CVPR2020 FairMOT:《A Simple Baseline for Multi-Object Tracking》

近年来，作为多目标跟踪的核心组件的目标检测和重新识别取得了显着进展。 但是，很少有人关注在单个网络中完成两项任务以提高推理速度。 沿着这条路径进行的最初尝试最终导致性能下降，这主要是因为重新识别分支的学习不正确。 在这项工作中，我们研究了故障背后的根本原因，并因此提出了解决问题的简单基准。 它以 30 fps 的速度远远超过了公共数据集的最新水平。 我们希望这个基准可以启发并帮助评估该领域的新想法。

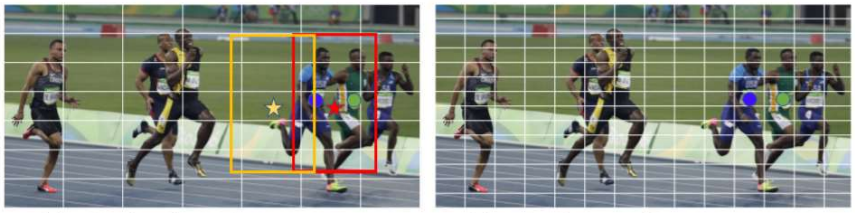
1 Introduction

多目标跟踪（MOT）已经成为计算机视觉领域的长期目标[3,37,6,40]。 目的是估计视频中多个感兴趣对象的轨迹。 成功完成任务可以受益于许多应用，例如动作识别，公共安全，体育视频分析，老人护理和人机交互。

最先进的方法[23,46,11,3,37,6,40]通常通过两个单独的模型来解决该问题：检测模型首先通过对图像中的框进行边界来定位感兴趣的对象，以及 然后，关联模型为每个边界框提取重新标识（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[27,12,44,26]和 Re-ID [43,6]上取得了显着进步，这反过来又提高了跟踪性能。 但是，这些方法无法以视频速率执行推理，因为两个网络不共享特征。

随着多任务学习的成熟[15]，联合检测物体并学习 Re-ID 特征的单发方法已开始引起更多关注[35,33]。由于两个模型共享大多数功能，因此它们有可能显着减少推理时间。然而，与两步法相比，单步法的准确性通常会显着下降。 特别是，ID 开关的数量增加了很多，如实验部分所示。 结果表明，将这两项任务结合起来并非易事，应谨慎对待。

我们没有使用大量技巧来提高跟踪精度，而是研究故障背后的原因，并提出了一个简单而有效的基准。 确定了对跟踪结果至关**重要的三个因素**。



(a) ***anchor-based*** Re-ID feature extraction　　(a) ***anchor-free*** Re-ID feature extraction

Fig. 1: (a) The yellow and red anchors are responsible for estimating the same ID (the person in blue shirt) although the image patches are very different. In addition, the anchor-based methods usually operate on a coarse grid. So there is a high chance that the features extracted at the anchor (red or yellow star) are not aligned with the object center. (b) The anchor-free approach suffers less from the ambiguities.

(1) **Anchors don't fit Re-ID**. 当前的 one-shot 跟踪器[35,33]都基于 anchor，因为它们是从对象检测器[26,12]修改而来的。 但是，anchor 有两个原因，不适合学习 Re-ID 功能。 首先，**对应于不同图像块的多个 anchor 可能负责估计同一对象的身份**。 这导致网络的严重歧义。 有关说明，请参见图 1。 此外，**通常会将特征图下采样 8 次以平衡精度和速度。 这对于检测是可以接受的，但对于 ReID 来说太粗糙了**，因为对象中心可能与在粗略 anchor 位置提取的用于预测对象身份的特征不对齐。 **我们通过将 MOT 问题视为高分辨率特征图顶部的像素级关键点（对象中心）估计和身份分类问题来解决该问题。**

(2) **Multi-Layer Feature Aggregation**. 这对于 MOT 尤其重要，因为 Re-ID 特征需要利用低级和高级特征来包含大小对象。 我们在实验中观察到，由于提高了处理尺度变化的能力，这有助于减少 one-shot 方法的身份切换。 请注意，对于两步方法而言，改进并不明显，因为在裁剪和调整大小操作之后，对象将具有相似的比例。

(3) **Dimensionality of the ReID Features**. 以前的 ReID 方法通常学习高维特征，并在其基准测试中取得了可喜的结果。但是，我们发现，**较低维度的特征实际上对 MOT 更好**，因为它的训练图像比 ReID 少（我们不能使用 ReID 数据集，因为它们仅提供裁剪后的人物图像）。 **学习低维特征有助于降低过度拟合小数据的风险，并提高跟踪的鲁棒性**。

我们提出了一个简单的基线，该基线共同考虑了以上三个因素。 请注意，我们不要求在算法上具有新颖性。 相反，**我们的贡献在于首先确定 one-shot 跟踪器背后的挑战**，然后将在计算机视觉的不同领域开发的多种技术和概念组合在一起，以解决以前的 MOT 工作中忽略的挑战。

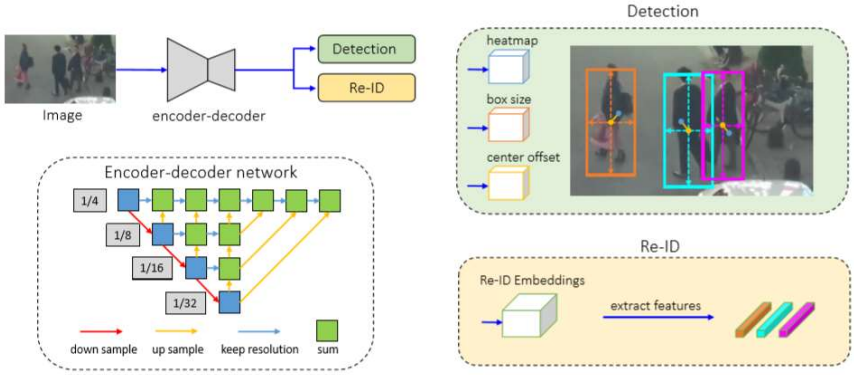


Fig. 2: Overview of our one-shot MOT tracker. The input image is first fed to an encoder-decoder network to extract high resolution feature maps (stride=4). Then we add two simple parallel heads for predicting bounding boxes and Re-ID features, respectively. The features at the predicted object centers are extracted for temporal bounding box linking.

我们的方法概述如图 2 所示。我们首先采用 anchor-free 对象检测方法来估计高分辨率特征图上的对象中心[44,17,45,9]。 消除锚点减轻了歧义性问题，并且高分辨率特征图的使用使 Re-ID 特征能够更好地与对象中心对齐。 然后，我们添加了一个并行分支，用于估算用于预测对象身份的逐像素 Re-ID 特征。 特别是，我们学习了低维 Re-ID 特征，这些特征不仅减少了计算时间，而且提高了特征匹配的鲁棒性。 我们为骨干网[13]配备了“深层聚合”运算符[41]，以融合来自多个层的要素，以便处理不同尺度的对象。

我们通过评估服务器评估我们在 MOT Challenge 基准测试中的方法。 它在 2DMOT15 [18]，MOT16 [24]，MOT17 [24]和 MOT20 [7]数据集的所有在线跟踪器中排名第一。 实际上，它在 2DMOT15，MOT17 和 MOT20 数据集上也胜过完整的跟踪器（MOT20 是最新的数据集，之前没有任何工作报告此结果）。尽管取得了很好的效果，但该方法还是非常简单，并且以 30 FPS 的速度运行。 我们希望可以将其用作该领域的强大基准。 该代码以及预训练的模型将被发布。

2 Related Work

在本节中，我们通过将 MOT 的相关工作分别分为两步法和单发方法来简要回顾它们。我们讨论了这些方法的优缺点，并将它们与我们的方法进行了比较。

2.1 Two-Step MOT Methods

最新的 MOT 方法（例如[37,40,23,46,11]）通常将对象检测和 Re-ID 视为两个单独的任务。 他们首先应用 CNN 检测器（例如[27,12,26]）通过多个框定位图像中所有感兴趣的对象。 然后，在一个单独的步骤中，他们根据框裁剪图像，并将其馈送到身份嵌入网络以提取 Re-ID 特征，然后将框链接以形成多个轨道。 这些作品通常遵循标准的盒式链接做法，

首先根据 Re-ID 特征和边界框的交集交点（IoU）计算成本矩阵，然后使用卡尔曼滤波器[36]和匈牙利算法[16] 完成链接任务。 少数作品[23,46,11]使用更复杂的关联策略，例如组模型和 RNN。

两步方法的优点在于，它们可以针对每个任务分别使用最合适的模型，而不会做出折衷。此外，他们可以根据检测到的边界框裁剪图像补丁，并在预测 Re-ID 功能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[40]在公共数据集上取得了最佳性能。 但是，它们通常非常慢，因为对象检测和 Re-ID 功能嵌入都需要大量计算，而没有它们之间的共享。 因此，很难实现许多应用中所需的视频速率推断。

2.2 One-Shot MOT Methods

随着深度学习中多任务学习的成熟[15,25,30]，单次 MOT 已开始引起更多研究关注。核心思想是在单个网络中同时完成对象检测和身份嵌入（Re-ID 功能），以通过共享大部分计算来减少推理时间。 例如，Track-RCNN(MOTS)在 Mask-RCNN [12]的顶部添加了一个 Re-ID 头，并为每个提案回归了边界框和 Re-ID 功能。 JDE [35]是在 YOLOv3 [26]框架之上引入的，该框架可实现接近视频速率的推断。

但是，单发方法的跟踪精度通常低于两步方法的跟踪精度。 我们发现这是**因为学习到的 ReID 特征不是最佳的**，这会导致大量的 ID 切换。 我们**深入研究了原因，并发现在锚点提取的身份嵌入特征与对象中心不对齐，这导致了严重的歧义。 为了解决该问题，我们建议对对象检测和身份嵌入使用 anchor-free 方法**，从而显着提高所有基准上的跟踪精度。

3 The Technical Approach

在本节中，我们分别介绍骨干网，对象检测分支和 Re-ID 特征嵌入分支的详细信息。

3.1 Backbone Network

我们采用 ResNet-34 [13]作为我们的骨干，以便在准确性和速度之间取得良好的平衡。为了适应不同规模的对象，如图 2 所示，将深层聚合（DLA）[44]的一种变体应用于骨干网。与原始 DLA [41]不同，它在底层和高层特征之间具有更多的 skip connections，类似于特征金字塔网络（FPN）[19]。 此外，上采样模块中的所有卷积层都由 deformable 卷积层代替，因此它们可以根据对象的比例和姿势动态地调整感受野。这些修改也有助于缓解对齐问题。生成的模型名为 DLA-34。 将输入图像的大小表示为$H_{image} \times W_{image}$，然后输出特征图的形状为$C \times H \times W$，其中$H = H_{image} / 4$和$W = W_{image} / 4$。

3.2 Object Detection Branch

继DLA变体之后，我们将对象检测视为高分辨率特征图上基于中心的包围盒回归任务。特别是，将三个并行回归头附加到骨干网络以分别估计 **heatmaps**，**对象中心偏移**和**边界框大小**。 通过对骨干网的输出特征图应用 3×3 卷积（具有 256 个通道）来实现每个 head，然后通过 1×1 卷积层生成最终目标。

Heatmap Head。该 head 负责估计对象中心的位置。 这里采用基于 heatmap 的表示法，这是界标点估计任务的事实上的标准。 尤其是，heatmap 的尺寸为$1 \times H \times W$。如果 heatmap 随 GT 中心响应，则在 heatmap 中某个位置的响应预计为 1。 随着 heatmap 中位置和对象中心之间的距离，响应呈指数衰减。

Center Offset Head。该 head 负责更精确地定位对象。 回想一下，特征图的 stride 为 4，这将引入不可忽略的量化误差。 注意，物体检测性能的好处可能是微不足道的。但是对于跟踪至关重要(因为应根据准确的对象中心提取 Re-ID 功能)。 我们在实验中发现，**ReID 特征与对象中心的仔细对齐对于性能至关重要**。

Box Size Head。该 head 负责估计每个 anchor 位置的目标边界框的高度和宽度。 该

head 与 Re-ID 特征没有直接关系，但是定位精度将影响对象检测性能的评估。

3.3 Identity Embedding Branch

身份嵌入分支的目标是生成可以区分不同对象的特征。 理想情况下，不同对象之间的距离应大于同一对象之间的距离。 为了实现该目标，我们在主干特征之上应用了具有 128 个内核的卷积层，以提取每个位置的身份嵌入特征。 生成的特征图为$E \in R^{128\times W\times H}$。 从特征图中提取对象在$(x,y)$处的 Re-ID 特征$E_{x,y} \in R^{128}$。

3.4 Loss Functions

Heatmap Loss. 对于图片中每个 GT box $b^i=(x_1^i,y_1^i,x_2^i,y_2^i)$，我们会计算出其目标中心点$(c_x^i,c_y^i)$，其中$c_x^i=\frac{x_1^i+x_2^i}{2}$，$c_y^i=\frac{y_1^i+y_2^i}{2}$。然后，通过划分步幅$(\tilde{c}_x^i,\tilde{c}_y^i)=(\left[\frac{c_x^i}{4}\right],\left[\frac{c_y^i}{4}\right])$来获得其在特征图上的位置。然后，将位置$(x,y)$处的热图响应计算为$M_{xy}=\sum_{i=1}^{N}exp^{-\frac{(x-\tilde{c}_x^i)^2+(y-\tilde{c}_y^i)^2}{2\sigma_c^2}}$，其中$N$表示图像中的对象数量，而$\sigma_c$表示标准偏差。 损失函数定义为具有 focal loss 的逐像素逻辑回归[20]:

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

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

Offset and Size Loss.我们定义 outputs size 和 offset heads 为 $\hat{S}\in R^{W\times H\times 2}$、$\hat{O}\in R^{W\times H\times 2}$。对于图像中每一个 GT box $b^i=(x_1^i,y_1^i,x_2^i,y_2^i)$, 我们计算其 size $s^i=(x_2^i-x_1^i,y_2^i-y_1^i)$。相似的，GT offset 可以这样计算$o^i=\left(\frac{c_x^i}{4},\frac{c_y^i}{4}\right)-(\left[\frac{c_x^i}{4}\right],\left[\frac{c_y^i}{4}\right])$。将在相应位置的估计 size 和 offset 分别表示为$\hat{s}^i$和$\hat{o}^i$。然后我们对两个 head 强加 L1 损失:

$$L_{\text{box}}=\sum_{i=1}^{N}\|\mathbf{o}^i-\hat{\mathbf{o}}^i\|_1+\|\mathbf{s}^i-\hat{\mathbf{s}}^i\|_1. \quad (2)$$

Identity Embedding Loss. 我们将对象身份嵌入视为分类任务。 特别是，训练集中具有相同 id 的所有对象实例都被视为一个类。对于图像中每一个 GT box $b^i=(x_1^i,y_1^i,x_2^i,y_2^i)$，我们从 heatmap 上获得对象中心$(\tilde{c}_x^i,\tilde{c}_y^i)$。我们在该位置提取 id 特征向量$E_{x^i,y^i}$，并学习将其映射到类分布向量$p(k)$。将 GT 类标签的 one-hot 表示形式表示为$L^i(k)$。然后我们将 softmax 损失计算为:

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

where K is the number of classes.

3.5 Online Tracking

在本节中，我们将说明模型的推论以及如何使用检测结果和身份嵌入来执行框跟踪。

Network Inference. 网络将输入大小为 1088×608 的图像作为输入，该图像与以前的工作 JDE [35]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。 我们还在估计的对象中心提取 id 嵌入。

Online Box Linking. 我们使用标准的在线跟踪算法来实现 box linking。 我们根据第一帧中的估计框来初始化多个小轨迹。 在随后的帧中，我们根据 Re-ID 特征和 IoU 测量的距

离将这些框链接到现有的 Tracklet。我们还使用卡尔曼滤波器预测轨迹在当前帧中的位置。如果离链接的检测距离太远，我们将相应的成本设置为无穷大，从而有效地防止将检测与大的运动链接在一起。 我们按照每个步骤更新跟踪器的外观特征，以处理外观变化，如[4,14]中所述。

4 Experiments

4.1 Datasets and Metrics

继[35]等之前的工作之后，我们通过组合来自六个公共数据集的训练图像进行人体检测和搜索，组成了一个大型训练数据集。 特别是，ETH [10]和 CityPerson [42]数据集仅提供边界框注释，因此我们仅对它们进行训练。 CalTech [8]，MOT17 [24]，CUHK-SYSU [39]和 PRW [43]数据集提供边界框和身份注释，我们在其上训练检测和身份嵌入分支。 由于 ETH 数据集中的一些视频也出现在 MOT16 数据集的测试集中，因此我们将它们从训练数据集中删除以进行公平比较。 在一些消融实验中，我们建议在较小的数据集上训练我们的模型，以节省计算成本，这将在后面进行详细描述。

我们在四个基准的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。 与[35]中一样，我们使用平均精度（AP）评估检测性能，并使用假阳性率 0.1 的真正率（TPR）评估 Re-ID 功能。 我们使用 CLEAR 度量[2]和 IDF1 [28]来评估跟踪精度。

4.2 Implementation Details

我们使用[44]中提出的 DLA-34 变体作为我们的默认主干。在 COCO 检测数据集[21]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 优化器训练模型 30 个时间段，起始学习率为 1e-4。 学习率分别在 20 和 27 个时元衰减到 1e-5 和 1e-6。 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。 输入图像的大小调整为 1088×608，特征图的分辨率为 272×152。 两个 RTX 2080 GPU 的培训时间约为 30 小时。

4.3 Ablative Study

Anchor-based vs. Anchor-free。先前的 one-shot 跟踪器基于 anchor，这些 anchor 会遇到前面各节中描述的未对准问题。 在本节中，我们通过在我们的方法之上构造一个基于 anchor 的基线，通过用[35]中使用的基于 anchor 的方法替换检测分支，在数字上验证该论点。 对于公平比较的两种方法，我们将其余因素保持相同。 请注意，本节中的模型是在大型训练数据集上进行训练的，因为当我们使用小型数据集进行训练时，基于 anchor 的方法会获得非常差的结果。 结果示于表 1。

Table 1: Evaluation of the anchor-based and anchor-free methods on the validation videos of the MOT15 dataset. The large training dataset is used and all models are trained for 10 epochs. ↑ means the larger the better and ↓ means the smaller the better. The best results are in **bold**.

Backbone	stride	Head	MOTA↑	IDF1↑	IDs↓	Prec↑	Rec↑	AP↑	TPR↑
DLA-34	2	anchor-free	71.9	70.3	93	91.7	79.8	87.2	56.5
DLA-34	4	anchor-based	64.9	62.1	137	87.9	76.4	81.9	73.6
DLA-34	4	anchor-free	**75.9**	**72.3**	**93**	94.2	**81.6**	**88.2**	80.8
DLA-34	8	anchor-based	65.5	66.3	139	91.8	73.1	83.4	75.3
DLA-34	8	anchor-free	67.3	64.9	109	**94.8**	72.2	85.1	**85.5**

我们可以看到，对于不同的步幅，基于 anchor 的方法获得的 MOTA 得分始终低于 anchor-free 方法。 例如，当跨度为 8 时，anchor-free 方法的 TPR 分数要比基于锚的基线

好得多（85.5%vs. 75.3%），这意味着 anchor-free 方法的 Re-ID 特征具有明显的优势。 主要原因是**锚点和对象中心之间的未对齐会严重影响网络的学习**。

值得注意的是，为 anchor-based 的方法增加特征图分辨率甚至会降低 MOTA 分数。这是因为当使用高分辨率特征图时，将会有更多未对齐的 positive anchors，这会使网络训练更加困难。 我们没有显示 stride 为 2 的结果，因为 anchor 的数量显着增加超过了我们 GPU 的存储容量。

相比之下，与 anchor-based 的方法相比，我们的 anchor-free 方法解决的对准误差问题更少，并且 MOTA 得分明显更高。 尤其是，对于四个跨度，ID 开关的数量从 137 个显着减少到 93 个。 更重要的是，当将步幅从 8 减小到 4 时，我们的方法会受益匪浅。将步幅进一步减小到 2 会降低结果的质量，因为引入了较低级别的特征会降低表示形式对外观变化的鲁棒性。

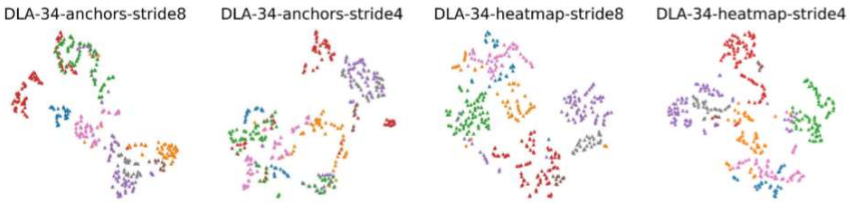


Fig. 3: We plot the Re-ID features of all persons in the testing set learned by four models using t-SNE [22]. The features of the same person are represented by the same color.

我们还可视化了图 3 中不同模型学习到的 Re-ID 特征。我们可以看到，对于 anchor-based 的方法，尤其是在跨度为 4 时，不同身份的特征是混合在一起的。相比之下，对于我们的 anchor-free 方法，它们是完全分开的。

Multi-Layer Feature Aggregation. 本部分评估骨干网中多层特征聚合的影响。 特别是，我们评估了许多主干，例如普通 ResNet [13]，特征金字塔网络（FPN）[19]，高分辨率网络（HRNet）[31]和 DLA-34 [44]。 为了公平比较，将方法的其余因素控制为相同。 对于本实验中的所有方法，最终特征图的步幅均为 4。 特别是，我们为香草 ResNet 添加了三个上采样操作，以获得 stride-4 特征图。 按照先前的工作[38]，我们将 2DMOT15 数据集的训练子集分为 5 个训练视频和 6 个验证视频。 为了减少计算成本，此处不使用大规模训练数据集。

Table 2: Evaluation of different backbones on the 2DMOT15 dataset. The best results are shown in **bold**.

Backbone	MOTA↑	IDF1↑	IDs↓	Prec↑	Rec↑	FPS↑	AP↑	TPR↑
ResNet-34	30.7	41.3	372	74.6	48.8	**47.3**	61.9	35.0
ResNet-34-FPN	34.0	45.2	320	77.1	50.3	36.1	67.3	40.9
ResNet-50	34.6	42.8	432	81.9	46.7	32.0	62.8	35.4
HRNet-W32	37.9	52.8	189	83.9	47.8	22.2	65.7	63.8
DLA-34	**40.4**	**53.9**	**136**	**83.9**	**50.7**	31.0	**68.3**	**67.3**

结果显示在表 2 中。我们可以看到，在 ResNet-34 之上构建的 DLA-34 与普通的 ResNet-34 相比，可显着提高 MOTA 得分。 特别是，TPR 从 35.0%显着增加到 67.3%，这反过来又将 ID 开关（ID）的数量从 372 个减少到 136 个。实验结果表明，由于具有多层融合结构，Re-ID 特征的描述能力有所提高。

通过比较 ResNet-34 和 ResNet-50 的结果，我们可以看到使用较大的网络也可以提高 MOTA 总体得分。 但是，如果我们查看详细的指标，我们发现改进主要来自 AP 测量的增强检测结果。 **但是，Re-ID 功能几乎不能从较大的网络中受益**。 例如，TPR 仅从 35.0%提

高到 35.4%。 相比之下，DLA-34 的数字为 67.3%。 结果表明，**在改善身份嵌入方面，多层融合相对于使用更深层的网络具有明显的优势**。

我们还比较了其他多层融合方法，例如 HRNet [31]和 FPN [19]。 两种方法都比 ResNet-34 获得更好的 MOTA 分数。 改进不仅来自增强的检测结果，还归因于 Re-ID 特征的改进的区分能力。 例如，HRNet 的 TPR 从 35.0%增加到 63.8%。

Table 3: The impact of backbones for objects of different scales. *Small*: area smaller than 6000; *Medium*: area from 6000 to 25000; *Large*: area larger than 25000.

Backbone	AP^S	AP^M	AP^L	TPR^S	TPR^M	TPR^L	IDs^S	IDs^M	IDs^L
ResNet-34	32.6	60.2	72.6	28.8	32.2	22.5	48	131	149
ResNet-34-FPN	**39.3**	**63.9**	75.1	38.3	41.5	34.2	49	121	104
ResNet-50	33.0	59.8	71.1	29.7	43.7	30.3	37	162	172
HRNet-W32	35.6	60.2	**78.7**	60.1	67.9	**59.7**	**23**	49	97
DLA-34	36.2	62.9	78.3	**61.9**	**71.2**	55.2	25	**47**	**41**

DLA-34 模型比 FPN 和 HRNet 具有更多优势。 我们发现 DLA-34 中的可变形卷积是产生间隙的主要原因，因为它可以减轻由小物体下采样引起的未对准问题。 如表 3 所示，我们可以看到 DLA-34 在中小型物体上的性能主要优于 HRNet。

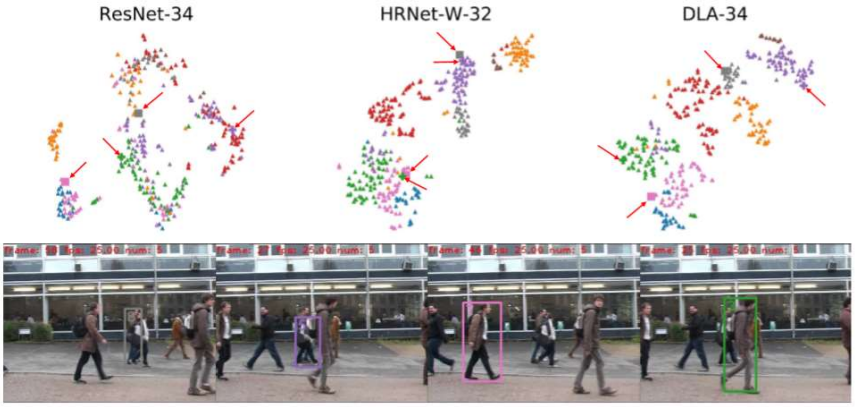


Fig. 4: We plot the Re-ID features of all persons learned by three backbone networks, respectively, using t-SNE [22]. The features of the same person are represented by the same color. The features learned by DLA-34 has clear advantages in terms of discriminative ability. We highlight the features of four different people by red arrows. The appearance of the four people correspond to the boxes of different colors as shown in the bottom images.

通过 t-SNE [22]，我们在图 4 的测试集中可视化了所有人的 Re-ID 特征。 我们可以看到，由于不同身份的特征大多混合在一起，因此，普通 ResNet-34 所学习的特征没有区别。这将在链接阶段导致大量的 ID 切换。 HRNet 学会的 Re-ID 功能变得更好，除了粉红色和绿色的点非常混乱。 此外，DLA-34 的 Re-ID 功能比两种基准方法更具区分性。

The Re-ID Feature Dimension. 先前的作品通常不进行消融研究就学习 512 维特征。但是，我们在实验中发现，特征维实际上起着重要的作用。 通常，为了避免过度拟合，训练高维 Re-ID 特征需要大量训练图像，这对于单次跟踪问题是不可用的。 以前的两步方法受此问题的影响较小，因为它们可以利用提供可裁剪人图像的大量 Re-ID 数据集。 包括我们在内的单发方法无法使用它们，因为它需要原始的未裁剪图像。 **一种解决方案是通过减小 Re-ID 特征的维数来减少其对数据的依赖性**。

我们在表 4 中评估了多个维度选择。我们可以看到，当维度从 512 减少到 128 时，TPR 持续提高，这表明使用低维度特征的优势。 将尺寸进一步减小到 64 会开始降低 TPR，因为 Re-ID 特征的代表能力很强。 尽管 MOTA 分数的变化很小，但 ID 切换的数量实际上从 210

个显着减少到 136 个。这实际上在改善用户体验方面起着至关重要的作用。 通过减小 Re-ID 特征的维数，推理速度也略有提高。 值得注意的是，只有在我们访问少量训练数据时，使用低维 Re-ID 功能的论据才成立。 当训练数据的数量增加时，由特征维数引起的间隙将变小。

Table 4: Evaluation of the Re-ID feature dimensions on the 2DMOT15 dataset.

Backbone	dim	MOTA ↑	IDF1 ↑	IDs ↓	FPS↑	TPR ↑
DLA-34	512	40.4	52.1	210	28.7	61.5
DLA-34	256	40.2	**55.1**	157	30.6	63.5
DLA-34	128	**40.4**	53.9	**136**	31.0	**67.3**
DLA-34	64	40.4	51.1	165	**31.7**	61.0

4.4 The State-of-the-arts

我们将我们的方法与最先进的方法进行了比较，包括一站式方法和两步方法。

One-Shot MOT Methods. 只有两个发表的作品，即 JDE [35]和 TrackRCNN（MOTS），可以共同执行对象检测和身份特征嵌入。 特别是，TrackRCNN 需要附加的细分注释，并使用针对细分任务的不同度量来报告结果。 因此，在这项工作中，我们**仅将其与 JDE 进行比较**。

Table 5: Comparison to the state-of-the-art one-shot trackers on two datasets. The results on MOT16-*test* are obtained from the MOT challenge server.

Dataset	Method	MOTA↑	IDF1↑	IDs↓	FPS↑
MOT15 *train*	JDE [35]	67.5	66.7	218	22.5
	FairMOT(ours)	**77.1**	**76.0**	**80**	**30.9**
MOT16 *test*	JDE [35]	64.4	55.8	1544	18.5
	FairMOT(ours)	**68.7**	**70.4**	**953**	**25.9**

为了公平比较，我们使用与[35]中相同的数据进行培训和测试。 具体来说，我们使用 2DMOT15-train 和 MOT16-test 进行验证。 CLEAR 度量[2]和 IDF1 [28]用于测量性能。 结果显示在表 5 中。我们可以看到，在两个数据集上，我们的方法均明显优于 JDE [35]。 特别是，ID 开关的数量从 218 减少到 80，这在改善用户体验方面有很大的提高。 结果证明了无锚方法比以前的基于锚方法的有效性。 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。

Two-Step MOT Methods. 我们将我们的方法与最先进的在线跟踪器进行了比较，包括表 6 中 MOT Challenge 数据集上的两步方法。由于我们不使用公共检测结果，因此采用了“专用检测器”协议。 我们分别报告 2DMOT15，MOT16，MOT17 和 MOT20 数据集的测试集的结果。 在进行测试之前，我们在每个数据集上对模型进行 10 个周期的拟合。 所有结果都是在 MOT 挑战评估服务器上获得的。

Table 6: Comparison to the state-of-the-arts under the "private detector" protocol. It is noteworthy that the computation time (Hz) only counts for the association step for the two-step trackers. But for the one-shot trackers, it counts for the whole system. The one-shot trackers are labeled by "*".

Dataset	Tracker	MOTA↑	IDF1↑	MT↑	ML↓	IDs↓	Hz↑
MOT15	MDP_SubCNN[38]	47.5	55.7	30.0%	18.6%	628	2.1
	CDA_DDAL[1]	51.3	54.1	36.3%	22.2%	544	1.3
	EAMTT[29]	53.0	54.0	35.9%	19.6%	7538	11.5
	AP_HWDPL[5]	53.0	52.2	29.1%	20.2%	708	6.7
	RAR15[11]	56.5	61.3	45.1%	14.6%	**428**	5.1
	Ours*	**59.0**	**62.2**	**45.6%**	**11.5%**	582	**30.5**
MOT16	EAMTT[29]	52.5	53.3	19.9%	34.9%	910	12.2
	SORTwHPD16[3]	59.8	53.8	25.4%	22.7%	1423	**59.5**
	DeepSORT_2[37]	61.4	62.2	32.8%	**18.2%**	781	17.4
	RAR16wVGG[11]	63.0	63.8	**39.9%**	22.1%	**482**	1.6
	VMaxx[34]	62.6	49.2	32.7%	21.1%	1389	6.5
	JDE*[35]	64.4	55.8	35.4%	20.0%	1544	18.5
	TAP[46]	64.8	**73.5**	38.5%	21.6%	571	39.4
	CNNMTT[23]	65.2	62.2	32.4%	21.3%	946	11.2
	POI[40]	66.1	65.1	34.0%	20.8%	805	9.9
	Ours*	**68.7**	70.4	39.5%	19.0%	953	25.9
MOT17	SST[32]	52.4	49.5	21.4%	30.7%	8431	6.3
	Ours*	**67.5**	**69.8**	**37.7%**	**20.8%**	**2868**	**25.9**
MOT20	Ours*	**58.7**	**63.7**	**66.3%**	**8.5%**	**6013**	**13.2**

我们的方法在四个数据集的所有在线跟踪器中排名第一。 实际上，在 2DMOT15 和 MOT17 数据集上，所有在线和在线跟踪器中的 MOTA 得分也最高。 考虑到我们的方法非常简单，这是一个非常好的结果。 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[11,40]）通常比我们的慢。

5 Conclusion

我们为单次多对象跟踪提供了一个简单的基准。 我们从研究为什么以前的方法（如[35]）无法获得与两步法类似的结果的原因开始。 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点可能负责估计相同的身份，这导致网络训练的歧义。 我们提供了一种简单的无锚方法，该方法在 30 fps 的几个基准数据集上表现优于先前的最新技术。 我们希望它能激发和评估这一领域的新想法。