[基本概念 2](#_Toc318968608)

[一 IoU（Intersection over Union） 3](#_Toc608753878)

[二 基于深度学习的目标检测 4](#_Toc1529250317)

[1、传统目标检测 5](#_Toc114055912)

[1.1 区域选择 5](#_Toc1229397415)

[1.2 特征提取 6](#_Toc1144513000)

[1.3 分类器 6](#_Toc2053342603)

[1.4 总结 7](#_Toc343694914)

[2、基于深度学习的目标检测 7](#_Toc1582027126)

[2.1相关研究 10](#_Toc1436260976)

[2.2、基于区域提名的方法 14](#_Toc2106734191)

[（1）R-CNN 14](#_Toc359316437)

目标识别有关

# 基本概念

1）CNN：Convolutional Neural Networks

2）FC：Fully Connected 全连接

3）IoU：Intersection over Union （IoU的值定义：Region Proposal与Ground Truth的窗口的交集比并集的比值，如果IoU低于0.5，那么相当于目标还 是没有检测到）

4）ICCV：International Conference on Computer Vision

5）R-CNN：Region-based Convolutional Neural Networks

6）AR：Average Recall

7）mAP：mean Average Precision 平均精度

8）RPN：Region Proposal Networks

基于区域提名的神经网络 or 基于候选区域的神经网络

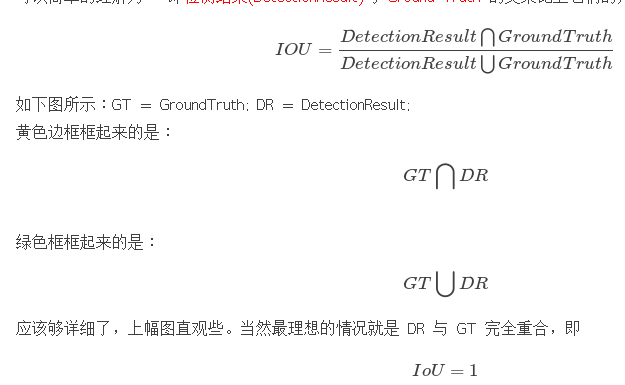
9）FAIR：Facebook AI Research

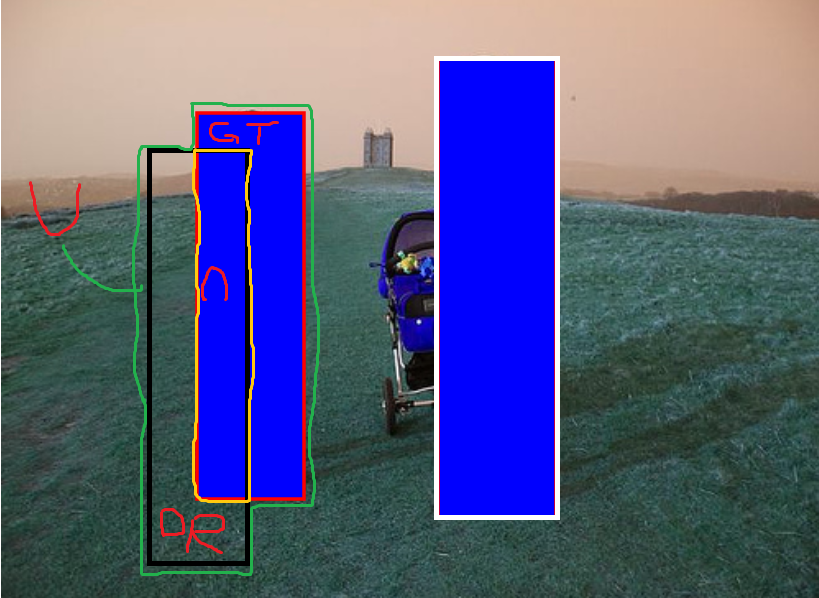
10）w.r.t.：with respect to

# 一 IoU（Intersection over Union）

检测评价函数IOU。

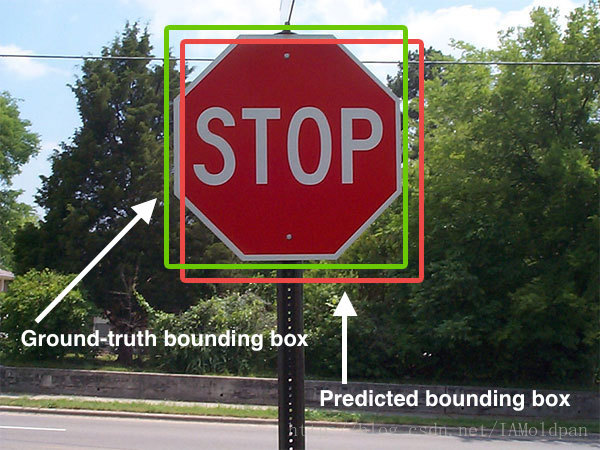
在目标检测的评价体系中，有一个参数叫做 IoU ，简单来讲就是模型产生的目标窗口和原来标记窗口的交叠率。具体我们可以简单的理解为： 即检测结果(DetectionResult)与 Ground Truth 的交集比上它们的并集，即为检测的准确率 IoU 。





通常我们在 HOG + Linear SVM object detectors 和 Convolutional Neural Network detectors (R-CNN, Faster R-CNN, YOLO)中使用该方法检测其性能。注意，这个测量方法和你在任务中使用的物体检测算法没有关系。

IoU是一个简单的测量标准，只要是在输出中得出一个预测范围(bounding boxex)的任务都可以用IoU来进行测量。为了可以使IoU用于测量任意大小形状的物体检测，我们需要：   
1、 ground-truth bounding boxes（人为在训练集图像中标出要检测物体的大概范围）；   
2、我们的算法得出的结果范围。



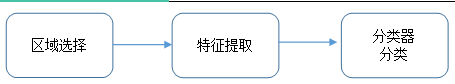
# 二 基于深度学习的目标检测

普通的深度学习监督算法主要是用来做分类，而在ILSVRC（ImageNet Large Scale Visual Recognition Challenge)竞赛以及实际的应用中，还包括目标定位和目标检测等任务。其中目标定位是不仅仅要识别出来是什么物体（即分类），而且还要预测物体的位置，位置一般用边框（bounding box）标记。而目标检测实质是多目标的定位，即要在图片中定位多个目标物体，包括分类和定位。

目标检测对于人类来说并不困难，通过对图片中不同颜色模块的感知很容易定位并分类出其中目标物体，但对于计算机来说，面对的是RGB像素矩阵，很难从图像中直接得到狗和猫这样的抽象概念并定位其位置，再加上有时候多个物体和杂乱的背景混杂在一起，目标检测更加困难。

## 1、传统目标检测

传统的目标检测一般使用滑动窗口的框架，主要包括三个步骤：



### 1.1 区域选择

利用不同尺寸的滑动窗口框住图中的某一部分作为候选区域。

这一步是为了对目标的位置进行定位。由于目标可能出现在图像的任何位置，而且目标的大小、长宽比例也不确定，所以最初采用滑动窗口的策略对整幅图像进行遍历，而且需要设置不同的尺度，不同的长宽比。这种穷举的策略虽然包含了目标所有可能出现的位置，但是缺点也是显而易见的：时间复杂度太高，产生冗余窗口太多，这也严重影响后续特征提取和分类的速度和性能。（实际上由于受到时间复杂度的问题，滑动窗口的长宽比一般都是固定的设置几个，所以对于长宽比浮动较大的多类别目标检测，即便是滑动窗口遍历也不能得到很好的区域）

### 1.2 特征提取

提取候选区域相关的视觉特征，比如人脸检测常用的Harr特征；行人检测和普通目标检测常用的HOG特征等；

由于目标的形态多样性，光照变化多样性，背景多样性等因素使得设计一个鲁棒的特征并不是那么容易。然而提取特征的好坏直接影响到分类的准确性。（这个阶段常用的特征有SIFT、HOG等）

### 1.3 分类器

利用分类器进行识别，比如常用的SVM模型。

传统的目标检测中，多尺度形变部件模型DPM（Deformable Part Model）[13]是出类拔萃的，连续获得VOC（Visual Object Class）2007到2009的检测冠军，2010年其作者Felzenszwalb Pedro被VOC授予”终身成就奖”。DPM把物体看成了多个组成的部件（比如人脸的鼻子、嘴巴等），用部件间的关系来描述物体，这个特性非常符合自然界很多物体的非刚体特征。DPM可以看做是HOG+SVM的扩展，很好的继承了两者的优点，在人脸检测、行人检测等任务上取得了不错的效果，但是DPM相对复杂，检测速度也较慢，从而也出现了很多改进的方法。正当大家热火朝天改进DPM性能的时候，基于深度学习的目标检测横空出世，迅速盖过了DPM的风头，很多之前研究传统目标检测算法的研究者也开始转向深度学习。

### 1.4 总结

传统目标检测存在的两个主要问题：一个是基于滑动窗口的区域选择策略没有针对性，时间复杂度高，窗口冗余；二是手工设计的特征对于多样性的变化并没有很好的鲁棒性。

## 2、基于深度学习的目标检测

基于深度学习的目标检测发展起来后，其实效果也一直难以突破。比如文献[6]中的算法在VOC 2007测试集合上的mAP只能30%多一点，文献[7]中的OverFeat在ILSVRC 2013测试集上的mAP只能达到24.3%。2013年R-CNN诞生了，VOC 2007测试集的mAP被提升至48%，2014年时通过修改网络结构又飙升到了66%，同时ILSVRC 2013测试集的mAP也被提升至31.4%。

那么，对于传统目标检测任务存在的两个主要问题，我们该如何解决呢？

对于滑动窗口存在的问题，region proposal提供了很好的解决方案。region proposal（候选区域）是预先找出图中目标可能出现的位置。但由于region proposal利用了图像中的纹理、边缘、颜色等信息，可以保证在选取较少窗口（几千个甚至几百个）的情况下保持较高的召回率。这大大降低了后续操作的时间复杂度，并且获取的候选窗口要比滑动窗口的质量更高（滑动窗口固定长宽比）。比较常用的region proposal算法有selective Search和edge Boxes，如果想具体了解region proposal可以看一下PAMI2015的[“What makes for effective detection proposals？”](https://www.mpi-inf.mpg.de/departments/computer-vision-and-multimodal-computing/research/object-recognition-and-scene-understanding/how-good-are-detection-proposals-really/)

有了候选区域，剩下的工作实际就是对候选区域进行图像分类的工作（特征提取+分类）。对于图像分类，不得不提的是2012年ImageNet大规模视觉识别挑战赛（ILSVRC）上，机器学习泰斗Geoffrey Hinton教授带领学生Krizhevsky使用卷积神经网络将ILSVRC分类任务的Top-5 error降低到了15.3%，而使用传统方法的第二名top-5 error高达 26.2%。此后，卷积神经网络占据了图像分类任务的绝对统治地位，微软最新的ResNet和谷歌的Inception V4模型的top-5 error降到了4%以内多，这已经超越人在这个特定任务上的能力。所以目标检测得到候选区域后使用CNN对其进行图像分类是一个不错的选择。

2014年，RBG（Ross B. Girshick）大神使用region proposal+CNN代替传统目标检测使用的滑动窗口+手工设计特征，设计了R-CNN框架，使得目标检测取得巨大突破，并开启了基于深度学习目标检测的热潮。

R-CNN是Region-based Convolutional Neural Networks的缩写，中文翻译是基于区域的卷积神经网络，是一种结合区域提名（Region Proposal）和卷积神经网络（CNN）的目标检测方法。Ross Girshick在2013年的开山之作《Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation》[1]奠定了这个子领域的基础，这篇论文后续版本发表在CVPR 2014[2]，期刊版本发表在PAMI 2015[3]。

其实在R-CNN之前已经有很多研究者尝试用Deep Learning的方法来做目标检测了，包括OverFeat[7]，但R-CNN是第一个真正可以工业级应用的解决方案，这也和深度学习本身的发展类似，神经网络、卷积网络都不是什么新概念，但在本世纪突然真正变得可行，而一旦可行之后再迅猛发展也不足为奇了。

R-CNN这个领域目前研究非常活跃，先后出现了R-CNN[1,2,3,18]、SPP-net[4,19]、Fast R-CNN[14, 20] 、Faster R-CNN[5,21]、R-FCN[16,24]、YOLO[15,22]、SSD[17,23]等研究。Ross Girshick作为这个领域的开山鼻祖总是神一样的存在，R-CNN、Fast R-CNN、Faster R-CNN、YOLO都和他有关。这些创新的工作其实很多时候是把一些传统视觉领域的方法和深度学习结合起来了，比如选择性搜索（Selective Search)和图像金字塔（Pyramid）等。

深度学习相关的目标检测方法也可以大致分为两派：

（1）基于区域提名的，如R-CNN、SPP-net、Fast R-CNN、Faster R-CNN、 R-FCN；

（2）端到端（End-to-End），无需区域提名的，如YOLO、SSD。

目前来说，基于区域提名的方法依然占据上风，但端到端的方法速度上优势明显，后续的发展拭目以待。

### 2.1相关研究

（1）region proposal

目标检测的第一步是要做区域提名（Region Proposal），也就是找出可能的感兴趣区域（Region Of Interest, ROI）。区域提名类似于-光学字符识别（OCR）领域的切分，OCR切分常用过切分方法，简单说就是尽量切碎到小的连通域（比如小的笔画之类），然后再根据相邻块的一些形态学特征进行合并。但目标检测的对象相比OCR领域千差万别，而且图形不规则，大小不一，所以一定程度上可以说区域提名是比OCR切分更难的一个问题

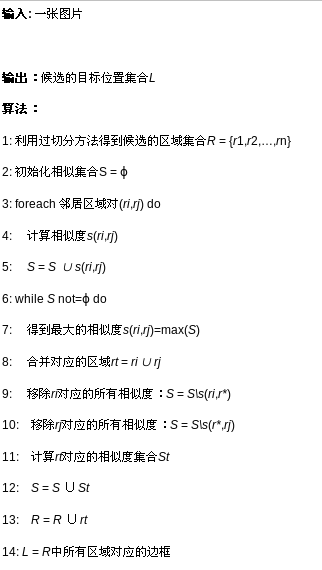
区域提名（候选区域）可能的方法有：

A）滑动窗口。滑动窗口本质上就是穷举法，利用不同的尺度和长宽比把所有可 能的大大小小的块都穷举出来，然后送去识别，识别出来概率大的就留下来。 很明显，这样的方法复杂度太高，产生了很多的冗余候选区域，在现实当中 不可行。

B）规则块。在穷举法的基础上进行了一些剪枝，只选用固定的大小和长宽比。 这在一些特定的应用场景是很有效的，比如拍照搜题APP小猿搜题中的汉 字检测，因为汉字方方正正，长宽比大多比较一致，因此用规则块做区域提 名是一种比较合适的选择。但是对于普通的目标检测来说，规则块依然需要 访问很多的位置，复杂度高。

C）选择性搜索。从机器学习的角度来说，前面的方法召回是不错了，但是精度 差强人意，所以问题的核心在于如何有效地去除冗余候选区域。其实冗余候 选区域大多是发生了重叠，选择性搜索利用这一点，自底向上合并相邻的重 叠区域，从而减少冗余。

选择性搜索的具体算法细节[8]如算法1所示。总体上选择性搜索是自底向上不断合并候选区域的迭代过程。



从算法不难看出，R中的区域都是合并后的，因此减少了不少冗余，相当于准确率提升了，但是别忘了我们还需要继续保证召回率，因此算法1中的相似度计算策略就显得非常关键了。如果简单采用一种策略很容易错误合并不相似的区域，比如只考虑轮廓时，不同颜色的区域很容易被误合并。选择性搜索采用多样性策略来增加候选区域以保证召回，比如颜色空间考虑RGB、灰度、HSV及其变种等，相似度计算时既考虑颜色相似度，又考虑纹理、大小、重叠情况等。

总体上，选择性搜索是一种比较朴素的区域提名方法，被早期的基于深度学习的目标检测方法（包括Overfeat和R-CNN等）广泛利用，但被当前的新方法弃用了。

区域提名并不只有以上所说的三种方法，实际上这块是非常灵活的，因此变种也很多，有兴趣的读者不妨参考一下文献[12]。

（2）OverFeat

OverFeat[7][9]是用CNN统一来做分类、定位和检测的经典之作，作者是深度学习大神之一————Yann Lecun在纽约大学的团队。OverFeat也是ILSVRC 2013任务3（分类+定位）的冠军得主[10]。

OverFeat的核心思想有三点：

A）区域提名：结合滑动窗口和规则块，即多尺度（multi-scale)的滑动窗口；

B）分类和定位：统一用CNN来做分类和预测边框位置，模型与AlexNet[12] 类似，其中1-5层为特征抽取层，即将图片转换为固定维度的特征向量，6-9 层为分类层(分类任务专用)，不同的任务（分类、定位、检测）公用特征抽 取层（1-5层），只替换6-9层；

C）累积：因为用了滑动窗口，同一个目标对象会有多个位置，也就是多个视角； 因为用了多尺度，同一个目标对象又会有多个大小不一的块。这些不同位置 和不同大小块上的分类置信度会进行累加，从而使得判定更为准确。

OverFeat的关键步骤有四步：

A）利用滑动窗口进行不同尺度的区域提名，然后使用CNN模型对每个区域进 行分类，得到类别和置信度。从图2中可以看出，不同缩放比例时，检测出 来的目标对象数量和种类存在较大差异；

B）利用多尺度滑动窗口来增加检测数量，提升分类效果，如图3所示；

C）用回归模型预测每个对象的位置，从图4中来看，放大比例较大的图片，边 框数量也较多；

D）边框合并。

Overfeat是CNN用来做目标检测的早期工作，主要思想是采用了多尺度滑动窗口来做分类、定位和检测，虽然是多个任务但重用了模型前面几层，这种模型重用的思路也是后来R-CNN系列不断沿用和改进的经典做法。

当然Overfeat也是有不少缺点的，至少速度和效果都有很大改进空间，后面的R-CNN系列在这两方面做了很多提升。

### 2.2、基于区域提名的方法

主要介绍基于区域提名的方法，包括R-CNN、SPP-net、Fast R-CNN、Faster R-CNN、R-FCN。

#### （1）R-CNN

如前面所述，早期的目标检测，大都使用滑动窗口的方式进行窗口提名，这种方式本质是穷举法，R-CNN[1,2,3]采用的是Selective Search。

以下是R-CNN的主要步骤：

A）区域提名：通过Selective Search从原始图片提取2000个左右区域候选 框；

B）区域大小归一化：把所有侯选框缩放成固定大小（原文采用227×227）；

C）特征提取：通过CNN网络，提取特征；

分类与回归：在特征层的基础上添加两个全连接层，再用SVM分类来做识别，用线性回归来微调边框位置与大小，其中每个类别单独训练一个边框回归器。

Overfeat可以看做是R-CNN的一个特殊情况，只需要把Selective Search换成多尺度的滑动窗口，每个类别的边框回归器换成统一的边框回归器，SVM换为多层网络即可。但是Overfeat实际比R-CNN快9倍，这主要得益于卷积相关的共享计算。

事实上，R-CNN有很多缺点：

A）重复计算：R-CNN虽然不再是穷举，但依然有两千个左右的候选框，这些 候选框都需要进行CNN操作，计算量依然很大，其中有不少其实是重复计 算；

B）SVM模型：而且还是线性模型，在标注数据不缺的时候显然不是最好的选 择；

C）训练测试分为多步：区域提名、特征提取、分类、回归都是断开的训练的过 程，中间数据还需要单独保存；

D）训练的空间和时间代价很高：卷积出来的特征需要先存在硬盘上，这些特征 需要几百G的存储空间；

E）慢：前面的缺点最终导致R-CNN出奇的慢，GPU上处理一张图片需要13 秒，CPU上则需要53秒[2]。

当然，R-CNN这次是冲着效果来的，其中ILSVRC 2013数据集上的mAP由Overfeat的24.3%提升到了31.4%，第一次有了质的改变。

（2）SPP-net

SPP-net[4,19]是MSRA何恺明等人提出的，其主要思想是去掉了原始图像上的crop/warp等操作，换成了在卷积特征上的空间金字塔池化层（Spatial Pyramid Pooling，SPP）。为何要引入SPP层 ，主要原因是CNN的全连接层要求输入图片是大小一致的，而实际中的输入图片往往大小不一，如果直接缩放到同一尺寸，很可能有的物体会充满整个图片，而有的物体可能只能占到图片的一角。传统的解决方案是进行不同位置的裁剪，但是这些裁剪技术都可能会导致一些问题出现，比如图7中的crop会导致物体不全，warp导致物体被拉伸后形变严重，SPP就是为了解决这种问题的。SPP对整图提取固定维度的特征，再把图片均分成4份，每份提取相同维度的特征，再把图片均分为16份，以此类推。可以看出，无论图片大小如何，提取出来的维度数据都是一致的，这样就可以统一送至全连接层了。SPP思想在后来的R-CNN模型中也被广泛用到。

参考文献：

[1] R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. ImageNet Large-Scale Visual Recognition Challenge workshop, ICCV, 2013. R-CNN

[2] R. Girshick, J. Donahue, T. Darrell, J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014. R-CNN

[3] R. Girshick, J. Donahue, T. Darrell, J. Malik. Region-Based Convolutional Networks for Accurate Object Detection and Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, May. 2015. R-CNN

[18] R-CNN: Region-based Convolutional Neural Networks: <https://github.com/rbgirshick/rcnn> R-CNN

[4] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. In ECCV. 2014. SPP-net

[19] SPP-net: <https://github.com/ShaoqingRen/SPP_net> SPP-net

[14] Girshick, R. Fast R-CNN. ICCV 2015. Fast R-CNN

[20] Fast R-CNN: <https://github.com/rbgirshick/fast-rcnn> Fast R-CNN

[5] S. Ren, K. He, R. Girshick, J. Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. Advances in Neural Information Processing Systems 28 (NIPS), 2015. Faster R-CNN

[21] Faster R-CNN: <https://github.com/rbgirshick/py-faster-rcnn> Faster R-CNN

[16] R-FCN: Object Detection via Region-based Fully Convolutional Networks. Jifeng Dai, Yi Li, Kaiming He, and Jian Sun. Conference on Neural Information Processing Systems (NIPS), 2016. R-FCN

[24] R-FCN: <https://github.com/daijifeng001/r-fcn> R-FCN

[15] Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: CVPR. (2016) YOLO

[22] YOLO: <http://pjreddie.com/darknet/yolo/> YOLO

[17] Liu W, Anguelov D, Erhan D, et al. SSD: Single Shot MultiBox Detector[J]. arXiv preprint arXiv:1512.02325, 2015. SSD

[23] SSD: <https://github.com/weiliu89/caffe/tree/ssd> SSD

[6] C. Szegedy, A. Toshev, D. Erhan. Deep Neural Networks for Object Detection. Advances in Neural Information Processing Systems 26 (NIPS), 2013.

[7] P. Sermanet, D. Eigen, X.Zhang, M. Mathieu, R. Fergus, and Y. LeCun. OverFeat: Integrated recognition, localization and detection using convolutional networks. In ICLR, 2014. OverFeat

[9] OverFeat source code:

<http://cilvr.nyu.edu/doku.php?id=software:OverFeat:start> OverFeat

[8] J.R. Uijlings, K.E. vandeSande, T. Gevers, and A.W. Smeulders. Selective search for object recognition. IJCV, 2013. 选择性搜索

[12] J. Hosang, R. Benenson, P. Dolla ́r, and B. Schiele. What makes for effective detection proposals? TPAMI, 2015. 区域提名变种

[13] Felzenszwalb P F, Girshick R B, McAllester D, et al. Object detection with discriminatively trained part-based models[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2010, 32(9): 1627-1645.

DPM 多尺度形变部件模型

[10] ILSVRC 2013 results:

http://www.image-net.org/challenges/LSVRC/2013/results.php

[11] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural net- works. In NIPS, 2012.

[25] Detection Results: VOC2012 Competition – “comp4” (train on own data): http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=4