IterDet: Iterative Scheme for Object Detection in Crowded Environments

拥挤环境中目标检测的迭代方案

摘要

以深度学习为基础的检测器对同样的目标多次检测通常产生冗余数量的目标边界框。这些矩形框之后使用 NMS 筛选。这种贪婪方案简单，对孤立目标提供了有效的精度，但对遮挡的场景经常失败，因为既需要为不同的目标保存矩形框也要抑制重复的检测结果。本文中提出了一种新的迭代方法，每次迭代都会检测新的对象子集。之前迭代检测到的矩形框会传递给下一次迭代，确保同一目标不被检测两次。这个迭代方法只需对迭代方案简单的修改就能用在单步和多步目标检测器中。

引言

现在所有的目标检测器的一个已知的问题是，处理拥挤环境下多个同类的重叠物体困难。造成这种影响的原因如下。

首先，同类的重叠物体很难区分是两个矩形框属于一个目标还是对应不同的重叠目标。

第二，高度拥挤的实例视觉线索变弱，很难提供有效的信息进行准确的目标检测。

所有 NMS 的变形都是召回和精度的平衡，并没有完全解决这个问题，因为既需要移除同一物体的冗余框，也要保留难检测的被遮挡物体的检测框。

本文提出了一种新的迭代框架（IterDet）用于目标检测。本文的框架迭代的提供检测结果，而不是同时检测图片中所有的目标。每次迭代，都会检测新的对象子集。前次迭代检测的矩形框传递给网络中下一次的迭代，避免重复检测。

图 1 展示了使用 iterDet 的 faster rcnn 在来自 CrowdHuman 的一张测试图片上的结果。置信度高于 0.1 的正确的正样本框被展示，错误的样本框为了展示清晰而被省略。在第二次迭代，从 137 个增加了 9 个额外的目标（黄色展示），超过基准 faster rcnn 了 5 个正

确的正样本并且 AP 提高 2.7%。在图片右上角展示了两个严重重叠的目标，基准检测不能发现，但是 IterDet 在两次迭代后就检测到了所有的目标。

图 1：原始的 faster rcnn（左图）和基于 faster rcnn 的 IterDet（右图）在同一张图片上比较，图片来自 CrowdHuman 测试集使用可见标注。在第一次和第二次迭代中检测到的矩形框分别用绿色和黄色标识。Recall 和 AP 的值在图片下方。

提出的方法

本文提出的迭代方法如图 2 展示。

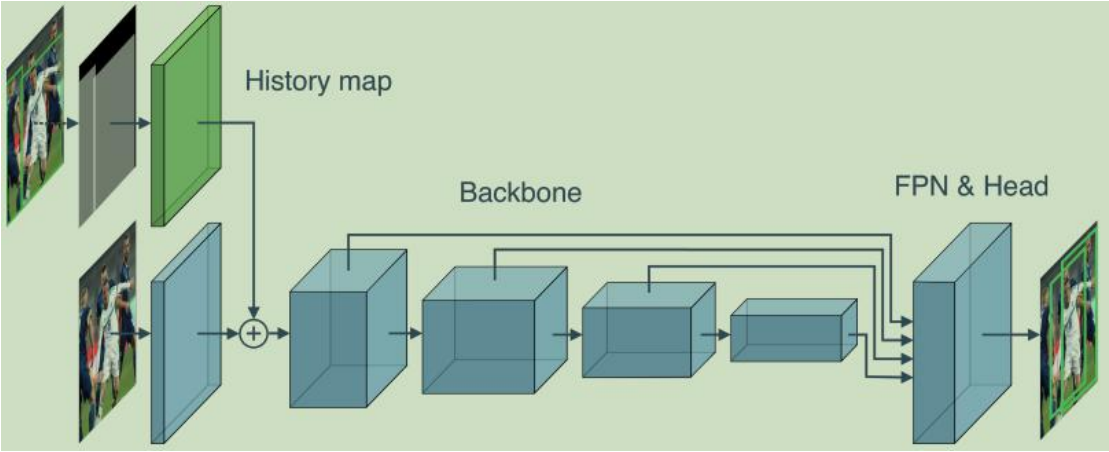

图 2：IterDet 框架。一个任意的检测器未改变的元结构被标记为蓝色。历史特征添加融合的块为绿色。图中 4 个重叠的目标，2 个在历史检测结果中，它们要么在训练过程中随机采样，要么在推理过程中前一次迭代检测到。剩下的两个在本次检测中检测到。

首先，引入符号和描述推理过程，之后，解释训练步骤中的修改。

推理过程。一个典型的目标检测器 D 是一个将图片$I \in \mathbb{R}^{w \times h \times 3}$映射到一组边界框$B = \{(x_k, y_k, w_k, h_k)\}_{k=1}^{n}$的算法。每个框由左上角的坐标（x，y），宽 w 和高 h 表示。对于给定的一组框，定义一个和输入图片相同尺寸的历史图片 H，其中每个像素记录覆盖该像素的检测框数目：

$$H_{xy} = \sum_{k=1}^{|B|} \mathbb{1}_{x_k \le x \le x_k + w_k,\ y_k \le y \le y_k + h_k} \tag{1}$$

图 2 展示了一个历史图片的例子，其中的值是颜色编码的。如果将历史图片 H 和输入图片 I 一起作为检测器输入，可以使检测器 D`对历史敏感。

现在介绍本文提出的检测框架 IterDet（D`），在一次迭代中输入图片 I 产生一组边界框 B。在第一次迭代 t=1 时历史 H1 是空，D`将图片 I 和 H0 映射到一组边界框 B1。第二次，B1 映射到历史 H2 中，历史 H2 接着在迭代 t=2 时通过 D`被映射到 B2。当达到限制的迭代次数或|Bm|=0 时这个步骤才停止。IterDet（D`）最后的预测结果是$B = \bigcup_{i=1}^{m} B_i$.

上述的方案需要两种设计选择：

1）如何修改一个任意的检测器 D 到历史敏感的 D`。

2）如何强迫 D`在每次迭代 t 时预测不同组的目标框 Bt。

历史感知的检测器结构。本文提出的历史感知检测器的结构是简单但有效的。将通过一层卷积层的历史图片和主干网络第一层卷积输出相加。这个方案可以被用于任何主干。对于和 ResNet 类似的主干，图片相加之前，经过了尺寸为 7 步长为 2 的 64 个滤波器的一层卷积，BN 层和 ReLU 激活层。接收历史图片的卷积层有尺寸为 3 步长为 2 的 64 个滤波器.

训练步骤。在训练时随机将真实的边界框 B 分为两个子集 Bold 和 Bnew，这样$B_{old} \cup B_{new} = B$和$B_{old} \cap B_{new} = \emptyset$将 Bold 映射为历史 H 并强迫 D`预测在历史图片中缺失的边界框 Bnew。因此，通过预测框 B 和目标框 Bnew 的误差使用反向传

播优化 D`的损失。一方面，这种训练方法强迫模型利用历史并在每次迭代推理时只预测新的目标。另一方面，通过采样不同的 Bold 和 Bnew 组合提供了额外的数据增强来源。

实验

为了评估本文提出的迭代方案的效果，在三个拥挤的数据集上进行了实验：AdaptIS ToyV1 和 ToyV2，CrowdHuman，WiderPerson。

AdaptIS。AdaptIS ToyV1 和 ToyV2 是两个合成的数据集，原本用于实例分割任务。可用的标注允许使用它们作为目标检测。数据集中的每张图片平均有 30 个目标，其中大多数是高度重叠的。统计结果展示在表 1.

	Toy V1	Toy V2	CrowdHuman	WiderPerson
object/image	14.88	31.25	22.64	29.51
pair/image				
IoU > 0.3	3.67	7.12	9.02	9.21
IoU > 0.4	1.95	3.22	4.89	4.78
IoU > 0.5	0.95	1.25	2.40	2.15
IoU > 0.6	0.38	0.45	1.01	0.81

表 1：实验中使用的四个数据集上对象平均数的比较和成对的两个物体的重叠。

CrowdHuman。从每张图人物的数量和 IoU 大于 0.5 相交的边界框数量来说，最近提出的 CrowdHuman 数据集是目前最复杂的人类图像数据集。每张图片平均有 23 个人，每个人有 3 个框：全部身体，可见身体和头部。最有挑战和经常被用于其它工作的是全部身体的标注，这种框不仅重叠最厉害，而且超出了图像边界。本文还使用了可见部分标注来实验。

WiderPerson。WiderPerson 是从不同来源收集到的另一个密集人类检测数据集。包含五类标注，行人，骑手，部分可见的人，拥挤和忽略的区域。在本文的实验中，将前四类合为一类。

实验结果

AdaptIS 数据集结果。表 2 展示了在 AdaptIS ToyV1 和 ToyV2 数据集上 IterDet 和基准度量。在两个数据集上 IterDet 都大大地提升了 AP。在 faster rcnn 上增加了 4%使最后的 AP 达到 99%.

Method	Detector	Toy V1		Toy V2	
		Recall	AP	Recall	AP
Baseline	RetinaNet	95.46	94.46	96.27	95.62
IterDet, 1 iter.		95.21	95.31	96.27	94.17
IterDet, 2 iter.		**99.56**	**97.71**	**99.35**	**97.27**
Baseline	Faster RCNN	94.05	93.96	94.88	94.81
IterDet, 1 iter.		94.34	94.27	94.97	94.89
IterDet, 2 iter.		**99.60**	**99.25**	**99.29**	**99.00**

Table 2: Experimental results on AdaptIS Toy V1 and Toy V2 dataset.

CrowdHuman 上的结果。CrowdHuman 数据集上全部身体部位和可见身体部位标注的结果分别显示在表 3 和表 4.

Method	Detector	Recall	AP	mMR
Baseline [17]	RetinaNet	**93.80**	80.83	63.33
IterDet, 1 iter.		79.68	76.78	**53.03**
IterDet, 2 iter.		91.49	**84.77**	56.21
Baseline [17]	Faster RCNN	90.24	84.95	50.49
Soft NMS [2, 11]		91.73	83.92	51.97
Adaptive NMS [11]		91.27	84.71	49.73
Repulsion Loss [4, 22]		90.74	85.71	-
PS-RCNN [4]		93.77	86.05	-
IterDet, 1 iter.		88.94	84.43	**49.12**
IterDet, 2 iter.		**95.80**	**88.08**	49.44

Table 3: Experimental results on CrowdHuman dataset with *full* body annotations.

Method	Detector	Recall	AP	mMR
Baseline [17]	RetinaNet	**90.96**	77.19	65.47
Feature NMS [16]		-	68.65	75.35
IterDet, 1 iter.		86.91	81.24	**58.78**
IterDet, 2 iter.		89.63	**82.32**	59.19
Baseline [17]	Faster RCNN	91.51	**85.60**	55.94
IterDet, 1 iter.		87.59	83.28	**55.54**
IterDet, 2 iter.		**91.63**	85.33	55.61

Table 4: Experimental results on CrowdHuman dataset with *visible* body annotations.

在训练时不使用额外的数据和标注，将之前的结果和本文提出的 IterDet 方案在两个检测器上比较：retinanet 和 faster rcnn。正如表 3 最后两行显示，在最具挑战的全身标注上全部三项指标取得了显著的提升。因此，与基线相比，IterDet 的 recall 提高了 5.5%，AP 提高了 3.1%，mMR 降低了 1%.即使与之前的先进方法如 adaptive NMS 和 PS-RCNN 比较，也是有显著的提高。在基础度量 mMR 上面，IterDet 在四个场景中超越了现有的方法：单步检测和多步检测，可见部分和全部身体标注。在两种类型的标注上，都在 retinanet 上超过了 6%.值得注意的是在第一次迭代时就有在 mMR 上有如此的提升，表明了历史敏感的训练对原始的检测器有正则化的效果。尽管随着迭代次数的增加 mMR 有轻微的下降，AP 仍显著地增长。因此，在两种类型的标注上超过了 retinanet 最好的结果 3.9%AP。

WiderPerson 上的结果。WiderPerson 上的结果显示在表 5.将基线的结果用于标注困难子集，这意味所有的框高度都大于 20.在测试时，通过在测试时使用所有的没有高度限制的边界框作为更具挑战的任务。在这两种检测器，本文提出的 IterDet 在 recall，AP 和 mMR 上都显著地超过了所有先前的结果。

Method	Detector	Recall	AP	mMR
Baseline [23]	RetinaNet	-	-	48.32
IterDet, 1 iter.		90.38	87.17	**43.23**
IterDet, 2 iter.		**95.35**	**90.23**	43.88
Baseline [23]	Faster RCNN	-	-	46.06
Baseline [4]		93.60	88.89	-
PS-RCNN [4]		94.71	89.96	-
IterDet, 1 iter.		92.67	89.49	**40.35**
IterDet, 2 iter.		**97.15**	**91.95**	40.78

Table 5: Experimental results on WiderPerson dataset.

迭代次数的选择。表 6 展示了不同检测器和数据集使用不同次数迭代的框架下的 AP 比较。可以看到在第二次迭代后 AP 没有继续增加。

# iter.	CrowdHuman		Toy V2	
	Faster RCNN	RetinaNet	Faster RCNN	RetinaNet
1	84.43	76.78	94.89	95.62
2	**88.08**	**84.47**	**99.00**	**97.27**
3	87.71	84.65	98.96	97.23
4	87.16	83.10	98.96	97.22

Table 6: Comparison of AP for different number of iterations for IterDet based on Faster RCNN or RetinaNet. Full body annotation is used for CrowdHuman.

图 3 展示了本实验中在四个数据集上使用了 IterDet 的 faster rcnn 的检测结果的例子。

可以看到如果当目标被显著地遮挡，第二次迭代确实有助于恢复许多被遮挡的对象。

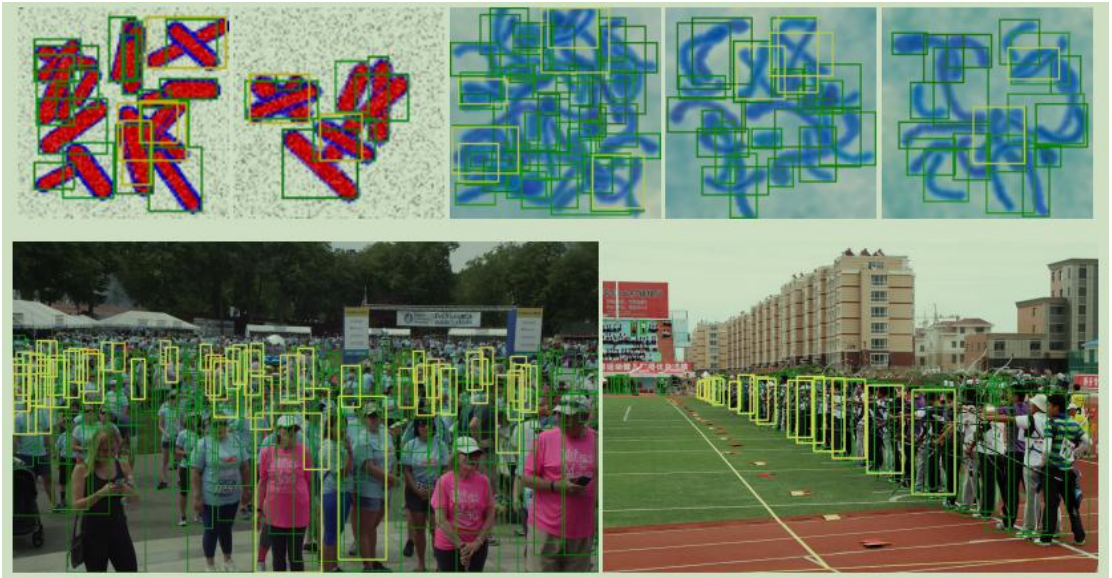

Figure 3: Examples of IterDet results on ToyV1, ToyV2, CrowdHuman (with *full* body annotataions), and WiderPerson. The boxes found on the first and second iterations are marked in green and yellow respectively. The scores thresholded for visualization are above 0.1.

总结

本文为拥挤的环境设计了一个迭代方案(IterDet)用于目标检测可以用于多步和单步检测。

在有多个重叠对象的 AdaptIS ToyV1 和 ToyV2 上的实验证明了 IterDet 能达到几乎完美的检测精度。在 CrowdHuman 和 WiderPerson 上的广泛比较表明，在两阶段 faster rcnn 和一阶段 retinanet 检测器上应用迭代方案都能获得比现有方案更高的精度。