ConInfer: Context-Aware Inference for Training-Free Open-Vocabulary Remote Sensing Segmentation

Supplementary Material

In this supplementary material, we provide the following contents: 1) more implementation details of datasets (§A), 2) the class prompt design (§B), 3) the details of the MaskCLIP* baseline (§C), 4) additional qualitative results and analyses (§D).

A. Datasets

A.1. Semantic Segmentation

- **OpenEarthMap** [14] 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** [12] 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** [13] 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 DOTAv1.0 dataset [69]. 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**[1] and **Vaihingen**[2] 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 MMSeg's[3] 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 (DJIPhantom 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*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.

A.2. Building extraction

- WHUAerial [5] 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.
- WHUSat.II [5] 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 × 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 its test set for evaluation.
- xBD [4] 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.

A.3. Road extraction

- CHN6-CUG [16] is a large-scale satellite image data set of representative cities in China, collected from Google Earth. It contains 4511 labeled images of 512 × 512 size with a spatial resolution of 0.5m. We use its test set for evaluation.
- DeepGlobe[4] 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 [6].
- 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 [11] 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

[1] https://www.isprs.org/education/benchmarks/UrbanSemLab
[2] https://www.isprs.org/education/benchmarks/UrbanSemLab
[3] https://github.com/open-mmlab/mmsegmentation
[4] http://deepglobe.org

SpaceNet challenge. We use the test set for evaluation according to the setting in [6].

A.4. Flood Detection

- WBS-SI[5] 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.

B. Class prompt

We define the prompt for each class as Tab. 1. For single-class extraction tasks, we define the background class as a miscellaneous category independent of the foreground class. This approach is reasonable for clustering and also enhances performance for other methods.

C. MaskCLIP* (Baseline)

Drawing inspiration from the hierarchical architecture of Vision Transformers [3, 7, 15], we divide the ViT into four stages to leverage the transition from local to global attention patterns. We extract attention maps from the middle layers (2nd, 5th, 8th, 11th) of each stage to form a multi-level attention composite, denoted as Attn_M. This composite is then integrated with the attention map of the final layer, Attn, which is modified according to the SegEarth-OV [6] setting: its FFN is removed, and attention weights are derived from the sum of query, key, and value self-similarities. The final fused attention, Attn', is computed as a weighted sum with a coefficient λ:

$$\mathrm{Attn}' = \lambda \cdot \mathrm{Attn} + (1 - \lambda) \cdot \mathrm{Attn}_M. \tag{1}$$

In the experiment, λ was set to 0.6. We report in Tabs. 2 and 3 the impact of varying λ, as well as the effect of removing Attn_M from the framework in Tabs. 4 and 5.

D. Qualitative Results

References

[1] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025. 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*, pages 347–359, 2018. 1

[3] Xiangxiang Chu, Zhi Tian, Yuqing Wang, Bo Zhang, Haibing Ren, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Twins: Revisiting the design of spatial attention in vision transformers. *NeurIPS*, 34:9355–9366, 2021. 2

[4] 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. arXiv preprint arXiv:1911.09296. 1

[5] 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

[6] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *CVPR*, pages 10545–10556, 2025. 1, 2

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, pages 10012–10022, 2021. 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 generalize to any city? the inria aerial image labeling benchmark. In *IEEE International geoscience and remote sensing symposium*, pages 3226–3229, 2017. 1

[10] Volodymyr Mnih. *Machine learning for aerial image labeling*. PhD thesis, University of Toronto, 2013. 1

[11] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series, 2018. arXiv preprint arXiv:1807.01232. 1

[12] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation, 2021. arXiv preprint arXiv:2110.08733. 1

[13] 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 *CVPRW*, pages 28–37, 2019. 1

[14] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *WACV*, pages 6254–6264, 2023. 1

[15] Weihao Yu, Mi Luo, Pan Zhou, Chenyang Si, Yichen Zhou, Xinchao Wang, Jiashi Feng, and Shuicheng Yan. Metaformer is actually what you need for vision. In *CVPR*, pages 10819–10829, 2022. 2

[16] 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

[5] https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images

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, roof, road, river, barren, forest, agricultural
iSAID	{background, grass, agriculture, water, forest, road, building, barren, urban}, boat, industrial cylindrical storage 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	{road, parking lot}, building, low vegetation, tree, car, {clutter, background}
Vaihingen	impervious surface, building, low vegetation, tree, car, clutter
UAVid	background, building, road, car, tree, vegetation, human
UDD5	vegetation, building, road, vehicle, background
VDD	{background, barren}, exterior, road, grass, car, roof, water
WHUAerial	{forest, vegetation, bareland, river, road, vehicle}, building
WHUSat.II	{forest, vegetation, bareland, river, route, vehicle}, building
Inria	{tree, vegetation, bareland, river, paved road, vehicle}, building
xBD	{background, forest, vegetation, bareland, river, paved road}, building
CHN6-CUG	{background, forest, vegetation, barren, water, building, vehicle, basketball court, agriculture, urban, rangeland, cropland}, road
DeepGlobe	{background, forest, vegetation, barren, water, building, vehicle, agriculture, urban, rangeland, cropland}, road
Massachusetts	{background, forest, vegetation, barren, water, building, vehicle, urban, rangeland, cropland}, road
SpaceNet	{background, forest, vegetation, barren, water, building, vehicle, agriculture, urban, rangeland, cropland}, road
WBS-SI	{background, forest, vegetation, barren, agriculture, urban}, water

Table 2. Comparisons of mIoU on different datasets.

Method	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid_img	UDD5	VDD	Average
Baseline (w/o Attn_M)	31.87	30.09	18.43	43.15	21.60	37.66	42.36	42.20	33.42
Baseline (w/ Attn_M)	32.46	31.58	18.17	44.49	19.94	37.20	42.17	42.54	33.57
Ours (w/o Attn_M)	41.35	37.89	19.29	49.20	31.36	45.92	46.54	47.95	39.94
Ours (w/ Attn_M)	41.95	39.33	20.08	49.99	31.37	46.40	46.86	50.29	40.78

Table 3. Comparison of Foreground IoU on different datasets.

Method	WHU_Aerial	WHU_Sat.II	Inria	xBD_pre	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI	Average
Baseline (w/o Attn_M)	45.09	17.00	40.52	34.27	27.30	14.13	8.04	19.07	48.98	28.27
Baseline (w/ Attn_M)	45.76	15.17	37.26	31.40	29.12	14.68	8.42	19.83	52.57	28.25
Ours (w/o Attn_M)	55.07	31.85	53.66	41.01	35.67	18.02	11.84	23.47	59.40	36.67
Ours (w/ Attn_M)	58.54	39.35	55.65	41.34	40.00	19.85	12.16	24.02	61.38	39.14

Table 4. Performance comparison with different λ values across multiple datasets.

Setting	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid_img	UDD5	VDD
$\lambda = 0.0$	41.35	37.89	19.29	49.20	31.36	45.92	46.54	47.95
$\lambda = 0.2$	41.82	38.19	19.73	49.78	31.15	46.12	46.71	48.23
$\lambda = 0.4$	42.09	38.90	19.97	49.93	30.62	46.14	46.82	49.31
$\lambda = 0.6$	41.95	39.33	20.08	49.99	29.73	45.60	46.86	50.29
$\lambda = 0.8$	40.06	39.61	19.86	49.83	27.74	44.59	46.93	48.43
$\lambda = 1.0$	39.92	39.31	19.35	49.35	25.92	43.43	44.58	46.39

Table 5. Performance comparison on multiple datasets with varying λ.

Setting	WHU_Aerial	WHU_Sat.II	Inria	xBD_pre	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
$\lambda = 0.0$	55.07	31.85	53.66	41.01	35.67	18.02	11.84	23.47	59.40
$\lambda = 0.2$	56.23	34.49	54.58	41.62	36.28	18.63	12.16	23.67	59.84
$\lambda = 0.4$	57.47	36.92	55.45	41.86	36.33	19.37	12.16	23.81	60.65
$\lambda = 0.6$	58.54	39.35	55.65	41.82	40.00	19.85	12.16	24.02	61.38
$\lambda = 0.8$	58.97	41.84	55.36	41.13	37.00	20.07	12.13	24.14	61.79
$\lambda = 1.0$	57.41	41.64	54.22	39.28	33.73	19.89	12.14	24.34	62.25

Figure 1. Qualitative comparison of different OVSS methods on the OpenEarthMap (land cover mapping) datasets.

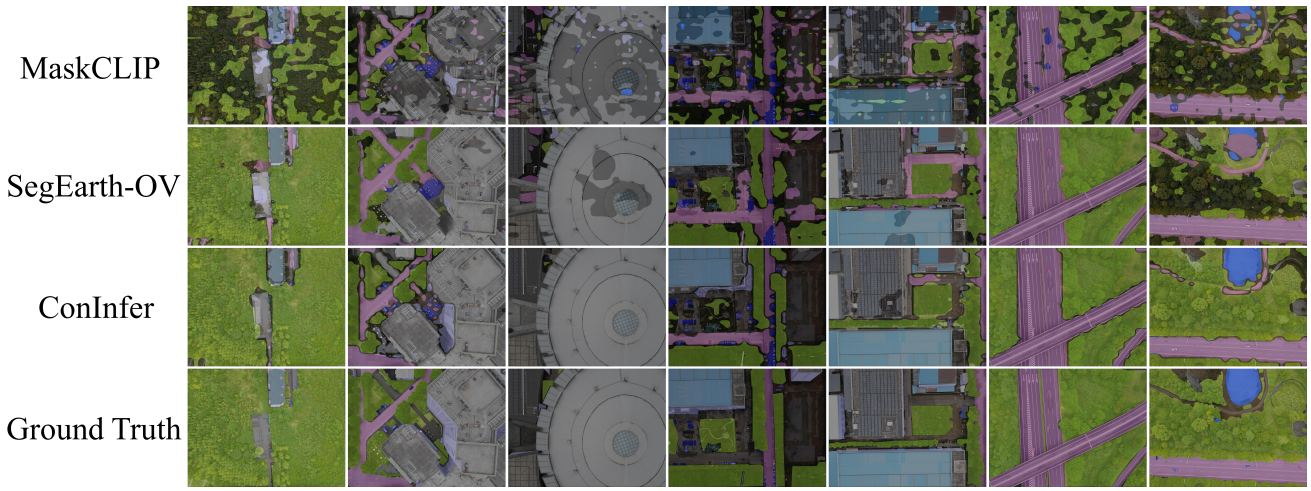

Figure 2. Qualitative comparison of different OVSS methods on the VDD (UAV aerial scene) datasets.

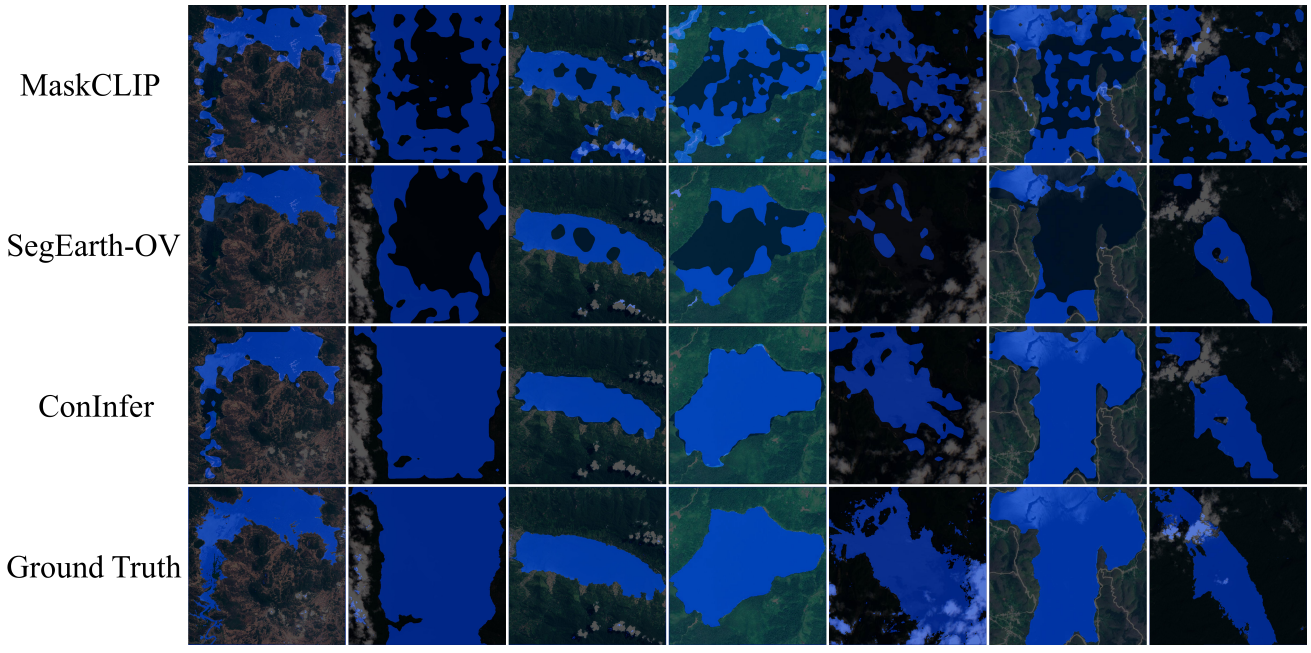

Figure 3. Qualitative comparison of different OVSS methods on the WBI-SI (water body segmentation) datasets.