From Visual Primitives to Semantic Masks: Fine-Grained Visual-Linguistic Alignment for Open-Vocabulary Remote Sensing Image Segmentation

Yuchen Deng, Yang Xu*, Zhihui Wei, and Zebin Wu

School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China
{yuchendeng,xuyangth90,gswei,wuzb}@njust.edu.cn
https://github.com/blackdyc/SANO3

Abstract. Open-vocabulary semantic segmentation (OVSS) makes segmentation driven by text descriptions, enabling generalization to novel classes. However, fine-grained OVSS in remote sensing is challenged by visual similarity and semantic ambiguity, with existing methods lacking explicit fine-grained visual-linguistic alignment and relying on a single forward pass. To address these issues, we propose SANO3, a training-free framework that leverages refined textual and spatial guidance to boost the performance of SAM 3. Leveraging the visual self-similarity priors of vision foundation models, we extract discriminative visual primitives and align them with text through visual-conditioned relation-aware alignment strategy. Next, we enhance text representations with hybrid prototype injection for deep visual-linguistic fusion. To refine boundaries, we treat spatial prompts generation as an iterative optimization process, gradually selecting the most informative points via a greedy entropy reduction strategy. Further, we introduce FG-OVSSRS Bench, the first fine-grained remote sensing OVSS benchmark, encompassing 8 diverse datasets across different sensors and domains. Evaluated on these datasets, SANO3 outperforms existing approaches and achieves state-of-the-art performance.

Keywords: Open-Vocabulary · Fine-Grained Semantic Segmentation · Remote Sensing

1 Introduction

Semantic segmentation is a core task in remote sensing, enabling pixel-level understanding for applications such as environmental monitoring [17], agriculture [26], urban analysis [37], and disaster response [1]. Traditional methods operate under a closed-set setting, requiring predefined categories and retraining for new targets. Open-vocabulary semantic segmentation (OVSS) overcomes

* Corresponding author.

these limitations by recognizing novel classes from textual descriptions, offering greater flexibility and scalability for real-world remote sensing scenarios.

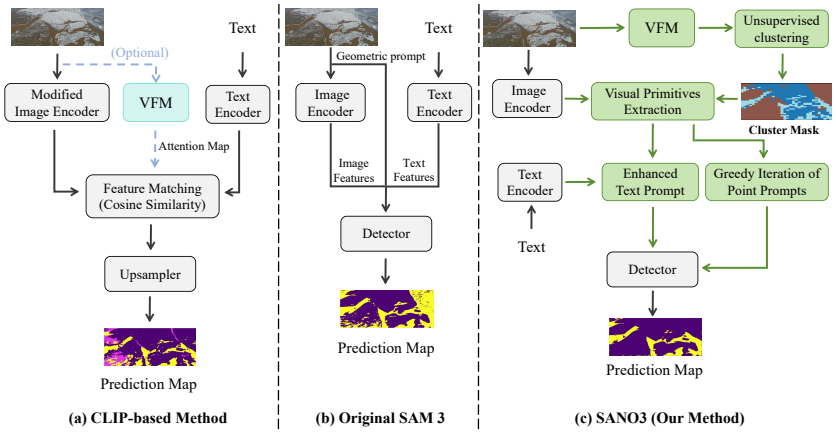

Fig. 1: Comparison of paradigms for OVSS. (a) CLIP-based methods modify attention pooling to obtain dense features, sometimes aided by VFMs, but still suffer from semantic confusion and blurry boundaries. (b) SAM 3 uses native multi-modal encoders, whose text features lack discriminability for fine-grained categories. (c) SANO3 introduces VFM-guided visual primitives and informative point prompts, enabling precise semantic and geometric segmentation.

Driven by the rise of Vision-Language Models (VLMs) such as CLIP [31], recent OVSS methods transfer global image-text alignment to the pixel level. Since CLIP is designed for global correspondence rather than dense prediction, recent works modify its attention pooling layer to extract denser features [22, 36, 38, 43] and sometimes introduce Vision Foundation Models (VFMs) for better spatial localization and local semantic aggregation [20, 23, 40]. Yet they still struggle with fine-grained remote sensing scenes, often producing semantic confusion and blurry boundaries (Fig. 1(a)). These limitations motivate the adoption of SAM 3 [9], whose promptable segmentation objective provides stronger geometric priors for OVSS.

However, directly applying SAM 3 to remote sensing OVSS remains insufficient. Fine-grained scenarios involve not only visual similarity but also ambiguous text queries (*e.g.*, *deciduous trees* vs. *coniferous trees*). Existing methods [5, 28] do not explicitly address this issue, leading to insufficiently discriminative text representations, as shown in Fig. 1(b). Moreover, most existing methods rely on a single forward prediction, failing to fully exploit the iterative and feedback-driven nature of prompt-based segmentation.

To address these issues, we introduce VFM-guided visual priors to enhance SAM 3, as illustrated in Fig. 1(c). Self-supervised VFMs learn topology-aware representations, where the features of patches naturally organize into semanti-

cally coherent clusters, revealing the intrinsic structure of visual scenes [15, 29, 35]. Building upon this key property, we leverage the visual self-similarity in VFMs to decompose an image into discriminative visual primitives via unsupervised clustering. These primitives refine text prompts, enlarging the separability between confusing queries and enabling finer visual-linguistic alignment. Furthermore, we exploit the promptable nature of SAM 3 through an greedy iterative prompt selection that progressively selects informative point prompts based on entropy reduction, refining segmentation boundaries step by step. This collaborative enhancement of textual and spatial guidance enables more accurate semantic and geometric segmentation in complex remote sensing scenes.

In parallel, existing remote sensing OVSS benchmarks [8, 25, 27] remain limited in both scenario diversity and semantic granularity. They typically focus on urban or rural settings with coarse land-cover labels that represent broad natural categories (*e.g.*, *water*, *tree*) and man-made structures (*e.g.*, *road*, *building*). To address this limitation, we introduce FG-OVSSRS Bench, a fine-grained open-vocabulary benchmark spanning diverse remote sensing scenarios with highly discriminative semantic labels.

Our contributions can be summarized as:

- We propose SANO3, a training-free OVSS framework that enhances SAM 3 by refining textual and spatial guidance. It extracts discriminative visual primitives from VFM's self-similarity priors, aligns them with text via visual-conditioned relation-aware alignment, and applies hybrid prototype injection for deep visual-linguistic fusion.
- We develop an iterative information-gain visual prompting strategy that formulates point selection as an optimization process, employing greedy entropy reduction to progressively minimize uncertainty and refine boundaries.
- We introduce FG-OVSSRS Bench, the first fine-grained remote sensing OVSS benchmark, comprising eight diverse datasets across sensors and domains for systematic evaluation under subtle semantic distinctions.

2 Related Work

2.1 Training-free Open-Vocabulary Semantic Segmentation

While training-based methods [8, 13, 25, 39] adapt VLMs with additional supervision to improve dense prediction, they risk reducing the open-vocabulary flexibility inherent in the pre-trained model. In contrast, training-free OVSS methods leverage the priors of VLMs without the need for additional learning, maintaining the inherent flexibility. Early CLIP-based techniques [7, 22, 36, 43] modify attention pooling layers to extract denser features, improving zero-shot segmentation. However, CLIP's pretraining objective focuses on global image-text alignment, limiting its ability to localize fine-grained structures accurately. To mitigate this, recent works [3, 19, 23, 34, 40] incorporate auxiliary foundation models, such as SAM [21], DINO [10], and Stable Diffusion [33], to enhance the localization ability of CLIP. These hybrid approaches aim to address CLIP's

weak spatial localization and local semantic aggregation, but remain suboptimal for distinguishing complex boundaries.

Recently, several studies, such as SegEarth-OV3 [28], have explored the use of SAM 3 for remote sensing OVSS. However, these approaches primarily rely on simple post-processing strategies, combining SAM 3's segmentation head with the Transformer decoder output to enhance mask quality. They do not explicitly model fine-grained visual-text alignment or adaptively incorporate spatial cues—both of which are crucial for accurately recognizing subtle categories and producing complete segmentation boundaries in complex remote sensing scenes.

2.2 Fine-Grained Semantic Segmentation in Remote Sensing

Fine-grained semantic segmentation in remote sensing presents unique challenges due to high inter-class similarity and complex intra-class variance in earth observation data. Most existing methods rely on training-based paradigms, where models are either fine-tuned on task-specific datasets or use domain adaptation to transfer knowledge across regions. Approaches like D2LS [44] and KTDA [41] leverage labeled data to capture subtle semantic differences, while GeoSAM [18] fine-tunes SAM with multi-modal prompts for precise urban infrastructure segmentation. Similarly, DG-Net [24] employs path-selective test-time adaptation to enhance performance by adapting models to unseen samples during inference.

Despite their effectiveness, these approaches share common limitations: they rely on extensive pixel-level annotations, are computationally expensive, and are limited to predefined classes seen during training. As a result, they fail to fully leverage the flexibility of open-vocabulary settings, where models must generalize to unseen categories without additional training. This limitation motivates the development of training-free frameworks like SANO3, designed to address the underexplored challenge of fine-grained OVSS in remote sensing, and the introduction of our proposed FG-OVSSRS Bench.

3 Benchmark: FG-OVSSRS Bench

To encourage standardized assessment in the emerging area of fine-grained remote sensing open-vocabulary semantic segmentation, we introduce FG-OVSSRS Bench. This benchmark integrates eight publicly available remote sensing datasets, spanning diverse domains, sensing platforms, spatial resolutions, and environmental conditions. FG-OVSSRS Bench is designed to assess fine-grained discrimination, cross-domain generalization, and robustness to open-vocabulary prompts in realistic scenarios.

3.1 Benchmark Construction

FG-OVSSRS Bench is constructed by integrating eight representative remote sensing datasets, each corresponding to a distinct application scenario while emphasizing fine-grained semantic characteristics, and follows an open-vocabulary

Table 1: Detailed statistics of the datasets in FG-OVSSRS Bench.

Dataset	Classes	Sensor Type	Scene	Samples	Size
Forest	10	UAV	Vegetation	4303	1280 × 768
Cloud	4	Satellite	Atmosphere	2643	512 × 512
Rescue	11	UAV	Disaster	450	4000 × 3000
YRCC	3	UAV	Cryosphere	363	1600 × 640
SkyScapes	20	Aerial	Urban Land	240	1008 × 1008
GIS	3	Aerial	Urban Infra.	2380	1024 × 1024
RSMI	3	Satellite	Minerals	2050	128 × 128, 256 × 256
Barley	3	Satellite	Agriculture	1271	1008 × 1008

paradigm, where models are required to generalize to previously unseen classes based solely on textual descriptions rather than task-specific retraining.

Specifically, Forest [4] provides pixel-level annotations for deciduous trees, coniferous trees, fallen trees, and ground vegetation, supporting detailed forest monitoring and deforestation analysis. Cloud [14] includes annotations for cloud, thin cloud, cloud shadow, and clear sky, enabling fine-grained atmospheric interpretation. Rescue [32] offers pixel-level labels for multiple post-disaster damage categories, such as building damage levels, road conditions, and affected infrastructure, facilitating detailed disaster assessment. Cryospheric and hydrological monitoring is represented by YRCC [42], which provides annotations for river ice, water, and shore.

Urban scenarios are covered from complementary perspectives. SkyScapes [2] focuses on fine-grained urban land-cover mapping, capturing detailed surface categories and road-related semantics, whereas GIS [18] focuses on urban infrastructure monitoring, such as roads and sidewalks, using high-resolution aerial imagery and GIS-derived annotations. Fine-RSMI [11] represents resource-related applications, focusing on pixel-level segmentation of three major mineral categories: limestone, sandstone, and shale. In contrast, Barley[1] is used for agricultural applications, providing UAV imagery and yield annotations for the fine-grained classification of coix, corn, tobacco, and built-up areas, supporting agricultural asset inventory and crop yield prediction.

To better reflect real-world open-vocabulary usage, we expand the label vocabulary of each dataset according to its semantic definitions and annotation rules by introducing synonyms or semantically equivalent terms for every category. In practical open-vocabulary scenarios, user-provided prompts can vary significantly (*e.g.*, *car*, *vehicle*, or *automobile*), and our vocabulary expansion enables systematic evaluation of model robustness to prompt variations with consistent semantics. Additionally, for categories like *background* or *clutter*, which lack specific semantic meaning, we expand them into their corresponding real-world categories based on the dataset's annotations, such as sky, rubbish, etc.

[1] https://tianchi.aliyun.com/dataset/dataDetail?dataId=74952

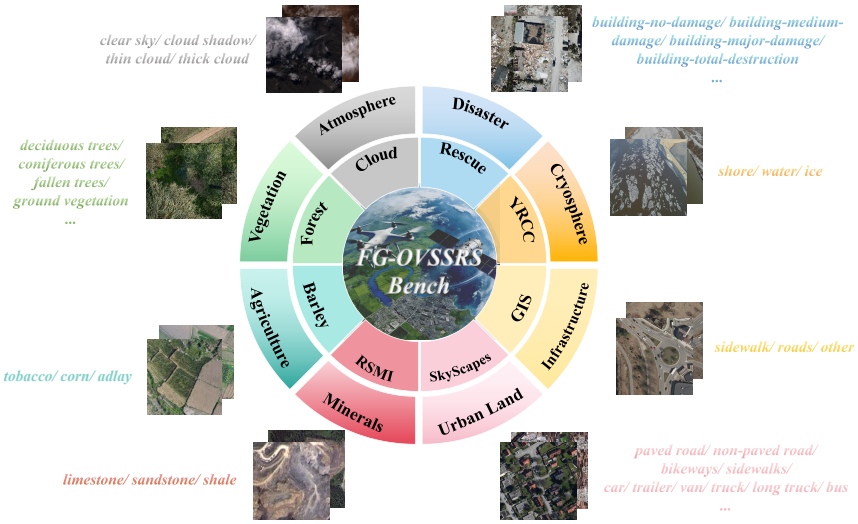

Fig. 2: Overview of the proposed FG-OVSSRS Bench. The inner ring shows the datasets and the outer ring their application domains. Representative fine-grained categories and sample images illustrate the diversity and semantic granularity of the benchmark.

3.2 Statistics and Analysis

FG-OVSSRS Bench exhibits substantial diversity in data scale, spatial resolution, semantic granularity, and application domains, as shown in Fig. 2. It integrates datasets from satellite, aerial, and UAV imagery across diverse regions and environmental conditions. Semantically, it captures fine-grained distinctions in coverage levels (cloud thickness), physical states (building damage), material types (mineral categories), and functional attributes (farmland usage), reflecting realistic variations across remote sensing applications. A detailed summary of each dataset, including image count, spatial resolution, semantic categories, and application domains, is provided in Tab. 1.

3.3 Benchmark Evaluation Protocol

FG-OVSSRS Bench is evaluated in a training-free OVSS setting, where models are applied directly without task-specific fine-tuning or additional supervision on the benchmark datasets. All compared methods are state-of-the-art training-free open-vocabulary segmentation approaches, ensuring a fair and consistent comparison under the same inference-only protocol.

For each dataset, models are provided with textual descriptions sampled from the expanded vocabulary associated with each semantic category and are required to generate pixel-level predictions accordingly. To assess robustness to prompt variation, multiple synonymous or semantically equivalent prompts

are used for each category, with no model parameters updated during evaluation. Performance is quantitatively measured using mean Intersection over Union (mIoU), computed between the predicted segmentation maps and ground-truth annotations. Results are reported per dataset and averaged across datasets to reflect both fine-grained discrimination and cross-domain generalization in a unified training-free open-vocabulary setting.

4 Methods

This section presents the SANO3 framework for training-free fine-grained OVSS in remote sensing. We first introduce SAM 3-based preliminaries (Sec. 4.1) and unsupervised visual primitive discovery (Sec. 4.2). Then, we describe visual-conditioned relation-aware alignment (Sec. 4.3), hybrid prototype injection (Sec. 4.4), and iterative information-gain visual prompting (Sec. 4.5), which together enable precise fine-grained segmentation under subtle semantic variations. An overview of the framework is shown in Fig. 3.

4.1 Preliminaries

We adopt SAM 3 as the foundational architecture, formulating open-vocabulary segmentation as a prompt-conditioned prediction problem. Given an image I and a text prompt t, SAM 3 predicts the spatial correspondence for the queried concept, supporting additional guidance from points, boxes, or masks. The architecture comprises a vision encoder $E_v(\cdot)$ and a text encoder $E_t(\cdot)$, extracting features F_v and F_t, which are fused by a fusion encoder to produce prompt-aware representations. SAM 3 outputs a scalar presence score $s_{\text{pres}} \in [0,1]$, a dense semantic probability map $P_{\text{sem}} \in [0,1]^{H\times W}$, and N instance-level predictions $\{(P_{\text{inst}}^{(k)}, s_{\text{conf}}^{(k)})\}_{k=1}^{N}$, where $P_{\text{inst}}^{(k)} \in [0,1]^{H\times W}$ represents the probability map for the k-th object query, and $s_{\text{conf}}^{(k)} \in [0,1]$ is its associated confidence score. This prompt-driven design offers strong geometric priors, making it well-suited for complex remote sensing tasks.

4.2 Unsupervised Visual Primitive Discovery

To overcome the limitations of cross-modal alignment in capturing fine-grained details, we leverage the intrinsic data topology of DINOv2 [29] to discover latent visual primitives, starting from the intra-modal self-similarity within the visual modality. Given an image I, we extract visual feature map $\mathbf{F}_{vfm}$ using a pre-trained DINOv2 backbone. To capture subtle intra-class variations, we employ an over-clustering strategy, partitioning the feature map into $K_{init} = Q \times C$ disjoint regions via K-Means clustering, where Q represents the number of text queries and C is a hyperparameter that controls the granularity of the clustering.

To reduce fragmentation, we apply a hierarchical merging mechanism. Initial prototypes are computed for each cluster, and pairs with cosine similarity exceeding a threshold (*e.g.*, 0.85) are iteratively merged, yielding a refined set

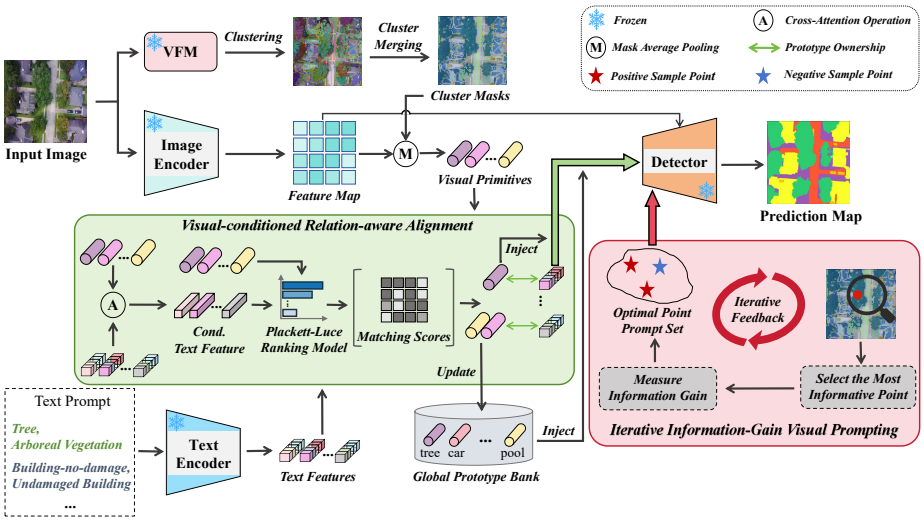

Fig. 3: The overall framework of SANO3. Given an image and a text query, SANO3 extracts visual prototypes via unsupervised clustering and performs relation-aware ranking alignment for semantic assignment. Dynamic and static prototypes are injected into the text stream to form a hybrid multi-modal prompt. An iterative information-gain mechanism then refines boundaries by actively selecting informative point prompts to reduce uncertainty. The framework is entirely training-free.

of masks $\mathcal{M} = \{M_1, \ldots, M_K\}$. Finally, to ensure compatibility with SAM 3's feature space, we align the masks to the resolution of the vision encoder's output $\mathbf{F}_{sam}$, and compute the visual prototypes $\boldsymbol{v}_k$ for each mask M_k by masked average pooling:

$$\mathcal{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_K\}, \quad \boldsymbol{v}_k = \frac{1}{|M_k|} \sum_{(x,y):M_k(x,y)=1} \mathbf{F}_{sam}(x,y) \tag{1}$$

These prototypes serve as concrete visual anchors for the subsequent semantic alignment.

4.3 Visual-conditioned Relation-aware Alignment

Directly matching the visual prototypes $\mathcal{V}$ to the textual representations via cosine similarity is inherently suboptimal for fine-grained alignment. SAM 3's text encoder suffers from feature collapse, where embeddings for semantically subtle concepts cluster tightly in the latent space. This issue is further discussed in Sec. 5.3. To resolve this, we propose a two-stage alignment strategy: visual-conditioned text refinement followed by dynamic relation-aware matching.

Visual-conditioned Text Refinement Instead of relying on the static sentence-level [EOS] embedding or global average pooling embedding, which often loses

granular details, we dynamically query token-level text features using the visual prototypes. Let $\mathbf{T}_c \in \mathbb{R}^{L\times d}$ be the token embeddings extracted by $E_t(\cdot)$ for the text query c, where L is the sequence length. For each visual prototype $\boldsymbol{v}_k \in \mathcal{V}$ and text query c, we compute a visual-conditioned text feature $\boldsymbol{t}_{k,c} \in \mathbb{R}^d$ via cross-attention:

$$\boldsymbol{t}_{k,c} = \sum_{l=1}^{L} \alpha_{k,c,l} \mathbf{T}_c^{(l)}, \quad \text{with} \quad \alpha_{k,c,l} = \mathrm{Softmax}_l\left(\frac{\boldsymbol{v}_k^\top \mathbf{T}_c^{(l)}}{\tau_{attn}}\right) \tag{2}$$

Here, $\alpha_{k,c,l}$ represents the attention weight of the l-th token for the k-th visual prototype and c-th query, and τ_{attn} is a temperature parameter. Collecting all features yields $\boldsymbol{t}_{cond} \in \mathbb{R}^{K\times Q\times d}$, where $\boldsymbol{t}_{cond}[k,c] = \boldsymbol{t}_{k,c}$. This mechanism selectively attends to informative tokens, enhancing fine-grained discriminability.

Dynamic Relation-aware Matching To achieve more robust prototype-to-text alignment, we propose aligning the distribution of order relationships between the visual prototype features and the conditioned text features rather than relying solely on absolute similarity scores. Specifically, we model the matching as a ranking alignment problem using the Plackett-Luce model [30].

For each visual prototype $\boldsymbol{v}_k$, we select the top-N candidate text queries $\mathcal{A}_k = \{\boldsymbol{t}_{k,1}, \ldots, \boldsymbol{t}_{k,N}\}$ based on cosine similarity, where we empirically set $N = 4$. We then define a probability distribution $\mathbf{P}_k^v \in \mathbb{R}^{N!}$ over the space of all possible permutations Π_N. The probability of a specific permutation $\pi \in \Pi_N$ is:

$$[\mathbf{P}_k^v]_\pi = P(\pi|\boldsymbol{v}_k) = \prod_{i=1}^{N} \frac{\exp(\cos(\boldsymbol{v}_k, \boldsymbol{t}_{k,\pi(i)})/\tau_t)}{\sum_{j=i}^{N} \exp(\cos(\boldsymbol{v}_k, \boldsymbol{t}_{k,\pi(j)})/\tau_t)}, \tag{3}$$

Similarly, for each conditioned text feature $\boldsymbol{t}_{k,c}$ associated with the prototype, we compute a similar probability distribution $\mathbf{P}_{k,c}^t \in \mathbb{R}^{N!}$, but this time based on its self-similarity within the text modality:

$$[\mathbf{P}_{k,c}^t]_\pi = P(\pi|\boldsymbol{t}_{k,c}) = \prod_{i=1}^{N} \frac{\exp(\cos(\boldsymbol{t}_{k,c}, \boldsymbol{t}_{k,\pi(i)})/\tau_t)}{\sum_{j=i}^{N} \exp(\cos(\boldsymbol{t}_{k,c}, \boldsymbol{t}_{k,\pi(j)})/\tau_t)}. \tag{4}$$

The final alignment score $S(k,c)$ is computed by measuring the consistency between the visual-anchored and text-anchored distributions:

$$S(k,c) = \langle \mathbf{P}_k^v, \mathbf{P}_{k,c}^t \rangle = \sum_{\pi\in\Pi_N} P(\pi|\boldsymbol{v}_k) \cdot P(\pi|\boldsymbol{t}_{k,c}). \tag{5}$$

Based on this score, each visual prototype is assigned to the text query with the maximum alignment score. By maximizing this structural consistency, our method robustly disambiguates fine-grained categories by verifying whether the neighborhood structure of the visual features matches the semantic neighborhood structure of the text features.

4.4 Hybrid Prototype Injection

To inject visual priors into the text prompt stream, we propose a hybrid prototype injection mechanism. This mechanism combines dynamic image-specific prototypes with static global prototypes, creating a multi-modal reference set that bridges the gap between semantic queries and visual features.

For a target text query c, we identify corresponding prototypes $\mathcal{P}_{dyn}^{(c)}$, capturing instance-specific features. Only prototypes whose final alignment score $S(k, c)$ exceeds a threshold τ are selected for injection, ensuring that unreliable matches are filtered out. To enhance robustness, we retrieve a static prototype $\boldsymbol{m}_c$ from a Global Memory Bank, updated online with a momentum mechanism:

$$\boldsymbol{m}_c^{(t)} = \eta \cdot \boldsymbol{m}_c^{(t-1)} + (1 - \eta) \cdot \bar{\mathcal{P}}_{dyn}^{(c)}, \tag{6}$$

where η is the momentum coefficient, empirically set to 0.9. The dynamic and static prototypes are then concatenated with the text token embeddings:

$$\mathbf{T}_{hybrid} = \text{Concat}(\mathbf{T}_c, \mathcal{P}_{dyn}^{(c)}, \boldsymbol{m}_c) \tag{7}$$

The hybrid prompt $\mathbf{T}_{hybrid}$ is fed into SAM 3's fusion encoder, enabling adaptive attention for improved local and class-level matching, enhancing visual-text alignment and segmentation accuracy.

4.5 Iterative Information-Gain Visual Prompting

While hybrid prototype injection aligns semantics, accurate boundary delineation requires spatial location guidance. We address this by constructing candidate pools for positive and negative prompts, followed by an entropy-based optimization loop that iteratively selects informative points to reduce segmentation uncertainty.

Candidate Pool Construction For each target query c, we begin by constructing two candidate pools: positive candidates $\mathcal{X}_{pos}$ and negative candidates $\mathcal{X}_{neg}$. Positive candidates are selected from the assigned visual clusters $M_k \in \mathcal{M}$, while negative candidates are selected from other clusters. To ensure representativeness, we calculate a prototype similarity score $S_{proto}(x)$ for each candidate point $x \in \mathcal{X}_{pos} \cup \mathcal{X}_{neg}$:

$$S_{proto}(x) = \frac{\boldsymbol{f}_x^\top \boldsymbol{v}_k}{\|\boldsymbol{f}_x\| \|\boldsymbol{v}_k\|}, \quad x \in \mathcal{X}_{pos} \cup \mathcal{X}_{neg}, \tag{8}$$

where $\boldsymbol{f}_x$ is the feature of point x, and $\boldsymbol{v}_k$ is its corresponding cluster prototype. This score prioritizes points that best represent their respective clusters.

Entropy-based Optimization Loop The iterative process begins by initializing the segmentation with the enhanced text prompt from Sec. 4.4, which generates the initial probability map P_t. The entropy $\mathcal{U}_{prev}$ is then computed from P_t, reflecting the model's uncertainty:

$$\mathcal{U}_{prev} = -\frac{1}{HW}\sum_{i}\left[P_t^{(i)}\log(P_t^{(i)}) + (1-P_t^{(i)})\log(1-P_t^{(i)})\right] \tag{9}$$

At each iteration t, we select the best positive candidate x^*_{pos} from $\mathcal{X}_{pos}$ and the best negative candidate x^*_{neg} from $\mathcal{X}_{neg}$ based on their prototype similarity S_{proto}, then simulate their effect by computing the new probability maps, P^{pos}_{new} and P^{neg}_{new}. The information gain for each candidate is calculated as the reduction in global entropy:

$$\Delta\mathcal{U} = \mathcal{U}_{prev} - \mathcal{U}_{new}, \tag{10}$$

where $\mathcal{U}_{new}$ corresponds to the updated entropy after adding the selected candidate, and $\mathcal{U}_{prev}$ is the entropy from the previous iteration.

Each iteration compares potential gains $\Delta\mathcal{U}_{pos}$ and $\Delta\mathcal{U}_{neg}$. The candidate with the maximum entropy reduction is added to the prompt set. After updating the prompt set, we update $\mathcal{U}_{prev} = \mathcal{U}_{new}$ and repeat until the gain drops below a threshold τ_e. This dynamic switching between positive and negative prompts refines boundaries and clarifies ambiguous regions.

5 Experiments

5.1 Experimental Details

For all experiments, we employ the official SAM 3 model with a Perception Encoder-Large+ (PE-L+) backbone [6] and DINOV2 ViT-L/14 [29] as the VFM. Input images are resized to 1008×1008, using class names as text prompts without test-time augmentation or post-processing. Performance is evaluated via mIoU. All experiments run on the MMSegmentation framework using a single NVIDIA RTX 4090 (24GB). More implementation details are provided in **Supplementary Materials**.

5.2 Comparison to the State-of-the-Arts

We compare SANO3 with representative state-of-the-art training-free OVSS methods from two categories: CLIP-based approaches and SAM 3-based approaches. The CLIP-based methods include GEM [7], ProxyCLIP [23], FSA [12], SCLIP [36], ClearCLIP [22], CorrCLIP [40], and SegEarth-OV [27], while the SAM 3-based methods include SAM 3 [9] and SegEarth-OV3 [28].

Quantitative results Tab. 2 presents quantitative comparisons with state-of-the-art methods on eight fine-grained remote sensing datasets. Our method achieves the best or second-best results on most datasets and obtains the highest average mIoU of 34.55%, clearly outperforming previous SOTA approaches. Compared with SAM 3-based methods, SANO3 significantly enhances the fine-grained semantic discrimination capability of the original SAM 3. Notably, substantial gains are observed on challenging datasets such as YRCC and Barley, where our method improves over SAM 3 by 5.12% and 8.23%, respectively.

Moreover, SANO3 maintains consistently strong performance across all eight application scenarios. In contrast, several competing methods exhibit clear performance drops in specific domains. For example, CorrCLIP achieves only 19.24% on Cloud, and SegEarth-OV reaches 9.06% on Barley. The stable performance of SANO3 across heterogeneous datasets highlights its robustness and strong cross-domain generalization ability.

Table 2: Performance comparison on FG-OVSSRS Benchmark. **Best** and 2nd-best results are highlighted.

Methods	Backbone	Forest	Cloud	Rescue	GIS	YRCC	SkyScapes	RSMI	Barley	Avg.
GEM$_{CVPR\ 24}$	CLIP	12.97	30.25	18.88	31.57	26.56	9.70	7.24	9.06	18.28
ProxyCLIP$_{ECCV\ 24}$	CLIP+DINO	27.37	32.51	33.22	32.82	39.68	15.43	10.61	8.06	24.96
FSA$_{ICCV\ 25}$	CLIP+DINO	27.73	32.70	33.40	33.04	43.23	15.41	10.32	8.03	25.48
SCLIP$_{ECCV\ 24}$	CLIP	9.22	28.10	14.11	30.02	32.02	6.40	12.22	11.05	17.89
ClearCLIP$_{ECCV\ 24}$	CLIP	13.89	**32.85**	22.31	31.46	26.96	13.23	14.68	9.38	20.60
CorrCLIP$_{ICCV\ 25}$	CLIP+SAM2+DINO	29.33	19.24	33.89	31.81	50.57	15.46	12.10	10.51	25.36
SegEarth-OV$_{CVPR\ 25}$	CLIP	22.91	24.41	26.83	33.78	56.10	14.79	17.95	9.06	25.73
SAM 3$_{ICLR\ 26}$	SAM3	33.10	29.56	35.76	41.46	52.66	**26.84**	29.17	10.55	32.39
SegEarth-OV3$_{Arvix\ 25}$	SAM3	33.05	28.80	35.85	41.46	52.24	26.77	**29.17**	10.53	32.23
Ours	SAM3+DINOv2	**33.90**	30.70	**36.84**	**43.06**	**57.78**	26.83	28.50	**18.78**	**34.55**

Qualitative results Fig. 4 presents qualitative comparisons with previous SOTA methods on the Forest [4], YRCC [42], and GIS [18] datasets. Compared with existing approaches, SANO3 produces clearer boundaries and more accurate recognition of fine-grained categories, demonstrating the effectiveness of our visual-linguistic alignment and spatial prompting strategy.

5.3 Ablation Studies

The Effectiveness of Different Components SANO3 improves the fine-grained segmentation performance of SAM 3 by jointly enhancing textual and spatial guidance. As shown in Tab. 3, we evaluate three components: dynamic prototype injection (DPI), which injects image-specific prototypes into the text stream; static prototype injection (SPI), which incorporates global bank prototypes into textual prompts; and iterative information-gain visual prompting (IGVP), which adds uncertainty-driven point prompts for spatial refinement.

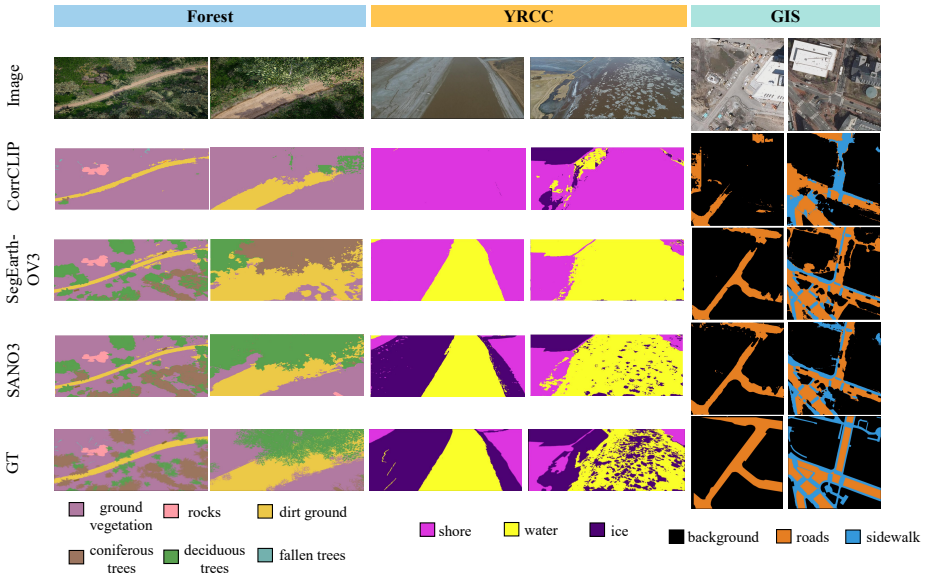

Fig. 4: Qualitative comparison with previous SOTA methods on Forest [4], YRCC [42], and GIS [18] datasets.

DPI consistently improves performance, SPI brings further gains, and IGVP yields the best overall results. The full model achieves the strongest performance across all datasets, demonstrating the effectiveness of collaborative textual and spatial enhancement.

Ablation Study of Hyperparameters In SANO3, the main hyperparameters include the clustering granularity C, the prototype injection threshold τ, and the point prompt information gain threshold τ_e. We analyze their impact on segmentation performance, as summarized in Tab. 4.

For C, smaller values lead to coarse primitives, while larger values may cause fragmentation. Performance is optimal when C is between 6 and 8, balancing discrimination and stability. The hierarchical merging mechanism mitigates over-clustering, improving robustness for larger C values.

For τ and τ_e, both thresholds control the quality of injected information in text and visual guidance. A low τ may introduce unreliable prototype matches, while an overly strict value can discard useful complementary cues; empirically, values between 0.4 and 0.8 achieve stable performance. Similarly, τ_e is set relatively strictly to ensure high-quality point selection for boundary refinement. Although the optimal τ_e varies slightly across datasets due to differences in boundary complexity, its sensitivity is low and typically fluctuates within a narrow range.

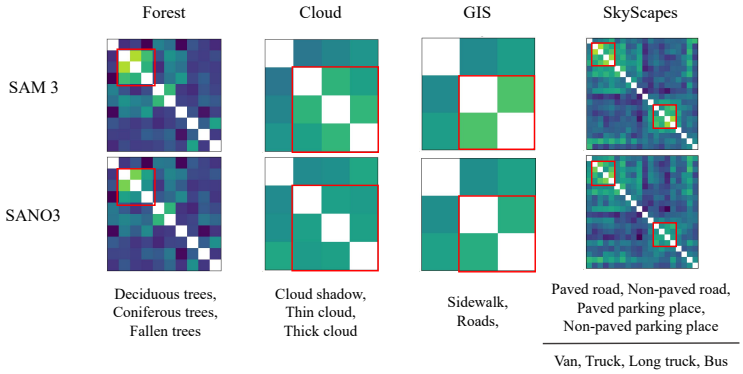

Fig. 5: Cosine similarity heatmaps of text features in Forest [4], Cloud [14], GIS [18], and SkyScapes [2]. The fine-grained category texts are listed below each heatmap. Red boxes highlight regions where the original SAM 3 encoder exhibits feature collapse, which is alleviated by visual-conditioned text refinement, improving the discriminability of text features.

Table 3: Detailed ablation results for each component.

DPI	SPI	IGVP	Cloud	YRCC	Rescue
			29.56	52.66	35.76
✓			29.68	55.25	36.67
✓	✓		30.11	56.46	36.82
✓	✓	✓	**30.70**	**57.78**	**36.84**

Table 4: Ablation study of hyperparameters on the Cloud dataset.

C	mIoU	τ	mIoU	τ_e	mIoU
2	30.15	0.2	29.99	0.05	30.60
4	30.39	0.4	30.11	0.10	30.47
6	30.47	0.6	30.28	0.15	**30.70**
8	**30.70**	0.8	**30.70**	0.20	30.62

Ablation Study of VFM To evaluate SANO3's sensitivity to the choice of VFM, we test different backbones, including MAE [16], DINOv2 [29], and DINOv3 [35]. Although each VFM provides varying levels of visual self-similarity priors and local semantic aggregation capability due to differences in pretraining, all can guide SAM 3 to improve segmentation, and SANO3 consistently achieves competitive results across backbones, demonstrating its robustness to the choice of feature extractor.

Ablation Study of Prototype Alignment Method We further compare the proposed visual-conditioned relation-aware alignment (VARA) with a max cosine similarity baseline that assigns prototypes based on the highest cosine similarity. As shown in Tab. 6, max cosine similarity relies on absolute similarity scores and performs independent matching. However, due to feature collapse in the native SAM 3's text encoder (Fig. 5), fine-grained categories often exhibit high inter-class similarity, making such naive matching unreliable. In contrast, VARA incorporates visual-conditioned text refinement and relation-aware rank-

Table 5: Ablation study on different VFM.

	Cloud	YRCC	Rescue
MAE	30.44	58.64	36.70
DINOv2	**30.70**	57.78	**36.84**
DINOv3	30.26	**58.84**	36.66

Table 6: Ablation study on different prototype matching method.

	Cloud	YRCC	Rescue
Max Cosine Simi.	30.30	57.11	36.42
VARA	**30.70**	**57.78**	**36.84**

ing, leading to more robust semantic assignment and consistently better performance across datasets.

6 Conclusion

In this work, we propose SANO3, a training-free framework for fine-grained open-vocabulary semantic segmentation in remote sensing. By leveraging the visual self-similarity priors in VFMs, SANO3 extracts discriminative visual primitives and enhances SAM 3 through collaborative refinement of textual and spatial guidance, including relation-aware alignment, hybrid prototype injection, and iterative entropy-driven spatial prompting. Extensive experiments across eight datasets demonstrate state-of-the-art performance. Additionally, we introduce FG-OVSSRS Bench, the first fine-grained remote sensing OVSS benchmark spanning eight diverse scenarios, which we hope will foster systematic evaluation and further research on fine-grained semantic understanding.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62471233 and Grant U23B2006, in part by the Jiangsu Provincial Innovation Support Program under Grant BZ2023046, in part by the Jiangsu Provincial Key Research and Development Program under Grant BE2022065-2, in part by Natural Science Foundation of Jiangsu Province under Grant BK20230440, and in part by the Fundamental Research Funds for the Central Universities under Grant 30925020214.

References

1. Al Shafian, S., Hu, D.: Integrating machine learning and remote sensing in disaster management: A decadal review of post-disaster building damage assessment. Buildings **14**(8), 2344 (2024)
2. Azimi, S.M., Henry, C., Sommer, L., Schumann, A., Vig, E.: Skyscapes fine-grained semantic understanding of aerial scenes. In: Int. Conf. Comput. Vis. pp. 7393–7403 (2019)
3. Barsellotti, L., Amoroso, R., Cornia, M., Baraldi, L., Cucchiara, R.: Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3689–3698 (2024)

4. Blaga, B.C.Z., Nedevschi, S.: Forest inspection dataset for aerial semantic segmentation and depth estimation. arXiv preprint arXiv:2403.06621 (2024)
5. Blushtein-Livnon, R., Rafaeli, O., Ioffe, D., Boger, A., Esquenazi, K.S., Svoray, T.: On the effectiveness of textual prompting with lightweight fine-tuning for sam3 remote sensing segmentation. IEEE Geoscience and Remote Sensing Letters (2026)
6. Bolya, D., Huang, P.Y., Sun, P., Cho, J.H., Madotto, A., Wei, C., Ma, T., Zhi, J., Rajasegaran, J., Bangalath, H., et al.: Perception encoder: The best visual embeddings are not at the output of the network. In: Adv. Neural Inform. Process. Syst. vol. 38, pp. 60884–60937 (2025)
7. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 3828–3837 (2024)
8. Cao, Q., Chen, Y., Ma, C., Yang, X.: Open-vocabulary high-resolution remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing (2025)
9. Carion, N., Gustafson, L., Hu, Y.T., Debnath, S., Hu, R., Suris, D., Ryali, C., Alwala, K.V., Khedr, H., Huang, A., et al.: Sam 3: Segment anything with concepts. arXiv preprint arXiv:2511.16719 (2025)
10. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Int. Conf. Comput. Vis. pp. 9650–9660 (2021)
11. Chen, Z., Xu, T., Pan, Y., Shen, N., Chen, H., Li, J.: Edge feature enhancement for fine-grained segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–13 (2024)
12. Chi, Z., Wu, Y., Gu, L., Liu, H., Wang, Z., Zhang, Y., Wang, Y., Plataniotis, K.: Plug-in feedback self-adaptive attention in clip for training-free open-vocabulary segmentation. In: Int. Conf. Comput. Vis. pp. 22815–22825 (2025)
13. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4113–4123 (2024)
14. Foga, S., Scaramuzza, P.L., Guo, S., Zhu, Z., Dilley Jr, R.D., Beckmann, T., Schmidt, G.L., Dwyer, J.L., Hughes, M.J., Laue, B.: Cloud detection algorithm comparison and validation for operational landsat data products. Remote sensing of environment **194**, 379–390 (2017)
15. Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., Freeman, W.T.: Unsupervised semantic segmentation by distilling feature correspondences. arXiv preprint arXiv:2203.08414 (2022)
16. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16000–16009 (2022)
17. Himeur, Y., Rimal, B., Tiwary, A., Amira, A.: Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. Information Fusion **86**, 44–75 (2022)
18. Ibn Sultan, R., Li, C., Zhu, H., Khanduri, P., Brocanelli, M., Zhu, D.: Geosam: Fine-tuning sam with sparse and dense visual prompting for automated segmentation of mobility infrastructure. arXiv e-prints pp. arXiv–2311 (2023)
19. Kang, D., Cho, M.: In defense of lazy visual grounding for open-vocabulary semantic segmentation. In: Eur. Conf. Comput. Vis. pp. 143–164. Springer (2024)
20. Kim, C., Ju, D., Han, W., Yang, M.H., Hwang, S.J.: Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15033–15042 (2025)

21. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Int. Conf. Comput. Vis. pp. 4015–4026 (2023)
22. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: Eur. Conf. Comput. Vis. pp. 143–160. Springer (2024)
23. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In: Eur. Conf. Comput. Vis. pp. 70–88. Springer (2024)
24. Lee, K., Lee, H., Park, J., Hwang, J.Y.: Fine-grained binary object segmentation in remote sensing imagery via path-selective test-time adaptation. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
25. Li, B., Dong, H., Zhang, D., Zhao, Z., Sun, H., Gao, J.: Exploring efficient open-vocabulary segmentation in the remote sensing. In: AAAI. vol. 40, pp. 5982–5991 (2026)
26. Li, J., Wei, Y., Wei, T., He, W.: A comprehensive deep-learning framework for fine-grained farmland mapping from high-resolution images. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–15 (2024)
27. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10545–10556 (2025)
28. Li, K., Zhang, S., Wang, Y., Deng, Y., Wang, Z., Meng, D., Cao, X.: Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images. arXiv preprint arXiv:2512.08730 (2025)
29. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
30. Plackett, R.L.: The analysis of permutations. Journal of the Royal Statistical Society Series C: Applied Statistics **24**(2), 193–202 (1975)
31. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Int. Conf. Mach. Learn. pp. 8748–8763. PmLR (2021)
32. Rahnemoonfar, M., Chowdhury, T., Murphy, R.: Rescuenet: a high resolution uav semantic segmentation dataset for natural disaster damage assessment. Scientific data **10**(1), 913 (2023)
33. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10684–10695 (2022)
34. Shi, Y., Dong, M., Xu, C.: Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In: Int. Conf. Comput. Vis. pp. 23487–23497 (2025)
35. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
36. Wang, F., Mei, J., Yuille, A.: Sclip: Rethinking self-attention for dense vision-language inference. In: Eur. Conf. Comput. Vis. pp. 315–332. Springer (2024)
37. Wang, L., Li, R., Zhang, C., Fang, S., Duan, C., Meng, X., Atkinson, P.M.: Unetformer: A unet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. ISPRS Journal of Photogrammetry and Remote Sensing **190**, 196–214 (2022)

38. Yang, Y., Deng, J., Li, W., Duan, L.: Resclip: Residual attention for training-free dense vision-language inference. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 29968–29978 (2025)
39. Ye, C., Zhuge, Y., Zhang, P.: Towards open-vocabulary remote sensing image semantic segmentation. In: AAAI. vol. 39, pp. 9436–9444 (2025)
40. Zhang, D., Liu, F., Tang, Q.: Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In: Int. Conf. Comput. Vis. pp. 24677–24687 (2025)
41. Zhang, S., Zou, X., Li, K., Lang, C., Wang, S., Tao, P., Cao, T.: Knowledge transfer and domain adaptation for fine-grained remote sensing image segmentation. In: Int. Conf. Multimedia and Expo. pp. 1–6. IEEE (2025)
42. Zhang, X., Jin, J., Lan, Z., Li, C., Fan, M., Wang, Y., Yu, X., Zhang, Y.: Icenet: A semantic segmentation deep network for river ice by fusing positional and channel-wise attentive features. Remote Sensing **12**(2), 221 (2020)
43. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: Eur. Conf. Comput. Vis. pp. 696–712. Springer (2022)
44. Zou, X., Li, Y., Zhang, S., Li, K., Wang, S., Tao, P., Xing, J., Lang, C.: Dynamic dictionary learning for remote sensing image segmentation. In: Int. Conf. Comput. Vis. pp. 22457–22466 (2025)