# 数据爬取与筛选实验报告

## 任务目标

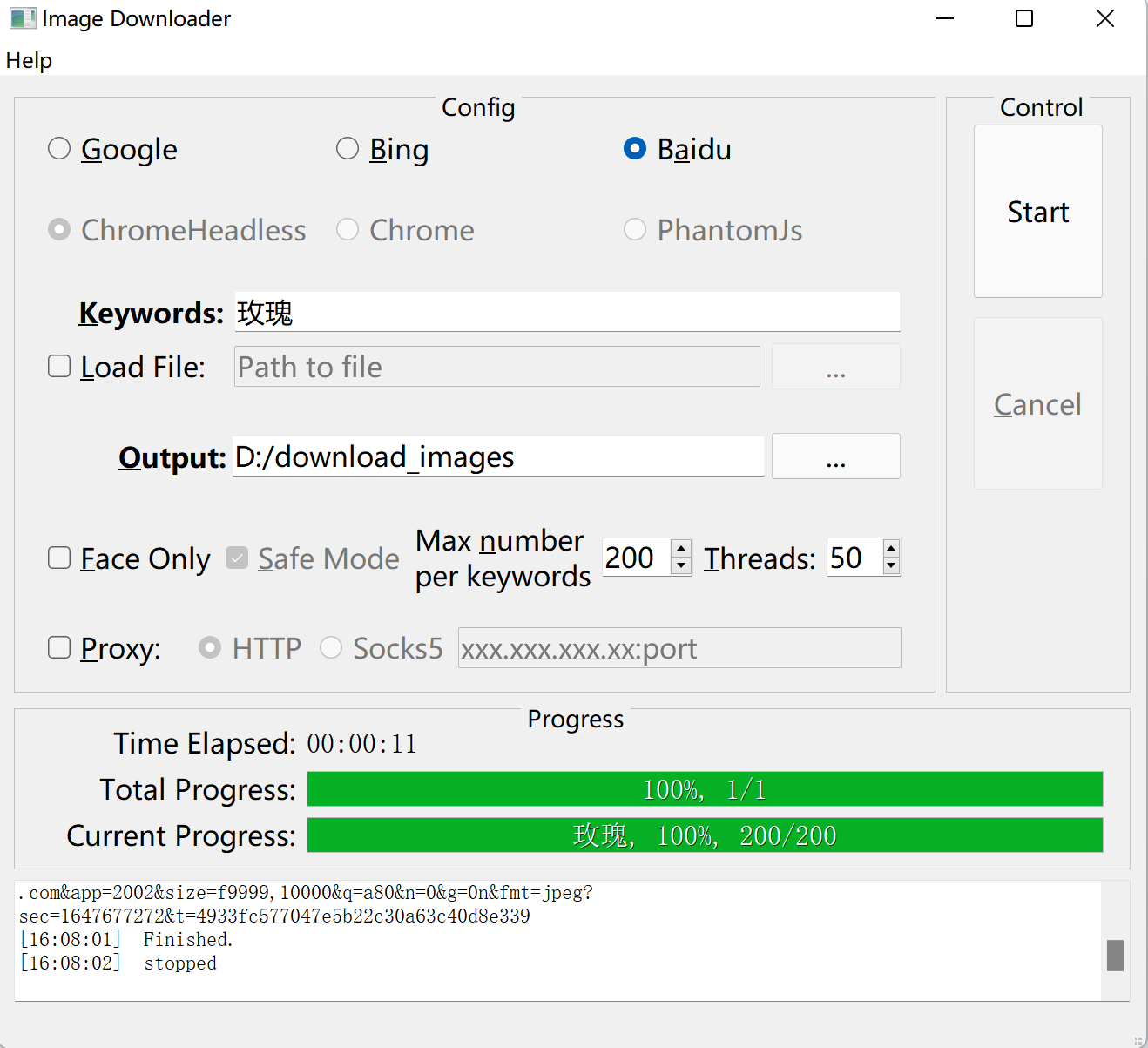
通过网络爬取获取以“玫瑰”为关键字的数据集，图片要求为真实场景下的玫瑰。

## 数据源

以百度为搜索引擎的图片爬取

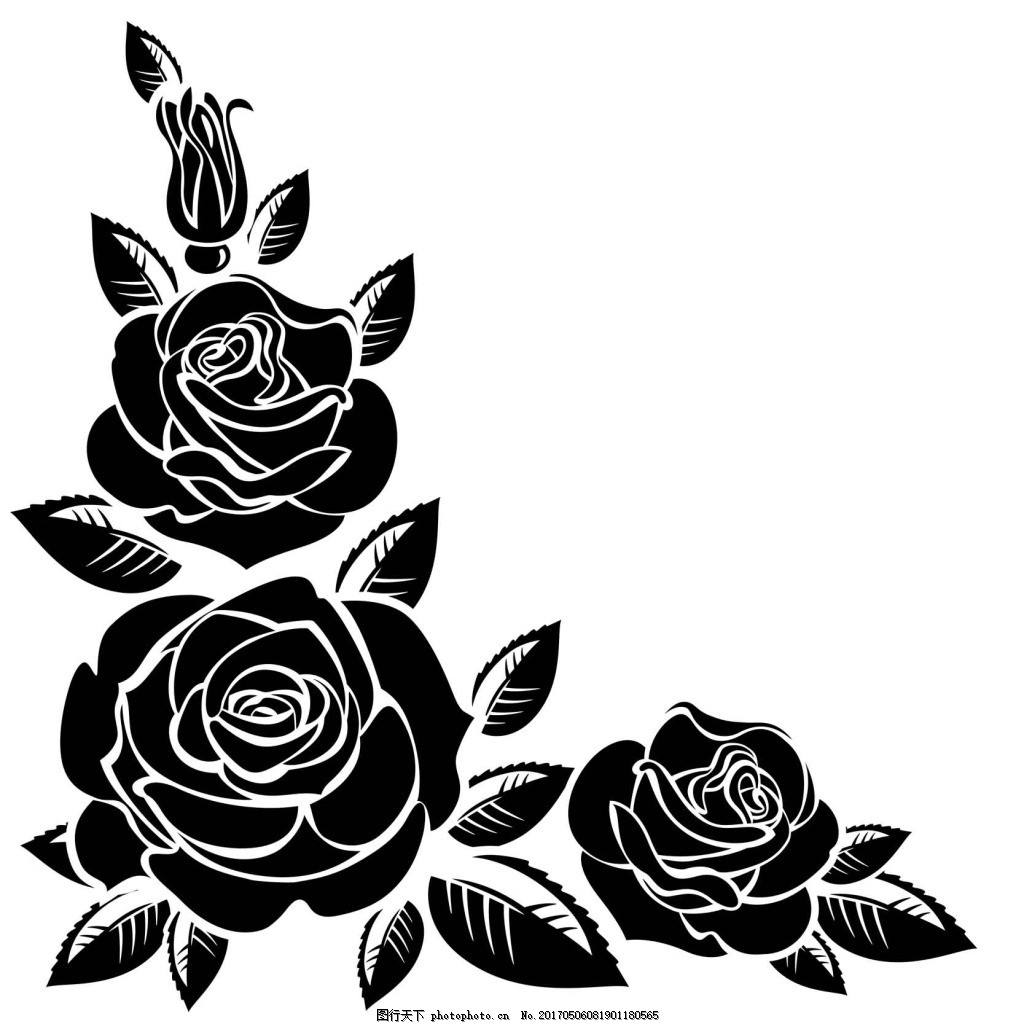
## 爬取方式

通过Github上爬取搜索引擎图片的python项目，运行后在GUI界面设置“玫瑰”关键字，总共试图从百度爬取200张图片。(https://github.com/sczhengyabin/Image-Downloader)



## 数据观察结果

爬取成功的图片有176张，但是并非所有图片中的玫瑰来自现实的场景均在一定量的**无效图片**、**绘画作品**。



真实的玫瑰图片也可分为单枝玫瑰和多枝玫瑰



以及根据玫瑰花色可以将玫瑰分为不同的种类



综上可以根据实际需要选择对玫瑰的图片进行不同目标、不同程度上的筛选，本实验只进行真实玫瑰和非真实玫瑰的筛选。

## 筛选目标

通过筛选获得真实玫瑰图片，忽略花朵数和花朵颜色的区别

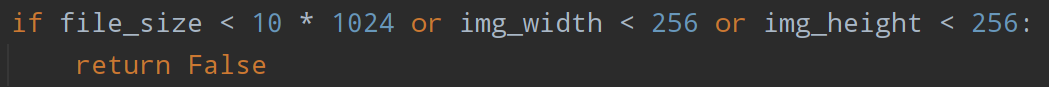
## 筛选方式

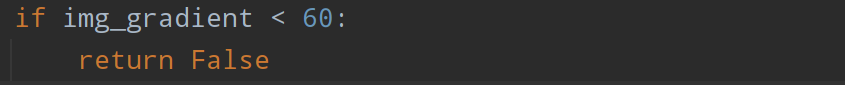
借鉴课堂上学到的两类方法

初步筛选：图片文件大小、图片梯度信息

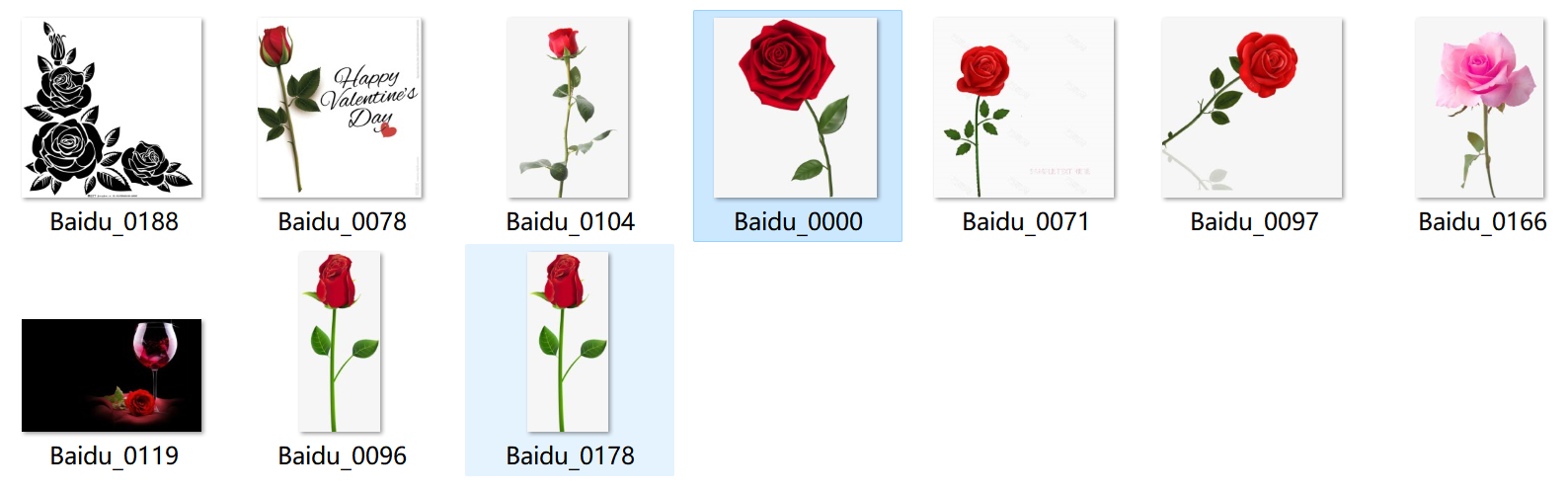
高级筛选：通过预训练的CNN获取图片的特征图，通过特征图的降维，分析降维后的二维分布图对图片进行筛选。

## 筛选结果

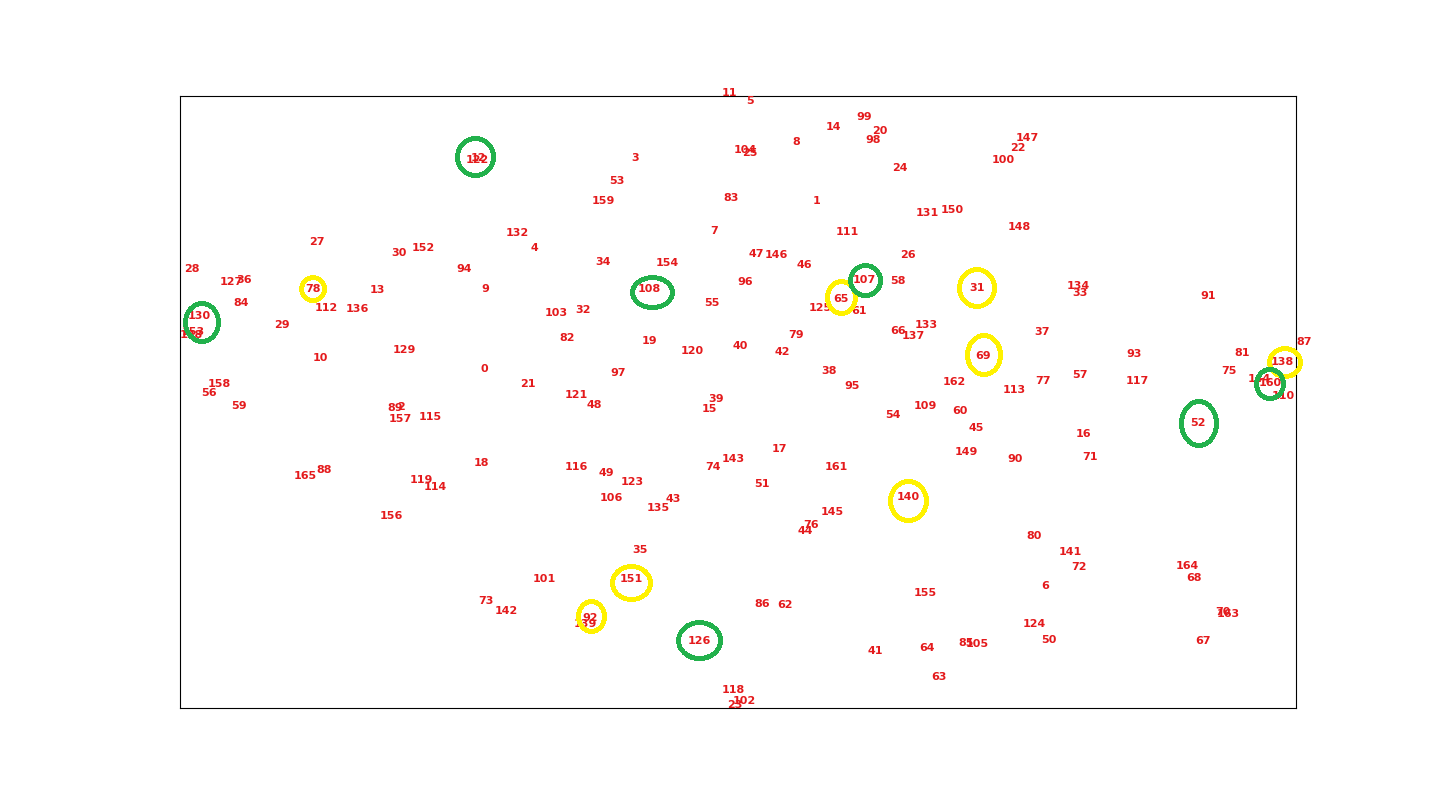
经过分析，初步筛选中的阈值设置如下：



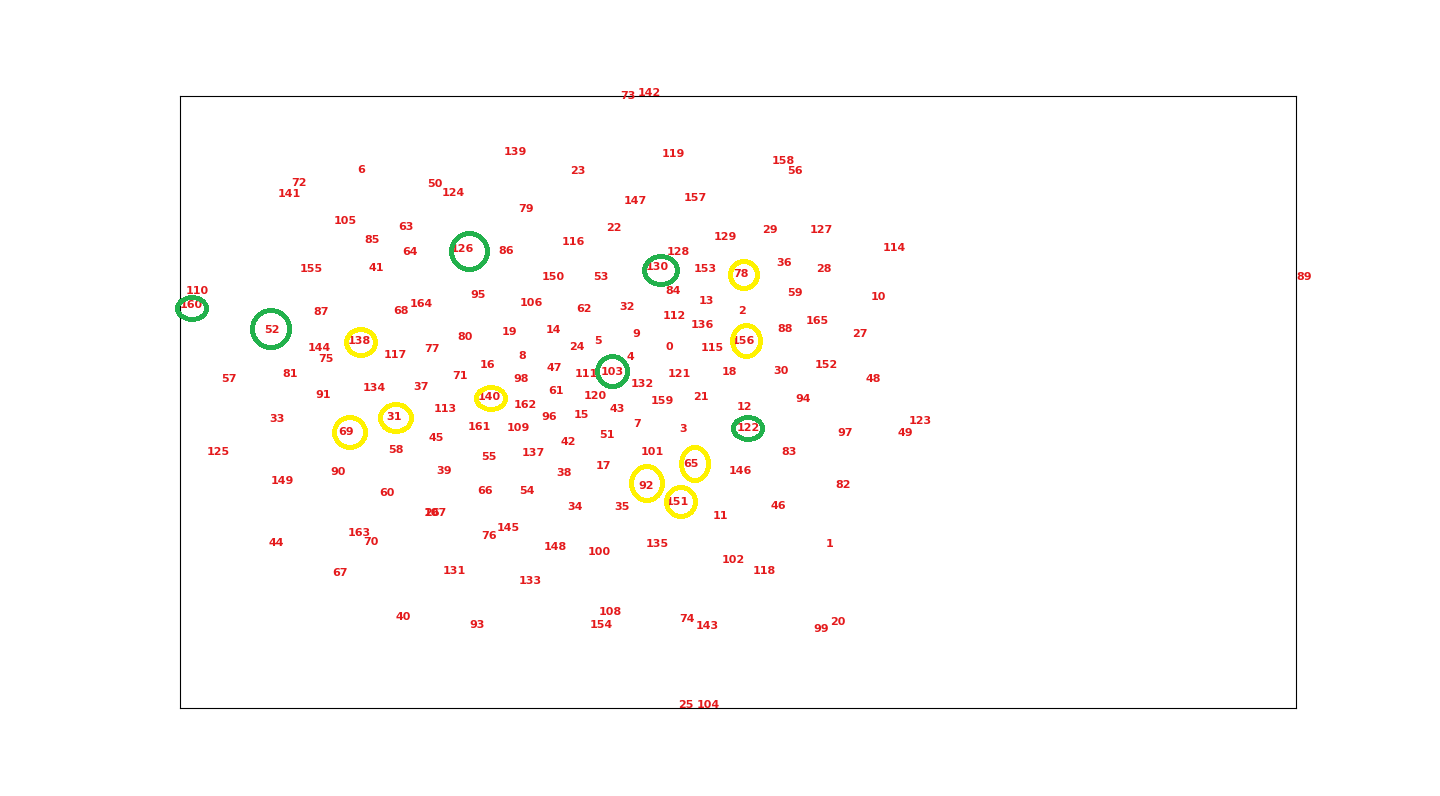
得到相应的初步筛选结果：



通过上述筛选步骤筛选出了一定的无效图片、绘画作品。



t-SNE降维



UMAP降维

根据ResNet50进特征提取后得到上述两种特征降维结果，示意图中以绿色标志典型的真实玫瑰图片，黄色标志典型的非真实玫瑰图片。由分布来看，两种图像中的真实和非真实玫瑰图片间均没有得到很好的区分。

考虑到玫瑰的特征简单，绘画逼真度高，故难以做出明显的区分。但是降维示意图仍然具有一定的借鉴意义，找到一张非真实的图片，在示意图上显示距离过近的图片仍有可能是非真实的，例如图片12和图片122。同时，在实际应用中真实和非真实玫瑰图片对于模型预测任务的影响可能不大，非真实图片对于提高模型泛化能力仍然具有价值，没有必要做严格的区分。

改进的思路：

1. 在降维为2时不显示区分度，在更高维空间是否存在显著的分布差别，可以考虑利用机器学习的方法对4096的特征维度数据进行训练，输出预测为真的概率。
2. 尝试其他的高级筛选办法。。。。。。

代码和报告的地址见：

https://github.com/ShijianRuan/Data-crawling-and-filtering