# 划分训练集和测试集

## 导入工具包

In [7]:
```
# 划分训练集和测试集
# 导入工具包
import os
import shutil
import random
import pandas as pd
```

## 获得所有类别名称

In [3]:
```
# 指定数据集路径
dataset_path = 'fruit81_full'
```

In [8]:
```
dataset_name = dataset_path.split('_')[0]
print('数据集', dataset_name)
```

```
数据集 fruit81
```

In [9]:
```
classes = os.listdir(dataset_path)
```

In [11]:
```
# classes
```

In [12]:
```
len(classes)
```

Out[12]:
```
81
```

## 创建训练集文件夹和测试集文件夹

In [13]:
```
# 创建 train 文件夹
os.mkdir(os.path.join(dataset_path, 'train'))

# 创建 test 文件夹
os.mkdir(os.path.join(dataset_path, 'val'))
```

In [15]:
```
# 在 train 和 test 文件夹中创建各类别子文件夹  已经存在会报错
for fruit in classes:
    os.mkdir(os.path.join(dataset_path, 'train', fruit))
    os.mkdir(os.path.join(dataset_path, 'val', fruit))
```

```
---------------------------------------------------------------------------
FileExistsError                           Traceback (most recent call last)
Cell In[15], line 3
      1 # 在 train 和 test 文件夹中创建各类别子文件夹  已经存在会报错
      2 for fruit in classes:
----> 3     os.mkdir(os.path.join(dataset_path, 'train', fruit))
      4     os.mkdir(os.path.join(dataset_path, 'val', fruit))

FileExistsError: [WinError 183] 当文件已存在时，无法创建该文件。: 'fruit81_full\\train\\人参果'
```

# 划分训练集、测试集，移动文件

In [16]:
```
test_frac = 0.2  # 测试集比例
random.seed(123) # 随机数种子，便于复现
```

In [15]:
```
print('{:^18} {:^18} {:^18}'.format('类别', '训练集数据个数', '测试集数据个数'))
```

```
        类别                训练集数据个数              测试集数据个数
```

In [18]:
```
df = pd.DataFrame()

print('{:^18} {:^18} {:^18}'.format('类别', '训练集数据个数', '测试集数据个数'))

for fruit in classes: # 遍历每个类别

    # 读取该类别的所有图像文件名
    old_dir = os.path.join(dataset_path, fruit)
    images_filename = os.listdir(old_dir)

    random.shuffle(images_filename) # 随机打乱

    # 划分训练集和测试集
    testset_numer = int(len(images_filename) * test_frac) # 测试集图像个数
    testset_images = images_filename[:testset_numer]      # 获取拟移动至 test 目录的
    trainset_images = images_filename[testset_numer:]     # 获取拟移动至 train 目录的

    # print(testset_images)
    # print(trainset_images)
    # 移动图像至 test 目录
    for image in testset_images:
        old_img_path = os.path.join(dataset_path, fruit, image)         # 获取原始文
        new_test_path = os.path.join(dataset_path, 'val', fruit, image) # 获取 test
        shutil.move(old_img_path, new_test_path) # 移动文件

    # 移动图像至 train 目录
    for image in trainset_images:
        old_img_path = os.path.join(dataset_path, fruit, image)           # 获取原始
        new_train_path = os.path.join(dataset_path, 'train', fruit, image) # 获取 tr
        shutil.move(old_img_path, new_train_path) # 移动文件

    # 删除旧文件夹
    assert len(os.listdir(old_dir)) == 0 # 确保旧文件夹中的所有图像都被移动走
    shutil.rmtree(old_dir) # 删除文件夹和里面文件，递归删除

    # 工整地输出每一类别的数据个数
    print('{:^18} {:^18} {:^18}'.format(fruit, len(trainset_images), len(testset_ima

    # 保存到表格中
    df = df.append({'class':fruit, 'trainset':len(trainset_images), 'testset':len(te
    # 无需自然索引
# 重命名数据集文件夹
shutil.move(dataset_path, dataset_name+'_split')

# 数据集各类别数量统计表格，导出为 csv 文件
df['total'] = df['trainset'] + df['testset']
df.to_csv('fruit81_full数据集划分统计.csv', index=False)
```

| 类别 | 训练集数据个数 | 测试集数据个数 |
|---|---|---|
| 人参果 | 146 | 36 |
| 佛手瓜 | 129 | 32 |
| 哈密瓜 | 157 | 39 |
| 圣女果 | 158 | 39 |
| 山楂 | 159 | 39 |
| 山竹 | 152 | 38 |
| 无花果 | 156 | 39 |
| 木瓜 | 156 | 38 |
| 李子 | 154 | 38 |
| 杏 | 158 | 39 |
| 杨桃 | 157 | 39 |
| 杨梅 | 153 | 38 |
| 枇杷 | 151 | 37 |
| 枣 | 156 | 38 |
| 柚子 | 156 | 39 |
| 柠檬 | 122 | 30 |
| 柿子 | 154 | 38 |
| 树莓 | 149 | 37 |
| 桂圆 | 159 | 39 |
| 桑葚 | 156 | 39 |
| 梨 | 155 | 38 |
| 椰子 | 159 | 39 |
| 榴莲 | 159 | 39 |
| 樱桃 | 133 | 33 |
| 橘子 | 145 | 36 |
| 毛丹 | 127 | 31 |
| 水蜜桃 | 141 | 35 |
| 沃柑 | 159 | 39 |
| 沙果 | 153 | 38 |
| 沙棘 | 147 | 36 |
| 油桃 | 160 | 39 |
| 牛油果 | 120 | 30 |
| 猕猴桃 | 158 | 39 |
| 甘蔗 | 158 | 39 |
| 甜瓜-伊丽莎白 | 75 | 18 |
| 甜瓜-白 | 68 | 17 |
| 甜瓜-绿 | 35 | 8 |
| 甜瓜-金 | 42 | 10 |
| 番石榴-百 | 105 | 26 |
| 番石榴-红 | 121 | 30 |
| 白兰瓜 | 103 | 25 |
| 白心火龙果 | 148 | 37 |
| 白萝卜 | 160 | 39 |
| 百香果 | 151 | 37 |
| 石榴 | 153 | 38 |
| 砂糖橘 | 148 | 36 |
| 粑粑柑 | 154 | 38 |
| 红心火龙果 | 159 | 39 |
| 红苹果 | 142 | 35 |
| 羊奶果 | 156 | 39 |
| 羊角蜜 | 157 | 39 |
| 胡萝卜 | 149 | 37 |
| 脐橙 | 154 | 38 |
| 腰果 | 160 | 40 |
| 芒果 | 139 | 34 |
| 芦柑 | 146 | 36 |
| 草莓 | 159 | 39 |
| 荔枝 | 158 | 39 |
| 莲雾 | 156 | 39 |
| 菠萝 | 158 | 39 |
| 菠萝莓 | 91 | 22 |
| 菠萝蜜 | 160 | 39 |
| 葡萄-白 | 125 | 31 |

| | | |
|---|---|---|
| 葡萄-红 | 160 | 39 |
| 蓝莓 | 158 | 39 |
| 蛇皮果 | 138 | 34 |
| 蟠桃 | 145 | 36 |
| 血橙 | 150 | 37 |
| 西柚 | 147 | 36 |
| 西梅 | 158 | 39 |
| 西瓜 | 156 | 38 |
| 西红柿 | 150 | 37 |
| 车厘子 | 136 | 33 |
| 酸角 | 153 | 38 |
| 金桔 | 145 | 36 |
| 青柠 | 119 | 29 |
| 青苹果 | 156 | 39 |
| 香橼 | 104 | 25 |
| 香蕉 | 155 | 38 |
| 黄桃 | 155 | 38 |
| 黑莓 | 150 | 37 |

# 查看文件目录结构

In [25]:
```
!pip install tree
```

```
Collecting tree
  Downloading Tree-0.2.4.tar.gz (6.5 kB)
  Preparing metadata (setup.py): started
  Preparing metadata (setup.py): finished with status 'done'
Requirement already satisfied: Pillow in d:\software\anaconda3\lib\site-packages (fr
om tree) (9.4.0)
Collecting svgwrite (from tree)
  Downloading svgwrite-1.4.3-py3-none-any.whl (67 kB)
     ---------------------------------------- 0.0/67.1 kB ? eta -:--:--
     ------------------ --------------------- 30.7/67.1 kB 1.3 MB/s eta 0:00:01
     ------------------ --------------------- 30.7/67.1 kB 1.3 MB/s eta 0:00:01
     ---------------------------------------- 67.1/67.1 kB 521.9 kB/s eta 0:00:00
Requirement already satisfied: setuptools in d:\software\anaconda3\lib\site-packages
(from tree) (68.0.0)
Requirement already satisfied: click in d:\software\anaconda3\lib\site-packages (fro
m tree) (8.0.4)
Requirement already satisfied: colorama in d:\software\anaconda3\lib\site-packages
(from click->tree) (0.4.6)
Building wheels for collected packages: tree
  Building wheel for tree (setup.py): started
  Building wheel for tree (setup.py): finished with status 'done'
  Created wheel for tree: filename=Tree-0.2.4-py3-none-any.whl size=7878 sha256=cad9
6567958f5a50fd7718b01851b992b806a8817c0f4ea7698f23213d31c2fb
  Stored in directory: c:\users\baisichang\appdata\local\pip\cache\wheels\e8\ed\fe\b
4c6a9b7a5b8df6d966ea673e26a46a7451b020af754eafa6b
Successfully built tree
Installing collected packages: svgwrite, tree
Successfully installed svgwrite-1.4.3 tree-0.2.4
```

In [19]:
```
!tree fruit81_split
```

```
卷 系统 的文件夹 PATH 列表
卷序列号为 000000CC FC7A:27E2
C:\USERS\BAISICHANG\IMG_CLASSIFICATION\IMG_CLA\1_BUILDING_DATASET\FRUIT81_SPLIT
├─train
│   ├─人参果
│   ├─佛手瓜
│   ├─哈密瓜
│   ├─圣女果
│   ├─山楂
│   ├─山竹
│   ├─无花果
│   ├─木瓜
│   ├─李子
│   ├─杏
│   ├─杨桃
│   ├─杨梅
│   ├─枇杷
│   ├─枣
│   ├─柚子
│   ├─柠檬
│   ├─柿子
│   ├─树莓
│   ├─桂圆
│   ├─桑葚
│   ├─梨
│   ├─椰子
│   ├─榴莲
│   ├─樱桃
│   ├─橘子
│   ├─毛丹
│   ├─水蜜桃
│   ├─沃柑
│   ├─沙果
│   ├─沙棘
│   ├─油桃
│   ├─牛油果
│   ├─猕猴桃
│   ├─甘蔗
│   ├─甜瓜-伊丽莎白
│   ├─甜瓜-白
│   ├─甜瓜-绿
│   ├─甜瓜-金
│   ├─番石榴-百
│   ├─番石榴-红
│   ├─白兰瓜
│   ├─白心火龙果
│   ├─白萝卜
│   ├─百香果
│   ├─石榴
│   ├─砂糖橘
│   ├─粑粑柑
│   ├─红心火龙果
│   ├─红苹果
│   ├─羊奶果
│   ├─羊角蜜
│   ├─胡萝卜
│   ├─脐橙
│   ├─腰果
│   ├─芒果
│   ├─芦柑
│   ├─草莓
│   ├─荔枝
│   ├─莲雾
│   ├─菠萝
```

```
│   ├──菠萝莓
│   ├──菠萝蜜
│   ├──葡萄-白
│   ├──葡萄-红
│   ├──蓝莓
│   ├──蛇皮果
│   ├──蟠桃
│   ├──血橙
│   ├──西柚
│   ├──西梅
│   ├──西瓜
│   ├──西红柿
│   ├──车厘子
│   ├──酸角
│   ├──金桔
│   ├──青柠
│   ├──青苹果
│   ├──香橼
│   ├──香蕉
│   ├──黄桃
│   └──黑莓
└──val
    ├──人参果
    ├──佛手瓜
    ├──哈密瓜
    ├──圣女果
    ├──山楂
    ├──山竹
    ├──无花果
    ├──木瓜
    ├──李子
    ├──杏
    ├──杨桃
    ├──杨梅
    ├──枇杷
    ├──枣
    ├──柚子
    ├──柠檬
    ├──柿子
    ├──树莓
    ├──桂圆
    ├──桑葚
    ├──梨
    ├──椰子
    ├──榴莲
    ├──樱桃
    ├──橘子
    ├──毛丹
    ├──水蜜桃
    ├──沃柑
    ├──沙果
    ├──沙棘
    ├──油桃
    ├──牛油果
    ├──猕猴桃
    ├──甘蔗
    ├──甜瓜-伊丽莎白
    ├──甜瓜-白
    ├──甜瓜-绿
    ├──甜瓜-金
    ├──番石榴-百
    ├──番石榴-红
    ├──白兰瓜
    ├──白心火龙果
```

```
├──白萝卜
├──百香果
├──石榴
├──砂糖橘
├──粑粑柑
├──红心火龙果
├──红苹果
├──羊奶果
├──羊角蜜
├──胡萝卜
├──脐橙
├──腰果
├──芒果
├──芦柑
├──草莓
├──荔枝
├──莲雾
├──菠萝
├──菠萝莓
├──菠萝蜜
├──葡萄-白
├──葡萄-红
├──蓝莓
├──蛇皮果
├──蟠桃
├──血橙
├──西柚
├──西梅
├──西瓜
├──西红柿
├──车厘子
├──酸角
├──金桔
├──青柠
├──青苹果
├──香橼
├──香蕉
├──黄桃
└──黑莓
```

In [ ]: