摘要

近年来，作为多目标跟踪的关键组成部分的目标检测和重新识别（re-ID）取得了显着进展。但是，很少有注意力集中在在单个网络中共同完成两项任务。我们的研究表明，先前的尝试最终导致准确性降低，这主要是因为 re-ID 任务的学习不充分，导致许多身份切换。不公平性有两个方面：（1）他们将 re-ID 视为次要任务，其准确性在很大程度上取决于主要检测任务。因此，训练在很大程度上偏向于检测任务，而忽略了重新 ID 任务。（2）他们使用 ROI-Align 提取直接从对象检测中借用的 re-ID 特征。但是，由于许多采样点可能属于令人不安的实例或背景，因此在表征对象时引入了很多歧义。为了解决这些问题，我们提出了一种简单的方法 FairMOT，该方法由两个同构分支组成，以预测像素级客观性得分和 re-ID 特征。任务之间实现的公平性使 FairMOT 能够获得高水平的检测和跟踪精度，并在多个公共数据集上大大超越了现有技术水平。

1 INTRODUCTION

多目标跟踪（MOT）是计算机视觉中的长期目标[1]，[2]，[3]，[4]，其目的是估计视频中感兴趣对象的轨迹。 问题的成功解决可以使许多应用受益，例如视频分析，动作识别，智能老人护理和人机交互。

现有的方法，例如[1]，[2]，[3]，[4]，[5]，[6]，[7]，通常通过两个单独的模型来解决该问题：检测模型首先将感兴趣的对象定位 通过在每个帧中包围框，然后关联模型为每个包围框提取重新标识（re-ID）特征，并根据在特征上定义的某些度量将其链接到现有轨道之一。 近年来，分别在对象检测[8]，[9]，[10]，[11]和 re-ID [3]，[12]方面取得了显着进步，这反过来又大大提高了整体跟踪性能。 但是，这些方法无法执行实时推断，尤其是在存在大量对象的情况下，因为这两个模型不共享特征，并且它们需要为视频中的每个边界框应用 re-ID 模型。

随着多任务学习的成熟[13]，使用单个网络估计对象并学习 re-ID 特征的单次跟踪器吸引了更多的关注[14]，[15]。 例如，Voigtlaender 等 [15]建议在 Mask R-CNN 的顶部添加一个 re-ID 分支，以使用 ROI-Align 获得建议的 re-ID 特征。 通过将骨干特征重新用于 re-ID 网络，可以减少推理时间。 不幸的是，与两步跟踪相比，跟踪精度显着下降。 特别地，ID 切换的数量大大增加。 结果表明，将这两项任务结合起来是一个不小的问题，应谨慎对待。 在本文中，我们旨在深入了解失败的原因，并提出一种简单而有效的方法。 特别地，确定了三个因素。

1.1 Unfairness Caused by Anchors

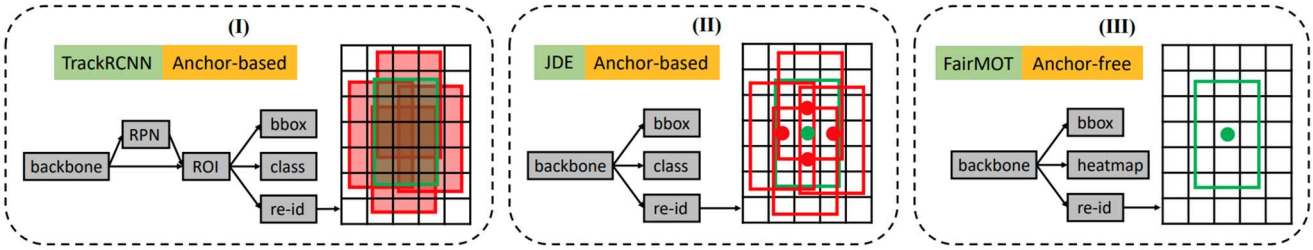
现有的单次跟踪器（例如 Track R-CNN [15]和 JDE [14]）大多基于锚，因为它们是直接从基于锚的对象检测器（如 YOLO [11]和 Mask R-CNN [9]）中修改而来的 。 然而，我们在这项研究中发现，基于锚的框架不适合于学习 re-ID 特征，尽管检测结果良好，但仍会导致大量 ID 切换。

Overlooked re-ID task: TrackR-CNN [15]以级联的方式运行，它首先估计目标的 proposal，然后从 proposal 中合并 reID 特征以估计相应的 reID 特征。 值得注意的是，reID 特征的质量在很大程度上取决于 proposal 的质量。 结果，在训练阶段，该模型严重偏向于去估计准确的目标框，而不是高质量的 re-ID 特征。 总而言之，这是种事实上的标准的“检测优先，reID 第二”的框架使得对 re-ID 网络的并不公平。

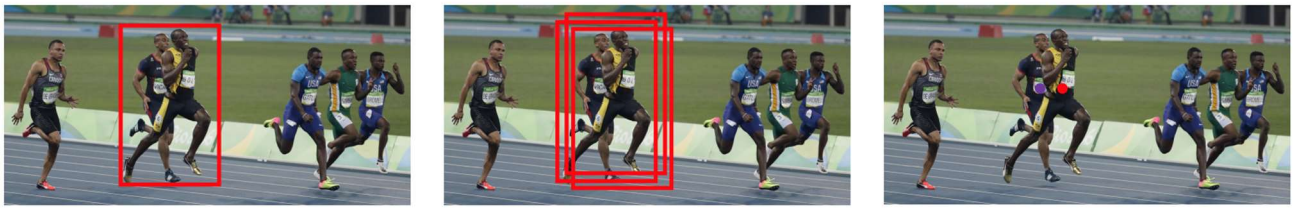
One anchor corresponds to multiple identities: 基于锚的方法通常使用 ROI-Pool 或 ROI-Align 从每个 proposal 中提取特征。 如图1所示，ROI-Align 中的大多数采样位置可能属于其他干扰实例或背景。结果，就准确而有区别地表示目标对象而言，提取的特征并不是最佳的。相反，我们在这项工作中发现，仅在估计的对象中心处提取特征要好得多。

Multiple anchors correspond to one identity: 在[15]和[14]中，只要它们的 IoU 足够大，可能会迫使对应于不同图像 patch 的多个相邻锚框估计相同的身份。 这就给训练带来了很大的歧

义。 有关说明，请参见图1。 另一方面，当图像经历小的扰动时，例如由于数据增强，有可能迫使相同的锚框估计不同的身份。 此外，对象检测中的特征图通常会下采样8/16/32次，以平衡精度和速度。 这对于物体检测是可以接受的，但是对于学习 re-ID 特征来说太粗糙了，因为在粗锚处提取的特征可能未与物体中心对齐。



(a) *Comparison of the existing one-shot trackers and FairMOT*



(b) *One anchor contains multiple identities* (c) *Multiple anchors response for one identity* (d) *One point for one identity*

图一、(a) Track R-CNN 将检测视为主要任务，将 re-ID 视为次要任务。 Track R-CNN 和 JDE 都是基于锚的。 红色框代表正锚，绿色框代表目标对象。 三种方法提取 re-ID 特征的方式有所不同。 Track R-CNN 使用 ROI-Align 提取所有正锚的 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) 红色锚点包含两个不同的实例。 因此，将不得不预测两个相互冲突的类。 (c) 具有不同图像斑块的三个不同锚点是用于预测相同身份的响应。 (d) FairMOT 仅在对象中心提取 re-ID 特征，并且可以缓解 (b) 和 (c) 中的问题。

1.2 Unfairness Caused by Features

对于单次跟踪器，大多数特征在目标检测和 re-ID 任务之间共享。 但是众所周知，它们实际上需要来自不同层的特征才能获得最佳效果。 特别是，目标检测需要深层和抽象的特征来估计对象的类和位置，而 re-ID 则更多地关注于低层外观特征以区分同一类的不同实例。 我们从经验上发现，多层特征聚合通过允许两个任务（网络分支）从多层聚合特征中提取他们需要的任何特征，从而有效地解决了这一矛盾。 如果没有多层融合，该模型将偏向主要检测分支并生成低质量的 re-ID 特征。 此外，多层融合融合了来自具有不同感受野的层中的特征，还提高了处理对象比例变化的能力，这在实践中非常普遍。

1.3 Unfairness Caused by Feature Dimension

先前的 re-ID 工作通常学习非常高的维度特征，并在其领域的基准上取得了可喜的成果。然而，我们发现学习低维特征对于 one-shot MOT 实际上更好，原因有以下三个：(1) 尽管学习高维 re-ID 特征可能会略微提高其区分对象的能力，但由于显着影响了物体检测的准确性，两项任务之间的竞争对最终的跟踪准确性也有负面影响。因此，考虑到对象检测中的特征维通常很低（类别数量+框位置），我们建议学习低维 re-ID 特征以平衡这两个任务； (2) 当训练数据较少时，学习低维 re-ID 特征可降低过拟合的风险。 MOT 中的数据集通常比 re-ID 区域中的数据集小得多。因此有利于减小特征维度。 (3) 学习低维 re-ID 特征可提高推理速度，这将在我们的实验中显示。

1.4 Overview of FairM

在这项工作中，我们提出一种称为 FairMOT 的简单方法，以共同解决三个公平问题。 它与先前的“检测优先，reID 次要”框架本质上不同，因为在 FairMOT 中检测和 reID 任务被平等对待。 我们的贡献是三倍。 首先，我们从经验上证明和讨论了以前的一次性跟踪框架所面临的挑战，这些框架已被忽视，但严重限制了它们的性能。 其次，在诸如[10]之类的 anchor-free 目标检测方法之上，我们引入了一个框架来公平地平衡检测和 reID 任务，该框架明显优于以前的方法。 最后，我们还提出了一种自我监督的学习方法，以在大规模检测数据集上训练 FairMOT，从而提高了泛化能力。 这具有重要的经验值。

图2展示了 FairMOT。 它采用非常简单的网络结构，该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[10]，[16]，[17]，[18]的启发，检测分支以无锚样式实现，该样式估计对象中心和大小（表示为位置感知测量图）。 类似地，reID 分支为每个像素估计 reID 特征，以表征以该像素为中心的对象。 请注意，这两个分支是完全同质的，这与以层叠方式执行检测和重新 ID 的先前方法本质上有所不同。 因此，FairMOT 消除了表3中反映的检测分支的不公平优势，有效地学习了高质量的 re-ID 特征，并在检测和 re-ID 之间获得了良好的折衷，以获得更好的 MOT 结果。

还值得注意的是，FairMOT 在步幅为4的高分辨率特征图上运行，而以前的基于锚的方法在步幅为32的特征图上运行。取消锚点以及使用高分辨率的特征图可以更好地对齐 具有对象中心的 reID 特征，可大大提高跟踪精度。 re-ID 特征的维数设置为仅64，这不仅减少了计算时间，而且通过在检测任务和 re-ID 任务之间取得良好的平衡来提高跟踪的鲁棒性。 我们为骨干网[19]配备了“深层聚合”运算符[20]，以融合多层的要素，以容纳分支并处理不同规模的对象。

我们通过评估服务器以 MOT Challenge 基准评估 FairMOT。 它在2DMOT15 [21]，MOT16 [22]，MOT17 [22]和 MOT20 [23]数据集的所有跟踪器中排名第一。 当我们使用我们提出的自我监督学习方法进一步对模型进行预训练时，它将在所有数据集上获得额外的收益。 尽管取得了不错的效果，但该方法非常简单，并且在单个 RTX 2080Ti GPU 上以30 FPS 的速度运行。它阐明了 MOT 中检测与 re-ID 之间的关系，并为设计单次视频跟踪网络提供了指导。

2 RELATED WORK

我们首先回顾有关 MOT 的相关工作，包括基于深度学习和基于非深度学习的 MOT。 然后，我们简要地讨论视频对象检测，因为它也与对象跟踪有关。 我们讨论方法的优缺点，并将其与我们的方法进行比较。

2.1 Non-deep Learning MOT Methods

多目标跟踪可以分为在线方法[1]，[24]，[25]，[26]，[27]和批处理方法[28]，[29]，[30]，[31]，[32]]，[33]基于它们是否依赖于将来的框架。 在线方法只能使用当前和以前的帧，而批处理方法则使用整个序列。

大多数在线方法都假定可以进行目标检测，并专注于数据关联步骤。例如，SORT [1]首先使用卡尔曼滤波器[34]来预测未来的对象位置，计算它们与未来帧中检测到的对象的重叠，最后采用匈牙利算法[35]进行跟踪。 IOU-Tracker [24]通过相邻帧的空间重叠直接关联检测，而无需使用卡尔曼滤波器，并且可以达到100K fps 的推理速度（不计算检测时间）。 SORT 和 IOU-Tracker 由于其简单性而在实践中得到了广泛使用。但是，由于缺少 re-ID 功能，它们可能无法用于充满挑战的场景，例如拥挤的场景和快速的摄像机运动。 Bae 等。 [26]应用线性判别分析来提取对象的 re-ID 特征，从而获得更鲁棒的跟踪结果。 Xiang 等。 [25]将在线 MOT 制定为马尔可夫决策过程（MDP），并利用在线单对象跟踪和强化学习来决定小轨迹的生/死和外观/消失。

由于在整个序列中有效地进行全局优化，因此与在线方法相比，该批方法的效果更好。 例如，Zhang 等。 [28]建立了一个图形模型，其节点代表了在多帧跟踪中所有帧中的检测。 使用最小成本流算法搜索全局最优，该算法利用图形的特定结构比线性规划更快地达到最优。 Berclaz 等。 [29]还将数据关联视为流优化任务，并使用 K 最短路径算法对其进行求解，从而显着加快了计算速度并减少了需要调整的参数。 米兰等。 [31]将多目标跟踪公式化为连续能量的最小化，并着重于设计能量函数。 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。

2.2 Deep Learning MOT Methods

深度学习的快速发展促使研究人员探索现代对象检测器，而不是使用基准数据集提供的基线检测结果。例如，某些性能最佳的方法，例如[2]，[4]，[5]，[6]，[7]将对象检测和 re-ID 视为两个单独的任务。他们首先应用基于 CNN 的物体检测器，例如 Faster R-CNN [8]和 YOLOv3 [11]，以定位输入图像中所有感兴趣的物体。然后，在一个单独的步骤中，它们根据框裁剪图像并将其馈送到身份嵌入网络，以提取 re-ID 特征，这些特征将用于随时间链接框。链接步骤通常遵循标准惯例，该惯例首先根据 re-ID 特征和边界框的并集交集（IoU）计算成本矩阵，然后使用卡尔曼滤波器[34]和匈牙利算法[35]完成链接任务。诸如[5]，[6]，[7]之类的少数作品也建议使用更复杂的关联策略，例如组模型和 RNN。

两步方法的主要优点是，它们可以为每个任务分别开发最合适的模型，而不会做出妥协。此外，他们可以根据检测到的边界框裁剪图像补丁，并在估计 reID 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。 结果，这些方法[4]在公共数据集上取得了最佳性能。但是，它们通常很慢，因为这两个任务需要分别完成而不共享。 因此，很难实现许多应用中所需的视频速率推断。

随着深度学习中多任务学习[13]，[36]，[37]的迅速成熟，单阶段 MOT 已开始吸引更多的研究关注。 核心思想是在单个网络中同时完成对象检测和身份嵌入(reID 功能），以减少推理时间。 例如，Track-RCNN [15]在 Mask R-CNN [9]的顶部添加了一个 re-ID 头，并为每个 proposal 回归了边界框和 reID 特征。 同样，JDE [14]建立在 YOLOv3 [11]的基础上，它可以实现接近视频速率的推断。 但是，单阶段追踪器的准确性通常低于两步追踪器。

我们的工作也属于单阶段跟踪器。 与以前的工作不同，我们深入研究了失败的原因，并从三个方面发现与检测任务相比，对 re-ID 任务的处理不公平。 最重要的是，我们提出了 FairMOT，它可以在两个任务之间达到良好的平衡。 我们表明，无需大量的工程工作即可大大提高跟踪精度。

视频对象检测（VOD）与 MOT 相关，因为它利用对象跟踪来改善具有挑战性的帧中的对象检测[38]，[39]。 例如，Tang 等。 [40]检测视频中的目标管，其目的是基于具有挑战性的帧的相邻帧来提高分类得分。 在基准数据集上，小物体的检测率大大提高。 在[40]，[41]，[42]，[43]，[44]中也探索了类似的想法。这些基于管的方法的主要局限性在于它们的速度非常慢，尤其是在存在大量的这种方法的情况下。 视频中的对象。

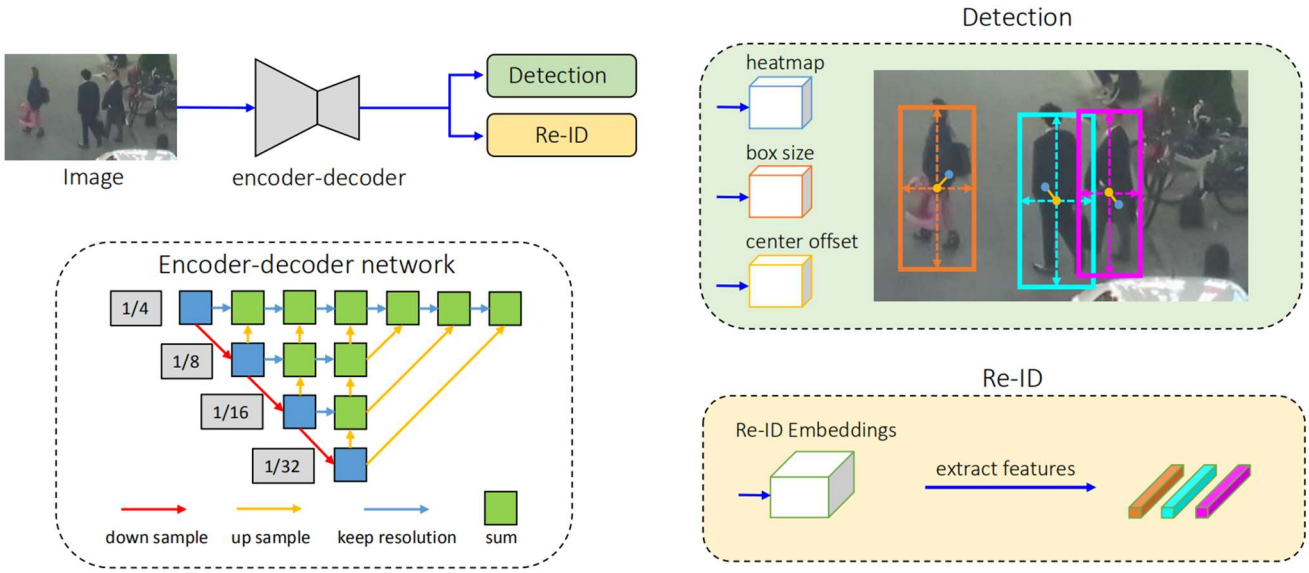


图2、我们的单阶段跟踪器 FairMOT 概述。 首先将输入图像馈送到编码器-解码器网络，以提取高分辨率特征图（步幅= 4）。 然后，我们添加两个同构分支，分别用于检测对象和提取 re-ID 特征。 预测对象中心的特征用于跟踪。

3 FAIRMOT

在本节中，我们介绍 FairMOT 的技术细节，包括骨干网，对象检测分支，re-ID 分支以及培训细节。

3.1 Backbone Network

我们采用 ResNet-34作为骨干，以便在准确性和速度之间取得良好的平衡。深层聚合（DLA）[10]的增强版应用于主干以融合多层特征，如图2所示。与原始 DLA [20]不同，它在低级和高级 DLA 之间具有更多的跳过连接。与功能金字塔网络（FPN）[45]类似的功能。此外，所有上采样模块中的卷积层均由可变形卷积代替，以便它们可以根据对象比例和姿势动态调整接收场。这些修改也有助于减轻对齐问题。生成的模型名为 DLA-34。将输入图像的大小表示为 Himage×Wimage，然后输出要素图的形状为 C×H×W，其中 H = Himage / 4和 W = Wimage / 4。除了 DLA 之外，其他具有多尺度卷积功能的深度网络，例如 Higher HRNet [46]，也可以在我们的框架中使用，以为检测和 reID 提供公平的特征。

3.2.1 Heatmap Head

该 head 负责估计对象中心的位置。 这里采用基于热图的表示法，这是界标点估计任务的事实上的标准。 特别地，热图的尺寸是1×H×W。 如果热图的某个位置随地面真实对象中心折叠，则在热图中某个位置的响应预计为1。 响应随着热图位置和对象中心之间的距离呈指数衰减。

GT 标注方法：

这样， $\hat{Y}x,y,c=1$ *就是一个检测到物体的预测值，对于* $\hat{Y}x,y,c=1$ ，表示对于类别 c ，在当前 (x,y) 坐标中检测到了这种类别的物体，而 $\hat{Y}_{x,y,c}=0$ 则表示当前当前这个坐标点不存在类别为 c 的物体。

在整个训练的流程中，CenterNet学习了CornerNet的方法。对于每个标签图(ground truth)中的某一 C 类，我们要将真实关键点(true keypoint) $p\in\mathcal{R}^2$ 计算出来用于训练，中心点的计算方式为 $p=\left(\frac{x_1+x_2}{2},\frac{y_1+y_2}{2}\right)$ ，对于下采样后的坐标，我们设为 $\tilde{p}=\left\lfloor\frac{p}{R}\right\rfloor$ ，其中 R 是上文中提到的下采样因子4。所以我们最终计算出来的中心点是对应低分辨率的中心点。

然后我们利用 $Y\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$ 来对图像进行标记，在下采样的[128,128]图像中将**ground truth point**以 $Y\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$ 的形式，用一个高斯核 $Y_{xyc}=\exp\left(-\frac{(x-\tilde{p}_x)^2+(y-\tilde{p}_y)^2}{2\sigma_p^2}\right)$ 来将关键点分布到特征图上，其中 σ_p 是一个与目标大小(也就是w和h)相关的标准差。如果某一个类的两个高斯分布发生了重叠，直接取元素间最大的就可以。

这么说可能不是很好理解，那么直接看一个官方源码中生成的一个高斯分布[9,9]：

6906821	0.02425801345428226	0.05103688810314776	0.07974465034866318	0.092535281158422	0.07974465034866318	0.05103688810314776	0.02425801345428226	0.008
5428226	0.06872199640635958	0.14458549326087305	0.225913452682986	0.26214880584576306	0.225913452682986	0.14458549326087305	0.06872199640635958	0.024
0314776	0.14458549326087305	0.30419612285238284	0.4753035374189698	0.5515397744971643	0.4753035374189698	0.30419612285238284	0.14458549326087305	0.051
4866318	0.225913452682986	0.4753035374189698	0.7426572389044386	0.8617756314171564	0.7426572389044386	0.4753035374189698	0.225913452682986	0.079
158422	0.26214880584576306	0.5515397744971643	0.8617756314171564	1.0	0.8617756314171564	0.5515397744971643	0.26214880584576306	0.09
4866318	0.225913452682986	0.4753035374189698	0.7426572389044386	0.8617756314171564	0.7426572389044386	0.4753035374189698	0.225913452682986	0.079
0314776	0.14458549326087305	0.30419612285238284	0.4753035374189698	0.5515397744971643	0.4753035374189698	0.30419612285238284	0.14458549326087305	0.051
5428226	0.06872199640635958	0.14458549326087305	0.225913452682986	0.26214880584576306	0.225913452682986	0.14458549326087305	0.06872199640635958	0.024
6906821	0.02425801345428226	0.05103688810314776	0.07974465034866318	0.092535281158422	0.07974465034866318	0.05103688810314776	0.02425801345428226	0.008

每个点 $Y\in[0,1]^{\frac{W}{R}\times\frac{H}{R}\times C}$ 的范围是0-1,而1则代表这个目标的中心点，也就是我们要预测要学习的点。

损失函数（https://zhuanlan.zhihu.com/p/66048276）:

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

where $\hat{M}$ is the estimated heatmap, and α,β are the predetermined parameters in focal loss.

3.2.2 Box Offset and Size Heads

Box offset head 旨在更精确地定位对象。 由于最终特征图的跨度为4，因此它将引入量化误差，最大可达4个像素。 该分支为每个像素估计相对于对象中心的连续偏移，以减轻下采样的影响。 Box size head 负责估计每个位置上目标框的高度和宽度。

详见原文，概括而言即损失函数为框大小的差异的 L1 loss 和框中心坐标的差异的 L1 loss 之和。

3.3 Re-ID Branch

Re-ID 分支旨在生成可以区分对象的特征。 理想情况下，不同对象之间的近似度应该小于相同对象。 为了实现此目标，我们在 backbone 生成的特征图上应用了具有128个内核的卷积层，以提取每个位置的 re-ID 特征。 将得到的特征图表示为 E∈R^128×W×H。 可以从特征图中提取以（x，y）为中心的对象的 re-ID 特征 Ex，y∈R128。

3.3.1 Re-ID Loss

我们通过分类任务学习 re-ID 特征。 训练集中具有相同标识的所有对象实例都被视为同一类。对于图像中每一个 GT box，我们获取其在热图上的目标中心坐标。我们提取 reID 特征向量并学习将其映射到一个类分布向量上（看原文）。定义一个 one-hot 表示的 GT 类标签 L^i(k)。最后 reIDloss 为：

$$L_{\text{identity}} = -\sum_{i=1}^{N}\sum_{k=1}^{K}\mathbf{L}^{i}(k)\log(\mathbf{p}(k)), \tag{3}$$

其中 K 是类别数。 在我们网络的训练过程中，仅使用位于对象中心的身份嵌入矢量进行训练，因为我们可以在测试中从对象热度图中获取对象中心。

3.4 Training FairMOT

我们通过将损失（即等式（1），等式（2）和等式（3））加在一起来共同训练检测和 reID 分支。 特别是，我们使用[50]中提出的不确定性损失来自动平衡检测和 reID 任务：

$$L_{\text{detection}} = L_{\text{heat}} + L_{\text{box}}, \tag{4}$$

$$L_{\text{total}} = \frac{1}{2}\left(\frac{1}{e^{w_1}}L_{\text{detection}} + \frac{1}{e^{w_2}}L_{\text{identity}} + w_1 + w_2\right), \tag{5}$$

其中 w1和 w2是可平衡两个任务的可学习参数。 具体来说，给定一个包含几个对象及其对应 ID 的图像，我们将生成 GT 热图、框偏移量和大小图以及这些对象的 one-hot 类表示。 将这些与估计的度量进行比较，以获取损失来训练整个网络。

除了上面介绍的标准训练策略外，我们还提出了一种弱监督学习方法，用于在图像级目标检测数据集（例如 COCO）上训练 FairMOT。 受[51]的启发，我们将数据集中的每个对象实例视为一个单独的类，并将同一对象的不同转换视为同一类中的实例。 所采用的转换包括 HSV 增强，旋转，缩放，平移和剪切。 我们在 CrowdHuman 数据集上对模型进行预训练[52]，然后在 MOT 数据集上进行微调。 通过这种自我监督的学习方法，我们进一步提高了最终表现。

3.5 Online Inference

在本节中，我们介绍如何执行在线推理，尤其是如何表现检测结果和 re-ID 特征的关联。

3.5.1 Network Inference

网络采用大小为1088×608的帧作为输入，与先前的工作 JDE [14]相同。 在预测的热图之上，我们根据热图得分执行非最大抑制（NMS），以提取峰值关键点。 我们保留热点图得分大

于阈值的关键点的位置。 然后，我们根据估计的偏移量和框大小来计算相应的边界框。 我们还将在估计的对象中心提取身份嵌入。 在下一部分中，我们将讨论如何使用 re-ID 特征在帧间将检测结果关联。

3.5.2 Online Association

我们遵循标准的在线跟踪算法来关联框。 我们首先根据第一帧中的估计框来初始化多个小轨迹。 然后，在随后的帧中，我们根据在 Re-ID 特征上计算出的余弦距离将检测到的框链接到现有的小轨迹，并通过二分匹配将它们的框重叠[35]。 我们还使用卡尔曼滤波器[34]来预测小帧在当前帧中的位置。 如果距离链接的检测距离太远，我们会将相应的成本设置为无穷大，这可以有效地防止将检测与大运动链接在一起。 我们按照每个步骤更新跟踪器的外观特征，以处理外观差异，如[53]，[54]中所述。

4 EXPERIMENTS

4.1 Datasets and Metrics

简要介绍了六个训练数据集，如下所示：ETH [55]和 CityPerson [56]数据集仅提供框注释，因此我们仅在它们上训练检测分支。 CalTech [57]，MOT17 [22]，CUHK-SYSU [58]和 PRW [12]数据集提供了框和身份注释，这使我们能够训练两个分支。 ETH 中的一些视频也出现在 MOT16的测试集中，这些视频已从训练数据集中删除以进行公平比较。 总体训练策略在第3.4节中进行了说明，与[14]相同。 对于我们方法的自我监督训练，我们使用 CrowdHuman 数据集[52]（该数据集仅包含对象边界框注释）。

我们在以下四个基准测试集上广泛评估了我们方法的各种因素：2DMOT15，MOT16，MOT17和最近发布的 MOT20。 遵循 MOT 的常规做法，我们使用平均精度（AP）来评估检测性能，并使用误报率为0.1的真正率（TPR）来通过 GT 检测来严格评估 reID 功能。 我们使用 CLEAR 指标[59]和 IDF1 [60]评估整体跟踪精度。

4.2 Implementation Details

我们使用[10]中提出的 DLA-34变体作为默认主干。 在 COCO 数据集[61]上预先训练的模型参数用于初始化我们的模型。 我们用 Adam 优化器[62]训练了30个时期的模型，起始学习率为 e ^ −4。 学习速率在20个时代下降到 e ^ -5。 批处理大小设置为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。 输入图像的大小调整为1088×608，特征图的分辨率为272×152。 在两个 RTX 2080 Ti GPU 上，训练步骤大约需要30个小时。

4.3 Ablative Studies

在本节中，我们将通过精心设计一些基准方法，对 FairMOT 中的三个关键因素进行严格的研究，包括无锚点 re-ID 特征提取，特征融合和特征尺寸。

4.3.1 Fairness Issue in Anchors

我们评估了之前的工作中经常使用的从检测到框中采样 re-ID 特征的四种策略[14] [15]。 第一种策略是在 Track R-CNN 中使用 ROI-Align [15]。 它使用 ROI-Align 从检测到的建议中采样特征。 如前所述，许多采样位置偏离了对象中心。 第二种策略是 JDE [14]中使用的 POS-Anchor。 它从正 anchor 采样特征，正 anchor 也可能偏离对象中心。 第三种策略是 FairMOT 中使用的“中心”。 它仅在对象中心采样特征。 回想一下，在我们的方法中，从离散的低分辨率地图中提取了 re-ID 特征。 为了在准确的对象位置上对特征进行采样，我们还尝试应用双线性插值（Center-BI）提取更准确的特征。

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，我们使用 ROI Align [9]根据检测到的边界框提取主干特征，然后使用 re-ID 头（全连接层）来获取 re-ID 特征。

结果显示在表1中。请注意，这五种方法都是基于 FairMOT 构建的。 唯一的区别在于他们如何从检测框中采样 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）获得了更高的 IDF1得分和真实阳性率（TPR）。 此度量标准与对象检测结果无关，并忠实反映了 re-ID 特征的质量。 另外，我们的方法的 ID 切换（IDs）的数量也大大少于两个基线。 结果证明，对象中心的采样特征比以前的工作中使用的策略更有效。 双线性插值（Center-BI）可以实现比 Center 更高的 TPR，因为它可以在更精确的位置采样特征。 两阶段方法会损害 re-ID 特征的质量。

TABLE 1
Comparison of different re-ID feature extraction (sampling) strategies on the validation set of MOT17. The rest of the models are kept the same for fair comparison. ↑ means the larger the better and ↓ means the smaller the better. The best results are shown in **bold**.

Feature Extraction	Anchor	MOTA↑	IDF1↑	IDs↓	TPR↑
FairMOT (ROI-Align)	✓	68.7	71.0	331	93.1
FairMOT (POS-Anchor)	✓	69.0	70.3	434	93.9
FairMOT (Center)		**69.1**	72.8	**299**	94.4
FairMOT (Center-BI)		68.8	**74.3**	303	**94.9**
FairMOT (Two-Stage)	✓	69.0	68.2	388	90.5

4.3.2 Fairness Issue in Features

我们旨在研究多层特征融合在解决特征不公平问题方面的有效性。 为此，我们比较了多个骨干网，例如原版 ResNet [19]，特征金字塔网络（FPN）[45]，高分辨率网络（HRNet）[63]和 DLA-34 [10]的 re ID 特征和检测精度。 请注意，为了公平比较，这些方法的其余因素（例如训练数据集）都被控制为相同。 特别是，对于所有方法，最终特征图的跨度为4。 我们为原版 ResNet 添加了三个上采样操作，以获得步幅为4的特征图。

TABLE 2
Comparison of different backbones on the validation set of MOT17 dataset. The best results are shown in **bold**.

Backbone	MOTA↑	IDF1↑	IDs↓	AP↑	TPR↑
ResNet-34	63.6	67.2	435	75.1	90.9
ResNet-50	63.7	67.7	501	75.5	91.9
ResNet-34-FPN	64.4	69.6	369	77.7	94.2
HRNet-W18	67.4	**74.3**	315	80.5	**94.6**
DLA-34	**69.1**	72.8	**299**	**81.2**	94.4

结果显示在表2中。通过比较 ResNet-34和 ResNet-50的结果，我们惊奇地发现，使用较大的网络只会稍微改善通过 MOTA 测量的总体跟踪结果。 特别是，reID 特征的质量几乎无法从较大的网络中受益。 例如，IDF1仅从67.2％提高到67.7％，TPR 从90.9％提高到91.9％。 此外，ID 开关的数量甚至从435个增加到501个。所有这些结果表明，使用较大的网络会给最终跟踪精度增加非常有限的值。

相比之下，实际上比 ResNet-50具有更少参数的 ResNet-34-FPN 则比 ResNet-50获得更大的 MOTA 分数。更重要的是，TPR 从90.9％显着提高到94.2％，这表明与仅使用较大的网络相比，多层特征融合具有明显的优势。此外，同样基于 ResNet-34构建的 DLA-34具有更高级别的特征融合功能，其 MOTA 得分更高。特别是，TPR 从90.9％显着增加到94.4％，这反过来又将 ID 切换（IDs）的数量从435减少到299。结果验证了特征融合（FPN 和 DLA 两者）有效地提高了 re-ID 特征的判别能力。另一方面，尽管 ResNet-34-FPN 获得与 DLA-34一样好的 re-ID 特征（TPR），但其检测结果（AP）明显比 DLA-34差。我们认为在 DLA-34中使用可变形卷积是主要原因，因为它可为不同大小的对象提供更灵活的感受野-这对我们的方法非常重要，因为 FairMOT 仅从对象中心提取特征而不使用任何区域特征。用 DLA-34中的正常卷积替换所有可变形卷积时，我们只能得到65.0 MOTA 和78.1 AP。如表4所示，我们可以看到 DLA-34在中型和大型对象上的性能主要优于 HRNet-W18。

TABLE 3
Demonstration of *feature conflict* between the detection and re-ID tasks on the validation set of the MOT17 dataset. "-det" means only the detection branch is trained and the re-ID branch is randomly initialized.

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**

TABLE 4
The impact of different backbones on objects of different scales. *Small*: area smaller than 7000 pixels; *Medium*: area from 7000 to 15000 pixels; *Large*: area larger than 15000 pixels. The best results are shown in **bold**.

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

为了验证检测任务和 reID 任务之间是否存在特征冲突，我们引入了一个基线 ResNet-34-det，该基线仅训练检测分支（随机对 re-ID 分支进行初始化）。从表3中可以看出，如果不训练表示两个任务之间冲突的 re-ID 分支，则 AP 测量的检测结果将有很大的提高。特别是，ResNet-34-det 甚至比 ResNet-34更高的 MOTA 分数，因为该指标比跟踪结果更易。相比之下，DLA-34在 ResNet-34上添加了多层功能融合，可实现更好的检测和跟踪结果。这意味着多层特征融合通过允许每个任务从融合特征中提取其自身任务所需的任何内容，来帮助缓解特征冲突问题。

4.3.3 Fairness Issue in Feature Dimensionality

先前的单阶段跟踪器通常按照两步法在没有消融研究的情况下学习512维 re-ID 特征。但是，我们在实验中发现，特征维实际上在平衡检测和跟踪精度方面起着重要作用。学习较低维度的 reID 特征对检测精度的危害较小，并提高了推理速度。

我们在表5中评估了 re-ID 特征维数的多种选择。我们可以看到512获得了最高的 IDF1和 TPR 分数，这表明更高维的 re-ID 特征会导致更强的辨别能力。但是，令人惊讶的是，当我们

将维数从512减小到64时，MOTA 分数持续提高。这主要是由于检测任务和重新 ID 任务之间的冲突引起的。 特别地，我们可以看到，当我们减小 re-ID 特征的尺寸时，检测结果（AP）会提高。 在我们的实验中，我们将特征尺寸设置为64，这在两个任务之间取得了很好的平衡。

TABLE 5
Evaluation of re-ID feature dimensions on the validation set of MOT17. The best results are shown in **bold**.

Backbone	dim	MOTA ↑	IDF1 ↑	IDs ↓	FPS↑	AP↑	TPR ↑
DLA-34	512	68.5	**73.7**	312	24.1	80.9	**94.6**
DLA-34	256	68.5	72.5	337	26.1	81.1	94.3
DLA-34	128	69.1	72.8	299	26.6	81.2	94.4
DLA-34	64	**69.2**	73.3	**283**	**26.8**	**81.3**	94.3

4.3.4 Data Association Methods

本节评估数据关联步骤中的三个要素，包括边界框 IoU，re-ID 特征和卡尔曼滤波器[34]。这些用于计算每对检测框之间的相似度。 这样，我们使用匈牙利算法[35]解决分配问题。 表6示出了结果。 我们可以看到，仅使用盒子 IoU 会导致很多 ID 切换。 对于拥挤的场景和快速的相机运动尤其如此。 单独使用 re-ID 特征会显着增加 IDF1并减少 ID 切换的数量。 此外，添加卡尔曼滤波器有助于获得平滑的（合理的）小轨迹，从而进一步减少 ID 切换的数量。 当一个对象被部分遮挡时，其 re-ID 特征将变得不可靠。 在这种情况下，重要的是要利用盒子的 IoU，reID 特征和卡尔曼滤波器来获得良好的跟踪性能。

TABLE 6
Evaluation of the three ingredients in the data association model. The backbone is DLA-34.

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**

4.3.5 Visualization of Re-ID Similarity

我们使用 re-ID 相似度图来展示图3中 re-ID 特征的判别能力。我们从验证集中随机选择两个帧。 第一帧包含查询实例，第二帧包含具有相同 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度来获得 re-ID 相似度图，如第4.3.1节和第4.3.2节所述。 通过比较 ResNet-34和 ResNet-34-det 的相似性图，我们可以看到训练 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多的判别性 re-ID 特征。 在所有采样策略中，建议的 Center 和 Center-BI 可以更好地将目标对象与拥挤场景中的周围对象区分开。

图3.可视化 re-ID 特征的描述能力。 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（例如，第4.3.1节中所述的 Center，Center BI，ROI-Align 和 POS-Anchor）和不同主干（例如 ResNet-34和 DLA- 34）。 查询帧和目标帧是从 MOT17-09和 MOT17-02序列中随机选择的。

4.4 Self-supervised Learning

我们首先在 CrowdHuman 数据集上对 FairMOT 进行了预训练[52]。 特别是，我们为每个边界框分配唯一的身份标签，并使用3.4节中描述的方法训练 FairMOT。 然后，我们对目标数据集 MOT17上的预训练模型进行微调。

TABLE 7

Effects of self supervised learning on the validation set of MOT17. "CH" and "MIX" stand for CrowdHuman and the composed five datasets introduced in Section 4.1, respectively. * means no identity annotations are used.

Training Data	MOTA ↑	IDF1 ↑	IDs ↓	AP↑	TPR ↑
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7示出了结果。 首先，通过对 CrowdHuman 进行 self-supervised learning 的预训练大大优于直接对 MOT17数据集进行训练。 其次， self-supervised learning 模型甚至胜过在" MIX"和

MOT17数据集上训练的完全监督模型。 结果验证了所提出的自我监督式预训练的有效性，这节省了大量注释工作，并使 FairMOT 在实际应用中更具吸引力。

4.5 Results on MOTChallenge

我们将我们的方法与最先进的（SOTA）方法进行了比较，包括一步式方法和两步方法。

4.5.1 Comparing with One-Shot SOTA MOT Methods

仅有 JDE [14]和 Track-RCNN [15]两篇发表的著作共同执行对象检测和身份特征嵌入。 我们将两者的方法进行比较。 按照先前的工作[14]，测试数据集包含2DMOT15的6个视频。 FairMOT 使用与论文中所述的两种方法相同的训练数据。 特别是，当我们与 JDE 比较时，FairMOT 和 JDE 都使用第4.1节中描述的大规模组成的数据集。 由于 Track R-CNN 需要分割标签来训练网络，因此它仅使用 MOT17数据集的4个具有分割标签的视频作为训练数据。 在这种情况下，我们还将使用这4个视频来训练我们的模型。 CLEAR 指标[59]和 IDF1 [60]用于衡量其性能。

结果显示在表8中。我们可以看到我们的方法明显优于 JDE [14]。 特别是，ID 切换的数量从218个减少到80个，这在用户体验方面有很大的提高。 结果证明了 anchor-free 方法比以前 anchor-base 的方法更具有效性。 两种方法的推理速度都接近视频速率，而我们的方法则更快。 与 Track R-CNN [15]相比，它们的检测结果略好于我们的检测结果（FN 较低）。 但是，FairMOT 可获得更高的 IDF1得分（64.0对49.4）和更少的 ID 切换（96对294）。 这主要是因为 Track R-CNN 遵循“检测优先，reID 其次”的框架，并且使用定位符，这也给 re-ID 任务带来了歧义。

TABLE 8

Comparison of the state-of-the-art one-shot trackers on the 2DMOT15 dataset. “MIX” represents the large scale training dataset and “MOT17 Seg” stands for the 4 videos with segmentation labels in the MOT17 dataset.

Training Data	Method	MOTA↑	IDF1↑	IDs↓	FP↓	FN↓	FPS↑
MIX	JDE [14]	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 [15]	69.2	49.4	294	1328	**2349**	2.0
	FairMOT(ours)	**70.2**	**64.0**	**96**	**1209**	2537	**30.9**

4.5.2 Comparing with Two-Step SOTA MOT Methods

我们将我们的方法与最新的跟踪器（包括表9中的两步方法）进行了比较。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2DMOT15，MOT16，MOT17和 MOT20数据集的测试集的结果。 请注意，所有结果都是直接从官方 MOT 挑战评估服务器获得的。

我们的方法在这四个数据集的所有在线和离线跟踪器中均排名第一。 特别是，它大大优于其他方法。 考虑到我们的方法非常简单，这是非常好的结果。 另外，我们的方法可以实现视频速率推断。 相反，大多数高性能跟踪器（例如[4]，[7]）通常比我们的慢。

4.5.3 Training Data Ablation Study

我们还使用不同数量的培训数据评估 FairMOT 的性能。 仅使用 MOT17数据集进行训练时，我们就可以达到69.8 MOTA，这已经超过了使用更多训练数据的其他方法的效果。 当我们使用与 JDE [14]相同的训练数据时，我们可以达到72.9 MOTA，这明显优于 JDE。 此外，当我们在 CrowdHuman 数据集上执行自我监督学习时，MOTA 得分提高到73.7。 结果表明，我们的方法不消耗数据，这在实际应用中是一个很大的优势。

TABLE 9
Comparison of the state-of-the-art methods under the "private detector" protocol. It is noteworthy that FPS considers both detection and association time. The one-shot trackers are labeled by "*".

Dataset	Tracker	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	FPS↑
MOT15	MDP_SubCNN [25]	47.5	55.7	30.0%	18.6%	628	<1.7
	CDA_DDAL [64]	51.3	54.1	36.3%	22.2%	544	<1.2
	EAMTT [65]	53.0	54.0	35.9%	19.6%	7538	<4.0
	AP_HWDPL [66]	53.0	52.2	29.1%	20.2%	708	6.7
	RAR15 [7]	56.5	61.3	45.1%	14.6%	**428**	<3.4
	TubeTK* [44]	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 [65]	52.5	53.3	19.9%	34.9%	910	<5.5
	SORTwHPD16 [1]	59.8	53.8	25.4%	22.7%	1423	<8.6
	DeepSORT_2 [2]	61.4	62.2	32.8%	18.2%	781	<6.4
	RAR16wVGG [7]	63.0	63.8	39.9%	22.1%	**482**	<1.4
	VMaxx [67]	62.6	49.2	32.7%	21.1%	1389	<3.9
	TubeTK* [44]	64.0	59.4	33.5%	19.4%	1117	1.0
	JDE* [14]	64.4	55.8	35.4%	20.0%	1544	18.5
	TAP [6]	64.8	**73.5**	38.5%	21.6%	571	<8.0
	CNNMTT [5]	65.2	62.2	32.4%	21.3%	946	<5.3
	POI [4]	66.1	65.1	34.0%	20.8%	805	<5.0
	CTrackerV1* [68]	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 [69]	52.4	49.5	21.4%	30.7%	8431	<3.9
	TubeTK* [44]	63.0	58.6	31.2%	19.9%	4137	3.0
	CTrackerV1* [68]	66.6	57.4	32.2%	24.2%	5529	6.8
	CenterTrack* [70]	67.3	59.9	34.9%	24.8%	**2898**	22.0
	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**

4.6 Qualitative Results

图4显示了在 MOT17的测试集上 FairMOT 的几个跟踪结果[22]。 从 MOT17-01的结果可以看出，当两个行人交叉时，我们的方法可以借助高质量的 re-ID 特征分配正确的身份。 在这些情况下，使用包围盒 IOU [1]，[24]的跟踪器通常会导致身份切换。 从 MOT17-03的结果来看，我们的方法在拥挤的场景下表现良好。 从 MOT17-08的结果可以看出，当行人被严重遮挡时，我们的方法既可以保留正确的身份，又可以保留正确的边界框。 MOT17-06和 MOT17-12的结果表明，我们的方法可以处理大规模的变化。 这主要归因于多层特征聚合的使用。 MOT17-07和 MOT17-14的结果表明，我们的方法可以准确地检测小物体。

5 CONCLUSION

从研究为什么以前的单发方法（例如[14]）无法获得与两步法类似的结果开始，我们发现在对象检测和身份嵌入中使用锚是导致结果降低的主要原因。特别地，对应于对象的不同部分的多个附近的锚可以负责估计导致网络训练的歧义的相同身份。此外，我们发现了以前的 MOT 框架中检测和 reID 任务之间的特征不公平问题和特征冲突问题。通过解决无锚单发深度网络中的这些问题，我们提出了 FairMOT。在跟踪准确度和推理速度方面，它在许多基准数据集上均优于以前的最新方法。此外，FairMOT 本质上是训练数据效率高的，因此我们建议

仅使用带有边界框的图像对多对象跟踪器进行自我监督训练，这两种方法都使我们的方法在实际应用中更具吸引力。