NOT IN ISOLATION: LEARNING FROM SCENE DEPENDENCIES FOR OPEN-VOCABULARY REMOTE SENSING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Although existing open-vocabulary semantic segmentation (OVSS) methods have made significant progress, segmenting complex remote sensing images remains challenging. Regions of the same semantic class often vary widely in scale and shape, while different classes may exhibit similar local appearances. These ambiguities can be reduced by using spatial relationships between regions and their surroundings as complementary cues. However, such scene dependencies remain insufficiently modeled in OVSS. Therefore, we propose *SCENE*, a scene dependency learning framework that incorporates scene dependencies into open-vocabulary visual representations through content-adaptive multi-granularity perception at the location level and spatial relation modeling at the region level. Specifically, *Content-Adaptive Spatial Perception* adaptively aggregates information across multiple spatial granularities for each position based on image content; *Scene-Relational Structure Learning* treats category-agnostic regions as entities, propagates region information through multiple types of spatial relations, and maps the resulting relational representations into pixel space to generate dense relational context; *Structure-Guided Fusion Decoder* further uses relational context to guide the refinement of spatial perception representations and dense visual representations and performs progressive decoding with multi-level features to produce segmentation predictions. Extensive experiments on four remote sensing benchmarks demonstrate that *SCENE* effectively improves OVSS performance and generalization across datasets. Code will be available at `https://anonymous.4open.science/r/Anonymous-DD21`.

1 INTRODUCTION

Vision-language pretraining maps visual content and natural language concepts into a shared semantic space, providing a new approach to alleviating the constraints imposed by fixed category sets (Radford et al., 2021). Building on this foundation, OVSS extends visual-text alignment at the image-level to dense prediction, enabling models to recognize classes unseen during training using text prompts (Rao et al., 2022; Dutta et al., 2025; Li et al., 2025; Ye et al., 2025).

Existing methods for OVSS primarily focus on visual–text semantic alignment. LSeg (Li et al., 2022) maps dense visual embeddings and textual labels to a shared semantic space, while DenseCLIP (Rao et al., 2022) extends the CLIP image-text alignment to pixel–text alignment. OVSeg (Liang et al., 2023) and SAN (Xu et al., 2023) further employ category-agnostic regions or mask proposals for open-vocabulary segmentation. These methods have established an open-vocabulary segmentation paradigm centered on visual–text alignment. However, regions belonging to different semantic classes in remote sensing scenes may exhibit similar local appearances, making it difficult for models to produce complete and reliable class responses. As shown in Fig. 1(a), in Case I, only portions of adjacent semantic regions with similar appearances are recognized; in Case II, class responses are concentrated in local regions, resulting in fragmented predictions. These observations further indicate the need for more comprehensive spatial visual information to support the prediction of semantic classes.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

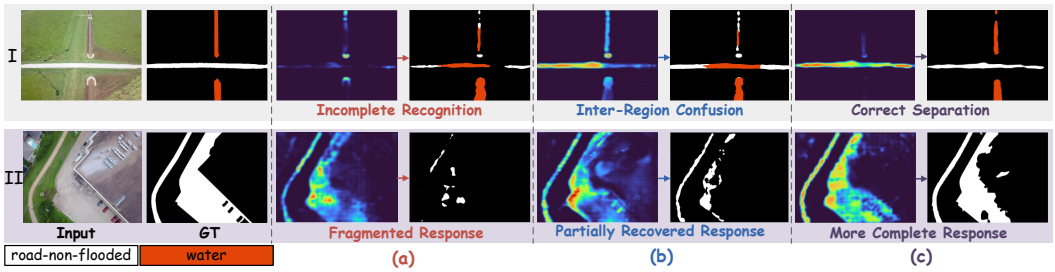

Figure 1: Comparison of class responses and segmentation predictions across visual-text modeling strategies. (a) Visual-Text Alignment. (b) Context-Enhanced Visual-Text Alignment. (c) Ours.

Regarding this demand, related studies have improved dense visual representations through correlation aggregation, spatial information recovery, feature fusion, and contextual interaction. CAT-Seg (Cho et al., 2024) aggregates an image–text cost volume along the spatial and class dimensions, while SED (Xie et al., 2024) supplements spatial information through hierarchical visual features and progressive decoding. For remote sensing imagery, SegEarth-OV (Li et al., 2025) recovers spatial details lost in deep representations through feature upsampling; RSKT-Seg (Li et al., 2026) enhances visual representations using multidirectional cost-map aggregation and remote sensing knowledge transfer; ROSS (Zhang et al., 2026) improves visual–text representations by incorporating rotational variations, remote sensing domain knowledge, and multi-scale feature fusion; ReAttnCLIP (Niu et al., 2026) enhances dense features by reconstructing CLIP attention; and ConInfer (Chen et al., 2026) introduces inter-region semantic dependencies through joint inference across spatial units. These methods expand the spatial information available for visual–text alignment from different perspectives. Specifically, contextual interactions can supplement local class responses by propagating information across spatial locations. However, as shown in Fig. 1(b), although context enhancement recovers more class responses, category confusion persists at the road–water intersection in Case I. In Case II, prediction continuity is improved, yet the target region remains incompletely covered. These results indicate that simply increasing spatial information does not ensure that it provides discriminative evidence suited to the current region; the surrounding visual information and its spatial associations with the current region still require effective modeling. Therefore, we propose SCENE, which jointly models content-adaptive spatial perception and inter-region structural relations. These provide complementary discriminative cues for segmentation. As shown in Fig. 1(c), SCENE produces more complete semantic responses and better distinguishes adjacent regions belonging to different semantic classes.

Specifically, SCENE consists of three main components. CASP adaptively modulates information aggregation across different spatial granularities based on the input visual content, producing location-dependent spatial perception representations. SRSL treats category-agnostic region proposals as entities, constructs a multi-relational graph based on the spatial relations among these entities, updates region representations through relational propagation, and generates dense relational context via region-to-pixel mapping. SGFD leverages dense relational context to refine spatial perception representations and dense visual representations, incorporates text embeddings for fusion, and produces segmentation results via progressive decoding guided by multi-level features. We evaluate the open-vocabulary segmentation performance and cross-dataset generalization of SCENE on four remote sensing benchmarks. Our main contributions are summarized as follows:

- We propose SCENE, a scene dependency learning framework that combines content-adaptive spatial perception with category-agnostic modeling of inter-region structural relations to enrich visual representations for open-vocabulary segmentation.
- We design CASP and SRSL for the adaptive acquisition and relational encoding of spatial visual evidence. CASP dynamically aggregates information at multiple spatial granularities for each location, while SRSL treats category-agnostic regions as entities, encodes inter-region spatial relations through multi-relational propagation, and maps the resulting relational representations into pixel space.

- We propose SGFD to integrate scene relational structure, spatial perception, and dense visual representations for class-conditioned decoding, reducing class confusion and fragmented predictions.

2 Related work

2.1 Open-vocabulary semantic segmentation

OVSS leverages the semantic transferability of vision–language models to recognize unseen classes during training using text prompts. To bridge open-vocabulary semantics with pixel-level prediction, CAT-Seg (Cho et al., 2024) constructs a vision–text cost volume and aggregates matching information along the spatial and class dimensions; SED (Xie et al., 2024) integrates cost maps with multi-level visual features through hierarchical visual encoding and progressive decoding; and DPSeg (Zhao et al., 2025) employs dual visual and textual prompts together with shallow visual features to improve dense cross-modal matching. Recent methods further leverage visual foundation models: ProxyCLIP (Lan et al., 2024) reconstructs CLIP attention using spatial feature correspondences, while ESC-Net (Lee et al., 2025) generates pseudo-prompts from vision–text correlations to guide the SAM decoder. Substantial variations in scale, orientation, and morphology in remote sensing imagery pose additional challenges to domain adaptation and preservation of spatial-details. GSNet (Ye et al., 2025) fuses features from remote-sensing-specific and general-purpose vision-language models through text-guided interaction; SegEarth-OV (Li et al., 2025) employs feature upsampling and bias suppression to restore spatial responses in CLIP features; ROSS (Zhang et al., 2026) combines rotation-invariant modeling with domain adaptation; and ReAttnCLIP (Niu et al., 2026) reconstructs CLIP attention to refine dense visual representations. In general, existing methods for OVSS focus primarily on domain knowledge transfer and spatial adaptation of vision-language features.

2.2 Spatial context modeling

Existing spatial context methods typically expand the visual information available for prediction through local-global feature interactions. Some methods (Fu et al., 2025; Ding et al., 2026) combine local attention with state-space models to capture global context while preserving local details. To address background interference, intra-class variation, and changes in scale and orientation in remote sensing imagery, LOGCAN++ (Ma et al., 2025b) uses local class representations as an intermediary between pixels and global context, and employs affine transformations to adaptively derive these representations. Other methods (Ma et al., 2025a; Ding et al., 2025) supplement contextual information through local–global semantic constraints or long-range dependency modeling, mitigating background interference and facilitating information use in broader spatial ranges. Beyond local–global information aggregation, contextual representations can also be affected by inconsistent feature responses (Wang et al., 2025; Kim et al., 2025) and fragmentation introduced by window-based processing (Lee et al., 2026; Moon et al., 2026). To address inconsistent intra-object responses, DeCLIP (Wang et al., 2025) decouples self-attention into content and context features and employs a visual foundation model to guide the context features in preserving spatial correlations. To address context fragmentation caused by sliding-window processing, GLA-CLIP (Lee et al., 2026) extends the keys and values of attention across all windows, allowing queries to aggregate information from other windows. In general, spatial context modeling is evolving from expanding information coverage toward preserving the consistency and integrity of context propagation.

3 Method

3.1 Overall architecture of SCENE

The overall framework of SCENE is shown in Fig. 2. A frozen RemoteCLIP encoder extracts remote sensing visual features, while a CLIP image encoder with trainable attention (**Appendix A.8**) layers generates dense visual features. A frozen SAM3 model and a frozen CLIP text encoder provide category-agnostic region proposals and class embeddings, respectively. Subsequently, CASP adaptively aggregates information at different spatial granularities, while SRSL models inter-region spatial relations based on the region proposals and generates dense relational context. Finally, SGFD

uses the relational context to refine visual representations and incorporates class embeddings and multi-level features to generate segmentation results.

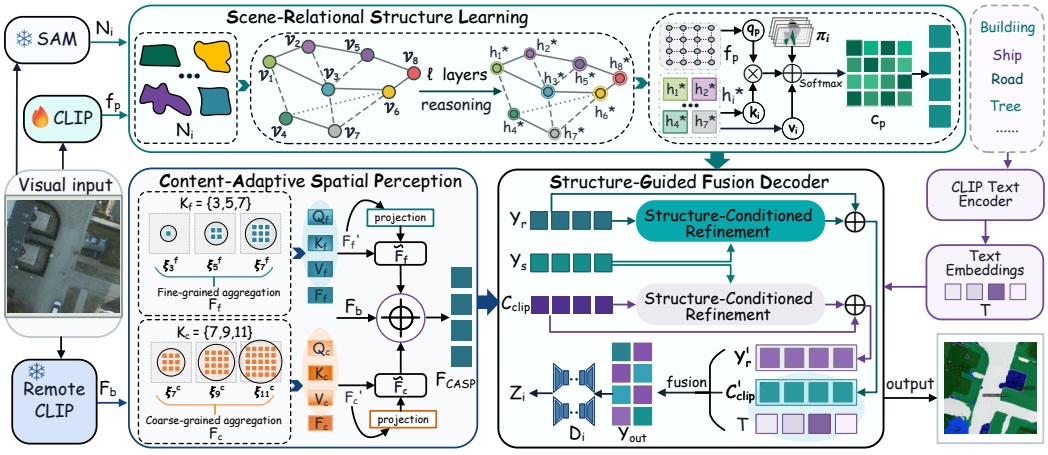

Figure 2: The overall architecture of SCENE.

3.2 SCENE-RELATIONAL STRUCTURE LEARNING

Inter-region spatial relations may provide additional context for inferring region semantics in complex scenes, particularly when local appearances are ambiguous. Accordingly, SRSL organizes category-agnostic region entities into a multi-relational graph. It performs dynamic multi-relational reasoning by weighting relation-specific messages according to current node states and edge attributes, encoding the scene relational structure.

Stage-1: Category-agnostic region entity construction. Given an image $I \in \mathbb{R}^{H\times W\times C}$, we adopt base prompts P_{base} (details in **Appendix A.1.1**) and geometric prompts P_{geo} to drive SAM3 (Li et al., 2025) to generate region candidate proposals:

$$\mathcal{R} = \text{Filter}\big(\text{SAM}(I, P_{\text{base}}) \cup \text{SAM}(I, P_{\text{geo}})\big) = \{\mathcal{R}_i = (M_i, l_i, B_i, s_i)\}_{i=1}^{n}, \tag{1}$$

where M_i and l_i denote the binary mask and continuous mask logits of the i-th candidate region, B_i represents the bounding box, and s_i is the proposal confidence. To jointly encode regional appearance and spatial attributes, we first mask the background within B_i using M_i and extract the primary CLIP regional feature $a_i = \text{Norm}\big(E_v\big(\text{Resize}(I_{B_i} \odot M_{i,M_i})\big)\big)$. It is then fused with geometric spatial attributes $g_i = [x_i, y_i, A_i, b_i, o_i]$ via $N_i = \text{MLP}_{fuse}([a_i, g_i])$. Here E_v denotes the CLIP image encoder. (x_i, y_i), A_i, b_i, and o_i represent the normalized mask center, mask area ratio, bounding-box aspect ratio, and principal orientation estimated from mask covariance, respectively.

Stage-2: Dynamic multi-relational reasoning. We construct a directed relational graph $\mathcal{G} = (\mathcal{H}, \mathcal{E})$ based on region entities, where $\mathcal{H} = \{h_i\}_{i=1}^{n}$ denotes the set of nodes mapped from region entities, and $\mathcal{E}$ represents relational connections between nodes. Following classical qualitative spatial representations (Cohn & Renz, 2008; Frank, 1996), two fundamental types of relations are defined to describe regional organization: topological relations and directional relations. Topological relations: $r_{ij}^{topo} \subseteq \{\text{inside}, \text{contains}, \text{overlap}, \text{adjacency}\}$. Directional relations: $r_{ij}^{dir} \subseteq \{\text{horizontal}(\text{left}, \text{right}), \text{vertical}(\text{above}, \text{below})\}$. To reduce the number of parameters, a relation-group parameter-sharing strategy is adopted, in which transformation parameters are shared among spatial relations with similar geometric semantics. The 8 types of primitive relations are mapped into 5 relation groups $g(r) \in \{\text{containment}, \text{horizontal}, \text{vertical}, \text{overlap}, \text{adjacent}\}$. For relation r from node i to node j at layer ℓ (ℓ analysis is reported in **Appendix A.3.1**), the joint relation representation and relation-aware weight are formulated as:

$$\bar{m}_{i,j}^{r,\ell} = W_{g(r),\ell}^{src} h_i^{\ell} + W_{g(r),\ell}^{tgt} h_j^{\ell} + e_r, \tag{2}$$

$$\alpha_{i,j}^{r,\ell} = \sigma\big(MLP_{att}^{\ell}\big([h_i^{\ell}; h_j^{\ell}; e_r; \tau_{ij}; c_{ij}; c_i; c_j]\big)\big), \tag{3}$$

where h_i^ℓ and h_j^ℓ denote the source and target node states at layer ℓ, respectively. $g(r)$ indicates the relation group to which relation r belongs. $W_{g(r),\ell}^{src}$ and $W_{g(r),\ell}^{tgt}$ are the source and target node projections, respectively. e_r is the relation embedding. $\sigma(\cdot)$ denotes the sigmoid function. τ_{ij} represents the spatial geometric attributes between the two nodes. c_{ij} is the edge relation confidence derived from the corresponding spatial criteria. Ablations on r_{ij}^{topo}, r_{ij}^{dir}, and $\alpha_{i,j}^{r,\ell}$ are reported in **Appendix A.6**. Subsequently, incoming messages to node j are weighted by their relation-aware coefficients and aggregated, and the node state is updated through a residual connection:

$$h_j^{\ell+1} = LN\left(h_j^\ell + MLP_{upd}^\ell\left[h_j^\ell, \frac{\sum_{(i,r)\to j}\alpha_{i,j}^{r,\ell}\bar{m}_{i,j}^{r,\ell}}{\max\left(\sum_{(i,r)\to j}\alpha_{i,j}^{r,\ell}, 1\right)}\right]\right), \tag{4}$$

Stage-3: Mask-guided region-to-pixel transfer. To transfer the relation-enhanced node representations h_i^* into dense space, the model maps pixel features f_p from the main CLIP into Query $q_p = W_Q f_p$, and maps the node h_i^* into Key $k_i = W_K h_i^*$ and Value $v_i = W_V h_i^*$. Next, the continuous mask logits and proposal confidence scores from SAM3 are leveraged to construct the soft spatial prior: $\pi_i(p) = s_i \cdot \text{AvgPool}\big(\sigma(\text{Resize}(l_i))\big)$. Subsequently, mask-prior-guided region-to-pixel cross-attention is performed to obtain pixel-level relational context c_p:

$$c_p = \sum_i \text{Softmax}\left(\frac{q_p^T k_i}{\sqrt{d}} + \log[\pi_i(p) + \epsilon]\right) v_i, \tag{5}$$

3.3 Content-Adaptive Spatial Perception

Semantic regions in remote sensing scenes vary in scale and shape, and the visual evidence required at different positions may be distributed across different spatial ranges. Therefore, CASP estimates the contribution of each candidate granularity based on the visual content at each position and uses the resulting weights to aggregate information from multiple spatial ranges into fine- and coarse-grained representations. The base remote sensing visual features further guide the fusion of the two representations into spatial perception representations.

Stage-1: Adaptive granularity aggregation. Let $F_b \in \mathbb{R}^{B\times 32\times 24\times 24}$ denote the base visual representation obtained by channel projection, spatial alignment, and convolutional fusion of two intermediate RemoteCLIP (Liu et al., 2024) features. Two candidate mode sets for fine and coarse perception, $\mathcal{K}_f = \{3, 5, 7\}$ and $\mathcal{K}_c = \{7, 9, 11\}$, are defined to aggregate spatial information over varying receptive ranges. For each perception mode $m \in \{f, c\}$ (ablations are reported in **Appendix A.7**), an independent position-wise linear mapping predicts the granularity weights:

$$\xi_k^m(p) = \frac{\exp\left((W_m^r F_b(p) + b_m^r)_k\right)}{\sum_{j\in\mathcal{K}_m}\exp\left((W_m^r F_b(p) + b_m^r)_j\right)}, \qquad k \in \mathcal{K}_m, \tag{6}$$

where p denotes a spatial location, and k denotes a candidate spatial perception granularity within mode m, corresponding to a $k\times k$ spatial aggregation window. For any position p, $\sum_{k\in\mathcal{K}_f}\xi_k^f(p) = 1$ and $\sum_{k\in\mathcal{K}_c}\xi_k^c(p) = 1$. The granularity weights $\xi_k^m(p)$ vary with the input image and spatial location but are shared across channels ($\xi_k^m(p)$ ablations are reported in **Appendix A.7**). Subsequently, visual information at different spatial granularities is aggregated using the granularity weights.

$$F_m(p) = \sum_{k\in\mathcal{K}_m}\xi_k^m(p)\cdot\frac{1}{|\Omega_k(p)|}\sum_{q\in\Omega_k(p)}F_b(q), \qquad m \in \{f, c\}, \tag{7}$$

where $F_f(p)$ denotes the fine-grained spatial perception representation, which focuses on local structural variations and boundary information. $F_c(p)$ denotes the coarse-grained spatial perception representation, which encodes regional structure and spatial layout over a larger spatial aggregation range. $\Omega_k(p)$ denotes the valid window neighborhood centered at p. Window averaging is implemented using stride-1 pooling while preserving the spatial resolution. Consequently, each location obtains content-adaptive fine- and coarse-grained representations, allowing the spatial aggregation range to vary dynamically with the visual content at that location.

Stage-2: Guided spatial perception. The perception representation at granularity m, F_m, generates the query $Q_m = W_q^m(\text{LN}(F_m))$, which characterizes the information required at

the current location. F_m and the base visual feature F_b jointly generate the key $K_m = W_k^m(S_m^K([\mathrm{LN}(F_m), \mathrm{LN}(F_b)]))$, allowing matching across locations to consider both granularity context and remote sensing visual information. The value $V_m = W_v^m(S_m^V(\mathrm{LN}(F_m)))$ is generated from F_m to convey the representation specific to the corresponding perception mode. Here, S_m^K and S_m^V are the spatial processing operations applied to the key and value, respectively. The fine-grained mode preserves the spatial resolution of the key and value, whereas the coarse-grained mode compresses them using convolutions with kernel size and stride both set to 4, reducing computational overhead. The aggregation result for the r-th attention head is:

$$A_m^r = \mathrm{Softmax}\left(\frac{Q_m^r (K_m^r)^\mathsf{T}}{\sqrt{d_h}}\right) V_m^r, \tag{8}$$

The outputs of all attention heads are concatenated and linearly projected to obtain A_m. The resulting attention feature is added to the representation at the current granularity through a residual connection, and the result is then layer-normalized: $F'_m = \mathrm{LN}\left(F_m + A_m\right)$.

To adjust channel responses and incorporate local spatial information, the normalized features are mapped into two expanded representations, with the activated responses of one modulating the other. Channel projection and depthwise convolution are then applied to perform the residual update:

$$\widetilde{F}_m = F'_m + \mathrm{DWConv}\left(L_3^m\left(\mathrm{ReLU}\left(L_1^m(F'_m)\right) \odot L_2^m(F'_m)\right)\right), \tag{9}$$

where L_1^m, L_2^m, and L_3^m are linear mappings with biases. The first two expand the channel dimension from C to $3C$, while the last restores it to C. Finally, the updated fine- and coarse-grained representations are fused with the base visual feature to obtain the content-conditioned spatial perception representation F_{CASP}:

$$F_{\mathrm{CASP}} = F_b + \rho\left(\mathrm{Conv}_{3\times3}^{C\to C}\left(\rho\left(\mathrm{Conv}_{3\times3}^{3C\to C}\left([F_b, \widetilde{F}_f, \widetilde{F}_c]\right)\right)\right)\right), \tag{10}$$

3.4 Structure-guided fusion decoder

SGFD uses structural relational context to refine spatial perception and dense visual representations for class prediction. Let Y_r, Y_s, and C_{clip} denote the spatial perception representation produced by CASP, the aligned dense relational context produced by SRSL, and the channel-projected CLIP dense visual representation, respectively. To incorporate information about region organization into the dense visual representation, we jointly estimate gating weights from C_{clip} and Y_s and use the structural context to refine the dense visual representation:

$$C'_{\mathrm{clip}} = C_{\mathrm{clip}} + \sigma\left(\mathrm{Conv}_{1\times1}\left(\rho\left(\mathrm{Conv}_{3\times3}\left([C_{\mathrm{clip}}, Y_s]\right)\right)\right)\right) \odot Y_s, \tag{11}$$

For the spatial perception representation, the relational context further provides region-organization guidance for visual responses across different spatial ranges. The refinement is formulated as:

$$Y'_r = Y_r + C_R^{(2)}\left(\rho\left(C_R^{(1)}\left([Y_r, Y_s]\right)\right)\right), \tag{12}$$

where $C_R^{(1)}$ and $C_R^{(2)}$ denote the two convolution operations. Y'_r ablations are reported in **Appendix A.9**. Text semantics are subsequently introduced, and the dense visual representation refined by the relational context is fused with the spatial perception representation to form category-relevant decoding features: $Y_{\mathrm{out}} = \Phi_f\left([C'_{\mathrm{clip}} + T, Y'_r]\right)$. Here, $\Phi_f(\cdot)$ denotes a feature fusion operation consisting of two convolutional layers, and T denotes the text embeddings. The fused representation is subsequently decoded through two progressive stages, each incorporating visual features from the corresponding levels of the main CLIP and RemoteCLIP:

$$Z_i = \mathcal{D}_i\left(\left[Z_{i-1}, G_{\mathrm{clip}}^i, G_{\mathrm{re}}^i\right]\right), \quad Z_0 = Y_{\mathrm{out}}, \quad i \in \{1, 2\}, \tag{13}$$

where $\mathcal{D}_i$ denotes the i-th upsampling decoding operation, while G_{clip}^i and G_{re}^i denote the visual guidance features from the main CLIP and RemoteCLIP, respectively.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Benchmark & Metric. All methods are trained on LandDiscover50K and evaluated on four remote sensing datasets: ISPRS Potsdam, FLAIR, FAST, and FloodNet (**Appendix A.2.1**). Metrics for evaluation include mIoU, mACC, and HM (**Appendix A.2.2**). Note that, unless otherwise specified, the **best** and **second-best** results in all experimental tables are highlighted in red and blue, respectively.

Training Details. All methods use CLIP ViT-B/16 (Dosovitskiy et al., 2020) and CLIP ViT-L/14 (Radford et al., 2021) as vision–language backbones and are trained for 20,000 iterations on two NVIDIA RTX 4500 GPUs, with a batch size of 4 and a learning rate of 2e-4. The remaining parameters of the compared methods follow the default settings in their respective public implementations.

Table 1: Comparison with state-of-the-art OVSS methods on four datasets.

Method	FLAIR		FAST		Potsdam		FloodNet		Avg	
	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
CLIP-ViT-B as the backbone										
CAT-Seg (Cho et al., 2024)	18.25	35.59	14.08	24.35	33.51	58.38	30.55	36.80	24.10	38.78
EBSeg (Shan et al., 2024)	14.29	30.51	11.10	19.18	19.89	38.40	9.03	9.77	13.58	24.47
OVRS (Cao et al., 2024)	16.59	34.83	15.33	**25.49**	36.68	**63.04**	31.17	38.57	24.94	40.48
TPOVSeg (Zhang et al., 2025)	17.62	34.74	14.40	23.16	31.45	56.60	23.26	27.38	21.68	35.47
GSNet (Ye et al., 2025)	**19.69**	**38.92**	15.44	24.91	**38.17**	58.74	**33.90**	**41.76**	**26.80**	**41.08**
RSKT-Seg (Li et al., 2026)	18.49	34.21	**16.62**	25.30	26.86	52.59	20.47	24.11	20.61	34.05
PCA-Seg (Yin et al., 2026)	17.04	33.44	14.80	24.25	32.64	59.98	31.13	38.12	23.90	38.95
DINODe (Yoon et al., 2026)	15.92	33.42	11.30	21.65	21.37	49.55	11.00	38.77	14.90	35.85
Ours	**18.52**	**35.69**	**17.68**	**26.71**	**41.97**	**66.16**	**34.91**	**44.22**	**28.27**	**43.20**
CLIP-ViT-L as the backbone										
CAT-Seg (Cho et al., 2024)	17.12	30.19	16.99	26.24	26.67	52.34	33.92	41.60	23.68	37.59
EBSeg (Shan et al., 2024)	15.04	31.60	14.70	23.83	10.49	34.70	11.97	13.11	13.05	25.81
OVRS (Cao et al., 2024)	18.94	34.20	17.00	26.40	26.42	52.48	35.09	47.67	24.36	40.19
TPOVSeg (Zhang et al., 2025)	16.05	34.91	16.02	25.56	28.49	57.39	23.48	37.06	21.01	38.73
GSNet (Ye et al., 2025)	**19.54**	33.20	**18.01**	27.15	27.53	52.36	34.94	40.45	**25.01**	38.29
RSKT-Seg (Li et al., 2026)	**20.81**	**38.93**	13.75	23.23	**29.49**	**57.94**	**35.99**	**47.93**	**25.01**	**42.01**
PCA-Seg (Yin et al., 2026)	16.58	32.78	17.74	**28.63**	26.03	52.39	21.71	33.79	20.52	36.89
DINODe (Yoon et al., 2026)	15.42	28.63	10.77	20.45	17.09	41.43	12.35	40.98	13.91	32.87
Ours	19.32	**40.07**	**18.84**	**28.28**	**33.19**	**62.28**	**36.47**	**48.45**	**26.96**	**44.77**

Table 2: Per-class IoU and HM (%) on the FLAIR and Potsdam datasets. Green and purple headers denote **seen** classes and **unseen** classes, respectively.

Method	FLAIR													Potsdam					
	Bui.	Agr.	Wat.	Bar.	Per.	Imp.	Con.	Dec.	Bru.	Vin.	Her.	Plo.	HM	Bui.	Tre.	Car.	Low.	Imp.	HM
CAT-Seg	45.47	29.58	51.67	15.86	0.44	18.44	8.68	20.43	9.90	10.87	3.55	4.22	15.08	54.43	25.79	20.39	25.64	41.28	33.50
EBSeg	30.73	19.89	38.61	0.80	10.53	5.63	8.84	0.78	0.37	41.55	0.21	13.53	14.02	33.05	0.00	29.23	29.48	7.70	19.62
OVRS	40.03	29.60	48.70	11.56	3.22	7.23	7.15	15.60	9.08	17.56	3.29	6.07	13.66	63.62	33.82	14.80	31.93	39.18	36.46
TPOVSeg	38.13	32.36	53.21	11.89	0.47	14.10	8.73	9.00	13.39	26.72	3.28	0.20	14.82	51.87	26.95	12.00	34.20	32.25	31.68
GSNet	45.34	28.90	47.88	28.14	1.51	22.37	7.32	20.67	9.96	5.24	10.68	8.32	**16.74**	59.93	14.73	37.36	32.26	46.60	**38.36**
RSKT-Seg	44.19	32.17	45.05	12.47	5.00	6.80	10.09	10.24	12.73	35.75	3.00	4.47	16.57	58.18	27.63	13.20	5.76	29.56	23.04
PCA-Seg	42.56	26.31	45.48	13.91	4.55	1.89	9.33	12.50	12.45	27.54	6.18	1.75	14.69	63.75	34.09	8.16	33.27	23.96	31.65
DINODe	47.21	36.10	63.28	17.37	2.44	10.86	9.46	1.04	0.92	0.00	0.00	2.40	6.26	56.98	30.52	5.89	0.00	13.47	11.09
Ours	41.54	29.91	52.25	4.72	4.91	5.75	5.20	26.33	18.83	22.65	3.97	6.35	**17.20**	64.05	28.14	29.85	38.98	48.80	**42.23**

4.2 COMPARISON WITH STATE-OF-THE-ART METHODS

Quantitative Evaluation. With both ViT-B and ViT-L backbones, our method achieves competitive m-mIoU and m-mACC (Table 1). Notably, all methods show lower mIoU on Potsdam when using ViT-L. Since only the backbone is replaced while all other training settings remain unchanged, this shared decline may reflect insufficient adaptation of the fixed training configuration to the larger backbone, together with the data distribution of Potsdam. These results suggest that increasing the backbone size does not necessarily improve cross-dataset performance.

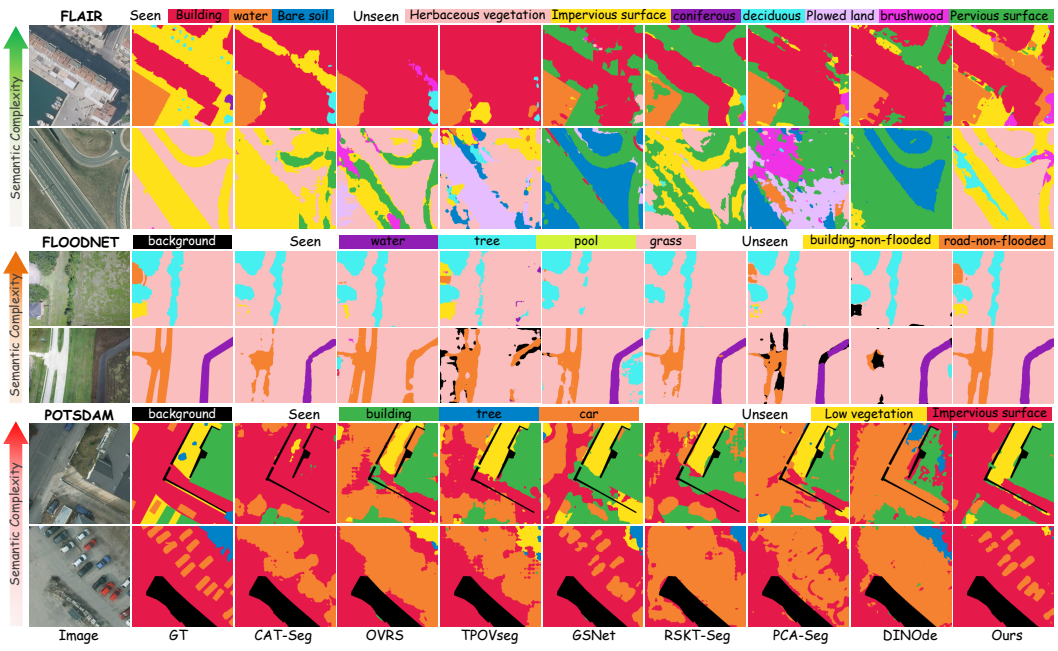

Figure 3: Qualitative comparison on FLAIR, FloodNet, and Potsdam under varying scene complexity. The selected samples cover different scene types and exhibit increasing semantic complexity within each dataset. Seen and unseen classes are marked in the figure.

Table 3: Ablation study of the components in SCENE. For SGFD, "×" denotes removing only the structural refinement paths while retaining the subsequent decoding process.

Components			FLAIR		FAST		Potsdam		FloodNet	
CASP	SRSL	SGFD	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC
×	×	×	15.17	30.99	14.72	22.73	31.81	54.22	29.14	35.13
✓	×	✓	15.55	32.57	15.35	23.90	33.57	59.85	30.90	36.80
✓	✓	×	16.23	32.33	15.56	23.80	34.18	60.84	31.96	39.62
×	✓	✓	16.74	33.50	15.98	23.97	35.95	60.98	32.34	39.84
✓	✓	✓	18.52	35.69	17.68	26.71	41.97	66.16	34.91	44.22

Qualitative Evaluation in Complex Scenes. As shown in Fig. 3, confusion between similar classes and interference between adjacent regions become more pronounced as scene composition becomes more diverse, particularly for unseen classes. In contrast, our method produces more coherent and complete predictions across different datasets and scenes. These results suggest that adaptive spatial perception and relational context modeling help improve spatial representations and class discrimination in complex remote sensing scenes.

Generalization Evaluation. As shown in Table 2, SCENE achieves HM scores of 17.20% and 42.23% on FLAIR and Potsdam, respectively, exceeding GSNet by 0.46 and 3.87 percentage points. These results indicate improved combined performance across seen and unseen classes. Additional results on open-vocabulary generalization are provided in **Appendix A.4**.

4.3 Ablation Studies and Analysis

Ablation Study of Different Components. The full model achieves the highest mIoU and mACC across all four datasets (Table 3). Removing either CASP or SRSL degrades performance, indicating that both content-adaptive spatial perception and spatial relational structure modeling benefit segmentation. Ablating SGFD while retaining both components also reduces performance, further suggesting that refinement conditioned on relational structural context benefits subsequent fusion and decoding.

Table 4: Ablation study of structure-guided refinement in SGFD. "S→C" and "S→CLIP" denote the refinement of CASP spatial perception representations and CLIP dense visual representations using relational structural context, respectively.

Refinement Paths		FLAIR		FAST		Potsdam		FloodNet	
S→C	S→CLIP	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC
×	×	16.23	32.33	15.56	23.80	34.18	60.84	31.96	39.62
×	✓	17.70	34.38	16.12	24.64	35.00	63.11	32.80	41.74
✓	×	17.79	33.74	15.41	23.95	35.26	62.41	32.51	41.41
✓	✓	18.52	35.69	17.68	26.71	41.97	66.16	34.91	44.22

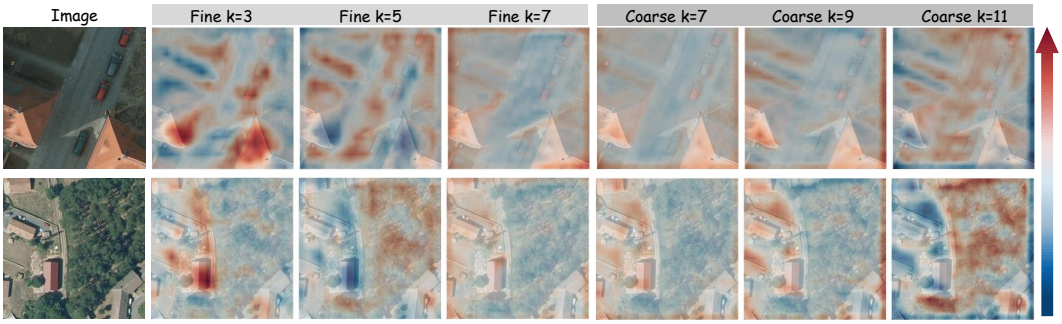

Figure 4: Response visualization of spatial granularity weights in CASP. Each subplot shows the spatial deviation of the corresponding granularity weight at each position relative to the average weight of the current image. Red and blue indicate values higher and lower than the image average, respectively, and darker colors represent larger deviation magnitudes.

Ablation Study of Structure-Guided Refinement. Compared with the setting where both paths are turned off, using either "S→C" or "S→CLIP" alone improves most metrics (Table 4). Using both paths together further increases mIoU and mACC on all four datasets to the best results. This shows the two paths complement each other, and dual-path refinement can better use structural context.

Spatial Granularity Perception Analysis of CASP. As shown in Fig. 4, the granularity weights under the Fine mode mainly vary at object boundaries and regions with complex local structures. In contrast, the Coarse mode exhibits more prominent region-level responses in large continuous areas such as roads. The weight residuals are generally small in regions with uniform textures or consistent semantics, demonstrating that the model maintains a relatively stable scale combination at these locations. This result indicates that CASP can adaptively adjust the relative contributions of local detail perception and long-range context aggregation according to different spatial ranges.

5 Conclusion

In this work, to derive discriminative evidence complementary to local appearance from spatial associations between semantic regions and their surrounding scene, we propose the scene dependency learning framework SCENE. SCENE adaptively aggregates spatial information at multiple granularities and generates dense relational context through multi-relational reasoning among category-agnostic region entities. This context is further used to refine spatial perception representations and dense visual representations, incorporating spatial associations within the scene into open-vocabulary class discrimination. Experiments on four remote sensing benchmarks demonstrate that SCENE improves open-vocabulary segmentation performance and cross-dataset generalization. The limitations and future research directions are discussed in **Appendix A.11**.

AI use statement

In this work, generative AI tools, including Gemini and Doubao, were used to check the grammar of selected sentences, provide guidance on visualizing class response maps, and assist with literature

searches (e.g., searching for recently published papers in the relevant field). These tools were also used to improve manuscript organization and LaTeX formatting (e.g., table construction and style customization) and provide general debugging suggestions (e.g., for environment-related errors and version conflicts among packages). Generative AI was not involved in the design of the core algorithm, the implementation of the proposed method, or the execution of experiments. All AI-assisted content was reviewed, revised, and verified by the authors. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

Ethics statement

This study was conducted in accordance with the ICLR Code of Ethics and Code of Conduct. Our experiments rely solely on publicly released remote sensing benchmarks and involve neither human participants, crowd workers, animal studies, nor the collection or use of sensitive personally identifiable information. Institutional review board approval is therefore not applicable. Each dataset is used within the scope permitted by its original license, and no data subject to redistribution restrictions are republished as part of this work.

Reproducibility statement

To ensure the reproducibility of our results, comprehensive implementation details are provided in the main text and appendix. The overall architecture of SCENE and its constituent modules are detailed in Sec 3. All datasets used in this study are publicly available benchmarks, and more information is provided in Sec 4.1 and Appendix A.2.1. In line with our commitment to open science, the source code and experimental configurations will be made publicly available.

References

Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary remote sensing image semantic segmentation. *arXiv preprint arXiv:2409.07683*, 2024.

Wenyang Chen, Zhanxuan Hu, Yaping Zhang, Hailong Ning, and Yonghang Tai. Coninfer: Context-aware inference for training-free open-vocabulary remote sensing segmentation. *arXiv preprint arXiv:2603.29271*, 2026.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4113–4123. IEEE, 2024.

Anthony G Cohn and Jochen Renz. Qualitative spatial representation and reasoning. *Foundations of artificial intelligence*, 3:551–596, 2008.

Shujie Ding, Xiaoli Ruan, Jing Yang, Chengjiang Li, Jie Sun, Xianghong Tang, and Zhidong Su. Lrdtn: Spectral–spatial convolutional fusion long-range dependence transformer network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–21, 2025.

Shujie Ding, Xiaoli Ruan, Jing Yang, and Zhiqiang Du. Mvch: Multiview collaborative hybrid network for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 64:5528315–5528315, 2026.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Saikat Dutta, Akhil Vasim, Siddhant Gole, Hamid Rezatofighi, and Biplab Banerjee. Aeroseg: Harnessing sam for open-vocabulary segmentation in remote sensing images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 2245–2255. IEEE, 2025.

Andrew U Frank. Qualitative spatial reasoning: Cardinal directions as an example. *International journal of geographical information science*, 10(3):269–290, 1996.

Yunxiang Fu, Meng Lou, and Yizhou Yu. Segman: Omni-scale context modeling with state space models and local attention for semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19077–19087. IEEE, 2025.

Anatol Garioud, Nicolas Gonthier, Loic Landrieu, Apolline De Wit, Marion Valette, Marc Poupée, Sébastien Giordano, et al. Flair: a country-scale land cover semantic segmentation dataset from multi-source optical imagery. *Advances in Neural Information Processing Systems*, 36:16456–16482, 2023.

ISPRS. 2d semantic labeling potsdam dataset. `https://www.isprs.org/education/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx`, 2013. Accessed: 2024-08-11.

Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for object-context aware open-vocabulary semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 15033–15042. IEEE, 2025.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Proxyclip: Proxy attention improves clip for open-vocabulary segmentation. In *European Conference on Computer Vision*, pp. 70–88. Springer, 2024.

ByeongCheol Lee, Hyun Seok Seong, Sangeek Hyun, Gilhan Park, WonJun Moon, and Jae-Pil Heo. Looking beyond the window: Global-local aligned clip for training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2603.23030*, 2026.

Minhyeok Lee, Suhwan Cho, Jungho Lee, Sunghun Yang, Heeseung Choi, Ig-Jae Kim, and Sangyoun Lee. Effective sam combination for open-vocabulary semantic segmentation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26081–26090. IEEE, 2025.

Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Hao Sun, and Junyu Gao. Exploring efficient open-vocabulary segmentation in the remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 5982–5991, 2026.

Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *arXiv preprint arXiv:2201.03546*, 2022.

Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10545–10556. IEEE, 2025.

Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7061–7070. IEEE, 2023.

Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–16, 2024.

Xiaowen Ma, Rongrong Lian, Zhenkai Wu, Renxiang Guan, Tingfeng Hong, Mengjiao Zhao, Mengting Ma, Jiangtao Nie, Zhenhong Du, Siyang Song, et al. A novel scene coupling semantic mask network for remote sensing image segmentation. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:44–63, 2025a.

Xiaowen Ma, Rongrong Lian, Zhenkai Wu, Hongbo Guo, Fan Yang, Mengting Ma, Sensen Wu, Zhenhong Du, Wei Zhang, and Siyang Song. Logcan++: Adaptive local-global class-aware network for semantic segmentation of remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–16, 2025b.

Seungjae Moon, Seunghyun Oh, and Youngmin Ro. Ov-stitcher: A global context-aware framework for training-free open-vocabulary semantic segmentation. *arXiv preprint arXiv:2604.08110*, 2026.

Xin Niu, Manqi Zhao, Dongsheng Jiang, Yingying Wu, and Bing Su. Reattnclip: Training-free open-vocabulary remote sensing image segmentation via re-defined attention in clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24980–24989, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Maryam Rahnemoonfar, Tashnim Chowdhury, Argho Sarkar, Debvrat Varshney, Masoud Yari, and Robin Roberson Murphy. Floodnet: A high resolution aerial imagery dataset for post flood scene understanding. *IEEE Access*, 9:89644–89654, 2021.

Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 18061–18070. IEEE, 2022.

Xiangheng Shan, Dongyue Wu, Guilin Zhu, Yuanjie Shao, Nong Sang, and Changxin Gao. Open-vocabulary semantic segmentation with image embedding balancing. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 28412–28421. IEEE, 2024.

Di Wang, Jing Zhang, Bo Du, Minqiang Xu, Lin Liu, Dacheng Tao, and Liangpei Zhang. Samrs: Scaling-up remote sensing segmentation dataset with segment anything model. *Advances in Neural Information Processing Systems*, 36:8815–8827, 2023.

Junjie Wang, Bin Chen, Yulin Li, Bin Kang, Yichi Chen, and Zhuotao Tian. Declip: Decoupled learning for open-vocabulary dense perception. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14824–14834. IEEE, 2025.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3426–3436. IEEE, 2024.

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: Side adapter network for open-vocabulary semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):15546–15561, 2023.

Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 9436–9444, 2025.

Jianjian Yin, Tao Chen, Yi Chen, Gensheng Pei, Xiangbo Shu, Yazhou Yao, and Fumin Shen. Pcaseg: Revisiting cost aggregation for open-vocabulary semantic and part segmentation. *arXiv preprint arXiv:2603.17520*, 2026.

Sung-Hoon Yoon, Hoyong Kwon, Changgyoon Oh, and Kuk-Jin Yoon. Dinode: Continuous vision-text alignment for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2607.21371*, 2026.

Da Zhang, Mingmin Zeng, and Xuelong Li. Towards open-vocabulary semantic segmentation for remote sensing images. *Pattern Recognition*, pp. 113120, 2026.

Xiaokang Zhang, Chufeng Zhou, Jianzhong Huang, and Lefei Zhang. Tpov-seg: Textually enhanced prompt tuning of vision-language models for open-vocabulary remote sensing semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–17, 2025.

Ziyu Zhao, Xiaoguang Li, Lingjia Shi, Nasrin Imanpour, and Song Wang. Dpseg: Dual-prompt cost volume learning for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25346–25356, 2025.

A APPENDIX

A.1 ADDITIONAL IMPLEMENTATION DETAILS OF SRSL

A.1.1 REGION PROPOSAL GENERATION

To construct class-agnostic region entities, we use SAM3 to generate base and geometry-guided proposals (ablation studies in **Appendix A.10**). Both sets undergo a common filtering procedure to obtain regions for scene relation learning.

Base proposal generation. Given an image I, features F_I are extracted using the SAM 3 image encoder. Its text encoder E_T produces a fixed prompt representation $\bar{P}_{\text{base}} = E_T(\text{“visual”})$. This prompt contains no training or test category names and remains unchanged across datasets during both training and testing. The base branch uses no additional geometric prompts and generates proposals as follows:

$$P_{\text{base}} = S(F_I, \bar{P}_{\text{base}}, \varnothing), \tag{14}$$

where S denotes the SAM3 proposal generation process, and $\varnothing$ indicates the absence of geometric prompts.

Geometry-guided proposal generation. To supplement proposal coverage at different locations and scales, we construct a set of multiscale box prompts B_g. Specifically, 1×1, 2×2, and 4×4 grids provide 21 boxes, supplemented by three strip-shaped boxes, yielding 24 geometric prompts. All boxes are treated as positive prompts and jointly supplied to a single proposal generation pass:

$$P_{\text{geo}} = S(F_I, \bar{P}_{\text{base}}, B_g), \tag{15}$$

Proposal filtering and merging. For each branch, proposals with confidence scores below τ_s or mask areas smaller than $a_{\min}$ are removed. Mask non-maximum suppression (NMS) is then applied with threshold τ_b, followed by selection of up to K_b proposals ranked by confidence. This yields the filtered sets $\widetilde{P}_{\text{vis}}$ and $\widetilde{P}_{\text{geo}}$. The two sets are subsequently merged and processed by global mask NMS with threshold τ_g. Finally, up to n proposals with the highest confidence scores are retained:

$$\mathcal{R} = \text{TopK}_n \left(\text{MaskNMS}_{\tau_g} \left[\widetilde{P}_{\text{vis}} \cup \widetilde{P}_{\text{geo}} \right] \right), \tag{16}$$

The resulting region set is denoted by $\mathcal{R} = \{R_i\}_{i=1}^{n}$, where $n \leq K$. Each retained proposal is represented as $\mathcal{R}_i = (M_i, l_i, B_i, s_i)$. These regions serve as class-agnostic entities for subsequent scene relation learning.

A.2 DATASETS AND EVALUATION METRICS

A.2.1 DATASET DETAILS

LandDiscover50K (Ye et al., 2025) is used as the common training dataset, and cross-dataset evaluation is conducted on ISPRS Potsdam (ISPRS, 2013), FLAIR (Garioud et al., 2023), FAST (Wang et al., 2023), and FloodNet (Rahnemoonfar et al., 2021). These datasets cover urban scenes, diverse land-cover types, fine-grained land features, and post-disaster environments. All datasets use RGB inputs. Dataset sizes and selection rationales are summarized in Table 5.

Seen and Unseen classes are defined according to overlap with the supervised training vocabulary of LandDiscover50K. Singular and plural forms of class names are normalized. Classes overlapping with the training vocabulary are designated as Seen; non-overlapping classes are designated as Unseen. The same class split are used across all compared methods to ensure consistent comparisons.

Background regions are retained in the visualizations. For quantitative evaluation, background scores are not reported, and the background class is excluded from the class averages used to compute mIoU, mACC, and HM.

Table 5: Dataset overview.

	Datasets	Selection Rationale	Data Volume	Classes	Seen/Unseen
Training	LandDiscover50K	Provides training samples from diverse sources and at multiple spatial scales.	51846	40	–
Evaluation	ISPRS Potsdam	Evaluates segmentation in dense urban scenes.	5472	6	3/2
	FLAIR	Evaluates segmentation across diverse land-cover types.	15700	12	4/8
	FAST	Evaluates segmentation of fine-grained classes and small land features.	3207	37	4/33
	FloodNet	Evaluates generalization to post-flood scenes.	898	10	5/4

A.2.2 Evaluation metrics

mIoU is the mean of the intersection-over-union values over all valid categories. mACC denotes the mean per-class pixel accuracy, where per-class accuracy is defined as the ratio of correctly predicted pixels to the total number of ground-truth pixels for that category. All above metrics are computed based on aggregated pixel statistics from the evaluation set.

To evaluate the OVSS generalization ability while taking the segmentation performance on seen classes into account, we use the harmonic mean (HM) of mIoU over seen and unseen classes as an auxiliary metric, which is defined as:

$$HM = \frac{2 \cdot \mathrm{mIoU_S} \cdot \mathrm{mIoU_U}}{\mathrm{mIoU_S} + \mathrm{mIoU_U}} \tag{17}$$

where $\mathrm{mIoU_S}$ and $\mathrm{mIoU_U}$ represent the mean intersection-over-union for seen and unseen classes, respectively. HM is more sensitive to the group with inferior performance and can comprehensively reflect the segmentation capability of the evaluated approach on both sets of classes.

A.3 Parameter sensitivity analysis

A.3.1 Analysis of reasoning layers ℓ

To investigate the impact of the propagation range of structural information on segmentation performance, we vary the number of reasoning layers ℓ in SRSL and evaluate four configurations with 1, 2, 3, and 4 layers (Fig. 5). In terms of parameter complexity, the number of SRSL parameters increases approximately linearly with the number of reasoning layers, from 1.19M to 4.35M. A single layer aggregates structural information from directly adjacent regions, while multiple layers progressively capture higher-order regional structural context. However, excessively deep reasoning can introduce redundant interactions. Therefore, we adopt a 2-layer relational reasoning configuration.

A.3.2 Analysis of regional proposals n

To investigate the impact of the number of region proposals on scene structure modeling, we set the maximum number of proposals retained for each image to 6, 12, 14, 16, 18, 20, and 22. As shown in Fig. 6, m-mIoU exhibits some variation with the number of proposals, but the overall change remains limited. When the number of proposals increases from 6 to 18, m-mIoU increases from 27.02% to 28.27%; with further increases to 20 and 22 proposals, m-mIoU decreases to 27.64% and 28.08%, respectively. Meanwhile, the average iteration time continuously increases from 0.94 s to 3.50 s. These results indicate that increasing the number of region proposals does not lead to sustained performance gains, while it increases the computational cost of region-relation modeling. Therefore, we adopt 18 region proposals as the default configuration.

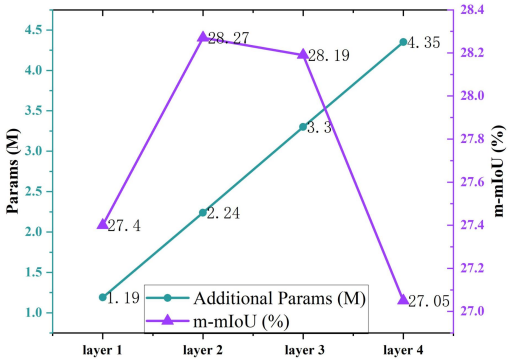

Figure 5: Performance and learnable parameters with different numbers of relational-reasoning layers in SRSL. m-mIoU denotes the average mIoU over four evaluation datasets. Params represent the additional learnable parameters introduced by SRSL.

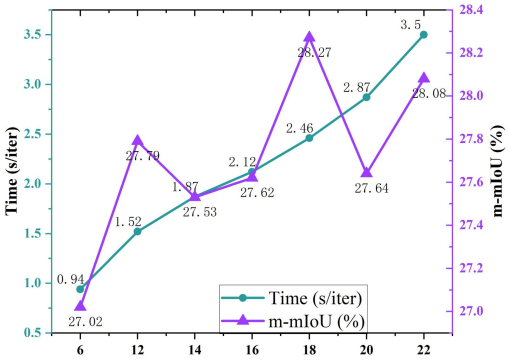

Figure 6: Performance and computational overhead under different numbers of regional proposals. m-mIoU denotes the average mIoU across the four evaluation datasets. Time represents the average time consumed by the model per training iteration.

A.4 Additional Generalization Evaluation

To further evaluate OVSS performance on FloodNet, Table 6 presents per-class IoU together with HM. As a dedicated metric, HM quantifies the balance between seen class recognition and unseen class prediction. It mitigates the masking effect whereby strong performance on one set of classes conceals weaknesses on the other. Based on this definition, our method obtains an HM of 24.02%, the best among the compared methods, exceeding GSNet (21.32%) and OVRS (21.13%), which indicates a better overall trade-off between the seen and unseen classes.

Table 6: Per-class IoU and HM (%) on the FLAIR and Potsdam datasets. Green and purple headers denote **seen** classes and **unseen** classes, respectively.

Method	FloodNet									
	Wat.	Tre.	Veh.	Poo.	Gra.	Bf.	Bnf.	Rf.	Rnf.	HM
CAT-Seg	65.17	42.36	18.09	36.78	60.58	0.04	8.54	32.50	10.84	20.11
EBSeg	13.78	18.39	25.38	6.56	16.95	0.00	0.07	0.18	0.03	0.14
OVRS	65.96	38.73	23.90	38.03	58.78	1.09	8.50	28.73	16.80	21.13
TPOVSeg	47.59	36.29	16.86	22.67	52.30	0.00	3.85	13.18	16.56	13.57
GSNet	66.25	56.35	39.59	17.34	71.90	0.00	3.96	16.37	33.48	**21.32**
RSKT-Seg	54.10	39.50	0.54	14.76	59.34	0.02	1.09	9.05	5.82	7.14
PCA-Seg	61.52	49.05	22.81	39.55	58.50	0.19	3.64	28.73	16.15	19.24
DINODe	48.04	18.32	2.12	1.38	0.02	0.26	5.26	1.97	21.64	9.56
Ours	68.01	49.00	32.46	35.57	65.98	0.13	12.03	32.42	18.64	**24.02**

A.5 Analysis of Boundary Prediction Quality

Fig. 7 compares the boundary prediction quality of different methods using Boundary IoU (BIoU). Among the evaluated methods, ours ranks first on FAST, Potsdam, and FloodNet and second on FLAIR, demonstrating competitive performance in different remote sensing scenes, with the margins of improvement varying by dataset. However, overall scores remain low for all methods, particularly on FAST, indicating that fine boundary prediction still requires further improvement.

A.6 Ablation Study of Relational Structure Modeling

As shown in Table 7, introducing shared relation propagation improves the m-mIoU from 22.83% to 23.94%, showing the benefit of information exchange between region nodes. Distinguishing topological or directional relations further increases it to 25.16% and 24.93%, respectively. Their

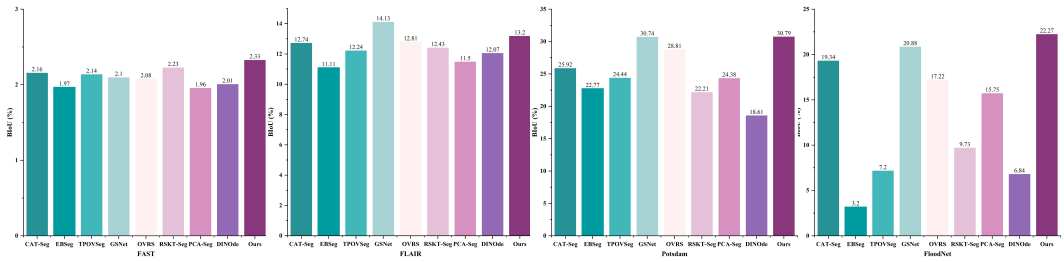

Figure 7: Comparison of segmentation boundary quality measured by BIoU (%) on four datasets.

combination with uniform weighting achieves 26.90%, indicating that the two relation groups provide complementary structural information. Full SRSL reaches 28.27% and consistently improves all four datasets over uniform weighting, with gains of 2.45 and 2.07 percentage points on Potsdam and FloodNet. These results support the joint use of typed spatial relations and adaptive relation weighting.

Table 7: Ablation of relational structure modeling in SRSL. All variants use the same region proposals, region-to-pixel propagation, and training settings. *Shared Relation* treats all connected region pairs as a single relation type. *Topology Only* and *Direction Only* retain only topological and directional relations, respectively. *Uniform Weight* assigns equal contributions to both relation groups, whereas *Full SRSL* adaptively estimates their contributions. All results are reported as mIoU (%), and m-mIoU is averaged over the four datasets.

Configuration	Region Nodes	Shared Relation	Topological Relations	Directional Relations	Relation-aware Weight	FLAIR	FAST	Potsdam	FloodNet	m-mIoU
Node Only	✓	×	×	×	–	15.71	14.66	31.56	29.40	22.83
Shared Relation	✓	✓	×	×	✓	16.47	14.81	34.40	30.09	23.94
Topology Only	✓	×	✓	×	✓	16.18	16.13	36.41	31.90	25.16
Direction Only	✓	×	×	✓	✓	16.93	16.03	37.01	29.76	24.93
Uniform Weight	✓	×	✓	✓	×	17.95	17.23	39.52	32.84	26.90
Full SRSL	✓	×	✓	✓	✓	18.52	17.68	41.97	34.91	28.27

A.7 Analysis of CASP granularity and perception weights

The complete configuration achieves the best m-mIoU among the evaluated settings (Table 8), reaching 28.27%. Replacing the adaptive perception weights of either mode with uniform weights or removing either granularity mode reduces the mean performance. Removing both modes further lowers the mean mIoU to 23.80%. These results indicate that the fine-grained and coarse-grained modes provide complementary spatial context, while content-adaptive perception weights help adjust the contributions of different granularities. Fixed weights may retain candidate granularities that are mismatched to the current location, introducing redundant context.

Table 8: Impact of Granularity Modes and Perception Weights. "A" denotes adaptive perception weights, "U" denotes uniform perception weights, and "–" indicates that the corresponding mode is disabled. All results are reported as mIoU (%), and m-mIoU is averaged over the four datasets.

Variant	Fine weight	Coarse weight	FLAIR	FAST	Potsdam	FloodNet	m-mIoU
Full	A	A	**18.52**	**17.68**	**41.97**	**34.91**	**28.27**
Fine-U	U	A	17.00	**15.85**	33.62	**33.54**	25.00
Coarse-U	A	U	16.53	14.98	36.28	31.89	24.92
Uniform	U	U	16.03	15.19	35.27	29.02	23.88
Coarse mode only	–	A	**17.31**	15.50	36.31	32.84	**25.49**
Fine mode only	A	–	16.21	15.77	**37.80**	29.94	24.93
None	–	–	14.22	14.90	33.54	32.55	23.80

A.8 Ablation study of encoder fine-tuning strategies.

Fine-tuning only the attention modules in CLIP while freezing RemoteCLIP records the best average mIoU and mACC (Table 9). Further fine-tuning the attention modules in RemoteCLIP lowers both average metrics, indicating that retaining the parameters pretrained in RemoteCLIP is more beneficial for overall segmentation performance under the current setting.

Table 9: Comparison of fine-tuning strategies for the CLIP and RemoteCLIP visual encoders.

Fine-Tuning Strategy		**FLAIR**		**FAST**		**Potsdam**		**FloodNet**		**Avg**	
CLIP	RemoteCLIP	mIoU	mACC	mIoU	mACC	mIoU	mACC	mIoU	mACC	m-mIoU	m-mACC
frozen	frozen	16.57	**37.03**	12.94	23.56	26.35	54.73	21.30	30.83	19.29	36.54
attention	frozen	**18.52**	35.69	**17.68**	**26.71**	**41.97**	**66.16**	**34.91**	**44.22**	**28.27**	**43.20**
frozen	attention	**18.71**	**38.42**	12.93	23.62	27.12	55.62	10.63	13.67	17.35	32.83
attention	attention	18.36	34.21	**16.16**	**24.19**	**39.71**	**60.92**	**32.17**	**40.49**	**26.60**	**40.00**

A.9 Effect of feature refinement strategies

To examine refinement strategies between the spatial perception representation and relational structural context, the CASP and SRSL architectures and training settings are kept unchanged, while only the refinement operations between their output features are varied. Four strategies are compared: retaining C without refinement between the two features (No refinement); using S as the residual base and C to learn a residual correction (C→S); using C as the residual base and S to learn a residual correction (S→C); and performing feature refinement in both directions (Bidirectional refinement). Structure-guided refinement of the CLIP dense representation remains active in all configurations. Therefore, No refinement corresponds to the configuration in Table 4 where structural context refines only the CLIP representation.

As shown in Table 10, "S→C" achieves the best average mIoU of 28.27%, exceeding No refinement by 2.86 percentage points. Although "C→S" performs slightly better on FloodNet, its m-mIoU of 26.52% is lower than that of "S→C". Bidirectional refinement obtains an average mIoU of 23.95%, below both unidirectional strategies and "No refinement", indicating that adding reverse refinement provides no performance benefit under the current setting. Overall, these results support using the spatial perception representation as the residual base and relational structural context to provide complementary corrections.

Table 10: Comparison of feature refinement strategies. "C" and "S" denote the spatial perception representation produced by CASP and the structural relational context produced by SRSL, respectively. "→" indicate the direction of feature refinement. Results are reported as mIoU (%), and m-mIoU denotes the average over the four datasets.

Refinement strategy	**FLAIR**	**FAST**	**Potsdam**	**FloodNet**	**m-mIoU**
No refinement	17.70	16.12	35.00	32.80	25.41
C→S	17.49	**17.32**	**36.03**	**35.24**	**26.52**
S→C	**18.52**	**17.68**	**41.97**	**34.91**	**28.27**
Bidirectional	15.04	15.21	35.11	30.42	23.95

A.10 Ablation study of SAM prompt strategies

As shown in Fig. 8, geometric prompts consistently outperform the base prompts when used alone, indicating the importance of spatial constraints for region proposal generation. Combining both prompt types achieves the best mIoU and mACC across all four datasets.

A.11 Limitations and future work

Although our framework achieved the highest BIoU on FAST (see Fig. A.7), indicating that scene dependency modeling helps delineate the spatial extents of small-scale objects, the BIoU and mIoU

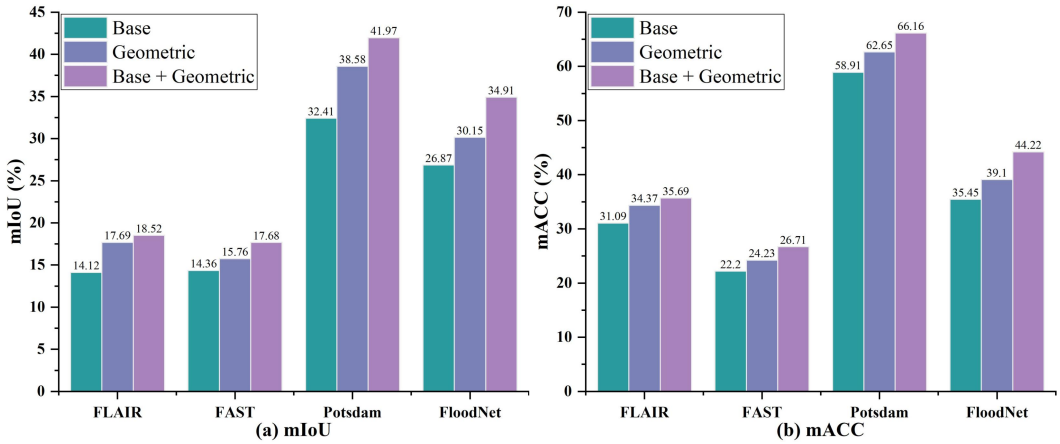

Figure 8: Comparison of three prompting strategies for SAM-based region proposal generation on FLAIR, FAST, Potsdam, and FloodNet using mIoU and mACC (%).

(see Table 1) scores of all methods on this dataset remain generally low. FAST contains numerous small-scale and semantically fine-grained categories, including different aircraft models such as "A220," "A321," and "A330," as well as lexically similar categories such as "Tractor," "Trailer," and "Truck-Tractor." These classes encode subtle distinctions through alphanumeric strings or exhibit substantial lexical overlap. Together with the blurred appearance of some objects, these characteristics can prevent vision–language models from learning sufficiently discriminative textual and visual representations. Moreover, although SCENE achieves a relatively high HM among all compared methods (see Tables 2 and 6), indicating better generalization, all methods still perform poorly on some Seen classes. This phenomenon may be associated with cross dataset shifts in visual distributions and the degradation of some pretrained knowledge during source-domain training. In general, existing methods still have room for improvement in fine-grained class vocabulary representation and pretrained knowledge preservation.

In future work, we will investigate domain-aware tokenization, compositional parsing of class names, and attribute-enriched class descriptions to construct more discriminative fine-grained textual representations. We will also explore knowledge-preserving adaptation strategies to better retain the transferability of vision–language models while learning remote-sensing-specific features.