SegEarth-OV: Towards Training-Free Open-Vocabulary Segmentation for Remote Sensing Images

Supplementary Material

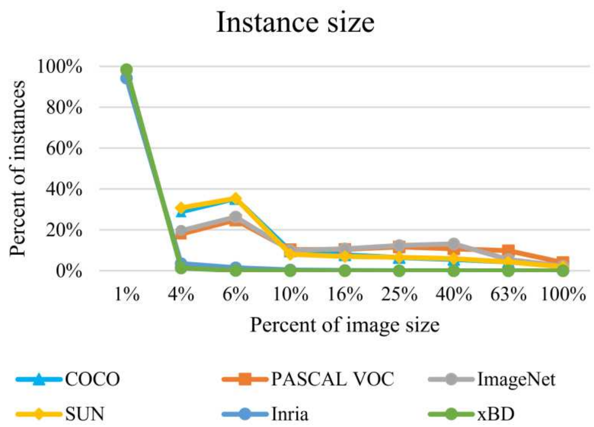

Figure 1. The distribution of instance sizes for natural image datasets (MS COCO, ImageNet Detection, PASCAL VOC and SUN) and remote sensing datasets (Inria and xBD). The data for the natural image is borrowed from [6], and the data for Inria and xBD are calculated with the image size at 1024 $\times$ 1024.

1. Datasets

1.1. Semantic Segmentation

- **OpenEarthMap** [16] includes worldwide satellite and aerial images with a spatial resolution of 0.25-0.5m. It contains 8 foreground classes and one background class. We use its validation set (excluding xBD data) for evaluation.
- **LoveDA** [13] is constructed using 0.3m images obtained from the Google Earth platform. It contains both urban and rural areas. It contains 6 foreground classes and one background class. We use its validation set for evaluation.
- **iSAID** [14] is manily collected from the Google Earth, some are taken by satellite JL-1, the others are taken by satellite GF-2. Its image data is the same as the DOTA-v1.0 dataset [15]. It contains 15 foreground classes and one background class. We use its validation set for evaluation, which is cropped to 11,644 images by default (patch_size=896, overlap_area=384).
- **Potsdam and Vaihingen** are for urban semantic segmentation used in the 2D Semantic Labeling Contest. Their spatial resolutions are 5cm and 9cm, respectively, and they contain 5 foreground classes and one background class. We use the validation set for evaluation according to MMSegmentation's setting.
- **UAVid** [8] consists of 30 video sequences capturing 4K HR images in slanted views. We treat them as images without considering the relationship between frames, and the classes "static car" and "moving car" are converted to "car". Therefore, it contains 5 foreground classes and one background class. We use its test set for evaluation, which is cropped to 1020 images (patch_height=1280, patch_width=1080, no overlap).
- **UDD5** [2] is collected by a professional-grade UAV (DJI-Phantom 4) at altitudes between 60 and 100m. It contains 4 foreground classes and one background class. We use its validation set for evaluation.
- **VDD** [1] is collected by DJI MAVIC AIR II, including 400 RGB images with 4000 $\times$ 3000 pixel size. All the images are taken at altitudes ranging from 50m to 120m. It contains 6 foreground classes and one background class. We use its test set for evaluation.

1.2. Building extraction

- **WHU**^{Aerial} [4] consists of more than 220k independent buildings extracted from aerial images with 0.075m spatial resolution and 450 km^2 covering in Christchurch, New Zealand. We use its validation set for evaluation.
- **WHU**$^{Sat.\text{II}}$ [4] consists of 6 neighboring satellite images covering 860 km^2 on East Asia with 0.45m ground resolution. We use its test set (3726 tiles with 8358 buildings) for evaluation. The original images are cropped to 1000 $\times$ 1000 without overlap.
- **Inria** [9] covers dissimilar urban settlements, ranging from densely populated areas (e.g., San Francisco's financial district) to alpine towns (e.g,. Lienz in Austrian Tyrol). It covers 810 km^2 with a spatial resolution of 0.3m. We use the test set for evaluation according to the setting in [5].
- **xBD** [3] covers a diverse set of disasters and geographical locations with over 800k building annotations across over 45k km^2 of imagery. Its spatial resolution is 0.8m. We use the pre-disaster satellite data of test set for evaluation.

1.3. Road extraction

- **CHN6-CUG** [17] is a large-scale satellite image data set of representative cities in China, collected from Google Earth. It contains 4511 labeled images of 512 $\times$ 512 size with a spatial resolution of 0.5m. We use its test set for evaluation.
- **DeepGlobe** covers images captured over Thailand, Indonesia, and India. Its available data cover 362 km^2 with a spatial resolution of 5m. The roads are precisely annotated with varying road widths. We use the validation set for evaluation according to the setting in [11].

Table 1. The prompt class name of the evaluation datasets. {} indicates multiple prompt vocabularies for one class.

Dataset	Class Name
OpenEarthMap	background, {bareland, barren}, grass, pavement, road, {tree, forest}, {water, river}, cropland, {building, roof, house}
LoveDA	background, {building, roof, house}, road, water, barren, forest, agricultural
iSAID	background, ship, store tank, baseball diamond, tennis court, basketball court, ground track field, bridge, large vehicle, small vehicle, helicopter, swimming pool, roundabout, soccer ball field, plane, harbor
Potsdam, Vaihingen	{road, parking lot}, building, low vegetation, tree, car, {clutter, background}
UAVid	background, building, road, car, tree, vegetation, human
UDD5	vegetation, building, road, vehicle, background
VDD	background, facade, road, vegetation, vehicle, roof, water
WHU^{Aerial}, WHU$^{Sat.\text{II}}$, Inria, xBD	background, building
CHN6-CUG, DeepGlobe, Massachusetts, SpaceNet	background, road
WBS-SI	background, water

Figure 2. Qualitative comparison between different training-free OVSS methods on OpenEarthMap.

- **Massachusetts** [10] covers a wide variety of urban, suburban, and rural regions and covers an area of over 2,600 km^2 with a spatial resolution of 1m. Its labels are generated by rasterizing road centerlines obtained from the OpenStreetMap project, and it uses a line thickness of 7 pixels. We use its test set for evaluation.

- **SpaceNet** [12] contains 422 km^2 of very high-resolution imagery with a spatial resolution of 0.3m. It covers Las Vegas, Paris, Shanghai, Khartoum and is designed for the SpaceNet challenge. We use the test set for evaluation according to the setting in [7].

1.4. Flood Detection

- **WBS-SI** is a satellite image dataset for water body segmentation. It contains 2495 images and we randomly divided 20% of the data as a test set for evaluation.

References

[1] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *arXiv preprint arXiv:2305.13608*, 2023. 1

[2] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, pages 347–359. Springer, 2018. 1

Figure 3. Qualitative comparison between different training-free OVSS methods on UDD5.

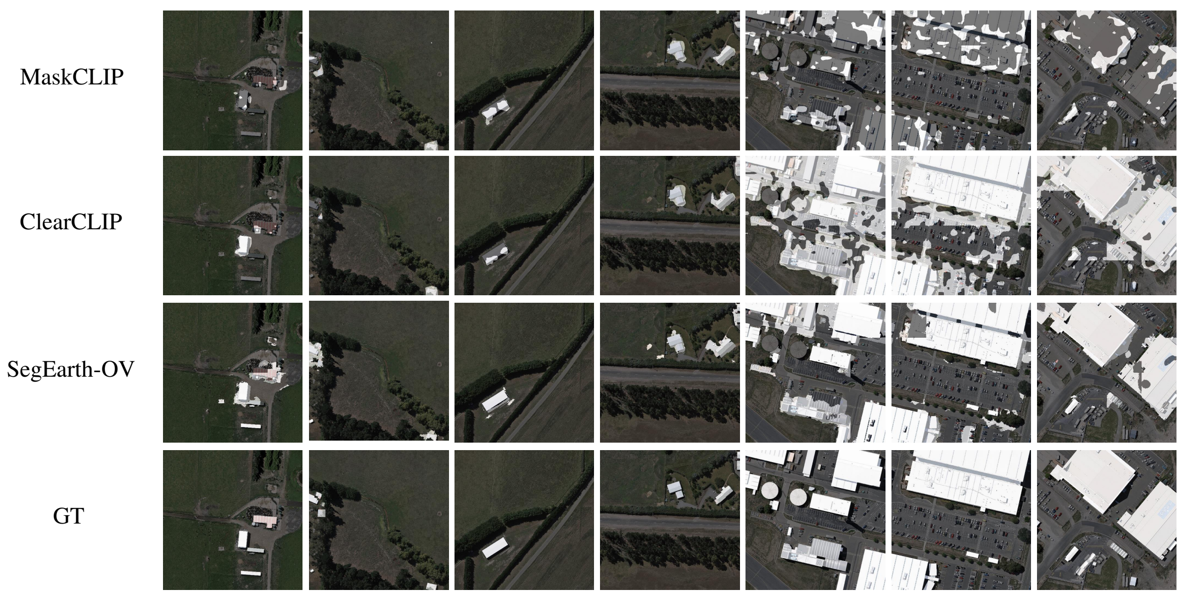

Figure 4. Qualitative comparison between different training-free OVSS methods on WHU^{Aerial}.

[3] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery, 2019. 1

[4] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on geoscience and remote sensing*, 57(1):574–586, 2018. 1

[5] Qi Li, Weixiang Yang, Wenxi Liu, Yuanlong Yu, and Shengfeng He. From contexts to locality: Ultra-high resolution image segmentation via locality-aware contextual correlation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7252–7261, 2021. 1

[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[7] Xiaoyan Lu, Yanfei Zhong, Zhuo Zheng, JunJue Wang, Dingyuan Chen, and Yu Su. Global road extraction using a pseudo-label guided framework: from benchmark dataset to cross-region semi-supervised learning. *Geo-spatial Information Science*, pages 1–19, 2024. 2

[8] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108 – 119, 2020. 1

[9] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods general-

ize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International geoscience and remote sensing symposium (IGARSS)*, pages 3226–3229. IEEE, 2017. 1

[10] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013. 2

[11] Suriya Singh, Anil Batra, Guansong PANG, Lorenzo Torresani, Saikat Basu, Manohar Paluri, and CV Jawahar. Self-supervised feature learning for semantic segmentation of overhead imagery. 2018. 1

[12] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 2

[13] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*. Curran Associates, Inc., 2021. 1

[14] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–37, 2019. 1

[15] Gui-Song Xia, Xiang Bai, Jian Ding, Zhen Zhu, Serge Belongie, Jiebo Luo, Mihai Datcu, Marcello Pelillo, and Liangpei Zhang. Dota: A large-scale dataset for object detection in aerial images. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[16] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6254–6264, 2023. 1

[17] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li. A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 175:353–365, 2021. 1