理图像和车辆速度。EAIPF具有基于动作条件交互的模型,PIE数据集在结果上Pre提高了3个百分点,Acc和F1各提高了1个百分点。尽管JAAD数据集没有提供交通对象边框坐标,而是通过Fast-RCNN检测定位,无法准确得到交通对象的属性和位置信息。且JAAD驾驶环境更为复杂,雨雪、黑夜等场景能见度低。场景对象交互模块难以获取影响行人过街的交互信息。但是,EAIPF也能捕获高级交互语义感知,能够获得比属性语义感知相比更好的效果。JAAD_all数据集在Acc上提升了3个百分点。JAAD_beh数据集在F1上提升了1个百分点,其他指标与SOTA模型不相上下。

表 4 EAIPF 与主流方法对比实验

模型名称	模型变体	PIE					JAAD_beh					JAAD_all				
		Acc	AUC	F1	Pre	Rec	Acc	AUC	F1	Pre	Rec	Acc	AUC	F1	Pre	Rec
ATGC[14]	VGG16	0.71	0.60	0.41	0.49	0.36	0.59	0.52	0.71	0.63	0.82	0.82	0.75	0.55	0.49	0.63
	ResNet50	0.70	0.59	0.38	0.47	0.32	0.46	0.45	0.54	0.58	0.51	0.81	0.72	0.52	0.47	0.56
ConvLSTM[27]	VGG16	0.58	0.55	0.39	0.32	0.49	0.53	0.49	0.64	0.64	0.64	0.63	0.57	0.32	0.24	0.48
	ResNet50	0.54	0.46	0.26	0.23	0.29	0.59	0.55	0.69	0.68	0.70	0.63	0.58	0.33	0.25	0.49
SingleRNN[28]	GRU	0.83	0.77	0.67	0.70	0.64	0.58	0.54	0.67	0.67	0.68	0.65	0.59	0.34	0.26	0.49
	LSTM	0.81	0.75	0.64	0.67	0.61	0.51	0.48	0.61	0.63	0.59	0.78	0.75	0.54	0.44	0.70
MultiRNN[29]	GRU	0.83	0.80	0.71	0.69	0.73	0.61	0.50	0.74	0.64	0.86	0.79	0.79	0.58	0.45	0.79
StackedRNN[30]	GRU	0.82	0.78	0.67	0.67	0.68	0.60	0.60	0.66	0.73	0.61	0.79	0.79	0.58	0.46	0.79
HierarchicalRNN[31]	GRU	0.82	0.77	0.67	0.68	0.66	0.53	0.50	0.63	0.64	0.61	0.80	0.79	0.59	0.47	0.79
SFRNN[15]	GRU	0.82	0.79	0.69	0.67	0.70	0.51	0.45	0.63	0.61	0.64	0.84	0.84	0.65	0.54	0.84
C3D[32]	3DConv	0.77	0.67	0.52	0.63	0.44	0.61	0.51	0.75	0.63	0.91	0.84	0.81	0.65	0.57	0.75
I3D[33]	3DConv	0.80	0.73	0.62	0.67	0.58	0.62	0.56	0.73	0.68	0.79	0.81	0.74	0.63	0.66	0.61
	Optical	0.81	0.83	0.72	0.60	0.90	0.62	0.51	0.75	0.65	0.88	0.84	0.80	0.63	0.55	0.73
PCPA[34]	3DConv	0.86	0.91	0.78	0.69	0.89	0.50	0.47	0.59	0.61	0.58	0.70	0.85	0.51	0.36	0.87
Ped Graph+[11]	GCN	0.89	0.90	0.81	0.83	0.79	0.70	0.70	0.76	0.77	0.75	0.86	0.88	0.65	0.58	0.75
Ours	GCN	0.90	0.87	0.82	0.86	0.78	0.69	0.65	0.77	0.73	0.81	0.89	0.84	0.79	0.86	0.73

2.7 意图预测定性分析

图 6 和图 7 示出 EAIPF、PCPA 和 Pedestrian Graph+ 在 JAAD 和 PIE 数据集中的定性结果。图中用红色边框标注每个样本的目标行人,用红色字体表示真实标签和真实标签一致的预测结果,用蓝色字体表示与真实标签不一致的预测结果。在提供的行人样本中,DPCIAN 能够准确地预测行人的过街意图。但是,PCPA 和 Pedestrian Graph+ 在一些样本中出现了预测误差。通过分析样本观测序列,从行人动作信息和场景信息两个方面将原因归纳以下几点。(1)图 6(a)~图 6(c)和图 7(c)样本所示,雨雪和夜晚等能见度低的交通场景中,PCPA 和 Pedestrian Graph+ 的行人姿态估计和动作编码性能明显下降。然而,EAIPF 中行人动作编码模块在低分辨率场景中挖掘骨架上下文信息,提升行人动作编码能力。(2)图 7(a)为行人遮挡样本,面对行人骨架不完备问题,EAIPF 利用场景对象交互模块建模目标行人和相关交通对象的交互信息,辅助行人动作编码模块补全场景理解信息,减少被遮挡行人过街意图预测误差。(3)图 7(b)为无关行人样本,在大部分情况下无关行人不会影响自动驾驶汽车的正常行驶,但无关行人本身具有穿越意图,且无关行人作为智慧交通场景的主要参与者,将影响其他行人的过街意图。EAIPF 以道路安全为准则,对场景对象交互关系进行建模,提升行人过街意图预测准确率的同时,也提升了道路安全。

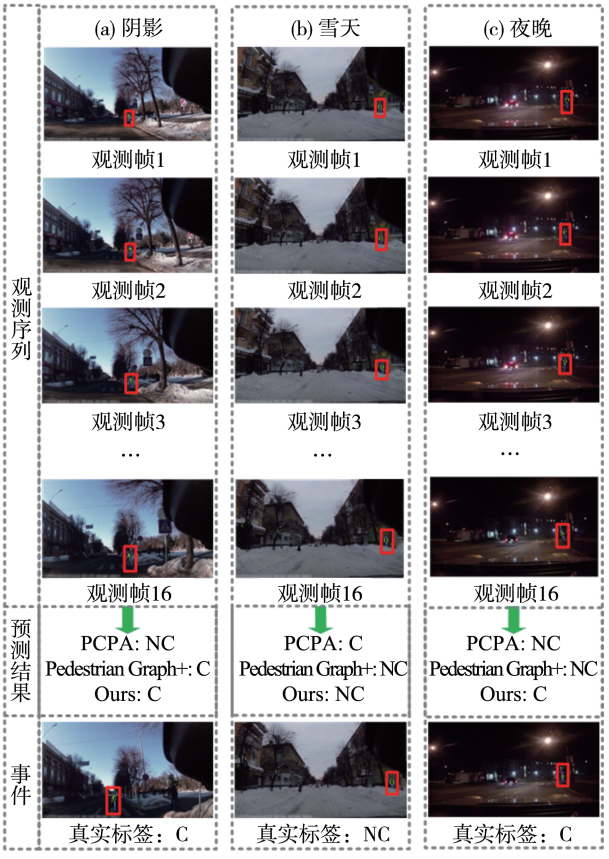

图 6 EAIPF 在 JAAD 数据集定性结果

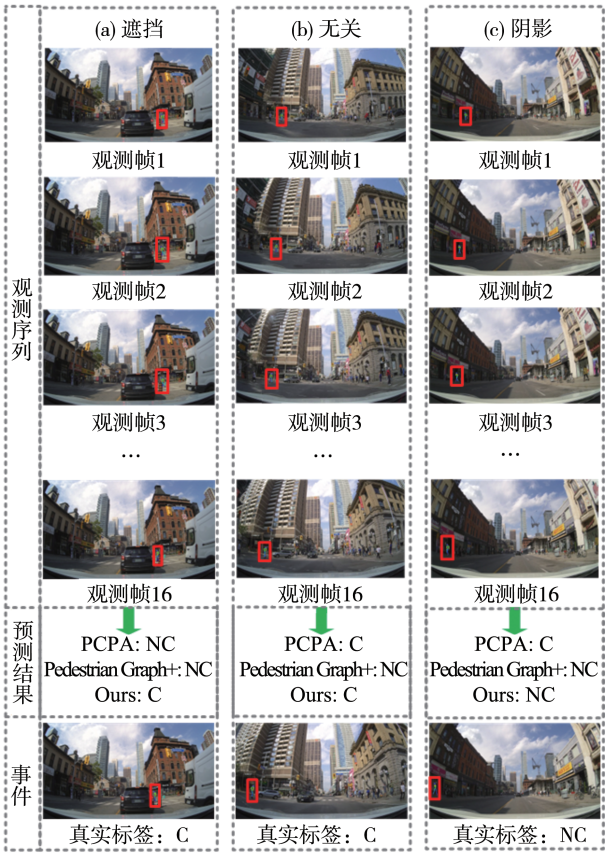

图7　EAIPF在PIE数据集定性结果

3　结论

提出一种EAIPF,融合行人动作信息和交通对象交互信息,用于预测行人过街意图。为在低分辨率场景下精确提取行人动作特征,行人动作编码模块使用人体关节点的向量表征方式和基于自注意力图卷积网络,增强多模态动作表征能力和挖掘骨架上下文信息。为解决交通对象交互信息瓶颈,利用场景对象交互模块建模目标行人和相关交通对象的交互信息,实现深入理解交通场景中高级语义线索。相比于其它主流算法,EAIPF在JAAD和PIE数据集上都取得了最佳性能。因此,EAIPF可应用于无人驾驶车辆或高级驾驶辅助系统,帮助预测人车冲突并提高驾驶舒适性。未来的工作重点是准确获取交通对象的属性和位置信息,并在雨雪、夜间等低能见度场景下精确建立场景对象交互,以提高行人过马路意图预测的准确性。

参考文献

[1] CHEN B, SUN D, ZHOU J, et al. A future intelligent traffic system with mixed autonomous vehicles and human-driven vehicles [J]. Information Sciences, 2020, 529: 59-72.

[2] 连静, 王欣然, 李琳辉, 等. 基于人-车交互的行人轨迹预测[J]. 中国公路学报, 2021, 34(5): 215.
LIAN J, WANG X R, LI L H, et al. Pedestrian trajectory prediction based on human-vehicle interaction [J]. China Journal of Highway and Transport, 2021, 34(5): 215.

[3] LI J, SU W, WANG Z. Simple pose: rethinking and improving a bottom-up approach for multi-person pose estimation [C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020: 11354-11361.

[4] LUO Z, WANG Z, HUANG Y, et al. Rethinking the heatmap regression for bottom-up human pose estimation[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021: 13264-13273.

[5] LIU J, ROJAS J, LI Y, et al. A graph attention spatio-temporal convolutional network for 3D human pose estimation in video[C]. 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021: 3374-3380.

[6] LI Y, YANG S, LIU P, et al. SimCC: a simple coordinate classification perspective for human pose estimation[C]. Computer Vision-ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VI. Cham: Springer Nature Switzerland, 2022: 89-106.

[7] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation [C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 5693-5703.

[8] YU B, YIN H, ZHU Z. Spatio-temporal graph convolutional networks: a deep learning framework for traffic forecasting[J]. arXiv preprint arXiv:1709.04875, 2017.

[9] SHI L, ZHANG Y, CHENG J, et al. Two-stream adaptive graph convolutional networks for skeleton-based action recognition[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 12026-12035.

[10] YE F, PU S, ZHONG Q, et al. Dynamic GCN: context-enriched topology learning for skeleton-based action recognition[C]. Proceedings of the 28th ACM International Conference on Multimedia, 2020: 55-63.

[11] 杨彪, 范福成, 杨吉成, 等. 基于动作预测与环境条件的行人过街意图识别[J]. 汽车工程, 2021, 43(7): 1066-1076.
YANG Biao, FAN Fucheng, YANG Jicheng, et al. Recognizing pedestrians' crossing intentions based on action prediction and environment context [J]. Automotive Engineering, 2021, 43(7): 1066-1076.

[12] CADENA P R G, QIAN Y, WANG C, et al. Pedestrian Graph+: a fast pedestrian crossing prediction model based on graph convolutional networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2022.

[13] YANG D, ZHANG H, YURTSEVER E, et al. Predicting pedestrian crossing intention with feature fusion and spatio-temporal attention[J]. IEEE Transactions on Intelligent Vehicles, 2022, 7

(2): 221-230.

[14] NI R, YANG B, WEI Z, et al. Pedestrians crossing intention anticipation based on dual-channel action recognition and hierarchical environmental context[J]. IET Intelligent Transport Systems, 2023, 17(2): 255-269.

[15] 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.

[16] RASOULI A, KOTSERUBA I, TSOTSOS J K. Pedestrian action anticipation using contextual feature fusion in stacked RNNs[J]. arXiv preprint arXiv:2005.06582, 2020.

[17] SHI J, LIU C, ISHI C T, et al. Skeleton-based emotion recognition based on two-stream self-attention enhanced spatial-temporal graph convolutional network[J]. Sensors, 2020, 21(1): 205.

[18] STERGIOU A, POPPE R. Adapool: exponential adaptive pooling for information-retaining downsampling[J]. IEEE Transactions on Image Processing, 2022, 32: 251-266.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[20] CHEN Y, ZHANG Z, YUAN C, et al. Channel-wise topology refinement graph convolution for skeleton-based action recognition[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 13359-13368.

[21] GIRSHICK R. Fast R-CNN[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 1440-1448.

[22] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 2117-2125.

[23] 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.

[24] ZHOU Z, FAN X, SHI P, et al. R-MSFM: recurrent multi-scale feature modulation for monocular depth estimating[C]. Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 12777-12786.

[25] BOUAZIZI A, HOLZBOCK A, KRESSEL U, et al. Motionmixer: MLP-based 3D human body pose forecasting[J]. arXiv preprint arXiv:2207.00499, 2022.

[26] KOTSERUBA I, RASOULI A, TSOTSOS J K. Benchmark for evaluating pedestrian action prediction[C]. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021: 1258-1268.

[27] SHI X, CHEN Z, WANG H, et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting[J]. Advances in Neural Information Processing Systems, 2015, 28.

[28] KOTSERUBA I, RASOULI A, TSOTSOS J K. Do they want to cross? understanding pedestrian intention for behavior prediction[C]. 2020 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2020: 1688-1693.

[29] 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.

[30] NG J Y, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: deep networks for video classification[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015: 4694-4702.

[31] LIN R, LIU S, YANG M, et al. Hierarchical recurrent neural network for document modeling[C]. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 899-907.

[32] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]. Proceedings of the IEEE International Conference on Computer Vision, 2015: 4489-4497.

[33] LI J, WANG C, ZHU H, et al. CrowdPose: efficient crowded scenes pose estimation and a new benchmark[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 10863-10872.

[34] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299-6308.