文章编号：1673-3193（2024）06-0832-11

基于改进 VisionTransformer 模型的团队体育视频多目标跟踪深度学习框架

曹　伟[1]，王晓勇[2]，刘咸祥[1]

（1. 淮南联合大学 公共教育学院，安徽 淮南 232038；2. 淮南联合大学 信息工程学院，安徽 淮南 232038）

摘　要：多目标跟踪（MOT）技术为团队体育视频监测和分析提供了全新的可能性，能够实时跟踪多个运动员并支持对比赛动态的多维度分析与理解。然而，在复杂的团队运动场景下，诸如运动员之间的相互遮挡、快速移动以及目标身份的频繁变换等问题，都可能降低跟踪性能。为此，本文提出了基于 VisionTransformer 的端到端深度学习 MOT 框架，主要包括检测网络和记忆网络两个部分。检测网络由卷积神经网络（CNN）骨干网、VisionTransformer 编码器和解码器组成，采用 ResNet50 作为特征提取器，并引入局部注意力（LA）模块替代传统前馈神经网络（FFN）层。通过全局注意力和局部卷积的结合，得到更全面的特征表示。记忆网络由记忆编码模块和时空记忆解码器组成。记忆编码模块负责聚合目标嵌入信息，其中，短时互注意力（CA）模块关注即时状态，而长时记忆 CA 模块则挖掘了记忆涵盖的时间跨度内的显著特征，捕捉长时间间隔内的依赖关系和关联，从而有效保留了跟踪对象的时间上下文信息。时空记忆解码器在嵌入融合过程中综合考虑了编码帧、候选嵌入和轨迹嵌入信息，解决了 MOT 中的多目标检测和身份关联。时空记忆机制能够有效地保留目标历史状态的观察结果，并结合注意力机制对目标状态进行准确预测。实验结果表明，所提框架在团队体育视频公开数据集 SportsMOT 上实现了 75. 7% 的 HOTA 和 98. 5% 的 MOTA 结果，优于其他先进的 MOT 方法。此外，所提框架在通用公开数据集 MOT17 和 MOT20 上的多个指标取得了最优或次优性能，进一步验证了所提方法的有效性和鲁棒性。

关键词：多目标跟踪；深度学习；团队体育视频；VisionTransformer；时空记忆；注意力机制
中图分类号：TP 391　　**文献标识码**：A　　**doi:** 10. 3969/j. issn. 1673-3193. 2024. 06. 012
引用格式：曹伟，王晓勇，刘咸祥. 基于改进 VisionTransformer 模型的团队体育视频多目标跟踪深度学习框架[J]. 中北大学学报（自然科学版），2024，45（6）：832-842.
CAO　Wei，WANG　Xiaoyong，LIU　Xianxiang. A deep learning framework for multi-object tracking of team sports videos based on the improved VisionTransformer model[J]. Journal of North University of China（Natural Science Edition），2024，45（6）：832-842.

A Deep Learning Framework for Multi-Object Tracking of Team Sports Videos Based on the Improved VisionTransformer Model

CAO Wei[1]，WANG Xiaoyong[2]，LIU Xianxiang[1]

（1. School of Public Education，Huainan Union University，Huainan 232038，China；
2. School of Information Engineering，Huainan Union University，Huainan 232038，China）

Abstract：The application of multi-object tracking（MOT）technology opens up new possibilities for team sports video monitoring and analysis，enabling real-time tracking of multiple athletes and supporting multi-

收稿日期：2024-01-27
基金项目：2021 年度高等学校省级质量工程项目（2021jxtd259）
作者简介：曹　伟（1982—），男，副教授，硕士，主要从事体育人工智能、数据挖掘等研究。E-mail：caowei@hnuu.edu.cn。

dimensional analysis and understanding of game dynamics. However, in complex team sports scenarios, issues such as mutual occlusion between athletes, rapid movements, and frequent changes in target identities may potentially degrade tracking performance. To address these challenges, an end-to-end deep learning MOT framework based on VisionTransformer was proposed, which mainly consisted of two parts: detection network and memory network. The detection network comprised a convolutional neural network (CNN) backbone, Vision Transformer encoder and decoder. The ResNet50 was adopted as a feature extractor, and the traditional feed-forward neural network (FFN) layer was replaced by a local attention (LA) module to obtain more comprehensive feature representations through the combination of global attention and local convolution. The memory network consisted of a memory encoding module and spatio-temporal memory decoder. The memory encoding module was responsible for aggregating the target embedding information, in which the short-term cross attention (CA) module focused on the immediate states, while the long-term CA module explored the significant features covered by memory over time spans, captured dependencies and associations over long time intervals to effectively preserve temporal context information of tracked objects. The spatio-temporal memory decoder integrated encoded frame embeddings, candidate embeddings and trajectory embeddings to address multi-object detection and identity association in MOT. The spatio-temporal memory mechanism efficiently retained observed historical states of targets and combined with an attention mechanism, accurately predicted target states. Experimental results demonstrate that the proposed framework achieves 75.7% HOTA and 98.5% MOTA on the team sports video public dataset SportsMOT, outperforming other state-of-the-art MOT methods. Additionally, the proposed framework achieves optimal or near-optimal performance on multiple metrics on the generalized public datasets MOT17 and MOT20, further validating the effectiveness and robustness of the proposed framework.

Key words: multi-object tracking; deep learning; team sports videos; VisionTransformer; spatio-temporal memory; attention mechanism

0　引　言

在计算机视觉领域，多目标跟踪（Multi-Object Tracking，MOT）是在一系列视频帧中持续跟踪多个目标的位置，在目标外观和位置改变时维持目标原有身份（Identity，ID），并预测目标在视觉场景中的移动轨迹[1]。MOT 通常分为在线跟踪和离线跟踪两种模式。在线跟踪要求在处理每一帧时，仅利用当前帧和历史帧的信息来确定当前帧的跟踪结果，离线跟踪则允许利用所有帧的信息以获得全局最优解[2]。在体育视频中，在线跟踪更具应用价值，能够在实时场景中对运动员进行即时监测和分析，从而为实时决策提供支持。

传统在线 MOT 方法通常包括目标检测和 ID 关联两个阶段，前者识别和定位每帧中的目标实例，后者模拟跟踪对象的状态变化，完成跟踪对象与检测结果之间的 ID 匹配，确保跟踪对象的时间连续性和一致性[3]。然而，两阶段 MOT 方法需要通过检测器获取当前帧中所有对象的边界框，提取每个边界框的重识别（Re-Identification，ReID）特征，计算开销极大。为此，研究人员提出了联合检测与嵌入（Joint Detection and Embedding，JDE）方法，在一个网络中同时预测目标位置并提取 ReID 特征[4]。但是，JDE 方法在单个模型内直接合并目标检测和 ReID 任务，触发了两个任务之间的竞争关系，降低了 MOT 的准确性。近期，VisionTransformer 模型在处理不同输入组件的依赖关系和整体决策方面表现出优越性能。由此，很好地解决了密集场景中目标之间相互作用的建模准确性问题。

相对于广泛研究的行人监控，体育视频，尤其是篮球、足球等复杂团队体育运动面临更大的挑战，如跟踪漂移、身份交换、运动模糊和频繁遮挡等问题。相较于个人体育赛事，如羽毛球、网球，团队体育比赛参与人数多、运动激烈、视频复杂。本文所提方法针对团队体育视频的复杂场景进行了以下创新：

1)提出端到端的MOT网络,同时学习目标检测和身份关联。检测网络利用基于Vision-Transformer的编码器-解码器结构,从帧序列输入中生成目标候选,得到结合多尺寸特征的目标嵌入向量。引入局部注意机制,实现局部和全局空间特征的融合。

2)通过时空记忆网络,同时对目标的时间和空间上下文建模。融合短时记忆和长时记忆,通过自注意机制得到跟踪目标的轨迹嵌入。利用目标和轨迹置信分,确定在当前帧中被跟踪对象的时空位置和可见性,提高频繁遮挡的复杂体育场景中ID关联的稳定性。

1 相关研究

传统MOT方法多采用基于检测的跟踪范式,基于高性能检测器提供的目标边界框,完成同一对象在不同帧之间的关联。Bewley等[5]提出的Sort使用卡尔曼滤波器预测后续帧中所有候选边界框的位置,其后基于当前边界框与下一帧中预测边界框之间的交并比(Intersection over Union,IOU),利用匈牙利算法完成匹配。在此基础上,Du等[6]提出的StrongSort进一步纳入了跟踪目标的外观信息,通过将目标的感兴趣区域(Region of Interest,RoI)传递给单独的卷积神经网络(Convolutional Neural Network,CNN)以提取目标的视觉特征。鄂贵等[7]利用区域全卷积神经网络(Region-based Fully Convolutional Network,R-FCN)得到可信候选框,再通过孪生网络将候选与轨迹相关联。Xu等[8]提出基于时空关系网络的相似性测量框架,通过线索编码和时空推理增强跟踪器性能。然而,两阶段MOT方法的计算效率较低,且跟踪算法中使用的状态变量的维度有限,在遮挡、旋转或动态背景等复杂场景下无法准确捕捉到对象的所有动态特征,从而影响了跟踪性能。

JDE方法可以在单个阶段同时完成检测和跟踪。Zhou等[9]提出的CenterTrack基于预测点的逐帧偏移量,将目标的逐帧传播简化为中心点跟踪问题,从而实现高效MOT。Zhang等[10]提出基于无锚框检测器的FairMOT,通过平衡检测和Re-ID任务的权重来提升MOT性能。但是,检测任务的目标是使网络能够最小化同一类别内不同对象之间的差异,以便准确地检测出每个目标。然而,ReID任务则更侧重于在同一类别内识别不同对象之间的差异,强调对象的身份关联。由于两个任务侧重点不同,优化网络以更好地完成其中一个任务,会导致在另一个任务上的性能下降。JDE方法在遮挡后牺牲了跟踪恢复,无法重新与长时间消失的对象建立连接,限制了MOT在遮挡频繁的团队体育场景中使用的性能。

近期,Transformer架构在各种视觉任务中表现优秀,其利用独特的查询-键机制,通过注意机制处理提取的深度特征[11]。Meinhardt等[12]提出了基于基于可变形注意力Transformer的TrackFormer框架,通过将对象和自回归轨迹查询连接为Transformer解码器的输入,同时执行检测和ID关联。Xu等[13]提出的TransCenter仅将Transformer用作特征提取器,并循环传递跟踪对象特征以学习每个对象的联合嵌入。Chu等[14]提出的TransMOT将跟踪目标的轨迹和检测候选对象排列成的稀疏加权图结合,由此捕捉检测、外观和运动特征的时空线索,提高复杂场景中的MOT性能。然而,上述方法侧重于使用动态嵌入来表示对象状态,缺乏对长时时空信息和自适应特征融合方法的充分建模。

2 本文方法

MOT可在视频每一帧中得到完整且准确排序的目标集合。令帧序列为$I=\{I_0,I_1,I_2,\cdots,I_T\}$,MOT的目标是在实时处理中识别和跟踪$N$个对象位置,同时保持其轨迹$T=\{T_0,T_1,T_2,\cdots,T_N\}$。为解决团队体育场景下MOT任务中目标交叉和遮挡、快速运动和身份频繁切换等难题,所提框架同时学习目标检测和身份关联,并结合存储跟踪对象长期依赖性的时空记忆块,通过记忆编码器-解码器机制,有效提取相关表示,从而在经过较长时间后的一系列帧中关联相同的对象。

本文所提框架主要由3个模块组成:

1)候选生成,在当前时间步对每帧中检测到的目标进行处理,并创建目标候选;

2)轨迹级记忆编码模块,负责聚合关联对象嵌入;

3)记忆解码器,将检测到的候选目标与已跟踪目标关联。

具体如图1所示。

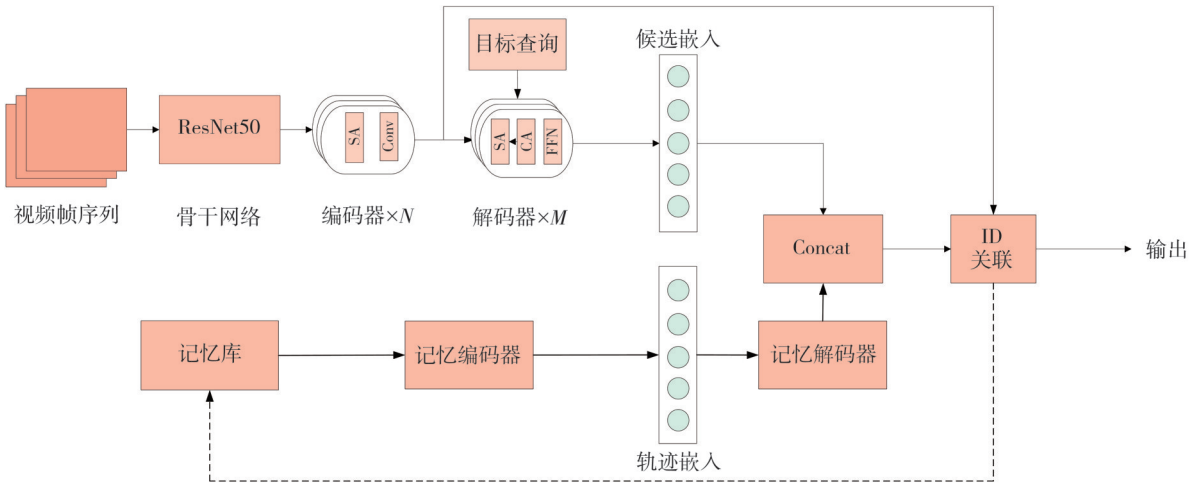


图 1　所提方法框架

Fig. 1　Framework of the proposed method

2.1　候选目标检测网络

2.1.1　结合局部注意力的编码器设计

传统 VisionTransformer 采用编码器-解码器结构，每个编码器包含多头自注意层（Self Attention，SA）和前馈神经网络层（Feedforward Neural Network，FFN）。Transformer 模型中的解码器除 SA 和 FFN 外，还包含互注意力（Cross Attention，CA）层，通过关注编码器输出的不同位置，获取输入序列的相关信息[15]。

VisionTransformer 通过自注意机制对序列中所有实体之间的相互作用进行建模，以结构化预测任务。令 $F\in R^{N\times d}$ 为 N 个实体序列，其中，d 表示每个实体的嵌入维度。将输入序列 F 投影到 3 个可学习的线性变换上，得到查询 $\boldsymbol{Q}$、键 $\boldsymbol{K}$ 和值 $\boldsymbol{V}$。自注意机制的输出为[16]

$$Att(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\cdot\boldsymbol{V}。\tag{1}$$

所提 MOT 框架中，候选生成网络包括 CNN 骨干网、VisionTransformer 编码器和解码器。首先，使用 ResNet50 作为特征提取器[17]，从两个连续帧 $\{I_t, I_{t-1}\}$ 中提取出特征 $\{X_t, X_{t-1}\}$。将特征拼接并输入 VisionTransformer 编码器 T_{Encoder}，计算特征关联

$$\delta_t=Concat(X_t,X_{t-1})。\tag{2}$$

将 δ_t 通过 N 个编码器层，得到输出嵌入 δ_θ。该嵌入将被投影到解码器的键和值矩阵中，并使用可训练目标查询与键和值矩阵进行交互。

传统 VisionTransformer 编码器由 SA 和 FFN 层组成，每层均使用 skip 连接和层归一化。在时间 t 下，对于目标 k 的注意力机制输出为

$$\delta_{t,k}^{\mathrm{att}}=Norm(Att(\delta_{t,k-1})+\delta_{t,k-1}),\tag{3}$$

式中：$Att(\cdot)$ 表示自注意力操作。

编码器的最终输出为目标 k 在时间 t 的状态

$$\delta_{t,k}=Norm(T_{\mathrm{Encoder}}(\delta_{t,k}^{\mathrm{att}})+\delta_{t,k}^{\mathrm{att}})。\tag{4}$$

传统 VisionTransformer 编码器将图像划分为固定数量分块，扁平化后作为输入序列。但是，这会导致模型忽略像素之间的局部关系。为此，对编码器进行调整，引入局部注意机制，使模型能够更好地捕捉输入图像中像素之间的局部关系[18]。图 2 给出了所提方法中 VisionTransformer 编码器结构，使用局部注意力（Local Attention，LA）模块代替 FFN 层。LA 中，Seq2Img 和 Imag2Seq 分别完成从序列到图像特征图和从图像到序列特征图的转换，以支持卷积层的局部特征提取，DWConv 表示逐深度卷积。

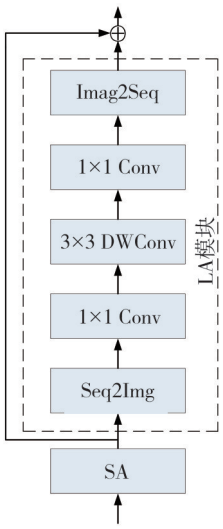


图 2　VisionTransformer 编码器结构

Fig. 2　Structure of the VisionTransformer encoder

LA模块使模型具备对输入图像的多个空间层次的感知能力，这意味着模型可以同时关注全局和局部信息，有助于更全面地理解图像内容。具体来说，将SA处理后的嵌入δ_t^{att}输入Seq2Img层，转换为二维特征图[19]

$$\delta_t^{\text{2D}} = Seq2Img(\delta_t^{\text{att}}), \tag{5}$$

式中：$\delta_t^{\text{2D}} \in R^{d\times h\times w}$，$d$为通道数，$h=H/p$，$w=W/p$，$H$、$W$分别为视频帧$I$的高度和宽度；$p$为卷积操作中卷积核在该特征图上的滑动步长。LA模块处理后，需要将特征图δ_t^{2D}转换回一维嵌入

$$\delta_t^{\text{att}} = Img2Seq(\delta_t^{\text{2D}}), \tag{6}$$

式中：$\delta_t^{\text{att}} \in R^{j\times d}$，$j=h\times w$。传统VisionTransformer编码器利用FFN对局部依赖性建模，所提方法利用LA模块替代FFN，通过全局注意力和局部卷积的结合，得到更全面的特征表示。

2.1.2 多尺度特征

在团队体育视频MOT任务中，引入多尺度特征能够提高模型对不同目标的适应性，更全面地捕捉上下文信息，有效处理尺度变化和减轻遮挡影响。

令$\{X_t^l\}^L$为骨干网络ResNet50在时间t的多尺度输出，共包含L个级别，其中每个X_t^l都对应于从骨干网络的不同级别(尺度)上获得的特征图。将$\{X_t^l\}^L$扁平化并拼接为δ_t，作为输入传递至SA模块。为计算不同尺度特征图的局部关联，所提方法在编码器的SA模块后部署L个LA模块，对应于每个尺度的特征图

$$\{X_t^l\}^L = split(\delta_{t,i}^{\text{att}})。 \tag{7}$$

最后，将不同LA模块的输出扁平化并拼接在一起。由此，提供更全面、多层次的特征表示，从而捕捉到不同尺度下的信息。这有助于模型更好地理解和处理输入数据中的空间层次结构，提高模型对复杂场景的感知能力。

2.1.3 VisionTransformer解码器

将VisionTransformer编码器生成的嵌入输入M个解码器T_{Decoder}组成的解码层。每个解码器包含处理查询的SA模块、处理键-值对的CA模块和提高非线性处理能力的FFN模块。解码器T_{Decoder}结合可训练目标查询q_t^{can}生成检测候选

$$Q_t^{\text{can}} = T_{\text{Decoder}}(\delta_\theta, q_t^{\text{can}})。 \tag{8}$$

2.2 时空记忆网络

团队体育视频中，MOT任务需要考虑目标在时空上的长期依赖性，即保留视觉运动特征的同时，捕捉在视觉和空间领域中同时出现的跨帧相似性。因为团队运动涉及复杂的运动轨迹和相互作用，需要对目标之间的时空关系进行建模。学习长期时空依赖性有助于更准确地预测目标的未来位置、运动轨迹和团队中的相互作用，提高MOT系统的性能和鲁棒性，实现对动态场景中复杂目标行为的理解和准确跟踪。

根据上述分析，本文所提框架创建记忆库M，存储所有N个被跟踪目标在过去T个时间步的历史状态(轨迹)N_{t-1}^{track}。M采用先进先出结构，最多保留$N_{\max}$个对象，每个跟踪对象最多保留$T_{\max}$个时间步。一旦T超过$T_{\max}$，则从记忆库中移除该轨迹最早的一个状态。针对不同体育项目，$N_{\max}$设定为场上运动员和裁判数量，$T_{\max}$则应在硬件限制内设为较大步数，以处理目标遮挡或长时间脱离跟踪范围的情况。该对象不存在于特定帧中时，其状态用0填充。由此，所提框架对一系列过去帧中活跃对象进行状态跟踪。

2.2.1 记忆编码器

时空记忆网络中，通过记忆编码器M_{Encoder}进行记忆编码，并利用注意力及模块提取轨迹嵌入，图3给出了记忆编码器结构图。短时记忆CA模块关注即时状态，汇聚来自连续帧的嵌入，显著减轻数据固有的潜在噪声，确保提取更清晰的嵌入。长时记忆CA模块挖掘记忆涵盖时间跨度内的显著特征，捕捉长时间间隔内的依赖关系和关联。SA模块整合短时和长时分支生成的嵌入，促进对候选嵌入的更全面理解。

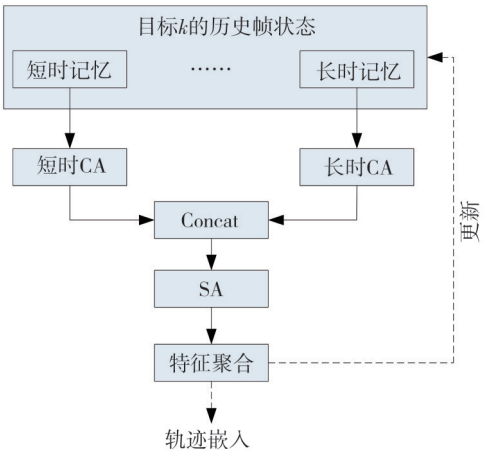


图3 记忆编码器结构

Fig. 3 Structure of the memory encoder

记忆库中，短时记忆和长时记忆分别包含过去T_S和T_L个状态，$T_S \ll T_L$。短时CA取最近状态

δ_{t-1}为查询输入$\boldsymbol{Q}$，确保模型能够迅速响应数据的最新变化。长时CA则取动态记忆聚合标记（Dynamic Memory Aggregation Tokens，DMAT）为查询输入$\boldsymbol{Q}$，DMAT代表着每个时间步更新的记忆经过聚合处理的嵌入向量[20]，表示为

$$Q_{t-1}^{\mathrm{DMAT}}=\{\delta_{t-1}^{k}\}_{k=1:N_{\mathrm{track}}}。\tag{9}$$

短时和长时CA的键输入$\boldsymbol{K}$和值输入$\boldsymbol{V}$分别为不同长度的历史状态，由此，模型能够同时关注目标候选在不同历史事件的状态，促进更丰富的上下文理解。初始时，所有轨迹的DMAT是相同的，代表每个跟踪对象具有相同的记忆表征。时间步$t>0$时，DMAT基于上一个时间步的状态迭代更新，以表征跟踪对象状态随时间的演变。

其后，通过SA模块合并短时和长时CA的输出，生成轨迹嵌入Q_t^{track}。此外，SA模块还将更新后的Q_t^{DMAT}返回长时记忆中。由此，记忆机制保留并利用了跟踪对象的时间上下文，增强了整体跟踪性能。

2.2.2　记忆解码器

记忆解码器M_{Decdoer}取检测网络编码器T_{Encoder}的编码帧特征，检测网络解码器T_{Decoder}的候选嵌入，以及记忆网络编码器的轨迹嵌入作为输入，预测最终的跟踪结果。M_{Decoder}的输入中，编码帧特征提供了结合局部和全局的上下文图像多尺度特征，候选嵌入给出了关于目标存在的初步假设，轨迹嵌入则捕捉历史和当前轨迹的细微差异。M_{Decoder}由多个VisionTransformer解码器堆叠组成，使用候选嵌入Q_t^{can}和轨迹嵌入Q_t^{track}为查询$\boldsymbol{Q}$，将T_{Encoder}输出的$\delta_{t,k}$作为键值对，完成跟踪对象和跟踪轨迹的匹配映射，输出更新后的融合嵌入$[Q_t'^{\mathrm{can}},Q_t'^{\mathrm{track}}]$。融合过程实质上是一种查询和键值对之间的匹配映射，由此将跟踪对象和跟踪轨迹的信息结合起来，生成更新后的融合嵌入。

VisionTransformer的注意力机制中，键$\boldsymbol{K}$负责提供上下文信息，以便帮助模型理解当前查询$\boldsymbol{Q}$的重要性，而值$\boldsymbol{V}$则负责提供与键$\boldsymbol{K}$相关的特定信息。通过将$\delta_{t,k}$用作键值对$\boldsymbol{K}$和$\boldsymbol{V}$，基于式(1)将包含候选嵌入信息和轨迹嵌入信息的每个查询映射到T_{Encoder}输出的丰富视觉表征中，模型能够在处理查询$\boldsymbol{Q}$时，充分利用T_{Encoder}提取的丰富视觉特征，确保生成的跟踪预测不仅基于历史数据，且考虑到了视觉背景下的图像特征，帮助模型更好地理解和推断输入数据，进而提高模型的性能和鲁棒性。Transformer模型中，经过注意力计算后生成的加权和被认为是查询$\boldsymbol{Q}$的一种表示。由此，更新后的融合嵌入$Q_t'^{\mathrm{can}}$和$Q_t'^{\mathrm{track}}$可视为查询$\boldsymbol{Q}$的更新版本，其保留了与键值对$\boldsymbol{K}$和$\boldsymbol{V}$相关的信息，经过了注意力计算后生成最终的跟踪结果。

基于最终得到的嵌入$[Q_t'^{\mathrm{can}},Q_t'^{\mathrm{track}}]$，计算跟踪目标的包围盒坐标$[B_t^{\mathrm{can}},B_t^{\mathrm{track}}]$和置信分$[C_t^{\mathrm{can}},C_t^{\mathrm{track}}]$。最后，利用跟踪目标的位置和状态进行轨迹和记忆的更新。

对于$[Q_t'^{\mathrm{can}},Q_t'^{\mathrm{track}}]$中的每个条目$q_t^k$，令$o_t^k\in[0,1]$为目标性得分，$u_t^k\in[0,1]$为唯一性得分。$o_t^k=1$表示识别出目标，$u_t^k=1$表示该目标是与其他目标均不重复的唯一目标。针对候选对象和轨迹，综合置信分计算为

$$C_t^k=o_t^k\cdot u_t^k。\tag{10}$$

跟踪过程中的置信预测包含候选对象置信C_t^{can}和轨迹查询置信C_t^{track}，评估候选对象或查询的有效性。对于每个条目q_t^k，解码器输出包含目标中心坐标、高度和宽度的包围框b_t^k。推理过程中，对$[Q_t'^{\mathrm{can}},Q_t'^{\mathrm{track}}]$中的每个条目设定置信度阈值，仅保留高置信度的对象和轨迹。将包围框输出和跟踪ID合并，得到最终跟踪结果。

模型训练中，整个框架的损失函数为

$$L=\lambda_{\mathrm{cls}}\cdot L_{\mathrm{cls}}+\lambda_{L1}\cdot L_{L1}+\lambda_{\mathrm{giou}}\cdot L_{\mathrm{giou}},\tag{11}$$

式中：L_{cls}为预测目标类别的焦点损失[21]；L_{L1}为预测目标包围框的$L1$损失；L_{giou}为包围框广义交并比损失[22]；λ为权重参数。

3　实验和讨论

3.1　数据集和参数设置

所提方法在16.04 Linux系统上使用Python3.8和PyTorch1.7，Cuda 11.0框架实现，硬件配置为i5-13400 CPU，32G RAM，GPU为NVIDIA Tesla V100。训练数据由SportsMoT、MOT20数据集的训练部分组成，在训练过程中使用随机水平翻转、裁剪、缩放等数据增强方法防止过拟合。使用AdamW优化器，采用以序列为导向的训练方案，初始序列长度为4帧，每20代增加4帧[23]。模型训练代数为200。骨干网络ResNet50的学习率为2.0×10^{-5}。Trasnformer的初始学习率为3.0×10^{-2}，权重衰减为1.0×10^{-2}，λ_{cls}、λ_{L1}和λ_{giou}分别设为3，6和3。

首先，使用体育视频SportsMOT公开数据集[24]，分析所提方法在体育场景中的性能。Sport-

MOT包含从英超和奥运会等赛事视频中采集的足球、篮球和排球的高质量比赛视频，该数据集共包括240个视频序列，每个序列的分辨率为720 P，共计包含150 379个图像帧。

此外，为评估所提框架的通用性，在MOT任务最新基准数据集MOT17[25]和MOT20[26]上对所提方法与其他方法进行性能比较。MOT17是高质量的大规模目标跟踪基准数据集，包含静态和动态摄像机拍摄的多个不同场景，共14个视频序列，11 235个视频帧，每帧均包含手工注释。MOT20数据集包含从无约束环境中提取的8个稠密人群序列，平均每帧246个行人，人群密度较大，为MOT任务带来更大的挑战。但要指出，所提方法针对的是体育场景中的MOT任务，通常要检测的目标密度要远小于地铁口、演唱会等异常拥挤的场景。

3.2 评估指标

为公平准确地评价算法性能，采用MOT中常用的标准度量作为性能指标[27]，列举如下：

1）多目标跟踪准确度(Multiple Object Tracking Accuracy，MOTA)：作为最广泛使用性能指标，MOTA在检测级别完成匹配。MOTA测量3种类型的跟踪错误：误报、漏报和ID切换，计算为

$$A_{\mathrm{MOT}}=1-\frac{N_{\mathrm{FN}}+N_{\mathrm{FP}}+N_{\mathrm{IDSW}}}{N_{\mathrm{GT}}},\quad(12)$$

式中：N_{FN}为未检测到的目标数；N_{FP}为错误的跟踪预测数；N_{IDSW}为ID切换次数；N_{GT}为实际目标总数。

2）高精度多目标跟踪准确度(Higher Order Tracking Accuracy，HOTA)：用于对跟踪器进行排名的统一度量，通过在检测级别执行匹配，同时在轨迹上对关联进行全局评分。HOTA计算为

$$A_{\mathrm{HOT}}=\sqrt{(1-A_{\mathrm{track}})\times(1-A_{\mathrm{ID}})},\quad(13)$$

式中：A_{track}为表示跟踪准确度，即正确匹配的跟踪预测数量除以总跟踪预测数量；A_{ID}为身份识别准确度，即正确匹配的跟踪对数量除以总跟踪对数量。

3）IDF1：该指标强调对目标轨迹的标识和准确性，计算为ID精度和ID召回率的调和平均值：

$$IDF1=\frac{2\times N_{\mathrm{TP}}}{2\times N_{\mathrm{TP}}+N_{\mathrm{IDSW}}+N_{\mathrm{FP}}},\quad(14)$$

式中：N_{TP}为正确匹配的跟踪预测数。

4）定位准确性(Location Accuracy，LocA)：将定位误差的统计度量表示为LocA分数，以反映模型在定位目标方面的性能，计算公式为

$$A_{\mathrm{Loc}}=\frac{1}{N}\sum_{i=1}^{N}\frac{D_i}{S_i},\quad(15)$$

式中：D_i为跟踪器预测的边界框与真实边界框之间的位置误差；S_i为目标的实际边界框尺寸。

5）ID切换(Identity Switches，IDSW)：在跟踪过程中发生目标ID切换的次数。ID切换表示模型在某帧中将一个目标与另一个目标错误地关联起来，导致对目标身份的混淆。

3.3 评估结果

首先，验证所提框架在复杂体育场景中的MOT性能，表1给出了在SportsMOT数据集上不同方法的测试结果。其中箭头向上表示指标数值越大越好，箭头向下表示指标数值越小越好。由于该数据集为高清晰度比赛视频，目标外观较易于辨认，且场景密度相对较低，但需要在检测结果中排除观众和裁判的影响。此外，运动场景中需要处理的难点主要是队友和对手之间的严重遮挡、目标快速运动后的轨迹关联、运动模糊和姿态显著变化等问题。因此，对于各方法的检测和轨迹的关联性能要求更高。从表1结果中可发现，基于JDE的MOT方法CenterTrack和FairMOT的整体性能较差。

表1 不同方法在SportMOT数据集上的性能比较
(粗体表示最好性能)

Tab. 1 Performance comparison of different methods on the SportMOT dataset (bold indicates the best performance)

方法	MOTA ↑/%	HOTA ↑/%	IDF1 ↑/%	LocA ↑/%	IDSW ↓
CenterTrack[9]	91.2	60.8	63.5	87.9	6 657
FairMOT[10]	86.4	49.3	53.5	83.9	9 928
TrackFormer[12]	93.7	69.9	68.5	90.1	4 282
TransCenter[13]	94.2	71.8	71.4	91.9	3 644
TransMOT[14]	96.5	73.1	72.2	92.1	3 277
所提方法	**98.5**	**75.7**	**78.2**	**96.1**	**1 974**

尽管检测器可提取出高区分度的外观特征，但仅靠外观特征和re-ID算法进行目标关联，在例如运动员号码被遮挡、剧烈姿态变化造成的外观差异极大等情况下的识别精度会大幅下降。TrackFormer、TransCenter和TransMOT利用基于Transformer的模型，提取出运动不变性特征，一定程度上提高了在复杂体育场景中的MOT性能。与其他方法相比，所提方法在绝大部分指标上都取得了最好性能，这主要得益于所提方法通过基于改进VisionTransformer的检测网络，有效融合了局部和全局空间特征，提

高了模型对目标的感知定位能力。同时，通过记忆网络实现准确的时空关联匹配，显著提升了复杂体育场景中的MOT性能。

为验证所提方法的通用性，表 2 和表 3 分别给出了在MOT17和MOT20数据集上所提方法与其他方法的实验结果。

表 2　不同方法在通用数据集MOT17上的性能比较（粗体表示最好性能）

Tab. 2　Performance comparison of different methods on the MOT17 dataset (bold indicates the best performance)

方法	MOTA ↑/%	HOTA ↑/%	IDF1 ↑/%	LocA ↑/%	IDSW ↓
CenterTrack[9]	76.8	60.5	74.9	78.2	3 218
FairMOT[10]	73.7	59.3	72.3	81.4	3 303
TrackFormer[12]	65.0	59.7	63.9	83.5	3 258
TransCenter[13]	73.2	54.5	62.2	80.9	3 663
TransMOT[14]	**76.4**	66.3	76.3	**88.9**	1 623
所提方法	76.1	**66.9**	**77.7**	88.7	**1 509**

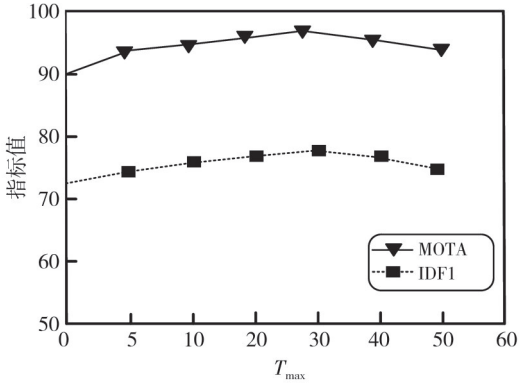
表 3　不同方法在通用数据集MOT20上的性能比较（粗体表示最好性能）

Tab. 3　Performance comparison of different methods on the MOT20 dataset (bold indicates the best performance)

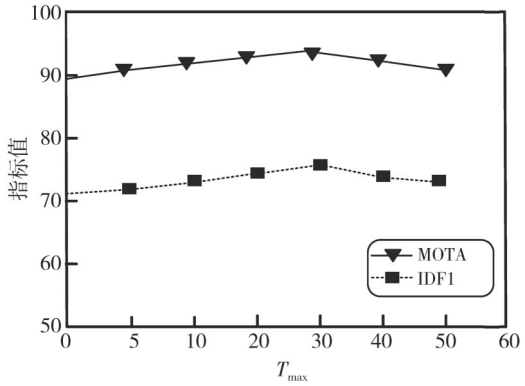
方法	MOTA ↑/%	HOTA ↑/%	IDF1 ↑/%	LocA ↑/%	IDSW ↓
CenterTrack[9]	65.3	52.1	66.7	76.4	3 277
FairMOT[10]	61.8	50.4	59.8	74.9	5 243
TrackFormer[12]	61.0	56.7	66.9	78.8	3 285
TransCenter[13]	68.6	52.4	64.5	80.9	3 277
TransMOT[14]	**77.3**	**65.5**	**75.2**	**89.8**	**1 601**
所提方法	72.4	63.3	71.7	86.4	2 097

由表 2 和表 3 可发现，所提方法在HOTA、IDF1和IDs等多个指标上均取得了相对较好的水平。特别是在MOT17数据集上，所提方法在HOTA、IDF1和IDSW指标上均取得了最好性能。这得益于所提方法在检测网络中通过LA模块替代传统的FFN，以及全局注意力和局部卷积的结合，使该方法能够获取更全面、更多层次的特征表示，从而更好地理解和处理输入数据，提高了模型的感知能力和鲁棒性。同时，通过记忆网络机制有效地捕捉目标之间的长期时空依赖性，使得模型能够更准确地预测目标的未来位置和运动轨迹，从而实现了对目标的准确跟踪。但要指出，由于所提方法主要针对复杂运动场景的MOT任务，所以并未考虑照明剧烈变化、大量目标拥挤等情况。相较于MOT17，MOT20数据集包含了更多更拥挤的场景，目标尺寸更小，因此所提方法在MOT20上的性能略低于TransMOT方法。从整体结果中可得出结论，尽管所提方法主要针对足球、篮球、排球等团队体育视频场景，但在通用公开数据集中的性能依然能够达到较好水平，验证了所提框架的通用性。

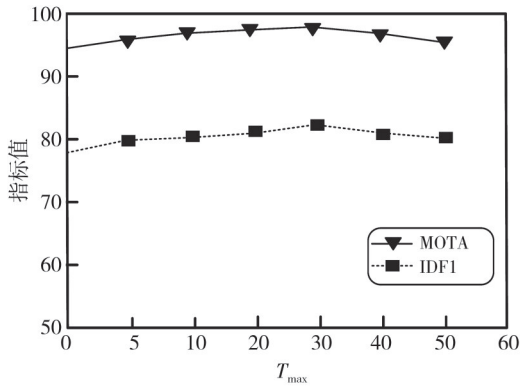
所提框架的记忆网络在记忆库M中保存所有被跟踪目标在过去T个时间步的轨迹，每个跟踪对象最多保留$T_{\max}$个时间步。在SportMOT数据集的排球、篮球和足球子集上分别设定不同的记忆长度，分析不同记忆长度对模型性能的影响，图 4 给出了MOTA和IDF1的实验结果。



（a）篮球子集



（b）足球子集



（c）排球子集

图 4　不同记忆长度在SportsMOT数据集各子集上的实验结果

Fig. 4　Experimental results of different memory lengths on each subset of the SportsMOT dataset

由图 4 可发现，当$T_{\max}<30$时，记忆网络能够显著提升模型的MOT性能，证明了所提框架的记忆

网络设计在MOT任务中的积极作用。但进一步增加记忆库容量时，记忆库中的存储信息会变得过于庞大，导致模型在处理长期依赖关系时的性能下降。

图5为所提方法在SportsMOT数据集不同体育场景下的跟踪结果示例。其中，不同颜色的包围框表示不同目标的跟踪结果，包围框上方数字为目标ID，图片右上角或右下角标注了视频帧序列号。所选取的每组视频帧序列的间隔为9帧，以更好地展示所提方法解决复杂团队体育场景下MOT任务中身份交换、运动模糊和严重遮挡等难题时的稳定性。从结果中可发现，在图5(a)的篮球场景下，所提方法始终能够保持对所有运动员的跟踪和ID关联，即使在第249帧中ID1目标肢体被ID9目标严重遮挡的情况下依然实现了准确检测。在图5(b)的排球场景下，第41帧中ID4目标被完全遮挡，但所提方法利用记忆机制，在第50帧中成功检测并关联到该目标。此外，同场景下ID9目标在不同视频帧中表现出剧烈姿态变化和运动模糊，但所提方法始终保持对该目标的准确跟踪。在图5(c)的足球场景下，第531帧中ID17目标被ID8目标完全遮挡，在540帧中所提方法准确完成了对重新出现的ID17目标的检测和ID关联。此外，该视频帧序列中目标出现的摔倒(ID13)、转体(ID8)、拼抢(ID20)等动作均未影响到跟踪稳定性。可视化结果证明，所提方法能够在包括篮球、足球和排球在内的多种不同复杂体育场景中实现有效稳定的多运动员跟踪。

(a) 篮球场景

(b) 排球场景

(c) 足球场景

图5 SportsMOT数据集不同子集的MOT可视化结果示例

Fig. 5 Examples of MOT visualization results for different subsets of the SportsMOT dataset

4 结 论

为解决复杂团队体育场景中多运动员跟踪问题，提出了基于改进VisionTransformer和记忆网络的MOT框架。其中，在检测网络中结合局部注意力模块和多尺寸特征融合机制，充分提取了视频帧的局部和空间特征。通过记忆网络的设计，通过短时和长时记忆机制，将帧内目标空间特征与全局时间特征融合，提高了在线MOT的性能。所提方法在SportsMOT体育视频数据集上取得了最优性能，验证了其在不同团队体育视频场景中的有效性。此外，MOT20数据集上的比较实验证明，所提方法具有良好的通用性，适用于各种不同场景下的MOT任务。未来，将尝试通过监视模型的损失曲线、梯度变化、注意力权重等手段，对模型进行更细致的调参，以支持更大的记忆库容量，进一步提高模型的MOT性能。

参考文献：

[1] PAL S K, PRAMANIK A, MAITI J, et al. Deep learning in multi-object detection and tracking: state of the art [J]. Applied Intelligence, 2021, 51(9): 6400-6429.

[2] 周雪，梁超，何均洋，等. 一体化多目标跟踪算法研究综述[J]. 电子科技大学学报，2022，51(5)：728-736.
ZHOU Xue, LIANG Chao, HE Junyang, et al. A survey on one-shot multi-object tracking algorithm[J]. Journal of University of Electronic Science and Technology of China, 2022, 51(5): 728-736. (in Chinese)

[3] YAO R, LIN G, XIA S, et al. Video object segmentation and tracking: A survey[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2020, 11(4): 1-47.

[4] 晏康，曾凤彩，何宁，等. 引入注意力机制的 JDE 多目标跟踪方法[J]. 计算机工程与应用，2022，58(21)：189-196.
YAN Kang, ZENG Fengcai, HE Ning, et al. JDE Multi-Object Tracking Method with Attention Mechanism [J]. Computer Engineering and Applications, 2022, 58(21): 189-196. (in Chinese)

[5] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking [C]//2016 IEEE International Conference on Image Processing (ICIP). IEEE, 2016: 3464-3468.

[6] DU Y, ZHAO Z, SONG Y, et al. Strongsort: Make deepsort great again[J]. IEEE Transactions on Multimedia, 2023, 25: 8725-8737.

[7] 鄂贵，王永雄. 基于 R-FCN 框架的多候选关联在线多目标跟踪[J]. 光电工程，2020，47(1)：29-37.
E Gui, WANG Yongxiong. Multi-candidate association online multi-target tracking based on R-FCN framework [J]. Opto-Electronic Engineering, 2020, 47(1): 29-37. (in Chinese)

[8] XU J, CAO Y, ZHANG Z, et al. Spatial-temporal relation networks for multi-object tracking [C]//2019 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019: 3987-3997.

[9] ZHOU X, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points [C]//2020 European Conference on Computer Vision (ECCV). Springer Science, 2020 : 474-490.

[10] ZHANG Y, WANG C, WANG X, et al. FairMOT: On the fairness of detection and re-identification in multiple object tracking[J]. International Journal of Computer Vision, 2021, 129(11): 3069-3087.

[11] 李清格，杨小冈，卢瑞涛，等. 计算机视觉中的 Transformer 发展综述[J]. 小型微型计算机系统，2023，44(4)：850-861.
LI Qingge, YANG Xiaogang, LU Ruitao, et al. Transformer in computer vision: A survey[J]. Journal of Chinese Computer Systems, 2023, 44(4): 850-861. (in Chinese)

[12] MEINHARDT T, KIRILLOV A, LEAL-TAIXÉ L, et al. Trackformer: Multi-object tracking with transformers [C]//2022 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 8834-8844.

[13] XU Y, BAN Y, DELORME G, et al. TransCenter: Transformers with dense representations for multiple-object tracking [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(6): 7820-7835.

[14] CHU P, WANG J, YOU Q, et al. TransMOT: Spatial-temporal graph transformer for multiple object tracking [C]//2023 Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2023: 4859-4869.

[15] 王宁，席茂，周文罡，等. 深度视觉目标跟踪进展综述[J]. 中国科学技术大学学报，2021，51(4)：335-344.
WANG Ning, Xi Mao, ZHOU Wengang, et al. Recent advance in deep visual object tracking[J]. Journal of University of Science and Technology of China, 2021, 51(4): 335-344. (in Chinese)

[16] ARKIN E, YADIKAR N, XU X, et al. A survey: Object detection methods from CNN to transformer [J]. Multimedia Tools and Applications, 2023, 82(14): 21353-21383.

[17] 张涛，张晓利，任彦. Transformer 与 CNN 融合的单目图像深度估计[J]. 哈尔滨理工大学学报，2022，27(6)：88-94.
ZHANG Tao, ZHANG Xiaoli, REN Yan. Monocular image depth estimation based on the fusion of transformer and CNN[J]. Journal of Harbin University of Science and Technology, 2022, 27(6): 88-94. (in Chinese)

[18] LI K, YU R, WANG Z, et al. Locality guidance for improving vision transformers on tiny datasets [C]//2022 European Conference on Computer Vision (ECCV), 2022: 110-127.

[19] SONG C H, YOON J, CHOI S, et al. Boosting

vision transformers for image retrieval[C]//2023 Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2023: 107-117.

[20] CAI J, XU M, LI W, et al. MeMOT: Multi-object tracking with memory[C]//2022 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2022: 8080-8090.

[21] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//2017 Proceedings of the IEEE International Conference on Computer vision (ICCV). IEEE, 2017: 2899-3007.

[22] REZATOFIGHI H, TSOI N, GWAK J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//2019 Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2019: 658-666.

[23] LLUGSI R, EL YACOUBI S, FONTAINE A, et al. Comparison between Adam, AdaMax and Adam W optimizers to implement a weather forecast based on neural networks for the andean city of quito[C]//2021 IEEE Fifth Ecuador Technical Chapters Meeting (ETCM). IEEE, 2021: 1-6.

[24] CUI Y, ZENG C, ZHAO X, et al. SportsMOT: A large multi-object tracking dataset in multiple sports scenes[C]//2023 Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2023: 9887-9897.

[25] MILAN A, LEAL-TAIXÉ L, REID I, et al. MOT16: A benchmark for multi-object tracking[DB/OL]. (2016-05-03)[2024-01-27]. http://arxiv.org/abs/1603.00831v2.

[26] DENDORFER P, REZATOFIGHI H, MILAN A, et al. MOT20: A benchmark for multi object tracking in crowded scenes[DB/OL]. (2020-03-19)[2024-01-27]. https://arxiv.org/abs/2003.09003.

[27] 潘昊, 刘翔, 赵静文, 等. 联合Transformer与BYTE数据关联的多目标实时跟踪算法[J]. 激光与光电子学进展, 2023, 60(6): 144-151.
PAN Hao, LIU Xiang, ZHAO Jingwen, et al. Multi-target real-time tracking algorithm based on transformer and BYTE Data[J]. Laser & Optoelectronics Progress, 2023, 60(6): 144-151. (in Chinese)