激光与光电子学进展 先进成像

基于视频行人重识别和时空特征融合的跟踪算法

惠冠程[1]，李开放[1]，辛明[3]，张苗辉[1,2*]

[1]河南大学人工智能学院，河南 开封 475004；

[2]河南大学河南省大数据分析与处理重点实验室，河南 开封 475004；

[3]北京航空航天大学计算机学院，北京 100191

摘要 针对多目标跟踪算法在现实拥堵场景容易引发行人身份交换频繁的问题，提出了一种融合目标检测与行人重识别两个任务的联合网络。同时引入一种用于融合重识别特征和时间信息的轨迹评分机制，该机制通过从检测结果和跟踪预测结果中收集候选目标，互相补充行人目标跟踪预测信息与重识别特征信息。针对视频画面中小目标难以被检测到的问题，对ResNet-34网络进行改进，在主干网络上通过结合深层聚合网络，同时将传统的残差块替换为多级特征卷积网络，实现了对小目标的着重关注，提高了检测准确率。在多目标跟踪数据集MOT16、MOT17、MOT20上进行实验，所提网络的多目标跟踪准确度(MOTA)分别达74.7、73.7、66.4，行人身份转换次数分别为210、209、1403。实验结果表明，所提网络取得了良好的检测跟踪效果。

关键词 目标检测；行人重识别；联合网络；多目标跟踪

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1215004

Tracking Algorithm Based on Video Person Reidentification and Spatiotemporal Feature Fusion

Hui Guancheng[1], Li Kaifang[1], Xin Ming[3], Zhang Miaohui[1,2*]

[1]*School of Artificial Intelligence, Henan University, Kaifeng* 475004, *Henan, China*;

[2]*Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng* 475004, *Henan, China*;

[3]*School of Computer Science and Engineering, Beihang University, Beijing* 100191, *China*

Abstract Multiobject tracking algorithms are frequently affected by the problem of the exchange of pedestrian identity in real congestion situations. To solve this problem, this study proposes a joint network that integrates target detection and person reidentification. Additionally, a track scoring mechanism is introduced to integrate the reidentified feature and time information. By collecting candidates from the detection results and tracking prediction results, the tracking prediction information and reidentified feature information of pedestrian targets can complement each other. To solve the problem of detecting small targets in video images, this study improves the ResNet-34 network by combining the deep aggregation network on the backbone network and replacing the traditional residual block with a multiscale convolutional network to focus on small targets and improve the detection accuracy. In this study, experiments were conducted on the multiobject tracking datasets MOT16, MOT17, and MOT20. The corresponding multiple object tracking accuracy (MOTA) of the proposed network reaches 74.7, 73.7, and 66.4,

收稿日期：2021-04-16；**修回日期**：2021-05-20；**录用日期**：2021-06-11

基金项目：国家自然科学基金(61802111，62002100)、河南省教育厅科学技术研究重点项目(19A50002)

通信作者：*zhmh@henu.edu.cn

respectively, and the conversion durations of pedestrian identity are 210, 209, and 1403, respectively. The results reveal that the proposed network has good detection and tracking performances.

Key words　target detection; person reidentification; joint network; multiobject tracking

1　引　　言

多目标跟踪(MOT)的主要任务是在给定视频中同时对多个感兴趣的目标进行检测定位及跟踪。在实际应用场景中,多目标跟踪算法经常面临着人群拥堵带来的行人间相互遮挡、行人姿态的频繁变换、远处小目标行人所产生的漏检误检、目标边界框不准确等诸多问题。因此,在实际场景中,提出鲁棒而又泛化能力强的在线多目标跟踪算法仍然是一项富有挑战性的任务。

经典的多目标跟踪算法[1-2]通常将行人目标的检测与重识别(re-ID)特征的提取分为两阶段进行,首先进行目标检测用于目标定位,再对检测到的目标进行表观特征的提取,并进一步实现跟踪目标与特征间的数据关联。为了达到实时追踪的效果,文献[3]使用单阶段网络同时完成目标的检测与重识别任务,并取得了较好的实时跟踪效果。基于视频的行人重识别从视频序列图像检索出特定的目标,对视频序列中的行人进行检测跟踪时,行人间的遮挡和目标外观的相似性会导致数据关联的模糊。针对这些问题,文献[4-5]通过对包括运动信息、外观特征和边界框交并比等多条线索进行加权融合来缓解数据关联模糊的问题。文献[2, 6-7]建议使用批处理方法解决不可靠的检测结果,这些方法通过使用整个视频帧或时间窗口中的检测结果来解决全局优化问题,并进一步将检测结果链接到目标的跟踪轨迹。李畅等[4]将跟踪器和目标检测器的输出结果分别作为两个独立的身份信息,并将它们作为候选目标,然后根据传统的诸如光流和颜色直方图等特征选择候选对象。

本文算法首先检测出每帧中感兴趣的目标,然后对目标的位置、尺度、重识别特征等多种线索进行融合,对本帧的检测结果与已有跟踪目标进行数据关联,以此确定视频中出现的每个目标关联的跟踪轨迹。研究内容主要体现在几个方面:首先,提出了一种融合目标检测与行人重识别两个任务的联合网络,该网络可以同时输出目标检测框和重识别特征信息,能够在实现实时追踪的同时,提高检测精度;第二,引入了一种用于融合重识别特征和时间信息的轨迹评分机制,该机制充分利用了目标的表观信息和运动信息的互补性,有效地降低了拥堵场景下行人身份转换次数;第三,针对视频画面中的小目标难以被检测到的问题,对 ResNet-34 网络进行了改进,通过在主干网络上结合深层聚合网络,同时将传统的残差块替换为多级特征卷积网络,实现对小目标的着重关注,提高了对小目标的检测效果。

2　基于多级特征卷积网络的行人跟踪方法

介绍联合网络的主干网络、目标检测分支、重识别分支以及轨迹评分机制。

2.1　联合网络主干结构

首先将 ResNet-34 网络[8]的残差块全部替换为多级特征卷积(MSC)网络之后将其作为联合网络的主干网络,如图 1 所示,并在主干网络融合了深层聚合网络(DLANet)[9]的一种变体。与原始 DLA 不同的是,所采用的聚合网络在低层聚合之

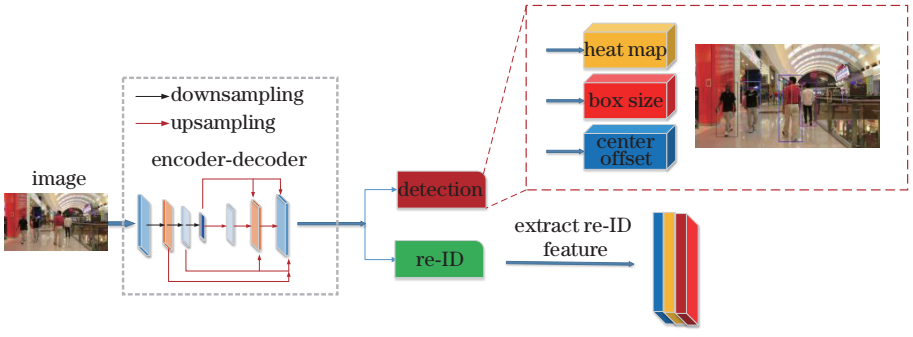

图 1　联合网络结构

Fig. 1　Joint network architecture

间具有更多类似于特征金字塔网络的跳跃连接，使得联合网络可以更好地融合语义特征和空间特征。联合网络将分辨率为1088×608的图像输入编码器-解码器网络，经过编码解码提取出高质量的特征图（步幅设置为4）之后，将特征图送入两个分别用来检测目标和提取重识别特征的平行分支，然后作进一步的数据关联，实现对多目标的跟踪。针对拍摄的视频画面远处行人目标小、像素低、行人目标遮挡引起行人身份转换更加频繁的问题，本文提出MSC网络，该网络使用三种不同尺度的卷积核提取不同大小感受野的特征，并对这些特征进行堆叠融合（concatenation），与传统的残差块相比，MSC网络实现了对小目标的着重关注，提高了目标检测的准确率。为了缓解目标对齐问题，联合网络使用可变形的卷积层替代上采样模块中的常规卷积层，使得联合网络可以根据目标的大小和姿态的变换动态调整感受野。

图2是进一步对编码器-解码器网络的详细描述，将ResNet-34网络stage中的残差块全部替换为MSC网络。图2中的iterative deep aggregation（IDA）模块使得上采样过程可以聚集不同分辨率的特征：

$$I(x_1,\cdots,x_n)=\begin{cases}x_1, & n=1\\ I\left[N(x_1,x_2),\cdots,x_n\right], & n\neq 1\end{cases},\quad(1)$$

式中：I表示IDA模块；$x_1,\cdots,x_n$表示聚集节点N的输入，例如$N(x_1,x_2)$表示一个输入为x_1和x_2的聚集节点N。IDA模块可以使特征由浅到深传播的同时聚集不同深度特征。

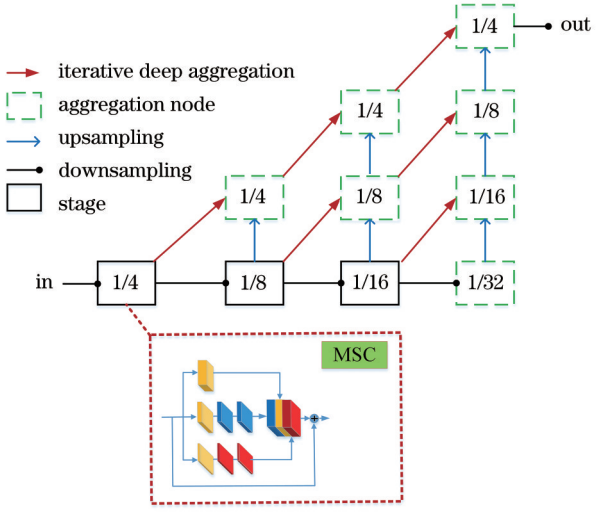

图2　编码器-解码器网络

Fig. 2　Encoder-decoder network

2.2　联合网络检测分支

联合网络的检测分支基于无锚框的CenterNet网络，包含了三个用来估计热图（heat map）、目标中心偏移（center offset）和边界框尺寸（box size）的平行分支，如图3所示，其中C为类别数，H和W分别为图像的高和宽。检测分支通过输出热图损失、边界框大小损失、偏移量损失来确定目标的边界框。目标检测可以看作一个高分辨率特征图上基于目标中心的边界框回归任务，检测分支在估计热图上，根据热图响应$M_{x,y}$执行非极大值抑制来提取峰值关键点，保留热点图得分大于阈值的关键点的位置，然后根据估计的偏移量和边界框大小来计算相应的边界框。

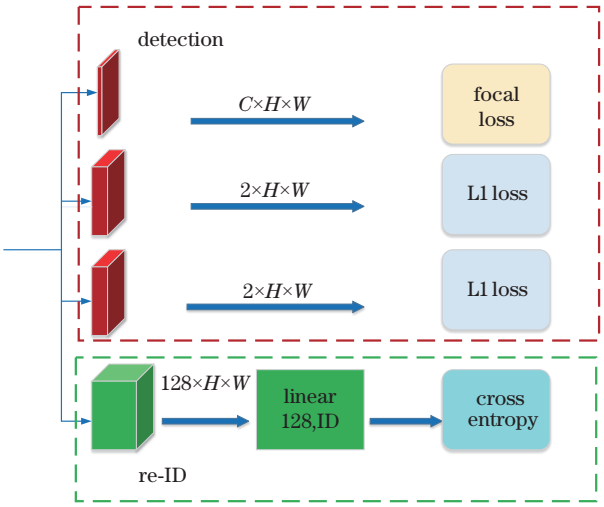

图3　输出heat map，center offset，box size确定边界框信息的检测分支和输出每个ID的分类概率的重识别分支

Fig. 3　Detection branch outputing the heat map, center offset, and box size to determine the information of the bounding box and the re-identify branch outputing the classification probability of each ID

应用图4（d）MSC网络来替换ResNet-34网络的残差块[如图4（c）所示]，通过对上采样之后的特征图应用三种不同大小（1×1，3×3，5×5）的卷积核（256个通道）来平衡每个平行的分支。为了减少参数量，提高推理速度，将3×3卷积核替换为级联的1×3和3×1卷积核，将5×5卷积核替换为两个级联的3×3卷积核，最后送入1×1卷积层。从图4（a）、（b）的输出结果可以看出，与原始残差网络相比，基于MSC网络的主干对远处小目标的检测效果更加优异，能够有效缓解漏检问题。

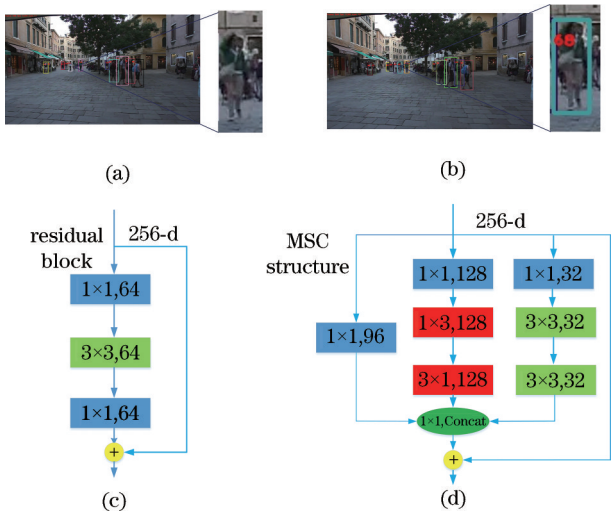

图 4　MSC 网络与原始 ResNet-34 网络跟踪结果的对比。(a)原始 ResNet-34 网络的输出结果;(b) MSC 网络的输出结果;(c)原始 ResNet-34 网络结构图;(d) MSC 网络结构图

Fig. 4　Tracking result comparison between the MSC network and the original ResNet-34 network. (a) Detection result of the original ResNet-34 network; (b) detection result of the MSC network; (c) structure diagram of the original ResNet-34 network; (d) structure diagram of the MSC network

2.3　联合网络检测分支的热图损失

联合网络检测分支的主要功能是估计出物体的中心位置。热图的维度是 $1\times h\times w$,其输出响应的物理意义是,如果热图中的某个点与其真实物体中心相重叠,那么该位置的响应应该为 1,响应会随热图位置与物体中心的距离增大呈指数衰减。

假设图像第 i 个边界框坐标为 $b_i=\left(x_i^1,y_i^1,x_i^2,y_i^2\right)$,计算出对象的中心坐标为 $\left(c_x^i=\frac{x_i^1+x_i^2}{2},c_y^i=\frac{y_i^1+y_i^2}{2}\right)$,它在特征图上对应的位置为 $\left(\overline{c_x^i},\overline{c_y^i}\right)=\left(\left\lfloor\frac{c_x^i}{4}\right\rfloor,\left\lfloor\frac{c_y^i}{4}\right\rfloor\right)$,其中 4 表示步长。在坐标 (x,y) 处的热图响应为

$$M_{x,y}=\sum_{i=1}^{I}\exp\left[-\frac{\left(x-\overline{c_x^i}\right)^2-\left(y-\overline{c_y^i}\right)^2}{2\sigma_c^2}\right],\quad(2)$$

式中:I 代表图像中物体的数量;σ_c 为标准差。

热图损失函数定义为

$$L_{\text{heat}}=-\frac{1}{I}\sum_{xy}\begin{cases}\left(1-\widehat{M_{xy}}\right)^{\alpha}\log_{10}\left(\widehat{M_{xy}}\right), & M_{xy}=1\\ \left(1-\widehat{M_{xy}}\right)^{\beta}\widehat{M_{xy}}^{\alpha}\log_{10}\left(1-\widehat{M_{xy}}\right), & M_{xy}\neq1\end{cases},\quad(3)$$

式中:$\widehat{M}$ 是估计热图的真实值;α 和 β 均为超参数。

2.4　联合网络检测分支的边界框大小和偏移量损失

计算边界框的偏移量是为了更加精确边界框位置,由于最终特征图的步长为 4,那就会有 4 个像素单位的量化误差,偏移量分支通过估计每个像素相对目标中心的连续偏移量来减轻下采样的影响。计算边界框的尺寸后可以估算出目标边界框的高度和宽度。边界框大小和偏移量分支的输出分别表示为 $\hat{\boldsymbol{s}}\in\mathbb{R}^{2\times h\times w}$ 和 $\hat{\boldsymbol{o}}\in\mathbb{R}^{2\times h\times w}$,$w$ 和 h 分别表示特征图的宽和高。对于图像中的第 i 个边界框 $b_i=\left(x_i^1,y_i^1,x_i^2,y_i^2\right)$,其边界框大小计算为 $\boldsymbol{s}_i=\left(x_i^2-x_i^1,y_i^2-y_i^1\right)$,边界框的偏移量计算为 $\boldsymbol{o}_i=\left(\frac{c_x^i}{4},\frac{c_y^i}{4}\right)-\left(\left\lfloor\frac{c_x^i}{4}\right\rfloor,\left\lfloor\frac{c_x^i}{4}\right\rfloor\right)$。分别将对应位置评估的尺寸和偏移表示为 $\hat{\boldsymbol{s}}_i$ 和 $\hat{\boldsymbol{o}}_i$,然后对两个分支实施 L1 损失:

$$L_{\text{box}}=\sum_{i=1}^{I}\left\|\boldsymbol{o}_i-\hat{\boldsymbol{o}}_i\right\|_1+\left\|\boldsymbol{s}_i-\hat{\boldsymbol{s}}_i\right\|_1。\quad(4)$$

2.5　联合网络重识别分支

重识别分支通过提取目标对象的 re-ID 特征,获得目标对象的表观信息,将不同身份的对象区分开来。在理想的情况下,不同类别目标之间的距离应当大于同一类别目标的距离[10],选择合适的度量方法可以区分出不同类别的目标。如图 3 所示,重识别分支使用了 128 个卷积核的卷积层,以提取主干特征图上每个位置的身份嵌入特征,之后生成的特征图为 $\boldsymbol{E}\in\mathbb{R}^{128\times w\times h}$。从特征图 $\boldsymbol{E}$ 中提取目标中心在 (x,y) 处的身份嵌入特征 $\boldsymbol{E}_{x,y}\in\mathbb{R}^{128}$。

学习表观特征完成行人身份分配可以看作是一项分类任务[11],训练集中具有相同身份的所有目标实例都被视为相同类别。对于图像中的第 i 个边界框 $b_i=\left(x_i^1,y_i^1,x_i^2,y_i^2\right)$,它在特征图上的中心位置是 $\left(\overline{c_x^i},\overline{c_y^i}\right)$,在这个位置上提取身份嵌入特征向量 $\boldsymbol{E}_{\overline{c_x^i},\overline{c_y^i}}$,并将它映射到一个类分布向量 $\boldsymbol{P}=\left\{p_{(k)},k\in[1,K]\right\}$,目标实例的类标签独热码表示为 $L_{(k)}^i$,重识别的损失函数计算为

$$L_{\text{identity}}=-\sum_{i=1}^{I}\sum_{k=1}^{K}L_{(k)}^{i}\log_{10}\left(p_{(k)}\right),\tag{5}$$

式中：K表示类别的数量。

2.6　联合网络训练

任务依赖型不确定性是一种偶然不确定性，指的是一个对于所有输入数据保持不变的量，它会在不同的任务之间变化。在多任务学习中，任务依赖型不确定性表明了各个任务之间的相对置信度，分类和回归问题中存在这种固有的不确定性，本文基于检测与重识别任务之间的任务依赖型不确定性，参照文献[12]的不确定性损失函数，构造一个整体的损失函数L_{total}来训练联合网络。

$$L_{\text{detection}}=L_{\text{heat}}+L_{\text{box}},\tag{6}$$

$$L_{\text{total}}=\frac{1}{2}\left[\frac{1}{\exp(a_1)}L_{\text{detection}}+\frac{1}{\exp(a_2)}L_{\text{identity}}+a_1+a_2\right],\tag{7}$$

式中：a_1和a_2是两个用来平衡检测与重识别任务的可学习参数。

2.7　轨迹评分机制

所提算法基于在线跟踪算法，由第一帧视频画面中的检测目标框来初始化多个追踪轨迹，在随后的时间帧中，根据从身份嵌入特征上计算出的余弦距离和边界框的双向匹配重合度，对当前帧检测得到的目标框与现有的跟踪轨迹进行数据关联[13]。除了表观信息，数据关联部分还应该考虑目标的运动信息，给定一个新视频帧，使用卡尔曼滤波器[14]预测每个现有目标在新视频帧的位置。这些预测用于确定目标表观信息的突变和拥堵场景的遮挡引发检测失败时的目标位置。但是算法如果仅靠预测来确认追踪的结果，就不能够满足长期追踪任务的需求，如果已被跟踪的目标长时间未被检测到，卡尔曼滤波器的参数就无法得到更新，跟踪预测的结果会变得不再可信[15-16]。所提数据关联策略对检测和跟踪结果进行对比，选择输出最优解，如图5所示，检测结果被可视化为实线矩形框，其输出候选框的评分函数曲线为实线；跟踪结果被可视化为虚线矩形框，其输出候选框的评分函数曲线为虚线。从左起第三列图像可以明显看出，在短期检测失败的情况下，跟踪预测可以有效解决轨迹跟丢问题。

目标跟踪轨迹是通过关联来自连续视频帧的检测候选框生成的，因为跟踪轨迹可能在其生命周期中被中断和检索多次，所以可以将单个跟踪轨迹

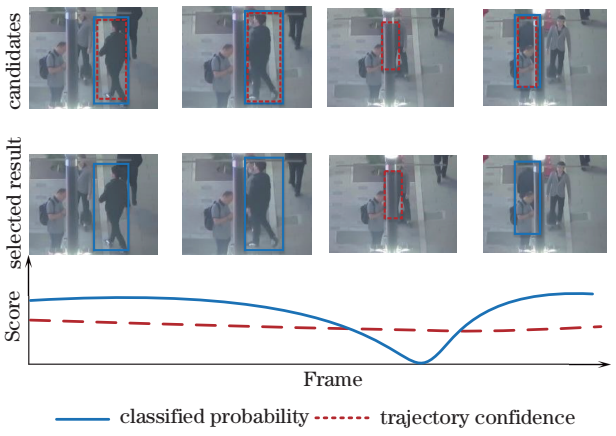

图5　基于统一评分机制的候选框选择

Fig. 5　Candidate box selection based on unified scoring mechanism

分成一组时间序列的小轨迹。每次从丢失状态检索轨迹时，卡尔曼滤波器将被重新初始化，因此可以仅利用最后一个小轨迹的信息来表示大轨迹的置信度。定义轨迹置信度为

$$S_{\text{trk}}=\begin{cases}\max\left[1-\log_{10}\left(1+\delta L_{\text{trk}}\right),0\right], & L_{\text{det}}\geqslant 2\\ 0, & L_{\text{det}}<2\end{cases},\tag{8}$$

式中：L_{det}表示一个目标轨迹生成之前有检测结果关联到该轨迹的次数；L_{trk}表示距离上一次有检测结果关联到该轨迹的视频帧数；δ是超参数。式(8)表明，L_{trk}越大，目标没有被再次检测到的次数越多，仅靠预测维持的轨迹帧数越多，轨迹置信度就越低，那么S_{trk}就越小，这样可以防止跟踪结果的漂移。评分函数由分类概率$P_{(k)}$和轨迹置信度S_{trk}组合得到，定义为

$$S=\omega p_{(k)}+(1-\omega)S_{\text{trk}},\tag{9}$$

式中：ω为平衡分类概率$P_{(k)}$和轨迹置信度S_{trk}的权重参数。

3　实验讨论与分析

3.1　数据集和度量方法

数据集CalTech[17]、MOT17[18]、CUHK-SYSU[19]、PRW[20]标注了身份信息和目标边界框，可以同时训练检测和重识别两个分支网络。训练完成后，在三个广泛使用的多目标跟踪数据集MOT16[18]、MOT17[18]、MOT20[21]上测试了所提算法，这三个数据集都是真实摄像头拍摄的画面，包括许多拥挤场景下的多目标追踪任务，与实际应用场景相符。采用多目标跟踪算法普遍使用的评价指标，多目标跟踪准确度(MOTA)、IDF1、身份转换次数(IDs)、轨

迹命中率(MT)和轨迹丢失率(ML)。

$$P_{\mathrm{MOTA}}=1-\frac{\sum_t\left(N_{\mathrm{m}t}+N_{\mathrm{fp}t}+N_{\mathrm{mme}t}\right)}{\sum_t N_{\mathrm{g}t}},\quad(10)$$

式中：$N_{\mathrm{m}t}$，$N_{\mathrm{fp}t}$，$N_{\mathrm{mme}t}$和$N_{\mathrm{g}t}$分别是t帧时的错误检测，遗漏检测，错误匹配的数量和总目标数。

3.2　实验环境和参数设置

本实验采用MSC网络作为联合网络的主干网络，在COCO检测数据集[22]上进行预训练来初始化模型参数，使用Adam优化器对模型进行了30个epoch的训练，起始学习率为0.0001，在20个epoch时下降到0.00001，batch size设置为6。实验使用了标准的数据增强方法，包括缩放、旋转、平移和剪切等。输入图像的分辨率大小调整为1088×608，下采样率为4，特征图的分辨率为272×152。实验的开发环境如表1所示。

表1　实验平台参数

Table 1　Experimental platform parameters

Configuration	Parameter
Operating system	Ubuntu 16.04
RAM(random processing unit)	128 G
CPU(central processing unit)	2.50 GHz E5-2678 v3
GPU(graphics processing unit)	Tesla T4 16 G
Software platform	Pytorch 1.1 Python 3.6

3.3　消融实验

大多数一步法跟踪器使用的是512维的重识别特征。出于对实时性能、平衡检测精度和跟踪准确度的考虑，对不同特征维度的重识别分支进行对比实验。如表2所示，512维的重识别特征获得了最高的IDF1分数，说明了使用高维的重识别特征可以获得更强的辨别能力，但是当特征维度降到64维的过程中，MOTA的分数一直在上升。因此，设置重识别特征的维度时要考虑到检测和重识别两个任务之间的冲突，高维的特征有利于重识别任务，低维度特征有利于检测任务及推理速度的提升。通过分析实验结果，将重识别特征维数设置为128，以平衡两个任务。

跟踪算法[13]数据关联部分通常使用三个要素，边界框交并比(Box IoU)、重识别特征(re-ID Features)、卡尔曼滤波器(Kalman Filter)，来计算每对检测到的目标边界框的相似度，然后采用匈牙利算法分配行人身份。表3展示了所提数据关联方法的实验结果，只使用边界框交并比会导致较多的行人身份转换次数，尤其是在拥挤和相机快速运动的场景下；单独使用重识别特征会显著增加IDF1并减少行人身份转换次数；此外，添加卡尔曼滤波器使得行人身份转换次数减少并且有助于获得平滑的跟踪轨迹。但是当一个物体被部分遮挡时，重识别特征变得不可靠，在这种情况下，所提轨迹评分机制可以使数据关联更加侧重可信的运动信息，可以进一步减少行人身份转换次数。

表2　MOT17验证集上评估重识别特征维度

Table 2　Recognition feature dimensions evaluated on the MOT17 validation set

Dimension	MOTA	IDF1	IDs	Time /s
512	68.5	73.7	312	24.1
256	68.5	72.8	337	26.1
128	69.1	72.5	299	26.6
64	69.2	72.3	283	26.8

表3　MOT17验证集上评估数据关联的三个要素的效果

Table 3　Evaluation of the three elements associated with the evaluation data on the MOT17 validation set

Box IoU	re-ID Features	Kalman Filter	MOTA	IDF1	IDs
√			67.8	67.2	648
	√		68.1	70.3	435
	√	√	68.9	71.8	342
√	√	√	69.1	72.8	299

通过表4可以看出，MSC网络相比于ResNet-34网络实现了更好的检测及跟踪性能，表明了所提联合网络对提取和融合行人多级特征的有效性。

表4　MOT17验证集上对比MSC网络与ResNet-34网络

Table 4　Comparison between MSC network and ResNet-34 network on MOT17 validation set

Network	MOTA	IDF1	IDs
ResNet-34	63.6	67.2	435
MSC	69.1	72.8	299

表5为使用不同数据集对联合网络进行训练的结果，表中的“MIX”表示使用CalTech，MOT17，CUHK-SYSU和PRW四个数据集混合训练后的结果。

表 5　使用不同数据集进行训练的结果

Table 5　Result using different datasets for training

Dataset	Number of images	Number of boxes	Number of identities	MOTA	IDF1	IDs
MOT17	5×10^3	112×10^3	0.5×10^3	69.1	72.8	299
MIX	54×10^3	270×10^3	8.7×10^3	73.7	80.1	209

3.4　实验结果

为了验证所提算法的有效性，与目前主流的方法进行了对比实验，实验结果如表 6 所示，可以明显地看出，所提算法的检测结果的可信度有较大提升，轨迹跟丢的情况也得到了很大改善。如图 6 所示，每行按视频序列的时间顺序显示采样帧的结果，图像中标记了边界框和身份，不同颜色的边框代表不同的身份。从 MOT17 测试集的输出结果上可以看出，有行人短暂重叠时，轨迹评分机制会偏向输出行人运动信息得到的预测结果框，避免了遮挡导致的漏检误检情况，在这种情况下，传统的跟踪策略，比如仅使用边界框交并比的跟踪器往往会导致行人身份切换。在图 6 MOT17 测试集上的输出结果表明，在行人较多的拥挤场景下，所提算法的跟踪效果表现良好，既能保持正确的边界框，又能保持身份不变，这主要归功于 MSC 网络对小目标的敏感性。

表 6　不同方法的结果对比

Table 6　Comparison of results of different methods

Dataset	Tracker	MOTA	IDF1	MT /%	ML /%	IDs	Time /s
MOT16	EAMTT[23]	52.5	53.3	19.9	34.9	910	<5.5
	SORTwHPD16[24]	59.8	53.8	25.4	22.7	1423	<8.6
	DeepSORT_2[25]	61.4	62.2	32.8	18.2	781	<6.4
	RAR16wVGG[26]	63.0	63.8	39.9	22.1	482	<1.4
	VMaxx[27]	62.6	49.2	32.7	21.1	1389	<3.9
	TubeTK[28]	64.0	59.4	33.5	19.4	1117	1.0
	JDE[3]	64.4	55.8	35.4	20.0	1544	18.5
	TAP[29]	64.8	73.5	38.5	21.6	571	<8.0
	CNNMTT[30]	65.2	62.2	32.4	21.3	946	<5.3
	POI[31]	66.1	65.1	34.0	20.8	805	<5.0
	CTackerVI[32]	67.6	57.2	32.9	23.1	1897	6.8
	Proposed method	74.7	80.2	38.10	21.47	210	13.3
MOT17	SST[33]	52.4	49.5	21.4	30.7	8431	<3.96
	TubeTK[28]	63.0	58.6	31.2	19.9	4137	3.0
	CTackerVI[32]	66.6	57.4	32.2	24.2	5529	6.8
	CenterTack[34]	67.3	59.9	34.6	24.6	2898	22.0
	Proposed method	73.7	80.1	36.99	22.89	209	12.70
MOT20	ArTIST-T[35]	53.6	51.0	31.6	28.1	1531	
	MPNTrack[36]	57.6	59.1	38.2	22.5	1210	
	Proposed method	66.4	72.8	46.87	14.84	1403	12.0

与传统的 ResNet-34 网络相比，MSC 网络参数量有所增加，必然会导致推理速度的下降，对此进行关于推理速度的对比实验。如图 7 所示，分别在数据集 MOT16，MOT17，MOT20 上对比了两种网络的推理速度，在 MOT16，MOT17 数据集上，行人目标尺寸分布比较分散，而 MOT20 数据集上行人目标尺寸较为单一。MSC 网络与 ResNet-34 在 MOT20 数据集的推理时间差别并没有在 MOT16，MOT17 那么明显。结合表 6 实验数据可以看出：MSC 网络在行人目标尺寸分布比较分散的视频画面中的检测准确率比较高，但是会导致推理速度的下降；在行人目标尺寸较为单一的场景中，MSC 网络的检测准确率提升没有在行人目标尺寸分布分散的场景中那么明显，所以推理速度下降不太显著。

图 6　所提方法在MOT17测试集上的输出结果

Fig. 6　Output results of the proposed method on MOT17 test set

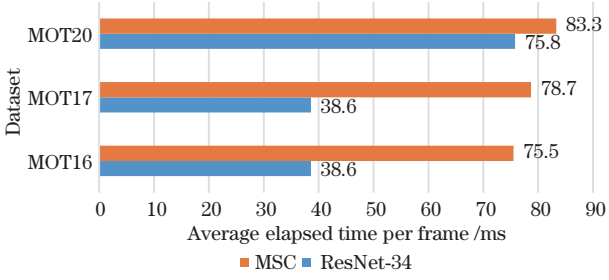

图 7　MSC网络和ResNet-34网络在三个数据集上的推理时间

Fig. 7　Reasoning time of MSC network and ResNet-34 network on three data sets

4　结　　论

以ResNet-34网络为基础，提出了一种融合行人检测与重识别的联合网络，同时引入基于时空信息互补性的轨迹评分机制。该机制通过对运动信息和行人重识别特征进行融合筛选，弥补短期运动信息缺失的目标表观信息，从而有效缓解了拥堵场景下行人遮挡导致行人身份转换频繁的问题。与传统ResNet-34网络相比，所提MSC网络能够针对不同大小的目标提取更加合理的多级特征，从而提高了小目标检测精度，缓解了视频画面远处行人难以被检测到的问题。实验结果表明，所提方法在数据集MOT16，MOT17，MOT20上的性能与目前主流方法相比检测精度有了明显的提高，身份转换次数显著下降，表明了所提方法的有效性。

参　考　文　献

[1] 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.

[2] Fagot-Bouquet L, Audigier R, Dhome Y, et al. Improving multi-frame data association with sparse representations for robust near-online multi-object tracking[M]//Leibe B, Matas J, Sebe N, et al. Computer vision-ECCV 2016. Lecture notes in computer science. Cham: Springer, 2016, 9912: 774-790.

[3] Wang Z D, Zheng L, Liu Y X, et al. Towards real-time multi-object tracking[EB/OL]. (2019-09-27)[2021-05-09]. https://arxiv.org/abs/1909.12605.

[4] 李畅, 杨德东, 宋鹏, 等. 基于全局感知孪生网络的红外目标跟踪[J]. 光学学报, 2021, 41(6): 0615002.
Li C, Yang D D, Song P, et al. Global-aware siamese network for thermal infrared object tracking [J]. Acta Optica Sinica, 2021, 41(6): 0615002.

[5] 李福进, 刘慧慧, 任红格, 等. 高置信度的尺度自适应核相关跟踪方法[J]. 激光与光电子学进展, 2021, 58(8): 0815004.
Li F J, Liu H H, Ren H G, et al. Scale adaptive kernel correlation tracking method with high

confidence[J]. Laser & Optoelectronics Progress, 2021, 58(8): 0815004.

[6] Kim C, Li F X, Ciptadi A, et al. Multiple hypothesis tracking revisited[C]//2015 IEEE International Conference on Computer Vision (ICCV), December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4696-4704.

[7] Tang S Y, Andriluka M, Andres B, et al. Multiple people tracking by lifted multicut and person re-identification[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3701-3710.

[8] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[9] Zhou X Y, Wang D Q, Krähenbühl P. Objects as points[EB/OL]. (2019-04-16)[2021-06-01]. https://arxiv.org/abs/1904.07850.

[10] 刘可文, 房攀攀, 熊红霞, 等. 基于多层级特征的行人重识别[J]. 激光与光电子学进展, 2020, 57(8): 081503.
Liu K W, Fang P P, Xiong H X, et al. Person re-identification based on multi-layer feature[J]. Laser & Optoelectronics Progress, 2020, 57(8): 081503.

[11] 罗浩, 姜伟, 范星, 等. 基于深度学习的行人重识别研究进展[J]. 自动化学报, 2019, 45(11): 2032-2049.
Luo H, Jiang W, Fan X, et al. A survey on deep learning based person re-identification[J]. Acta Automatica Sinica, 2019, 45(11): 2032-2049.

[12] Cipolla R, Gal Y, Kendall A. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7482-7491.

[13] Kuhn H W. The Hungarian method for the assignment problem[J]. Naval Research Logistics Quarterly, 1955, 2(1/2): 83-97.

[14] Welch G, Bishop G. An introduction to the Kalman filter[EB/OL]. (2006-07-24)[2021-05-06]. https://is.muni.cz/el/1431/podzim2014/Bi0440/um/kalman_intro.pdf?lang=cs.

[15] Bolme D S, Beveridge J R, Draper B A, et al. Visual object tracking using adaptive correlation filters[C]//2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, June 13-18, 2010, San Francisco, CA, USA. New York: IEEE Press, 2010: 2544-2550.

[16] 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.

[17] Dollar P, Wojek C, Schiele B, et al. Pedestrian detection: a benchmark[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE Press, 2009: 304-311.

[18] Milan A, Leal-Taixé L, Reid I, et al. MOT16: a benchmark for multi-object tracking[EB/OL]. (2016-03-02)[2021-05-12]. https://arxiv.org/abs/1603.00831.

[19] Xiao T, Li S, Wang B C, et al. Joint detection and identification feature learning for person search[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3376-3385.

[20] Zheng L, Zhang H H, Sun S Y, et al. Person re-identification in the wild[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE Press, 2017: 3346-3355.

[21] Dendorfer P, Rezatofighi H, Milan A, et al. MOT20: a benchmark for multi object tracking in crowded scenes[EB/OL]. (2020-03-19)[2021-05-01]. https://arxiv.org/abs/2003.09003.

[22] Bewley A, Ge Z Y, Ott L, et al. Simple online and realtime tracking[EB/OL]. (2016-02-02) [2021-05-02]. https://arxiv.org/abs/1602.00763.

[23] Sanchez-Matilla R, Poiesi F, Cavallaro A. Online multi-target tracking with strong and weak detections [M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 84-99.

[24] Lin T Y, Maire M, Belongie S, et al. Microsoft COCO: common objects in context[M]//Fleet D, Pajdla T, Schiele B, et al. Computer vision-ECCV 2014. Lecture notes in computer science. Cham: Springer, 2014, 8693: 740-755.

[25] Wojke N, Bewley A, Paulus D. Simple online and realtime tracking with a deep association metric[C]//2017 IEEE International Conference on Image Processing, September 17-20, 2017, Beijing, China. New York: IEEE Press, 2017: 3645-3649.

[26] Fang K, Xiang Y, Li X C, et al. Recurrent autoregressive networks for online multi-object tracking[C]//2018 IEEE Winter Conference on Applications of Computer Vision, March 12-15, 2018, Lake Tahoe, NV, USA. New York: IEEE Press, 2018: 466-475.

[27] Wan X Y, Wang J J, Kong Z F, et al. Multi-object tracking using online metric learning with long short-term memory[C]//2018 25th IEEE International Conference on Image Processing, October 7-10, 2018, Athens, Greece. New York: IEEE Press, 2018: 788-792.

[28] Pang B, Li Y Z, Zhang Y F, et al. TubeTK: adopting tubes to track multi-object in a one-step training model[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6307-6317.

[29] Zhou Z W, Xing J L, Zhang M D, et al. Online multi-target tracking with tensor-based high-order graph matching[C]//2018 24th International Conference on Pattern Recognition (ICPR), August 20-24, 2018, Beijing, China. New York: IEEE Press, 2018: 1809-1814.

[30] 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.

[31] Yu F W, Li W B, Li Q Q, et al. POI: multiple object tracking with high performance detection and appearance feature[M]//Hua G, Jégou H. Computer vision-ECCV 2016 workshops. Lecture notes in computer science. Cham: Springer, 2016, 9914: 36-42.

[32] 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[M]//Vedaldi A, Bischof H, Brox T, et al. Computer vision-ECCV 2020. Lecture notes in computer science. Cham: Springer, 2020, 12349: 145-161.

[33] 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.

[34] Zhou X Y, Koltun V, Krähenbühl P. Tracking objects as points[EB/OL]. (2020-04-02) [2021-05-09]. https://arxiv.org/abs/2004.01177.

[35] Saleh F, Aliakbarian S, Rezatofighi H, et al. Probabilistic tracklet scoring and inpainting for multiple object tracking[C]//2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 20-25, 2021, Nashville, TN, USA. New York: IEEE Press, 2021: 14324-14334.

[36] Brasó G, Leal-Taixé L. Learning a neural solver for multiple object tracking[C]//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 13-19, 2020, Seattle, WA, USA. New York: IEEE Press, 2020: 6246-6256.