·开发研究与工程应用·　　文章编号：1000-3428(2022)09-0305-09　　文献标志码：A　　中图分类号：TP391

多特征融合的端到端链式行人多目标跟踪网络

周海赟[1]，项学智[2]，王馨遥[2]，任文凯[2]

（1.南京森林警察学院 治安学院，南京 210023；2.哈尔滨工程大学 信息与通信工程学院，哈尔滨 150001）

摘　要：目标检测、特征提取与数据关联作为多目标跟踪网络中重要的组件，独立或部分联合地发挥作用，这种组件分离的方法虽取得了良好的跟踪效果，但增加了跟踪网络的复杂性，影响了跟踪速度。为提升行人多目标跟踪速度及维持跟踪精度，提出一种端到端链式行人多目标跟踪网络。将目标检测、特征提取与数据关联集成到一个统一的框架中，将连续2帧图片组成一个节点作为输入，直接回归出节点之间相同目标的成对边界框，利用相邻节点之间公共帧的强相似性，仅使用交并比匹配进行数据关联，以提高跟踪速度。使用多特征融合的双向特征金字塔，并在金字塔网络中引用改进可变形卷积，提高模型对目标形变的适应性。为解决正负样本不平衡及梯度贡献的差异，将focal loss与BalancedL1 Loss组成多任务学习损失函数以促进网络的均衡学习。在MOT17数据集上的实验结果表明，与DeepSORT、TubeTK、CenterTrack等网络相比，该网络可有效实现跟踪速度与精度的平衡，多目标跟踪精度为69.6，跟踪速度保持为21.6 frame/s。

关键词：多目标跟踪；链式跟踪；多特征融合；特征金字塔；多任务损失函数

开放科学（资源服务）标志码（OSID）：

中文引用格式：周海赟，项学智，王馨遥，等.多特征融合的端到端链式行人多目标跟踪网络[J].计算机工程，2022，48(9)：305-313.

英文引用格式：ZHOU H Y，XIANG X Z，WANG X Y，et al. Chained end-to-end pedestrian multi-object tracking network with multi-feature fusion[J].Computer Engineering，2022，48(9)：305-313.

Chained End-to-End Pedestrian Multi-Object Tracking Network with Multi-Feature Fusion

ZHOU Haiyun[1]，XIANG Xuezhi[2]，WANG Xinyao[2]，REN Wenkai[2]

（1.Institute of Public Security，Nanjing Forest Police College，Nanjing 210023，China；
2.School of Information and Communication Engineering，Harbin Engineering University，Harbin 150001，China）

【Abstract】Object detection，feature extraction，and data association as important components in multi-target tracking network，work independently or partially jointly. Despite the improved tracking performance，separated components increase the tracking network complexity and decrease the tracking speed. An end-to-end chained network with multifeature fusion is proposed to increase the speed of pedestrian multi-object tracking while maintaining tracking accuracy.The network integrates object detection，feature extraction，and data association into a framework.Two adjacent frames form a node as the input. The network regresses the bounding box pairs of the same target in the node. The common frames across nodes have a strong correlation such that using Intersection over Union（IoU）matching for data association improves the tracking speed.In addition，the multi-feature fusion pyramid is adopted to fully integrate the high-level semantic information and low-level position information. The pyramid adopts deformable convolution v2，which increases adaptability to the deformation of objects.Focal loss and balanced L1 loss form multitask learning loss for promoting the balanced learning to improve the tracking performance，owing to the imbalance in the positive and negative samples and the differences in the gradient contributions.The experimental results for the MOT17 dataset show that compared with DeepSORT，TubeTK，CenterTrack，and other networks，this network can effectively achieve the trade-off between the tracking speed and accuracy.The tracking accuracy Mota value is 69.6，and the tracking speed is maintained at 21.6 frame/s.

【Key words】multi-object tracking；chained-tracker；multi-feature fusion；feature pyramid；multi-task loss function

DOI：10. 19678/j. issn. 1000-3428. 0062296

基金项目：中央高校基础科研业务费项目（LGY201802，LGZD202102）；国家自然科学基金（61401113）；黑龙江省科学基金项目（LH2021F011）；华为MindSpore学术基金。

作者简介：周海赟（1980—），女，副教授、博士，主研方向为计算机视觉、模式识别；项学智，副教授；王馨遥、任文凯，硕士研究生。

收稿日期：2021-08-03　　**修回日期：**2021-09-20　　**E-mail：**zhouhy@nfpc.edu.cn

0 概述

多目标跟踪是指在视频中持续对目标进行准确定位,在场景发生变化时仍能维持目标身份信息不变,最后输出目标完整运动轨迹的技术。在复杂场景中,跟踪目标数目不定、目标之间存在频繁的遮挡以及交互、目标之间包含相似的外观特性等因素都会给多目标跟踪的实现带来挑战。由于行人是非刚体目标,且现有数据集中包含大量行人的视频,因此当前多目标跟踪中行人跟踪的算法占多数[1]。行人多目标跟踪主要分为离线跟踪与在线跟踪。在线跟踪只能使用当前帧及之前的信息来进行跟踪,而离线跟踪对每一帧的预测都可以使用整个视频帧的信息,因此离线跟踪可以看成是一个全局优化的问题,常见解决方法是基于图论的方式,将多目标跟踪建模为网络最大流问题[2]或距离最小成本问题[3]。由于离线跟踪的全局优化方式增加了对算力的要求,且离线跟踪不能应用于对跟踪实时性有要求的场景,因此本文主要研究行人多目标在线跟踪。

传统多目标跟踪网络主要通过滤波算法来预测目标在下一帧的位置进行目标跟踪,卡尔曼滤波器利用连续帧中相同目标的速度及协方差相关性最大原理进行目标状态的预测与更新[4]。使用核相关算法训练相关滤波器,并通过计算目标相关性获得置信图来预测跟踪结果[5]。当前多目标跟踪网络主要采用基于检测的跟踪方法,即先对视频中每一帧的目标进行目标检测,之后利用各种数据关联算法将检测结果与跟踪轨迹进行匹配,从而进行轨迹更新。

近年来,越来越多的研究人员致力于基于深度学习的多目标跟踪网络的研究。BEWLEY等[6]提出Sort网络,检测部分采用Faster R-CNN网络,利用卡尔曼滤波预测结果与检测结果之间的交并比(Intersection over Union,IoU)进行匈牙利匹配来完成数据关联。由于Sort仅使用IoU进行数据关联,导致在人流较密集的场景下会产生大量的身份切换。因此,WOJKE等[7]提出Deepsort,在Sort网络IoU匹配的基础上增加级联匹配,并使用一个行人重识别(Person Re-identification,ReID)网络提取目标的外观特征辅助数据关联,有效解决身份切换问题。BAE等[8]也利用预训练的ReID网络提取可区分的行人特征,并将轨迹分为可靠轨迹与不可靠轨迹,再与检测结果进行分级关联。上述这些研究仅根据检测结果进行轨迹更新,受检测器性能的影响很大,当出现不可靠的检测时,跟踪性能也会下降。因此,CHEN等[9]将检测框与跟踪的预测框同时作为轨迹更新的候选框,设计一种评分函数统一衡量所有的候选框,再利用空间信息和ReID特征进行数据关联。尽管这些基于检测进行跟踪的网络取得了良好的效果,但这些网络的检测部分与跟踪部分是完全独立的,这直接增加了跟踪的复杂性,不利于满足实时性的要求。为解决该问题,BERGMANN等[10]提出Tracktor++,利用检测器的边界框回归思想直接预测目标在下一帧中的位置,完成检测与跟踪的联合,并融入运动模型与ReID网络,以减少帧间身份切换。ZHOU等[11]在CenterNet检测器的基础上输出当前帧中目标的尺寸、目标中心点的热力图及相较于上一帧的偏移量,依靠贪婪匹配实现数据关联。WANG等[12]提出JDE网络,将ReID网络与检测网络整合到一个网络中,使网络同时输出检测结果和相应的外观嵌入,再根据目标的外观信息与运动信息进行数据关联。ZHAN等[13]提出FairMOT网络,网络中包含检测与ReID两个同质分支,使用编解码架构提取网络的多层融合特征,提高网络对物体尺度变换的适应能力。尽管上述方法进一步改善了目标跟踪的性能,但上述方法不使用端到端的网络,文献[10-11]在一个网络中联合学习检测与跟踪,文献[12-13]将检测与ReID网络集成到一起,这些方法中的数据关联过程仍被视为后处理部分,是一种部分端到端的网络,仍然无法做到全局优化,需要复杂的数据关联机制来处理不同模块的特征,不利于满足在线跟踪的实时性要求。

本文基于链式结构[14]提出一种多特征融合的端到端链式行人多目标跟踪网络,利用链式特性降低数据关联的复杂性。在链式结构中引入双向金字塔,在传统特征金字塔的基础上增加一条聚合路径以获得更深入的融合特征。为适应目标形状和尺度的改变,在双向金字塔中采用具有采样特征加权的改进可变形卷积。使用联合注意力提高目标框的准确性,重点突出2帧图片中属于同一目标的区域[14]。最后,设计多任务学习损失函数,优化成对目标边界框回归的准确性,提升整体跟踪的性能。

1 多特征融合的链式跟踪网络

1.1 本文网络架构

本文基于链式网络结构提出多特征融合的跟踪网络,将目标检测、特征提取和数据关联融入到一个统一的框架中。与其他网络不同,常见的在线多目标跟踪逐帧进行检测与数据关联,网络的输入仅为单个帧,本文将相邻的两帧组成链节点作为网络的输入,完成链式跟踪,链式跟踪的整体流程如图1所示。给定一个共有N帧的图像序列$\{F_t\}_{t=1}^{N}$,F_t表示第t帧的图像,每一个链节点由相邻两帧图像组成,第1个链节点为(F_1,F_2),第N个节点为(F_N,F_{N+1}),由于图像序列最多只有N帧,将F_{N+1}用F_N表示,即将第N个节点改写为(F_N,F_N)。将节点(F_{t-1},F_t)输入到网络中,网络会输出2帧中属于相同目标的成对边界框$\{(D_{t-1}^{i},\hat{D}_t^{i})\}_{i=1}^{n_{t-1}}$,其中$n_{t-1}$表示相同目标对的数量,$D_{t-1}^{i}$与$\hat{D}_t^{i}$分别表示节点内$F_{t-1}$与$F_t$中相同目标的两个边界框。同理,下一个节点经过网络的输出

$\{(D_t^j,\hat{D}_{t+1}^j)\}_{j=1}^{n_t}$。$\hat{D}_t^i$与$D_t^j$表示相邻节点的公共帧中相同目标的边界框,本质上它们来自同一帧图像,理论上仅存在微小的差异,故不需要复杂的数据关联机制。计算$\hat{D}_t^i$与D_t^j之间的帧间交并比以获取亲和力矩阵,从而链接2个相邻的节点。应用匈牙利算法完成$\hat{D}_t^i$与D_t^j中相同目标检测框的最优匹配任务,对于成功匹配上的边界框对应用D_t^j对$\hat{D}_t^i$所在的轨迹进行更新。针对目标消失的情况,若目标出现在F_{t-1}帧而在F_t帧消失,节点(F_{t-1},F_t)与(F_t,F_{t+1})均不会检测到该目标,因此可以认为该目标在F_{t-1}帧甚至是F_{t-2}帧就已消失,避免误检噪声引起的跟踪器的漂移现象。针对目标可能连续几帧消失在可视范围内导致检测失败的情况,保留消失目标的轨迹和身份σ帧,在这期间利用物体的匀速运动模型进行运动估计,持续预测目标位置并与当前检测结果不断进行匹配,尝试把丢失的目标重新链接至轨迹中,保证在强遮挡情况下目标仍可以被有效跟踪,减少身份切换的现象发生。若在σ帧之后仍没有匹配成功,则认为该目标离开了场景,此时将该目标的相关轨迹以及身份信息删除。

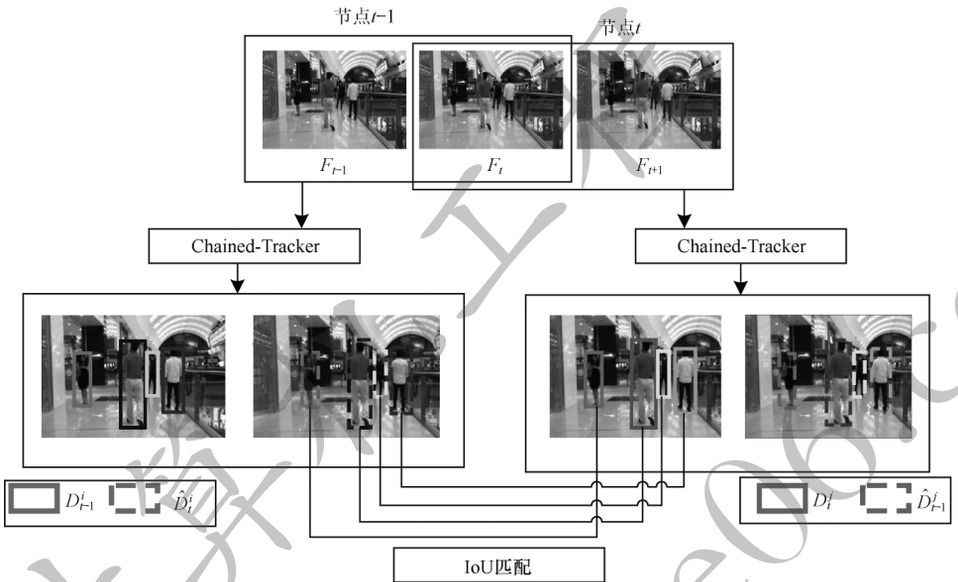

图1　链式跟踪的整体流程

Fig.1　Overall process of chain tracking

针对场景中新目标出现的问题,在进行IoU匹配时,将D_t^j未匹配上的检测框认为是新出现的目标,对其分配新的身份并且初始化新的轨迹。若目标不在F_{t-1}帧而出现在F_t帧,节点(F_{t-1},F_t)旨在输出相同目标的边界框对,因此不会识别该目标,但如果该目标稳定出现在场景中,该目标在节点(F_t,F_{t+1})的输出就会被检测到,并获得初始化的新轨迹和身份标识。模型利用IoU匹配进行数据关联,同时运动估计保证了长轨迹的生成,增加模型应对遮挡的鲁棒性。

为获得每个节点中的边界框对,网络利用了目标检测中的边界框回归思想,直接回归出两帧图像中相同目标的边界框对,网络的整体架构如图2所示。

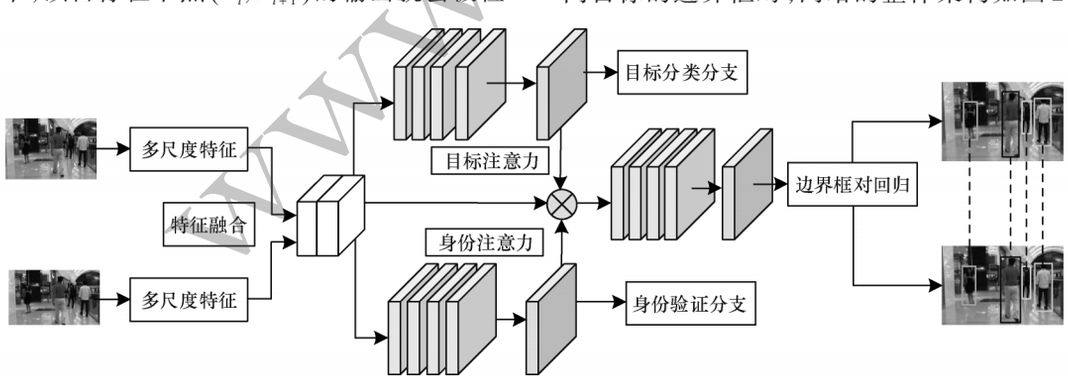

图2　网络整体架构

Fig.2　Overall architecture of network

由图2可知,网络采用孪生网络结构将连续两帧图像共同输入至网络中,分别利用Resnet50作为骨干网提取深层语义特征,并利用多特征融合的双向金字塔结构输出多尺度的特征表示。多特征融合的特征金字塔结构如图3所示。为获得两帧图像中相同目标的位置,首先将骨干网络生成的相邻帧多

尺度特征图进行拼接,然后送入预测网络中,以直接回归出边界框对。预测网络由3个分支组成,包括目标分类分支、身份验证分支以及边界框对回归分支。目标分类分支针对每个检测框预测前景区域置信度分数,以判断该区域中是目标还是背景。身份验证分支用于判断成对的检测框中是否包含同一个目标。若包含同一个目标,边界框回归分支同时预测两个边界框中该目标的坐标。

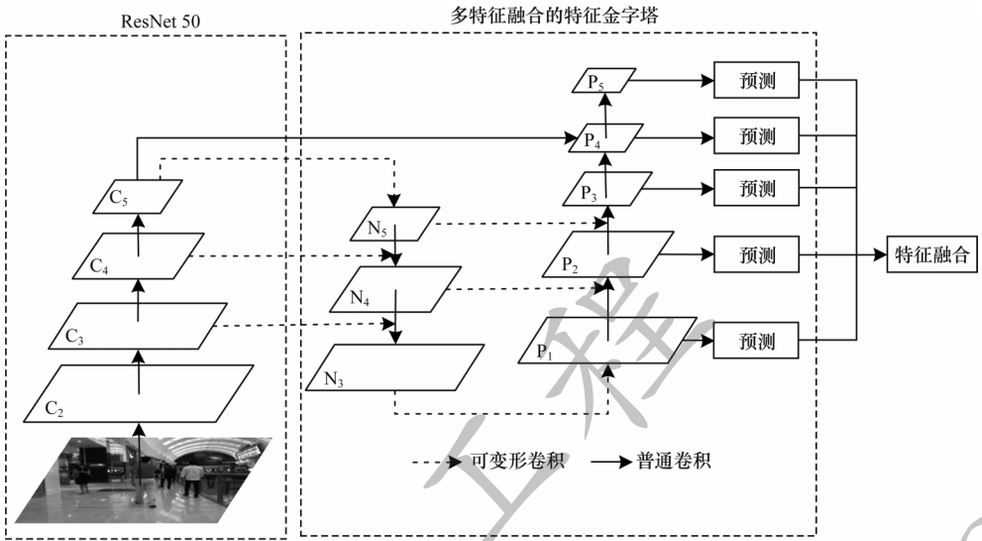

图3 多特征融合的金字塔网络

Fig.3 Feature pyramid network with multi-feature fusion

为促进边界框回归过程可以集中于两帧图像中的相同目标,并且避免被无关信息干扰,预测网络中使用联合注意力模块,使回归过程更加关注组合特征中的有效信息区域[14],身份验证分支和目标分类分支的预测置信度图均被用作注意力图,将注意力图与组合特征相乘后再输入到边界框对回归分支,来自2个分支的注意力起互补作用。利用预测网络结构中3个分支的特性构造联合注意力模块,充分利用分类分支与身份验证分支的信息。相比只用于单个分支的常规注意力,联合注意力可以共同利用2个分支的结果作用于回归过程。在执行边界框回归前,其他2个分支的结果联合作用于回归分支,2个分支的结果可以通过损失函数的设计来调节回归过程,其中分类分支的注意力图促进回归过程更加关注包含有效信息的前景区域,身份验证分支的注意力图使网络集中于相同目标检测框对的回归,能充分利用2个分支的有效信息并且更好地监督回归过程,在一定程度上促进了网络中相同目标边界框回归的准确性。

网络中将相邻两帧图片组成一个节点作为输入,网络回归出两帧图片中相同目标的边界框对,不同节点之间由于存在公共帧,因此差异较小,故使用简单的IoU匹配完成节点之间的关联,使用基础的匈牙利算法就可以完成检测框之间的最优匹配,从而完成帧间数据关联过程。数据关联的简化有利于提高跟踪的速度,满足实时性的要求。根据网络的输出特性设计轨迹管理机制,若节点间的公共帧成功匹配,则更新轨迹状态;若匹配失败则进入轨迹丢失状态,保存当前运动轨迹以及身份,同时使用运动估计尝试重新关联轨迹与目标。网络节点间的链式特性降低了误检的影响,也降低了关联机制的复杂度,实现了端到端的跟踪过程。

1.2 多特征融合的特征金字塔结构

行人目标在视频帧中处于移动状态,目标尺度变化很大,如果利用检测器的回归思想直接回归出图像对的边界框,就需要充分利用目标的语义信息保证回归的边界框坐标准确,同时增加小目标识别的准确性。常见的目标检测网络如Faster R-CNN仅利用了骨干网提取的顶层特征来进行目标的识别与定位,图像中小目标在下采样过程中包含的有用信息会进一步减少甚至消失,这种方法不利于对小目标进行预测[15]。SSD网络使用了多尺度特征融合的方法,从骨干网的不同层中提取不同尺度的特征进行融合,但这仍没有充分融合低层的语义信息[16]。因此,提出在骨干网后接入特征金字塔网络(Feature Pyramid Networks,FPN)[17],骨干网采用ResNet50完成自底向上前向传播的过程,将特征图尺寸不变的层归为一个阶段。提取每个阶段最后一层的输出来完成特征融合,同时加入自顶向下的过程,来自顶层的特征图经过上采样与骨干网提取的相同尺寸的特征图横向连接并进行特征融合,以同样的方式逐层进行特征融合获得多尺度的特征图,充分融合高层与低层的语义信息,进而适应目标的尺度变化。

本文为提高网络对目标尺度的适应能力,在链式跟踪网络架构的基础上引入多特征融合的双向金字塔网络,借鉴了PANet[18]的思想,在特征金字塔FPN自顶向下的聚合路径后增加一条自底向上的特征聚合路径,形成多特征融合的特征金字塔结构,更加充分地利用

网络的浅层信息,有利于获得目标的更多位置信息。同时,为适应目标的形状变化特性,将原金字塔结构中的传统卷积替换为采样特征加权的改进可变形卷积(Deformable ConvNets v2,DCN v2),DCN v2的具体结构见1.3节。利用多特征融合的特征金字塔可同时适应目标的尺寸与形状变化特性。

1.3　改进可变形卷积

传统卷积方式使用尺寸固定的卷积核,对输入特征图采用规则采样,一般采样方式以矩形结构为主,提取的特征也是矩形框内的特征。然而,视频序列中行人目标在不停移动,不同帧中目标的形状以及尺度变化很大,传统卷积方式对这种几何变换的适应性较差,导致特征提取不完整,引入背景噪声也容易影响最后的结果。为适应不同目标形状和尺度的改变,在双向金字塔网络中引入DCN v2卷积,DCN v2卷积实现过程如图4所示。

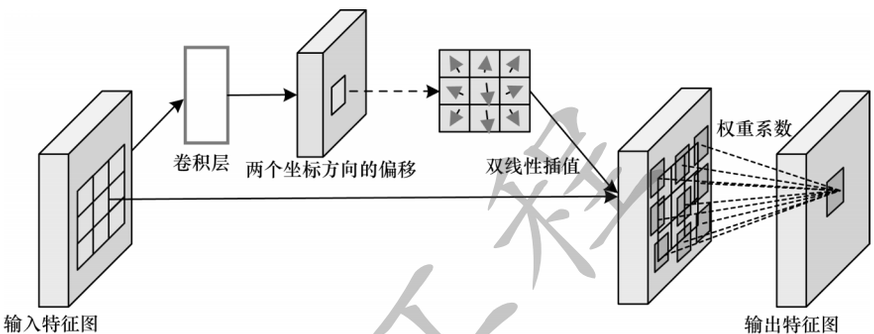

图4　DCN v2卷积的实现过程

Fig.4　Implementation process of DCN v2 convolution

在生成多尺度特征图的过程中将传统卷积替换为改进可变形卷积,根据当前目标的形状和尺寸自适应地调整采样点的位置,从而更加准确地提取目标特征。可变形卷积的主要思想是在标准卷积的规则网格采样位置添加2D偏移量,偏移量的计算通过另一个标准卷积过程实现,故偏移量可以在训练过程中一起被学习,卷积核的大小和位置则可以根据学习到的偏移量进行动态调整,达到根据目标形状与尺度自适应调整的目的[19]。

可变形卷积在实现的过程中并非根据偏移量直接改变卷积核,而是通过对卷积前图片的像素值进行重新整合后再进行一般卷积操作,达到卷积核扩张的效果。可变形卷积的实现需要利用普通卷积中共N个采样点提取特征图,通过额外的卷积层从特征图中进行偏移量的学习,偏移量具有和输入特征图相同的空间分辨率,且为2D分量,每个采样位置叠加1个偏移量,分别包含x与y两个坐标方向的偏移,故将输出通道数设置为$2N$,表示N个采样点的2D偏移,偏移量随着网络训练过程一起被学习。偏移量的变化以局部、自适应的方式取决于输入特征图中目标的形变,将获得的偏移量叠加到原来的特征图上获得加入偏移后的坐标位置,通过双线性插值的方式计算坐标位置对应的像素值。当获得所有像素值后便得到一个新的图像,将这个新图像作为输入,并进行常规的卷积操作。可变形卷积的使用对目标的形变等更具适应性,但偏移量的叠加倾向于使采样点聚集在目标对象周围,对物体的覆盖不精确,且会引入无关的背景信息造成干扰。为每一个采样点增加权重系数,通过给每一个偏移后的采样位置赋予权重来区分当前采样位置是否包含有效信息。采样权重取值在[0,1]之间,采样权重与偏移量一样,都是通过对输入特征图采用卷积运算来获得,即采样权重也是可学习参数,可以根据采样位置的变化在训练中学习得到。对输入特征图进行卷积操作,输出通道数由$2N$增加为$3N$,其中前$2N$个通道仍表示N个采样点的2D偏移量,剩余的N个通道被进一步送入Sigmoid层以获得采样权重,将偏移量与采样权重的初始值分别设置为0和0.5,并在网络训练过程中不断被优化。如果叠加了偏移后的采样区域没有目标信息,则通过学习使权重降低,从而使网络可以更加集中于目标区域[20]。

2　损失函数

网络中的损失函数主要由预测网络中3个分支的损失函数进行加权得到。在目标分类分支中,为避免正负样本比例失衡的问题,采用focal loss[21]的表示形式,用L_{cls}表示分类任务的损失函数,对于任意节点(F_t,F_{t+1})用$A_t^i=(x_a^{t,i},y_a^{t,i},w_a^{t,i},h_a^{t,i})$表示其第$i$个链锚。其中:$(x_a^{t,i},y_a^{t,i})$表示锚盒中心点的位置坐标;$w_a^{t,i}$与$h_a^{t,i}$分别表示锚盒的宽与高;$G_t^j$表示$F_t$中的真值目标框。首先针对每个真值框$G_t^j$找到与其IoU最大的$A_t^i$,将该锚框与其匹配,保证每个真值框均可以被匹配到。为了防止正负样本的比例差距过大,将剩余的锚框与真值框继续匹配。若A_t^i与G_t^j之间的IoU大于阈值$T_{\max}$,则将当前的锚框与该真值框匹配,将所有匹配成功的锚框位置标签c_{cls}^i记为1,表示A_t^i所在区域为前景区域且与G_t^j成功匹配;反之,记标签为0,表示背景。将分类分支网络最后一层经Sigmoid激活函数输出的分类概率记为$p_{t,\mathrm{cls}}^i$,以focal loss表示分类分支的损失函数,如式(1)所示:

$$L_{\text{cls}}=\begin{cases}-\alpha(1-p^i_{t,\text{cls}})^{\gamma}\ln p^i_{t,\text{cls}}, c^i_{\text{cls}}=1\\-(1-\alpha)p^{i\ \gamma}_{t,\text{cls}}\ln(1-p^i_{t,\text{cls}}), c^i_{\text{cls}}=0\end{cases}\tag{1}$$

其中：γ参数可以减少简单样本的损失，使网络专注于困难样本；α参数用于平衡正负样本的比例。在身份验证分支中，损失函数也采用focal loss的表示形式使网络集中于包含相同目标的成对边界框。用L_{id}表示该分支的损失函数，对于链锚A^i_t预测的检测框对$(D^i_t,\hat{D}^i_{t+1})$，用(G^j_t,G^k_{t+1})表示与其相匹配的真值框，若此时预测框为前景区域且与其匹配的真值框对应的身份相同，则将其身份标签c^i_{id}记为1，表示此时网络预测的检测框对为同一目标。身份标签c^i_{id}的表达式如式(2)所示：

$$c^i_{\text{id}}=\begin{cases}1, c^i_{\text{cls}}=1\text{且}\Gamma(G^j_t)=\Gamma(G^k_{t+1})\\0, \text{其他}\end{cases}\tag{2}$$

其中：$\Gamma(G^j_t)$表示G^j_t对应的身份标签，身份验证分支与目标分类分支具有相同的网络结构。

记最后一层Sigmoid输出为$p^i_{t,\text{id}}$，则身份验证分支的损失函数为：

$$L_{\text{cls}}=\begin{cases}-\alpha(1-p^i_{t,\text{id}})^{\gamma}\ln p^i_{t,\text{id}}, c^i_{\text{id}}=1\\-(1-\alpha)p^{i\ \gamma}_{t,\text{id}}\ln(1-p^i_{t,\text{id}}), c^i_{\text{id}}=0\end{cases}\tag{3}$$

边界框对$(D^i_t,\hat{D}^i_{t+1})$回归损失由前一帧的回归损失与后一帧的回归损失相加组成，总的回归损失由L_{reg}表示。回归任务中存在巨大的样本不平衡问题，常用的回归损失函数如SmoothL1 Loss在网络训练过程中由简单样本产生的梯度影响仅为困难样本的30%，网络的回归任务没有得到均衡的学习[15]。为避免简单样本与困难样本不平衡，本文使用BalancedL1 Loss使简单样本与困难样本产生的梯度贡献相同，从而促进网络的均衡学习[22]。对于链锚$A^i_t=(x^{t,i}_a,y^{t,i}_a,w^{t,i}_a,h^{t,i}_a)$回归出的检测框对$(D^i_t,\hat{D}^i_{t+1})$，用$(\Delta^{t,i}_d,\Delta^{t+1,i}_{\hat{d}})$表示成对的预测框相对于锚盒的回归偏移，$(\Delta^{t,j}_g,\Delta^{t+1,k}_g)$表示真值框相对于锚盒的偏移。以$\Delta^{t,i}_d$的计算为例，检测框$D^i_t$可用$(x^{t,i}_d,y^{t,i}_d,w^{t,i}_d,h^{t,i}_d)$表示，则偏移量$\Delta^{t,i}_d$包括这4个分量的偏移，表达式如式(4)所示：

$$\begin{aligned}&\Delta^{t,i}_d=(\Delta^{t,i}_{d,x},\Delta^{t,i}_{d,y},\Delta^{t,i}_{d,w},\Delta^{t,i}_{d,h})\\&\Delta^{t,i}_{d,x}=(x^{t,i}_d-x^{t,i}_a)/w^{t,i}_a\\&\Delta^{t,i}_{d,y}=(y^{t,i}_d-y^{t,i}_a)/h^{t,i}_a\\&\Delta^{t,i}_{d,w}=\ln(w^{t,i}_d/w^{t,i}_a)\\&\Delta^{t,i}_{d,h}=\ln(h^{t,i}_d/h^{t,i}_a)\end{aligned}\tag{4}$$

同理得到4个偏移量$(\Delta^{t,i}_{\hat{d}},\Delta^{t+1,i}_{\hat{d}})$与$(\Delta^{t,j}_g,\Delta^{t+1,k}_g)$，通过BalancedL1 Loss来最小化预测框与真值框间的偏移差，使预测框尽可能接近于真值框。BalancedL1 Loss用$L_b(x)$表示：

$$\begin{aligned}&L_b(x)=\begin{cases}\dfrac{\lambda}{b}(b|x|+1)\ln(b|x|+1)-\lambda|x|, |x|<1\\\eta|x|+c, \text{其他}\end{cases}\\&\text{s.t. }\lambda\ln(b+1)=\eta\end{aligned}\tag{5}$$

其中：参数λ可以控制简单样本的梯度变化，当λ较小时可以使简单样本产生较大的梯度，以平衡简单样本与困难样本的梯度贡献；参数η用于调整回归误差的上限；参数b保证了在$x=1$时，损失函数有相同的值。3个参数共同作用以满足约束条件，使得当偏移差接近于0时，梯度迅速下降，接近于1时梯度缓慢上升，解决SmoothL1 Loss在偏移差为1时的突变问题，使网络训练可以更平衡。

由$L_b(x)$可得回归分支成对检测框的回归损失L_{reg}的表示式如式(6)所示：

$$L_{\text{reg}}=\sum_{l\in\{x,y,w,h\}}\left[L_b(\Delta^{t,i}_{d,l}-\Delta^{t,i}_{g,l})+L_b(\Delta^{t+1,i}_{\hat{d},l}-\Delta^{t+1,k}_{g,l})\right]/8\tag{6}$$

在获得目标分类分支、身份确认分支与边界框回归3个分支的损失函数后，以一定权重对3个分支损失函数进行加权，获得网络总的损失函数L_{total}，其表达式如式(7)所示：

$$L_{\text{total}}=L_{\text{reg}}+mL_{\text{cls}}+nL_{\text{id}}\tag{7}$$

其中：参数m与n分别表示分类损失与身份确认损失在L_{total}的权重。

3 实验结果与分析

本文所设计的网络在MOT17数据集上进行训练与测试。MOT17数据集发布于MOTChallenge上，相较于之前版本的视频序列有更高的行人密度，共包括1 342个身份标识及292 733个目标框，总计11 235帧。MOT17数据集包含14个视频序列，既有静态摄像机场景也有动态摄像机场景，还包含不同的光照场景，例如晚间人群密集的商业街、光线昏暗的公园、明亮的商场中运动摄像机的跟拍、街道上模拟自动驾驶场景等。本文将MOT17数据集中7个视频序列用于训练，其余7个用于测试。

本文使用多目标跟踪中最常用的CLEAR Metric[23]与IDF1[24]指标来评估模型的性能，其中CLEAR Metrics主要包括多目标跟踪准确度(Multiple-Object Tracking Accuracy，MOTA)、多目标跟踪精度(Multiple-Object Tracking Precision，MOTP)、主要跟踪轨迹(Mostly Tracked Trajectories，MT)、主要丢失目标轨迹(Mostly Lost Trajectories，ML)、身份切换总数(Identity Switches，IDS)、跟踪速度等指标。

1)MOTA是融合了误检、漏检与身份切换3种因素的综合性指标，衡量模型在检测目标和关联轨迹时的整体性能，体现多目标跟踪的准确度；

2)MOTP为目标检测框与真值框在所有帧之间的平均度量距离，衡量多目标跟踪的精度，主要是检测器的定位精度；

3)MT指标衡量了目标存在期间与真值轨迹匹配高于80%的预测轨迹数目占轨迹总数目的比例；

4)ML指标衡量目标存在期间与真值轨迹匹配低于20%的预测轨迹占总轨迹数的比例，MT与ML两个指标均不考虑目标是否发生身份切换，仅衡量目标跟踪的完整性；

5)IDS衡量整个跟踪过程身份切换的数目，衡量跟踪算法的稳定性；

6）跟踪速度指标用帧率（Frame Per Seconds，FPS）来衡量，FPS数值越大，跟踪速度越快；

7）IDF1指标衡量轨迹中身份标识的准确性。

以上指标中，MOTA为最受关注的指标，体现了跟踪整体的性能。

在网络训练过程中为防止过拟合，一般会利用4种方法进行数据增强：以0.5的概率随机对图像进行亮度调整；色彩与饱和度调整；水平翻转；以[0.3,0.8]的尺度范围对图像进行随机裁剪。将模型在MOT17训练集上训练时的批量大小设置为8，采用标准的Adam优化器对网络训练100轮，初始的学习率设为5×10^{-5}，在网络训练过程中连续3轮损失不下降则衰减学习率，学习率衰减因子为0.1。为平衡训练过程中的回归损失与分类损失，将损失函数L_{total}中的参数m与参数n均设置为1.4，目标分类损失L_{cls}与身份验证损失L_{id}中参数α与参数γ分别设置为0.25与2.0，回归损失L_{reg}中参数λ与参数η分别设置为0.5与1.5。在锚框与真值框匹配阶段将IoU匹配阈值T_{max}设置为0.5，在节点链接阶段，根据IoU匹配的链接阈值设置为0.4，消失的目标保留其身份与轨迹σ帧，此处σ设置为10。

本文设计消融实验探究模型中各模块对整体性能的影响，实验结果如表1所示。

表1　消融实验结果

Table 1　Ablation experiment results

方法	MOTA	MOTP	IDF1	MT/%	ML/%	IDS
基础链式结构	66.6	78.2	57.4	32.2	24.2	5 529
基础链式结构+多特征融合金字塔	68.4	79.5	61.5	34.1	23.6	4 770
基础链式结构+多特征融合金字塔+Loss	69.6	81.0	63.9	33.9	24.5	4 668

由表1可以看出，由于基础链式结构中不包含多特征融合以及多任务损失模块，在加入多特征融合的特征金字塔结构之后，MOTA指标从66.6提升到了68.4，MOTP指标也提升了1.3，MT指标、IDF1指标均得到大幅提升，但ML指标与IDS指标均有不同程度的下降。改进后的特征金字塔网络增加了一条从下到上的特征融合路径，将其与可变形卷积融合，使其在特征提取阶段获得适应多目标形变与多尺度变化的融合特征。充分利用浅层语义信息提取更多的目标位置信息，可以根据目标的变化动态调整感受野，使网络更能适应目标的形变。由于实验所用数据集人流密度较大且处于动态变化，因此网络可以在目标发生形变时自适应地提取动态变化的特征，增强对目标形变的适应能力以及对小目标的检测能力，进而提升MOTA、MOTP、MT和ML等指标。此外，检测到的回归框能够进一步保证节点链接的准确性，因此相同目标的数据关联过程更准确，能够减少身份切换现象的发生，进一步优化了IDF1与IDS指标。本文网络引入BalancedL1 Loss替换传统的SmoothL1 Loss损失函数，并进一步调整了损失函数的权重。虽然MT、ML指标有一定波动，但是其他指标均有不同程度的提升，MOTA指标获得了1.2的增益、MOTP、IDF1指标分别提高了1.5、2.4，这表明在网络训练过程中平衡简单样本与困难样本的梯度影响更有利于网络回归任务的均衡学习，提高了回归边界框对的准确性，进一步改善了跟踪过程的准确度与精度。

多目标跟踪网络受检测器的影响很大，为了公平地评价多目标跟踪网络的性能，将网络分为Private与Public两种。Public方法使用数据集中提供的固定检测器完成整个跟踪模型的搭建，Private方法可以使用任意检测器。由于Private方法可以使用任何一个性能更好的检测器，因此同等条件下的Private方法比Public方法效果更好。MOT17数据集中的公共检测器为DPM、SDP与Faster R-CNN3种检测器，而本文网络结构中检测部分利用了RetinaNet结构，属于Private方法。为公平比较，本文仅将所设计网络与其他使用Private方法的网络比较，结果如表2所示，表中加粗数字为该组数据的最大值。

表2　不同网络在MOT17数据集下的实验结果

Table 2　Experiment results of different networks under MOT17 date set

网络	MOTA	MOTP	IDF1	MT/%	ML/%	IDS	帧率/(frame·s^{-1})
DeepSORT[7]网络	60.3	79.1	61.2	31.5	20.3	**2 442**	20.0
CenterTrack[11]网络	67.8	78.4	64.7	34.6	24.6	3 039	3.8
SST[25]网络	52.4	76.9	49.5	21.4	30.7	8 431	<3.9
TubeTK[26]网络	63.0	78.3	58.6	31.2	19.9	4 137	3.0
GSDT[27]网络	66.2	79.9	**68.7**	**40.8**	**18.3**	3 318	4.9
Quasi_Dense[28]网络	68.7	79.0	66.3	40.6	21.9	3 378	20.3
TransCenter[29]网络	**70.0**	79.6	62.1	38.9	20.4	4 647	1.0
本文网络	69.6	**81.0**	63.9	33.9	24.5	4 668	**21.6**

由表2可知，基于本文网络的方法具有较高的MOTA值以及MOTP值。分析原因可能是使用基于可变形卷积的多特征融合网络增强了模型特征提取能力，提高了模型对行人目标尺度以及形变的适应能力，进而提高了整体跟踪的精度与准确度。由表2还可知，本文网络的MT指标、ML指标与IDS指标相较于其他网络效果略有降低，但是具有最高的帧率。这是因为本文网络在链式跟踪时节点之间仅使用IoU进行匹配，利用节点之间公共帧的相似性进行数据关联，省去了复杂的数据关联算法，因此大幅提高了跟踪算法整体的速度。但是本文网络仅使用IoU关联，与其他复杂的关联算法进行对比，关联的准确率与精度有所下降，影响跟踪过程的完整性，或者容易出现身份切换的现象，这表现在MT指标、ML指标与IDS指标值的降低。测试结果中，本文网络的帧率最高，MOTA指标相较于其他网络略有降低。这是因为数据关联阶段使用的IoU匹配属于简单的基础匹配方法，在一定程度上影响了匹配的准确性，从而降低了跟踪精度。在通常情况下采用更复杂的匹配方式替换IoU匹配可以提高跟踪精度，但数据关联需要对输入的视频帧进行逐帧匹配，数据关联算法的复杂性增加后，整个视频帧的跟踪速度就会大幅降低。从跟踪速度与精度权衡的角度考虑，本文选取了复杂性较低的基础IoU匹配方法。为了降低简单匹配方式带来的影响，本文还采用了链式结构，利用节点之间的公共帧保证匹配双方具有强相似性，降低对复杂匹配方式的依赖性。此外，本文利用多特征融合结构与多任务损失提高边界框回归的精确性，进一步保证匹配过程的准确性。实验结果表明，所设计网络实现了速度与精度的权衡。

本文选取了测试集中2个不同场景下连续3帧的跟踪结果进一步展示多目标跟踪算法的实际效果。图5所示为本文网络在MOT17-03数据集下的可视化跟踪结果示例（彩色效果见《计算机工程》官网HTML版本），此场景为静止的摄像机场景，地处晚间的商业街，光线较昏暗且人流密集度较大，具有一定的跟踪难度。图6为本文网络在MOT17-12数据集下的跟踪结果示例（彩色效果见《计算机工程》官网HTML版本），此场景为运动摄像机视角下的场景。由图5可知，MOT17-03数据集下大部分的目标都可以被成功检测且在跟踪过程中保持身份不变，在人流密集处也能有效地被检测到，但是在严重遮挡环境下存在发生身份切换的可能性，例如图6中第1帧图像身份标识593的目标在第2帧中未被检测到，在第3帧身份标识切换成了594，即在严重遮挡情况下出现了漏检、误检以及身身份切换的现象。在图6中身份标识为140的目标在第1帧没有被检测到，在后2帧才被正确地识别与跟踪。由于运动摄像机下相机与行人均处于运动状态，目标的形状与位置信息在通常情况下具有一定的模糊性，所以会发生误检以及漏检的现象，除了140号小目标之外其余目标都被成功跟踪，这进一步说明了检测结果对整体跟踪性能的影响。实验结果表明，本文设计的网络可以较好地应对场景中的动态变化，在人流较高、光照改变、运动摄像机等复杂场景下仍具有一定的鲁棒性。

(a)跟踪结果1 (b)跟踪结果2 (c)跟踪结果3

图5 本文网络在MOT17-03数据集下的可视化跟踪结果示例

Fig.5 Visual tracking results example of network in this paper under MOT17-03 date set

(a)跟踪结果1 (b)跟踪结果2 (c)跟踪结果3

图6 本文网络在MOT17-12数据集下的可视化跟踪结果示例

Fig.6 Visual tracking results example of network in this paper under MOT17-12 date set

4 结束语

本文设计一种多特征融合的端到端链式多目标跟踪网络，将目标检测、外观特征提取与数据关联集成到一个框架中，并将多特征融合的双向金字塔网络引入框架中，在特征金字塔结构中融入具有采样加权的改进可变形卷积，进一步增加对目标形变的适应能力。本文网络可以根据目标的变化动态调整

感受野,从而提升模型特征提取能力,从整体上改善跟踪的性能。引入focalloss与BalancedL1 Loss两种损失函数进行多任务学习,进一步解决回归任务中正负样本不平衡、简单样本与困难样本梯度贡献差距大的问题,实现网络的均衡学习,提升跟踪的精度与准确度。实验结果表明,本文网络实现了速度与精度的权衡,具有较高的应用价值。但本文网络在数据关联阶段仅使用了IoU匹配,虽然简单的数据关联算法可以提高整体的跟踪速度,但是会影响关联的准确性,导致身份切换的现象发生。下一步将使用级联匹配、图卷积等方法对数据关联阶段进行优化,设计更合理的关联方法,并尝试将该网络应用于其他特定场景中。

参考文献

[1] LUO W H, XING J L, MILAN A, et al. Multiple object tracking: a literature review[EB/OL]. [2021-07-07]. https://arxiv. org/abs/1409. 7618.

[2] BERCLAZ J, FLEURET F, TURETKEN E, et al. Multiple object tracking using *K*-shortest paths optimization[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 33(9): 1806-1819.

[3] PIRSIAVASH H, RAMANAN D, FOWLKES C C. Globally-optimal greedy algorithms for tracking a variable number of objects[C]//Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2011: 1201-1208.

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

[5] HENRIQUES J F, CASEIRO R, MARTINS P, et al. High-speed tracking with kernelized correlation filters[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3): 583-596.

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

[7] WOJKE N, BEWLEY A, PAULUS D. Simple online and realtime tracking with a deep association metric[C]//Proceedings of IEEE International Conference on Image Processing. Washington D. C., USA: IEEE Press, 2017: 3645-3649.

[8] BAE S H, YOON K J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(3): 595-610.

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

[10] BERGMANN P, MEINHARDT T, LEAL-TAIXÉ L. Tracking without bells and whistles[C]//Proceedings of IEEE/CVF International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2019: 941-951.

[11] ZHOU X Y, KOLTUN V, KRÄHENBÜHL P. Tracking objects as points[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 474-490.

[12] WANG Z D, ZHENG L, LIU Y X, et al. Towards real-time multi-object tracking[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 107-122.

[13] ZHANG Y F, WANG C Y, WANG X G, et al. FairMOT: on the fairness of detection and re-identification in multiple object tracking[EB/OL]. [2021-07-07]. https://arxiv. org/abs/2004. 01888.

[14] PENG J L, WANG C G, WAN F B, et al. Chained-tracker: chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2020: 145-161.

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

[16] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector[C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 21-37.

[17] LIN T Y, DOLLÁR P, GIRSHICK R, et al. Feature pyramid networks for object detection[C]//Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 936-944.

[18] LIU S, QI L, QIN H F, et al. Path aggregation network for instance segmentation[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2018: 8759-8768.

[19] DAI J F, QI H Z, XIONG Y W, et al. Deformable convolutional networks[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 764-773.

[20] ZHU X Z, HU H, LIN S, et al. Deformable ConvNets V2: more deformable, better results[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 9300-9308.

[21] LIN T Y, GOYAL P, GIRSHICK R, et al. Focal loss for dense object detection[C]//Proceedings of IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2017: 2999-3007.

[22] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: towards balanced learning for object detection[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2019: 821-830.

[23] BERNARDIN K, STIEFELHAGEN R. Evaluating multiple object tracking performance: the CLEAR MOT metrics[J]. Journal on Image and Video Processing, 2008, 28: 1-11.

[24] RISTANI E, SOLERA F, ZOU R, et al. Performance measures and a data set for multi-target, multi-camera tracking[C]//Proceedings of the 2016 European Conference on Computer Vision. Berlin, Germany: Springer, 2016: 17-35.

[25] SUN S J, AKHTAR N, SONG H S, et al. Deep affinity network for multiple object tracking[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 43(1): 104-119.

[26] PANG B, LI Y Z, ZHANG Y F, et al. TubeTK: adopting tubes to track multi-object in a one-step training model[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2020: 6307-6317.

[27] WANG Y X, KITANI K, WENG X S. Joint object detection and multi-object tracking with graph neural networks[EB/OL]. [2021-07-07]. https://arxiv. org/abs/2006. 13164.

[28] PANG J M, QIU L L, LI X, et al. Quasi-dense similarity learning for multiple object tracking[EB/OL]. [2021-07-07]. https://arxiv. org/abs/2006. 06664.

[29] XU Y H, BAN Y T, DELORME G, et al. TransCenter: transformers with dense representations for multiple-object tracking[EB/OL]. [2021-07-07]. https://arxiv. org/abs/2103. 15145.

编辑　赖玉玲