DOI:10.19651/j.cnki.emt.2209796

基于多特征匹配的轻量化行人跟踪算法研究*

安胜彪 刘新宇 白 宇
(河北科技大学信息科学与工程学院 石家庄 050018)

摘 要: 行人跟踪是深度学习研究中的热点内容。目前的跟踪算法存在无法满足实时性和因跟踪目标相似度太高、目标间的遮挡、运动不规律造成 ID 频繁转换的问题。为了提高运行速度,在目标检测阶段使用 CNN 和 transformer 相结合的轻量化网络,采用联合检测的方式,共享特征权重,并行计算检测、重识别、人体姿态估计分支,同时调整各个分支卷积通道数。跟踪部分则利用卡尔曼滤波预测的目标运动信息,目标重识别信息,和目标姿态的各个关键点位置信息共同完成目标身份匹配,减少了同一 ID 的频繁转换。实验部分采用 MOT16 数据集训练和测试。本算法的多目标跟踪准确度(MOTA)为 48.5%,多目标跟踪精确度(MOTP)为 78.17%,FPS 为 20,模型大小为 18.4 M。实验表明,提出的跟踪算法提高了整体的跟踪性能,实时性和准确性达到了预期要求。

关键词: 轻量化网络;多特征匹配;行人跟踪
中图分类号: TP391 **文献标识码:** A **国家标准学科分类代码:** 510.4050

Research on lightweight pedestrian tracking algorithm based on multi-feature matching

An Shengbiao Liu Xinyu Bai Yu
(School of Information Science and Engineering, Hebei University of Science and Technology, Shijiazhuang 050018, China)

Abstract: Pedestrian tracking is a hot topic in deep learning research. The current tracking algorithm has the problems that it cannot meet the real-time performance and frequent ID conversion due to the high similarity of the tracking targets, the occlusion between the targets, and the irregular motion. In order to improve the running speed, a lightweight network combining CNN and transformer is used in the target detection stage, and a joint detection method is adopted to share feature weights, calculate detection, re-identification, and human pose estimation branches in parallel, and adjust the number of convolution channels of each branch at the same time. . The tracking part uses the target motion information predicted by Kalman filtering, the target re-identification information, and the position information of each key point of the target pose to complete the target identity matching, which reduces the frequent conversion of the same ID. The experimental part uses the MOT16 dataset for training and testing. The multi-target tracking accuracy (MOTA) of this algorithm is 48.5%, the multi-target tracking accuracy (MOTP) is 78.17%, the FPS is 20, and the model size is 18.4 M. Experiments show that the proposed tracking algorithm improves the overall tracking performance, and the real-time performance and accuracy meet the expected requirements.

Keywords: lightweight network; multi-feature matching; pedestrian tracking

0 引 言

行人跟踪是智能视频监控系统的重要环节,在智慧交通、安全监控等方面有着重要的应用价值[1-3]。行人跟踪第一步是行人检测,即检测图片或者视频帧中是否有行人,如果存在行人,则标记出行人的位置坐标[4]。不同于单纯的行人检测,行人跟踪是在检测任务上的扩展,它最终的目的,是利用行人检测阶段提取到的不同目标的特征信息,匹配前后两帧中具有极高相似度的目标,从而对视频中的行人目标保持长时间的跟踪,获得行人运动的速度、轨迹和方向等信息[5]。

现实场景中如需跟踪的目标,具有形变、相互遮挡的特点[6]。目前的目标跟踪框架多为 DBT(detection by tracking)模式[7],该模式首先通过神经网络检测出行人的

收稿日期:2022-04-29
* 基金项目:河北省自然科学基金(F2019208305)、国家自然科学基金(61902108)项目资助

位置，然后利用 SORT[8] 或者 DeepSORT[9] 算法完成数据级联匹配。在 DBT 模式下，Faster R-CNN[10]、YOLO[11] 和 DETR[12] 目标检测算法表现出良好的性能。Faster R-CNN 得益于区域生成网络（RPN）网络的加入，它首先产生候选框，再对候选框做分类和回归，这种两步检测算法具有非常高的精确度。YOLO 不同于 Faster-RCNN 选取感兴趣区域的两步检测方法，检测图片被划分为多个格子，通过检测目标中心所在的格子，便可得到检测框和类别信息目标识别的速度大大提高。DETR 采用了 Transformer 的结构，考虑到了目标和图像间全局上下文的关系，不需要 YOLO 中的先验框信息，提高了识别精度。DeepSORT 利用检测算法提取到的特征信息，再融入每个目标独特的表观特征完成级联匹配，更擅长处理目标重叠情况，但也降低运行效率。

DBT 模式虽然有着广泛的应用，但是从它的结构来看，DBT 模式中的行人特征提取，需要使用额外的网络模型。虽然有些工作努力使这个网络轻量化，例如，Shufflenet[13] 网络和 Mobilenet[14] 网络，但是仍无法保证实时性。最近两年发展的新工作，整合了 DBT 模式中单独用来提取深度特征的部分，即 JDT（joint detection tracking）模式。

JDT 模式的经典网络是 JDE[15]。它使用了基于 Anchor 聚类的 Yolov3[16] 作为检测网络，并且增强不同职能模块间的耦合度。Fairmot[17] 算法则采用了基于 Anchor-free 检测网络[18]，改善了人群密度太大时，基于 Anchor 的方法无法对齐行人特征的问题，进一步增强了 JDT 模式的准确性。此外，还有把行人目标的运动特征和检测网络结合的工作。Tracktor＋＋[19] 框架，不再利用卡尔曼滤波对目标轨迹进行预测，在检测的同时回归出跟踪框，再通过简单的关联算法完成轨迹的匹配，这种模式带来了效率的提升。CTrack[20] 在同一个网络结构中融合了目标定位、特征提取、匹配关联任务，借助视频序列连续的特点，以及多种注意力机制，使目标跟踪变得简单快捷。

本论文的主要贡献如下：

1）在 JDT 模式联合检测行人深度特征的基础上，检测行人关键点信息，增强模型面对复杂现实环境时的鲁棒性。

2）采用 Anchor-free 的网络框架以及轻量化的骨干网络，确保了精确度和实时性。

3）优化多种特征匹配方式，解决人员密集和遮挡严重的问题。

1　轻量化网络模型

为了降低网络参数，同时保持检测准确率，本文采用基 Anchor-free 的中心关键点检测模型作为检测部分的主体框架，骨干网络采用 CNN 和 Transformer[21] 相结合的 Mobilevit。Mobilevit 结构兼顾了局部和全局特征，而且避免了以往 Transformer 参数量太大的缺点。上采样采用三层反卷积结构，增大了特征图的分辨率。同时，对骨干网络和反卷积使用短路连接，以获得更加丰富的语义信息。模型头部采用多个不同维度的平行分支，分别预测出行人的位置信息，重识别信息和每个行人目标 17 个关键点的位置信息。同时，调整各个头部分支的卷积通道数，达到减少参数计算量的目的。

1.1　骨干网络

骨干网络负责从图像中提取目标的特征信息，后续的网络结构利用这些特征信息实现分类和定位。Mobilevit 中的轻量级卷积神经网络，由于空间归纳偏置的特性可以在不同的视觉任务中学习局部的表征。Mobilevit 中基于自注意力（Self-attention）的 Transfomer 能够反映复杂的空间变换和长距离依赖，从而获得一种全局特征的表示。在行人跟踪中，通过轻量级神经网络和 Transformer 的结合，可以极大的提高行人目标局部特征的感知能力，以及在全局中表示的行人目标的局部细节。

Mobilevit 的结构如图 1 所示，首先输入的图片经过一个 3＊3 的卷积，然后经过 5 个 MV2 卷积模块进行下采样，紧接着由 Mobilevit block 和 MV2 卷积模块交叉堆叠。MV2 模块先使用扩张层把纬度从 C 扩展到 4C，深度可分离卷积提取特征，然后映射层来降维让网络重新变小[22]。从高维向低维转换的过程中，为了防止信息丢失或破坏，因此采用了非线性激活函数 Relu6，该激活函数还限制了激活范围，减少了精度损失。Mobilevit block 将尺寸为 H（高）$\times W$（宽）$\times C$（通道数）的特征图进行展平，得到维度为 P（$H\times W$ 所切图片块的大小）$\times N$（$H\times W/P$ 图片块总数）$\times C$（通道数）的输出，然后经过数个堆叠的基于 self-attention 的 Transformer 进行全局信息关注，最后经过堆积操作把输出尺寸还原为输入特征图的 $H\times W\times C$。其中自注意力的计算公式为：

$$\mathrm{Attention}(Q,K,V)=\mathrm{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_K}}\right)V \tag{1}$$

1.2　基于中心点的多分支检测网络

基于中心点的多分支的行人检测总体网络结构如图 2 所示。通过设计多个不同功能的检测头，可以同时预测出精准的行人的位置，重识别特征，以及姿态关键点的位置。这种联合的检测方法，极大的提高了推理的速度，与网络设计时的要求相符合。基于中心点的检测，避免了基于框的检测在进行行人重识别特征提取时，因提取的特征和对象中心未对齐所带来的歧义。骨干网络把提取到的特征图通过短路连接送入 3 个堆叠的反卷积网络，反卷积层把浅层和深层的特征图结合，可以提高小目标的分辨率，经过反卷积层后的特征图变为输入尺寸的 1/4。相比于经过骨干网络之后的特征图尺寸，反卷积层后的特征图尺寸是其 8 倍。这样的设计，使得行人重识别特征更加丰富，相同身份的行人目标可以聚集在一起，不同身份的行人目标则相互远离。

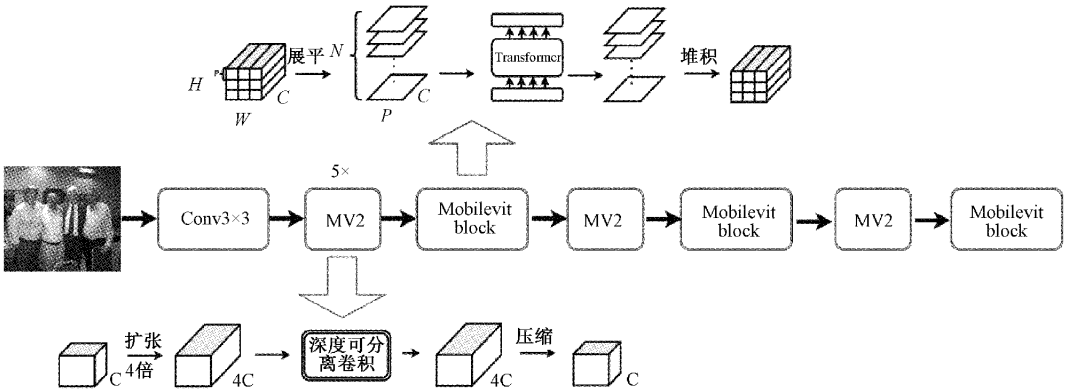


图 1 Mobilevit 结构图

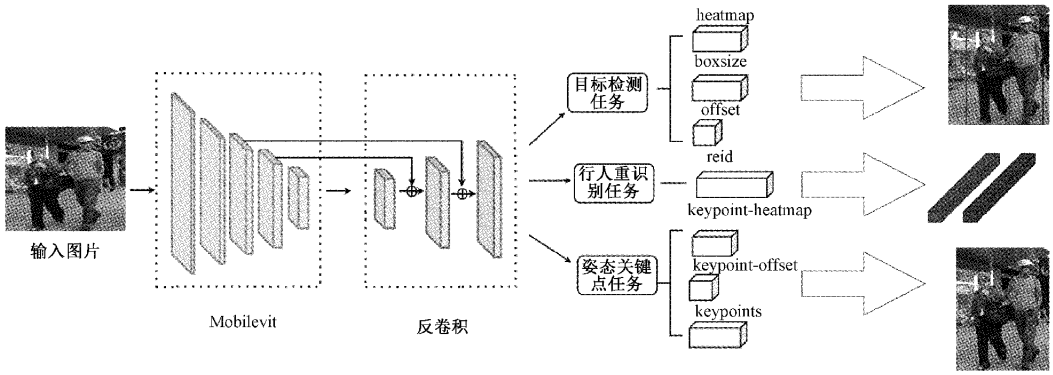


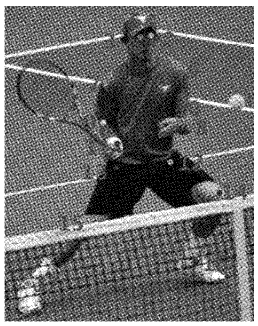
图 2 基于中心点的多分支检测网络总体结构

本文设计的检测头分为目标检测,行人重识别,关键点检测 3 个任务,每个任务由多个不同维度的分支完成。其中,目标检测任务对应 heatmap,boxsize,offset 分支。行人重识别对应 reid 分支。关键点检测任务对应 keypoint-heatmap, keypoint-offset, keypoints 分支。heatmap 是每个行人目标所在位置的中心点;boxsize 是行人框宽和高;offset 是对行人中心点的矫正,提高检测的准确度;行人重识别用 128 维特征向量来表征行人。keypoint-heatmap 表示人体关键点的位置信息,这 17 个人体关键点如图 3 所示。keypoints 是每个行人目标中心点到 17 个关键点在 x,y 两个方向上的偏移值,所以它是 34 维的。keypoint-offset 表示每个行人目标 17 个姿态关键点的关于关键点自身的二维偏移量。

1)目标检测任务

heatmap 分支负责预测中心点的位置。如果预测的中心点与真实值的中心点重合,heatmap 应该为 1。heatmap 中位置和对象中心之间的距离越远,响应也就越小。对于每一个行人目标的检测框,计算检测框的中心位置,然后计算中心点的热力图响应 M_{xy}。这个分支使用焦点损失,公式为:

$$L_{heatmap} = -\frac{1}{N}\sum_{xy}\begin{cases}(1-\hat{M}_{xy})^{\alpha}\log(\hat{M}_{xy}), & M_{xy}=1\\(1-M_{xy})^{\beta}(\hat{M}_{xy})^{\alpha}\log(1-\hat{M}_{xy}), & \text{其他}\end{cases} \tag{2}$$



1. 鼻子
2. 左耳
3. 右耳
4. 左眼
5. 右眼
6. 左肩
7. 右肩
8. 左肘
9. 右肘
10. 左手腕
11. 右手腕
12. 左臀
13. 右臀
14. 左膝
15. 右膝
16. 左脚踝
17. 右脚踝

图 3 人体 17 个关键点

offset 分支是对中心点位置的修正,中心点位置不准确,会给行人重识别任务带来困扰,无法提取行人最具代表力的特征,进而造成跟踪阶段无法正确关联相同身份的行人。boxsize 可以预测出检测框的大小,使用 $\hat{s}$ 和 $\hat{o}$ 分别表示尺寸和偏移头的输出。对于每一个真值框,我们计算它的真实尺寸 s,同样的计算真实的偏移 o。这两个分支都使用了 L1 loss 做为损失函数,它们的公式为:

$$L_{box} = \sum_{i=1}^{N} \| o^{i} - \hat{o}i \|_{1} + \| s - \hat{s}^{i} \|_{1} \tag{3}$$

2)行人重识别任务

行人重识别分支的目的是生成特征以便于区分不同的物体,在理想情况下,不同类别物体特征维度上的距离

比相同物体特征维度上的距离大。为了实现此目标，在检测头上面添加了一个输出通道数为 128 维的卷积层，对每个物体进行特征提取。输出的特征图为 $R^{128\times W\times H}$。

行人重识别是对不同身份行人的多分类问题，训练集中具有相同 ID 的行人目标被视为一个类别。对于每一个表示行人所在位置的真值框，在获取其 heatmap 上对应的中心点坐标后，把这个点提取到的特征，映射到一个类的分布向量中，类似于多目标分类的输出。$P(k)$表示真实的类标签，使用交叉熵损失函数，公式为：

$$L_{id}=-\sum_{k=1}^{K}\log(P(k)) \tag{4}$$

K 为训练集中的 ID 总数。

3)姿态关键点任务

keypoints 分支是 17 个关键点相对于中心点偏移的预测，keypoint-heatmap 分支检测出了所有的定义了 17 个关键点，keypoint-offset 是关键点它们各自的 offset。keypoint-heatmap 分支加 keypoint-offset 分支对关键点的预测有很高的置信度，但是无法把这些关键点关联成一个行人，keypoints 分支解决了这个问题。首先，目标检测任务已经确定了了一个行人，行人中心点加上 keypoints 分支回归出的各个关键点到中心点的偏移，得到了 17 个关键点相对于行人中心的大致位置。然后，根据距离关联到 keypoint-heatmap 分支预测的 17 个点上，同时，keypoint-heatmap 分支加 keypoint-offset 分支补充了 keypoints 分支没有回归出的的关键点。keypoint-heatmap 分支、keypoint-offset 损失函数分别和 heatmap、offset 类似。

本文将检测部分 3 个分支、姿态部分 3 个分支与重识别分支共同训练，并且用不确定性损失来自动平衡检测、重识别任务和关键点任务，根据式(2)～(4)得：

$$L_{total}=\frac{1}{2}\left(\frac{1}{e^{w_1}}(L_{heatmap}+L_{box})+\frac{1}{e^{w_2}}(L_{id})+\frac{1}{e^{w_3}}(L_{keypoint-heatmap}+L_{keypoint-offset}+L_{keypoints})+w_1+w_2+w_3\right) \tag{5}$$

式中：$\omega1$、$\omega2$ 和 $\omega3$ 是平衡参数。

2　行人多特征匹配

行人跟踪的环境往往非常复杂，需要考虑很多因素，基于单一特征的匹配会造成极低的精确度。我们在跟踪阶段充分利用检测网络得到的多种特征，以及卡尔曼滤波预测的行人运动特征，把它们作为判断依据，减少因遮挡、相似性太高等因素造成的轨迹中断或者身份频繁切换。另一方面，优化匹配方式，保证跟踪过程中的时实性。

2.1　重识别特征

在输入视频序列后，通过检测网络的 reid 分支提取每个行人目标的 128 维表观特征，我们把同一身份行人目标的表观特征聚集在一起，通过改正后的动态指数加权平均方法归一化这些特征，θ_t 是第 t 帧新检测出的行人特征，是该行人轨迹 t 帧以前的加权特征，β 是超参数，我们设置为 0.9。相比于未改正的指数加权平均法，添加了偏差修正，可以减少轨迹检测的初期特征加权带来的误差，避免特征损失影响表观特征匹配的精确度。该方法的公式如下所示。

$$v_t=\frac{v_t}{1-\beta^t}=\frac{\beta v_{t-1}+(1-\beta)\theta_t}{1-\beta^t} \tag{6}$$

其中，与传统的 DeepSORT 算法相比，DeepSORT 中使用的是某个行人轨迹所在视频帧中的所有特征，进行特征匹配时，按照跟踪对象距离上一次更新的远近作为匹配顺序，再把每个检测对象和一段轨迹中代价最小的作为代价矩阵中的值。本文的加权特征，每个轨迹的行人目标都只有一个特征，每次只计算每个检测对象和每个特征距离的最小值即可。特征匹配时的相似度量采用余弦距离，它可以体现特征间的相对差异，也适合行人跟踪这种特征数量较多的任务。公式如下：

$$\text{similarity}=\frac{\boldsymbol{A}\cdot\boldsymbol{B}}{\|\boldsymbol{A}\|\ \|\boldsymbol{B}\|}=\frac{\sum_{i=1}^{n}A_iB_i}{\sqrt{\sum_{i=1}^{n}A_i^2}\sqrt{\sum_{i=1}^{n}B_i^2}} \tag{7}$$

式中：$\boldsymbol{A}$ 是行人在当前帧提取到的表观特征向量，$\boldsymbol{B}$ 是行人跟踪轨迹对应的加权特征向量。

2.2　运动特征

当两个人外观相似度极高时，仅凭外观特征计算的相似度来判定两个物体为同一物体，仍会造成错误匹配，所以还需要结合目标间的位置来进一步判断，相邻两帧之间时间间隔很短，人不能会在两帧之间的物理位置差异特别大。卡尔曼滤波可以基于行人目标前一时刻的位置，预测下一时刻的位置它需要每个轨迹的均值和方差。均值有一个 8 维的向量组成，是目标的位置信息。协方差是一个 8×8 的矩阵，矩阵中数字和位置不确定性呈正比。通过对轨迹在当前帧的预测位置，与目标的检测位置进行比较，判断是否为同一行人。

2.3　人体姿态关键点相似度计算

仅通过卡尔曼滤波预测的行人运动特征，在运动轨迹没有规律，相机的位移时，会带来跟踪的不可靠性。在视频序列的前后两帧，比较通过检测网络得到的更加细粒度的，每个行人 17 个关键点的关系，来改善行人身份切换。本文采用了 OKS(object keypoint similarity)，即关键点相似度[23]，计算两个行人目标之间的关键点位置上的相似度，当视频序列中的某帧有 M 个行人目标，下一帧预测出 N 个行人目标，那么 M 个人中的每个人都会和预测的 N 个人的关键点进行相似度计算，最后我们会得到一个 $M\times N$ 的损失矩阵，对于其中的某个行人轨迹来说，它会选择指数最大的作为最优解。公式如下：

$$OKS_p = \frac{\sum_i \exp\{-d_{pi}^2/(2S_p^2\sigma_{pi}^2)\}\delta(v_{pi}=1, v'_{pi}=1)}{\sum_i \delta(v_{pi}=1)} \tag{8}$$

其中，p 表示行人身份编号，pi 代表关键点编号，范围是 1～17。v_{pi} 表示关键点的可见性。d 表示行人目标前后两帧某个关键点的欧式距离。当出现上一帧和下一帧两个预测的行人目标欧式距离相等的情况时，需要加入尺寸关系，S_p^2 表示前后两帧行人目标框面积平均值的平方。σ 表示行人第 i 个关键点的对目标整体影响因子。最后把取值范围规范在 0～1 之间。

3 实验分析

3.1 数据集准备及预处理

本文使用了 MOT Challenge 中的 MOT16 数据集。MOT16 数据集充分展现了行人跟踪的难点，非常具有代表性，是常用的用来评估行人跟踪算法的数据集之一。使用 MOT16 数据集 train 部分的 7 个视频序列训练，test 部分的 7 个视频测试，并把结果上传到 MOT Challenge 官网进行反馈。MOT16 数据集提供的真实值只有行人目标的检测框，缺失训练需要的行人目标姿态关键点信息，我们采取了给行人目标添加姿态关键点添加伪标签的形式。主流的姿态估计方式为自上而下，即首先检测出每个行人位置，再预测各个关键点位置，这种二阶段的方式有着较高的准确率。遵循这种模式，把代表着行人位置信息的检测框坐标作为输入，使用高分辨率网络 Hrnet[24] 预测，同时调整预测结果中有遮挡部分，或者超出画面关键点的可见性，正常为 1，遮挡严重为 0，这样训练过程中会减少错误数据信息的干扰。

训练时，对加载的图片采用旋转、尺度变换、仿射变换、颜色变换方式进行数据增强，旨在提高网络模型的泛化能力。考虑到网络模型需要学习人体姿态关键点信息，首先采用 Letterbox 方法处理图片到输入尺寸，传统 Resize 方法，会使图片产生形变，导致网络模型无法正确学习到关键点相对中心点的偏移关系。

3.2 实验说明

本文使用 Ubuntu18.04 系统，GeForce GTX 3090 GPU 显卡进行训练和测试。训练和测试阶段输入尺寸均为(576,320)。训练时，所有网络均未使用预训练模型，学习率设为 $1e^{-4}$，Epoch 设置为 50，前 5 个 Epoch 使用 Warmup 预热，后面恢复为预设学习率，并采用余弦退火方式周期性衰减学习率。测试时，匹配置信度为 0.4。

3.3 评价指标

本文采用行人跟踪领域最常用的 MOT Challenge 评价指标来评价算法，各个指标及含义如表 1 所示。

其中，MOTA、MOTP、MT、FPS 的值越高越好。FP、FN、IDs、ML 的值越低越好。MOTA 是最重要的指标，直观的反映跟踪时行人目标的检测和轨迹的保持能力，公式如下：

表 1 MOT Challenge 评价指标及含义

指标	含义
MOTA	多目标跟踪准确度。跟踪过程中所有行人目标发生错误匹配的百分比。
MOTP	多目标跟踪精确度。跟踪过程中，行人目标检测框和真值框之间在所有帧中的平均度量距离。
MT	大概率跟踪指数。超过 80%时间跟踪上的轨迹占所有行人目标的比例。
ML	大概率跟丢指数。小于 20%时间跟踪上的轨迹占所有行人目标的比例。
FP	误检。被误认为是正样本的比例。
FN	漏检。被误认为是负样本的比例。
IDs	身份切换。所有行人目标在跟踪过程中的身份转换次数，反映是否尽可能长的跟踪目标。
FPS	帧率。

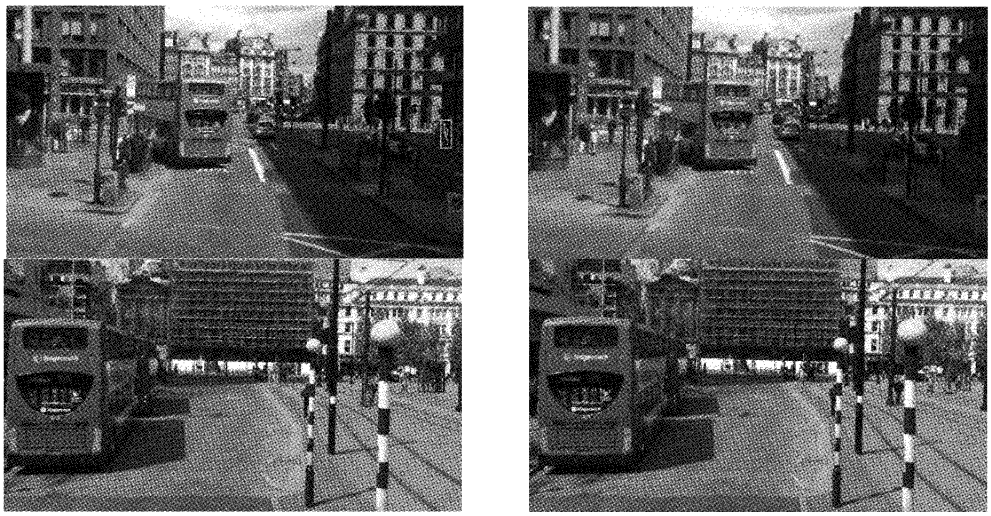
$$MOTA = 1 - \frac{\sum(FN + FP + IDs)}{\sum GT} \in (-\infty, 1] \tag{9}$$

3.4 实验结果分析

在对跟踪算法性能评估时，我们还把模型大小作为评价指标。因为在实际应用中，往往需要部署到资源有限的移动终端和边缘嵌入式计算设备中，更小的模型意味着占用的内存更小，更低的延迟，对不同平台的适用性更高。首先，验证了不同骨干网络对跟踪效果的影响，对比结果如表 2 所示。相对于没有加入短连接的 Mobilevit，本文算法 MOTA 高了 1.5%，MOTP 高了 1.38%。短连接的设计兼顾了大目标和小目标语义信息，占像素少的小目标不会因网络的下采样造成信息丢失，但同时，也会增加推理时间和参数量。图 4 通过视频序列的检测结果展示了两种网络的区别，虽然有漏检情况，但明显有短路连接的 Mobilevit 效果更好。与同样轻量化的采用短连接的 Mobilenet 相比，MOTA 和 IDs 都不如本文算法，这显示出 CNN 和 Transformer 结合带来的显著优势，其更加具备在特征图捕获全局和局部信息的能力。图 5、6 对比了两种网络结构在浅层和深层的热度图，热度图越接近红色说明提取到了置信度高的特征，从浅层和深层综合来看，Mobilenet

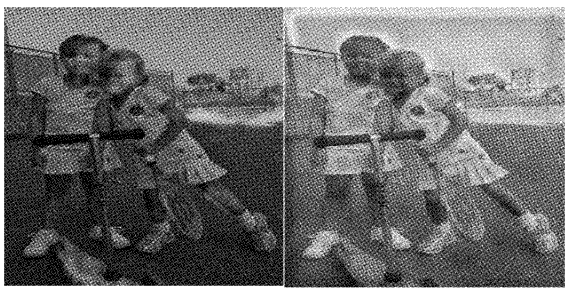
表 2 不同骨干网络对比

	MOTA	MOTP	IDs	FPS	模型大小/M
Mobilevit 无短路连接	47.08	76.79	872	21	18.1
Mobilenet 有短路连接	31.02	73.05	2 134	21	15.1
DLA34[25]	50.10	76.06	1 154	15	76.6
Ours	48.50	78.17	900	20	18.4



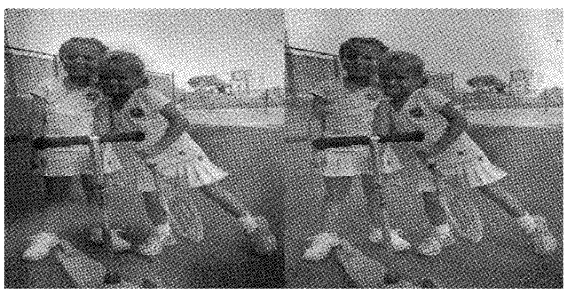
(a) Mobilevit有短路连接　　(b) Mobilevit无短路连接

图 4　小目标检测结果对比



(a) Mobilenet　　(b) 本文算法

图 5　网络浅层热度图对比



(a) Mobilenet　　(b) 本文算法

图 6　网络深层热度图对比

网络只获得了目标的局部特征，并且引入了噪声，本文算法则更完整的保存了目标的特征信息。

对目标跟踪阶段的匹配方式也作了消融实验，主要考察 FPS 和 IDs 两项，表 3 直观的展示了不同匹配方式的差别。单纯通过重识别特征匹配，ID 切换次数较高，这是由重识别特征对检测部分过度依赖造成的，对于两个相似度极高的特征或复杂的身体变化时，无法准确判断。加入运动特征，则综合考虑了目标运动轨迹，在短期和长期内都有很好的效果。人体姿态关键点特征，在大范围的人体被遮挡时，通过聚焦某个关键点也能准确完成匹配，可以明显看出 IDs 减少，但是，这也造成了计算成本略有增加，导致推理速度降低。

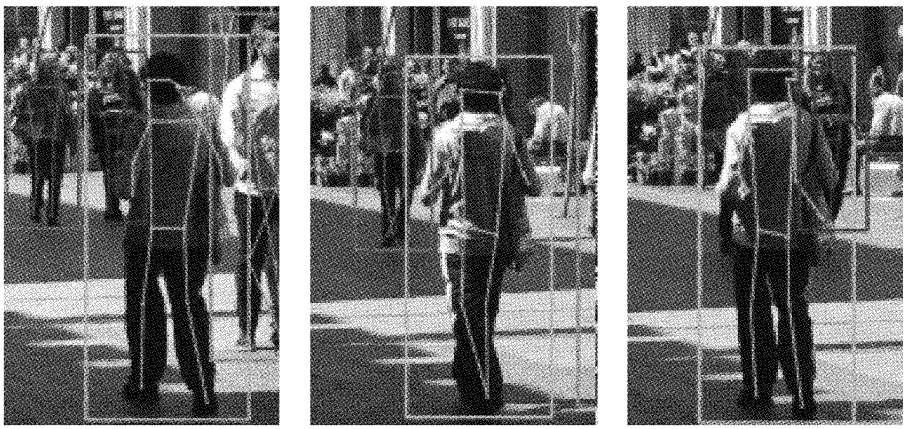
表 3　匹配特征消融实验

重识别特征	运动特征	关键点特征	IDs	FPS
√			1 203	24.6
√	√		1 127	23.1
√		√	982	21.8
√	√	√	900	20

为了更直观的说明加入姿态关键点特征在处理遮挡情况时的优势，我们对视频序列跟踪结果可视化。图 7 中的(a)为第一组，代表使用重识别和运动特征进行匹配，(b)为第二组，代表加入了姿态关键点特征。在第一组中，ID 编号为 12 的黑色上衣目标在(a)中可被正确识别，在(b)中被灰色上衣目标遮挡，(c)中再次出现时，身份切换为 72。第二组中，同样的编号为 12 的目标，被遮挡后依然可以被正确关联，ID 没有发生切换。由此可见，更多的特征信息会使匹配结果更加准确。

本文算法和其他算法在 MOT16 测试集上各项指标的对比如表 4 所示。本文算法和其他 4 个算法相比，在模型大小和推理速度上占据优势，表现出不错的跟踪性能。DeepSORT 算法在实际应用中还需要加上检测部分消耗的时间，本文算法属于联合一步检测方法，FPS 计算的是检测到匹配所有步骤的时间。MOTDT 算法的 MOTA 指标比本文算法低了 1%，高性能的 Transformer 结构使整体模型稍大，但得益于联合检测多种行人特征的设计，无需引入额外的模块，平衡了模型大小和推理速度，我们也计划将这一结构的优化纳入未来的工作。TubeTK 和 DeepMOT 模型大小是本文的数倍，不满足实际部署中的要求。

(a) 第一组



(b) 第二组

图 7 跟踪结果可视化

表 4 不同算法对比

	MOTA	IDS	MT	ML	FPS	模型大小/M
TubeTK[26]	64.0	1 117	254	147	1.0	536
DeepSORT	61.4	781	249	138	17.4	167.3
DeepMOT[27]	56.2	648	157	272	1.6	233
MOTDT[28]	47.6	792	115	272	20.6	16
Ours	48.52	900	121	284	20	18.4

4 结 论

本文提出了一种 CNN 和 Transformer 相结合的多特征跟踪范式,充分利用被跟踪目标的多种线索,实现高可靠度的跨帧关联。实验表明,本算法独特的人体姿态关键点匹配方式,能够有效应对目标间的遮挡、重叠,减少 ID 身份切换。同时,本算法降低了通道的维度使模型足够轻量化,联合检测的设计使推理速度满足实时性的要求,针对难以检测的小目标,使用短路连接的结构改善了性能。在未来的工作中,我们考虑把该算法部署到边缘计算设备中,并针对实际场景的需要进一步优化速度和准确度,展现出本算法的实用性。

参考文献

[1] 马金鹏.改进的基于人头检测的行人跟踪算法[J].电子测量技术,2017,40(12):233-237.

[2] 董美琳,任安虎.基于深度学习的高速公路交通事件检测研究[J].国外电子测量技术,2021,40(10):108-116.

[3] 李建良,张婷婷,陶知非,等.基于改进 Camshift 与 Kalman 滤波融合的领航汽车跟踪算法[J].电子测量与仪器学报,2021,35(6):131-139.

[4] 曹自强,赛斌.行人跟踪算法及应用综述[J].物理学报,2020,69(8):84-203.

[5] 候建华,张国帅,项俊.基于深度学习的多目标跟踪关

联设计[J]. 自动化学报,2020,46(12):2690-2700.
[6] 杨梅,贾旭,殷浩东,等. 基于联合注意力孪生网络目标跟踪算法[J]. 仪器仪表学报,2021,42(1):127-136.
[7] LUO W H, XING J L, MILAN A, et al. Multiple object tracking: A literature review [J]. ArXiv Preprint,2014,ArXiv: 1409.7618.
[8] BEWLEY A, GE Z, OTT L, et al. Simple online and realtime tracking[C]. IEEE International Conference on Image Processing, 2016: 3464-3468.
[9] WOJKE N,BEWLEY A,PAULUS D. Simple online and real-time tracking with a deep association metric[C]. 2017 IEEE International Conference on Image Processing (ICIP),2017:3645-3649.
[10] 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 and Machine Intelligence, 2017, 39 (6): 1137-1149.
[11] REDMON J, FARHADI A. YOLO9000: Better, faster, stronger[C]. IEEE Conference on Computer Vision and Pattern Recognition, IEEE Press, 2017: 6517-6525.
[12] NICALOS C, FRANCISCO M, GABRIEL S, et al. End-to-end object detection with transformers[J]. ArXiv Preprint,2020,ArXiv: 2005.12872.
[13] ZHANG X, ZHOU X, LIN M, et al. ShuffleNet: An extremely efficient convolutional neural network for mobile devices [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018: 6848-6856.
[14] HOWARD A, ZHU M, CHEN B, et al. MobileNets: Efficient convolutional neural networks for mobile vision applications [C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017:1-9.
[15] WANG Z, ZHENG L, LIU Y, et al. Towards real-time multiobject tracking[J]. European Conference on Computer Vision,2020:107-122.
[16] REDMN J, FARHADI A. YOLOv3: An incremental improvement [J]. ArXiv Preprint, 2018, ArXiv: 1804.02767.
[17] 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(3):1-19.
[18] ZHOU X Y, WANG D Q, KRÄHENBÜHL P. Objects as points [J]. ArXiv Preprint, 2019, ArXiv: 1904.07850.
[19] BERGMANN P, MEINHARDT T. Tracking without bells and whistles[C]. International Conference on Computer Vision, 2019:941-951.
[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]. European Conference on Computer Vision, 2020: 145-161.
[21] ASHIH V A, NOAM S, NIKI P, et al. Attention is all you need [J]. ArXiv Preprint, 2017, ArXiv: 1706.03762.
[22] JI Z, HUA Y, NIAN L, MIN K, et al. Online multi-object tracking with dual matching attention networks. InEur[C]. Proceedings of the European Conference on Computer Vision, 2018:366-382.
[23] TSUNG Y L, MICHAAEL M, SERGE B, et al. Microsoft COCO: Common objects in context[J]. Proceedings of the European Conference on Computer Vision, 2014:740-755.
[24] SUN K, XIAO B, LIU D, et al. Deep high-resolution representation learning for human pose estimation[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2019:5693-5703.
[25] YU F, WANG D, SHELHAMER E, et al. Deep layer aggregation [C]. In IEEE Confonerence on Computter Visionand Pattern Recognition, 2018:2403-2412.
[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 the IEEE/CVF Conference on Computer Visionand Pattern Recognition, 2020: 6308-6318.
[27] XU Y, BAN Y. Deepmot: A differentiable framework for training multiple object trackers [J]. ArXiv Preprint, 2019, ArXiv: 1906.06618.
[28] CHEN L, AI H, ZHUANG Z J, et al. Real time multiple people tracking with deeply learned candidate selection and person re-id entification [J]. ArXiv Preprint, 2018, ArXiv: 1809.04427.

作者简介

安胜彪,硕士,副教授,主要研究方向为集成电子系统和集成电路的研究。
E-mail:33588253@qq.com

刘新宇,硕士,主要研究方向为计算机视觉。
E-mail:390494785@qq.com

白宇,博士,讲师,主要研究方向为信息物理系统(CPS)、同步系统、深度学习、基于模型的系统设计、形式化方法。
E-mail:baiyu@hebust.edu.cn