2018小目标检测文章总结

作者：wq604887956@CSDN
原文：https://blog.csdn.net/wq604887956/article/details/83053927

介绍

在现有的目标检测的文献中，大多数是针对通用的目标来进行检测，如经典的单阶段方法yolo和ssd，两阶段方法faster-rcnn等，这些方法主要是针对通用目标数据集来设计的解决方案，因此对于图像中的小目标来说，检测效果不是很理想。因此就有大神提出针对小目标检测的一些方法，这些方法是建立在现有的目标检测基础之上提出的一些改进或者优化。**接下来主要对存在的优秀的小目标检测算法进行简单介绍。**

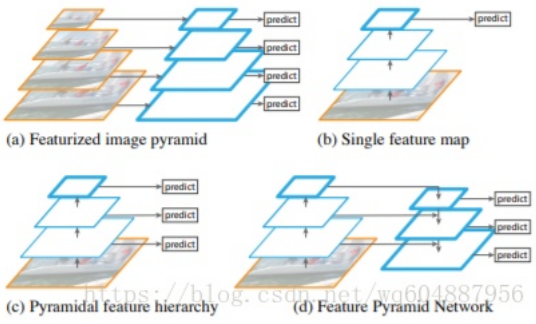
小目标的介绍：有两种定义方式，一种是相对尺寸大小，如目标尺寸的长宽是原图像尺寸的0.1，即可认为是小目标，另外一种是绝对尺寸的定义，即尺寸小于32*32像素的目标即可认为是小目标。

FPN（feature pyramid networks)

论文：feature pyramid networks for object detection
论文链接：https://arxiv.org/abs/1612.03144

这是cvpr2017年出的一篇文章，该文主要是针对通用目标的检测方法，但是在小目标检测中起到了关键作用，以至于之后的很多小目标检测方法都用到了类似的该方法，如faster-rcnn+fpn，yolo3中的特征融合。

该文的主要思想：在fpn之前目标检测的大多数方法都是和分类一样，使用顶层的特征来进行处理。但是这种方法只是用到了高层的语义信息，但是位置信息却没有得到，尤其在检测目标的过程中，位置信息是特别重要的，而位置信息又是主要在网络的低层（换种说法：低层的特征语义信息比较少，但是目标位置准确，高层的特征语义信息比较丰富，但是目标的位置粗略）。因此在这片文章中采用了多尺度特征融合的方式，采用不同特征层特征融合之后的结果来做预测。



该图非常形象的说明了fpn和其他特征融合方式的区别。(a)图是典型的图像金字塔形象，该方法主要是将图像生成不同的尺寸，在每一个尺寸上生成对应的特征图，再在对应的特征图上做相应的预测。这种方法所需要占用的内存和时间比较大，因此没有多少算法使用该方法。在(b)中，在最经典的目标检测方法，只在最后一层的特征图上做预测，常见的rcnn,faster-rcnn都是这种方法。图(c)中是使用了多层的特征图，每一个特征图来做一个新的预测，典型的是ssd中使用的方法。(d)中即使使用fpn的方法，它在图(c)的基础上得到每一层的特征图，之后采用自顶向下的方法将小的特征图上采样之后与下一个特征图融合，融合之后再做预测，依次如此，即可得到多个预测结果。

该方法与faster-rcnn，ssd结合，通过融合高层的语义信息和低层的位置信息，预测在不同的特征图上进行。在测试结果上的结果是令人喜悦的，尤其在小目标检测的提升是比较明显的。

An Analysis of Scale Invariance in Object Detection – SNIP

论文链接：https://arxiv.org/abs/1805.09300
代码链接：https://github.com/mahyarnajibi/SNIPER

这是cvpr2018的一篇文章，主要是针对目标检测中的多尺寸问题提出的方法。

首先有一个中心思想：就是要让输入的分布尽可能地接近模型预训练地分布。作者通过实验1得到通过在Imagenet网络上得到的预训练结果不利于检测小目标，因此提出一种方法，先用ImageNet做预训练，之后使用原图上采样得到的图像来做微调，使用微调的模型来预测原图经过上采样的图像。该方法的提升效果比较显著。

作者的第二个实验想说明将原图放大到什么大小有助于检测小目标。

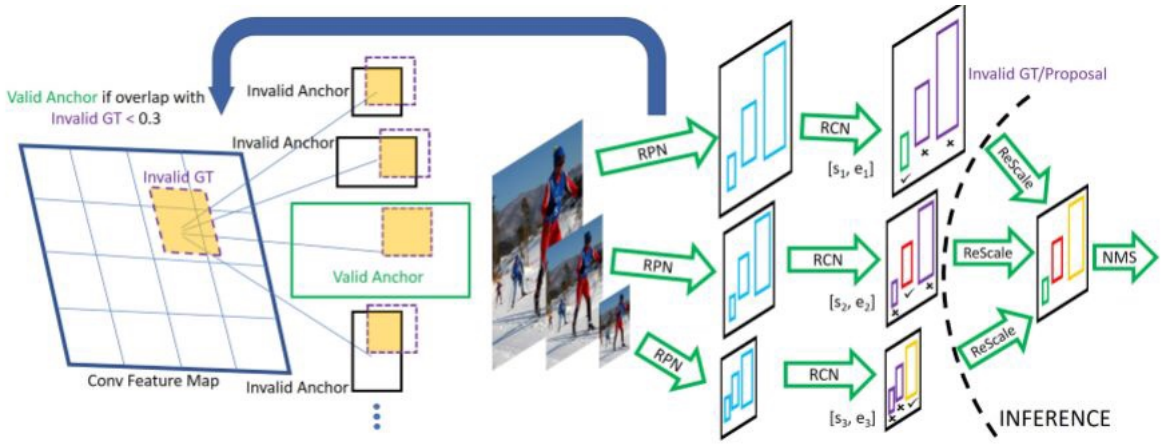


Figure 6. SNIP training and inference is shown. Invalid RoIs which fall outside the specified range at each scale are shown in purple. These are discarded during training and inference. Each batch during training consists of images sampled from a particular scale. Invalid GT boxes are used to invalidate anchors in RPN. Detections from each scale are rescaled and combined using NMS.

在上图中，对rfcn提出的改进主要有两个地方，一是多尺寸图像输入，针对不同大小的输入，在经过rpn网络时需要判断valid GT和invalid GT，以及valid anchor和invalid anchor，通过这一分类，使得得到的预选框更加的准确；二是在rcn阶段，根据预选框的大小，只选取在一定范围内的预选框，最后使用soft-nms来得到最终结果。

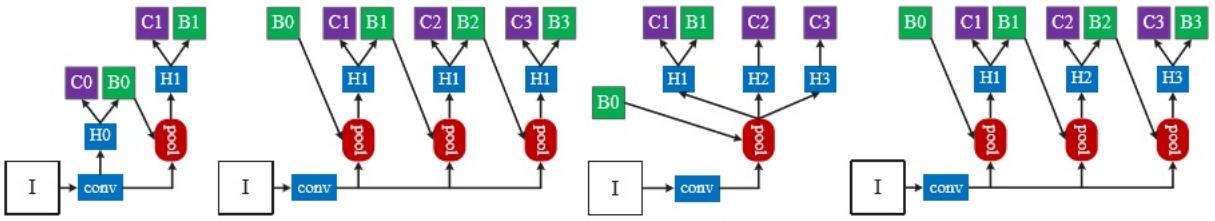
该文的实验结果出其的好，并且在小目标检测的结果上也有着显著的提升。

Cascade R-CNN: Delving into High Quality Object Detection

论文链接：https://arxiv.org/abs/1712.00726
代码链接：https://github.com/zhaoweicai/cascade-rcnn

这是cvpr2018上的最新文章，这篇文章和上一篇文章都是通过许多实验来探索对小目标检测的提升效果。本文主要是针对目标检测中的两阶段方法中的阈值来做文章。我们都知道检测种的IoU阈值对于样本的选取是至关重要的，如果IoU阈值过高，会导致正样本质量很高，但是数量会很少，会出现样本比例不平衡的影响；如果IoU阈值较低，样本的数量就会增加，但是样本的质量也会下降。因此如何平衡这一关系，即如何选取好的IoU，对于检测结果来说很重要。

作者为了解决这一问题，提出了多阶段的结构，通过不断提高IoU的阈值，使得在保证样本数量的同时，也能使得样本的质量不下降，最后训练出高质量的检测器，如文章题目所言。



(a) Faster R-CNN (b) Iterative BBox at inference (c) Integral Loss (d) Cascade R-CNN

Figure 3. The architectures of different frameworks. "I" is input image, "conv" backbone convolutions, "pool" region-wise feature extraction, "H" network head, "B" bounding box, and "C" classification. "B0" is proposals in all architectures.

在上图中主要是介绍了几种不同的类似于级联方法的比较。(a)图中是常规的fatser-rcnn方法；(b)图是在推理中使用迭代的方式不断地改进推理结果，但是在检测时使用的是相同的检测头，这会出现两个问题，单一阈值这对于检测高质量的目标并不有利和每次检测出来的分布都是差别较大的；(c)图中虽然使用了多个检测头，但是预选框只有一个分布，以及只做了一个边界框回归，这显然是不行的；(d)图即是本文提出的方法，融合了前面的几个方法，在得到B0之后，通过每一个不同的检测头，该检测头的IoU阈值只逐渐提升的，在通过检测头之后，得到的B1会和前一个B0结合生成新的预选框，用作下一个检测头，依次处理，就可以得到最终的结果。

总结起来就是：

- cascaded regression不断改变了proposal的分布，并且通过调整阈值的方式重采样
- cascaded在train和inference时都会使用，并没有偏差问题
- cascaded重采样后的每个检测器，都对重采样后的样本是最优的，没有mismatch问题

该方法最终测试的结果是优秀的，在coco数据集上取得了最好的检测准确度。并且在小目标的检测结果上也有着显著的提升。

Learning Efficient Single-stage Pedestrian Detectors by Asymptotic Localization Fitting(ALFNet)

论文地址：
http://openaccess.thecvf.com/content_ECCV_2018/html/Wei_Liu_Learning_Efficient_Single-stage_ECCV_2018_paper.html

论文代码：https://github.com/VideoObjectSearch/ALFNet

这是ECCV2018的一篇文章，这篇文章的主旨是向高效行人检测迈进。

不同于上述的几种两阶段的方法，该方法是基于单阶段的一种方法，但是该方法又和上述的级联网络思想相同，也用到了级联网络，作者称这一模块为渐进定位拟合模块。该文的主要研究对象不是通用的目标，而是指定的目标对象——行人。

文章的主要思想是利用不断提升的IoU阈值训练多个定位模块，来达到提升定位精度的目的。

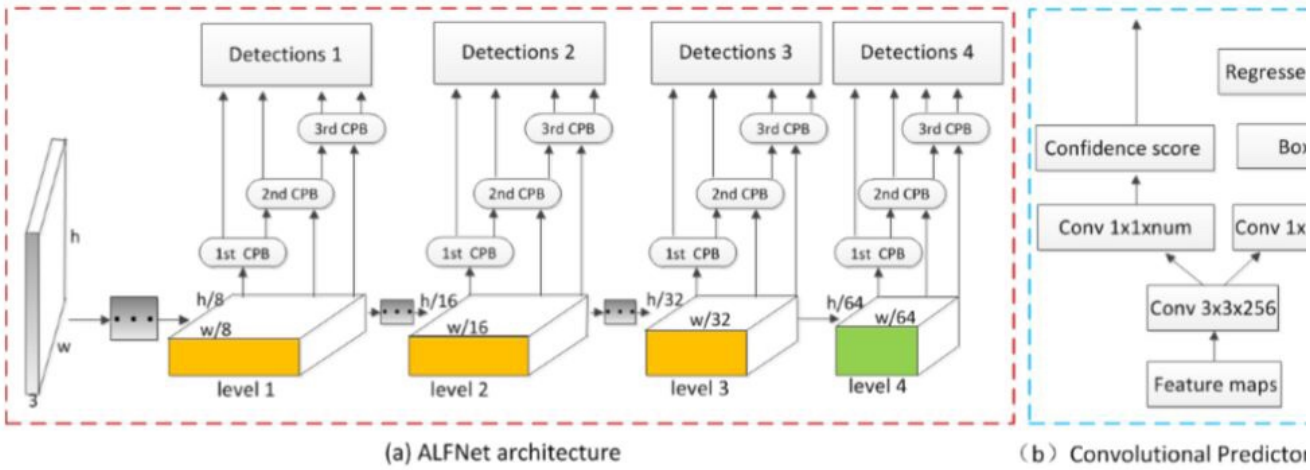


Fig. 3. (a) ALFNet architecture, which is constructed by four levels of featu for detecting objects with different sizes, where the first three blocks in yel from the backbone network, and the green one is an added convolutional laye end of the truncated backbone network. (b) Convolutional Predictor Block which is attached to each level of feature maps to translate default anchor b corresponding detection results.

文章主要是以ssd为框架，在此基础上提出的一些改进。作者以ResNet-50为基础网络为例，给出了ALFNet检测器的网络架构，如上图所示所示，选用ResNet-50的第3、4、5个stage的最后一层（黄色部分）以及新接的一层（绿色部分）作为多尺度特征图（分别较原始图像降采样了8、16、32和64倍），在这些特征图上添加ALF模块，也即堆叠多个CPB（上图(b)）。

该方法在行人检测的数据集上效果显著，达到最先进的结果，并且速度也是很快的。

该文和发表于CVPR2018的Cascade R-CNN有共通之处，一是训练采用提升的IoU阈值能够或者更好的检测性能，二是并非堆叠越多步数检测器性能越好，级联一定步数时检测器性能会趋向饱和。区别在于：前者基于SSD单阶段的检测框架，目的在于提升定位精度的同时保证算法的速度优势，后者基于Faster R-CNN两阶段的检测框架，目的在于提升检测器的accuracy。尽管采用的检测框架不同，但均证明了多步预测是提升检测器性能的一个非常行之有效的方式。

Perceptual Generative Adversarial Networks for Small Object Detection

论文链接：https://arxiv.org/abs/1706.05274v2
论文代码：无

这篇文章是cvpr2017的一篇文章，该文主要是针对小目标而设计的。（自我感觉上述几篇主要还是解决通用的目标检测，顺带解决小目标检测的方法。）

该文先是提出小目标检测的一般方法：提高输入图像的分辨率，会增加运算量；多尺度特征表示，结果不可控。

该文主要内容：论文使用感知生成式对抗网络（Perceptual GAN）提高小物体检测率，generator将小物体的poor表示转换成super-resolved的表示，discriminator与generator以竞争的方式分辨特征。Perceptual GAN挖掘不同尺度物体间的结构关联，提高小物体的特征表示，使之与大物体类似。包含两个子网络，生成网络和感知分辨网络。生成网络是一

个深度残差特征生成模型，通过引入低层精细粒度的特征将原始的较差的特征转换为高分变形的特征。分辨网络一方面分辨小物体生成的高分辨率特征与真实大物体特征，另一方面使用感知损失提升检测率。（转载自：https://blog.csdn.net/cv_family_z/article/details/77332721）(由于没有源码，所以文章没有仔细阅读)

DetNet: A Backbone network for Object Detection

论文链接：https://arxiv.org/abs/1804.06215?context=cs
论文代码：https://github.com/tsing-cv/DetNet

这篇文章是eccv2018上的一篇文章，文章的主要贡献是建立一个专门用于检测的骨干网络。它先总结了传统网络的三个缺点：网络阶段数不同(预训练不一致)，大目标的回归弱，小目标很难发现。为了解决这种问题，会出现两个关键挑战：需要保证特征图的尺寸足够大，但是所占用的内存也会很大；减少下采样的数目，但是感受野的大小无法保证。

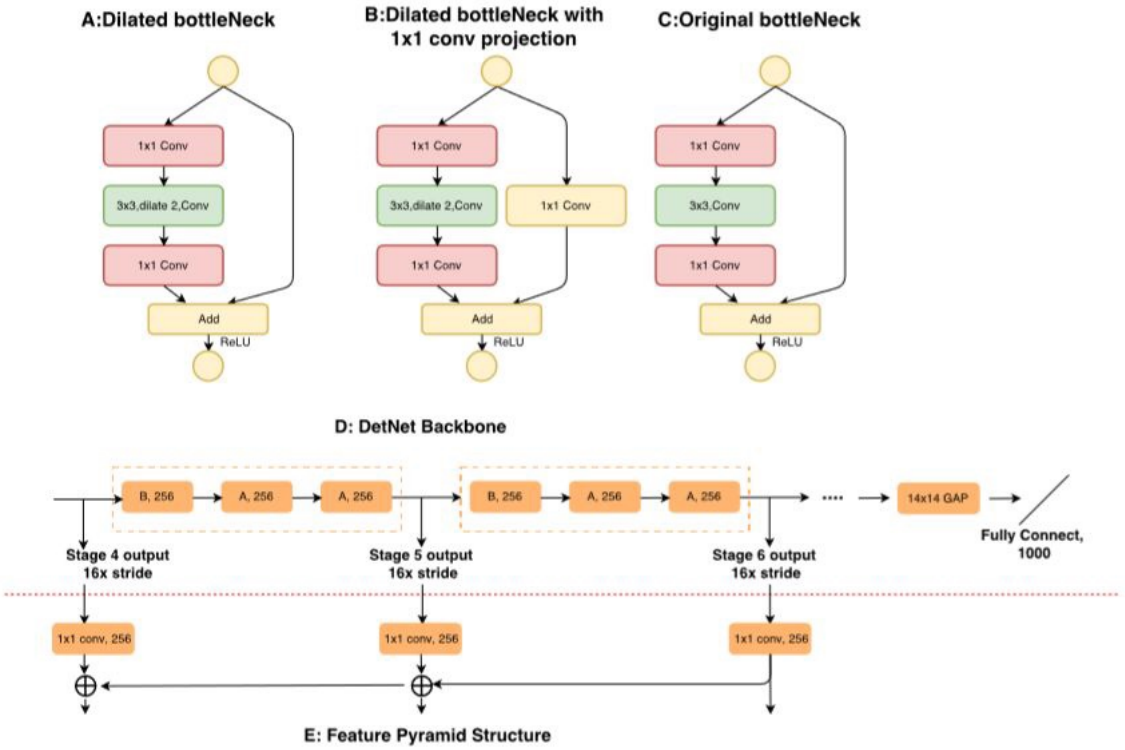


Fig. 2. Detail structure of DetNet (D) and DetNet based Feature Pyramid Net-Work (E). Different bottleneck block used in DetNet is illustrated in (A, B). Original bottleneck is illustrated in (C). DetNet follows the same design as ResNet before stage 4, while keeps spatial size after stage 4 (e.g. stage 5 and 6).

上图是detnet的网络结构，是在resnet50的基础上提出的改进。与resnet50相比，保留了resnet50的前三个阶段，增加了第六个阶段，修改了第4，5阶段。

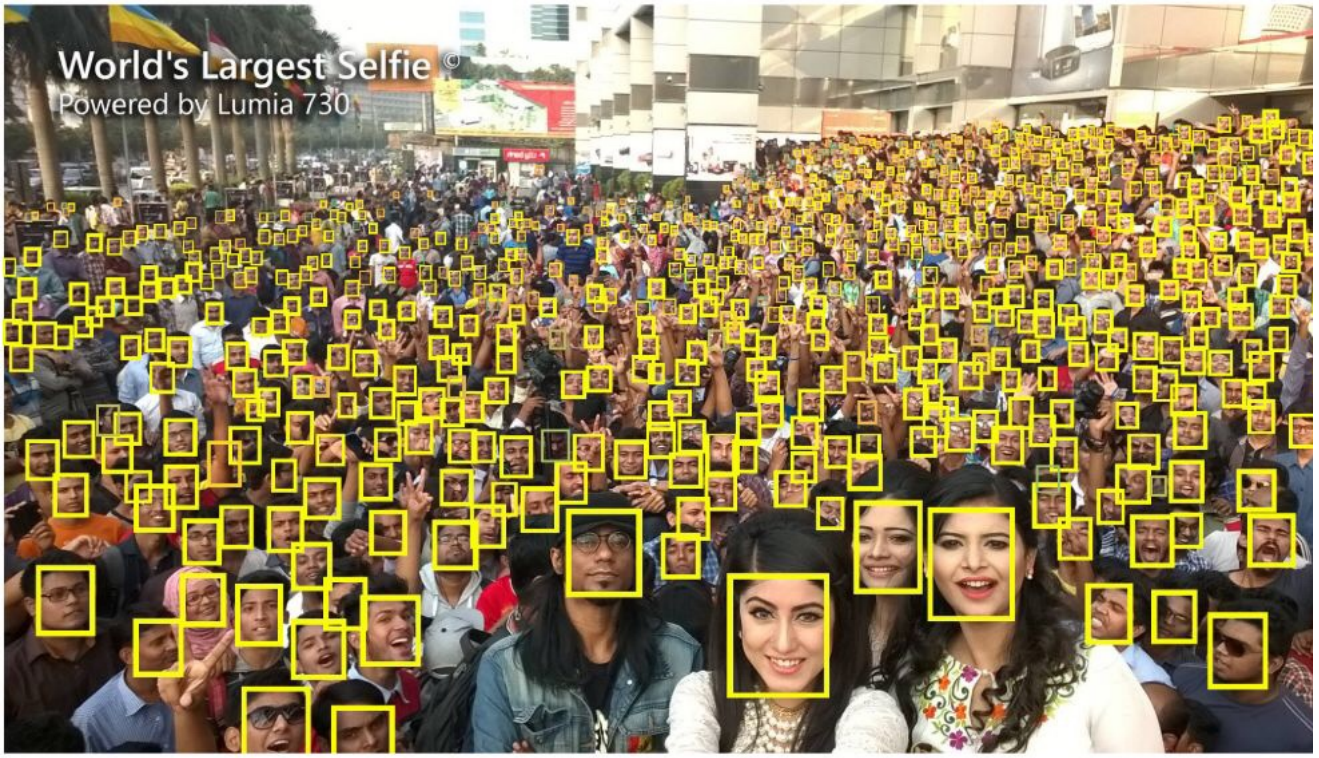
1. 原来的resnet50的阶段5的特征图是原图的1/32，但是在detnet中，阶段4，5，6的特征图都是原图的1/16。
2. 使用了dilated bottleneck，分成A和B两种，在上图中的A和B所示，D图中显示了完整的结构。
3. 使用dilated 的技术是为了增加感受野(即保证特征图的大小)，但是为了考虑计算量的问题，在第5，6阶段保持了相同的通道数。

这种网络结构设计在实验中经过测试，得到的结果相比较于resnet50，提升了几个百分点，说明这种方法对于检测网络还是有效的。同时在检测小目标上的提升也是比较显著的。

Finding Tiny Faces

论文链接：https://www.cs.cmu.edu/~peiyunh/tiny/
论文代码：https://github.com/peiyunh/tiny

这是cvpr2017的一篇寻找小人脸的文章，该文的结果在人脸的数据集上得到的结果是爆炸性的，看下图就知道了：



该文主要对小人脸的识别上，主要从三个方面来考虑：尺度不变性，图像分辨率和上下文推理。

在这片文章中也得到了三个结论：

1. 由于小目标的信息太少，因此需要使用小目标的周围信息来扩充小目标的感受野，即利用人脸周围的头发，耳朵，肩膀等信息；
2. 对于小人脸，不是越大的感受野得到的结果越好，对于大人脸，更大的感受野对提升效果不明显；
3. 不同大小的特征图的融合是检测小人脸的关键，但是对于大人脸效果不明显。

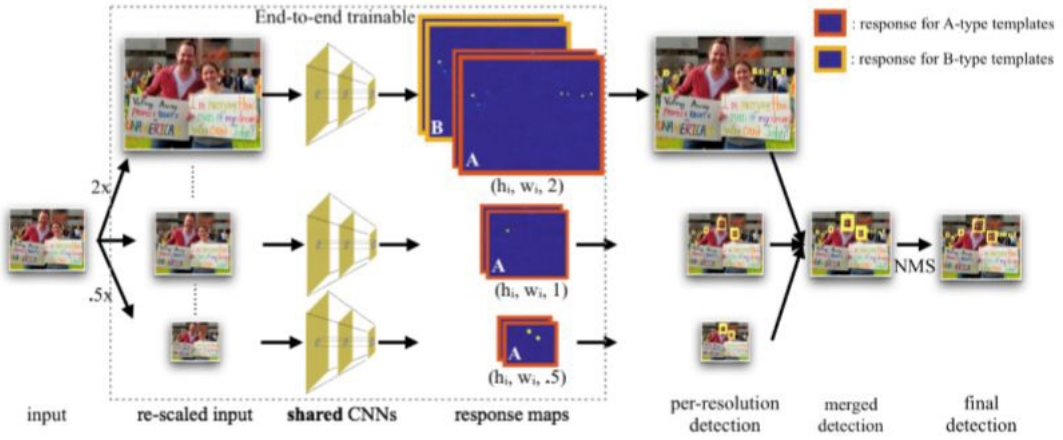


Figure 10: Overview of our detection pipeline. Starting with an input image, we first create a coarse image pyramid (includ 2X interpolation). We then feed the scaled input into a CNN to predict template responses (for both detection and regressi at every resolution. In the end, we apply non-maximum suppression (NMS) at the original resolution to get the final detect results. The dotted box represents the end-to-end trainable part. We run A-type templates (tuned for 40-140px tall faces) the coarse image pyramid (including 2X interpolation), while only run B-type (tuned for less than 20px tall faces) templa on only 2X interpolated images (Fig. 9)

上图中是整个检测的结构，对于一张输入图片，先建立一个粗略的图像金字塔，之后使用CNN在每一个像素上预测结果，最后使用非最大抑制方法来得到结果。主要思想是通过在不同大小的模板上分别寻找对应大小的目标，即在大的模板上寻找小的目标，在小的模板上寻找大的目标。

该方法在小人脸检测数据集的结果是惊人的，但是可能不太适合于通用的小目标检测，却可以作为一种思路的参考值得借鉴。

以上是对已经阅读过的文章做的一些简单的总结，自我感觉在小目标检测这一方面文章的数量不是很多，并且大多数也不是专门解决小目标检测的文章，因此呢，在这一方面机遇还有是许多的，但另一方面挑战也是同时存在的。前途漫漫，不屈不饶！

*延伸阅读

- **港中文开源视频动作分析库MMAction，目标检测库算法大更新**
- **解读目标检测新范式：Segmentations is All You Need**
- **Kaggle实战目标检测奇淫技巧合集**

觉得有用麻烦给个在看啦~