# 数据爬取报告

**任务目标：**爬取100张狗的图片

**数据源：**百度，必应，谷歌

**爬取方式：**通过爬虫项目Image Downloader爬取，计划爬取100张“狗”的图片，实际爬取88张

**数据观察结果：**

1.部分图片尺寸过小（编号0577, 0883）

2.某些图片背景不是自然场景（编号0939）



**筛选目标：**

筛选出尺寸小于10kb图片，以及一些背景分布异常的图片

**筛选方式：**

1. 初级筛选：

根据图像大小及宽高筛选出尺寸小于10kb图片

筛选出梯度过低的图片

1. 高级筛选：

将图片送入模型resnet50，得到图片特征，利用UMAP或者TSNE降维，根据图片分布，剔除边缘数据，即离群点。

**筛选结果：**

筛选出了小尺寸图片



通过UMAP降维结果，找到了一些“异常数据”，手动移除这些脏数据，可能对训练有好处。

