分类号_____ 密级_____

U D C_____ 编号_____

华中师范大学

硕 士 学 位 论 文

基于姿态混合的动态图卷积群体行为识别

学位申请人姓名：谢作权

申请学位学生类别：全日制硕士

申请学位学科专业：计算机技术

指 导 教 师 姓 名：姚华雄 副教授

硕士学位论文

基于姿态混合的动态图卷积群体行为识别

论文作者：谢作权
指导教师：姚华雄　副教授
学科专业：计算机技术
研究方向：群体行为识别

华中师范大学计算机学院

2023 年 5 月

硕士学位论文
MASTER'S THESIS

Pose Mixed Dynamic Graph Convolution for Group Activity Recognition

A Thesis

Submitted in Partial Fulfillment of the Requirement

For the M.S. Degree in Computer Technology

By

Zuoquan Xie

Postgraduate Program

School of Computer

Central China Normal University

Supervisor: Huaxiong Yao

Academic Title: Associate Professor Signature_____

Approved

May. 2023

华中师范大学学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的研究成果。除文中已经标明引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的研究成果。对本文的研究做出贡献的个人和集体，均已在文中以明确方式标明。本声明的法律结果由本人承担。

作者签名：　　日期：　　年　　月　　日

学位论文版权使用授权书

学位论文作者完全了解华中师范大学有关保留、使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属华中师范大学。学校有权保留并向国家有关部门或机构送交论文的复印件和电子版，允许学位论文被查阅和借阅；学校可以公布学位论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存、汇编学位论文。（保密的学位论文在解密后遵守此规定）

保密论文注释：本学位论文属于保密，在 _____年解密后适用本授权书。

非保密论文注释：本学位论文不属于保密范围，适用本授权书。

作者签名：　　导师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

本人已经认真阅读“CALIS 高校学位论文全文数据库发布章程”，同意将本人的学位论文提交“CALIS 高校学位论文全文数据库”中全文发布，并可按“章程”中的规定享受相关权益。同意论文提交后滞后：□半年；□一年；□二年发布。

作者签名：　　导师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘　要

群体行为识别是一个重要而且具有挑战性的问题，最近在视频理解领域中已经引起了越来越多的关注。群体行为识别旨在识别在多人场景中的总体行为，它具有广阔的应用前景，例如体育视频分析、安防系统、社会行为理解、视频搜索和检索等。群体行为识别的核心不仅在于识别单个行为者的动作，还在于完全地探索场景信息以及个体之间的交互关系。然而，一些之前的工作只致力于在个体行为者特征上面进行推理，却忽视了场景信息的建模，这些场景信息往往对于推理该群体的行为具有一些线索。在个体行为者之间的空间和时间关系建模上，之前的方法或者单独地去捕获空间和时间关系，或者直接聚合个体的特征来生成该群体的表征，这样很难同时在空间和时间维度上使得模型达到最优，并且生成的群体特征不具有丰富的语义性和相关性。

为了解决上述的一些问题，本文提出了基于跨时间步动态图卷积的群体行为识别模型，该分支模型将个体的外观特征作为输入。具体来说，首先利用基于 Transformer 的场景编码模块将场景上下文信息编码进个体的外观特征中，使得个体的外观特征与所在的场景关联从而得到增强。随后，将进一步地探索场景上下文编码过的个体特征的交互关系。具体来说，由不同帧所有的个体形成一个跨时间步的时空图，经过动态图卷积网络建模所有个体的时空关系。最后，在两个广泛使用的数据集上，即 Volleyball Dataset 和 Collective Activity Dataset，提出的方法达到了与最先进模型竞争的结果，实验结果论证了每个提出模块的有效性。

个体行为者的姿态特征通常对群体行为的学习起着引导作用，不仅可以促进个体动作的准确识别，而且还包含了群体行为的关键线索。所以，本文提出了混合姿态特征的群体行为识别模型 PDGCN，该模型旨在充分学习姿态特征中的时空关系，得到具有丰富信息的群体表征来进行行为预测。具体来说，首先用姿态估计骨干网络预测数据集标注的所有个体的关节点坐标，经过线性投影变换为姿态特征。同样地，将场景上下文信息编码进个体的姿态特征中，然后使用跨时间步的动态图卷积模块推理姿态特征中的时空关系。进一步地，分别使用基于姿态特征的分支与基于外观特征的分支学习场景上下文和推理时空交互关系，然后将这两个分支融合进行最终的群体行为预测。大量的实验验证了每个分支的有效性和合理性，该方法在两个数据集上达到了与最先进模型竞争的结果，对比单分支取得了精度的提升。

关键词：群体行为识别；图神经网络；注意力机制；视频分析

Abstract

Group activity recognition is a crucial and challenging problem that has recently attracted increasing attention in the field of video understanding. Group activity recognition aims to recognize overall activity in multi-person scenes, and it has promising applications, such as sports video analysis, security systems, social behavior understanding, video search and retrieval, etc. The core of group activity recognition lies not only in recognizing the actions of individual actors, but also in fully exploring the scene information and the interactions among individuals. However, some previous methods have only dedicated to reasoning on individual actor features, but neglected to model scene information that often has some clues for reasoning about the group's activity. For modeling the spatial and temporal relationships between individual actors, previous approaches either capture the spatial and temporal relationships separately or aggregate individual features directly to generate group representations, which makes it difficult to optimize the model in both spatial and temporal dimensions, and the generated group features are not sufficient in semantics and relevance.

In order to solve the mentioned problems, this paper proposes a group activity recognition model based on cross-time step dynamic graphs, and this branch takes the appearance features of individuals as input. Specifically, the scene context information is first encoded into the appearance features of individuals using the Transformer-based scene encoding module, so that the appearance features of individuals are associated with the scene they are in and thus enhanced. Subsequently, the interaction of the individual features encoded by the scene context is further explored. Specifically, a spatial-temporal graph is formed from all individuals in different frames across time steps, and the spatial-temporal relationships of all individuals are learned by a dynamic graph convolutional network. Finally, these individual features with spatial-temporal relationships are globally pooled to obtain the group features. Moreover, on two widely used datasets, namely Volleyball Dataset and Collective Activity Dataset, the proposed method achieves results that compete with state-of-the-art methods, and experimental results demonstrate the effectiveness of each proposed module.

The pose features of individual actors usually play a guiding role in learning of group activity, which not only facilitate the accurate recognition of individual actions, but also

contain key cues for group activity. Therefore, this paper proposes a group activity recognition model mixed with pose features, which aims to fully learn the spatial-temporal relationships in pose features and obtain informative group representations for activity prediction. Firstly, the pose estimation backbone network is employed to predict the key-point coordinates of all individuals annotated in the dataset, which are transformed into pose features by linear projection. Similarly, the scene context information is encoded into the pose features of individuals, and then the cross-time dynamic graph convolution module is utilized to reason about the spatial-temporal relationships. Furthermore, the pose-based branch and the appearance-based branch are used to learn the scene context and reason about spatial-temporal interactions respectively, and then the two branches are fused for the final group activity prediction. Extensive experiments verify the effectiveness and rationality of each branch, and the method achieves results that compete with state-of-the-art models on both datasets, while achieving accuracy improvements over single branch.

Keywords: group activity recognition; graph neural networks; attention mechanism; video analysis

目　录

1 绪论

1.1 研究背景与意义

随着科技与经济的高速发展，人们已经进入了大数据时代，每天都有海量的视频、图片、文本等数据产生。特别对于视频数据，它主要来自于监控设备，所占据的体量最大包含的信息量更多，分析研究这些视频数据的过程比文本图片更为复杂和困难。近年来随着深度学习和机器学习的发展，以及算力设备的不断更迭，处理大量的视频数据逐渐变得简便，视频理解和分析由于其广阔的应用前景已经成为了研究热点，吸引了越来越多的国内外学者的注意。

群体行为识别 (Group Activity Recognition, GAR) 作为视频理解和分析的一个分支，是一个重要且具有挑战性的问题，在很多领域具有广泛的社会影响，例如体育视频分析、安防系统、社会行为理解、视频搜索和检索、人机交互等[1–2]。群体行为识别旨在检测出一个多人场景的总体行为[3–5]，识别的对象是一个未经剪辑的视频片段。群体行为识别不同于传统的动作识别，因为仅仅识别出某个个体的动作不足以推断出整个群体的行为，个体的动作只是在整个群体行为中承担某一个角色。群体行为识别不仅需要识别单个行为者的动作，还需要学习个体所在的场景上下文信息，以及这些个体行为者之间的时间和空间的交互关系，据此推理出该群体的表征[6–9]。这项任务的处理流程更为复杂。

由于监控设备部署的成本不断降低，视频数据呈爆炸式地增长，将群体行为识别应用到海量的视频数据可以提取到很多有用的信息，例如对监控视频中的人群行为进行识别，可以检测出异常的行为，并及时发出预警阻止一些负面事情的发生；对体育比赛视频进行实时的识别，可以自动分析出当前赛场的局势，不需要解说员就可以对比赛进行自动解说和分析。在过去由于硬件设备和方法技术的限制，对视频数据的处理大多依靠人工提取，这种方式不仅时间效率低下、价格成本高昂，而且提取的视频特征包含的信息相关性低可靠性不强，没有很好的可解释性和泛化能力。因此手工处理视频数据的方式无法应对大数据时代带来的挑战，从而需要一种高效且智能的视频分析技术。在未来的人工智能和大数据时代，群体行为识别技术将会充分利用深度学习的优势，将会继续在视频分析理解领域占据重要的作用，会不断地受到学术界和工业界的广泛关注。

群体行为识别任务处理的视频主要来自监控设备和体育比赛记录，研究的对象是视频中的目标群体。对于来自监控设备的视频，识别的目标是检测出参与人数最

多的行为，例如一个视频片段里面涉及到了马路的场景，对应一群人在穿越马路，该视频片段应该被识别为过马路 (crossing) 而不是排队 (queueing)，识别的结果与对应的场景相关联。对于来自体育比赛的视频，目标是识别出场上所有运动员的总体行为。例如在一场排球比赛中，场上的运动员在空间上分为左右两侧，如果左侧有一个运动员在扣球，而右侧的运动员都在防守状态中，此时视频片段应该被识别为左侧扣球 (left spiking)，识别结果是由场上所有运动员的交互关系决定的。上述两个例子论证了群体行为识别的两个重要方面，即对场景信息的理解和对个体之间时空关系的推理。

然而，在实际应用场景中原始视频的每一帧图像中没有提供个体行为者的边界框 (bounding box)，这与群体行为识别研究所要求的实验条件有所差异。通常群体行为识别的视频数据是经过预先标注的，即每一帧图片中所有目标个体的边界框都是给定的 (或者所有个体的轨迹是通过某种方法[10]给定的)，例如 Volleyball 数据集[11]。因此，现实场景中的视频数据需要经过上游的处理才可以应用群体行为识别技术，该上游处理的过程可能包含一些人类先验知识来决定哪些是目标群体，哪些是无关对象。于是，直接在现实场景中应用该技术可能会导致算法性能和其理论效果有所差异，甚至是大打折扣。

1.2 国内外研究现状

群体行为识别的发展可以被深度学习分为两个阶段，即基于手工特征的方法和基于深度学习的方法。

1.2.1 基于手工特征的方法

传统的基于手工特征的群体行为识别方法可以分为两种：自顶向下的方法和自底向上的方法。

自顶向下的方法致力于分析整个群体或每个子群体的全局运动模式，研究群体的轨迹和互动，而场景中特定行为者的个体动作则不那么重要。这样，它们对遮挡和低分辨率的视频场景识别具有更好的鲁棒性。根据 [4]，自顶向下的方法可以分为：基于轨迹的方法，基于子群体交互的方法和基于多镜头上下文的方法。

基于轨迹的方法集中于分析个体轨迹间的相互作用来识别群体行为。Vaswani 等人[12]将移动物体建模为二维平面上的点物体。他们建议按照 Kendall's 形状理论，将群体活动表现为这些点的配置在时间框架内的多边形变化，而不是跟踪每个点并识别它们的互动。相似地，Khan 和 Shah[13]提出了一种检测群体活动的方法，这种

活动可以用刚性信息来描述。他们将每个实体表示为三维多边形的一个角，每个实体在三维多边形平面上的轨迹被视为轨迹特征。最终的分类结果是由轨迹和参与实体之间的相互作用组成的结构推断出来的。Zhou 等人[14]设计了一组特征来衡量两个轨迹之间的因果关系的强度，另一组特征描述了因果关系的类型。这两组特征与轨迹对的常规速度和位置特征相融合，以探索两个实体之间的关系。

为了应对一个场景中多个群体进行不同活动的复杂情况，基于子群体交互的方法首先检测子群体，然后分析不同群体的交互和每个群体的活动。Yin 等人[15]首先通过最小生成树算法将每个个体聚类为几个子群，然后使用基于社会网络分析的特征描述来提取结构特征，这些特征包含每个子群的全局模式以及每个群体中个体的局部运动信息。最后，训练一个高斯过程动态模型来分别模拟不同的群体行为。子群信息是在复杂场景下识别群体行为的一个有用线索，然而如何识别有意义的子群仍然是一个具有挑战性的问题。Kim 等人[16]提出了检测群体交互区并随着时间的推移进行更新，这样可以抑制噪声信息，增强行为的活跃区。为了表示群体交互区内的交互，他们进一步提出了两个特征，即群体交互能量特征、吸引和排斥特征。Tran 等人[17]通过社会信号线索度量个体之间的交互程度。然后，他们利用图聚类算法来发现场景中相互作用的子群，并抛弃了非主导的群体。为了更好地理解群体行为，他们提出了一个描述符，对社会交互线索和活跃子群体中的个体运动信息进行编码。

基于多镜头上下文的方法使用多个摄像头预测群体活动，在多机位场景中，机位内和机位间的上下文是重要的信息。在 [18] 中，多摄像头中的多个轨迹被用来提取个体的时空特征。他们考虑了两种层次聚类方法对个体进行分组，即 agglomerative clustering 和 decisive clustering，使用异同度来衡量跟踪目标之间的关系。Zha 等人[19]提出了一个带有隐藏变量的图模型，从中提取摄像机内和摄像机间的语境。通过优化图模型的结构，自动探索语境。此外，他们提出了一个时空特征，即警戒区，以编码一个区域的运动信息，这被证明对群体行为的表示是有效的。

自下而上的方法适用于识别个体数量有限的群体行为，这些个体有与其他人不同的角色，该方法需要具有识别每个个体的动作及其结构的能力。根据 [4]，自底向上的群体行为识别模型可以分为：基于隐藏马尔可夫的模型 (Hidden Markov model, HMM)，基于描述符的模型 (descriptor based model)，交互上下文 (interaction context) 模型和基于轨迹的方法 (tracklets based method)。

基于 HMM 的模型[20]适用于解决分层结构。在底层，个人的原子动作从序列中被识别出来，而第二层则对群体行为进行建模。场景中的上下文信息有助于区分模

棱两可的行为，如站立和排队。基于描述符的方法[21–22]提出了从焦点个体及其周围区域提取各种特征描述符来整合上下文信息。与基于描述符的方法 (它提供焦点个体特征与群体内所有个体之间的上下文信息) 不同的是，交互上下文模型[23–24]提供了人与人、人与群体、群体与群体之间的互动信息，使其有可能解决复杂的互动场景。对于自下而上的方法，识别每个人的连贯轨迹是群体行为识别的一个预处理步骤。以前的方法是将跟踪和识别的任务分开，但是个体的运动与其活动有时是相互关联的。基于轨迹的方法[25–26]的目标是联合执行两项任务，并使其相互促进。

总结来说，自上而下的方法是从群体层面的运动和互动来分析活动。这些方法的缺点是缺乏对活动的详细描述，不能充分地利用个人层面的特征。自下而上的方法侧重于识别每个个体，并根据个体特征的集合和它们的统计数据来描述行为。因此，它对由于遮挡或遗漏检测导致提取失败的个体特征很敏感。然而，基于手工特征的方法处理起来难度大，模型效果差，很容易受到场景偏差的影响。随着深度学习和算力的发展，基于深度学习的群体行为识别方法已经成为了主流，并取得了显著的提升效果。

1.2.2 基于深度学习的方法

基于深度学习的群体行为识别方法根据侧重的关键点可以分为：层次时序建模 (hierarchical temporal modeling)，关系建模 (relationship modeling)，注意力建模 (attention modeling) 和统一建模框架 (a unified modeling framework)。

群体行为识别任务的一个挑战是如何设计适当的网络，使学习算法专注于区分更高层次的行为类别，它是关于群体行为的空间和时间演变。长短期记忆网络 (LSTM)[27]，一种特殊类型的循环神经网络 (RNNs)，已经在序列任务中取得了巨大的成功，包括语音识别和图像标题生成。对于群体行为的识别，一些研究者试图应用 LSTM 构建一个层次结构表征来推断个人行动和群体活动。早期的基于深度学习的方法使用卷积神经网络 (CNNs) 提取低层次的视觉特征，然后使用 LSTM 来进行时序建模[11,28–33]。

Ibrahim 等人[11]提出了一个两阶段的层次化深度时间模型 (HDTM)。第一阶段在每个个体的轨迹应用个人层面的 LSTM 来建模个人行为。在第二阶段，采用群体级的 LSTM 来结合个人层面的信息，形成群体行为的群体级特征。该方法是第一个结合深度 LSTM 框架来解决群体行为识别的工作。

Shu 等人[30]认为现有的群体行为基准数据集 (Collective Activity dataset[21]和 Volleyball dataset[11]) 太小，无法训练出健壮的 LSTMs 框架。为了解决这个问题，他们提

出了置信度-能量循环网络 (CERN)，该网络通过加入置信度和基于能量的模型扩展了 LSTMs 框架的两级层次结构，并且通过最小化预测能量来保证准确识别群体行为类别。紧接着，HRN[34]使用 LSTM 引入了 relational layer 捕获场景中每个人的空间关系，然后生成群体表征。注意力机制也在 GAR 中证明了它的有效性。例如，Qi 等人[35]提出了一个注意力的语义循环神经网络 stagNet，从单词标签和视觉数据中建立一个语义图。个人行为和时序上下文信息被一个 structural-RNN 模型整合，个人之间的空间关系是通过信息传递机制在语义图中推断出来的。除此之外，个体级的空间注意力和帧级的时间注意力被用来自动发现关键人物和关键帧。但是 RNNs 难以训练以及无法考虑到全局且长程的信息，所以逐渐地被注意力机制取代。

因为学习个体与个体之间的交互关系对 GAR 是至关重要的，很多现在的研究工作致力于探索如何捕捉行为者在场景中的上下文信息以及他们之间的时空关系。一些工作基于图的视角来解决上述问题，例如使用图卷积网络 (GCN)[36]来进行深层关系的建模。更近期的一些工作使用注意力机制建模，包括使用 Transformers[37]进行关系推理，主要集中在决定场景中最具影响力的个体，子群体或者交互关系。现存的方法主要使用基于 RGB 和/或基于光流的特征，并应用 RoIAlign[38]来表征行为者。还有一些近期的工作使用行为者的关节点/姿态来代替或者增强上述的 RoIAlign 特征。一些方法仅仅使用基于数值坐标的关节点表征，然而其他一些方法使用从深度姿态骨干网络中提取的高维向量特征，但该做法不是那么高效。

最近，图卷积网络 (GCN)[36]已经成为深度学习的一个新兴课题，已被应用于计算机视觉的许多领域，如视觉跟踪和单一个体动作识别。图卷积网络是解决群体行为识别的合适模型，其中每个人都可以被看作是一个节点。Wu 等人[39]将 GCN 引入群体行为识别中，并提出了 ARG 模型。通过卷积神经网络提取个人层面的特征，并基于视觉相似性和个人之间的空间位置距离建立一个行为者关系图。图卷积网络被用来对行为者关系图进行关系推理以获得每个人的关系特征。

与 ARG[39]不同的是，DIN[8]提出了在特定于个体的交互场 (interaction field) 来更新时空信息，从而实现群体行为的动态推理。它对时空图中的每一个个体，首先用 2D 卷积神经网络预测该个体交互场的关系矩阵，然后用双线性插值法预测关系矩阵的动态游走偏移量。SACRF[7]使用堆叠方式的时空注意力去提取特征，随后添加到 mean-field 的条件随机场去强化这些特征。不同地，Yan 等人[40]分别建立空间和时间关系图，随后建模行为者之间的关系。

Hu 等人[41]运用了一种新的方法即 PRL，将深度强化学习用于群体行为识别中的关系学习。他们首先构建了一个语义关系图来建模场景中的个体关系。然后，两

个基于马尔科夫决策过程的代理被用来完善该关系图。关系门控代理负责执行相关的关系学习和抛弃不相关的关系，另一个特征提炼代理提炼特征的关键帧，这类似于一个时间注意力机制。

Transformer 最早是在 [37] 中提出的，用于序列到序列的机器翻译任务，此后在各种自然语言处理任务中流行起来，最近在机器视觉任务中被广泛使用。Gavrilyuk 等人[6]受 Transformer 的启发，提出了一个 Actor-Transformers 网络，该网络学习行为者之间的相互作用，自适应地提取行为识别的重要信息，并且使用 I3D[42]来编码时序信息。

随后，TCE-STBiP[43]模型提出了使用 Transformer 的编码器去编码全局场景的上下文信息，然后在构造的时空图上采用时空双线性池化模块动态更新节点特征，并取得了很好的效果。GroupForme[44]引入了一种聚类的交叉注意力机制，来获取具有更好的群体信息的特征，并且第一次将 Transformer 的解码器引入了来同时提取时空特征，打破了以往模型采用堆叠或者并行融合时空注意力的范式。

在最近的工作中，Dual-AI 框架[45]提出互补的空间-时间和时间-空间双路径来学习行为者的交互关系，并且用一种自监督多尺度的行为者对比损失来加强两条路径之间行为者的一致性，以此进行有效的表征学习。

COMPOSER[46]只使用轻量级基于坐标的关节点表征作为输入，将一个视频建模成为多个 tokens 来表示视频中多尺度的语义概念，然后用提出的多尺度 Transformer 模块在不同的语义尺度上进行分层地注意力推理，通过提炼多尺度的表征来学习群体行为。

1.3 主要研究内容

本文研究基于姿态混合的动态图卷积群体行为识别，群体行为识别是一项复杂且具有挑战的任务，它的关键在于充分挖掘行为者在场景中的上下文信息以及他们之间的时空关系，推理出他们之间关系的重要性得到群体表征进行预测。经过对国内外工作的研究，本文首先提出了利用跨时间步的动态图卷积在个体的外观特征进行关系推理，然后融合行为者的姿态特征分支进行推理，主要工作如下：

(1) 提出基于跨时间步动态图卷积的群体行为识别方法。现存的一些方法在进行群体行为识别的时候，忽视了对全局场景上下文信息的提取，而仅在行为者的外观特征上进行推理，这样会丢失一些有用的线索。此外，在探索个体之间的时空关系过程中，有些方法采用异步的方式分别地建模空间和时间关系；还有些方法没有跨时间步考虑空间关系，个体关系特征的感受野太小；还有些直接利用注意力建模

空间和时间关系，引起一些不相关信息的干扰。因此，本文提出了一种基于跨时间步动态图的群体行为识别方法，来学习全局的场景上下文信息和更具语义相关性的时空关系。具体来说，首先将利用场景编码模块将上下文信息编码到行为者的外观特征中，然后用动态图卷积模块在创建的时空图上进行关系推理。算法在两个基准数据集上的消融实验验证了模块的有效性。

(2) 提出混合姿态特征的群体行为识别方法。行为者的姿态特征对识别个体动作和群体行为起着关键作用，它包含了个体更细粒度的运动信息，同时不会受到视频背景偏差带来的影响。然而，一些方法仅仅使用外观特征，没有考虑姿态信息。因此，在前文方法 (外观特征分支) 的基础上进一步融合了姿态特征分支，充分地利用个体的姿态信息建模时空关系。相似地，姿态特征分支也包含场景编码模块和动态图卷积模块。在两个数据集上的实验结果表明，两个分支互补地学习群体行为，并且性能优于单个分支。

1.4 本文的组织结构

本文基于姿态混合的动态图卷积进行群体行为识别研究，一共包括五个章节，本文的组织结构如下：

第一章为绪论，主要介绍了群体行为识别的研究背景和意义，并且对该领域的国内外研究现状做了总结和分析，阐述了本文的主要研究内容和方法动机。

第二章介绍了与本文研究工作相关的理论和技术，便于展开后面的章节，包括提取图片特征的骨干网络，图神经网络和时空注意力机制。

第三章提出了基于跨时间步动态图卷积的群体行为识别方法，介绍了网络的总体框架，详细阐述了算法的每一个模块，最后分析比较在两个广泛使用的数据集上的实验结果，通过对比实验验证每一个模块的有效性。

第四章提出了混合姿态特征的群体行为识别方法，它在第三章方法的基础上进一步融合了姿态特征推理分支。该章节介绍了网络总体设计，姿态特征提取模块和混合姿态特征的推理过程，最后在两个基准数据集上进行实验。

第五章是总结与展望，对本文的研究内容进行总结，对未来群体行为识别的工作提出展望。

2 相关理论技术

2.1 提取图片特征的骨干网络

在本研究中，下面三种骨干网络 (backbone) 可以作为选择来提取图片的特征，这些骨干网络都是基于 2D 卷积神经网络 (2D CNNs)。之前的工作有的将提取图片的特征和群体行为的推理分为两个阶段执行，但最近的工作将两个阶段合并以端到端的方式进行群体行为识别。下面介绍涉及到的骨干网络总体框架，以及在本研究中提取特征的过程。

2.1.1 ResNet-18

ResNet[47]网络是 2015 年由微软实验室的何恺明提出，获得 ImageNet[48] 2015 图像分类竞赛第一名。在 ResNet 网络提出之前，传统的卷积神经网络都是将一系列的卷积层和池化层堆叠得到的，但当网络堆叠到一定深度时，就会出现退化问题。ResNet 引入了一种残差网络结构，使用这种结构可以避免出现模型性能退化问题，可以实现搭建较深的网络结构 (突破 1000 层)。ResNet 中的残差结构可以表示为：

$$\mathbf{y} = \mathcal{F}(\mathbf{x}, \{W_i\}) + \mathbf{x}, \tag{2.1}$$

这里 $\mathbf{x}$ 和 $\mathbf{y}$ 表示给定层的输入和输出向量，函数 $\mathcal{F}(\mathbf{x}, \{W_i\})$ 表示要学习的残差映射，例如前馈全连接网络。$\mathcal{F}+\mathbf{x}$ 的操作通过短接 (shortcut connection) 和 element-wise 加法来实现，其中短接代表跳过一个或多个网络层，上式的短接只是简单地执行了单位变换 (identity mapping)。

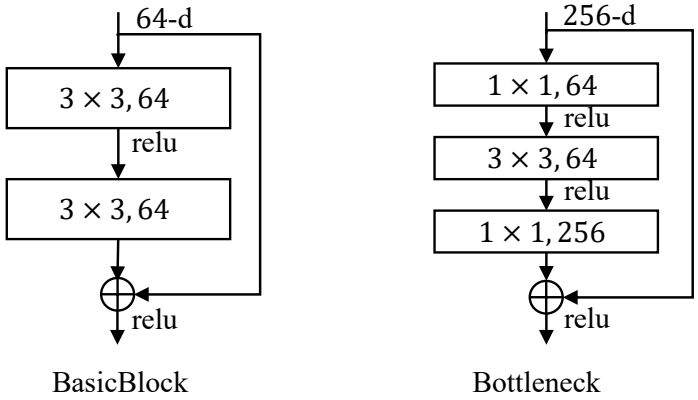


图 2.1 ResNet 网络中两种不同的残差块

图 2.1 表示深度 ResNet 网络用到的两种残差块 (residual block) 类型的示例，左边的普通残差块 (basic block) 由两个 3×3 的卷积层构成，用于 ResNet-18/34；右边的称之为 bottleneck 残差块用于 ResNet-50/101/152，它由 3 个卷积层构成，分别是 1×1，3×3 和 1×1 的卷积，这里 1×1 的卷积层用来降低然后增加维度，3×3 的层作为一个瓶颈具有更小的输入/输出维度。

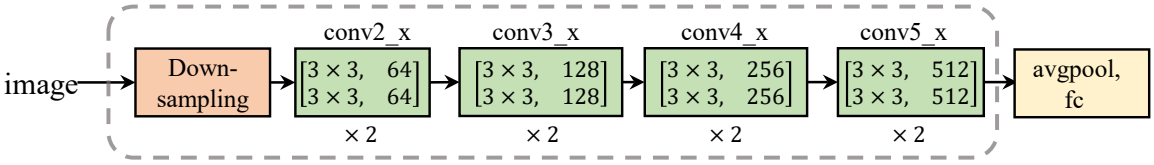


图 2.2 ResNet-18 网络的简化结构图

ResNet-18 的结构如图 2.2 所示，其中 downsampling 代表 7×7 步长为 2 的卷积和 3×3 步长为 2 的最大池化的下采样操作，后面的 conv2_x conv5_x 代表残差块，其结构如图 2.1 左边所示，对应下采样操作在 conv3_1, conv4_1 和 conv5_1 以第一层步长为 2 的卷积实现。

在本工作中，采用的 ResNet-18 骨干网络结构如图 2 中的虚线框所示，它不包括最后一个子块即均值池化和全连接层，然后加载在 ImageNet 上预训练的权重进行迁移学习。图片输入到该骨干网络后，将最后一个残差块的输出作为图片的特征，即 conv5_2 的输出，此时的下采样因子为 32。

2.1.2 VGG-16

VGG 网络[49]在 2014 年由牛津大学著名研究组 VGG(Visual Geometry Group) 在论文 *Very Deep Convolutional Networks for Large-scale Image Recognition* 中提出，斩获该年 ImageNet 竞赛中 Localization task(定位任务) 第一名和 Classification Task(分类任务) 第二名。由于 VGG 网络的模型参数较多以及准确率不高，现在作为骨干网络并不占据优势，但是在之前的群体行为识别模型中，有的仍采用 VGG 网络作为骨干网络。

图 2.3 为 VGG-16 网络的结构 (包含了 16 个权重层)，其中 conv3-x 表示卷积核大小为 3×3，步长为 1，padding 为 1 的卷积层和 ReLU 激活层，max pooling 表示 size 为 2，步长为 2 的最大池化下采样层，FC 表示全连接层及其中间的 ReLU 激活层。

在本研究中，采用的 VGG-16 骨干网络结构不包括最后一个 max pooling 层和 FC 块 (如图 2.3 中的虚线框所示)，加载在 ImageNet 上预训练的权重进行迁移学习。

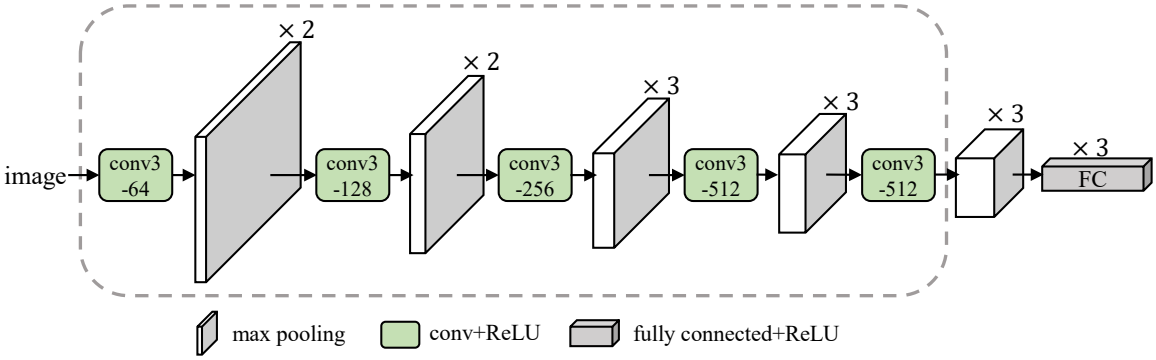


图 2.3 VGG-16 网络的简化结构图

视频图片输入到该骨干网络后，将最后一个卷积层的输出作为图片的特征，此时经过了 4 个 max pooling 层，所以得到的特征图下采样因子为 16。

2.1.3 Inception-v3

Inception-v3[50]在论文 *Rethinking the Inception Architecture for Computer Vision* 中由 Google 团队提出，获得 2015 年 ImageNet 竞赛中分类任务的第二名 (第一名是 ResNet)，它在 GoogLeNet(Inception-v1)[51]的基础上做出了很多改进。由于权衡 Inception-v3 模型体量大小和识别准确率，目前在各种视觉任务中作为骨干网络仍被广泛使用。

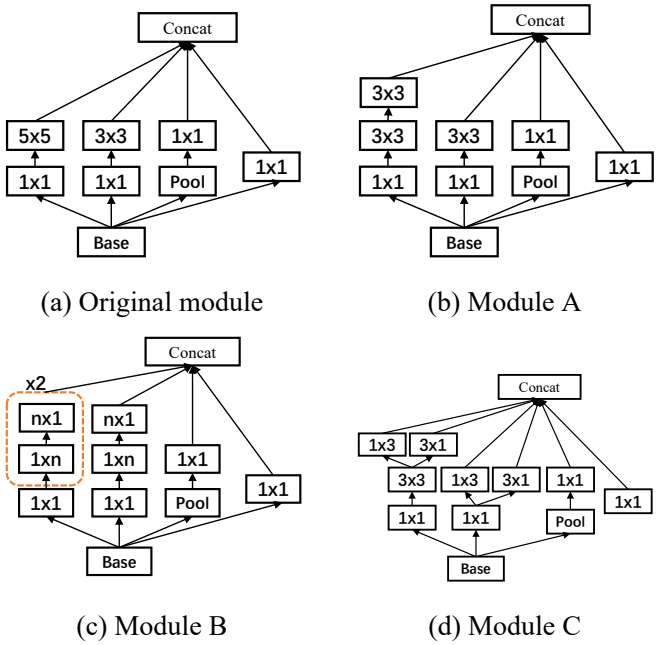


图 2.4 不同的 Inception 模块

GoogLeNet[51]在 2014 年由 Google 团队提出，斩获当年 ImageNet 竞赛中分类任务的第一名，发表在 *Going Deeper with Convolutions* 这篇论文中。GoogLeNet 提出了 Inception 模块，它采用 4 条路径使用不同大小卷积核的卷积，从不同层面抽取信息，不改变输入的高宽，然后在输出通道维度拼接，如图 2.4a 所示。GoogLeNet 大量地使用 1×1 的卷积核进行降维以及映射处理，并且丢弃了全连接层，使用平均池化层，相比 VGG 网络大大减少模型参数。

Inception-v3 利用卷积分解改变了 GoogLeNet 中的 Inception 模块结构，提出了 3 种新的 Inception 模块，如图 2.4(b)(c)(d) 所示，Inception-v3 中的卷积分解方式如下：

- 将大卷积核用多层堆叠的小卷积核代替，例如将一层 5×5 的卷积替换成两层 3×3 的卷积，在减少参数量的同时保持相同的感受野；
- 将 $N\times N$ 的卷积分解成 $1\times N$ 和 $N\times1$ 的非对称卷积。

上述两种分解卷积的方法可以使得参数量更少，计算量减少，非线性增加。具体来说，Inception-v3 得到 3 种 Inception 模块的过程如下：

(1) 将图 2.4a 中的 5×5 卷积分解成两个堆叠的 3×3 卷积得到 Inception 模块 A，如图 2.4b 所示；

(2) 先将图 2.4a 中的 5×5 卷积分解成两组 $1\times n$ 和 $n\times1$ 的非对称卷积，再将 3×3 卷积分解成堆叠的 $1\times n$ 和 $n\times1$ 的非对称卷积得到 Inception 模块 B(在实现过程中 $n=7$)，如图 2.4c 所示；

(3) 先将图 2.4a 中的 5×5 卷积分解成一个 3×3 的卷积和一组并行的 1×3 和 3×1 的非对称卷积，再将 3×3 卷积分解成一组并行的 1×3 和 3×1 的非对称卷积得到 Inception 模块 C，如图 2.4d 所示。

其中，前两种模块以堆叠的方式分解卷积，第三种以并行的方式分解来促进高维度的表征。

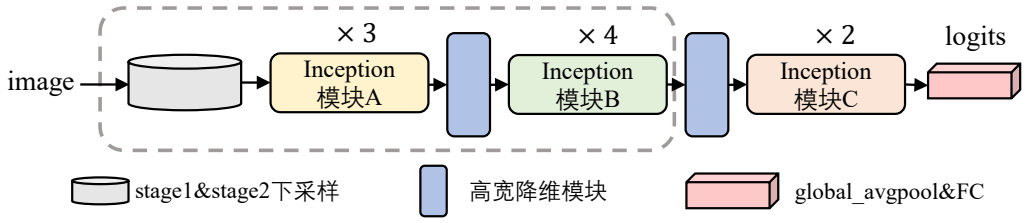


图 2.5 Inception-v3 网络的简化结构图

图 2.5 为 Inception-v3 的总体网络框架，其中 stage1&stage2 下采样包含了步长

为 2 的一次 7×7 卷积层和两次 3×3 最大池化层的下采样，高宽降维模块表示 Inception-v3 中提出的高效的特征图尺寸 (指高和宽) 降维模块，global_avgPool&FC 表示全局平均池化层和全连接层。

在本工作中，采用的 Inception-v3 骨干网络结构只包括图 2.5 中的虚线框部分，依旧加载在 ImageNet 上预训练的权重进行迁移学习。图片输入到该骨干网络后，将图 4 中最后一个 Inception 模块 A 输出的特征图和最后一个 Inception 模块 B 输出的特征图，通过线性插值法在通道维度上拼接作为输入图片最终的特征图，其中后者作为全局场景上下文特征图，在第 3.2.1 节中会详细介绍该过程。

2.2 图神经网络

普通卷积神经网络 (CNN) 研究的对象是具备欧氏空间 (Euclidean domains) 的网格数据 (如图片，语音等)，欧氏空间数据最显著的特征是具有规则的空间结构。CNN 卷积操作配合池化操作在结构规则的图像等数据上效果显著，但对于不规则的图数据对象，普通卷积网络的效果不尽人意，因为难以选取固定的卷积核来适应整个图的不规则性，如邻居节点数量和节点顺序的不确定。所以图神经网络 (GNN) 应运而生，它将卷积操作泛化到图 (graph) 数据结构，并取得了显著的效果。

下面首先介绍普通的图卷积神经网络，然后在此基础上阐述动态图卷积神经网络的方法。

2.2.1 图卷积神经网络

图神经网络可以分为两类，一类是空间域 (spatial domain) 或顶点域 (vertex domain) 方法，例如 GraphSAGE[52]和 GAT[53]；另一类是谱域 (spectral domain) 方法，例如 ChebNet[54]和 GCN[36]。通俗点解释，空域方法可以类比到直接在图片的像素点上进行卷积，而谱域方法可以类比到对图片进行傅里叶变换后，再进行卷积。

谱图理论将卷积操作从基于网格的数据推广到图结构数据，它将图转化为代数形式来分析图的拓扑属性，如图结构中的连通性。在谱图分析理论里面，一个图被表示为其对应的拉普拉斯矩阵 (Laplacian matrix)，图结构的属性可以通过分析 Laplacian 矩阵及其特征值得到。

给定一个图 G，G 的节点个数为 N，它的 Laplacian 矩阵被定义为 $L = D - A$，归一化形式为 $L = I_N - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} \in \mathbb{R}^{N\times N}$，这里 A 是邻接矩阵，I_N 代表单位矩阵，$D \in \mathbb{R}^{N\times N}$ 为度矩阵，D 是一个对角矩阵由节点的度组成，并且 $D_{ii} = \sum_j A_{ij}$。Laplacian 矩阵的特征分解为 $L = U\Lambda U^T$，这里 $\Lambda = diag([\lambda_0, \ldots, \lambda_{N-1}]) \in \mathbb{R}^{N\times N}$ 是

特征值矩阵并且为对角阵，U 为 Fourier 基 (Fourier basis)。考虑图 G 上对应于每个顶点的信号 $x \in \mathbb{R}^N$，信号在图上的 Fourier 变换被定义为 $\hat{x} = U^T x$。根据 Laplacian 矩阵的性质，U 是一个正交矩阵，所以对应的 Fourier 逆变换为 $x = U\hat{x}$。图卷积是通过使用在傅里叶域中对角化的线性算子实现的卷积运算，它取代了经典卷积算子[55]。由此，图 G 上的信号 x 被一个卷积核 g_θ 滤波的操作如下：

$$g_\theta *_G x = g_\theta(L)x = g_\theta(U\Lambda U^T)x = Ug_\theta(\Lambda)U^T x, \tag{2.2}$$

这里 $*_G$ 代表图卷积算子。因为图信号的卷积操作等于通过图 Fourier 变换转换到谱域的信号的乘积，上式可以被理解为 g_θ 和 x 分别经过 Fourier 变换到频域，然后将变换后的结果相乘，最后做 Fourier 逆变换得到卷积操作的结果。然而，当图的规模很大时，给 Laplacian 矩阵直接进行特征分解需要花费很大的代价。因此，Defferrard 等人[54]提出 ChebNet，定义特征向量对角矩阵的 Chebyshev 多项式为滤波器：

$$g_\theta = g_\theta(\Lambda) \approx \sum_{k=0}^{K-1} \theta_k T_k(\tilde{\Lambda}) \tag{2.3}$$

其中，$\tilde{\Lambda} = 2\Lambda/\lambda_{max} - I_N$ 为缩放后的特征向量矩阵，λ_{max} 是 Laplacian 矩阵最大的特征值，$\theta \in \mathbb{R}^K$ 为多项式的系数向量。Chebyshev 多项式的递归定义为：$T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$，其中 $T_0(x) = 1, T_1(x) = x$。回到对信号 x 与滤波器 g_θ 卷积的定义，现在有：

$$g_\theta *_G x = g_\theta(L)x = \sum_{k=0}^{K-1} \theta_k T_k(\tilde{L})x, \tag{2.4}$$

其中，$\tilde{L} = 2L/\lambda_{max} - I_N$。上式利用 Chebyshev 多项式拟合卷积核的方法，避免计算图 Fourier 的基，来降低计算复杂度。随后，Kipf 等人[36]引入了一种一阶近似 ChebNet(1stChebNet)，假设 $K = 1, \lambda_{max} = 2$，则 ChebNet 卷积公式简化近似为：

$$g_\theta *_G x = \theta_0 x - \theta_1 D^{-\frac{1}{2}} A D^{-\frac{1}{2}} x, \tag{2.5}$$

为了抑制参数数量防止过拟合，1stChebNet 假设 $\theta = \theta_0 = -\theta_1$，图卷积的定义就近似为：

$$g_\theta *_G x = \theta(I_N + D^{-\frac{1}{2}} A D^{-\frac{1}{2}})x, \tag{2.6}$$

为了融合多维图输入信号，给定具有 C 个输入通道的信号 $X \in \mathbb{R}^{N \times C}$，对上式进行修正提出了图卷积公式如下：

$$Z = \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} XW, \tag{2.7}$$

其中，$\tilde{A} = A + I_N, \tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$，即图中加上自环；$W \in \mathbb{R}^{C\times F}$ 是一个滤波器参数矩阵，$Z \in \mathbb{R}^{N\times F}$ 是图卷积层的输出矩阵，上式即为图卷积网络 (GCN) 的一般形式。从上式可以看出，GCN 计算的复杂度大大降低，不需要再计算 Laplacian 矩阵的特征分解，并且具有权值共享，局部连接，感受野正比于卷积层层数等优点。

2.2.2 动态图卷积神经网络

在 2.2.1 节介绍的图卷积神经网络中，所用到的邻接矩阵 A 是预定义的和静态的，即根据图结构的先验知识 (prior knowledge) 得来。例如，在交通道路网络中，每个道路的传感器可以当作图中的一个节点，对应的邻接矩阵根据传感器之间的欧氏距离得到；在文献引用网络中，每一篇文献可以当作一个节点，可以根据文献之间是否引用导出邻接矩阵。上述这种邻接矩阵在模型训练的过程中不会改变，并且依赖于数据集的先验知识。然而，在群体行为识别的视频数据里面，除了个体动作标签和整个片段的行为标签是已知的，没有其他任何先验知识。所以，需要动态地计算邻接矩阵，在构建的时空图上利用动态图卷积操作进行关系推理。

在公式 2.7 中的邻接矩阵 A 是预定义的，为了构造自适应的邻接矩阵，Wu 等人[56]在 Graph WaveNet 中用可学习的参数随机初始化两个节点嵌入字典 $E_1, E_2 \in \mathbb{R}^{N\times c}$，他们将自适应的邻接矩阵表示如下：

$$\tilde{A}_{adp} = \text{softmax}(\text{ReLU}(E_1 E_2^T)). \tag{2.8}$$

其中，E_1 被称为源节点嵌入 E_2 被成为目标节点嵌入，通过 E_1 和 E_2 相乘可以得到源节点和目标节点之间空间依赖的权重。使用 ReLU 激活函数可以消除一些弱连接，使用 softmax 函数归一化自适应的邻接矩阵。根据 Li 等人[57]提出的扩散卷积模型，$\tilde{A}_{adp}$ 被当作隐藏扩散过程中的转移矩阵，Graph WaveNet[56]定义如下的图卷积层：

$$Z = \sum_{k=0}^{K} \tilde{A}_{apt}^k X W_k, \tag{2.9}$$

这里，$X \in \mathbb{R}^{N\times C}$ 表示输入信号，$Z \in \mathbb{R}^{N\times M}$ 表示卷积层的输出，$W \in \mathbb{R}^{C\times M}$ 表示参数矩阵，上式只使用自适应的邻接矩阵建模空间关系，而不是依赖于预定义的邻接矩阵。

如图 2.6 所示，Guo 等人[58]在 ASTGCN 中将空间注意力机制与 ChebNet[54]相结合来动态地调整节点之间的空间相关性。在图 2.6 中，给定某一时间步 t 的图信号 $X \in \mathbb{R}^{N\times d}$，按如下公式计算空间关系权重矩阵 S，

$$Q = XW_q, K = XW_k, S = \text{softmax}(\frac{QK^T}{\sqrt{d}}) \in \mathbb{R}^{N\times N}, \tag{2.10}$$

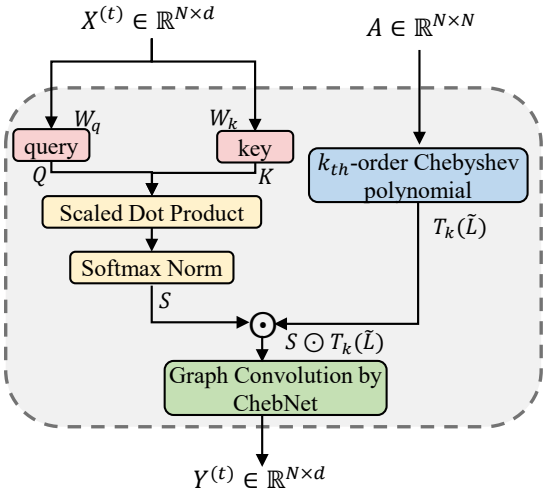


图 2.6 原始的动态图卷积实现过程示意图

其中，W_q, W_k 都是线性变换参数矩阵。在 ASTGCN[58]中，利用 element-wise 点积 (亦称 Hadamard 乘积) 将 S 与公式 2.4 中的 $T_k(\tilde{L})$ 结合得到 $T_k(\tilde{L}) \odot S$，这里 $\odot$ 代表 Hadamard 乘积。因此公式 2.4 中的图卷积操作变为：

$$g_\theta *_G x = g_\theta(L)x = \sum_{k=0}^{K-1} \theta_k (T_k(\tilde{L}) \odot S)x. \tag{2.11}$$

随后又在 ASTGNN[59]中，将空间注意力机制与 GCN[36]相结合，即利用空间关系权重矩阵 S 通过 Hadamard 乘积来调整静态权重矩阵 A，公式 2.7 变化为：

$$Z = (\tilde{A} \odot S)XW. \tag{2.12}$$

公式 2.11 和 2.12 都是用学习的注意力关系矩阵来调整静态邻接矩阵信息，从而进行动态图卷积。在本研究工作中，仅仅使用空间注意力关系矩阵，并且使其经过一些变换去表达跨时间步的个体之间的关系。

2.3 Transformer

原始的 Transformer[37]最开始被用于序列到序列的自回归 (auto-regressive) 任务。与以前的序列直推模型 (transduction model) 相比，Transformer 继承了 encoder-decoder 结构，但是通过使用多头注意力机制和 point-wise 前馈网络 (feed-forward networks, FFN)，完全摒弃了循环和卷积单元[27,60]。在下面的小节，将会描述 Transformer 的关键部分以及总体框架。

2.3.1 注意力机制

作为 Transformer 的基本组成部分，注意力机制 (attention mechanism) 可以被划分为下面两个部分。

2.3.1.1 线性转换层

线性转换层将输入序列 $X \in \mathbb{R}^{n_x \times d_x}, Y \in \mathbb{R}^{n_y \times d_y}$ 转换成三个不同的序列向量 (query Q, key K 和 value V)，这里 n 和 d 分别代表输入序列的长度和特征维度。Q, K, V 通过如下方式得到：

$$Q = XW^Q, K = YW^K, V = YW^V, \tag{2.13}$$

这里 $W^Q \in \mathbb{R}^{d_x \times d^k}, W^K \in \mathbb{R}^{d_y \times d^k}, W^V \in \mathbb{R}^{d_y \times d^v}$ 都是线性矩阵，d^k 是 query 和 key 的维度，d^v 是 value 的维度。query 从 X 投影得来，同时 key 和 value 从 Y 投影而来，这种两个序列的输入方案被称为交叉注意力机制 (cross-attention mechanism)。特别地，当 $Y = X$ 的时候，注意力机制可以被当作 self-attention。此外，self-attention 被同时用在 encoder 和 decoder 里面，然而 cross-attention 作为一个纽带仅被用在 decoder 里面。

2.3.1.2 注意力层

注意力层显式地将 query 与相应的 key 聚合在一起，将它们分配给 value，并更新输出向量。该过程可以被如下公式描述：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{2.14}$$

这里注意力权重通过 query 和 key 之间的点积 (dot-product) 操作产生，缩放因子 d_k 和 softmax 操作将注意力权重转变成归一化的分布。得到的注意力权重被分配给 value 中对应的元素，从而产生最终的输出向量。

2.3.2 多头注意力机制

由于特征子空间的限制，单头注意力模块的建模能力是粗糙的。为了解决这个问题，Vaswani 等人[37]提出了一个多头自注意力机制 (multi-head self-attention, MHSA)，它将输入线性地投影到多个特征子空间里面，并且通过一些独立的注意力头并行地处理它们。每个注意力头得到的结果向量被拼接在一起，并且被映射为最终的输出。

MHSA 的过程可以被描述如下：

$$\begin{aligned} Q_i = XW^{Q_i}, K_i = XW^{K_i}, V_i = XW^{V_i}, \\ Z_i = \text{Attention}(Q_i, K_i, V_i), i = 1 \ldots h, \\ \text{MultiHead}(Q, K, V) = \text{Concat}(Z_1, Z_2, \ldots, Z_h)W^O, \end{aligned} \tag{2.15}$$

这里 h 为注意力头的数量，$W^O \in \mathbb{R}^{hd_v \times d_{model}}$ 表示输出投影矩阵，Z_i 表示每个头的输出向量，$W^{Q_i} \in \mathbb{R}^{d_{model} \times d_k}, W^{K_i} \in \mathbb{R}^{d_{model} \times d_k}, W^{V_i} \in \mathbb{R}^{d_{model} \times d_v}$ 是三组不同的线性矩阵。

与卷积的稀疏连接相似，多头注意力将输入分割成 h 个独立的注意力头，每个注意力头具有 d_{model}/h 维的向量，然后并行地融合每个头的特征。多头注意力在没有额外的计算代价情况下，丰富了特征子空间的多样性。

2.3.3 FFN 和 Positional Encoding

MHSA 的输出随后被送入两层连续的前馈网络 (feed-forward networks, FFN)，并且经过 ReLU 激活函数，该过程表示如下：

$$\text{FFN} = \text{ReLU}(W_1 x + b_1)W_2 + b_2. \tag{2.16}$$

该 position-wise 前馈层可以被看作一个 point-wise 卷积，它同等地对待每一个位置，但是在每层间使用不同的参数。

因为 Transformer/注意力机制同时且同等地在输入的嵌入上操作，所以忽视了序列的顺序。为了充分利用序列的信息，一种普通的做法是在输入后面追加一个额外的位置 (positional) 向量，因此该过程称为位置编码 (positional encoding)。有很多种 positional encoding 的方法，一种典型的选择是使用不同频率的 sine 和 cosine 函数，表示如下：

$$PE_{(pos,i)} = \begin{cases} \sin(pos \cdot w_k) & \text{if } i = 2k \\ \cos(pos \cdot w_k) & \text{if } i = 2k+1, \end{cases} \quad w_k = \frac{1}{10000^{2k/d}}, k = 1, \ldots, d/2, \tag{2.17}$$

这里 i 和 d 分别表示向量的索引和长度，pos 表示每个元素在序列里面的位置。

2.3.4 Transformer 架构

Transformer 模型总体为 encoder-decoder 架构。特别地，Transformer 由 $N = 6$ 个连续的 encoder 块组成，每个 encoder 由两个子层构成。一个是 MHSA 子层，它

在 encoder 的嵌入里面聚合关系；另一个是 position-wise FFN 子层，用来提取特征的表征。至于 decoder，它也是由 6 个连续的块组成，并跟随在堆叠的 encoder 后面。相比于 encoder，每个 decoder 块追加了一个多头的交互注意力层用来聚合 decoder 的嵌入和 encoder 的输出，在公式 2.13 中，Y 对应前者，X 对应了后者。此外，在 encoder 和 decoder 中的所有子层都采用了残差连接 (residual connection)[47]和 Layer Normalization[61]，用来增强 Transformer 的可扩展性。为了加入序列中的位置信息，在堆叠 encoder 和 decoder 的开始每个输入嵌入都附加了 positional encoding。最后，一个线性层和 softmax 操作被用来下个单词的预测。

作为一个自回归的语言模型，Transformer 最开始起源于机器翻译任务。在本研究工作中，使用了 Transformer 的 encoder 模块将场景上下文信息编码到个体的外观或姿态特征中，利用 MHSA 机制学习个体间的时空关系。

2.4 本章小节

本章介绍了一些与群体行为识别相关的理论技术，将会在后面的章节里面引用到。首先介绍了本研究中涉及到的提取图片特征的骨干网络，即 ResNet-18，VGG-19 和 Inception-v3，并且给出了每种骨干网络在迁移学习时的结构。随后，介绍了图神经网络的相关理论，包括 ChebyNet、GCN 中图卷积操作的具体形式，然后将其推广到动态的图卷积神经网络，并给出了几种具体方法。最后，详细介绍了 Transformer 模型的具体组件和总体框架，包括注意力机制，FFN 和 Positional Encoding。

3 基于跨时间步动态图卷积的群体行为识别

本章首先叙述了基于跨时间步动态图卷积的群体行为识别模型提出的动机，然后描述了该模型的总体设计框架，接着对模型的一些重要组成模块进行了详细的说明，最后通过系统的实验验证了本章算法的有效性和合理性。

3.1 引言

在前面的章节中，回顾了群体行为识别的相关理论技术，并介绍了两种输入模态的方法，即只有 RGB 图片输入的方法和 RGB 图片与对应的关节点同时作为输入的方法，本章将介绍第一种方法。

群体行为识别的关键点在于充分挖掘行为者在场景中的上下文信息以及学习他们之间的动态时空关系，推理出他们之间关系的重要性得到群体表征进行预测。之前的一些工作直接在个体的外观特征上进行关系推理，忽视了对全局场景上下文信息的提取，这样会丢失一些有用的线索，如 ARG[39]，DIN[8]，ActorFormer[6]等方法。例如，在排球比赛中，周围的观众会对赛场上球员的不同表现做出不同的反应，如在关键防御和关键得分时周围的观众可能会站立加油。类似观众的表现这种场景上下文信息，可能会包含赛场运动员群体行为的一些有用线索。所以，在建模行为者的时空关系之前，需要充分探索场景中相关的上下文信息。于是，本文通过场景上下文编码模块，使用 Transformer 中的 encoder 从全局场景特征图中通过自注意力机制学习相关性高的场景信息，然后将其编码进个体的外观特征中，使得个体的外观特征与所在的场景关联从而得到增强，随后将进一步地探索场景上下文编码过的个体特征间的交互关系。

编码了行为者的场景上下文信息后，就需要学习行为者之间空间和时间的交互关系。然而，之前有的方法采用异步的方式分别地建模空间和时间关系，如 Actor-Former[6]；还有的方法没有跨时间步地考虑空间关系，个体关系特征的感受野太小；还有的方法直接利用空间或时间注意力机制中的点积操作计算稠密的个体关系，忽视了空间或时间关系稀疏性的事实，引起一些不相关信息的干扰。为了能够全面地考虑所有时间步的个体相互之间的关系，构建了一个跨时间步的时空图，然后利用动态图卷积同时学习所有时间步个体之间的时空关系。这样，每一个行为者不仅可以关注到同一个时间步内个体的空间关系，而且可以关注到跨时间步的个体的时序动态，进而可以得到具有丰富语义和信息的个体时空特征表示，最终得到更加稳定

的群体表征。

值得注意的是，本章介绍的基于跨时间步动态图卷积的群体行为识别方法只使用了 RGB 图像作为输入，没有使用额外的其他模态输入，如图片的光流 (optical flow) 特征和个体的关节点信息。此外，本章方法仅仅使用 2D 卷积骨干网络提取特征，如 VGG[49]，ResNet[47]和 Inception[50]网络，而不是利用 3D 卷积骨干网络去提取特征，例如 I3D[42]网络。

在本章中首先介绍了网络的总体框架以及损失函数，接着描述了 RoIAlign[38]提取个体行为者特征的过程，然后详细阐述了模型中的场景上下文编码模块，跨时间步动态图卷积模块和特征池化模块。最后在两个广泛使用的公开数据集上进行实验对比分析，实验结果验证了本章算法的有效性，提出的跨时间步动态图卷积模块可以有效地学习个体间的时空交互关系。

3.2 网络总体设计

3.2.1 基于个体特征的网络框架

在本章中的基于跨时间步动态图卷积的群体行为识别模型总体架构如图 3.1 所示，该模型将 RGB 视频片段作为输入。首先以视频片段里面标注的帧 (frame) 为中心选取 T 帧，表示为 $X_{\text{img}} \in \mathbb{R}^{T\times H\times W\times 3}$，其中 H, W 分别表示输入帧的高和宽，3 代表颜色通道的数量。

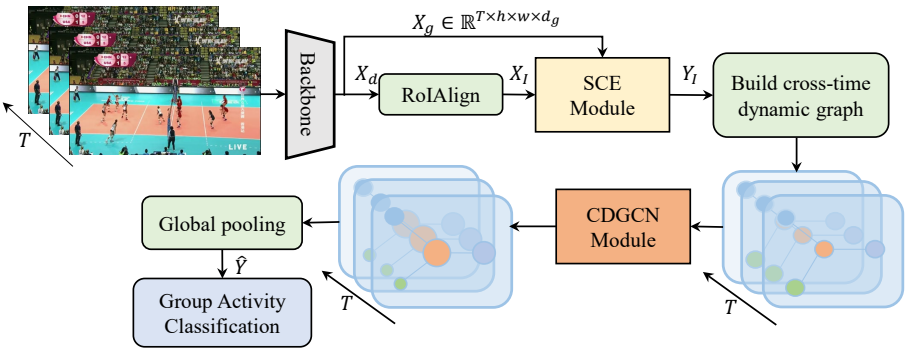


图 3.1 本章模型的总体框架

对于 2.1 小节提到的 3 种骨干网络 (backbone)，本章实验主要以 Inception-v3 为 backbone 进行，来验证网络中模块的有效性以及与之前的方法形成公平的对比。在实践中，将 X_{img} 输入到 Inception-v3，从 Inception-v3 的最后一个卷积层提取特征图然后改变其大小得到 $X_g \in \mathbb{R}^{T\times h\times w\times d_g}$，$X_g$ 可以被看作整个视频片段的场景上下文特

征。此外，从 Inception-v3 中间层 (Mixed_5d) 的输出产生更高分辨率的特征图 X'_d，将 X'_d 与 X_g 通过线性插值法在通道维度拼接得到视觉特征图 X_d。然后，在 X_d 上利用 RoIAlign 提取每个个体的特征，将这些特征与对应个体的边界框对齐。最后，一个全连接层被用来将对齐的每个个体特征嵌入到一个 D 维的空间，这些个体特征堆叠在一起形成 $X_I \in \mathbb{R}^{T\times N\times D}$，其中 T 表示输入的时间步 (即时序维度)，N 表示视频每一帧中标注的个体数量 (即空间维度)。

随后，将得到的场景特征 X_g 和个体特征 X_I 作为场景上下文编码模块 (SCE Module) 的输入，输出的 $Y_I \in \mathbb{R}^{T\times N\times D}$ 表示场景上下文编码过的个体特征，注意 Y_I 的空间维度由所有个体的边界框坐标排列，这 $T\times N$ 个个体形成了一个时空图 (spatial-temporal graph, ST graph)。接着，提出的跨时间步的动态图卷积模块 (CDGCN Module) 在构建的 ST graph 进行个体之间的时空关系推理，经过全局池化 (Global pooling) 模块得到最终的群体表征 $\tilde{Y}$，最后基于 $\tilde{Y}$ 进行群体行为的预测。其中，全局池化模块包括在空间维度上的最大池化，和在时间维度上的平均池化。

3.2.2 损失函数

本章的模型以端到端的方式进行训练。在该模型中，通过最终的群体表征可以得到群体行为预测的概率分布 $\tilde{y}_g$；同样地，另外一个分类器使用个体表征预测个体动作的概率分布 $\tilde{y}_a$。对于这两个任务，选择交叉熵损失函数来优化网络参数，如下所示：

$$\mathcal{L} = \mathcal{L}_1(y_g, \tilde{y}_g) + \lambda\mathcal{L}_2(y_a, \tilde{y}_a) \tag{3.1}$$

公式 3.1 中 $\mathcal{L}_1$ 和 $\mathcal{L}_2$ 分别是群体行为识别和个体动作识别的交叉熵损失；y_g 和 y_a 分别是群体行为和个体动作的真实标签。λ 是一个超参数，用来平衡上述两项损失。

3.3 RoIAlign 模块

给定一帧视频图片中个体的边界框和该帧图片对应的特征图 X_d，RoIAlign 模块在 X_d 上通过边界框对齐提取每个个体的特征图，该过程如图 3.2 所示。在图 3.2 的视频图片中给定了某个行为者的归一化边界框坐标 (x_1, y_1, x_2, y_2)，该行为者在特征图 X_d 上的边界框坐标可以如下计算：

$$x'_1 = x_1 * OW,\ y'_1 = y_1 * OH,\ x'_2 = x_2 * OW,\ y'_2 = y_2 * OH, \tag{3.2}$$

这里 OH 和 OW 分别是特征图 X_d 的高和宽，得到的边界框坐标为 (x'_1, y'_1, x'_2, y'_2)。随后，RoIAlign 模块根据边界框坐标裁剪出该个体对应的特征图，然后重新调整其

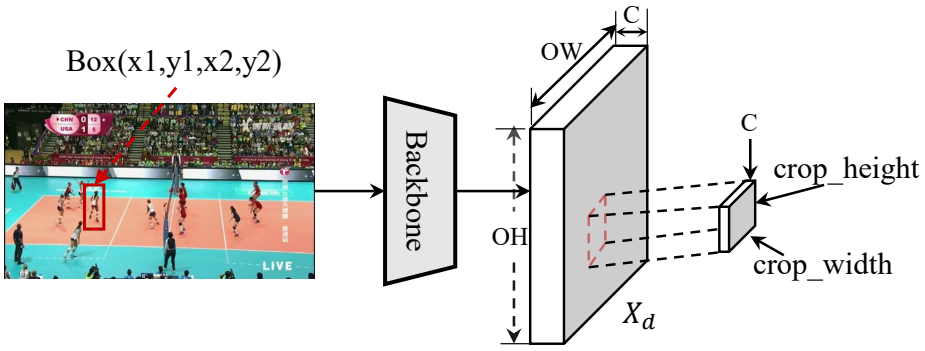


图 3.2 RoIAlign 模块实现原理示意图

大小为 (crop_height, crop_weight, C)，其中 crop_height 和 crop_width 分别为提取出的个体特征图的高和宽，C 为通道数量并且在前后保持不变。上述 RoIAlign 模块提取个体特征的过程可以总结为两个阶段，即裁剪和调整大小 (crop and resize)。

3.4 场景上下文编码

为了将个体特征与所在的场景上下文联系起来，使用场景上下文编码 (Scene Context Encoding, SCE) 模块将场景上下文信息编码到个体的特征中，得到输出 Y_I。SCE 模块的输入为不同时间步堆叠的个体特征 $X_I \in \mathbb{R}^{T\times N\times D}$ 和全局场景特征图 $X_g \in \mathbb{R}^{T\times h\times w\times d_g}$，这里将时序维度 T 看作是 batch 维度。为了简便，后文将以第 t 帧的个体特征 $X_I^{(t)} \in \mathbb{R}^{N\times D}$ 和全局场景特征图 $X_g^{(t)} \in \mathbb{R}^{h\times w\times d_g}$ 为例，详细介绍 SCE 模块生成上下文信息编码过的个体特征 $Y_I^{(t)} \in \mathbb{R}^{N\times D}$ 的过程。

3.4.1 SCE 模块

SCE 模块的架构如图 3.3 所示，它被分为两个部分：注意力聚合 (attention aggregation) 和前馈网络 (Feed-Forward Network, FFN) 嵌入 (FFN embedding)，其中 attention 表示注意力分数计算，softmax 表示归一化，⊗ 表示矩阵乘法，add & LN 表示残差连接 (residual connection)[47]和 Layer Normalization(LayerNorm)[61]，dropout 表示随机丢弃神经元。原始的 Transformer 中的 encoder 学习一组 query，key 和 value，通过 attention 机制计算 query 和 key 之间的注意力权重分数，然后通过 value 的加权求和得到输出。具体来说，在执行注意力聚合部分之前，分别对全局场景特征图 $X_g^{(t)}$ 和个体特征 $X_I^{(t)}$ 进行位置信息编码，位置编码的过程将会在 3.4.2 小节详细阐述，同时在该过程中将 $X_g^{(t)}$ 的空间维度折叠起来得到 $\tilde{X}_g^{(t)} \in \mathbb{R}^{hw\times d_g}$，因为 encoder 的输入要求是一个序列。

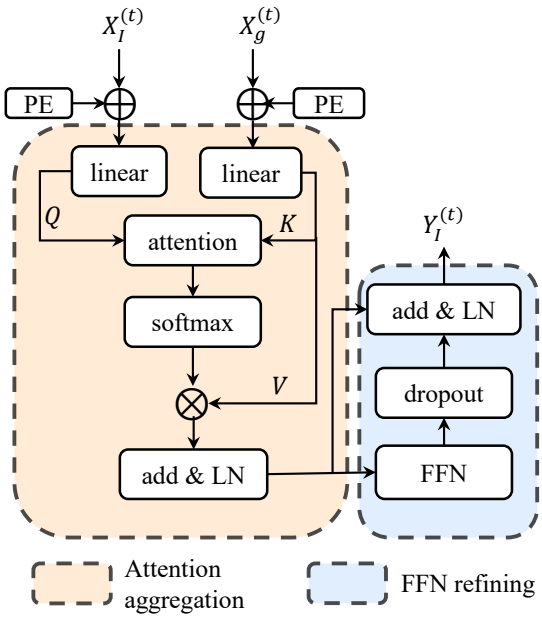


图 3.3 SCE 模块的结构图

在 SCE 模块的注意力聚合部分中，query 对应个体特征的线性投影，key 和 value 对应全局场景特征图的线性投影。为了得到场景上下文编码后的表征，执行缩放的点积注意力分数 (scaled dot-product attention) 计算，随后使用 softmax 函数归一化，加权求和，再进行残差连接和 LayerNorm，该过程可以表示如下：

$$\begin{aligned} Q^{(t)} &= X_I^{(t)} W_{tq}, K^{(t)} = X_g^{(t)} W_{tk}, V^{(t)} = X_g^{(t)} W_{tv}, \\ V'^{(t)} &= \text{softmax}(\frac{Q^{(t)} K^{(t)^T}}{\sqrt{D}}) V^{(t)}, \\ X_I'^{(t)} &= \text{LN}\left(X_I^{(t)} + \text{dropout}\left(V'^{(t)}\right)\right), \end{aligned} \tag{3.3}$$

这里，$W_{tq}, W_{tk}, W_{tv} \in \mathbb{R}^{D \times D}$ 都是可学习的参数矩阵，LN 代表 LayerNorm 层，$X_I'^{(t)} \in \mathbb{R}^{N \times D}$ 是包含场景上下文信息的个体特征。

随后，在 SCE 模块中执行 FFN 嵌入来进一步细化已编码的表征。这一步包含一个前馈特征上的 LayerNorm 层和一个残差连接，该部分可以被表述为：

$$\begin{aligned} \text{FFN}(X_I'^{(t)}) &= \text{linear}_2(\text{dropout}(\text{ReLU}(\text{linear}_1(X_I'^{(t)})))), \\ X_I''^{(t)} &= \text{LN}(X_I'^{(t)} + \text{dropout}(\text{FFN}(X_I'^{(t)}))), \end{aligned} \tag{3.4}$$

上式中 $\text{linear}_1, \text{linear}_2$ 代表两个不同的线性层，$X_I''^{(t)}$ 是 FFN 嵌入后的包含上下文的特征。

SCE 模块可以被很容易地扩展到多头的 (multi-head) 形式，为了增强该模块的表达能力，使用拼接作为多个头特征的融合方式，并且加上残差连接。综上所述，SCE 模块的输出 $Y_I^{(t)} \in \mathbb{R}^{N\times D}$ 可以被公式表示如下：

$$Y_I^{(t)} = X_I^{(t)} \| \{ \|_{i=1}^{h_s} \mathrm{sce}_i(X_I^{(t)}, X_g^{(t)}) \}, \tag{3.5}$$

其中 $\|$ 表示通道维度上的拼接操作，$\mathrm{sce}_i(X_I^{(t)}, X_g^{(t)})$ 表示 SCE 模块的第 i 个头，h_s 表示 SCE 模块注意力头的数量，这里每个注意力头都有独立的参数。

3.4.2 位置编码

在场景中个体的位置信息是重要的,因为个体之间的相对位置 (如个体的前驱和后继) 可以给时空关系的推理带来帮助，因此在 SCE 模块中加入位置编码 (positional encoding, PE)。在实际中，对于一个给定的个体特征，使用 sine 和 cosine 函数去编码该个体原始的边界框中心坐标 (l_w, l_h)。得到的位置特征中一半的维度用 l_w 来编码，另一半用 l_h 来编码，该过程可以如下表示：

$$\begin{aligned} \mathrm{PE}_{(l,2k)} &= \sin\left(\frac{l}{10000^{2k/D}}\right), \\ \mathrm{PE}_{(l,2k+1)} &= \cos\left(\frac{l}{10000^{2k/D}}\right) \end{aligned} \tag{3.6}$$

这里 l 表示坐标 l_w 或者 l_h，如果 $l = l_w$，则维度 $k \in \{0, 1, \ldots, \frac{D}{4} - 1\}$；如果 $l = l_h$，则维度 $k \in \{\frac{D}{4}, \frac{D}{4} + 1, \ldots, \frac{D}{2} - 1\}$。

对于全局场景特征图 $X_g^{(t)}$，使用公式 3.6 中的方法编码特征图上的坐标，并且乘以对应的下采样因子 (如对于 Inception-v3 为 16)，乘以下采样因子是为了与个体特征的位置编码放大阶数相匹配。

3.5 跨时间步动态图卷积

在群体行为识别任务中需要捕获个体之间动态的空间关系，这种空间关系不仅存在于同一个时间步的个体之间，而且存在于跨时间步的个体之间，我们将前者称之为 intra-spatial 关系，后者称之为 inter-spatial 关系。然而，有些方法分别在每个时间步中使用空间注意力来捕捉 intra-spatial 关系，却没有考虑到学习跨时间步的 inter-spatial 关系，如 GroupFormer[44]中的 encoder 学习空间上下文的部分。因此，在建模空间关系时，需要同时考虑同一时间步和跨时间步的空间关系。值得注意的是，在跨时间步的 inter-spatial 关系中包含了同一个体的时序动态信息，所以个体之间的空间和时间关系是同步学习的。

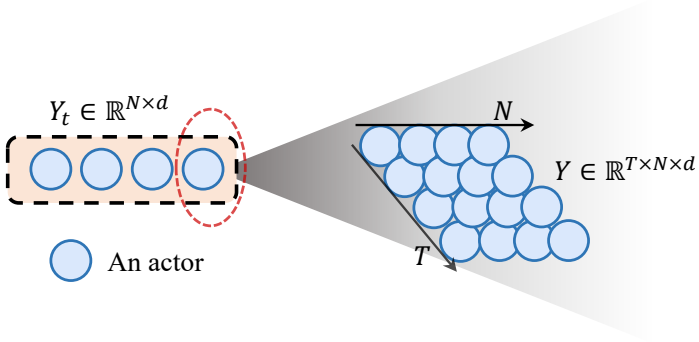


图 3.4 捕捉 intra-spatial 关系和 inter-spatial 关系的示意图

如图 3.4 所示，对于在时间步 t 的个体特征 $Y_t \in \mathbb{R}^{N\times D}$, Y_t 来自于 SCE 模块的输出 $Y_I \in \mathbb{R}^{T\times N\times D}$ 的第 t 个时间步，其中的一个行为者 (如图 3.4 虚线圆圈划出来的 actor) 的空间关系来自于时空图中所有的 $T \times N$ 个个体，同时包括了 intra-spatial 关系和 inter-spatial 关系。下面首先介绍动态图卷积网络 (dynamic graph convolution network, DGCN) 提取空间关系的过程，接着阐述跨时间步动态图卷积网络 (cross-time dynamic graph convolution network, CDGCN) 的框架。

有些方法，例如 GroupFormer[44]中的 encoder 和 Dual-AI[45]中的 S-Trans 部分，在每一个时间步利用空间注意力机制中的点积操作计算稠密的空间关系矩阵，如公式 2.14 所示，却忽视了空间关系的稀疏性质。因此，通过一个门控机制来稀疏化从点积操作得到的稠密空间关系矩阵，并加上一个单位矩阵使得个体的自表达能力增强，对于在时间步 t 的所有个体特征 $Y_t \in \mathbb{R}^{N\times D}$，该过程公式化如下：

$$\begin{aligned} Q = Y_t W^Q, K = Y_t W^K, V = Y_t W^V, \\ \tilde{A} = \mathrm{ReLU}\left(QK^T\right) + I_N, \end{aligned} \tag{3.7}$$

这里 $W^Q, W^K, W^V \in \mathbb{R}^{D\times D}$ 均为可学习的参数矩阵，I_N 为单位矩阵，ReLU(·) 为非线性激活函数，它可以对空间关系起到稀疏门控作用。

此外，由于自注意力机制中的指数操作会导致梯度消失或者爆炸，所以取消对空间关系矩阵 $\tilde{A}$ 的 softmax 归一化操作，而是通过 $\tilde{A}$ 除以它的度矩阵 $\tilde{\mathbf{D}} \in \mathbb{R}^{N\times N}$ 进行归一化。综上所述，动态图卷积网络提取空间关系的过程可以描述如下：

$$\mathrm{DGCN}(Q, K, V) = \tilde{\mathbf{D}}^{-1}\tilde{A}V. \tag{3.8}$$

该 DGCN 模块可以被进一步地扩展为多头的方式，将其称之为 multi-head DGCN(MHDGCN)，从而可以从不同的子空间关注不同的信息。MHDGCN 的过程

如下：

$$\text{MHDGCN}(Q,K,V) = \|_{i=1}^{h_d} \text{dgcn}_i(Q,K,V)W^O, \tag{3.9}$$

这里 $W^O \in \mathbb{R}^{D\times D}$ 为输出投影的参数，$\|$ 表示通道维度上的拼接操作，$\text{dgcn}_i(Q,K,V)$ 表示 MHDGCN 模块的第 i 个头，h_d 表示 MHDGCN 模块注意力头的数量，这里每个注意力头都有独立的参数。

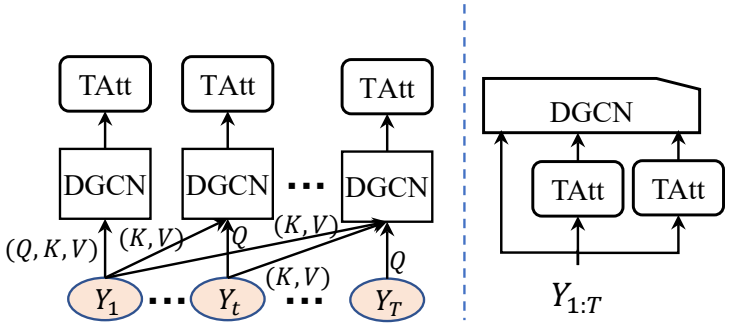


图 3.5 左边：原始的 CDGCN 模块；右边：最终版本的 CDGCN 模块

原始 CDGCN(original CDGCN) 的结构如图 3.5 左边所示，其中 DGCN 模块不仅被用在每一个时间步内来捕获 intra-spatial 关系，而且被用在跨时间步之间来捕获 inter-spatial 关系，该过程描述如下：

$$\begin{aligned} &\text{CDGCN}_{\text{Original}} = \|_{t=1}^{T}(\text{cdgcn}_1,\ldots,\text{cdgcn}_t,\ldots,\text{cdgcn}_T) \\ &\text{where} \quad \text{cdgcn}_t = \text{TAtt}(\text{MHDGCN}(Y_t, Y_{1:T}, Y_{1:T})), \end{aligned} \tag{3.10}$$

这里 $Y_t \in \mathbb{R}^{N\times D}$ 为在时间步 t 的所有个体特征，$Y_{1:T} \in \mathbb{R}^{T\times N\times D}$ 就是前文中 SCE 模块输出的 Y_I，代表所有时间步个体的特征；$\text{TAtt}(\cdot)$ 代表沿着时间维度对 T 个含有跨时间步空间关系的个体特征加权求和。假定 $S^{(t)} = S^{(t)}_{1:T} \in \mathbb{R}^{T\times N\times D}$ 为上式中 MHDGCN 在第 t 个时间步的输出，则 $\text{TAtt}(\cdot)$ 的计算过程展开如下：

$$\begin{aligned} &T_{sum} = \sum_{i=1}^{T} S_i^{(t)} \in \mathbb{R}^{N\times D},\ T'_{sum} = \frac{S^{(t)}}{T_{sum}} \odot S^{(t)} \in \mathbb{R}^{T\times N\times D}, \\ &\text{TAtt}(S^{(t)}) = \sum_{i=1}^{T} T'_{sumi} \in \mathbb{R}^{N\times D}. \end{aligned} \tag{3.11}$$

尽管利用原始的 CDGCN 来提取跨时间步的 inter-spatial 关系，但是公式 3.10 的计算复杂度为 $O(T^5N^3D^2)$ 是难以负担的。如图 3.5 右边所示，将公式 3.10 中的

TAtt(·) 移到 MHDGCN(·) 中来简化 CDGCN 的计算过程。具体来说，首先沿着时间维度使用 TAtt(·) 将来自于跨时间步个体特征的 key 和 value 加权求和，然后在压缩后的 key 和 value 上执行动态图卷积来提取 inter-spatial 关系。最终版本的 CDGCN 的计算复杂度被降低到 $O(T^3N^3D^2)$，它的公式表示如下：

$$\text{CDGCN} = \text{MHDGCN}(Y_I, \text{TAtt}(Y_I), \text{TAtt}(Y_I)). \tag{3.12}$$

3.6 实验

本小节将介绍基于跨时间步动态图卷积的群体行为识别模型在实验中所用到的数据集，实验的具体环境和实现细节，随后将本章方法的实验结果与其他基线模型结果进行比较分析，以及将实验结果可视化分析，最后进行消融研究分析来验证本章模型中所有模块的有效性。

3.6.1 数据集

本章实验在两个常用的公开数据集上进行，即 Volleyball dataset(VD)[11]和 Collective Activity dataset(CAD)[21]，下面分别介绍两个数据集的基本信息。

Volleyball dataset 收集了排球比赛的 55 个视频记录，它们被剪辑分成 3493 个训练集视频片段和 1337 个测试集片段 (总共 4830 个片段)。每个视频片段的中间一帧提供了 3 种标注：(1) 所有选手的边界框坐标；(2) 被标记选手的个体动作标签，即：blocking, digging, falling, jumping, moving, setting, spiking, standing 和 waiting；(3) 给定视频片段的群体行为标签，即：right set, right spike, right pass, right winpoint, left set, left spike, left pass 和 left winpoint。该数据集的个体动作类别数量为 9 种，群体行为类别数量为 8 种。对于那些没有被标注的视频帧，为了提取整个片段的特征，使用 Bagautdinov 等人[10]提供的个体轨迹 (tracklets)。在该数据集上，两种指标被用来评估模型的性能，即多类别分类准确率 (Multi-class Classification Accuracy, MCA) 和平均每类准确率 (Mean Per Class Accuracy, MPCA)。

Collective Activity dataset 由 44 个视频中的 2481 个行为片段组成，每个视频包含的帧数从 194 到 1814 帧不等。该数据集的训练集和测试集的划分根据 Qi 等人[35]而来，即选择 1/3 的视频序列作为测试集，其余的用作训练集。和数据集 VD 类似，该数据集也有 3 种类型的标注：(1) 每十帧的中间一帧上所有个体的边界框坐标；(2) 对应被标注个体的动作标签，即：NA, crossing, waiting, queueing, walking 和 talking；(3) 对应每十帧的群体行为标签，即：crossing, waiting, queueing, walking 和 talking。该数据集提供的个体动作类别数量为 6 种，群体行为类别数量为 5 种，其

中群体行为类别定义为场景中数量最多的个体动作类别。根据 [29,62] 合并群体行为类别 *crossing* 和 *walking* 为 *moving*，同样地，使用 [10] 提供的个体轨迹数据。由于 CAD 数据集每类样本数量的不均衡，应该仅使用 MPCA 作为性能的评价指标，但是之前许多方法仍沿用 MCA，所以本实验同时采用这两种评价指标。

表 3.1 数据集 VD 和 CAD 的统计数据

Dataset	#samples	#training	#testing	#actions	#activities
VD	4830	3493	1337	9	8
CAD	2481	1746	765	6	4

综上所述，数据集 VD 和 CAD 的统计信息如表 3.1 所示。这里，#samples 代表数据集中样本总数，#training 代表训练样本数量，#testing 代表测试集样本数量，#actions 代表个体动作类别数量，#activities 代表群体行为类别数量。

3.6.2 实验配置和实现细节

本章实验所用到的环境配置如表 3.2 所示，模型网络的实现是基于 PyTorch 深度学习框架。

表 3.2 本章实验的运行环境配置

Operation System	GPU	Memory size	Python version	PyTorch version	CUDA version
CentOS7	NVIDIA Tesla V100	16G	3.7.10	1.9.0	10.2

根据之前的方法[6,8,39,44]，调整数据集 VD 中的视频图片的分辨率大小为 $H \times W = 720 \times 1280$，调整数据集 CAD 中的视频图片的分辨率大小为 $H \times W = 480 \times 720$。为了和之前的方法公平比较，对于每个剪辑片段使用 $T = 10$ 帧图片，分别为中间一帧的前面 5 帧和后面 4 帧。随后，对于两个数据集，首先将 T 帧输入的视频片段划分为 $K = 3$ 个时序片段，然后从这 K 个片段里面均匀地采样 K 帧。对于数据集 VD，场景中个体的最大数量为 $N = 12$；对于数据集 CAD 为 $N = 13$。

本章模型使用在 ImageNet 上预训练的 Inception-v3 作为骨干网络提取图片的特征图，得到的全局场景特征图 X_g 通过一个 1×1 的卷积将通道维度降为 $d_g = 128$。通过 [10] 提供的个体边界框，在 X_d 上使用裁剪大小为 5×5 的 RoIAlign 模块来提取个体特征图，然后个体特征图被嵌入到维度为 $D = 1024$ 的向量。对于多头的 SCE 模块，将其注意力头的数量设置为 $h_s = 4$，每个注意力头的隐藏维度设置为

$d_c = 128$，dropout 随机丢弃的概率为 0.1。对于 CDGCN 模块，将其中的 MHDGCN 部分的注意力头数量设置为 $h_d = 4$，则每个注意力头的维度为 $D/h_d = 256$，dropout 的比例设置为 0.1。

对于数据集 VD 的训练过程，采用 Adam[63]优化器学习网络的参数，它的超参数被固定为 $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$。训练的 mini-batch 大小为 6，epoch 数量为 60，初始学习率为 1e-4，并且学习率每 20 个 epoch 以因子 3 衰减。对于数据集 CAD 的训练过程，采用同样设置超参数的 Adam 优化器，训练的 mini-batch 大小为 6，epoch 数量为 30，并且使用一个固定的学习率 5e-5。在本章所有实验中，挑选不同的训练损失函数里面的权重因子 $\lambda \in \{0.2, 0.5, 0.8, 1.0, 1.5\}$，得到当 $\lambda = 1$ 的时候模型表现性能最好。所有的实验在 3 个 V100 GPUs 上进行，并使用 PyTorch 的 DDP 模块进行多 GPU 分布式训练。

3.6.3 实验结果比较与分析

在本小节，分别在数据集 VD 和 CAD 上比较了本章方法和之前 state-of-the-art 方法的性能，如表 3.3 所示。为了公平地进行对比，表 3.3 报告了之前的方法在不同模态输入和不同 backbone 条件下的结果。

在表 3.3 中，'-' 符号表示不存在，对于数据集 CAD 中带'*' 上标的结果表示 MPCA 数值，因为有些文献只提供了 MPCA 而没有提供 MCA；输入模态 (modality) 一列中的 RGB 表示视频 RGB 图像特征，Flow 表示光流 (optical flow) 特征，Keypoint 表示姿态特征。从表 3.3 可以看出，基于注意力/Transformer 的 (attention/Transformer-based) 方法取得了最优秀的结果，例如 GroupFormer[44]和 Dual-AI[45]。对于数据集 VD，骨干网络为 Inception-v3 的本章方法比基于循环神经网络的 (RNN-based) 方法表现要好，例如 stagNet[35], SSU[64], SBGAR[31], CERN[30], HDTM[11]，这归功于对空间和时间交互关系的同步建模，而 RNN-based 方法没有充分考虑空间关系。此外，本章模型的性能超出了一些基于图神经网络的 (GNN-based) 方法，例如 ARG[39]和 HiGCIN[62]，这得益于对时空关系的稀疏化和动态建模。但是，在相同模态的输入前提下，本章方法的性能与最先进的方法仍有较大差距 (MCA 分数和 Dual-AI 差距最大，达到了 1.8%)，可能是由于模型参数过多或者数据集样本不足导致模型过拟合。

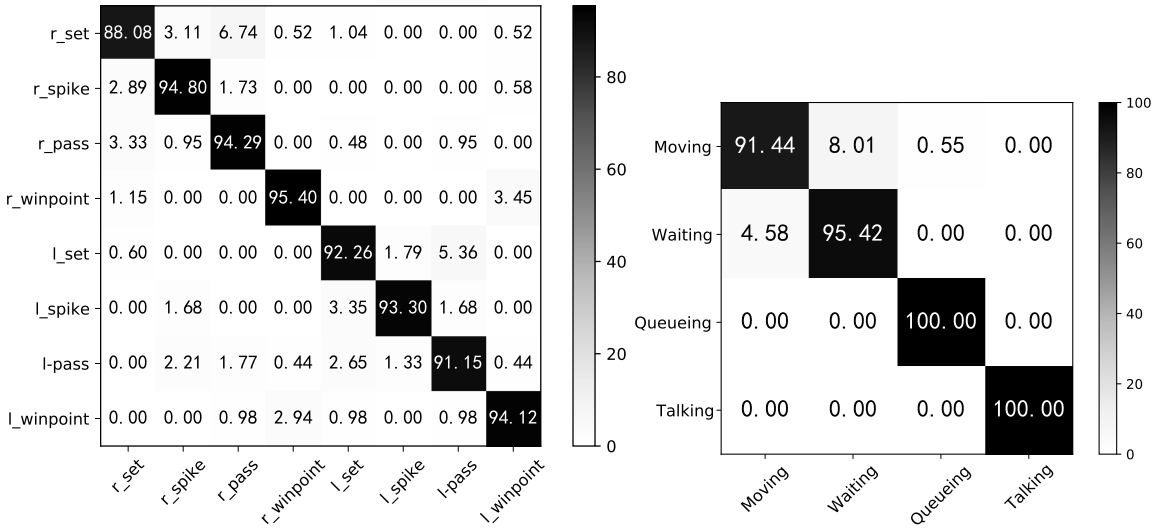
对于数据集 CAD，基于 Inception-v3 骨干网络的本章模型达到了最先进的 MPCA 分数 96.7%，超出了基于相同模态输入和 backbone 的其他方法至少 0.8% MPCA，甚至还超过一些使用额外光流输入的方法如 Dual-AI[45]和 GroupFormer[44]，

表 3.3 本章模型在数据集 VD 和 CAD 上与先进方法比较的结果

Method	Backbone	Modality			Dataset	
		RGB	Flow	Keypoint	VD-MCA(%)	CAD-MCA(%)
HDTM[11]	AlexNet	✓			81.9	89.7
stagNet[35]	VGG-16	✓			89.3	89.1
SSU[64]	Inception-v3	✓			90.6	-
HiGCIN[62]	ResNet-18	✓			91.4	-(93.0*)
CERN[30]	VGG-16	✓			83.3	87.2
ARG[39]	Inception-v3	✓			92.5	91.0
Actor-Transformer[6]	I3D	✓			91.4	-
PRL[41]	VGG-16	✓			91.4	-(93.8*)
DIN[8]	VGG-16	✓			93.6	-(95.9*)
TCE+STBiP[43]	Inception-v3	✓			93.3	-(95.1*)
GroupFormer[44]	Inception-v3	✓			94.1	**93.6**
Dual-AI[45]	Inception-v3	✓			**94.4**	-
SBGAR[31]	Inception-v3	✓	✓		67.6	-(89.9*)
CRM[65]	I3D	✓	✓		93.0	85.8
Dual-AI[45]	Inception-v3	✓	✓		95.4	-(96.5*)
GroupFormer[44]	I3D+AlphaPose	✓	✓	✓	**95.7**	**96.3**
Actor-Transformer[6]	I3D+HRNet	✓		✓	93.5	91.0
TCE+STBiP[43]	Inception-v3+HRNet	✓		✓	94.1	-(95.4*)
Ours(SCE+CDGCN)	Inception-v3	✓			92.6 (92.9*)	95.2 (96.7*)
	ResNet-18	✓			91.8 (92.1*)	93.6 (91.5*)

这论证了本章方法利用动态图卷积推理个体间时空交互关系的有效性。特别地，与基于相同 RGB 图片输入和 Inception-v3 骨干网络的 GroupFormer 相比，本章模型超出了 1.6%MCA 分数。但是，在数据集 CAD 上的 MCA 分数与最先进的方法仍有差距 (MCA 分数和多模态输入的 GroupFormer 差距最大，达到了 1.1%)，这可能受到数据集 CAD 中每类样本数量不均衡的影响。

基于 Inception-v3 骨干网络的模型，在数据集 VD 和 CAD 上的混淆矩阵如图 3.6a 和 3.6b 所示。对于数据集 VD，对应结果的混淆矩阵表明，由于该模型对空间信息的建模能力，所以能够比较好地区分右边队伍和左边队伍执行的动作。此外，该模型对行为类别 *winpoint* 的预测表现最好，这可能归功于 SCE 模块，它可以帮助区分类别 *winpoint* 中与其他行为相对不同的空间布局。在数据集 VD 中大多数预测失败的情形，是把一支队伍执行的一种行为误当作同一支队伍执行的另一种



(a) 数据集 VD 上的混淆矩阵　　(b) 数据集 CAD 上的混淆矩阵

图 3.6 本章模型在数据集 VD 和 CAD 上的混淆矩阵

行为。例如，一些 *right set*(*left set*) 行为被误当作 *right pass*(*left pass*)，这可能是由于它们具有相似的时空关系和场景上下文信息，导致它们没有被很好地区分。

对于数据集 CAD，对应的混淆矩阵表明，大多数预测失败的情形是把行为类别 *moving* 误当作 *waiting*，这可能是由于这两个行为类别具有相似的视觉上下文信息，再加上视频片段的时序动态不是很充足，导致没有很好地捕捉这两个类别的差异。

3.6.4 消融研究分析

在本小节，对本章的模型进行消融研究分析，验证模型中主要组成模块的有效性。消融研究中所有的实验在数据集 VD 上进行，对应模型的 backbone 网络设置为 Inception-v3。下面是对消融模型的介绍：

- Base model：它由 backbone 网络，RoIAlign 模块，全局池化层和最后的分类层组成。

- w/ SCE ：它由 backbone 网络，RoIAlign 模块，SCE 模块，全局池化层和分类层组成，它去掉了 CDGCN 模块，只使用场景上下文编码模块利用增强的个体特征进行预测。

- w/ CDGCN：它由 backbone 网络，RoIAlign 模块，CDGCN 模块，全局池化层和分类层组成。它去掉了 SCE 模块，只使用跨时间步动态图卷积建模个体的时空关系进行预测。

- w/ SCE+CDGCN：它由 backbone 网络，RoIAlign 模块，SCE 模块，CDGCN

模块，全局池化层和分类层组成，即为最终的完整模型，它同时利用 SCE 模块和 CDGCN 模块进行关系推理。

- w/o PE：它和 SCE+CDGCN 模型的组成类似，只是移除了其中 SCE 模块的位置编码 (PE) 部分。

表 3.4 本章不同变体模型的消融实验结果

Model	Base model	w/ SCE	w/ CDGCN	w/ SCE+CDGCN	w/o PE
MCA(%)	91.1	92.1	92.3	**92.6**	92.4
MPCA(%)	91.4	92.4	92.7	**92.9**	92.7

以上模型的实验结果如表 3.4 所示，表 3.4 表明加上任何一个模块都可以提升模型的性能。具体来说，w/ SCE 和 w/ CDGCN 相比于 base model 取得了相近的提升，分别说明了 SCE 模块提取场景上下文信息和 CDGCN 模块推理时空关系的有效性。进一步地，w/ CDGCN 的结果略好于 w/ SCE，说明 CDGCN 模块可以充分地建模个体之间的空间和时间关系。将 SCE 模块和 CDGCN 模块结合在一起，w/ SCE+CDGCN 可以学习更具语义和信息的表征，因此得到的效果进一步提升。特别地，w/o PE 的结果对比 w/ SCE+CDGCN 的有所降低，表明了 SCE 模块中的位置编码可以促进个体上下文信息的提取。

3.7 本章小结

本章针对群体行为识别的研究，提出了基于跨时间步动态图卷积的方法，来提取场景上下文信息和推理个体之间的时空关系。首先介绍利用 RoIAlign 模块提取个体的外观特征，然后使用 SCE 模块提取个体所在场景的上下文信息，接着使用 CDGCN 模块学习跨时间步的时空关系，随后使用全局池化层生成群体表征进行群体行为预测。最后在两个广泛使用的数据集上进行系统的实验，在只有 RGB 图像模态的输入条件下，本章方法在群体识别任务中取得了较好的性能，同时也验证了方法中组成模块的有效性。

4 混合姿态特征的双分支群体行为识别

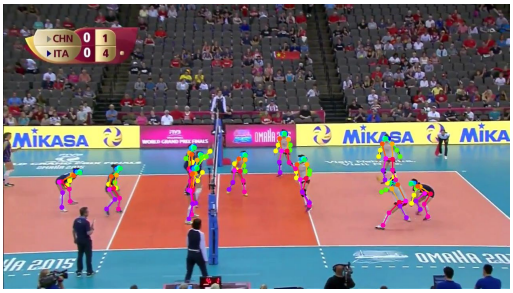
本章在上一章的基础上介绍了混合姿态特征的双分支群体行为识别模型，首先阐述了模型提出的动机，然后描述了该模型的总体设计框架，随后对混合姿态特征分支的流程进行了详细的说明，最后通过大量的实验验证了本章算法的有效性以及合理性。

4.1 引言

在第 3 章中介绍了基于跨时间步动态图卷积的群体行为识别方法，该方法仅仅利用个体的外观特征进行关系推理，因此将其称为外观分支 (box branch)，在本章中介绍第二种输入模态的方法。同样地，在之前的许多方法中仅仅利用 RGB 图像作为输入预测群体行为，并且有些取得了不错的效果。这些方法将 RGB 图像输入 backbone 得到视觉特征图，然后利用 RoIAlign[38]和边界框从特征图上提取个体的外观特征，基于这种外观特征推理个体之间的时空关系，最后从个体表征中得到群体表征。可以看出，个体动作识别是群体行为识别过程中的重要环节。但是，仅仅利用个体的视觉外观特征进行动作识别是不准确的，还可以融合个体的姿态特征进行识别，因为人类的动作大部分和身体关节点 (keypoint) 的位置以及运动 (即姿态特征) 是高度相关的，例如手部和腿部的关节点。

此外，使用 RGB 图像表达个体特征会带来场景偏差 (scene biases)，因为数据集中的场景相对有限，在不同场景中进行训练可能导致结果差异很大；另一方面，相似场景得到的个体外观特征可能会更相近，而不同场景中的个体外观特征可能差异很大，尽管这些个体执行相同的动作。例如，在两个不同场地的排球比赛中，一个在室内另外一个在户外，我们希望在室内和在户外执行扣球的运动员具有相似的个体外观特征，而不会受到场地背景差异的影响。除此之外，使用视频 RGB 图片帧作为输入，可能会包含个体的一些隐私或者偏差信息的视觉数据，甚至会导致相关的隐私和道德问题。但是，使用个体的姿态特征就不会受到上述问题的影响，因为姿态特征只与关节点的位置变化有关。

如图 4.1 所示分别为排球比赛场景和餐厅排队场景中目标人群关节点的示例，根据每个运动员在多帧中的关节运动和位置可以判断其动作，如扣球 (spiking) 和垫球 (digging)；同样地，根据餐厅人群在多帧中的关节位置可以判断每个人的动作，如排队 (queueing) 和说话 (talking)。这说明身体关节点不仅可以运用到细微动作的



(a) 数据集 VD 上目标个体的关节点位置

(b) 数据集 CAD 上目标个体的关节点位置

图 4.1 在数据集 VD 和 CAD 上目标个体的关节点位置示意图

识别，如在体育比赛的动作，还可以用于日常动作识别之中。因此，为了得到准确且健壮的个体语义表征，不仅需要捕捉个体关节点的位置变化 (即姿态特征)，而且需要学习他们外观特征中包含的时空动态。基于这个目的，在本章中将使用两种输入模态进行群体行为识别，即将 RGB 图像和个体的关节点位置信息作为输入，并且使用两种不同的 backbone 网络。

本章将会融合姿态分支 (pose branch) 和第 3 章中外观分支 (box branch) 进行群体行为识别，提出了基于姿态特征的动态图卷积网络 (Pose-based Dynamic Graph Convolution Network, PDGCN)。具体来说，姿态分支除了生成姿态特征的过程之外，它与第 3 章中外观分支具有相似的结构。对于姿态分支，首先利用姿态估计模型和个体边界框生成所有帧个体的 2D 关节点坐标，然后将关节点坐标嵌入生成姿态特征向量，随后利用第 3 章外观分支中的模块进行个体关系推理。两个分支分别进行训练，最后将两个分支融合进行群体行为识别。本章研究工作的对应代码已发布在 https://github.com/0shelter0/PDGCN。

4.2 混合姿态特征的网络框架

本章提出的 PDGCN 模型总体框架如图 4.2 所示，该模型由两个分支组成，即姿态分支 (pose branch) 和外观分支 (box branch)；模型的输入为一个 RGB 视频片段，该片段包含 T 帧，每帧标注的最大个体数量为 N。对于姿态分支，给定输入的 RGB 图片 $X_{\text{img}} \in \mathbb{R}^{T\times H\times W\times 3}$ 和个体边界框，利用姿态估计网络 HRNet[66](图 4.2中 backbone2) 提取出个体的 2D 关节点坐标 $C \in \mathbb{R}^{T\times N\times 17\times 2}$，其中 17 代表个体中预测关节点的数量。随后将这些关节点坐标嵌入生成姿态特征向量 $X_P \in \mathbb{R}^{T\times N\times D_p}$，其中 D_p 表示姿态特征向量的维度。对于外观分支，通过 3.2.1 小节的过程得到所有

个体的特征表示 $X_I \in \mathbb{R}^{T\times N\times D}$。

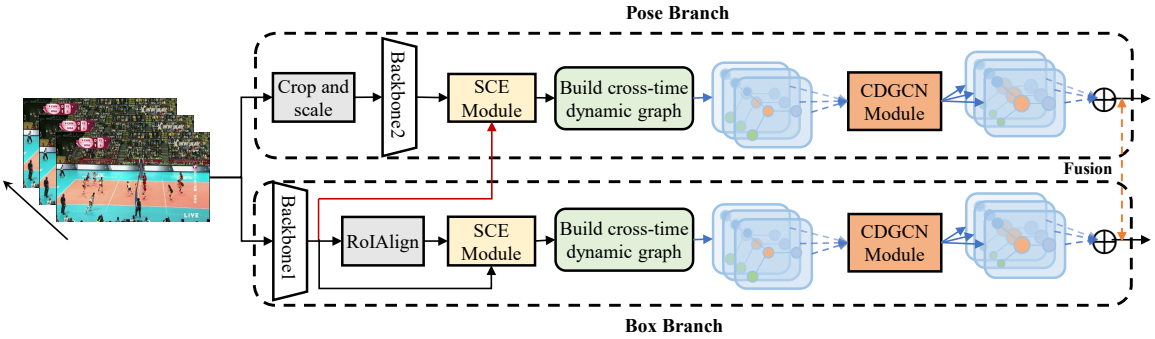


图 4.2 PDGCN 模型的总体框架

在姿态分支中，将得到的姿态特征 X_P 和 3.2.1 小节中的全局场景上下文特征图 X_g 作为场景上下文编码 (SCE) 模块的输入，输出的 $Y_P \in \mathbb{R}^{T\times N\times D_P}$ 表示场景上下文编码过的姿态特征，类似地 Y_P 中的这 $T\times N$ 个个体形成了一个时空图 (ST graph)。随后，利用动态图卷积模块在 ST graph 上推理个体之间的时空关系，经过全局池化 (图 4.2 中的 $\oplus$) 模块得到最终的群体表征 $\tilde{Y}_P$，最后基于 $\tilde{Y}_P$ 得到姿态分支的群体行为预测分数。在外观分支中，对个体特征 X_I 的关系推理已经在 3.2.1 小节中说明，此处不再赘述。最后，将两个分支得到的群体行为预测分数进行融合，如图 4.2 中最右侧连接两个分支的虚线所示。

4.3 姿态特征提取

本小节将介绍 PDGCN 中姿态分支提取姿态特征的过程，该过程分为两个步骤：首先利用姿态估计网络 HRNet 提取每帧中个体的关节点坐标，然后基于这些关节点坐标生成个体的姿态特征向量。

4.3.1 姿态分支的关节点提取

和之前许多使用关节点作为输入模态的方法一样，PDGCN 中的姿态分支使用 HRNet 作为姿态估计网络，来提取个体的关节点坐标。HRNet[66]在 CVPR2019 论文 *Deep High-Resolution Representation Learning for Human Pose Estimation* 中提出，选择 HRNet 作为姿态估计网络是因为它具有相对简单的设计，同时在姿态估计 (即关节点检测) 基线上达到很好的性能。具体来说，在姿态分支中使用最小版本的 HRNet 模型 *pose_hrnet_w32*，对应输入模型的单一个体区域固定高宽为 256×192，并采用原论文作者提供的在 COCO 关节点检测数据集上 (COCO keypoint detection dataset)[67]训练好的模型权重。

现在基于深度学习的方法主要有两种方式进行人体姿态估计，即：基于回归 (regressing) 的方式，直接预测每个关节点的位置坐标；基于热力图 (heatmap) 的方式，即针对每个关节点预测一张热力图 (预测出现在每个位置上的分数)。当前检测效果最好的一些方法基本都是基于 heatmap 的，所以 HRNet 也是基于 heatmap 的方式。

HRNet 由多个并行的从高到低分辨率的子网络组成，在整个过程中始终保持高分辨率表征；这些并行的多分辨率子网络之间不断地进行信息交换，以此来增强每个阶段的高分辨率表征。HRNet 以并行的方式连接这些从高到低分辨率的子网络，它可以使得高分辨率一直贯穿整个过程，而不是通过由低到高的方式恢复分辨率。由于高分率表征会包含更多的原始图片位置信息，所以 HRNet 可以保持对位置的敏感性 (position sensetivity)。最终仅仅使用 HRNet 输出的最高分辨率表征估计关节点位置，因此预测的关节点热力图可能在空间上更准确。

具体来说，将 HRNet 最后一个 stage 分辨率最高的特征图，即下采样 4 倍子网络的输出，输入一个步长和卷积核大小都为 1、卷积核个数为 17 的卷积层。最终得到大小为 $\frac{256}{4} \times \frac{192}{4} \times 17$ 的特征层，它即为每个关节点的热力图，这里 17 为 COCO 数据集对人体标注的关节点数量。在本章的姿态分支中，将会使用这 17 个关节点预测的坐标生成相应的姿态特征，这些关节点在 COCO 数据集[67]中对应的顺序如下：nose, left_eye, right_eye, left_ear, right_ear, left_shoulder, right_shoulder, left_elbow, right_elbow, left_wrist, right_wrist, left_hip, right_hip, left_knee, right_knee, left_ankle, right_ankle。得到了对每个关节点预测的 heatmap 之后，将每张 heatmap 中分数最大的位置作为该关节点的预测位置，然后映射回原图就能得到原图上单一个体关键点的坐标。

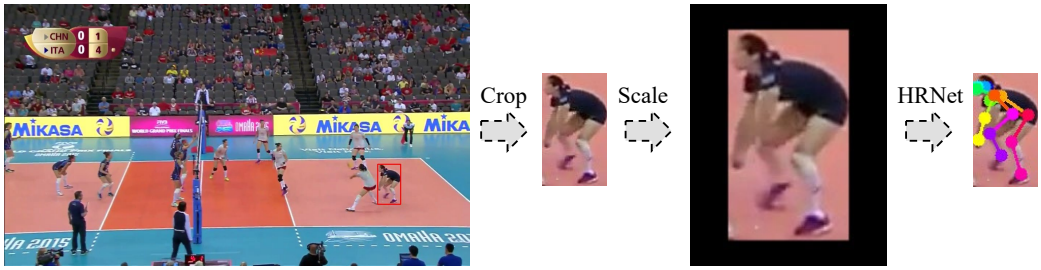


图 4.3 数据集中的视频图片经过裁剪、缩放以及得到关节点位置的过程

然而，HRNet 针对的是单一个体的姿态估计 (即输入网络的图像中应该只有一个人体目标)，但本章所用到的数据集都涉及到多人的场景，所以需要利用个体的边界框将其从视频帧中裁剪出来，亦即定位单一个体在图片中的位置。裁剪定位的单

一个体区域还需要缩放到固定尺寸大小 256×192，以满足 HRNet 网络输入图像的尺寸，该过程可以被归结为裁剪和缩放。如图 4.3 所示，为在 Valleyball 数据集[11]中的一帧视频图片上进行裁剪 (Crop) 和缩放 (Scale) 的过程，然后将缩放后的个体区域输入 HRNet，得到目标个体最终的关节点位置。

对单一个体区域的缩放是通过仿射变换 (affine transform) 实现的，下面详细解释该过程实现的原理。仿射变换指一个向量空间进行线性变换和平移变成另外一个向量空间，利用仿射变换可以实现旋转、平移、缩放等操作。一个任意的仿射变换都能表示为乘以一个矩阵 (线性变换) 再加上一个向量 (平移)，在二维空间中给定一个线性变换矩阵 $A_{2\times2} = \begin{bmatrix} a_{00} & a_{01} \\ a_{10} & a_{11} \end{bmatrix}$ 和平移向量 $B_{2\times1} = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$，则一个仿射变换矩阵可以表示为，利用矩阵 A 和向量 B 对二维坐标向量 $S = \begin{bmatrix} x & y \end{bmatrix}^T$ 做仿射变换，该过程表示如下：

$$T = A \cdot S + B = \begin{bmatrix} A & B \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = M \cdot \begin{bmatrix} x & y & 1 \end{bmatrix}^T, \tag{4.1}$$

上式中 T 表示变换后的二维坐标，$M_{2\times3} = \begin{bmatrix} A & B \end{bmatrix} = \left[\begin{array}{cc|c} a_{00} & a_{01} & b_0 \\ a_{10} & a_{11} & b_1 \end{array}\right]$ 是一个仿射变换矩阵。但是一般很难直接找出变换矩阵 $M_{2\times3}$，于是在 OpenCV 中提供了对应的接口，通过原平面中任意三个点的坐标和在变换后相应的三个点坐标，解出变换矩阵 $M_{2\times3}$。如图 4.4 为变换前后平面中三个点的映射示意图，即分别将原平面 (Source) 中的点 $P_1(\frac{w_i}{2} + w_{pad}, \frac{h_i}{2}), P_2(\frac{w_i}{2} + w_{pad}, 0), P_3(w_i + 2w_{pad}, \frac{h_i}{2})$ 代入公式 4.1 中的 S，同时分别将目标平面 (Target) 中相应的点 $P'_1(\frac{w_o}{2}, \frac{h_o}{2}), P'_2(\frac{w_o}{2}, 0), P'_3(w_o, \frac{h_o}{2})$ 代入公式 4.1 中的 T，通过线性方程组求解出 $M_{2\times3}$；图 4.4 中的 w_i, h_i, w_{pad} 分别表示原平面的宽、高和填充 (padding) 宽度，w_o, h_o 分别表示变换后平面的宽和高。

考虑对 Volleyball 数据集中的 RGB 图像进行仿射变换的情形，假设输入图像某个通道像素值 I 和坐标 (x, y) 的映射关系为 $I = \mathrm{src}(x, y)$，变换后该通道的映射关系为 $I' = \mathrm{dst}(x, y)$，则有如下关系：

$$\begin{aligned} I' = \mathrm{dst}(x, y) &= \mathrm{src}(M_{2\times3} \cdot \begin{bmatrix} x & y & 1 \end{bmatrix}^T) \\ &= \mathrm{src}(a_{00}x + a_{00}y + b_0, a_{10}x + a_{11}y + b_1). \end{aligned} \tag{4.2}$$

在图 4.3 中展示了 Volleyball 数据集中具体的目标个体经过公式 4.2 实现缩放的效果，该变换不仅保持了目标个体的横纵比例，而且满足了输出的固定尺寸大小 256×192，所以在输出区域中会有黑边填充。

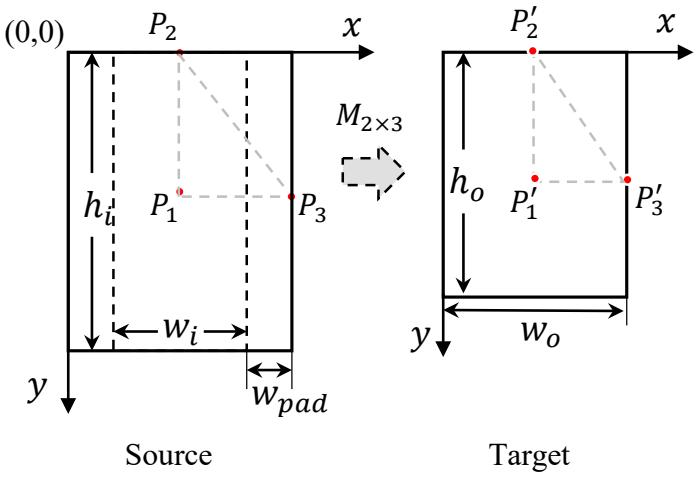


图 4.4 求仿射变换矩阵的示意图

另一方面，数据集中的视频图片经过 HRNet 得到预测的 heatmap，然后将 heatmap 预测得到的关节点坐标再映射回对应的数据集图片中，是一个与公式 4.1 逆向的仿射变换过程。可以翻转图 4.4 中的映射关系，将公式 4.1 中的 S 和 T 的位置交换 (亦即令 P_i 代入 T 中，P_i' 代入 S 中，$i = 1, 2, 3$)，求出逆仿射变换矩阵 $M_{2\times3}'$，然后将 $M_{2\times3}'$ 左乘 heatmap 上预测的关节点坐标，即可映射回原始视频图片中。

4.3.2 姿态特征生成

通过 4.3.1 小节可以得到数据集图片中单一个体边界框中的关节点坐标，随后将这些关节点坐标转化成对应的姿态特征。该过程可以分为两个步骤，首先是对个体的关节点坐标进行归一化处理，然后通过线性变换将其嵌入到 D 维的特征空间。给定第 i 个个体的非归一化边界框 $B_i = (x_1, y_1, x_2, y_2)$ 和其中的 17 个关节点坐标 $C_i \in \mathbb{R}^{17\times2}$，可以得到边界框的中心点坐标为 $O_i = (\frac{x_1+x_2}{2}, \frac{y_1+y_2}{2})$，定义对关节点坐标的缩放因子为 $s = \sqrt{\frac{(x_2-x_1)(y_2-y_1)}{4}}$，则对 $C_i \in \mathbb{R}^{17\times2}$ 进行如下的归一化处理：

$$C_i' = \frac{C_i - O_i}{s}, \tag{4.3}$$

上式中 C_i' 为归一化后第 i 个个体的关节点坐标。经过公式 4.3 中的归一化处理后，使得关节点坐标数值的偏差更小，从而让训练更加稳定。

随后，首先将 C_i' 的形状展平，然后通过两次线性变换将 C_i' 嵌入到 D_p 维的空间中，可以表示如下：

$$X_{Pi} = (C_i'W_1 + b_1)W_2 + b_2, \tag{4.4}$$

这里，$X_{Pi} \in \mathbb{R}^{D_P}$ 为得到的第 i 个个体的姿态特征向量，$W_1 \in \mathbb{R}^{(17\times2)\times d_w}, W_2 \in \mathbb{R}^{d_w \times D_P}$ 为可学习的线性变换参数矩阵，$b_1 \in \mathbb{R}^{d_w}, b_2 \in \mathbb{R}^{D_P}$ 为可学习的偏置向量，d_w 代表嵌入的隐藏层维度。将输入视频帧中的所有 $T \times N$ 个个体的姿态特征堆叠在一起形成姿态表征 $X_P \in \mathbb{R}^{T\times N\times D_P}$。

4.4 混合姿态特征的关系推理

在本小节将会介绍混合姿态特征的时空关系推理，即 PDGCN 中利用姿态分支和外观分支推理个体之间时空关系的过程。由于 PDGCN 的外观分支已经在上一章中详细阐述，故在下面的 4.4.1 和 4.4.2 小节只介绍姿态分支的关系推理过程，然后在 4.4.3 小节说明如何将这两个分支融合来预测群体行为。

4.4.1 场景上下文编码

为了将全局场景的上下文信息融入到个体的姿态特征中，使用 3.4 节中的场景上下文编码 (SCE) 模块对姿态特征 $X_P \in \mathbb{R}^{T\times N\times D_P}$ 进行编码。同时，将来自于外观分支 backbone 中的全局场景特征图 $X_g \in \mathbb{R}^{T\times h\times w\times d_g}$ 作为姿态分支 SCE 模块的输入，如图 4.2 中连接两个分支的实线箭头所示。根据 3.4 节中对 SCE 模块的介绍，PDGCN 中姿态分支的 SCE 模块将场景上下文信息编码到 X_P 中的过程表示如下：

$$Y_P^{(t)} = X_P^{(t)} \| \{ \|_{i=1}^{h'_s} \mathrm{sce}_i(X_P^{(t)}, X_g^{(t)}) \} \tag{4.5}$$

其中，t 表示时间步的索引，$Y_P^{(t)} \in \mathbb{R}^{N\times D_P}$ 为在第 t 个时间步中上下文信息编码过的个体姿态特征，h'_s 表示姿态分支中 SCE 模块注意力头的数量。将 T 个时间步的 SCE 模块输出堆叠，可以得到 $Y_P \in \mathbb{R}^{T\times N\times D_P}$，表示所有 $T \times N$ 个个体的上下文信息编码过的姿态特征。

4.4.2 动态图卷积推理

在 3.5 节中提到，在群体行为识别任务中需要动态地捕获个体之间的空间关系，这种空间关系包括 intra-spatial 关系和 inter-spatial 关系，利用跨时间步的动态图卷积网络 (DGCN) 可以同时建模上述两种个体间的空间关系。同样地对于 PDGCN 中的姿态分支，在得到上下文信息编码过的个体姿态特征 $Y_P \in \mathbb{R}^{T\times N\times D_P}$ 之后，使用多头的 DGCN(即 MHDGCN) 在 Y_P 上推理个体间的时空关系，最终得到具有丰富语义的和相关性的个体表征 $\tilde{Y}_P \in \mathbb{R}^{T\times N\times D_P}$。该过程通过公式 3.12 可以描述如下：

$$\tilde{Y}_P = \mathrm{MHDGCN}(Y_P, \mathrm{TAtt}(Y_P), \mathrm{TAtt}(Y_P)) \tag{4.6}$$

同样地，在 PDGCN 的外观分支中，将场景上下文信息编码后的个体外观特征 $Y_I \in \mathbb{R}^{T\times N\times D}$，通过公式 3.12 中的动态图卷积进行时空关系推理得到 $\tilde{Y}_I \in \mathbb{R}^{T\times N\times D}$。

4.4.3 分支融合和损失函数

从 4.4.2 小节中可以得到，姿态分支和外观分支中个体的最终表征分别为 $\tilde{Y}_P$ 和 $\tilde{Y}_I$，利用 $\tilde{Y}_P$ 和 $\tilde{Y}_I$ 可以在对应分支上进行个体动作和群体行为的预测任务，得到对应的预测分数 (这里指 softmax 分数或概率分布)。根据现有方法[6,43]的结论，应该选取对两个分支预测分数的融合 (晚期融合)，而不是融合早期的姿态特征和外观特征再进行时空关系推理以及预测。具体来说，一般的晚期融合采用的方式为：分别单独地训练两个分支，然后在测试阶段将两个分支的预测分数加权求和作为最后类别的概率分布。

PDGCN 也采用晚期融合的方式将姿态分支和外观分支进行融合，将 $\tilde{y}_P^g$ 和 $\tilde{y}_P^a$ 分别表示为姿态分支对群体行为和个体动作预测的分数，同时将 $\tilde{y}_I^g$ 和 $\tilde{y}_I^a$ 分别表示为外观分支对群体行为和个体动作预测的分数；则在模型测试阶段 (即 inference)，最终个体动作和群体行为的预测分数 (概率分布) 表示为：

$$\tilde{y}^a = \alpha_1\tilde{y}_P^a + \alpha_2\tilde{y}_I^a,\ \tilde{y}^g = \alpha_1\tilde{y}_P^g + \alpha_2\tilde{y}_I^g \tag{4.7}$$

其中，$\tilde{y}^a$ 和 $\tilde{y}^g$ 分别表示最终个体动作和群体行为的概率分布，α_1, α_2 分别为姿态分支和外观分支预测分数的权重，根据现有方法[6,43]挑选出的两个分支的最佳权重比例，在本章所有实验中设置 $\alpha_1 = \frac{1}{3}, \alpha_2 = \frac{2}{3}$。

PDGCN 以端到端的方式进行训练，在两个分支中同时预测个体动作和群体行为，分别将这两个分类任务在不同分支中的交叉熵损失相加，得到个体动作分类总损失 $\mathcal{L}^a$ 和群体行为分类总损失 $\mathcal{L}^g$，最后将这两个总损失加权求和作为训练的目标 $\mathcal{L}$，该过程计算如下：

$$\begin{aligned}
\mathcal{L}^a &= \mathcal{L}_P^a(y^a, \tilde{y}_P^a) + \mathcal{L}_I^a(y^a, \tilde{y}_I^a), \\
\mathcal{L}^g &= \mathcal{L}_P^g(y^g, \tilde{y}_P^g) + \mathcal{L}_I^g(y^g, \tilde{y}_I^g), \\
\mathcal{L} &= \lambda_1\mathcal{L}^a + \lambda_2\mathcal{L}^g,
\end{aligned} \tag{4.8}$$

这里，$\mathcal{L}_P^a$ 和 $\mathcal{L}_I^a$ 分别为姿态分支和外观分支中预测个体动作的交叉熵损失，$\mathcal{L}_P^g$ 和 $\mathcal{L}_I^g$ 分别为姿态分支和外观分支中预测群体行为的交叉熵损失，y^g 和 y^a 是群体行为和个体动作的真实标签。λ_1 和 λ_2 为两个总损失的权重因子，经过实验发现个体动作和群体行为总损失权重在相等的时候模型表现最好，所以在所有的实验中设置 $\lambda_1 = \lambda_2 = 1$。

4.5 实验

本小节首先介绍 PDGCN 模型在实验中的详细配置以及实现细节，接着将 PDGCN 的实验结果与其他模型进行比较，以及对结果的可视化分析，最后进行消融研究来验证模型组成的有效性。

4.5.1 实验配置

4.5.1.1 数据集

本节实验在两个广泛使用的群体行为识别数据集上进行，即 Volleyball dataset (VD)[11] 和 Collective Activity dataset (CAD)[21]，下面将会简要地介绍这两个数据集，有关数据集更详细的信息可以参考 3.6.1 小节。

Volleyball dataset. 该数据集包含 55 个排球比赛视频，这些视频被分为 4830 个片段，其中 3493 个片段作为训练集，1337 个作为测试集。每个片段被标注为 8 个群体行为类别中的一个，这些群体行为包括：right set, right spike, right pass, right winpoint, left set, left spike, left pass, left winpoint。此外，每个片段只有中间一帧被标注，包括选手的边界框及其 9 种个体动作标签 (waiting, setting, digging, falling, spiking, blocking, jumping, moving and standing) 之一的个体动作类别。

Collective Activity dataset. 该数据集包含来自 44 个视频序列中的 2481 个群体行为片段，这些视频通过手持相机在街区和室内场景拍摄得到。该数据集包括 5 种群体行为标签 (crossing, waiting, queuing, walking, talking)，以及 6 种个体动作标签 (NA, crossing, waiting, queuing, walking, talking)，其中群体行为的标签被定义为场景中最大参与人数的个体动作类别。根据之前工作[35]的数据集划分方案，使用 1/3 的视频序列用作测试，其余的用作训练。此外，将行为类别 *walking* 和 *crossing* 合并为类别 *moving*。

4.5.1.2 实验运行环境

本章实验所在的运行环境配置如表 4.1 所示，PDGCN 网络的实现是基于 PyTorch 深度学习框架。所有的实验在 3 块 GPUs 上进行，通过 PyTorch 的 DDP 模块 (DistributedDataParallel) 进行多 GPUs 的分布式训练，并且采用多 GPUs 间的通信库为 NCCL(NVIDIA Collective Communication Library)。

表 4.1 本章实验的运行环境配置

Operation System	GPU	Memory size	Python version	PyTorch version	CUDA version
CentOS7	NVIDIA Tesla V100	16G	3.7.13	1.10.1	10.2

4.5.1.3 评价指标

为了方便和现有方法的结果进行全面对比，本章实验采用两种指标来评估模型的性能，即多类别分类准确率 (Multi-class Classification Accuracy, MCA) 和平均每类准确率 (Mean Per Class Accuracy, MPCA)。MCA 即为一般的多分类准确率，MPCA 为每个类别准确率 (亦即召回率，recall) 的平均值，它们的计算方法如下：

$$\text{MCA} = \frac{N_{\text{hits}}}{N_{\text{preds}}} \times 100\%, \tag{4.9}$$

$$\text{MPCA} = \frac{1}{M} \sum_{i=1}^{M} acc_i \times 100\%. \tag{4.10}$$

上式中，N_{hits} 和 N_{preds} 分别表示预测正确的样本数量和预测的样本总数，M 表示类别总数，acc_i 为类别 i 的预测准确率。具体来说，对于数据集 VD 只使用 MCA(%) 作为评价指标；对于数据集 CAD，由于其类别数量的不均衡，故同时使用 MCA(%) 和 MPCA(%) 作为评价指标。

4.5.2 PDGCN 实现细节

对于数据集 VD，将其视频图片的分辨率调整到 $H \times W = 720 \times 1280$；对于数据集 CAD，将视频每一帧的分辨率调整到 $H \times W = 480 \times 720$。为了和之前的方法进行公平的比较，在每个视频片段里面挑选 $T = 10$ 帧，其中 5 帧在被标注帧之前，4 帧在之后。随后对于这两个数据集，首先将 T 帧输入的视频片段划分为 $K = 3$ 个时序片段，然后从这 K 个片段里面均匀地采样 K 帧，作为 PDGCN 模型的输入。

对于外观分支，使用在 ImageNet[48]上预训练的 Inception-v3/ResNet-18 作为 backbone 来提取图片的特征，得到全局场景特征图 X_g 和个体视觉特征图 X_d，随后利用个体边界框和裁剪大小为 5×5 的 RoIAlign[38]模块，在 X_d 上提取个体外观特征 X_I。对于姿态分支，使用在 COCO 关节点检测数据集[67]上训练好的模型权重 *pos_hrnet_w32*，并利用个体的边界框数据，分别在数据集 VD 和 CAD 上预测关节点的位置，随后生成对应的个体姿态特征 X_P。将个体外观特征和姿态特征的通道数量分别设置为 $D = 1024$ 和 $D_p = 256$。在两个分支中，将多头 SCE 模块注意力头的数量均设置为 $h_s = 4$，dropout 丢弃概率设置为 0.1，编码的维度设置为 128；将 MHDGCN 模块注意力头的数量均设置为 4。

在数据集 VD 的训练过程中，采用初始学习率为 10^{-4} 的 Adam[63]优化器学习网络的参数，它的超参数设置为 $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$。PDGCN 训练的 mini-batch 大小为 6，epoch 数量为 30，学习率每 10 个 epoch 衰减为原来的 $\frac{1}{3}$。在数

据集 CAD 的训练过程中，使用同样设置的 Adam 优化器，训练的 mini-batch 大小为 6，epoch 数量为 50，采用固定大小的学习率 5×10^{-5} 进行训练。

4.5.3 实验结果比较分析

在本小节中，在数据集 VD 和 CAD 上分别比较了 PDGCN 和其他先进的方法，比较的结果如表 4.2 所示。为了公平的比较，表 4.2 报告了这些方法在不同输入模态和不同 backbone 网络情形下的结果。

表 4.2 在数据集 VD 和 CAD 上与先进方法比较的结果

Method	Backbone	Modality			Dataset	
		RGB	Flow	Keypoint	VD-MCA(%)	CAD-MCA(%)
HDTM[11]	AlexNet	✓			81.9	89.7
stagNet[35]	VGG-16	✓			89.3	89.1
SSU[64]	Inception-v3	✓			90.6	-
CERN[30]	VGG-16	✓			83.3	87.2
ARG[39]	Inception-v3	✓			92.5	91.0
Actor-Transformer[6]	I3D	✓			91.4	-
PRL[41]	VGG-16	✓			91.4	-(93.8*)
DIN[8]	VGG-16	✓			93.6	-(95.9*)
TCE+STBiP[43]	Inception-v3	✓			93.3	-(95.1*)
GroupFormer[44]	Inception-v3	✓			94.1	**93.6**
Dual-AI[45]	Inception-v3	✓			**94.4**	-
SBGAR[31]	Inception-v3	✓	✓		67.6	-(89.9*)
CRM[65]	I3D	✓	✓		93.0	85.8
Dual-AI[45]	Inception-v3	✓	✓		95.4	-(96.5*)
SACRF[7]	I3D+Pose+FPN	✓	✓	✓	95.0	95.2
GroupFormer[44]	I3D+AlphaPose	✓	✓	✓	**95.7**	96.3
Actor-Transformer[6]	I3D+HRNet	✓		✓	93.5	91.0
TCE+STBiP[43]	Inception-v3+HRNet	✓		✓	94.1	-(95.4*)
COMPOSER[46]	HRNet			✓	94.6	96.2
Ours(Box-branch‡)	Inception-v3	✓			92.6 (92.9*)	95.2 (96.7*)
Ours(Pose-branch†)	HRNet			✓	91.3 (91.2*)	91.6 (89.1*)
Ours(PDGCN)	Inception-v3+HRNet	✓		✓	93.2 (93.4*)	**96.9** (96.0*)
	ResNet-18+HRNet	✓		✓	92.0 (92.3*)	93.6 (91.9*)

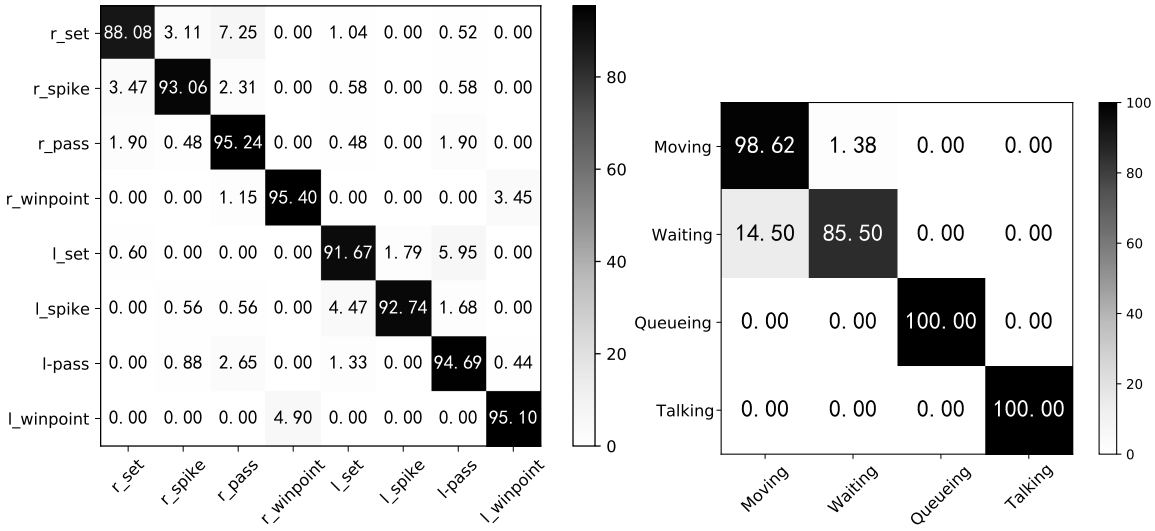
在表 4.2 中，'-' 符号表示不存在，对于带有上标'*' 的结果表示 MPCA 数值，因为

有些方法只提供了 MPCA 而没有提供 MCA；上标'†' 表示仅用姿态分支预测的结果，上标'‡' 表示仅用外观分支预测的结果，即对应第 3 章中的模型；输入模态 (modality) 一列中的 RGB 表示视频 RGB 图像特征，Flow 表示光流 (optical flow) 特征，Keypoint 表示姿态特征。从表 4.2 中可以得到，对于数据集 VD，在只有 RGB 和 Keypoint 作为输入模态的条件下，我们的 PDGCN 方法能够取得和一些先进方法竞争的结果，例如方法 Actor-Transformer[6]和 TCE+STBiP[43]，它们都是基于 Transformer 的方法。但是与最先进的方法仍有较大的差距，例如与 3 种模态输入的 GroupFormer[44] MCA 分数相差最大，达到了 2.5%，这说明了 PDGCN 在数据集 VD 上不能很好地区分某些行为类别的时序动态变化。一方面可能是由于 CDGCN 推理模块对时空关系的处理不够精细，导致学习到不可靠的模式；另一方面对姿态特征的生成可能有些粗糙，只是简单的将个体关节点坐标进行线性嵌入，导致没有学习到具有强大辨别能力的姿态表征。由于这些不足，使得 PDGCN 对这些时空交互关系的建模具有局限性。

对于数据集 CAD，我们的 PDGCN 方法显示出了它的优越性，并达到了最先进的 MCA 结果。PDGCN 将 Inception-v3+HRNet 作为 backbone，使用 RGB 特征和姿态特征作为输入，对比之前的大多数方法取得了明显的提升，甚至还超过了一些使用额外光流输入的方法，如 GroupFormer[44]和 SACRF[7]。特别地，在数据集 CAD 上，PDGCN 对比 GroupFormer[44]超出了 0.6% 的 MCA 分数，对比具有相同 backbone 的 TCE+STBiP[43]超出了 0.6% 的 MPCA 分数，这说明了 PDGCN 利用双分支进行时空关系推理的优越性。此外，在这两个数据集上，基于 Inception-v3+HRNet 配置的 PDGCN 同时超过外观分支或者姿态分支单独预测的结果，这表明了两种分支可以互补地融合在一起，提升最终群体行为预测的准确率。

基于 Inception-v3+HRNet 配置的 PDGCN 模型，在数据集 VD 和 CAD 上的混淆矩阵分别如图 4.5a 和 4.5b 所示。对于数据集 VD，对应预测结果的混淆矩阵表明，由于 PDGCN 中的两个分支对场景上下文信息的学习和对空间关系的动态推理，使得模型可以准确地区分左边的群体行为和右边的群体行为。与图 3.6a 中的混淆矩阵相比，PDGCN 对行为类别 *pass* 的预测表现更好，这可能是因为姿态分支增强了个体之间时空交互的语义信息，因为群体行为 *pass* 涉及到一个传球的人和一个接球的人。此外，在数据集 VD 中大多数预测失败的案例与图 3.6a 中相似，即将一支队伍执行的行为误当作同一支队伍执行的另一种行为，这可能是由于这些行为类别中的个体具有相似的场景上下文和交互关系，导致姿态分支没有起到很多积极作用。

对于数据集 CAD，从预测结果的混淆矩阵中可以看出，大多数预测失败的情形是将类别 *waiting* 预测为 *moving*，一方面可能由于这两种行为中的个体具有相似



(a) 在数据集 VD 上的混淆矩阵　　(b) 在数据集 CAD 上的混淆矩阵

图 4.5 PDGCN 在数据集 VD 和 CAD 上的混淆矩阵

的外观特征，另一方面由于 CAD 数据集中的视频片段没有足够的时序动态，导致难以区分这两类行为。

4.5.4 消融研究分析

在本小节中，为了验证 PDGCN 中姿态分支的有效性，对 PDGCN 进行消融研究分析。消融研究中所有的实验在数据集 VD 上进行，PDGCN 的 backbone 设置为 Inception-v3+HRNet，消融对比实验的结果如表 4.3 所示。

表 4.3 PDGCN 不同变体模型的消融实验结果

h_s for PB	h_c for PB	MCA(%)	MPCA(%)
-	-	92.6	92.9
1	1	92.9	93.1
2	2	92.8	93.0
4	4	**93.2**	**93.4**
8	8	92.8	93.2

在表 4.3 中，h_s 指 SCE 的注意力头数量，h_c 指 MHCDGCN 的注意力头数量，PB 表示姿态分支。表 4.3 中的第一行表示外观分支基线模型，即对应第三章的模型，它的最佳配置为 $h_s = 4$ 和 $h_c = 4$。我们将不同注意力头数量 (h_s 和 h_c) 的姿态

分支和最佳的外观分支融合，发现当姿态分支上的 $h_s = 4$ 和 $h_c = 4$ 的时候 PDGCN 表现最好。此外，根据表 4.3 可以得到：(1) 姿态分支的注意力头数量过少，通过姿态特征学习全局上下文信息以及推理交互关系的能力会不足，同时注意力头的数量过多会导致信息冗余。(2) 具有两个分支的 PDGCN 的 MCA 分数超出单个的外观分支，这论证了姿态分支可以有效地在上下文编码过的姿态特征上推理时空关系，使得 PDGCN 可以在姿态分支和外观分支上学习到互补的特征，从而提升了模型的预测性能。

4.6 本章小结

本章通过混合姿态特征提出了 PDGCN，利用姿态分支和外观分支分别学习场景上下文信息以及推理个体之间的时空关系，然后将两个分支融合进行最终的群体行为预测。此外，在两个广泛使用的数据集上的实验，论证了提出方法的有效性，并且取得了较好的结果。

5 总结与展望

5.1 研究总结

群体行为识别是视频理解和分析中一项重要且具有挑战的任务，最近已经获得了越来越多研究者的关注。随着深度学习理论技术的发展和硬件算力的提升，群体行为识别算法在基准数据集上的预测精度得到了显著的提升。本文的研究对象是群体行为识别，该任务比视频中对个体动作的识别更加复杂，因为要同时考虑个体所在的场景上下文和他们之间的时空交互关系。本文首先对群体行为识别领域的国内外研究现状进行介绍，接着对该领域中涉及到的算法和一些相关理论技术进行总结，加深对群体行为识别任务的理解，为后面模型的改进做铺垫。通过对国内外已有成果的研究，本文首先从学习场景上下文信息和推理个体之间的时空关系出发，提出了基于跨时间步动态图卷积的方法；然后，融合了个体的姿态特征，提出了混合姿态特征的双分支群体行为识别方法 PDGCN。具体的研究内容总结如下：

(1) 由于之前的一些工作直接在个体的外观特征上进行关系推理，然后从这些具有时空关系的个体特征生成群体表征进行预测，而忽视了对全局场景上下文信息的学习，这样会丢失一些有用的线索。所以，在推理个体之间的时空关系之前，需要充分地探索场景中相关的上下文信息，于是利用场景上下文编码模块将该信息编码到行为者的外观特征中。另一方面，在推理个体的时空交互关系时，不仅需要考虑同一时间步内个体之间的 intra-spatial 关系，而且需要考虑跨时间步个体之间的 inter-spatial 关系。因此本文基于动态图卷积来学习这两种空间关系，这样可以关注到跨时间步个体的时序动态，更加有效且相关地推理他们之间的时空交互关系，进而得到具有丰富语义和信息的个体表征。在两个广泛使用的数据集即 Volleyball[11]和 Collective Activity[21]上的实验结果，验证了该方法 (将其称之为外观分支) 的有效性。

(2) 因为人类的动作大部分和关节点的位置以及运动高度相关，通过关节点的动态变化可以识别出视频中个体的具体动作，所以可以使用关节点的位置来表达图片中个体的特征 (即姿态特征)，从而利用姿态特征在姿态分支上进行预测。本文提出了基于姿态特征的动态图卷积网络 PDGCN，分别使用姿态分支和外观分支学习场景上下文信息和推理个体的时空交互关系，随后将两个分支融合进行最终的群体行为预测。在 Volleyball 数据集和 Collective Activity 数据集上的实验结果表明了 PDGCN 的有效性和合理性。

5.2 未来工作展望

尽管现有的很多方法可以利用多种模态输入进行群体行为识别，并且能达到很高的精度 (在数据集 Volleyball 达到 95% 以上)，但是对于实际应用仍有较长距离，这由于实际场景的复杂性以及对实时性需求。另一方面，对于使用光流输入的方法一般具有很高的计算复杂度，对应的模型体量较大，所以群体行为识别的性能仍有一些提升空间。下面对未来的群体行为识别 (GAR) 方向做出展望。

(1) 大规模的 GAR 数据集。目前主流的群体行为数据集为 Volleyball 数据集和 Collective Activity 数据集，但是相比于目标检测、图像分类等任务，这两个数据集包含的可用样本数量并不大，对于最大的 Volleyball 数据集仅有不到 5K 个数据样本，使用到的图片数量不超过 50K。另一方面，数据集包含的群体行为类别数量不多，例如 Volleyball 数据集中只有 8 种群体行为。这对于那些参数量超多的模型，很难训练出最好的结果，或者陷入过拟合。所以，在未来研究一个大规模的 GAR 数据集是一个比较重要的需求。

(2) 统一的建模框架。GAR 是一项复杂的任务，它可能包含多个子任务，例如通过目标检测任务提取个体边界框，通过姿态估计预测关节点坐标。目前大多数的方法都是分阶段进行这些子任务，然后再进行 GAR 这个下游任务。显然，这些子任务方法的性能会影响到最终 GAR 模型的表现，这也意味着 GAR 模型最终的结果容易达到局部最优，而不是和这些子任务一起达到全局最优。所以，在未来 GAR 的一个可能发展方向是，通过一个统一的建模框架同时训练上游的子任务和最终的 GAR 任务。

参考文献

[1] 吴建超, 王利民, 武港山. 视频群体行为识别综述[J]. 软件学报, 2023, 34(02): 964–984.

[2] 裴利沈, 赵雪专. 群体行为识别深度学习方法研究综述[J]. 计算机科学与探索, 2022, 16(04): 775–790.

[3] CHOI W, SAVARESE S. Understanding collective activitiesof people from videos[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 36(6): 1242–1257.

[4] WU L F, WANG Q, JIAN M, et al. A comprehensive review of group activity recognition in videos[J]. International Journal of Automation and Computing, 2021, 18: 334–350.

[5] 裴利沈, 赵雪专. 融合时间和空间上下文特征的群体行为识别[J]. 智能计算机与应用, 2022, 12(09): 45–49.

[6] GAVRILYUK K, SANFORD R, JAVAN M, et al. Actor-transformers for group activity recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 839–848.

[7] PRAMONO R R A, CHEN Y T, FANG W H. Empowering relational network by self-attention augmented conditional random fields for group activity recognition[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. 2020: 71–90.

[8] YUAN H, NI D, WANG M. Spatio-temporal dynamic inference network for group activity recognition[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 7476–7485.

[9] EHSANPOUR M, ABEDIN A, SALEH F, et al. Joint learning of social groups, individuals action and sub-group activities in videos[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16. 2020: 177–195.

[10] BAGAUTDINOV T, ALAHI A, FLEURET F, et al. Social scene understanding: End-to-end multi-person action localization and collective activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4315–4324.

[11] IBRAHIM M S, MURALIDHARAN S, DENG Z, et al. A hierarchical deep temporal model for group activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 1971–1980.

[12] VASWANI N, CHOWDHURY A R, CHELLAPPA R. Activity recognition using the dynamics of the configuration of interacting objects[C]//2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings. Vol. 2. 2003: II–633.

[13] KHAN S M, SHAH M. Detecting group activities using rigidity of formation[C]// Proceedings of the 13th annual ACM international conference on Multimedia. 2005: 403–406.

[14] ZHOU Y, NI B, YAN S, et al. Recognizing pair-activities by causality analysis[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2011, 2(1): 1–20.

[15] YIN Y, YANG G, XU J, et al. Small group human activity recognition[C]//2012 19th IEEE International Conference on Image Processing. 2012: 2709–2712.

[16] KIM Y J, CHO N G, LEE S W. Group activity recognition with group interaction zone[C]//2014 22nd international conference on pattern recognition. 2014: 3517–3521.

[17] TRAN K N, GALA A, KAKADIARIS I A, et al. Activity analysis in crowded environments using social cues for group discovery and human interaction modeling[J]. Pattern Recognition Letters, 2014, 44: 49–57.

[18] CHANG M C, KRAHNSTOEVER N, LIM S, et al. Group level activity recognition in crowded environments across multiple cameras[C]//2010 7th IEEE International Conference on Advanced Video and Signal Based Surveillance. 2010: 56–63.

[19] ZHA Z J, ZHANG H, WANG M, et al. Detecting group activities with multi-camera context[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2012, 23(5): 856–869.

[20] DAI P, DI H, DONG L, et al. Group interaction analysis in dynamic context[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2008, 38(1): 275–282.

[21] CHOI W, SHAHID K, SAVARESE S. What are they doing?: Collective activity classification using spatio-temporal relationship among people[C]//2009 IEEE 12th international conference on computer vision workshops, ICCV Workshops. 2009: 1282–1289.

[22] LAN T, WANG Y, MORI G, et al. Retrieving actions in group contexts[C]//Trends and Topics in Computer Vision: ECCV 2010 Workshops, Heraklion, Crete, Greece, September 10-11, 2010, Revised Selected Papers, Part I 11. 2012: 181–194.

[23] AMER M R, LEI P, TODOROVIC S. Hirf: Hierarchical random field for collective activity recognition in videos[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VI 13. 2014: 572–585.

[24] CHANG X, ZHENG W S, ZHANG J. Learning person–person interaction in collective activity recognition[J]. IEEE Transactions on Image Processing, 2015, 24(6): 1905–1918.

[25] KHAMIS S, MORARIU V I, DAVIS L S. Combining per-frame and per-track cues for multi-person action recognition[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part I 12. 2012: 116–129.

[26] CHOI W, SAVARESE S. A unified framework for multi-target tracking and collective activity recognition[C]//Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part IV 12. 2012: 215–230.

[27] GRAVES A. Long short-term memory[J]. Supervised sequence labelling with recurrent neural networks, 2012: 37–45.

[28] WU L, YANG Z, HE J, et al. Ontology-based global and collective motion patterns for event classification in basketball videos[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2019, 30(7): 2178–2190.

[29] WANG M, NI B, YANG X. Recurrent modeling of interaction context for collective activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 3048–3056.

[30] SHU T, TODOROVIC S, ZHU S C. CERN: confidence-energy recurrent network for group activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 5523–5531.

[31] LI X, CHOO CHUAH M. Sbgar: Semantics based group activity recognition[C]// Proceedings of the IEEE international conference on computer vision. 2017: 2876–2885.

[32] KIM P S, LEE D G, LEE S W. Discriminative context learning with gated recurrent unit for group activity recognition[J]. Pattern Recognition, 2018, 76: 149–161.

[33] SHU X, ZHANG L, SUN Y, et al. Host–parasite: Graph LSTM-in-LSTM for group activity recognition[J]. IEEE transactions on neural networks and learning systems, 2020, 32(2): 663–674.

[34] IBRAHIM M S, MORI G. Hierarchical relational networks for group activity recognition and retrieval[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 721–736.

[35] QI M, QIN J, LI A, et al. Stagnet: An attentive semantic rnn for group activity recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 101–117.

[36] KIPF T N, WELLING M. Semi-supervised classification with graph convolutional networks[J]. ArXiv preprint arXiv:1609.02907, 2016.

[37] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[38] HE K, GKIOXARI G, DOLLÁR P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961–2969.

[39] WU J, WANG L, WANG L, et al. Learning actor relation graphs for group activity recognition[C]//Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. 2019: 9964–9974.

[40] YAN R, XIE L, TANG J, et al. Social adaptive module for weakly-supervised group activity recognition[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VIII 16. 2020: 208–224.

[41] HU G, CUI B, HE Y, et al. Progressive relation learning for group activity recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020: 980–989.

[42] 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.

[43] YUAN H, NI D. Learning visual context for group activity recognition[C]// Proceedings of the AAAI Conference on Artificial Intelligence: vol. 35: 4. 2021: 3261–3269.

[44] LI S, CAO Q, LIU L, et al. Groupformer: Group activity recognition with clustered spatial-temporal transformer[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021: 13668–13677.

[45] HAN M, ZHANG D J, WANG Y, et al. Dual-AI: dual-path actor interaction learning for group activity recognition[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 2990–2999.

[46] ZHOU H, KADAV A, SHAMSIAN A, et al. COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality[C]//Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV. 2022: 249–266.

[47] 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.

[48] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6): 84–90.

[49] SIMONYAN K, ZISSERMAN A. Very deep convolutional networks for large-scale image recognition[J]. ArXiv preprint arXiv:1409.1556, 2014.

[50] SZEGEDY C, VANHOUCKE V, IOFFE S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2818–2826.

[51] SZEGEDY C, LIU W, JIA Y, et al. Going deeper with convolutions[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 1–9.

[52] HAMILTON W, YING Z, LESKOVEC J. Inductive representation learning on large graphs[J]. Advances in neural information processing systems, 2017, 30.

[53] VELIČKOVIĆ P, CUCURULL G, CASANOVA A, et al. Graph attention networks[J]. ArXiv preprint arXiv:1710.10903, 2017.

[54] DEFFERRARD M, BRESSON X, VANDERGHEYNST P. Convolutional neural networks on graphs with fast localized spectral filtering[J]. Advances in neural information processing systems, 2016, 29.

[55] HENAFF M, BRUNA J, LECUN Y. Deep convolutional networks on graph-structured data[J]. ArXiv preprint arXiv:1506.05163, 2015.

[56] WU Z, PAN S, LONG G, et al. Graph wavenet for deep spatial-temporal graph modeling[J]. ArXiv preprint arXiv:1906.00121, 2019.

[57] LI Y, YU R, SHAHABI C, et al. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting[J]. ArXiv preprint arXiv:1707.01926, 2017.

[58] GUO S, LIN Y, FENG N, et al. Attention based spatial-temporal graph convolutional networks for traffic flow forecasting[C]//Proceedings of the AAAI conference on artificial intelligence: vol. 33. 2019: 922–929.

[59] GUO S, LIN Y, WAN H, et al. Learning dynamics and heterogeneity of spatial-temporal graph data for traffic forecasting[J]. IEEE Transactions on Knowledge and Data Engineering, 2021.

[60] CHUNG J, GULCEHRE C, CHO K, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[C]//NIPS 2014 Workshop on Deep Learning, December 2014. 2014.

[61] BA J L, KIROS J R, HINTON G E. Layer normalization[J]. ArXiv preprint arXiv:1607.06450, 2016.

[62] YAN R, XIE L, TANG J, et al. HiGCIN: Hierarchical graph-based cross inference network for group activity recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

[63] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. ArXiv preprint arXiv:1412.6980, 2014.

[64] BAGAUTDINOV T, ALAHI A, FLEURET F, et al. Social scene understanding: End-to-end multi-person action localization and collective activity recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4315–4324.

[65] AZAR S M, ATIGH M G, NICKABADI A, et al. Convolutional relational machine for group activity recognition[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 7892–7901.

[66] 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.

[67] LIN T Y, MAIRE M, BELONGIE S, et al. Microsoft COCO: Common objects in context[C]//Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. 2014: 740–755.

攻读学位期间发表的学术论文

[1] YAO H, YANG J, XIE Z, et al. PCKT: Problem Composition in Knowledge Tracking[C]//Proceedings of the 14th International Conference on Education Technology and Computers. 2022: 442–448.

[2] YAO H, CHEN R, XIE Z, et al. MRA-DGCN: Multi-Range Attention-Based Dynamic Graph Convolutional Network for Traffic Prediction[C]//2022 IEEE International Conference on Big Data (Big Data). 2022: 1613–1621.

致　谢

樱花飘落，论文临付梓之际，谨向所有帮助、关心和支持我的人致以衷心的谢意。

首先感谢的是我的导师姚华雄老师，姚老师在我的研究生阶段一直给予我悉心的指导，他是一位非常负责任和态度严谨的导师。姚老师给我的学术生涯指明了研究方向，提供了良好的科研条件，我在和他平常的交流中受益匪浅。祝姚老师桃李满天下。同时还要感谢学院辅导员蔡庆昱老师，蔡老师处理了很多繁琐的学生工作，才可以让我们每一个人专心地进行科研任务。

其次感谢室友李响同学，我和他经常一起交流生活和学习上的问题，他是一个积极向上具有规划的人，和他的相处让我受到了很多正面的影响，同时给我的生活增添了几分乐趣。还要感谢我的一些同门，虽然与他们的研究方向各有差异，但是从他们研究的领域可以让我获得更加宽广的研究思路。

我还要感谢我的家人，感谢他们一直辛勤的付出和一直以来的支持关心，他们是我最坚强的后盾。同时也感谢自己，能够一直积极地生活，努力地完成学业。

此外，还感谢华中师范大学大学给我们提供了美丽的校园环境，以及配置了高端的综合楼学习环境。

最后，感谢各位评审专家宝贵的评审意见，在此致以由衷的感谢。

ShelterX

South Lake in May 2023