**宫颈细胞自动检测**

一、问题描述

宫颈癌是最常见的妇女恶性肿瘤，也是恶性肿瘤中唯一病因明确、可以早期预防和治疗的癌症。细胞学检查是目前最常见的早期宫颈癌筛查技术。由于人口众多，自动化阅片成为满足我国妇女定期宫颈癌筛查需求的必要条件，研发计算机辅助阅片系统迫在眉睫，其中最关键的一个步骤是从宫颈细胞学图像的海量细胞中快速准确地定位异常区域。

项目基于宫颈癌液基薄层细胞检测（TCT）数据，综合运用目标检测、深度学习等方法对宫颈癌细胞学异常鳞状上皮细胞进行定位，辅助医生进行诊断。。

本项目只进行二分类（阳性和非阳性），标注出阳性细胞。

二、参考阅读

1. Faster R-CNN：《Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks》

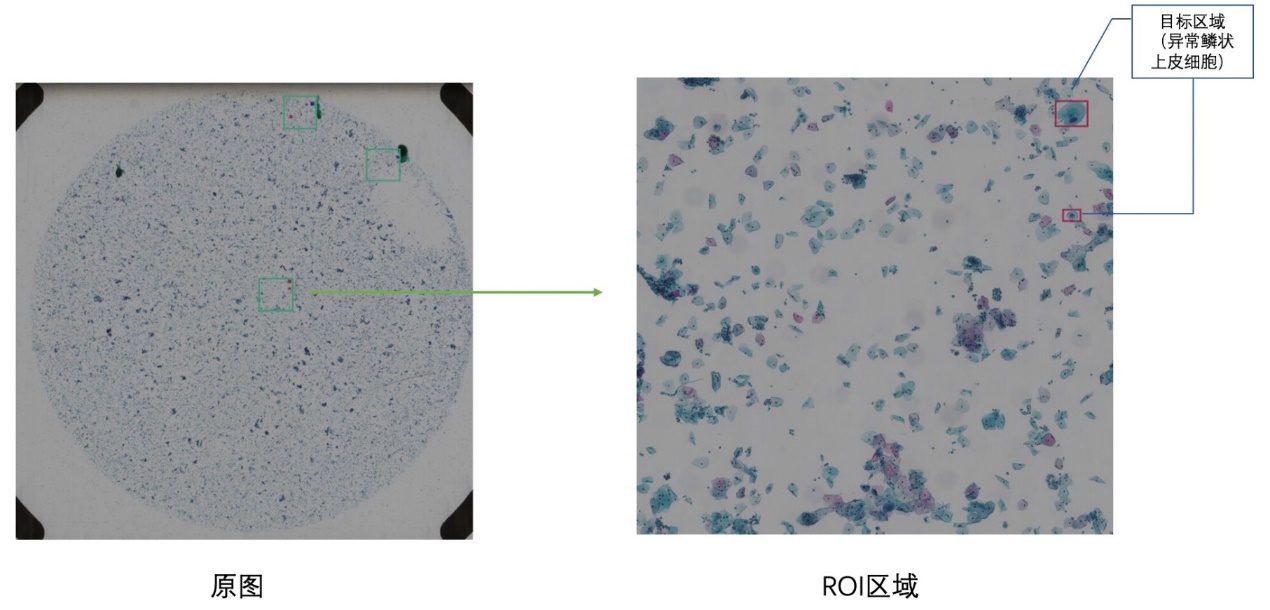
2. HRNet：《Deep High-Resolution Representation Learning for Human Pose Estimation》

3. Cascade R-CNN:《Cascade R-CNN: Delving into High Quality Object Detection》

以上论文均可从ArXiv网站下载

三、数据集合

数据集（来源于阿里天池宫颈癌挑战赛网站）包括宫颈细胞学整体切片成像（WSI）图片800张，其中阳性图片500张，阴性图片300张。每张数据在20倍数字扫描仪下获取，大小300～400M。WSI阳性图片会提供多个高分辨率ROI区域，ROI区域中标注异常鳞状上皮细胞位置，阴性图片中不包含异常鳞状上皮细胞，无标注。ROI区域的个数共1212块，阳性异常鳞状上皮细胞标注的个数共22673个。



JSON标注文件内容是一个list文件，记录了每个ROI区域的位置和异常鳞状上皮细胞的位置坐标（细胞所在矩形框的左上角坐标和矩形宽高）。类别roi表示感兴趣区域，pos表示异常鳞状上皮细胞。json标注文件示例如下：

|  |
| --- |
| [{"x": 33842, "y": 31905, "w": 101, "h": 106, "class": "pos"},  {"x": 31755, "y": 31016, "w": 4728, "h": 3696, "class": "roi"},  {"x": 32770, "y": 34121, "w": 84, "h": 71, "class": "pos"},  {"x": 13991, "y": 38929, "w": 131, "h": 115, "class": "pos"},  {"x": 9598, "y": 35063, "w": 5247, "h": 5407, "class": "roi"},  {"x": 25030, "y": 40115, "w": 250, "h": 173, "class": "pos"}] |

数据集下载地址：

kfbreader.zip 59.8MB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/kfbreader.zip

labels.zip 210.29KB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/labels.zip

neg\_0.zip 10.74GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_0.zip

neg\_1.zip 10.24GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_1.zip

neg\_2.zip 11.55GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_2.zip

neg\_3.zip 11.11GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_3.zip

neg\_4.zip 11.25GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_4.zip

neg\_5.zip 11.57GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/neg\_5.zip

pos\_0.zip 13.01GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_0.zip

pos\_1.zip 12.68GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_1.zip

pos\_2.zip 12.85GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_2.zip

pos\_3.zip 12.74GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_3.zip

pos\_4.zip 12.81GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_4.zip

pos\_5.zip 12.84GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_5.zip

pos\_6.zip 13.36GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_6.zip

pos\_7.zip 13.41GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_7.zip

pos\_8.zip 12.97GB http://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_8.zip

pos\_9.zip 14.48GB http ://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/pos\_9.zip

test\_0.zip 12.03GB http ://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/test\_0.zip

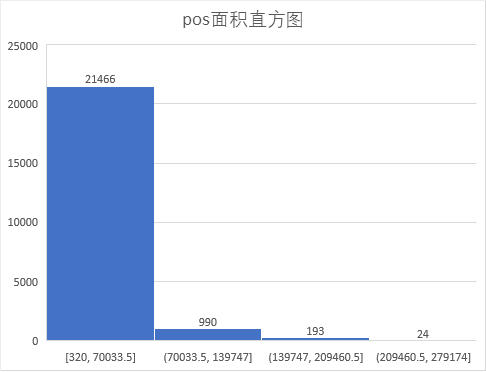
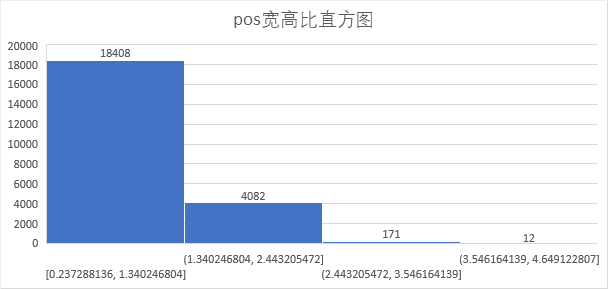
test\_1.zip 12.5GB http ://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/test\_1.zip

test\_2.zip 12.64GB http ://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/test\_2.zip

test\_3.zip 11.4GB http ://tianchi-competition.oss-cn-hangzhou.aliyuncs.com/231757/test\_3.zip

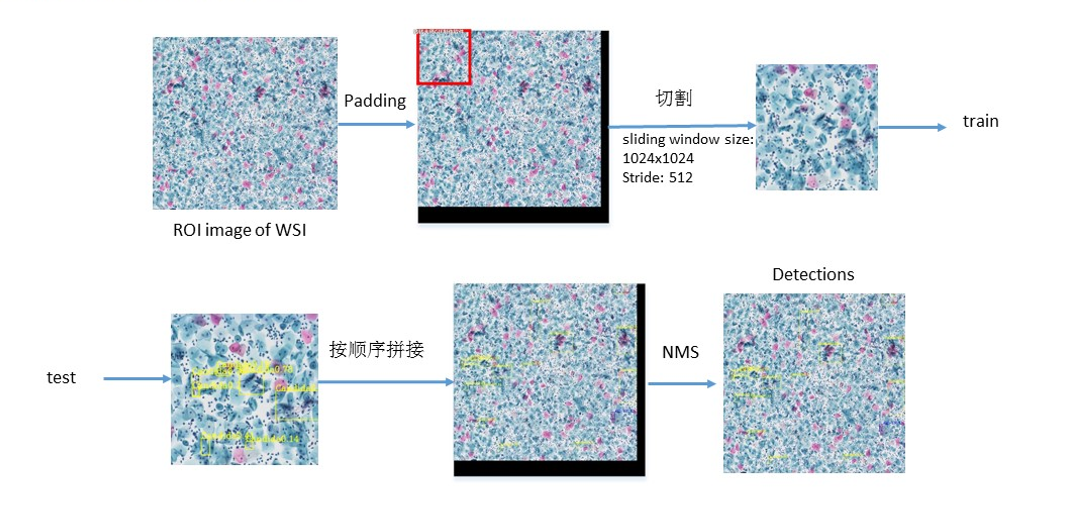
四、建议方法

项目难点主要在于异常鳞状细胞尺度分布范围大，不均衡，目标大小不一，小目标异常细胞占比很大，但也存在较大异常细胞。统计数据集图片尺度如下图pos面积直方图和pos宽高比直方图，分析可知绝大部分pos的面积很小(相对于ROI)。

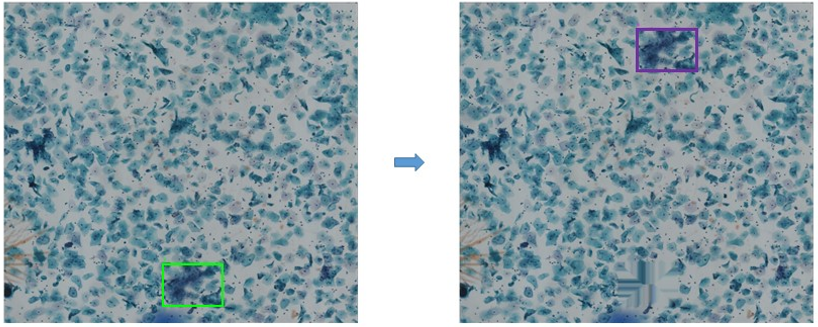
 

1. 预处理

针对ROI图片分辨率大的问题，采用裁图方法，用1024x1024，步长（Stride）为512的滑动窗口（Sliding Window）裁剪ROI图片并对边缘补0。我们对500张阳性WSI的ROI区域进行裁剪。如下图：

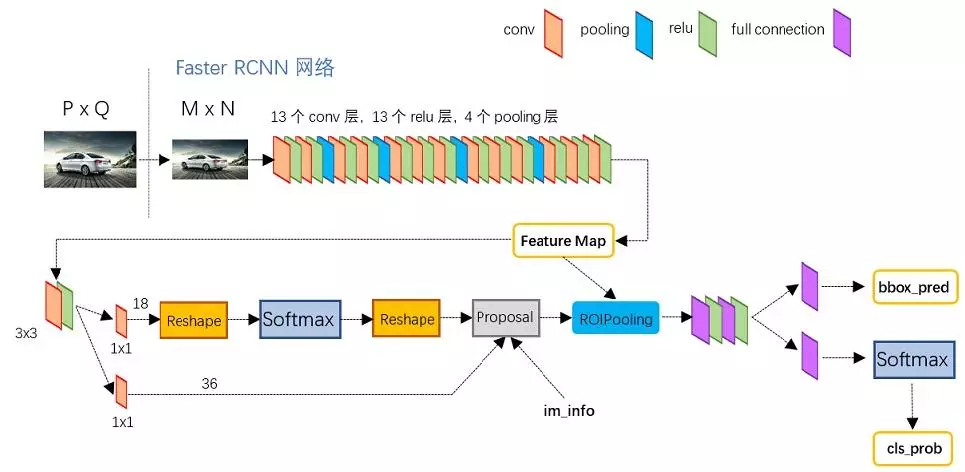


为扩充数据，对裁剪后的图片，做随机移动标注框得到14886张1024x1024尺寸的图像，对图像进行随机平移、旋转、翻转、特征标准化等常用数据增广策略。



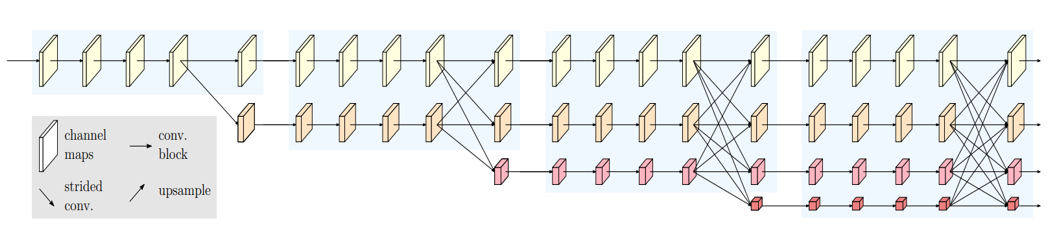
2. 网络

1. 基础版本：采用Faster-RCNN网络

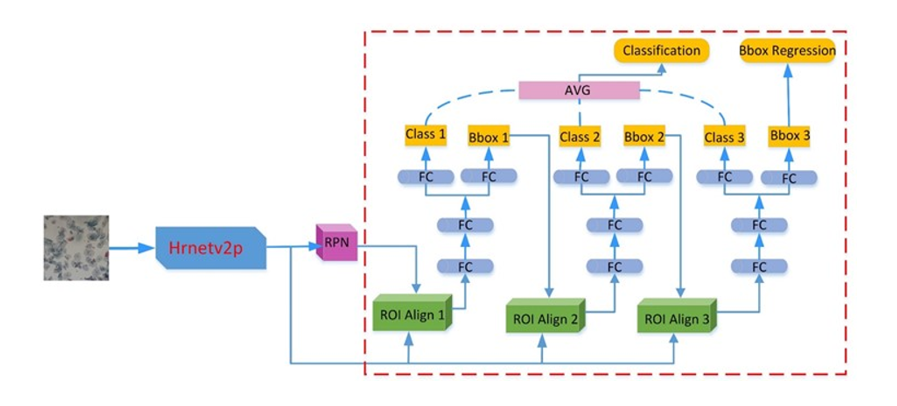


1. 进阶版本：

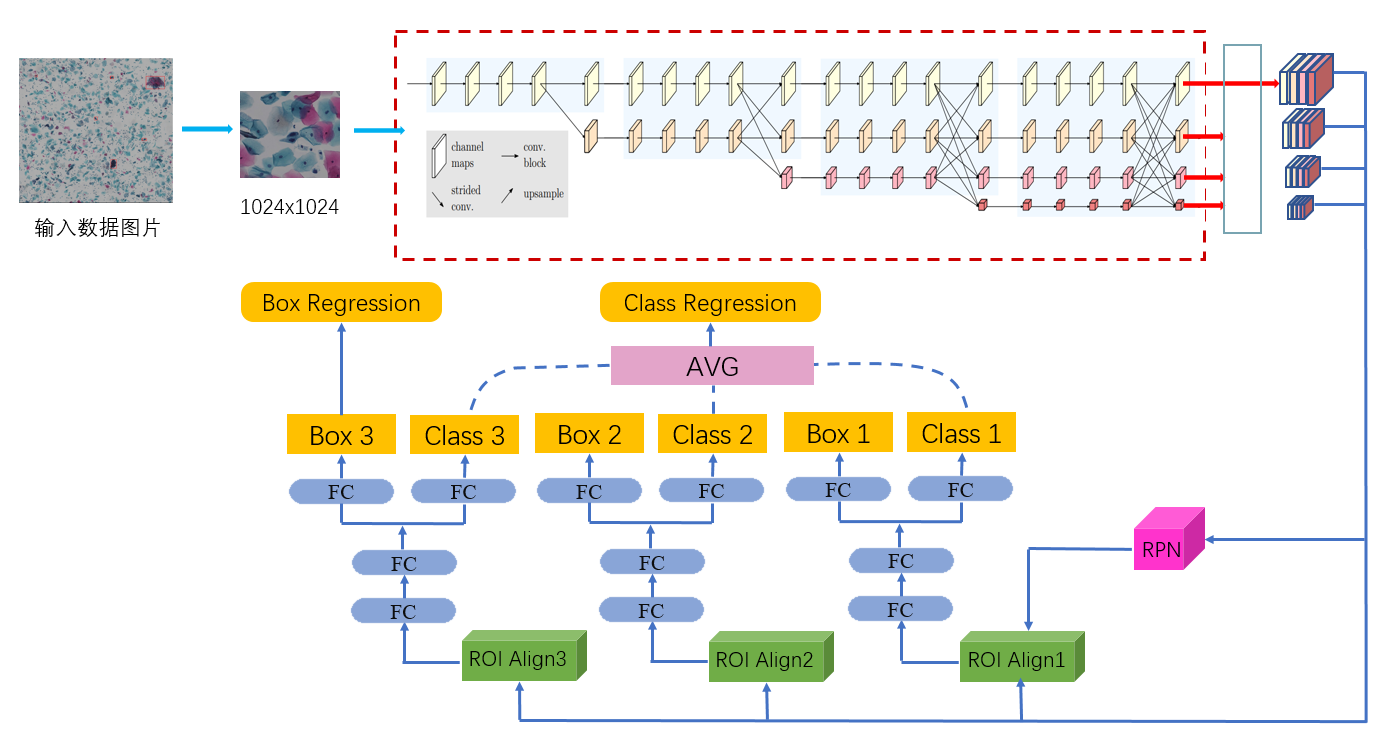
针对小目标异常细胞问题，HRNet通过并行高分辨率和低分辨率网络，能够保持高分辨率的特征，对小目标的识别有更好的效果。如图所示，HRNet以高分辨率子网作为第一阶段始，逐步增加高分辨率到低分辨率的子网，形成更多的阶段，并将多分辨率子网并行连接。



Cascade R-CNN通过并行级联多个检测头，每个检测头采取不同的IOU阈值，逐步提高RPN网络提出的proposal的质量，抑制Fasle Negative样本。



最终网络结构如下：



注：相关解决方案可参照：

https://blog.csdn.net/u011797832/article/details/103909521

五、评估方法

1.指标要求：

mAP（mean Average Precision）达到79%以上

recall达到80%以上

注：采用两个IoU阈值（0.5，0.7）分别来计算AP，再综合平均作为最终的评测结果。

AP计算过程：首先固定一个IoU阈值，计算每个预测框和真实标签的IoU大小，根据阈值判断预测框是否正确。然后在对预测框根据置信度排序，设定不同的置信度阈值得到一系列召回率和精确率值，在不同的召回率下对准确率求平均，即为AP。

Recall（召回率）= TP/ (TP+FN)Recall（召回率）=TP/(TP+FN)

Precision（精确率）= TP/ (TP+FP)Precision（精确率）=TP/(TP+FP)

2.在测试图片上用矩形框框出检查的阳性细胞，如下图所示。

