

基于 Transformer 的双分支检测和重识别的多行人追踪

黄丹丹¹, 张新茹¹, 刘智^{1,2}, 彭刚³

(1. 长春理工大学电子信息工程学院, 长春 130022; 2. 长春理工大学空间光电技术国家地方联合工程研究中心, 长春 130022; 3. 长春师凯科技产业有限责任公司研发技术部, 长春 130015)

摘要: 针对密集行人场景下多目标跟踪存在的目标错检、漏检、关联不准确、重识别错误等问题, 提出了一种基于 Transformer 的多行人跟踪网络。算法包含检测、数据关联和追踪 3 个模块, 其中检测模块采用选择性查询收集方法增强解码器对关键特征的收集, 提高模型对目标的表征能力, 有效减少目标错检漏检问题; 数据关联模块采用双线性长短期记忆网络 (BLSTM) 与二次数据关联的融合策略, 解决密集行人由于相似外观导致关联不准确的问题; 最后在追踪模块上将注意力金字塔嵌入金字塔时空聚合模块以捕获不同尺度特征图的时空信息, 提高了目标重识别的准确性。本文网络在公开数据集 MOT16、MOT17 上进行了性能测试, 实验结果表明: 相较于其他方法, 本文方法能够实现更准确的多行人追踪。

关键词: 计算机视觉; 多行人追踪; 重识别; 数据关联

中图分类号: TP391 **文献标志码:** A **文章编号:** 1671-5497(2025)10-3384-10

DOI: 10.13229/j.cnki.jdxbgxb.20231415

Multi-pedestrian tracking based on Transformer double branch detection and re-identification

HUANG Dan-dan¹, ZHANG Xin-ru¹, LIU Zhi^{1,2}, PENG Gang³

(1. School of Electronics and Information Engineering, Changchun University of Science and Technology, Changchun 130022, China; 2. National and Local Joint Engineering Research Center of Space Photoelectric Technology, Changchun University of Science and Technology, Changchun 130022, China; 3. Changchun Shikai Technology Industry Co., Research and Development Center, Changchun 130015, China)

Abstract: Aiming at the problems of target misdetection and omission, inaccurate association, and re-identification error in multi-target tracking in dense pedestrian scenarios, this study proposes a multi-pedestrian tracking network based on Transformer. The algorithm consists of three modules: detection, data association and tracking, in which the detection module adopts the selective query recollection method to enhance the decoder's collection of key features, improve the model's ability to characterize the target, and effectively reduce the problem of target misdetection and omission; the data association module adopts the fusion strategy of bilinear LSTM and quadratic data association, to solve the inaccurate association of

收稿日期: 2023-12-17.

基金项目: 吉林省科技厅重点研发项目 (20230201071GX).

作者简介: 黄丹丹 (1984-), 女, 讲师, 博士. 研究方向: 图像处理与机器视觉. E-mail: hdd@cust.edu.cn

dense pedestrians due to the similarity of the appearance of the target; Finally, the attention pyramid is embedded into the pyramid spatio-temporal aggregation module on the tracking module to capture the spatio-temporal information of the feature map at different scales, which improves the accuracy of target re-identification. The performance of the proposed network is tested on the publicly available datasets MOT16, MOT17, and the experimental results show that the method in this study is able to achieve more accurate multi-pedestrian tracking compared to other methods.

Key words: computer vision; multi-pedestrian tracking; re-identification; data association

0 引言

多目标跟踪(Multiple object tracking, MOT)是计算机视觉领域的研究热点之一,在视频监控、人机交互、自动驾驶等领域^[1]有广泛应用。传统的MOT方法大多基于高斯混合模型构建目标外观,然后采用贝叶斯理论或数据关联模型进行跟踪^[2]。随着深度学习的发展,基于卷积神经网络的方法逐渐取代了传统的多目标跟踪方法,并且极大改善了跟踪效果。但在密集行人跟踪方面,由于场景复杂多样和人体的柔性变化,仍存在遮挡导致的漏检、错检,以及相似外观和重识别错误问题。

针对上述问题,Wojke等^[3]提出了Simple online and realtime tracking(SORT),使用匈牙利算法在图像空间和逐帧数据关联中执行卡尔曼滤波,提高数据关联的准确性。Zhang等^[4]则联合检测与重识别,提出了FairMOT,通过消除检测分支的不公平,有效学习了ReID特征。注意力机制的提出,使得基于Transformer框架的多目标跟踪方法得到了深入研究。Xu等^[5]提出了TransCenter方法,通过图像相关的密集检测查询和稀疏跟踪查询解决了MOT问题。此外,GTR算法^[6]为所有对象生成全局轨迹并与对象检测器联合训练。Cai等^[7]提出了MeMOT,通过时空内存存储被跟踪对象的ID嵌入实现MOT。这类方法将Transformer的自注意力机制及多头自注意力层广泛地应用于特征提取网络,能够有效缓解严重遮挡情况下漏检和错检的问题。

上述算法虽然提高了多行人跟踪的准确性,但缺乏对解码器层数间的信息传递以及不同尺度特征图时空特征的考虑,因此,未能有效解决行人间的严重遮挡和相似外观问题。为此,本文对基准网络TransTrack进行改进,提出了一个具有双分支检测与重识别模块的网络结构,在基于查询对象的检测上引入了选择性查询收集模块^[8],该

模块帮助解码器层有针对性地选择中间特征,提高目标特征的表达能力;在ReID部分引入了金字塔时空聚合(Pyramid spatial-temporal aggregation, PSTA)^[9]与注意力金字塔(Attention pyramid network, APNet)^[10]融合模块,精准捕获不同尺度特征图的时空特征,实现目标ID的准确识别;在关联部分采用了双线性长短期记忆网络(Bilinear LSTM, BLSTM)^[11]和二次数据关联^[12]融合模块,存储历史帧轨迹并通过高低阈值的方式对检测框进行筛选,提高算法的关联性能,最终实现复杂场景中对密集人群的准确跟踪。

1 多目标追踪模型

1.1 整体框架

基于Transformer的多目标追踪方法仍存在诸多缺陷,如TrackFormer^[13]只能应对短期遮挡,MOTR^[14]将分类置信度低的目标判定为退出,且无法处理新出现的目标。本文采用TransTrack^[15]为基准网络,该方法不但可以有效捕捉新出现的目标,而且可以通过编码器-解码器结构对连续两帧的组合特征映射进行处理,保持了目标的一致性和连续性。本文在此基础上对基准网络进行改进,旨在提高算法对复杂场景下密集人群的跟踪性能。本文保留了TransTrack网络的检测与重识别分支作为主体框架,主要从检测网络、数据关联和行人重识别3个方面对基准算法进行改进,改进网络的总体框架如图1所示。

从图1中可知,本文方法首先将视频中相邻两帧的图像输入骨干网络进行特征提取,然后将提取的相邻特征图送入编码器,最后将编码后的特征图分别送入检测分支和跟踪分支进行后续处理。在检测分支中,图像的特征图经过解码后再进行分类和回归,然后生成行人检测结果,其中解码器添加了选择性查询收集模块,以增强特征表达,提高检测器的性能;在跟踪分支中,特征图在

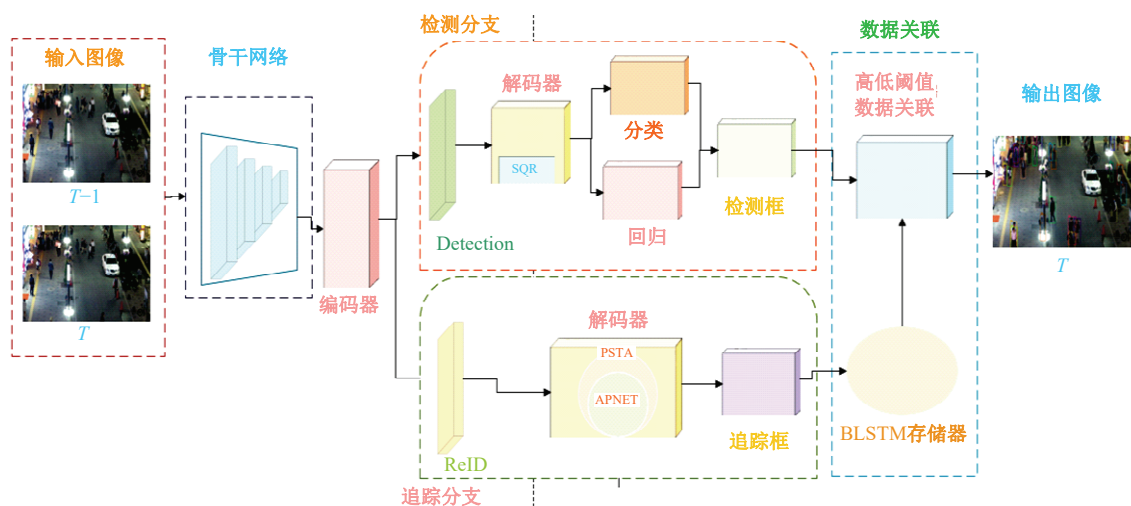


图 1 整体框架

Fig. 1 Overall framework

解码器中进行 ReID, 最后生成跟踪框, 其中 ReID 模块引入了注意力金字塔模块和金字塔时空聚合模块使其能够对不同尺度时空特征进行识别和利用, 保证重识别的准确性。接下来, 将检测结果和跟踪结果进行数据关联, 本文使用双线性 LSTM 进行历史帧轨迹的高效存储以减少轨迹缺失, 采用二次关联算法将无重叠的检测框分为高分和低分的目标检测框, 使得低分框的图像也可以与无匹配对象的轨迹进行关联, 减少数据关联的错漏, 得到更准确的行人跟踪结果。

1.2 检测分支

本文采用 Deformable DETR^[16] 作为检测器, 通过编码器提取特征, 采用解码器进行分类和回归, 得到检测结果。编码器和解码器包括多尺度可变形注意力层、残差连接正则化层和前馈网络层。编码器利用多尺度可变形自注意力交换不同尺度的信息, 解码器则通过多尺度可变形的交叉注意力将对象查询与图像特征交互。虽然多尺度可变形注意力机制能有效利用目标的多尺度特征, 但增加了模型的计算复杂性, 导致推理时间增加, 其主要原因在于各层次间的信息传递方式不够高效。

Deformable DETR 的解码器总共有 6 层, 但由于每层信息的单向传递设计, 后面的层无法接收前面层的信息, 导致检测结果的精度下降。现有文献表明, Transformer 解码器的第 6 层受第 4 层和第 5 层的影响较大。具体而言, 对于真阳性 TP (模型的预测与真实对象的位置和类别都正确匹配) 结果, 第 4 层和第 5 层提供了大部分更好的

真阳性; 同样对于假阳性 FP (错误的预测), 这两个层次也提供了大部分更好的假阳性。因此, 为了减少无效信息的干扰并提高检测准确性, 本文采用了选择性查询收集模块 (Selective query recollection, SQR)。SQR 模块使得解码器的每一层都能同时接收到前两层的输出作为输入, 从而使后续层能够获取前两层的信息, 其结构如图 2 所示。

这种选择性查询收集的方法在实验中表现出很好的性能, 尤其是能够有效减轻计算负担, 该方法可以由式 (1) 表达为:

$$\begin{cases} C^0 = \{q^0\} \\ C^1 = \{q^0, q^{0-1}\} \\ C^S = \{D^S(q) | q \in C^{S-1}\} \cup \text{select}(C^{S-1}) = \\ \quad \{D^S(q) | q \in C^{S-1}\} \cup \{D^{S-1}(q) | q \in C^{S-2}\} \end{cases} \quad (1)$$

式中: q 为查询集; q^{0-1} 为两阶段细化的最终查询, 每经过一个阶段处理都会形成一个集合 C 。集合 C 中一半的查询是由当前阶段新生成的, 即来自 $\{D^S(q) | q \in C^{S-1}\}$, 而另一半来自前一阶段 C^{S-1} 。从式 (1) 可以看出, 在开始阶段 D^S 前, 从最近的两个阶段 (即 D^{S-1} 和 D^{S-2}) 收集 q 作为 D^S 的输入, 使解码器能够更有效地利用关键特征, 从而提高解码的准确性和鲁棒性。此外, 通过调整开始阶段, 可以进一步减少每个集合中的查询总数, 减轻计算负担。这种选择的灵活性使得 SQR 方法能够更好地适应不同情景和计算资源的需求。因此, 本文在解码器部分引入选择性查询收集模块, 不

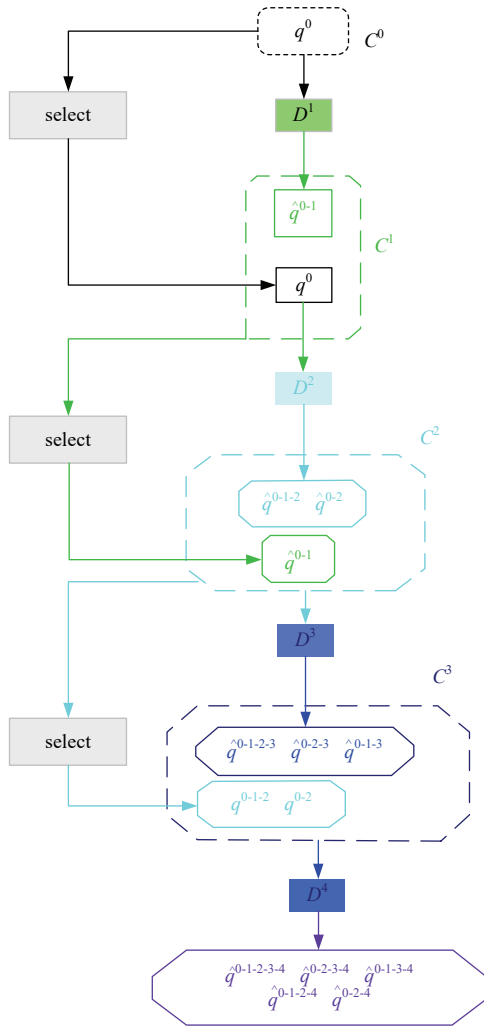


图2 选择性查询收集模块结构图

Fig. 2 Structure of selective query
recollection module

仅可以提高模型的计算效率,而且使查询收集更加精准和有效,有助于提升检测精度。

1.3 跟踪分支

如上节所述,检测分支对输入图像进行分类和回归生成检测框,跟踪分支则使用前一帧的目标信息定位当前帧的目标位置,并通过重识别网络为每个目标分配唯一 ID 以生成跟踪框。在跟踪分支中,目标的重识别是处理的核心。虽然现有方法可以通过注意力机制学习行人图像的显著性特征,但是仍缺乏对时间特征、空间特征和不同尺度特征图细节特征的统筹考虑^[17],因此, ID 重识别的性能仍然具有提升空间。

本文对基准网络中的重识别模块进行了改进,采用“检测器+ReID 跟踪器”的方式保持联合检测和跟踪范式。在解码器的独立 ReID 通道中,引入了金字塔时空聚合 PSTA 模块,通过分层聚

合建立长期依赖关系,以提高行人重识别性能。同时,结合注意力金字塔 APNet 模块,将特征图拆分为多个局部部分,并学习每个部分的注意力图,以有效捕捉不同尺度的特征图^[10]。

PSTA 模块通过逐步聚合帧级特征和分层时间特征融合,从而提供更丰富和互补的时空信息。其采用时空聚合模块 (Spatial-temporal aggregation module, STAM) 进一步聚合相邻特征。STAM 包含空间参考注意 (Spatial reference attention, SRA) 和时间参考注意 (Temporal reference attention, TRA), SRA 探索帧内的空间相关性, TRA 利用相邻帧之间的相关性增强判别特征并抑制干扰特征,从而提取更多的判别信息并解决遮挡问题。与此同时,为了更好地适应不同尺度的特征图, PSTA 融合了 APNet 模块,将输入的特征图分割为多个局部部分,并学习每个部分的注意力图,从而提供更强的表示能力,适应不同尺度、遮挡程度的特征图。这种结合相互增强,使得多目标跟踪算法在应对遮挡、丢失或切换等挑战时表现更出色。

重识别整体流程描述如下:首先,轨迹集通过特征提取得到一组特征图集合 $F' = \{F_1^0, \dots, F_T^0\}$ 。将相邻的特征图 $\{F_{2t-1}^0, F_{2t}^0 | i \in \{1, 2, \dots, T/2\}\}$ 传递给时空聚合模块 (STAM), 用于聚合局部时间特征,直到 STAM 输出一个特征图, $F_t^{n'}$ 生成的具体公式为:

$$F_t^{n'} = \text{STAM}^n(F_{2t-1}^{n-1}, F_{2t}^{n-1}), t = 1, 2, \dots, t/2^n \quad (2)$$

STAM 由空间参考注意力 SRA 和时间参考注意力 TRA 组成。将 STAM 输出的特征图 F_t', F_{t+1}' 经过 APNet 分解为局部部分,然后局部注意力图合并为全局图像的注意力,从而捕捉不同尺度的特征图。图 3 为 APNet 结构,通过将注

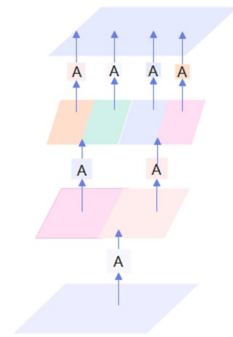


图3 APNet 结构图

Fig. 3 APNet Structure diagram

意图从粗到细堆叠为金字塔结构,从而将来自不同粒度的线索聚合。如图 3 所示,经过 APNet 处理后得到的特征图 F_t, F_{t+1} , 分别送入 SRA 和 TRA, 生成对应的注意力图 $A_t^n, A_{t+1}^n, A_{t,t+1}^n, A_{t+1,t}^n$ 。其中, A_t^n 为 SRA 对 F_t^n 生成的注意力图, $A_{t,t+1}^n$ 为 TRA 对输入 F_t^n 和其参考 F_{t+1}^n 生成的注意力图。最后,通过将注意力图与

相应的特征图按元素相乘,得到细化的特征图。这种架构充分利用帧内信息,增强了前景特征,并通过帧间相关性来增强目标人物的信息^[9]。

融合 APNet 和 STAM 的架构如图 4 所示,本文将命名为 ASTAM。这样的设计更好地应对了不同程度的遮挡,实现了 ID 的准确识别。

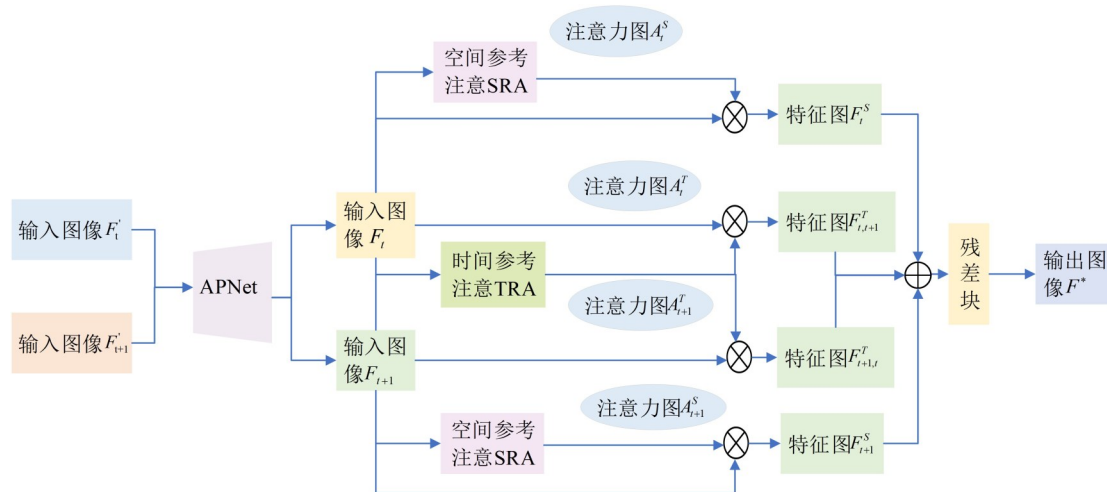


图 4 ASTAM 结构图

Fig. 4 ASTAM structure diagram

1.4 数据关联

传统的 MOT 方法通过建立外观模型和运动模型跟踪每个目标,但是这种方法难以区分外观相似的行人。并且如果直接删除严重遮挡的行人会导致大量目标丢失,影响多目标跟踪的准确性。

因此,针对上述问题,本文引入双线性 LSTM,使所有轨迹共享相同的 LSTM 网络,其外观将在内存中动态更新,用于在跟踪过程中存储状态信息,从而在外观相似时提高匹配可靠性;同时联合二次关联策略,将无重叠的检测框分为高分和低分的目标检测框^[18],其中低分框的图像不会被删除,而是与无匹配对象的轨迹进行关联,提高数据关联性能,进而提高多目标跟踪准确性。

二次数据关联的工作方式如下:在跟踪开始后,使用第一帧的检测初始化轨迹,同时为每个轨迹生成并存储 BLSTM 内存;在下一帧,计算每个轨迹和检测之间的相似度,然后采用二次数据关联方法进行关联。具体的关联过程:首先将高分检测框与轨迹相关联,对于那些无法匹配的轨迹(通常是由于遮挡、运动模糊或大小变化等情况),则将低分检测框与之关联,以便找回低分检测框中的目标同时滤除背景。在此过程中,轨迹

会根据数据关联的结果不断更新内存,如果轨迹遗漏太多检测结果或连续遗漏检测数量达到一定阈值,轨迹将被终止。被终止的轨迹将不再用于数据关联,也不会用于计算非目标对象存储器。上述过程在整个多行人跟踪中重复进行。

数据关联框图如图 5 所示,整体步骤为:首先读取输入视频序列 T , 利用高效存储的 BLSTM 存储历史轨迹;遍历视频序列的每一帧图片并且初始化预测检测框、高分检测框、低分检测框,如果检测框分数大于 T_{high} , 将其放入高分检测集合 D_{high} , 否则当检测框分数大于 T_{low} 小于 T_{high} 时,将其放入低分检测集合 D_{low} ;使用卡尔曼滤波器 KF 预测 T 中每个轨迹的新位置,使用 IoU 将检测框 D_{high} 和轨迹的预测框 T 进行匹配,将未匹配的检测结果保存至 D_{remain} , 将未匹配的轨迹保存至 T_{remain} , 再次使用 IoU 匹配低分数检测框 D_{low} 和剩余轨迹 D_{remain} , 将未匹配的轨迹保存至 T_{remain} , 删除所有未匹配的低置信度的检测框。对于第二次关联后保留的无匹配轨迹 T_{remain} , 将其放入 T_{lost} 中。对于 T_{lost} 中的每个轨迹,只有当它存在超过 30 帧时,才会将其从轨迹 T 中删除。否则,保留 T 中丢失的轨迹 T_{lost} 。遍历第一次关联后从不匹配的高分检

测框 D_{remain} 中的检测框,如果置信度大于 x ,那么初始化一个新轨迹。

1.5 损失函数

本文应用集合预测损失监督分类和回归的检测框和跟踪框,这种损失函数能够产生预测和真实对象之间的最优二分匹配。同时引入了注意力分散损失 ADL^[19],以提高不同标记(即注意力图中的行)所接收的注意力的多样性。为此,计算每一层中行之间的注意力相似并在训练过程中将其最小化。定义匹配损失为:

$$L = \lambda_{\text{cls}} \cdot L_{\text{cls}} + \lambda_{L1} \cdot L_{L1} + \lambda_{\text{giou}} \cdot L_{\text{giou}} + \lambda_{\text{ADL}} \cdot L_{\text{ADL}} \quad (3)$$

式中: L_{cls} 为预测分类和回归标签的损失; L_{L1} 和 L_{giou} 分别为归一化中心坐标与预测框和真实框的高度和宽度之间的 L_1 损失和广义 IoU 损失; L_{ADL} 为注意力分散损失; λ_{cls} 、 λ_{L1} 和 λ_{giou} 、 λ_{ADL} 为每个分量的系数,最后的损失是由训练批中的对象数量归一化的所有对的总和。

2 实验分析

本部分将详细介绍实验的具体设置、实现细节、使用的数据集以及相关评价指标。同时在 MOT17 测试集上对本文提出的多行人跟踪方法的各个模块进行消融分析,以验证其有效性。此外,根据 MOT Challenge 上给出的结果将本文算法与其他 MOT 算法在 MOT16、MOT17 数据集上进行比较,以验证算法性能。

2.1 数据集、评价指标和实验设置

2.1.1 数据集

本文采用多行人跟踪数据集 MOT16 和 MOT17 作为测试视频,二者都由 14 个视频序列组成,但是 MOT17 数据集场景中的行人更加密集、遮挡情形较多。模型预训练采用 CrowdHuman 数据集,该数据集人群较密集,并且同时存在各种各样的遮挡。

2.1.2 算法评价指标

为了全面评估跟踪算法的性能,本文选取 MOTA、HOTA、DetA、AssA 和 IDF1 对模型性能进行评价,MOTA(多目标跟踪准确度)是一个综合性指标,可以帮助衡量跟踪算法在误检、漏检和身份切换的性能表现。IDF1(识别 F 值)是衡量每个目标跟踪框中对目标 ID 识别能力判断的指标,

判断轨迹是否属于同一目标。HOTA(高阶跟踪精度)将精确检测、关联和定位的效果平衡到一个统一的用于比较跟踪器的度量中,从而能够清楚地分析跟踪性能。DetA(检测精度)评估检测结果的准确性,而 AssA(关联精度)评估关联数据的性能。

2.1.3 实验设置

本文使用在 ImageNet 上预训练的 ResNet-50 作为骨干网络,并将从预训练中获得的权重作为模型的初始值。同时采用数据增强技术提高模型的鲁棒性。算法在 Windows 系统下运行,计算显卡为 NVIDIA GeForce RTX 3090,使用 Py-

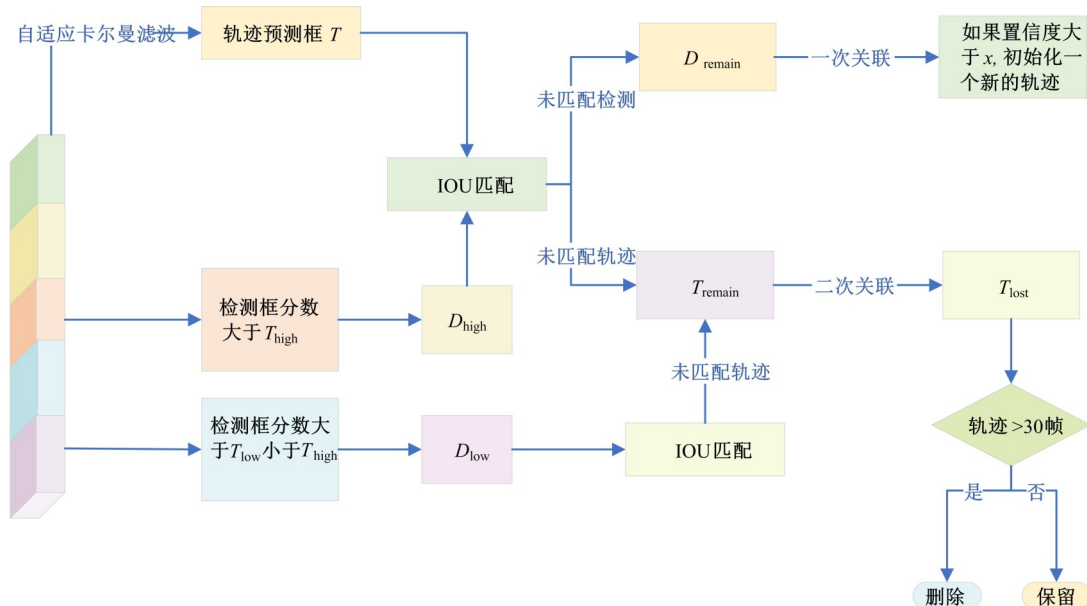


图 5 数据关联框图

Fig. 5 Data association block diagram

Torch 深度学习框架,并在 Python 3.8、CUDA 11.3 和 CuDNN 11.3 的环境下搭建和训练多目标跟踪模型。该方法在 CrowdHuman 数据集上进行模型预训练,在测试阶段,使用公开 MOT16 和 MOT17 多目标跟踪基准数据集对网络进行评估。

2.2 消融实验

为了证明本文网络的优越性和添加模块的有效性,本文在 MOT17 数据集上进行了消融实验,并对加入的模块进行了分析。将 SQR 和 BL_Byte(双线性 LSTM 和数据关联融合模块)以及融合 APNet 的 PSTA 模块(表 1 中缩写为 PA)依次添加到网络中与基准网络进行比较。表 1 中的实验结果表明:添加的单独 SQR 模块使高阶跟踪精度和检测精度分别提高了 1.57% 和 0.83%,这是因为在目标检测阶段的解码器中引入了选择性查询收集模块,增强了特征的表达能力,所以检测的精度得以提升。

表 1 添加了 SQR、BL_Byte 和 PA 在 MOT17 数据集上的变化情况

Table 1 Added changes in SQR with BL_Byte and PA on the MOT17 dataset

算 法	MOTA	HOTA	AssA	DetA	IDF1
基准	66.86	49.18	46.07	53.14	60.79
+SQR	67.43	50.75	48.20	53.97	61.57
+SQR+BL_Byte	68.27	54.23	53.95	55.07	67.02
+SQR+BL_Byte+PA	68.77	55.15	55.66	55.10	68.94

同时添加两个模块可以使训练后的网络具有更好的跟踪效果,使跟踪准确度检测精度和关联精度分别提高了 1.41% 和 1.93%、7.88%,IDF1 也提高了 6.23%。这是因为在数据关联模块采用了双线性 LSTM 和二次数据关联模块的融合策略,不但收集了相似外观信息而且使得低分框的图像也能与不匹配的轨迹关联,从而提高了关联精度。

同时添加 3 个模块,跟踪准确度检测精度和关联精度分别提高了 1.91%、1.96%、9.59%,IDF1 也提高了 8.15%。这是因为在 ReID 分支引入了金字塔时空聚合模块 PSTA 与注意力金字塔模块 APNet 的融合结构,使网络更好地捕获了不同尺度特征图的时空信息,从而提高了 ID 识别的准确性。

通过上述消融实验可以得出以下结论:同时将这 3 个模块合并添加到网络中可以提高检测和

关联精度,进而提高多目标跟踪的准确性,实验结果证明了本文改进模块的有效性。

2.3 对比实验

本节通过定量分析的方法,将本文提出的方法与本领域内的相关研究进行性能对比,实验在公开数据集 MOT16 和 MOT17 上进行,为了保证实验的全面性和有效性,对比算法的跟踪结果采用 MOT challenge 官网公布的数据,实验对比结果如表 2 和表 3 所示。

表 2 MOT16 测试数据集实验结果对比

Table 2 Comparison of experimental results of MOT16 test dataset

算 法	MOTA	HOTA	AssA	DetA	IDF1
CTracker-V1 ^[20]	67.60	48.80	43.70	54.90	57.20
GSDT ^[21]	66.70	55.90	54.90	57.20	69.20
Tube_TK ^[22]	64.00	48.70	45.50	52.50	59.40
KDNT ^[23]	68.20	50.10	45.20	56.00	60.00
Hugmot2 ^[24]	70.20	52.10	49.50	55.30	65.40
TraDes ^[25]	70.10	53.20	50.90	56.20	64.70
CNNMTT ^[26]	65.20	49.40	47.00	52.20	62.20
SmartSORT ^[27]	60.40	46.10	41.90	51.00	56.10
TransTrack	68.84	47.41	40.93	55.53	56.60
本文	70.52	55.86	55.76	56.53	69.73

表 2 给出的是 MOT16 数据集上的对比结果,根据表 2 的对比实验结果可知,本文方法相对于 GSDT 在 DetA 指标上稍微逊色一点,但在 MOTA 指标上提高了 3.82%,因为本文改进了数据关联和重识别模块,很好地提高了多目标跟踪准确度。实验结果表明:本文提出的模块能够实现更准确的多行人跟踪。

表 3 给出的是 MOT17 数据集上的对比结果,

表 3 MOT17 测试数据集实验结果对比

Table 3 Comparison of experimental results of MOT17 test dataset

算 法	MOTA	HOTA	AssA	DetA	IDF1
CTracker-V1 ^[20]	66.60	49.00	45.20	53.60	57.40
GSDT ^[21]	66.20	55.50	54.80	56.40	68.70
Tube_TK ^[22]	63.00	48.00	45.10	51.40	58.60
CJTracker ^[28]	58.70	48.40	48.00	49.10	58.20
Hugmot2 ^[24]	68.80	51.50	49.40	54.10	64.60
TraDes ^[25]	69.10	52.70	50.80	55.20	63.90
MENDER ^[29]	65.00	53.90	54.40	53.60	67.10
QuasiDense ^[30]	68.70	53.90	52.70	55.60	66.30
CTTrack17 ^[31]	67.80	52.20	51.00	53.80	64.70
TransTrack	66.86	49.18	46.07	53.14	60.79
本文	68.77	55.15	55.66	55.10	68.94

根据表3的对比实验结果可知,尽管MOT17行人更加密集,并且由于遮挡导致了漏检和关联缺失,但本文方法在大部分指标上仍然取得了不错的成绩。本文的方法在IDF1指标上超过了TraDes 5.04%,展示了本文提出的改进重识别模块的有效性。

从表2、表3的实验结果可知,本文方法在多个指标上表现出色,尽管在某些指标上稍低于其他算法的最高得分,但综合考虑,本文方法在多目标跟踪任务中具有更大的优势,能够在复杂的拥挤场景中准确地识别和追踪目标。

2.4 结果可视化

本节通过对比本文算法与基准算法的可视化结果分析算法的有效性,并展示了算法改进前后MOT17数据集的跟踪效果图,结果如图6所示。



图6 MOT17-02可视化图像

Fig. 6 MOT17-02 Visualization image

在图6的MOT17-02视频序列中,紫衣行人遮挡了后面的红衣行人。从图6中可以看出,左面未加改进的数据关联模块之前丢失了目标,而右面加入了改进的数据关联后则能够保持跟踪轨迹,即使红衣行人在两三个行人的遮挡下仍然能够保持跟踪。这充分说明加入改进数据关联模块后,低分框的图像也能与不匹配的轨迹进行关联,从而在行人受到遮挡时仍然能够被准确追踪。

通过分析MOT17数据集的可视化结果,证明了本文提出的模块在面对遮挡时的有效性。此外,本文还对MOT16数据集进行了可视化分析,图7给出了MOT16数据集上的测试效果。

在图7的MOT16-11视频帧中,浅蓝色衣服



图7 MOT16-11可视化图像

Fig. 7 MOT16-11 Visualization image

的行人被遮挡,未对重识别模块改进之前算法无法继续跟踪该目标,而在算法加入了改进的重识别模块后可以更好地捕捉不同尺度特征图,使得小尺寸的目标也可以被检测到,保持了跟踪轨迹的连续性,减少了发生漏检的情况。

通过对MOT16和MOT17数据集进行可视化分析,展示了本文改进的模块在应对遮挡和漏检等问题时仍能够准确检测到行人,直观地展示了本文提出的改进网络在复杂街道场景下对密集行人追踪的有效性。

3 结束语

本文针对密集行人场景下多目标跟踪存在的目标漏检、错检、关联不准确以及重识别错误等问题,提出了一种基于Transformer的双分支检测与重识别的网络结构。该方法在目标检测的解码器部分引入了选择性查询收集模块,帮助解码器层有针对性地选择中间特征,使后面的层数更精准收集到前面层数的信息,在降低计算成本的同时提高目标特征的表达能力,解决了目标漏检、错检的问题;在数据关联部分,采用双线性LSTM和二次数据关联模块融合的策略,通过存储器存储每一帧图像的轨迹信息,并采用高低阈值的方式对检测框重新打分,使得低分框也能关联到不匹配的轨迹,解决了密集行人由于相似外观导致的目标混淆和严重遮挡等问题;最后,本文还设计了金字塔时空聚合与注意力金字塔融合模块,以增强空间和判别特征,同时精准提取多尺度特征图,解决了重识别错误的问题。消融实验及可视化结果证明了本文设计模块的有效性,对比实验证明了本文网络相较其他MOT算法,在多目标跟踪准确性方面取得了更好的结果。

参考文献:

- [1] 丁贵鹏,陶钢,庞春桥,等.基于无锚的轻量化孪生网络目标跟踪算法[J].吉林大学学报:理学版,2023,61(4):890-898.
Ding Gui-peng, Tao Gang, Pang Chun-qiao, et al. Anchorless target tracking algorithm for lightweight siamese network[J]. Journal of Jilin University (Science Edition), 2023, 61(4):890-898.
- [2] 徐涛,马克,刘才华.基于深度学习的行人多目标跟踪方法[J].吉林大学学报:工学版,2021,51(1):27-38.

- Xu Tao, Ma Ke, Liu Cai-hua, et al. Multi-object pedestrian tracking based on deep learning[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(1): 27-38.
- [3] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//IEEE International Conference on Image Processing (ICIP), Beijing, China, 2017: 3645-3649.
- [4] 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: 3069-3087.
- [5] 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, 2022, 45(6): 7820-7835.
- [6] Zhou X, Yin T, Koltun V, et al. Global tracking transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8771-8780.
- [7] Cai J, Xu M, Li W, et al. Memot: multi-object tracking with memory[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8090-8100.
- [8] Chen F, Zhang H, Hu K, et al. Enhanced training of query-based object detection via selective query recollection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, Canada, 2023: 23756-23765.
- [9] Wang Y, Zhang P, Gao S, et al. Pyramid spatial-temporal aggregation for video-based person re-identification[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, Canada, 2021: 12026-12035.
- [10] Chen G, Gu T, Lu J, et al. Person re-identification via attention pyramid[J]. IEEE Transactions on Image Processing, 2021, 30: 7663-7676.
- [11] Kim C, Li F X, Alotaibi M, et al. Discriminative appearance modeling with multi-track pooling for real-time multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 9553-9562.
- [12] Zhang Y, Sun P, Jiang Y, et al. Bytetrack: multi-object tracking by associating every detection box[C]//The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 1-21.
- [13] Meinhardt T, Kirillov A, Leal-Taixe L, et al. Trackformer: multi-object tracking with transformers[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, USA, 2022: 8844-8854.
- [14] Zeng F, Dong B, Zhang Y, et al. Motr: end-to-end multiple-object tracking with transformer[C]//The 17th European Conference on Computer Vision, Tel Aviv, Israel, 2022: 659-675.
- [15] Sun P, Cao J, Jiang Y, et al. Transtrack: multiple object tracking with transformer[J/OL]. [2023-11-20]. <https://arxiv.org/pdf/2012.15460>.
- [16] Zhu X, Su W, Lu L, et al. Deformable detr: deformable transformers for end-to-end object detection[J/OL]. [2023-11-21]. <https://arxiv.org/pdf/2010.04159>.
- [17] 庄珊娜, 王君帅, 白晶, 等. 基于三维卷积与自注意力机制的视频行人重识别[J]. 吉林大学学报: 工学版, 2025, 55(7): 2409-2417.
- Zhuang Shan-na, Wang Jun-shuai, Bai Jing, et al. Video-based person re-identification based on three-dimensional convolution and self-attention mechanism[J]. Journal of Jilin University (Engineering and Technology Edition), 2025, 55(7): 2409-2417.
- [28] 涂淑琴, 黄正鑫, 梁云, 等. 改进 TransTrack 多目标生猪行为跟踪方法[J]. 农业工程学报, 2023, 39(15): 172-180.
- Tu Shu-qin, Huang Zheng-xin, Liang Yun, et al. Improvement of the TransTrack multi-objective hog behavior tracking method[J]. Transactions of the Chinese Society of Agricultural Engineering, 2023, 39(15): 172-180.
- [19] Guo Y, Stutz D, Schiele B. Robustifying token attention for vision transformers[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, Paris, France, 2023: 17511-17522.
- [20] Peng J, Wang C, Wan F, et al. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//Computer Vision-ECCV 2020: 16th European Conference, Glasgow, UK, 2020: 145-161.
- [21] Wang Y, Kitani K, Weng X. Joint object detection and multi-object tracking with graph neural networks[C]//IEEE International Conference on Robotics and Automation(ICRA), Xi'an, China, 2021: 13708-13715.
- [22] Pang B, Li Y, Zhang Y, et al. Tubetk: adopting

- tubes to track multi-object in a one-step training model[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, USA, 2020: 6307-6317.
- [23] Yu F, Li W, Li Q, et al. Poi: multiple object tracking with high performance detection and appearance feature[C] //European Conference on Computer Vision: amsterdam, The Netherlands, 2016: 36-42.
- [24] Cao J, Zhang J, Li B, et al. RetinaMOT: rethinking anchor-free YOLOv5 for online multiple object tracking[J]. *Complex & Intelligent Systems*, 2023, 9(5): 5115-5133.
- [25] Wan X, Zhou S, Wang J, et al. Multiple object tracking by trajectory map regression with temporal priors embedding[C]//Proceedings of the 29th ACM International Conference on Multimedia, Chengdu, China, 2021: 1377-1386.
- [26] Wu J, Cao J, Song L, et al. Track to detect and segment: an online multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 12347-12356.
- [27] Nguyen P, Quach K G, Kitani K, et al. Type-to-track: retrieve any object via prompt-based tracking[J/OL]. [2023-11-20]. <https://arxiv.org/pdf/2305.13495v3>.
- [28] Mahmoudi N, Ahadi S M, Rahmati M. Multi-target tracking using CNN-based features: CNNMTT[J]. *Multimedia Tools and Applications*, 2019, 78(6): 7077-7096.
- [29] Meneses M, Matos L, Prado B, et al. Learning to associate detections for real-time multiple object tracking[J/OL]. [2023-11-22]. <https://arxiv.org/pdf/2007.06041>.
- [30] Pang J, Qiu L, Li X, et al. Quasi-dense similarity learning for multiple object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, USA, 2021: 164-173.
- [31] Zhou X, Koltun V, Krähenbühl P. Tracking objects as points[C]//European Conference on Computer Vision, Glasgow, UK, 2020: 474-490.