**地铁车站监控视频内容分析与**

**技术要求(Ver 0.2)**

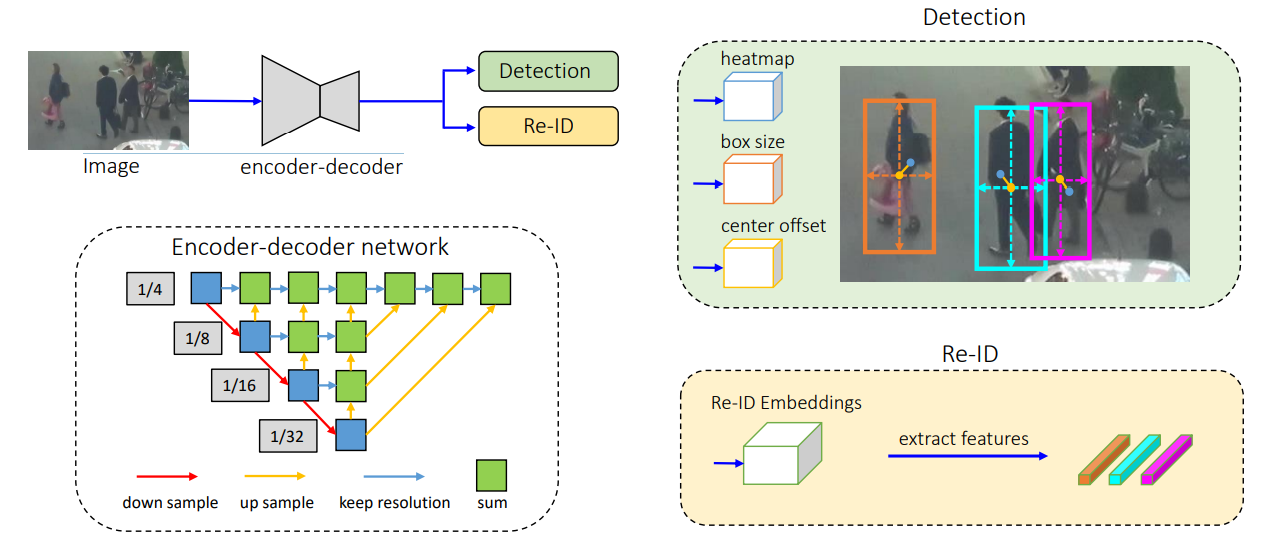
1. **多目标跟踪简介**

**1.1 MOT简介**

多目标跟踪是计算机视觉中的一个重要问题，具有广泛的应用。将MOT描述为在单个网络中进行对象检测和重新识别的多任务学习很有吸引力，因为它允许两个任务的联合优化，并且具有较高的计算效率。然而，我们发现这两项任务往往相互竞争，需要仔细处理。特别是，以前的工作通常将re-ID视为次要任务，其准确性受到首次检测因素的严重影响。因此，网络主要偏向于检测任务而不是跟踪任务，这对re-ID任务不公平。为了解决这个问题，我们提出了一种简单而有效的方法，称为基于无锚目标检测体系结构CenterNet的FairMOT。请注意，它不是CenterNet和re-ID的简单组合。相反，我们提出了一系列详细的设计，这些设计对于通过深入的实证研究获得良好的跟踪结果至关重要。该方法实现了高精度的检测和跟踪。在几个公共数据集上，该方法的性能大大优于最先进的方法。

多目标跟踪一直是计算机视觉领域的一个长期目标。目标是估计视频中显示的感兴趣对象的轨迹。该问题的成功解决可以立即造福于许多应用，如智能视频分析、人机交互、人类活动识别，甚至社会计算。现有的方法大多试图通过两个独立的模型来解决这个问题：检测模型首先通过每帧中的边界框来检测感兴趣的对象，然后关联模型从每个边界框对应的图像区域中提取重新识别（re-ID）特征，将检测链接到现有轨迹之一，或根据特征上定义的某些指标创建新轨迹。最近，在目标检测和重新识别方面分别取得了显著进展，这反过来又提高了整体跟踪精度。然而，这两步方法存在可扩展性问题。当环境中存在大量对象时，它们无法实现实时推理速度，因为这两个模型不共享特征，并且需要在视频中独立地为每个边界框应用re-ID模型。随着多任务学习的成熟，利用单个网络估计对象和学习识别特征的单次跟踪器受到了越来越多的关注。例如向掩码R-CNN添加re-ID分支，以提取每个提案的re-ID特征。它通过重新使用re-ID网络的主干功能来减少推理时间。但与两步模型相比，性能显著下降。事实上，检测精度仍然很高，但跟踪性能下降了很多。例如，ID开关的数量大幅增加。结果表明，将这两项任务结合起来是一项非常重要的任务，应谨慎对待。

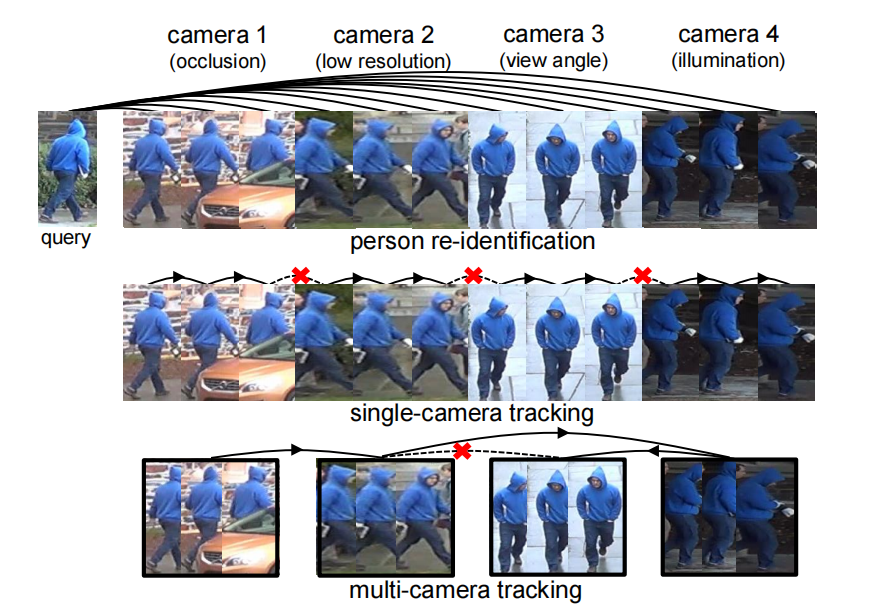
在本文中，我们调查了失败背后的原因，并提出了一个简单而有效的解决方案。确定了导致故障的三个因素。第一个问题是由锚引起的。锚最初是为目标检测而设计的。然而，我们发现锚不适合提取re-ID特征，原因有二。首先，基于锚的单次跟踪器（如Track R-CNN）忽略了re-ID任务，因为它们需要锚首先检测对象，然后根据检测结果提取re-ID特征（当检测结果不正确时，re-ID特征无效）。因此，当两个任务之间发生竞争时，它将有利于检测任务。锚在训练re-ID特征时也会引入很多模糊性，因为一个锚可能对应多个身份，而多个锚可能对应一个身份，尤其是在拥挤的场景中。第二个问题是由两个任务之间的功能共享引起的。检测任务和重新识别任务是两个完全不同的任务，它们需要不同的特征。一般来说，re-ID特征需要更多的低级特征来区分同一类的不同实例，而检测特征需要对不同实例相似。一次性跟踪器中的共享特征将导致特征冲突，从而降低每个任务的性能。第三个问题是由特征维度引起的。re-ID特征的维数通常高达512或1024，远高于目标检测。我们发现维度之间的巨大差异会影响两个任务的性能。更重要的是，我们的实验表明，学习用于“联合检测和re-ID”网络的低维re ID特征可以实现更高的跟踪精度和效率。这也揭示了MOT任务和re-ID任务之间的差异，这在MOT领域被忽视。在这项工作中，我们提出了一种称为FairMOT的简单方法，它解决了图1所示的三个问题。FairMOT构建在CenterNet之上。特别是，检测和重新识别任务在FairMOT中得到同等对待，这与之前的“检测优先，重新识别次要”框架有本质区别。

如图1显示了FairMOT的概述。它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。它揭示了MOT中检测和re-ID之间的关系，并为设计单次视频跟踪网络提供了指导。我们的贡献如下：

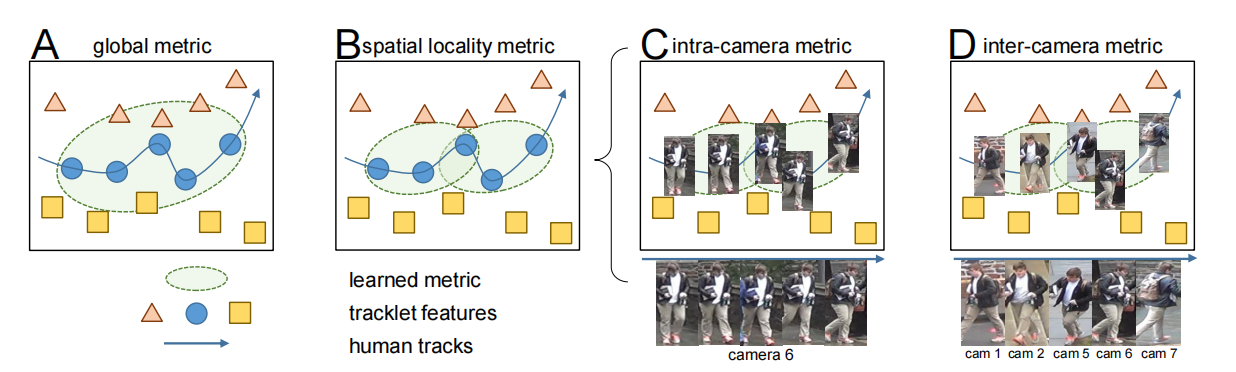
1. 我们根据经验证明，流行的基于锚点的一次性MOT架构在学习有效的re-ID特征方面存在局限性，而这些特征被忽略了。这些问题严重限制了这些方法的跟踪性能。
2. 我们提出FairMOT来解决公平问题。FairMOT构建在CenterNet之上。虽然所采用的技术本身大多不是新颖的，但我们有一些新的发现对MOT很重要。这些是既新颖又有价值。
3. 实现的公平性使我们的FairMOT能够获得高水平的检测和跟踪精度，并在多个数据集（如2DMOT15）上大大优于以前的最先进方法，MOT16、MOT17和MOT20。

**1.2 MTMCT简介**

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。我们认为，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，我们引入了局部感知外观度量（LAAM）。具体来说，我们设计了用于单摄像机跟踪的摄像机内度量和用于多摄像机跟踪的摄像机间度量。这两个指标都使用从其相应的局部邻域采样的数据对进行训练，而不是从re-ID角度进行全局采样。数据表明，局部学习的度量可以成功地应用于几个全局学习的里德特征之上。利用该方法，我们在DukeMTMC数据集上报告了最新的性能，并对CityFlow数据集进行了实质性改进。

图2: re-ID和MTMCT之间的差异。给定一个查询，re-ID在全局库中搜索所有摄像机的真实匹配。相比之下，在单摄像机跟踪（SCT）和多摄像机跟踪（MCT）中，MTMCT在局部邻域内匹配。具体来说，在MCT中，当目标位于摄像机2中时，我们不考虑摄像机3，因为目标永远不会连续出现在这两个摄像机中（摄像机可能太远）。

多目标多摄像机跟踪（MTMCT）旨在始终识别和定位多摄像机系统中的所有目标。在一个密切相关的任务中，给定一个探针图像，重新识别（re ID）系统搜索库以检索相同身份的图像。MTMCT系统由多个组件组成，包括检测、相似性估计和数据关联。基于相似性估计，首先在单摄像机跟踪（SCT）中关联检测到的对象边界框，然后在多摄像机跟踪（MCT）中关联检测到的对象边界框。由于目标（例如行人、车辆）具有连续的轨迹，大多数跟踪系统仅搜索局部邻域以获取数据关联。例如，在许多MTMCT系统中采用了时间滑动窗口技术。在SCT中，时间滑动窗口将局部匹配邻域限制为连续帧在摄像机内。在MCT中，这些窗口将局部匹配邻域限制为目标可能连续出现的相邻摄像机。外观特征是MTMCT的驱动力。目前，跟踪社区与re-ID社区共享非常相似的外观表示和深度学习架构。也就是说，该特征从整个列车集全局学习，然后应用于SCT和MCT。

图3：（A）从整个列车集学习的全局度量考虑了所有数据。该指标相对稳健，但具有松弛的决策边界（通常存在错误）。（B）本文提出了一种局部学习度量。它具有严格的决策边界，并且更敏感。在MTMCT中，数据关联通常在局部邻域内，而不是re-ID中的全局匹配。因此，局部度量学习更适合。提出的位置感知外观度量（LAAM）具有（C）用于SCT的摄像机内度量和（D）用于MCT的摄像机间度量。前者是在同一摄像机的一段时间内在轨迹对上学习的。后者在相邻摄像机的轨迹对上学习（目标可能会连续出现）。

然而，MTMCT和re-ID有其不同之处。首先，Re-ID系统通常无法访问轨迹、摄像机拓扑和其他时空线索。其次，SCT是MTMCT的一个非常重要的部分。相反

Re-ID忽略与探针初始化评估相同摄像机的候选摄像机。在本研究中，我们进一步探讨了第三个也是最重要的区别：跟踪中的局部匹配与re-ID中的全局匹配。如图2所示，MTMCT仅关联局部邻域内的数据（较小的外观方差）。具体来说，在SCT中，仅搜索单摄像机视频中的连续帧。在MCT中，我们在有限的摄像机池中进行匹配，因为时间滑动窗口缩小了搜索范围。另一方面，给定查询图像，re-ID搜索覆盖所有摄像机的全局库（外观差异较大）。这种局部与全局的差异是不平凡的。当直接应用re ID特征时，跟踪中的局部匹配和全局re-ID外观特征之间的不匹配可能会影响MTMCT性能。我们发现高性能的re-ID特性并不一定会导致良好的MTMCT性能。事实上，re-ID模型学习处理各种环境差异。然而，在SCT中，我们只需要匹配外观变化相对较小（与交叉摄像机相比）的连续帧。在MCT中，我们仍然不需要考虑所有环境变化。例如，MCT的特征不需要同时对视点方差和低分辨率具有鲁棒性（图2），因为这些摄像机中永远不会连续出现目标。在这种情况下，更强的re-ID外观特征不一定会导致更高的MTMCT结果。为了适应跟踪中的局部匹配过程，本文提出了一种局部感知外观度量（LAAM）。具体来说，对于SCT，我们从单个摄像机内的连续帧中采样训练数据对。对于MCT，从相邻摄像机中选择训练数据对（目标可能会连续出现）。使用两种采样策略，我们得到了用于SCT的摄像机内度量和用于MCT的摄像机间度量（见图3）。我们表明，LAAM可以有效地提高多个数据集的跟踪精度，包括行人数据集DukeMTMC和车辆数据集CityFlow。它还可以应用于多个ID功能，如IDE、PCB和三重态功能。通过竞争跟踪器，我们在DukeMTMC数据集上报告了最先进的准确性。

1. **多目标跟踪发展**

性能最好的MOT方法通常遵循检测跟踪范式，即首先检测每帧中的对象，然后随着时间的推移将其关联。我们根据使用单个模型还是单独模型来检测对象和对象，将现有工作分为两类提取关联特征。我们讨论了这些方法的优缺点，并将其与我们的方法进行了比较。

**2.1通过分离模型进行检测和跟踪**

**2.1.1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。

**2.1.2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。

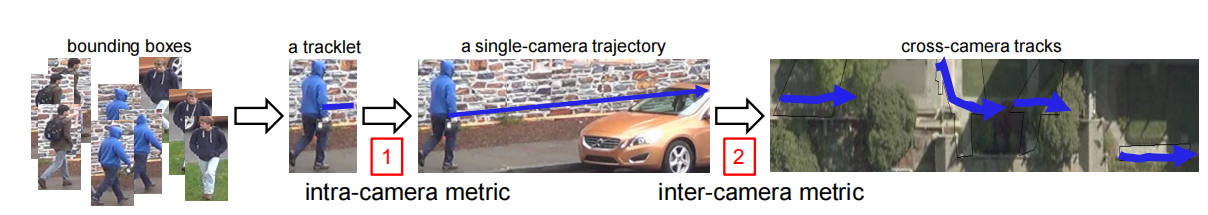
**2.2单个模型的检测和跟踪**

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，该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，这些方法仅关联相邻帧中的对象，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。

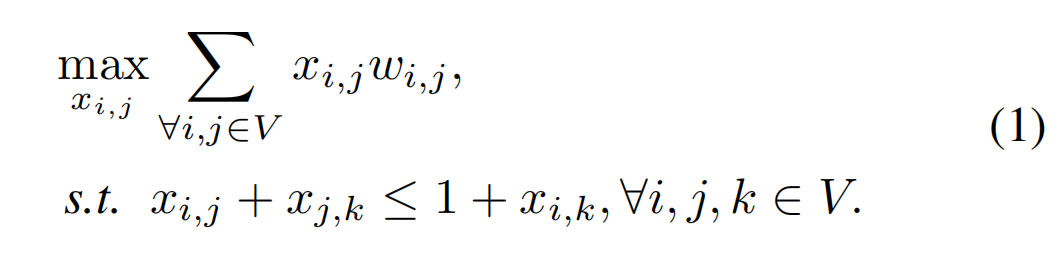
**2.3视频对象检测**

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。

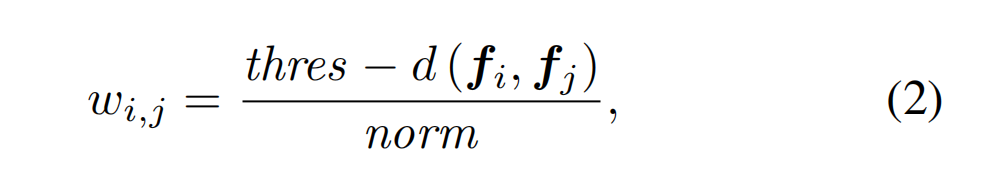
图4:MTMCT系统概述。给定对象边界框，我们首先将边界框连接成短而可靠的轨迹。然后，将轨迹合并为单摄像机轨迹。最后，将单摄像机轨迹关联起来，形成交叉摄像机轨迹。所提出的LAAM包括摄像机内度量和摄像机间度量。摄像机内/摄像机间度量分别用于生成单摄像机轨迹和跨摄像机轨迹。

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。
5. **MTMCT系统概述**

**问题公式** 我们遵循中介绍的基于图的问题lem公式。我们将观察值（包围盒、轨迹、轨迹）表示为节点，并将它们之间的相似性表示为图G=（V，E）中的加权边。对于一对节点i，j∈ V、wi、j∈ E是指它们之间的估计相似性，xi，j∈ {−1，1}表示它们是否具有相同的身份。

优化问题公式如下：，

等式1最大化了组内相似性，最小化了组间相似性，并加强了传递性（如果两个数据都与第三个数据点具有相同的身份，则两个数据应该具有相同的身份）。事实上，性能更好的相似性估计wi，j将使该优化问题更容易，并提高关联精度。

**相似性估计** 在基线中，给定一对CNN特征fi和fj，其外观相似性分数wi，j计算为，

其中d（·，·）是一个距离度量，我们在这里只使用欧几里得距离。thres=（µn+µp）/2，norm=（µn）−µp）/2。µp和µn分别表示相同和不同身份的平均特征距离。

**数据关联** 图4描述了整个数据关联过程。首先，将对象边界框连接到tracklet中。然后，将轨迹匹配为单摄像机轨迹。最后，将单摄像机轨迹关联起来，形成跨摄像机轨迹。对于SCT，我们使用短时间滑动窗口来关联轨迹。对于MCT，由于行走时间长，在数据关联中使用了更长的时间滑动窗口目标穿过摄像机的时间。

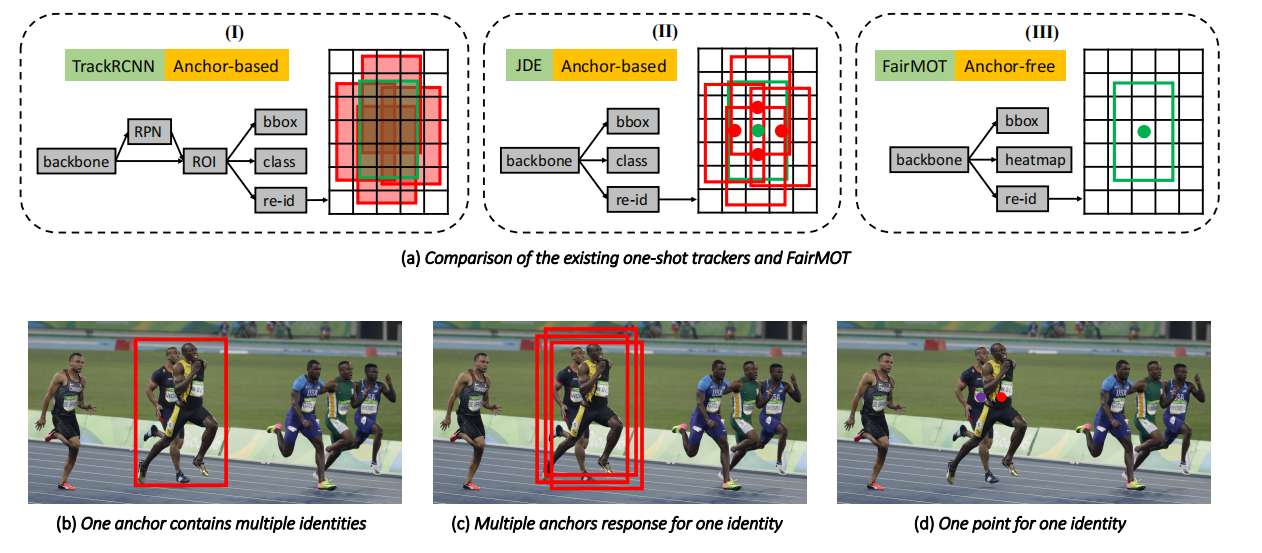
**4.跟踪器中的不平衡问题**

在本节中，我们讨论了现有跟踪器中出现的三个不平衡问题，这些问题通常会导致跟踪性能下降。

**4.1锚引起的不平衡**

现有的跟踪器，如Track R-CNN和JDE，大多是基于锚点，因为它们是直接从基于锚点的对象检测器（如YOLO和Mask R-CNN）修改而来的。然而，我们发现基于锚的设计不适合学习re-ID特征，尽管检测结果良好，但这会导致大量ID切换。下面我们从三个角度来解释这个问题。

1. **忽略re-ID任务**轨迹R-CNN以级联方式运行，首先估计对象建议（框），然后从中汇集特征估计相应的re-ID特征。Re-ID特征的质量在很大程度上取决于训练期间提议的质量（如果提议不准确，则re-ID特征无效）。因此，在训练阶段，该模型严重偏向于估计准确的对象建议，而不是高质量的re-ID特征。因此，现有跟踪器的标准“检测优先，re-ID二次”设计使得re-ID网络没有被公平地学习。

图5（a）轨迹R-CNN将检测视为主要任务，re ID视为次要任务。轨道R-CNN和JDE都是基于锚的。红色框表示正定位，绿色框表示目标对象。这三种方法提取re-ID特征的方式不同。Track R-CNN使用ROI Align为所有正定位提取re ID特征。JDE在所有正锚点的中心提取re ID特征。FairMOT在对象中心提取re ID特征。（b） 红色锚包含两个不同的实例。因此，它将被迫预测两个相互冲突的类。（c） 具有不同图像块的三个不同锚是预测相同身份的响应。（d） FairMOT仅在对象中心提取re ID特征，可以缓解（b）和（c）中的问题。

1. **一个锚对应多个身份** 基于锚定的方法通常使用ROI-Align从提案中提取特征。ROI-Align中的大多数采样位置可能属于其他干扰实例或背景，如图5所示。因此，提取的特征在准确和有区别地表示目标对象方面并不理想。相反，我们在这项工作中发现，仅在单个点（即估计的对象中心）提取特征明显更好。
2. **多个锚对应于一个身份** 多个相邻锚对应于不同的图像块，只要它们的IOU足够大，就可能被迫估计相同的身份。这给训练带来了严重的模糊性。见图5。另一方面，当图像受到小扰动时，例如由于数据增加，同一个anchor可能被迫估计不同的恒等式。另外，目标检测中的特征图通常下采样8/16/32倍，以平衡精度和速度。这对于对象检测来说是可以接受的，但对于学习re-ID特征来说太粗糙了，因为在粗定位点提取的特征可能与对象中心不对齐。

**4.2特征引起的不平衡**

对于单次跟踪器，大多数特征在目标检测和重新识别任务之间共享。但众所周知，它们实际上需要来自不同层的功能才能达到最佳效果。特别是，目标检测需要深度特征来估计对象类别和位置，但需要重新识别，需要低级外观特征来区分同一类的不同实例。从多任务丢失优化的角度来看，检测和重新识别的优化目标存在冲突。因此，平衡这两项任务的损耗优化策略非常重要。

**4.3特征尺寸引起的不平衡**

以前的re-ID工作通常学习非常高维的特征，并在其领域的基准上取得了有希望的结果。然而，我们发现学习低维特征实际上对MOT是一次更好的尝试，有三个原因：（1）由于两个任务的竞争，高维re-ID特征显著损害了目标检测精度，这反过来也对最终的跟踪精度产生负面影响。因此，考虑到目标检测中的特征维数通常很低（类数+方框位置），我们建议学习低维re-ID特征来平衡这两项任务；（2）MOT任务不同于re-ID任务。MOT任务仅在两个连续帧之间执行少量的一对一匹配。re-ID任务需要将查询与大量候选匹配，因此需要更具辨别力和高维的re-ID特征。所以在MOT中，我们不需要高维特征；（3）如我们的实验所示，学习低维re-ID特征可以提高推理速度。

**5.实验分析**

**5.1数据集和指标**

以下简要介绍了六个训练数据集：ETH和CityPerson数据集仅提供框注释，因此我们仅在其上训练检测分支。加州理工学院、MOT17、CUHK-SYSU和PRW数据集提供了框注释和身份注释，允许我们进行训练两个分支。ETH中的一些视频也出现在MOT17的测试集中，这些视频从训练数据集中删除以进行公平比较。第4.4节描述了总体培训策略，与相同。对于我们方法的自监督训练，我们使用了仅包含对象边界框注释的CrowdHuman数据集。我们在四个基准测试集上评估了我们的方法：2DMOT15、MOT16、MOT17和MOT20。我们使用平均精度（AP）来评估检测结果。接下来，我们使用假接受率为0.1的真阳性率（TPR）来评估re ID特征。特别是，我们提取对应于地面真值框的re-ID特征，并使用每个特征进行检索N最相似的候选人。我们报告的真阳性率为假接受率0.1(TPR@FAR=0.1)。 注意，TPR不受检测结果的影响，并且反映了re-ID特征的质量。我们使用CLEAR度量（即MOTA、IDs）和IDF1来评估整体跟踪精度。

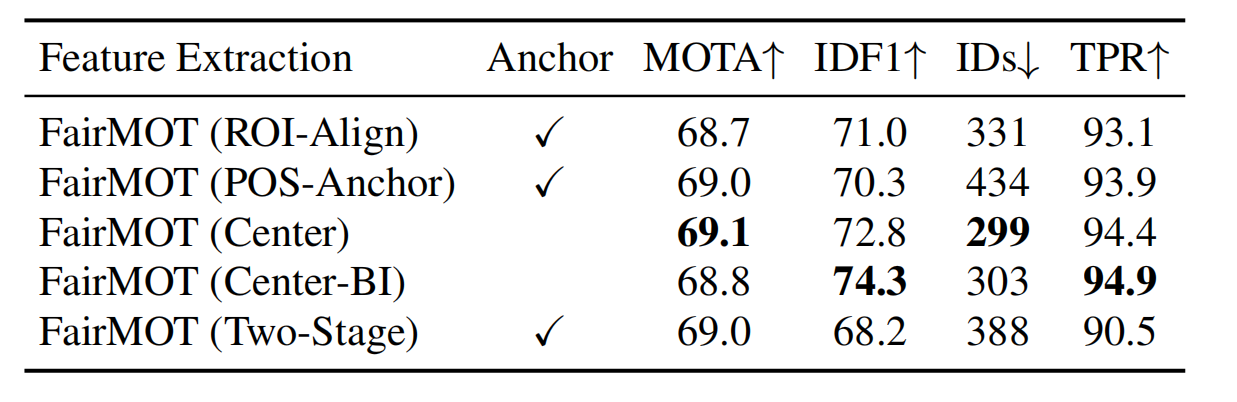
**5.2实验细节**

我们使用中提出的DLA-34变体作为默认主干。在COCO数据集上预先训练的模型参数用于初始化我们的模型。我们使用Adam优化器对模型进行30个阶段的训练，起始学习率为10−4、学习率下降到10−5分20个时代。批量大小设置为12。我们使用标准的数据增强技术，包括旋转、缩放和颜色抖动。输入图像大小调整为1088×608，特征图分辨率为272×152，在两个RTX GPU上执行步骤大约需要30个小时。

**5.3分离研究**

在本节中，我们通过仔细设计一些基线方法，对FairMOT中的三个关键因素进行了严格的研究，包括无锚特征提取、特征融合和特征维数。

**表1.**不同re ID特征提取的比较（采样）MOT17验证集的策略。其余模型为保持不变以进行公平比较。↑ 意味着越大越好↓意味着越小越好。最佳结果以粗体显示。

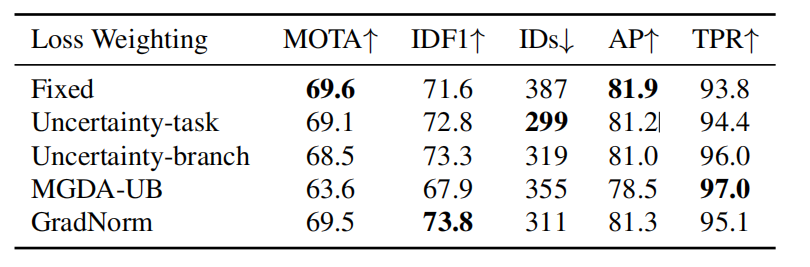
**5.3.1锚**

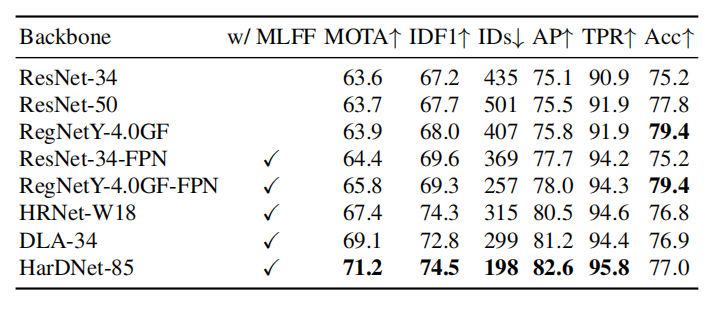
我们评估了四种从之前工作中经常使用的检测框中采样re-ID特征的策略。第一种策略是用于跟踪R-CNN的ROI对齐。它使用ROI-Align从检测到的提议中采样特征。如前所述，许多采样位置偏离对象中心。第二种策略是JDE中使用的位置锚。它从也可能偏离对象中心的正锚点采样特征。第三种策略是FairMOT中使用的“中心”。它仅对对象中心的特征进行采样。回想一下，在我们的方法中，re-ID特征是从离散化的低分辨率地图中提取的。为了对特征进行采样在精确的目标位置，我们还尝试应用双线性插值（中心Bi）来提取更精确的特征。我们还评估了一种两阶段方法，首先检测对象边界框，然后提取re-ID特征。在第一阶段，检测部分与我们的FairMOT相同。在第二阶段，我们基于检测到的边界盒使用ROI对齐来提取主干特征，然后使用re-ID头来获得re-ID特征。两阶段方法和一阶段“ROI对齐”方法之间的主要区别在于，两阶段方法的re-ID特征依赖于检测结果，而一阶段方法的re-ID特征在训练期间不依赖于检测结果。结果如表1所示。请注意，这五种方法都建立在FairMOT之上。唯一的区别在于它们如何从检测到的框中采样re-ID特征。首先，我们可以看到，与ROI-Align、POS-Anchor和两阶段方法相比，我们的方法（中心）获得了显著更高的IDF1分数和真阳性率（TPR）。该度量独立于目标检测结果，并且忠实地反映了re-ID特征的质量。此外，我们方法的ID开关（ID）数量也明显小于两个基线。结果验证了在对象中心的采样特征比以前工作中使用的策略更有效。双线性插值（中心Bi）实现了比中心更高的TPR，因为它在更精确的位置对特征进行采样。两阶段方法损害了re-ID特征的质量。

**5.3.2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2.**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。

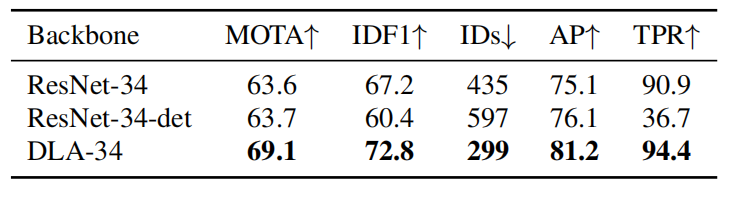
**表3.**MOT17数据集上不同主干的比较验证。“MLFF”是多层特征融合的简称。“Acc”是ImageNet分类精度的缩写。ImageNet分类精度的结果来自骨干网络的原始论文。最佳结果以粗体显示。

结果如表2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。这意味着该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。

**5.3.3多层特征融合**

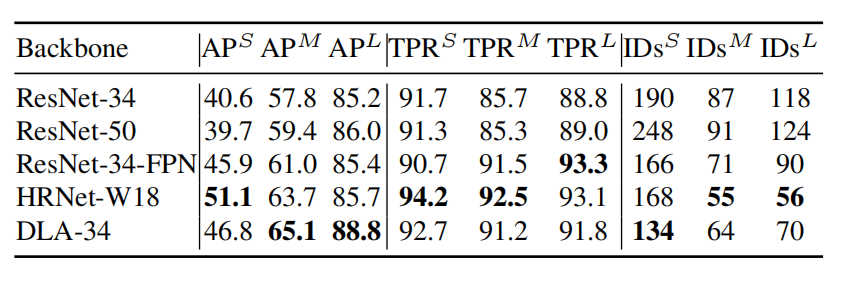
我们比较了一些主干网，如vanilla ResNet、特征金字塔网络（FPN）、高分辨率网络（HRNet）、DLA、HarDNet和RegNet。请注意，这些方法的其余因素（如训练数据集）均被控制为相同，以便进行公平比较。特别是，对于所有方法，最终特征图的步长为4。我们为vanilla ResNet和RegNet添加了三个向上采样操作，以获得第四步的特征图。我们将这些主干分为两类，一类没有多层融合，另一类有（即FPN、HRNet、DLA和HarDNet）。结果如表3所示。我们还列出了ImageNet分类精度Acc，以证明在一项任务中具有强大的主干，但这并不意味着它也会在MOT中取得好成绩。因此，对MOT进行详细的研究是必要和有益的。通过比较ResNet-34和ResNet-50的结果，我们发现盲目使用较大的网络并不显著

改善MOTA测量的整体跟踪结果。特别是，re-ID功能的质量几乎没有从更大的网络中受益。例如，IDF1仅从67.2%提高到67.7%，TPR分别从90.9%提高到91.9%。此外，ID交换机的数量甚至从435个增加到501个。通过比较ResNet-50和RegNetY-4.0GF，我们可以发现，使用更强大的主干网也可以获得非常有限的增益。RegNetY-4.0GF的re-ID度量TPR与ResNet-50（91.9）相同。而ImageNet分类精度提高了很多（79.4比77.8）。所有这些结果表明，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。

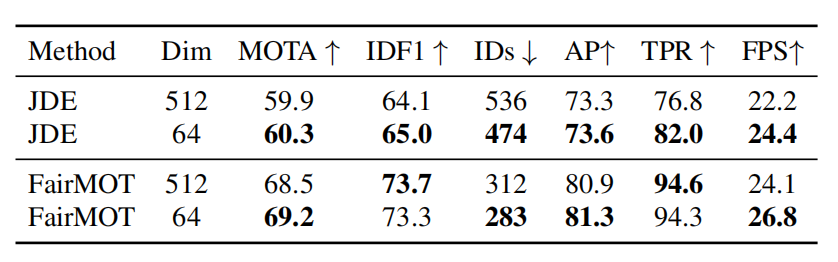
**表4.**检测和检测之间的特征冲突演示重新识别MOT17数据集验证集上的任务。“-det”是指只训练检测分支，随机训练re-ID分支已初始化。最佳结果以粗体显示。

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4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5.**不同脊柱对不同比例物体的影响。小：面积小于7000像素；中等：面积从7000到15000像素；大：大于15000像素的区域。最佳结果以粗体显示。

**5.3.4特征尺寸**

以前的单次跟踪器（如JDE）通常采用两步方法学习512维re-ID特征，无需进行烧蚀研究。然而，我们在实验中发现，特征维数实际上在平衡检测和跟踪精度方面起着重要作用。学习低维re-ID特征对检测精度的危害较小，并提高了推理速度。我们在不同的单次跟踪器上进行了实验，发现低维re-ID特征比高维（即512）re-ID特征获得更好的性能是一个普遍规律。我们在表6中评估了JDE和FairMOT的re-ID特征维度的多项选择。对于JDE，我们可以看到64在所有指标上都比512获得更好的性能。对于FairMOT，我们可以看到512实现了更高的IDF1和TPR分数，这表明更高维度的re-ID特征导致更强的辨别能力。然而，当我们将维数从512降到64时，MOTA分数会提高。这主要是检测和重新识别任务之间的冲突。特别是，我们可以看到，当我们降低re-ID特征的维数时，检测结果（AP）得到了改善。与re-ID任务不同，低维re-ID特征在MOT任务上实现了更好的性能和效率。

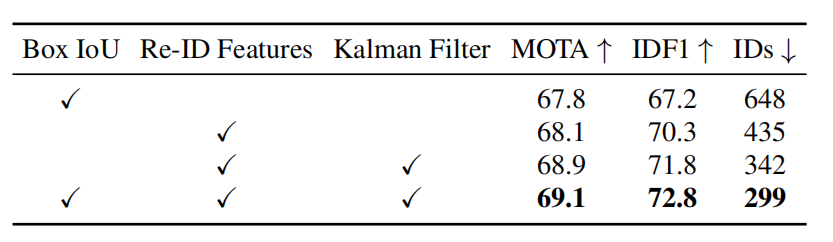
表6在MOT17的验证集上评估JDE和FairMOT的re ID特征维度。相同方法的最佳结果以粗体显示。

**5.3.5数据关联方法**

本节评估数据关联步骤中的三个成分，包括边界框IoU、re ID特征和卡尔曼滤波器。这些用于计算每对检测框之间的相似性。我们使用匈牙利算法来解决分配问题。表7显示了结果。我们可以看到，仅使用box-IoU会导致大量的ID切换。这对于拥挤的场景和快速的摄像机运动尤其如此。单独使用re ID功能可以显著增加IDF1并减少ID开关的数量。此外，添加卡尔曼滤波器有助于获得平滑（合理）的轨迹，从而进一步减少ID开关的数量。当对象部分被遮挡，其re ID特征变得不可靠。在这种情况下，重要的是利用盒IoU、re ID特征和卡尔曼滤波器来获得良好的跟踪性能。我们还提供了不同组件的详细运行时分解，包括检测、re ID匹配、卡尔曼滤波器和IoU匹配。我们在不同密度（每帧平均行人数）的序列上测试运行时。用于联合检测和重新识别的时间受密度的影响最小。卡尔曼滤波和IoU匹配花费的时间约为1ms或2ms，可以忽略不计。re-ID匹配所花费的时间随着密度的增加而线性增加。这是因为更新每个tracklet的re-ID特性需要花费大量时间。

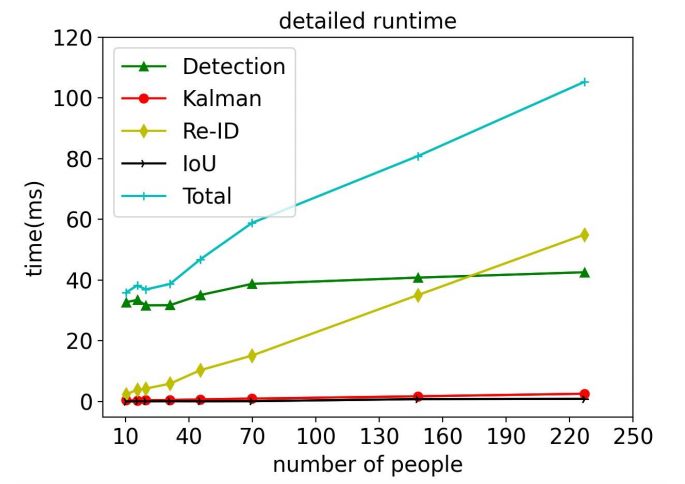
**图6.**re-ID特征识别能力的可视化。查询实例标记为红色框，目标实例标记为绿色框。使用基于不同策略（如第5.3.1节所述的中心、中心BI、ROI对齐和POS定位）和不同主干（如ResNet-34和DLA-34）提取的re ID特征计算相似性图。查询帧和目标帧从MOT17-09和MOT17-02序列中随机选择。

**表7.**数据关联模型中三个成分的评估。主干是DLA-34。最佳结果以粗体显示。

**5.3.6 Re-ID相似性的可视化**

我们使用re-ID相似图来证明图6中re-ID特征的区分能力。我们从验证集中随机选择两帧。第一个框架包含查询实例，第二个框架包含具有相同ID的目标实例。我们通过计算查询实例的re-ID特征与目标帧的整个re-ID特征图之间的余弦相似性来获得re-ID相似性图，分别如第5.3.1节和第5.3.3节所述。通过比较ResNet-34和ResNet-34-det的相似图，我们可以看出训练re-ID分支很重要。通过比较DLA-34和ResNet-34，我们可以看到多层特征聚合可以获得更具区分性的re-ID特征。在所有采样策略中，所提出的中心和在拥挤的场景中，中心BI可以更好地区分目标对象和周围对象。

**图7.**我们整个MOT系统不同部分花费的时间。我们从MOT17数据集中跟踪不同密度的序列和MOT20数据集。

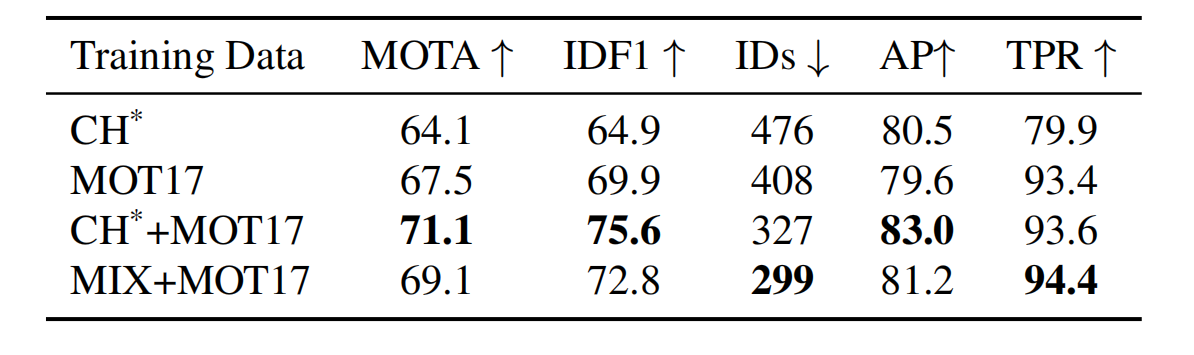
**5.4单幅图像训练**

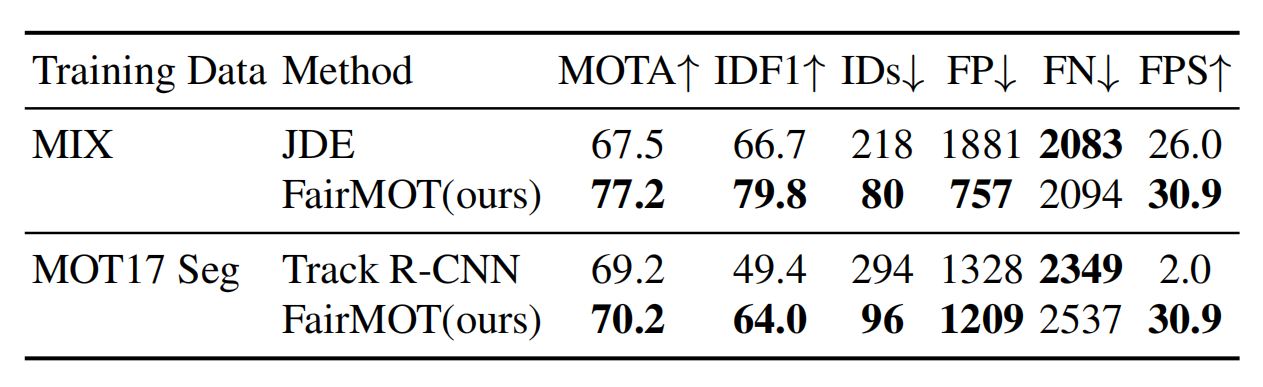
我们首先在CrowdHuman数据集上预训练FairMOT。特别是，我们为每个边界框指定一个唯一的标识标签，并使用第4.4节中描述的方法训练FairMOT。然后我们微调预训练的目标数据集MOT17上的模型。表8显示了结果。首先，预训练的模型可以直接用作跟踪器，并在上获得可接受的结果MOT数据集，例如MOT17。这是因为CrowdHuman数据集可以提高人体检测性能，并且具有很强的领域泛化能力。我们对re-ID特征的训练进一步增强了跟踪器的关联能力。第二，对CrowdHuman进行预培训大大优于MOT17数据集上的直接训练。第三，单图像训练模型甚至优于在具有身份标注的“混合”和MOT17数据集上训练的模型。结果验证了提出的单图像预训练的有效性，节省了大量注释工作，使FairMOT在实际应用中更具吸引力。

**5.5 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。

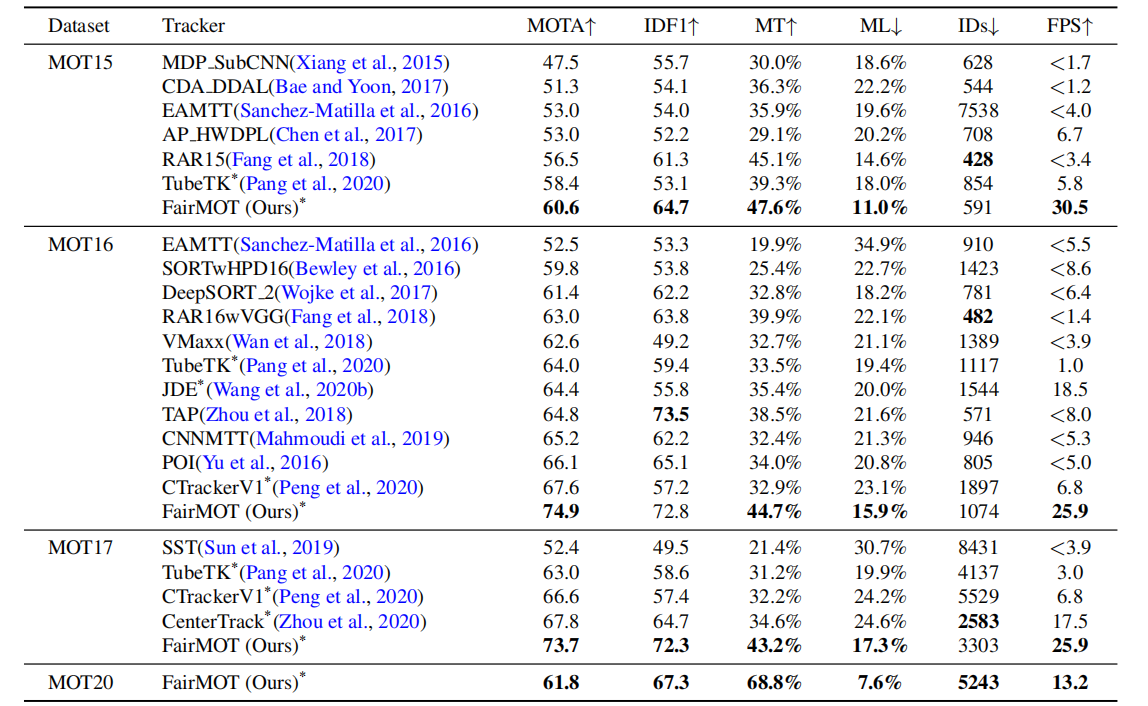
**表8.**单幅图像训练对MOT17验证集的影响。“CH”和“MIX”分别代表CrowdHuman和第5.1节中介绍的五个数据集。\*表示未使用标识注释。最佳结果以粗体显示。

**表9**.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。

**5.5.1与单次SOTA MOT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9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10.** “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。

**5.5.2与其他SOTA MOT方法的比较**

我们将我们的方法与最先进的跟踪器进行了比较，包括表10中的两步方法。由于我们不使用公共检测结果，因此采用了“私有检测器”协议。我们分别报告了2DMOT15、MOT16、MOT17和MOT20数据集的测试集的结果。请注意，所有结果均直接从官方MOT挑战评估服务器获得。我们的方法在所有在线和离线方法中排名第一，四个数据集上的跟踪器。特别是，它大大优于其他方法。这是一个非常强大的结果，尤其是考虑到我们的方法非常简单。此外，我们的方法实现了视频速率推断。相比之下，大多数高性能跟踪器（如）通常比我们的速度慢。我们的方法在最近的一次当地调查中排名第二，这进一步表明我们的方法实现了非常高的跟踪性能（表10）。

**5.5.3训练数据消融研究**

我们还使用表11中不同数量的训练数据评估了FairMOT的性能。当仅使用MOT17数据集进行训练时，我们可以实现69.8 MOTA，这已经优于使用更多训练数据的其他方法。当我们使用与JDE相同的训练数据时，我们可以实现72.9 MOTA，这显著优于JDE。此外，当我们在CrowdHuman数据集上执行单图像训练时，MOTA分数提高到73.7。结果表明，我们的方法不需要数据，这在实际应用中具有很大优势。**表11**

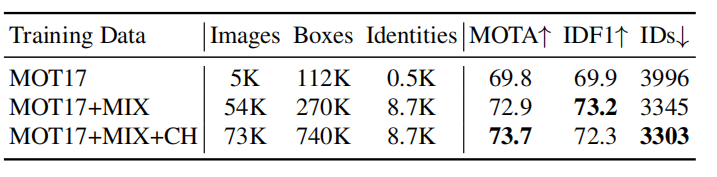
**5.6定性结果**

图8显示了FairMOT在MOT17测试集上的多个跟踪结果。从MOT17-01的结果可以看出，当两个行人相互交叉时，我们的方法可以借助高质量的re-ID特征来分配正确的身份。在这些情况下，使用边界框IOU的跟踪器通常会导致身份切换。从MOT17-03的结果可以看出，我们的方法在拥挤场景下表现良好。从MOT17-08的结果可以看出，当行人被严重遮挡时，我们的方法可以保持正确的身份和正确的边界框。MOT17-06和MOT17-12的结果表明，我们的方法可以处理大尺度变化。这主要归因于使用多层特征聚合。如MOT17-07和MOT17-14的结果所示，我们的方法可以准确地检测小物体。

从研究为什么以前的一次性方法无法获得与两步方法类似的结果开始，我们发现在目标检测和身份嵌入中使用锚是导致结果下降的主要原因。特别是，对应于对象不同部分的多个相邻锚可能负责估计相同的身份，这会导致网络训练的模糊性。此外，我们发现了以前MOT框架中检测和重新识别任务之间的特征不公平问题和特征维数问题。通过在无锚单炮深度网络中解决这些问题，我们提出了FairMOT。在跟踪精度和推理速度方面，它在多个基准数据集上大大优于以前最先进的方法。此外，FairMOT固有的训练数据高效性，我们提出了仅使用边界框注释图像的多目标跟踪器的单图像训练，这两种方法都使我们的方法在实际应用中更具吸引力。**图8**

