基于Transformer的跨空间特征关联多目标追踪

沈锦荣

（南京邮电大学 通信与信息工程学院，江苏 南京 210003）

摘　要：多目标追踪算法在复杂场景中常会出现目标识别不精确与追踪效果差的问题，在追踪目标数量多且相互遮蔽情况严重的环境中追踪目标丢失现象更为明显。为此，提出一种基于Transformer跨空间特征关联的多目标追踪算法，利用Transformer结构提取全局特征与多头注意力机制的优势提升目标特征的提取能力。此外，为解决追踪目标之间相互遮蔽而导致的追踪目标丢失问题，利用互注意力机制映射追踪目标与干扰目标之间的特征进行增强与抑制，以提高追踪算法的准确性和可靠性；同时基于追踪目标与干扰目标之间的特征相似度确定特征增强与抑制的程度。在MOT16与MOT17数据集上进行实验，所提算法分别取得了58.81%和60.05%的多目标追踪准确性，相较其他主流算法性能更优。

关键词：多目标追踪；Transformer；特征关联；目标检测；注意力机制

DOI：10.11907/rjdk.231278　　开放科学（资源服务）标识码（OSID）：

中图分类号：TP391　　**文献标识码：**A　　**文章编号：**1672-7800(2024)003-0165-07

Multi-Object Tracking with Cross-Spatial Feature Association Based on Transformer

SHEN Jinrong

(*School of Communications and Information Engineering*, *Nanjing University of Posts and Telecommunications*, *Nanjing* 210003, *China*)

Abstract: Multi target tracking algorithms often suffer from inaccurate target recognition and poor tracking performance in complex scenes. In environments with a large number of tracked targets and severe mutual occlusion, the phenomenon of target loss is more pronounced. To this end, a multi-target tracking algorithm based on Transformer cross spatial feature association is proposed, which utilizes the advantages of Transformer structure to extract global features and multi head attention mechanism to improve the extraction ability of target features. In addition, to solve the problem of tracking target loss caused by mutual occlusion between tracking targets, a mutual attention mechanism is used to map the features between tracking targets and interfering targets for enhancement and suppression, in order to improve the accuracy and reliability of the tracking algorithm; At the same time, the degree of feature enhancement and suppression is determined based on the feature similarity between the tracking target and the interfering target. Experiments were conducted on the MOT16 and MOT17 datasets, and the proposed algorithm achieved multi-target tracking accuracy of 58.81% and 60.05%, respectively, with better performance compared to other mainstream algorithms.

Key Words: multi-object tracking; Transformer; feature association; object detection; attention mechanism

0　引言

随着视频处理技术与移动终端的快速发展，视频逐渐取代文字成为主流的信息传输载体，被广泛应用于机器人、监控和自动驾驶领域。多目标追踪的目的是在目标数量及运动轨迹尚不可知的情况下对连续视频帧中不断运动的潜在目标进行位置检测，再赋予其ID进行追踪，以预测多个目标的运动轨迹并精确查找到所需目标。多目标追踪技术在视频处理方面作用巨大，受到机器视觉领域学者们的广泛重视。

目前多目标追踪算法通常采用卷积神经网络（Convo-

收稿日期：2023-03-17

专利编号：202310175401X

作者简介：沈锦荣（1997-），男，南京邮电大学硕士研究生，研究方向为计算机视觉、多目标追踪。

lutional Neural Networks，CNN）架构，其在特征提取时无法从全局特征的视角关联图像特征，当目标被遮蔽时往往会提取到其运动轨迹上其他目标的特征，造成目标边界框漂移，目标预测位置亦会开始跟随相邻对象。因此，CNN 在预测目标运动轨迹时无法有效应对遮蔽问题，这种错误的边界框预测将会在连续视频帧中传播，使得运动轨迹的预测损害关联嵌入。此外，复杂场景通常包含多个运动目标与复杂多变的运动状态，例如目标尺度变化、目标之间相互遮蔽、目标之间特征相似度高等，该类场景下多目标跟踪算法的准确率一般无法保证。

为解决以上问题，本文提出一种基于 Transformer 的跨空间特征关联多目标追踪算法，利用 Transformer 的全局视野与多头注意力机制改善目标的特征提取能力，并基于互注意力机制对目标特征加以映射，从而加强目标匹配关联、提高目标追踪准确度。

1 相关研究

1.1 基于检测的多目标追踪

基于检测的多目标追踪采用成熟的目标检测算法在每一帧视频中检测出追踪目标，然后通过数据关联在连续视频帧中连接目标运动轨迹，主要包括两阶段方法和一阶段方法两种。

1.1.1 一阶段方法

该类方法可以看作是多任务学习在多目标追踪领域的应用，主要特点为特征提取网络同时服务于目标检测与目标追踪两项任务，将二者统一起来，从而达到追踪速度与精度的平衡。例如，Wang 等[1]将目标检测环节与外观特征信息提取环节融合设计为一个网络，采用 anchor-based 方法首先输出目标检测框，然后在目标检测网络的基础行添加嵌入学习模型[2]；Zhang 等[3]采用 anchor-free 的目标检测方式，以高分辨率图像中的像素点为中心提取特征；Zhou 等[4]提出的 CenterTrack 是以 CenterNet[5]为基础的改进算法，其包括当前帧图像、前一帧图像和前一帧预测的 heatmap 3 个输入，对目标在时间上进行关联匹配。

1.1.2 两阶段方法

该类方法将多目标追踪分为两个步骤进行：首先在每个视频帧中检测追踪目标，然后通过数据关联跨帧连接形成运动轨迹，通常采用身份嵌入区分关联对象。例如，Bewley 等[6]首先使用 Faster R-CNN[7]对视频帧进行目标检测；然后利用卡尔曼滤波器[8]预测目标运动轨迹，从而对检测框进行跟踪；最后以交并比作为匹配标准，通过匈牙利算法对检测框与追踪框进行匹配。Wojke 等[9]设计了级联匹配模块，其基于深度卷积网络在交并比匹配前提取并保存目标图像的深层特征，缓解了多目标相互遮挡而导致的追踪目标丢失问题；Zhang 等[10]提出的 ByteTrack 利用检测框与跟踪轨迹之间的相似性保留高置信度的检测结果，在低置信度的检测结果中去除背景，挖掘出追踪目标更具辨识度的图像特征，从而避免了目标丢失，提高了轨迹连贯性。

以上基于检测的多目标追踪方法基本都采用 CNN 进行特征提取与关联匹配，但 CNN 仅具有局部感受野，只能对特征进行局部提取与目标匹配关联，在复杂场景中，该类方法无法将目标的全局特征关联起来并从中找出关键特征，目标跟踪效果不佳。而本文算法采用 Transformer 结构提取目标的全局特征，利用注意力机制找出关键特征，强化了目标的特征提取环节。在关联匹配环节，本文采用 Transformer 结构的互注意力机制将目标的身份特征映射到目标原图像特征之上，与常规方法相比可以在全局范围内更准确地映射目标特征，并在此过程中寻找到目标关联匹配之间的关键特征。

1.2 基于孪生网络的多目标追踪

基于孪生网络的多目标追踪算法使用参数相同的两个神经网络提取两个输入图像的特征，其中一个作为基准模板，另一个作为要选择的样本。该方法通过对模板与样本提取到的不同特征进行互相关操作得到相似度最高的区域，从而确定追踪目标的位置。例如，Shuai 等[11]通过估计相邻帧之间目标的移动关联检测目标，在取得较高准确性的同时具有一定的实时性。

孪生神经网络在单目标追踪任务中取得了重大突破，但其运用于多目标追踪任务中时有着难以解决的问题。孪生神经网络会先保存目标特征，再在视频帧中通过相似度计算预测位置，因此当视频帧中出现新目标时，其特征并不在孪生神经网络预先存储的特征中，此时便会检测错误。如果想要解决这个问题，就需要在前后视频帧中提取目标特征加以更新。然而为保证追踪的准确性，基于孪生网络的多目标追踪算法需要设计非常复杂的网络结构，随着视频帧中追踪目标数量的增长，网络参数量会呈指数增长。

本文算法可在长时间目标追踪过程中找出目标的关键特征并加以保存，同时基于互注意力机制可更好地进行目标匹配关联，减少多目标追踪过程中的误检与漏检现象。此外，基于 Transformer 的网络结构不需要构建孪生网络，在模型结构上更简单。

1.3 Transformer

Transformer 结构最早被运用于自然语言处理领域，并取得了极好的效果[12]。其全注意力的结构不仅增强了特征提取能力，而且保持了并行计算的优势，可以又快又好地完成自然语言处理领域内几乎所有的下游任务，极大推动了该领域的发展。例如，Carion 等[13]、Zhu 等[14]率先在计算机视觉领域采用 Transformer 结构，其采用一种特殊的 query-key 机制，将目标检测视为一个集合检测问题，将 CNN 提取到的特征输入编码器模块得到 key 与 value，并输入到解码器模块；将解码器模块的 query 定义为一个可学

习的参数,由解码器输出检测结果;Dosovitskiy 等[15]率先在计算机视觉领域提出纯Transformer结构的网络ViT,其首先将输入图片切分重排,输入全连接层进行矩阵乘法;然后输入到编码模块;最后输入全连接层得到最后的分类得分。图1为ViT的一个子模块。对于维度为(C, H, W)的特征,通过全连接层后维度变换为(C, H, 3 * inner*dim)。式中,inner_dim=heads * head_dim;heads为多头注意力机制的头数,决定了多头注意力机制的注意力可能性;head_dim为设定好的为每个注意力头分配的维度数。特征经过维度变换后与位置编码相加,然后通过编码器得到子模块的输出特征。

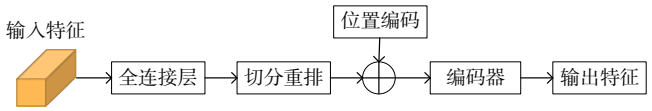

Fig. 1　ViT子模块

图1　ViT子模块

后续研究在ViT的基础上推出了改进模型PVT(Pyramid Vision Transformer)[16]与Swin Transformer[17]。PVT摒弃了ViT中使用的直筒型网络结构,借鉴特征金字塔结构将特征图分为大小不同的块,对每块执行Max Pooling,最后拼接起来。Swin Transformer则使用分层结构提取不同维度的特征图,提出window multiscale self attention,将注意力机制的计算限制到同一个窗口内,降低了计算复杂度;同时利用相对编码赋予特征图中的像素点以位置信息。

1.3.1　位置编码

位置向量会对每个位置编号,每个编号对应一个位置向量。将特征图的特征信息与位置编码提供的向量信息相结合,重新赋予切分重排后丧失位置信息的图像特征以位置信息,编码器可以通过位置编码分辨出不同位置的图像特征。位置编码为编码器提供了全局特征的位置信息,其计算公式为:

$$PE\left(pos,\ 2i\right)=\sin\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right) \tag{1}$$

$$PE\left(pos,\ 2i+1\right)=\cos\left(\frac{pos}{10\,000^{\frac{2i}{d_{model}}}}\right) \tag{2}$$

1.3.2　自注意力机制

注意力机制为Transformer结构的核心。作为Transformer独有的一种注意力机制,自注意力机制可以跨越长距离地关联特征图中不同区域的特征,从大量图像特征中自动分辨出关键特征,并将算力聚集到这些关键特征上。自注意力机制网络复杂性小,采用的矩阵乘法可以保证并行计算。其计算公式为:

$$Attention\left(Q,\ K,\ V\right)=softmax\left(\frac{QK^{T}}{\sqrt{d_k}}\right)V \tag{3}$$

式中:Q、K和V均由模块的输入特征经过矩阵乘法得到。由Q与K的转置进行点积,得到注意力得分,然后对分数取模,除以$\sqrt{d_k}$,以确保梯度稳定,避免由于点积得到的注意力得分过大而导致经过softmax后梯度过小,从而有碍于反向传播。然后采用softmax激活函数对注意力得分进行归一化操作,点乘V得到每个输入向量的加权得分。自注意力机制的缩放点积模块结构如图2所示。

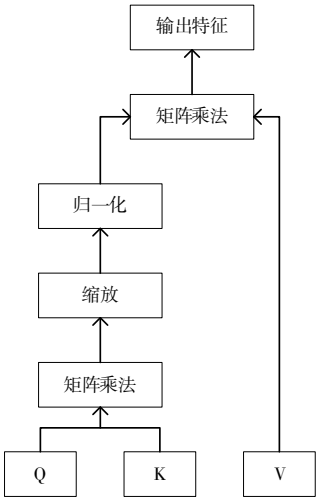

Fig. 2　Self-attention scaled dot-product block structure

图2　自注意力缩放点积模块结构

1.3.3　多头注意力机制

Transformer使用的多头注意力机制为自注意力机制的一种应用,可以对相距很远但存在相关性的特征之间进行关联,是一种非局部信息统计的注意力机制。其结构中存在多个相乘矩阵,可对模块输入的特征图进行多次自注意力机制运算从而获取多个Q、K和V,从而使输入信息获得更多注意力可能性,对输入特征进行更多样化的相似度计算,将最后结果拼接到一起,将某个特征于更多地与其他特征关联起来。多头注意力机制计算公式为:

$$Q^{i}=QW_{i}^{Q},\ K^{i}=KW_{i}^{K},\ V^{i}=VW_{i}^{V} \tag{4}$$

$$head_{i}=Attention\left(QW_{i}^{Q},KW_{i}^{K},VW_{i}^{V}\right) \tag{5}$$

$$MultiHead\left(Q,\ K,\ V\right)=Concat\left(head_{i},\ \cdots,head_{n}\right)W^{0} \tag{6}$$

式中:W为权重矩阵。

本文算法结合了CNN与Transformer结构的优点,利用CNN高效提取局部特征,利用Transformer结构提取全局特征。相较于纯Transformer结构的网络,本文算法参数量更少,不需要大量数据集与训练时间便可取得较好性能。此外,CNN可以提供位置信息,不需要额外添加位置编码提供一种抽象的数据型位置信息。

2　本文算法

本文提出的多目标追踪算法包括特征提取与关联匹配两部分。针对多目标在复杂场景,尤其是多目标之间相互遮蔽严重的场景下追踪困难的问题,将Transformer结构与CNN相结合,提出目标特征提取网络,强化特征提取阶

段；基于 Transformer 结构提出跨空间特征强化与抑制模块，强化关联匹配环节；基于追踪目标与干扰目标之间的特征相似度计算自适应参数，从而控制特征强化与抑制的程度。

2.1 基于 Transformer 与 CNN 的目标特征提取网络

以往多目标追踪网络结构多基于传统 CNN，但传统 CNN 是基于物理世界的归纳偏置，其局部连接的特点导致其只能拥有有限的感受野，扩展感受野只能依靠卷积层数的不断堆叠，以及特征提取的不断深入，从而提取全局信息，直至覆盖整幅图像。与传统 CNN 不同，Transformer 具有全局视野，其全注意力机制结构可以基于特征之间的相似度提取全局特征的上下文信息，跨空间地关联起不同区域的特征，从而自动在特征图中找出更加具有表征性的特征，有利于在长时间且复杂的环境下提升多目标追踪的准确性。然而 Transformer 结构被用于计算机视觉领域以来一直都存在网络参数量大、训练时间长和需要大量训练数据等问题。因此，本文提出一种融合 Transformer 与 CNN 的特征提取网络，将 Transformer 的全局视野和自注意力机制与 CNN 提取局部特征的高效性相结合。结构如图 3 所示。

在特征提取阶段，输入图像特征首先经过点卷积降低特征维度，从而降低运算复杂度；然后将点卷积输出分为两个支路，一个支路输入深度可分离卷积提取局部特征，另一个支路输入 Transformer 结构提取全局特征。将经过深度可分离卷积的局部特征通过全连接层调整至与全局特征相同的特征维度，然后与全局特征相加并经过前馈网络融合。前馈网络包括两个全连接层，中间有一个 ReLU 激活层，其可对局部特征与全局特征进行线性转换；最后将深度可分离网络的输出与融合后的特征拼接，通过点卷积融合。由于局部特征提取模块的输出保持了图像的位置信息，本文提出的目标特征提取网络不需要在 Transformer 结构中提供位置信息。

2.2 基于注意力机制的跨空间特征增强与抑制模块

多目标追踪任务的目标检测阶段会在当前帧中得到每个目标的检测框，在之后的目标追踪阶段会基于上一帧该追踪目标的检测框提取目标特征，并根据提取到的特征进行关联匹配以确定在新一帧中追踪目标的运动轨迹。当追踪目标周围存在其他干扰目标，并且追踪目标与干扰目标的检测框存在较大重叠时，就会难以预测追踪目标的运动轨迹。假设干扰目标遮挡在追踪目标之前，那么根据检测框提取到的特征将会有很多来自于干扰目标，这将使得追踪目标的运动轨迹预测更加偏向于干扰目标，这也是严重遮蔽环境会导致追踪目标丢失的主要原因。为克服这种偏移趋势，本文提出一种基于 Transformer 的跨空间特征关联模块。该模块会增强遮蔽情况下追踪目标未被遮蔽的区域，同时抑制追踪目标被遮蔽区域的特征，以更好地保留追踪目标在长时间追踪过程中更具表征性的特征，减小在严重遮蔽情况下干扰目标对追踪目标的不良影响。

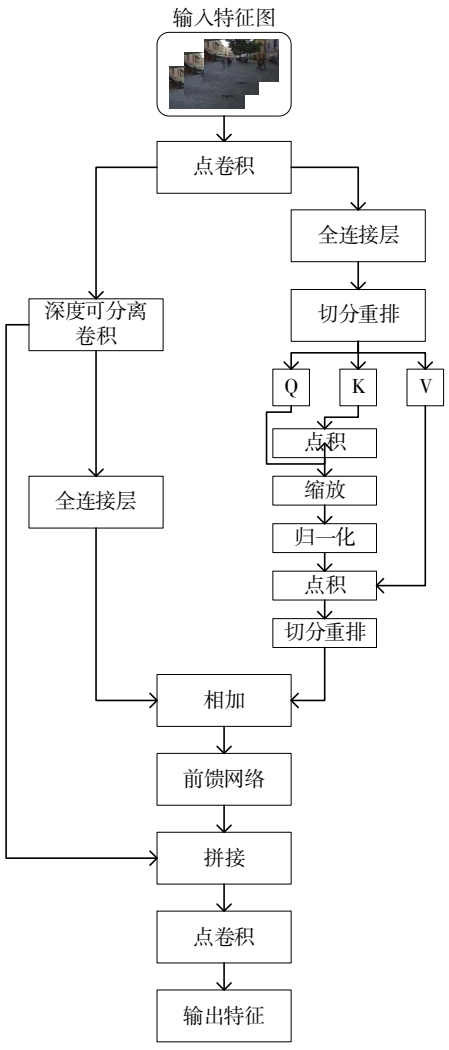

Fig. 3 Target fusion feature extraction network

图 3 目标融合特征提取网络

基于 Transformer 的跨空间特征关联模块会在与追踪目标发生重叠的其他追踪目标中选择一个 IoU 最大的对象作为干扰目标，对于追踪目标与干扰目标，利用注意力机制映射，找到被遮蔽的特征与未被遮蔽的特征，并对目标追踪器内保存的身份特征进行抑制与增强，得到更具表征性的特征，从而增强追踪器中目标特征记忆模块的鲁棒性。

互注意力机制类似于自注意力机制，不同点在于互注意力机制多出了数据交互模块，该模块多头注意力机制中的 K、V 来自于编码器输出，Q 来自于其他输入特征。本文提出的基于 Transformer 的跨空间特征关联模块采用注意力机制映射追踪目标与干扰目标的特征。结构如图 4 所示。

通过 Transformer 结构的注意力机制在当前帧中提取追踪目标的特征生成 K 和 V，将当前帧追踪目标的身份特征作为 Q 输入到解码器。基于互注意力机制关联特征图上下文信息，得到当前特征图中的关键特征作为增强特征。同理，在当前帧中提取干扰目标的特征生成 K 和 V，将

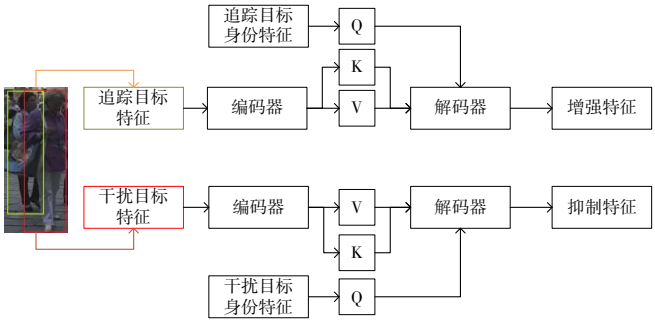

Fig. 4　Cross-spatial feature association module

图 4　跨空间特征关联模块

当前帧干扰目标的身份特征作为 Q 输入解码器,基于互注意力机制找到干扰目标在追踪过程中的关键特征,从而找出追踪目标被遮蔽的特征作为抑制特征。将增强特征加到目标追踪器存储的目标特征中以增强追踪目标未被遮蔽的特征。同时,目标追踪器中存储的目标特征减去抑制特征以减弱追踪目标被遮蔽的特征。通过这种方式,目标追踪器会加强追踪目标更具表征性的特征,同时削弱可能导致目标追踪器偏移的附近干扰目标的特征。

2.3　基于相似度的特征增强与抑制自适应参数

在得到目标相互遮蔽时未被遮蔽与被遮蔽的特征后,针对追踪目标与干扰目标之间的特征相似度对追踪目标特征进行增强与抑制。这一过程将不可避免地改变目标原本的图像特征,如果对目标特征增强或抑制得太过会加大追踪器进行目标匹配的难度,因此本文提出一种基于相似度的特征增强与抑制方法,根据追踪目标与干扰目标之间的特征相似度决定特征增强与抑制的程度。对于特征相似度大的追踪目标与干扰目标,追踪器在进行匹配时往往难以分辨,容易造成目标丢失或偏移,因此需要加大增强与抑制的程度;对于特征相似度小的追踪目标与干扰目标,追踪器往往可以正确匹配,因此应该降低增强与抑制的程度,从而保证目标特征的完整性。

为解决以上问题,本文采用自适应权重 weight。计算公式为:

$$weight = \frac{\vec{A}\cdot\vec{B}^{T}}{\left|\vec{A}\right|\cdot\left|\vec{B}\right|} + 0.3 \tag{7}$$

式中:A 为追踪目标的特征图,B 为干扰目标的特征图,$|A|$与$|B|$表示对 A 与 B 取模。在对目标特征进行增强与抑制时,跨空间特征关联模块输出的增强特征与抑制特征会受到该权重的自适应调节,即在追踪目标与干扰目标难以区分的情况下会加大特征增强与抑制的程度,而在追踪目标与干扰目标容易区分的情况下会适当降低特征增强与抑制的程度。

3　实验方法与结果分析

3.1　实验环境

服务器操作系统为 64 位 Ubuntu 20.04,显卡型号为 NVIDIA GeForce RTX 2070 开启 GPU 加速,CPU 为 Intel® CoreTM i7-8700K CPU @3.70GHZ,内存 32 GB。网络模型搭建使用 Python3.6 编程语言、Pytorch 1.3.1 深度学习框架。选择多目标追踪领域常用的性能测试公开数据集 MOT16 与 MOT17[18]。

在 MOT 数据集上训练网络之前,首先在 COCO 目标检测数据集上预训练骨干网络 ResNet101[19]、目标检测网络 RPN(Region Proposal Network)与 PAN(Pyramid Attention Network)[20]。批尺寸设置为 2,使用 SGD 作为优化器,先进行 3 轮预训练加快收敛,其中学习率设为 1×10^{-2}。最大训练轮次设置为 50 次,骨干网络的学习率为 2×10^{-5},其他部分网络学习率为 2×10^{-4}。

3.2　评价指标

为验证本文算法性能,使用 CLEAR MOT 相关指标(MOTA ↑、MT ↑、ML ↓、IDF1 ↑、FP ↓、FN ↓、IDS ↓)进行评价,“↑”表示数值越高越好,“↓”表示数值越低越好。

MOTA 为多目标追踪精确度,通常作为衡量多目标追踪的综合指标。计算公式为:

$$MOTA = 1 - \frac{\sum_t \left(FN_t + FP_t + IDSW_t\right)}{\sum_t GT_t} \tag{8}$$

式中:t 表示时刻,GT_t 表示 t 时刻视频帧中所有追踪目标的数量,$IDSW_t$ 表示 t 时刻出现 ID 切换的次数,FN_t 表示 t 时刻的漏检次数,FP_t 表示 t 时刻的误检次数。

MT 表示成功跟踪的帧数占总视频帧数 80% 以上的轨迹数量;ML 表示成功跟踪的帧数占总视频帧数 20% 以下的轨迹数量。

IDF1 衡量了多目标追踪算法在追踪目标 ID 分配方面的性能。计算公式为:

$$IDF1 = \frac{2*TP}{2*IDTP + IDFP + IDFN} \tag{9}$$

式中:TP 表示正确检测的目标数量。

FP 表示误检次数,即预测存在但事实上不存在的次数;FN 表示漏检次数,即没有被检测出来但实际上存在的次数;IDS 表示在连续视频帧中进行多目标追踪时跟踪目标 ID 的切换次数,该指标可以用于衡量多目标追踪算法的稳定性。

3.3　性能验证实验

图 5 为在追踪目标之间相互遮蔽情况下本文算法的追踪效果。图中共有 4 个跟踪对象,图 5(左)中 ID=8 与 ID=27 的对象正在迎面走来,被红色框标注;图 5(中)中二者发生严重重叠,ID=27 的对象出现跟踪丢失的现象,只剩下 ID=8 的跟踪框;图 5(右)中 ID=8 的对象与 ID=27 的对象分开,二者跟踪框恢复,没有发生误检或漏检的问题。

3.4　比较实验

选择 MOTDT、DeepMOT[21]、Tracktor++V2、GSM 算法作为对照,在 MOT16 数据集上与本文算法进行性能比较。结果见表 1。

选择 MOTDT[22]、FAMNET[23]、Tracktor++V2[24]、GSM

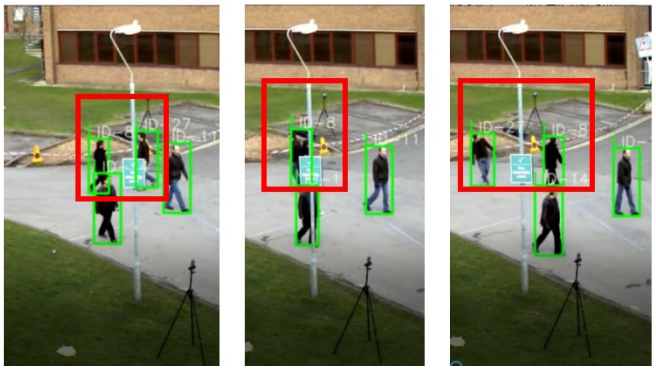

Fig. 5 The tracking effect of the proposed algorithm

图 5 本文算法追踪效果

Table 1 Performance comparison of various algorithms on the MOT16 dataset

表 1 各算法在MOT16数据集上的性能比较

算法	MOTA ↑	MT ↑	ML ↓	IDF1 ↑	FP ↓	FN ↓	IDS ↓
MOTDT	47.60	15.2	38.3	50.90	9 253	85 431	792
DeepMOT	54.80	19.1	37.0	53.40	2 955	78 765	645
Tracktor++V2	56.20	20.7	35.8	54.90	2 394	76 844	617
GSM	57.00	22.0	34.5	58.20	4 332	73 573	475
本文算法	58.81	23.6	34.7	55.74	2 908	71 396	802

(Graph Similarity Model)[25]算法作为对照,在MOT17数据集上与本文算法进行性能比较。结果见表2。

Table 2 Performance comparison of various algorithms on the MOT17 dataset

表 2 各算法在MOT17数据集上的性能比较

算法	MOTA ↑	MT ↑	ML ↓	IDF1 ↑	FP ↓	FN ↓	IDS ↓
MOTDT	50.90	17.5	35.7	52.70	24 069	250 768	2 474
FAMNET	52.00	19.1	33.4	48.70	14 138	253 616	3 072
Tracktor++V2	56.30	21.1	35.3	55.10	8 866	235 449	1 987
GSM	56.40	22.2	34.5	57.80	14 379	230 174	1 485
本文算法	60.05	23.4	33.5	55.92	7 353	215 346	2 700

可以看出,本文算法在MOT16与MOT17数据集上取得了最好的MOTA、MT与FN。这是由于本文算法采用融合Transformer与CNN的特征提取模块,可以更好地提取到目标的全局与局部特征;同时采用基于注意力机制的跨空间特征增强与抑制模块,协助目标追踪器关注更具表征性的特征,并减少周边干扰目标的影响。在整个追踪过程中,每个被追踪目标都会尽可能得保留自身特征,并削弱可能误导目标追踪器跟随其他干扰目标的特征,因此可以取得最好的目标追踪效果,并在此过程中尽可能少地跟踪错误目标。本文算法在ML、FP、IDF1、IDS指标方面没有取得最好效果,但与最好效果相差不大,这是由于取得最好效果的对照算法比本文算法更加复杂。综合而言,本文算法各项指标均较为优秀。

3.5 消融实验

在MOT17的训练集上进行消融实验,通过在网络模型中删除提出的模块,以查看它们对整体网络结构的贡献。实验结果如表3所示。

Table 3 Results of ablation experiment

表 3 消融实验结果

模块	MOTA ↑	MT ↑	ML ↓	IDF1 ↑	FP ↓	FN ↓	IDS ↓
w/o enhance & reduce	68.0	35.0	23.1	70.4	33 111	142 256	1 505
w/o feature extraction	68.3	35.3	22.4	71.0	10 563	143 564	651
w/o adaptive weight	68.1	35.6	22.6	69.5	10 654	139 485	684
Full model	68.9	36.4	21.5	71.3	31 323	133 165	643

可以看出,在去掉跨空间特征增强与抑制模块后,模型MOTA、MT、IDF1下降,FN、FP、ML、IDS上升,表明该模块提高了多目标追踪精度,在区分追踪目标与干扰目标方面表现较好。在去掉特征提取模块后,模型MOTA、MT、IDF1、FP下降,FN、ML、IDS上升,但各指标变化幅度更弱。在去掉基于相似度的特征增强与抑制自适应参数后,模型MOTA、MT、IDF1、FP下降,ML、FN、IDS上升,说明自适应参数可以帮助整体网络在跟踪目标过程中更好地抑制追踪目标与干扰目标之间相似度高的特征,并且保留追踪目标在长时间运动过程中的原本特征。

3.6 网络结构实验

在MOT17的训练集上进行网络结构实验,以验证本文提出的基于Transformer与CNN的目标特征提取网络的有效性。结果如表4所示。

Table 4 Results of network structure experiment

表 4 网络结构实验结果

网络结构	MOTA ↑	MT ↑	ML ↓	IDF1 ↑	FP ↓	FN ↓	IDS ↓	Million Parameters
CNN	68.3	36.0	22.6	71.0	31 563	143 564	651	302.5
Transformer	69.0	35.8	22.3	71.4	30 791	133 573	662	400.6
CNN+Transformer	68.9	36.4	21.5	71.3	31 323	133 165	643	320.5

可以看出,本文采用的CNN+Transformer网络结构兼顾了多目标追踪的准确性与网络参数量。相较纯Transformer结构,本文网络结构精度略微下降,参数量明显下降;相较于纯CNN结构,本文网络结构精度有一定提升,且没有付出过大的参数量增长代价。

4 结语

在复杂场景中,追踪目标在运动过程中经常被其他目标遮挡,导致多目标追踪算法容易出现丢失、漏检与误检追踪目标的现象,最终降低追踪精度。针对该问题,本文提出一种基于Transformer的跨空间特征关联多目标追踪算法,将Transformer结构对全局特征的提取能力与CNN对局部特征的提取效率优势相结合,同时利用Transformer结构的互注意力机制映射出需要增强与抑制的特征,根据自适应参数决定其增强与抑制的程度,以加强目标之间的匹配关联。经实验验证,本文算法具有较好的多目标追踪准确度,可以长期跟踪多个目标的运动轨迹,可以降低误检与漏检概率。然而,该算法在进行多目标追踪时的稳定性

有待提升,后续考虑采用基于Transformer的骨干网络进一步改善算法性能。

参考文献:

[1] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multi-object tracking [C]//16th European Conference on Computer Vision, 2020: 107-122.

[2] CHEN L, AI H, ZHUANG Z, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification [C]// IEEE International Conference on Multimedia and Expo, 2018: 1-6.

[3] ZHANG Y, WANG C, WANG X, et al. A simple baseline for multi-object tracking[DB/OL]. https://arxiv. org/pdf/2004. 01888v2. pdf.

[4] ZHOU X, KOLTUN V, KRAHENBUHL P. Tracking objects as points [C]//16th European Conference on Computer Vision, 2020: 474-490.

[5] ZHOU X, WANG D, KRHENBÜHL P. Objects as points[DB/OL]. https://arxiv. org/abs/1904. 07850.

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

[7] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6): 1137-1149.

[8] KALMAN R E. A new approach to linear filtering and prediction problems [J]. Journal of Basic Engineering, 1960, 82(1): 35-45.

[9] WOJKE N, BEWLEY A, PAULUS D, et al. Simple online and realtime tracking with a deep association metric[C]//24th IEEE International Conference on Image Processing, 2017:3645-3649.

[10] ZHANG Y, SUN P, JIANG Y, et al. ByteTrack: multi-object tracking by associating every detection box[C]//European Conference on Computer Vision, 2022:1-21.

[11] SHUAI B, BERNESHAWI A, LI X, et al. Siammot: siamese multi-object tracking[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021:12372-12382.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//31st Annual Conference on Neural Information Processing Systems, 2017: 5999-6099.

[13] CARION N, MASSA F, SYNNAEVE G, et al. End-to-end object detection with Transformer[C]//16th European Conference on Computer Vision, 2020: 213-229.

[14] ZHU X, SU W, LU L, et al. Deformable DETR: deformable transformers for end-to-end object detection [DB/OL]. https://arxiv. org/abs/2010. 04159.

[15] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16x16 words: Transformers for image recognition at scale [DB/OL]. https://arxiv. org/abs/2010. 11929.

[16] WANG W, XIE E, LI X, et al. Pyramid vision transformer: a versatile backbone for dense prediction without convolutions [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 568-578.

[17] LIU Z, LIN Y, CAO Y, et al. Swin transformer: hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021: 10012-10022.

[18] MILAN A, LEAL-TAIXÉ L, REID I, et al. MOT16: a benchmark for multi-object tracking [DB/OL]. https://web3. arxiv. org/abs/1603. 00831.

[19] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[20] LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//30th IEEE Conference on Computer Vision and Pattern Recognition, 2017: 936-944.

[21] XU Y, OSEP A, BAN Y, et al. How to train your deep multi-object tracker[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020: 6787-6796.

[22] CHEN L, AI H, ZHUANG Z, et al. Real-time multiple people tracking with deeply learned candidate selection and person re-identification[C]// 2018 IEEE International Conference on Multimedia and Expo, 2018: 1-6.

[23] CHU P, LING H. Famnet: joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 6172-6181.

[24] BERGMANN P, MEINHARDT T, LEAL-TAIXE L. Tracking without bells and whistles[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019: 941-951.

[25] LIU Q, CHU Q, LIU B, et al. GSM: graph similarity model for multi-object tracking[C]//Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence, 2020: 530-536.

(责任编辑:尹晨茹)