基于多特征融合的行人过街意图推理方法

尹守国[1]， 杜泉成[2]， 李灵犀[3]， 王晓[1,4]， 孙长银[5]

（1. 安徽大学人工智能学院，安徽 合肥 230601；
2. 北京科技大学计算机与通信工程学院，北京 100083；
3. 美国普渡大学电子与计算机工程系，印第安纳州 西拉法叶 IN 46204；
4. 光电信息获取与防护技术全国重点实验室，安徽 合肥 230031；
5. 安徽大学，安徽 合肥 230601）

摘 要：准确理解和预测行人过街意图对自动驾驶汽车的行驶安全至关重要。现有方法多依赖于行人轨迹或整体身体姿态等视觉运动特征，而对行人与车辆之间的交互信号关注不足，因此难以捕捉行人通过手势、头部朝向等细微信号所传递的通行意图。为了解决这些限制，提出了准确推理行人过街意图（accurate reasoning for pedestrian crossing intent，ARPCI）框架，这是一个多特征融合框架。具体而言，设计了一个行人特征模块，该模块首先关注行人的骨架特征以捕捉行人的运动趋势，在此基础上利用 MobileNet 提取头部姿态特征，结合 YOLOv8n 识别手部动作，从而获得行人与车辆间的交互信号。此外，还引入了场景编码模块和自车特征模块，这能够有效融合环境上下文与车辆动态信息，增强模型对复杂交通场景的适应能力，提高对行人过街意图的预测准确率。在广泛使用的 JAAD 数据集上进行的实验表明，该方法准确率达到了 88%，优于多个同类模型 SOTA（state of the art），消融实验也进一步验证了各输入特征的有效性。

关键词：行人过街意图；多模态特征融合；交互信号；行驶安全
中图分类号：TP391.4
文献标志码：A
doi: 10.11959/j.issn.2096−6652.202601

Pedestrian crossing intent inference method based on multi-feature fusion

Yin Shouguo[1], Du Quancheng[2], Li Lingxi[3], Wang Xiao[1,4], Sun Changyin[5]

1. School of Artificial Intelligence, Anhui University, Hefei 230601, China
2. School of Computer and Communication Engineering, University of Science and Technology Beijing, Beijing 100083, China
3. School of Electrical and Computer Engineering, Purdue University, West Lafayette IN 46204, USA
4. National Key Laboratory of Optoelectronic Information Acquisition and Protection Technology, Hefei 230031, China
5. Anhui University, Hefei 230601, China

Abstract: Accurately understanding and predicting pedestrian crossing intent is crucial for ensuring the safety of autonomous vehicles. Existing approaches are often limited to visual motion cues such as pedestrian trajectories or body poses, while overlooking interactive signals like gestures and head orientations, making it difficult to capture key cues of pedestrian-vehicle interaction. To address these limitations, ARPCI (accurate reasoning for pedestrian crossing intent) was proposed, a multi-feature fusion framework designed for pedestrian intent inference. Specifically, a pedestrian feature module was developed that first focused on skeleton-based features to capture motion trends, and further leveraged MobileNet to extract head pose features. Combined with YOLOv8n for gesture recognition, pedestrian-vehicle interaction

收稿日期：2025−09−27；**修回日期：**2025−12−01
通信作者：王晓，xiao.wang@ahu.edu.cn
基金项目：国家自然科学基金项目（No.62522601）
Foundation Item: The National Natural Science Foundation of China (No.62522601)

signals were captured more comprehensively by the model. In addition, a scene encoding module and a self-vehicle feature module were introduced to integrate contextual and ego-dynamic information, thereby enhancing adaptability to complex traffic environments and improving prediction accuracy. Extensive experiments on the widely used JAAD dataset show that the approach achieves an accuracy of 88%, surpassing several state-of-the-art counterparts. Moreover, the ablation studies provide further evidence of the effectiveness of the proposed input features.

Key words: pedestrian crossing intention, multi-feature fusion, interaction signal, traffic safety

0 引言

根据《中国统计年鉴》，2023年中国交通事故总数达25.47万起，造成直接财产损失117.933亿元。行人作为弱势交通参与者[1-2]，其行为具有一定的随机性[3]，准确识别并理解其行为意图，是提升道路安全性的关键基础[4-6]。行人的行为意图会随时间推移而呈现动态变化[7-8]，因此，将过街意图识别问题视为时序建模任务，能够更合理地揭示其随时间演化的规律[9]，并挖掘潜在的语义层次特征[10-12]。

早期研究采用传统的模型驱动方法对行人运动模式进行建模，如运动学模型，然而，这些方法在捕捉复杂的运动交互模式方面存在局限性，难以有效处理行人之间的动态交互及与场景上下文之间的耦合关系[13]。这些方法通常依赖于运动学或基于物理的模型，国外研究者Helbing等[14]首次提出社会力模型，通过能量势场（排斥力与吸引力）来描述行人之间的交互关系，它使用能量势场的概念来表征行人运动，然后利用排斥力与吸引力的合力驱动行人运动进而预测行人轨迹，但该模型往往忽视环境语义信息（如人行横道、遮挡物等）对行人决策的关键影响，仅依赖运动轨迹和简单物理量，导致其在真实复杂道路场景中泛化能力较差。为提升模型的预测性能，后续研究提出了基于行人行为分析的方法，通过解析人体关键点的运动模式与动力学特征，推断其是否具有横穿道路的意图。Fang等[15]提出了一种基于单目视觉的行人过街意图预测方法，该方法首先对视觉图片进行行人检测和骨骼关键点提取，然后从骨骼关键点中提取高层次特征并采用支持向量机（support vector machine，SVM）对这些特征进行融合和分类，以判断行人是否具有“横穿”“停止”“转向”或“起步”等行为意图。然而，由于行人过街行为本身具有高度的随机性，加之城市交通通行标识牌不完善及交通流量环境复杂多变[16-17]，该方法在实际应用中仍面临分类准确性不足的问题。为解决这一问题，后续研究引入了深度学习模型，如循环神经网络（recurrent neural network，RNN）与长短期记忆网络（long short-term memory，LSTM），以利用其强大的非线性特征提取能力，更好地建模行人动态行为中的复杂时空模式。Li等[18]采用门控循环单元（gated recurrent unit，GRU）处理从2D姿态数据提取的静态时空特征，包括位置、距离和关节之间的角度及其时序差分动态特征，通过序列建模实现行人状态识别。近期一些研究表明，行人在穿越马路前的某些行为，如步态变化、手势示意及视线注视与其过街意图之间存在显著关联。这一发现推动了基于人体姿态特征的行人过街意图预测方法的兴起。在早期研究中，此类预测主要依赖于单一模态的特征信息，如行人整体姿态，或是针对上半身动作的细化分类。Varytimidis等[19]使用了卷积神经网络（convolutional neural network，CNN）结合分类器，通过分析行人头部方向和运动状态来预测行人是否要过街。Muhamma等[20]提出了一种基于头部姿态估计的行人过街意图早期预警方法。该方法首先使用YOLOv3检测行人头部，再采用宽范围头部姿态估计网络（wide-range head pose estimation network，WHENet）模型估计头部的偏航角、俯仰角和翻滚角，最后利用聚类算法对这些角度特征进行分类，以判断行人是否具有过街意图。现有的行人过街意图预测方法大多依赖于单一模态的特征，如整体姿态、局部动作等，这导致它们难以全面捕捉行人行为中的多层次交互信号。在复杂的交通环境中，尤其是在行人与自车、环境之间存在动态交互时，现有方法的准确性和鲁棒性常常受到限制。为了解决这一问题，本文提出了一种新的行人过街意图预测框架——准确推理行人过街意图（accurate reasoning for pedestrian crossing intent，ARPCI）。ARPCI框架的创新之处在于，它不仅依赖于行人的骨架特征，还着重考虑了行人头部姿态和手势动作所传递的交互信号，并将环境信息和自车特征进行有效融合。本文还设计了一个加权融合机制，根据不同特

征在不同场景下的重要性自动调整其权重，最终综合判断行人是否具有过街意图。这一策略能够更准确地捕捉行人在动态交通环境中的行为模式，尤其是在复杂的道路场景和行人与车辆的交互关系中，显著提升了过街意图的预测性能。本文主要贡献有以下3个方面。

（1）提出一种基于多特征融合的行人过街意图预测框架ARPCI。与以往多模态方法的简单模态叠加不同，ARPCI将头部偏航角、手势动作、骨架动态、自车状态以及局部/全局场景语义视为互补的意图线索，并在统一的推理框架中进行结构化特征提取与语义建模。模型能够捕捉行人与车辆在过街决策中的交互因素（如行人注视方向、行人手势与车辆交互意图等），从而提升模型在复杂交通场景下对行人意图的判别能力。基于JAAD数据集的实验结果表明，本文方法在准确率方面显著优于现有多模态融合基线方法，达到88%，准确率提升了4%。

（2）对YOLOv8n手势识别模块进行了结构增强，结合轻量化网络模块VanillaNet以重构特征提取路径，并加入卷积块注意力模块（convolutional block attention module，CBAM）增强时序数据的处理能力，使模型能够显式聚焦于与意图相关的关键手势区域，提高复杂场景下行人手势交互行为分析的准确性。

（3）提出一种加权融合机制，通过可学习的权重参数ω_n建模不同模态在不同交通场景下的重要性，根据不同模态在特征表达中的有效性，通过权重参数ω_n自动调整各模态的贡献，从而实现注意力机制难以提供的跨模态语义级别的调节与场景自适应性。

1　相关工作

根据目前主流研究方法的思路，行人过街意图推理方法主要可分为两类：一类是基于轨迹预测的方法，另一类是基于二元分类的方法。

1.1　基于轨迹预测的方法

基于轨迹预测的方法通常利用行人过去一段时间的运动序列（如位置、速度等[21-22]），利用时序模型编码其运动模式，通过捕捉运动历史中的细微变化（如减速、朝向调整）实现对其未来行为的推断，从而判断行人是否有横穿马路的意图[23-24]。Gupta等[25]使用生成对抗网络（generative adversarial network，GAN）来针对拥挤场景下行人轨迹的预测，提出一种基于编码器-解码器的架构，先编码目标行人的过去轨迹，再解码输出多条未来轨迹；判别器以“真实/虚假”对抗损失迫使生成轨迹符合社会规范，从而提高预测准确性。Lin等[26]提出了基于锚点的Transformer网络，通过引入可学习的意图锚点并结合历史轨迹上下文信息，在统一的编码器-解码器框架中同时进行行人轨迹预测和过街意图预测。该方法在JAAD数据集上取得了较优性能，轨迹预测误差降低9%，过街意图预测的精确度提升11%。Wang等[27]提出了一种基于时空注意力机制的行人轨迹与意图预测方法。该方法通过时空特征提取模块捕获行人运动模式，利用多头注意力机制编码轨迹序列中的长短时依赖关系，并设计专门的意图识别模块将轨迹特征与过街意图（直行、左转、右转、停止等）显式关联，进而准确地推理行人过街意图。Mohamed等[28]提出了社交时空图卷积网络（social spatio-temporal graph convolutional network，Social-STGCN）模型，通过将行人轨迹建模为时空图，利用图卷积网络直接捕捉行人之间的社会交互，该方法结合自定义的基于距离的核函数构建加权邻接矩阵，量化行人间的相互影响，并使用时域外推卷积网络一次性预测所有行人的未来轨迹。尽管基于轨迹预测的方法能够较好地刻画行人运动的时序特征，但仍存在一定不足：此类方法往往依赖较长时间的历史轨迹，对短时观测或存在遮挡的场景适应性较弱；此外，模型主要关注行人历史轨迹的变化，缺乏对行人意图和环境语义的深入建模，因此在复杂交通场景下预测精度受到限制。

1.2　基于二元分类的方法

基于二元分类的方法通常直接预测行人在未来数秒内是否会横穿马路，该类方法因其高效与直接的特点，适用于自动驾驶系统中车载视角下的实时意图推理任务。Tang等[29]提出一种可分离的密度辅助行人检测（density-aware decomposable anchor-based detector，DDAD）的多任务学习方法，用于提升复杂场景下的行人检测性能。DDAD在传统检测器的基础上增加了一个可分离的“人群密度估计”分支，通过现有边界框生成近似密度标注，迫使模型更关注行人头部和上边界，通过深层特征融合加强两个任务的协同作用，进一步提升预测准确率。胡远志等[30]提出了一种双流自适应图卷积网

络，该方法充分利用了行人的动态姿态特征，并将其与过街意图信息进行融合。在建模过程中，一方面捕捉行人身体关键点的运动模式，另一方面结合其语义层面的过街意图，从而实现了空间与时间特征的深度关联建模，从而提高预测精确度。Chaabane 等[31]首先利用编码器-解码器网络根据车辆视角的历史视频帧预测未来视频帧，其次将预测的未来帧通过时空深度网络识别行人是否会穿过车辆行驶路径。Xie 等[32]也明确将行人过街意图预测定义为二元分类任务：判断行人是否会在未来 1~2 s 内横穿马路，并提出一种名为具有位置解耦机制的图嵌入式 Transformer（graph-embedded Transformer with positional decoupling module，GTransPDM）的模型，该模型结合了图卷积网络和 Transformer 来预测行人过街意图。Cadena 等[33]提出了一种基于图卷积网络的快速行人过街意图预测模型，该模型融合了行人姿态、图像上下文等信息，通过其全卷积架构实现了对输入序列时间长度的动态灵活处理，进而预测行人意图。Liu 等[34]通过构建以行人为中心的动态场景图，整合行人、车辆、交通标志等信息，并引入位置中心化预测视角来预测行人是否过街。Piccoli 等[35]提出了一种名为全时空融合网络（fully spatio-temporal fusion network，FuSSI-Net）的端到端行人过街意图预测网络，通过融合目标检测的边界框信息与人体姿态的骨骼关键点特征，有效降低了误报率。基于二元分类的方法虽然能够直接判断行人是否会横穿马路，而且推理速度快、易于部署，但也存在模型对数据分布较为敏感、跨场景泛化能力不足等局限性。

2　基于多特征融合的行人过街意图推理研究方法

行人过街意图推理的目的是在复杂交通环境中预测行人在未来短时间内是否具有过街的意图，以便为驾驶员提供提前决策的依据，从而有效降低人车冲突的风险。本文将行人过街意图定义为一个二元分类任务，包括有过街意图（C）和无过街意图（NC）两类。本文提出的行人过街意图推理整体模型框架如图 1 所示，该模型框架主要由自车特征模块、行人特征模块、场景特征模块和多特征融合模块组成。下面对该模型中各输入特征的获取以及各个模块进行详细介绍。

2.1　自车特征模块

本文提出的模型主要在 JAAD 数据集上进行实验验证。该数据集提供了自车的显式标签信息（如速度状态）以及行人的边界框位置标注。因此，自车速度可以直接由数据集标签获得，用于表征自车在不同时刻的运动状态（如加速、减速、匀速或静止）。自车速度序列 $\boldsymbol{V}_{\mathrm{v}}$ 可以表示为：

$$\boldsymbol{V}_{\mathrm{v}}=\left\{v_{\mathrm{v}}^{t-n},v_{\mathrm{v}}^{t-n+1},\cdots,v_{\mathrm{v}}^{t}\right\} \tag{1}$$

其中，v_{v}^{t} 表示自车在第 t 帧的速度大小。为反映自车在连续帧间的移动方向，定义自车从第 t–1 帧到第 t 帧的位移向量为：

$$\boldsymbol{m}_{\mathrm{v}}^{t}=\left(x_{\mathrm{v}}^{t}-x_{\mathrm{v}}^{t-1},y_{\mathrm{v}}^{t}-y_{\mathrm{v}}^{t-1}\right) \tag{2}$$

其中，（x_{v}^{t}，y_{v}^{t}）表示自车在第 t 帧的中心坐标，基于上述单帧位移向量，可构建自车在最近 n 帧内的轨迹方向序列，该序列反映了自车在连续时间上的

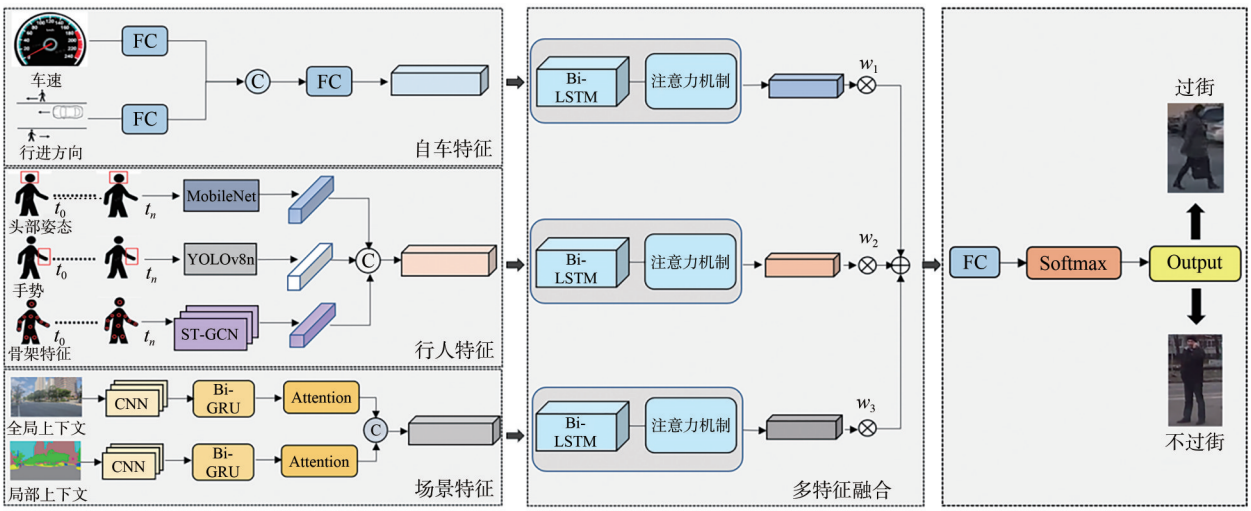

图 1　整体模型框架

移动方向变化：

$$\boldsymbol{M}_{\mathrm{v}}=\left\{\boldsymbol{m}_{\mathrm{v}}^{t-n},\boldsymbol{m}_{\mathrm{v}}^{t-n+1},\cdots,\boldsymbol{m}_{\mathrm{v}}^{t}\right\} \tag{3}$$

目标行人 i 的二维边界框位置由数据集提供的标注直接获取，表示为：

$$\boldsymbol{L}_i=\left\{\boldsymbol{L}_i^{t-m},\boldsymbol{L}_i^{t-m+1},\cdots,\boldsymbol{L}_i^{t}\right\} \tag{4}$$

其中，$\boldsymbol{L}_i^{t-m}=\left\{x_{i,\mathrm{tl}}^{t-m},y_{i,\mathrm{tl}}^{t-m},x_{i,\mathrm{br}}^{t-m},y_{i,\mathrm{br}}^{t-m}\right\}$，$x_{i,\mathrm{tl}}^{t-m}$ 和 $y_{i,\mathrm{tl}}^{t-m}$ 表示边界框的左上角点，$x_{i,\mathrm{br}}^{t-m}$ 和 $y_{i,\mathrm{br}}^{t-m}$ 表示边界框的右下角点。

行人从第 $t-1$ 帧到第 t 帧的位移向量定义为：

$$\boldsymbol{m}_{\mathrm{p}}^{t}=\left(x_{\mathrm{p}}^{t}-x_{\mathrm{p}}^{t-1},y_{\mathrm{p}}^{t}-y_{\mathrm{p}}^{t-1}\right) \tag{5}$$

其中，$\left(x_{\mathrm{p}}^{t},y_{\mathrm{p}}^{t}\right)$ 为第 t 帧行人边界框的中心位置，由此可构建行人在最近 m 帧内的运动方向序列：

$$\boldsymbol{M}_{\mathrm{P}}=\left\{\boldsymbol{m}_{\mathrm{p}}^{t-m},\boldsymbol{m}_{\mathrm{p}}^{t-m+1},\cdots,\boldsymbol{m}_{\mathrm{p}}^{t}\right\} \tag{6}$$

基于自车与行人的运动方向向量，可以进一步计算它们之间的夹角，夹角通过下式计算：

$$\theta=\arccos\left(\frac{\boldsymbol{m}_{\mathrm{p}}^{t}\cdot\boldsymbol{m}_{\mathrm{v}}^{t}}{\left\|\boldsymbol{m}_{\mathrm{p}}^{t}\right\|\cdot\left\|\boldsymbol{m}_{\mathrm{v}}^{t}\right\|}\right) \tag{7}$$

其中，·表示向量点积，$\boldsymbol{m}_{\mathrm{p}}^{t}$ 和 $\boldsymbol{m}_{\mathrm{v}}^{t}$ 分别表示行人与自车在最新一帧的位移向量，利用二者的夹角判断行进方向是否一致。如果夹角小于设定的阈值角度，则认为行人和车辆的行进方向一致；否则，认为方向不一致。

2.2　行人特征模块

（1）行人头部特征编码

行人的头部朝向与其注视方向密切相关，因此可作为判断其是否具有过街意图的重要依据。然而，在包含多个行人目标且距离较远的复杂交通环境中，要精确获取目标行人的头部特征较为困难。传统的头部朝向建模通常用偏航角 Yaw、俯仰角 Pitch 和翻滚角 Roll 3 种转角来表征头部方向特征。在实际交通场景中，偏航角 Yaw 被普遍认为是最能反映行人注意力方向的角度特征，而俯仰角 Pitch 与翻滚角 Roll 对行人意图推断的贡献有限。Yang 等[36]与 Zhang 等[37]的研究表明，行人在步行或道路交互中头部主要发生水平方向的转动，而俯仰角与翻滚角通常较小，且多与低头使用手机、点头等短时动作相关，并不直接关联行人是否关注来车方向。因此，偏航角 Yaw 是判断行人是否观察车辆、是否准备横穿马路的关键指标。

基于上述分析，本文选择能够有效反映行人在水平面上头部转向情况的偏航角 Yaw 作为头部方向分类的主要特征。本文偏航角的定义如下：以自车行驶方向为参考方向，当行人面部正对该参考方向时记为 0°，头部向左偏转时角度取正，向右偏转时角度取负，取值范围为[−180°，180°]，为实现离散化建模，将整个 360°空间范围均匀划分为 8 个角度区间，每个区间跨度为 45°，分别对应 8 个头部朝向类别。8 类方向既能提供足够的区分度，又能显著降低角度噪声，实现准确且稳定的头部方向分类。行人头部偏航角分类如图 2 所示。本文采用 MobileNet 作为分类网络，对行人头部区域图像进行特征提取与角度分类。分类器的输出用于识别行人是否面向行驶中的车辆，并根据其凝视方向推测未来的行为意图。

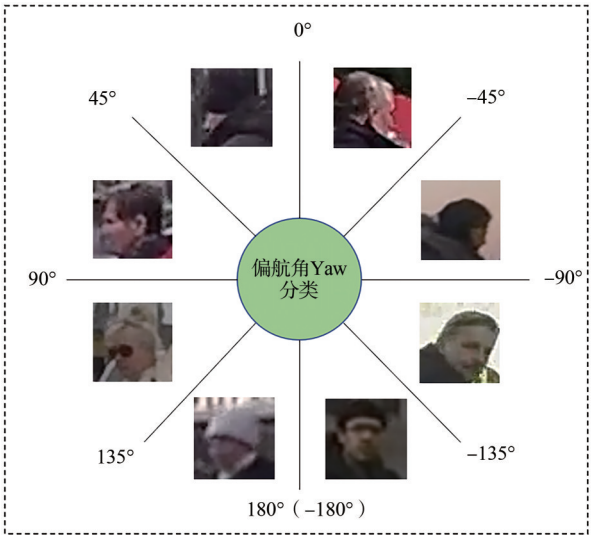

图2　行人头部偏航角分类

（2）行人手势特征编码

行人与驾驶员的交互信号也是影响其过街行为的因素之一。为向驾驶员明确传递通行意图，行人常采用眼神接触或特定手势进行示意，如通过手臂摆动传递“你可以通行，我已决定让你先行”的信息。本文针对当前行人手势识别方法在复杂背景适应性以及模型效率与准确率平衡方面的局限性，采用改进的 YOLOv8n 轻量级手势识别模型来提取行人的手势动作。本文改进工作主要包括以下两点：首先，在骨干网络中采用 VanillaNet[38]轻量化结构，显著降低了参数量和模型复杂度，从而提升推理速度；其次，引入计算高效的 CBAM[39]注意力

机制，引导模型聚焦于手势关键区域，降低复杂背景及环境因素对检测任务的干扰，从而提升手势识别的准确性。改进的 YOLOv8n 网络结构如图 3 所示。骨干网络由 VanillaNet 和空间金字塔池化（spatial pyramid pooling-fast，SPPF）组成，颈部网络核心由 C2f（CSP bottleneck with 2 convolutions）模块构成，头部网络由检测头组成。

VanillaNet 网络仅由标准卷积层和池化层构成，未引入跳跃连接或多分支设计，显著降低了计算负载与参数规模。由于其极简的模块化结构能够有效减少整体模型的参数量与计算复杂度，在保持检测性能的同时提升部署效率。基于上述优势，本文将 VanillaNet 作为骨干网络进行集成。VanillaNet 结构如图 4 所示。该网络主要由 3 个部分构成。首先，输入三通道 RGB（red-green-blue）的彩色图像经由第一部分的 Steam Block 模块通过卷积操作扩充通道数至 C，并完成下采样。其次，第二主体部分包含 4 个阶段：前 3 个阶段均仅由单层网络层构成，采用“1×1 卷积接步长为 2 的最大池化”结构，每经过一个阶段，特征图通道数翻倍；第 4 个阶段则采用“1×1 卷积接全局平均池化”结构，将特征图的空间尺寸压缩至 1×1，同时保留通道维度，从而将前层提取的特征聚合为类别数目对应的一维向量，以适配分类任务。最后，网络通过一个全连接层输出不同类别的预测概率。

在实际交通场景中，手势识别常面临目标行人手势与其他皮肤区域重叠的干扰，这类复杂背景易导致识别性能下降。为抑制环境噪声与背景干扰，增强模型对关键手势特征的捕捉能力，本文通过引入 CBAM，使模型更聚焦于目标行人手势区域，从而提升检测精度与鲁棒性。CBAM 结构如图 5 所示，包含通道注意力模块和空间注意力模块两个子部分。

通道注意力模块结构如图 6 所示。给定一个尺寸为 $H\times W\times C$ 的输入特征图，该模块首先分别执行全局平均池化与全局最大池化操作，以聚合空间信息，生成两个一维通道特征向量。再将这两个向量送入同一个由多层感知机构成的共享权重网络中进行非线性变换。将多层感知机输出的两个结果逐元素相加后，通过 Sigmoid 激活函数生成最终的通道注意力权重。该权重与原始输入特征图逐通道相乘，即可得到细化后的特征图 $\boldsymbol{P}_c$，其计算过程如下：

$$\boldsymbol{P}_c=\sigma\left(\mathrm{MLP}\left(\mathrm{Concat}\left(\mathrm{AvgPool}(\boldsymbol{F}),\mathrm{MaxPool}(\boldsymbol{F})\right)\right)\right) \tag{8}$$

再将特征图 $\boldsymbol{P}_c$ 与输入图 $\boldsymbol{F}$ 进行逐元素乘法，即可得空间注意力模块输入 $\boldsymbol{F}_c$：

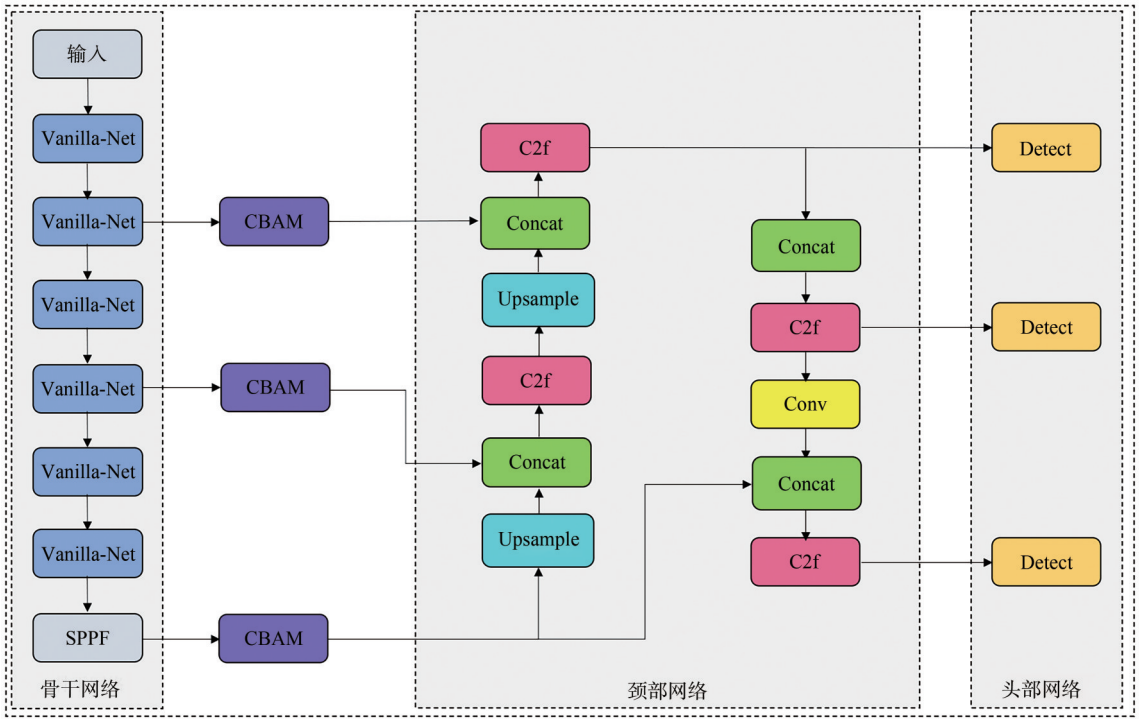

图 3 改进的 YOLOv8n 网络结构

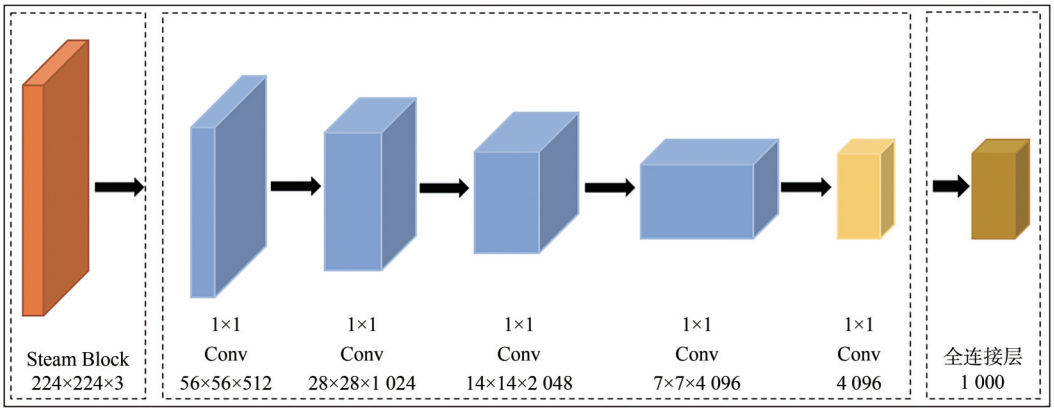

图4　VanillaNet结构

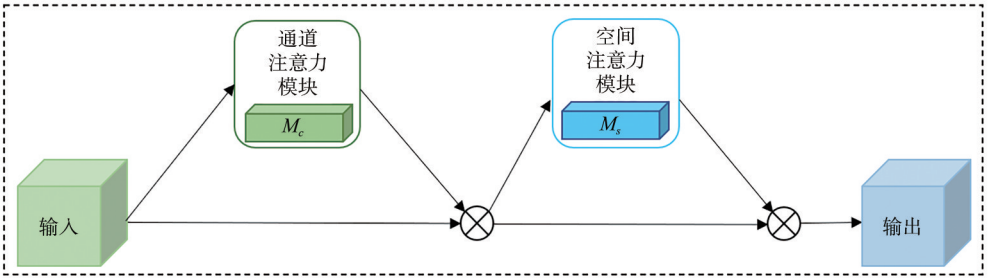

图5　CBAM结构

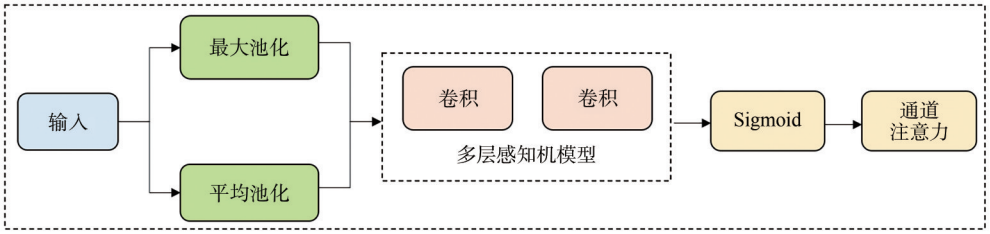

图6　通道注意力模块结构

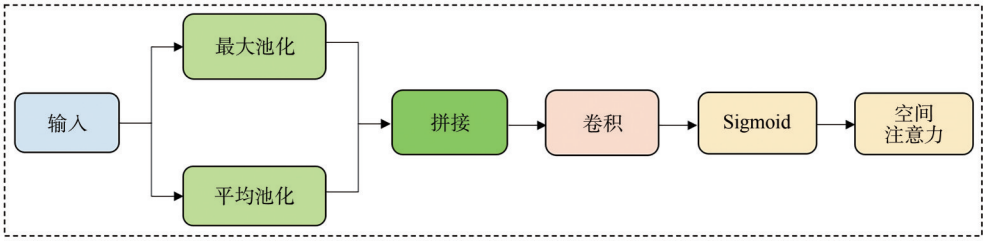

图7　空间注意力模块结构

$$\boldsymbol{F}_c=\boldsymbol{P}_c\otimes\boldsymbol{F} \tag{9}$$

空间注意力模块结构如图7所示。该模块以通道注意力模块的输出为输入，首先分别经过全局最大池化和全局平均池化处理，得到两个空间特征图；接着将这两个特征图在通道维度上进行拼接，并通过卷积操作实现特征融合与降维；最后经Sigmoid函数激活，生成最终的空间注意力特征图$\boldsymbol{M}_s$，其计算过程如下：

$$\boldsymbol{M}_s=\sigma\left(\mathrm{CONV}\left(\mathrm{Concat}\left(\mathrm{AvgPool}\left(\boldsymbol{F}_c\right),\mathrm{MaxPool}\left(\boldsymbol{F}_c\right)\right)\right)\right) \tag{10}$$

最终，将空间注意力特征图M_s与输入特征$\boldsymbol{F}_c$逐元素相乘，得到增强后的特征图$\boldsymbol{F}_s$，计算式表示如下：

$$\boldsymbol{F}_s=\boldsymbol{M}_s\otimes\boldsymbol{F}_c \tag{11}$$

在YOLOv8n的骨干网络及特征金字塔模块之

间引入CBAM注意力机制，以实现通道与空间维度信息的有效融合。该方法在不引入额外计算负担的前提下，增强了模型对关键特征的提取与整合能力，从而同时提高了检测精度与推理效率。

为验证所引入的双层注意力机制在特征提取中的有效性，本文对CBAM模块的注意力权重分布进行了分析。结果表明，通道注意力层能够自适应地增强与手势相关的语义特征通道，并抑制背景噪声通道；空间注意力层能够使模型聚焦于与手势识别相关的关键空间区域。此外，在后续时序注意力层中，模型在行人做出动作准备、转头等关键帧上的注意力权重明显较高，表明该机制能够自适应地捕捉具有行为意义的时间片段，实现对关键动态的聚焦，从而进一步提高模型对动态过街意图的识别准确性与鲁棒性。

（3）行人骨架特征编码

行人骨架姿态反映了与过街行为高度相关的动作，它也是判断行人意图的关键指标。本文采用OpenPose模型提取目标行人的18个骨骼关键点序列，记该输入矩阵为$\boldsymbol{X}\in \mathrm{R}^{T\times J\times C}$，其中$T$、$J$、$C$分别表示时间步长、关节点数量和关节点的维度特征。再使用得到的骨骼点的坐标信息作为输入数据，通过时空图卷积网络（spatio-temporal graph convolutional network，ST-GCN）来提取行人姿态的时空特征，通过一系列的ST-GCN层对输入数据进行特征提取，设第l层的卷积核数量为$\boldsymbol{K}^{(l)}$，每一层的输出特征图记为$\boldsymbol{H}^{(l)}=\mathrm{R}^{T\times J\times K^{(l)}}$。在ST-GCN中，空间卷积操作用于捕获骨骼点之间的空间关系，定义为：

$$\boldsymbol{H}^{(l+1)}=\sigma\left(\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}\boldsymbol{W}^{(l)}\right) \tag{12}$$

其中，$\hat{\boldsymbol{A}}=\boldsymbol{A}+\boldsymbol{I}$是邻接矩阵$\boldsymbol{A}$加上自环边的结果，$\boldsymbol{I}$是单位矩阵；$\hat{\boldsymbol{D}}$是$\hat{\boldsymbol{A}}$的度矩阵；$\sigma$是激活函数，$\boldsymbol{W}^{(l)}$是第$l$层的学习权重矩阵；$\boldsymbol{H}^{(l)}$是第$l$层的节点特征矩阵，$\hat{\boldsymbol{D}}^{-\frac{1}{2}}\hat{\boldsymbol{A}}\hat{\boldsymbol{D}}^{-\frac{1}{2}}\boldsymbol{H}^{(l)}$是骨骼点的归一化邻接矩阵，捕捉骨骼点间的空间依赖关系。时间卷积操作旨在捕捉同一骨骼点跨不同时间帧的动态变化，采用一维卷积定义为：

$$\boldsymbol{H}^{(l+1)}=\mathrm{Conv1D}\left(\boldsymbol{H}^{(l)}\right) \tag{13}$$

通过若干ST-GCN层，卷积核数量分别设置为16、32、96和64，每一层后接池化层以提取最终的时空特征。

2.3 场景特征模块

行人的行为决策通常受周围环境因素的影响[40]，如路面上的行人密度、交通流量以及道路布局等因素，都对行人是否选择过马路起着重要作用。为了充分建模行人与周围交通环境之间的时空语义关系，本文从车载摄像头图像中提取局部上下文和全局上下文两种场景特征。其中，局部上下文指以目标行人为中心裁剪得到的邻域图像区域，主要包含行人周围的局部环境信息，如相邻车辆、斑马线以及人行道边界等。该特征能够反映行人短距离交互环境，对于捕捉行人即时通行意图具有较强指示性。而全局上下文则对应整幅交通场景图像，用于表征更大范围内的环境语义信息，如车道分布、交通流密度、信号灯状态及整体道路布局等，有助于模型理解行人所处的宏观交通语境。

本文首先采用预训练的DeepLabV3模型对时刻τ的输入图像$\boldsymbol{X}_\tau$进行语义分割，从而得到对应的分割结果$\boldsymbol{Y}_\tau$，在这一过程中，输入图像中的每个像素都会被映射到一个语义类别标签，以表征其对应的目标类型。该过程可以被表示为：

$$\boldsymbol{Y}_\tau=\mathrm{DeepLabV3}\left(\boldsymbol{X}_\tau\right) \tag{14}$$

其次，为了提取目标行人与其周围环境之间的交互特征，对目标行人区域进行高斯模糊处理，削弱该区域的细节特征，避免其对整体场景语义的干扰，从而获得处理后的图像。随后，利用卷积神经网络对处理后的图像进行特征提取，并通过最大池化操作得到语义特征，将其作为场景特征，进而生成语义特征映射$\varnothing_\tau$。该过程可表示为：

$$\varnothing_\tau=\mathrm{MP}\left(\mathrm{CNN}\left(\mathrm{g}_\sigma\left(\boldsymbol{Y}_\tau\right)\right)\right) \tag{15}$$

其中，MP表示最大池化函数，g_σ表示进行高斯模糊处理。

最后，本文通过线性变换层将场景信息映射到特征空间，得到最终的场景特征表示，如式（16）所示：

$$\boldsymbol{E}_f=\boldsymbol{W}_f\cdot\boldsymbol{f}+\boldsymbol{b}_f \tag{16}$$

其中，$\boldsymbol{W}_f$为嵌入层的权值向量，$\boldsymbol{b}_f$为偏置向量，$\boldsymbol{E}_f$为编码后的场景特征向量。

2.4 多特征融合模块

不同模态特征（如手势、骨架、头部姿态等）在视频序列中可能存在帧级延迟或采样频率差异。为保证多模态输入的时序一致性，本文基于视频帧

时间戳对各模态特征进行同步。对缺失帧采用前向填充策略补齐，对高频模态进行下采样，以保证在每个时间步 t，多模态特征能够对应同一帧语义时刻，从而确保双向长短期记忆网络（bidirectional long short-term memory，Bi-LSTM）在时间维度上的对齐与可比性。

另外，在该模块中，Bi-LSTM 层分别从序列的前向与后向建模时间依赖关系，形成更加完整的时序特征表示。反向序列建模在模型离线训练阶段引入，用于提升时间上下文的建模能力；在推理阶段，模型利用当前及历史观测信息进行预测，保证在交通场景中的实时可用性。为了进一步突出序列中的关键信息，本文在 Bi-LSTM 输出之后引入了注意力机制。该注意力机制模块能够沿着时间维度对特征序列中的每一个元素赋予不同的权重，使模型能够聚焦于最相关的部分。对于第 n 个分支，注意力权重向量 $\boldsymbol{r}_n$ 的计算式如下：

$$\boldsymbol{r}_n = \mathrm{softmax}\left(\tanh\left(\boldsymbol{P}\boldsymbol{\alpha}_k + \boldsymbol{b}\right)\right) \tag{17}$$

再通过加权得到注意力输出向量 $\boldsymbol{z}_n$：

$$\boldsymbol{z}_n = \sum_{k=1}^{d} \boldsymbol{r}_n \boldsymbol{\alpha}_k \tag{18}$$

其中，$\boldsymbol{\alpha}_k$ 表示在第 k 个时间步的隐藏状态向量，$\boldsymbol{P}$ 和 $\boldsymbol{b}$ 是可训练参数。d 表示在时间维度上执行注意力权重 $\boldsymbol{r}_n$ 计算的长度，是对应于 Bi-LSTM 输出的隐状态维度。本文研究采用的 Bi-LSTM 层隐藏单元数设置为 64。

由自车编码模块、行人特征编码和场景特征编码模块生成的特征，对行人意图推理有不同的影响。因此，这里引入自适应融合模块对所有编码特征进行加权。所提融合的这些隐藏表示累加在一个向量表示“$\boldsymbol{F}$”中，计算式如下：

$$\boldsymbol{F} = \sum_{n=1}^{3} \omega_n \boldsymbol{z}_n \tag{19}$$

其中，$\boldsymbol{z}_n$ 表示融合了时间序列关键信息的注意力输出向量，而 ω_n 是具有 HeNormal 初始化的可训练权重。为实现权重的自适应学习，将 ω_n 设计为可训练参数，并通过模态注意力机制动态计算，计算式如下：

$$\omega_n = \frac{\exp\left(\boldsymbol{W}_m^{\mathrm{T}} \boldsymbol{f}_n\right)}{\sum_{k=1}^{3} \exp\left(\boldsymbol{W}_m^{\mathrm{T}} f_k\right)} \tag{20}$$

其中，$\boldsymbol{f}_n$ 是第 n 模态的特征表示，$\boldsymbol{W}_m$ 表示模态注意力层的可学习参数向量。该融合特征随后通过全连接层和 Softmax 分类器，输出“有过街意图”或“无过街意图”的预测。在训练过程中，参数 ω_n 与网络其他部分一起通过反向传播进行更新，使模型能够根据不同场景自动调整各模态的权重分布。

2.5　损失函数

行人过街意图预测本质上是一个二元分类问题。因此本文采用交叉熵损失函数进行优化。对于模型的预测概率 p_i（行人过街的置信度）及其对应的真实标签 y_i，损失函数被定义为所有样本损失的平均值，具体表达式定义如下：

$$L = -\frac{1}{n}\left(y_i \log\left(p_i\right) + \left(1 - y_i\right)\log\left(1 - p_i\right)\right) \tag{21}$$

其中，n 代表样本总数，该损失函数通过优化预测概率 p_i 来最小化真实标签与预测值之间的差异，使得模型提高分类精度。

3　实验设计

3.1　数据集和评价指标

本文使用在行人意图推理领域广泛使用的 JAAD[41] 数据集来验证模型。该数据集专门用于调查驾驶员和行人在过马路时的行为。它由 346 个短视频片段组成，每个片段持续 5~10 s，这些视频片段是由视频编辑从一台 30 fps 摄像机拍摄的 240 多个小时的驾驶镜头中提取出来的，其中每一帧都包含上下文信息，如人行横道和交通标志的出现、道路类型以及天气状况和一天中的时间。数据集也为行人提供边界框及描述行人行为的标签（如站立、慢速或快速移动、观看、扫视、停止），还有给车辆的行为标签（移动缓慢或快速，减速或加速）。

本文采用 PyTorch 在 Python3.6 环境下构建，并使用 NVIDIA GeForce RTX 4070 Super GPU 对所提模型进行全面的训练和评估。在训练过程中，本文使用 Adam 优化器调整超参数：对于 JAAD 数据集，模型训练了 60 个 epoch，学习率为 5×10^{-5}，保持 32 个批量的规模。本文使用了先前工作中常用的评估指标（包括 ACC、AUC、F1 分数、精度和召回率）来评估模型性能，其中 TP 为真阳性，TN 为真阴性，FP 为假阳性，FN 为假阴性。评估指标计算式为：

$$\mathrm{ACC} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{22}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{23}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{24}$$

$$\text{F1} = \frac{2\times\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{25}$$

AUC 是 ROC 曲线下的面积，是衡量模型性能的重要指标。一般来说，AUC 值越高，模型预测性能越好，其取值范围为 0.5~1。

3.2 定量分析

本文将所提模型与在 JAAD 数据集上表现优异、结构上具有相似性的先进方法进行比较，这些方法均以多模态数据为输入，利用 CNN 和 RNN 提取时空特征，并融合这些特征进行行人过街意图预测。

（1）MultiRNN[42]：一种多流 RNN 架构，其中每个独立的 RNN 分支分别处理不同类型的输入特征，通过并行编码保持特征特异性，再将各分支的隐藏状态进行拼接，并输入全连接层进行融合与预测。

（2）SF-GRU[43]：一种分层循环网络，用于逐步融合并编码多源异构信息，完成行人过街意图预测。共引入 5 类数据：目标外观、场景上下文、行人骨架位姿、空间坐标序列以及自车速度。此外，还设计了一种 SF-GRU 的改进结构，即在底层将全部模态一次性整合，再进行统一的空间建模。

（3）TrouSPI-Net[44]：一种轻量化多分支推理网络，该网络采用上下文无关的架构设计，该架构首先从骨骼关节位置构建的伪图像序列中提取时空特征，同时，模型独立处理相对关节距离和边界框位置等附加动态特征，再进行多特征融合，实现对行人行为意图的预测。

（4）PCPA[45]：该模型采用多分支架构，利用多个独立的 RNN 分支分别处理动态时序数据，同时采用 3D 卷积分支提取空间–上下文视觉特征，再使用模态关注层对从这些分支中提取的特征进行合并，以提高预测性能。

（5）FSTA[46]：一种新型神经网络架构，该架构融合了 RGB 图像、语义分割掩模以及自车速度。通过注意力机制及循环神经网络来实现对行人过街意图的预测。

（6）MMH-PAP[47]：一种混合预测架构，综合利用视觉与动态两类信息。视觉部分包括语义地图，通过卷积网络进行编码；动态部分涵盖行人运动轨迹，借助循环网络建模。两类特征进行融合后用于预测行人是否会过街。

（7）PPCI_{att}[48]：一种基于长短期记忆网络和注意力机制的轻量级模型，该模型使用非视觉特征包括人体姿态关键点、边界框坐标和自车速度作为输入进行行人意图推理。

本文模型与其他模型性能比较见表 1，最好的结果用粗体表示，第二好的结果用下划线表示，这表明本文模型在 JAAD 数据集上优于其他模型。与其他模型不同的是，本文提出的模型在输入源中引入了行人手势和头部姿态，并着重关注行人与车辆的行进方向相同与否等因素，有效地学习了行人的运动特征以及人车交互信号。实验结果表明，与各指标表现较优的模型相比，ACC 提高了 4%，AUC 提高了 1%，Precision 分数提高了 11%。尽管 PCPA 模型在 Recall 指标上更高，但其更倾向于预测为正例，虽然捕捉到更多真实正例，却同时带来了更多误报，导致 Precision 相对较低。

表 1 本文模型与其他模型性能比较

模型	ACC	AUC	*F*1	Precision	Recall
MultiRNN	0.79	0.79	0.58	0.45	0.79
SF-GRU	0.76	0.77	0.53	0.40	0.79
TrouSPI-Net	0.82	0.77	0.58	0.49	0.70
PCPA	0.76	0.79	0.55	0.41	**0.83**
FSTA	0.83	<u>0.82</u>	0.63	0.51	0.81
MMH-PAP	<u>0.84</u>	0.80	0.62	<u>0.54</u>	0.72
PPCI_{att}	0.81	0.78	**0.75**	—	—
本文模型	**0.88**	**0.84**	<u>0.72</u>	**0.65**	<u>0.80</u>

本文也对 ARPCI 的内存占用和推理时间与几个最先进的模型进行了对比，本文模型与其他模型的大小和推理时间比较见表 2，最好的结果用粗体表示，第二好的结果用下划线表示。

表 2 本文模型与其他模型的大小和推理时间比较

模型	大小/MB	推理时间/ms
FSTA	374.2	70.83
PCPA	118.8	38.60
TEP	12.8	<u>2.85</u>
PedGraph+	**0.28**	5.47
本文模型	<u>3.89</u>	**1.83**

3.3 可视化结果

JAAD 数据集中行人过街意图预测结果如图 8 所示，实验结果展示了本文方法在 JAAD 数据集中

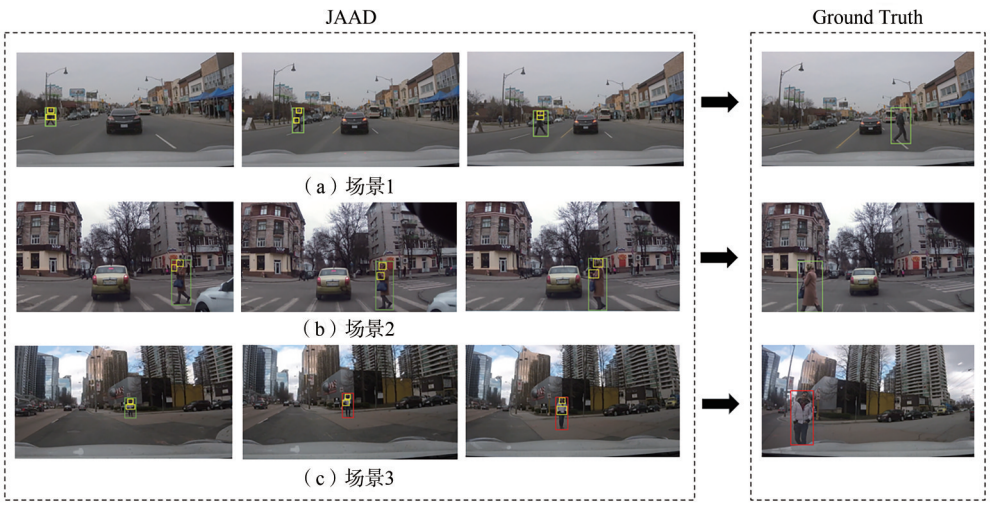

图8　JAAD数据集中行人过街意图预测结果

行人与车辆驾驶员有交互信号的典型场景下的预测性能。可视化结果中，红色边界框表示标识出行人不过街的预测，绿色边界框则对应行人过街的预测。在图8（a）和图8（b）场景中，行人的头部方向关注自车车辆的行驶方向，并通过手势对驾驶员表达明显的“我想先行”信号，ARPCI推断出行人具有强烈的过街意图。在图8（c）场景中，行人一开始表现出过街行为（如头部看向迎面车辆方向、跨步姿态），随后通过明确的向车辆挥手示意让车辆先行，最终表现为不过街行为。ARPCI通过动态融合细粒度姿态特征（如头部视线方向、行人骨架、手势特征）和实时车速数据，在开始时推断出行人穿越的意图，随后及时推断行人不过街意图的变化即不穿越意图。分析表明，ARPCI在不同交通场景中均表现出稳定和准确的行人过街意图推理能力。

3.4　消融实验

输入特征消融实验见表3，验证了本文提出的框架中每个组件的贡献。去除头部姿态（ID1）导致JAAD上的ACC下降5%（0.88→0.83），这表明头部姿态是理解行人行为和意图的关键因素之一，它能够提供有关行人朝向和注意力方向的重要线索。同样，忽略骨架姿态以及手势动作（ID2、ID3）不仅会使准确率下降，还会使召回率降低4%（0.80→0.76）和8%（0.80→0.72），F1分数降低3%（0.72→0.69）和4%（0.72→0.68），这表明行人骨架姿态特征和手势动作特征提供了丰富的动态信息，能够提高模型对行人行为识别的准确性。虽然场景特征（ID4）对总体准确率的贡献较小（0.88→0.86），但它们的移除影响了模型在JAAD数据集上表现的召回率（0.80→0.71），这可能是因为场景特征在提供行人行为发生的环境上下文方面发挥了关键作用。速度和方向（ID5、ID6）仍然是必不可少的，去掉该因素使ACC降低了5%和4%，其他指标均有不同程度的下降。这些信息在行人行为预测中有关键作用，它们为模型提供了动态环境的详细信息，有助于准确预测行人的运动趋势和潜在行为。完整模型（ID7）在所有指标上实现了最佳平衡，证明了综合利用所有特征（头部姿态、骨架姿态、手势动作、场景特征、速度、方向）能够提高模型预测准确性和鲁棒性，从而在复杂的实际应用场景中提供稳定可靠的性能。

为验证改进的YOLOv8n中各改进模块对模型性能的影响，本文针对VanillaNet与CBAM两个模块分别进行消融实验，改进的YOLOv8n消融实验对比见表4，二者结合能够在保持高实时性的同时获得较优的检测精度，为复杂交通环境下的手势识别任务提供了更均衡的性能表现。

4　结束语

本文提出了一种多特征融合的行人过街意图推理框架。该框架综合利用行人特征、自车特征以及场景语义信息，能够在复杂交通环境下对行人未来行为进行更加准确的预测。通过在JAAD数据集上的实验验证，结果表明所提模型在预测行人过街意图方面优于现有多种SOTA模型，尤其在应对多样化交通场景和行人发出交互信号时，该框架能够较好地捕捉行人与车辆、道路元素之间的交互关系，

表3 输入特征消融实验

ID	头部姿态	骨架姿态	手势动作	场景特征	速度	方向	ACC	AUC	F1	Precision	Recall
ID 1	×	√	√	√	√	√	0.83	0.79	0.67	0.62	0.73
ID 2	√	×	√	√	√	√	0.80	0.80	0.69	0.63	0.76
ID 3	√	√	×	√	√	√	0.82	0.78	0.68	0.64	0.72
ID 4	√	√	√	×	√	√	0.86	0.80	0.67	0.63	0.71
ID 5	√	√	√	√	×	√	0.83	0.73	0.65	0.59	0.74
ID 6	√	√	√	√	√	×	0.84	0.81	0.68	0.62	0.78
ID 7	√	√	√	√	√	√	**0.88**	**0.84**	**0.72**	**0.65**	**0.80**

表4 改进的YOLOv8n消融实验对比

模型	VanillaNet	CBAM	Parameters	GFLOPs	mAP	FPS
YOLOv8n	×	×	3 007 208	8.1	96.8%	67.5
YOLOv8n	√	×	1 994 187	1.5	95.5%	91.7
YOLOv8n	×	√	3 040 892	8.35	97.1%	66.5
YOLOv8n	√	√	2 028 263	1.6	95.8%	90.3

为自动驾驶系统提供更可靠的决策依据。总体而言，本文为行人过街意图推理提供了一种新的方法，未来工作中，可在保持整体框架不变的前提下，进一步围绕模型中的组件开展更细粒度的对比实验和机制分析，以更深入地揭示多模态特征之间的作用机理并完善模型结构。此外，还可以进一步探索如何将该方法扩展至更复杂的真实交通场景，并结合实时计算优化，实现对大规模复杂交通环境的快速响应，从而降低交通事故发生率，为提高道路交通安全提供更有力的技术支撑。

参考文献：

[1] Hu J W, Flannagan C, Ganesan S, et al. Understanding the new trends in pedestrian injury distribution and mechanism through data linkage and modeling[J]. Accident; Analysis and Prevention, 2023, 188: 107095.

[2] Wang X, Tang K, Dai X, et al. S4TP: social-suitable and safety-sensitive trajectory planning for autonomous vehicles[EB]. 2024.

[3] Liu J M, Lin H, Wang X D, et al. Reliable trajectory prediction in scene fusion based on spatio-temporal structure causal model[J]. Information Fusion, 2024, 107: 102309.

[4] Wang X D, Liu J M, Lin H, et al. A multi-modal spatial-temporal model for accurate motion forecasting with visual fusion[J]. Information Fusion, 2024, 102: 102046.

[5] Landry F G, Akhloufi M A. Predicting pedestrian crossing intention in autonomous vehicles: a review[J]. Neurocomputing, 2025, 618: 129105.

[6] Du Q C, Xu L L, Wu Q, et al. SafeCrossNet: multi-modal fusion with social-aware for pedestrian crossing intention prediction[J]. Information Fusion, 2026, 126: 103609.

[7] Li Z R, Gong C, Lin Y L, et al. Continual driver behaviour learning for connected vehicles and intelligent transportation systems: framework, survey and challenges[J]. Green Energy and Intelligent Transportation, 2023, 2(4): 100103.

[8] Du Q C, Wu Q, Li L X, et al. Review and perspectives on pedestrian trajectory prediction for safe transportation[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, PP(99): 1-27.

[9] Wang X, Liu J G, Mei T, et al. CoSeg: cognitively inspired unsupervised generic event segmentation[J]. IEEE Transactions on Neural Networks and Learning Systems, 2024, 35(9): 12507-12517.

[10] Wang X, Wu Z Z, Jiang B, et al. HARDVS: revisiting human activity recognition with dynamic vision sensors[C]//Proceedings of the AAAI Conference on Artificial Intelligence. Piscataway: AAAI Press, 2024, 38(6): 5615-5623.

[11] Azarmi M, Rezaei M, Wang H. PIP-net: pedestrian intention prediction in the wild[J]. IEEE Transactions on Intelligent Transportation Systems, 2025, 26(7): 9824-9837.

[12] Wang X, Tang K, Dai X Y, et al. Safety-balanced driving-style aware trajectory planning in intersection scenarios with uncertain environment[J]. IEEE Transactions on Intelligent Vehicles, 2023, 8(4): 2888-2898.

[13] Du Q C, Wang X, Yin S G, et al. Social force embedded mixed graph convolutional network for multi-class trajectory prediction[J]. IEEE Transactions on Intelligent Vehicles, 2024, 9(9): 5571-5580.

[14] Helbing D, Molnár P. Social force model for pedestrian dynamics[J]. Physical Review E, Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 1995, 51(5): 4282-4286.

[15] Fang Z J, Vázquez D, López A M. On-board detection of pedestrian intentions[J]. Sensors, 2017, 17(10): 2193.

[16] Zhang W, Zhu F H, Chen Y Y, et al. Differential time-variant traffic flow prediction based on deep learning[C]//Proceedings of the 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). Piscataway: IEEE Press, 2020: 1-6.

[17] Wang J G, Wang X, Shen T Y, et al. A long-tail regularization method for traffic sign recognition based on parallel vision[J]. IEEE Journal of Radio Frequency Identification, 2022, 6: 957-961.

[18] Li F, Fan S W, Chen P Z, et al. Pedestrian motion state estimation from 2D pose[C]//Proceedings of the 2020 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2020: 1682-1687.

[19] Varytimidis D, Alonso-Fernandez F, Duran B, et al. Action and intention recognition of pedestrians in urban traffic[C]//Proceedings of the 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS). Piscataway: IEEE Press, 2018: 676-682.

[20] Perdana M I, Anggraeni W, Sidharta H A, et al. Early warning pedestrian crossing intention from its head gesture using head pose estimation[C]//Proceedings of the 2021 International Seminar on Intelligent Technology and Its Applications (ISITIA). Piscataway: IEEE Press, 2021: 402-407.

[21] 杜泉成, 王晓, 李灵犀, 等. 行人轨迹预测方法关键问题研究: 现状及展望[J]. 智能科学与技术学报, 2023, 5(2): 143-162.
Du Q C, Wang X, Li L X, et al. Key problems and progress of pedestrian trajectory prediction methods: the state of the art and prospects[J].

Chinese Journal of Intelligent Science and Technology, 2023, 5(2): 143-162.

[22] 李琳辉, 周彬, 任威威, 等. 行人轨迹预测方法综述[J]. 智能科学与技术学报, 2021, 3(4): 399-411.
Li L H, Zhou B, Ren W W, et al. Review of pedestrian trajectory prediction methods[J]. Chinese Journal of Intelligent Science and Technology, 2021, 3(4): 399-411.

[23] Guo J R, Ding Y T, Tian A S. Multimodal feature fusion for pedestrian crossing intention prediction based on hybrid attention mechanism[C]//Proceedings of the 2024 6th International Conference on Internet of Things, Automation and Artificial Intelligence (IoTAAI). Piscataway: IEEE Press, 2024: 70-74.

[24] Dey D, Terken J. Pedestrian interaction with vehicles: roles of explicit and implicit communication[C]//Proceedings of the 9th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. New York: ACM Press, 2017: 109-113.

[25] Gupta A, Johnson J, Li F F, et al. Social GAN: socially acceptable trajectories with generative adversarial networks[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 2255-2264.

[26] Lin Y W, Hu C, Zhao B X, et al. Anchor-based multi-modal transformer network for pedestrian trajectory and intention prediction[C]//Proceedings of the 2023 7th CAA International Conference on Vehicular Control and Intelligence (CVCI). Piscataway: IEEE Press, 2023: 1-6.

[27] Wang Y, Wan W X, Zhang H Q, et al. Pedestrian trajectory intention prediction in autonomous driving scenarios based on spatio-temporal attention mechanism[C]//Proceedings of the 2024 4th International Conference on Electronic Information Engineering and Computer Communication (EIECC). Piscataway: IEEE Press, 2024: 1519-1522.

[28] Mohamed A, Qian K, Elhoseiny M, et al. Social-STGCNN: a social spatio-temporal graph convolutional neural network for human trajectory prediction[C]//Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2020: 14412-14420.

[29] Tang W X, Liu K, Shakeel M S, et al. DDAD: detachable crowd density estimation assisted pedestrian detection[J]. IEEE Transactions on Intelligent Transportation Systems, 2023, 24(2): 1867-1878.

[30] 胡远志, 蒋涛, 刘西, 等. 基于双流自适应图卷积神经网络的行人过街意图识别[J]. 汽车安全与节能学报, 2022, 13(2): 325-332.
Hu Y Z, Jiang T, Liu X, et al. Pedestrian-crossing intention-recognition based on dual-stream adaptive graph-convolutional neural-network[J]. Journal of Automotive Safety and Energy, 2022, 13(2): 325-332.

[31] Chaabane M, Trabelsi A, Blanchard N, et al. Looking ahead: anticipating pedestrians crossing with future frames prediction[C]//Proceedings of the 2020 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2020: 2286-2295.

[32] Xie C, Lin C Y, Zheng X Y, et al. GTransPDM: a graph-embedded transformer with positional decoupling for pedestrian crossing intention prediction[J]. IEEE Signal Processing Letters, 2025, 32: 2109-2113.

[33] Cadena P R G, Qian Y Q, Wang C X, et al. Pedestrian graph: a fast pedestrian crossing prediction model based on graph convolutional networks[J]. IEEE Transactions on Intelligent Transportation Systems, 2022, 23(11): 21050-21061.

[34] Liu B B, Adeli E, Cao Z J, et al. Spatiotemporal relationship reasoning for pedestrian intent prediction[J]. IEEE Robotics and Automation Letters, 2020, 5(2): 3485-3492.

[35] Piccoli F, Balakrishnan R, Perez M J, et al. FuSSI-Net: fusion of spatio-temporal skeletons for intention prediction network[EB]. 2020.

[36] Yang T Y, Chen Y T, Lin Y, et al. FSA-net: learning fine-grained structure aggregation for head pose estimation from a single image[C]//Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE Press, 2019: 1087-1096.

[37] Zhang J Y, Xu Y N, Chen X. Multimodal-based pedestrian crossing intention recognition method[C]//Proceedings of the 2023 China Automation Congress (CAC). Piscataway: IEEE Press, 2023: 3508-3513.

[38] Chen H, Wang Y, Guo J, et al. Vanillanet: the power of minimalism in deep learning[J]. Advances in Neural Information Processing Systems, 2023, 36: 7050-7064.

[39] Woo S, Park J, Lee J Y, et al. CBAM: convolutional block attention module[C]//Proceedings of the European Conference on Computer Vision (ECCV) 2018. Cham: Springer, 2018: 3-19.

[40] Wang X, Huang J, Tian Y L, et al. Parallel driving with big models and foundation intelligence in cyber-physical-social spaces[J]. Research, 2024, 7: 0349.

[41] Kotseruba I, Rasouli A, Tsotsos J K. Joint attention in autonomous driving (JAAD)[J]. 2016.

[42] Bhattacharyya A, Fritz M, Schiele B. Long-term on-board prediction of people in traffic scenes under uncertainty[C]//Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Piscataway: IEEE Press, 2018: 4194-4202.

[43] Rasouli A, Kotseruba I, Tsotsos J K. Pedestrian action anticipation using contextual feature fusion in stacked rnns[EB]. 2020.

[44] Gesnouin J, Pechberti S, Stanciulcscu B, et al. TrouSPI-Net: spatio-temporal attention on parallel atrous convolutions and U-GRUs for skeletal pedestrian crossing prediction[C]//Proceedings of the 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021). New York: ACM Press, 2021: 1-7.

[45] Kotseruba I, Rasouli A, Tsotsos J K. Benchmark for evaluating pedestrian action prediction[C]//Proceedings of the 2021 IEEE Winter Conference on Applications of Computer Vision (WACV). Piscataway: IEEE Press, 2021: 1257-1267.

[46] Yang D F, Zhang H L, 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.

[47] Rasouli A, Yau T, Rohani M, et al. Multi-modal hybrid architecture for pedestrian action prediction[C]//Proceedings of the 2022 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2022: 91-97.

[48] Alofi A, Greer R, Gopalkrishnan A, et al. Pedestrian safety by intent prediction: a lightweight LSTM-attention architecture and experimental evaluations with real-world datasets[C]//Proceedings of the 2024 IEEE Intelligent Vehicles Symposium (IV). Piscataway: IEEE Press, 2024: 77-84.

[作者简介]

尹守国（2002– ），男，安徽大学人工智能学院硕士生，主要研究方向为意图推理、轨迹预测、智能交通系统。

杜泉成（1994– ），男，北京科技大学计算机与通信工程学院博士生，主要研究方向为轨迹预测和车辆规划决策。

王晓（1988– ），女，安徽大学人工智能学院教授，主要研究方向为社会计算、群体行为建模、无人自主系统及其平行测试。

李灵犀（1977– ），男，博士，美国普渡大学电子与计算机工程系教授，主要研究方向为复杂系统的建模、分析、控制与优化、智能交通系统、离散事件动态系统。

孙长银（1975– ），男，安徽大学校长，主要研究方向为智能控制与优化、强化学习、神经网络和数据驱动控制。