地铁车站监控视频调查报告

目录

1. 多目标跟踪算法（MTSCT）

1.1 目标跟踪面临的挑战

（一）遮挡问题：

在目标检测中，遮挡问题是比较常见的，主要分为两种，第一种是其他物体对行人的遮挡，这往往会带来目标信息的缺失，进而导致漏检；第二种是行人个体之间的相互遮挡，这往往会引入大量的干扰信息，进而导致更多的虚检。

第一种遮挡，由于目标干扰物体遮挡，而算法只能学习待检测物体的特征，因此第一种遮挡只能通过增加样本来优化检测效果。第二种遮挡又分为类间遮挡和类内遮挡，理解为小孩与大人重叠时，小孩的头部与大人的衣物间遮挡为类间遮挡，类内遮挡理解为是头部重叠，身体部位重叠，衣物重叠等。类内遮挡产生于同类物体，也被称为密集遮挡，由于密集遮挡的两个目标的类别是相同的，所以两个目标之间的特征是相似的，检测器很可能无法定位。重点解决的是后一种情况导致的遮挡问题。

解决思路有两种：（1）利用检测机制判断目标是否被遮挡，从而决定是否更新模板，保证模板对遮挡的鲁棒性。（2）把目标分成多个块，利用没有被遮挡的块进行有效的跟踪。

（二）形变问题：

指目标表现的不断变化，通常导致跟踪发生漂移。解决漂移问题常用的方法是更新目标的表观模型，使其适应表现的变化。

（三）背景斑杂问题：

指的是要跟踪的目标周围有非常相似的目标对跟踪造成了干扰，还有在光照不均匀的复杂情况下获得的前景目标不完整，不准确，导致跟踪目标失败。解决

办法：利用目标的运动信息，预测运动的大致轨迹，防止跟踪器跟踪到相似的其他目标上。或利用目标周围的大量样本框对分类器进行更新训练，提高分类器对背景与目标的辨别能力。解决光照问题常将 RGB 颜色信息与纹理信息以置信度相融合方法来抑制阴影，提高运动目标跟踪在光照变换情况下的鲁棒性。

（四）尺度变换：

指的是在运动过程中的由远及近或由近及远而产生的尺度大小变换的现象。预测目标框大小也是目标跟踪中的一项挑战。通常的做法有：在运动模型产生候选样本的时候，生成大量尺度大小不一的候选框，或是在多个不同尺度目标上进行目标跟踪，产生多个预测结果，选择其中最优的作为最后的预测目标。

（五）运动模糊：

指目标或摄像机的运动导致的目标区域变模糊，导致跟踪效果不佳。常用均值偏移跟踪方法来进行跟踪，无需去模糊，利用从运动模糊中得到的信息，就能够完成跟踪目标。

（六）低分辨率：

指 ground-truth 边界框内的像素点个数少于 tr(tr=400) ，可采用非负矩阵分解的方法来建立目标模型，通过非负矩阵分解迭代计算提取目标重要轮廓信息，以一个字典矩阵的形式表示目标，进而完成跟踪。

（七）快速运动：

指 ground-truth 边界框内的像素点个数少于 tr(tr=400) ，可采用非负矩阵分解的方法来建立目标模型，通过非负矩阵分解迭代计算提取目标重要轮廓信息，以一个字典矩阵的形式表示目标，进而完成跟踪。

（八）超出视野：

指目标的一部分离开视野，通过引入一个检测器（TLD 算法提出跟踪和检测是可以互相促进的），用于在跟踪失败时的补充，跟踪为检测器提供正样本，检测器在跟踪失败时重新初始化跟踪器。使得跟踪鲁棒性增强。

1.2 生成式视频目标跟踪算法

生成式目标跟踪算法通过提取待跟踪目标的特征建立目标模型，利用生成的模型对待检测图像进行搜索，寻找与目标模型最匹配区域，该区域即为目标区域。因此生成式目标跟踪算法主要框架包含四个部分，目标选择、目标特征提取、目标建模、目标定位，如图 1 所示。具体来说，人工或使用目标检测算法对第一帧图像进行处理，勾选出目标并标记，当前常用的目标检测算法有帧差法、背景差法等；然后对选中的目标的特征进行建模，通常关注目标的灰度特征、边缘特征、梯度特征、颜色特征、纹理特征等，常用的模型有混合高斯模型、贝叶斯网络模型、马尔可夫模型等；目标定位则是完成跟踪。

表 1 VOT 每年排名前五算法汇总

排名	VOT2013	VOT2014	VOT2015	VOT2016	VOT2017	VOT2018	VOT2019	VOT2020
1	PLT	DSST	MDNet	C-COT	LSART	LADCF	DRNet	RPT
2	FoT	SAMF	DeepSRDCF	TCNN	CFWCR	MFT	Trackyou	OceanPlus
3	EDFT	KCF	EBT	SSAT	CFCF	SiamRPN	ATP	AlphaRef
4	LGT++	DGT	SRDCF	MLDF	ECO	UPDT	DiMP	AFOD
5	LT-FLO	PLT 14	LDP	Staple	Gnet	RCO	Cola	LWTL

图 1 生成式目标跟踪框架

（一）均值漂移算法

均值漂移算法是沿向量方向连续迭代候选目标帧，使其与模板的相似度最大，并收敛到目标的真实位置。该方法计算复杂度低，提取特征时会根据空间距离对中心位置周围的点进行加权。为解决光照变化导致的鬼影问题，王凯等将色度、梯度以及运动矢量预测引入 Meanshift 算法中提高了监控系统场景下目标跟踪算法准确性。如表 2 所示是均值漂移算法改进与效果。

算法名称	改进方式	改进效果
Camshif[19]	每个帧图像进行Meanshift、自动调节搜索窗口	解决了目标尺度变化问题
ASMS[21]	引入了尺度估计、经典颜色直方图特征、两个先验和一个检测	进一步优化目标尺度变化问题，保证速度的情况下提高了跟踪准确度。

表 2 均值漂移算法改进与效果

（二）贝叶斯滤波算法

递归贝叶斯滤波（Bayesian filtering）算法是基于贝叶斯估计理论的基础提出的目标跟踪方案。该算法包含预测和更新两个步骤，通过这两个步骤反复迭代估计图像中目标的位置。递归贝叶斯滤波概率在实际目标跟踪中很难获得最优解，为解决这一问题提出了卡尔曼滤波目标跟踪算法和粒子滤波目标跟踪算法。

卡尔曼滤波是用状态空间法描述系统的，由状态方程和量测方程所组成。卡尔曼滤波用前一个状态的估计值和最近一个观测数据来估计状态当前值，并以状态变量的估计值的形式给出。其具体形式如下:

假设某系统 k 时刻的状态变量为 X_k， 状态方程和量测方程表示为：

$$X_{k+1} = A_k X_k + \omega_k \quad (1)$$

$$y_k = C_k X_k + v_k \quad (2)$$

其中，k 表示时间;ω_k是一种白噪声输入信号，输出信号的观测噪声 v_k 也是一个白噪声，输入信号到状态变量的支路增益等于 1;A 表示状态变量之间的增

益矩阵;C 表示状态变量与输出信号之间的增益矩阵。

卡尔曼滤波算法在多目标跟踪系统中具有速度快、存储量小、消耗资源少等优点，但其要求观测方程必须是高斯形式的。然而在实际应用中，几乎所有的线性系统的观测方程都是非高斯形式的。在线性系统、非高斯观测方程领域内比较经典的一种算法是扩展卡尔曼滤波算法。

粒子滤波算法将蒙特卡洛思想引入贝叶斯滤波中。该算法核心思想是将随机采样与重要性重采样相结合。通过对图像随机散布粒子并采样特征，将采样结果与目标特征对比，计算出每个粒子的相似度，对相似度高的区域投入更多的粒子，迭代操作最终确定目标位置。粒子滤波从一定程度上，属于卡尔曼滤波的拓展，解决了卡尔曼滤波只适用于线性高斯分布概率问题，为分析非线性模型提供了一种有效的解决方案。赵宗超等利用引导图像滤波(guided image filer, GIF)对待检测图像滤波处理增强目标区域，使编码器增加训练样本，提高粒子置信度准确性，实现在线跟踪。如表 3 所示为贝叶斯滤波算法改进方式与效果。

算法名称	改进方式	改进效果
卡尔曼滤波[23]	采用最小均方误差的最优线性递归滤波方法	实现观测值不准确情况下对状态真实值的最优估计，只适用于高斯模型
粒子滤波[26]	蒙特卡洛思想引入贝叶斯滤波	适用于非线性模型，提高跟踪精度

表 3 贝叶斯滤波算法改进与效果

1.3 判别式视频目标跟踪算法

判别式目标跟踪算法认为目标跟踪问题是关于目标和背景的分类问题。该类算法将图像中将目标区域作为正样本，背景区域作为负样本进行训练并生成分类器，生成的分类器可以在下一帧图像中找到最优区域，该区域为目标区域。目前判别式目标跟踪算法主要可以分为相关滤波类、深度学习类以及孪生神经网络类。

（一）相关滤波目标跟踪算法

基于相关滤波的目标跟踪算法可以近似看成两个信号寻找最大相关值。通过对第一帧样本图片进行训练，输出一个具有区分背景和目标能力的滤波器，使用该滤波器对后面的每一帧图片进行运算获取相关值，根据运算后相关值的大小判断目标位置，相关值越大，说明该区域与目标的相似度越高，同时将每一回合响应结果返回滤波器对滤波器进行更新以提高下回合跟踪的准确性。如图 2 所示为相关滤波结构框图。

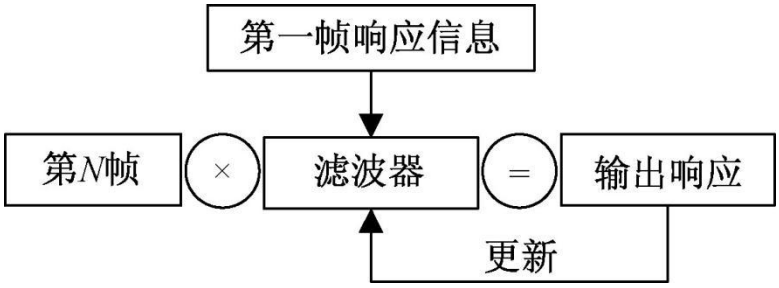

图 2 相关滤波框图

（二）深度学习目标跟踪算法

从 VOT 目标跟踪大赛比赛结果可以看出，竞赛的前五名大多使用了深度学习算法。目前主流的深度学习算法在目标跟踪领域仍然存在两个明显的缺陷：①目标跟踪中正样本通常只有初始帧中的目标，没有大量的数据支撑很难训练出良好的分类器；②为改善目标跟踪效果，通常需要提高网络的复杂度，这导致了算法的复杂度提高，实时性大幅下降。当前基于深度学习的目标跟踪算法主要可以分为两大类，一类是将深度学习技术引入相关滤波目标跟踪算法，对原有的算法进行改进；另一类是直接使用深度学习技术对目标进行端到端的跟踪。

深度学习与相关滤波相结合的目标跟踪算法主要是使用深度学习技术对目标进行特征提取，并将提取的特征代替原相关滤波算法中的特征。如表 5 为部分算法在 OTB50 和 OTB100 标准集上的表现情况。

算法名称	采用特征	特点	缺点
MOSSE[28]	Gray	利用傅里叶变换使滤波过程在频域执行跟踪速度快	采用的图像特征均为单通道灰度特征，因此在很大程度上限制了这两种跟踪方法在面对复杂背景或目标与背景颜色相近时的跟踪能力，无法应对目标尺度变化
CSK[29]	Gray	利用核函数和傅里叶变换计算相邻两帧图像的相关性，在提高跟踪效果的同时保证了较快的跟踪速度	
KCF[31] DCF	HOG	提升了跟踪器在复杂环境下对运动模糊和光照变化的鲁棒性	对于目标本身尺度变化导致的跟踪漂移问题的处理能力仍然有待加强
SAMF[33]	Gray、Color、HOG	使用7个粗略尺度的尺度池，单个滤波器	全局寻优的过程，响应最高的全局最优不能保证是局部最优
DSST[34]	HOG、Intensity	分别训练平移滤波器和尺度滤波器，特征选择上保持高度的灵活性，其对运动模 糊、光照变化具有很强的鲁棒性	算法复杂度高，计算量大，实时性差，对于目标发生形 变时跟踪效果差
SRDCF[40]	HOG	用空域正则化对滤波器边界函数加大权重约束，并进行迭代优化，分类器能够更准确地进行追踪	算法复杂度高，计算量大，实时性差

表 4 部分相关滤波目标跟踪算法

端到端的分类深度网络跟踪最早由 Nam 等提出多域卷积神经网络跟踪器(multi-domain convolutional neural network tracker, MDnet)算法并取得了 VOT2015 大赛冠军。该网络由特征提取通用的共享层和多分支检测的全连接层两部分组成。MDnet 算法包括两个阶段，训练阶段和目标跟踪阶段。训练阶段通过对数据集训练得到全连接层，共享层对所有数据通用以便获得动态目标通用性深度特征；目标跟踪保留固定共享层，并根据新的数据建立新的全连接层，二者结合组成新的端到端网络。MDnet 利用卷积操作将目标跟踪作为目标与背景的二分类问题处理，导致其在跟踪过程中易受目标相似物的干扰，并且对目标遮挡的鲁棒性较差。

针对该问题，后面都有相关的提升[1]。

（三）孪生网络目标跟踪算法

孪生网络结构是由两个网络结构或多个网络结构共同组成，并且该两个网络结构参数共享，一种特殊的神经网络架构。在目标跟踪领域，孪生网络算法同时

接收两个图片并分别输入两个子神经网络进行训练，同时两个子神经网络权值共享，通过对不同子神经网络获得的图像特征用于获取相关响应图像进行分析计算，以此判断目标位置并完成目标跟踪。

（四）Transformer 框架目标跟踪算法

Transformer 框架不同于传统深度学习框架，是一种基于注意力机制的框架。该框架早期服务于自然语言，近年才被应用于计算机视觉。Transformer 框架本质是一个编码解码的结构。在计算机视觉中主要被用于捕捉图像上的目标感受野。

Wang 等[2]提出 TrDiMP,将 Transformer 作为中间模块用于特征提取，并有效地提升了特征质量。同时对传统的 Transformer 进行改良，将编码和解码两部分成两个并行分支。编码部分对模板部分进行提取并使用注意力机制进行增强；搜索部分有解码进行处理。该算法虽然在检测精度方面取得较好效果，但是对于目标遮挡、目标消失仍需要优化。Chen 等[3]提出 Transt 算法，在孪生网络的基础上引入 Transformer 框架，利用该框架中的注意力机制用于避免目标跟踪过程中的语义丢失问题。该算法虽然已经取得较好效果，但是并没有充分利用背景信息，还具有较大的提升空间。

目前基于 Transformer 框架的目标跟踪算法尚处于起步阶段，但是已经在现有的数据集上取得了较好的效果，潜力巨大。

2. RE-ID

人员重识别去冗余问题是在双目摄像机检测到的视图有重叠部分，要将重叠部分的人员冗余去掉。在本项目中存在的去冗余问题有很多外界因素影响，由于行人数据来源于异时异地的不同设备，存在不同程度的行人姿势变化、目标遮挡、照明差异、视角差异、背景变化、设备像素差异以及开放性问题等，给 Re-ID

研究带来了巨大的挑战。

多目标多摄像头跟踪(MTMCT)旨在从多个摄像头拍摄的视频流中确定每个人在任何时候的位置。由此产生的多摄像头轨迹实现了包括视觉监控、可疑活动和异常检测、运动员跟踪和人群行为分析等应用。MTMCT 是一个众所周知的难题:为了降低成本，摄像机通常放置得很远，而且它们的视野并不总是重叠的。这导致了遮挡的延长时间和大的变化视点和照明在不同的视野。此外，通常无法提前知道人数，需要处理的数据量是巨大的。人员重新识别(Re- Id)与 MTMCT 密切相关:给定一个人的快照(查询)，Re- Id 系统从数据库中检索其他人的其他快照列表，通常是在不同的相机和不同的时间拍摄的，并通过减少与查询的相似性对它们进行排序。这样做的目的是，数据库中与查询中的人相同(即描述同一个人)的任何快照的排名都很高。

MTMCT 和 Re-ID 有细微但本质上的区别,因为 Re-ID 对查询的距离进行排序，而 MTMCT 将一对图像分类为同一性或非同一性,因此它们的性能由不同的度量标准来衡量:Re-ID 的排序性能，MTMCT 的分类错误率。这种差异似乎表明，用于两个问题的外观特征必须通过不同的损失函数学习。理想情况下，Re-ID 损失应该确保对于任何查询，a 和与它相同的特征之间的最大距离小于与它不相同的特征之间的最小距离。这将保证任何给定查询的正确功能排名。相比之下，MTMCT 损失应该确保任意两个同一性特征之间的最大距离小于任意两个非同一性特征之间的最小距离，以保证同一性内距离和同一性间距离的边际。根据这些标准，零 MTMCT 损失意味着零 Re-ID 损失，但反之则不然。然而，训练 MTMCT 类型的损失代价非常高，因为它需要使用所有特性对作为输入。更重要的是，在身份对内和身份对之间的数量会出现严重的不平衡。将 Re-ID 类型的三重损失函数与基于硬

数据挖掘的训练过程相结合，获得了 Re-ID 和 MTMCT 的高性能特征[4]。

3. 跨相机多目标跟踪算法(MTMCT)

3.1 基于锚定一次性 MOT 架构[5]

多目标多摄像机跟踪（MTMCT）系统跨摄像机跟踪目标。由于目标轨迹的连续性，跟踪系统通常限制其在局部邻域内的数据关联。在单摄像机跟踪中，局部邻域是指连续帧；在多摄像机跟踪中，它是指目标可能连续出现的相邻摄像机。对于相似性估计，跟踪系统通常采用从重新识别角度学习的外观特征。与跟踪不同，re-ID 通常无法访问轨迹线索，从而将搜索空间限制为局部邻域。由于其全局匹配特性，re-ID 透视图需要学习全局外观特征。跟踪中的局部匹配过程与 re-ID 外观特征的全局性质之间的不匹配可能会影响 MTMCT 性能。为了适应 MTMCT 中的局部匹配过程，在这项工作中，可以引入局部感知外观度量（LAAM）。

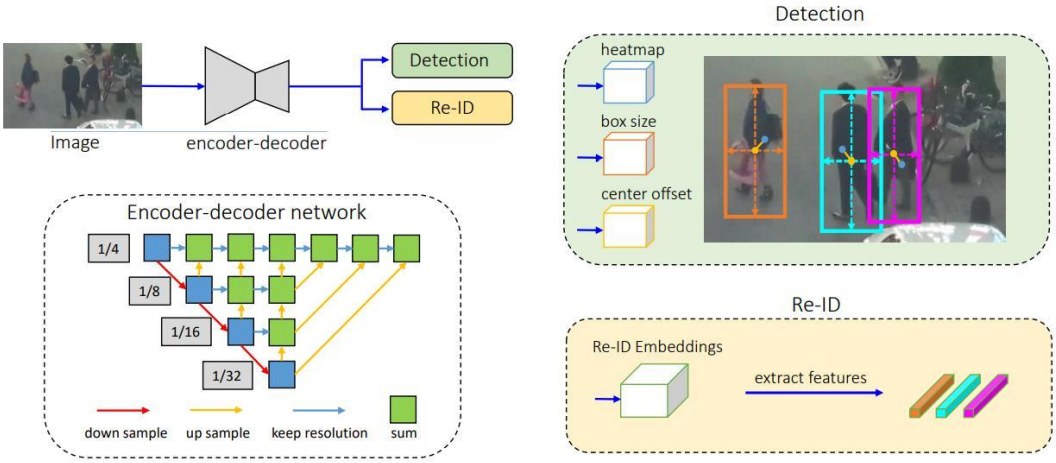

图 3 FairMOT 模型框架图

模型具有简单的网络结构，由两个同质分支组成，分别用于检测对象和提取 re-ID 特征。受此启发，检测分支以无锚点的方式实现，该方式估计以位置测量地图表示的对象中心和大小。类似地，re-ID 分支估计每个像素的 re-ID 特征，以表征以像素为中心的对象。注意，这两个分支是完全同质的，这与以前以两级

级联方式执行检测和重新识别的方法有本质区别。因此，FairMOT 消除了检测分支不平衡的缺点，有效地学习了高质量的 re-ID 特征，并在检测和 re-ID 之间取得了良好的权衡。我们通过评估服务器在 MOT 挑战基准上评估 FairMOT。它在 MOT15、MOT16、MOT17 和 MOT20 数据集的所有跟踪器中排名第一。当我们使用我们提出的单图像训练方法进一步预训练我们的模型时，它在所有数据集上实现了额外的增益。尽管结果很好，但该方法非常简单，在单个 RTX GPU 上以 30 FPS 的速度运行。

3.1.1 多目标跟踪发展

（一）通过分离模型进行检测和跟踪

检测方法：

大多数基准数据集（如 MOT17）提供了通过 DPM、更快的 R-CNN 和 SDP 等流行方法获得的检测结果，因此可以在相同的目标检测上公平地比较关注跟踪部分的工作。一些工作，例如使用大型私人行人检测数据集来训练以 VGG-16 为主干的更快的 R-CNN 检测器，从而获得更好的检测性能。少量工作，例如使用最近开发的更强大的检测器，例如级联 R-CNN，以提高检测性能。

跟踪方法：

现有的大多数工作都集中在问题的跟踪部分。我们根据用于关联的线索类型将其分为两类。

（1）基于位置和运动线索的排序方法

首先使用卡尔曼滤波器预测小轨迹的未来位置，计算其和检测的重叠，并使用匈牙利算法将检测分配给小轨迹。IOU 跟踪器直接计算（前一帧的）小轨迹和检测之间的重叠，而不使用卡尔曼滤波器来预测未来的位置。该方法实现了 100K fps 的推理速度（检测时间不计算），在目标运动较小时效果良好。SORT 和 IOU

跟踪器由于其简单性而在实践中得到广泛应用。然而，在拥挤场景和快速运动的挑战性情况下，它们可能会失败。一些工作，例如利用复杂的单目标跟踪方法来获得准确的目标位置并减少误报。然而，这些方法非常缓慢，尤其是当场景中有大量人员时。为了解决轨迹碎片问题， 提出了一种运动评估网络来学习小轨迹的长距离特征以进行关联。MAT 是一种增强的排序，它对摄像机运动，并使用动态窗口进行远程重新关联。

（2）基于外观线索的方法

最近的一些工作；建议裁剪检测的图像区域，并将其馈送到 re-ID 网络以提取图像特征。然后，他们根据 re ID 特征计算轨迹和检测之间的相似性，并使用匈牙利算法完成分配。该方法对快速运动和遮挡具有鲁棒性。特别是，由于外观特征，它可以重新初始化丢失的轨迹，随着时间的推移相对稳定。还有一些作品侧重于增强外观特征。例如，Bae 等人提出了一种在线外观学习方法来处理外观变化。Tang 等人利用身体姿势特征来增强外观特征。一些方法提出融合多个线索以获得更可靠的相似性。MOTDT 提出了一种分层数据关联策略，当外观特征不可靠时，使用 IoU 关联对象。少数工作，如也提出使用更复杂的关联策略，例如组模型和 RNN。

（3）离线方法

离线方法（或批处理方法）通常通过在整个序列中执行全局优化来获得更好的结果。例如，张等人建立了一个图形模型，其中节点表示所有帧中的检测。使用最小成本流算法搜索最优分配，该算法利用图的特定结构以比线性规划更快的速度达到最优。Berclaz 等人还将数据关联视为一项流优化任务，并使用 Kshortest 路径算法进行求解，这大大加快了计算速度，减少了需要调整的参数。

Milan 等人将多目标跟踪表述为连续能量的最小化，并专注于设计能量函数。能量取决于所有帧中所有目标的位置和运动以及物理约束。MPNTrack 提出了可训练的图神经网络，以执行整个检测集的全局关联，并进行 MOT 完全可微。Lif-T 将 MOT 描述为提升不相交路径问题，并引入提升边以进行长距离时间交互，这显著减少了 id 切换和重新识别丢失。

（4）优点和缺点

对于通过分离模型执行检测和跟踪的方法，其主要优点是可以分别为每个任务开发最合适的模型，而不会妥协。此外，他们可以根据检测到的边界框裁剪图像面片，并在估计 re-ID 特征之前将其调整为相同大小。这有助于处理对象的比例变化。因此，这些方法在公共数据集上取得了最佳性能。然而，它们通常非常慢，因为这两项任务需要单独完成，而不需要共享。因此，很难实现许多应用中所需的视频速率推断。

3.1.2 单个模型的检测和跟踪

随着深度学习中多任务学习的快速成熟，使用单个网络的联合检测和跟踪开始受到更多的研究关注。我们将其分为两类，如下所述。

（1）联合检测和 Re-ID

第一类方法在单个网络中执行对象检测和重新识别特征提取，以减少推理时间。例如，轨迹 RCNN 在掩码 RCNN 的顶部添加一个 re-ID 头，并为每个提案回归一个边界框和 re-ID 特征。类似地，JDE 构建在 YOLOv3 的基础上，YOLOv3 实现了接近视频速率的推断。然而，这些一次性跟踪器的精度通常低于两个步骤中的一个。

（2）联合检测和运动预测

第二类方法在单个网络中学习检测和运动特征。D&T 提出了一种暹罗网络，改网络接收相邻帧的输入，并预测边界框之间的帧间位移。 Tracktor 直接利用边界盒回归头传播区域建议的身份，从而消除盒关联。链式跟踪器提出了一种端到端模型，使用相邻帧对作为输入，并生成代表相同目标的盒对。这些基于框的方法假设边界框在帧之间有很大的重叠，这在低帧速率视频中是不正确的。与这些方法不同，CenterTrack 通过成对输入预测对象的中心位移，并通过这些点距离进行关联。它还向网络提供了基于点的附加热图输入轨迹，然后能够在任何地方匹配对象，即使框完全没有重叠。然而，这些方法仅关联邻帧中的对象，而不重新初始化丢失的轨迹，因此难以处理遮挡情况。我们的工作属于一流。我们研究了单次跟踪器关联性能下降的原因，并提出了一种简单的方法来解决这些问题。我们表明，在无需大量工程努力的情况下，跟踪精度显著提高。并发工作 CSTrack 还旨在缓解从特征的角度分析了两个任务之间的冲突，并提出了一个互相关网络模块，使模型能够学习与任务相关的表示。与 CSTrack 不同的是，我们的方法试图从三个角度系统地解决这个问题获得比 CSTrack 更好的性能。CenterTrack 也与我们的工作相关，因为它还使用基于中心的对象检测框架。但 CenterTrack 不提取外观特征，仅链接相邻帧中的对象。相反，FairMOT 可以和外观特征进行长期关联，并处理遮挡情况。

（3）多任务学习

有大量关于多任务学习的文献，可用于平衡目标检测和重新识别特征提取任务。不确定性使用任务相关的不确定性来自动平衡单个任务的损失。MGDA 通过在任务特定梯度中找到一个共同方向来更新共享网络权重。GradNorm 通过模拟任务特定梯度的大小来控制多任务网络的训练。我们在实验部分对这些方法进行

了评估。

3.1.3 视频对象检测

视频对象检测（VOD）与 MOT 相关，因为它利用跟踪来提高挑战性帧中的对象检测性能。虽然这些方法没有在 MOT 数据集上进行评估，但其中一些想法可能对该领域有价值。因此，我们在本节中简要回顾了它们。Tang 等人检测视频中的目标管，其目的是基于相邻帧在挑战帧中提高分类分数。在基准测试中，小对象的检测率大幅提高。在中也探讨了类似的想法。这些基于管的方法的一个主要限制是速度非常慢，尤其是当视频中有大量对象时。

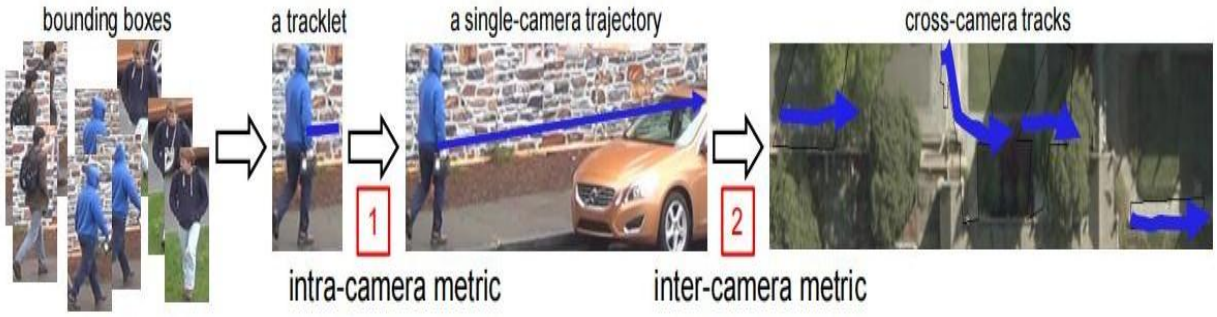

图 4 MTMCT 系统概述。给定对象边界框，我们首先将边界框连接成短而可靠的轨迹。然后，将轨迹合并为单摄像机轨迹。最后，将单摄像机轨迹关联起来， 形成交叉摄像机轨迹。所提出的 LAAM 包括摄像机内度量和摄像机间度量。摄像机内/摄像机间度量分别用于生成单摄像机轨迹和跨摄像机轨迹。

（1）单摄像机中的多目标跟踪

在 MTMCT 中，SCT 步骤的灵感来自多对象跟踪（MOT）。有在线和离线两种方法。在线跟踪方法不应使用来自未来时隙的数据。他们通常以贪婪的方式将检测与轨迹关联起来。离线方法可以从未来的信息中受益。他们通常将问题表述为批优化，例如最短路径、二部图和成对项。为了降低计算复杂度，一些采用了分层方法或时间滑动窗口。

（2）MTMCT 中的交叉摄像机跟踪

跨摄像机跟踪是 MTMCT 的一个独特功能。另一方面，离线方法通常采用批

量优化技术以获得更高的精度，这与 MOT 跟踪器类似。还研究了车辆 MTMCT。Tang 等人在车辆跟踪中使用多个线索来适应相似的外观、严重的遮挡和较大的视角变化。

（3）Re-ID 特征及其在 MTMCT 中的应用

Re-ID 源于跨摄像机跟踪。最近，这一领域出现了许多竞争性的 CNN 结构。研究了损失函数和训练技术，如对比损失、三重损失和硬负挖掘。探索了数据扩充方法，以丰富数据库。Re-ID 的进步推动了 MTMCT 的最新发展。Ristani 等人提出了一种全局特征学习方法，以提高 re-ID 和 MTMCT 的性能。

（4）多摄像机任务的度量学习

在 re-ID 中研究了度量学习算法。此外，还研究了跟踪中的度量学习，以计算观测值之间的相似性。与预定义的距离不同度量，这些学习的度量可以自动适应特定场景，并产生更高的准确性。例如，Leal 等人训练暹罗网络以聚集像素值和光流。Xiang 等人共同学习了用于多目标跟踪的全局特征表示和距离度量。Thoreau 等人从 re-ID 数据集学习暹罗网络，用于在线跟踪中的相似性估计。本文从现有文献出发，研究了 MTMCT 和 re-ID 之间的内在差异。相反为了直接学习全局特征表示/度量，我们研究了局部感知外观度量（LAAM），以满足 MTMCT 数据关联中的局部匹配。

3.1.4 模型实验数据

（一）平衡多任务损失

我们评估了平衡不同任务损失的不同方法，包括不确定性、梯度范数和 MGDA-UB。我们还使用网格搜索获得的固定权重评估基线。我们为基于不确定性的方法。第一种是“不确定性任务”，它分别学习检测损失和 re-ID 损失的两

个参数。第二个是“不确定性分支”，它分别学习热图损失、箱体尺寸损失、偏移损失和 re-ID 损失的四个参数。

表 5 MOT17 数据集验证集上不同损失加权策略的比较。最佳结果以粗体显示。

Loss Weighting	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
Fixed	**69.6**	71.6	387	**81.9**	93.8
Uncertainty-task	69.1	72.8	**299**	81.2	94.4
Uncertainty-branch	68.5	73.3	319	81.0	96.0
MGDA-UB	63.6	67.9	355	78.5	**97.0**
GradNorm	69.5	**73.8**	311	81.3	95.1

表 6 MOT17 数据集上不同主干的比较验证。“MLFF”是多层特征融合的简称。“Acc”是 ImageNet 分类精度的缩写。ImageNet 分类精度的结果来自骨干网络的原始论文。最佳结果以粗体显示。

Backbone	w/ MLFF	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑	Acc↑
ResNet-34		63.6	67.2	435	75.1	90.9	75.2
ResNet-50		63.7	67.7	501	75.5	91.9	77.8
RegNetY-4.0GF		63.9	68.0	407	75.8	91.9	**79.4**
ResNet-34-FPN	✓	64.4	69.6	369	77.7	94.2	75.2
RegNetY-4.0GF-FPN	✓	65.8	69.3	257	78.0	94.3	**79.4**
HRNet-W18	✓	67.4	74.3	315	80.5	94.6	76.8
DLA-34	✓	69.1	72.8	299	81.2	94.4	76.9
HarDNet-85	✓	**71.2**	**74.5**	**198**	**82.6**	**95.8**	77.0

结果如表 5 所示。我们可以看到，“固定”方法得到的 MOTA 和 AP 最好，但 IDs 和 TPR 最差。这意味着改模型偏向于检测任务 MGDA-UB 的 TPR 最高，但最低 MOTA 和 AP，这表明该模型偏向于 re-ID 任务。类似的结果可以在中找到。GradNorm 获得了最佳的整体跟踪精度（最高的 IDF1 和第二高的 MOTA），这意味着确保不同任务具有相似的梯度幅值有助于处理特征冲突。然而，GradNorm 需要更长的训练时间。因此，我们使用了更简单的不确定度方法，该方法在其余实验中略差于梯度范数。

（二）多层特征融合

我们比较了一些主干网，如 vanilla ResNet、特征金字塔网络（FPN）、高分辨率网络（HRNet）、DLA、HarDNet 和 RegNet。请注意，这些方法的其余因素（如训练数据集）均被控制为相同，以便进行公平比较。结果表明，直接使用更大或更强大的网络并不总能提高最终跟踪精度。相比之下，ResNet-34-FPN 的参数实际上比 ResNet-50 少，其 MOTA 分数比 ResNet-50 高。更重要的是，TPR 从 90.9% 显著提高到 94.2%。通过比较 RegNetY-4.0GFFPN 和 RegNetY-4.0GF，我们可以看到在 RegNet 中添加多层特征融合结构带来了可观的收益（+1.9 MOTA，+1.3 IDF1，-36.9%IDs，+2.2 AP，+2.3 TPR），这表明多层特征融合比简单使用更大或更强大的网络具有明显的优势。

表7 检测和检测之间的特征冲突演示重新识别 MOT17 数据集验证集上的任务。“-det”是指只训练检测分支，随机训练 re-ID 分支已初始化。最佳结果以粗体显示。

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-34-det	63.7	60.4	597	76.1	36.7
DLA-34	**69.1**	**72.8**	**299**	**81.2**	**94.4**

为了验证检测和 re-ID 任务之间是否存在特征冲突， 我们引入了基线 ResNet-34-det，它只训练检测分支（re-ID 分支随机初始化）。从表 7 可以看出，如果我们不训练显示两个任务之间冲突的 re-ID 分支，则 AP 测量的检测结果提高了 1 个点。特别是，ResNet-34-det 甚至比 ResNet-34 获得更高的 MOTA 分数，因为该度量更倾向于比跟踪结果更好的检测。相比之下，DLA-34 在 ResNet-34 上添加了多层特征融合，实现了更好的检测和跟踪结果。这意味着多层特征融合允许每个任务提取其需要的任何内容，从而有助于缓解特征冲突问题。

表 8 不同脊柱对不同比例物体的影响。S：面积小于 7000 像素；M：面积从 7000 到

15000 像素；L：大于 15000 像素的区域。最佳结果以粗体显示。

Backbone	AP^S	AP^M	AP^L	TPR^S	TPR^M	TPR^L	IDs^S	IDs^M	IDs^L
ResNet-34	40.6	57.8	85.2	91.7	85.7	88.8	190	87	118
ResNet-50	39.7	59.4	86.0	91.3	85.3	89.0	248	91	124
ResNet-34-FPN	45.9	61.0	85.4	90.7	91.5	**93.3**	166	71	90
HRNet-W18	**51.1**	63.7	85.7	**94.2**	**92.5**	93.1	168	**55**	**56**
DLA-34	46.8	**65.1**	**88.8**	92.7	91.2	91.8	**134**	64	70

（三）MOTChallenge 结果

我们将我们的方法与最先进的方法（SOTA）进行了比较，包括一次法和两步法。

表 9 单幅图像训练对 MOT17 验证集的影响。“CH” 和“MIX” 分别代表 CrowdHuman 等五个数据集。*表示未使用标识注释。最佳结果以粗体显示。

Training Data	MOTA ↑	IDF1 ↑	IDs ↓	AP↑	TPR ↑
CH*	64.1	64.9	476	80.5	79.9
MOT17	67.5	69.9	408	79.6	93.4
CH*+MOT17	**71.1**	**75.6**	327	**83.0**	93.6
MIX+MOT17	69.1	72.8	**299**	81.2	**94.4**

表10 2DMOT15 数据集上最先进的单次跟踪器的比较。“MIX”代表大规模训练数据集，“MOT17 Seg”代表MOT17 数据集中具有分割标签的 4 个视频。相同训练数据的最佳结果以粗体显示。

Training Data	Method	MOTA↑	IDF1↑	IDs↓	FP↓	FN↓	FPS↑
MIX	JDE	67.5	66.7	218	1881	**2083**	26.0
	FairMOT(ours)	**77.2**	**79.8**	**80**	**757**	2094	**30.9**
MOT17 Seg	Track R-CNN	69.2	49.4	294	1328	**2349**	2.0
	FairMOT(ours)	**70.2**	**64.0**	**96**	**1209**	2537	**30.9**

（四）与单次 SOTAMOT 方法的比较

JDE 和 TrackRCNN 有两个已发表的作品，它们共同执行目标检测和身份特征嵌入。我们将我们的方法与这两种方法进行比较。在之前的工作之后，测试数

据集包含来自 2DMOT15 的 6 个视频。FairMOT 使用与他们论文中描述的两种方法相同的训练数据。特别是，当我们与 JDE 进行比较时，FairMOT 和 JDE 都使用第 5.1 节中描述的大规模合成数据集。由于 Track R-CNN 需要分割标签来训练网络，因此它只使用具有分割标签的 MOT17 数据集的 4 个视频作为训练数据。在本例中，我们还使用 4 个视频来训练我们的模型。CLEAR 指标和 IDF1 用于衡量其性能。结果如表 11 所示。我们可以看到，我们的方法明显优于 JDE。特别是，ID 开关的数量从 218 减少到 80，这在用户体验方面是一个很大的进步。结果验证了无锚方法相对于以前基于锚的方法的有效性。这两种方法的推理速度接近视频速率，而我们的方法更快。与 Track R-CNN 相比，它们的检测结果略优于我们的（FN 较低）。然而，FairMOT 获得了更高的 DF1 分数（64.0 比 49.4）和更少的 ID 开关（96 比 294）。这主要是因为跟踪 R-CNN 遵循“先检测，后识别”的框架，并使用锚点，这也会给识别任务带来歧义。

表 11 “专用检测器”协议下最先进方法的比较。值得注意的是，FPS 同时考虑了检测和关联时间。一次性跟踪器标有“*”。每个数据集的最佳结果以粗体显示。

Dataset	Tracker	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	FPS↑
MOT15	MDP_SubCNN(Xiang et al., 2015)	47.5	55.7	30.0%	18.6%	628	<1.7
	CDA_DDAL(Bae and Yoon, 2017)	51.3	54.1	36.3%	22.2%	544	<1.2
	EAMTT(Sanchez-Matilla et al., 2016)	53.0	54.0	35.9%	19.6%	7538	<4.0
	AP_HWDPL(Chen et al., 2017)	53.0	52.2	29.1%	20.2%	708	6.7
	RAR15(Fang et al., 2018)	56.5	61.3	45.1%	14.6%	**428**	<3.4
	TubeTK*(Pang et al., 2020)	58.4	53.1	39.3%	18.0%	854	5.8
	FairMOT (Ours)*	**60.6**	**64.7**	**47.6%**	**11.0%**	591	**30.5**
MOT16	EAMTT(Sanchez-Matilla et al., 2016)	52.5	53.3	19.9%	34.9%	910	<5.5
	SORTwHPD16(Bewley et al., 2016)	59.8	53.8	25.4%	22.7%	1423	<8.6
	DeepSORT_2(Wojke et al., 2017)	61.4	62.2	32.8%	18.2%	781	<6.4
	RAR16wVGG(Fang et al., 2018)	63.0	63.8	39.9%	22.1%	**482**	<1.4
	VMaxx(Wan et al., 2018)	62.6	49.2	32.7%	21.1%	1389	<3.9
	TubeTK*(Pang et al., 2020)	64.0	59.4	33.5%	19.4%	1117	1.0
	JDE*(Wang et al., 2020b)	64.4	55.8	35.4%	20.0%	1544	18.5
	TAP(Zhou et al., 2018)	64.8	**73.5**	38.5%	21.6%	571	<8.0
	CNNMTT(Mahmoudi et al., 2019)	65.2	62.2	32.4%	21.3%	946	<5.3
	POI(Yu et al., 2016)	66.1	65.1	34.0%	20.8%	805	<5.0
	CTrackerV1*(Peng et al., 2020)	67.6	57.2	32.9%	23.1%	1897	6.8
	FairMOT (Ours)*	**74.9**	72.8	**44.7%**	**15.9%**	1074	**25.9**
MOT17	SST(Sun et al., 2019)	52.4	49.5	21.4%	30.7%	8431	<3.9
	TubeTK*(Pang et al., 2020)	63.0	58.6	31.2%	19.9%	4137	3.0
	CTrackerV1*(Peng et al., 2020)	66.6	57.4	32.2%	24.2%	5529	6.8
	CenterTrack*(Zhou et al., 2020)	67.8	64.7	34.6%	24.6%	**2583**	17.5
	FairMOT (Ours)*	**73.7**	**72.3**	**43.2%**	**17.3%**	3303	**25.9**
MOT20	FairMOT (Ours)*	**61.8**	**67.3**	**68.8%**	**7.6%**	**5243**	**13.2**

3.2 基于深度学习的融合方向和分块 HSV 直方图特征 MTMCT 算法[6]

在第一章中介绍了多目标跟踪算法，匹配不同摄像头下的跟踪轨迹时，跨摄像头多目标跟踪（MTMCT）其实可以看成重识别问题，因此外观特征是非常重要的判别依据。为了准确快速计算外观特征，国内学者提出了分块 HSV 颜色直方图方法，本章采用分块 HSV 颜色直方图来描述特征，母的是排除局部干扰，同时保留全局的结构信息，从而实现对相似外观的人的区分以减少误匹配。

为了减轻摄像头下的视角不同带来的行人外观变化，本文提出了一个结合方向信息和分块 HSV 特征的跨摄像头多目标跟踪算法，通过比较不同轨迹中各个方向的外观特征来计算轨迹之间的外观相似度，从而实现对不同摄像头下的行人进行快速准确的跟踪。

采用分块 HSV 直方图作为外观特征，增强局部特征的判断能力。同时将行人的方向信息引入到轨迹的外观特征中，对相同方向的不同行人轨迹加以约束，对相同方向的相同行人轨迹增大权重，最终减轻行人方向变化对轨迹外观特征的影响，从而达到提高跟踪进度的目的。

分块 HSV（色调，饱和度，明度）颜色模型将色调，饱和度等与传统的颜色融合，更多地基于人类视觉来设计色彩的属性。对 HSV 进行非均匀量化，最后按照一定的权值线性合成一维特征向量。将 HSV 直方图特征作为目标的外观特征，既有局部特征的统计，又可以保留目标的全局外观特征信息。

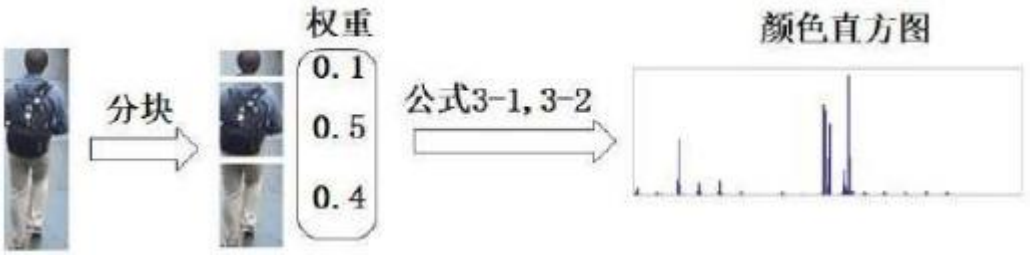

图 5 分块 HSV 颜色直方图流程图

方向梯度直方图（HOG）特征对光照变化和几何形变有较好的抗干扰能力，并降低了对表达图像特征所需的维度。利用 HOG 特征生成检测框的特征向量，训练独立的 SVM 分类器，每个分类器对应一个角度方向（行人）。

3.2.1 结合方向和分块 HSV 特征的 MTMCT 算法

算法流程是，输入单摄像头多目标跟踪轨迹，使用 DPM 检测器提取检测框，对检测框提取分块 HSV 颜色直方图特征，在对检测框通过前面的行人方向估计模型得到检测框内行人的方向，将带有方向信息的 HOG 特征将方向与分块 HSV 特征结合作为行人的外观特征，在进行不同摄像头下轨迹的匹配时，轨迹中所有帧图像的外观特征的平均值作为轨迹的外观特征，利用此外观特征进行聚类形成外观相似的轨迹的分组，在对分组后的轨迹进行二值整数规划求解，将同一个身份的

轨迹划分到同一个组里面，即生成同一个目标的所有摄像头下的轨迹。

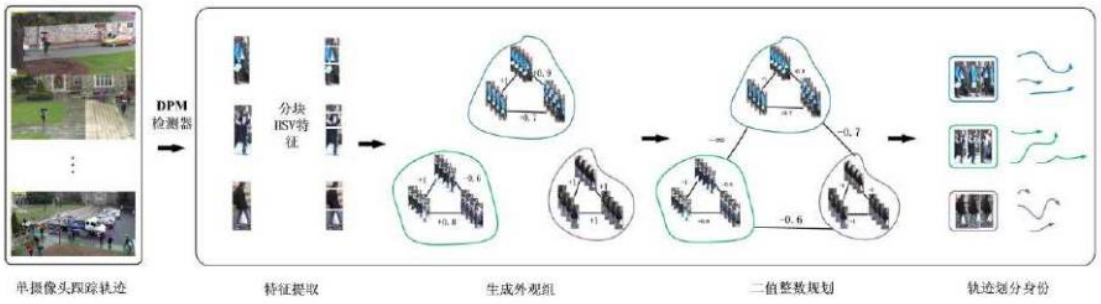

图 6 融合方向信息特征的跨摄像头多目标跟踪算法流程

3.2.2 实验结果分析

（一）实验环境和评价指标

本章以单目多摄像头目标跟算法（BIPCC）[7]为基础框架，使用 PETS2009，Town Center 和 Parking Lot 三个具有代表性的数据集，另外在 DukeMTMCT 数据集上进行测试效果。

实验环境在 Ubuntu 16.04 系统下的 MATLAB2016b 上进行的，显卡容量 12GB，计算机主频为 3.5GHz，内存为 128G。

评价指标可以看第五章内容。

（二）实验结果比较与分析

为了证明使用分块 HSV 颜色直方图作为外观特征和引入行人方向信息对轨迹的外观匹配有促进作用，我们在 DukeMTMCT 数据集的 Test-easy 测试集上进行和基线[8]的对比实验，以 BIPCC 算法框架为基础，构建跨摄像头多目标跟踪框架，外观特征从 HSV 颜色空间变换到分块 HSV 颜色直方图特征再变换到加入方向信息。

Multi-Easy	IDF1	IDP	IDR
Baseline[26]	56.2	67.0	48.4
Ours(HSV)	56.4	67.5	48.5
Ours(分块 HSV)	56.7	68.7	48.3
Ours(分块 HSV+dir)	**57.2**	**69.3**	**48.8**

表 12 各算法在 DukeMTMCT 数据集 Test-easy 上的实验结果比较

Multi-Hard	IDF1	IDP	IDR
Baseline[26]	47.3	59.6	39.2
Ours(HSV)	47.4	59.8	39.3
Ours(分块 HSV)	47.7	60.0	39.6
Ours(分块 HSV+dir)	**48.3**	**60.4**	**40.2**

表 13 在 DukeMTMCT 数据集上 Test-hard 上的结果

图 7 本章算法的部分正确实验结果展示

3.3 基于深度学习的融合方向和整体外观特征的跨摄像头多目标跟踪算法

3.2 介绍了一种结合方向和分块 HSV 特征的跨摄像头多目标跟踪算法，以第 1 章的图分区模型为基础，从单摄像头多目标跟踪延伸到跨摄像头多目标跟踪，也是一种行之有效的算法。分块 HSV 特征快速有效，既能对目标遮挡有一定的鲁棒性，也能更好地区分外观相似的特征。结合方向信息的整体特征能有效应对不同视角造成的行人外观变化。虽然 3.2 的跨摄像头多目标跟踪算法在 DukeMTMCT 数据集上的 IDF1 跟 Baseline 比起来有所提升，但是最新的跨摄像

头多目标跟踪算法在 DukeMTMCT 数据集上的 IDF1 提升却更为明显，这种提高得益于行人重识别在深度学习领域的飞速发展。

跨摄像头多目标跟踪任务和行人重识别任务要解决的问题类似，只不过跨摄像头多目标跟踪是将每个人在不同摄像头下的轨迹检索出来，而行人重识别则是将行人从不同摄像头下的图片中检索出来。它们之间的关联在于我们可以把跨摄像头多目标跟踪任务看作是单个摄像机内的多目标跟踪和多个摄像机之间基于图片的行人重识别作外观特征的轨迹匹配。Beyer 等人[9]使用基于行人重识别特征的最优贝叶斯滤波作为跟踪器，这种基于概率优化的方法虽然看起来简单，但是跟踪精度受到前一状态的影响，因此跟踪效果并未达到最佳，不过也为后面的研究人员使用行人重识别作为跨摄像头多目标跟踪的外观特征提供了新思路。

在 3.2 算法框架的基础上，考虑本章算法是基于检测的跟踪，检测正确率会影响跟踪精度，因此本章使用基于行人姿态估计的 OpenPose[10]检测器降低虚检、漏检和误检。考虑到手工设计的特征对相似外观的行人判别能力较差，本文使用基于深度学习的行人重识别特征作为轨迹相似性度量的外观特征，考虑到不同视角下行人外观的差异性，本文设计了一个加入目标方向信息的三元组损失函数训练神经网络，并与目标本身的全局外观特征相融合，减轻方向变化对行人外观特征的影响，从而进行更好的跨摄像头多目标跟踪轨迹的数据关联。

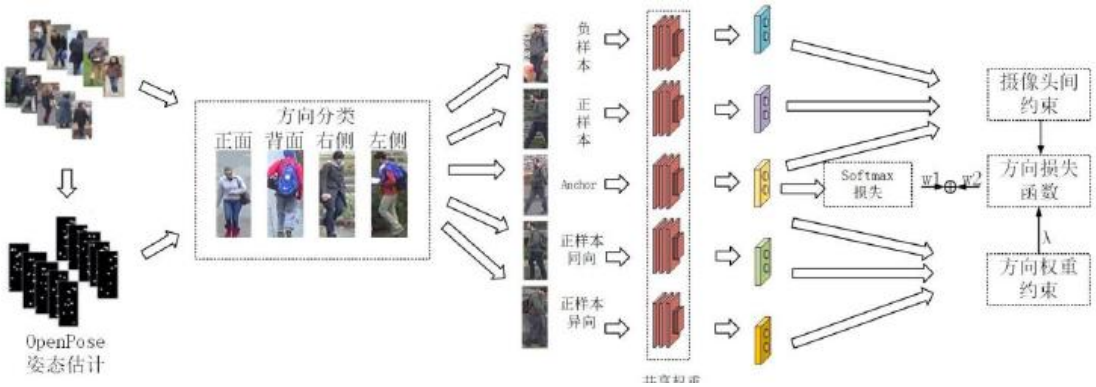

图 8　融合行人整体特征和行走方向的算法流程

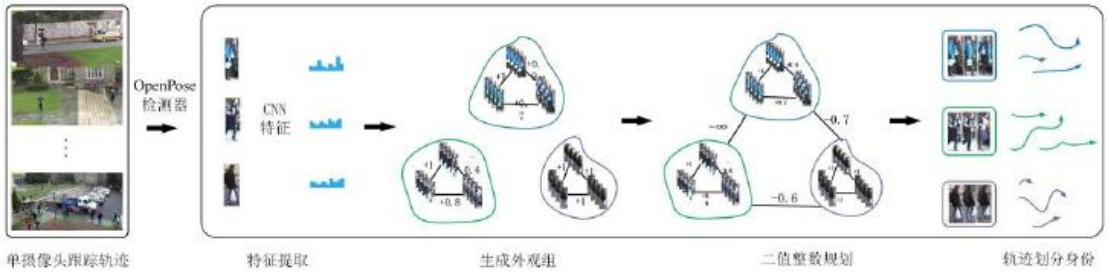

图 9 算法流程

3.3.1 与各行人重识别算法的实验结果对比

为了证明我们提出的融合行人方向信息的行人重识别算法的有效性，我们将其与多个行人重识别算法在 Duke-reID 和 Market1501 数据集上进行比较，比较结果如表 14 和表 15 所示。

methods	Duke-reID	
	Rank-1	mAP
BoW+KISSME[54]	25.13	12.17
LOMO+XQDA[56]	30.75	17.04
Baseline[57]	65.22	44.99
PAN[58]	71.59	51.51
TriHard[51]	73.24	54.60
Ours	**74.68**	**56.52**

表 14 与各行人重识别算法在 Duke-reID 数据集的比较结果

methods	Market1501	
	Rank-1	mAP
DNS[59]	55.43	29.87
GatedSiamese [60]	65.88	39.55
PointSet [61]	70.72	44.27
PAN[58]	82.81	63.35
TriHard[51]	82.99	**66.63**
Ours	**84.78**	65.32

表 15 与主流算法在 Market1501 数据集上的比较结果

3.3.2 与各跨摄像头多目标跟踪算法的实验结果对比

为了证明本文提出的基于深度学习的融合行人方向信息和整体特征的跨摄像头多目标跟踪算法的有效性，我们在 DukeMTMCT 数据集的 Test-easy 测试集上与当下的一些跨摄像头多目标跟踪算法进行比较，实验结果如表 16 所示。

methods	Multi-Camera Easy		
	IDF1	IDP	IDR
Baseline[26]	56.2	67.0	48.4
W. Liu et al.[62]	55.5	78.89	44.6
MTMC_CDSC[63]	60.0	68.3	53.5
MYTRACKER[64]	64.8	70.8	59.8
Ours(W 均为 1)	61.7	71.0	60.1
Ours(W 带方向)	**68.9**	**72.8**	**65.4**

表 16 与各算法在 DukeMTMCT Test-easy 上的比较结果

为了证明本文提出的基于深度学习的融合行人方向信息和整体特征的跨摄

像头多目标跟踪算法的有效性，我们在 DukeMTMCT 数据集的 Test-hard 测试集上与当下的一些跨摄像头多目标跟踪算法进行比较，实验结果如表 17 所示。

methods	Multi-Camera Hard		
	IDF1	IDP	IDR
Baseline[26]	47.3	59.6	39.2
MTMC_CDSC[63]	50.9	63.2	42.6
MYTRACKER[64]	47.3	55.6	41.2
Ours(W 均为 1)	54.6	59.1	50.8
Ours(W 带方向)	**59.7**	**65.9**	**54.4**

表 17 与各算法在 DukeMTMCT Test-hard 上的比较结果

3.3.3 本章算法与 3.2 算法在 Test-easy 数据集的结果比较

此处的 IDF1 比 3.2 的 IDF1 高 8.5%，比使用 OpenPose+（分块 HSV+方向）的 IDF1 高 2.9%。本章的算法采用 OpenPose 作为检测器，加上融合方向信息的深度学习特征作为外观特征，在 Test-easy 数据集上的 IDF1 是 68.9%，IDP 是 72.8%，IDR 是 65.4%，此处的 IDF1 比 3.2 的 IDF1 高 11.7%，比 OpenPose+（分块 HSV+方向）的 IDF1 高 6.1%，比 DPM+（CNN+方向）的 IDF1 高 3.2%。整体来说，本章的跨摄像头多目标跟踪算法取得了最佳的效果。

检测器	特征	Multi-Camera Easy		
		IDF1	IDP	IDR
DPM[42]	分块 HSV+方向	57.2	69.3	48.8
OpenPose[75]	分块 HSV+方向	62.8	68.4	58.1
DPM	CNN+方向	65.7	70.1	61.9
Openpose	CNN+方向	**68.9**	**72.8**	**65.4**

表 18 本章算法与 3.2 算法在 Test-easy 数据集的结果比较

3.3.4 本章实验结果图

展现了四个正确的跨摄像头多目标跟踪结果，分别对应每个目标在摄像头下出现的轨迹和该目标在哪个摄像头下出现。

图 10 本章算法部分正确实验结果展示

4. 数据集

（1）MOT17 数据集

MOT17 数据集内容和 MOT16 一致，但是使用了更精确的标注框，同时使用三个检测器得出三种检测结果供研究者使用，包括 11235 帧、1342 个行人和 292733 个行人检测框。MOT17 数据集输入尺寸在 640×480-1920×1080 范围内，数据集总共包含 14 个视频，7 个用于训练 、7 个用于测试，视频时长在 20s-1min25s 之间，人群分布较稀疏，只在 MOT17-03、MOT17-04 两个子数据集中，人群较为密集，最高人群密度可达到 69.7%，含了室内/室外，白天/晚上等多种场景。

（2）MOT20 数据集

MOT20 数据集从 3 个场景中提取了 8 份全新的稠密人群序列。这些序列包含了室内/室外，白天/晚上场景。输入尺寸在 1173×880-1920×1080 范围内，数据集总共包含 8 个视频，4 个用于训练 、4 个用于测试，MOT20 数据集以密集行人跟踪为背景，行人密度极高，最拥挤的视频平均每帧可达 245 人，数据集包含 13410 帧、6869 个行人和 2259143 个行人检测框。

5. 评估指标

(一）MOTA

多目标跟踪的准确度,计算跟踪所有帧中所有目标的误检、漏检和错误匹配，其中 FN_t、FP_t和 $IDSW_t$分别是 t 帧时漏检、误检和错误匹配的数量，g_t是地面真值目标矩形框的数量。计算公式为

$$MOTA = 1 - \frac{\sum_t (FN_t + FP_t + IDSW_t)}{\sum_t g_t}$$

(二）MOTP

多目标跟踪的精度：用来量化检测器的定位精度，d^i_t代表第 i 个检测目标

与给它分配的真值之间在所有帧中的平均度量距离,c_t代表在当前帧匹配成功的数目;计算公式为：

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t}$$

(三）FPS

是指画面每秒传输帧数。

(四）IDsW

目标身份切换的总次数：即 ID 改变的次数。

(五）MT

最多跟踪的目标数量,（Mostly Tracked）是指跟踪目标在 80%的时间以上都能够成功匹配的轨迹数量。

(六）ML

最少丢失的目标数量：（Mostly Lost）是指跟踪目标在 20%的时间以下能成功匹配的轨迹数量。

(七）IDF_1

IDF_1即是指每个目标框中目标 ID 识别的 F 值。计算公式：

$$IDF1 = \frac{2}{\frac{1}{IDP} + \frac{1}{IDR}} = \frac{2IDTP}{2IDTP + IDFP + IDFN}$$

其中 IDP 识别精确度是指每个行人框中行人 ID 识别的精确度，IDR 识别召回率是指每个行人框中行人 ID 识别的召回率。 IDTP 是真正 ID 数， IDFP 是假正 ID 数，IDFN 是假负 ID 数。

6. 预搭建框架

算法框架可以分为三个部分：行人检测、行人重识别和行人数据关联。首先

利用行人检测算法检测给定视频中出现的行人，然后将已检测行人的图片与给定的人脸图库相关联构建行人图库，再利用行人重识别算法为已检测图片分配一个标签，最后整合行人检测和行人重识别算法得到的行人信息，生成出最终包含所有行人信息的 JSON 文件。其中行人数据关联部分的主要作用是对不同摄像头中出现的同一目标进行关联。

7. 参考文献

[1] 彭建盛,许恒铭,李涛涛,侯雅茹.生成式与判别式视觉目标跟踪算法综述[J].科学技术与工程,2021,21(35):14871-14881.

[2] Wang N,Zhou W,Wang J,et al.Transformer meets tracker:exploiting temporal context for robust visual tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE Press,2021:2103.11681.

[3] Chen X,Yan B,Zhu J,et al.Transformer Tracking[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE Press,2021:2103.15436.

[4] Ergys Ristani Carlo Tomasi,et al.Features for Multi-Target Multi-Camera Tracking and Re-Identification[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE Press,2018.6

[5] Yifu Zhang • Chunyu Wang,et al.FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking arXiv:2004.01888v6 [cs.CV] 19 Oct 2021

[6] 熊月. 跨摄像头多目标跟踪的研究[D]. 华中科技大学,2019.DOI:10.27157/d.cnki.ghzku.2019.003841.

[7] Ergys Ristani, Carlo Tomasi. Tracking Multiple People Online and in Real Time[C]. Proceedings of Asian Conference on Computer Vision (ACCV), pp.444-459, November 1-5,2014, Singapore

[8] Ergys Ristani, Francesco Solera, Roger S. Zou, Rita Cucchiara, Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking[C]. Proceedings of European Conference on Computer Vision (ECCV) Workshop on Benchmarking Multi-Target Tracking, October 8-16, 2006, Amsterdadm, Netherlands

[9] Lucas Beyer, Stefan Breuers, Vitaly Kurin, Bastian Leibe. Towards a Principled Integration of Multi-Camera Re-Identification and Tracking through Optimal Bayes Filters[C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp.1444-1453, July 21, 26, 2017, Honolulu, Hawaii, USA

[10] Zhe Cao, Tomas Simon, Shih-En Wei, Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7291-7299,

July 21-26, 2017, Honolulu, Hawaii, USA