分 类 号_____　　学号__M2019xxxxx__
学校代码__10487__　　密级_____

華中科技大學

硕士学位论文

（学术型□　　专业型□）

标题：宋体，英文 Times New Roman，一号，加粗，不超 30 字

(中英文标题、学科专业、导师姓名正确、一致)

学位申请人：　XXX

学 科 专 业：　XXXXX

指 导 教 师：　XXX　教授

答 辩 日 期：　202X 年 X 月 X 日

A Thesis Submitted in Partial Fulfillment of the Requirements for the Master Degree in Engineering

English Title，Times New Roman，小二号，实词的首字母大写

中英文标题、学科专业、导师姓名正确、一致

Candidate ： **xxx（中文习惯，姓在前且姓全部大写）**

Major ： **Control Science and Engineering**

Supervisor ： **Prof. Xxxx**

Huazhong University of Science and Technology

Wuhan 430074, P. R. China

February, 2026

独创性声明

本人声明所呈交的学位论文是我个人在导师的指导下进行的研究工作及取得的研究成果。尽我所知，除文中已标明引用的内容外，本论文不包含任何其他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。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：

日期：　　年　　月　　日

学位论文版权使用授权书

本学位论文作者完全了解学校有关保留、使用学位论文的规定，即：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，允许论文被查阅和借阅。本人授权华中科技大学可以将本学位论文的全部或部分内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

本论文属于　保　密 □，在 ____ 年解密后适用本授权书。

　　不保密 □。

（请在以上方框内打“√”）

学位论文作者签名：　　指导教师签名：

日期：　　年　　月　　日　　日期：　　年　　月　　日

摘 要

摘要是学位论文极为重要、不可缺少的组成部分，它是论文的窗口，并频繁用于国内外资料交流、情报检索、二次文献编辑等。其性质和要求如下：

[1] 摘要即摘录论文要点，是论文要点不加注释和评论的一篇完整的陈述性短文，具有很强的自含性和独立性，能独立使用和被引用。

[2] 博士学位论文的摘要应包含全文的主要信息，并突出创造性成果。

[3] 内容范围应包含以下基本要素：

(1) 目的：研究、研制、调查等的前提、目的和任务以及所涉及的主题范围。

(2) 方法：所用原理、理论、条件、对象、材料、工艺、手段、装备、程序等。

(3) 结果：实验的、研究的、调查的、观察的结果、数据，被确定的关系，得到的效果、性能等。

(4) 结论：结果的分析、研究、比较、评价、应用；提出的问题，今后的课题，建议，预测等。

(5) 其他：不属于研究、研制、调查的主要目的，但就其见识和情报价值而言也是重要的信息。

[4] 摘要的详简度视论文的内容、性质而定，硕士学位论文摘要一般为 500 －600 汉字。

[5] 摘要及全文中均建议不出现“我们”等字样。摘要中主语（作用）常常省略，因而一般使用被动语态；应使用正确的时态，并要注意主、谓语的一致，必要的冠词不能省略。

[6] 一般不用图、表、化学结构式、计算机程序，不用非公知公用的符号、术语和非法定的计量单位。

[7] 摘要中一般不使用缩写词，若实在需要，在第一次使用前，需给出中文全称（缩写词）；在使用英文缩写词之前，需给出英文全称（英文全称，缩写词），再次出现时可以采用中文或英文缩写词。

[8] 关键词应有 3 至 8 个，另起一行置于摘要下方，领域从大到小排列。关键词之间用分号隔开，最后一个关键词后面无标点。

[9] 摘要、关键词采用中文宋体；英文 Times New Roman；小四号。

[10] 应有与中文摘要和关键词相对应的英文摘要和关键词。英语摘要用词应准确，使用本学科通用的词汇。

关键词：关键词 1；关键词 2；关键词 3

Abstract

This is abstract.

英文摘要字体为 Times New Roman，小四，1.5 倍行距。

英文摘要和关键词应与中文相对应。英语摘要用词应准确，使用本学科通用的词汇；摘要中主语（作用）常常省略，因而一般使用被动语态；应使用正确的时态，并要注意主、谓语的一致，必要的冠词不能省略。

Keywords: Keyword1, Keyword2, Keyword3

目 录

主要符号对照表

xue	我的姓
ruini	我的名
W.M. Zheng	我的老师
Tsinghua	学校名
Long	来个比较长的，看看会出现什么情况。
劝学	君子曰：学不可以已。青，取之于蓝，而青于蓝；冰，水为之，而寒于水。木直中绳。（车柔）以为轮，其曲中规。虽有槁暴，不复挺者，（车柔）使之然也。故木受绳则直，金就砺则利，君子博学而日参省乎己，则知明而行无过矣。吾尝终日而思矣，不如须臾之所学也；吾尝（足齐）而望矣，不如登高之博见也。登高而招，臂非加长也，而见者远；顺风而呼，声非加疾也，而闻者彰。假舆马者，非利足也，而致千里；假舟楫者，非能水也，而绝江河，君子生非异也，善假于物也。积土成山，风雨兴焉；积水成渊，蛟龙生焉；积善成德，而神明自得，圣心备焉。故不积跬步，无以至千里；不积小流，无以成江海。骐骥一跃，不能十步；驽马十驾，功在不舍。锲而舍之，朽木不折；锲而不舍，金石可镂。蚓无爪牙之利，筋骨之强，上食埃土，下饮黄泉，用心一也。蟹六跪而二螯，非蛇鳝之穴无可寄托者，用心躁也。—— 荀况

1 绪论

1.1 研究背景与意义

随着移动设备的普及和互联网技术的快速发展，视频内容数据呈现爆炸式增长。视频由于其直观、生动的表现形式，已成为信息传播和交流的重要媒介。根据中国视听大数据（CVB）[1] 统计显示，全国卫视频道体育赛事播出总场次 43329 场，其中直播赛事 4901 场，全国累计收视规模超 247.5 亿人次，累计收视时长突破 66.6 亿小时。面对海量的体育场景视频，传统的人工视频分析技术已无法满足人们的需求，针对体育场景的智能化视频内容理解技术已成为研究人员关注的重点。

人工智能技术与体育竞技进行深度融合是行业发展的趋势。根据国家体育总局相关报道[2]，“体育 + 人工智能”行动将被纳入《“十五五”体育科教发展规划》，将汇聚各方科技力量大力推进人工智能在体育行业的应用。在体育运动领域中，基于人工智能的视频分析技术已经在运动员训练辅助、体育赛事分析、赛事转播等多个场景都有丰富的实际应用并发挥着重要作用。

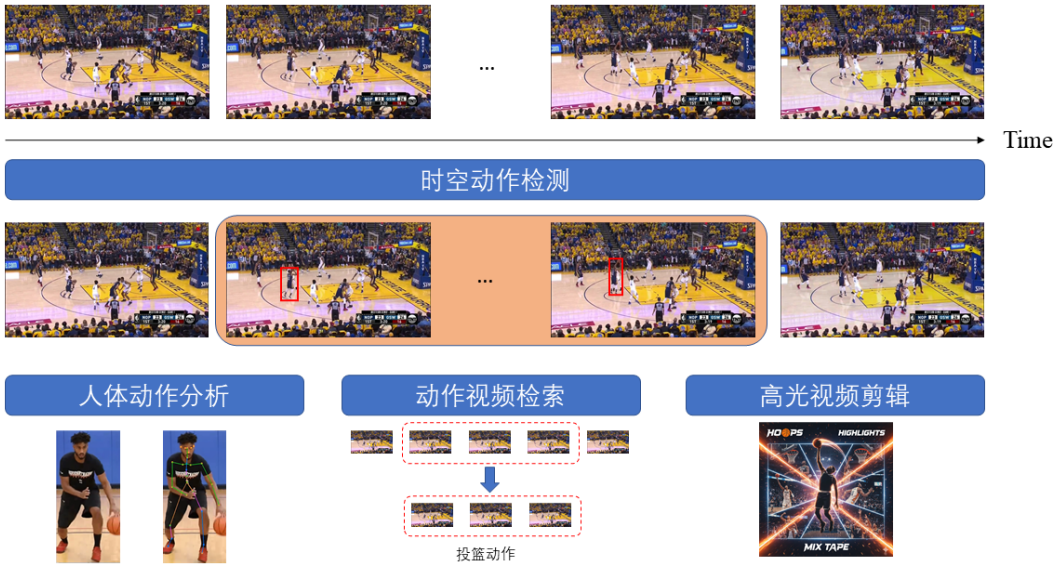


图 1.1 体育时空动作检测的应用

时空动作检测作为视频理解领域下的关键技术之一，能在未经剪辑的视频中识别出关注的动作类别、定位动作发生的起始帧与结束帧，并给出动作主体的空间位

置，同时实现时间和空间维度的定位，已在包括智能安防、自动驾驶、虚拟现实等领域有着广泛的应用。近年来，针对体育视频的时空动作检测技术得到了广泛关注，由于其可以从海量未剪辑的体育运动视频中，精确地定位出特定的技术动作视频片段，并识别出运动员的动作类别和运动主体的位置，为运动员训练、体育赛事分析、高光视频生成等提供有力的数据支持，为诸多下游应用提供了技术基础。

不同于常规的安防场景，体育运动场景下的时空动作检测有着多重挑战。对于体育运动场景，运动的动作通常复杂多变、运动员运动剧烈、速度快，动作过程中伴随大范围的形变和位移，且在运动竞技过程中难免存在运动主体相互遮挡的情况。根据相关统计[3]，体育动作场景的动作内复杂度和动作间复杂度是日常动作的 3-8 倍，为此，如何提高复杂运动情况下的帧间目标一致性是体育视频时空动作检测的一大关键。此外，体育运动视频中动作类别繁多，部分技术动作之间的视觉细微差异也会对时空动作检测带来困难，如足球中的传球和射门、篮球中的二分球和三分球，其动作特征十分相似的，主要的区别在于动作发生的位置和动作的意图，这要求模型在理解动作运动特征之外，还需要对全局视频的时空视觉线索和动作间的细节差异进行充分的挖掘，只有对包括场地信息、运动员信息和专业动作高级语义等信息进行充分关系建模，才能准确检测体育场景下的动作类型。

综上所述，针对体育运动场景的时空动作检测技术的研究，不仅有具体的实用价值，而且具有重要的学术研究意义。在应用方面，体育运动场景的时空动作检测技术可以提升体育视频分析的效率和精度，不仅可以通过对运动人员的动作进行精确的分析，切实地指导训练而提高运动员的竞技能力，为赛事转播提供智能化的技术支持，提升观众的观赛体验；在学术研究方面，深入研究体育视频时空动作检测技术，有助于推动视频理解相关技术在处理复杂运动场景视频的能力。因此，本文针对体育运动场景下的时空动作检测技术展开深入研究。

1.2 国内外研究现状

1.2.1 时空动作检测研究现状

时空动作检测是动作识别任务和时序动作检测任务的延伸。动作识别旨在对一段视频片段中发生的动作进行分类，仅关注这段视频发生了什么动作，属于基础的视频分类任务；时序动作检测则需要在动作识别的基础上，确定动作在视频中发生

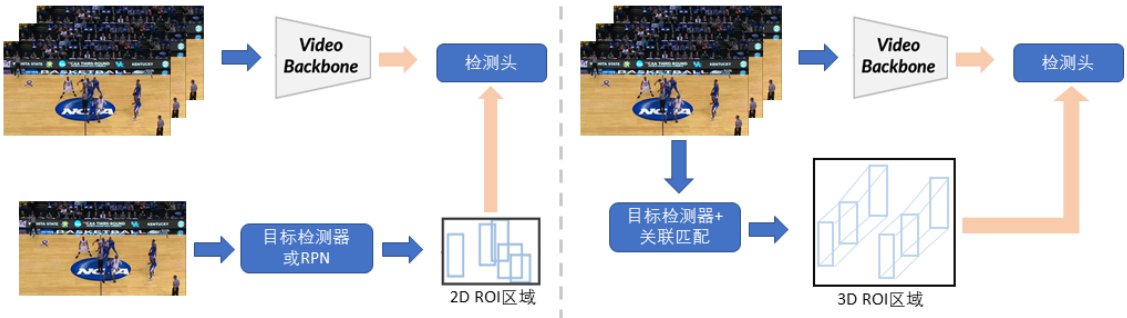
的时间区间，同时实现动作分类和时间维度的定位；而时空动作检测不仅仅需要知道视频动作的类别和发生区间，还需要在空间维度上对动作目标主体进行定位，给出动作发生过程中在视频画面中的位置，从而实现时间维度和空间维度的定位。

近年来，随着深度学习技术的快速发展，时空动作检测技术取得了显著进展。现有的时空动作检测方法根据算法结构范式可以主要分为双阶段方法和单阶段方法两大类，本节将分别介绍双阶段和单阶段时空动作检测方法的研究现状。

（1）双阶段时空动作检测

双阶段时空动作检测检测方法通常会依赖于额外的目标检测器或区域候选网络（Region Proposal Network，RPN）[4]，预先生成的候选动作主体的 ROI（Region of interest）区域，再通过对 ROI 区域进行时空特征提取和环节关系建模完成时空动作检测任务。

如图**??**所示, 根据所使用的 ROI 区域的是否存在时间维度，可以将双阶段方法分为帧级和片段级。前者通常会利用关键帧上的所产生的 ROI 区域尝试与全局视觉特征进行交互融合，从而提升模型的环境理解能力；后者则更加关注于如何生成高质量的动作管级别 3D 候选区域，以提升模型的动作理解能力。



（a）帧级双阶段时空动作检测器　　（b）片段级双阶段时空动作检测器

图 1.2 双阶段时空动作检测算法可分为（a）帧级检测方法，（b）片段级检测方法

帧级的双阶段方法受益于目标检测任务的发展，SAHA 等人[5] 首次在时空动作检测任务中引入了基于 Fast R-CNN[4] 的 RPN 网络来代替传统的无监督区域生成算法，实现了对动作主体的高效定位，他们将 RGB 图像和光流图像分别输入到两个独立的 RPN 网络，以输出检测框和动作类别得分。ACAM[6] 通过设计一种关系建模模块，对由 RPN 网络生成的候选区域特征与全局特征图进行交互建模，提升了模型对环境相关动作的检测能力。MRSN[7] 则是基于 Vision transformer[8] 架构，将全局视

频特征与局部候选区域特征进行 Patch 化处理，通过 Transformer 机制[9] 进行运动员特征与全局视觉特征的交互融合，增强了模型的场景理解能力。HIT[10] 额外引入了人体关键点和手部区域信息，通过融合人体-手部-关键点三重特征，提升了模型对细粒度动作的识别能力。EVAD[11] 通过设计了基于关键帧的 Token dropout 机制，来提升模型的推理效率。

片段级的双阶段方法则是更加关注于如何在进行关系建模之前进行帧间目标关联，以生成高质量的动作管级别 3D 候选区域。CFAD[12] 提出了一种从粗到精获取 3D ROI 的新范式，通过粗略模块对长时域信息进行参数化建模，先从视频流中估算出初步的动作管，随后再利用精细模块，在关键时间戳的引导下，对初步估算的管柱位置进行选择性调整和细化。TrAD[13] 则是通过利用额外的目标跟踪器生成 TOI 区域获得高质量的动作管候选区域，为动作分类提供了高质量的空间信息先验。ART[3] 则是参考了目标跟踪中的相关性学习方法，通过计算帧间的 query 相似度，设计了 query 层级的匹配机制，进而生成 3D 片段级的 query，隐式地实现了帧间目标关联和动作管生成。

（2）单阶段时空动作检测

近年来单阶段目标检测器的快速发展，如 YOLO[14]、CenterNet[15]、Sparse R-CNN[16]、DETR[17] 等范式通过端到端的模型设计，高效地实现了视觉定位任务。时空动作检测方法受此启发，也产生了很多单阶段时空动作检测方法，直接对视频帧进行时空特征提取和动作检测，无需额外的候选区域生成步骤。

YOWO[18] 是参考 YOLO 首次将单阶段目标检测器思想引入时空动作检测任务的方法，其设计了双分支视觉特征提取网络，分别使用一个 2D backbone 和一个 3D backbone 来提取空间和时序特征，并通过信息融合模块将两者进行融合，最终通过单个检测头实现动作分类和位置信息的回归。MOC[19] 则是参考 CenterNet[15] 的思想，设计了一种自上而下的基于中心点时空动作检测方法，通过预测输入片段关键帧的中心点和其它帧中心点的相对偏移量，以实现了对目标运动信息的提取并直接输出片段级的预测结果。Tuber[20] 则是首个将 DETR 范式引入时空动作检测任务的方法，通过设计时空动作查询机制，利用 Transformer 的自注意力机制对时空特征进行建模，实现了对动作类别和位置的直接预测。在此基础上，STAR[21] 则是受益于新型的视频预训练模型 Vivit[22]，并设计了时空解耦的时空动作查询并解耦时空注意力计算模块，在相关基线上取得了显著提升。STMixer[23] 通过设计自适应采样模块

和双分支时空特征融合模块，在高效地提取时空特征的同时，实现了对时空特征的充分融合，高效地实现了单阶段时空动作检测。STDet[24] 则是的核心在于使用可学习的管柱查询直接进行时空动作检测，它彻底舍弃了传统方法中繁琐的手工预设锚点和低效的帧间关联操作，通过直接对可以在更长的窗口内进行全局回归，这种长时域建模能力使得模型能够更精确地利用长期信息，并显式预测动作的时间边界。

1.2.2 体育视频时空动作检测研究现状

体育场景的视频理解算法研究不仅可以辅助运动员训练，提高观众的观赛体验，带来明显的经济效益，而且由于体育运动的特殊性，相比于一般的视频分析任务更具挑战性，因此近年来受到了广泛关注。

Khurram Soomro 等人[25] 最先关注到体育视频研究价值，根据 UCF101[26] 数据集衍生出了针对体育场景的时空动作检测数据集 UCF Sports，该数据集包含了 10 类体育运动动作，并提供了每个动作的时空标注信息，主要研究了基于传统手工特征的体育视频动作识别方法。随着相关研究的深入，越来越多更具挑战性的体育视频数据集被提出，如 Sports-1M[27]、FineGym[28]、SoccerNet[29] 等，涵盖了包括篮球、足球、体操等多种体育运动场景，为体育视频理解算法的研究提供了丰富的数据资源。

当前针对体育视频理解分析的研究多种多样[30–35]，如体育视频质量评估、体育视频问答大模型、AI 足球裁判等，极大推动了人工智能技术在体育视频理解方面的应用。LiuZiao 等人[35] 提出了 smartboard 视频助理裁判模型，结合大模型对足球比赛进行结合可视分析与体育数据分析，引入一个大模型智能作为中介，连接“人的意图”与“底层数据/可视化组件”，该系统会自动生成并排列多个可视化组件，形成一个直观的战术板用来帮助决策。Matchtime[33] 则是一个基于解说文本和视频画面信息进行时序对齐的大模型解说专家，针对解说内容通常滞后于画面信息的问题提出了一种新的时序匹配机制，通过对齐解说文本和视频画面，实现更加精准的解说内容定位。SportsGPT[34] 则是一个面向体育视频理解的大模型，主要针对图片视频的问答场景，通过引入多模态预训练模型和大规模体育视频数据集，实现了对体育视频内容的深度理解和智能问答。

针对体育场景下的时空动作检测任务研究，近些年也取得了显著的进展。TAAD[36] 针对体育视频场景下的时空动作检测任务，基于当前先进的目标跟踪器 YOLOv5-DeepSort[37]，并使用预训练的行人重识别基础模型 OsNet-x0-25[38] 作为特

征提取器进一步提高对运动员的外观判别能力，在进行时空动作检测前便实现了对运动员的高质量跟踪，从而提升了模型在复杂多人运动场景下的动作检测能力。PoSTAL[39] 则是创新性地引入了 BLIP[40] 多模态预训练模型，通过对体育视频中运动员的衣物颜色和号码等文本信息进行提取和融合，并通过设计的基于提示的动作编码器和动作管解码器直接预测动作管，提高了模型对细粒度篮球动作类别的区分能力。HHIDet[41] 则是针对篮球场景和网球场景中运动员交互关系复杂的问题，使用人体检测器获取视频中所有的人体提议，并提出了显式的交互提议生成和提议间的信息交换机制，以更好地捕捉主体和客体之间的空间几何关系。

1.3 存在的问题

虽然目前针对体育视频的时空动作检测技术已经取得了一定的进展，但仍然存在一些亟待解决的关键问题：

（1）专业体育技术动作的相似性导致准确检测存在困难。在体育运动场景下，很多技术动作从视觉上看来差异较小，比如网球场景的扣球和吊球，足球场景的传球和射门等，这些动作在视觉上往往存在较大的相似性，对于人类是基于一定的专业知识才能较为准确判别此类动作，而现在的时空动作的检测方法基本依赖于视觉特征进行动作理解，难以充分挖掘动作类别背后的高复杂语义信息，比如动作的意图和规则，这限制了模型对细粒度体育动作类别的性能。

（2）复杂运动状态的帧间信息关联存在困难。对于体育场景下动作的运动特点，通常存在高速运动、巨大形变、相互遮挡等复杂情况，现有的时空动作检测方法大多依赖于预训练的目标检测器或 RPN 网络生成候选区域在时间维度进行复制，难以应对大位移、强遮挡等情况，甚至会带来噪声干扰。虽然近年来有一些方法尝试通过目标跟踪等手段进行实例级别帧间关联，但大多难以适应体育场景下复杂的运动情况，导致生成的动作管质量不高，影响了后续的动作检测性能，且对于遮挡情况，3D ROI 由于几何约束的限制，十分容易引入干扰信息。

（3）运动主体和场景间的关系信息建模困难。体育运动视频中动作类别繁多，部分技术动作之间的视觉细微差异也会对时空动作检测带来困难，很多动作只有结合场景信息才能进行准确识别，尤其是在多人运动场景下，运动员与运动员之间、运动员与环境直接的交互关系复杂。现在的方法大多以运动员为中心，与环境进行注

意力计算以实现交互建模，或通过图神经网络等方式对视频中关键元素的 RPN 特征进行关系建模但是此类方法仅局限于局部区域的信息交互，难以实现对全局时空视觉信息的充分挖掘。

1.4 本文主要内容

针对以上时空动作检测技术在体育场景下存在的关键问题，本文进行了以下具体工作：

（1）基于多模态知识库增强的时空动作检测算法研究。针对专业体育技术动作的相似性导致准确检测存在困难的问题，本文设计了一种基于多模态知识库增强的时空动作检测算法，包括设计了一种基于多模态大模型构建体育知识库的方法，并基于多模态特征检索机制提升模型对专业体育动作的理解能力。

（2）基于时空一致性的时空动作检测算法研究。针对复杂运动状态下的帧间信息关联问题，本算法提出了通过设计了一种视频主题感知模块 TAM（Topic Aware Module）对输入视频片段进行全局主题特征压缩和提取，并通过主题特征引导的帧间信息关联实现对复杂运动状态下的帧间信息关联。针对体育动作的关系信息建模困难的问题，本文设计了一种自适应特征时空特征采样策略和可变形时空注意力模块，自适应提取视觉特征中的关键特征点，以提升模型对动作主体的特征提取能力。由此实现特征级的帧间信息关联，提升了模型在复杂运动状态下的动作检测能力。

（3）基于多模态特征引导的时空动作检测算法研究。结合前两项工作，本文提出了一种基于多模态特征引导的帧间一致性建模时空动作检测算法，通过多模态知识库提升模型对专业体育动作的理解能力，且通过帧间一致性建模提升模型在复杂运动状态下的动作检测能力，从而提升整体的体育视频时空动作检测性能。

本文共分为 6 章内容，章节内容之间的关系图如图 1.3所示。

第一章：绪论。在本章中，首先介绍了体育视频时空动作检测技术的研究背景与意义。其次，综述了国内外的体育运动场景下时空动作检测领域的研究现状，并分析和总结了现有相关研究中仍然存在的关键问题。最后针对相关关键问题，阐述了本文的研究内容，包括基于时空动作主题引导的时空动作检测算法和基于多模特征解耦的体育视频时空动作检测算法。

第二章：时空动作检测相关理论和技术。本章详细介绍了与本文工作相关的时

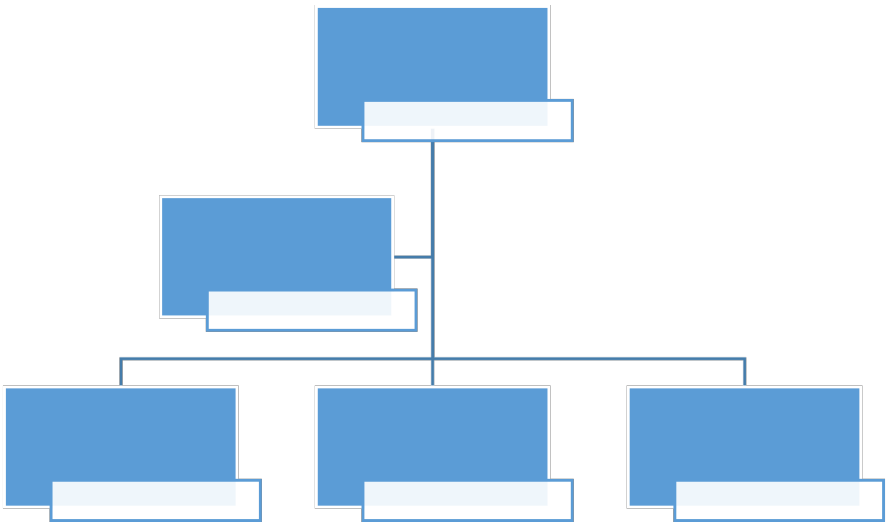


图 1.3 组织结构图

空动作检测相关理论和技术，包括时空动作检测的任务定义、基于查询的时空动作检测方法多模态信息知识库等技术。然后，我们通过对现有体育场景下时空动作检测数据集进行详细讨论，细致说明体育场景动作检测的特点。最后，介绍了时空动作检测的性能评价指标。

第三章：基于时空动作主题引导的时空动作检测方法。针对复杂运动状态的帧间空间信息关联，本章提出了基于时空动作主题特征引导策略，并结合自适应时空动作采样和可变形时空注意力灵活地进行帧间信息关联，以解决常规 ROI 几何约束引入噪声的问题。时空动作主题特征引导策略通过一个时空动作主题特征感知模块对动作主题特征进行提取，并通过指导时空特征采样来引导特征级别信息关联。自适应时空动作采样和可变形时空注意力提升了模型对动作主体的特征提取能力，进而提升了复杂运动状态下的时空动作检测性能。最后，通过对比实验验证了所提方法的优越性和有效性。

第四章：基于多模特征解耦的体育视频时空动作检测方法。针对多人体育动作情况的关系信息挖掘和体育运动视频的细粒度时空特征建模，本章提出了一种利用多模态大模型提取多模态特征构建多模态特征库的方法，以提升模型对复杂体育动作的理解能力。随后，为了使多模态特征可以更好地适配时空动作检测任务，设计了一种双分支时空特征解耦模块，以分别空间分支通过空间编码器提取关键帧的空间特征，时序分支通过时序编码器提取时序特征，并分别通过时间对齐和空间对齐模块进行特征对齐，显著提升了模型对细粒度动作的检测能力。最后，通过大量实

验验证了所提方法在体育视频时空动作检测任务中的有效性和优越性。

第五章：

第六章：总结与展望。本章总结了论文的主要内容和创新点，指出了当前工作的不足之处，并提出了未来研究的方向。

2 体育动作时空动作检测相关理论和技术

2.1 引言

本章首先明确了时空动作检测的任务定义，说明其相比于动作检测和时序动作检测任务的优势与挑战，阐释了其在视频中实现分类、时序定位和空间定位的任务目标（第 2.1 节）。随后，本章介绍了现有视频特征提取方法（第 2.2 节）。然后，基于现有的检测范式，本章分析重点基于查询机制的动作检测器的理论基础和相关方法（第 2.3 节）。最后，本章对相关领域的基准数据集和评价指标进行了介绍（第 2.4 节），并具体分析了体育场景下的时空动作检测的难点（第 2.5 节），并对本章内容进行了总结（第 2.6 节）。

2.2 时空动作检测任务定义

从任务目标来看，时空动作检测是动作识别和时序动作检测的进一步延伸。动作识别任务是对一段指定视频数据中所包含的动作类别进行分类，通常假设视频数据已经经过剪辑处理，确保视频中仅包含单一动作类别的信息。时序动作检测任务则进一步要求模型不仅能够识别视频中的动作类别，还需要对动作发生的时间段进行定位，即确定动作的起始时间和结束时间。然而，时空动作检测任务在两者的基础上更进一步，通常处理的是未经剪辑的长视频数据，视频中可能包含多个动作类别，并且这些动作可能在时间和空间上交织在一起。因此，时空动作检测任务的目标不仅包括动作分类和时序定位，还需要实现对动作在空间维度的定位出特定的运动动作片段。

具体来说，为了统一表达，假设视频序列为 $V = \{I_t\}_{t=1}^{T}$，动作类别集合为 $\mathcal{C}$，动作识别的任务目标是判定整段视频所属的类别，不涉及具体的时间定位和空间定位。

$$\Phi(V) = c_i \tag{2.1}$$

其中 $c_i \in \mathcal{C}$ 是动作标签。

时序动作检测的任务目标则是确定输入视频片段中动作发生的时间范围（何时

图 2.1 时空动作检测任务目标：从未剪辑视频中获取完成动作分类、时间定位和空间定位

发生）以及动作类别，如下所示：

$$\Phi(V) = (c_i, t_b, t_e) \tag{2.2}$$

其中 t_b 为开始帧，t_e 为结束帧。

而时空动作检测的任务目标是需要同时确定动作的类别、时间范围以及每一帧中的空间位置（何处发生）。

$$\Phi(V) = \left(c_i, \{R_t^i\}_{t=t_b}^{t_e}\right) \tag{2.3}$$

$$\{R_t^i\}_{t=t_b}^{t_e} = \{(x_{min}, y_{min}, x_{max}, y_{max})_t \mid t \in [t_b, t_e]\} \tag{2.4}$$

$\{R_t^i\}$ 是从开始时间 t_b 到结束时间 t_e 每一帧图像 I_t 中对应的边界框或区域集合。

在图2.1 具体展示了时空动作检测相比于动作检测和时序动作检测的任务目标的区别。根据任务目标的差异，可以看出时空动作检测相较于动作检测和时序动作检测具有更高的复杂性和挑战性，需要模型具备更强的时空理解能力和精确的定位能力，而由于其能够提供更丰富的动作信息，因此在实际应用中具有更广泛的适用性和价值。表2.1 对比了三种任务类型及其典型应用场景。

表 2.1 动作检测任务类型及其适用场景对比

任务类型	任务目标	典型应用场景
动作识别	动作类别	视频分类、内容审核
时序动作检测	动作类别 + 时序定位	视频检索、录像回溯
时空动作检测	动作类别 + 时序定位 + 空间定位	高光视频生成、运动竞技分析

2.3 视频特征提取方法

对视频数据中时空特征的提取与建模直接影响到时空动作检测模型的效果和性能，视频的时空特征包含空间信息（环境场景、人员外观等）与时序信息（运动特征、演变信息等）。近些年，3D 卷积基础模型（C3D[?]、CSN）和基于 Transformer 基础模型（Timesformer、ViViT）逐渐已成为主流的视频特征提取模型，随着多模态大模型的兴起，基于视频-文本预训练模型（LLaVA-Video，Qwen-vl）的视频特征提取方法也逐渐受到关注。本节将介绍这些方法的基本原理和特点。

2.3.1 基于 3D 卷积的视频特征提取器

虽然有一些基于 2D 卷积的方法（MOC[19]、TSM[42]）通过在时间维度堆叠特征图辅以时序信息提取模块来捕捉时序信息，但是这些方法对时空信息的挖掘通常存在局限。3D 卷积能够在空间和时间维度上同时进行卷积操作，从而捕捉视频中的时空特征，因此，通过引入 3D 卷积核来同时处理空间和时间维度的信息成为一种解决方案。

C3D（Convolutional 3D）是视频理解领域的开创性模型之一，它确立了将 $3\times3\times3$ 的 3D 卷积核作为时空特征学习的基本计算单元，证明了 3D 卷积网络能够同时从视频中学习空间和时序特征。具体来说，对于输入视频片段 $X \in \mathbb{R}^{C_{in}\times T\times H\times W}$ 和 3D 卷积核 $W \in \mathbb{R}^{C_{out}\times C_{in}\times k\times k\times k}$，其中 C_{in} 表示输入通道数，C_{out} 表示输出通道数，k 为卷积核的空间与时间尺寸，T, H, W 分别表示时间帧数、高度和宽度。在进行特征提取时，标准 3D 卷积核将在输入特征图上滑动，输出特征图 $Y \in \mathbb{R}^{C_{out}\times t\times h\times w}$。其第 c_{out} 个输出通道在位置 (t, h, w) 的元素计算公式为：

$$Y_{c_{out},t,h,w} = \sum_{c_{in}=0}^{C_{in}-1}\sum_{i=0}^{k-1}\sum_{j=0}^{k-1}\sum_{l=0}^{k-1} W_{c_{out},c_{in},i,j,l}\cdot X_{c_{in},t+i,h+j,w+l} + b_{c_{out}} \tag{2.5}$$

完成一次该操作，其总计算量（FLOPs）约为 $O(C_{in} \cdot C_{out} \cdot k^3 \cdot THW)$。

为了降低计算复杂度并减轻过拟合，CSN（Channel-Separated Convolutional Networks）引入了分组卷积的思想，将标准的 $k \times k \times k$ 3D 卷积分解为逐点卷积（Pointwise Convolution）和深度卷积（Depthwise Convolution）。首先，利用 $1 \times 1 \times 1$ 的逐点卷积核 $W_{pw} \in \mathbb{R}^{C_{out} \times C_{in} \times 1 \times 1 \times 1}$ 进行通道维度的信息融合（Channel Interaction）。该步骤不涉及相邻时空信息的聚合，计算公式为：

$$Y'cout, t, h, w = \sum_{c_{in}=0}^{C_{in}-1} W_{pw, c_{out}, c_{in}} \cdot X_{c_{in}, t, h, w} + b_{c_{out}} \tag{2.6}$$

随后，为了捕获时空特征，CSN 使用深度卷积。深度卷积核 W_{dw} 的维度为 $C_{out} \times 1 \times k \times k \times k$（即组数等于通道数），输出的第 c 个通道仅依赖于输入（即上一层输出 Y'）的第 c 个通道，实现了通道分离（Channel Separation）：

$$Y_{c,t,h,w} = \sum_{i=0}^{k-1} \sum_{j=0}^{k-1} \sum_{l=0}^{k-1} W_{dw,c,0,i,j,l} \cdot Y'c, t+i, h+j, w+l+bc \tag{2.7}$$

通过这种分解，总计算量变为 $O(C_{in} \cdot C_{out} \cdot THW + C_{out} \cdot k^3 \cdot THW)$。与标准 3D 卷积的 $O(C_{in} \cdot C_{out} \cdot k^3 \cdot THW)$ 相比，当输入通道数 C_{in} 较大时（通常 $C_{in} \gg 1$），计算量减少了约 k^3 倍，在显著降低计算复杂度的同时保持了模型的表达能力。

2.3.2 基于 Transformer 的视频特征提取器

Transformer 架构已经在图像处理领域也得到了广泛的应用，Vision Transformer（ViT）[8] 作为最具代表性的工作之一，通过将图片进行 Patch 化为不同的图像块，并通过空间位置编码将视觉特征提取转化为一个序列学习问题。ViT 的成功表明了 Transformer 在视觉任务中的潜力，然而，由于视频数据中时间维度的存在，如何利用 Transformer 有效地进行时空建模成为问题的关键。

ViViT（Video Vision Transformer）[22] 创新性地提出了管状嵌入视频 patch 化的方法。如图所示，它从视频中切出“时空管嵌入”作为基本的时空 patch 单元，这保证了每个 patch 都跨越了空间和时间维度，这种方式能让模型在最初阶段就捕捉到时空演变信息。假设输入的视频为一个四维张量 $V \in \mathbb{R}^{T \times H \times W \times C}$，对于大小为 $t \times h \times w$ 的一个 3D 窗口，这个过程将视频 V 切分为 $N = \lfloor \frac{T}{t} \rfloor \times \lfloor \frac{H}{h} \rfloor \times \lfloor \frac{W}{w} \rfloor$ 个不重叠的时空管嵌入 p_i, 每个 p_i 的维度为 $\mathbb{R}^{t \times h \times w \times C}$。由于 Transformer 只能计算序列化的 token，

因此 ViViT 将每一个时空管嵌入 p_i 展平成一个一维的向量 $\mathbf{x}_i \in \mathbb{R}^{thwC}$，然后通过一个可学习的权重矩阵 $\mathbf{E} \in \mathbb{R}^{d\times(thwC)}$ 进行线性投影，得到维度为 d 的 Token:

$$\mathbf{z}_i = \mathbf{E}\mathbf{x}_i \in \mathbb{R}^d \tag{2.8}$$

结合位置编码，输入到 Transformer 编码器的第 0 层序列可表示为:

$$\mathbf{Z}_0 = [\mathbf{z}_{cls}; \mathbf{z}_1 + \mathbf{e}_1; \mathbf{z}_2 + \mathbf{e}_2; \ldots; \mathbf{z}_N + \mathbf{e}_N] \tag{2.9}$$

其中 $\mathbf{z}_{cls}$ 是用于分类的特殊 Token，$\mathbf{e}_i$ 是学习到的时空位置编码，用于保留 Token 在原视频中的位置信息，这种设计不仅降低计算复杂度，而且保留了局部时空特征，有利于模型建立时空长程依赖关系。

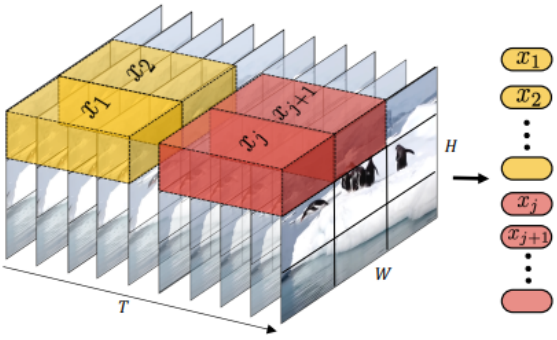


图 2.2 ViViT 的时空管道嵌入

2.3.3 基于多模态大模型的视频特征提取器

近些年，随着多模态大模型（如 GPT-4、PaLM-E 等）的兴起，基于视频-文本预训练模型的视频特征提取方法逐渐受到关注。这些模型通过在极大规模视频和文本数据上进行联合预训练，不仅学习到丰富的时空特征表示，并能从文本中获得丰富的语义信息，具有强大的跨模态理解能力。

LLaVA-Video 通过动态窗口采样与线性投影将视频看作带有时序标记的高分辨率图像序列，它通过在空间维度将每一帧图像根据分辨率切分为 N 个子网格（Sub-patches），每个网格独立提取特征，而为了防止 Token 数量爆炸，它会对连续帧的相同空间位置进行特征聚合。对于一个有 T 帧的视频，经过视觉编码器后得到特征矩阵 $X \in \mathbb{R}^{T\times N\times D}$，其中 N 是每帧的 Token 数，D 是特征维度，LLaVA 会应用一

个步长为 s 的池化操作来压缩时间维度：

$$X'_{i,j} = \text{Linear}(\text{Pool}(X_{t:t+s,i,j})) \tag{2.10}$$

Pool 表示池化操作。由于相邻帧之间的空间冗余度极高，通过在投影层中引入 1D 卷积或池化，它能将 T 帧压缩为 T/s 个时序特征块，同时保留物体移动的轨迹。

Qwen2.5-VL 通过动态分辨率（Naive Dynamic Resolution）机制和多模态旋转位置编码（M-RoPE），有效保留了视频 Token 的原生时空信息。在输入处理阶段，模型将视频视为一个连续的三维时空体，为每一个视觉 Token 分配唯一的三维坐标 (t, h, w)。M-RoPE 的核心在于通道解耦（Channel Decomposition），它不将位置编码简单叠加，而是将 Query 和 Key 的特征向量在通道维度上切分为三个子空间，分别对应时间、高度和宽度。位置信息 $f(t, h, w)$ 通过以下方式注入：

$$\mathbf{q}_{rot} = \text{Concat}\left(\mathbf{q}t \cdot \mathbf{R}\Theta, t, \quad \mathbf{q}h \cdot \mathbf{R}\Theta, h, \quad \mathbf{q}w \cdot \mathbf{R}\Theta, w\right) \tag{2.11}$$

其中，$\mathbf{q}_t, \mathbf{q}_h, \mathbf{q}_w$ 是原向量 $\mathbf{q}$ 切分后的三个分量，$\mathbf{R}_{\Theta,p}$ 表示在位置 p 处的旋转矩阵。通过这种机制，Attention 在计算两个任意 Token 的相关性时，实际上是在计算它们在时空三个维度上的相对位置的综合投影，从而捕捉到视频中真实的物理距离和时序依赖。

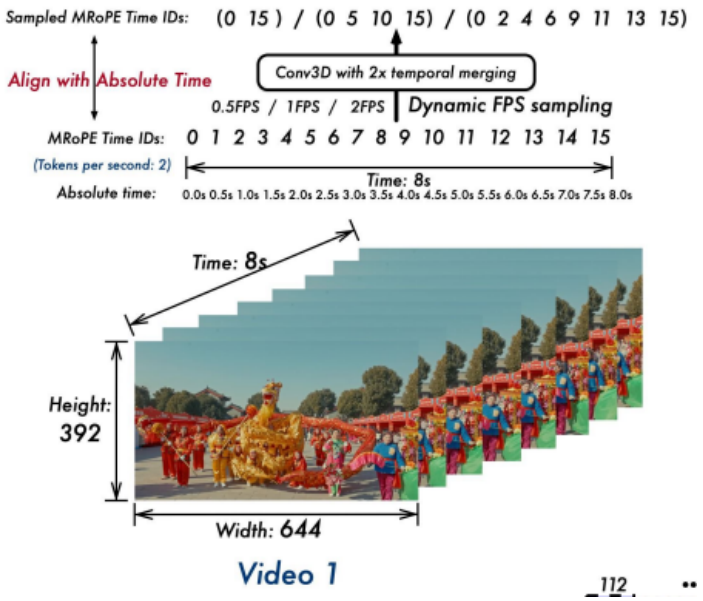


图 2.3 Qwen 的原生分辨率机制

2.4 基于查询的时空动作检测方法

时空动作检测算法可以分为双阶段检测范式和单阶段检测范式两大类方法，然而，由于双阶段检测范式依赖于目标检测器来生成候选的主体区域，这使得其性能受到目标检测器的限制，在处理体育场景这种复杂情况存在困难。早期单阶段检测方法参考 YOLO、CenterNet 等基于锚框的范式，但这些方法通常需要设计复杂的锚框生成策略，近年来，受到 DETR，AdaMix，DEIM 等基于查询的单阶段目标检测器的影响，时空动作检测方法也得到了长足发展。本节将重点介绍基于查询机制的时空动作检测方法的理论基础和相关方法。

2.4.1 DETR 设计概述

基于查询的检测任务方法的核心思想是通过一组可学习的查询（Queries）来直接预测图像或视频中的目标，最早由目标检测任务提出了 DETR（Detection Transformer）实现。DETR 通过引入 Transformer 架构，利用自注意力机制来建模图像中不同区域之间的关系，并通过一组可学习的查询向量来表示潜在的目标位置和类别。

2.4.2 基于查询的时空动作检测方法

TubeR 最早将基于查询的检测方法引入到时空动作检测任务中。TubeR 提出的 Tubelet Queries 将 DETR 的学习目标从二维的图像扩展为了三维的动作管，Tubelet Queries 记为 $\mathcal{Q} = \{Q_1, ..., Q_N\}$，其中 N 是预设的查询数量，而 Q_i 不再是一个单独的查询向量，而是一个查询序列：

$$Q_i = \{q_{i,1}, q_{i,2}, ..., q_{i,T_{out}}\} \tag{2.12}$$

这里，$q_{i,t} \in \mathbb{R}^C$ 是对应于第 i 个动作管在第 t 帧的查询嵌入，T_{out} 是输出的帧数，这种设计使得 Q_i 能够显式地对动作的时序演变进行建模。由于直接对所有 $N \times T_{out}$ 个查询进行全自注意力计算计算量巨大，为此 TubeR 设计了 Tubelet-Attention (TA) 模块，TA 模块将注意力机制分解为两个正交的步骤，通过分别在时间维度和空间维度进行自注意力计算。具体来说，空间自注意力 (Spatial Self-Attention) 该层进行同一

帧种不同动作实例的关系建模，其输入为同一时刻 t 的所有查询：

$$\text{Input}_{SSA} = \{q_{1,t}, q_{2,t}, ..., q_{N,t}\} \quad \forall t \in \{1, ..., T_{out}\} \tag{2.13}$$

这使得模型能够理解同一画面中不同动作实例之间的交互。而时间自注意力 (Temporal Self-Attention, TSA) 学习一个动作实例沿时间维度的状态变化，其输入为同一个动作实例在不同时间步 i 的所有查询：

$$\text{Input}_{TSA} = \{q_{i,1}, q_{i,2}, ..., q_{i,T_{out}}\} \quad \forall i \in \{1, ..., N\} \tag{2.14}$$

时间自注意力要求同一实例在时间维度 $q_{i,t}$ 和 $q_{i,t+1}$ 关注同一个目标对象，隐式地学习目标的运动轨迹。

PoSTAL 则是设计了一种结合视觉感知与语言引导的多模态学习框架，利用动作实例外观的自然语言描述作为额外信息的补充辅助动作检测和识别。具体地，PoSTAL 的模型设计可以由下式概况：

$$Y, \hat{y} = \mathcal{D}(\mathcal{P}(X, text)) \tag{2.15}$$

其中，$X \in \mathbb{R}^{T \times H \times W \times C}$ 表示输入视频，$text$ 表示包含球员的球衣颜色和号码结构化句子，$\mathcal{P}$ 表示提示驱动的目标动作编码器, 它是一个由基于 BLIP 的文本特征编码器和视频特征编码器组成的多模态特征提取模块，负责将原始视频张量 X 与文本提示 $text$ 进行融合，生成富含目标语义信息的特征表示，最后输出目标动作管 Y 和其对应的动作类别 $\hat{y}$，$\mathcal{D}$ 表示动作管解码器，负责从编码特征中直接预测出动作管道的时空坐标和动作类别。

2.5 相关数据集和评价指标

本文使用的数据集有 UCF101-24、multisports，使用的评价指标包括性能评价指标和误差分析指标，其中性能评价指标包括帧平均精度（Frame-level Mean Average Precision，Frame-map）和视频平均精度（Video-level Mean Average Precision，Video-map），误差分析指标包括分类误差（Classification Error，EC）、定位误差（Localization Error，EL）时间误差（Time Error，ET）、检测误差（Missed Detection，EM）和其它误差（Other Error，EO）。

2.5.1 数据集简介

（1）UCF101-24

UCF101-24 数据集源于 2012 年发布的 UCF101 数据集，UCF101 数据集针对动作分类任务，共包含 101 类动作，2013 年在 THUMOS 2013 挑战赛中从中选取了其中的 24 类动作进行了额外逐帧边界框标注和时间区间标注，构建了 UCF101-24 数据集用于时空动作检测任务。UCF101-24 涉及的动作类别主要为体育项目和少量的日常生活行为，其中体育动作包括篮球、足球、排球、网球、潜水和滑冰等体育项目，以及遛狗、跳绳和骑自行车等日常生活行为，UCF101-24 视频采集自 YouTube，视频的分辨率为 320×240，具有复杂背景、多变的视角和不同光照条件。视频总数约 3207 个视频片段，其中训练集包含 2293 个视频，测试集包含 914 个视频。

(c) UCF101-24: Trampoline Jumping

(d) Our something-STAR: Putting something next to something

图 2.4 UCF101-24 数据集样例

（2）MultiSports

MultiSports（Multi-person Sports Actions）数据集由南京大学的 Lei Chen 等人于 2021 年提出，它专注于竞技体育场景的多人时空动作检测任务，涵盖了篮球、排球、足球和竞技健美操 4 类运动项目。MultiSports 数据集同样采集自 YouTube，视频样本分辨率为 1080p，并包含 3200 个视频片段并定义了 66 个精细动作类别，这些运动具有多人参与、动作类别定义明确且边界清晰的特点，是目前体育运动场景时空动作检测领域最大的公开数据集。

Multisports 数据集充分体现了体育场景下时空动作检测的挑战性，为了评估数据集中动作管的复杂性，本节利用管内 IoU（IoU_Intra）来衡量单个动作管内部的复杂度，并利用管间 IoU（IoU_Inter）来评估视频中动作管间的相互作用。给定一

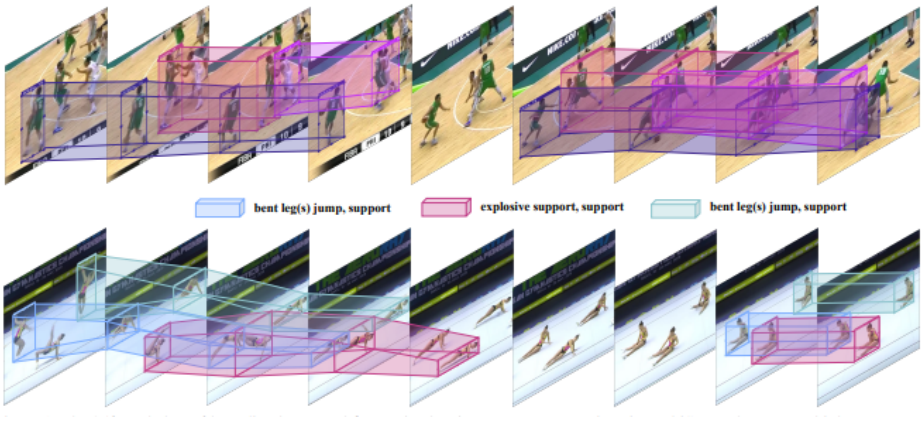


图 2.5 Multisports 数据集样例

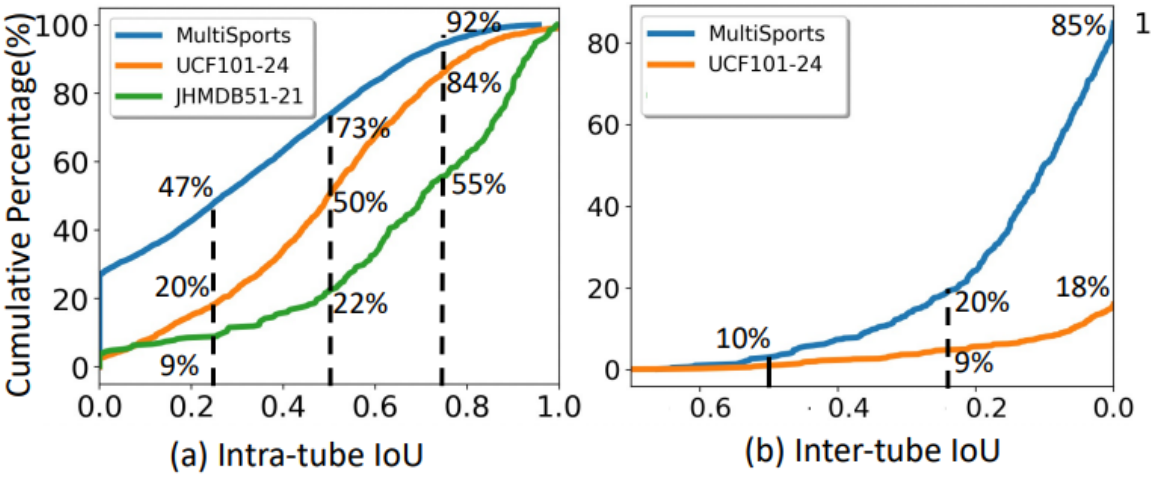


图 2.6 Multisports 复杂度

个包含 M 个动作管 $T_1, T_2, \ldots, T_M$ 的视频，其中动作管 $T_j = \{B_j^1, B_j^2, \ldots, B_j^l\}, j \in \{1, 2, \ldots, M\}, B_j^i (i \in \{1, 2, \ldots, l\})$ 是第 i 帧的边界框。动作管 T_m 的管内 IoU 定义为管内相邻框对 IoU 的平均值:

$$\text{IoU\_Intra} = \sum_{i=1}^{l-1} \text{IoU}(B_m^i, B_m^{i+1})/(l-1) \tag{2.16}$$

IoU_Intra 越低，表示该动作管的形状复杂度越高。为了衡量 T_m 由于与视频中其他管柱相互作用而产生的复杂度，本文首先计算 T_m 与视频中每个其他管柱 T_j 之间的管间 IoU:

$$\text{IoU\_Inter} = \text{TIoU}(T_m, T_j), j \in \{1, 2, \ldots, M\} \& j \neq m \tag{2.17}$$

其中 TIoU 表示两个动作管之间的 IoU。由图可知 UCF101-24 仅有 50% 的管内 IoU 低于 0.5，相比之下 MultiSports 高达 73%，管柱之间更高的重叠度意味着更复杂的相互作用。MultiSports 中 85% 的管柱与其他管柱存在重叠，而 UCF 数据集仅为 18%。统计结果表明，MultiSports 和 UCF 均含大量形状复杂的动作管柱，而 MultiSports 在多人数并发和动作精细度方面的挑战性显著更高。

2.5.2 评价指标

（1）性能评价指标

时空动作检测的性能评价指标需要能评估算法在进行动作分类、时序定位和空间定位三方面的能力，目前领域内常用评价 frame-mAP 和 video-mAP 来对模型的性能进行评估。frame-mAP 通过计算标注结果与预测结果在每一帧上的空间交并比（sIoU）来评估模型在空间定位方面的能力，sIoU 的计算方式与传统的 IoU 相同，具体计算公式如下所示:

$$\text{sIoU}(B_p, B_g) = \frac{Area(B_p \cap B_g)}{Area(B_p \cup B_g)} \tag{2.18}$$

其中 B_p 表示预测边界框，B_g 表示真实边界框。frame-mAP 通过计算每一帧上所有动作类别的平均精度来评估模型的整体性能，其计算方式如下所示:

$$\text{f-mAP} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \text{AP}_c \tag{2.19}$$

video-mAP 是以每一个完整动作管作为计算单元来评估模型整体的性能，通过计算

每个预测动作管与真实动作管在时间和空间维度上的时空交并比（IoU_{st}）来进行每个具体动作实例的匹配。具体来说，IoU_{st} 首先计算预测管道与真实管道在时间维度上的重叠区间，然后对于重叠区间内的每一帧计算空间 IoU 进行平均，具体计算公式如下所示：

$$\text{IoU}_{st} = \frac{1}{N}\sum_{i=1}^{N}\text{IoU}(B_i, B_i') \tag{2.20}$$

其中，B_i 表示预测边界框，B_i' 表示真实边界框，N 表示重叠区间内的帧数。video-mAP 则是通过计算所有预测结果的（IoU_{st}）与真实结果进行匹配后，计算每个动作类别的平均精度，video-mAP 引入了三维的（IoU_{st}）进行动作实例的匹配，更加注重模型在时间和空间维度上对动作区间的精准定位能力，其计算公式如下所示:

$$\text{video-mAP} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\text{AP}_c \tag{2.21}$$

（2）误差分析指标

由于时空动作检测本质上涉及到动作分类、时序定位和空间定位三方面的任务，因此仅通过 frame-mAP 和 video-mAP 并不能够全面反映模型在各个方面的表现。为了更细致地分析模型在不同任务上的误差情况，本文引入了误差分析指标，我们主要讨论以下五种误差类型：分类误差（EC）、定位误差（EL）、时间误差（ET）、检测误差（EM）和其他误差（EO）。（1）分类误差（EC）: 当预测的动作类别与真实类别不匹配时，产生分类误差。分类误差反映了模型在动作识别方面的能力, 动作管时空（IoU_{st}）大于与真实值的阈值，但其动作类别与真实值的类别不同。

（2）定位误差（EL）: 当预测的动作类别正确，但时空位置与真实位置不匹配时，产生定位误差。定位误差反映了模型在空间定位方面的能力，动作管的时空（IoU_{st}）介于两个阈值之间。预测动作管与某个真实值具有相同的动作类别和时间 IoU，且时间 IoU 大于阈值，但其在真实值与检测结果的时间交集区域的平均空间边界框 IoU 较低，导致检测结果的 IoU 低于所需阈值。

（3）时间误差（ET）: 当预测的动作类别正确且空间位置匹配，但时间区间与真实区间不匹配时，产生时间误差。时间误差反映了模型在时序定位方面的能力，动作管的时间 IoU 介于两个阈值之间。检测结果与动作类别相同，且平均空间边界框 IoU 大于阈值，在时间交集区域内与某些真实值相同，但时间 IoU 较低，使得动作管时空（IoU_{st}）低于所需阈值。

（4）检测误差（EM）：当模型未能检测到真实存在的动作实例时，产生检测误差。检测误差反映了模型在动作实例发现方面的能力。检测结果与任何真实值均不匹配，导致动作管时空（IoU_{st}）低于所需阈值。

（5）其他误差（EO）：除上述四种误差外的其他类型误差，例如由于视频质量差、遮挡严重等原因导致的误差。

各类误差的计算方式和关系具体如下图所示：

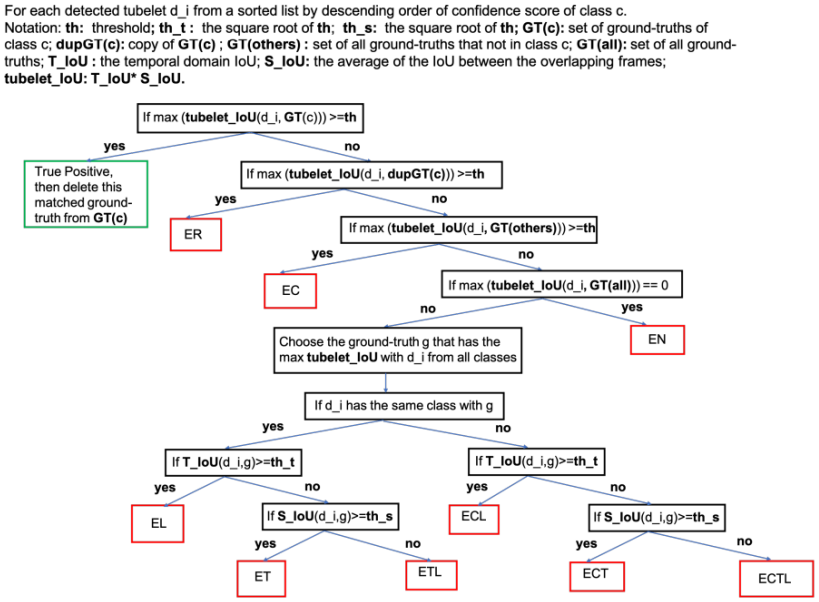


图 2.7 误差分析指标示意图

2.6 本章小结

本章主要介绍系统与控制理论类论文正文章节的框架结构。在每章的最后，都需要对该章的内容进行小结，不宜太长，建议 1/2-2/3 页版面较好。主要小结一下本章用什么理论或方法、做了什么事、得到的重要结果或结论。

3 基于多模态特征增强的体育时空动作检测算法研究

与日常行为场景不同，体育场景下的动作检测具有高度的专业性和规则依赖性，仅依靠视觉信息进行时空动作检测面临显著挑战。对于区分度较低的复杂体育动作，往往需要结合深层的语义信息和领域专家的先验知识才能准确理解动作含义并完成精确辨析。鉴于体育运动的复杂性和多样性，单一的视觉特征往往难以充分捕捉动作背后的规则约束和上下文关系。为此，本章提出了一种基于多模态特征增强的体育时空动作检测方法。该方法通过构建包含“场景-文本”与“运动-文本”的双元素多模态知识库，并利用对齐与融合机制，将领域专业知识有效地注入视觉检测框架中，显著提升了模型对复杂体育动作的理解与检测能力。

本章的组织结构安排如下：3.1 节首先回顾基于外部知识增强的时空动作检测研究现状，并阐述本章的研究动机与思路；3.2 节详细介绍体育视频双元素多模态知识库的构建方法；3.3 节概述基于多模态特征增强的时空动作检测模型的整体架构；3.4 节和 3.5 节分别深入阐述查询模块和多模态特征门控融合模块的设计细节；3.6 节定义模型的损失函数；3.7 节介绍实验设置，并在 UCF101-24 和 MultiSports 数据集上进行对比实验与消融分析；3.8 节对本章工作进行总结。

3.1 引言

近年来，时空动作检测因其在智能体育分析、视频监控等领域的广泛应用前景，受到了学术界和工业界的高度关注。该任务旨在同时实现动作的空间定位与时间分类，因此，早期的研究主要致力于设计更高效的时空特征建模方法以提升性能。例如，EVAD 和 VideoMAE 等工作通过设计时空自掩码机制来学习更具鲁棒性的时空表征。然而，时空动作检测不仅依赖于底层的视觉特征，人体的运动模式、场景上下文以及动作的深层语义同样对检测结果起着决定性作用。尽管 HIT 和 2in1 等方法尝试通过引入人体关键点或光流特征来补充运动信息，但在面对复杂的体育场景时，这些方法仍然存在不足。

相比于日常场景，体育场景的时空动作检测面临着更为严峻的挑战：（1）细粒度动作的视觉差异微弱。许多体育动作在视觉外观上高度相似，主要区别在于动作

意图或力度等细微特征。例如，在排球场景中，“扣球”和“吊球”在起跳和挥臂的初期轨迹上几乎一致，其核心区别在于击球瞬间的力度控制和战术意图，单纯依赖像素级的视觉特征难以对这两类动作进行准确界定。（2）动作定义的强规则约束性。体育动作的判定往往严格遵循特定的竞技规则和技术标准，而非简单的肢体运动。以体操运动中的“俯卧撑”为例，其标准定义要求运动员在腾空阶段呈屈体姿势，躯干与双腿间形成约 60° 夹角，且落地时需呈伸展后撑姿势。这种包含特定角度、时序逻辑和接触状态的判别性语义，构成了体育动作的核心特征。单一的视觉特征往往难以捕捉这种隐含在规则背后的高层语义，从而限制了模型在复杂规则场景下的检测精度。

近年来，随着多模态大模型（如 CLIP, BLIP, GPT-4, Qwen 等）的兴起，多模态特征融合技术在诸多视觉感知任务中展现出卓越性能。这些大模型通过大规模预训练，能够提取包含丰富语义的“视觉-文本”综合特征，为深入理解视频内容提供了新的途径。多模态大模型强大的泛化能力与场景理解能力，与体育动作检测的需求高度契合。部分研究已开始探索将其应用于该领域，例如 PoSTAL 通过将篮球运动员的衣着颜色和号码作为文本提示输入 BLIP 模型，利用文本编码引导模型关注特定视觉区域；HCBS 则面向足球场景，利用 LLava 生成从粗到细的多层级问答，并将文本编码与视觉特征融合。

然而，现有的多模态体育动作检测方法仍存在以下局限性：（1）缺乏领域专业信息的有效引导。现有方法多依赖通用的大模型生成描述，或仅使用简单的标签、外观特征（如球衣颜色）作为提示。这种通用描述缺乏对体育动作技术规范的深入理解，且生成式模型容易产生“幻觉”，导致文本信息不准确，进而误导检测模型。（2）多模态交互特征挖掘不充分。现有方法大多将多模态大模型视为单纯的文本编码器，仅利用其输出的文本嵌入，而忽略了多模态模型在预训练阶段习得的深层“视频-文本”对齐特征与交互关系。这种浅层的利用方式限制了多模态信息对视觉特征的增强效果。

综上所述，为了解决上述问题，本章提出了一种基于多模态特征增强的时空动作检测算法。核心思路在于解决两大关键挑战：一是如何结合体育领域的专业知识，构建高可信度的多模态知识库，以替代不稳定的生成式描述；二是如何设计有效的融合机制，充分挖掘多模态大模型中视频与文本的综合特征，以增强对复杂动作的表征能力。

具体而言，为了引入精确的领域知识，本章对现有数据集的动作类别定义进行了专业的搜集，构建了一个“场景-文本 + 运动-文本”的双元素多模态知识库。对于每一类动作原型，该知识库不仅提供宏观的场景上下文描述，还提供微观的动作技术规范描述，实现了从粗粒度到细粒度的全面语义覆盖。在此基础上，本章设计了一个包含特征对齐模块与门控融合模块的检测框架，能够自适应地将多模态语义信息注入视觉特征中，实现了性能的显著提升。

本章的主要贡献总结如下：

（1）提出了一种基于“场景 + 运动”双元素驱动的体育多模态知识库构建方法。区别于传统的标签式提示，本章构建的知识库深度融合了体育动作的专业定义与规则描述。通过解耦“场景上下文”与“运动技术规范”两类特征，该方法不仅解决了大模型生成的幻觉问题，还为模型提供了更具判别性的语义先验，显著增强了知识库的鲁棒性与准确性。

（2）设计了一种基于多模态深层交互增强的时空动作检测算法。针对现有方法利用多模态信息肤浅的问题，本章提出了一套新的特征增强架构。通过设计语义-视觉对齐模块与自适应门控融合模块，该算法能够根据样本的难易程度，动态调节外部语义信息对视觉特征的修正力度，实现了多模态信息的高效注入与互补。

（3）在标准数据集上验证了方法的有效性。本章在 UCF101-24 和 MultiSports 数据集上进行了广泛的对比实验与消融研究。实验结果表明，所提方法在处理复杂体育动作时表现优异，显著优于现有的单模态及部分多模态检测方法。

4 基于时空一致性建模的体育时空动作检测算法研究

第三章从特征增强的角度出发，利用多模态大模型构建外部知识库，旨在为复杂的体育运动时空动作检测补充丰富的语义信息，而本章专注于时空特征的深度挖掘，致力于提升模型自身的时空一致性建模能力。鉴于体育动作具有运动剧烈、位移显著以及与场景元素高度耦合等特性，本章提出了一种基于动作感知引导的时空一致性建模方法，该方法通过生成动作相关的动作管查询，结合正交自适应采样机制和解耦时空注意力模块，有效实现了跨帧特征关联和上下文提取，从而有效提升模型在复杂体育场景下的检测性能。

本章的组织结构安排如下：4.1 节首先剖析现有方法在时空一致性建模含义和目前存在的局限性，并阐述本章的研究动机与思路；4.2 节详细介绍本章模型的整体架构；随后，4.3 节、4.4 节和 4.5 节分别阐述本章设计的三个核心模块：动作相关的动作管查询生成模块、动作特征引导的正交自适应采样模块以及解耦时空注意力模块；最后，4.6 节介绍了实验设置并对实验结果进行深入分析，4.7 节对本章工作进行了总结。

4.1 引言

在时空动作检测任务中，“时空一致性建模”指的是一个包含双重目标的联合问题：（1）同一动作实例在时序维度上的帧间关联；（2）动作实例与时空上下文（如交互物体、场景信息）的动态关系建模。

帧间关联是时空动作检测面临的关键挑战之一。早期研究通常基于帧级检测结果，将帧间关联视为独立的前处理或后处理步骤，主要依赖相邻帧检测框的空间重叠度（IoU）和视觉相似性进行连接，然而，这类基于局部启发式规则的方法忽略了跨帧特征的连续性，人为割裂了检测与关联过程，难以处理非连续的时空轨迹。随后，部分工作尝试设计三维锚框以实现特征层面的关联，但这些方法通常建立在“目标运动平滑”“小位移变化”的假设之上，预定义的三维锚框在时序维度缺乏灵活性，导致其在处理大位移运动时性能显著下降。相比之下，基于查询（Query-based）的方法利用可学习的动作管查询和全局自注意力机制，能够直接建立端到端的长程时

空依赖，即使在目标发生剧烈位移时，模型仍然能够较好地保持时空特征的一致性。图4.1展示了现有的几种帧间动作关联方法之间的区别。

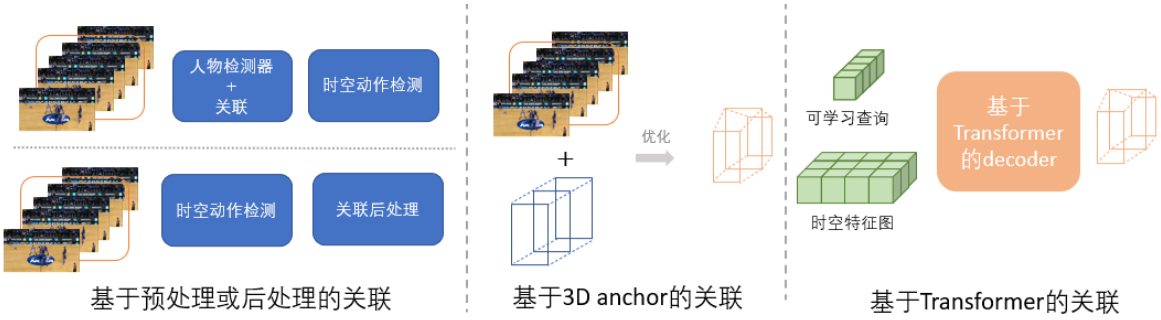


图 4.1 主流帧间动作关联方法示意图

具体来说，Tuber 提出的动作管查询设计验证了时空正交解耦的自注意力机制有助于在查询层面实现同一动作实例的一致性关联，然而，Tuber 完全依赖于 Transformer 架构的长程建模能力，缺乏对跨帧时空特征一致性的显式引导，在面对复杂运动场景时，这种隐式建模存在信息瓶颈。在此基础上，STAR 设计了正交解耦的交叉注意力机制，限制动作管查询仅与当前时间步的特征进行交互，这虽然缓解了计算负担，但同样未解决跨帧特征一致性显式建模的问题。ART 方法则利用人物检测器和 RoI Align 聚合关键帧特征，经时序扩展生成人物相关的动作管查询，但该查询仅包含关键帧的人物外观先验，缺乏对运动趋势和场景上下文等关键信息的利用。

综上所述，受限于体育运动的剧烈性和环境复杂性，现有基于查询的方法在处理体育时空动作检测时仍面临两大难题：（1）动作实例间的帧间信息关联鲁棒性不足，难以适应大位移和剧烈形变；（2）对时空上下文特征的挖掘存在局限，导致对相似体育动作的辨别能力较弱。

针对上述问题，本章提出了一种基于动作感知引导的时空一致性建模方法（Action-guided Spatiotemporal Consistency Modeling, AGCM）。具体来说，为了增强帧间动作实例关联，设计了动作感知模块（Action-Aware Module, AAM），该模块显式地通过从全局时空特征中挖掘潜在的动作相关的上下文特征，并以此引导解码器中的信息聚合过程，从而保证了时间维度上的主题一致性。为了提升时空上下文特征的挖掘能力，本章进一步设计了动作特征引导的正交自适应采样模块（Action-guided Adaptive Sampling Module, AGAS）以及解耦时空注意力模块

（Decoupled Spatiotemporal Attention Module, DSTA）。前者通过自适应采样策略，灵活地提取与聚合时空特征图中与动作相关的特征；后者则通过时空解耦增强了模型对时空依赖的挖掘能力。

本章在 JHMDB51-21、UCF101-24 和 Multisports 数据集上与现有方法进行了对比实验，并通过各模块间的消融实验，验证了所提方法的有效性。本章的主要贡献包括以下三点：

（1）提出了一种基于动作感知引导的时空一致性建模框架（AGCM）。针对体育场景下时空一致性建模困难的问题，本章打破了以往隐式建模的局限，提出利用生成动作相关的相关动作管查询，显式引导时空特征的一致性建模。

（2）设计了动作感知模块（AAM），显著增强了帧间动作关联的鲁棒性。为了解决同一动作实例在不同时间步的时空特征关联问题，该模块通过监督学习感知全局时空特征中的潜在动作特征，生成了动作相关的动作管查询，引导解码器自适应地关联动作相关的时空特征。

（3）提出了动作特征引导的自适应采样模块（AGAS）与解耦时空交叉注意力模块（DSTA），AGAS 通过自适应采样策略灵活聚焦与动作高度相关的时空特征，DSTA 则通过解耦交叉注意力高效地建模了时空依赖关系，提升了对时空上下文特征的挖掘能力

4.2 模型整体架构

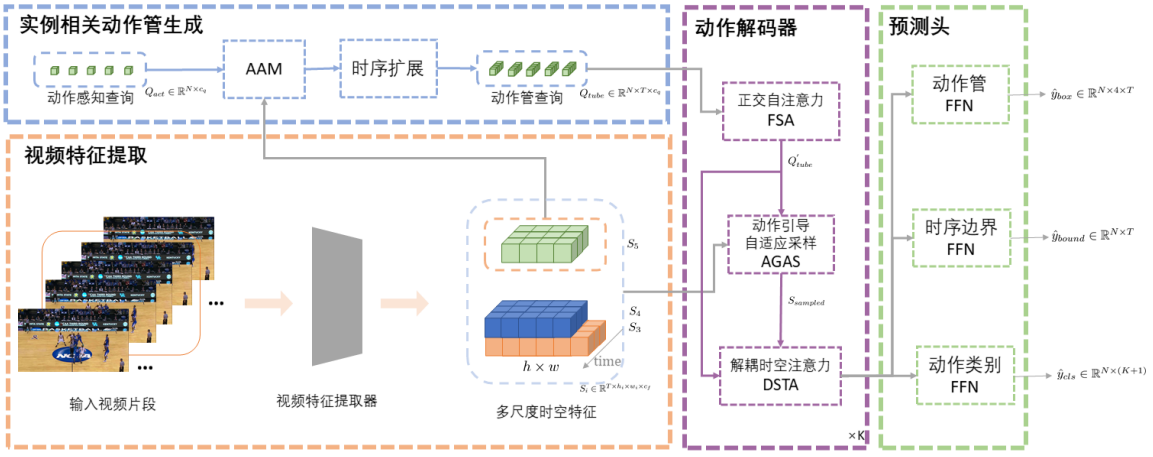


图 4.2 基于动作感知引导的时空一致性建模方法整体框架

如图4.2所示，本章提出的基于动作感知引导的时空一致性建模方法整体架构由视频特征提取、动作相关动作管查询生成模块、时空动作解码器以及检测头四个主要部分组成。在时空特征提取阶段，首先利用基于 3D CNN 的预训练特征提取网络从输入视频片段 $X \in \mathbb{R}^{T\times H\times W\times 3}$ 中提取多尺度的时空视觉特征，构建出时空特征金字塔 $\mathcal{S}=\{S_i\}_{i=3}^{5}$，其中第 i 层特征图 $S_i \in \mathbb{R}^{T\times h_i\times w_i\times c_f}$。紧接着，为了解决随机初始化查询难以捕捉复杂运动的问题，动作相关动作管查询生成模块引入了动作感知模块（AAM），在训练阶段，AAM 通过额外的辅助分类头，使可学习的动作感知查询 $Q_{act} \in \mathbb{R}^{N\times c_q}$（$N$ 为查询数量）具备挖掘潜在的“动作先验”的能力，在推理阶段，由 AAM 基于全局时空特征感知“动作相关”的时空信息并生成具有动作先验的查询 Q_{act}；随后，将 Q_{act} 会在时间维度上进行复制与时序扩展，并叠加时序位置编码，生成贯穿整个时序的动作管查询 $Q_{tube} \in \mathbb{R}^{N\times T\times c_q}$。在时空动作解码器中，这些动作管查询首先经过正交分解的自注意力机制（Factorised Self Attention，FSA）进行内部交互，再输入到动作特征引导的自适应采样模块 AGAS，通过该模块可以自适应获得动作相关的采样特征 $S_{sampled} \in \mathbb{R}^{N\times n\times T\times c_f}$（其中 n 为采样点数量）；最后，通过解耦时空交叉注意力模块（DSTA）进行动作管查询 Q_{tube} 与采样特征 $S_{sampled}$ 之间的交互，将视频上下文信息注入动作管查询中，得到更新后的特征 Q_{update}，在下一轮循环解码计算时，会将 Q_{update} 与 Q_{tube} 进行加和以生成新的 Q_{tube}。经过解码器的多层迭代更新后，最终的动作管特征 Q_{update} 被送入检测头，该模块由三个任务分支组成，均由前馈神经网络 FFN 构成：一个动作分类头，用于预测动作实例的类别，输出维度为 $\hat{y}_{cls} \in \mathbb{R}^{N\times(K+1)}$（$K$ 为类别数）；时序定位头和空间定位头分别负责时序边界判断和矩形框回归，前者输出时序边界概率 $\hat{y}_{bound} \in \mathbb{R}^{N\times T}$，后者输出归一化的时空边界框坐标 $\hat{y}_{box} \in \mathbb{R}^{N\times 4\times T}$，从而实现对时空动作的精确定位与识别。

4.3 动作相关的动作管查询生成

为了增强帧间动作实例关联的鲁棒性，本节设计了动作相关的动作管查询生成模块，该模块包含两个主要部分：动作感知模块（AAM）和时序拓展模块，具体结构如图4.3所示。前者旨在基于全局时空特征挖掘潜在的动作相关特征，实现视频片段中动作实例信息的预提取，这使得动作感知查询能够包含丰富的动作实例先验，从而指导后续的时空特征采样；后者则负责将动作实例感知特征在时间维度上进行

扩展，并通过补充位置编码信息注入时序位置先验，最终生成既包含动作实例语义，又能感知时间维度运动变化的动作管查询。下面首先介绍 AAM。现有的时空动作检

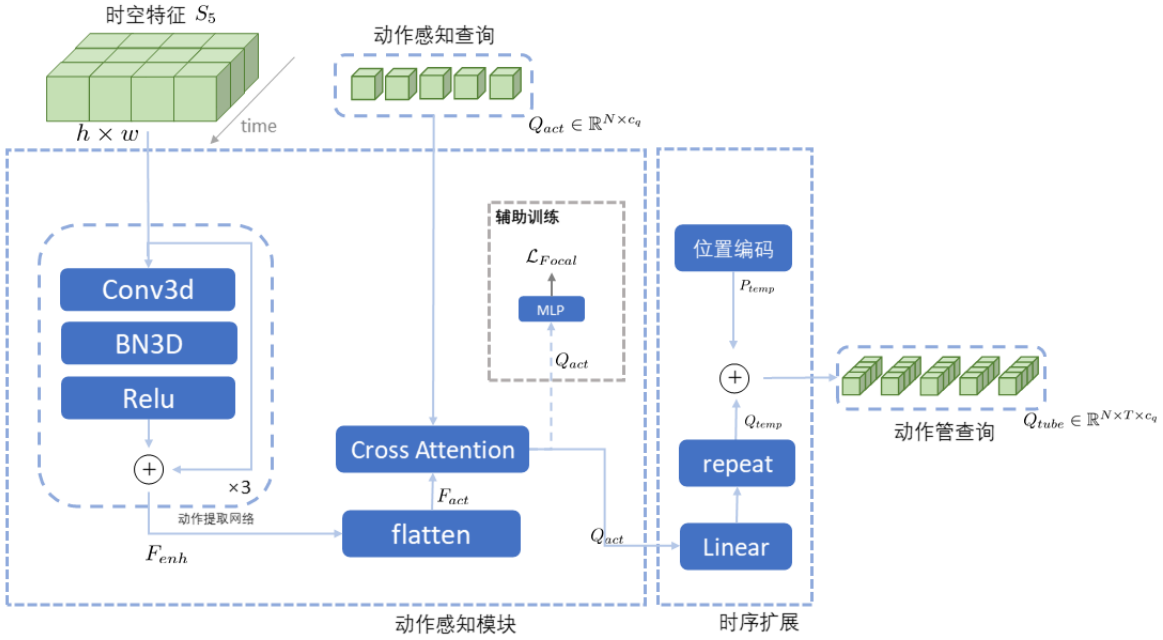


图 4.3 动作实例相关的动作管查询生成模块

测方法（如 TubeR、STAR）中，动作管查询通常采用随机初始化，或仅随机初始化空间查询后在时间维度上复制，这种方式忽略了视频片段中潜在的特定动作实例信息，导致查询难以迅速捕捉复杂运动特征；另有一些方法（如 ART、STMixer）依赖预生成的关键帧特征，忽略了跨帧时空一致性的显式建模，导致在复杂运动场景下，动作实例间的帧间关联鲁棒性不足。为了解决这个问题，本节设计了动作感知模块（AAM），该模块在训练阶段，AAM 通过一个额外的辅助分类头，监督动作感知查询学习潜在的动作先验信息；在推理阶段，AAM 基于全局时空特征感知获取动作实例相关的动作查询。

具体来说，对于由视觉特征提取器得到的深层时空特征 $S_5 \in \mathbb{R}^{T \times h_5 \times w_5 \times c_5}$，在 AAM 中，首先通过一个轻量的动作提取模块 $\mathcal{F}_{ext}(\cdot)$ 对时空特征进行增强。该模块由三个堆叠的（Conv3D, BN3D, ReLU）单元组成，并引入残差连接以缓解梯度消失问题，增强后的特征 F_{enh} 计算如下：

$$F_{enh} = \mathcal{F}ext(S_5) + S_5 \tag{4.1}$$

随后，为了保留丰富的时空细节，模型并未对 F_{enh} 进行全局池化压缩，而是将其在时空维度上进行展平，得到序列化的动作实例特征 F_{act}，表示为：

$$F_{act} = \text{Flatten}(F_{enh}) \in \mathbb{R}^{L \times c_5} \tag{4.2}$$

其中，$L = T \times h_5 \times w_5$ 表示时空特征序列的长度，c_5 为特征通道数。接着，为了从 F_{act} 中解耦出 N 个独立的潜在动作实例，模型定义了一组可学习的动作感知查询 $Q_{act} \in \mathbb{R}^{N \times c_q}$。AAM 利用多头注意力模块建立动作感知查询与时空特征之间的交互，在计算过程中将 Q_{act} 作为查询，将包含丰富时空细节的 F_{act} 同时作为键和值。最终生成具有实例区分性的动作感知查询 Q_{act}：

$$Q_{act} = \text{MHA}(Q_{act}, F_{act}, F_{act}) \in \mathbb{R}^{N \times c_q} \tag{4.3}$$

为了确保动作感知查询能够聚焦于视频片段中的真实动作实例，在训练阶段，AAM 会对 Q_{act} 进行监督学习，使其能够预感知全局时空特征中的动作实例信息。具体来说，动作感知查询 Q_{act} 会被送入一个辅助分类头，用来预测输入片段中真实存在的动作类别，并通过与真实标签进行匹配监督，而在推理阶段，该分类头则被舍弃，动作感知查询 Q_{act} 直接用于后续的时空特征采样与建模。通过这种设计，AAM 能够促使动作感知查询挖掘视频片段中潜在的动作实例先验信息，为后续的时空特征采样和时空建模提供了有力的指导。

接下来，时序拓展环节负责将动作实例感知查询 Q_{act} 在时间维度上进行扩展，以生成贯穿整个时序的动作管查询。具体来说，首先将 Q_{act} 在时间维度上进行复制扩展，得到初始的动作管查询 $Q_{temp} \in \mathbb{R}^{N \times T \times C}$。为了让 Q_{temp} 具备感知动作在不同帧之间的动态变化，模型引入了时序位置编码（Temporal Positional Embedding）P_{temp}，将其与 Q_{temp} 进行加和运算，最终生成了包含动作实例先验且具备时序位置信息的动作管查询 $Q_{tube} \in \mathbb{R}^{N \times T \times C}$。

4.4 动作引导的自适应采样模块

在得到动作相关的动作管查询 Q_{tube} 后，时空动作解码器需要基于这些查询从多尺度特征图中提取与当前动作实例高度相关的时空特征。为了增强对复杂运动场景下的时空上下文特征挖掘能力，本节设计了动作引导的自适应采样模块（ADAS），其结构如图 4.4 所示。该模块通过生成与动作实例特征相关的采样点偏移量，实现对

多尺度时空特征图的自适应采样，从而灵活地聚焦于与动作高度相关的区域，提升模型对局部细节和动态变化的感知能力。对视觉特征进行自适应采样的思想在计算

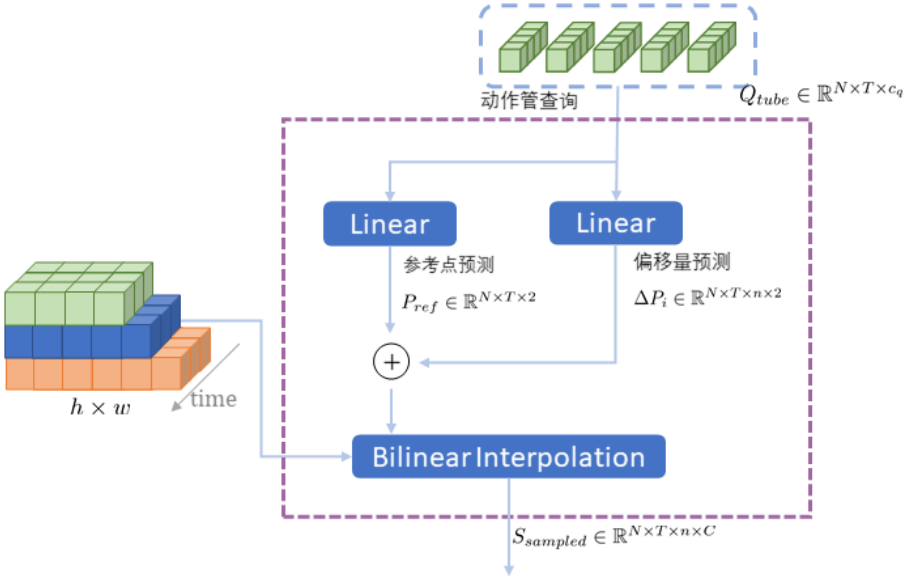


图 4.4 实例特征引导的自适应采样模块

机视觉领域由来已久（如 DCN、Deformable DETR、DAT），通过预测采样点偏移引导模型关注重要区域，不仅可以扩大有效感受野、提高信噪比，还能更好地关注时空上下文，而通过动作引导的自适应采样模块，更能确保采样特征与当前动作高度相关。首先，与现有范式（如 TubeR、STAR）相同，在动作管查询进入 AGAS 之前，首先通过一个正交时空自注意力模块（FSA）进行内部交互，以增强查询间的信息传递与协同。IASM 接收经过 FSA 增强后的动作管查询 $Q'_{tube} \in \mathbb{R}^{N\times T\times C}$ 以及多尺度时空特征图 $\mathcal{S} = \{S_i\}_{i=3}^{5}$ 作为输入。随后，Q'_{tube} 会通过两个线性层，其中，其中一个线性层会生成该动作管查询在每一时间步的参考点（Reference Point）$P_{ref} \in \mathbb{R}^{N\times T\times 2}$，另一个线性层动作管查询特征预测采样偏移量, 具体计算如下：

$$P_{\text{ref}} = \text{Linear}_{\text{ref}}(Q_{\text{tube}}) \tag{4.4}$$

$$\Delta P_i = \text{Linear}_{\text{offset}}(Q_{\text{tube}}) \tag{4.5}$$

其中，$\Delta P_i \in \mathbb{R}^{N\times T\times n\times 2}$ 表示 N 个实例在 T 个时间步上、每个步长的 n 个采样点的坐标偏移。接下来，模块在特征图 S_i 上执行自适应采样。对于每个动作管查询在时刻 t 的第 k 个采样点，其绝对采样位置计算为 $P_{sample} = \phi(P_{ref} + \Delta P_{ik})$（$\phi$ 为坐标归

一化函数）。模块通过双线性插值从 S_i 中提取该位置的时空特征，得到层级采样特征 $S^i_{sampled}$：

$$S^i_{sampled} = \text{BilinearSample}(S_i, P_{ref} + \Delta P_i) \in \mathbb{R}^{N\times T\times n\times C} \tag{4.6}$$

通过这种实例特征引导的自适应采样策略，IASM 能够基于动作管查询预测的参考轨迹，灵活地从多尺度特征图中抓取与当前动作高度相关的局部细节，显著提升了模型对复杂运动场景的感知能力。

4.5 解耦时空交叉注意力模块

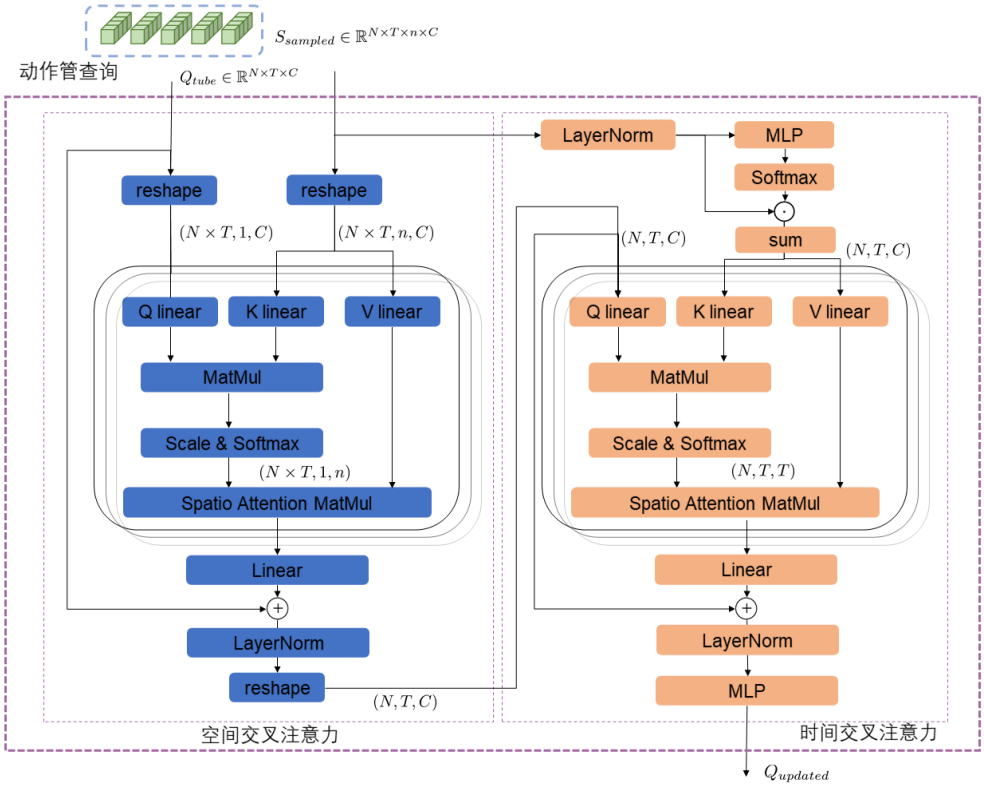


图 4.5 解耦时空交叉注意力模块

接下来，对于采样特征 $S_{sampled} \in \mathbb{R}^{N\times T\times n\times c_f}$ 和动作管查询 $Q_{tube} \in \mathbb{R}^{N\times T\times c_q}$，模型需要将二者进行有效融合，以便将关键的视觉上下文信息注入到动作管查询中。为此，本节设计了解耦时空交叉注意力模块（DSTA），其结构如图 4.5 所示。

不同于（deformable detr）所简单地采样线性加权来进行特征融合，为了能够基

于候选特征自适应地从这 n 个候选点中筛选并聚合最关键的视觉信息，并进一步建模时序依赖，本节设计了时空解耦注意力模块（DSTA），该模块包含空间交叉注意力（Spatial Cross Attention，SCA）和时序自注意力（Temporal Cross Attention，TCA）两个部分，SCA 和 TSA 都是基于多头注意力实现。在 $S_{sampled}$ 和 Q_{tube} 进行交互之前，会先通过线性层将它们的通道数映射到统一的维度 C，以确保后续注意力计算的一致性。首先，SCA 旨在将 n 个离散的采样特征聚合为紧凑的实例特征，对于动作管查询 $Q_{tube} \in \mathbb{R}^{N\times T\times C}$ 和未聚合的采样特征 $S_{sampled} \in \mathbb{R}^{N\times T\times n\times C}$，SCA 先将 Q_{tube} 重塑为 $N \cdot T \times 1 \times C$ 作为查询，并将 $S_{sampled}$ 重塑为 $N \cdot T \times n \times C$ 作为键和值。SCA 通过如下公式进行计算：

$$Q_{spatial} = LayerNorm(\text{SCA}(Q_{tube}, V_{sampled}, V_{sampled}) + Q_{tube}) \tag{4.7}$$

其中，输出特征 $Q_{spatial} \in \mathbb{R}^{N\cdot T\times 1\times C}$，并在计算后恢复为 $N \times T \times C$。在这一过程中，注意力机制会计算查询与每个采样点之间的语义相似度，从而赋予包含显著动作细节的采样点更高的权重。接下来，时序自注意力模块（TCA）负责在时间维度上建模动作管查询的时序依赖，在进行 TCA 计算之前，对于采样特征 $S_{sampled}$ 会先在空间维度进行聚合，得到时序特征 $S_{temporal} \in \mathbb{R}^{N\times T\times c_f}$，具体计算如下：

$$S_{sampled} = LayerNorm(S_{sampled}) \tag{4.8}$$

$$\text{Weights} = \text{Softmax}(\text{MLP}(S_{sampled})) \tag{4.9}$$

$$S_{temporal} = \sum(S_{sampled} \odot \text{Weights}) \tag{4.10}$$

其中，$\odot$ 表示逐元素乘法操作。随后，TCA 将 $Q_{spatial}$ 作为查询，将 $S_{temporal}$ 作为键和值，计算公式如下：

$$Q_{update} = MLP(LayerNorm(\text{TCA}(Q_{spatial}, S_{temporal}, S_{temporal}) + Q_{spatial})) \tag{4.11}$$

最后经过一个前馈神经网络（MLP）得到最终的更新特征 $Q_{update} \in \mathbb{R}^{N\times T\times C}$。

$$Q_{update} = MLP(LayerNorm(Q_{update})) \tag{4.12}$$

可以观察到，在 SCA 中注意力矩阵的形状为 $1 \times n$，而在 TCA 中注意力矩阵的形状为 $T \times T$，这表明 DSTA 通过解耦的方式分别在空间和时间维度上进行注意力计算，在 SCA 中关注空间位置间的关系，而在 TCA 中关注时间步间的依赖关系，通过这

种解耦的时空交叉注意力机制，DSTA 能够有效地融合动作管查询与采样特征，提升模型对时空上下文的挖掘能力，从而更好地捕捉复杂运动场景下的动作特征。

4.6 损失函数和匹配机制

（1）损失函数

AGCM 的总训练损失 $\mathcal{L}_{total}$ 由两部分组成：主损失 $\mathcal{L}_{train}$ 和辅助损失 $\mathcal{L}_{aux}$。

其中，主损失分别包括分类损失 $\mathcal{L}_{class}$、边界框回归损失 $\mathcal{L}_{box}$、管级广义交并比损失 $\mathcal{L}_{giou}$ 以及时序掩码损失 $\mathcal{L}_{tempmask}$ 四部分构成，定义如下：

$$\mathcal{L}_{main} = \lambda_{class}\mathcal{L}_{class} + \lambda_{box}\mathcal{L}_{box} + \lambda_{giou}\mathcal{L}_{giou} + \lambda_{tempmask}\mathcal{L}_{tempmask} \tag{4.13}$$

其中，λ 表示各损失分量的权重系数。

AGCM 在基线方法的基础上对主损失函数进行了调整优化。具体而言，对于分类损失，在 deformable detr 的相关研究中表明，由于查询数量 N 一般大于实际存在的正样本数量 M，这会导致训练时的正负样本不平衡的问题，而焦点损失（Focal loss）可以通过 $(1-p)^{\gamma}$ 项自动降低了那些负样本的损失权重，让模型专注于那些难分类的样本。对于 N 个预测查询，分类损失的计算公式如下：

$$\mathcal{L}_{class} = -\sum_{j=1}^{N}[\alpha(1-\hat{p}_j(c_j))^{\gamma}\log(\hat{p}_j(c_j))] \tag{4.14}$$

其中，c_j 表示第 j 个查询对应的真实类别标签，$\hat{p}_j(c_j)$ 为模型预测该类别的概率，α 和 γ 分别为平衡因子和调节因子。边界框回归损失 $\mathcal{L}_{box}$ 采用 L_1 损失，直接惩罚匹配样本中预测框与真实框在中心点坐标及长宽上的绝对误差。管级广义交并比损失 $\mathcal{L}_{giou}$ 定义为当前实例在所有有效帧上 2D GIoU 的均值：

$$\mathcal{L}_{giou} = \sum_{i=1}^{M}\left(1 - \frac{1}{T^i_{valid}}\sum_{t\in\mathcal{T}_i}\text{GIoU}(b^i_t, \hat{b}^{\hat{\sigma}(i)}_t)\right) \tag{4.15}$$

其中，M 为真实动作实例的总数，$\hat{\sigma}(i)$ 表示与第 i 个真值匹配的预测索引，$\mathcal{T}_i$ 为第 i 个动作实例存在的帧集合，T^i_{valid} 为该集合的帧数，b 与 $\hat{b}$ 分别代表真值框与预测框。此外，为提升动作起止时间的定位精度，模型引入 STDet 所设计的时序掩码损失 $\mathcal{L}_{tempmask}$，利用二元交叉熵（BCE）监督每一帧是否真实存在动作，而不仅仅通

过分类损失的概率预测隐式地预测动作边界：

$$\mathcal{L}_{mask} = -\frac{1}{N \cdot T} \sum_{j=1}^{N} \sum_{t=1}^{T} [m_{j,t} \log(\hat{m}_{j,t}) + (1 - m_{j,t}) \log(1 - \hat{m}_{j,t})] \tag{4.16}$$

其中，$m_{j,t} \in \{0, 1\}$ 为真值标签，表示第 j 个实例在第 t 帧的动作是否存在；$\hat{m}_{j,t}$ 为模型输出的动作存在概率。

而辅助损失则同样采用的焦点损失作为辅助分类损失，用于监督动作感知模块（AAM）中动作感知查询的学习，可表示为：

$$\mathcal{L}_{aux} = \lambda_{class}^{aux} \mathcal{L}_{class}^{aux} \tag{4.17}$$

最终，AGCM 的总训练损失定义为：

$$\mathcal{L}_{total} = \mathcal{L}_{main} + \mathcal{L}_{aux} \tag{4.18}$$

（2）匹配机制

AGCM 采用和基线方法相同的样本匹配机制，对模型输出的固定数量的动作管预测 $\hat{y} = \{\hat{y}_j\}_{j=1}^{N}$，通过匈牙利匹配算法（Hungarian Match Algorithm），在预测集合与真实动作集合 $y = \{y_i\}_{i=1}^{M}$ 之间构建最优二分图匹配。匹配代价综合考量了类别预测准确性、空间位置回归精度以及时序对齐质量。对于第 i 个真值和第 j 个预测，其匹配代价定义为：

$$\mathcal{L}_{match} = \lambda_{class} \mathcal{L}_{class} + \lambda_{box} \mathcal{L}_{box} + \lambda_{giou} \mathcal{L}_{giou} \tag{4.19}$$

最终，通过最小化总匹配代价获得最优映射 $\hat{\sigma}$，并据此计算上述训练损失。

4.7 实验结果与分析

4.7.1 实验设置

本章的所有实验均在 JHMDB、UCF101-24 和 Multisports 数据集上进行评估，以验证所提方法在复杂体育场景下的时空动作检测性能。模型使用 PyTorch 2.1.0 和 Python 3.8 实现，并在 GeForce RTX 4090 GPU 上完成，为了对比公平，所有的对比模型使用 3D CNN 的 CSN152 预训练模型作为视觉特征提取器，它们均在 Kinetics-700 数据集上进行了预训练，模型处理帧数 $T = 32$，输入视频片段的短边被 resize 为

256。查询向量的数量 N 设定为 32，每个动作管查询在各时间步上的采样点 n 设置为 32。查询向量的维度 D_q 设定为 256。在动作解码器中堆叠了 $k=6$ 层解码器层。模型使用 AdamW 优化器，权重衰减设为 0.01，骨干网络的初始学习率设为 1e-5，解码器的初始学习率设为 1e-6。在训练阶段，采用了颜色抖动、随机裁剪和水平翻转等数据增强策略。损失函数的权重配置如下：

$\lambda_{class}=4$，$\lambda_{box}=1$，$\lambda_{giou}=1$，$\lambda_{tempmask}=1$，$\lambda_{class}^{aux}=0.25$。

实验使用 frame-mAP 和 video-mAP 作为性能评价指标，并在通过误差分析综合评估模型，各指标的具体计算方式见 2.5.2 节。

4.7.2 对比实验结果及分析

（1）对比方法的选择为了更加全面地评估本章提出方法的有效性，本节将提出的 ISCM 与当前主流的时空动作检测方法进行了广泛的对比实验，在对比方法中，选择了当前最先进的双阶段方法和单阶段方法进行对比。

对于双阶段方法，本节选择了 HIT 和 TAAD 作为对比对象，HIT 使用 ResNet-50 作为骨干网络的 Fast-RCNN 目标检测器获取人体检测框，并定位运动员的手部区域，通过增强人体和手部信息的关系建模增强模型的动作理解能力；TAAD 利用 YOLOv5 对逐帧进行人体检测，并结合预训练的行人重识别模型 OsNet 以及跟踪算法 deepsort 获得动作管候选，在此基础上判断动作的类别。对于单阶段方法，本节选择了 YOWOv3、SAMOC、TubeR、STMixer、STAR 和 ART 作为对比对象。其中，YOWOv3 和 SAMOC 是基于候选锚框设计的单阶段时空动作检测方法，YOWOv3 通过深度整合卷积和自注意力机制，在通道维度上对 2D 和 3D 特征进行混合，实现了高效的特征融合，SAMOC 通过引入运动分支实现了候选框在时间维度上的扩展，并通过跟踪分支加强了帧间的一致性建模。Tuber、STMixer 和 STAR 则均是基于查询的单阶段时空动作检测方法，与本章方法采用同样的模型范式，其中，Tuber 可以视为此类方法的基准模型，其关键设计在于动作管级的查询设置和正交时空自注意力，将基于查询的范式成功引入到时空动作检测任务中；STMixer 则是面向稀疏时空动作检测，每次推理仅预测关键帧的结果，通过设计一种 4D 特征空间采样方法，将多尺度时空特征图通过双线性插值进行扩展，可以更加高效地进行时空特征采样；STAR 则是在动作解码器中设计了正交时空注意力模块，每个动作管查询仅仅与对应时间步的特征图进行交叉注意力计算，在减少计算量的基础上减少了时序信息对

空间定位的干扰；ART 则是通过设计运动员相关的查询生成模块，通过关键帧中人体检测框提取运动员特征，并将其注入到动作管查询中，引导模型关注与运动员高度相关的时空区域。

（2）定量对比实验结果本节将 ISCM 在 JHMDB51-24、UCF101-24 和 Multisports 三个常用的基准数据集上与这些方法进行了对比实验，以验证所提方法在复杂体育场景下的时空动作检测性能。其中，黑体字部分表示在各项指标上取得的最佳结果，下横线表示在该组对比中次优的结果。表格中绘制 * 表示该模型进行了本地训练评估，未绘制 * 表示结果来源于原论文公布的结果。

表 4.1 JHMDB51-21 数据集上的定量对比结果

方法	检测器	frame-mAP	video-mAP		
			@0.2	@0.5	0.50:0.95
TAAD	YOLO-v5	-	-	82.8	56.4
HIT	fast-RCNN	83.8	89.7	88.1	-
YOWOv3	-	73.1	79.2	78.3	58.7
SAMOC	-	73.1	79.2	78.3	58.7
TubeR	-	84.2	87.3	83.3	59.4
STMixer	-	85.1	88.1	84.0	60.2
STAR	-	86.9	89.5	88.2	-
ART	-	86.9	89.5	88.2	-
Ours	CSN-152	**88.3**	**90.3**	**89.5**	**60.9**

表 4.2 UCF101-24 数据集上的定量对比结果

方法	视频特征提取	frame-mAP	video-mAP		
			@0.2	@0.5	0.50:0.95
TAAD	CSN-152	-	85.6	59.4	29.1
HIT	SlowFast-50	84.8	88.8	74.3	-
YOWOv3	-	73.1	79.2	78.3	58.7
SAMOC	DLA-34	73.1	79.2	78.3	58.7
TubeR	CSN-152	83.2	83.3	58.4	28.9
STMixer	SlowFast-50	84.1	90.1	89.0	60.2
STAR	CSN-152	86.7	87.0	65.4	30.6
ART	-	86.9	89.5	88.2	-
Ours	SlowFast-50	**88.4**	**83.3**	**87.3**	**67.6**

表 4.3 Multisports 数据集上的定量对比结果

方法	视频特征提取	frame-mAP	video-mAP		
			@0.2	@0.5	0.50:0.95
TAAD	SlowFast-50	-	62.8	36.0	-
HIT	SlowFast-50	33.3	27.8	8.8	-
YOWOv3	-	73.1	79.2	78.3	58.7
SAMOC	DLA-34	31.5	16.5	8.9	-
TubeR	SlowFast-50	-	59.4	31.7	-
STMixer	SlowFast-101	50.4	57.7	35.2	-
STAR	CSN-152	45.4	50.1	18.6	-
ART	-	86.9	89.5	88.2	-
Ours	SlowFast-50	**88.4**	**83.3**	**87.3**	**67.6**

根据表格中的结果可以看出，ISCM 在三个数据集上均取得了一致性的性能提升，相比而言，ISCM 在 JHMDB51-21 数据集上相对次优的结果 frame-mAP 提升了 1.4%，video-mAP@0.2,video-mAP@0.5 提升了 1.2%,video-mAP@0.50:0.95 提升了 0.7%; 在 UCF101-24 数据集上，ISCM 相对次优的结果 frame-mAP 提升了 1.7%，video-mAP@0.2 提升了 2.3%，video-mAP@0.5 提升了 12.9%，video-mAP@0.50:0.95 提升了 7.0%; 在 Multisports 数据集上，ISCM 的 frame-mAP 提升了 3.0%，video-mAP@0.2 提升了 5.2%，video-mAP@0.5 提升了 18.7%，video-mAP@0.50:0.95 提升了 7.4%。对于数据集间表现的差异，可以从运动剧烈程度和场景复杂度来看，JHMDB51-21 数据集中的动作相对较为简单，运动幅度和遮挡情况较少，ISCM 在该数据集上提升相对有限，而 UCF101-24 和 Multisports 数据集包含了更多复杂的体育动作和多样化的场景，这一点第 2 章数据集分析部分已有详细说明。对于不同方法直接的差异，可以从模型范式的特点来看，基于候选框的单阶段方法 YOWOv3 和 SAMOC 虽然通过设计 3 维锚框来端到端的时空动作检测，但是形状固定的锚框在处理复杂运动和遮挡时存在局限性，导致其性能相对较低，尤其是在 UCF101-24 和 Multisports 数据集上表现不佳。而带有额外检测器的双阶段方法 TAAD 和 HIT 的表现则优于基于锚框的方法，但由于其依赖于预训练的检测器，且未能充分利用时空上下文信息，对于复杂动作的时空边界定位仍存在不足。相比之下，基于查询的单阶段方法如 Tuber、STMixer、STAR 和 ART 的表现更好。但是，不同于 Tuber、STMixer 和 STAR 的动作管查询仅仅通过随机初始化获得，从物理含义上说，这些动作管查询其实也相当

于一种预设的锚框，限制了模型对复杂运动的适应性。ART 虽然通过运动员相关的查询生成模块引入了运动员特征，但运动员的外观信息并不能完全反映动作的动态变化，所提供的先验信息属实有限，而 ISCM 通过对视频实例进行感知，直接将动作实例的信息注入到动作管查询中，使得每个查询都能够动态地适应当前视频中的复杂运动。

下图表示的是 ISCM 和 Tuber 在 JHMDB51-21、UCF101-24 和 Multisports 数据集上，各个动作类别的 video-mAP@20 的对比结果。从图中可以看出 ISCM 在复杂运动场景下，能够更准确地定位动作的时空边界，并且在处理快速运动和遮挡等挑战时表现出更强的鲁棒性。

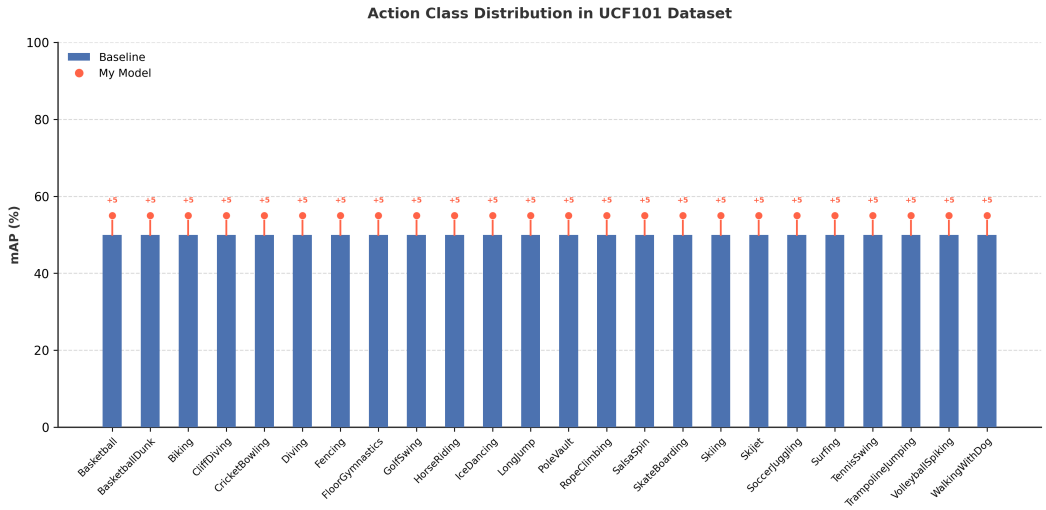


图 4.6 test1

4.7.3 消融实验结果

为了进一步评估所提出方法的有效性，本节在 UCF101-24 和 Multisports 数据集上进行了消融实验，具体评估了 ISCM 中各个模块对整体性能的贡献。基线模型为不包含 ISCM 的 Tuber 模型，主要评估了时空动作感知模块（TAM）、正交自适应采样模块（）以及解耦时空注意力模块（DSTA）。为了加强对比的说服力，本节分别将各个模块替换为对应的对比设计，具体来说：对于时空动作感知模块（），本节通过随机初始化的动作管查询替换为由 TAM 生成的查询向量，但是会为动作管查询加入时间位置编码，以补充时序信息，对于正交自适应采样模块（），本节将其替换为

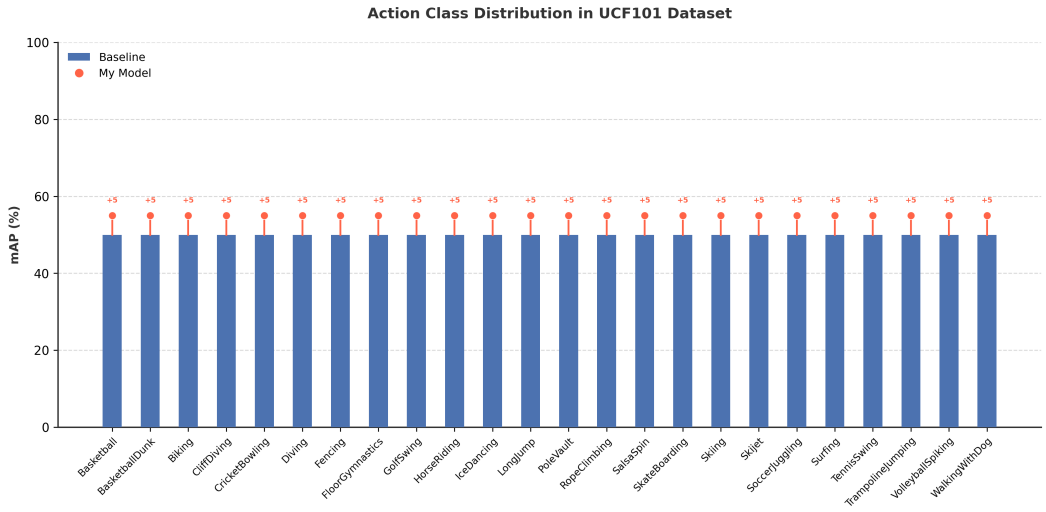


图 4.7 test2

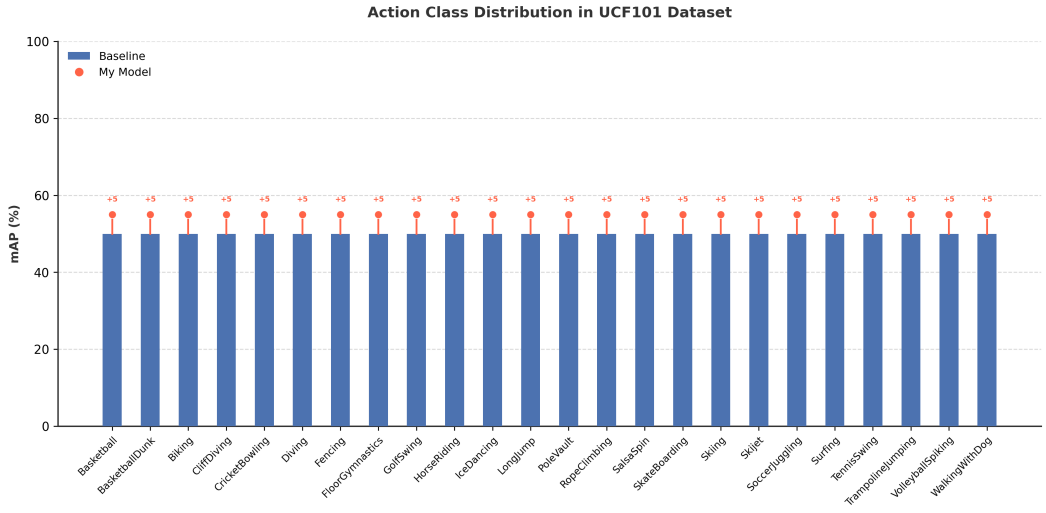


图 4.8 test3

固定网格采样，采样点的数量同样设置为 32；对于解耦时空注意力模块，本节将为标准的时空交叉注意力机制。每个消融模型保证训练超参数和数据预处理完全一致，以确保对比的公平性。以下是各个模块的消融实验结果：

表 4.4 在 UCF101-24 数据集上进行消融实验

TAM	ADSM	DSTA	frame-mAP	video-mAP		
				@0.2	@0.5	0.50：0.90
B	B	B	31.2	14.2	33.6	45.1
B	B	B	33.0	15.4	34.7	45.7
B	B	B	34.8	16.7	35.5	47.4
B	B	B	36.0	**18.8**	37.5	46.0
B	B	B	**37.0**	17.9	**38.1**	**47.3**

表 4.5 在 UCF101-24 数据集上进行消融实验

TAM	ADSM	DSTA	frame-mAP	video-mAP		
				@0.2	@0.5	0.50：0.90
B	B	B	31.2	14.2	33.6	45.1
B	B	B	33.0	15.4	34.7	45.7
B	B	B	34.8	16.7	35.5	47.4
B	B	B	36.0	**18.8**	37.5	46.0
B	B	B	**37.0**	17.9	**38.1**	**47.3**

根据模型在 UCF101-24 和 Multisports 数据集上的消融实验结果可以看出，随着各个模块的逐步引入，模型的性能得到了显著提升。具体来说，引入时空动作感知模块（TAM）后，模型的 frame-mAP 提升了 1.8% 和 1.7%，video-mAP@0.2 提升了 1.2% 和 1.3%，video-mAP@0.5 提升了 1.1% 和 1.2%，video-mAP@0.50:0.95 提升了 1.7% 和 1.5%。这表明 TAM 能够有效地为动作管查询提供与视频内容相关的先验信息，从而提升了模型对复杂动作的理解能力。引入正交自适应采样模块（ADSM）后，模型的 frame-mAP 提升了 1.8% 和 1.6%，video-mAP@0.2 提升了 1.3% 和 1.2%，video-mAP@0.5 提升了 0.8% 和 1.0%，video-mAP@0.50:0.95 提升了 1.6% 和 1.4%。这表明 ADSM 通过动态调整采样点的位置，更好地捕捉了动作的时空特征，从而提升了动作的定位精度。引入解耦时空注意力模块（DSTA）后，模型的 frame-mAP 提升了 1.2% 和 1.4%，video-mAP@0.2 提升了 1.1% 和 1.0%，video-mAP@0.5 提升了

0.6% 和 0.8%，video-mAP@0.50:0.95 提升了 0.9% 和 1.2%。这表明 DSTA 通过解耦时空信息，提升了模型对时空特征的建模能力，从而进一步提升了动作的检测性能。

为了进一步评估所提方法各个模块之间的效果，在 UCF101-24 和 Multisports 数据集上将测试集样本按照运动强度划分为三个子集，具体划分方式参考 TAAD，分别对应小运动、中运动、剧烈运动以下是 ISAM 在各子集上的与现有方法的对比结果：

表 4.6 在 UCF101-24 数据集上进行消融实验

TAM	ADSM	DSTA	frame-mAP			video-mAP		
			Small	Medium	Large	Small	Medium	Large
B	B	B	49.6	54.9	31.2	14.2	33.6	45.1
B	B	B	50.6	56.3	33.0	15.4	34.7	45.7
B	B	B	53.9	57.7	34.8	16.7	35.5	47.4
B	B	B	54.4	58.4	36.0	**18.8**	37.5	46.0
B	B	B	**53.4**	**60.4**	**37.0**	17.9	**38.1**	**47.3**

表 4.7 在 Multisports 数据集上进行消融实验

TAM	ADSM	DSTA	frame-mAP			video-mAP		
			Small	Medium	Large	Small	Medium	Large
B	B	B	49.6	54.9	31.2	14.2	33.6	45.1
B	B	B	50.6	56.3	33.0	15.4	34.7	45.7
B	B	B	53.9	57.7	34.8	16.7	35.5	47.4
B	B	B	54.4	58.4	36.0	**18.8**	37.5	46.0
B	B	B	**53.4**	**60.4**	**37.0**	17.9	**38.1**	**47.3**

根据表格中的结果可以看出，随着各个模块的逐步引入，模型在不同运动强度子集上的性能均得到了显著提升。具体来说，引入时空动作感知模块（TAM）后，模型在小运动、中运动和剧烈运动子集上的 frame-mAP 分别提升了 1.0%、1.4% 和 1.8%，video-mAP 分别提升了 1.2%、1.1% 和 1.1%。这表明 TAM 能够有效地为动作管查询提供与视频内容相关的先验信息，从而提升了模型对复杂动作的理解能力。引入正交自适应采样模块（ADSM）后，模型在小运动、中运动和剧烈运动子集上的 frame-mAP 分别提升了 3.3%、1.4% 和 1.8%，video-mAP 分别提升了 1.3%、0.8% 和 1.0%。这表明 ADSM 通过动态调整采样点的位置，更好地捕捉了动作的时空特征，

从而提升了动作的定位精度。引入解耦时空注意力模块（DSTA）后，模型在小运动、中运动和剧烈运动子集上的 frame-mAP 分别提升了 0.5%、1.7% 和 1.0%，video-mAP 分别提升了 1.1%、0.6% 和 0.8%。这表明 DSTA 通过解耦时空信息，提升了模型对时空特征的建模能力，从而进一步提升了动作的检测性能。且对于剧烈运动子集上的提升尤为显著，表明所提方法在处理复杂运动场景下具有较强的适应性和鲁棒性。

4.7.4 可视化分析

为了进一步说明所提出的模块的有效性，本节对 ISCM 在 UCF101-24 和 Multisports 数据集上的进行可视化分析，具体包括时空动作感知模块（TAM）生成的动作管查询的可视化结果，以及正交自适应采样模块（ADSM）采样点位置的可视化结果。

对于时空动作感知模块（TAM），本节将生成的动作管查询

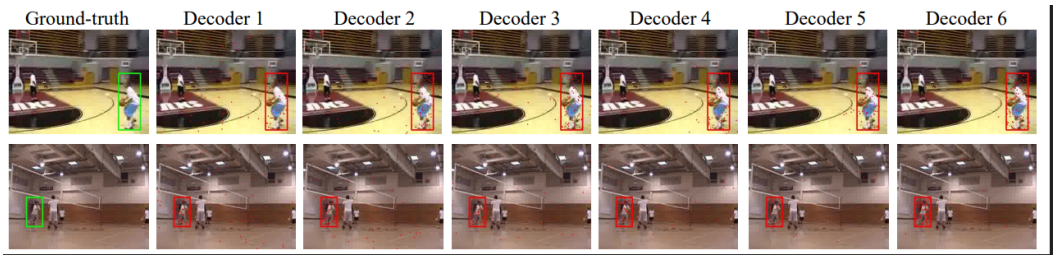


图 4.9 test4

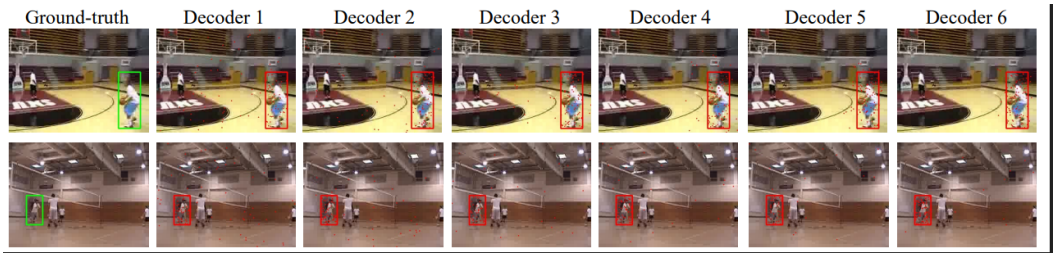


图 4.10 test4

4.8 本章小结

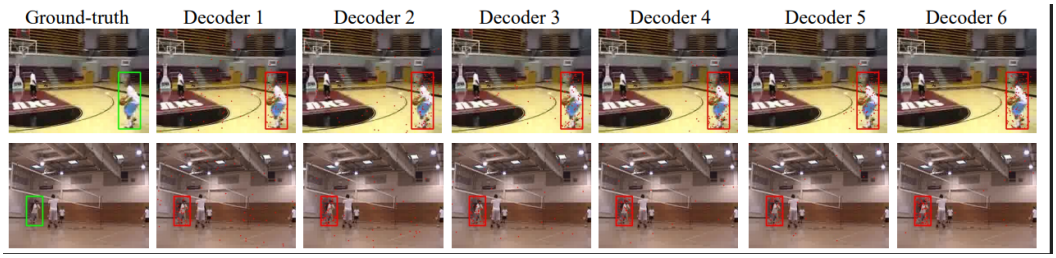

Ground-truth
Decoder 1
Decoder 2
Decoder 3
Decoder 4
Decoder 5
Decoder 6

图 4.11 test4

5 总结与展望

5.1 本文主要内容及结论

对全文进行全面地总结，并根据各章节归纳出若干有机联系的论点。按正文的内容分段描述，包括本研究“做了什么（提出 ** 新理论/算法、设计或研发 ** 工艺/仪器）、获取什么结果、得出什么结论”。

请特别注意，全文总结与摘要及各章的小节要有所区分，不能简单的拷贝。这里的重点是结论，结论应该准确、完整、明确、精练。

5.2 本文主要创新点

通常情况下，学位论文的创新点应放在最后一章。

创新点要凝炼，表述要清晰明了，如提出了什么创新的思路，主要特点是什么，相比现有理论或技术的提高是什么、或者有什么新的发现，是否具有重要的科学意义或应用前景。既不能过于简单，也不要太细。

硕士学位论文创新点不宜太多，一般为 2 个左右即可，要注意归纳创新点，千万不要以为越多越好。论文的创新不以创新点的多少来评定的，而以其创新性的价值来评定。几章的工作合在一起凝炼成一个创新点也不是不可以的。

5.3 展望

对本研究成果的意义、推广应用的现实性或可能性加以论述。同时，描述本文研究中尚存在的不足，或因时间尚未完成但又必须继续的工作，对进一步的工作进行展望。

致　　谢

对在课题研究及论文写作过程中给予指导和帮助的导师、校内外专家、实验技术人员、同学等表示感谢。

在致谢时建议具体，不同的人如何助力完成你的论文，都需要特别注明。如导师、其他老师或实验技术人员、以及同学对你论文的贡献是不一样的，有指引课题方向、修改论文，也有具体教会实验操作，也有协助你做了哪方向的实验，或者给你精神安慰、陪你度过紧张的研究生生涯。

越具体越能表达你真实的感受，否则就是毫无意义的套话。

参考文献

[1] 规划院信息网络中心. 《中国视听大数据（CVB）2025 年体育赛事收视报告》. Technical report, 国家广播电视总局, 12, 2025. https://www.nrta.gov.cn/art/2025/12/26/art_114_72195.html.

[2] 傅潇雯. 《AI 无所不在让体育更智能》. Technical report, 中国体育报, 4, 2025. https://www.sport.gov.cn/n20001280/n20745751/c28601905/content.html.

[3] J. Zhao, S. Liao, X. Li, B. Shuai, C. Chen. ART: Actor-Related Tubelet for Detecting Complex-shaped Action Tubes, 2024. https://openreview.net/forum?id=ICr9KMxa1K.

[4] R. Girshick. Fast r-cnn. in: Proceedings of the IEEE international conference on computer vision, 2015:1440–1448.

[5] S. Saha, G. Singh, M. Sapienza, P. H. Torr, F. Cuzzolin. Deep learning for detecting multiple space-time action tubes in videos. arXiv preprint arXiv:1608.01529, 2016.

[6] O. Ulutan, S. Rallapalli, M. Srivatsa, C. Torres, B. Manjunath. Actor conditioned attention maps for video action detection. in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020:527–536.

[7] Y.-D. Zheng, G. Chen, M. Yuan, T. Lu. Mrsn: Multi-relation support network for video action detection. in: 2023 IEEE International Conference on Multimedia and Expo (ICME). IEEE, 2023:1026–1031.

[8] A. Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. Attention is all you need. Advances in neural information processing systems, 2017, 30.

[10] G. J. Faure, M.-H. Chen, S.-H. Lai. Holistic interaction transformer network for action detection. in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023:3340–3350.

[11] L. Chen, Z. Tong, Y. Song, G. Wu, L. Wang. Efficient video action detection with token dropout and context refinement. in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023:10388–10399.

[12] Y. Li, W. Lin, J. See, N. Xu, S. Xu, K. Yan, et al. Cfad: Coarse-to-fine action detector for spatiotemporal action localization. in: European Conference on Computer Vision. Springer, 2020:510–527.

[13] J. Luo, Y. Yang, R. Liu, L. Chen, H. Fei, C. Hu, et al. A Tracking-Based Two-Stage Framework for Spatio-Temporal Action Detection. Electronics, 2024, 13(3):479.

[14] J. Redmon, S. Divvala, R. Girshick, A. Farhadi. You only look once: Unified, real-time object detection. in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:779–788.

[15] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, Q. Tian. Centernet: Keypoint triplets for object detection. in: Proceedings of the IEEE/CVF international conference on computer vision, 2019:6569–6578.

[16] P. Sun, R. Zhang, Y. Jiang, T. Kong, C. Xu, W. Zhan, et al. Sparse r-cnn: End-to-end object detection with learnable proposals. in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021:14454–14463.

[17] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko. End-to-end object detection with transformers. in: European conference on computer vision. Springer, 2020:213–229.

[18] S. Chen, P. Sun, E. Xie, C. Ge, J. Wu, L. Ma, et al. Watch only once: An end-to-end video action detection framework. in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021:8178–8187.

[19] Y. Li, Z. Wang, L. Wang, G. Wu. Actions as moving points. in: European Conference on Computer Vision. Springer, 2020:68–84.

[20] J. Zhao, Y. Zhang, X. Li, H. Chen, B. Shuai, M. Xu, et al. Tuber: Tubelet transformer for video action detection. in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022:13598–13607.

[21] A. A. Gritsenko, X. Xiong, J. Djolonga, M. Dehghani, C. Sun, M. Lucic, et al. End-to-end spatio-temporal action localisation with video transformers. in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024:18373–18383.

[22] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, C. Schmid. Vivit: A video vision transformer. in: Proceedings of the IEEE/CVF international conference on computer vision, 2021:6836–6846.

[23] T. Wu, M. Cao, Z. Gao, G. Wu, L. Wang. Stmixer: A one-stage sparse action detector. in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023:14720–14729.

[24] Y. Li, Z. Wang, Z. Li, L. Wang. Sparse action tube detection. IEEE Transactions on Image Processing, 2024, 33:1740–1752.

[25] K. Soomro, A. R. Zamir. Action recognition in realistic sports videos. in: Computer vision in sports, Springer, 2015:181–208.

[26] K. Soomro, A. R. Zamir, M. Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402, 2012.

[27] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei. Large-scale video classification with convolutional neural networks. in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014:1725–1732.

[28] D. Shao, Y. Zhao, B. Dai, D. Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020:2616–2625.

[29] A. Cioppa, S. Giancola, V. Somers, F. Magera, X. Zhou, H. Mkhallati, et al. SoccerNet 2023 challenges results. Sports Engineering, 2024, 27(2):24.

[30] J. Rao, H. Wu, H. Jiang, Y. Zhang, Y. Wang, W. Xie. Towards universal soccer video understanding. in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025:8384–8394.

[31] J. Rao, Z. Li, H. Wu, Y. Zhang, Y. Wang, W. Xie. Multi-agent system for comprehensive soccer understanding. in: Proceedings of the 33rd ACM International Conference on Multimedia, 2025:3654–3663.

[32] H. Yang, J. Rao, H. Wu, W. Xie. SoccerMaster: A Vision Foundation Model for Soccer Understanding. arXiv preprint arXiv:2512.11016, 2025.

[33] J. Rao, H. Wu, C. Liu, Y. Wang, W. Xie. Matchtime: Towards automatic soccer game commentary generation. arXiv preprint arXiv:2406.18530, 2024.

[34] W. Tian, R. Lin, H. Zheng, Y. Yang, G. Wu, Z. Zhang, et al. SportsGPT: An LLM-driven Framework for Interpretable Sports Motion Assessment and Training Guidance. arXiv preprint arXiv:2512.14121, 2025.

[35] Z. Liu, X. Xie, M. He, W. Zhao, Y. Wu, L. Cheng, et al. Smartboard: Visual Exploration of Team Tactics with LLM Agent. IEEE Transactions on Visualization and Computer Graphics, 2024.

[36] G. Singh, V. Choutas, S. Saha, F. Yu, L. Van Gool. Spatio-temporal action detection under large motion. in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023:6009–6018.

[37] Y. Gai, W. He, Z. Zhou. Pedestrian target tracking based on DeepSORT with YOLOv5. in: 2021 2nd International Conference on Computer Engineering and Intelligent Control (ICCEIC). IEEE, 2021:1–5.

[38] T.-T. Cao, X.-H. Manh, A.-H. Kieu, T.-T. Nguyen, V.-H. Dao, H. Vu, et al. OSNet-DCN: Integrating Deformable Feature Learning for Tracking Lesion in Endoscopic Videos. in: 2025 International Conference on Multimedia Analysis and Pattern Recognition (MAPR). IEEE, 2025:1–6.

[39] J. Xu, G. Zhao, S. Yin, W. Zhou, Y. Peng. Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024:21773–21782.

[40] J. Li, D. Li, C. Xiong, S. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. in: International conference on machine learning. PMLR, 2022:12888–12900.

[41] T. Wu, R. He, G. Wu, L. Wang. Sportshhi: A dataset for human-human interaction detection in sports videos. in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024:18537–18546.

[42] J. Lin, C. Gan, S. Han. Tsm: Temporal shift module for efficient video understanding. in: Proceedings of the IEEE/CVF international conference on computer vision, 2019:7083–7093.

附录 1　攻读硕士学位期间取得的研究成果

发表与接收论文

[1] Linqiang Pan, **Lianghao Li**, Ran Cheng, Cheng He, Kay Chen Tan.[J]. IEEE Transactions on Cybernetics, vol. 58, no. 6, pp. 3325-3337, 2019. (SCI 源刊; IF:11.448; 署名单位: 华中科技大学)

[2] 参照参考文献列出学术论文相关信息（含期刊、会议、或参编书稿），但无论有多少个作者，都必须列出全部作者名；若为英文论文，则名在前、姓在后，姓名均为全称；在本人的名字加粗，以示区别（若为第一作者，则需在最后特别注明署名华中科技大学是否为第一单位）

[3] 若已发表，按参考文献给出页码；若只是 online, 给出链接；若接受或修改或投稿或拟投，也必须分别注明

[4] 一般情况，一作或重要的论文放在前面

专　　利

[1] 全部作者的姓名全称，本人的名字加粗. 专利题名. 专利国别，专利文献种类，专利号或申请号

标　　准

[1] 全部作者的姓名全称，本人的名字加粗. 标准题名. 哪种层次的标准，发表年

科技奖励

[1] 全部作者的姓名全称，本人的名字加粗. 题目. 国家级/省部级科技类奖，获奖年

[2] 全部作者的姓名全称，本人的名字加粗. 题目. 国际/国内竞赛类奖，获奖年

附录 2　攻读硕士学位期间参与的科研项目

1. 项目类型

项目名称: 项目名称

项目编号: No. 88888888

起止时间: 2018 年 8 月至 2018 年 8 月

担任角色：担任角色

附录 3　其他附录

可包括详细的公式推导、实验数据、计算程序、援引他人的原始资料、数据及其设备条件等。