

改进 RT-DETR 的足球比赛目标检测^①

蔡逸磊, 王 雷

(中国科学技术大学 信息科学与技术学院 自动化系, 合肥 230026)

通信作者: 蔡逸磊, E-mail: cmoon@mail.ustc.edu.cn



摘 要: 足球比赛视频的高精度目标检测对于智能战术分析与赛事转播具有重要应用价值. 然而, 场景中目标尺度跨度大, 运动模糊以及球员密集遮挡等问题, 严重制约了检测的精度与鲁棒性. 本文提出一种改进 RT-DETR 的检测算法. 针对微小目标因运动模糊导致的特征混叠与丢失, 使用残差哈尔离散小波变换 (residual Haar discrete wavelet transform, RHDWT) 下采样模块, 利用小波频域特性有效保留高频边缘细节. 引入内容感知特征重组 (content-aware reassembly of features, CARAFE) 模块替代线性插值上采样, 通过动态预测重组核调整采样区域, 提升多尺度特征的语义对齐精度. 最后在解码器输入端构建多尺度动态特征精炼模块 (multi-scale dynamic feature refiner module, MDFRM), 融合卷积注意力机制, 自适应增强目标区域特征响应并抑制背景噪声干扰. 在 SoccerNet-Tracking 数据集上的实验结果表明, 与 RT-DETR 基准模型及主流 YOLO 系列模型相比, 该算法在各项精度指标上均取得了最优结果, 精确率与召回率分别达到 89.9% 和 85.4%, $mAP@50$ 达到 84.9%. 该方法有效解决了复杂足球比赛场景下的漏检与误检问题, 为足球比赛智能分析提供可靠支撑.

关键词: RT-DETR; 目标检测; 体育分析; 小波变换; 注意力机制

引用格式: 蔡逸磊, 王雷. 改进 RT-DETR 的足球比赛目标检测. 计算机系统应用, 2026, 35(9): 196–207. <http://www.c-s-a.org.cn/1003-3254/10239.html>

Soccer Match Object Detection Based on Improved RT-DETR

CAI Yi-Lei, WANG Lei

(Department of Automation, School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China)

Abstract: High-precision object detection in soccer match videos is of great application significance for intelligent tactical analysis and event broadcasting. However, challenges such as large variations in object scale, motion blur, and dense player occlusion severely limit detection accuracy and robustness. This study proposes a detection algorithm based on improved RT-DETR. A residual Haar discrete wavelet transform (RHDWT) downsampling module is employed to address feature aliasing and loss in tiny objects caused by motion blur. This module utilizes wavelet frequency domain properties to effectively preserve high-frequency edge details. Additionally, the content-aware reassembly of features (CARAFE) module is introduced to replace linear interpolation upsampling. By dynamically predicting reassembly kernels to adjust sampling regions, the semantic alignment accuracy of multi-scale features is improved. Finally, a multi-scale dynamic feature refiner module (MDFRM) is constructed at the input of the decoder. It integrates a convolutional attention mechanism to adaptively enhance feature responses of object regions and suppress background noise interference. Results from experiments conducted on the SoccerNet-Tracking dataset demonstrate that the proposed algorithm outperforms the baseline model RT-DETR and mainstream YOLO series models, yielding the best performance across all accuracy metrics. The precision and recall reach 89.9% and 85.4% respectively, and $mAP@50$ reaches 84.9%.

① 收稿时间: 2026-01-16; 修改时间: 2026-02-05, 2026-02-27; 采用时间: 2026-03-06; csa 在线出版时间: 2026-06-18

CNKI 网络首发时间: 2026-06-22

This method effectively solves the problems of missed detection and false detection in complex soccer match scenes, providing reliable support for intelligent analysis of soccer matches.

Key words: RT-DETR; object detection; sport analytics; wavelet transform; attention mechanism

1 引言

随着计算机视觉技术的飞速发展,智能体育视频分析已成为人工智能领域的重要研究方向。足球作为全球最受欢迎的体育运动,其比赛视频蕴含着巨大的分析价值,不仅对职业球队的战术研究、球员表现评估至关重要,也为体育媒体、训练指导和商业分析提供了关键的技术支撑^[1,2]。在足球比赛中,对足球、球员及裁判等关键目标的高精度、实时检测,是构建高层语义分析,实现战术自动分析、赛事转播增强及辅助判罚系统的基础^[3]。然而,足球比赛场景具有极高的动态性和复杂性,视频图像往往面临目标尺度跨度大、足球和球员高速运动产生模糊、肢体密集遮挡以及观众席与广告牌造成的复杂背景干扰等多重挑战,这对目标检测算法的鲁棒性提出了严苛要求^[4]。

早期的体育赛事视频分析主要依赖背景差分法、光流法及手工设计的特征(如HOG、SIFT)结合支持向量机进行识别,难以适应光照变化剧烈的赛场环境^[5]。随着深度学习的发展,目标检测算法逐渐取代传统方法成为主流。以R-CNN^[6]和Faster R-CNN^[7]为代表的双阶段算法虽在检测精度上表现出色,但其基于大量预选锚框进行二次特征提取的级联架构限制了运算速度,导致推理延迟较高,难以满足直播流处理的实时需求。

以YOLO^[8]系列为代表的单阶段算法因在速度上占据优势而被广泛应用。Komorowski等人^[9]提出的FootAndBall采用全卷积架构与特征金字塔设计,专为足球场景优化,以极低的参数量实现了实时检测,但其在复杂遮挡下的鲁棒性已显不足。Moutselos等人^[10]通过对比实验验证了YOLOv8n模型在远景实时球员与球检测场景下的性能显著优于FootAndBall。于兆廷等人^[11]基于YOLOv8n架构提出的ECABiF-Y5n模型虽然提升了足球场景的精度,但在处理高速运动模糊时仍表现不佳。近期,YOLOv10^[12]移除了非极大值抑制(non-maximum suppression, NMS)以降低延迟,YOLO-11^[13]与YOLO12^[14]则进一步引入注意力机制以增强特征提取能力。然而,CNN架构固有的局部感受野限制了其捕捉长距离依赖的能力,在推断被遮挡球员或利

用全局上下文信息识别微小目标时,仍存在难以突破的瓶颈。

针对卷积神经网络难以捕捉长距离依赖的缺陷,Carion等人^[15]提出的DETR(detection Transformer)引入了全新的视觉范式。借助Transformer核心的自注意力机制,模型能够在全图范围内交互特征信息,这种全局感受野对于推断赛场上的严重遮挡与微小目标至关重要。同时,基于二分图匹配的集合预测策略彻底摒弃了非极大值抑制等人工先验操作,实现了真正的端到端检测。面对原始DETR收敛缓慢且计算复杂度高的问题,Zhu等人^[16]随后引入多尺度可变形注意力(deformable attention)模块,迫使注意力仅聚焦于参考点周围的关键采样位置,在大幅降低计算量的同时加速了模型收敛。直至2024年,Zhao等人^[17]推出的RT-DETR通过设计高效混合编码器解耦多尺度特征处理,在保留全局建模优势的同时显著提升了推理速度,成功兼顾了高精度与实时性,成为当前实时目标检测领域的先进基准。

尽管RT-DETR在COCO等通用数据集上表现优异,但其核心设计仍主要面向静态图像的通用检测任务,直接迁移到足球比赛等体育赛事视频场景中仍面临严峻挑战。体育赛事视频普遍存在多视角切换与高速运动带来的运动模糊,使得目标边缘细节退化^[18];同时远景镜头下足球与球员像素占比极低,小目标特征易被下采样与多尺度融合过程削弱,难以兼顾实时性与精细识别^[19]。近期研究也从不同角度揭示了RT-DETR的不足与改进空间,Madan等人^[20]的工作指出RT-DETR在小目标上的鲁棒性仍受限,并通过对混合编码器与损失函数等结构的增强提升小目标检测的性能。Yeung等人^[21]在SoccerNet评测对比中指出,RT-DETR在微调后虽能显著提升召回率,但精确率会下降,表明其在密集遮挡场景下仍易产生较多误检,检测稳定性不足。在球类运动的复杂场景中,李萌萌等人^[22]基于上下文感知融合的改进模型相较RT-DETR基准模型取得了精度增益,进一步表明RT-DETR在此类动态场景下仍存在可优化空间。

针对上述短板与挑战,本文提出一种改进的 RT-DETR 算法,旨在在保证实时性的前提下,提升复杂足球场景下的多尺度目标检测性能. 本文的主要贡献如下.

(1) 针对高速运动模糊问题,在主干网络与编码器的连接处引入残差哈尔离散小波变换 (residual Haar discrete wavelet transform, RHDWT) 下采样模块. 利用小波变换的频域特性保留图像的高频边缘细节,有效解决足球等高速微小目标在下采样过程中的特征模糊与混叠问题.

(2) 针对小目标和特征对齐问题,在跨尺度特征融合网络中引入内容感知特征重组 (content-aware reassembly of features, CARAFE) 模块替代传统的插值上采样. 通过预测自适应重组核,根据输入特征的语义内容动态调整采样区域,实现特征图的精准对齐与细节恢复.

(3) 针对密集遮挡和背景干扰问题,在解码器输入前设计了多尺度动态特征精炼模块 (multi-scale dynamic feature refiner module, MDFRM). 自适应地增强感

兴趣区域的特征响应并抑制背景噪声,显著提升了模型在密集遮挡与复杂背景下的鲁棒性.

2 算法设计与实现

2.1 改进的 RT-DETR 算法

针对足球比赛视频中普遍存在的目标尺度跨度大、高速运动模糊导致的边缘退化,以及密集遮挡与复杂背景干扰 (如观众席、广告牌) 引发的误检漏检问题,本文在 RT-DETR-r18 的基础上,提出了一种改进的目标检测算法. 该算法针对足球比赛场景的特殊性进行针对性改进,改进后的整体网络架构如图 1 所示. 本文在保留 RT-DETR-r18 基础骨干与混合编码器整体架构的前提下,对其特征融合与输出路径进行了改进. 图 1 展示了 3 个关键改进模块的嵌入位置: RHDWT 与 CARAFE 模块分别替换了原跨尺度特征融合模块 (cross-scale feature fusion module, CCFM) 中的下采样与上采样算子,而 MDFRM 模块则作为独立的特征增强单元嵌入在编码器输出与解码器输入之间.

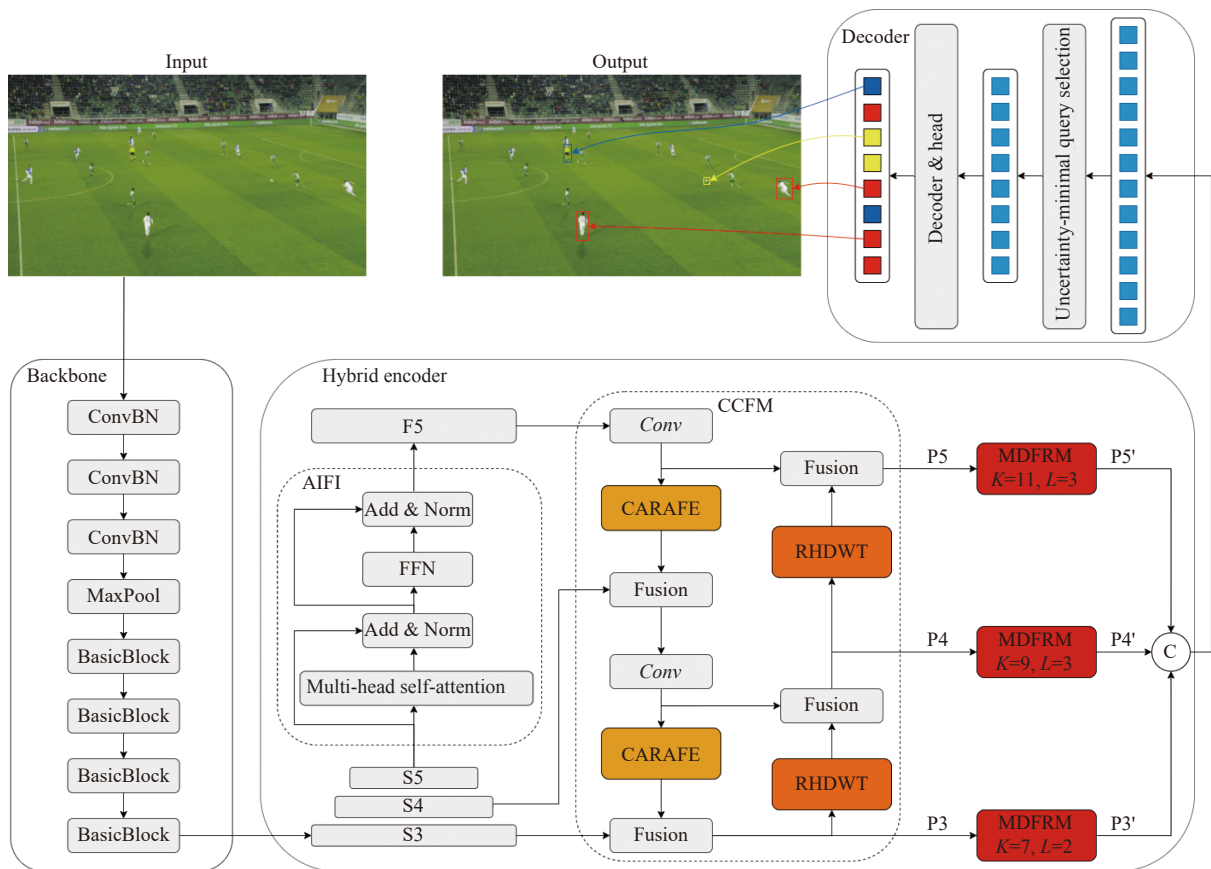


图1 算法框架结构图

作为改进的基础, RT-DETR-r18 遵循经典的主干-颈部-头部设计模式. 在特征提取阶段, 算法沿用 ResNet-18 作为主干网络. 输入图像经过层级卷积处理后, 输出 3 个不同尺度的特征层 S3、S4 和 S5. 其中, S3 保留了较多空间细节与边缘纹理信息, 对小目标的足球检测更为关键; S4 位于细节与语义之间的中间尺度, 兼具一定的空间分辨率与更强的语义表征能力, 能够为近中景球员、裁判等中等尺度目标提供更稳定的特征描述; S5 包含高度抽象的语义信息与更大的感受野, 有助于建模全局上下文并提升复杂背景下的类别判别能力.

在特征编码与融合阶段, 尺度内特征交互模块 (AIFI) 对 S5 特征进行自注意力运算以捕捉全局长距离依赖, 而基于 CNN 的跨尺度特征融合模块 (CCFM) 则负责在 S3-S5 之间进行双向信息融合. 最后, 融合后的特征 P3、P4 和 P5 被送入 Transformer 解码器, 配合预测头完成端到端的类别分类与边界框回归.

本文在 RT-DETR 基础架构的基础上, 对特征融合中的上采样和下采样层以及编码器的输出路径做出针对性改进.

(1) 基于残差小波变换的抗模糊下采样: 针对运动模糊问题, 引入残差哈尔离散小波变换 (RHDWT) 模块, 具体而言, 在 CCFM 的自底向上融合路径中, 使用 RHDWT 模块替换原网络中用于降低分辨率的 3×3 步长卷积层. 该模块利用离散哈尔小波变换将特征图分解为高低频子带, 在降低分辨率的同时保留目标轮廓的高频分量, 并通过残差连接融合空间结构信息, 有效解决足球等高速目标在下采样过程中的特征模糊与混叠问题.

(2) 基于内容感知重组的语义对齐上采样: 针对小目标和特征对齐问题, 本文引入了内容感知特征重组 (CARAFE) 模块替代原网络中使用的最近邻插值上采样算子. CARAFE 模块能够根据输入特征的语义内容, 动态生成实例特异性的上采样核, 从而自适应地调整采样区域与权重. 这一改进确保了高层语义特征在映射回高分辨率时, 能够与底层细节特征实现精准的语义对齐与细节恢复.

(3) 基于混合动态注意力的多尺度特征精炼: 针对密集遮挡和背景干扰问题, 在混合编码器输出与 Transformer 解码器输入之间, 设计了多尺度动态特征精炼模块 (MDFRM). 该模块接收 CCFM 输出的 3 层多尺度特征 P3、P4 和 P5 作为输入, 分别进行特征去噪与

增强处理, 精炼后的特征序列 P3'、P4' 和 P5' 随后被送入解码器. MDFRM 通过多尺度金字塔获取局部与全局不同语义层次的特征, 以及输入自适应的混合注意力机制, 结合动态小核卷积和静态大核卷积, 突出关键的结构区域并抑制冗余信息, 并且针对不同分辨率特征层的特点采用了差异化的参数配置, 显著提升了模型在密集遮挡与复杂背景下的鲁棒性.

本文方法在保持 RT-DETR-r18 端到端检测优势的同时, 通过 RHDWT 下采样保真、CARAFE 上采样对齐和 MDFRM 特征精炼的协同设计, 显著增强了模型在足球比赛视频中对运动模糊、微小目标、密集遮挡与复杂背景的适应能力. 后续将分别在第 2.2-2.4 节对 3 个关键模块的结构细节与实现过程进行详细说明.

2.2 基于残差小波变换的抗模糊下采样

在足球比赛视频的高动态场景下, 球体与运动员的高速位移常引发剧烈的运动模糊, 导致目标边缘的高频梯度信息严重衰减. RT-DETR 的混合编码器采用步长卷积下采样, 虽然高效且具备可学习性, 但在足球转播的高速运动与强背景纹理条件下容易产生采样混叠: 运动模糊使小球与肢体边缘的有效高频梯度进一步变弱, 而步长抽取会将残余细节与草地纹理等背景高频信息共同压缩并混合, 导致微小目标的可分辨线索在早期阶段被不可逆地削弱. 针对该瓶颈, 引入动态滤波器或可学习小波可增强自适应性, 以适配不同模糊形态, 但这类算法相应地会增加参数规模与计算复杂度, 并对数据分布与训练设置更敏感.

相比之下, 哈尔离散小波变换 (Haar discrete wavelet transform, HDWT) 作为一种无参数的固定变换, 具备优异的局部化分析能力与正交重构特性. 它能以极低的算力成本, 通过无损的可逆变换将图像分解为不同频率的子带, 在降低空间分辨率的同时完整保留图像的频谱信息, 有效解决高频边缘 (如模糊球体轮廓) 在下采样中的丢失问题. 本文引入残差哈尔离散小波变换模块 (RHDWT)^[23], 以频带分解和可学习残差补偿的双路并行结构重构下采样层, 使其更适配运动模糊与小目标并存的足球场景.

如图 2 所示, 频域主路径使用固定哈尔滤波器组将特征显式分解为低频与 3 个方向性高频子带, 并通过可学习卷积进行子带融合; 并行残差路径保留步长为 2 的 3×3 可学习卷积以补充数据驱动的空间语义结构. 这种设计使 RHDWT 在机制上同时具备频域先验

的抗混叠特性与卷积下采样的语义适配能力. 前者更有利于在运动模糊下保留微小足球及肢体边缘的残余

高频线索, 后者则有助于维持对球员和裁判等中大尺度目标的语义连续性, 并提升训练过程的稳定性.

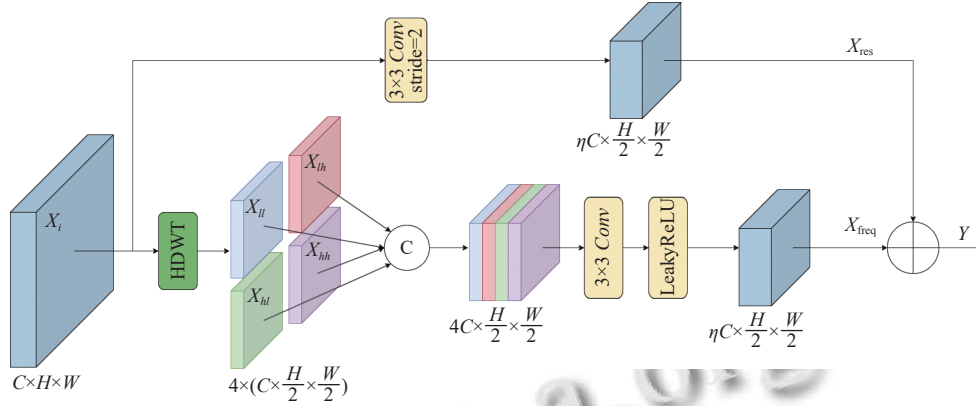


图2 RHDWT 模块结构图

频域变换主路径利用 HDWT 的频域先验处理输入特征. 给定输入特征张量 $X \in R^{C \times H \times W}$, 该路径通过一组预定义的哈尔小波滤波器组与输入特征进行步长为 2 的卷积运算, 将图像分解为 4 个互不重叠的频率子带:

$$X_a = \lambda_a \otimes_{s=2} X, a \in \{ll, lh, hl, hh\} \quad (1)$$

其中, $\lambda_{ll}, \lambda_{lh}, \lambda_{hl}, \lambda_{hh} \in R^{2 \times 2}$ 分别对应低频近似、水平细节、垂直细节与对角线细节算子, 其核权重定义如下:

$$\begin{cases} \lambda_{ll} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, & \lambda_{lh} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix} \\ \lambda_{hl} = \frac{1}{2} \begin{bmatrix} 1 & 1 \\ -1 & -1 \end{bmatrix}, & \lambda_{hh} = \frac{1}{2} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \end{cases} \quad (2)$$

X_{ll} 作为低频分量, 保留了图像的主体结构与光照分布; X_{lh} 、 X_{hl} 、 X_{hh} 这 3 个高频分量则分别捕捉水平、垂直和对角方向的纹理突变与边缘梯度. 在足球检测场景中, 即便球体轮廓因运动模糊而退化, 其残留的高频信号仍能被上述细节分量精准捕获. 随后, 4 个子带在通道维度拼接, 并通过 1 个带激活函数的 3×3 卷积层进行特征融合与通道映射, 生成频域特征 X_{freq} :

$$X_{freq} = \sigma(\text{Conv}_{3 \times 3}(\text{Concat}(X_{ll}, X_{lh}, X_{hl}, X_{hh}))) \quad (3)$$

其中, $\text{Concat}(\cdot)$ 表示通道拼接操作, $\text{Conv}_{3 \times 3}(\cdot)$ 表示步长为 1 的 3×3 卷积, $\sigma(\cdot)$ 为 LeakyReLU 激活函数.

为维持特征图的空间语义连续性并保证梯度传播的稳定性, 模块保留了一条基于步长卷积的残差分支. 该分支利用步长为 2 的 3×3 卷积直接对输入 X 进行空间下采样, 提取常规的空间结构特征 X_{res} :

$$X_{res} = \text{Conv}_{3 \times 3}^{s=2}(X) \quad (4)$$

此路径能够自适应学习空间语义信息, 补偿频域变换可能遗漏的低频结构信息. 最终, RHDWT 模块将频域分支的细节特征与空域分支的结构特征通过逐元素相加进行融合, 输出下采样特征图 Y :

$$Y = X_{freq} + X_{res} \quad (5)$$

这种双分支设计结合了离散小波变换的方向选择性先验与卷积网络的空间感知能力. 不仅有效提取了模糊场景下的高频边缘信号, 更通过残差连接增强了模型的表征鲁棒性, 确保特征信息在下采样过程中的完整传递.

2.3 基于内容感知重组的语义对齐上采样

在 RT-DETR 的混合编码器架构中, 跨尺度特征融合模块 (CCFM) 负责将深层的高级语义与浅层的空间细节进行双向融合. 在自顶向下的路径中, 为了最大限度地降低计算负载以满足实时检测的需求, RT-DETR 原生采用了最近邻插值作为上采样算子. 这种静态采样策略虽然计算高效, 但其仅基于像素间的几何欧氏距离进行离散采样, 局限于像素邻域, 感受野固定, 难以聚合周围的上下文信息. 在足球比赛的多尺度目标检测场景中, 球体、裁判与远景球员在图像中通常占据极小的像素比例, 且背景多为纹理单一的草地. 最近邻插值在进行空间放大时, 无法感知目标的语义边界, 极易将周围的背景噪声直接复制到目标区域, 在层级传递中导致特征图发生模糊或语义错位, 进而影响后续检测头对目标中心点的精确定位.

为了在保持算子轻量化优势的同时解决上述难题, 本文引入 CARAFE^[24] 替代 CCFM 中的上采样层. 不同

于增加繁重的卷积层或复杂的注意力机制, CARAFE 采用轻量级的双分支架构, 包含核预测模块与内容感知重组模块, 能够在保持较大的感受野的同时, 根据输

入特征的语义内容动态生成实例特定的重组核, 显著提升性能的同时仅引入极小的计算开销. 其具体架构如图 3 所示.

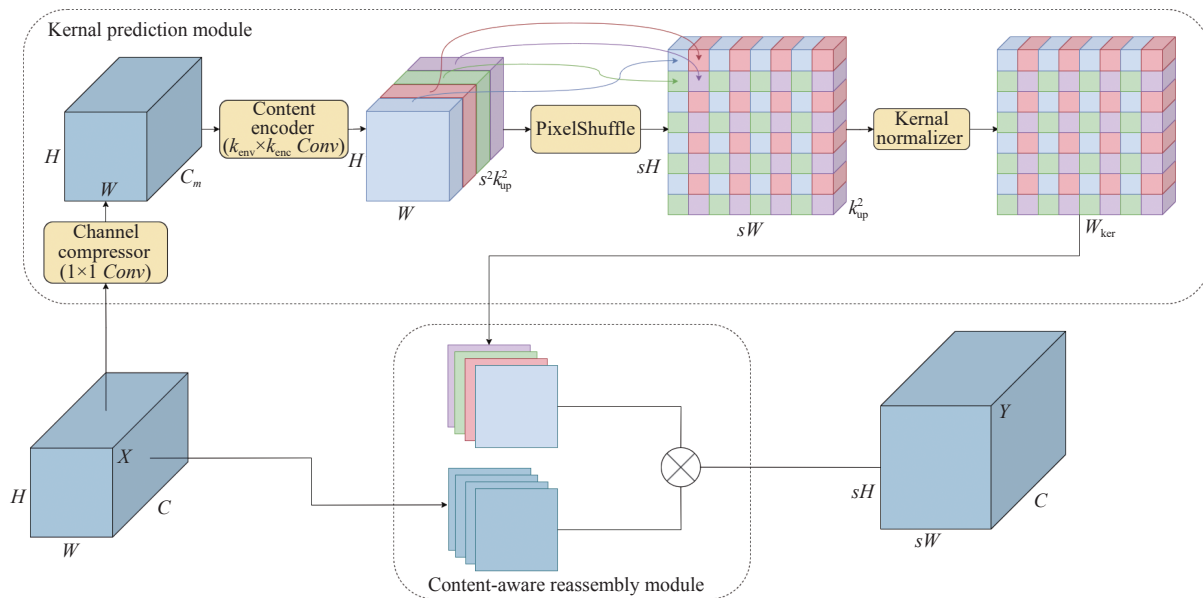


图3 CARAFE 模块结构图

核预测模块负责生成具有实例特异性的上采样核. 给定输入特征张量 $X \in \mathbb{R}^{C \times H \times W}$ 与上采样倍率 s , 模块首先利用 1×1 卷积作为通道压缩器, 将输入通道 C 降维至 C_m , 以在减少计算冗余的同时聚焦关键语义信息. 随后, 压缩后的特征图输入内容编码器, 通过卷积核大小为 k_{enc} 的卷积层感知局部上下文, 预测生成未归一化的重组核权重. 为适应上采样后的空间尺寸, 输出通道被扩展为 $s^2 k_{up}^2$, 其中 k_{up} 为重组核的尺寸. 最后, 通过核归一化器, 利用 PixelShuffle 操作将通道维度展开至空间维度, 并通过 Softmax 函数进行空间归一化, 得到最终的重组核 W_{ker} :

$$W_{ker} = \text{Softmax}(\psi(N(X, k_{enc}))) \quad (6)$$

其中, ψ 表示内容编码与通道展开操作, $N(X, k_{enc})$ 表示以输入特征为中心的局部感受野区域. Softmax 归一化强制核权重之和为 1, 确保了特征重组过程中局部数值分布的稳定性, 避免了数值漂移.

内容感知重组模块则利用生成的自适应核将特征映射至高分辨率空间. 对于输出特征图 Y 上的每一个目标位置 l' , 模块将其映射回输入特征图的源位置 l , 并提取以 l 为中心、大小为 $k_{up} \times k_{up}$ 的局部邻域区域. 重组过程通过计算该邻域特征与对应位置重组核的点积完成:

$$Y_{l'} = \sum_{n=-r}^r \sum_{m=-r}^r W_r(n, m) \cdot X_{(l_x+n, l_y+m)} \quad (7)$$

其中, $r = \lfloor k_{up}/2 \rfloor$, (n, m) 为邻域内的相对坐标偏移. 这种设计使得上采样过程不再局限于固定的几何位置, 而是能够根据目标的形状和纹理进行非对称加权. 在足球检测任务中, 即便目标因运动模糊而轮廓不清, CARAFE 也能利用较大的感受野聚合周围上下文信息, 自适应地锐化目标边缘并抑制背景噪声, 从而显著提升小目标在深层特征中的语义表征能力.

2.4 基于混合动态注意力的多尺度特征精炼

尽管前端网络已经通过 RHDWT 与 CARAFE 改善了特征的频域保真度与空间对齐性, 但输出的 P3-P5 特征图作为 Transformer 解码器中交叉注意力机制的键与值, 其语义质量直接决定了对象查询能否精准检索到目标. 特征图中残留的复杂背景噪声极易与前景目标混淆, 且密集遮挡导致的语义缺失会严重影响解码器中查询匹配的效果.

为了强化特征在进入解码器前的语义显著性与上下文完整性, 本文在混合编码器末端设计了多尺度动态特征精炼模块 MDFRM. 该模块采用金字塔结构提取多尺度特征, 以适应不同的空间感受野. 在特征交互

的核心环节,受 ConvAtt^[25]利用卷积算子模拟自注意力的启发,提出一种输入自适应的混合动态注意力机制 DynAtt,融合了依赖输入的动态卷积与层级独立的静态大核卷积,在高效捕捉长距离依赖的同时自适应地突出关键特征并抑制冗余信息.其具体结构如图4所示.

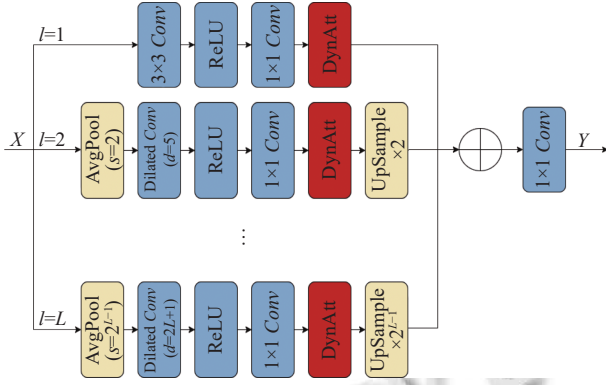


图4 MDFRM 结构图

为了同时兼顾微小目标与大尺度目标, MDFRM 首先构建 L 层并行的感知分支以解耦不同感受野下的特征表达. 假设输入特征为 $X \in R^{C \times H \times W}$, 对于第 l 层分支 ($l \in \{1, 2, \dots, L\}$), 特征图 X_l 的生成过程定义为:

$$X_l = \text{Conv}_{1 \times 1}(\sigma(\text{Conv}_{3 \times 3}^{d_l}(\mathcal{P}_{\text{avg}}^l(X)))) \quad (8)$$

其中, $\mathcal{P}_{\text{avg}}^l(\cdot)$ 表示步长为 2^{l-1} 的平均池化操作,旨在将特征图投影至不同空间尺度; $\text{Conv}_{3 \times 3}^{d_l}$ 为膨胀率 $d_l = 2l + 1$ 的空洞卷积, σ 为 ReLU 激活函数. 除第1层采用标准卷积保留微小细节外, 其余各层通过指数级扩大的膨胀率, 扩大有效感受野捕捉广域上下文信息. 这种设计使得模型能够在单一特征层级内部, 同时获得局部纹理细节与全局语义层次的特征, 增强了对大结构与微小细节的双重感知能力.

各分支特征随后进入混合动态注意力单元进行自适应精炼, 结构如图5所示. 为了在不显著增加计算成本的同时复现自注意力的表征优势, DynAtt 将输入特征在通道维度拆分为注意力部分 X_{att} 与恒等映射部分 X_{id} , 并通过动态小核与静态大核两条并行路径分别模拟自注意力的关键特性.

● 动态小核路径: 引入输入自适应的动态卷积核生成机制, 以模拟自注意力根据输入内容动态分配权重的特性. 首先通过全局平均池化聚合空间信息, 随后利用多层感知机映射生成动态权重 W_{dyn} :

$$W_{\text{dyn}} = \psi(\text{Conv}_{\text{up}}(\delta(\text{Conv}_{\text{down}}(\text{GAP}(X_{\text{att}})))) \quad (9)$$

其中, ψ 为 Reshape 操作, δ 为 GELU 激活函数. 生成的动态权重 W_{dyn} 与特征图进行卷积操作, 使得模型能够根据每个样本的模糊程度自适应地锐化边缘特征:

$$Y_{\text{dyn}} = \text{Conv}(X_{\text{att}}, W_{\text{dyn}}) \quad (10)$$

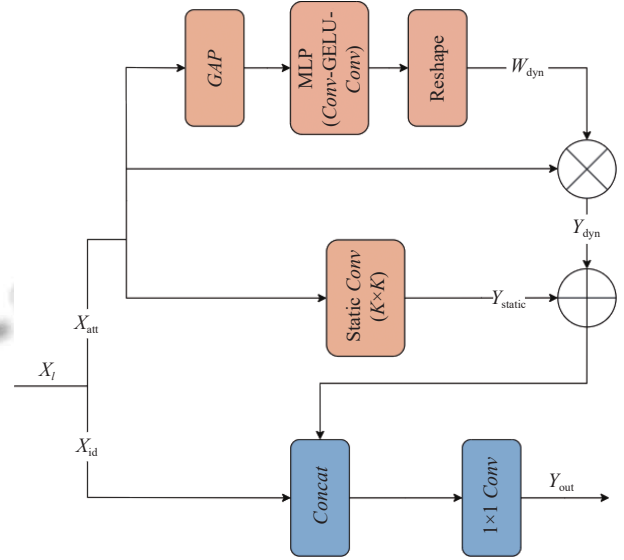


图5 DynAtt 结构图

● 静态大核路径: 针对球员密集遮挡导致的语义缺失, 需要依赖全局上下文信息进行推理. 该路径采用超大尺寸卷积核, 建立类似于自注意力的长距离空间依赖, 使模型能利用周围环境上下文辅助推断遮挡区域的潜在语义:

$$Y_{\text{static}} = \text{Conv}(X_{\text{att}}, K_{\text{static}}) \quad (11)$$

在完成特征精炼后, MDFRM 对各尺度信息进行统一融合. 两条路径的输出通过逐元素相加进行融合, 并与恒等映射部分 X_{id} 拼接, 经由 1×1 卷积层进行通道融合:

$$Y_{\text{out}} = \text{Conv}_{1 \times 1}(\text{Concat}(Y_{\text{dyn}} + Y_{\text{static}}, X_{\text{id}})) \quad (12)$$

随后, 各尺度分支经双线性插值上采样对齐至统一尺寸, 并通过残差连接输出最终的精炼特征. 这一过程确保了全尺度信息的有效集成, 得到了语义更强、噪声更低的增强输出.

考虑到 RT-DETR 输出的 P3-P5 特征层在空间分辨率与语义抽象程度上存在显著差异, 针对不同特征层的任务属性, 本文精细调整了金字塔层数 L 与静态大核尺寸 K_{static} . P3 层拥有最高的分辨率, 主要负责捕捉微小的足球目标, 为防止过度下采样导致高频细节丢失, 本文将感知金字塔的深度限制为 2 层, 并采用相对

较小的7×7静态卷积核. P4与P5层主要负责表征球员、裁判等大尺度语义, 针对此类目标常面临的肢体遮挡与分类混淆问题, 本文将金字塔层数扩展至3层, 并将静态卷积核尺寸分别增大至9×9与11×11. 通过逐层递增的核尺寸与更深的金字塔结构, MDFRM能够在抑制背景噪声的同时, 最大化不同尺度目标的语义显著性, 为后续检测头的分类与回归任务提供高质量的特征表达.

3 实验分析

3.1 数据集介绍

本文实验选用 SoccerNet-Tracking^[26]数据集作为实验基准. 该数据集源自大规模足球视频数据库 SoccerNet, 包含2019年瑞士超级联赛的12场完整比赛录像, 专注于足球比赛场景下的目标检测与跟踪任务. 数据集总计涵盖106个视频序列, 每个序列包含750帧连续的高清比赛画面, 总计79500张图像. 为了保证实验结果的公正性, 本文严格按照视频序列为单位进行数据集划分, 避免时序相关性导致的数据泄露. 将106个视频序列随机打乱后, 按照8:1:1的比例划分为训练集(84个序列)、验证集(11个序列)与测试集(11个序列). 数据集中的每个目标都经过精细的人工标注, 包含准确的边界框坐标和唯一的跟踪ID. 本文对原始标注信息进行了清洗与类别整合, 将目标统一映射为足球(Ball)、球员(Player)与裁判(Referee)这3大类别.

3.2 实验环境和参数配置

本实验的参数配置为: Python 3.10、PyTorch 2.9.1、CUDA 13.0, 训练轮数(epochs)为200, 优化器为AdamW, 预热(warmup)轮数为20, 学习率为0.0002, 批次大小为32, 输入图像大小为640×640. 本实验中使用的硬件配置如表1所示.

表1 实验使用硬件配置

类型	参数
系统	Ubuntu 22.04.5 LTS
CPU	Intel(R) Xeon(R) Platinum 8470Q
GPU	NVIDIA GeForce RTX 5090 32 GB
内存	90 GB

3.3 实验结果

为全面衡量模型在足球比赛场景下的检测性能, 实验选取精确率(Precision)、召回率(Recall)以及平均精度均值(mean average precision, mAP)作为核心评价指标. Precision反映了模型预测的正样本中真正正

样本的比例, Recall则衡量了模型对真正正样本的检出能力. 二者的计算公式如式(13)和式(14):

$$Precision = \frac{TP}{TP + FP} \tag{13}$$

$$Recall = \frac{TP}{TP + FN} \tag{14}$$

其中, TP(true positive)为正确检测的目标数量, FP(false positive)为误检的背景数量, FN(false negative)为漏检的目标数量.

mAP是综合评估不同类别检测精度的关键指标, 通过计算所有类别平均精度(average precision, AP)的均值得到. 计算公式为:

$$AP = \int_0^1 P(R) dR \tag{15}$$

$$mAP = \frac{1}{N_C} \sum_{i=1}^{N_C} AP_i \tag{16}$$

其中, N_C 表示目标类别总数(本文为3类: 足球(Ball)、球员(Player)、裁判(Referee)), AP_i 为第*i*类目标的平均精度. 实验重点关注 mAP@50 (IoU阈值为0.5时的mAP)和 mAP@50:95 (IoU阈值为0.5–0.95、步长为0.05时的平均值), 前者反映模型的检出能力, 后者更能体现高精度的定位性能. 此外, 引入参数量和计算量(GFLOPs)来评估模型的空间复杂度和计算复杂度, 以验证其在实际部署中的可行性.

3.3.1 对比实验

为了验证改进算法的优越性, 本文在 SoccerNet-Tracking数据集上与当前主流的单阶段目标检测算法进行了横向对比. 对比模型涵盖了YOLOv9、YOLOv10、YOLO11、YOLO12以及基准模型RT-DETR的不同规格版本. 所有模型均统一输入图像尺寸为640×640像素, 训练轮数设定为200轮. 对于学习率等其余超参数, YOLO系列均采用 Ultralytics 官方的默认配置, RT-DETR系列均采用与本文算法相同的配置. 对比实验结果如表2所示.

分析表2数据可知, RT-DETR架构在足球场景下的表现普遍优于YOLO系列算法. YOLO系列虽然在参数量控制上较为出色, 但在处理远景小目标与密集遮挡时显露出结构性短板. 其mAP@50均处于78%以下, Recall指标普遍停留在73.3%–74.5%, 显著低于RT-DETR系列近10个百分点. 这意味着在高速运动和复杂遮挡环境下, YOLO系列算法存在严重的漏检

现象,难以捕捉到所有关键目标.这主要归因于 CNN 架构固有的局部感受野限制,在缺乏全局上下文信息

支撑的情况下,难以对像素占比极低的足球或严重遮挡的球员进行有效推断.

表2 对比实验结果

模型	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50:95 (%)	参数量 (M)	计算量 (GFLOPs)
YOLOv9c	82.7	74.5	76.7	47.9	25.53	103.7
YOLOv10b	82.4	73.3	76.6	48.7	20.46	98.7
YOLO11m	83.2	74.4	77.0	48.5	20.06	68.2
YOLO11l	84.0	74.4	77.7	49.2	25.31	87.3
YOLO12m	85.4	73.9	77.1	48.7	20.14	67.7
YOLO12l	82.7	74.5	77.0	48.2	26.39	89.4
RT-DETR-r18	89.1	83.6	83.1	48.8	20.09	58.3
RT-DETR-r34	89.1	85.1	84.4	49.5	31.32	90.2
本文算法	89.9	85.4	84.9	50.0	30.86	84.4

注:加粗数据表示最优结果

相比之下,引入全局注意力机制的 RT-DETR 架构展现出更强的场景适应性.基准模型 RT-DETR-r18 凭借其全局建模能力,将 Recall 提升至 83.6%,mAP@50 达到 83.1%,确立了较高的性能基线.在此基础上,改进算法在各项精度指标上均取得了最优结果. Precision 和 Recall 分别达到 89.9% 和 85.4%,相较于基准模型分别提升了 0.8 和 1.8 个百分点,证明了改进策略在抑制背景误检和降低漏检率方面的有效性. mAP@0.50 指标达到 84.9%,不仅超越了 RT-DETR-r18,甚至优于

参数量更大的 RT-DETR-r34. 这一结果表明,本文算法通过针对性的模块设计,显著提升了模型对足球比赛场景的特征提取与融合能力,实现了检测精度与推理效率的更优平衡.

3.3.2 消融实验

为了探究 RHDWT、CARAFE 和 MDFRM 这 3 个改进对模型性能的具体贡献,以 RT-DETR-r18 为基准进行了逐项消融实验.实验重点关注了模型对不同类别目标的检测能力变化.消融实验结果如表 3 所示.

表3 消融实验结果

序号	RHDWT	CARAFE	MDFRM	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50:95 (%)	AP@50 (%)			参数量 (M)	计算量 (GFLOPs)
								Ball	Player	Referee		
1	×	×	×	89.1	83.6	83.1	48.8	55.5	97.6	96.2	20.09	58.3
2	√	×	×	89.7	84.0	83.7	49.2	57.7	97.4	96.0	24.80	67.7
3	×	√	×	89.5	84.1	83.6	49.2	57.0	97.5	96.4	20.23	58.6
4	×	×	√	89.9	84.8	84.1	49.7	57.9	97.8	96.4	25.99	74.7
5	√	√	×	88.9	84.9	83.8	48.7	59.2	97.4	94.9	24.94	68.0
6	√	×	√	89.1	84.9	84.1	49.6	58.8	97.9	95.5	30.70	84.1
7	×	√	√	89.2	85.3	84.3	49.7	58.9	97.5	96.4	26.13	75.0
8	√	√	√	89.9	85.4	84.9	50.0	60.4	97.7	96.6	30.84	84.4

注:加粗数据表示最优结果

由表 3 可知,基准模型在球员和裁判两类大尺度目标上表现稳健,AP@50 分别达到 97.6% 和 96.2%,但对足球类别的检测能力严重不足,AP@50 仅为 55.5%,显著低于其他类别.这证实了在缺乏针对性设计的情况下,通用模型难以应对足球赛场中微小目标的高速模糊与背景干扰,成为制约整体性能的瓶颈.

引入 RHDWT 模块后,足球类别的 AP@50 显著提升了 2.2%,带动整体 mAP@50 提升 0.6%.这一结果验证了第 2.2 节的理论分析:在处理高速运动的足球时,常规卷积下采样容易丢失高频边缘信息,而 RHDWT 利用哈尔小波的正交重构特性,成功在频域保留了球

体的模糊轮廓与边缘梯度,缓解下采样过程中的特征混叠问题.

单独加载 CARAFE 模块使整体 mAP@50 提升至 83.6%.足球类别的 AP@50 提升至 57.0%,而参数量和计算量仅分别增加了 0.14M 和 0.3 GFLOPs.这说明该模块在几乎不增加计算负担的情况下,有效弥补了原生上采样的不足,有力地支持了第 2.3 节的改进动机. CARAFE 通过动态核预测,有效改善了小目标在特征融合过程中的语义对齐,减少了背景噪声对目标边缘的侵蚀.

MDFRM 模块带来了最大的单模块性能增益,将

整体 $mAP@50$ 提升至 84.1%, $Recall$ 提升了 1.2%, 同时足球 $AP@50$ 提升了 2.4%, 球员和裁判的检测精度也得到了单模块最大的增幅. 这说明该模块位于解码器前的特征精炼发挥了全局性作用, 有效提升了对象查询在复杂场景下的判别力, 从而全面提升了各类目标的检测精度.

多模块组合实验显示, 各模块之间存在着良好的协同效应. RHDWT 与 CARAFE 联合使用使足球 $AP@50$ 跃升至 59.2%, 显著高于二者单独使用. 说明高保真的下采样与精准的上采样对齐相结合, 能最大程度保留小目标的结构信息. 而将 MDFRM 分别与 RHDWT 或 CARAFE 组合时, 模型性能均优于单模块, 证明了特征精炼模块与其他模块的兼容性. 当 3 个模块协同工作时, 模型在各项指标上均达到最优, $mAP@50$ 攀升至 84.9%, $mAP@50:95$ 突破 50.0%. 足球 $AP@50$ 最终达到 60.4%, 较基准模型大幅提升了 4.9 个百分点. 这一结果证实了 RHDWT 的频域保真、CARAFE 的空间对齐与 MDFRM 的语义精炼形成了良性互补: RHDWT 确保了微小目标的特征不丢失, CARAFE 保证了特征位置的准确性, 而 MDFRM 则在解码前滤除了背景干扰, 强化了语义特征的显著性, 共同解决了复杂足球场景下的检测难题.

3.3.3 泛化性实验

为了验证本文算法在不同体育场景下的跨领域泛化能力, 本文构建了一个篮球比赛目标检测数据集进行验证. 数据来源于 Roboflow 开源社区, 包含多场 NBA 篮球比赛, 部分数据集图像如图 6 所示.

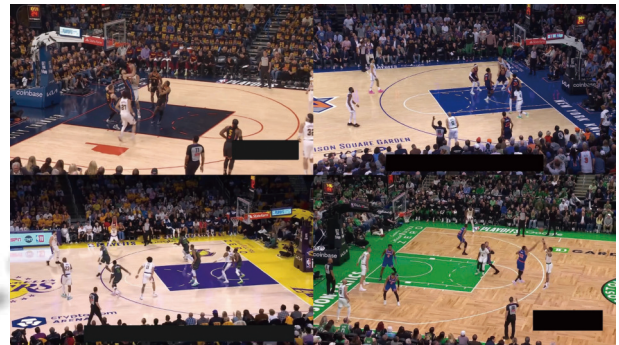


图 6 部分数据集图像

经清洗与整理, 共得到 2 052 张篮球比赛图像, 划分为训练集 1 507 张, 验证集 282 张, 测试集 263 张, 并保证测试集的比赛场次图像没有混入训练集和验证集中. 数据集标注类别与 SoccerNet-Tracking 保持一致, 包含篮球 (Ball)、球员 (Player) 和裁判 (Referee) 这 3 类. 实验采用与第 3.2 节一致的硬件环境与训练超参数配置, 实验结果如表 4 所示.

表 4 泛化性实验结果 (%)

模型	<i>Precision</i>	<i>Recall</i>	$mAP@50$	$mAP@50:95$	$AP@50$		
					Ball	Player	Referee
RT-DETR-r18	94.6	89.2	93.4	68.0	85.7	98.1	96.4
RT-DETR-r34	94.8	89.2	93.6	68.1	85.2	98.3	97.4
本文算法	97.5	91.2	95.7	72.4	90.2	98.6	98.3

注: 加粗数据表示最优结果

实验结果显示, 本文算法在篮球场景下依然保持了最优性能. *Precision* 和 *Recall* 相较于基准模型 RT-DETR-r18 分别提升 2.9 和 2.0 个百分点, $mAP@50$ 达到 95.7%, 且超越了参数量更大的 RT-DETR-r34 (93.6%). 本文算法在 3 个类别的 $AP@50$ 上均为最优, 尤其是最困难的篮球目标, $AP@50$ 高达 90.2%, 相比基准模型大幅提升了 4.5%.

篮球与足球同属体育转播场景, 虽然两者均面临球类小目标、多人遮挡以及镜头运动模糊的核心挑战, 但在视觉特征上存在显著差异. 篮球场地为木质地板且伴有反光, 其背景纹理、色彩分布与足球场景的室外草地截然不同; 篮球比赛多为半场阵地战, 存在密集的半场跑位与腾空起跳, 带来了更复杂的遮挡问题;

篮球转播镜头更多在中近景切换, 虽然目标比足球场景更大, 但镜头的平移与切换更为剧烈.

尽管视觉挑战发生改变, 本文算法依然取得了显著的性能提升. 这得益于本文各模块在机制设计上的通用性: RHDWT 提取的频域高频信号不依赖特定的背景色彩, 故在木地板反光下仍能精准捕捉模糊的篮球轮廓; CARAFE 的内容感知能力与 MDFRM 的动态特征精炼则自适应地过滤了室内光影与场地线的干扰, 为严重遮挡的篮球运动员提供了准确的上下文推理.

实验结果证明, 本文提出的改进策略并非仅针对足球数据的过拟合设计, 能够有效泛化至具有显著差异的同类复杂体育竞技场景, 具备极强的鲁棒性与跨体育场景应用潜力.

3.4 实验结果对比

为直观验证本文改进算法在复杂足球比赛场景下的检测性能,选取了4种典型的困难场景进行可视化对比分析。

图7展示了基准模型 RT-DETR-r18 与本文算法在禁区密集重叠、远景镜头、高速运动模糊以及严重遮挡的极端条件下的检测结果对比。其中红色检测框为足球,蓝色与黄色检测框分别为运动员与裁判。图7中第1列为原始图像,第2列为 RT-DETR-r18 检测结果,第3列为本文算法检测结果。



图7 实验结果对比

在图7(a)所示的禁区混战场景中,6名球员在狭小空间内发生严重肢体重叠,且有1名球员倒地蜷缩,足球被人群和球员腿部遮挡。基准模型在处理此类高密度遮挡时表现出明显的局限性,产生了大量重叠的冗余预测框,未能有效区分个体,且直接漏检了关键的足球目标。本文算法成功抑制了拥挤场景下的特征混淆,精准定位了所有6名球员及被遮挡的足球,展现了在密集遮挡下的鲁棒性。

图7(b)展示了远景镜头下的极限小目标检测挑战。足球在画面中像素占比极低,且位于白色球衣守门员脚下,目标特征与背景极易混淆。RT-DETR-r18未能识别出足球,反而将两名运动员的白色球鞋误检为足球,显示出其对微小目标特征提取能力的不足。本文算法成功识别出极微小的足球目标,且无误检产生。

在图7(c)的高速运动场景中,整体画面受运动模糊影响,图像清晰度大幅下降。基准模型受模糊噪声干

扰,错误地将观众席中的白色模糊斑块识别为足球,导致误检。本文算法检测出了与运动员身体重叠且轮廓模糊的正确足球目标。

图7(d)呈现了足球被多名运动员包围,且被其中一人头部严重遮挡的极端情况,仅露出极小部分的侧面轮廓。RT-DETR-r18彻底丢失了足球目标,并将身穿白色球衣运动员的肩部误判为足球。本文算法充分捕捉了目标的局部碎片化特征与全局上下文关联,在足球仅存微弱视觉线索的情况下,依然实现了精准定位,体现了模型对严重遮挡目标的强大推理与感知能力。

上述典型复杂场景的对比实验结果表明,本文改进算法在处理足球比赛中的密集遮挡、微小目标、运动模糊及严重遮挡等复杂场景时,均表现出优于基准模型的检测能力与抗干扰性能,有效解决了漏检与误检问题,验证了算法的实际应用价值。

4 结论与展望

本文针对足球比赛视频目标检测中面临的微小目标、运动模糊、密集遮挡及背景噪声干扰等挑战,在 RT-DETR-r18 架构基础上做出了一系列改进。通过引入残差哈尔离散小波变换重构下采样路径,利用频域先验有效缓解了模糊高速运动目标的特征退化;采用内容感知特征重组模块优化上采样过程,显著提升了多尺度特征融合时的语义对齐精度;设计多尺度动态特征精炼模块,在解码前实现了特征去噪与全局上下文增强。实验结果表明,该算法在 SoccerNet-Tracking 数据集上展现出优异的性能,各类别目标的检测精度均优于主流模型,在复杂动态场景下具有较强的鲁棒性。未来工作将致力于探索基于当前模型的衍生工作,例如基于检测模型的多目标跟踪、比赛事件分析等实用场景,扩展工作的应用价值;另外当前模型依赖计算平台进行推理,在移动端或边缘设备部署困难,未来可通过模型剪枝或量化技术降低计算资源消耗,提升模型的可部署性。

参考文献

- 1 Zhao ZH, Chai WH, Hao SY, *et al.* A survey of deep learning in sports applications: Perception, comprehension, and decision. *IEEE Transactions on Visualization and Computer Graphics*, 2025, 31(10): 9368–9386. [doi: 10.1109/TVCG.2025.3554801]
- 2 Akan S, Varlı S. Use of deep learning in soccer videos analysis: Survey. *Multimedia Systems*, 2023, 29(3): 897–915. [doi: 10.1007/s00530-022-01027-0]

- 3 Deliège A, Cioppa A, Giancola S, *et al.* SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Nashville: IEEE, 2021. 4503–4514. [doi: [10.1109/CVPRW53098.2021.00508](https://doi.org/10.1109/CVPRW53098.2021.00508)]
- 4 Markappa PSS, O’Leary C, Lynch C. A review of YOLO models for soccer-based object detection. Proceedings of the 6th International Conference on Intelligent Computing in Data Sciences. Marrakech: IEEE, 2024. 1–7. [doi: [10.1109/ICDS62089.2024.10756443](https://doi.org/10.1109/ICDS62089.2024.10756443)]
- 5 Dalal N, Triggs B. Histograms of oriented gradients for human detection. Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. San Diego: IEEE, 2005. 886–893. [doi: [10.1109/CVPR.2005.177](https://doi.org/10.1109/CVPR.2005.177)]
- 6 Girshick R, Donahue J, Darrell T, *et al.* Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition. Columbus: IEEE, 2014. 580–587. [doi: [10.1109/CVPR.2014.81](https://doi.org/10.1109/CVPR.2014.81)]
- 7 Ren SQ, He KM, Girshick R, *et al.* Faster R-CNN: Towards real-time object detection with region proposal networks. Proceedings of the 29th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 91–99.
- 8 Redmon J, Divvala S, Girshick R, *et al.* You only look once: Unified, real-time object detection. Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE, 2016. 779–788. [doi: [10.1109/CVPR.2016.91](https://doi.org/10.1109/CVPR.2016.91)]
- 9 Komorowski J, Kurzejamski G, Sarwas G. FootAndBall: Integrated player and ball detector. Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications. Valletta: VISAPP, 2020. 47–56. [doi: [10.5220/0008916000470056](https://doi.org/10.5220/0008916000470056)]
- 10 Moutselos K, Maglogiannis I. Setting a baseline for long-shot real-time player and ball detection in soccer videos. Proceedings of the 14th International Conference on Information, Intelligence, Systems & Applications. Volos: IEEE, 2023. 1–7. [doi: [10.1109/IISA59645.2023.10345947](https://doi.org/10.1109/IISA59645.2023.10345947)]
- 11 于兆廷, 杨建柏. 利用 ECABiF-Y5n 提高足球比赛中的目标检测精度. 计算机科学与应用, 2025, 15(2): 83–93. [doi: [10.12677/csa.2025.152035](https://doi.org/10.12677/csa.2025.152035)]
- 12 Wang A, Chen H, Liu LH, *et al.* YOLOv10: Real-time end-to-end object detection. Proceedings of the 38th International Conference on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2024. 3429.
- 13 Khanam R, Hussain M. YOLOv11: An overview of the key architectural enhancements. arXiv:2410.17725, 2024.
- 14 Tian YJ, Ye QX, Doermann D. YOLOv12: Attention-centric real-time object detectors. arXiv:2502.12524, 2025.
- 15 Carion N, Massa F, Synnaeve G, *et al.* End-to-end object detection with Transformers. Proceedings of the 16th European Conference on Computer Vision. Glasgow: Springer, 2020. 213–229. [doi: [10.1007/978-3-030-58452-8_13](https://doi.org/10.1007/978-3-030-58452-8_13)]
- 16 Zhu XZ, Su WJ, Lu LW, *et al.* Deformable DETR: Deformable Transformers for end-to-end object detection. Proceedings of the 9th International Conference on Learning Representations. Vienna: OpenReview.net, 2021.
- 17 Zhao Y, Lv WY, Xu SL, *et al.* DETRs beat YOLOs on real-time object detection. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Seattle: IEEE, 2024. 16965–16974. [doi: [10.1109/CVPR52733.2024.01605](https://doi.org/10.1109/CVPR52733.2024.01605)]
- 18 Sun LZ, Lin XP, Li ZH. BSTR: A Transformer-based framework for robust and efficient basketball object detection. Proceedings of the 2025 International Conference on Advances in Computer Vision Research and Applications. Nanjing: SPIE, 2025. 1369012. [doi: [10.1117/12.3068152](https://doi.org/10.1117/12.3068152)]
- 19 Zhang WT, Yang YC. FoT: An efficient Transformer framework for real-time small object detection in football videos. Scientific Reports, 2025, 15(1): 30875. [doi: [10.1038/s41598-025-16795-8](https://doi.org/10.1038/s41598-025-16795-8)]
- 20 Madan M, Reich C. Strengthening small object detection in adapted RT-DETR through robust enhancements. Electronics, 2025, 14(19): 3830. [doi: [10.3390/electronics14193830](https://doi.org/10.3390/electronics14193830)]
- 21 Yeung C, Ide K, Fujii K. AutoSoccerPose: Automated 3D posture analysis of soccer shot movements. Proceedings of the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. Seattle: IEEE, 2024. 3214–3224. [doi: [10.1109/CVPRW63382.2024.00327](https://doi.org/10.1109/CVPRW63382.2024.00327)]
- 22 李萌萌, 张雪松, 郭莉, 等. 基于上下文感知融合的轻量球类运动检测方法. 信息与控制, <https://link.cnki.net/urlid/21.1138.TP.20251230.1708.001>. (2025-12-31). [doi: [10.13976/j.cnki.xk.2025.1933](https://doi.org/10.13976/j.cnki.xk.2025.1933)]
- 23 Yuan S, Qin HL, Yan X, *et al.* ASCNet: Asymmetric sampling correction network for infrared image destriping. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 5001815. [doi: [10.1109/TGRS.2025.3534838](https://doi.org/10.1109/TGRS.2025.3534838)]
- 24 Wang JQ, Chen K, Xu R, *et al.* CARAFE: Content-aware reassembly of features. Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision. Seoul: IEEE, 2019. 3007–3016. [doi: [10.1109/ICCV.2019.00310](https://doi.org/10.1109/ICCV.2019.00310)]
- 25 Lee D, Yun S, Ro Y. Emulating self-attention with convolution for efficient image super-resolution. Proceedings of the 2025 IEEE/CVF International Conference on Computer Vision. IEEE, 2025. 24467–24477.
- 26 Cioppa A, Giancola S, Deliège A, *et al.* SoccerNet-Tracking: Multiple object tracking dataset and benchmark in soccer videos. Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. New Orleans: IEEE, 2022. 3490–3501. [doi: [10.1109/CVPRW56347.2022.00393](https://doi.org/10.1109/CVPRW56347.2022.00393)]

(校对责编: 李慧鑫)