CONCORD-SAM3: NATIVE EVIDENCE COORDINATION FOR TRAINING-FREE OPEN-VOCABULARY REMOTE SENSING SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Open-vocabulary remote sensing semantic segmentation aims to assign pixel-level labels to arbitrary textual categories beyond predefined land-cover taxonomies. However, independent prompt responses lack explicit cross-class coordination, while textual descriptions alone do not provide image-specific visual references for diverse within-class appearances. We introduce CONCORD-SAM3, a training-free framework built on frozen SAM 3 that coordinates native evidence across competing classes and within-image appearances. Dual-Head Evidence Coordination (DHEC) calibrates class scores using shared visual–text descriptors and accumulates pixelwise instance support through a bounded noisy-OR rule. Positive semantic–instance soft-union residuals are gated by query presence and semantic support within native masks. Multi-Prototype Visual Calibration (MPVC) constructs multiple prototypes per class from the current image's multi-scale features. A prediction-margin gate strengthens visual calibration at ambiguous pixels while reducing intervention where scores are well separated. The framework uses one image encoding per image or crop and one prompt evaluation per supplied query, without target-domain annotations, parameter updates, auxiliary backbones, re-querying, or transformed views. Evaluations on 16 datasets demonstrate the effectiveness of our method.

1 INTRODUCTION

Remote-sensing segmentation must combine vocabulary flexibility with spatial precision. Aerial scenes juxtapose compact vehicles and buildings with extended roads, waterways, and vegetation, while target categories vary across applications (Yuan et al., 2021). Collecting new pixel annotations whenever the vocabulary changes is costly and difficult to scale. Open-vocabulary semantic segmentation (OVSS) addresses this constraint by assigning pixel labels from textual category descriptions at inference time (Liang et al., 2023; Cho et al., 2024). The training-free setting is especially appealing for Earth observation, enabling frozen foundation models to process requested categories without target-domain parameter updates (Li et al., 2025b; 2026b).

Vision–language models such as CLIP (Radford et al., 2021) provide transferable image–text alignment, but image-level pretraining does not directly ensure precise dense localization. Training-free methods bridge this gap by revising attention (Wang et al., 2024; Lan et al., 2024a), recovering intermediate features (Yang et al., 2025), or introducing external spatial priors (Lan et al., 2024b; Zhang et al., 2025). SAM 3 (Carion et al., 2026) instead provides concept-conditioned semantic responses, instance masks, and presence scores within one architecture. These outputs offer complementary evidence, but query-wise inference does not explicitly coordinate scores across the requested vocabulary. Figure 1 illustrates the resulting gap: individual objects can be captured while regional coverage and class assignments remain inconsistent. Our central question is: *how can native predictions and features be coordinated across classes and within-image appearances without additional image encoding, auxiliary backbones, or repeated prompt evaluations?* We identify two coupled challenges.

First, query-wise evidence requires class-level coordination. A strong prompt response supports one concept without explicitly weighing alternatives in the vocabulary. Argmax compares response

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.

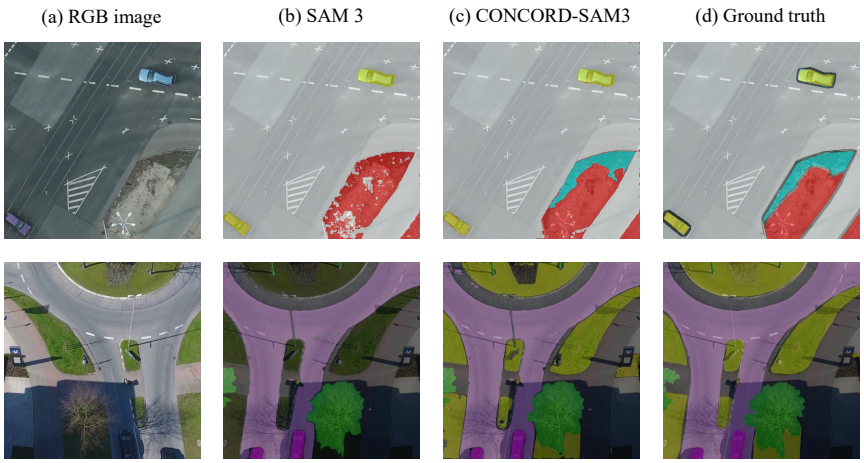

Figure 1: Qualitative comparison on two remote-sensing datasets: (a) RGB image, (b) SAM 3, (c) CONCORD-SAM3, and (d) ground truth. CONCORD-SAM3 improves regional class assignments and spatial coverage in these examples.

magnitudes but does not recalibrate class preferences using shared visual–text evidence. Maximum fusion likewise retains only the strongest semantic or instance response at each pixel, rather than accumulating mask-supported evidence and regulating it through semantic agreement. The challenge is to combine explicit cross-concept competition with reliability-gated instance support.

Second, semantic categories require image-specific visual references. Metallic industrial roofs and weathered tile roofs can share a category despite different textures and reflectance. Text identifies the concept but does not explicitly provide image-specific visual references; a single pooled prototype may also under-represent multiple appearances. The challenge is to derive several representative references without annotations and regulate their influence according to prediction ambiguity.

We introduce **CONCORD-SAM3**, a two-stage framework built on frozen SAM 3. The objective is not to obtain another set of foundation-model predictions, but to coordinate complementary semantic, instance, and visual evidence: class-level calibration establishes relative preferences, while image-level prototypes supply appearance references tailored to the scene. **Dual-Head Evidence Coordination (DHEC)** calibrates competing class scores using visual–text descriptors and supplements them with pixelwise instance support. A bounded noisy-OR accumulation and semantic–instance soft union produce positive residuals controlled by query presence and mask-level semantic agreement. **Multi-Prototype Visual Calibration (MPVC)** then constructs three confidence-weighted prototypes per class from native multi-scale features. A prediction-margin gate gives these visual references greater influence at ambiguous pixels while reducing intervention elsewhere. Both stages reuse existing predictions and features without additional image encoding or prompt evaluations. Experiments demonstrate improvements across diverse remote-sensing segmentation and extraction benchmarks.

Our contributions are threefold:

- We introduce CONCORD-SAM3 for training-free remote-sensing OVSS, coordinating native evidence at the class and image levels without additional image encoding or prompt evaluations.
- We develop DHEC to couple cross-concept score calibration with agreement-gated semantic–instance residual fusion.
- We propose MPVC to represent within-image class variation by constructing multiple prototypes per class from the current image's multi-scale features and applying margin-gated visual calibration.

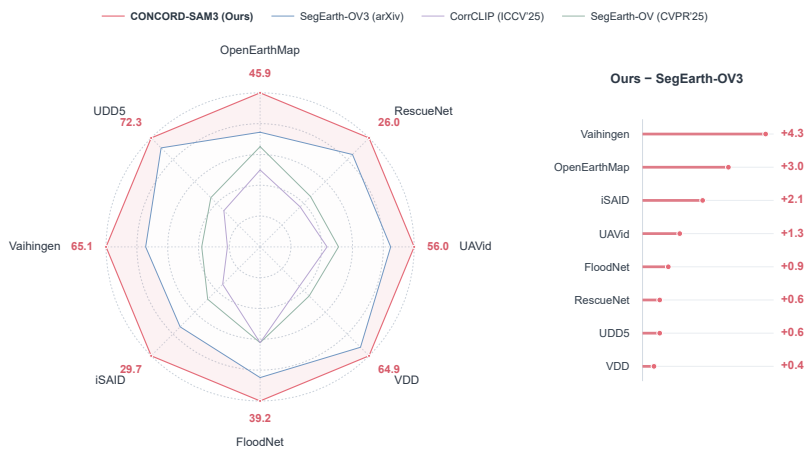

Figure 2: Comparison on eight selected remote-sensing benchmarks. Left: radar-style comparison of CONCORD-SAM3 with SegEarth-OV3, CorrCLIP, and SegEarth-OV; numeric labels give CONCORD-SAM3 mIoU (%). Right: absolute mIoU gains over SegEarth-OV3, in percentage points. Complete results on all ten benchmarks are reported in Table 1.

2 Related Work

Open-vocabulary dense prediction. OVSS assigns pixel labels from textual categories supplied at inference time. A major line transfers CLIP's image–text alignment to dense localization. Training-based methods include CAT-Seg (Cho et al., 2024), which aggregates image–text matching costs across spatial positions and categories; DCP-CLIP (Wang et al., 2026), which combines dynamic category selection, text-guided alignment, and coarse-to-fine decoding; and X-Agent (Li et al., 2025a), which enhances unseen-category semantics through agent-mediated cross-modal attention. Training-free methods instead reuse frozen representations: MaskCLIP (Zhou et al., 2022) extracts final-layer value features; SCLIP (Wang et al., 2024) uses query–query and key–key attention; ClearCLIP (Lan et al., 2024a) combines self-self attention with removal of the final residual and feed-forward paths; and ResCLIP (Yang et al., 2025) incorporates intermediate cross-correlations and semantic feedback. CONCORD-SAM3 operates on native SAM 3 predictions and features rather than adapting CLIP's image-level representations.

Visual priors and test-time evidence refinement. ProxyCLIP (Lan et al., 2024b) uses visual correspondences as proxy attention. Trident (Shi et al., 2025) combines CLIP and DINO (Caron et al., 2021) features across crops, uses SAM (Kirillov et al., 2023) features for global aggregation, and refines coarse predictions through SAM prompting. Meanwhile, CorrCLIP (Zhang et al., 2025) uses SAM-derived regions and DINO similarities to reconstruct patch interactions before region-wise majority correction. These approaches combine semantic and spatial evidence from heterogeneous encoders. CONCORD-SAM3 instead reuses existing SAM 3 outputs, with MPVC deriving image-specific visual references and regulating their contribution through prediction-margin gating. These references come from the current image rather than an auxiliary visual encoder.

Remote-sensing segmentation with foundation models. Overhead imagery combines scale variation, dense small objects, and heterogeneous land-cover appearance (Yuan et al., 2021). RemoteCLIP (Liu et al., 2024a) and RS5M/GeoRSCLIP (Zhang et al., 2024) strengthen image–text alignment through remote-sensing pretraining or adaptation. SegEarth-OV (Li et al., 2025b) combines learned feature upsampling with local–global token subtraction, while SegEarth-OV3 (Li et al., 2026b) explores SAM 3 for training-free remote-sensing OVSS by leveraging its native semantic, instance, and presence outputs. CONCORD-SAM3 explicitly coordinates native SAM 3 evidence across competing classes and image-specific appearances through cross-concept calibration, agreement-gated residual fusion, and multi-prototype visual calibration, without target-domain parameter updates or additional foundation-model evaluations.

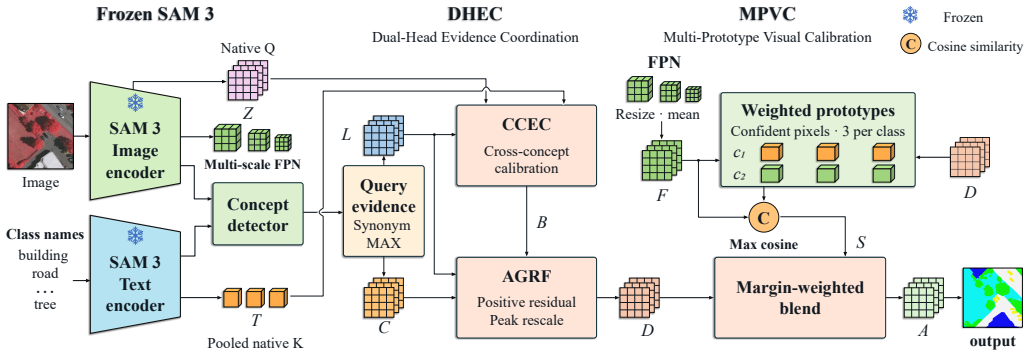

Figure 3: Overview of CONCORD-SAM3. Cross-Concept Evidence Calibration (CCEC) maps L to B; Agreement-Gated Residual Fusion (AGRF) combines B with $[C - L]_+$ to produce D. MPVC uses D and aggregated features F to form class prototypes, computes similarity evidence S, and blends it with D into final scores A. The circled C denotes cosine similarity.

3 METHOD

Given an aerial or satellite image $I \in \mathbb{R}^{H\times W\times 3}$ and $K \geq 2$ classes, each represented by text queries $\mathcal{Q}_c$, we seek a dense semantic label map without updating model parameters. CONCORD-SAM3 coordinates native frozen SAM 3 (Carion et al., 2026) evidence through DHEC and MPVC: the former compares class concepts and integrates instance-region support; the latter calibrates scores using image-specific visual prototypes. Each image or crop is encoded once, with one prompt call per supplied query. Both stages reuse these predictions and features without learnable parameters or additional model evaluations. Figure 3 outlines the pipeline.

3.1 DUAL-HEAD EVIDENCE COORDINATION

Accumulating reliable instance evidence. For query q, let $P_q(x)$ and $I_{qj}(x)$ denote semantic and instance-mask probabilities, p_q the query presence confidence, and $\bar{o}_{qj}$ the instance score supplied by the inference interface. To obtain dense class scores from these native SAM 3 outputs, we combine semantic and instance responses with the query presence confidence and max-merge synonym queries:

$$\ell_q(x) = p_q \max\left\{P_q(x), \max_j \bar{o}_{qj} I_{qj}(x)\right\}, \qquad L_c(x) = \max_{q\in\mathcal{Q}_c} \ell_q(x). \tag{1}$$

Here, L denotes the scores of the frozen SAM 3 baseline used in our ablation study. The index j denotes retained instances, and an empty instance maximum is zero. This readout keeps only the strongest response at each pixel. We additionally accumulate support from native binary masks M_{qj} through a bounded noisy-OR rule, using raw object confidences o_{qj}, distinct from $\bar{o}_{qj}$. Let $\mathcal{J}_q^+$ contain retained instances with nonempty masks and available raw confidences. Their accumulated instance evidence and region-level semantic agreement are

$$U_q(x) = 1 - \prod_{j\in\mathcal{J}_q^+} \left[1 - o_{qj} I_{qj}(x) M_{qj}(x)\right], \qquad a_q = \frac{1}{|\mathcal{J}_q^+|} \sum_{j\in\mathcal{J}_q^+} \frac{\sum_x M_{qj}(x) P_q(x)}{\sum_x M_{qj}(x)}. \tag{2}$$

Probabilities entering the product are clipped to $[0, 1]$. The agreement a_q averages semantic support within each mask, then equally across instances. We combine semantic and accumulated instance support through the bounded soft union $P_q + U_q - P_q U_q$. Presence and agreement then control a positive candidate residual:

$$c_q = \ell_q + p_q a_q \left[p_q \left(P_q + U_q - P_q U_q\right) - \ell_q\right]_+, \qquad [z]_+ = \max(z, 0). \tag{3}$$

Only the positive excess of this presence-weighted union over ℓ_q contributes; weak presence or semantic agreement reduces its influence. If $\mathcal{J}_q^+$ is empty, $c_q = \ell_q$. Candidate class scores are $C_c = \max_{q\in\mathcal{Q}_c} c_q$.

Cross-concept evidence calibration (CCEC) and residual fusion (AGRF). Evaluating prompts separately does not explicitly compare competing class names. We introduce cross-class competition through visual–text evidence before adding the region residual. Visual descriptors are extracted from vision block 18 (zero-based), projected through SAM 3's frozen unit-scale neck and the first fusion-encoder layer's cross-attention query projection. Current text tokens use the corresponding key projection and mean pooling over non-padding positions. For the resulting ℓ_2-normalized descriptors $Z(u)$ and T_q, class evidence on the visual grid is

$$e_c(u) = \tau_s \log \left[\frac{1}{|\mathcal{Q}_c|} \sum_{q \in \mathcal{Q}_c} \exp\left(\frac{Z(u)^\top T_q}{\tau_s} \right) \right], \qquad \tau_s = 0.10. \tag{4}$$

LogMeanExp emphasizes stronger matches while normalizing for query count. Bilinear resizing yields E on the score grid, followed by $\pi = \mathrm{softmax}(E/\tau_{\mathrm{cls}})$ with $\tau_{\mathrm{cls}} = 1$. Here, τ_s controls aggregation across synonym queries, whereas τ_{cls} controls the softmax across classes. The base prior is normalized across classes at each pixel: $b_c = ([L_c]_+ + \epsilon)/\sum_k ([L_k]_+ + \epsilon)$, where $\epsilon = 10^{-6}$.

Define the pixelwise rescaling operator $\mathcal{N}(X;Y)_c = X_c \max_k Y_k / \max(\max_k X_k, \epsilon)$. We first calibrate class preferences, then add the positive class residual at pixels where $[C - L]_+$ has a nonzero entry:

$$\begin{aligned} z &= \mathrm{softmax}(\log b + \lambda \log \pi), \qquad B = \mathcal{N}(z; L), \\ D &= \mathcal{N}(B + [C - L]_+; B)\,. \end{aligned} \tag{5}$$

In Eq. equation 5, the first line implements CCEC, and the second implements AGRF. Here, λ controls the strength of cross-concept evidence relative to the native class prior. Softmax acts across classes. Pixels with zero class residual retain B unchanged. Both steps preserve the incoming per-pixel maximum score within numerical tolerance. Missing or invalid projected features trigger identity fallback for concept calibration.

3.2 Multi-Prototype Visual Calibration

Image-adaptive prototype discovery. A class can exhibit varied appearances within one scene, such as weathered and newly built roofs. A single prototype may under-represent this variation. MPVC constructs three confidence-weighted visual prototypes per class from the current image. Available SAM 3 feature pyramid network (FPN) levels are bilinearly resized to the finest available grid and averaged into $F(u)$; channelwise ℓ_2 normalization gives $f(u)$. Let $w_c(u)$ be the DHEC score clipped to $[0, 1]$ and then resized to this grid. For each class independently, Ω_c contains the $\lceil 0.30N \rceil$ highest-scoring locations, where N is the grid size. The first center is a weighted mean; two further centers ($r = 2, 3$) use confidence-weighted farthest-point sampling:

$$\begin{aligned} \mu_{c1} &= \frac{\sum_{u \in \Omega_c} w_c(u) F(u)}{\max\left(\sum_{u \in \Omega_c} w_c(u), \epsilon\right)}, \\ u_r^\star &= \arg\max_{u \in \Omega_c} w_c(u) \left[1 - \max_{k<r} \cos(f(u), \mu_{ck}) \right]_+, \qquad \mu_{cr} = F(u_r^\star). \end{aligned} \tag{6}$$

Two assignment–update iterations assign the selected pixels to their most similar centers under cosine similarity, then recompute confidence-weighted means of F. Clusters with total weight at most ϵ retain their previous centers. The resulting prototypes provide multiple image-specific visual references without changing network weights.

Margin-gated visual calibration. Prototype evidence is $S_c(u) = \max_{r=1,2,3} \cos(f(u), \mu_{cr})$, bilinearly resized to the score grid. Let $\Delta_D(x) = D_{(1)}(x) - D_{(2)}(x)$ denote the gap between the highest and second-highest DHEC scores. We blend visual evidence through a margin-sensitive gate:

$$A_c(x) = D_c(x) + \alpha \exp\left(-\frac{\Delta_D(x)}{T_m} \right) [S_c(x) - D_c(x)]\,. \tag{7}$$

Here, α controls the maximum contribution of prototype evidence, while T_m determines how rapidly this contribution decays as the prediction margin increases. The gate reduces prototype influence when class scores are well separated. As the margin approaches zero, the blending weight approaches α, allowing stronger visual adjustment.

Final prediction. For sliding-window inference, calibrated crop scores are averaged in overlapping regions. Final labels follow classwise argmax; pixels whose maximum score is below the dataset-specific threshold are assigned to the configured background class.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and metrics. We evaluate multi-class segmentation on ten remote-sensing benchmarks: OpenEarthMap (Xia et al., 2023), LoveDA (Wang et al., 2021), Potsdam (ISPRS, a), RescueNet (Rahnemoonfar et al., 2023), UAVid (Lyu et al., 2020), VDD (Cai et al., 2025), FloodNet (Rahnemoonfar et al., 2021), iSAID (Waqas Zamir et al., 2019), Vaihingen (ISPRS, b), and UDD5 (Chen et al., 2018), covering land-cover, urban, UAV, object-centric, and disaster-related scenes. We report mean Intersection-over-Union (mIoU) over evaluated classes. Building extraction uses WHU-Aerial (Ji et al., 2019), Inria (Maggiori et al., 2017), and xBD-pre (Gupta et al., 2019); road extraction uses DeepGlobe (Demir et al., 2018), Massachusetts Roads (Mnih, 2013), and SpaceNet (Van Etten et al., 2018). These binary tasks report foreground-class IoU, not foreground–background mIoU. All scores are percentages; Avg. is the unweighted mean over all datasets in a table, omitted for incomplete rows.

Implementation details. We use frozen SAM 3 with dataset category names as queries and max-merge synonyms. Unless otherwise specified, DHEC and MPVC use block 18 descriptors, $\tau_s = 0.10$, $\tau_{\text{cls}} = 1$, three prototypes per class, top-30% selection, two assignment–update iterations, and $T_m = 0.20$. The choices of λ and α are examined in the parameter sensitivity study in Section 4.4. No prompt learning, target-domain parameter updates, auxiliary backbone, re-querying, or transformed-view evaluation is used.

4.2 REMOTE-SENSING SEGMENTATION AND EXTRACTION

Multi-class semantic segmentation. Table 1 compares CONCORD-SAM3 with supervised reference methods, including SCAN (Liu et al., 2024b), SAN (Xu et al., 2023), SED (Xie et al., 2024), CAT-Seg (Cho et al., 2024), OVRS (Cao et al., 2025), GSNet (Ye et al., 2025), RSKT-Seg (Li et al., 2026a), SkySense-O (Zhu et al., 2025), and ROSS (Zhang et al., 2026), as well as representative training-free approaches, including CLIP (Radford et al., 2021), MaskCLIP (Zhou et al., 2022), SCLIP (Wang et al., 2024), GEM (Bousselham et al., 2024), ClearCLIP (Lan et al., 2024a), ProxyCLIP (Lan et al., 2024b), SegEarth-OV (Li et al., 2025b), Trident (Shi et al., 2025), CorrCLIP (Zhang et al., 2025), ResCLIP (Yang et al., 2025), and SegEarth-OV3 (Li et al., 2026b). Missing entries remain unavailable rather than being estimated or averaged over inconsistent dataset subsets.

CONCORD-SAM3 achieves 50.5% average mIoU versus 49.1% for SegEarth-OV3 (Li et al., 2026b), with the highest reported training-free score on all ten listed datasets. Gains include 4.3 points on Vaihingen, 3.0 on OpenEarthMap, and 2.1 on iSAID. RescueNet and FloodNet also improve, from 25.4 to 26.0 and 38.3 to 39.2, respectively. These results support coordinating class scores and image-specific visual evidence beyond maximum fusion. Section 4.3 examines component effects.

Qualitative comparison. Figure 4 compares CONCORD-SAM3 with CLIP, CorrCLIP, SegEarth-OV, and SegEarth-OV3 on five remote-sensing scenes. In the displayed examples, our predictions combine coherent regions with localized object detail. In particular, the UDD5 and OpenEarthMap examples show fewer scattered responses than SegEarth-OV3 along the road and over water, respectively. The Potsdam and iSAID rows also illustrate the preservation of vehicle and ship structures. These observations complement the quantitative results in Table 1.

Building and road extraction. CONCORD-SAM3 obtains the highest reported foreground IoU in every column of Table 2. Improvements over SegEarth-OV3 (Li et al., 2026b) are 0.1, 0.1, and 0.7 points on WHU-Aerial, Inria, and xBD-pre, and 0.5, 2.9, and 8.8 on DeepGlobe, Massachusetts,

Method	Source	OEM	LoveDA	Pots.	Rescue	UAVid	VDD	Flood	iSAID	Vaih.	UDD5	Avg.
Training on remote-sensing segmentation data												
SCAN	CVPR'24	–	18.2	20.2	–	18.6	26.3	–	34.2	5.4	34.1	–
SAN	CVPR'23	–	23.2	30.3	–	22.3	34.8	–	30.6	31.9	37.2	–
SED	CVPR'24	–	21.3	19.5	–	20.1	31.4	–	21.5	29.4	35.7	–
CAT-Seg	CVPR'24	–	25.5	26.8	–	24.6	36.2	–	23.6	32.3	40.2	–
OVRS	TGRS'25	–	28.7	27.5	–	25.2	37.3	–	39.1	43.5	40.8	–
GSNet	AAAI'25	–	29.3	26.5	–	25.4	38.10	–	42.00	33.7	40.9	–
RSKT-Seg	AAAI'26	–	32.5	34.5	–	28.1	41.2	–	44	35.2	42.1	–
SkySense-O	CVPR'25	40.8	38.3	54.1	–	–	–	–	43.9	51.6	–	–
ROSS	PR'26	–	–	30.4	–	–	–	–	40.8	34.0	–	–
Training-free												
CLIP	ICML'21	21.5	20.9	28.0	4.0	21.8	26.7	8.7	9.7	13.6	25.7	18.1
MaskCLIP	ECCV'22	25.1	26.8	34.9	13.8	32.8	34.9	25.5	14.5	22.3	35.4	26.6
SCLIP	ECCV'24	30.8	28.1	41.4	15.4	35.6	39.9	26.1	15.9	20.6	39.5	29.3
GEM	CVPR'24	30.8	28.3	41.1	15.9	35.9	39.4	27.2	16.4	21.7	40.0	29.7
ClearCLIP	ECCV'24	29.8	27.7	40.2	15	34.5	38.2	25.7	15.8	20.6	38.2	28.6
ProxyCLIP	ECCV'24	33.3	30.9	42.1	7.9	37.2	39.3	16.8	16.4	25.8	36.3	28.6
SegEarth-OV	CVPR'25	40.3	36.9	48.5	18.5	42.5	45.3	33.6	21.7	40.0	50.6	37.8
Trident	ICCV'25	32.0	30.3	38.3	15.7	34.9	36.7	33.6	14.3	24.6	37.5	29.8
CorrCLIP	ICCV'25	34.4	32.7	42.6	15.6	38.1	38.4	33.6	16.9	24.7	40.1	31.7
ResCLIP	CVPR'25	30.3	27.7	34.9	9.9	32.0	29.5	26.6	14.3	19.9	32.9	25.8
SegEarth-OV3	arXiv	42.9	47.4	57.8	25.4	54.7	64.5	38.3	27.6	60.8	71.7	49.1
CONCORD-SAM3	Ours	**45.9**	**47.5**	**58.0**	**26.0**	**56.0**	**64.9**	**39.2**	**29.7**	**65.1**	**72.3**	**50.5**

Table 1: Remote-sensing segmentation (mIoU, %). Bold denotes the best training-free result; – denotes unavailable entries. Avg. uses all ten datasets. OEM/Pots./Rescue/Flood/Vaih. denote OpenEarthMap/Potsdam/RescueNet/FloodNet/Vaihingen.

Method	Building extraction			Road extraction		
	$\text{WHU}^{\text{Aerial}}$	Inria	xBD^{pre}	DeepGlobe	Massachusetts	SpaceNet
CLIP	17.7	19.6	16.0	3.9	4.9	7.1
MaskCLIP	29.8	33.4	29.2	13.2	10.6	20.8
SCLIP	33.4	34.9	25.9	7.0	7.4	14.9
GEM	24.4	28.5	20.8	4.7	5.1	11.9
ClearCLIP	36.6	39.0	30.1	5.7	6.4	16.3
SegEarth-OV	49.2	44.6	37.0	17.8	11.5	23.8
SegEarth-OV3	86.9	72.4	64.3	39.3	27.7	35.6
CONCORD-SAM3	**87.0**	**72.5**	**65.0**	**39.8**	**30.6**	**44.4**

Table 2: Building and road extraction (foreground IoU, %). Bold marks the best listed result; Massachusetts denotes the road benchmark.

and SpaceNet. The results extend the observed benefit to category-specific foreground extraction, with the largest gain on SpaceNet.

4.3 Ablation Study

Table 3 reports a progressive ablation on Potsdam, RescueNet, and FloodNet, starting from the frozen SAM 3 baseline, followed by DHEC and the complete DHEC–MPVC pipeline. CCEC and AGRF are enabled together as DHEC.

Adding DHEC improves the frozen SAM 3 baseline by 2.2, 2.5, and 20.0 mIoU points on Potsdam, RescueNet, and FloodNet, respectively. Incorporating MPVC after DHEC provides further gains of 0.3, 0.6, and 0.4 points, yielding the best result in each column. The complete pipeline improves the baseline by 2.5, 3.1, and 20.4 points. These progressive results support the two-stage design:

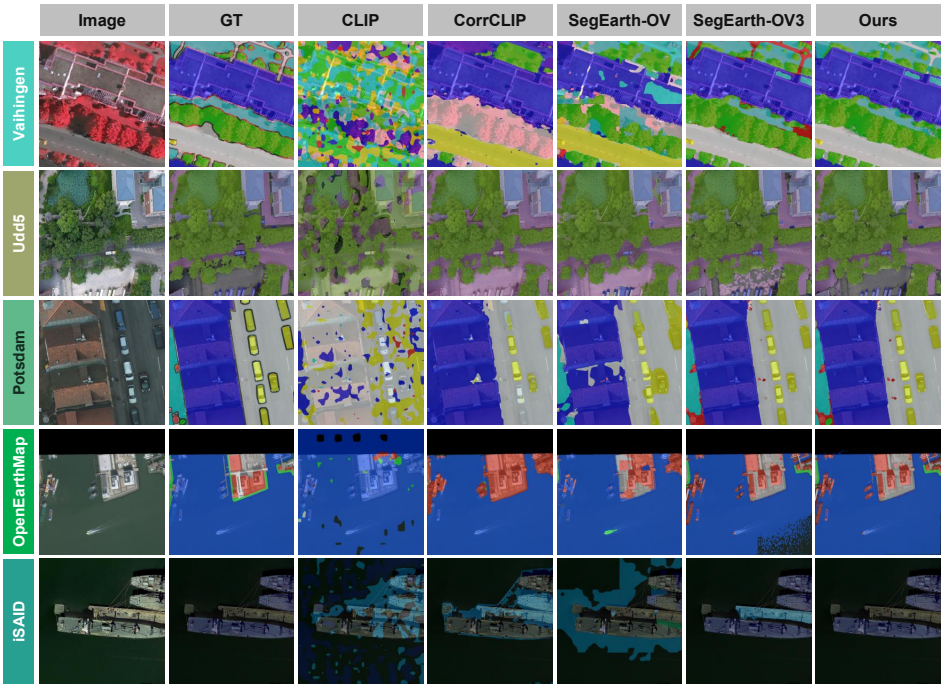

Figure 4: Qualitative comparison on five remote-sensing benchmarks. Columns show the input image, ground truth (GT), CLIP, CorrCLIP, SegEarth-OV, SegEarth-OV3, and CONCORD-SAM3 (Ours), from left to right. Rows correspond to Vaihingen, UDD5, Potsdam, OpenEarthMap, and iSAID, from top to bottom.

Config	DHEC	MPVC	Potsdam	RescueNet	FloodNet
SAM 3	–	–	55.5	22.9	18.8
+ DHEC	✓	–	57.7	25.4	38.8
+ MPVC	✓	✓	**58.0**	**26.0**	**39.2**

Table 3: Progressive component ablation (mIoU, %), starting from the frozen SAM 3 baseline. Check marks denote enabled modules; bold marks the best result.

DHEC supplies coordinated class scores for prototype weighting, while MPVC adds image-specific visual evidence that further improves the DHEC output.

4.4 Parameter Sensitivity

We investigate the sensitivity of CONCORD-SAM3 to its two principal calibration parameters: λ in CCEC and α in MPVC. The former controls the contribution of cross-concept evidence relative to the native class prior, while the latter determines the maximum strength of prototype-based visual calibration. We vary one parameter at a time while keeping the other fixed at its adopted value. Each tested (λ, α) configuration is shared across VDD, Vaihingen, Potsdam, and OpenEarthMap, without dataset-specific adjustment of these two parameters.

Unless otherwise specified, we use $\lambda = 0.70$ and $\alpha = 0.05$ throughout the main experiments. The following analysis examines the effectiveness of this configuration and its sensitivity to nearby parameter choices.

Figure 5(a) shows that $\alpha = 0.05$ achieves the highest observed mIoU on all four displayed datasets. Performance progressively decreases as α increases beyond this value, indicating that stronger prototype-based calibration does not necessarily improve the final predictions. These results fa-

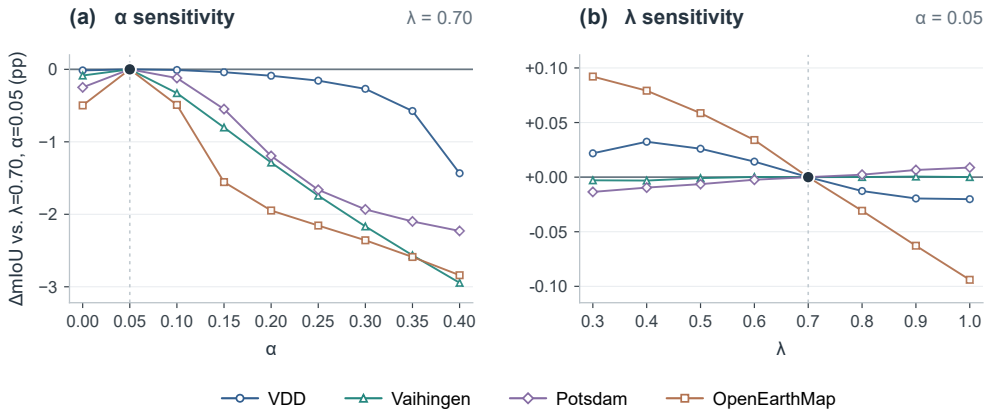

Figure 5: Parameter sensitivity on VDD, Vaihingen, Potsdam, and OpenEarthMap. (a) Varying α at fixed $\lambda = 0.70$. (b) Varying λ at fixed $\alpha = 0.05$. Curves show full-dataset mIoU changes, in percentage points, relative to the adopted configuration $(\lambda, \alpha) = (0.70, 0.05)$. The panels use different vertical-axis ranges.

vor a modest calibration weight that introduces image-specific visual evidence while preserving the existing class-score structure.

With $\alpha = 0.05$ fixed, Figure 5(b) shows that performance is comparatively insensitive to λ over the tested range. The adopted $\lambda = 0.70$ achieves a four-dataset mean mIoU of 58.5%, close to the highest observed mean in the sweep. Neighboring choices, including $\lambda = 0.60$ and $\lambda = 0.80$, yield very similar results on each displayed dataset, indicating a locally stable performance region around the adopted setting.

Together, these observations support $(\lambda, \alpha) = (0.70, 0.05)$ as an effective operating configuration: $\alpha = 0.05$ provides the best measured performance in the α sweep on these four datasets, while $\lambda = 0.70$ maintains a near-best observed average with little sensitivity to neighboring tested values.

5 Conclusion

We presented **CONCORD-SAM3**, a training-free framework that coordinates native frozen SAM 3 evidence across class and image levels. DHEC couples cross-concept calibration with agreement-gated semantic–instance residual fusion. MPVC complements these scores with confidence-weighted class prototypes and margin-gated visual calibration to represent within-image appearance variation. Both stages reuse existing predictions and features without parameter updates, auxiliary backbones, or additional image encoding and prompt evaluations. Experiments show improvements across remote-sensing segmentation and extraction benchmarks, while component ablations support MPVC's complementary role after DHEC. The central insight is to jointly organize semantic–instance evidence and regulate the contribution of image-specific visual references, rather than relying on additional foundation-model predictions.

AI Use Statement

Generative AI tools were used for language polishing and LaTeX assistance. All research ideas, methodology, experiments, results, and conclusions were developed and verified by the authors. The authors take full responsibility for the final content of this work.

Reproducibility Statement

We provide anonymized DHEC and MPVC core code with $\lambda = 0.70$ and $\alpha = 0.05$. The archive supports method inspection; full benchmark reproduction additionally requires model integration, pretrained weights, dataset configurations, and evaluation scripts, which are not included.

References

Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3828–3837, 2024.

Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *Journal of Visual Communication and Image Representation*, 109:104429, 2025.

Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary high-resolution remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–14, 2025.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, RISHI HAZRA, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollar, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=r35clVtGzw.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *Chinese Conference on Pattern Recognition and Computer Vision*, pp. 347–359. Springer, 2018.

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

Ilke Demir, Krzysztof Koperski, David Lindenbaum, Guan Pang, Jing Huang, Saikat Basu, Forest Hughes, Devis Tuia, and Ramesh Raskar. DeepGlobe 2018: A challenge to parse the earth through satellite images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 172–181, June 2018. URL https://openaccess.thecvf.com/content_cvpr_2018_workshops/w4/html/Demir_DeepGlobe_2018_A_CVPR_2018_paper.html.

Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xBD: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019. URL https://arxiv.org/abs/1911.09296.

ISPRS. 2D Semantic Labeling Contest – Potsdam. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx, a. Accessed: 2026-06-08.

ISPRS. 2D Semantic Labeling Contest – Vaihingen. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-vaihingen.aspx, b. Accessed: 2026-06-08.

Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE Transactions on Geoscience and Remote Sensing*, 57(1):574–586, 2019. doi: 10.1109/TGRS.2018.2858817.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, October 2023.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *European Conference on Computer Vision*, pp. 143–160. Springer, 2024a.

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, 2024b.

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, 2026a.

Jiahao Li, Yang Lu, Yachao Zhang, Fangyong Wang, Yuan Xie, and Yanyun Qu. Novel category discovery with X-Agent attention for open-vocabulary semantic segmentation. *arXiv preprint arXiv:2509.01275*, 2025a. URL https://arxiv.org/abs/2509.01275.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10545–10556, 2025b.

Kaiyu Li, Shengqi Zhang, Yujie Wang, Yupeng Deng, Zhi Wang, Deyu Meng, and Xiangyong Cao. Segearth-ov3: Exploring sam 3 for open-vocabulary semantic segmentation in remote sensing images, 2026b. URL https://arxiv.org/abs/2512.08730.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7061–7070, 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, 2024a.

Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3491–3500, 2024b.

Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 165:108–119, 2020.

Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*, pp. 3226–3229. IEEE, 2017.

Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

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.

Maryam Rahnemoonfar, Tashnim Chowdhury, and Robin Murphy. Rescuenet: A high resolution uav semantic segmentation dataset for natural disaster damage assessment. *Scientific data*, 10(1): 913, 2023.

Yuheng Shi, Minjing Dong, and Chang Xu. Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23487–23497, 2025.

Adam Van Etten, Dave Lindenbaum, and Todd M. Bacastow. SpaceNet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. URL https://arxiv.org/abs/1807.01232.

Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *European conference on computer vision*, pp. 315–332. Springer, 2024.

Jing Wang, Huimin Shi, Quan Zhou, Qibo Liu, Suofei Zhang, and Huimin Lu. DCP-CLIP: A coarse-to-fine framework for open-vocabulary semantic segmentation with dual interaction. *arXiv preprint arXiv:2603.13951*, 2026. URL https://arxiv.org/abs/2603.13951.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In J. Vanschoren and S. Yeung (eds.), *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021. URL https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/4e732ced3463d06de0ca9a15b6153677-Paper-round2.pdf.

Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 28–37, 2019.

Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6254–6264, 2023.

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

Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2945–2954, 2023.

Yuhang Yang, Jinhong Deng, Wen Li, and Lixin Duan. Resclip: Residual attention for training-free dense vision-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29968–29978, 2025.

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.

Xiaohui Yuan, Jianfang Shi, and Lichuan Gu. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 169:114417, 2021. doi: 10.1016/j.eswa.2020.114417. URL https://www.sciencedirect.com/science/article/pii/S0957417420310836.

Da Zhang, Mingmin Zeng, and Xuelong Li. Towards open-vocabulary semantic segmentation for remote sensing images. *Pattern Recognition*, 175:113120, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2026.113120. URL https://www.sciencedirect.com/science/article/pii/S003132032600083X.

Dengke Zhang, Fagui Liu, and Quan Tang. Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 24677–24687, 2025.

Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024. ISSN 1558-0644. doi: 10.1109/tgrs.2024.3449154. URL http://dx.doi.org/10.1109/TGRS.2024.3449154.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pp. 696–712. Springer, 2022.

Qi Zhu, Jiangwei Lao, Deyi Ji, Junwei Luo, Kang Wu, Yingying Zhang, Lixiang Ru, Jian Wang, Jingdong Chen, Ming Yang, Dong Liu, and Feng Zhao. Skysense-o: Towards open-world remote sensing interpretation with vision-centric visual-language modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14733–14744, June 2025.