# MM-OVSeg: Multimodal Optical–SAR Fusion for Open-Vocabulary Segmentation in Remote Sensing

## Supplementary Material

In this appendix, we present additional experimental results and analyses. Section A provides supplementary training details beyond Section 4.1.2 of the main paper. Section B reports the efficiency and sustainability comparison. To further demonstrate the generalization capabilities of MM-OVSeg, we provide Section C with more detailed split performance of seen/unseen classes across six datasets. Section D reports results of MM-OVSeg using the ViT-L/14 backbone. Section E provides further studies on CMU, including its extension to CLIP-SAR alignment. Section F includes additional qualitative visualizations.

## A. Implementation Details

In Table A1, we present more detailed implementations of our MM-OVSeg and other methods, including trainable parameters, batch size and learning rate.

| Method | Backbone | Param. | Batch. | Lr |
|---|---|---|---|---|
| CAT-Seg [5] | ViT-B/16 | **25M** | 4 | 2e-4 |
| FGAseg [20] | ViT-B/16 | 33M | 8 | 2e-4 |
| GSNet [51] | ViT-B/16 | 29M | 4 | 2e-4 |
| EBSeg [38] | ViT-B/16 | 26M | 16 | 1e-4 |
| MM-OVSeg | ViT-B/16 | 31M | 8 | 2.5e-4 |

Table A1. Model implementations. The table reports trainable parameters, batch size and learning rate (Lr).

## B. Efficiency and Sustainability Comparison

In Table A2, we report training/inference latency, estimated carbon emissions [14], and parameter counts for MM-OVSeg and GSNet on DDHR-SK→DDHR-SK. MM-OVSeg incurs additional cost due to SAR integration and dual encoders, but this overhead is accompanied by improved segmentation performance and robustness under adverse conditions.

| | Time (ms) | | $CO_2$ (kg) | | Parameters | mIoU |
|---|---|---|---|---|---|---|
| | Train | Infer | Train | Infer | | |
| CMU module | 343.7 | 31.6 | 1.05 | 0.006 | 85M | – |
| DEF module | 377.2 | 46.8 | 1.14 | 0.008 | 92M | – |
| *full* MM-OVSeg | 955.7 | 116.9 | 3.90 | 0.014 | 331M | **73.1** |
| GSNet [51] | 700.6 | 83.2 | 3.48 | 0.011 | 244M | 55.0 |

Table A2. Efficiency and sustainability comparison.

## C. Seen vs. Unseen Class Evaluation

In Table A3, we provide a more detailed breakdown of seen and unseen class performance, supplementing Table 2 from the main paper. All methods show reduced performance on unseen classes, which is expected in OVS. Prior methods often exhibit imbalanced behavior, performing well on either seen or unseen classes but not both, with inconsistent trends across datasets. In contrast, MM-OVSeg shows more balanced and consistently strong performance on both seen and unseen classes across datasets, indicating improved robustness in the OVS setting.

## D. Larger Backbones on MM-OVSeg

We also evaluate how backbone capacity affects the performance of MM-OVSeg. While the main paper uses ViT-B/16, here we replace both the CLIP and DINO encoders with ViT-L/14, using pretrained weights from CLIP [34] and DINO v3 [40], respectively. As before, multi-scale features are taken from the 8th, 16th, and 24th transformer blocks. During full MM-OVSeg training, we train for 120k iterations using AdamW with a batch size of 4 and an initial learning rate of $2.5 \times 10^{-4}$. All other settings follow those used for the ViT-B/16 backbone.

Following the evaluation protocol in Table 2 of the main paper, Table A4 compares MM-OVSeg (ViT-L/14) with recent state-of-the-art OVS methods across all benchmark datasets. Mean IoU is reported for each dataset and the overall average.

Consistent with the trend observed using ViT-B/16, MM-OVSeg (ViT-L/14) achieves the best overall performance, obtaining 55.0% mIoU across six benchmarks, outperforming the ViT-B/16 version (51.7%) due to its increased model capacity. This improvement further validates the strength of our multimodal fusion design for open-vocabulary segmentation in remote sensing. Moreover, MM-OVSeg (ViT-L/14) maintains a substantial lead on setting ⑥: DDHR-SK→DDHR-CH, demonstrating strong cross-domain robustness.

Figure A1 provides visual comparisons between MM-OVSeg (ViT-L/14) and MM-OVSeg (ViT-B/16). Consistent with the quantitative results in Table A4, the ViT-L/14 variant produces clearer boundaries, more stable predictions under cloud cover, and more accurate responses on both seen and unseen categories. This intuitive improvement further demonstrates how increasing model capacity strengthens multimodal fusion, reinforcing the effectiveness of our

| Method | Publication | ① | | ② | | ③ | | ④ | | ⑤ | | ⑥ | | Mean | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Unseen | Seen | Unseen | Seen | Unseen | Seen | Unseen | Seen | Unseen | Seen | Unseen | Seen | Unseen | Seen |
| CAT-Seg [5] | CVPR'24 | 14.9 | 74.2 | 41.1 | 62.9 | 32.2 | 34.6 | 17.0 | 36.9 | 17.1 | 75.1 | 7.5 | 41.5 | 21.6 | 54.2 |
| EBSeg [38] | CVPR'24 | 2.5 | 74.8 | 26.7 | 67.4 | 12.7 | 35.9 | 5.5 | 37.6 | 0.8 | 76.0 | 8.5 | 38.8 | 9.5 | 55.1 |
| GSNet [51] | AAAI'25 | 18.6 | 76.1 | 13.0 | 83.1 | 32.0 | 37.1 | 32.8 | 39.4 | 18.0 | 76.8 | 8.8 | 48.1 | 20.5 | 60.1 |
| SegEarth-OV [22] | CVPR'25 | 19.3 | 57.9 | 7.9 | 24.1 | 33.1 | 26.3 | 26.9 | 13.5 | 23.0 | 66.2 | 16.4 | 29.4 | 21.1 | 36.2 |
| FGAseg [20] | arXiv'25 | 13.4 | 70.6 | 47.5 | 54.4 | 7.2 | 37.2 | 23.6 | 38.2 | 13.4 | 71.4 | 18.7 | 55.1 | 20.6 | 54.5 |
| **MM-OVSeg** | - | **19.6** | **76.7** | **53.0** | **86.5** | **35.2** | **37.4** | **40.0** | **40.4** | **23.5** | **77.7** | **22.7** | **56.1** | **32.3** | **62.5** |

Table A3. Performance splits for unseen and seen classes. The table reports mIoU scores for each setting and the overall mean.

| Method | Backbone | Publication | ① | ② | ③ | ④ | ⑤ | ⑥ | Mean |
|---|---|---|---|---|---|---|---|---|---|
| CAT-Seg [5] | ViT-L/14 | CVPR'24 | 60.7 | 69.8 | 38.1 | 40.2 | 62.2 | 41.5 | 52.1 |
| EBSeg [38] | ViT-L/14 | CVPR'24 | 48.9 | 50.6 | 23.6 | 26.9 | 50.7 | 31.5 | 38.7 |
| FGAseg [20] | ViT-L/14 | arXiv'25 | 57.1 | 48.8 | 23.3 | 36.5 | 52.3 | 41.5 | 43.3 |
| **MM-OVSeg (ours)** | ViT-L/14 | – | **64.5** | **75.5** | **39.4** | **41.2** | **66.0** | **43.5** | **55.0** |

Table A4. Comparison of OVS methods with ViT-L/14 backbone across all evaluation settings as illustrated in Table 1 of the main paper. The table reports mIoU scores for each setting and the overall mean. Settings correspond to: ①: PIE-cloud→PIE-cloud; ②: DDHR-SK→DDHR-SK; ③: OEM-thick→OEM-thick; ④: OEM-thin→OEM-thin; ⑤: PIE-clean→PIE-clean; ⑥: DDHR-SK→DDHR-CH. MM-OVSeg achieves the highest accuracy in all settings and obtains the best overall mean score, demonstrating strong robustness under cloudy conditions and superior cross-domain generalization.

design for open-vocabulary segmentation in remote sensing.

## E. CMU for CLIP-SAR Alignment

As discussed in Section 3.2 of the main paper, we also investigate whether a CLIP-style visual encoder can be trained for SAR using the CMU procedure, analogous to the DINO-SAR setup. Following the same strategy, we distill multi-scale ViT features from the RGB CLIP encoder into a SAR-specific CLIP encoder using the InfoNCE loss. Table A5 reports the performance of four model variants:

- Model #1: baseline without CMU;
- Model #2: CMU applied to CLIP only (CLIP-SAR);
- Model #3: CMU applied to DINO only (DINO-SAR), which corresponds to MM-OVSeg;
- Model #4: CMU applied to both DINO and CLIP for SAR.

All CMU variants improve over the baseline, confirming the value of cross-modal alignment. However, DINO-SAR alone (Model #3) achieves the best performance, while adding a CLIP-SAR encoder (Model #4) results in a performance drop. This behavior can be explained as follows: DINO provides dense, locally discriminative features that are crucial for pixel-level segmentation, whereas CLIP encoders produce coarse global embeddings optimized for image-level alignment rather than spatial precision. Training a CLIP-SAR encoder substantially increases the number of global embeddings without providing new local information, which introduces redundancy and complicates the fusion process.

| Model index | CLIP | DINO | mIoU |
|---|---|---|---|
| #1 | ✗ | ✗ | 64.1 |
| #2 | ✓ | ✗ | 65.4 |
| #3 | ✗ | ✓ | **73.1** |
| #4 | ✓ | ✓ | 66.5 |

Table A5. Ablation on applying CMU to different visual encoders on the ②: DDHR-SK→DDHR-SK segmentation task.

## F. Additional Visualization Results

Similar to Figure 4 of the main paper, we provide more qualitative comparisons of MM-OVSeg (as in Table 2 "SOTA performance") for both intra-domain and cross-domain settings. Figure A2 shows additional intra-domain examples, and Figure A3 presents cross-domain results. These results further demonstrate the superiority of MM-OVSeg for multimodal open-vocabulary segmentation across diverse weather conditions.

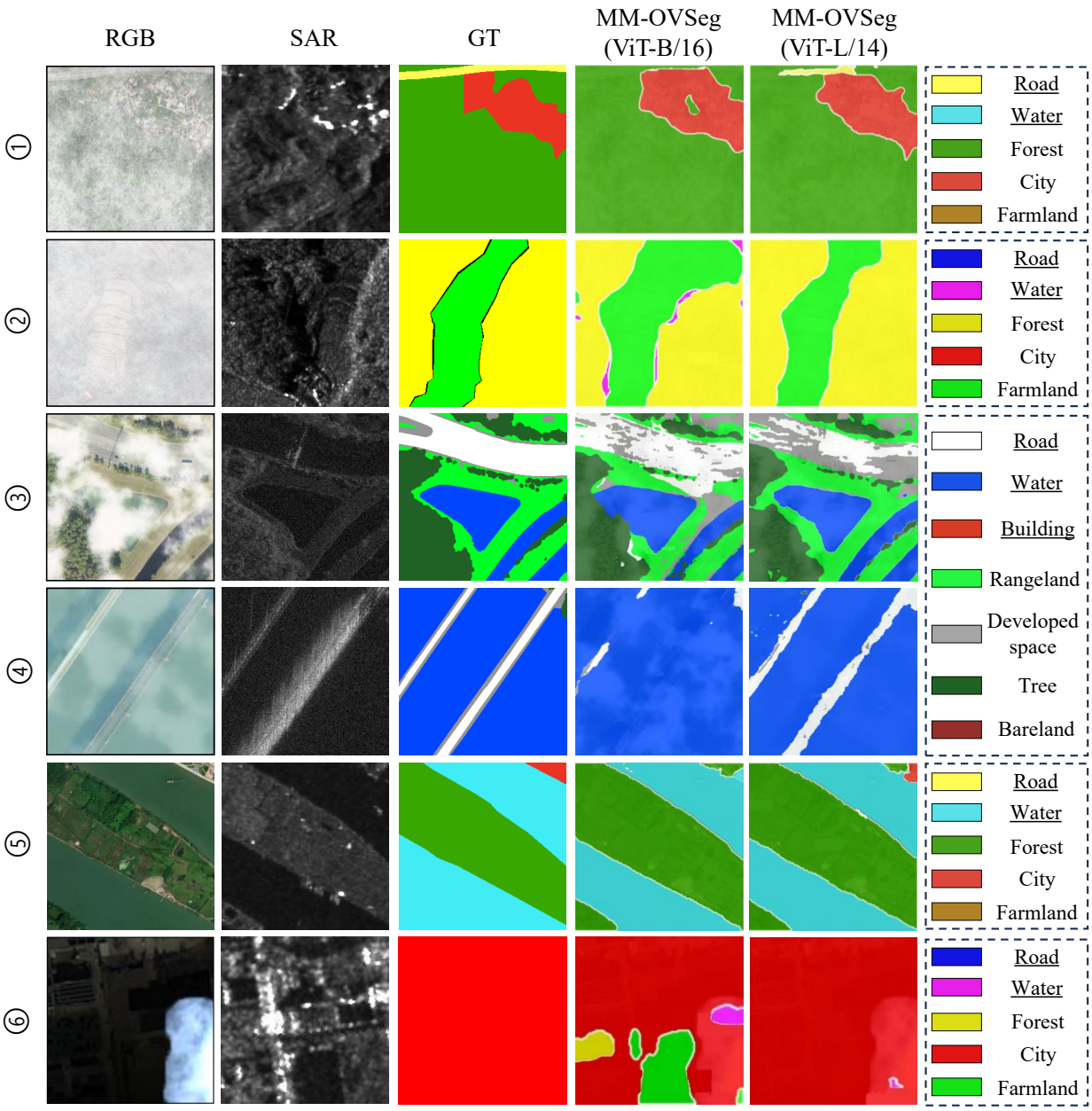


Figure A1. Visualization of OVS results. From left to right: input RGB image, input SAR image, ground truth, and segmentation outputs from MM-OVSeg (ViT-B/16) and MM-OVSeg (ViT-L/14). In the legend, underlined categories represent *unseen* classes and the remaining categories are *seen* classes.

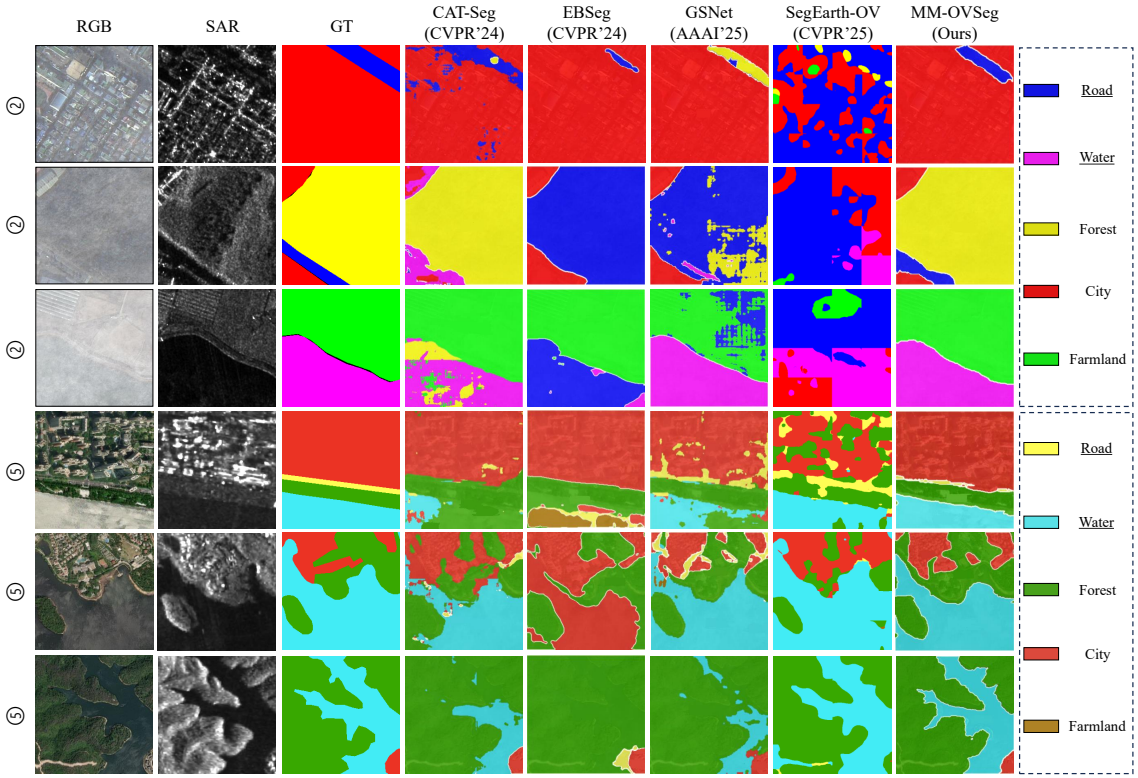


Figure A2. Intra-domain visualization of OVS results, including ②: DDHR-SK→DDHR-SK and ⑤: PIE-clean → PIE-clean. From left to right: input RGB image, input SAR image, ground truth, and segmentation outputs from CAT-Seg, EBSeg, GSNet, SegEarth-OV, and our MM-OVSeg. In the legend, underlined categories represent *unseen* classes and the remaining categories are *seen* classes.

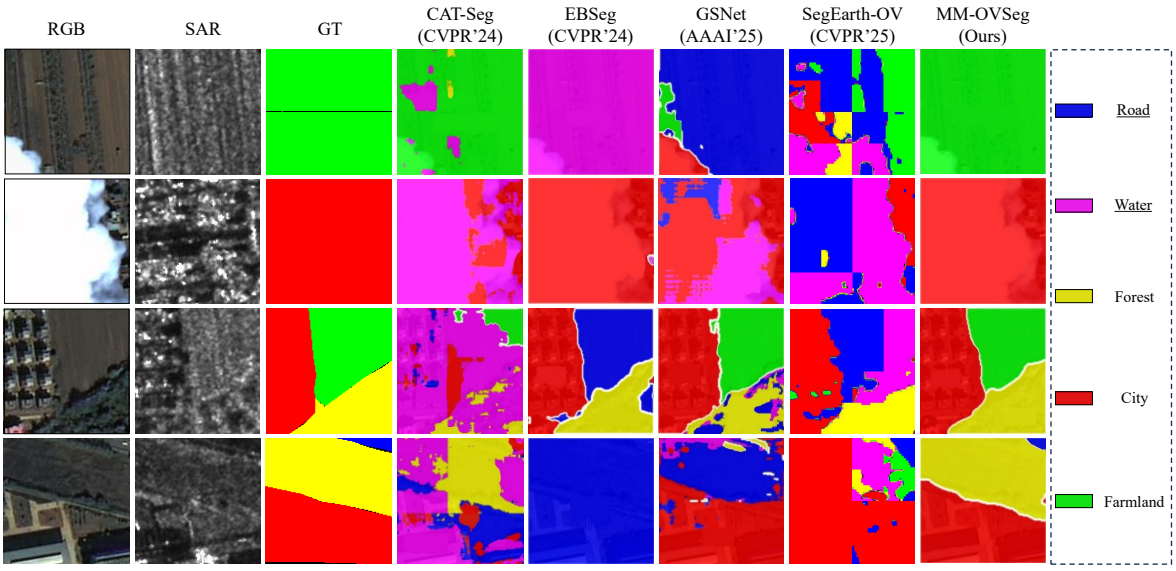


Figure A3. Cross-domain visualization of OVS results for ⑥: DDHR-SK→DDHR-CH. From left to right: input RGB image, input SAR image, ground truth, and segmentation outputs from CAT-Seg, EBSeg, GSNet, SegEarth-OV, and our MM-OVSeg. In the legend, underlined categories represent *unseen* classes and the remaining categories are *seen* classes.