统计图像尺寸、比例分布

导入工具包

In [6]:
```
import os
import numpy as np
import pandas as pd
import cv2
from tqdm import tqdm

import matplotlib.pyplot as plt
%matplotlib inline
import warnings
warnings.filterwarnings('ignore')
```

指定数据集路径

In [7]:
```
os.getcwd()
```

Out[7]:
```
'C:\\Users\\baisichang\\img_classification\\img_cla\\1_building_dataset'
```

In [8]:
```
# 指定数据集路径
dataset_path = 'fruit81_full' #还没有划分训练集和测试集
os.chdir(dataset_path) #重新指定工作目录
os.listdir()
```

Out[8]:
```
['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 [9]:

```
df = pd.DataFrame()
for fruit in tqdm(os.listdir()): # 遍历每个类别
    os.chdir(fruit)  # 改变当前的工作目录到这个名称下面
    for file in os.listdir(): # 遍历每张图像
        try:
            img = cv2.imread(file)
            # print(img.shape)
            df = df.append({'类别':fruit, '文件名':file, '图像宽':img.shape[1], '图像
        except:
            print(os.path.join(fruit, file), '读取错误')
    os.chdir('../')
os.chdir('../')
```

```
100%|████████████████████████████████████████████████████████████████████████████| 81/81 [01:02<00:00,  1.30it/s]
```

In [6]:

```
df
```

Out[6]:

	图像宽	图像高	文件名	类别
0	500.0	500.0	1.jpeg	人参果
1	500.0	329.0	10.jpg	人参果
2	749.0	500.0	100.jpg	人参果
3	500.0	500.0	101.jpg	人参果
4	300.0	200.0	102.jpg	人参果
...	...	...	...	...
14428	750.0	500.0	95.jpg	黑莓
14429	700.0	467.0	96.jpg	黑莓
14430	500.0	482.0	97.jpg	黑莓
14431	667.0	500.0	98.jpg	黑莓
14432	600.0	476.0	99.jpg	黑莓

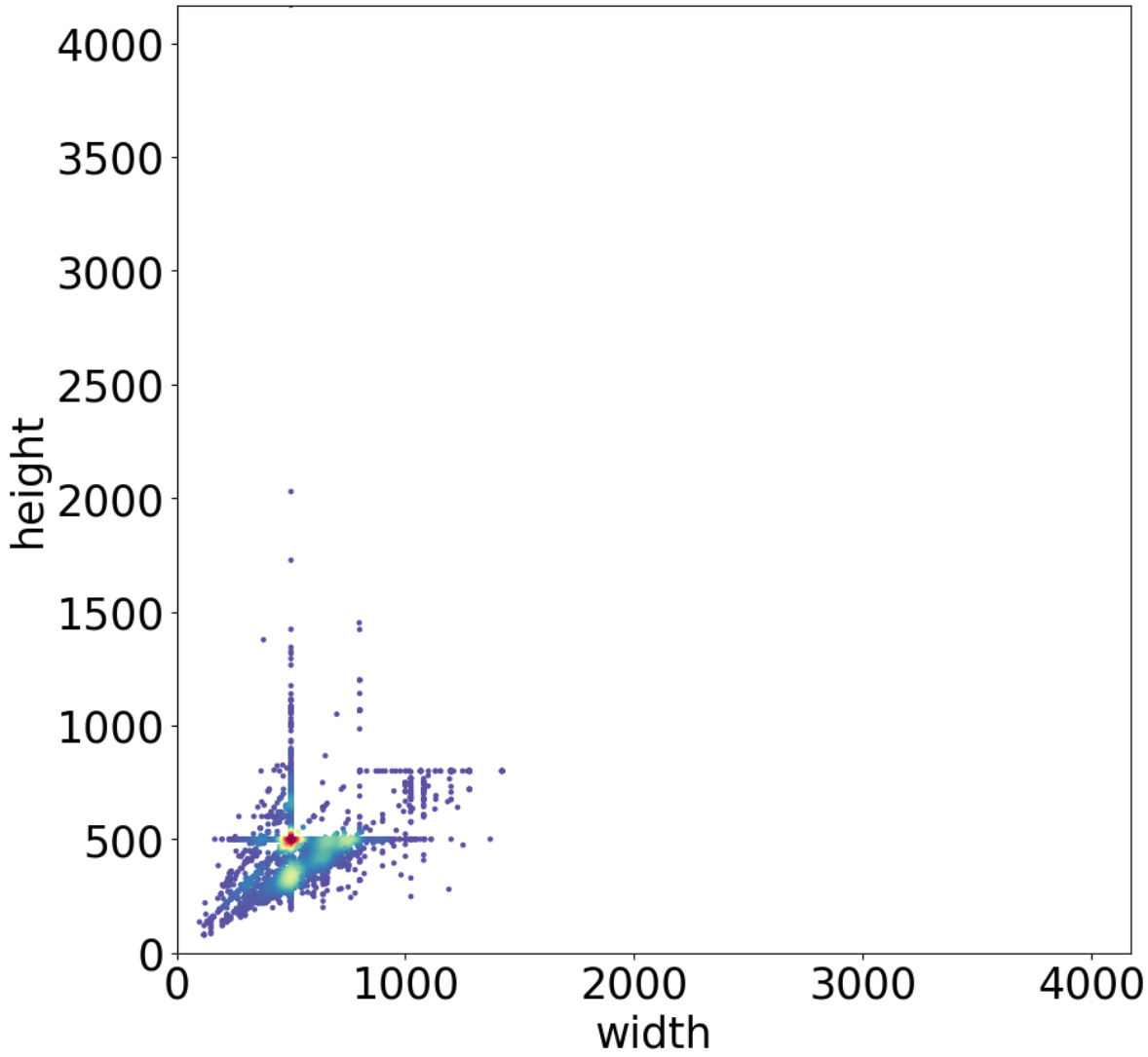
14433 rows × 4 columns

In [7]:

```
df.to_csv('C:\\Users\\baisichang\\img_classification\\img_cla\\1_building_dataset\\f
```

可视化图像尺寸分布

In [10]:
```
from scipy.stats import gaussian_kde
from matplotlib.colors import LogNorm

# df_path = 'C:\\Users\\baisichang\\img_classification\\img_cla\\1_building_dataset\
# df = pd.read_csv(df_path)

x = df['图像宽']

y = df['图像高']

xy = np.vstack([x,y]) # 垂直方向叠放
z = gaussian_kde(xy)(xy)

# Sort the points by density,so that the densest points are plotted last
# 按密度对点进行排序，以便最后绘制最密集的点。
idx = z.argsort()
x, y, z = x[idx], y[idx], z[idx]

plt.figure(figsize=(10,10)) # 创建图像
# plt.figure(figsize=(12,12))
plt.scatter(x, y, c=z,  s=5, cmap='Spectral_r')  # 函数用于生成一个scatter散点图,是-
# plt.colorbar()
# plt.xticks([])
# plt.yticks([])

plt.tick_params(labelsize=25) # 参数labelsize用于设置刻度线标签的字体大小

xy_max = max(max(df['图像宽']), max(df['图像高']))
plt.xlim(xmin=0, xmax=xy_max)
plt.ylim(ymin=0, ymax=xy_max)

plt.ylabel('height', fontsize=25) # xy轴标签字体大小
plt.xlabel('width', fontsize=25)

plt.savefig('图像尺寸分布.pdf', dpi=120, bbox_inches='tight')

plt.show()
```


4000
3500
3000
2500
2000
1500
1000
500
0
height
0
1000
2000
3000
4000
width

In [ ]: