data_manage项目说明文档

项目文件目录树：

```
data_manage
│  cookies.pickle    百度云账号的cookies数据
│  run_gen.py    运行data_manage总程序入口
│  run_pan.py    运行百度云盘自动化程序入口（创建分享链接功能有问题）
│
├─BaiDuPan    百度云盘自动化与钉钉自动发送百度云分享链接
│  │  baiDuPan.py
│  │  bypy_operate.py
│  │  logid.py
│  │  login_pan.py 百度云盘模拟登录（userName为百度云账号，password为百度云账号密码）
│  │
│  └─utils    工具类
│      │  fileUtil.py
│      │  pathUtil.py
│      │  profileUtil.py
│
├─config    配置文件存放路径
│      baidupan.conf    百度云相关配置文件
│      data_manage.conf    视频切分图片及数据预处理配置文件（step1-step8）
│      utils_db.conf    数据库操作相关配置文件（step8）
│
├─database    生成数据库及相关数据操作
│  │  create_db.py    创建数据库
│  │  delete_tables_data.py    删除表
│  │  insert_tables_data.py    插入表
│  │  update_tables_data.py    更新表
│  │
│  └─utils 工具类
│      │  connect_db_util.py
│      │  delete_data_util.py
│      │  insert_data_util.py
│      │  mkdir.py
│      │  pathUtil.py
│      │  profileUtil.py
│      │  select_data_util.py
│      │  update_data_util.py
│      │
│      └─utils_sql_create_DBImg    存放创建数据库的sql语句配置文件夹（见数据库字段说明表）
│              utils_sql_class_state.conf    创建class_state表
│              utils_sql_image_basic.conf    创建image_basic表
│              utils_sql_img_bbox.conf    创建img_bbox表
│              utils_sql_scene_basic.conf    创建scene_basic表
│              utils_sql_spot_basic.conf    创建spot_basic表
│
├─gen_data 数据预处理
│  │  check_data_step2.py    对标注公司标注好的文件进行二次质检，确保xml文件和jpg文件一一对应
```

step2*检查是否存在漏标的图片*
│ │ *classFeatureImages_step3.py　将标注后的图像中的特征图像提取出来后按照类别分类 step3*截取图片中标注后的特征图像，并对其分类处理
│ │ class_operate.py　小工具功能：将本地xml文件中多余的类删除，或者是修改已经存在的旧类
│ │ genAnn_step5.py　生成图片的具体信息：文件路径、特征图像的坐标等关键信息 step5*生成图片信息文件*
│ │ *gen_class_index_step4.py　生成图片的所有分类信息 step4*生成类别信息文件
│ │ gen_pb_tfrecord_step7.py　生成pbtxt文件和Tfrecord文件 step7*生成特定的文件*
│ │ *splitTrainVal_step6.py　在根据类别信息文件对图片进行分类后的基础上进行训练集和测试集的划分操作 step6*训练集和验证集的划分
│ │ video2image_step1.py　视频转换为图片 step1_视频切分为图片（已将此步骤利用pyqt生成了可交互的exe程序，请在test_util/QT5目录下查看）
│ │
│ └─utils　工具类
│　　coorUtil.py　检查标注框在原图片中的位置是否出现超越背景图片总长宽的情况
│　　fileUtil.py　文件筛选工具类
│　　pathUtil.py　路径操作工具类
│　　profileUtil.py　配置文件操作工具类
│　　strUtil.py　字符串工具类
│
└─test_util
　│ filter_file.py　获取某一目录下面所有子目录中前一半的视频文件（洛河电厂项目数据中存在红外镜头拍摄的视频，只取正常镜头拍摄的视频，文件可忽略）
　│
　├─find_diff　特殊场景下批量生成相同内容的xml文件（此文件为特殊情况，可忽略）
　│　　find_different.py
　│
　├─img_mask　利用旧的特征图像为前景将旧图片为背景，生成新的训练数据集
　│　　imgmask.py
　│　　pathUtil.py
　│
　├─merger_tfrecord　将两个tfrecord文件合并为一个
　│　　merger_tfrecord.py
　│
　├─other2jpg　将其他类型的图片转化为jpg格式（支持36类）
　│　　other2jpg.py
　│
　├─Qt5 可生成exe程序
　│ ├─excel 将总标注数据整理成表格的形式，利用程序实现鼠标点击即可一键导出选中的部分数据生成某个项目的标注数据说明文档交给标注公司，指导其标注工作
　│ │ │ ModifyTree.py 主程序
　│ │ │
　│ │ └─data
　│ │　　│ image.xlsx 存放标注数据的excel表
　│ │　　│ excel.conf 程序的配置文件
　│ │　　│
　│ │　　└─images 存放标注数据的示例图片
　│ │
　│ └─videos2imgs 视频切分图片
　│　　form1.3.py
　│
　└─replace_color 语义分割图中的颜色归并，将多色块图中的多种颜色归并为一种颜色（多分类语义

图中只保留树，将其他分类的色块都归并为背景）

parse_xml.py

replace_color.py

1.运行run_pan.py:

1.1 更新cookies：（配置文件：baidupan.conf)

当程序显示无效cookies时，需要运行此功能更新cookies。

首先检查login_pan.py文件中userName和password是否正确。然后运行该功能，模拟登录程序会自动填写账号和密码信息，但是需要人工完成辅助验证功能，完成后在程序运行命令行按回车键即可更新cookies。

1.2 上传文件（配置文件：baidupan.conf)

用于将本地文件上传到百度云路径下。

更改config/baidupan.conf配置文件[upload]中的相关信息，其中网盘的路径是相对路径，网盘上传路径会以/app/bypy文件夹为根目录，且不能更换。更改好配置文件后运行该功能即可实现文件的上传。

1.3 下载文件（配置文件：baidupan.conf)

用于将网盘文件下载到本地路径。

更改config/baidupan.conf配置文件[download]中的相关信息，网盘路径同样为相对路径，修改好配置文件运行该功能即可实现文件下载。

1.4 创建分享链接（配置文件：baidupan.conf)

将百度云盘中的文件分享给其他人

更改config/baidupan.conf配置文件[download]中的相关信息，网盘路径同样为相对路径，修改好配置文件运行该功能即可实现创建分享链接。（创建分享链接功能会经常出问题，该功能慎用）

2.运行run_gen.py:（)

运行流程：步骤1：video2image_step1>标注公司

步骤2：check_data_step2>classFeatureImages_step3>gen_class_index_step4>genAnn_step5>splitTrainVal_step6>gen_pb_tfrecord_step7

步骤3：存入数据库。

2.1 视频切分图片操作（step1）（配置文件：data_manage.conf)

更改config/data_manage.conf配置文件[vidoe2image]中的相关信息， video_dir为视频文件夹路径，save_dir为切分图片的保存路径。视频可以按帧数切分也可以按总数切分，此功能只支持按帧数切分，且建议使用test_util/QT5/videos2imgs目录下的程序，该程序是此功能的升级版，支持指定照片总数。有很友好的可视化界面及其对应的exe执行程序。

2.2 数据预处理（step2-step7）

此功能一般情况下只需要修改config/data_manage.conf配置文件[vidoe2image]中dir字段即可，其他即为默认属性。step2-step7为默认一整套功能。

step2：质检 （配置文件：data_manage.conf）

对标注公司返回的标注结果进行质检,检查xml文件和jpg文件是否一一对应。如果有不对应的文件则会被提出来放到redundant_data文件夹下

step3: 截取标注框并分类 （配置文件：data_manage.conf）

截取标注的矩形框并按照类别分别放在个类别的文件夹下面，将所有类别图片统一放在class_img_dir文件夹下，在此期间还会检查标注框的边界问题，如果标注框有问题如坐标超过对应背景图像宽或高则将此图片及其对应的xml文件转移到redundant_data文件夹下

step4：生成类别信息文件 （配置文件：data_manage.conf）

对应的类别信息文件有：存放标注种类的json文件夹和存放标注种类的txt文件夹。里面包含了该项目中所以标注框的set集合

step5：生成图片信息文件 （配置文件：data_manage.conf）

提取图片的具体信息：文件路径、特征图像的坐标等关键信息，将信息数据分别以txt和csv两种格式保存，其中txt分为两种形式，一种为普通txt文件，另一种以方便yolo5读取的txt格式保存。目前默认yolo_txt格式。具体可以在config/data_manage.conf中[data_manage]的flag字段中更改。

step6：训练集和验证集划分 （配置文件：data_manage.conf）

在根据类别信息文件对图片进行分类后的基础上进行训练集和测试集的划分操作，确保了训练集和验证集在数据类别上的完整性。

step7: 生成训练数据和测试数据的tfrecord文件，生成ob.pbtxt文件

没啥可说的，就是生成了三个文件。路径都是默认即可。一般情况下不需要改动。

2.3 数据库操作

```
img_basic :图片文件信息存放表
        +--------------------+--------------+------+-----+---------+-------------
-----+
        | Field              | Type         | Null | Key | Default | Extra
     |
        +--------------------+--------------+------+-----+---------+-------------
-----+
        | id                 | int(11)      | NO   | PRI | NULL    |
auto_increment   | 图片文件ID
        | path               | varchar(128) | NO   |     | NULL    |
     | 图片路径
        | filename           | varchar(128) | YES  |     | NULL    |
     | 文件名称
        | width              | int(11)      | YES  |     | NULL    |
     | 图片宽
        | height             | int(11)      | YES  |     | NULL    |
     | 图片高
        | depth              | int(11)      | YES  |     | NULL    |
     | 图片深度
        | spot_id            | int(11)      | YES  |     | NULL    |
     | 项目ID(地点ID)
```

```
        | create_time        | datetime     | YES  |     | NULL    |
CURRENT_TIMESTAMP| 创建时间
        | update_time        | datetime     | YES  |     | NULL    |
CURRENT_TIMESTAMP| 更新时间
        +--------------------+--------------+------+-----+---------+-------------
-----+


img_bbox :图片标注框信息存放表
        +--------------------+--------------+------+-----+---------+-------------
-----+
        | Field              | Type         | Null | Key | Default | Extra
     |
        +--------------------+--------------+------+-----+---------+-------------
-----+
        | id                 | int(11)      | NO   | PRI | NULL    |
auto_increment   | 标注框ID
        | file_id            | int(11)      | NO   |     | NULL    |
     | 图片文件ID
        | xmin               | int(11)      | YES  |     | NULL    |
     | xmin坐标
        | ymin               | int(11)      | YES  |     | NULL    |
     | ymin坐标
        | xmax               | int(11)      | YES  |     | NULL    |
     | xmax坐标
        | ymax               | int(11)      | YES  |     | NULL    |
     | ymax坐标
        | class_name         | varchar(128) | YES  |     | NULL    |
     | 类别名称
        | state_name         | varchar(128) | YES  |     | NULL    |
     | 状态名称
        | bbox_ratio         | double       | YES  |     | NULL    |
     | 标注框转背景图片比例
        | create_time        | datetime     | YES  |     | NULL    |
CURRENT_TIMESTAMP| 创建时间
        | update_time        | datetime     | YES  |     | NULL    |
CURRENT_TIMESTAMP| 更新时间
        +--------------------+--------------+------+-----+---------+-------------
-----+


class_state :状态类别信息表
        +-----------+-------------+------+-----+---------+-------+
        | Field     | Type        | Null | Key | Default | Extra |
        +-----------+-------------+------+-----+---------+-------+
        |state_name | varchar(40) | NO   | PRI | NULL    |       | 状态名称（E1-0）
        |class_name | varchar(40) | NO   |     | NULL    |       | 类别名称(E1)
        |is_common  | varchar(20) | NO   |     | NULL    |       | 是否通用
        +-----------+-------------+------+-----+---------+-------+

scene_basic :场景信息表
        +-------------+--------------+------+-----+---------+-----------------+
       | Field       | Type         | Null | Key | Default | Extra           |
        +-------------+--------------+------+-----+---------+-----------------+
        | scene_id    | int(11)      | NO   | PRI | NULL    | auto_increment  |
场景ID
        | description | varchar(128) | YES  |     | NULL    |                 |
描述
```

```
        | scene       | varchar(128) | YES  |     | NULL    |                 |
场景名称
        | create_time | datetime     | YES  |     | NULL    | CURRENT_TIMESTAMP|
创建时间
        | update_time | datetime     | YES  |     | NULL    | CURRENT_TIMESTAMP|
更新时间
        +-------------+--------------+------+-----+---------+-----------------+

spot_basic:项目信息表
        +--------------+--------------+------+-----+---------+-----------------
+
        | Field        | Type         | Null | Key | Default | Extra
|
        +--------------+--------------+------+-----+---------+-----------------+
        | spot_id      | int(11)      | NO   | PRI | NULL    | auto_increment   |
项目ID（地点ID）
        | scene_id     | int(11)      | NO   |     | NULL    |                 |
场景ID
        | spot         | varchar(128) | NO   | PRI | NULL    |                 |
项目名称
        | spot_CN      | varchar(128) | NO   |     | NULL    |                 |
项目中文名称
        | project_path | varchar(128) | NO   | PRI | NULL    |                 |
项目路径
        | create_time  | datetime     | YES  |     | NULL    | CURRENT_TIMESTAMP|
创建时间
        | update_time  | datetime     | YES  |     | NULL    | CURRENT_TIMESTAMP|
更新时间
        +--------------+--------------+------+-----+---------+-----------------
+
```

首先如果是第一次运行，在相应数据库没有建立的情况下，需要去database目录下运行create_db.py。数据库的IP、user、password等信息需要在config/utils_db.conf中的[connect]下修改相关字段，来确保数据库的正常连接

a. 数据插入（直接通过解析所有图片的xml文件进行数据的插入操作）（配置文件：data_manage.conf, utils_db.conf)

需要修改config/data_manage.conf配置文件[database]中相关字段，里面保存着项目的路径以及数据相对于项目路径下的文件夹

第一次插入数据时，场景数据表（scene_basic）和项目数据表（spot_basic）为空，相关程序会引导用户对这两张表进行字段的插入操作，如果需要添加新应用场景或者新项目名称，这种方式同样适用。插入的所有数据都是以项目ID（spot_id）来作为索引的。所以当需要新增原来已经存在的项目数据时，数据库会通过相关命令将原有的数据进行删除，然后再插入现有的数据。

b.状态类别删除（配置文件：utils_db.conf)

该功能是删除原来项目中不正确的分类，如S1-0。程序会引导用户输入项目ID和需要删除的状态分类。注意：状态类别只能一个一个删除，例如需要删除S0小分类下的所有状态，只能分别通过删除S0-0和S0-1来实现。

c. 状态类别修改（配置文件：utils_db.conf)

该功能是修改原来项目中需要替换的分类。和上面的删除一样，程序会引导用户输入需要修改的状态类别及新的状态类别

2.4 小工具

1. 状态类别删除（配置文件：data_manage.conf）

   该功能是直接通过解析本地xml文件来删除其中的类别来实现类别删除。请在操作前仔细确认需要删除的状态类别，删除的状态类别需要通过修改config/data_manage.conf配置文件[data_manage]中的remove_class_dir和remove_class字段来完成。

2. 类别修改

   该功能和上边的类别删除相似，需要修改对应路径的配置文件[data_manage]中的update_class_dir、update_new_class和update_old_class字段来完成。

3. 转换图片格式为jpg

   该功能可以批量将一些常见的图像格式转换成jpg格式。具体支持转换的图片格式见代码（支持36种）程序会引导用户填写图片所在的路径