实验记录　　序号：15

摘要：三叶青块根各类别照片数量的统计，以及第三轮拍摄

1. 实验准备

将三叶青块根照片划分为训练集、测试集、验证集；对三叶青块根照片进行第三次补拍

2. 实验步骤

2.1. 第三轮拍摄

由于是按照省份分类，导致之前拍的根据每个具体产地的样本量由原来的均一，变为不均一，样本量差距有点大。故要补充拍摄三叶青块根照片

最终实现按照省份分类的三叶青块根照片每一类的样本照片数量差不多在 1500 左右。

2.2. 划分为训练集、测试集、验证集

编写 python 代码，先将数据集按照 8：2 的比例划分预训练集和测试集；然后按同样比例将预训练集划分为训练集和验证集

将测试集划分出来

```python
import os
import random
import shutil
def random_select_images(source_folder, target_folder, ratio):
    # 遍历源文件夹及其子文件夹，收集图片文件
    images = []
    for root, dirs, files in os.walk(source_folder):
        for file in files:
            if file.lower().endswith(('.png', '.jpg', '.jpeg', '.bmp')):
                images.append(os.path.join(root, file))

    # 计算需要选择的图片数量
    num_images = int(len(images) * ratio)

    # 随机挑选指定数量的图片
    selected_images = random.sample(images, num_images)

    # 将挑选出的图片复制到目标文件夹，保留文件名
    if not os.path.exists(target_folder):
        os.makedirs(target_folder)
    for image in selected_images:
        # 获取图片文件名（不包含路径）
        filename = os.path.basename(image)
        # 构造目标文件路径
        target_path = os.path.join(target_folder, filename)
        # 移动文件到目标路径
        shutil.move(image, target_path)

    return selected_images


if __name__ == '__main__':
    # 使用示例
    source_folder = r"D:\SanYeQing_Project\hunhe-wht-sanyeqing\train_hun_weizhi"  # 源文件夹路径
    target_folder = r"D:\SanYeQing_Project\hunhe-wht-sanyeqing\test_hun_weizhi"  # 目标文件夹路径
    ratio = 0.2  # 需要挑选的图片比例
    random_select_images(source_folder, target_folder, ratio)
```

```python
                data_counts[dir_name][data_counts['class'].index(class_name)] += len(os.listdir(class_sub_dir))
        df = pd.DataFrame(data_counts)
        return df


hua_train_val = MyImageClassifier(data_dir="D:\SanYeQing_Project\sanyeqing_hun_weizhi_finally",target_dir="D:\linshi_mulu",
                                  valid_ratio=0.2,train_folder = 'train_hun_finally',
                                  test_folder='test_hun_finally')
hua_train_val.reorg_san_data('labels_hun.csv')
hua_train_val.classes()
df = hua_train_val.count_samples()
print(df)
test_df = pd.read_csv("数据量统计.csv")
print(test_df)
merged_df = pd.merge(df, test_df, on=['class'], how='outer')
merged_df['total'] = merged_df[['train', 'valid', 'test']].sum(axis=1)    #总和统计
print(merged_df)
merged_df.to_csv('数据量统计.csv', index=False)
```

2.3. 统计各类别图像数量

统计每个类别图像总数量、训练集、验证集、测试集数量，并形成一个表格，最后用柱状图显示。

```python
#total
plt.figure(figsize=(22, 7))

x = df['class']
y = df[feature]

plt.bar(x, y, facecolor='#1f77b4', edgecolor='k')

plt.xticks(rotation=90)
plt.tick_params(labelsize=15)
plt.xlabel('类别', fontsize=20)
plt.ylabel('图像数量', fontsize=20)

# plt.savefig('各类别图片数量.pdf', dpi=120, bbox_inches='tight')

plt.show()
```

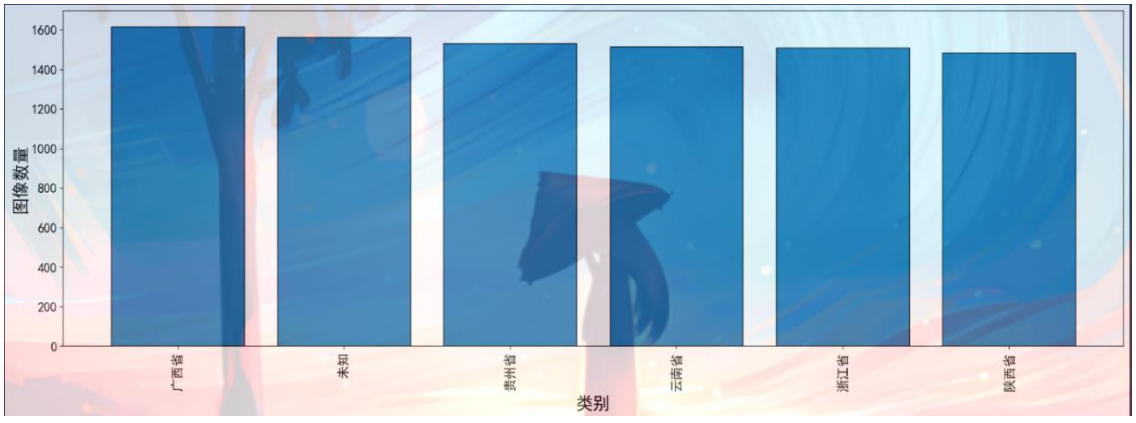
✓ 0.2s

3. **实验结果**

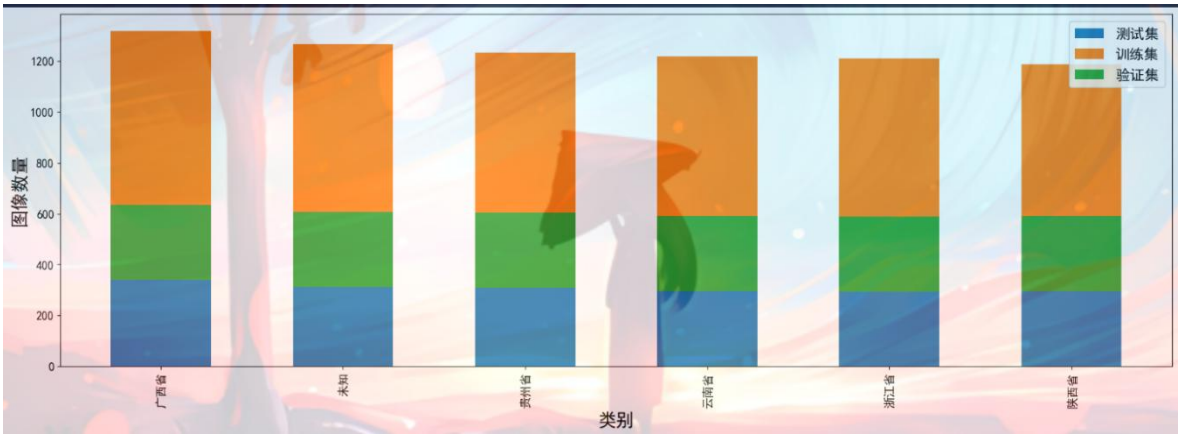
统计图像数量：

	class	train	valid	test	total
1	广西省	981	296	339	1616
2	未知	957	296	310	1563
4	贵州省	926	296	308	1530
0	云南省	922	296	296	1514
3	浙江省	919	296	293	1508
5	陕西省	891	296	296	1483

图像数量可视化：



实验记录　　序号：15


测试集
训练集
验证集
图像数量
0
200
400
600
800
1000
1200
广西省
未知
贵州省
云南省
浙江省
陕西省
类别