数据处理模块产品设计说明

目录

[1 概述 3](#_Toc2594427)

[2 登录 3](#_Toc2594428)

[3 数据集管理 3](#_Toc2594429)

[3.1 数据集浏览 3](#_Toc2594430)

[3.2 新建数据集 5](#_Toc2594431)

[3.3 数据集图片操作 5](#_Toc2594432)

[3.3.1 图片上传 5](#_Toc2594433)

[3.3.2 爬虫配置 7](#_Toc2594434)

[3.3.3 图片标注 7](#_Toc2594435)

[3.3.4 图片预标注 8](#_Toc2594436)

[3.4 数据集版本操作 8](#_Toc2594438)

[3.4.1 版本详情 9](#_Toc2594439)

[3.4.2 新建版本 9](#_Toc2594440)

[3.4.3 删除版本 10](#_Toc2594441)

[4 日志管理 10](#_Toc2594442)

[4.1 日志支持按照特定标记进行检索 10](#_Toc2594443)

[4.2 日志支持设置保存天数 10](#_Toc2594444)

[4.3 日志支持设置日志级别 10](#_Toc2594445)

[4.4 日志支持管理页面查看 10](#_Toc2594446)

[5 兼容性 10](#_Toc2594447)

[6 附录 11](#_Toc2594448)

[6.1 数据集 11](#_Toc2594449)

[6.2 数据集版本 12](#_Toc2594450)

[6.3 标注文件格式 12](#_Toc2594451)

# 概述

数据处理模块用于存储、维护图像训练数据与数据集，实现数据的上传、爬取、标注、清洗等功能。提供web交互页面与可以向外对接的API接口。

本文档用于定义数据处理模块的产品功能。

# 登录

用户可以通过用户名和密码登录到系统中。

1. 用户权限分为普通用户、标注用户2种；
   1. 普通用户：普通用户可以对数据集进行增删改查，对图像数据进行上传、爬取、标注、清洗等功能
   2. 标注用户：标注用户只能对图像进行标注
2. 用户登录后，可以选择退出系统，回到登录页面。
3. 用户信息需要与训练模块同步
4. 用户可以对其他用户共享/公开数据集权限。
   1. 可以选择向某几个普通用户开放数据集权限，包括增改查，不共享删除权限
   2. 可以选择向某几个标注用户开放对整个数据集的标注权限



# 数据集管理

## 数据集浏览

普通用户登录后进入数据集管理页面，数据集管理页面支持数据集列表展示

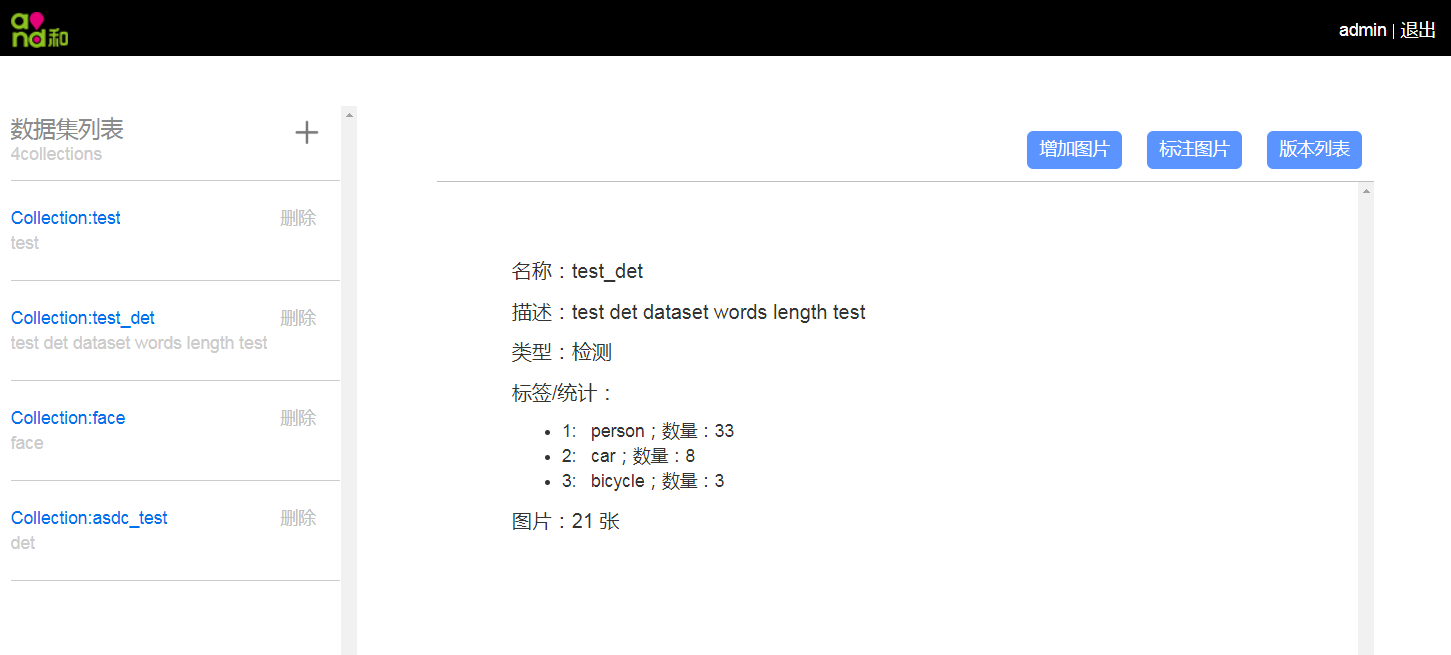


数据集列表支持：

1. 分类显示，用户创建数据集/其他用户共享数据集
2. 以列表形式展现数据集，显示名称、类型等信息
3. 数据集新建、删除等操作

标注用户登录后进入该界面，仅可以查看共享数据集信息，无新建删除权限。

选择某一数据集后，进入数据集详情界面



详情界面支持：

1. 详细信息查看，包括名称、描述、类型、版本信息等
2. 进行图片操作，包括上传、爬虫、标注、预标注等
3. 进行版本操作，包括查看所有版本详情等

标注用户仅能进行图片操作-标注

## 新建数据集

在数据集浏览页面点击新建数据集按钮进入数据集新建页面



1. 数据集名称：

数据集的名称，必填

1. 数据集描述：

描述数据集基本信息，50字以内，可选

1. 数据集类型：

互斥选项，必填。类型列表及说明参考附录6.1-数据集类型

1. 标注类别：

可以通过上传文件指定，也可以点击新增按钮人工输入

生成一组 { id:label } 列表。

1. 点击完成按钮完成数据集创建，进入该数据集详情页面；

## 数据集图片操作

## 图片上传

在数据集详情页点击图片操作-上传按钮进入图片上传页面

不同类型的数据集上传有所不同

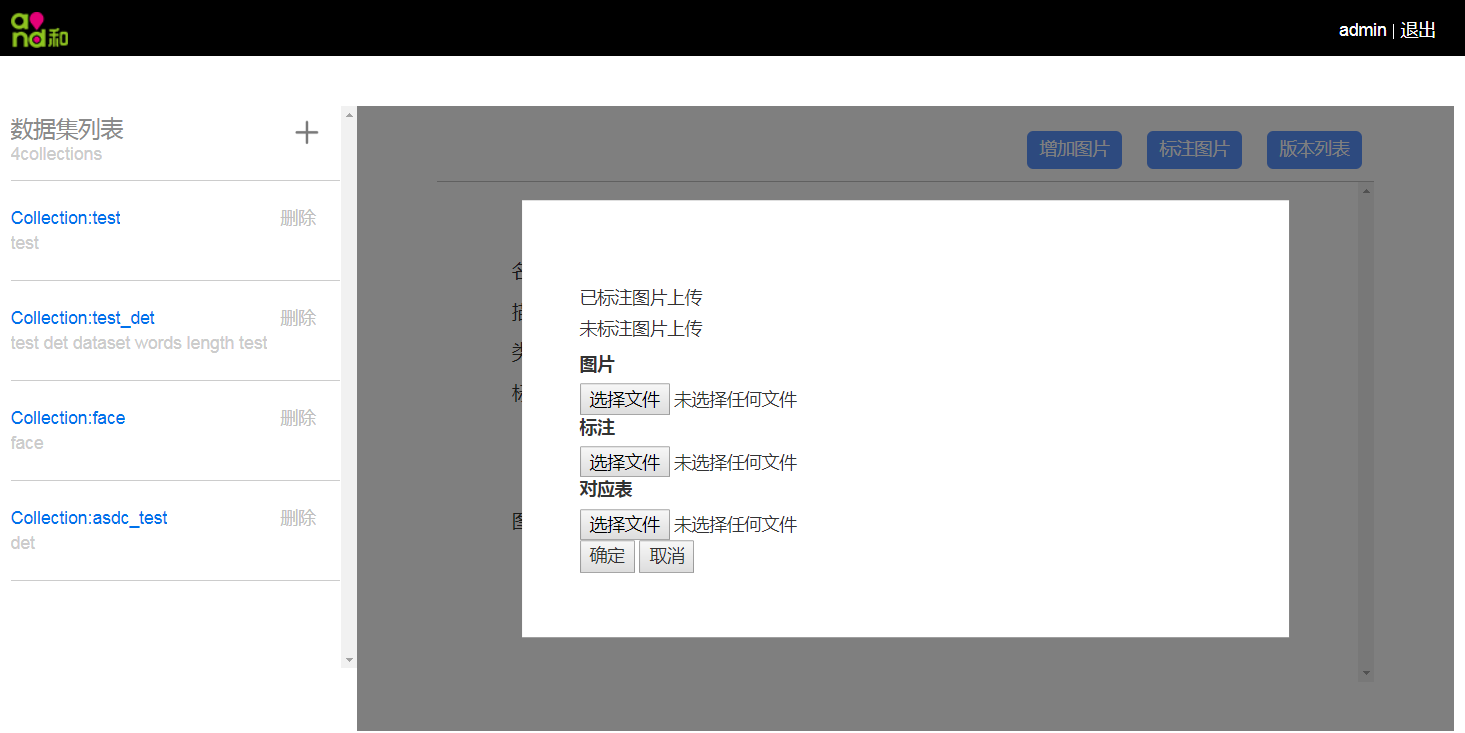
**分类/人脸识别**：



分类/人脸识别数据集上传时，可以选择上传已标注图片/未标注图片，可以单张/批量上传。

1. 已标注图片上传：需指定上传图片的标签，通过：
   1. 选择标签/新建标签方式指定
   2. 上传图片-标签对应表文件，文件每一行为“图片文件名 标签”格式。
2. 未标注文件上传：无需指定上传图片标签，只上传图片即可。

**检测/分割**：



检测/分割数据集上传时，可以选择上传已标注图片/未标注图片，可以单张/批量上传。

1. 已标注图片上传：需指定上传图片的标签，通过：
   1. 上传与图片有相同文件名的标注文件方式。如0001.jpg – 0001.xml
   2. 上传图片-标注对应表文件，文件每一行为“图片文件名 标注文件名”格式。
2. 未标注文件上传：无需指定上传图片标签，只上传图片即可。

## 爬虫配置

在数据集详情页点击图片操作-爬虫按钮进入爬虫配置页面

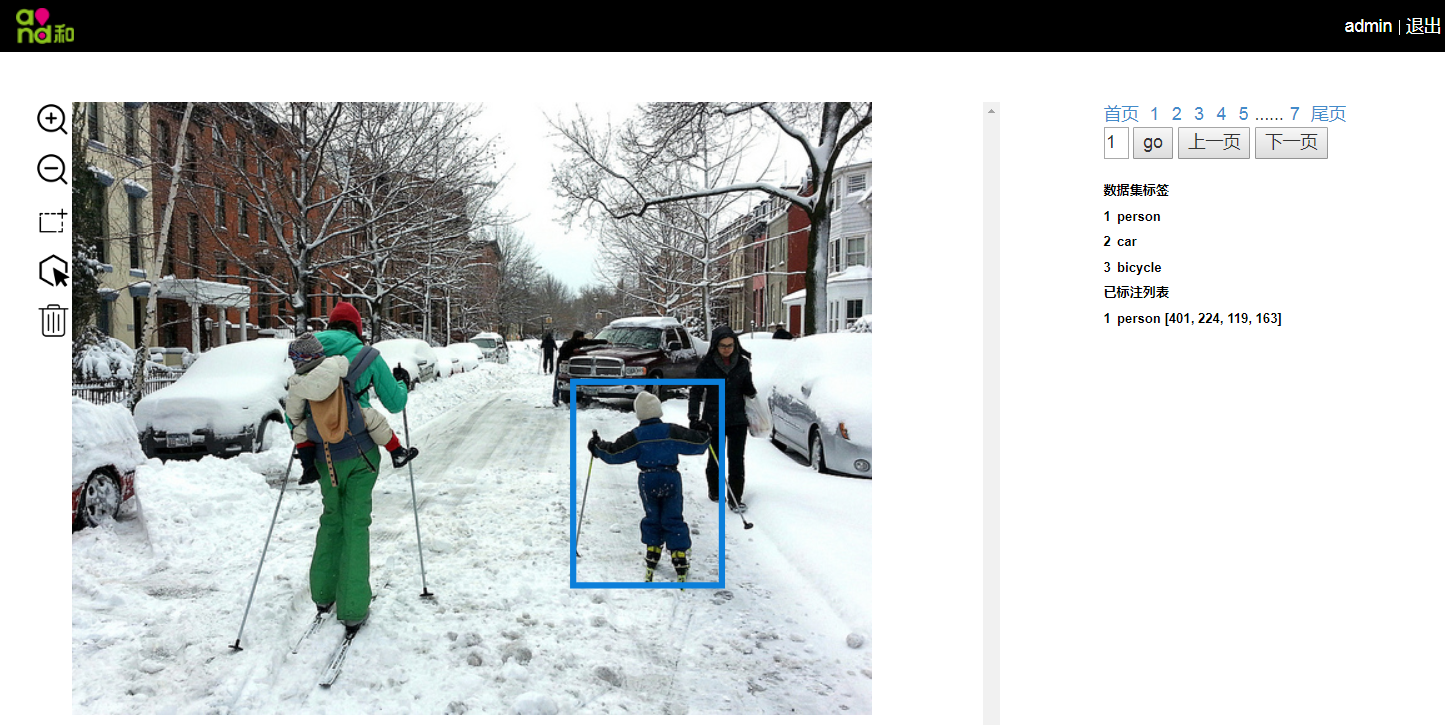


爬虫爬取百度图片搜索中图片，支持关键字、以图搜图方式

* 1. 关键字搜索：多个关键字以分号分隔；关键字组合以空格分隔
  2. 以图搜图：支持本地上传/网络url指定1张图片。关键字是可选选项

## 图片标注

在数据集详情页点击图片操作-标注按钮进入图片标注页面



图片标注功能包括：

1. 数据集标注说明：说明该数据集中包含的标签
2. 图片上一张、下一张
3. 图像缩放
4. 标注工具：
   1. 检测数据集：画四边形框，指定标签
   2. 分类数据集：不画框，指定唯一标签
   3. 分割数据集：画多边形框，指定标签
   4. 人脸数据集：不画框，指定唯一标签
5. 标注修改工具：对图片中某个标注，修改其标签，标注框，删除
6. 标注列表：列出该图片中所有的标注
7. 确认标注完成：
   1. 标注用户用于完成标注，提交审核
   2. 普通用户用于确认标注正确

## 图片预标注

在数据集详情页点击图片操作-预标注按钮进入图片预标注页面



后台保存智能系统（外部系统）提供的模型列表与预标注标签列表。

将本数据集标签类别与预标注标签进行对应，其中本数据集标签类别可以对应多个模型中的多个预标注标签

## 数据集版本操作

数据集版本为数据集中可供训练模块使用一部分图片的集合；

对版本操作包括新建、删除版本，查看某一版本详情

## 版本详情

在数据集详情页点击版本操作-版本详情按钮进入版本详情页面



版本信息以列表形式展现，其中默认包含“未归入版本”和“未标注”两行。

展示版本名称、版本描述、版本包含图片数等信息

可以进行版本新建（合并）、删除操作

## 新建版本

在版本详情页点击新建版本按钮进入新建版本页面



1. 版本名称：

版本的名称，必填

1. 版本描述：

描述版本基本信息，50字以内，可选

1. 图片数量

该新增版本包含的图片数量，最大为当前未归入版本图片数量

1. 合并之前版本

可选择将之前版本合并至该版本

## 删除版本

在版本详情页点击删除版本按钮，将某版本删除，并选择是否保留该版本包含的图片。若保留，则将图片归入“未归入版本”中

# 日志管理

## 日志支持按照特定标记进行检索

日志支持按照数据集id检索所有相关日志，

按照时间段列出所有相关日志，

按照用户名列出所有相关日志。

## 日志支持设置保存天数

超过保存天数的日志自动删除日志

## 日志支持设置日志级别

级别分为：错误、告警、调试。

## 日志支持管理页面查看

日志可以在后台管理员登录管理页面时分级别查看日志，以及设定日志保存天数和日志级别。

# 兼容性

兼容浏览器： chrome

兼容屏幕分辨率，根据分辨率不同，字体相应变化：

1. 基准分辨率是 1366\*768，
2. 最低支持1280（宽度），低于该分辨率的页面出现横向滚动条；
3. 支持1440/1680/1920/2450（宽度）

# 附录

## 数据集

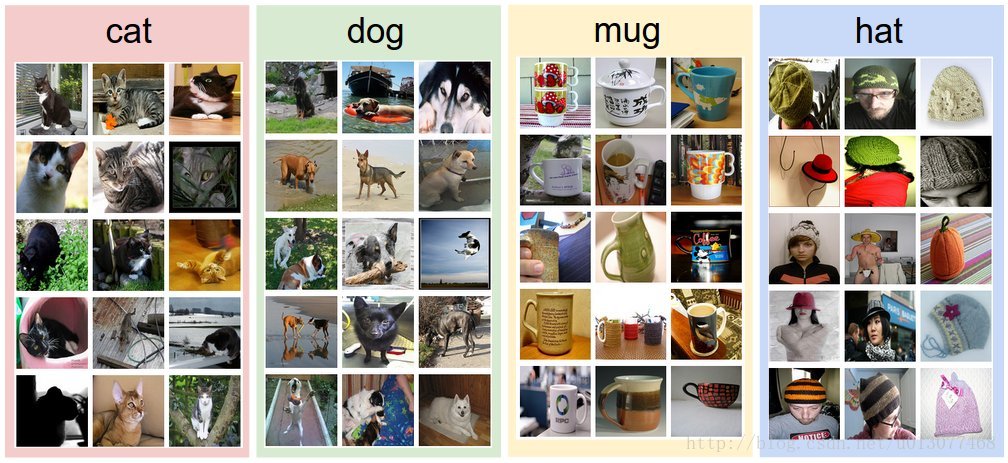
数据集是一组图片及其标注的集合。

**图片**：图片文件

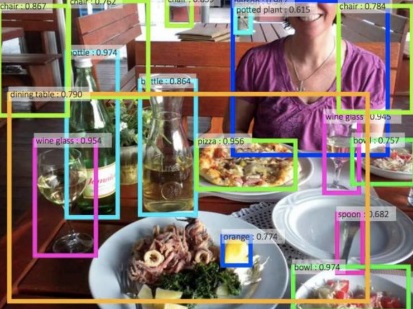
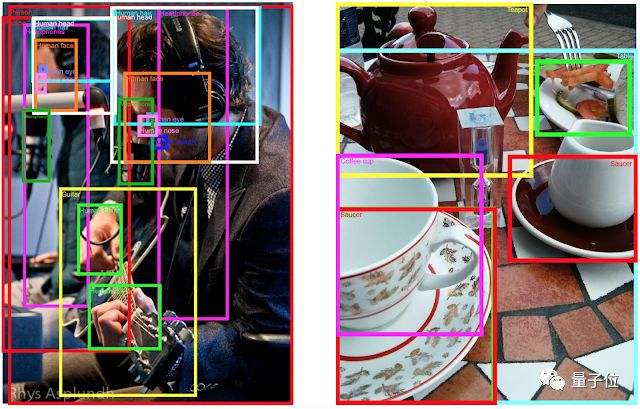
**标注**：对图片信息、内容进行说明的文件，为xml/json格式文本；

**数据集类型**：区别在于图片标注方式

分类：



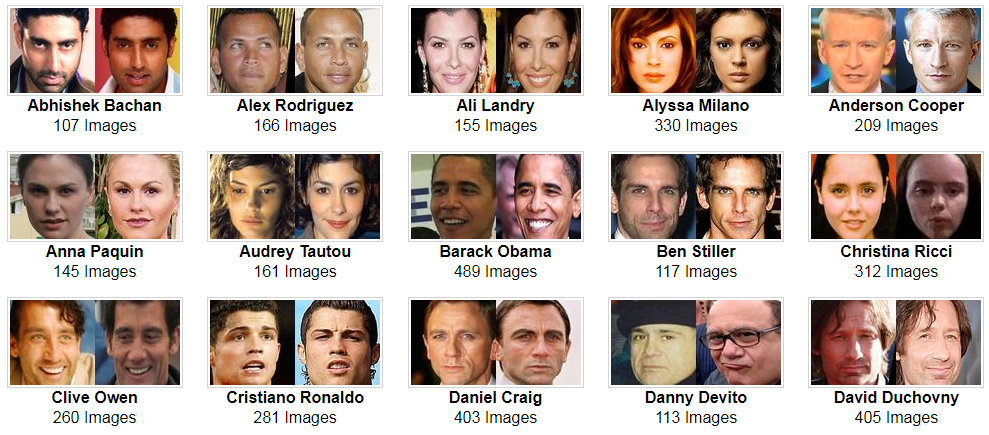
检测：

分割：



人脸识别：

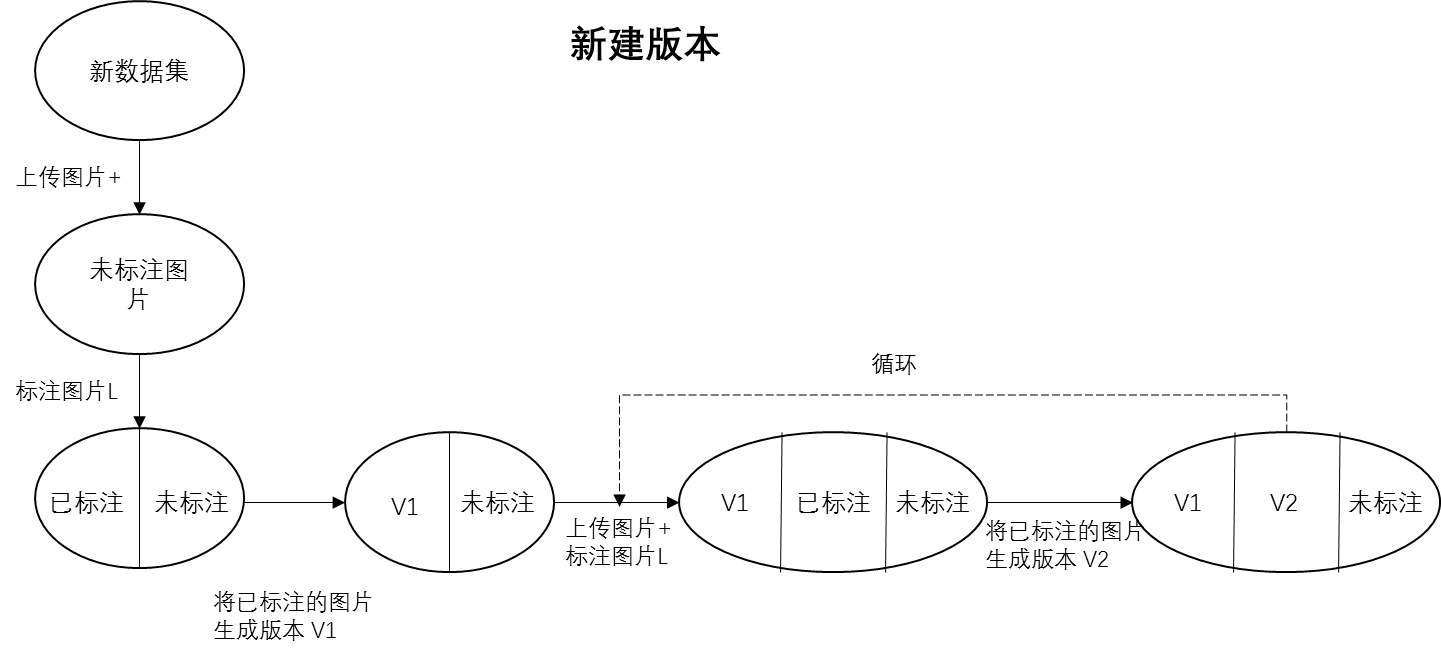


## 数据集版本

数据集版本为数据集中可供训练模块使用一部分图片的集合；

对版本操作包括新建版本、删除版本

**新建版本**：可以选择将数据集中一部分图片生成可供训练模块使用的版本。例：

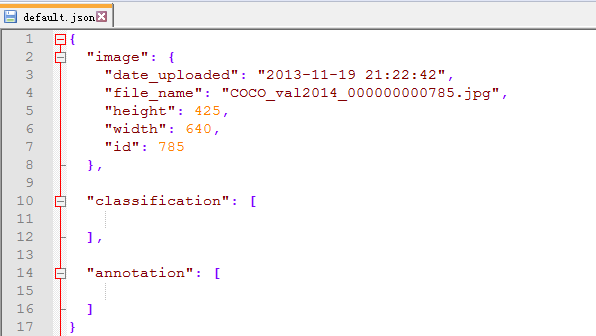


此时数据集可选2个版本V1,V2，V1与V2无重叠。

## 标注文件格式

标注文件使用json格式，MSCOCO数据集标注标准。

当上传一张未标注图片时，生成默认标注文件为：



‘image’字段包含图片信息：

1. ‘date\_uploaded’为上传时间
2. ‘file\_name’为文件名称
3. ‘height’为图片高度像素数
4. ‘width’ 为图片宽度像素数
5. ‘id’ 为图片id，在数据集中唯一

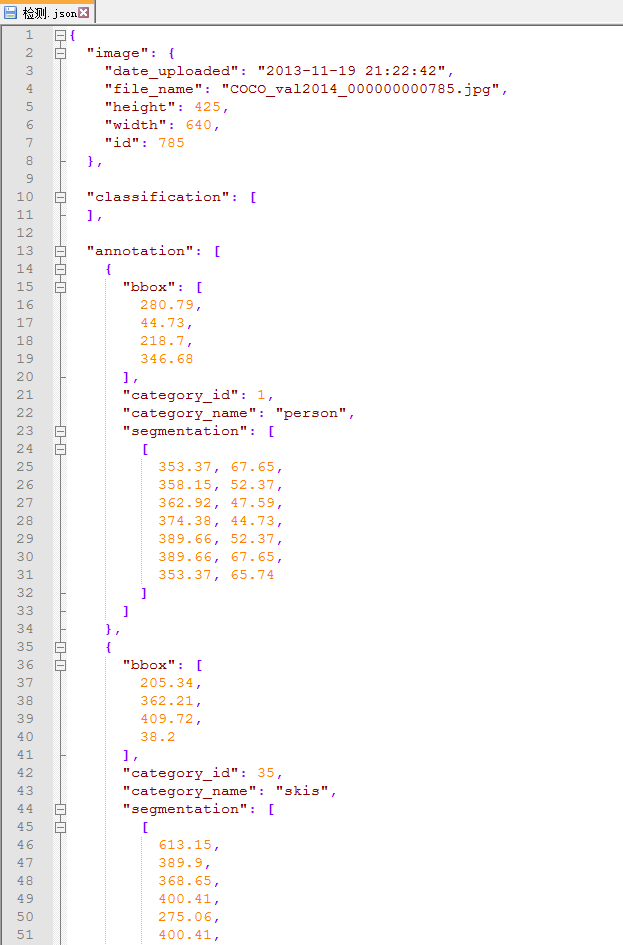
‘classification’字段包含分类/人脸识别标注信息，在分类/人脸识别数据集中使用，在检测/分割数据集中为空，例：

1. ‘category\_id’：label对应id
2. ‘category\_name’：标签名称



‘annotation’ 字段包含检测/分割标注信息，在检测分割数据集中使用，在分类/人脸识别数据集中为空，例

1. ‘bbox’为位置框坐标，格式[左上角x，左上角y，宽，高]
2. ‘category\_id’：label对应id
3. ‘category\_name’：标签名称
4. ‘segmentation’：分割标注，格式为[x1,y1, x2,y2, … xN,yN]在分割数据集中使用，检测数据集中可为空



样例文件参考：

