CN-MSLU-DEMO-1K 数据集介绍

基于 CN-MSLU-100K，我们制作了示例数据集 CN-MSLU-DEMO-1K，以供大家更好地了解数据集的特点和适用性。在本说明文件中，我们提供了数据集的基本信息，以及探索数据的示例代码。

总览

CN-MSLU-100K 数据集由 100k 张不规则遥感地块图像组成。结合《城市用地分类与规划建设用地标准》（GB 50137-2011）以及阿里巴巴高德地图 POI，我们将遥感图像所涵盖的地物按主要用途分为 5 个大类：“居住用地”、“商业服务业设施用地”、“工业产业用地”、“公共管理与公共服务设施用地”、以及“农业自然”，每个大类下又细分二级类别，共计 22 个小类。

此外，在标注过程中我们还获得了数量较少的“交通设施用地”，以及地块信息不足，难以判断的“未知土地利用”类别，一并包含在数据集中。因此，最终数据集所包含类别共计 7 大类 28 小类，具体一二级类别描述以及数量统计请参考附录所示表格。

我们从 CN-MSLU-100K 数据集的 5 个主要类别中每个类别提取了 200 张图片，制作成了 CN-MSLU-DEMO-1K 数据集。

数据集探索

文件结构

文件结构图 1 所示。

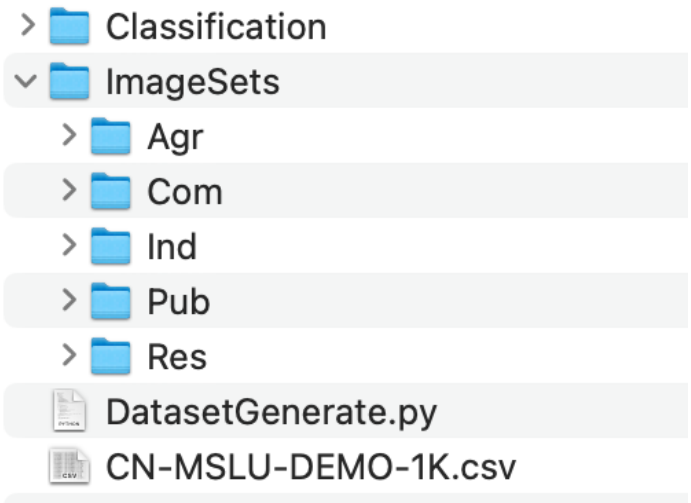

图 1 数据集文件结构

每个文件夹/文件的功能如表 1 所示。

表 1 数据集文件组织结构

文件/文件夹名		格式	说明
Classification		文件夹	存储了所有的数据说明文件（均为 xml 格式），包含了类别、路径、图像大小等信息
ImageSets		文件夹	按照类别存储的数据集原始图像
	Agr	文件夹	农业用地
	Com	文件夹	商业用地
	Ind	文件夹	工业用地
	Pub	文件夹	公共服务用地
	Res	文件夹	居住用地
DatasetGenerate.py		Python 脚本	将所有 xml 文件组织成数据集的示例代码
CN-MSLU-DEMO-1K.csv		csv	使用 DatasetGenerate.py 生成的存储数据集信息的表格。包含所有数据的类别、文件名、存储路径、图像宽度、图像高度、地理信息、一级类名、二级类名

示例代码

示例代码 DatasetGenerate.py 通过多线程的方式读取 xml 文件，提取其中所描述的每条数据的文件名、存储路径、土地利用类别等基本信息，并保存成 csv 文件。在制作数据集时，可以通过直接读取 CN-MSLU-DEMO-1K.csv 来快速对数据进行操作。

在使用代码时，可以根据所使用计算机的配置调整线程数量，以加快运行速度。

文件预览

Classification 文件夹中的数据如图 2 所示，为结构化的 XML 格式文件。XML 文件中包含了基本信息，如地块的图像路径、地块图像的大小、地块的地理信息等等，如图 3 所示。

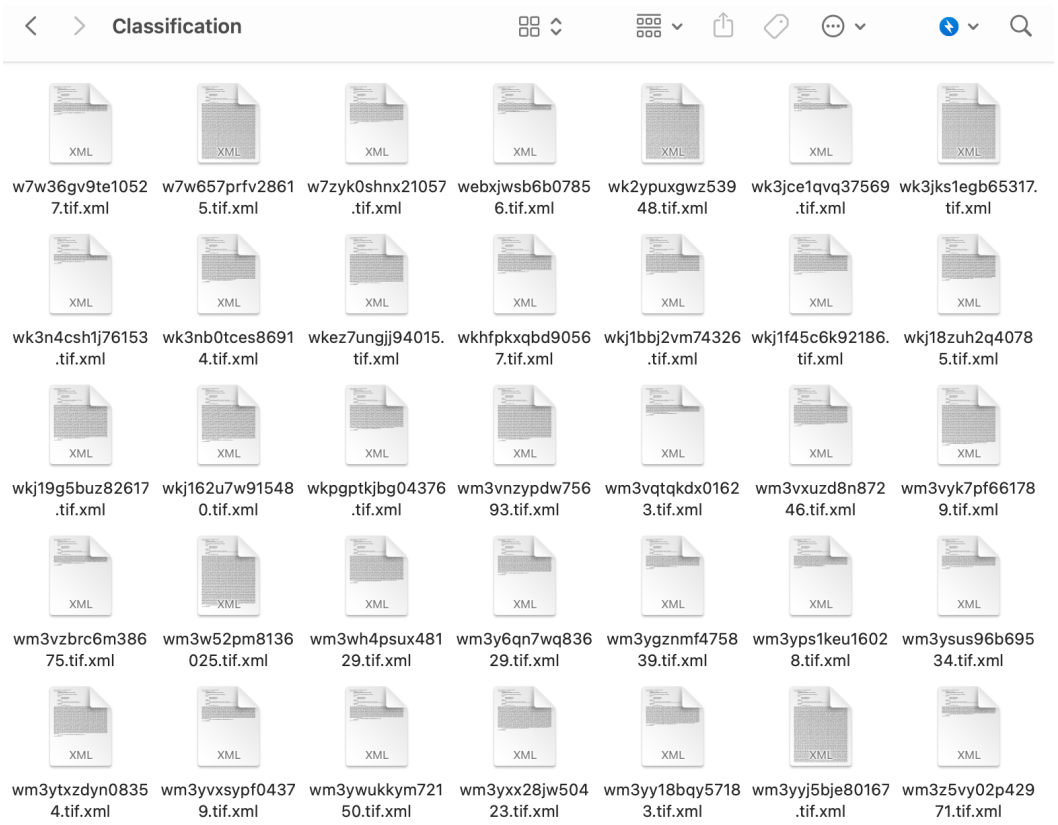

图 2 Classification 文件夹内容总览

```xml
<?xml version="1.0" encoding="utf-8"?>
<classification>
    <folder>Agr</folder>
    <filename>w7w36gv9te10527.tif</filename>
    <source>
        <database>CN-MSLU-DEMO-1K</database>
    </source>
    <size>
        <width>121</width>
        <height>123</height>
        <depth>3</depth>
    </size>
    <category>
        <firstlevel>Agriculture and Nature</firstlevel>
        <secondlevel>Forestland and Grassland</secondlevel>
    </category>
    <geoinfo>
        <coord>GCJ02</coord>
        <polygon>110.1664151,19.9286467;110.1664217,19.9286553;110.1664273,19.9286673;110.1664342,19.9287142
        ;110.1664382,19.9287241;110.1664466,19.9287361;110.1664484,19.9287416;110.1664629,19.928901;110.1664
        65,19.928912;110.1664722,19.9289269;110.1665626,19.9290575;110.1665693,19.9290654;110.1665774,19.929
        0719;110.1667532,19.9291876;110.1668248,19.929238;110.1668425,19.9292462;110.1668778,19.9292496;110.
        1689608,19.9293838;110.1689765,19.9293825;110.1689912,19.9293767;110.1690036,19.9293669;110.1690101,
        19.9293585;110.1690164,19.9293441;110.1690182,19.9293337;110.169015,19.9291995;110.168999,19.9287737
        ;110.169031,19.927771;110.1690649,19.9271387;110.1690633,19.9271225;110.169057,19.9271076;110.169050
        4,19.9270989;110.1690423,19.9270918;110.1690328,19.9270864;110.1690171,19.9270821;110.1665748,19.926
        8008;110.1665643,19.926801;110.1665491,19.926805;110.1664252,19.9268579;110.1664129,19.9268671;110.1
        664036,19.9268793;110.1663995,19.9268887;110.1663969,19.9269038;110.1663761,19.9276046;110.1664012,1
        9.9286119;110.1664026,19.9286227;110.1664061,19.9286331;110.1664151,19.9286467</polygon>
        <center>110.16771178050843,19.92810001608732</center>
    </geoinfo>
</classification>
```

图 3 XML 格式文件记录数据内容总览

`ImageSets` 文件夹中的数据如下图所示，遥感影像地块根据类别存储在不同的文件夹中，每个文件夹的名称都是对应类别的简称。

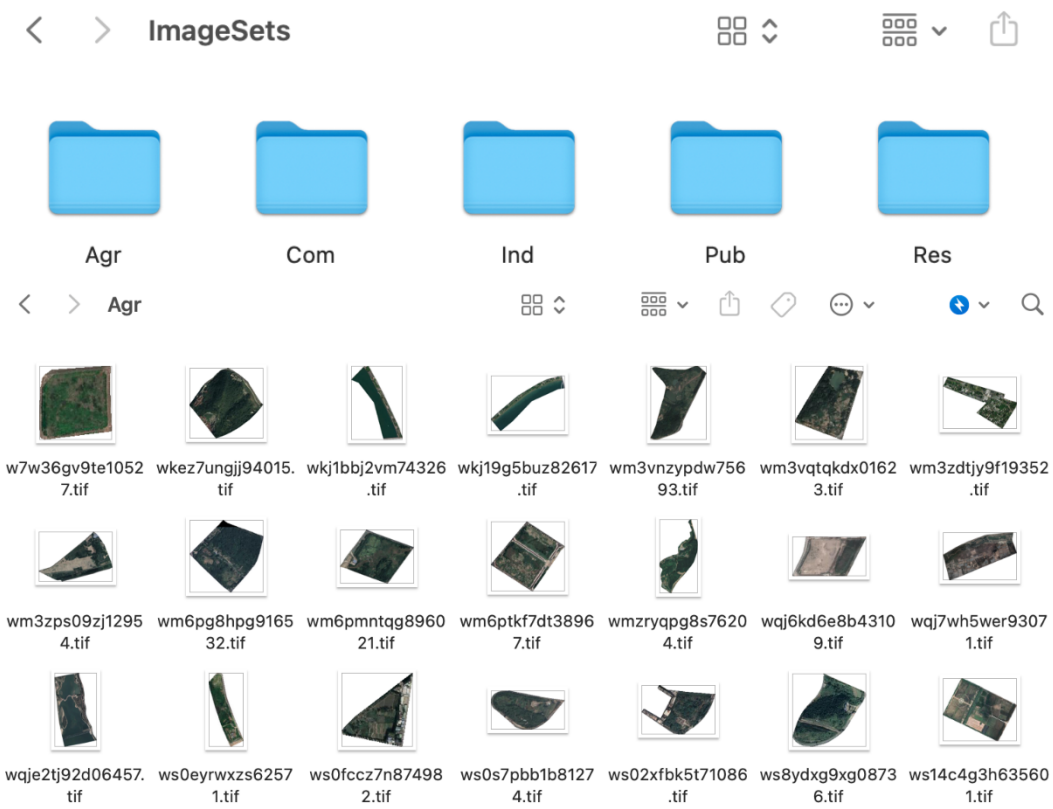

图 4 遥感影像存储组织形式总览

相关研究

以下推文将会详细介绍我们正在使用的数据集，以及我们在该数据集基础上开展的项目。推文中包含了相关负责人的电子邮箱，如果您有任何问题，欢迎联系我们！

CN-MSLU-100K：可支持多源时空大数据的地块（社区）尺度全国土地利用类别数据集 - 城市之光 – City of Light (urbancomp.net)

附录：CN-MSLU-100K 数据集一二级类别描述以及数量统计

一级类别	二级类别	数据数量
居住用地 (Res) 40682	农村宅基地	1549
	农村建筑与耕地	14148
	多层和高层住宅	20884
	别墅;高档住宅	1864
	城中村	2237
商业服务业设施用地 (Com) 6684	写字楼;商务大厦	978
	商业娱乐	588
	商务办公楼;园区	2708
	商贸市场	1125
	购物中心;商业街	1125
	酒店宾馆	160
工业产业用地 (Ind) 24498	工业园;工厂	21593
	建设用地	2904
公共管理与公共服务设施用地 (Pub) 6286	党政机关;事业单位	719
	公共服务场所（博物馆; 体育馆; 医院)	917
	教育科研院所	2580
	公园广场	2070
农业自然 (Agr) 21411	山体	2484
	林地;草地	6916
	水体	2260
	耕地	7293
	荒地	2458
交通设施用地 (Tra) 799	交通场所（停车场;加油站;服务区)	290
	交通枢纽（地铁站;汽车站;火车站;机场)	366
	道路	143
跳过 (Unk) 25069	信息不足	5753
	地块无效（狭长地块)	2776
	混合类型	16540