l清 华 大 学

综 合 论 文 训 练

题目：基于深度学习的多物体实时跟踪方法研究与实现

系 别：计算机科学与技术系

专 业：计算机科学与技术专业

姓 名：何蔚然

指导教师：姚海龙

2016年 06月 02日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

关键词：

ABSTRACT

Keywords：

目 录

第1章 引言 1

1.1 选题背景与意义 1

1.2 研究内容 1

1.3 论文结构 1

第2章 现有视频跟踪方法概述 2

2.1 小标题 2

第3章 单目标跟踪算法 3

3.1 小标题 3

第4章 多目标跟踪算法 4

4.1 小标题 4

第5章 行人数据集的应用 5

5.1 小标题 5

第6章 实验结果与分析 6

6.1 小标题 6

第7章 总结与展望 7

7.1 小标题 7

插图索引 8

表格索引 9

参考文献 11

致 谢 13

声 明 15

附录A 书面翻译 17

主要符号对照表

如果论文中使用了大量的物理量符号、标志、缩略词、专门计量单位、自定义名词和术语等，应编写成注释说明汇集表。若上述符号和缩略词使用数量不多，可以不设专门的汇集表，但必须在论文中出现时加以说明。

SOT 单物体追踪(Single Object Tracking)

MOT 多物体追踪(Multiple Objects Tracking)

CNN

FCN

RCNN

KITTI

MALF

INRIA

VGG

ReLU

# 引言

## 选题背景与意义

## 研究内容

## 论文结构

# 现有视频跟踪方法概述

## 小标题

# 单目标跟踪算法

## 小标题

# 多目标跟踪算法

## 小标题

# 行人数据集的应用

## 小标题

# 实验结果与分析

## 小标题

# 总结与展望

## 小标题

插图索引

No headings found. **This is an automatic table of contents. To use it, apply heading styles (on the Home tab) to the text that goes in your table of contents, and then update this table. If you want to type your own entries, use a manual table of contents (in the same menu as the automatic one).**

表格索引

No headings found. **This is an automatic table of contents. To use it, apply heading styles (on the Home tab) to the text that goes in your table of contents, and then update this table. If you want to type your own entries, use a manual table of contents (in the same menu as the automatic one).**

参考文献

[1]

致 谢

衷心感谢姚海龙老师在本次毕业设计及综合论文训练中对我的帮助和悉心指导。姚海龙老师认真负责的研究态度以及严谨的作风都让我受益匪浅。

最后感谢班里同学在我完成综合论文训练阶段给予我的帮助。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日 期：\_\_\_\_\_\_\_\_\_\_\_\_\_\_

附录A 书面翻译

稠密盒：基于端到端物体识别的统一标志定位

A1 摘要

一个FCN（全卷积神经网络）要怎么在物体检测上工作？我们发明了稠密盒——一个统一的端到端FCN框架——可以直接预测单张图片中任意位置及大小物体的包围盒以及物体分类置信度。我们的贡献主要有两方面。一是我们证明了一个FCN，只要经过小心的设计以及优化，可以极其准确且高效地检测多个不同物体；二是我们证明在多任务学习时引入标志定位，稠密盒可以更好地提高物体检测精确度。我们展示了在包括MALF人脸检测以及KITTI车辆检测的公开数据集上的实验结果，结果表明我们的稠密盒系统在检测如人脸、车辆等有挑战性的物体方面是目前最先进的。

A2 引言

我们的日常生活充斥着大量的物体检测实例。驾驶时注意最近的车辆、找一个人以及定位一张熟悉的脸，这些都是物体检测的例子。物体检测是计算机视觉中最核心的问题之一。在CNN成功以前，物体检测常常会归类为基于移动窗的方法，也就是在图像中所有可能的位置及尺度中运用分类器抽取人工特征。最近，基于FCN的方法给物体检测领域带来了一场革命。这些FCN框架也遵循移动窗的特点，但他们在学习模型参数及抽取图像特征时端到端的方式显著提升了检测表现。

RCNN相对基于FCN的方法在物体检测准确率上有了进一步提高。理论上，RCNN包含两个部分。一是使用区域提议方法生成图像中所有潜在的候选包围盒；二是使用了CNN分类器为每个提议分辨出不同的物体。虽然RCNN在普适物体识别上成为了新的最先进系统，但它仍很难检测如人脸、远处的车等小物体，因为低分辨率及缺少背景信息的候选包围盒会显著降低分类准确率。更多地，RCNN流程上这两个不同部分不可以被合并地优化，为端到端训练RCNN留下了麻烦。

我们的工作关注一个问题：在物体检测上一个单级FCN能达到什么程度？为此我们提出了一个基于FCN的物体检测器——稠密盒——不需要提议生成并且可以在训练时被端到端地优化。尽管与很多已有移动窗FCN检测框架相似，稠密盒还是在小尺度及高污染的情况下被更仔细地设计。我们训练稠密盒并使用精细、困难的挖掘技术去压榨它的表现。为了使它表现更好，我们更通过多任务学习在系统中整合了标志定位。为了验证标志定位是有用的，我们在KITTI车辆检测上手工标注了一批关键点数据，并会在将来公开它们。

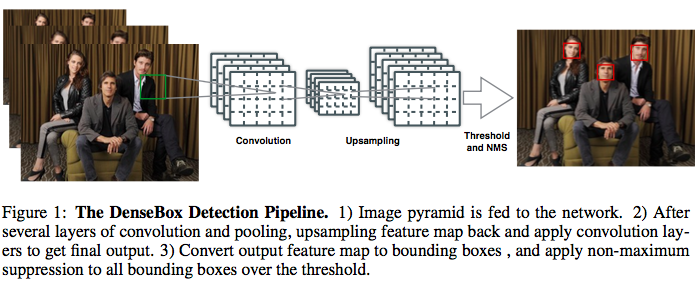
A3 稠密盒检测

图 1

一个完整的检测系统如图 1所示。一个卷积网络同时输出多个预测包围盒以及分类置信度。稠密盒中物体检测的所有部件都被建模成一个FCN（除了非极大值抑制的部分），所以区域提议生成不是必要的。在测试中，系统把一张图片（尺寸为）作为输入，并输出一张尺寸为的5通道特征图。如果在输出坐标空间中我们分别定义包围盒的左上和右下角为和，那么在输出特征图中，每一个位于的像素点i会以一个5维向量 描述一个包围盒，其中是作为一个物体的置信度分值，表明输出像素位置距离目标包围盒的边界的距离。最后每个输出图中的像素被转化为带置信度的包围盒，并对其中置信度达到阈值的包围盒应用非极大值抑制。

A3.1 生成标注值

把整个图像输入到网络里训练是不必要的，因为这会花费大多数计算时间在对背景的卷积操作上。一个聪明的策略是剪切含有人脸以及足够背景信息的子图进行训练。在论文中，我们在单尺度上训练网络，并使用多尺度进行校验。

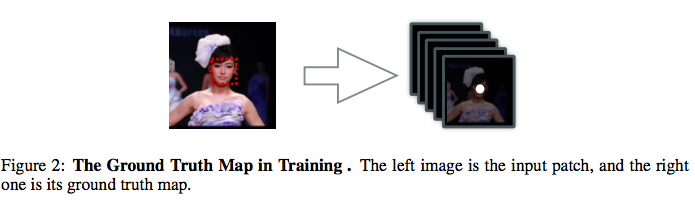
通常来说，我们提出的网络是使用类似分割的方法训练的。在训练时，子图会被剪切并缩放到，并包含位于正中间的高度大致50像素的人脸。训练时输出标注值时一个被缩放到的5通道图，向下采样系数为4。标注图第一个通道内的正例标注区域是一个半径为的填充圆，圆心在人脸包围盒的中心。半径对包围盒尺寸是正定的，输出坐标空间中相对盒子尺寸的缩放系数被设为0.3，如图 2所示。剩下的4通道被输出图像素坐标距离最近包围盒的左上、右下角的距离填满。

图 2

注意如果在单个子图中出现多张人脸，如果他们在相对于子图中心缩放范围内（在我们的设置中设为0.8到1.25），我们会保持这些脸为正例.其他脸会被视为负例。第一个通道中的像素点——表明了分类的置信度——在标注图中会被初始化为1，之后只要在正例标注区域内则设为1。我们也发现我们的标注值生成与Pinheiro等提出的分割工作比较相似。在他们的方法中，像素标注被物体在子图中的位置所决定，而在稠密盒中，像素标注被感应区域决定。具体地，如果像素满足感应区域大致包含一个物体在中心且在给定尺度的限制，则它会被标注为1。每个像素会被视为一个抽样，因为每个5通道像素描述一个包围盒。

A3.2 模型设计

我们展示在图 3的网络架构是从用于图像分类的VGG 19模型演化过来的。整个网络有16个卷积层，前12个卷积层被VGG 19模型初始化。conv4\_4的输出被送进四个的卷积层，其中前两个卷积层输出表示分类值的单通道图，另两个通过4通道图预测包围盒的相对位置。最后的卷积层就像移动窗里的全连接层一样工作。

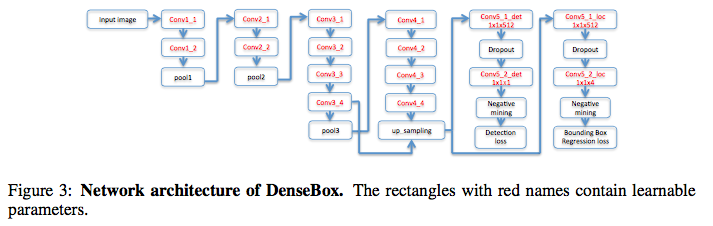
多层特征融合。最近的工作表明使用不同卷积层得到的特征可以增强如边界检测与分割的效果。旁支级别特征关注物体的本地细节以找到有判断性的外观部件，物体级别和高级别特征为了识别物体通常有更大的感应区域。更大的感应区域还未预测更准确的结果带来背景信息。在我们的实现中，我们把conv3\_4和conv4\_4的特征图连接在一起。conv\_3\_4的感应区域（或称移动窗口尺寸）是，几乎和训练中脸的尺寸一样，并且conv4\_4有一个大得多的感应区域，大概，可以利用全局纹理和上下文用于检测。注意conv4\_4的特征图尺寸是conv3\_4生成的图的一般，所以我们使用双向上采样层去把它们转换成相同的分辨率。

图 3

A3.3 多任务训练

我们使用ImageNet提前训练的VGG 19网络去初始化稠密盒。实际上，在初始化时，我们只保留前12个卷积层（从conv1\_1到conv4\_4），并且VGG 19中的其他层都被四个带有xavier初始化的新卷积层所替代。

像Fast R-CNN一样，我们的网络有两个兄妹输出分支。第一个输出（输出图中的每个像素）是一个目标物体的置信度分值。给定标注值，分类的损失函数可以被定义如下：

|  |  |  |
| --- | --- | --- |
|  |  | (‑) |

这里我们在人脸和车辆检测任务中都使用L2损失函数。我们没有尝试像hinge损失或者交叉熵损失这样其他的损失函数，这看起来是一个更正确的选择，因为我们发现简单的L2损失函数在我们的任务中表现很出色。

第二个分支输出包围盒的回归损失，标记为。它的目标是最小化预测位置偏移和目标间的L2损失，用公式表示为：

|  |  |  |
| --- | --- | --- |
|  |  | (‑) |

A3.3.1 平衡抽样

选择负例样本的过程是学习中的一个关键部分。如果在一个迷你批中简单地使用所有负例样本，由于他们统治了所有样本，将会使得对负例样本的预测有偏差。更多地，如果我们在那些位于正例与负例区域间的样本上进行损失惩罚，检测器将会降级。这里我们对每个输出像素使用一个二值遮盖来表明它是否在训练中被选中。

无视灰色区域。灰色区域是定义在正例与负例区域之间的地方。它不应该被看作正例或者负例，并且它的损失权重应该被设为0。对每个在输出坐标空间中的非正例标注像素，当且仅当存在在个像素长度之内任意带正例标注的像素，它的无视标志才被设为1。

困难负例挖掘。对在SVM中的困难负例挖掘过程的分析，通过搜索预测不好的样本而不是随机样本，我们使得学习更有效。在负例挖掘之后，预测不好的样本会有很大的几率被选中，于是在这些样本上的梯度下降学习引向了更鲁棒的预测同时还有更少的噪音。具体地，负例挖掘可以通过在线引导被有效地实施。在向前传播过程中，我们对输出像素损失(7‑1)进行倒序排序，并且对前1%分配为困难负例。在所有实验中，我们保留了所有正例标注像素（样本）并且正例与负例之比为1:1。在所有负例样本中，他们的一半是从困难负例样本中被抽样，并且剩下的一半会随机地从非困难负例中抽样。为了方便，我们对那些在迷你批中的像素（样本）设标志。

带遮盖的损失。现在我们可以为每个样本定义遮盖为一个关于上述提到的标志的函数：

|  |  |  |
| --- | --- | --- |
|  |  | (‑) |

然后如果我们合并分类(7‑1)和带遮罩包围盒回归(7‑2)损失，我们的全多任务损失可以被表示成，

|  |  |  |
| --- | --- | --- |
|  |  | (‑) |

其中θ是网络中参数的集合，并且艾弗森括号函数只有当标注值得分为正时才会被激活。明显的是对负例样本（背景）的包围盒回归损失应该被无视，因为他们并没有任何记号。分类和回归任务之间的平衡被参数所控制。在我们的实验中，我们通过除标准物体高度，也就是在标注图中50/4，对回归目标进行归一化，并且在这种归一化的所有实验中，表现良好。

其他实现细节。训练中，当一个输入批在特定尺寸下在正中间包含一个居中的物体，那它会被视为正例批。这些批的正例样本周围仅仅包含负例样本。为了在整个数据集中完全探索负例样本，我们也从训练图像中随机剪切了随机尺寸的子图，并把它们缩放到相同的尺寸并把它们送进网络。我们把这种批叫做“随机批”，并且训练时“正例批”与“随机批”之间的比例为1:1。还有，为了更好地增加我们模型的鲁棒性，我们也在送进网络前对每个批进行随机抖动。具体地，我们应用了左-右翻转、位移变换（在25像素范围内）以及尺度变形（从0.8到1.25）。

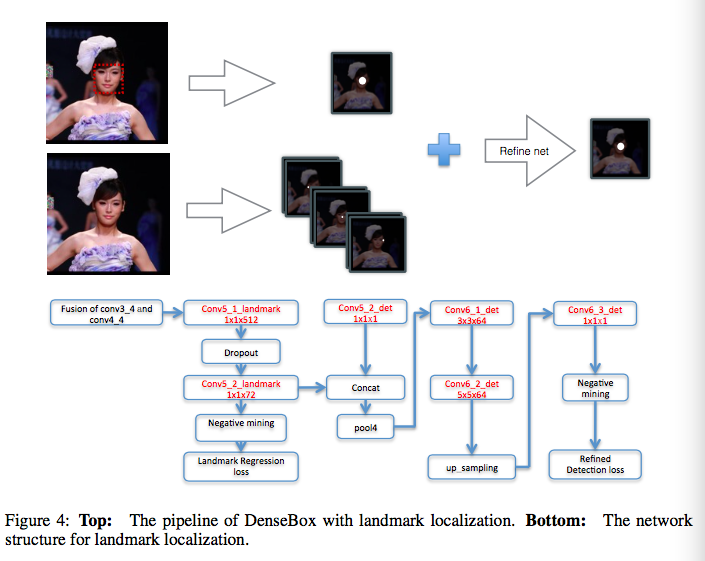
我们在训练中使用迷你批SGD并且批大小被设为10。损失以及输出梯度必须被有贡献像素的数量所缩放，使得损失和输出梯度都能在多任务学习中变得可比较。全局学习率一开始为0.001，并在每10万次迭代中被参数为10地衰减。我们保持默认动量项权重为0.9以及权重衰减系数为0.0005。

图 4

A3.4 改善标志定位

在这一部分，我们展示了仅仅通过累加几个全卷积网络结构中的网络层，标志定位就可以在稠密盒中被实现。更多地，我们可以通过融合标志热力图和人脸分值图来改善检测结果。如图 4所示，我们引入了另一个兄妹分值输出给标志定位。假设有N个标志点，则标志定位分支输出N个对应的图，上边每个像素点代表那个位置是一个标志点的置信度分值。表面上看，标注值图在这个任务上被使用和检测的标注值很相似。对一个标志实例，标志k的第i个实例，它的标注值是一个位于输出坐标空间中第k张响应图对应位置的正例标注区域。注意半径应该相对地小（比如）以避免精度的损失。与分类任务很相似，标志定位损失被定义成一个预测值与标注的L2损失，并且我们仍然应用负例挖掘且无视之前章节讨论过的区域。

最后的输出改善分支，得到分类分值图以及标志定位图作为输入，目标是改善分类结果。一个正确的方案应该是使用高层空间模型去学习标志置信度以及包围盒分值的限制，以进一步增强检测的表现。Tompson等提出的一个类MRF模型使用修改过的卷积（SoftPlus卷积）以及非负输出去连接每一个部位的空间信息分布。不管怎样，他们的模型也包含了和层级，使得模型非常难以训练。在我们的实现中，我们使用带ReLU激活的卷积操作去近似空间模型。如果我们把改善检测损失记为，几乎与之前提到但预测图是来自改善分支的分类损失中的一样，完整的损失变成了，

|  |  |  |
| --- | --- | --- |
|  |  | (‑) |

其中和控制三个任务之间的平衡。在我们的实验中，他们被分别分配为1和0.5。

A3.5 对比

稠密和的闪光点在于它把物体检测建构为一个回归问题，并且提供了一种端到端的识别框架。例如YOLO和Faster R-CNN的几个最近的工作都把区域提议生成和分类绑定在一起。这里我们比较了稠密和和其他相关检测系统，指出其中的关键相似点和不同点。

传统基于神经网络的人脸检测器。那些使用在最近取得突破性结果的图片分类卷积神经网络之前的神经网络的人脸检测系统，使用的是基于神经网络的人脸检测器。在人脸检测中使用神经网络已经有了很长的历史，可以追溯到1990年代的早期工作。Rowley等训练只在特定尺寸的人脸上可以被激活的基于神经网络的检测器，并且在带移动窗的图片金字塔中应用检测器。我们的稠密和与他们在检测流程上非常相似，除了我们使用了现代卷积神经网络作为检测器。因此稠密和可以被同样地称为“现代基于神经网络的检测器”。

OverFeat。被Sermanet等设计的OverFeat可能是在深度卷积神经网络在图片分类上成功应用后，第一个训练卷积神经网络去同时完成分类和定位的工作。它也在测试时应用全卷积神经网络，一个等价但更加有效得多的实现移动窗检测的方式。不管怎样，它仍然在训练时分离了分类和定位，并且需要复杂的后处理去处理检测结果。我们的方法和OverFeat非常相似但是一个多任务合并训练的端到端检测网络。

深度稠密人脸检测器（DDFD）。被Farfade等提出的DDFD是一个基于卷积神经网络的人脸检测系统。它声称在人脸检测任务上有超过R-CNN的表现，因为R-CNN中的区域生成可能会丢失一些人脸区域。尽管DDFD是一个完整的检测流程，但DDFD并不是一个端到端框架因为它把分类可能性预测和包围盒定位分离成了两个任务和两个层级。我们的稠密盒可以为了检测直接优化，并且可以通过引入标志信息被轻易地提高。

Faster R-CNN。Faster R-CNN仍然使用区域提议去在一张图片中寻找物体。不像它的前任变数，在Faster R-CNN中的区域提议是被区域提议网络（RPN）在第二级中与分类共享卷积特征计算生成的。RPN与我们的方法稠密盒共享了许多相似点。不管怎样，RPN需要提前定义锚点但我们的方法不需要。RPN是在多尺度物体上被训练的，但本论文展示的稠密盒是在经过抖动增广的单尺度上训练的，这意味着我们的方法需要在多尺度上验证特征。更多地，RPN与稠密盒的训练方式是非常不同的。

多重盒。多重盒训练了一个卷积神经网络去生成提议而不是通过选择搜索。多重盒和稠密盒都被训练来在一张图片中预测包围盒，它们以不同的方式生成包围盒。多重盒方法生成了800个具有非平移不变形的锚点，而我们的稠密盒像RPN一样直接输出平移不变的包围盒。因为输出图的向下采样系数为4，稠密盒会稠密地每4个像素点生成一个带分值的包围盒。

YOLO。Redmon等提出了一个统一物体检测流程，叫YOLO。稠密盒和YOLO都可以被从图片端到端地训练，但在输出层的模型设计上有所不同。YOLO系统把大小为的图片作为输入，并输出的网格，每张图片上只有49个包围盒。我们的稠密盒使用向上采样层去保持相对高分辨率的输出，在我们的模型中有一个缩放系数为4的向下采样。这让我们的网络能够检测非常小的物体以及高度重合的物体，这是YOLO不可能能够处理的。

A4 实验

在这一节中，我们阐述了在多属性标注人脸（MALF）数据集以及KITTI车辆检测任务上稠密盒的表现。不管有没有标志标注的帮助，我们都在那些任务上校验了我们的方法，显示带标志定位的多任务学习能显著地提高表现。我们比较了当前世界最好的系统与我们的方法，显示我们的方法在物体检测任务上实现了有竞争力的结果。注意到我们并没有直接比较我们的稠密盒与原始R-CNN的表现，但是我们强调了声称使用R-CNN或者其他已经被与R-CNN比较过的方法的方法的表现。

A4.1 MALF识别任务

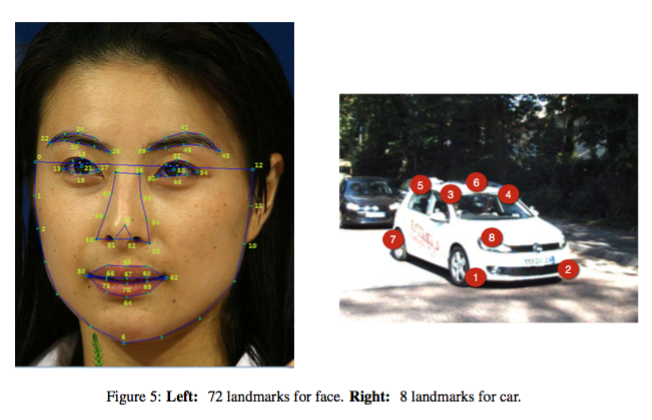
MALF检测测试数据集包含从互联网中搜集过来的5000张图片。不像被广泛使用的FDDB人脸检测数据集——这是从新闻图片和海报正面中搜集过来的额——MALF中的人脸图片有更大的变化度，使得它比FDDB更接近真实世界的应用。

图 5

训练与测试。我们在包含81024张有72个标志点的人脸31337张从互联网搜集到的图片上，训练了在第三节中描述过的两个模型，见图 5。一个模型至使用了包围盒信息，而另一个模型使用了包围盒以及标志信息，以进行比较。他们都被在ImageNet上提前训练好的VGG 19模型所初始化。在训练中人脸被大致缩放到高度为50像素，而缩放抖动方位为[0.8,1.25]，与在3.3节中描述的一样。在测试时，我们首先选择性地向下采样图片以对于每张图片最长的图片边不会超过800像素点。然后我们在每张图片上以多个尺度测试我们的模型。测试尺度以的步长从开始到结束。这种设置使得我们的模型能够检测高度从20像素点到400像素点的人脸。在人脸检测上的非极大值抑制IOU阈值被设为0.5。在这种配置下，在MALF数据集上一张英伟达K40显卡处理一张图片需要几秒时间。结果。我们描述了在MALF数据集上三种版本的稠密盒的结果。“无标志点稠密盒”显示了训练中没有标志点的稠密盒。“稠密盒标志点”是引入标志定位的模型，并且“稠密盒组装”是10个有从不同批迭代得来的标志点的稠密盒的结果。如图 6所示，标志定位给出了一个在人脸检测上显著的表现提高。我们也注意到使用不同批迭代训练的模型仍然有很高的变化度，因为另外的显著提高已经在模型组装中被看见了。然后我们比较了在MALF上我们最好的模型与其他世界最好的方法。令人意外的，我们的模型实现了最好的表现，平均召回率在87.26%，几乎比DDFD表现好了10%，说明我们在人脸检测任务上比R-CNN有更好的表现。

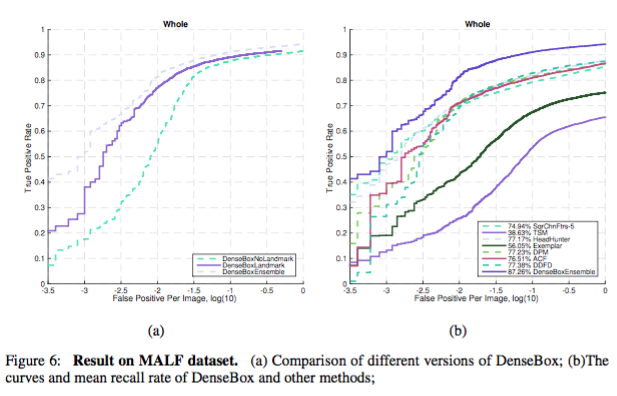
A4.2 KITTI车辆检测任务

图 6

KITTI物体检测数据集有7481张训练图片以及7518张测试图片。在训练中物体总数达到51867个，其中车辆仅仅占了28742个。KITTI车辆检测任务的主要难点在很大量的汽车是小尺寸的（高度小于40像素点）并且被遮挡的。为了克服这个难点，前人工作需要仔细的旁路及遮盖建模。

训练与测试。与人脸检测任务一样，我们在KITTI物体检测训练集上训练了两个网络（一个没有标志点，另一个有标志点）。因为KITTI没有为车辆提供标志点，我们选择性地为大车辆（高度大于50像素点）标注了8个标志点。车辆的标志点如图 5所示，我们最后标注了7790辆汽车，大概占车辆总数的27%。测试程序与在人脸检测中一样，除了我们并没有对车辆图片进行向下采样。KITTI车辆检测任务的评价方式与普适物体检测并不相同。KITTI的真正例包围盒需要有70%的重叠面积，而其他像人脸检测的任务只需要50%的重叠面积。这个严格的边界条件要求更高的车辆定位准确度。在KITTI上，我们设非极大值抑制IOU阈值为0.75。

结果。表 1显示了稠密盒以及其他方法的结果。我们可以看见部分标注标志信息（27%）仍然可以提高检测表现。平均上，有标志定位的模型轻微地比没有标志点的模型在平均精确度上提高了0.9%。在人脸检测上，表现的提高并没有那么好。原因可能是标志点信息并不充分。不充分性体现在总量（车辆上27%然而人脸是100%）和质量（车辆上8个标志点而人脸上是74）上。与其他方法相比，稠密盒仍然实现了有竞争力的结果。稠密盒以较大的差距打败了如Regionlets和spCov等传统检测系统。我们在经济适用车辆上的平均精确度为85.74%，比使用了在ImageNet上提前训练的GoogLeNet的R-CNN框架的DeepInsight轻微地好。我们的模型直到一个标题为“NIPS ID 331”匿名提交前已经4个月排名第一了，在训练和测试中使用了立体信息。最近一个叫“DJML”的方法超越了其他所有方法。

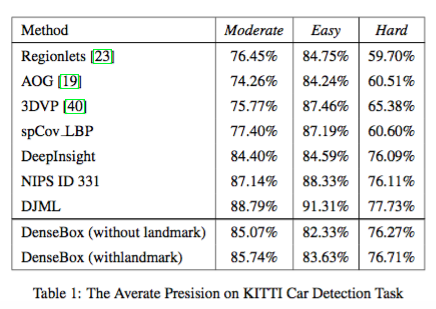
A5 总结

表 1

我们已经展示了稠密盒——一个针对检测的统一端到端检测流程。它的表现可以通过纳入标志信息被轻易地加速。我们也分析了我们的方法以及其他相关的物体检测系统，强调了其中稠密盒的不同以及贡献。稠密盒在人脸检测和车辆检测任务上都展现了令人印象深刻的表现，表明它对那些区域生成可能失败的情况的高适应性。稠密盒主要的问题是速度。论文中展示的原始的稠密盒需要几秒去处理一张图片。但是这已经在我们的后一个版本中被处理。我们会展示另一篇描述一个在KITTI和人脸检测的实时检测系统论文，叫稠密盒2。