摘要

作为多目标追踪的重要组件，目标检测和重识别近些年取得了巨大的进步。然而，很少有人将注意力放在在同一个网络中同时完成这两项工作。我们的研究表明之前的尝试以降低准确度告终主要是因为再识别任务没有被公正地学习，这导致了许多身份转换。不公平性体现在两方面：1）他们将再识别任务视为一个次要任务，它的准确性严重依赖于主要的检测任务。所以训练严重偏向于检测任务，忽视了再识别任务。2）他们使用直接从检测中借来的 ROI-Align 来提取再识别特征。然而，由于许多样本点可能属于令人不快的实例或者背景，这在表征物体中引入了许多歧义。为了解决这个问题，我们提出了一种包含两个同质分支来预测像素级的检测置信度和再识别特征的方法——FairMOT。任务间公平性的实现使得 FairMOT 达到了高水平的检测和追踪准确性并在多个公开数据及上大幅度超过以往 sota。

1 introduction

致力于在视频中估计感兴趣物体的轨迹的多物体追踪是计算机视觉中存在已久的目标。这个问题的成功解决可以为包括视频分析、行为识别、智能老年人关怀和人机交互等多个应用提供帮助。

例如 []，现有方法经常通过两个独立的模型来解决这个问题——检测模型首先通过边界框在每一帧中定位感兴趣的物体，之后关联模型提取每一个边界框的再识别特征并根据定义在特征上指标将它连接到一个已有的轨迹上。近年来物体检测 [] 都有了巨大的进展，因此也提高了追踪性能。然而，由于这两种方法不共享特征，需要为视频中的每一个边界框应用再识别模型，所以这些方法无法做到实时推理，尤其是在存在大量物体的情况下。

随着多任务学习的成熟，在一个网络中估计物体和学习再识别特征的 one-shot 追踪器吸引了越来越多的注意力。例如，Voigtlaender 等人提出，通过在 Mask R-CNN 上加入再识别分支通过 ROI-Align 来得到提案的再识别特征。它通过为再识别复用主干网络的特征减少了推理时间。然而，追踪的准确度与两阶段方法相比显著减低。特别的，身份交换的次数显著增加。这个结果表明将两个任务结合起来并不是一个简单的问题，这需要被精心处理。在这篇文章的目标是深度理解失败背后的原因，并展示一个简单然而有效的方法。特别的，三个因素被提了出来：

1.1 锚造成的不公平

由于例如 Track R-CNN 和 JDE 等现有的 one-shot 追踪器是直接从基于锚的检测器例如 YOLO 和 Mask R-CNN 修改得到的，所以大部分是基于锚的。然而，我们发现基于锚的框架并不适合学习再识别特征。虽然它有好的检测结果，但是导致了大量的身份交换。

被忽视的再识别任务：Track R-CNN 通过一种级联的方式操作，首先估计物体的提案之后在提案中提取再识别特征。值得注意的是再识别特征的质量严重取决于提案的质量。自然而然的，在训练阶段，模型严重偏向于估计准确的物体提案而不是高质量的再识别特征。简而言之，“检测第一，再识别居次”的事实上的标准框架使得再识别网络没有被公正习得。

一个锚对应多个身份：基于锚的方法通常使用 ROI-Pool 或者 ROI-Align 来从每个提案提取特征。ROI-Align 中的大部分采样位置可能属于其他的令人懊恼的实例或者背景。自然而然的，提取到的特征就准确并分明地表示目标物体而言并不是最优的。然而，我们在研究中发现只提取估计物体的中心点特征显著更优。

多个锚对应一个身份：在 [] 和 [] 中，属于不同图片部分的多个相邻的锚可能会被强制估计相同的物体，只要它们的 IOU 足够大。这为训练引入了严重的歧义。在另一方面，当图片经历了小的扰动，例如由于数据增强，同一个锚估计不同的身份是可能的。另外，为了平衡准确度和速度，物体检测中的特征图经常会被下采样 8/16/32 倍。这对于物体检测是可以接受的，但是由于在粗糙锚下提取的特征可能和物体中心不对齐，所有对于学习再识别特征，它太粗糙了。

1.2 特征造成的不公平

对于 one-shot 追踪器，大部分特征在物体检测和再识别任务间共享。然而众所周知的是，为了达到最好的结果，它们实际上需要来自不同层的特征。特别的，物体检测要求深和抽象的特征来估计物体类别和位置但是为了区分同一类别的不同实例，再识别更加专注于低层级的表观特征。我们经验性地发现，通过允许两个任务的网络枝干从多层聚合的特征中提取它们需要的特征，多层特征聚合对于解决矛盾是有效的。如果没有多层聚合特征，模型会偏向于首要的检测分支而产生低质量的再识别特征。此外，融合了来自拥有不同接收域的不同层的特征的多层特征，也提高了解决现实中十分常见的尺度变化的能力。

1.3 特征维度造成的不公平

之前的再识别工作通常学习非常高维的特征，并在它们的场景下取得了不错的结果。然而，我们发现处于三个原因，在 one-shot MOT 中学习更低维度的特征实际上更好。1）虽然学习高维再识别特征可能会轻微地提高它们辨别物体的能力，但是由于两个任务间的竞争，它显著地损害物体检测的准确率，进而对最终的追踪准确率产生负面影响。所以考虑到物体检测中的特征（类别数量 + 检测框位置）通常非常低，为了平衡这两个任务，我们提出学习低维的再识别特征。2）当训练数据较少时，学习低维再识别特征减少了过拟合的风险。MOT 数据集通常比再识别领域的数据集小得多。所以最好减少特征维度。3）正如我们将在实验中展示的，学习低维再识别特征提高了推理速度。

1.4 FairMOT 的全貌

在这篇工作中，我们展示了一个名为 FairMOT 的简单方式来同时解决三个公平性问题。它与之前的“检测第一，再识别第二”框架显著不同，因为在 FairMOT 中，检测和再识别被平等对待。我们的贡献有三个方面。首先，我们经验性地阐述并讨论了之前 one-shot 检测框架面对的挑战，它们被忽视但是严重限制框架性能。其次，在例如 [Objects as points] 不需要锚的检测器之上，我们引入了框架来公平平衡检测和再识别任务，在没有花里胡哨的情况下显著提高了之前的方法。最后，我们提出了一种自监督学习的方法来在大规模检测数据集上训练 FairMOT，这提升了它的泛化性能。这有重要的经验价值。

图二展示了 FairMOT 的全貌。它采用了一个非常简单的网络结构，网络结构由两个分别负责检测和提取再识别特征的同质分支组成。受 [] 启发，检测分支在 anchor free 的风格下实现，估计物体的中心和大小，它们被表示为位置感知测量图。注意这两个分支是完全同质的，这将它与以前的通过级联方式进行检测和再识别方式区分开来。因此正如表 3 显示的，FairMOT 消除了检测分支的不公平优势，有效地习得了高质量的再识别特征，并为取得更好的 MOT 结果达到了检测和再识别间的一个平衡。

另外一个值得注意点的是 FairMOT 在一个步长为 4 的高质量特征图上进行操作，而以前的基于锚的方法在步长为 32 的特征图上操作。锚的消除和高分辨率特征图的使用使得再识别特征与物体中心更加对齐，显著提高了追踪准确率。再识别的特征维度被设置为 64，这不仅减少了计算时间同时

也通过在检测和再识别任务中取得一个好的平衡来提高追踪的鲁棒性。为了适应两个分支并处理不同尺度的物体，我们为主干网络添加了深层聚合操作来将来自多层的特征聚合。
我们通过评测服务器在 MOT 基准上评估了 FairMOT。它在 2DMOT15，MOT16，MOT17 和 MOT20 数据集上排名第一。当我们更进一步，通过我们提出的自监督方法来预训练我们的模型，它在所有数据集上取得了额外的提高。不仅结果很好，这个方法非常简单，在单个 RTX2080Ti 上运行速度达到了 30FPS。它使得 MOT 中的检测和再识别的关系变得清晰，并且为设计 one-shot 视频追踪网络提供了指导。

2 FAIRMOT

在这个章节中，我们展示了包括主干网络、物体检测分支、再识别分支和训练细节在内的 *FairMOT* 的技术细节。

2.1 主干网络

为了取得准确率和速度间的平衡，我们采用 ResNet-34 作为主干网络。正如图 2 所示，为了融合多层特征，我们在主干网络上应用了增强版本的深层聚合 (DLA, Deep Layer Aggregation)。与原本的 DLA 不同的是，它在底层特征和高层特征间涌更多的跳跃连接，这与特征金字塔网络是类似的。另外，所有上采样模块中的卷基层都被替换为了可变形卷积，这样，它们就可以根据物体的形状和姿势来动态调整接受域。这样的修改也有助于缓解对齐问题。最终的模型名为 DLA-34。将输入图像的大小表示为 $H_{image} \times W_{image}$，那么输出特征图的形状就是 $C \times H \times W$，其中 $H = H_{image}/4, W = W_{image}/4$。包括 DLA，其他提供多尺度卷积特征的深度网络，例如 Higher HRNet，也可以被用来为检测和重识别提供公平的特征。

2.2 检测分支

我们的检测分支是建立在 CenterNet 之上的，但是也可以使用其他不基于锚的方法，例如 []。为了使我们的方法是独立的，我们对其进行简要介绍。特别的，我们在 DLA-34 上加了三个平行的头部来分别估计热力图、物体中心偏移和边界框大小。每个头都是通过对 DLA-34 的输出特征应用有 256 个通道的 3×3 卷积，跟随使用 1×1 卷积来生成最终的目标。

2.2.1 热力图头部

这个头是负责估计物体中心位置的。我们采用基于热力图的表示，这实际上是地标点估计任务的现有标准。特别的，热力图的维度是 $1 \times H \times W$。对于一个位置，如果它与真实物体中心重合，那么它的响应值应该为 1。响应值随着热力图位置和物体中心的距离指数衰减。

对于图片中的每一个边界框 $\boldsymbol{b}^i = (x_1^i, y_1^i, x_2^i, x_2^i)$，我们计算物体中心 (c_x^i, c_y^i)，其中 $c_x^i = \frac{x_1^i + x_2^i}{2}, c_y^i = \frac{y_1^i + y_2^i}{2}$。这是它在特征图上的位置就可以通过除以步长得到 $(\tilde{c}_x^i, \tilde{c}_y^i) = (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$。这样，热力图在位置 (x, y) 的响应就可以通过 $M_{xy} = \sum_{i=1}^{M} \exp(-\frac{(x-\tilde{c}_x^i)^2 + (y-\tilde{c}_y^i)^2}{2\sigma_c^2})$ 计算得到，其中 N 表示图片中的物体数量，σ_c 表示标准差。损失函数则定义为焦点误差下像素级别的逻辑回归：

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

这里 $\hat{M}$ 是估计得到的热力图，α 和 β 则是预先决定的焦点损失的参数。

2.2.2 框偏移和大小头部

框的偏移头部是为了更加准确的定位物体。由于最终特征图的步长是 4，所以它引入的量化误差最多为 4 个像素点。为了缓解下采样的影响，这个分支为每一个像素点估计一个相对于物体中心的连续偏移。边界框大小头部负责在每一个位置估计目标框的高和宽。

记 $size$ 和 $offset$ 头部的输出分别为 $\hat{S} \in R^{W \times H \times 2}$ 和 $\hat{O} \in R^{W \times H \times 2}$。对于图片中的每一个边界框 $\boldsymbol{b}^i = (x_1^i, y_1^i, x_2^i, x_2^i)$，我们计算它的大小为 $\boldsymbol{s}^i = (x_2^i - x_1^i, y_2^i - y_1^i)$。类似的，真实偏移值为 $\boldsymbol{o}^i = (\frac{c_x^i}{4}, \frac{c_y^i}{4}) - (\lfloor \frac{c_x^i}{4} \rfloor, \lfloor \frac{c_y^i}{4} \rfloor)$。记估计得到的对应位置大小和偏移 $\hat{\boldsymbol{s}}^i$ 和 $\hat{\boldsymbol{o}}^i$，这是我们为两个头部施加 l_1 损失：

$$L_{box} = \sum_{i=1}^{N} \left\| \boldsymbol{o}^i - \hat{\boldsymbol{o}}^i \right\|_1 + \left\| \boldsymbol{s}^i - \hat{\boldsymbol{s}}^i \right\|_1 \tag{2}$$

2.3 再识别分支

再识别分支的目标是生成可以区分物体的特征。理想情况下，不同物体间的亲和度应该比相同物体间的亲和度小。为了实现这一目标，我们在主干特征上应用有 128 个核的卷基层来为每一个位置提取再识别特征。记得到的特征图为 $\boldsymbol{E} \in R^{128 \times W \times H}$。可以在特征图中提取中心在 (x, y) 的物体的再识别特征 $\boldsymbol{E}_{x,y} \in R^{128}$。

2.3.1 再识别损失

我们通过一个分类任务来学习再识别特征。训练集中的所有具有相同身份的物体被视为同一类。对于图片中的每一个边界框 $\boldsymbol{b}^i = (x_1^i, y_1^i, x_2^i, x_2^i)$，我们在热力图 $(\tilde{c}_x^i, \tilde{c}_y^i)$ 得到物体中心。我们提取再识别特征向量 $\boldsymbol{E}_{\tilde{c}_x^i, \tilde{c}_y^i}$ 并学习将它映射为一个类别分布向量 $\boldsymbol{P} = \boldsymbol{p}(k), k \in [1, K]$。记真实类别标记的单热点表示为 $\boldsymbol{L}^i(k)$。那么我们可以计算再识别损失为：

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

其中 K 是类别数目。在网络的训练过程中，由于我们可以在测试中从置信度热力图中得到物体中心，所以只有位于物体中心的身份嵌入向量被用来训练。

2.4 训练 FairMOT

我们通过将损失相加来同时训练检测和重识别分支。特别的，我们使用不确定损失来自动平衡检测和再识别任务：

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

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

其中 w_1 和 w_2 是平衡两个任务的可学习参数。特别的，对于给定的包含一些物体和它们的对应身份的图片，我们生成真实热力图、边界框偏移图、边界框大小图和物体的单热点表示。将这些与估计值想比较的得到损失来训练整个网络。

除了上述标准的训练策略，我们还提出了一个在例如 COCO 的图片级别的物体检测数据集上训练 FairMOT 的弱监督方法。受 [] 启发，我们将数据集中的每个物体实例视为一个单独的类别，并将对同一物体的不同变化视为相同类别的实例。我们采用的变化包括 HSV 增强、旋转、缩放、变换和剪切。我们在 CrowdHuman 数据集上预训练后在 MOT 数据集上微调。通过这种自监督的学习方法，我们进一步提高了模型的性能。

3 在线推理

在这一节中，我们展示了我们如何进行在线推理，特别的，我们如何通过检测结果和再识别特征进行关联。

3.1 网络推理

与 JDE 相同，网络将大小为 1088×608 的帧作为输入。在预测的热力图之上，我们基于热力图的得分进行非最大抑制来提取尖峰关键点。我们保留热力图得分比阈值更大的位置的关键点。之后，我们基于估计得到的偏移和大小得到对应的边界框。我们同时提取在估计的物体中心位置的身份嵌入向量。在下一节中，我们讨论如何使用再识别特征将不同时刻的检测框关联起来。

3.2 在线关联

我们遵循标准的在线跟踪算法来关联边界框。我们首先根据在第一帧得到的检测结果来初始化一些追踪轨迹。在接下来的帧中，我们根据检测框和已有轨迹的再识别特征的余弦距离和它们的边界框重叠，通过二分图匹配来将检测到的边界框连接到已有的追踪轨迹上。我们也是用 Kalman 滤波来预测追踪轨迹在当前帧的位置。如果预测的位置和连接的框相距太远，我们通过将对应的损失设置为无穷来高效地防止连接移动很远的检测。和 [] 一样，我们在每次时间变化后更新追踪器的表观特征来处理表观变化。