Exploring Efficient Open-Vocabulary Segmentation in the Remote Sensing

Bingyu Li [1, 3*], **Haocheng Dong** [1, 3], **Da Zhang**[2, 3], **Zhiyuan Zhao**[3], **Hao Sun**[3], **Junyu Gao**[3, 2†]

[1]Department of Electronic Engineering and Information Science, University of Science and Technology of China, China
[2]School of Artificial Intelligence, OPtics and ElectroNics (iOPEN), Northwestern Polytechnical University, China
[3]Institute of Artificial Intelligence (TeleAI), China
libingyu0205@mail.ustc.edu.cn, haocheng-dong@mail.ustc.edu.cn, dazhang@mail.nwpu.edu.cn, tuzixini@163.com, sun.010@163.com, gjy3035@gmail.com

Abstract

Open-Vocabulary Remote Sensing Image Segmentation (OVRSIS), an emerging task that adapts Open-Vocabulary Segmentation (OVS) to the remote sensing (RS) domain, remains underexplored due to the absence of a unified evaluation benchmark and the domain gap between natural and RS images. To bridge these gaps, we first establish a standardized OVRSIS benchmark (**OVRSISBench**) based on widely-used RS segmentation datasets, enabling consistent evaluation across methods. Using this benchmark, we comprehensively evaluate several representative OVS/OVRSIS models and reveal their limitations when directly applied to remote sensing scenarios. Building on these insights, we propose **RSKT-Seg**, a novel open-vocabulary segmentation framework tailored for remote sensing. RSKT-Seg integrates three key components: (1) a Multi-Directional Cost Map Aggregation (RS-CMA) module that captures rotation-invariant visual cues by computing vision-language cosine similarities across multiple directions; (2) an Efficient Cost Map Fusion (RS-Fusion) transformer, which jointly models spatial and semantic dependencies with a lightweight dimensionality reduction strategy; and (3) a Remote Sensing Knowledge Transfer (RS-Transfer) module that injects pre-trained knowledge and facilitates domain adaptation via enhanced upsampling. Extensive experiments on the benchmark show that RSKT-Seg consistently outperforms strong OVS baselines by +3.8 mIoU and +5.9 mACC, while achieving 2× faster inference through efficient aggregation.

Code — https://github.com/LiBingyu01/RSKT-Seg
Extended version — https://arxiv.org/pdf/2509.12040

Introduction

Semantic segmentation, a classic task in computer vision, aims to achieve pixel-level category prediction (Chen et al. 2017). Traditional semantic segmentation models are based on manually annotated datasets, which usually cover only a limited number of categories. Similarly, traditional remote sensing image segmentation has long focused on a fixed set of predefined categories (Kotaridis and Lazaridou 2021; Diakogiannis et al. 2020). However, as application demands grow, especially in scenarios like tracking new urban infrastructure or identifying rare geological features, the shortcomings of this limited-category-based approach have become evident. This has spurred the development of open-vocabulary remote sensing image segmentation (OVRSIS) (Ye, Zhuge, and Zhang 2025).

OVRSIS, builded on the general OVS concept(Li et al. 2025a; Cho et al. 2024; Xie et al. 2024b; Xu et al. 2023), is tailored to the unique characteristics of remote sensing imagery. By leveraging the power of VLMs and cross-modal learning (Radford et al. 2021; Jia et al. 2021), OVRSIS breaks the constraints of traditional training categories, enabling the segmentation of new classes not presented in the original datasets. This significantly improves the adaptability and generalization ability of remote sensing image segmentation.

However, unlike the recent advances in OVS (Xu et al. 2023; Radford et al. 2021; Jia et al. 2021), its extension to the remote sensing (RS) domain (OVRSIS) remains largely underexplored. A critical bottleneck hindering progress in this emerging field is the absence of a standardized evaluation benchmark. Most prior works assess their models on limited datasets or under inconsistent experimental setups, making it difficult to draw fair comparisons or systematically analyze model behavior in remote sensing scenarios. To address this issue, we construct a unified benchmark for OVRSIS, named **OVRSISBench**, by reformulating several widely-used remote sensing segmentation datasets under open-vocabulary settings. By leveraging established RS datasets with open-vocabulary configurations, OVRSISBench retains the domain-specific challenges of traditional RS tasks while incorporating the flexibility and generalization demands of open-vocabulary segmentation. This benchmark facilitates fair, consistent, and scalable evaluation of OVS models in realistic RS environments.

Using this **OVRSISBench** benchmark, we comprehensively evaluate several representative OVS models and observe a significant drop in performance when these models are directly transferred from natural images to the RS domain (see Fig. fig. 1(a-b)). Furthermore, we survey recent works on OVRSIS model and evaluate them under the proposed **OVRSISBench**. Our experiments reveal that while some approaches demonstrate incremental improvements compared with OVS model, they often fail to achieve

*Work done during an internship at TeleAI
†Corresponding author

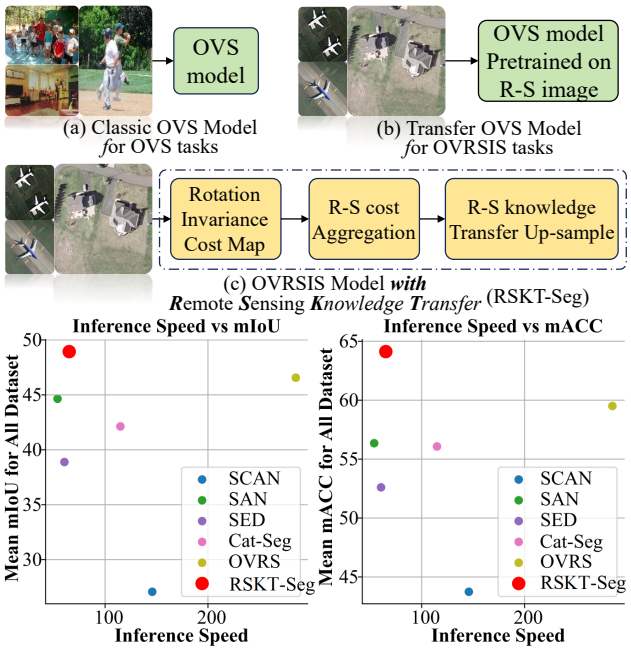

Figure 1: **(a-c):** Comparison of RSKT-Seg with classic OVS and OVRSIS model. **(d):** Comparison of RSKT-Seg with different models in terms of inference speed against mean Intersection over Union (mIoU) on the left and against mean Accuracy (mACC) on the right.

effective knowledge transfer to the RS domain. For example, OVRS (Cao et al. 2024), which introduces only minor modifications to OVS model (Cho et al. 2024), shows performance gains but lacks mechanisms to incorporate RS-specific priors, leading to limited adaptability. Similarly, models such as GSNet (Ye, Zhuge, and Zhang 2025), although partially incorporating RS knowledge, overlook key characteristics of RS imagery such as large-scale context, object rotation, and spectral diversity.

Overall, these findings highlight two key limitations of existing OVS and OVRSIS approaches: (1) classic OVS models exhibit limited transferability when applied directly to open-vocabulary tasks in the remote sensing domain; (2) existing OVRSIS methods fail to adequately model remote sensing-specific characteristics such as rotation invariance and large-scale spatial context.

To fill these gaps, we propose **RSKT-Seg**, an efficient framework for open-vocabulary segmentation in remote sensing that achieves both high segmentation accuracy and fast inference speed (see Fig. fig. 1(d)). First, we introduce a *Multi-Direction Remote Sensing Cost Map Aggregation* (RS-CMA) module to capture the rotation-invariant characteristics of RS images by computing vision-text similarities from multiple directions. Second, we design an efficient *Cost Map Fusion* (RS-Fusion) strategy that simultaneously considers spatial and class-wise interactions, while incorporating a dimensionality reduction mechanism to accelerate inference without sacrificing performance. Finally, we propose a *Remote Sensing Knowledge Transfer* (RS-Transfer) upsampling module that leverages pre-trained model knowledge to facilitate effective domain adaptation to RS imagery.

Our main contributions are summarized as follows:

- **Benchmark.** We construct **OVRSISBench**, an unified benchmark for open-vocabulary remote sensing image segmentation.
- **Evaluation.** Based on OVRSISBench, we comprehensively evaluate representative OVS models and recent OVRSIS methods. Our analysis reveals their potential limitations and provides a standardized reference for future research.
- **Framework.** We propose **RSKT-Seg**, an efficient and effective framework for open-vocabulary segmentation in remote sensing, which achieves both high segmentation accuracy and fast inference speed (shown in fig. 1(d)).

Related Works

Semantic Segmentation

Semantic segmentation, a crucial task in computer vision, focuses on pixel-level classification. Over the years, it has seen significant progress with various methods and models(Xiao et al. 2025). Fully convolutional networks (FCNs) (Long, Shelhamer, and Darrell 2015) were an early milestone. As end-to-end models, they enabled direct pixel-wise predictions. Later, SegNet (Badrinarayanan, Kendall, and Cipolla 2017) and U-Net (Ronneberger, Fischer, and Brox 2015; Wang et al. 2024) evolved from FCNs. These encoder-decoder architectures effectively captured low-level and high-level features, gaining popularity. ResNet is widely used as a feature encoder in semantic segmentation, forming the basis of many models (Lin et al. 2018; Li et al. 2025b, 2024d). Recently, Vision Transformers (ViT) (Dosovitskiy 2020; Wu et al. 2024, 2025) have brought new ideas. By using self-attention, ViT models capture long-range dependencies. Adaptations like SETR (Zheng et al. 2021) improve efficiency. New paradigms based on ViT have emerged for semantic segmentation (Li et al. 2024b; Hu et al. 2024; Yu et al. 2025), which requires precise pixel-level feature handling (Xie et al. 2024a; Zhu et al. 2024; Li et al. 2024a). Notable works such as MaskFormer (Cheng, Schwing, and Kirillov 2021) and Mask2Former (Cheng et al. 2022) unify pixel-level and mask classification, enhancing performance. SegFormer addresses resolution issues by removing positional encodings and presenting multi-resolution features (Xie et al. 2021). We summarize works about OVS in appendix A.

Open-Vocabulary Remote Sensing Image Segmentation

OVRSIS is an emerging field in a rapid development stage, and preliminary exploratory research work has been carried out. In the process of migrating from the OVS field to the remote sensing field, OVRS (Cao et al. 2024) takes into account the basic characteristics of the remote sensing field, makes some fundamental improvements on cat-seg, and has achieved certain results. GSNet (Ye, Zhuge, and Zhang 2025)introduces a specialist-generalist model, which

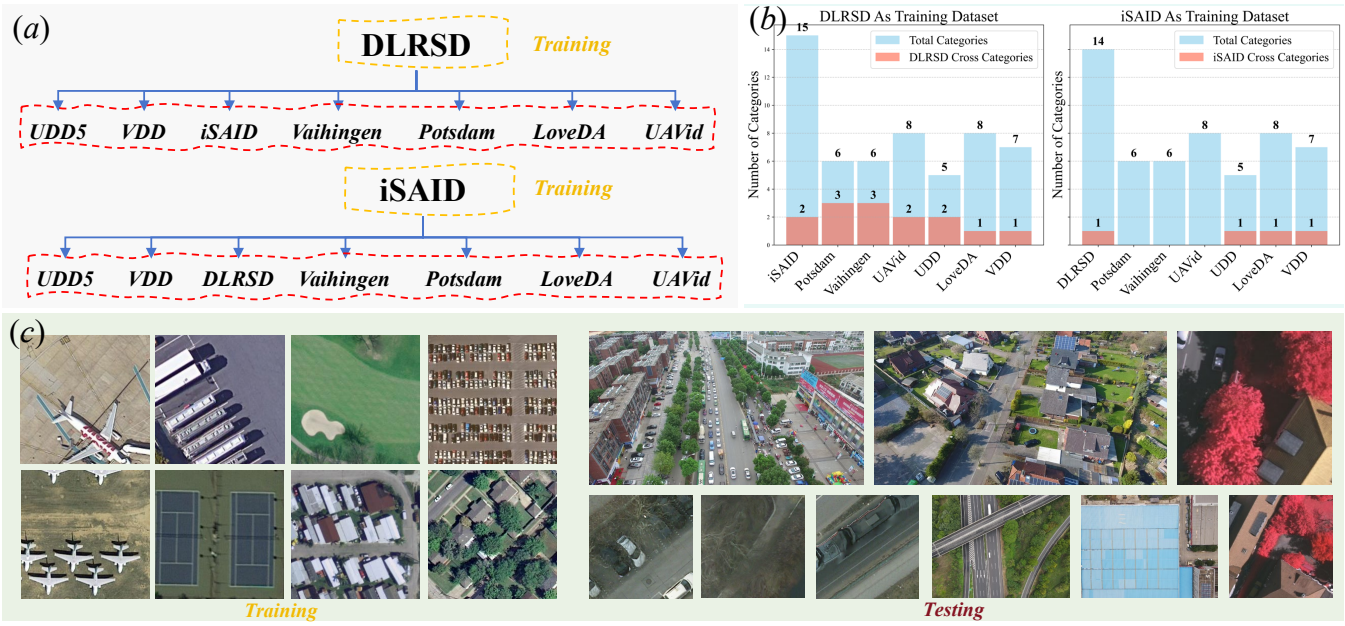

Figure 2: **Schematic diagram of OVRSISBench** (a) Dataset division based on the open-vocabulary protocol (b) Vocabulary (class) overlap number between training and test datasets under two division scenarios (c) Examples display of training and test sets. The more information is in the appendix.

can better integrate remote sensing domain knowledge into the model. SegEarth-OV (Li et al. 2024c) proposes a train-free model, which shows better performance than the OVS model on 17 datasets. Overall, as an emerging field, in OVRSIS, few existing models can consider remote sensing features and introduce remote sensing domain knowledge for migration. Most of them only modify and improve the OVS model. In view of this, we propose the RSKT-Seg model. By considering the rotational invariance of remote sensing images and introducing remote sensing domain knowledge, this model has obtained satisfactory performance on multiple datasets.

Benchmark: OVRSISBench

To promote standardized evaluation in the emerging field of open-vocabulary remote sensing image segmentation (OVRSIS), we construct **OVRSISBench**, a unified and scalable benchmark that reformulates several widely-used remote sensing datasets under an open-vocabulary setting.

Benchmark Construction

Dataset Construction **OVRSISBench** follows an open-vocabulary paradigm like (Cao et al. 2024), where the model must generalize to previously unseen classes based on text descriptions (fig. 2(a)). To construct this benchmark, we adapt 8 representative remote sensing datasets: DLRSD (Chaudhuri et al. 2017), iSAID (Yao et al. 2021), Potsdam, Vaihingen, UAVid(Lyu et al. 2020), UDD5(Chen et al. 2018), LoveDA(Wang et al. 2021) and VDD(Cai et al. 2025). These datasets cover a diverse range of scenes, including urban layouts, agricultural regions, and high-resolution aerial imagery, the examples of the datasets are illustrated in the fig. 2(c). We introduce the datasets in the appendix C.

Dataset Usage and Splits As shown in fig. 2(a), we utilize DLRSD and iSAID as training sets as (Cao et al. 2024) due to their large scale and diversity, enabling robust learning of remote sensing visual patterns. For evaluation, we test across all 8 datasets to ensure the model's generalization under varying scene distributions and resolutions. The training and evaluation splits are consistent across all methods to ensure fair comparison.

Open-Vocabulary Protocol Analysis

To rigorously assess the open-vocabulary generalization capability of the models, particularly its performance on unseen classes. We adopt a cross-dataset transfer protocol in our **OVRSISBench**. In this setting, the models are trained on the DLRSD and iSAID datasets and evaluated on the separate target datasets with partially disjoint category sets. This protocol aligns with the standard open-vocabulary segmentation (OVS) setting(Xu et al. 2022; Cho et al. 2024). As shown in fig. 2(b), we further perform a statistical analysis of the cross-category distribution to substantiate the open-vocabulary nature of our setup.

Evaluation on OVRSISBench

Model Selection and Experiments We first evaluated the performance of several classic OVS model on **OVRSISBench**, and subsequently selected recently published OVRSIS methods for evaluation. For the evaluation metrics, we adopted mean Intersection over Union (mIoU) and mean Accuracy (mACC). We introduce the Evaluation metric in the

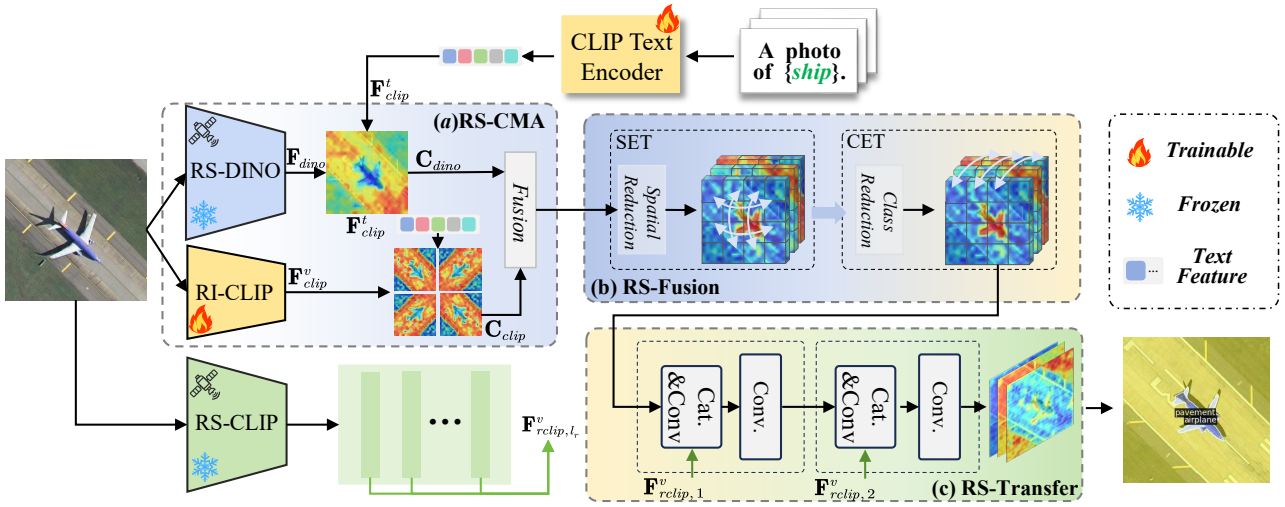

Figure 3: **The overall framework of RSKT-Seg** includes: (a) the overall procedure of RS-CMA module; (b) the workflow of the RS-Fusion Module; (c) the framework of the RS-Transfer Upsample. The more detailed framework is in appendix J

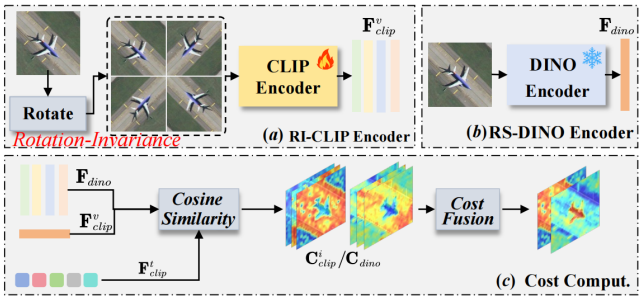

Figure 4: **(a)** Multi-rotation feature encoding using CLIP **(b)** feature encoding using RS-DINO and **(c)** cost map construction using CLIP and DINO.

appendix D. We utilized the pre-trained ViT-B/16 (Dosovitskiy 2020) and ViT-L/14@336 (Radford et al. 2021) as the vision-language models.

Analysis of Existing Model Limitations Table 1 shows classic OVS methods and existing OVRSIS methods perform poorly overall, with clear room for improvement. Classic OVS methods, designed for natural scenes, lag behind dedicated OVRSIS methods under the same backbone. In contrast, although existing OVRSIS methods perform better than classic OVS methods, they still have limitations in fully exploiting remote sensing domain knowledge.

Method: RSKT-Seg

Although previous OVRSIS models have achieved certain advancements, as we analyzed in the introduction section, we attribute the low accuracy observed in the tests on OVRSISBench to the lack of effective transfer of remote sensing domain knowledge. Therefore, we have designed the **RSKT-Seg** with remote sensing domain knowledge transfer capabilities.

Overall Architecture of RSKT-Seg

The overall framework of our proposed RSKT-Seg is shown in Figure fig. 3(a). We design the OVRSIS architecture with remote sensing knowledge transfer. The preparatory knowledge for training pre-trained VLMs such as CLIP has been extensively introduced in previous methods (Hu et al. 2024). The definition of OVRSIS is the same as that of OVS. We will not elaborate on it here, and researchers can refer to the corresponding sections of relevant literature (Li et al. 2025a; Cho et al. 2024; Xu et al. 2023).

Remote Sensing Cost Map Aggregation (RS-CMA)

As shown in fig. 2(c), unlike natural images, remote sensing (RS) imagery often captures scenes from top-down perspectives, where object orientation is arbitrary due to aerial rotation or satellite orbit variations. Consequently, the same semantic class (e.g., "bridge", "airplane") may appear under drastically different orientations across samples. This intrinsic variability poses a significant challenge for open-vocabulary recognition. To address this, we explicitly introduce **rotation-invariance** through our *Multi-Direction Remote Sensing Cost Map Aggregation (RS-CMA)* module.

Specifically, we augment the input image $\mathbf{I}$ with the height H and width W in four directions, obtaining $\mathbf{I}_i$ ($i = 0, 1, 2, 3$) with the same shape (shown in fig. 4(a)). Among them, $\mathbf{I}_0$ is the original orientation, and 1, 2, 3 respectively represent the enhanced images in the orientations rotated by $90 \times i$.

As fig. 4(a-b), we use the CLIP image encoder $\mathcal{E}^v_{clip}$ to encode $\mathbf{I}_i$, yielding $\mathbf{F}^{v,i}_{clip} = \mathcal{E}^v_{clip}(\mathbf{I}_i)$. To embed remote-sensing domain knowledge, we employ a pre-trained DINO encoder $\mathcal{E}_{dino}$ on a large-scale dataset (Ye, Zhuge, and Zhang 2025). Encoding $\mathbf{I}_0$ with $\mathcal{E}_{dino}$, we get $\mathbf{F}_{dino} = \mathcal{E}_{dino}(\mathbf{I}_0)$.

Given a set of classes $\mathcal{C} \in \mathbb{R}^{N_t}$, we obtain text embeddings $\mathbf{F}^t_{clip} = \mathcal{E}^t_{clip}(\mathcal{C}) \in \mathbb{R}^{N_t \times C_f}$. Then, for each $\mathbf{I}_i$, we

DLRSD as Training Dataset

Method	Backbone	Type	DLRSD		iSAID		LoveDA		Potsdam		UAVid		UDD5		Vaihingen		VDD		Mean of All Datasets	
			mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
SCAN$_{CVPR2024}$	ViT-B	OVS	48.52	68.68	34.18	49.05	18.23	40.12	20.22	34.70	18.56	30.31	32.12	40.45	5.38	22.54	26.25	43.67	25.43	41.19
SAN$_{CVPR2023}$	ViT-B	OVS	85.73	91.03	30.63	44.03	23.15	48.26	30.30	44.98	22.34	37.56	36.87	47.21	31.92	45.36	34.76	52.42	36.96	51.36
SED$_{CVPR2024}$	ConvNeXt-B	OVS	85.13	91.36	21.54	36.28	21.32	45.17	19.47	33.40	20.12	33.34	34.65	44.10	29.40	49.38	31.43	50.25	32.88	47.91
Cat-Seg$_{CVPR2024}$	ViT-B	OVS	85.84	91.44	23.56	38.48	25.45	50.32	26.79	44.72	24.56	39.20	38.23	49.56	32.32	49.65	36.18	54.30	36.62	52.21
OVRS$_{TGRS2025}$	ViT-B	OVRSIS	85.98	91.52	39.09	54.43	28.67	52.10	27.47	42.07	25.23	40.18	39.10	50.65	33.71	49.01	37.34	55.15	39.57	54.51
GSNet$_{AAAI2025}$	ViT-B	OVRSIS	84.12	90.53	42.00	59.19	29.32	53.02	26.46	43.20	25.42	40.70	40.05	51.72	35.15	52.62	38.10	56.01	40.08	55.87
RSKT-Seg	ViT-B	OVRSIS	**90.60**	**94.89**	**44.04**	**61.23**	**32.49**	**55.67**	**34.53**	**50.71**	**28.14**	**42.86**	**42.99**	**57.81**	**37.16**	**54.81**	**41.22**	**58.09**	**43.90**	**59.51**
SCAN$_{CVPR2024}$	ViT-L	OVS	52.42	72.43	44.28	67.25	23.17	35.36	27.45	39.22	20.28	34.43	34.14	43.25	15.23	29.45	29.24	45.57	30.78	45.87
SAN$_{CVPR2023}$	ViT-L	OVS	86.45	91.25	49.56	67.25	25.33	37.54	37.25	46.28	23.53	38.14	37.23	48.45	39.22	48.33	35.83	53.25	41.80	53.81
SED$_{CVPR2024}$	ConvNeXt-L	OVS	87.68	91.24	51.23	68.24	24.55	36.83	29.35	37.95	21.33	35.64	35.73	45.15	39.02	58.62	32.53	51.34	40.18	53.13
Cat-Seg$_{CVPR2024}$	ViT-L	OVS	88.68	93.34	53.34	70.86	28.64	38.73	35.75	49.03	25.73	40.54	40.24	51.65	42.30	60.65	39.14	55.85	44.23	57.58
OVRS$_{TGRS2025}$	ViT-L	OVRSIS	88.85	93.64	52.65	69.59	31.53	59.82	36.44	50.17	24.13	34.83	40.82	54.24	43.50	**63.31**	37.23	56.34	44.39	60.24
GSNet$_{AAAI2025}$	ViT-L	OVRSIS	86.02	91.48	53.73	71.57	32.52	**60.23**	37.85	52.35	24.22	35.03	40.92	57.04	**44.13**	62.38	37.34	57.04	44.56	60.89
RSKT-Seg	ViT-L	OVRSIS	**93.49**	**96.49**	**54.32**	**71.72**	**33.23**	57.41	**38.44**	**56.69**	25.72	**40.89**	**42.10**	56.39	42.69	63.29	**39.69**	**58.60**	**46.21**	**62.69**

iSAID as Training Dataset

Method	Backbone	Type	DLRSD		iSAID		LoveDA		Potsdam		UAVid		UDD5		Vaihingen		VDD		Mean of All Datasets	
			mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
SCAN$_{CVPR2024}$	ViT-B	OVS	16.09	38.25	62.34	76.48	12.56	32.10	18.25	33.17	9.87	15.30	8.45	28.60	8.72	27.20	15.32	35.10	18.95	35.78
SAN$_{CVPR2023}$	ViT-B	OVS	18.82	42.36	85.43	90.36	18.45	40.25	14.82	34.84	12.32	18.32	12.11	32.45	16.23	34.38	22.10	40.25	25.04	41.65
SED$_{CVPR2024}$	ConvNeXt-B	OVS	21.48	45.15	93.31	96.66	20.10	42.30	5.78	17.52	14.80	18.90	10.23	31.50	9.36	21.62	17.32	38.12	24.05	38.97
Cat-Seg$_{CVPR2024}$	ViT-B	OVS	20.41	44.08	94.16	96.72	23.50	41.55	15.23	37.17	15.47	23.57	12.10	38.55	14.03	38.61	19.62	**51.38**	26.82	46.45
OVRS$_{TGRS2025}$	ViT-B	OVRSIS	21.06	45.48	**94.60**	**96.87**	25.30	42.32	15.57	38.94	16.22	24.05	11.90	38.80	14.66	38.68	21.42	51.25	27.59	47.05
GSNet$_{AAAI2025}$	ViT-B	OVRSIS	**26.20**	**57.07**	90.00	93.60	26.80	42.84	15.12	36.16	15.93	24.33	12.44	39.62	14.25	41.15	22.22	42.33	27.87	47.14
RSKT-Seg	ViT-B	OVRSIS	24.80	55.94	93.16	96.37	**28.07**	**46.58**	**20.28**	**46.71**	**17.15**	**34.78**	**13.94**	**43.47**	**17.47**	**50.84**	**25.34**	47.18	**30.03**	**52.73**
SCAN$_{CVPR2024}$	ViT-L	OVS	21.44	53.26	64.28	85.46	18.50	38.20	28.32	52.47	14.32	26.00	16.22	32.51	14.23	34.25	20.18	36.70	24.69	44.86
SAN$_{CVPR2023}$	ViT-L	OVS	20.54	49.32	87.22	92.54	22.33	44.22	24.72	**56.54**	15.01	26.50	21.35	37.66	22.49	50.78	26.42	42.30	30.01	49.98
SED$_{CVPR2024}$	ConvNeXt-L	OVS	23.80	50.36	94.32	96.84	23.24	45.13	11.85	23.87	15.21	26.80	22.42	39.15	12.61	**25.73**	28.50	44.20	28.99	44.01
Cat-Seg$_{CVPR2024}$	ViT-L	OVS	28.80	59.56	94.77	96.96	25.11	48.25	23.90	49.49	16.10	26.90	24.32	42.77	21.74	51.25	30.16	47.20	33.11	52.80
OVRS$_{TGRS2025}$	ViT-L	OVRSIS	**32.25**	60.35	**94.86**	**97.06**	27.98	49.01	26.39	50.15	16.24	27.00	31.88	55.32	**28.80**	**54.20**	31.01	55.30	36.18	56.05
GSNet$_{AAAI2025}$	ViT-L	OVRSIS	31.50	**63.05**	93.11	95.98	27.21	49.53	28.50	52.00	15.98	27.10	32.24	56.20	25.10	52.80	32.07	55.21	35.71	56.48
RSKT-Seg	ViT-L	OVRSIS	31.57	61.66	93.96	96.63	**29.62**	**51.01**	**28.57**	51.69	**17.36**	**35.75**	**34.01**	**57.38**	25.55	52.66	**33.40**	**56.50**	**36.76**	**57.91**

Table 1: **Main Results of the RSKT-Seg on 8 OVRSIS Datasets.** Evaluation metrics: mIoU, mACC. Best results are bold, second-best underlined. m-mIoU and m-mACC denote average values across all datasets.

compute a cost volume $\mathbf{C}^i_{clip} \in \mathbb{R}^{(H_f \times W_f) \times N_t}$ using cosine similarity:

$$\mathbf{C}^i_{clip}(j,n) = \frac{\mathbf{F}^{v,i}_{clip}(j) \cdot \mathbf{F}^t_{clip}(n)}{\|\mathbf{F}^{v,i}_{clip}(j)\| \|\mathbf{F}^t_{clip}(n)\|} \quad (1)$$

where j represents 2D spatial positions of the image embedding and n is an index for a class.

Similarly, for the DINO-based features, we compute a cost volume $\mathbf{C}_{dino} \in \mathbb{R}^{(H_f \times W_f) \times N_t}$ for $\mathbf{I}$ as:

$$\mathbf{C}_{dino}(j,n) = \frac{\mathbf{F}_{dino}(j) \cdot \mathbf{F}^t_{clip}(n)}{\|\mathbf{F}_{dino}(j)\| \|\mathbf{F}^t_{clip}(n)\|} \quad (2)$$

Next, we apply a cost map fusion function across the four rotated variants and the DINO-based cost map to generate a rotation-invariant and domain-aware fused cost map (as shown in fig. 4(c)):

$$\mathbf{C}_{fused} = Fusion(\mathbf{C}^i_{clip}, \mathbf{C}_{dino}), \quad i = 0,1,2,3 \quad (3)$$

The fused cost map is further organized under multiple prompt templates, forming a tensor of shape $H_f \times W_f \times N_t \times P$, where P denotes the number of templates. $\mathbf{C}_{fused}$ is projected into $\mathbf{C}_s$ with the shap $H_f \times W_f \times N_t \times C_f$ by a linear layer.

Efficient Cost Map Fusion (RS-Fusion)

The cost map $\mathbf{C}_s$ is used to characterize the correlation between visual features and text features. Building upon this, we gradually predict the segmentation map by enhancing the cost map's **spatial discriminative ability** ($H_f W_f$ dimension) and **class discriminative ability** (N_t dimension). To achieve this, we have designed two dedicated modules: the Spatial Enhancement Transformer (SET) and the Class Enhancement Transformer (CET), as illustrated in fig. 3(b) and (c). This idea aligns with that presented in (Cho et al. 2024). To further strengthen the integration of remote sensing domain knowledge, we embed knowledge from the pre-trained DINO model together with the cost map. Additionally, considering the requirement for faster inference speed, we have developed two feature dimension reduction methods and a lightweight transformer architecture.

Spatial Enhancement Transformer(SET) To enhance the semantic representation with richer spatial information, we concatenate the cost map with intermediate-level features from both CLIP and DINO. These features are first projected to a unified dimension and then passed through a convolutional layer $\mathcal{R}_s$ that reduces the spatial resolution, yielding a compressed tensor for fast inference (as shown in fig. 3(b)):

$$\mathbf{C}_{sr} = \mathcal{R}_s([\mathbf{C}_s; \mathbf{F}^v_{clip}; \mathbf{F}_{dino}]) \quad (4)$$

Here, $\mathcal{R}_s$ denotes the spatial reduction module and $[\cdot;\cdot;\cdot]$ is concatenate operation. Then the reduced feature $\mathbf{C}_{sr}$ is used as key and value in a Transformer block designed to aggregate spatial context via cross-attention manner:

$$\mathbf{C}_{so} = \text{Transformer}_s(\mathbf{C}_s, \mathbf{C}_{sr}, \mathbf{C}_{sr}) \quad (5)$$

Class Enhancement Transformer(CET) As shown in fig. 3(b), we further refine the fused features $\mathbf{C}_{so}$ in the class dimension by integrating the text features $\mathbf{F}^t_{clip}$ from CLIP. After applying average pooling $Avg(\cdot)$ to downsample the spatial resolution:

$$\mathbf{C}_{cr} = \text{Avg}([\mathbf{C}_{so}; \mathbf{F}^t_{clip}]), \tag{6}$$

we reshape the cost map to match the shape of repeated text features. The two are concatenated and passed through a second Transformer block to process the class dimension:

$$\mathbf{C}_{co} = \text{Transformer}_c(\mathbf{C}_{cr}, \mathbf{C}_{cr}, \mathbf{C}_{cr}) \tag{7}$$

This allows the model to capture interactions across different categories.

Through iterative processing across N layers, the aforementioned SET and CET enhance the features of the cost map in both spatial and class dimensions. This strengthens the discriminability of the cost map for objects belonging to different categories as well as for objects within the same class, ultimately forming the fused cost map $\mathbf{C}_{agg}$. The analysis for Dimension Reduction is in the appendix B.

Dataset	Proposed Components				Metrics		
	R-I	DINO	RST	RSF	mIoU	fwIoU	mAcc
DLRSD					46.42	51.51	60.23
	✓				48.06	53.00	61.72
		✓			47.47	54.11	62.16
			✓		47.68	52.92	61.60
	✓	✓			48.72	54.70	62.56
	✓	✓	✓		48.76	54.94	63.11
	✓	✓	✓	✓	**48.94**	**55.46**	**64.12**
iSAID					35.51	35.54	57.36
	✓				38.42	37.96	58.86
		✓			37.59	37.66	58.42
			✓		35.36	36.30	57.91
	✓	✓			38.93	38.86	58.47
	✓	✓	✓		39.07	40.31	61.24
	✓	✓	✓	✓	**39.80**	**41.59**	**62.55**

Table 2: **The Effectiveness of Proposed Components.** R-I: R-I Cost Map, DINO: DINO Cost Map, RST: RS-Transfer, RSF: RS-Fusion.

RS-Transfer Upsample

The fused cost map $\mathbf{C}_{agg}$ provides spatial and class-discriminative signals but lacks detailed textures due to the low spatial resolution. To address this, as shown in fig. 3(c), we integrate intermediate-layer features from RemoteCLIP vision encoder $\mathbf{F}^v_{rclip}$, CLIP vision encoder $\mathbf{F}^v_{clip}$, and DINO encoder $\mathbf{F}_{dino,l_d}$ during the upsampling process. We concatenate these features with the upsampled cost map $\mathbf{C}_{agg}$ and pass them through a projection module:

$$\mathbf{C}_{agg} = \text{Proj}([\mathbf{C}_{agg}; \mathbf{F}^v_{rclip,l_r}; \mathbf{F}^v_{clip,l}; \mathbf{F}_{dino,l_d}]) \tag{8}$$

Here, Proj$(\cdot)$ is a feature projection module. After N_d layers of refinement, the final prediction O is generated in the original image resolution, with shape $N_t \times H \times W$.

Pretrained / Augmented Models				Metrics		
rotate	RS-DINO	N-DINO	RCLIP	mIoU	fwIoU	mAcc
				47.17	52.26	60.48
	✓			**48.22**	**54.86**	**62.41**
		✓		46.46	53.87	62.19
✓	✓			**49.47**	**55.45**	**62.81**
✓		✓		46.90	54.96	62.40
✓	✓		✓	**49.37**	**55.29**	**62.74**
✓		✓	✓	47.17	54.11	61.70

Table 3: **Effectiveness of DINO with Remote Sensing Knowledge Using DLRSD as Training Dataset.** RCLIP: RemoteCLIP. RS-DINO: Remote Sensing DINO. N-DINO: Natural Image DINO

Training Loss

We adopt standard cross-entropy loss for supervision. Given a one-hot segmentation mask $\mathrm{M} \in \mathbb{R}^{H \times W \times N_t}$ and the prediction O, the loss is defined as:

$$\mathcal{L}_{ce} = \text{CrossEntropyLoss}(\mathrm{O}, \mathrm{M}) \tag{9}$$

Minimizing this loss encourages the model to produce accurate pixel-level class predictions.

$Fusion$	m-mIoU	m-fwIoU	m-mACC
mean	47.12	51.99	**66.56**
cat	**48.74**	**52.50**	66.24
separate	47.83	51.89	65.17

Table 4: The Influence of **Different Cost Map Fusion Strategies** $Fusion$.

Dataset/ N	2	3	4	5	6
DLRSD	46.87	49.25	48.84	**51.58**	48.20
DLRSD	62.36	64.06	63.11	**65.41**	64.11
iSAID	**39.09**	38.93	37.22	37.84	37.34
iSAID	62.21	**62.47**	60.17	60.29	60.24

Table 5: **The Influence of Different Layers of the Cost Aggregation Module.** We test on 8 datasets in this experiment.

Experiment

Experimental Analysis of RSKT-Seg

Implementation Details of RSKT-Seg Our approach is implemented based on the PyTorch and Detectron2 (Wu et al. 2019) frameworks. The number of layers in RSKT-Upsample N_d is configured as 2. When using the ViT-B/16 architecture as encoder, l is set to the values $\{3, 7\}$, and when using the ViT-L/14 architecture, l is set to $\{7, 15\}$. We use DINO ViT-B/32 pretrained in (Ye, Zhuge, and Zhang 2025) and l_d is set to $\{3, 7\}$. l_r for RemoteCLIP ViT-B/16 is

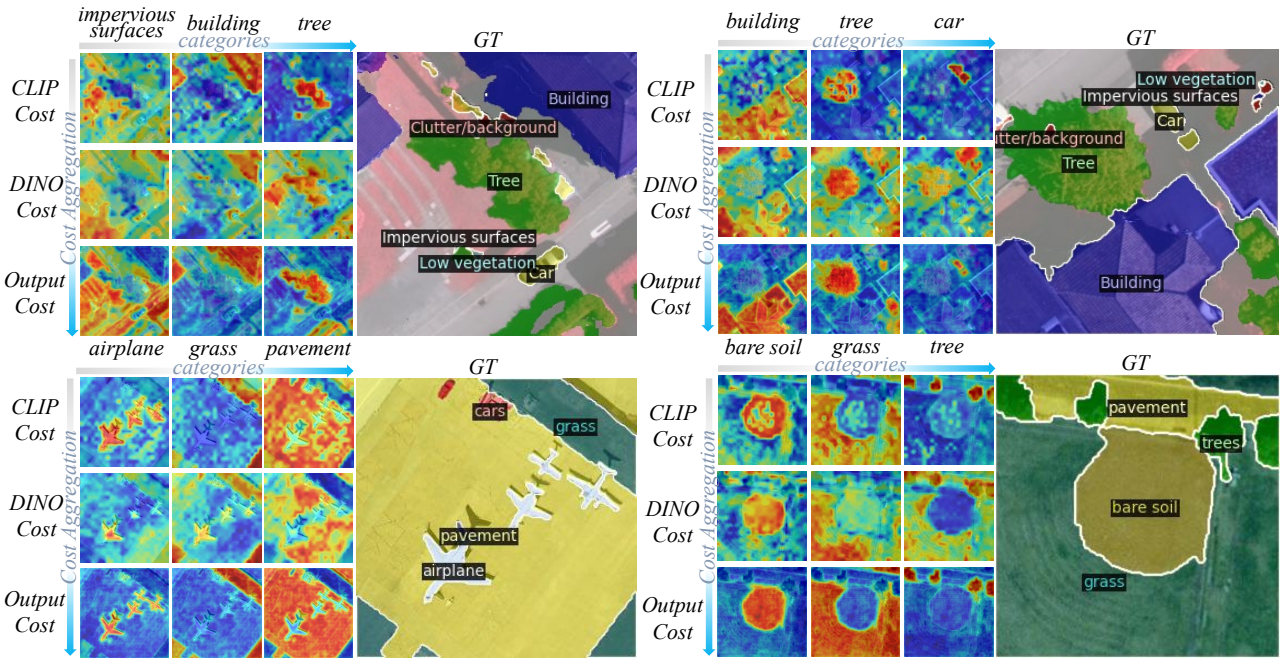

Figure 5: **Comparasion of different cost map and effectiveness of efficient cost map aggregation** (vertical) on different classes(horizontal).

set to $\{3, 7\}$. AdamW is utilized as the optimizer, with an initial learning rate of 2×10^{-4} and a weight decay of 1×10^{-4}. The batch size is fixed at 8. All the experiments are carried out on 4 Nvidia 4090 GPUs, each having 24GB of memory. Our experimental results is that the optimal number of layers N is 5 when using DLRSD as the training set, and 2 when using iSAID.

CLIP	DINO	m-mIoU	m-fwIoU	m-mACC
$\mathcal{F}$	$\mathcal{F}$	28.80	29.70	45.69
$\mathcal{F}$	$\mathcal{A}$	27.42	29.31	46.79
$\mathcal{A}$	$\mathcal{F}$	**39.07**	**40.31**	**61.24**
$\mathcal{A}$	$\mathcal{A}$	37.96	38.98	61.21

Table 6: **The Ablation Study of Different Fine-tuning Strategies** on ISAID Dataset. The trainable($\mathcal{A}$) CLIP and the frozen($\mathcal{F}$) DINO demonstrate better performance.

Method	Mean (ms)↓	FPS↑
SCAN	145.83	6.86
SAN	53.79	18.59
SED	60.47	16.54
Cat-Seg	114.88	8.70
OVRS	285.37	3.50
RSKT-Seg	65.11	15.36

Table 7: The Comparative Experiment on Inference Speed and FPS.

Results of the Proposed Method on Different Datasets Table 1 shows that RSKT-Seg outperforms both classic OVS methods and existing OVRSIS methods across 8 OVRSIS datasets, demonstrating superior effectiveness in remote sensing open-vocabulary segmentation. Under different training datasets (DLRSD and iSAID) and backbone configurations (ViT-B and ViT-L), it consistently leads in key metrics like m-mIoU and m-mACC RSKT-Seg exhibits strong robustness when switching training datasets, excelling in diverse scenarios such as complex land cover (Potsdam), small objects (UAVid), and challenging datasets (VDD, UDD5). Its performance scales effectively with stronger backbones, with ViT-L bringing significant gains over ViT-B. It overcomes the limitations of classic OVS methods in handling remote sensing-specific challenges.

Ablation Study On RSKT-Seg

Our ablation studies are conducted under the same default settings. All reported average values are obtained from inference results on the four datasets: DLRSD, iSAID, Potsdam, and Vaihingen.

The Effectiveness of Different Components Table 2 validates the effectiveness of each proposed component. Adding R-I Cost Map or DINO Cost Map alone improves metrics on datasets, with R-I Cost Map performing slightly better. Combining them yields further gains (e.g., DLRSD m-mIoU: 48.72 vs. 46.42). Incorporating RS-Transfer brings marginal improvements and adding RS-Fusion achieves the best results, boosting performance.

The importance of introducing remote sensing knowledge The table 3 compares the performance of "rsDINO"

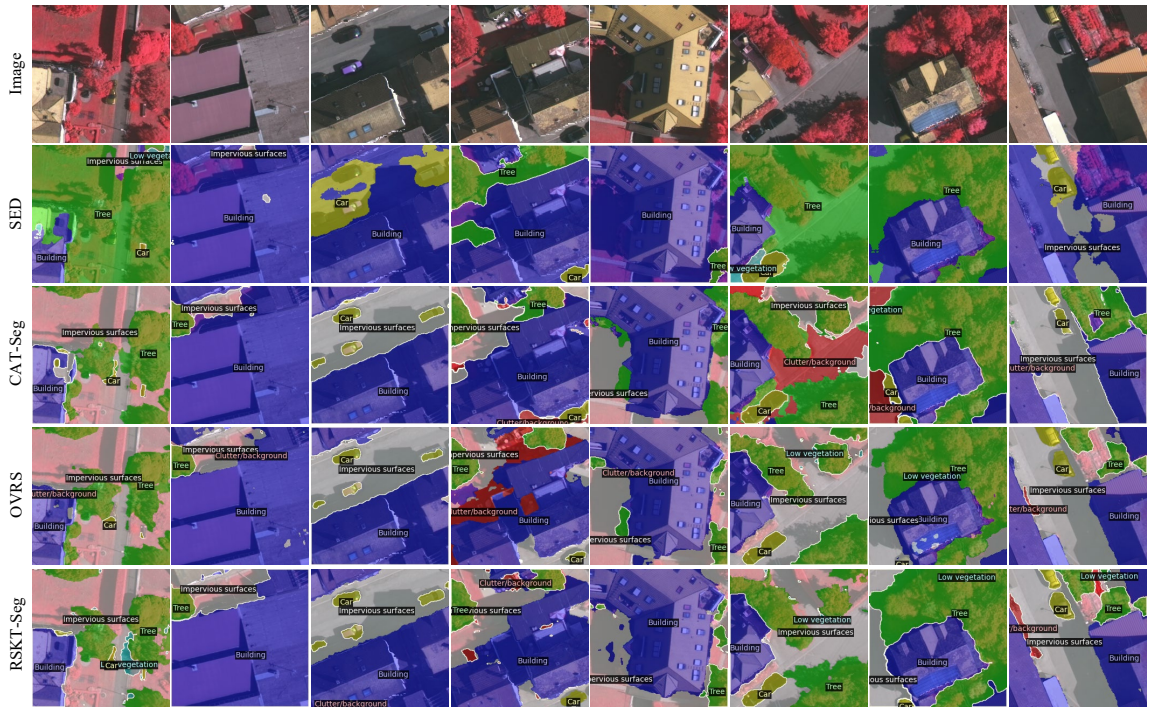

Figure 6: **The comparison of the segmentation visualization of different models.** The model (ViT-L) trained on DLRSD is used to perform inference on Vaihingen.

and "natureDINO" on the DLRSD dataset. When the "DINO" component is added to the model, it improves the model's performance significantly. Evidently, the "rsDINO" pre-trained on remote sensing data is more effective in enhancing the model's performance for the DLRSD dataset.

Ablation of Different cost map fusion strategy The table 4 shows evaluation metrics for different cost map fusion strategies. The values of m-mIoU, m-fwIoU, and m-mACC for "mean", "cat", and "separate" strategies vary slightly. This indicates that these strategies have similar effects on model performance.

The impacts of different layers of cost map aggregation This subsection analyzes the impacts of different layers in the cost map aggregation module using data from table 5 on DLRSD and iSAID datasets, with metrics like m-mIoU, m-fwIoU, and m-mACC. Based on the experimental results, we selected N=5 for the DLRSD dataset and N=2 for the iSAID dataset as the optimal configurations for the Cost Aggregation Module.

The impacts of different finetuning strategies This subsection analyzes the impacts of different finetuning strategies using data from table 6 on the iSAID dataset. The key components are CLIP and DINO, each with frozen and attention states. when CLIP is in attention and DINO frozen (highlighted), the model achieves the best performance among all combinations.

The analysis of inference speed This table presents the inference speed (in ms) and FPS across multiple datasets for different methods. Among them, RSKT-Seg stands out for its significantly improved inference speed compared to Cat-Seg and OVRS. RSKT-Seg only needs 65.11 ms on average. This shows that RSKT-Seg can achieve faster inference, which is beneficial for real-time applications.

Visualization

Cost map visualization on diffetent categories

As seen in fig. 5, cost maps from CLIP Cost and DINO Cost show how models assign costs to different classes. These cost reveal areas of high or low classification confidence.

Comparison of visualization

As depicted in fig. 6, a comparison is made among the segmentation visualizations of multiple models. Compared to other models, RSKT-Seg shows a more precise identification of object boundaries.

Conclusion

In this paper, we address OVRSIS challenges via **OVRSIS-Bench**, a unified benchmark. Evaluations show classic OVS methods perform poorly in remote sensing, and existing OVRSIS methods lack RS-specific modeling. Thus, we propose the **RSKT-Seg** that integrates RS-CMA, RS-Fusion, and RS-Transfer, outperforming baselines with 2× faster inference, advancing OVRSIS research. In the appendix H, we carefully analyzed the gap of open-vocabulary segmentation from natural images to remote-sensing images, as well as the **limitations** of our model, and our **future work**.

Acknowledgments

This work was supported in part by the Natural Science Foundation of China under Grant 62306241, and in part by grants from the Innovation Foundation for Doctor Dissertation of Northwestern Polytechnical University (No.CX2025109).

References

Badrinarayanan, V.; Kendall, A.; and Cipolla, R. 2017. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12): 2481–2495.

Cai, W.; Jin, K.; Hou, J.; Guo, C.; Wu, L.; and Yang, W. 2025. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109: 104429.

Cao, Q.; Chen, Y.; Ma, C.; and Yang, X. 2024. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*.

Chaudhuri, B.; Demir, B.; Chaudhuri, S.; and Bruzzone, L. 2017. Multilabel remote sensing image retrieval using a semisupervised graph-theoretic method. *IEEE Transactions on Geoscience and Remote Sensing*, 56(2): 1144–1158.

Chen, L.-C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; and Yuille, A. L. 2017. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(4): 834–848.

Chen, Y.; Wang, Y.; Lu, P.; Chen, Y.; and Wang, G. 2018. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision (PRCV)*, 347–359. Springer.

Cheng, B.; Misra, I.; Schwing, A. G.; Kirillov, A.; and Girdhar, R. 2022. Masked - attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1290–1299. IEEE.

Cheng, B.; Schwing, A.; and Kirillov, A. 2021. Per - pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34: 17864–17875.

Cho, S.; Shin, H.; Hong, S.; Arnab, A.; Seo, P. H.; and Kim, S. J. 2024. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4113–4123.

Diakogiannis, F. I.; Waldner, F.; Caccetta, P.; and Wu, C. 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. *ISPRS Journal of Photogrammetry and Remote Sensing*, 162: 94–114.

Dosovitskiy, A. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Hu, Z.; Gao, J.; Yuan, Y.; and Li, X. 2024. Contrastive Tokens and Label Activation for Remote Sensing Weakly Supervised Semantic Segmentation. *IEEE Transactions on Geoscience and Remote Sensing*.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 4904–4916. PMLR.

Kotaridis, I.; and Lazaridou, M. 2021. Remote sensing image segmentation advances: A meta-analysis. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173: 309–322.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2024a. StitchFusion: Weaving Any Visual Modalities to Enhance Multimodal Semantic Segmentation. *arXiv preprint arXiv:2408.01343*.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2024b. U3M: Unbiased Multiscale Modal Fusion Model for Multimodal Semantic Segmentation. *arXiv preprint arXiv:2405.15365*.

Li, B.; Zhang, D.; Zhao, Z.; Gao, J.; and Li, X. 2025a. FGAseg: Fine-Grained Pixel-Text Alignment for Open-Vocabulary Semantic Segmentation. *arXiv preprint arXiv:2501.00877*.

Li, K.; Liu, R.; Cao, X.; Bai, X.; Zhou, F.; Meng, D.; and Wang, Z. 2024c. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. *arXiv preprint arXiv:2410.01768*.

Li, S.; Ma, W.; Guo, J.; Xu, S.; Li, B.; and Zhang, X. 2024d. Unionformer: Unified-learning transformer with multi-view representation for image manipulation detection and localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12523–12533.

Li, S.; Xing, Z.; Wang, H.; Hao, P.; Li, X.; Liu, Z.; and Zhu, L. 2025b. Toward Medical Deepfake Detection: A Comprehensive Dataset and Novel Method. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 626–637. Springer.

Lin, B.; Xie, J.; Li, C.; and Qu, Y. 2018. Deeptongue: Tongue segmentation via ResNet. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1035–1039. IEEE.

Long, J.; Shelhamer, E.; and Darrell, T. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3431–3440.

Lyu, Y.; Vosselman, G.; Xia, G.-S.; Yilmaz, A.; and Yang, M. Y. 2020. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165: 108 – 119.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III*, 234–241. Springer.

Wang, J.; Zheng, Z.; Ma, A.; Lu, X.; and Zhong, Y. 2021. LoveDA: A Remote Sensing Land-Cover Dataset for Domain Adaptive Semantic Segmentation. In Vanschoren, J.; and Yeung, S., eds., *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1. Curran Associates, Inc.

Wang, X.; Ma, K.; Zhong, R.; Wang, X.; Fang, Y.; Xiao, Y.; and Xia, T. 2024. Towards dual transparent liquid level estimation in biomedical lab: Dataset, methods and practices. In *European Conference on Computer Vision*, 198–214. Springer.

Wu, S.; Liu, Z.; Zhang, B.; Zimmermann, R.; Ba, Z.; Zhang, X.; and Ren, K. 2024. Do as I Do: Pose Guided Human Motion Copy. *IEEE Trans. Dependable Secur. Comput.*, 21(6): 5293–5307.

Wu, S.; Zhang, H.; Liu, Z.; Chen, H.; and Jiao, Y. 2025. Enhancing Human Pose Estimation in Internet of Things via Diffusion Generative Models. *IEEE Internet Things J.*, 12(10): 13556–13567.

Wu, Y.; Kirillov, A.; Massa, F.; Lo, W.-Y.; and Girshick, R. 2019. Detectron2. https://github.com/facebookresearch/detectron2.

Xiao, J.; Chen, Y.; Feng, X.; Wang, R.; and Wu, Z. 2025. RecNet: Optimization for Dense Object Detection in Retail Scenarios Based on View Rectification. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xie, B.; Cao, J.; Anwer, R. M.; Xie, J.; Nie, J.; Yang, A.; and Pang, Y. 2024a. Multi - query and multi - level enhanced network for semantic segmentation. *Pattern Recognition*, 156: 110777.

Xie, B.; Cao, J.; Xie, J.; Khan, F. S.; and Pang, Y. 2024b. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3426–3436.

Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; and Luo, P. 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 34: 12077–12090.

Xu, M.; Zhang, Z.; Wei, F.; Hu, H.; and Bai, X. 2023. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2945–2954.

Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; and Bai, X. 2022. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, 736–753. Springer.

Yao, X.; Cao, Q.; Feng, X.; Cheng, G.; and Han, J. 2021. Scale-aware detailed matching for few-shot aerial image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1–11.

Ye, C.; Zhuge, Y.; and Zhang, P. 2025. Towards Open-Vocabulary Remote Sensing Image Semantic Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Yu, J.; Ma, Z.; Ma, Y.; Liu, K.; Wang, Y.; and Li, J. 2025. MILD: Multi-Layer Diffusion Strategy for Complex and Precise Multi-IP Aware Human Erasing. arXiv:2508.06543.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6881–6890.

Zhu, C.; Li, L.; Wu, Y.; and Sun, Z. 2024. Saswot: Real - time semantic segmentation architecture search without training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 7722–7730.