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 A.1。

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。

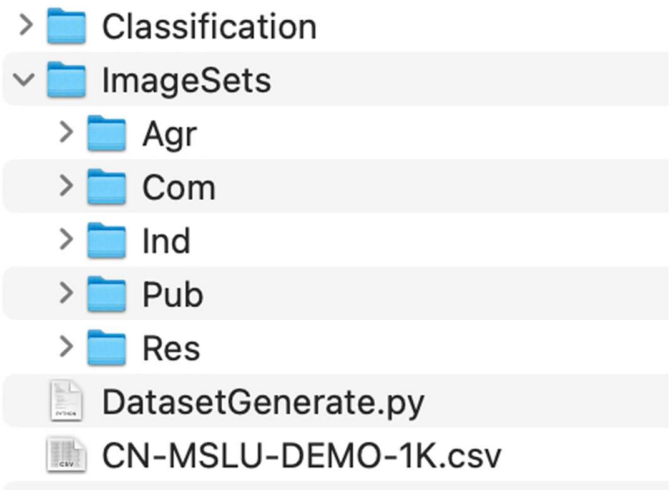

图 1 数据集文件结构

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	存储样本的元数据文件（xml 格式）。文件中包含了样本的类别、路径、图像大小等信息
ImageSets		文件夹	样本数据集，包含各土地利用类别的遥感影像
	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	根据 xml 文件生成数据集表格的示例代码
CN-MSLU-DEMO-1K.csv		csv	数据集表格。运行 DatasetGenerate.py 生成。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

样本元数据

Classification 文件夹中存储所有样本的元数据（XML 格式），如图 2 所示。XML 文件名对应样本数据名，文件中存储了样本的基本信息，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。XML 文件所包含的属性及其含义如附录中表 A.2 所示。

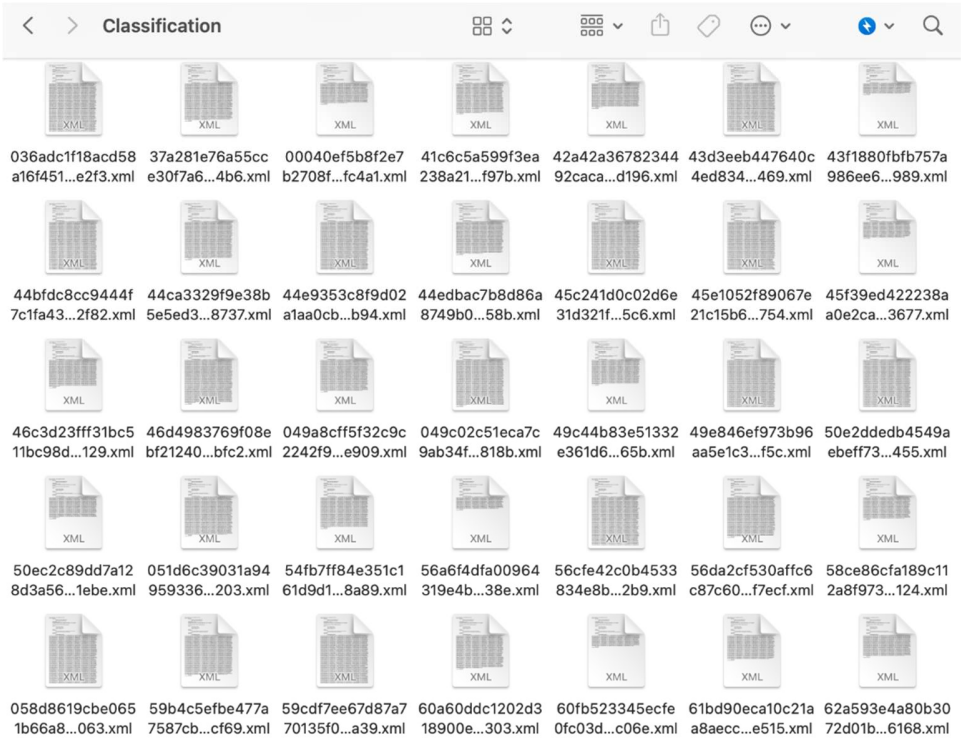

图 2 Classification 文件夹内容总览

```
<?xml version='1.0' encoding='utf-8'?>
<classification>
    <quality>
        <modelscore>4.0</modelscore>
        <softscore>5.0</softscore>
    </quality>
    <folder>Agr</folder>
    <filename>f74b9fb321a2a70f5ba270f919092006.tif</filename>
    <source>
        <database>CN-MSLU-1K</database>
    </source>
    <size>
        <width>213</width>
        <height>248</height>
        <depth>3</depth>
    </size>
    <category>
        <firstlevel>Agriculture and Nature</firstlevel>
        <secondlevel>Forestland and Grassland</secondlevel>
    </category>
    <geoinfo>
        <coord>GCJ02</coord>
        <coord>116.6118104806397, 30.6278174706768;116.6129050233956, 30.6268
30.623756406032946;116.61178828601524, 30.626791072896875;116.6117878861096
30.627230718467597;116.61156213943006, 30.627233818772435;116.6113737839843
30.627691448334335;116.61179388456064, 30.627812073959372;116.6118104806397,
    </geoinfo>
</classification>
```

图 3 XML 格式文件记录数据内容总览

数据质量评级：考虑到数据集的可用性，我们使用已有模型和软分类方法对数据集质量进行了评级，并将评级结果存储到 XML 文件的 quality 字段中。评级取值范围从 0 到 5，表示数据的质量逐渐提高：0 级代表未评级；1 级代表 AOI 过大，人难以识别；2 级代表模型难以正确识别；3 级代表基于 100K 数据集构建的模型可以正确识别；4 级代表基于 10K 数据集构建的模型可以正确识别；5 级代表基于 1K 数据集构建的模型可以正确识别。两种方法得到的分类情况详见附录表 A.3 和表 A.4。

样本数据集

ImageSets 文件夹中的数据如图 4 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。

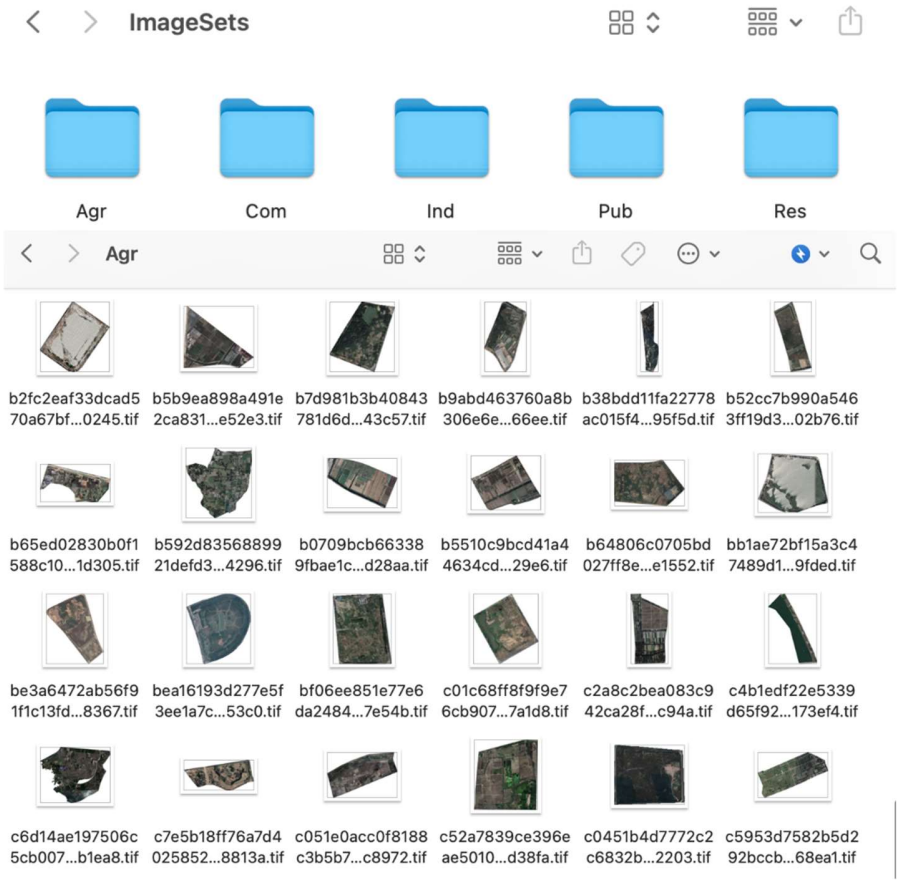

图 4 遥感影像存储组织形式总览

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。

代码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。

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

表 A.1 一二级类别对应关系以及数据数量说明

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

表 A.2 XML 文件所包含的属性及其含义

属性名		含义
quality	modelscore	用已有模型对数据集质量进行分类得到的等级
	softscore	用软分类方法对数据集质量进行分类得到的等级
folder		数据所在文件夹名称
filename		数据名称
source	database	数据集名称
size	width	数据宽度
	height	数据高度
	depth	数据深度
category	firstlevel	数据集一级类别
	secondlevel	数据集二级类别
polygon		包括边数、顶点数、内角和、外角和以及对角线数，用于描述多边形的形状和特征

表 A.3 已有模型对数据集质量进行分类得到的等级在各类别中的分布

一级类别	二级类别	5.0	4.0	3.0	2.0	1.0	0.0
居住用地 (Res)	农村宅基地	317	40	79	447	6	660
	农村建筑与耕地	773	145	348	9603	1391	1888
	多层和高层住宅	1105	42	40	230	0	447
	别墅;高档住宅	9768	583	594	4118	2	5819
	城中村	620	64	72	561	0	920
商业服务业设施用地 (Com)	写字楼;商务大厦	124	65	40	593	0	156
	商业娱乐	26	8	6	428	32	88
	商务办公楼;园区	279	89	55	1611	0	674
	商贸市场	101	26	25	708	0	265
	购物中心;商业街	218	55	38	612	0	202
	酒店宾馆	6	3	4	98	0	49
工业产业用地(Ind)	工业园;工厂	9265	939	1166	2985	9	7230
	建设用地	545	74	117	1378	0	790
公共管理与公共服务设施用地 (Pub)	党政机关;事业单位	163	66	23	326	0	141
	公共服务场所（博物馆; 体育馆; 医院）	247	104	28	367	0	171
	教育科研院所	922	343	38	734	1	542
	公园广场	460	91	16	1191	16	296
农业自然 (Agr)	山体	1851	151	12	31	0	439
	林地;草地	2040	2436	330	125	0	1985
	水体	1004	643	124	30	0	459
	耕地	4495	856	147	129	0	1666
	荒地	728	615	136	113	0	866
交通设施用地(Tra)	交通场所（停车场;加油站;服务区）	0	0	0	198	0	92
	交通枢纽（地铁站;汽车站;火车站;机场）	0	0	0	322	4	40
	道路	0	0	0	51	0	49
跳过(Unk)	信息不足	0	0	0	0	4	5749
	地块无效（狭长地块）	0	0	0	0	0	2776
	混合类型	0	0	0	0	7	16533
		35057	7438	3438	26989	1472	51035

表 A.4 软分类方法对数据集质量进行分类得到的等级在各类别中的分布

一级类别	二级类别	5.0	4.0	3.0	2.0	1.0	0.0
居住用地 (Res)	农村宅基地	311	50	80	442	6	660
	农村建筑与耕地	774	161	369	9564	1390	1890
	多层和高层住宅	1106	42	39	230	0	447
	别墅;高档住宅	9776	601	592	4094	2	5819
	城中村	620	64	77	556	0	920
商业服务业设施用地 (Com)	写字楼;商务大厦	126	68	40	588	0	156
	商业娱乐	26	9	6	427	32	88
	商务办公楼;园区	279	98	64	1593	0	674
	商贸市场	104	30	25	701	0	265
	购物中心;商业街	220	56	38	609	0	202
	酒店宾馆	6	3	4	98	0	49
工业产业用地(Ind)	工业园;工厂	8227	1535	1745	2848	9	7230
	建设用地	321	198	237	1358	0	790
公共管理与公共服务设施用地 (Pub)	党政机关;事业单位	134	82	41	321	0	141
	公共服务场所 (博物馆; 体育馆; 医院)	232	111	38	365	0	171
	教育科研院所	887	365	60	725	1	542
	公园广场	390	125	57	1186	16	296
农业自然 (Agr)	山体	1902	104	11	28	0	439
	林地;草地	4115	603	100	113	0	1985
	水体	1497	227	50	27	0	459
	耕地	4742	635	123	127	0	1666
	荒地	1061	316	107	108	0	866
交通设施用地(Tra)	交通场所 (停车场;加油站;服务区)	0	0	0	198	0	92
	交通枢纽 (地铁站;汽车站;火车站;机场)	0	0	0	322	4	40
	道路	0	0	0	51	0	92
跳过(Unk)	信息不足	0	0	0	0	0	5753
	地块无效 (狭长地块)	0	0	0	0	0	2776
	混合类型	0	0	0	0	0	16540
		36856	5483	3903	26679	1460	51048