南京航空航天大學
NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

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

Yimin Wei[1,2,*] Aoran Xiao[2,*] Hongruixuan Chen[1,2] Junshi Xia[2] Naoto Yokoya[1,2,†]
[1]The University of Tokyo [2]RIKEN AIP
4959184626@edu.k.u-tokyo.ac.jp, {xiaoaoran94, Qschrx}@gmail.com,
junshi.xia@riken.jp, yokoya@k.u-tokyo.ac.jp
*Equal contribution, †Corresponding author

CVPR 2026

# Introduction

- Open-vocabulary segmentation enables pixel-level recog nition from an open set of textual categories, allowing generalization beyond fixed classes.
- Despite great potential in remote sensing, progress in this area remains largely limited to clear-sky optical data and struggles under cloudy or haze-contaminated conditions.
- We present MM-OvSeg, a multimodal Optical-SAR fusion framework for resilient open-vocabulary segmentation under adverse weather conditions.

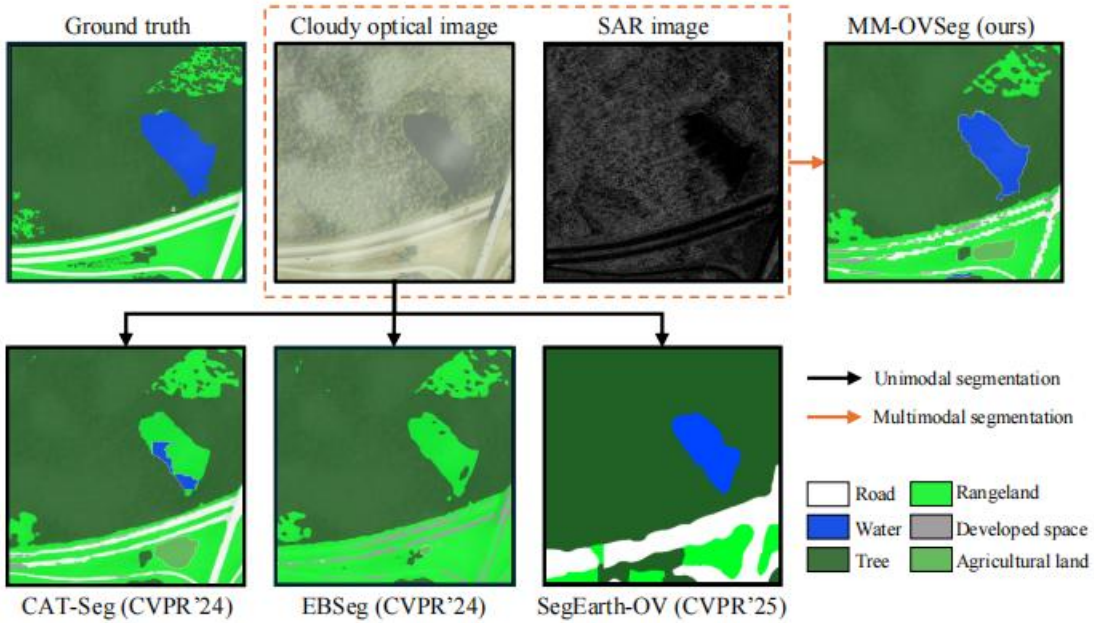
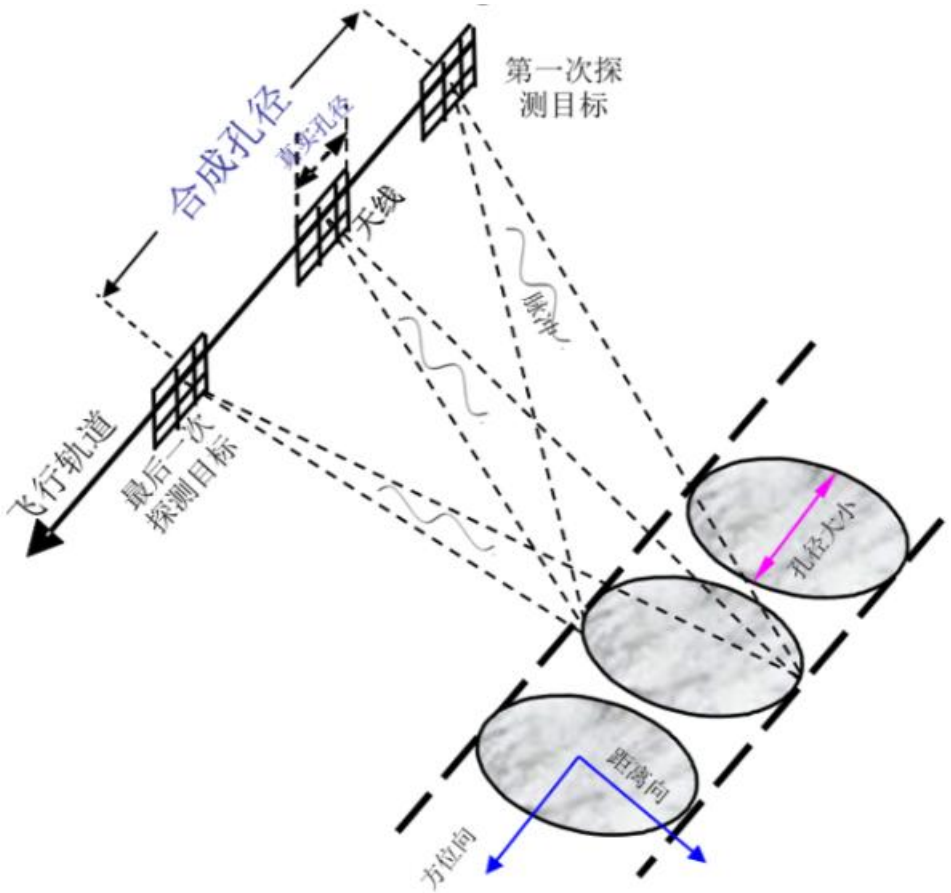



Figure 1. Existing unimodal OVS methods fail in cloudy environments due to severely degraded optical inputs. By incorporating SAR, which penetrates clouds and haze, MM-OVSeg produces significantly more accurate and consistent segmentation results.

# SAR（Synthetic Aperture Radar）




# Methods

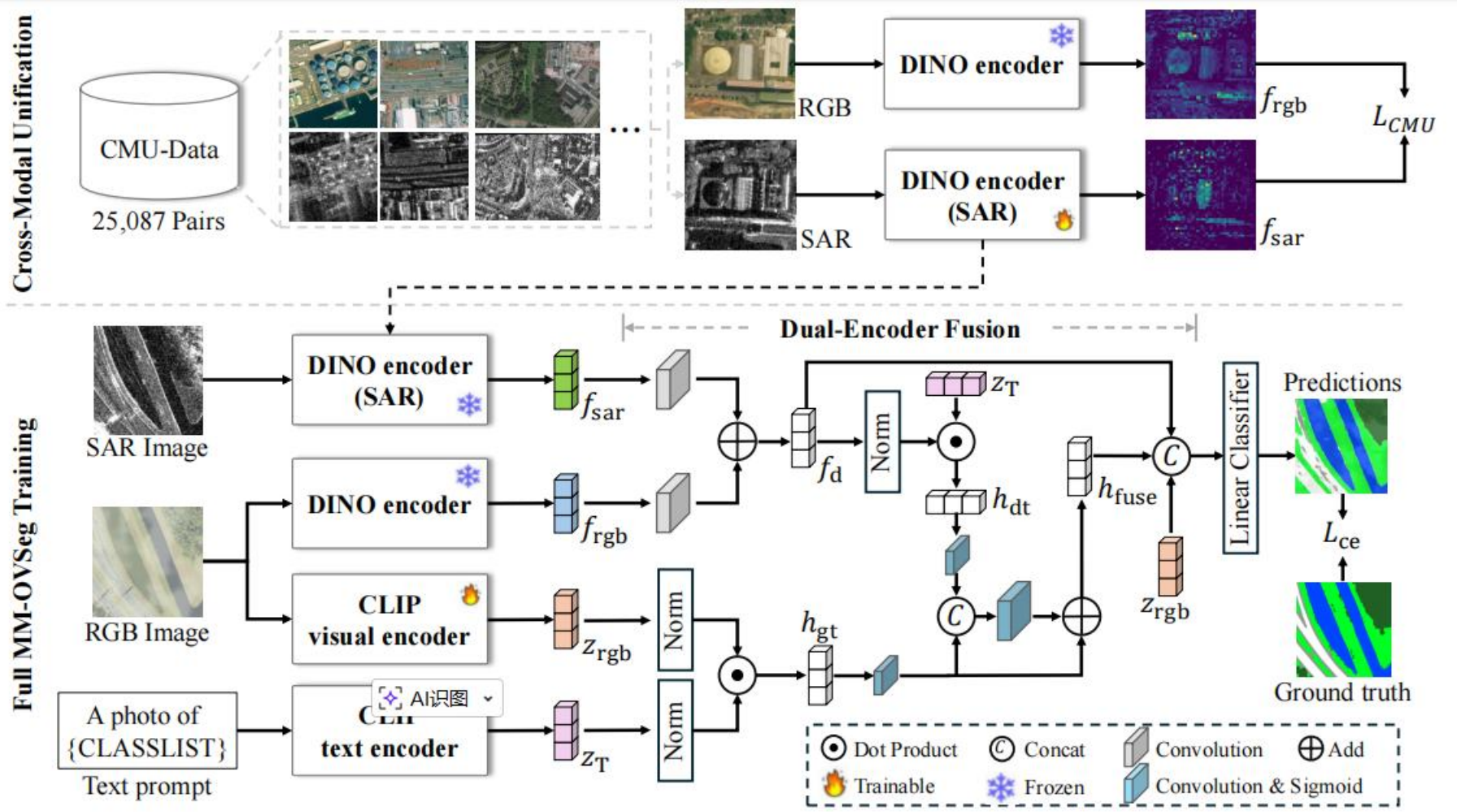
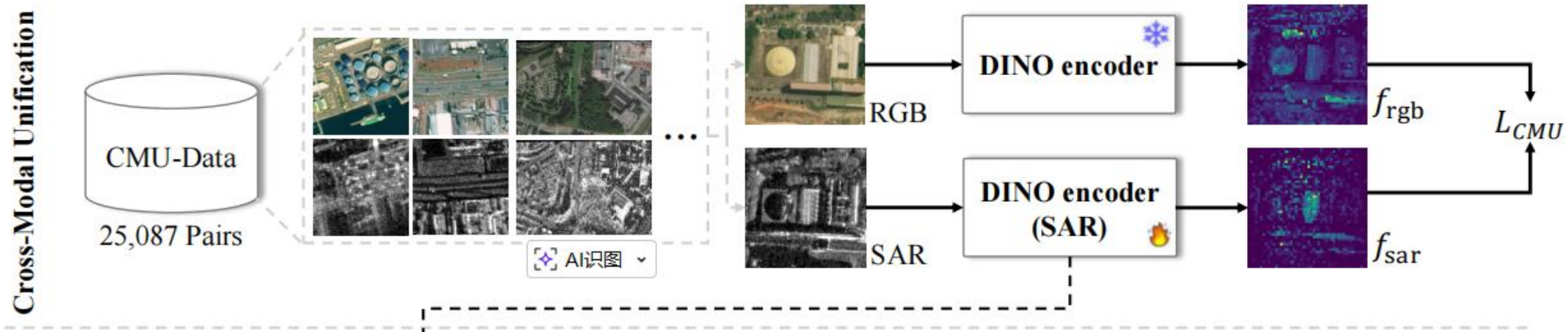



Figure 2. Overall optimization framework of MM-OVSeg. The training pipeline consists of two stages. (1) In the Cross-Modal Unification stage, the SAR DINO encoder is trained to align SAR features with the fixed RGB DINO features using the CMU-Data collection of 25,087 RGB and SAR image pairs. (2) In the full MM-OVSeg training stage, the model jointly processes optical and SAR inputs for multimodal open-vocabulary segmentation. The Dual-Encoder Fusion module integrates RGB and SAR dense features and aligns them with CLIP text embeddings, after which a linear classifier predicts the final segmentation map.

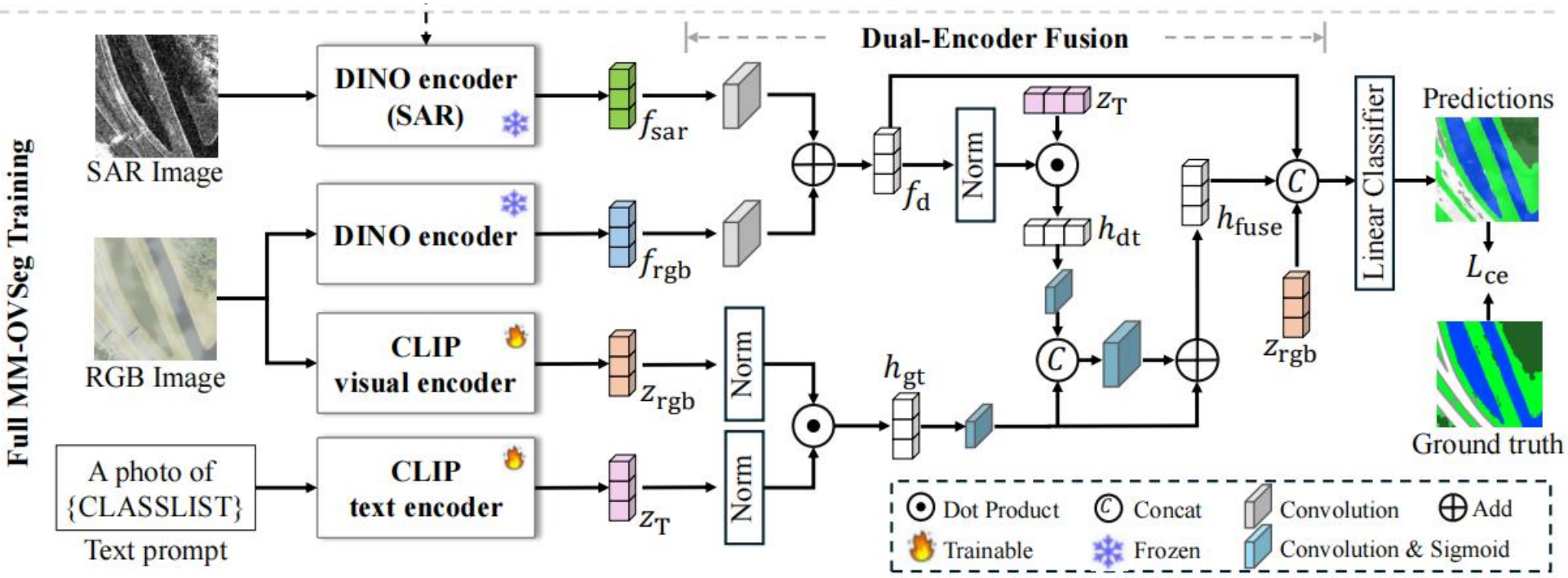
# Methods




$$L_{\mathrm{CMU}} = -\log \frac{\exp(f_{\mathrm{sar}} f_{\mathrm{rgb}}^{+}/\tau)}{\exp(f_{\mathrm{sar}} f_{\mathrm{rgb}}^{+}/\tau) + \sum_{j=1}^{N} \exp(f_{\mathrm{sar}} f_{\mathrm{rgb}}^{-j}/\tau)}$$

# Methods

$$f^i_{\text{rgb}} = \mathcal{F}^i_{\text{rgb}}(I_{\text{rgb}}), \qquad f^i_{\text{sar}} = \mathcal{F}^i_{\text{sar}}(I_{\text{sar}}), \quad i \in \{1,2,3\}, \qquad f^i_d = \sigma_i(f^i_{\text{rgb}}) + \sigma_i(f^i_{\text{sar}}).$$




$$z_{\text{rgb}} = \Phi_V(I_{\text{rgb}}), \quad z_T = \Phi_T(T)$$

$$h^i_{dt} = f^i_d \cdot z_T, \quad h_{gt} = z_{\text{rgb}} \cdot z_T$$

$$h^i_{dt}, h_{gt} = \varphi(\sigma_7(h^i_{dt})), \varphi(\sigma_7(h_{gt})).$$

$$h^i_{\text{fuse}} = \varphi(\sigma_7([h^i_{dt}; h_{gt}])) + h_{gt}$$

# Experiments

| Index | Datasets (*train* → *test*) | weather | cloud cover | cloud type | generalization |
|---|---|---|---|---|---|
| ① | PIE-cloud → PIE-cloud | cloudy | varied | synthetic | intra-domain |
| ② | DDHR-SK → DDHR-SK | cloudy | varied | synthetic | intra-domain |
| ③ | OEM-thick → OEM-thick | cloudy | thick | synthetic | intra-domain |
| ④ | OEM-thin → OEM-thin | cloudy | thin | synthetic | intra-domain |
| ⑤ | PIE-clean → PIE-clean | clear sky | *none* | *none* | intra-domain |
| ⑥ | DDHR-SK → DDHR-CH | cloudy | varied | synthetic | cross-domain |

Table 1. Evaluation settings for MM-OVSeg. The experiments cover clear sky and cloudy weather, synthetic cloud cover with different opacity levels (thin or thick or varied), and both intra-domain and cross-domain generalization scenarios.

| Method | Publication | ① | ② | ③ | ④ | ⑤ | ⑥ | Mean |
|---|---|---|---|---|---|---|---|---|
| CAT-Seg [5] | CVPR'24 | 54.5 | 54.2 | 33.8 | 29.5 | 55.8 | 27.8 | 42.6 |
| EBSeg [38] | CVPR'24 | 50.8 | 51.1 | 27.2 | 25.6 | 51.0 | 26.7 | 38.7 |
| GSNet [51] | AAAI'25 | 57.0 | 55.0 | 35.2 | 37.0 | 57.2 | 32.4 | 45.6 |
| SegEarth-OV [22] | CVPR'25 | 45.1 | 17.6 | 28.9 | 18.5 | 51.8 | 24.2 | 31.0 |
| FGAseg [20] | arXiv'25 | 51.6 | 51.6 | 26.0 | 32.8 | 52.1 | 40.6 | 42.5 |
| **MM-OVSeg (ours)** | – | **57.7** | **73.1** | **36.6** | **40.2** | **59.7** | **42.6** | **51.7** |

Table 2. Comparison of OVS methods across all evaluation settings defined in Table 1. 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.

# Experiments

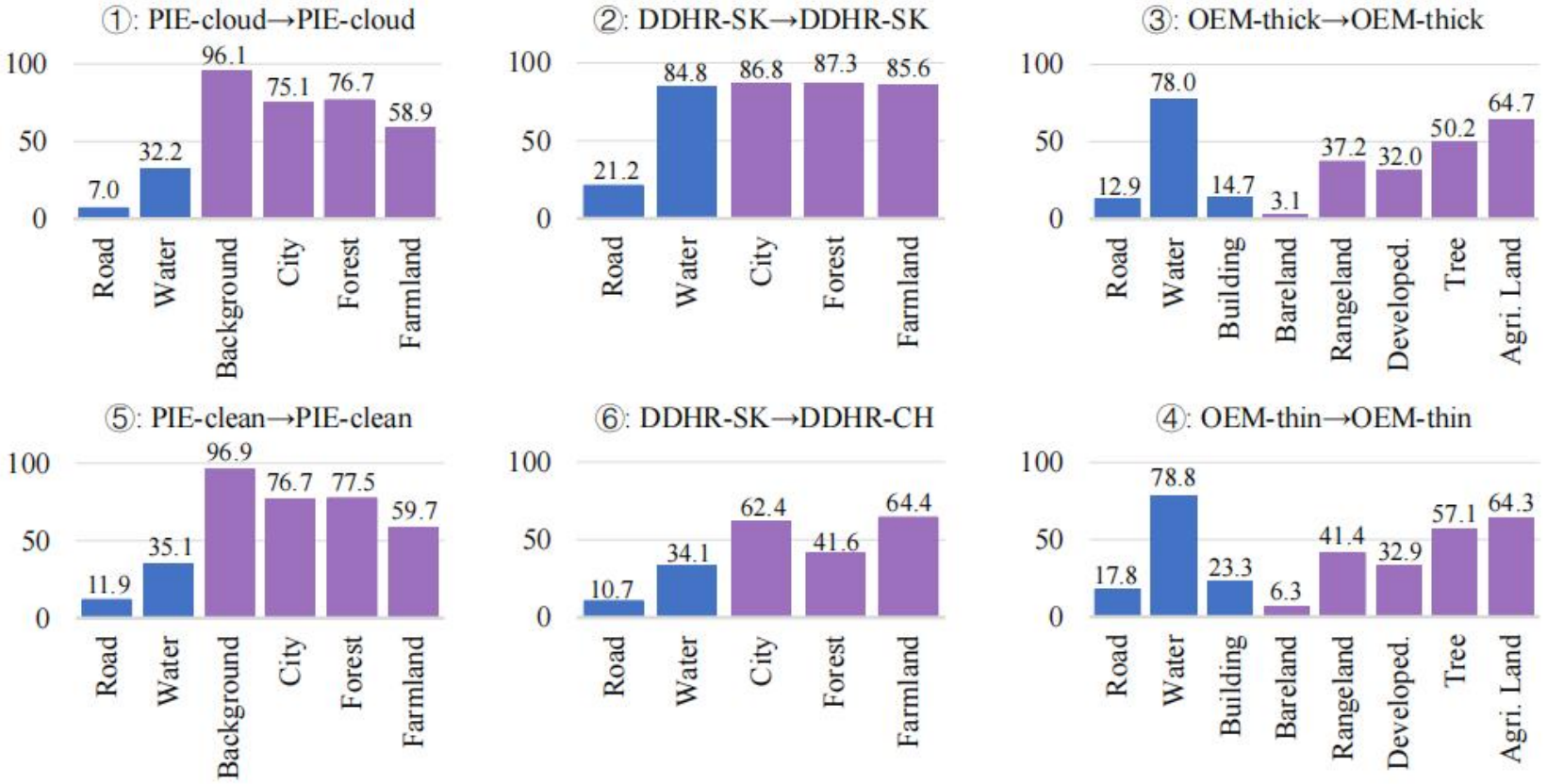



Figure 3. IoU performance for each individual class under the six evaluation settings defined in Table 1. Purple bars and blue bars represent *seen* and *unseen* classes, respectively.

# Experiments

南京航空航天大學 NANJING UNIVERSITY OF AERONAUTICS AND ASTRONAUTICS

| 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.

# Experiments

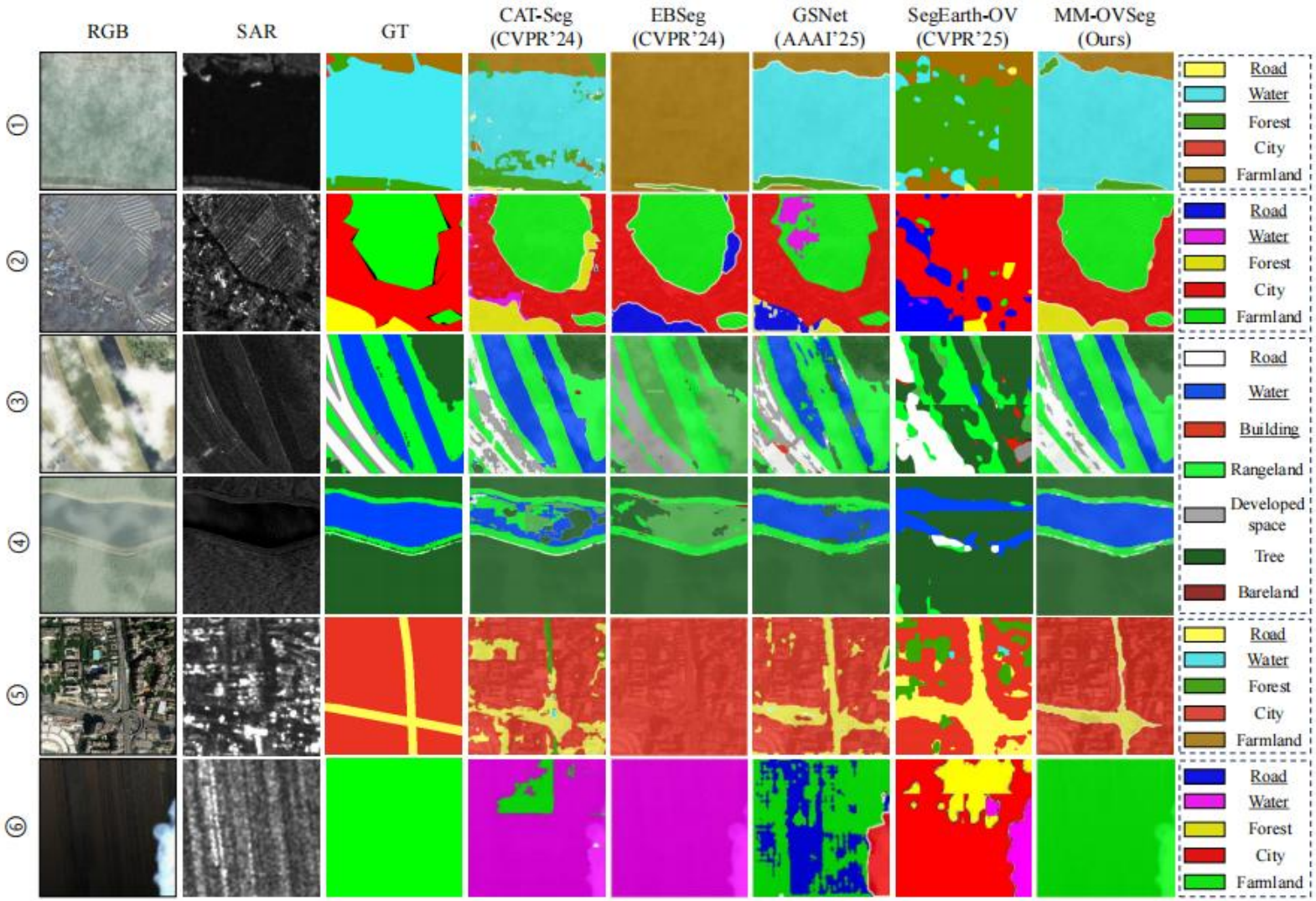



Figure 4. Visualization of OVS results. From left to right: input RGB image, input SAR image, ground truth, and segmentation outputs from CAT-Seg, EBSeg, GSNet, SegEarth-OV, and our MM-OVSeg. In the legend, underlined categories represent *unseen* classes and the remaining categories are *seen* classes.

# Experiments



| Model Variant | Forest | City | Farmland | Road | Water | mIoU |
|---|---|---|---|---|---|---|
| MM-OVSeg | 87.3 | 86.8 | 85.6 | 21.2 | 84.8 | 73.1 |
| *w/o* CMU | 57.2 | 83.7 | 81.2 | 16.8 | 81.4 | 64.1 |
| *w/o* CMU&DEF | 80.0 | 90.3 | 79.0 | 6.8 | 19.1 | 55.0 |

Table 3. Ablation study of MM-OVSeg on the DDHR-SK→DDHR-SK segmentation task under cloudy conditions. The proposed DEF module enables effective multimodal fusion, substantially improving over the single-modality baseline. In combination with CMU, the full model achieves the best performance, demonstrating that CMU and DEF are complementary.

# Thanks