Computer Science and Application 计算机科学与应用, 2025, 15(11), 220-233
Published Online November 2025 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2025.1511299

Hans汉斯

融合多核图卷积与注意力池化的行人过街意图预测

周兴鹏

西华大学汽车测控与安全四川省重点实验室，四川 成都

收稿日期：2025年10月17日；录用日期：2025年11月17日；发布日期：2025年11月25日

摘　要

自动驾驶场景中，行人过街意图的准确预测是保障行驶安全和提升交互效率的重要问题。针对行人动态行为复杂且与环境高度交互的特点，提出了一种基于多核图卷积与注意力池化的行人过街意图预测模型(Multi-kernel Attention Graph Network for Pedestrian Crossing Intention, MAGNet-PCI)。该模型采用并行的多分支时空编码器，利用多核时空图卷积模块(Multi-kernel Spatio-Temporal Graph Convolution, MKGC-ST)从多角度提取行人骨架序列中的动态特征。为减轻特征展平过程中的信息丢失问题，引入注意力池化机制(Attention Pooling Transformer, APT)，通过节点选择与多头注意力聚合生成结构感知的图级表示，用于意图分类。在公开的JAAD和PIE数据集上的实验结果表明，该方法在准确率上分别较PedGNN提升3%和10%。消融实验进一步验证了并行多分支结构、多核卷积机制及注意力池化模块的有效性。

关键词

行人过街意图预测，图神经网络，时空图卷积，注意力池化，多核机制

Pedestrian Crossing Intention Prediction with Multi-Kernel Graph Convolution and Attention Pooling

Xingpeng Zhou

Vehicle Measurement, Control and Safety Key Laboratory of Sichuan Province, Xihua University, Chengdu Sichuan

Received: October 17, 2025; accepted: November 17, 2025; published: November 25, 2025

文章引用: 周兴鹏. 融合多核图卷积与注意力池化的行人过街意图预测[J]. 计算机科学与应用, 2025, 15(11): 220-233.
DOI: 10.12677/csa.2025.1511299

Abstract

Accurate prediction of pedestrian crossing intention is critical for driving safety and interaction efficiency in autonomous driving. To address the complexity of pedestrian dynamics and strong interactions with the environment, a Multi-kernel Attention Graph Network for Pedestrian Crossing Intention (MAGNet-PCI) is proposed. The model employs a parallel multi-branch spatio-temporal encoder, where the Multi-kernel Spatio-Temporal Graph Convolution (MKGC-ST) module extracts motion features from pedestrian skeleton sequences from multiple perspectives. To mitigate information loss during feature flattening, an Attention Pooling Transformer (APT) mechanism is introduced. It selects key joints through graph convolution and aggregates global context with multi-head attention, generating structure-aware graph-level representations for intention classification. Experiments on the JAAD and PIE datasets show that the proposed method achieves accuracy improvements of 3% and 10% over PedGNN, respectively. Ablation studies further verify the effectiveness of the parallel multi-branch structure, multi-kernel convolution module, and attention pooling mechanism.

Keywords

Pedestrian Crossing Intention Prediction, Graph Neural Network, Spatio-Temporal Graph Convolution, Attention Pooling, Multi-Kernel Mechanism

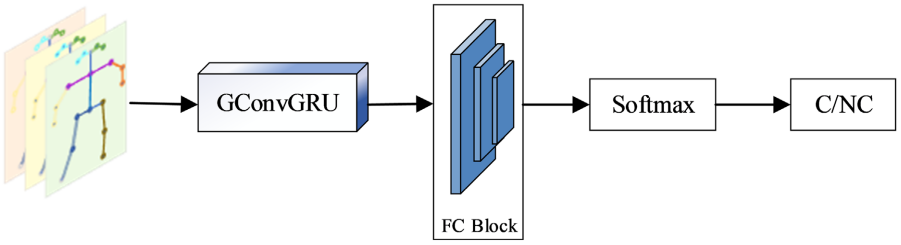
1. 引言

自动驾驶系统在城市道路中安全运行，需要预测行人是否存在过街意图，这对车辆的预判与决策至关重要[1]。研究表明，行人预测不准确是自动驾驶事故的重要诱因之一[2]-[4]。然而，现有模型在复杂环境下的泛化能力有限，常常在新场景或恶劣条件下性能下降[5]。

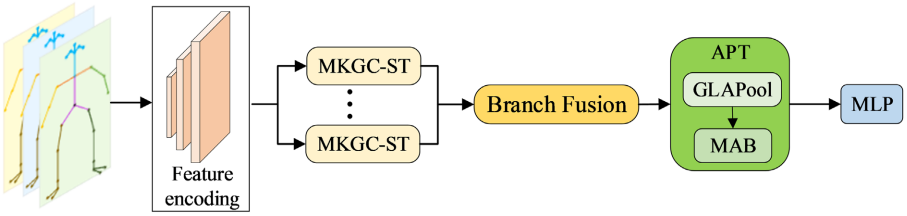
现有研究大体分为多模态和单模态两类[6]。多模态方法融合外观、语义和车辆信息，精度较高但计算开销大[7] [8]；骨骼数据因其稀疏与结构化特性，单模态方法更具轻量化和实时性优势[9]。然而，现有骨架方法多依赖边界框轨迹[4]或单路径编码，难以刻画细粒度时空动态和多尺度拓扑关系，在复杂场景下表现有限。Ahmed 等[10]利用视频骨架关键点序列训练 LSTM 分类器，能捕捉时序特征，但未能建模关节间结构依赖。近年来，图神经网络(GNN)逐渐成为主流。如 Shi 等[11]的 ST-GCN 和 Yan 等[12]的 STGCN 实现了时空统一建模，但依赖范围有限。在意图预测方面，Riaz 等[13]提出的 PedGNN 结合 GNN 与 GRU，提升了效率，但缺乏多尺度建模与关键节点筛选，易受噪声干扰。因此，实现多尺度时空建模与自适应聚合是提升性能的关键挑战。

为此，本文提出名为 MAGNet-PCI (Multi-kernel Attention Graph Network for Pedestrian Crossing Intention)的新型预测框架。其核心在于设计了并行的多分支多核时空图卷积编码器(Multi-kernel Spatio-Temporal Graph Convolution, MKGC-ST)。与传统单尺度模型，如图 1(a)所示不同，本文的编码器通过并行设置多尺度卷积核，分解并捕捉不同层级的动态特征。如图 1(b)所示，这种并行设计使模型既能捕捉手势、头部转动等局部细微动作，又能把握身体前倾、跨步等全局姿态，从而获得对行人意图更全面地理解。进一步引入注意力池化 Transformer 模块(Attention Pooling Transformer, APT)。该模块能够自适应筛选对

决策最关键的核心关节并进行高效的信息聚合，不仅有效滤除了冗余噪声，也使模型在面对遮挡等复杂情况时更具泛化能力。该方法在提升精度的同时保持轻量化设计，满足自动驾驶对实时性的需求。



(a) 传统的单尺度基线模型 PedGNN



(b) 本文提出的 MAGNet-PCI 框架

Figure 1. Comparison of single-scale baseline with MAGNet-PCI
图 1. 单尺度基线与 MAGNet-PCI 的对比

2. 相关研究进展

2.1. 行人意图预测

行人过街意图预测是智能驾驶感知系统的关键任务。现有研究主要分为三类：其一是基于轨迹的方法，早期通过卡尔曼滤波或隐马尔可夫模型建模位置序列[7]。后续引入 RNN 及其变体，如 SocialLSTM 利用“社交池化层”提升了拥挤场景下的预测精度[8]。但此类方法仅依赖(x, y)坐标，忽视了姿态与朝向等细粒度线索。其二是基于多模态视觉上下文的方法，融合行人外观、场景图像及车辆状态等信息[14] [15]。但这类方法通常依赖庞大而复杂的模型，导致计算开销大、推理延迟高，不利于车载实时部署[16]。其三是基于骨架图动态的方法，利用头部朝向、肢体动作等关键点特征进行预测[17] [18]。早期 LSTM 方法已取得一定效果[19]。而近年来时空图神经网络(STGNN)进一步推动了该方向的发展。例如 PedGNN 结合 GNN 与 GRU 直接建模时空骨架图，仅依赖骨架特征便在轻量化与实时性方面展现出优势。基于此，本文旨在通过设计更高效的图网络结构，深入挖掘骨架序列所蕴含的意图信息。

2.2. 图神经网络架构

图神经网络(GNN)为图结构数据提供了统一框架。早期工作从谱域展开：SCNN [20]将卷积由欧氏域映射到图拉普拉斯谱空间，ChebNet [21]以切比雪夫多项式实现局部化高效滤波；随后 GCN [22]以一阶近似在保证效果的同时显著降低复杂度并稳定训练。为提升归纳与自适应聚合能力，GraphSAGE [23]采用“采样–聚合”，GAT [24]以可学习权重调节邻域贡献。在此基础上，面向图级表示与下游任务的变体不断出现：VG-GCN [25]引入分层粗化与变分卷积，GMKEA [26]在再生核希尔伯特空间中实现多核注意力加权，MuchGNN [27]通过多通道卷积与多视图池化增强表达，CGAT [28]以通道感知注意力融入隐式语义。但是，这些方法大多局限于静态图场景，缺乏对时间维度的刻画，对于骨架序列等随时间演化的动态数据，标准 GNN 在建模能力上仍显不足，这也推动了时空图神经网络的兴起。

2.3. 时空图神经网络

时空图神经网络(STGNN)专为图结构的动态序列数据设计。其基本思路是在每个时间步执行图卷积以汇聚空间邻域，同时结合循环单元或注意力机制建模时间演化，从而实现时空一体化表示[12] [22]。典型方法如 GConvGRU，将图卷积嵌入 GRU 门控单元，在捕捉空间依赖的同时保留时间记忆，尤其适用于“拓扑固定、节点特征随时间变化”的骨架数据[29]。围绕 C/NC 判别的骨架方法也沿此范式演进：PedGraph 在骨架拓扑上执行图卷积并配合时间建模[30]，PedGraph+进一步融合自车速度与局部外观信息，取得更优表现[31]。吕超等[32]探索了基于图表示的行人意图识别框架，验证了图结构在建模行人骨架动态方面的有效性。尽管取得了进展，主流 STGNN 仍存在两方面不足：一是空间建模多依赖单尺度卷积核，难以兼顾局部细节与全局姿态；二是图级表示缺乏有效的读出机制，常见的全局平均或展平操作易造成信息损失和结构弱化。如何突破这两方面的限制，正是本文拟重点解决的问题。

3. MAGNet-PCI 模型

3.1. 模型总体结构

本章阐述本文为行人意图预测任务提出的新型图神经网络架构——MAGNet-PCI。该模型通过并行的、多分支的时空特征提取框架，并结合注意力池化机制，旨在有效捕捉复杂的行人动态并进行意图分类。图 2 为行人 26 骨骼节点的拓扑结构。

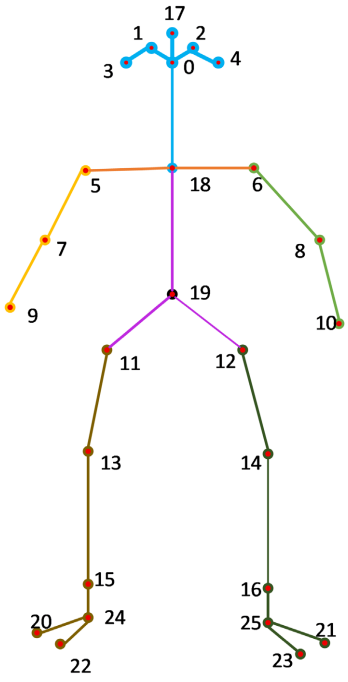


Figure 2. Pedestrian 26 bone node chart
图 2. 行人 26 骨骼节点图

如图 3 所示，MAGNet-PCI 由四个核心模块组成：(1) 节点特征编码器；(2) 并行多核时空图卷积模块；(3) 自适应融合与注意力池化模块；(4) 意图分类器。其训练流程包括：输入骨架序列的节点特征经编码器映射至高维嵌入；编码特征并行送入 B 个 MKGC-ST 分支学习时空特征；自适应融合模块加权整合各分支输出，并通过 APT 筛选关键节点生成图级嵌入；最后，意图分类器预测结果。

3.2. 并行多核时空图卷积

传统的单路径图卷积网络由于仅使用单一核函数，往往难以全面捕获行人动作中潜在的多样化运动模式，从而造成特征表达的局限性[27]。针对这一不足，本文提出一种并行多核时空图卷积(MKGC-ST)结构，以更有效地提取行人骨架图序列中的空间和时间特征，本文具体实现采用双分支结构($B = 2$)。

对于给定的行人骨架序列，将其表示为一段长度为T的图结构序列$\left\{G^{(t)}=\left(V,X^{(t)},A\right)\right\}_{t=1}^{T}$。其中，节点集合$V$表示人体骨架上的关节点(如头部、肩膀、肘部等)，节点总数为N，第t帧的节点特征矩阵记为$X^{(t)}\in\mathbb{R}^{N\times d_0}$，其关节点$i$的三维特征向量为$\left(x_i,y_i,c_i\right)$。其中$\left(x_i,y_i\right)$为由骨架模型提供的二维坐标，$c_i$为置信度。根据骨架序列建模的推荐方案[25]，每帧中的坐标均会被归一化至[0, 1]区间。以增强在不同距离条件下的鲁棒性。由此，MAGNet-PCI 的输入维度为$\left(T,N,d_0\right)$，其中d_0则代表每个关节的信息量，即$\left(x_i,y_i,c_i\right)$。邻接矩阵$A\in\mathbb{R}^{N\times N}$，反映骨架节点之间的拓扑结构，若两个节点$i$和$j$之间存在骨骼连接，则$A_{ij}=1$，否则为 0。该关系在序列中保持不变。为获取高维嵌入特征，本文将原始的节点特征序列$\left\{X^{(t)}\right\}_{t=1}^{T}$通过共享的节点特征编码器，使用线性层实现映射到高维特征空间，记为：

$$X_{emb}^{(t)}=\mathrm{Linear}\left(X^{(t)}\right)\in\mathbb{R}^{N\times d},t=1,\cdots,T \tag{1}$$

其中，d表示编码后的特征维度。随后，编码后的序列被同时输入到B个并行的 MKGC-ST 分支，每个分支独立学习互补的时空特征表示。

如图 3 左上角所示，在每个 MKGC-ST 模块内部设置M个并行且参数独立的图卷积门控循环单元，以增强对多样动态模式的捕捉。第m个图卷积门控循环核(GConvGRU$_m$)，其在时间步t的状态更新为：

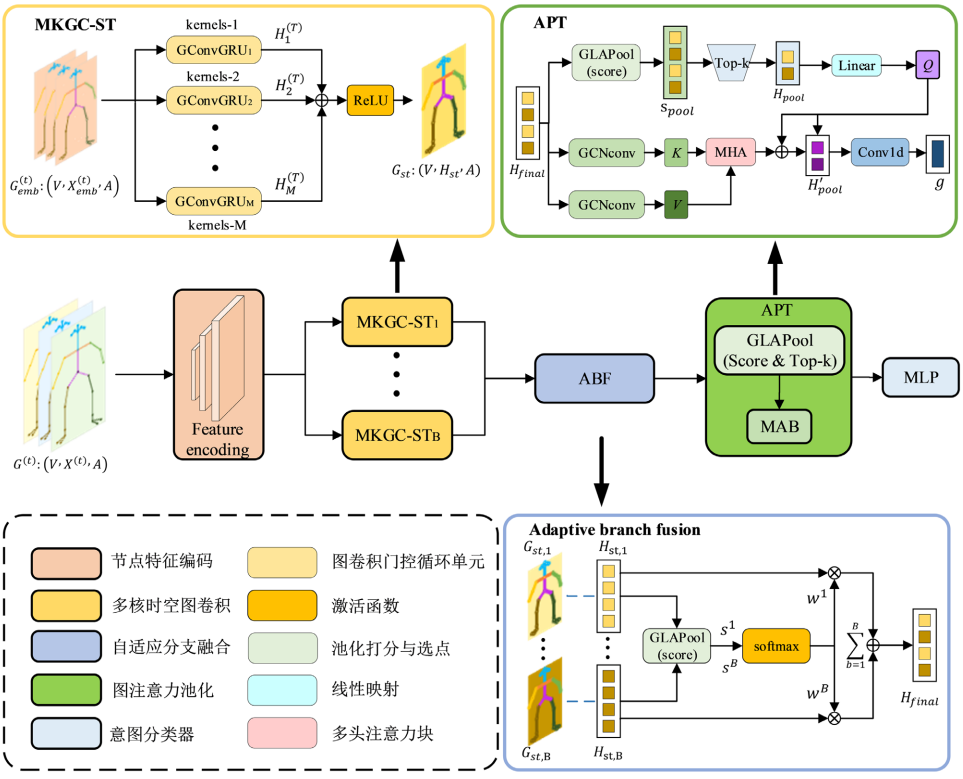


Figure 3. Overall architecture of MAGNet-PCI
图 3. MAGNet-PCI 模型整体架构

其中，$H_m^{(t)}$ 表示第 m 个核在时间 t 的隐藏状态，$X_{emb}^{(t)} \in \mathbb{R}^{N\times d}$ 为第 t 帧的节点嵌入特征，E 为节点邻接关系的边集合。GConvGRUk(·)通过图卷积实现空间邻域聚合，同时具备时序记忆能力。

在整个序列处理完毕后，所有 M 个核在最后一个时间步 T 的隐藏状态输出被求和聚合，并经过 ReLU 激活，形成一个融合多种动态模式的节点特征表示 H_{st}：

$$H_{st} = \mathrm{ReLU}\left(\sum_{m=1}^{M} H_m^{(T)}\right) \tag{3}$$

由于各 GConvGRU 核参数不共享，其输出 $H_1^{(T)}, H_2^{(T)}, \cdots, H_M^{(T)}$ 包含不同的语义信息。经求和与激活函数融合生成的节点特征矩阵 H_{st}，能包含更丰富的信息。在本文的 MAGNet-PCI 模型中，B 个并行的 MKGC-ST 模块分别输出一组特征图记为 $\{H_{st,1}, H_{st,2}, \cdots, H_{st,B}\}$。这种多分支设计使得模型能够在一个层次上并行地学习 B 组互补的高级时空特征，为后续的决策提供更强大的特征基础。

3.3. 自适应融合与注意力池化

3.3.1. 自适应特征融合

在从 B 个并行分支获得多组高级时空特征图 $H_{st,B}$ 后，如何有效融合并解码为紧凑的图级表示，是影响决策质量的关键。现有方法多采用直接求和或拼接，再经全局池化生成图级向量[33]-[36]。但这种策略存在两方面不足：其一，简单的融合操作同等对待所有分支，无法根据输入数据的特性动态地调整各分支特征的贡献度；其二，全局池化会平均处理所有节点，导致关键信息易被冗余或噪声节点掩盖，形成信息瓶颈。已有研究指出，更合理的特征融合与池化机制是提升图表示学习性能的关键[37] [38]。

针对上述问题，本文提出一种由自适应分支融合(Adaptive Branch Fusion)与注意力池化 Transformer (APT)构成的级联解码结构，如图 3 右下角所示：前者对多分支特征进行智能加权融合，后者进一步筛选并聚合关键信息，最终生成判别性更强的图级表示。

该模块的输入 B 个并行分支输出的节点特征为 $H_{st,b} \in \mathbb{R}^{N\times d}$。首先通过图卷积评分模块(GLAPool score)为每个节点特征赋予重要性分数。给定节点特征 $H_{st,b}$ 和图的边集 E (邻接矩阵 A 对应的边索引集合)，为每个节点计算一维评分用于后续选点与池化。GLAPool 中的节点 i 重要性分数计算公式为：

$$s^i = \alpha\left(H_{st,b}W_{a1}\right)_i + (1-\alpha)\sum_{j\in\mathcal{N}(i)}\left(H_{st,b}W_{a2}\right)_j \tag{4}$$

式中，$\alpha \in [0,1]$ 为可调的权衡参数，用于控制“自身通道”和“邻域通道”的占比，$W_{a1}, W_{a2} \in \mathbb{R}^{N\times 1}$ 为可学习参数矩阵，j 表示与节点 i 相邻的索引。求和集合 $\mathcal{N}(i)$ 为节点 i 的邻域节点集合。由图的边集(或邻接矩阵 A)给出：若 $(i,j)\in E$ (或 $A_{ij}=1$)，则 $j\in\mathcal{N}(i)$。因此，$\sum_{j\in\mathcal{N}(i)}(\cdot)$ 是对所有与 i 相连的邻居节点在标量打分上的聚合。式中，$\left(H_{st,b}W_{a1}\right)_i$ 是节点 i 的自身打分，由其特征向量与权重 W_{a1} 的内积得到，反映该节点单独的重要性。第二项 $\sum_{j\in\mathcal{N}(i)}\left(H_{st,b}W_{a2}\right)_j$ 是邻域打分，体现局部一致性与上下文支撑。通过该机制，节点的重要性不仅依赖自身特征，还能综合考虑邻居信息，从而获得更加稳健的特征表征。

接下来，针对每个节点 i，在 B 个分支上的得分 $(s^1, s^2, \cdots, s^B)$ 进行拼接，并通过 Softmax 函数将上述分数进行归一化，从而得到该节点在各个分支上的融合权重 $(w_i^1, w_i^2, \cdots, w_i^B)$：

$$w_i^b = \mathrm{Softmax}\left(s_i^b\right) \tag{5}$$

然后，融合后的特征图 H_{final} 通过对所有分支特征图进行逐节点的加权求和得到：

$$H_{final} = \sum_{m=1}^{M} w^b * H_{st,b} \tag{6}$$

其中，w^b 是包含了所有节点权重 w_i^b 的权重向量，符号*表示逐元素乘积。该机制能够在节点级别动态调整各分支的贡献，使融合后的特征表示更具判别力，避免简单拼接或平均融合带来的信息损失。

3.3.2. 注意力池化

当骨架图包含大量节点时，若直接对全部节点特征进行全局读出，往往会引入冗余甚至无关信息，削弱表示的判别性并导致信息瓶颈[39]。为解决这一挑战，学术界已提出多种先进的图池化机制，为此，已有研究提出多种图池化机制，旨在通过学习保留最关键的结构特征[40]。基于这一思路，本文设计了注意力池化 Transformer (APT)模块，如图 3 右上角所示，该模块采用“筛选–聚合–读出”的池化策略。

阶段一：基于图卷积的关键节点选择(GLAPool)。采用 GLAPool 从所有节点特征中选择对任务最关键的 k 个节点，构建信息更丰富的子图。再利用给定融合后的节点特征 H_{final} 与骨架拓扑 E，GLAPool 计算每个节点的重要性评分，再根据分数降序，选择得分最高的 k 个节点。保留节点集合记为：

$$s_{pool} = \mathrm{GLAPool}\left(H_{final}, E, \alpha\right) \tag{7}$$

$$H_{pool} = \mathrm{Top}_k\left(s_{pool}, k\right) \tag{8}$$

阶段二：基于多头注意力的信息聚合(MAB)。为降低因节点丢弃而造成的信息损失，在 APT 模块中引入多头注意力机制(MAB)，将关键节点的信息和图的整体信息进行聚合。将阶段一筛选出的 k 个关键节点的特征矩阵 $H_{pool} \in \mathbb{R}^{k\times d}$，通过线性投影生成查询矩阵 Q：

$$Q = H_{pool}W^{Q} \tag{9}$$

其中，$W^{Q} \in \mathbb{R}^{d\times d}$ 为可学习权重矩阵。同时，对子图进行图卷积操作获得键矩阵 K 和值矩阵 V：

$$K = \mathrm{GCNConv}\left(H_{final}, E\right)W^{K} \tag{10}$$

$$V = \mathrm{GCNConv}\left(H_{final}, E\right)W^{V} \tag{11}$$

式中，$W^{K}, W^{V} \in \mathbb{R}^{d\times d}$ 为学习权重矩阵。随后，通过标准的多头注意力机制计算 Q 与 K，V 之间的加权和。注意力的计算公式为：

$$\mathrm{Attention}\left(Q,K,V\right) = \mathrm{Softmax}\left(\frac{QK^{\top}}{\sqrt{d}}\right)V \tag{12}$$

$$H'_{pool} = \mathrm{MHA}\left(Q,K,V\right) \tag{13}$$

其中，H'_{pool} 为注意力增强后的节点特征矩阵。该机制以局部关键节点为查询，汇聚全图上下文信息，有助于捕捉潜在的判别性语义。最后，对经过 MAB 更新后的 k 个节点特征，采用 1×1 卷积核进行特征读出：

$$g = \mathrm{Conv1d}\left(H'_{pool}\right) \tag{14}$$

得到最终的图级表示向量 $g \in \mathbb{R}^{d}$。与传统的求和平均读出相比，1×1 卷积能在节点维上实现可学习加权与通道重排，缓解读出阶段的信息瓶颈与过度平滑。需要指出，APT 内部的 GLAPool 与前述分支融合的共享评分器在职责上有所不同：前者用于降采样与子图重构，后者仅在不改变图规模的前提下调整分支融合权重。二者解耦后，信息流更为顺畅。整体上，APT 通过自适应融合与注意力池化相结合，在精细整合多源时空特征的同时，有效提升了模型对行人过街意图的判别能力与泛化性能。

3.4. 意图分类

由 APT 模块输出的图级表示向量 g 被送入多层感知机(MLP)分类器以预测意图，输出未经 Softmax 激活的 logits。其具体形式如下：

$$\mathrm{logits} = \mathrm{MLP}\left(g\right) \tag{15}$$

该分类器由两个全连接层构成，在第一线性层后采用 ReLU 激活与 Dropout 正则化以抑制过拟合，分类器输出未经 Softmax 的 logits。训练时采用类别加权的交叉熵作为优化目标，以缓解类别不均衡问题。最终，模型通过第二层全连接(Linear2)输出二维的向量 $\text{logits} \in R^2$，该向量的两个元素分别对应行人“不过街(Non-Crossing)”和“过街(Crossing)”这两类的预测分数。在训练阶段，本文直接将此 logits 向量输入到带类别权重的交叉熵损失函数中进行优化。选择直接输出 logits 是因为交叉熵损失函数在内部集成了 LogSoftmax 操作，这样做可以提升数值计算的稳定性。

4. 实验与结果分析

4.1. 实验设置

4.1.1. 数据集

为评估所提出的 MAGNet-PCI 模型，本文选取 JAAD 与 PIE 两个车载视角的真实场景数据集作为基准。JAAD 数据集主要用于探索行人与车辆驾驶者之间的互动行为，并对行人过街时间进行标记。其记录了车载摄像头拍摄的行人在多种城市环境、天气和光照情况下的过街行为，包含 346 段高分辨率视频，视频长度从 60 帧到 930 帧不等，总计 82,032 帧。并提供了帧中 2D 位置的真实标签及 9 个动作标签。PIE 数据集则聚焦于城市交通场景下的行人意图预测，该数据集提供了在加拿大多伦多城市道路上晴朗天气条件下连续 6 小时的车载视角镜头，包括不同街道结构、不同人群密度地区的 1842 段的路侧行人样本与 293K 个带有注释的帧。并提供包括道路信号标志、交通信号灯、斑马线、道路路沿和交互车辆等在内的详细交通场景标注信息。

4.1.2. 评估指标

为报告研究结果，本文采用计算机视觉领域常用的 C/NC 预测指标体系，包括 Accuracy、Precision、Recall、F1-score。公式为：

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{16}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{17}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{18}$$

$$\text{F1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{19}$$

其中，TP (True Positive)为正确预测为正例的数量，FN (False Negative)为正例预测为负例的数量，FP (False Positive)为负例预测为正例的数量，TN (True Negative)为正确预测的负例数量。

4.1.3. 实现细节

在模型输入端，本文使用 AlphaPose 为每个行人帧提取 26 个关节点的二维骨架。每个关节点的特征由其归一化后的二维坐标(x, y)及其对应的检测置信度(cs)构成，形成三维的特征向量。所有实验基于 PyTorch 与 PyTorch Geometric 实现，并在配备 NVIDIA3090 显卡的服务器上完成。除特别说明外，训练配置保持一致。优化器采用 AdamW，并使用带权重的交叉熵损失函数以缓解数据不均衡问题。时间建模采用滑动窗口机制。

4.2. 模型性能对比

为全面验证 MAGNet-PCI 在行人过街意图预测任务中的有效性，本文在 JAAD 和 PIE 数据集上与多

种已发表的基线模型进行了对比，结果如表 1 所示。为确保比较的公平性，基线分为两类：一类仅基于单模态骨架序列，另一类为融合额外信息的多模态方法(表格中以*标注)；同时纳入非骨骼单模态消融模型(以†标注，输入为行人上下文或环境背景)作为参考。这样的划分既保证了对比维度的一致性，也便于评估 MAGNet-PCI 在纯姿态动态建模任务中的实际表现。

Table 1. Performance comparison of mainstream baseline methods on JAAD and PIE
表 1. JAAD 与 PIE 上主流基线方法的性能对比

模型	JAAD				PIE			
	Accuracy	Precision	Recall	F1-score	Accuracy	Precision	Recall	F1-score
TwoStream [41]*	0.56	0.66	0.66	0.66	0.64	0.33	0.31	0.32
SingleRNN [42]†	0.51	0.63	0.59	0.61	0.60	0.56	0.57	0.55
SFRNN [43]†	0.82	0.54	0.84	0.65	0.69	0.41	0.55	0.47
SST-GCNs [44]	0.68	0.78	0.80	0.75	0.74	0.80	0.71	0.75
TASAR [45]	0.78	0.80	0.87	0.83	0.67	0.78	0.66	0.62
STGCN [46]	0.63	0.66	0.83	0.74	-	-	-	-
ConvLSTM [47]	0.59	0.68	0.70	0.69	0.58	0.32	0.49	0.39
SPI-Net [48]	0.58	0.67	0.65	0.66	0.66	0.35	0.27	0.30
ATGC [49]	0.67	0.72	0.80	0.76	0.59	0.33	0.47	0.41
PedGNN [13]	0.80	0.84	0.87	0.85	0.68	0.66	0.68	0.69
MAGNet-PCI (Ours)	0.83	0.88	0.93	0.90	0.78	0.64	0.67	0.72

在 JAAD 数据集上，MAGNet-PCI 在所有纯骨骼输入模型中取得最佳性能，准确率(Accuracy)达 0.83，较次优模型 PedGNN (0.80)提升 3%。其精确率(Precision)与召回率(Recall)分别达到 0.88 和 0.93，F1-Score 为 0.90，显示模型在捕捉多尺度运动模式和聚焦关键关节方面的优势。在更复杂的 PIE 数据集上，MAGNet-PCI 同样表现领先，准确率为 0.78，F1-Score 为 0.72，全面超越其他骨骼基线。尤其在噪声与遮挡场景下，APT 模块通过“筛选–聚合”机制有效抑制干扰，提升了判断可靠性。尽管部分多模态方法性能更高，MAGNet-PCI 仅基于稀疏骨骼输入仍展现出强大竞争力，验证了其在骨骼动态信息挖掘方面的潜力。

4.3. 消融实验

为系统评估各关键模块对整体性能的贡献，本文在与第 4.1 节一致的训练与评测协议下开展了三组消融实验。所有实验均基于相同的数据划分与骨架预处理流程，除被考察模块外，其余超参数保持不变，以确保结果的公平性与可比性。

本文首先考察 MKGC-ST 模块内部并行核数的影响。如表 2 所示，将核数从 1 提升至 3，模型的 Accuracy 从 0.75 显著提升至 0.83，F1-Score 由 0.83 提升至 0.90。结果表明，传统单核 kernels = 1 的设计，由于其固定的感受野，难以充分建模复杂的人体动态，而多核设计能够并行提取不同尺度的时空特征，更好地覆盖从局部关节细微动作到整体姿态变化的多样模式，从而有效增强表征能力。

Table 2. Ablation study on the multi-kernel mechanism
表 2. 消融实验 - 多核机制

模型	Accuracy	F1-Score
kernels = 1	0.75	0.83
kernels = 2	0.78	0.87
kernels = 3	0.83	0.90

本文评估了图级表示生成方式对模型性能的影响。如表 3 所示，采用本文提出的 APT 模块，Accuracy 达到 0.83，而传统的全局平均池化(Mean Pool)和简单展平(Flatten)方法的 Accuracy 分别降低 0.06 和 0.09。这充分验证了本文的核心观点：简单的全局池化易引入大量背景噪声与无关节点，稀释关键特征。而 APT 模块通过“筛选 - 聚合”机制，有效聚焦对意图判断最关键的关节点(如头部和脚踝)，实现高效的信息整合，从而显著提升了模型的准确率与鲁棒性。

Table 3. Ablation study on attention pooling
表 3. 消融实验 - 注意力池化

模型(图级表示方法)	Accuracy	F1-Score
Flatten + MLP	0.74	0.84
Global Mean Pooling	0.77	0.86
APT	0.83	0.90

最后，本文验证了顶层并行分支设计的有效性。如表 4 所示，原始架构支持任意分支数 B，本文采用双分支结构($B = 2$)。为评估其贡献，我们将其退化为单分支($B = 1$)进行对比。结果表明，单分支模型 Accuracy 为 0.77，明显低于双分支的 0.83，差距达到 0.06。这表明并行分支能够捕获互补的时空特征，并通过分支融合模块实现自适应整合，为最终决策提供更丰富、更可靠的依据，尤其在意图模糊的边界样本中表现更佳。

Table 4. Ablation study on the parallel branch structure
表 4. 消融实验 - 并行分支结构

模型	Accuracy	F1-Score
Single-Branch	0.77	0.86
Dual-Branch	0.83	0.90

从表 2~4 可以看出，三项关键设计：多核时空编码、注意力池化、并行分支融合，均带来正向提升，并且性能增益可叠加。最终完整的 MAGNet-PCI 在 Accuracy 上提升了 0.09，在 F1-Score 上提升了 0.07，充分验证了模型结构的合理性与有效性。

4.4. 跨数据集泛化实验

为评估本文提出的 MAGNet-PCI 模型在未知新场景中的鲁棒性和泛化能力，设计了跨数据集的泛化实验。实验设置包括：在 JAAD 数据集上训练并在 PIE 数据集上测试，以及在 PIE 数据集上训练并在 JAAD 数据集上测试，结果如表 5 所示：

Table 5. Cross-dataset generalization experiments
表 5. 跨数据集泛化实验

Train	Test	Accuracy	F1-score
JAAD	PIE	0.62	0.60
PIE	JAAD	0.66	0.77

实验结果表明，相较于同域测试，跨域预测性能普遍下降，但模型在不同指标上表现出差异化特征。当以 JAAD 为源域时，模型在 PIE 上的 Accuracy 为 0.62，F1-score 为 0.60，表明在场景复杂度更高的 PIE 中仍能维持基本的识别能力。反之，当以 PIE 为源域在 JAAD 上测试时，整体性能更佳，Accuracy 达到 0.66，F1-score 提升至 0.77。这表明 MAGNet-PCI 在跨域迁移中能够较好地捕捉时空动态特征，尤其在目标域为 JAAD 时展现出更强的泛化能力。我们推测，这主要得益于 PIE 数据集本身包含更丰富、多样化的城市场景及行人交互模式。在该数据集上训练，使模型学得对复杂场景更具鲁棒性和通用性的动态骨架特征表征。因此，当应用于相对复杂度较低的 JAAD 数据集时，模型展现出更高的适应性和泛化性能。总体来看，尽管跨域性能低于同域测试，但模型在 Accuracy 和 F1-score 上的稳定表现，验证了并行多核时空编码与注意力读出机制在域移位场景下的鲁棒性，为实际应用提供了可靠依据。

4.5. 定性分析

图 4 展示了 MAGNet-PCI 在 JAAD 与 PIE 数据集上的定性预测结果。其中，(a)与(d)行源自 JAAD 数据集，(b)与(c)行源自 PIE 数据集。绿色字体与包围框表示预测正确，红色表示预测错误。为直观呈现模型的动态决策过程，每一行均展示一个独立案例，并给出最终决策时刻 t 及其后两个关键时刻(t + 30, t + 60)。

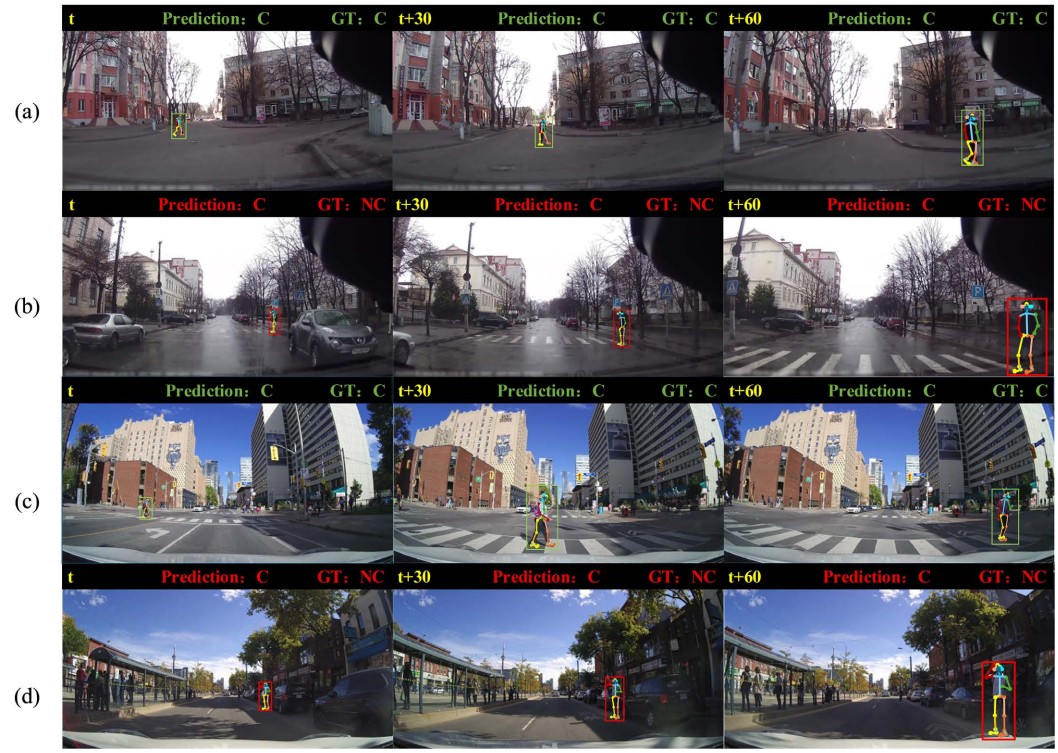


Figure 4. Qualitative results of MAGNet-PCI on the JAAD and PIE
图 4. MAGNet-PCI 在 JAAD 与 PIE 上的定性结果

在成功案例(a)与(c)中，模型展现了优异的鲁棒性与敏感度。案例(a)中，系统能够从自行车远距离接近路口的早期阶段起，稳定预测目标行人的过街意图。案例(c)则更具挑战性：尽管行人在 t 时刻部分被车辆遮挡，但模型仍能凭借多核时空编码机制 MKGC-ST 捕捉其起步的细微动态，并通过 APT 模块聚焦于关键关节，最终输出正确的“过街”预测。然而，案例(b)与(d)反映了意图模糊场景下的误判情况。案例(b)中，行人在序列中有轻微的朝向马路的身体转动；案例(d)中，行人在公交站台附近出现短暂的身体晃动。虽然两者真实意图均为“不过街”，但模型均将这些非功能性姿态解读为“过街”信号，并在整个序列中持续输出错误预测。这说明，虽然 MAGNet-PCI 在捕捉细微动态方面具备显著优势，但其高敏感度也可能导致对非关键动作的过度响应，从而引发误判。如何在保持敏感度的同时，更准确地区分“功能性起步动作”与“无意图的姿态扰动”，并结合更丰富的场景上下文信息，如公交站台区域属性，以实现更全面的意图理解，将是未来的重要研究方向。

5. 结论与展望

本文针对基于骨架序列的行人过街意图预测任务，提出并验证了一种名为 MAGNet-PCI 的新型图神经网络架构。为突破传统单路径模型在特征表达与信息聚合上的局限，本文设计了两大核心模块：多核时空图卷积(MKGC-ST)，通过多个并行的 GConvGRU 核，从不同尺度捕捉并编码复杂的行人动态模式；注意力池化 Transformer (APT)，在解码阶段智能筛选关键节点并结合全局上下文聚合，有效缓解传统池化方式带来的信息瓶颈问题。在 JAAD 与 PIE 两个公开数据集上的全面实验验证显示，MAGNet-PCI 在纯骨骼输入条件下，在所有关键指标上均优于包括 PedGNN 在内的先进基线。在 JAAD 数据集上，模型在 Accuracy 和 F1-Score 上分别达到了 0.83 和 0.90 的最佳水平，表明其在安全关键场景下对“过街”意图的识别更为准确与敏捷，同时具备满足实时应用的计算效率。尽管取得了显著成果，本研究仍存在一定局限。当前模型仅依赖单一骨骼模态，未来可探索多模态融合，将视觉特征或场景上下文信息与骨骼动态结合，以提升复杂场景下的鲁棒性；可引入可解释人工智能(XAI)技术，通过可视化注意力机制揭示模型决策依据，从而增强可信度与透明度；此外，进一步提升模型在意图突变与长尾行为下的响应能力与泛化能力，也是后续研究的重要方向。

参考文献

[1] Fang, J., Wang, F., Xue, J. and Chua, T. (2024) Behavioral Intention Prediction in Driving Scenes: A Survey. *IEEE Transactions on Intelligent Transportation Systems*, **25**, 8334-8355. https://doi.org/10.1109/tits.2024.3374342

[2] Razali, H., Mordan, T. and Alahi, A. (2021) Pedestrian Intention Prediction: A Convolutional Bottom-Up Multi-Task Approach. *Transportation Research Part C*: *Emerging Technologies*, **130**, Article 103259. https://doi.org/10.1016/j.trc.2021.103259

[3] 吕伟, 郭伏, 刘莉, 等. 行人与自动驾驶汽车的交互研究[J]. 中国机械工程, 2023, 34(5): 515-523.

[4] Rasouli, A., Kotseruba, I., Kunic, T. and Tsotsos, J. (2019) PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27 October 2019-2 November 2019, 6262-6271. https://doi.org/10.1109/iccv.2019.00636

[5] Ning, C., Menglu, L., Hao, Y., Xueping, S. and Yunhong, L. (2021) Survey of Pedestrian Detection with Occlusion. *Complex & Intelligent Systems*, **7**, 577-587. https://doi.org/10.1007/s40747-020-00206-8

[6] 陈龙, 杨晨, 蔡英凤, 等. 基于多模态特征融合的行人穿越意图预测方法[J]. 汽车工程, 2023, 45(10): 1779-1790.

[7] Schneider, N. and Gavrila, D.M. (2013) Pedestrian Path Prediction with Recursive Bayesian Filters: A Comparative Study. In: Weickert, J., Hein, M. and Schiele, B., Eds., *German Conference on Pattern Recognition*, Springer, 174-183. https://doi.org/10.1007/978-3-642-40602-7_18

[8] Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Li, F.-F. and Savarese, S. (2016) Social LSTM: Human Trajectory Prediction in Crowded Spaces. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Las Vegas, 27-30 Jun 2016, 961-971. https://doi.org/10.1109/cvpr.2016.110

[9] Achaji, L., Moreau, J., Fouqueray, T., Aioun, F. and Charpillet, F. (2022) Is Attention to Bounding Boxes All You Need for Pedestrian Action Prediction? 2022 *IEEE Intelligent Vehicles Symposium* (*IV*), Aachen, 4-9 June 2022, 895-902. https://doi.org/10.1109/iv51971.2022.9827084

[10] Ahmed, S., Bazi, A.A., Saha, C., Rajbhandari, S. and Huda, M.N. (2023) Multi-Scale Pedestrian Intent Prediction Using 3D Joint Information as Spatio-Temporal Representation. *Expert Systems with Applications*, **225**, Article 120077. https://doi.org/10.1016/j.eswa.2023.120077

[11] Shi, L., Zhang, Y., Cheng, J. and Lu, H. (2020) Skeleton-Based Action Recognition with Multi-Stream Adaptive Graph Convolutional Networks. *IEEE Transactions on Image Processing*, **29**, 9532-9545. https://doi.org/10.1109/tip.2020.3028207

[12] Yan, S., Xiong, Y. and Lin, D. (2018) Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, **32**, 7444-7452. https://doi.org/10.1609/aaai.v32i1.12328

[13] Riaz, M.N., Wielgosz, M., Romera, A.G. and López, A.M. (2023) Synthetic Data Generation Framework, Dataset, and Efficient Deep Model for Pedestrian Intention Prediction. 2023 *IEEE* 26*th International Conference on Intelligent Transportation Systems* (*ITSC*), Bilbao, 24-28 September 2023, 2742-2749. https://doi.org/10.1109/itsc57777.2023.10422401

[14] Kotseruba, I., Rasouli, A. and Tsotsos, J.K. (2021) Benchmark for Evaluating Pedestrian Action Prediction. 2021 *IEEE Winter Conference on Applications of Computer Vision* (*WACV*), Waikoloa, 3-8 January 2021, 1258-1268. https://doi.org/10.1109/wacv48630.2021.00130

[15] 杨彪, 韦智文, 倪蓉蓉, 等. 基于动作条件交互的高效行人过街意图预测[J]. 汽车工程, 2024, 46(1): 29-38.

[16] Yang, D., Zhang, H., Yurtsever, E., Redmill, K.A. and Ozguner, U. (2022) Predicting Pedestrian Crossing Intention with Feature Fusion and Spatio-Temporal Attention. *IEEE Transactions on Intelligent Vehicles*, **7**, 221-230. https://doi.org/10.1109/tiv.2022.3162719

[17] Cao, Z., Hidalgo, G., Simon, T., Wei, S. and Sheikh, Y. (2019) Openpose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **43**, 172-186. https://doi.org/10.1109/tpami.2019.2929257

[18] Fang, H., Li, J., Tang, H., Xu, C., Zhu, H., Xiu, Y., *et al*. (2023) Alphapose: Whole-Body Regional Multi-Person Pose Estimation and Tracking in Real-Time. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **45**, 7157-7173. https://doi.org/10.1109/tpami.2022.3222784

[19] Fang, Z. and Lopez, A.M. (2020) Intention Recognition of Pedestrians and Cyclists by 2D Pose Estimation. *IEEE Transactions on Intelligent Transportation Systems*, **21**, 4773-4783. https://doi.org/10.1109/tits.2019.2946642

[20] Bruna, J., Zaremba, W., Szlam, A., *et al*. (2025) Spectral Networks and Locally Connected Networks on Graphs. arXiv:1312.6203 https://arxiv.org/abs/1312.6203

[21] Kipf, T.N. and Welling, M. (2017) Semi-Supervised Learning with Graph Convolutional Networks. *International Conference on Learning Representations* (*ICLR*), Toulon, 24-26 April 2017.

[22] Kipf, T.N. (2025) Semi-Supervised Classification with Graph Convolutional Networks. arXiv:1609.02907 https://arxiv.org/abs/1609.02907

[23] Hamilton, W., Ying, Z. and Leskovec, J. (2017) Inductive Representation Learning on Large Graphs. *Advances in Neural Information Processing Systems* (*NIPS*), Long Beach, 4-9 December 2017, 1025-1035.

[24] Velickovic, P., Cucurull, G., Casanova, A., *et al*. (2018) Graph Attention Networks. *International Conference on Learning Representations* (*ICLR*), Vancouver, 30 April-3 May 2018.

[25] Hong, X., Zhang, T., Cui, Z. and Yang, J. (2021) Variational Gridded Graph Convolution Network for Node Classification. *IEEE/CAA Journal of Automatica Sinica*, **8**, 1697-1708. https://doi.org/10.1109/jas.2021.1004201

[26] Zhang, H. and Xu, M. (2021) Graph Neural Networks with Multiple Kernel Ensemble Attention. *Knowledge-Based Systems*, **229**, Article 107299. https://doi.org/10.1016/j.knosys.2021.107299

[27] Zhou, K., Song, Q., Huang, X., Zha, D., Zou, N. and Hu, X. (2021) Multi-Channel Graph Neural Networks. *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, 7-15 January 2021, 1352-1358. https://doi.org/10.24963/ijcai.2020/188

[28] Lin, L. and Wang, H. (2020) Graph Attention Networks over Edge Content-Based Channels. *Proceedings of the* 26*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 23-27 August 2020, 1819-1827. https://doi.org/10.1145/3394486.3403233

[29] Seo, Y., Defferrard, M., Vandergheynst, P. and Bresson, X. (2018) Structured Sequence Modeling with Graph Convolutional Recurrent Networks. In: *Lecture Notes in Computer Science*, Springer International Publishing, 362-373. https://doi.org/10.1007/978-3-030-04167-0_33

[30] Cadena, P.R.G., Yang, M., Qian, Y. and Wang, C. (2019) Pedestrian Graph: Pedestrian Crossing Prediction Based on

2D Pose Estimation and Graph Convolutional Networks. 2019 *IEEE Intelligent Transportation Systems Conference* (*ITSC*), Auckland, 27-30 October 2019, 2000-2005. https://doi.org/10.1109/itsc.2019.8917118

[31] Cadena, P.R.G., Qian, Y., Wang, C. and Yang, M. (2022) Pedestrian Graph +: A Fast Pedestrian Crossing Prediction Model Based on Graph Convolutional Networks. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 21050-21061. https://doi.org/10.1109/tits.2022.3173537

[32] 吕超, 崔格格, 孟相浩, 等. 基于图表示的智能车行人意图识别方法[J]. 北京理工大学学报自然版, 2022, 42(7): 688-695.

[33] Zhang, M., Cui, Z., Neumann, M., *et al*. (2018) An End-to-End Deep Learning Architecture for Graph Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, New Orleans, 2-7 February 2018, 2968-2975.

[34] Lee, J., Lee, I. and Kang, J. (2019) Self-Attention Graph Pooling. *International Conference on Machine Learning* (*ICML*), Long Beach, 9-15 June 2019, 3734-3743.

[35] Zhang, Z., Bu, J., Ester, M., *et al*. (2025) Hierarchical Graph Pooling with Structure Learning. arXiv:1911.05954 https://arxiv.org/abs/1911.05954

[36] Baek, J., Kang, M. and Hwang, S.J. (2025) Accurate Learning of Graph Representations with Graph Multiset Pooling. arXiv:2102.11533 https://arxiv.org/abs/2102.11533

[37] 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.

[38] Gao, H. and Ji, S. (2019) Graph U-Nets. *International Conference on Machine Learning* (*ICML*), Long Beach, 9-15 Jun 2019, 2083-2092.

[39] 桑海峰, 刘玉龙, 刘泉恺. 基于混合注意力机制的多信息行人过街意图预测[J]. 控制与决策, 2024, 39(12): 3946-3954.

[40] Ying, Z., You, J., Morris, C., *et al*. (2018) Hierarchical Graph Representation Learning with Differentiable Pooling. *Advances in Neural Information Processing Systems*, **31**, 4805-4815.

[41] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, **27**, 568-567.

[42] Kotseruba, I., Rasouli, A. and Tsotsos, J.K. (2020) Do They Want to Cross? Understanding Pedestrian Intention for Behavior Prediction. 2020 *IEEE Intelligent Vehicles Symposium* (*IV*), Las Vegas, 19 October 2020-13 November 2020, 1688-1693. https://doi.org/10.1109/iv47402.2020.9304591

[43] Rasouli, A., Kotseruba, I. and Tsotsos, J.K. (2025) Pedestrian Action Anticipation Using Contextual Feature Fusion in Stacked RNNs. arXiv:2005.06582. https://arxiv.org/abs/2005.06582

[44] Xie, J., Zhao, Y., Meng, Y., Zhao, H., Nguyen, A. and Zheng, Y. (2025) Are Spatial-Temporal Graph Convolution Networks for Human Action Recognition Over-Parameterized? 2025 *IEEE/CVF Conference on Computer Vision and Pattern Recognition* (*CVPR*), Nashville, 10-17 June 2025, 24309-24319. https://doi.org/10.1109/cvpr52734.2025.02264

[45] Diao, Y., Wu, B., Zhang, R., *et al*. (2025) TASAR: Transfer-Based Attack on Skeletal Action Recognition. arXiv:2409.02483 https://arxiv.org/abs/2409.02483

[46] Zhang, X., Angeloudis, P. and Demiris, Y. (2022) ST Crossingpose: A Spatial-Temporal Graph Convolutional Network for Skeleton-Based Pedestrian Crossing Intention Prediction. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 20773-20782. https://doi.org/10.1109/tits.2022.3177367

[47] Shi, X., Chen, Z., Wang, H., *et al*. (2015) Convolutional LSTM Network: A Machine Learning Approach for Precipitation Now-Casting. *Advances in Neural Information Processing Systems*, **28**, 802-810.

[48] Gesnouin, J., Pechberti, S., Bresson, G., Stanciulescu, B. and Moutarde, F. (2020) Predicting Intentions of Pedestrians from 2D Skeletal Pose Sequences with a Representation-Focused Multi-Branch Deep Learning Network. *Algorithms*, **13**, Article 331. https://doi.org/10.3390/a13120331

[49] Rasouli, A., Kotseruba, I. and Tsotsos, J.K. (2017) Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. 2017 *IEEE International Conference on Computer Vision Workshops* (*ICCVW*), Venice, 22-29 October 2017, 206-213. https://doi.org/10.1109/iccvw.2017.33