UrbanComp 位置智能和城市感知课题组
UrbanComp Lab for Location Intelligence and Urban Sensing

CN-MSLU-DEMO-1k Dataset Description

Based on CN-MSLU-100K, we have created the example dataset CN-MSLU-DEMO-1k for you to better understand the characteristics and applicability of the dataset. In this description file, we provide basic information about the dataset, as well as sample code for exploring the data.

Overview

The CN-MSLU-100k dataset consists of over 100,000 irregular remote-sensing land parcel images. Combining the "Classification and Planning Standards for Urban Land Use" (GB 50137-2011) and Alibaba's "AMAP POI", we have categorized the main features depicted in the remote sensing images into 5 major classes as "Residential Districts", "Commercial Zones", "Industrial Land", "Public Services", "Agriculture and Nature". Each major category is subdivided into secondary categories, totaling 22 sub-categories.

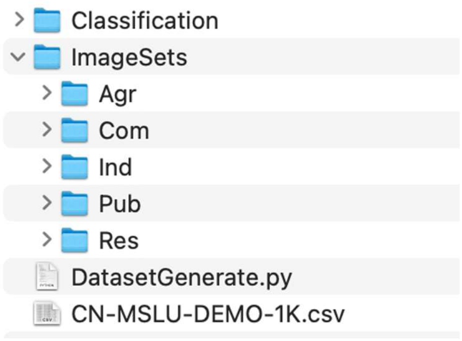
In addition, during the labelling process, we also obtained a smaller number of "Transportation Facilities", and large amount of "Unknow Landuse" categories which are difficult to judge due to insufficient information on land parcels, and included them in the dataset. The final dataset contains 7 categories and 28 sub-categories, please refer to the Table A.1 in the Appendix for the description and count number of first and second level categories.

In the CN-MSLU-100k dataset, we extracted 200 images from each of the five main categories to produce the CN-MSLU-DEMO-1k dataset for a better understanding of the characteristics and applicability of the dataset.

Dataset stats

File directory structure

The file structure is shown in Figure 1 as follows:



```
> Classification
v ImageSets
    > Agr
    > Com
    > Ind
    > Pub
    > Res
  DatasetGenerate.py
  CN-MSLU-DEMO-1K.csv
```

Figure 1 Dataset File Structure

The description for each folder and file is shown in Table 1。

Table 1 The description for each folder and file

Folder or file name		**Format**	**Description**
Classification		Folder	The metadata file stores information about the samples in XML format, including the sample category, path, and image size
ImageSets		Folder	A sample dataset containing remote sensing images for various land use categories
	Agr	Folder	Agriculture and Nature
	Com	Folder	Commercial Zones
	Ind	Folder	Industrial Land
	Pub	Folder	Public Services
	Res	Folder	Residential Districts
DatasetGenerate.py		Python Scrpit	code for generating a dataset table from XML file
CN-MSLU-DEMO-1K.csv		csv	Dataset table. Run DatasetGenerate.py build Contains all data categories, file names, storage paths, image widths, image heights, geographic information, first-level class names, and second-level class names

Sample metadata

The metadata (in XML format) for all samples is stored in the `Classification` folder, as shown in Figure 2. The XML file name corresponds to the sample data name, and the basic information of the sample is stored in the file, such as the image path of the plot, the size of the plot image, and the geographic information of the plot, as shown in Figure 3. The attributes contained in the XML file and their meanings are shown in Table A.2 in the Appendix.

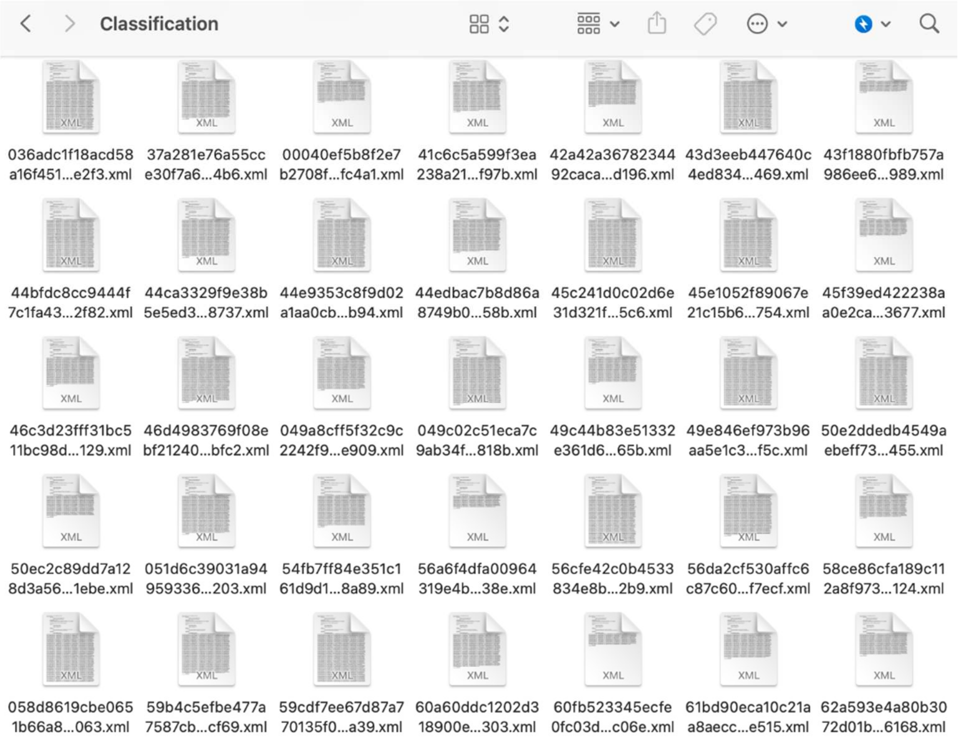
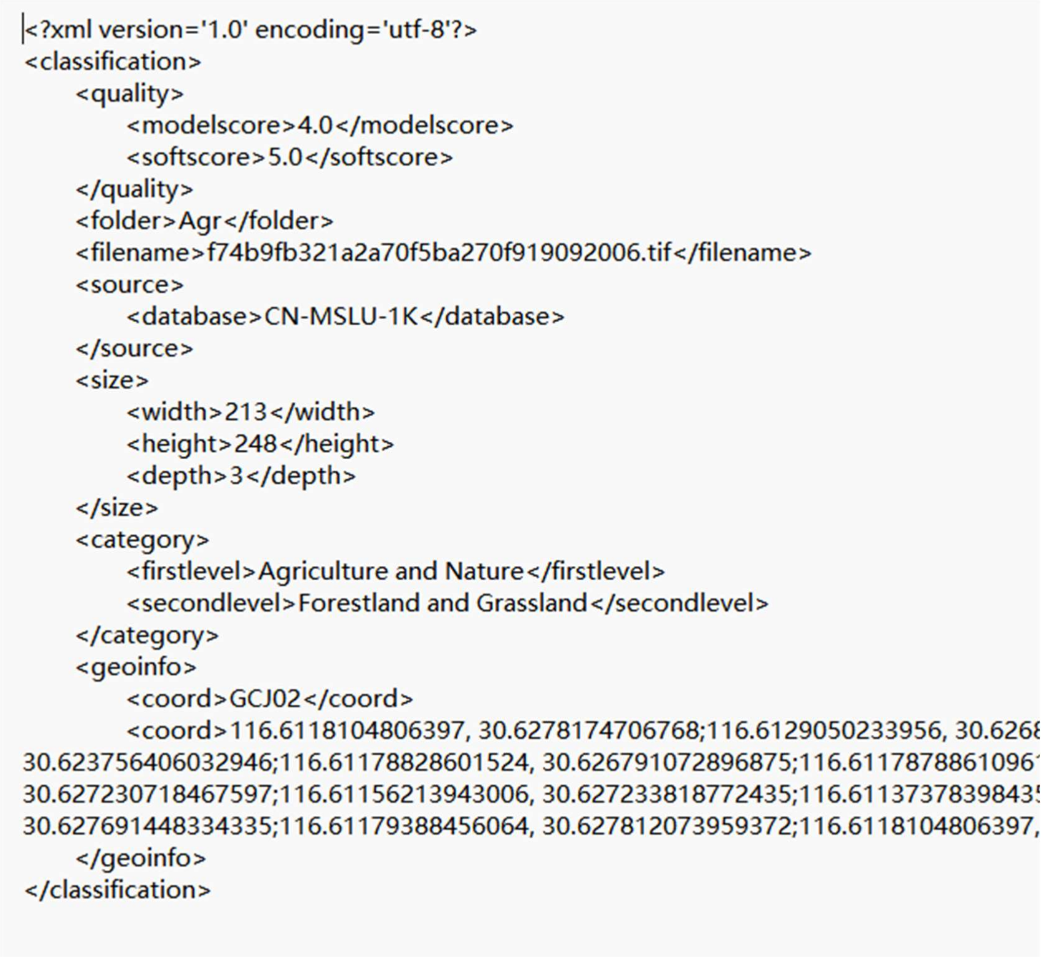

Classification
036adc1f18acd58a16f451...e2f3.xml
37a281e76a55cce30f7a6...4b6.xml
00040ef5b8f2e7b2708f...fc4a1.xml
41c6c5a599f3ea238a21...f97b.xml
42a42a3678234492caca...d196.xml
43d3eeb447640c4ed834...469.xml
43f1880fbfb757a986ee6...989.xml
44bfdc8cc9444f7c1fa43...2f82.xml
44ca3329f9e38b5e5ed3...8737.xml
44e9353c8f9d02a1aa0cb...b94.xml
44edbac7b8d86a8749b0...58b.xml
45c241d0c02d6e31d321f...5c6.xml
45e1052f89067e21c15b6...754.xml
45f39ed422238aa0e2ca...3677.xml
46c3d23fff31bc511bc98d...129.xml
46d4983769f08ebf21240...bfc2.xml
049a8cff5f32c9c2242f9...e909.xml
049c02c51eca7c9ab34f...818b.xml
49c44b83e51332e361d6...65b.xml
49e846ef973b96aa5e1c3...f5c.xml
50e2ddedb4549aebeff73...455.xml
50ec2c89dd7a128d3a56...1ebe.xml
051d6c39031a94959336...203.xml
54fb7ff84e351c161d9d1...8a89.xml
56a6f4dfa00964319e4b...38e.xml
56cfe42c0b4533834e8b...2b9.xml
56da2cf530affc6c87c60...f7ecf.xml
58ce86cfa189c112a8f973...124.xml
058d8619cbe0651b66a8...063.xml
59b4c5efbe477a7587cb...cf69.xml
59cdf7ee67d87a770135f0...a39.xml
60a60ddc1202d318900e...303.xml
60fb523345ecfe0fc03d...c06e.xml
61bd90eca10c21aa8aecc...e515.xml
62a593e4a80b3072d01b...6168.xml

Figure 2 Overview of xml files in Classification folder



```
<?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>
```

Figure 3 Overview of xml format file record data content

Data quality rating: Considering the availability of the dataset, we rated the quality of the dataset using existing models and soft classification methods, and stored the rating results in the quality field of the XML file. The rating range is from 0 to 5, indicating that the quality of the data is gradually improving: a grade of 0 indicates that it is not rated; Level 1 means that the AOI is too large for people to identify; Level 2 representative models are difficult to identify correctly; Level 3 means that the model built based on the 100K dataset can be correctly identified; Level 4 means that the model built based on the 10K dataset can be correctly identified; Level 5 means that the model built on the 1K dataset can be correctly recognized. The classifications obtained by the two methods are detailed in Table A.3 and Table A.4 in the Appendix.

Sample datasets

The data in the `ImageSets` folder is shown in Figure 4. Remote sensing image parcels are stored in different folders according to their categories, and the name of each folder is an abbreviation of the corresponding category.

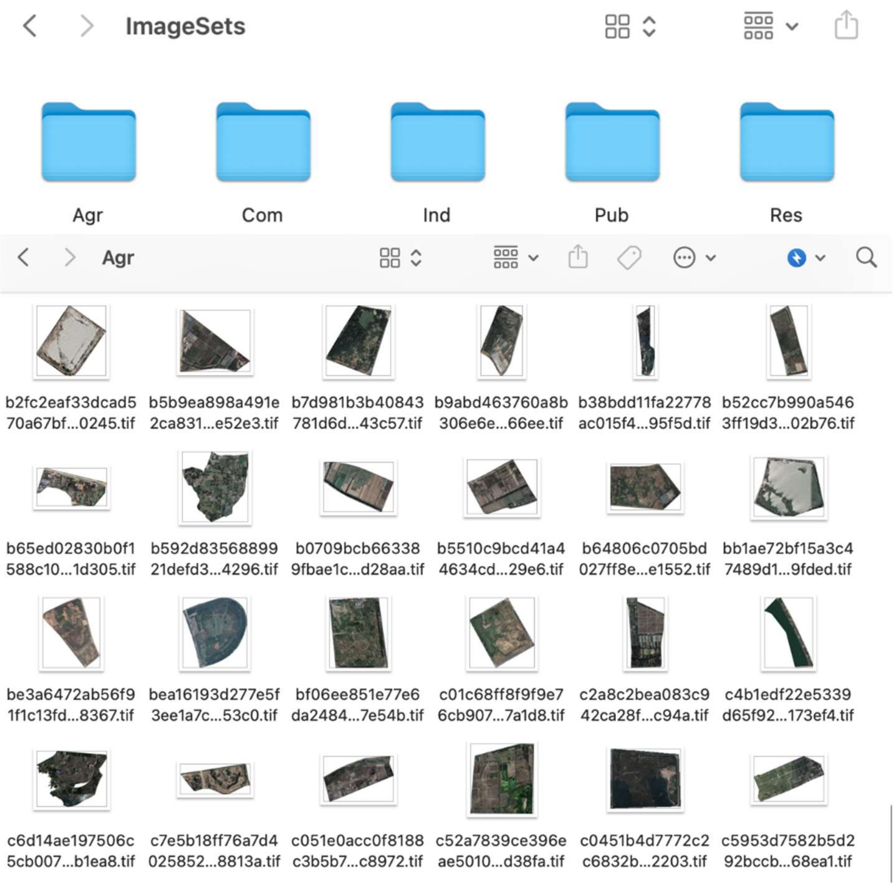

ImageSets
Agr
Com
Ind
Pub
Res
Agr
b2fc2eaf33dcad5 70a67bf...0245.tif
b5b9ea898a491e 2ca831...e52e3.tif
b7d981b3b40843 781d6d...43c57.tif
b9abd463760a8b 306e6e...66ee.tif
b38bdd11fa22778 ac015f4...95f5d.tif
b52cc7b990a546 3ff19d3...02b76.tif
b65ed02830b0f1 588c10...1d305.tif
b592d83568899 21defd3...4296.tif
b0709bcb66338 9fbae1c...d28aa.tif
b5510c9bcd41a4 4634cd...29e6.tif
b64806c0705bd 027ff8e...e1552.tif
bb1ae72bf15a3c4 7489d1...9fded.tif
be3a6472ab56f9 1f1c13fd...8367.tif
bea16193d277e5f 3ee1a7c...53c0.tif
bf06ee851e77e6 da2484...7e54b.tif
c01c68ff8f9f9e7 6cb907...7a1d8.tif
c2a8c2bea083c9 42ca28f...c94a.tif
c4b1edf22e5339 d65f92...173ef4.tif
c6d14ae197506c 5cb007...b1ea8.tif
c7e5b18ff76a7d4 025852...8813a.tif
c051e0acc0f8188 c3b5b7...c8972.tif
c52a7839ce396e ae5010...d38fa.tif
c0451b4d7772c2 c6832b...2203.tif
c5953d7582b5d2 92bccb...68ea1.tif

Figure 4 Overview of Remote sensing image parcels in ImageSets folder

Demo code

The code `DatasetGenerate.py` reads the xml file in a multi-threaded way, extracts the basic information such as file name, storage path, land use category, etc. of each piece of data described in it, and saves it as a csv file. When making the dataset, the data can be quickly manipulated by reading `CN-MSLU-DEMO-1K.csv` directly.

Related projects

The following Link will go into more detail about the dataset we are using and the projects we are working on based on that dataset. It also include the email addresses of the people responsible, so if you have any questions, please feel free to contact us!

CN-MSLU-100K: Land Use Classification Dataset at Block Scale for Multi-source Spatio-temporal Data - 城市之光 – City of Light (urbancomp.net)

Appendix

Table A.1 correspondence between first and second-level categories and explanation of data quantity

First Level Category	Second Level Category	The amount of data
Residential Districts (Res) 40682	Rural Homestead	1549
	Rural Architecture and Farmland	14148
	High-rise Residential Buildings	20884
	Villas and High-end Residences	1864
	Urban Villages	2237
Commercial Zones (Com) 6684	Business Tower	978
	Commercial Entertainment	588
	Office Campus	2708
	Commercial Market	1125
	Shopping Center and Commercial Street	1125
	Hotel	160
Industrial Land (Ind) 24498	Industrial Park and Factory	21593
	Construction Site	2904
Public Services (Pub) 6286	Party and Government Institutions	719
	Non-profit Public Institutions (Museum; Stadium; Hospital)	917
	Educational and Research Institutions	2580
	Parks and Squares	2070
Agriculture and Nature (Agr) 21411	Mountain	2484
	Forestland and Grassland	6916
	Water	2260
	Farmland	7293
	Wasteland	2458
Transportation Facilities (Tra) 799	Transport facilities (Car Park; Gas Station; Service Station)	290
	Transportation hub (Subway; Bus or Train Station; Airport)	366
	Highway & Track	143
Unknow Landuse (Unk) 25069	Lack of Information	5753
	Invalid Land Parcel (Small-sized & Narrow)	2776
	Mixed Landuse	16540

Table A.2 XML attributes and their meanings

The name of the property		Meaning
quality	modelscore	The grade obtained by classifying the quality of the dataset using an existing model
	softscore	The grade obtained by classifying the quality of the dataset by using the soft classification method
folder		The name of the folder where the data resides
filename		The name of the data
source	database	Dataset name
size	width	Data width
	height	Data height
	depth	Data depth
category	firstlevel	Dataset Level 1 category
	secondlevel	Dataset Level 2 category
polygon		Includes the number of sides, vertices, sum of interior corners, sum of outer corners, and number of diagonals, and is used to describe the shape and features of a polygon

Table A.3 Model-derived quality ratings distribution across different categories

First Level	Second Level	5.0	4.0	3.0	2.0	1.0	0.0
(Res)	Rural Homestead	317	40	79	447	6	660
	Rural Architecture and Farmland	773	145	348	9603	1391	1888
	High-rise Residential Buildings	1105	42	40	230	0	447
	Villas and High-end Residences	9768	583	594	4118	2	5819
	Urban Villages	620	64	72	561	0	920
(Com)	Business Tower	124	65	40	593	0	156
	Commercial Entertainment	26	8	6	428	32	88
	Office Campus	279	89	55	1611	0	674
	Commercial Market	101	26	25	708	0	265
	Shopping Center and Commercial Street	218	55	38	612	0	202
	Hotel	6	3	4	98	0	49
(Ind)	Industrial Park and Factory	9265	939	1166	2985	9	7230
	Construction Site	545	74	117	1378	0	790
(Pub)	Party and Government Institutions	163	66	23	326	0	141
	Non-profit Public Institutions (Museum; Stadium; Hospital)	247	104	28	367	0	171
	Educational and Research Institutions	922	343	38	734	1	542
	Parks and Squares	460	91	16	1191	16	296
(Agr)	Mountain	1851	151	12	31	0	439
	Forestland and Grassland	2040	2436	330	125	0	1985
	Water	1004	643	124	30	0	459
	Farmland	4495	856	147	129	0	1666
	Wasteland	728	615	136	113	0	866
(Tra)	Transport facilities (Car Park; Gas Station; Service Station)	0	0	0	198	0	92
	Transportation hub (Subway; Bus or Train Station; Airport)	0	0	0	322	4	40
	Highway & Track	0	0	0	51	0	49
(Unk)	Lack of Information	0	0	0	0	4	5749
	Invalid Land Parcel (Small-sized & Narrow)	0	0	0	0	0	2776
	Mixed Landuse	0	0	0	0	7	16533
		35057	7438	3438	26989	1472	51035

Table A.4 Soft classification method quality ratings distribution across different categories

First Level	Second Level	5.0	4.0	3.0	2.0	1.0	0.0
(Res)	Rural Homestead	311	50	80	442	6	660
	Rural Architecture and Farmland	774	161	369	9564	1390	1890
	High-rise Residential Buildings	1106	42	39	230	0	447
	Villas and High-end Residences	9776	601	592	4094	2	5819
	Urban Villages	620	64	77	556	0	920
(Com)	Business Tower	126	68	40	588	0	156
	Commercial Entertainment	26	9	6	427	32	88
	Office Campus	279	98	64	1593	0	674
	Commercial Market	104	30	25	701	0	265
	Shopping Center and Commercial Street	220	56	38	609	0	202
	Hotel	6	3	4	98	0	49
(Ind)	Industrial Park and Factory	8227	1535	1745	2848	9	7230
	Construction Site	321	198	237	1358	0	790
(Pub)	Party and Government Institutions	134	82	41	321	0	141
	Non-profit Public Institutions (Museum; Stadium; Hospital)	232	111	38	365	0	171
	Educational and Research Institutions	887	365	60	725	1	542
	Parks and Squares	390	125	57	1186	16	296
(Agr)	Mountain	1902	104	11	28	0	439
	Forestland and Grassland	4115	603	100	113	0	1985
	Water	1497	227	50	27	0	459
	Farmland	4742	635	123	127	0	1666
	Wasteland	1061	316	107	108	0	866
(Tra)	Transport facilities (Car Park; Gas Station; Service Station)	0	0	0	198	0	92
	Transportation hub (Subway; Bus or Train Station; Airport)	0	0	0	322	4	40
	Highway & Track	0	0	0	51	0	92
(Unk)	Lack of Information	0	0	0	0	0	5753
	Invalid Land Parcel (Small-sized & Narrow)	0	0	0	0	0	2776
	Mixed Landuse	0	0	0	0	0	16540
		36856	5483	3903	26679	1460	51048