📖 README - 数据集预处理工具

🚀 项目概述

本工具用于自动化处理图像分类任务的数据集准备，主要功能包括：

1. **自动解析文件夹结构**：根据预设的类别标签生成标注文件
2. **数据集划分**：按指定比例分割训练集和验证集
3. **生成标准标签文件**：输出JSON格式的标注文件

📂 数据准备要求

1. **目录结构**：

```
dataset/
└── images/
    ├── apple/
    │   ├── img001.jpg
    │   └── ...
    ├── banana/
    ├── battery/
    └── battle/
```

2. **文件命名**：
   - 每个子文件夹名称对应`label_dict`中的键值
   - 支持.jpg和.png格式图片

⚙️ 使用方法

1. **安装依赖**：

```
pip install scikit-learn
```

2. **配置参数**： 修改代码开头的配置部分：

```
# 类别定义（必须与文件夹名称一致）
label_dict = {"apple":0, "banana":1, ...}

# 路径配置
images_folder = "./dataset"
```

3. **运行脚本**：

```
python3 ./tools/dataset_preprocess.py
```

📊 输出文件格式

生成的标签文件示例（JSON格式）：

```
{
  "apple/img001.jpg": [0],
  "banana/img123.jpg": [1],
  ...
}
```

⚠ 注意事项

1. **类别匹配**：确保文件夹名称与`label_dict`完全一致
2. **数据平衡**：建议每个类别至少有100+样本以获得较好效果
3. **扩展支持**：可通过修改代码添加更多图片格式支持（如.webp）

```
# 📌 常见问题解答

**Q1：如何调整训练集/验证集的比例？**
> 修改代码中的`test_size`参数，例如设置为0.2表示20%作为验证集

**Q2：如何添加新的类别？**
> 1. 在images文件夹下新建类别文件夹
> 2. 在label_dict中添加对应的键值对
> 3. 重新运行脚本

**Q3：出现"未定义类别"警告怎么办？**
> 检查文件夹名称是否与label_dict中的键完全匹配，包括大小写

**Q4：如何实现分层抽样？**
> 修改train_test_split调用，添加`stratify=labels`参数（需单独提取labels列表）
```