Datasheet for LoveDA dataset

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, Yanfei Zhong
State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote Sensing
Wuhan University, Wuhan 430074, China
{kingdrone,zhengzhuo,maailong007,luxiaoyan,zhongyanfei}@whu.edu.cn

1 Motivation

- **For what purpose was the dataset created?** The Land-cOVEr Domain Adaptive semantic segmentation (LoveDA) dataset was created to provide land-cover semantic segmentation and unsupervised domain adaptation (UDA) tasks. Exploring the use of deep transfer learning methods on this dataset will be a meaningful way to promote city-level or national-level land-cover mapping.
- **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?** The initial version of the dataset was created by Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong, most of whom were researchers at the Wuhan University.

2 Composition

- **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** Each instance is an 0.3m image and corresponding semantic mask. The pixel value in the mask represents land-cover types: background – 1, building – 2, road – 3, water – 4, barren – 5,forest – 6, agriculture – 7. And the no-data regions were assigned 0 which should be ignored.
- **How many instances are there in total (of each type, if appropriate)?** The LoveDA contains 5987 HSR images with 166768 annotated objects from Nanjing, Changzhou and Wuhan cities. There are 2713 images for urban scenes and 3274 images for rural scenes.
- **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** The dataset does not contain all possible instances. The larger set covers the total Nanjing, Changzhou and Wuhan cities. The LoveDA is representative of the larger set for two reasons: **1) Representative sampling.** The images were collected from 18 spatially independent areas, covering 18 administrative districts in different cities. **2) Large geographic coverage.** The LoveDA dataset was constructed using 0.3 m HSR images obtained from Nanjing, Changzhou and Wuhan in July 2016, covering $536.15\mathrm{km}^2$. In future research, more countries and typical cities need to be considered.
- **What data does each instance consist of?** Each instance contains an image and corresponding semantic mask that are 1024 by 1024 pixels in PNG format.
- **Is there a label or target associated with each instance?** Each image has one corresponding semantic mask.
- **Is any information missing from individual instances?** Everything is included in the dataset.
- **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** The instances were obtained from 18 representative areas, so the images in the same area are spatially related.

Preprint. Under review.

- **Are there recommended data splits (e.g., training, development/validation, testing)?** The dataset comes with specified train, val, and test splits with the collected areas. The data is split into two tasks, semantic segmentation, and unsupervised domain adaptation. **1) Semantic segmentation**. There are eight areas for training, and the others are for validation and testing. The training, validation, and test sets cover both urban and rural areas.**2) Unsupervised domain adaptation**. The UDA process considers two cross-domain adaptation sub-tasks: *a) **Urban** → Rural.* The images from the Qinhuai, Qixia, Jianghan, and Gulou areas are included in the source training set. The images from Liuhe and Huangpi are included in the validation set. The Jiangning, Xinbei, and Liyang images included in the test set. The *Oracle* setting is designed to test the upper limit of accuracy in a single domain [4]. Hence, the training images were collected from the Pukou, Lishui, Gaochun, and Jiangxia areas. *b) **Rural** → Urban.* The images from the Pukou, Lishui, Gaochun, and Jiangxia areas are included in the source training set. The images from Yuhuatai and Jintan are used for the validation set. The Jiangye, Wuchang, and Wujin images are used for the test set. In the *Oracle* setting, the training images cover the Qinhuai, Qixia, Jianghan, and Gulou areas.

Domain	City	Region	#Images	Train	Val	Test
Urban	Nanjing	Qixia	320	✓		
		Gulou	320	✓		
		Qinhuai	336	✓		
		Yuhuatai	357		✓	
		Jianye	357			✓
	Changzhou	Jintan	320		✓	
		Wujin	320			✓
	Wuhan	Jianghan	180	✓		
		Wuchang	143			✓
Rural	Nanjing	Pukou	320	✓		
		Gaochun	336	✓		
		Lishui	336	✓		
		Liuhe	320		✓	
		Jiangning	336			✓
	Changzhou	Liyang	320			✓
		Xinbei	320			✓
	Wuhan	Jiangxia	374	✓		
		Huangpi	672		✓	
		Total	5987	2522	1669	1796

Table 1: The division of the LoveDA dataset

- **Are there any errors, sources of noise, or redundancies in the dataset?** This dataset has been checked with the "Surveying and Mapping Data of Geography and National Conditions" to ensure its effectiveness. The possible errors come from annotators cognitive bias, which are within control.
- **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** The dataset is self-contained.
- **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctorpatient confidentiality, data that includes the content of individuals' non-public communications)?** No. All data was obtained from Google Earth platform, and do not contain any coordinate location information.
- **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** No. The dataset only contains remote sensing images and land-cover annotations.
- **Does the dataset relate to people?** No.

3 Collection Process

- **How was the data associated with each instance acquired?** Based on the advanced *ArcGIS* geo-spatial software, all the images were annotated by professional remote sensing annotators, supervised with a comprehensive annotation pipeline [5].
- **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** The remote sensing images were obtained from the Google Earth platform.
- **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?** The larger set covers the total Nanjing, Changzhou and Wuhan cities. The LoveDA is representative of the larger set for two reasons: **1) Representative sampling.** The images were collected from 18 spatially independent areas, covering 18 administrative districts in different cities. The sampling methods were shown in the Figure 1.

3.1 Image Distribution and Division

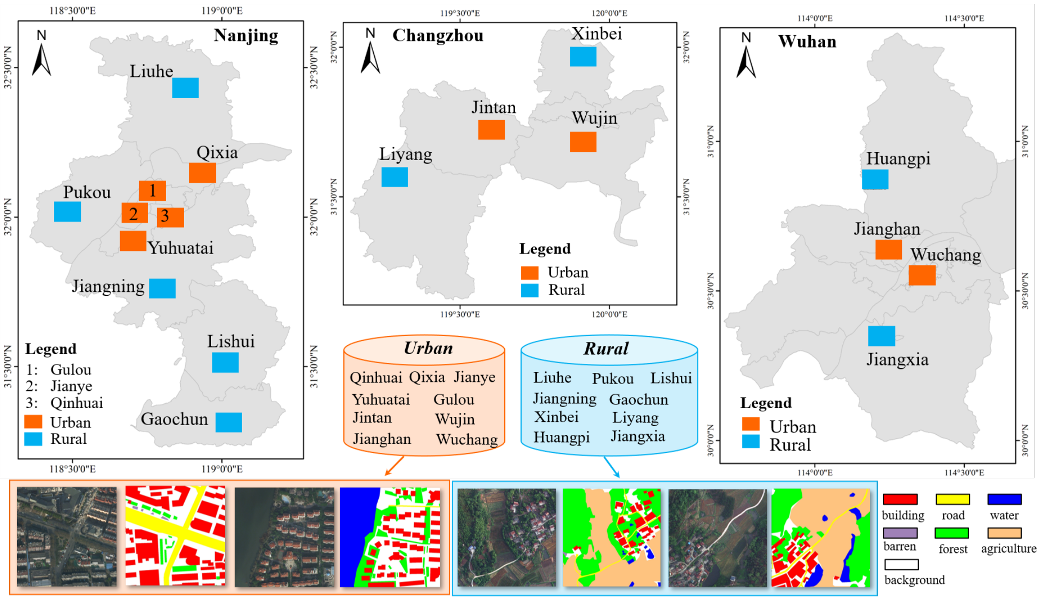

Figure 1: Overview of the dataset distribution. The images were collected from 3 representative cities, covering 18 administrative districts.

- **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?** The annotators and supervisors were students and teachers in the RSIDEA research group.
- **Over what timeframe was the data collected?** The LoveDA dataset was constructed using 0.3 m HSR images obtained from Nanjing, Changzhou and Wuhan in July 2016, covering 536.15km^2.
- **Were any ethical review processes conducted (e.g., by an institutional review board)?** Unknown.
- **Does the dataset relate to people?** No.

4 Preprocessing/cleaning/labeling

- **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)?** The raw image covers 18 large-scale areas in three representative cities shown in Figure 1. Each large-scale image was clipped into 1024 $\times$ 1024 patches without overlap. Besides, we converted all the images from TIFF format into PNG format for compressing storage.

- **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** The raw unprocessed data (consisting of large-scale images and annotations) is saved.
- **Is the software used to preprocess/clean/label the instances available?** All software used to process the data are available and has been specified above.

5 Uses

- **Has the dataset been used for any tasks already?** No, the dataset has not yet been used for any tasks.
- **Is there a repository that links to any or all papers or systems that use the dataset?** The initial benchmark of this dataset was shared at `https://github.com/Junjue-Wang/LoveDA`.
- **What (other) tasks could the dataset be used for?** The dataset can be used for semantic segmentation and unsupervised domain adaptation.
- **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** We do not foresee any harmful risks in future use.
- **Are there tasks for which the dataset should not be used?** The dataset is prohibited from being used for illegal tasks (e.g. illegal surveying and mapping).

6 Distribution

- **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** Yes. The dataset will be public at LoveDA Semantic Segmentation and LoveDA Unsupervised Domain Adaptation. as challenging competitions.
- **How will the dataset will be distributed (e.g., tarball on website, API, GitHub)?** Following the successful cases (e.g. US3D dataset [2], MS COCO Image Captioning dataset [1], LSMDC dataset [3], etc.), we will publish the LoveDA dataset on Codalab.
- **When will the dataset be distributed?** The dataset will be released in October 2021.
- **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** The LoveDA dataset copyright belongs to the authors of the reviews unless otherwise stated. The LoveDA dataset will be released under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International license (CC BY-NC-SA 4.0)
- **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** There are no fees or restrictions.
- **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** Unknown

7 Maintenance

- **Who is supporting/hosting/maintaining the dataset?** The dataset is hosted at the Wuhan University.
- **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?** All questions and comments can be sent to Junjue Wang: kingdrone@whu.edu.cn
- **Is there an erratum?** All changes to the dataset will be announced through the LoveDA mailing list. Those who would like to sign up should send an email to kingdrone@whu.edu.cn.

- **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** All changes to the dataset will be announced in the project page.
- **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** No.

References

[1] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[2] S. Kunwar, H. Chen, M. Lin, H. Zhang, P. D'Angelo, D. Cerra, S. M. Azimi, M. Brown, G. Hager, N. Yokoya, R. Hänsch, and B. Le Saux. Large-scale semantic 3-d reconstruction: Outcome of the 2019 ieee grss data fusion contest—part a. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 14:922–935, 2021.

[3] Y. Liu, S. Albanie, A. Nagrani, and A. Zisserman. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487*, 2019.

[4] X. Peng, B. Usman, N. Kaushik, D. Wang, J. Hoffman, and K. Saenko. Visda: A synthetic-to-real benchmark for visual domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 2021–2026, 2018.

[5] S. Waqas Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F. Shahbaz Khan, F. Zhu, L. Shao, G.-S. Xia, and X. Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019.