S^{2}PROMPT: SPARSE SEMANTIC-TO-STRUCTURAL PROMPTING FOR SEMI-SUPERVISED REMOTE SENSING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

CLIP provides transferable class semantics for low-label remote-sensing segmentation, whereas DINOv3 preserves local image structure. However, some existing dual-student training methods couple these complementary priors mainly through confidence-filtered pseudo-labels, leaving soft class uncertainty and its structure-aware propagation underused. We introduce S^{2}Prompt, a sparse semantic-to-structural prompting framework that uses the full CLIP class distribution as a semantic seed and a local top-k graph derived from detached DINOv3 features as structural transport. S^{2}Prompt propagates each class independently, retains the original seed through residual mixing, and injects the raw and refined priors into the DINO decoder through a zero-initialized, gradient-isolated prompt. Both branches remain deployable, with a fixed validation-guided rule selecting the stronger candidate. Across 12 low-label protocols on WHDLD, LoveDA, and Potsdam, the selected predictor achieves the highest validation mIoU among all compared methods. Under fully matched settings, it outperforms the strong baseline by up to 2.9 points, while prompting improves the DINO branch by 14.4–36.4 mIoU. Controlled experiments on three designated protocols show that the raw CLIP prior drives most of this recovery. Feature-derived DINO affinity provides a further average gain of 2.2 mIoU over a parameter-matched random local graph.

1 INTRODUCTION

Remote-sensing semantic segmentation supports land-cover mapping, urban monitoring, disaster assessment, and environmental observation. Its dense labels are nevertheless expensive: large scenes contain small objects, fine boundaries, repeated textures, and severe class imbalance, while the appearance of one class can change with sensor, spatial resolution, season, and geography (Gong et al., 2026). Remote-sensing foundation models reduce the representation gap through domain-specific masked pre-training and adapted vision transformers (Sun et al., 2023; Wang et al., 2023), but downstream segmentation still needs pixel-wise supervision. Semi-supervised methods therefore combine limited labels with abundant unlabelled imagery through perturbation consistency (Tarvainen & Valpola, 2017; Sohn et al., 2020), cross-model pseudo-label exchange (Chen et al., 2021), or explicit use of uncertain pixels (Wang et al., 2022b).

These methods depend on the quality of their pseudo-labels. In remote-sensing images, a high-confidence prediction can still omit parts of an object or disagree with its boundaries, and limited labelled data may not correct these errors during training. Reliability-aware methods use prediction stability, paste consistency, confidence reweighting, or consistency self-training (Yang et al., 2022; Wang et al., 2022a; Li et al., 2022; Huang et al., 2024). These mechanisms filter or reweight supervision according to confidence or consistency, but they do not specify how class probabilities should be propagated between spatially related tokens.

CLIP, DINO, and SAM provide different information for this problem. CLIP supplies transferable category semantics from image–text supervision (Radford et al., 2021); self-supervised DINO features encode object structure without class labels (Caron et al., 2021; Siméoni et al., 2025); and SAM

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.

provides class-agnostic masks and boundaries (Kirillov et al., 2023). Recent systems combine these priors, reconstruct CLIP patch correlations, or improve its spatial detail (Shi et al., 2025; Zhang et al., 2025; Sun et al., 2025). CLIP predictions can be semantically correct but spatially coarse, whereas DINO features can delineate regions without assigning classes. Existing combinations mainly address open-vocabulary inference or representation refinement, not low-label dual-student training.

Co2S jointly trains CLIP and DINOv3 as heterogeneous students using weak-to-strong consistency and confidence-based mutual teaching (Zhou et al., 2026). Its branch interaction occurs at the loss level and does not use DINO similarities to propagate the class probabilities produced by CLIP. A dense global affinity matrix would require $O(N^2)$ storage and could link distant instances with similar textures. We instead use a local, one-way transfer to limit such connections and avoid coupling the two branches through a trainable bridge. The transfer direction does not determine which branch is selected for inference.

Local affinity has been used to expand sparse semantic evidence in weakly supervised segmentation (Ahn & Kwak, 2018; Ru et al., 2022) and to suppress background noise around small objects in aerial imagery (Li et al., 2021). Simple-ViLMedSAM uses visual affinity to refine CLIP-derived positional cues before prompt-based medical segmentation (Qian et al., 2026). We adapt this idea by propagating each CLIP class channel on a local DINO graph before prompt injection. The injection is initialized at zero so that it does not change the decoder output at the start of training.

Figure 1 shows S^{2}Prompt. Its *class-wise affinity prior* (CAP) starts from the full CLIP class distribution and propagates each channel over local top-k neighbours defined by detached DINOv3 features. Residual mixing retains the unpropagated probabilities. A shared 1×1 projection maps the raw and refined probabilities to a class-wise prompt in the DINO decoder. Both the CLIP probabilities and DINO affinities are detached from this path, and the projection is initialized at zero. The first prompted forward pass is therefore identical to the unprompted baseline. During prompting, CLIP provides class probabilities and DINO provides local neighbourhood structure. This division does not fix the deployed predictor: validation selects either standalone CLIP or the joint CAP-DINO path for each protocol.

Our contributions are threefold:

- We propose class-wise sparse semantic–structural propagation, which retains CLIP's full predictive distribution and transports each class independently over a local top-k DINO graph. This design preserves soft class uncertainty without materializing a dense $N\times N$ affinity matrix.
- We develop directionally isolated residual prompting, which injects the raw and refined semantic priors through a class-shared, zero-initialized projection. Stop-gradient boundaries keep the transfer one-way, preserve the unprompted baseline at initialization, and require no additional loss.
- Relative to a strong architecture-matched baseline, S^{2}Prompt's selected output improves by up to 2.9 mIoU and its DINO branch by up to 36.4 points. Controlled ablations attribute 23.2 of the 25.7-point mean DINO recovery to the raw CLIP prior and a further 2.2 points to DINO affinity.

2 Related Work

Semi-supervised semantic segmentation. Semi-supervised segmentation primarily exploits unlabelled images through self-training and consistency regularization: Mean Teacher and FixMatch stabilize or threshold teacher targets (Tarvainen & Valpola, 2017; Sohn et al., 2020), while CPS, U2PL, ST++, and UniMatch control confirmation bias through dual networks, negative evidence, sample selection, or complementary perturbations (Chen et al., 2021; Wang et al., 2022b; Yang et al., 2022; 2023). Remote-sensing extensions address multiscale scenes, imbalanced classes, and fine boundaries through paste consistency, confidence-aware self-training, weak-to-strong consistency, class-relative weighting, or multiscale cross-teacher attention (Wang et al., 2022a; Li et al., 2022; Lu et al., 2023; Huang et al., 2024; Wang et al., 2025). These methods mainly improve training targets or losses; S^{2}Prompt instead retains the existing objectives of the dual-branch architecture and adds an explicit structure-aware path inside the decoder.

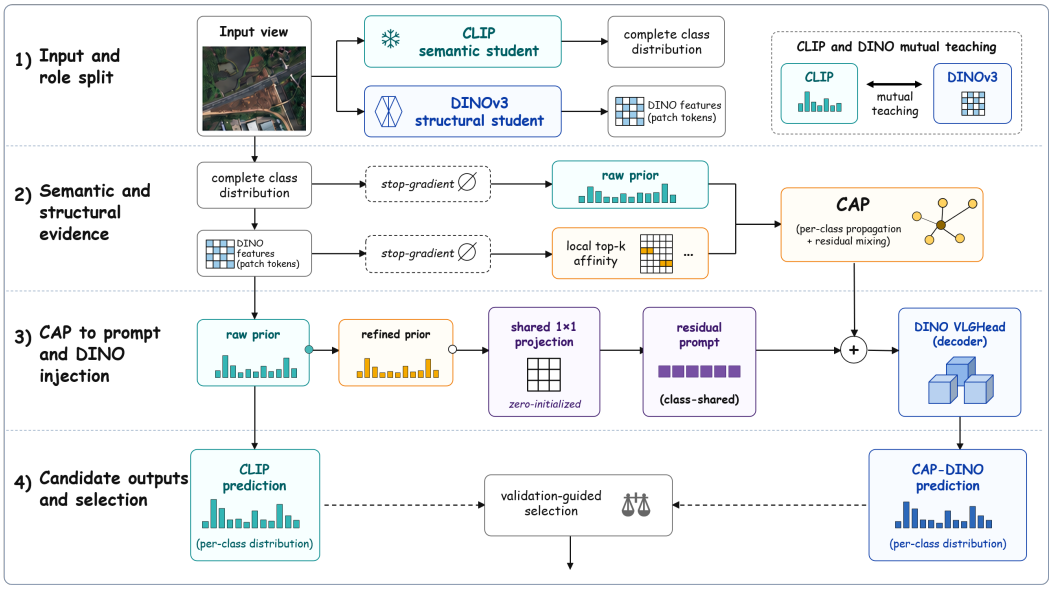

Figure 1: Overview of S^{2}Prompt. A shared input view is processed by a CLIP semantic student and a DINOv3 structural student. The complete CLIP class distribution and local DINO token affinity are detached and aligned to form a raw semantic prior and a sparse structural transport rule. CAP propagates each class independently over the local top-k graph and mixes the result residually with the raw prior. The raw and refined priors are jointly mapped by a shared zero-initialized 1×1 projection to a residual prompt in the DINO VLGHead, while gradients through the prompting path stop at both source branches. CLIP and CAP-DINO remain separate candidate outputs; a validation-guided rule selects the stronger complete branch for deployment.

Foundation models for dense and remote-sensing prediction. CLIP supplies language-aligned semantics but weak pixel localization (Radford et al., 2021); DINO offers dense object structure (Caron et al., 2021; Siméoni et al., 2025), and SAM provides promptable, class-agnostic masks (Kirillov et al., 2023). Remote-sensing foundation models such as RingMo and RVSA adapt pre-training or attention to overhead imagery, while CrossEarth targets cross-condition generalization (Sun et al., 2023; Wang et al., 2023; Gong et al., 2026).

Existing compositions mainly modify CLIP for open-vocabulary inference: Trident, CorrCLIP, and CLIPer combine heterogeneous foundation cues (Shi et al., 2025; Zhang et al., 2025; Sun et al., 2025); ProxyCLIP replaces CLIP's final attention with global VFM correspondences without training (Lan et al., 2024); and CLIP-DINOiser distils DINO patch correlations into lightweight CLIP-side heads (Wysoczańska et al., 2024). More closely, Simple-ViLMedSAM refines a BiomedCLIP attribution map with SAM affinity and projects the raw and refined maps into a bidirectional decoder (Qian et al., 2026). Co2S jointly trains CLIP and DINOv3 students (Zhou et al., 2026). In this low-label setting, S^{2}Prompt propagates every CLIP class channel on a local top-k DINO graph and injects the raw and refined probabilities through a zero-initialized, stop-gradient residual prompt. Unlike methods that modify CLIP features or fuse both branches, this path conditions the trainable DINO student with detached CLIP probabilities.

Affinity refinement and structural propagation. AffinityNet learns local pixel relations and expands discriminative regions with random walks (Ahn & Kwak, 2018). AFA derives semantic affinity from transformer attention to refine weak labels (Ru et al., 2022), while PointFlow shows that sparse point-wise propagation can suppress background noise in aerial segmentation (Li et al., 2021). These methods use affinity to expand labels or refine features. In S^{2}Prompt, detached DINOv3 similarities define local neighbours over which each channel of the CLIP distribution is propagated before DINO decoding.

3 METHOD

Let $\mathcal{D}_l = \{(x_m, y_m)\}_{m=1}^{n_l}$ and $\mathcal{D}_u = \{u_m\}_{m=1}^{n_u}$ denote the labelled and unlabelled sets, respectively, where $n_l \ll n_u$ and every pixel belongs to one of C classes. Our starting point is Co2S, which jointly trains a CLIP semantic student f_g and a DINOv3 structural student f_l with supervised, pseudo-label, feature-perturbation, and stability losses (Zhou et al., 2026). S^{2}Prompt leaves these objectives unchanged and modifies only the DINO decoder forward path. It consists of two components: (i) class-wise sparse semantic–structural propagation, which uses local DINO affinity to refine a detached CLIP distribution; and (ii) directionally isolated residual prompt injection, which conditions the DINO decoder on both the original and refined distributions.

We first establish notation shared by both components. Consider a batch of one geometrically aligned image view $I \in \mathbb{R}^{B\times 3\times H_0\times W_0}$. CLIP produces logits $Z_g \in \mathbb{R}^{B\times C\times H_g\times W_g}$, and DINOv3 produces its deepest feature map $F_l \in \mathbb{R}^{B\times D\times H_t\times W_t}$. Let $\Omega = \{1,\dots,H_t\}\times\{1,\dots,W_t\}$ be the DINO token grid and $N = |\Omega| = H_tW_t$. We use $b \in \{1,\dots,B\}$, $c \in \{1,\dots,C\}$, and $p, q \in \Omega$ for batch, class, and token indices. The c-th CLIP channel and the c-th DINO class query must refer to the same semantic class. Equations below operate independently for each b; we retain b where it prevents ambiguity. The operator $\mathrm{sg}(\cdot)$ is the identity in the forward pass and has zero derivative in the backward pass.

3.1 CLASS-WISE SPARSE SEMANTIC–STRUCTURAL PROPAGATION

Semantic seed. We retain the complete CLIP class distribution rather than converting it to a hard label. Applying softmax along the class dimension gives

$$P_g[b,c,r] = \frac{\exp Z_g[b,c,r]}{\sum_{c'=1}^{C}\exp Z_g[b,c',r]}, \qquad r \in \{1,\dots,H_g\}\times\{1,\dots,W_g\}. \tag{1}$$

We detach P_g, bilinearly resize each class channel to Ω, and renormalize across classes:

$$\begin{aligned} Q &= \mathrm{Resize}_{H_t,W_t}(\mathrm{sg}(P_g))\,, \\ \widetilde{P}_g[b,c,p] &= \frac{Q[b,c,p]}{\max\Big(\sum_{c'=1}^{C} Q[b,c',p], \epsilon_{\mathrm{prob}}\Big)}, \qquad \epsilon_{\mathrm{prob}} = 10^{-6}. \end{aligned} \tag{2}$$

Both tensors have shape $B\times C\times H_t\times W_t$. Thus $\widetilde{P}_g$ is aligned with the DINO grid and satisfies $\sum_{c=1}^{C}\widetilde{P}_g[b,c,p] = 1$ up to numerical precision.

Sparse structural affinity. For each $p \in \Omega$, let $\mathcal{W}_w(p)$ denote the $w\times w$ window centred at p, including p itself, let $\mathcal{V}_w(p) = \mathcal{W}_w(p)\cap\Omega$ be its valid in-image candidates, and let $K = w^2$. We normalize the detached DINO vector at every token with an ϵ_f-stabilized ℓ_2 operation,

$$\widehat{f}_{b,p} = \frac{\mathrm{sg}(F_l[b,:,p])}{\max(\|\,\mathrm{sg}(F_l[b,:,p])\|_2, \epsilon_f)} \in \mathbb{R}^D, \qquad \epsilon_f = 10^{-12}, \tag{3}$$

and define the temperature-scaled normalized inner-product score (equal to cosine similarity for nonzero feature vectors)

$$s_{b,pq} = \widehat{f}_{b,p}^{\top}\widehat{f}_{b,q}/\tau, \qquad q \in \mathcal{V}_w(p), \quad \tau > 0. \tag{4}$$

Let $k_p = \min(k, |\mathcal{V}_w(p)|)$ and let $\mathcal{T}_k(b,p)$ be the indices of the k_p largest scores in $\mathcal{V}_w(p)$. This index-set definition, rather than a set of score values, makes the subsequent gathering operation explicit. The sparse aggregation weight assigned by target token p to source token q is

$$A_l[b,p,q] = \begin{cases} \dfrac{\exp(s_{b,pq})}{\sum_{q'\in\mathcal{T}_k(b,p)}\exp(s_{b,pq'})}, & q \in \mathcal{T}_k(b,p), \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

Because the centre token is always valid, the denominator in Equation 5 is nonzero, and $\sum_{q\in\Omega} A_l[b,p,q] = 1$. In code, every window is represented with K slots: out-of-image slots receive score $-\infty$ before top-k and therefore zero softmax weight. This padded implementation is equivalent to the valid-set definition above. Conceptually, $A_l \in \mathbb{R}^{B\times N\times N}$ has at most k nonzero entries per row, but this dense tensor is never materialized.

Class-wise propagation. The same structural graph transports each semantic channel independently:

$$P_{\text{prop}}[b, c, p] = \sum_{q \in \mathcal{T}_k(b,p)} A_l[b, p, q]\, \widetilde{P}_g[b, c, q]. \tag{6}$$

The affinity therefore controls *where* evidence moves, while $\widetilde{P}_g$ determines *which* class the evidence supports. We retain the original seed through a convex residual update,

$$P_g^{\text{ref}} = \rho \widetilde{P}_g + (1 - \rho) P_{\text{prop}}, \qquad 0 \le \rho \le 1. \tag{7}$$

Both P_{prop} and P_g^{ref} belong to $\mathbb{R}^{B \times C \times H_t \times W_t}$. Because both terms in Equation 7 are class distributions, their convex combination is also a class distribution in exact arithmetic. The implementation renormalizes it across c to suppress floating-point drift. We use $w = 5$, $k = 8$, $\tau = 0.1$, and $\rho = 0.5$ by default, with the valid domain $w \in \{1, 3, 5, \ldots\}$, $1 \le k \le w^2$, $\tau > 0$, and $\rho \in [0, 1]$.

Sparse implementation and complexity. Let $K = w^2$ denote the number of candidates in each local window. CAP examines NK candidate pairs and retains Nk edges, instead of evaluating the N^2 pairs of a global graph. Spatial shifts compute these similarities without materializing a dense $N \times N$ matrix, and the retained indices and weights have shape $B \times N \times k$. The current probability-gathering step temporarily materializes $B \times N \times K \times C$ values before top-k selection, so peak temporary storage still depends on K.

3.2 Directionally Isolated Residual Prompt Injection

Class-shared projection and residual injection. For every class and token, we preserve both stages of the prior as a two-dimensional vector

$$M[b, c, :, p] = \left[\widetilde{P}_g[b, c, p],\ P_g^{\text{ref}}[b, c, p]\right] \in \mathbb{R}^2. \tag{8}$$

Collecting these vectors gives $M \in \mathbb{R}^{B \times C \times 2 \times H_t \times W_t}$. Let D_p be the channel width of the DINO VLGHead. A pointwise 1×1 projection $\phi_\theta(m) = W_\theta m + a_\theta$, with $W_\theta \in \mathbb{R}^{D_p \times 2}$ and $a_\theta \in \mathbb{R}^{D_p}$, is shared over all batch elements, classes, and spatial positions:

$$F_p[b, :, c, p] = \phi_\theta(M[b, c, :, p])\,. \tag{9}$$

Consequently, $F_p \in \mathbb{R}^{B \times D_p \times C \times H_t \times W_t}$. Sharing ϕ_θ prevents class-specific parameter growth while retaining a separate spatial prompt for every class.

The DINO VLGHead produces class-response features $X \in \mathbb{R}^{B \times D_p \times C \times H_t \times W_t}$ after its spatial reasoning blocks. Before semantic reasoning, we inject

$$X' = X + F_p, \qquad X' \in \mathbb{R}^{B \times D_p \times C \times H_t \times W_t}. \tag{10}$$

Both W_θ and a_θ are initialized to zero. Hence $F_p = 0$ and $X' = X$ at initialization, making the initial DINO forward pass identical to the unprompted baseline. The segmentation objectives subsequently learn θ jointly with the ordinary DINO decoding path.

Directional gradient isolation. The prompt path is intentionally one-way. Equation 2 prevents DINO-side losses from updating CLIP through the prior, while Equation 3 prevents those losses from updating the DINO backbone through affinity construction. These stop-gradient operations apply only to the new prompting path: CLIP still receives gradients from the original objectives of the dual-branch architecture, and DINO still receives gradients through its ordinary encoder–decoder path. Thus S^{2}Prompt introduces neither bidirectional feature fusion nor a new training loss.

View correspondence and flexible deployment. For every training view $I^{(v)}$, we compute $Z_g^{(v)} = f_g(I^{(v)})$ and $P_g^{(v)} = \text{softmax}_C(Z_g^{(v)})$ from that same geometric view. Labelled, weak, strong, and CutMix views therefore have separate aligned priors. Clean and feature-perturbed DINO passes share a prior only when their image geometry is identical. Validation and inference retain the same joint path,

$$I \longrightarrow f_g(I) \longrightarrow P_g \longrightarrow f_l(I; P_g). \tag{11}$$

We retain two deployable candidates: the standalone CLIP output and the joint CLIP→CAP-DINO output. Let $M_g(e)$ and $M_l(e)$ denote their validation mIoU at epoch e. The selected branch and epoch are

$$(b^*, e^*) = \underset{b\in\{g,l\},\, e}{\operatorname{argmax}}\, M_b(e), \tag{12}$$

with ties resolved in favour of CLIP because it avoids the additional joint inference path. If $b^* = l$, deployment loads the CLIP and DINO weights from the same epoch e^* so that the prior and conditioned decoder remain paired; if $b^* = g$, the selected CLIP checkpoint is deployed directly. This rule is applied identically to every protocol. It makes branch flexibility an explicit system property, while the fixed-DINO ablation separately tests whether CAP improves the structural branch under controlled conditions.

4 Experiments

4.1 Experimental Setup

Datasets, protocols, and implementation details. Table 6 summarizes the datasets, low-label protocols, and complete shared implementation, optimization, and evaluation protocol.

Evaluated output. The main comparison reports the CLIP candidate, CAP-DINO candidate, and the validation-selected output from Equation 12. Core ablations instead fix the DINO decoder: A0 is unprompted, whereas A1, A4, and A6 use raw, random-refined, and DINO-refined prompts, respectively. We write P_{rand} for the distribution obtained by propagating $\widetilde{P}_g$ over the parameter-matched random local graph used by A4.

Metric. Validation mIoU is the sole aggregate evaluation metric. Per-class IoUs are reported only as a decomposition of this same validation calculation for the branch analysis. From validation curves, we report the branch-best diagnostic gap

$$\Delta_{\mathrm{branch}} = \max_e M_l(e) - \max_e M_g(e). \tag{13}$$

Learning-dynamics plots instead use the paired per-epoch gap $M_l(e) - M_g(e)$.

4.2 Comparison with State-of-the-Art

We compare against FixMatch (Sohn et al., 2020), U2PL (Wang et al., 2022b), WSCL (Lu et al., 2023), UniMatch (Yang et al., 2023), DWL (Huang et al., 2024), MUCA (Wang et al., 2025), and Co2S (Zhou et al., 2026). Table 1 consolidates the 12 selected protocols. Historical baseline rows are copied from the Co2S comparison tables. We reproduce both Co2S branches with the same data splits, trainable parameter set, training budget, and evaluation protocol as S^{2}Prompt. For each system, the same rule in Equation 12 selects between its CLIP and DINO candidates. This design separates contextual comparison with prior methods from fully matched branch- and system-level comparisons with Co2S.

Overall comparison. S^{2}Prompt achieves the highest validation mIoU on all 12 protocols. Relative to the strongest non-S^{2}Prompt entry in each column, the selected output improves WHDLD by 0.7, 0.9, 0.7, and 1.0 mIoU from $1/24$ to $1/4$ labels; LoveDA by 1.6, 1.3, 0.7, and 1.1; and Potsdam by 1.0, 0.7, 0.7, and 0.4. These margins span 0.4–1.6 mIoU and average 0.9. The matched Co2S selector is the strongest reference in eight columns, and DWL is strongest in the remaining four. The improvements therefore cover all three datasets and all four label ratios.

Fully matched system comparison. With the same candidate set and selection rule, S^{2}Prompt–Selected exceeds Co2S-Selected in all 12 protocols by 0.7–2.9 mIoU, with an average gain of 1.2. The dataset means are 0.8 on WHDLD, 1.3 on LoveDA, and 1.4 on Potsdam. Co2S selects CLIP in every protocol, whereas S^{2}Prompt selects CLIP in nine protocols and CAP-DINO in three LoveDA protocols. The S^{2}Prompt–CLIP candidate also exceeds Co2S-CLIP in every protocol by 0.7–2.9 points, averaging 1.1. Because the new prompting path is detached from CLIP, any change in the CLIP branch must be mediated by the original mutual-teaching objective.

Table 1: Main comparison across three benchmarks and 12 low-label protocols. Values are validation mIoU (%), rounded to one decimal. OnlySup–MUCA are historical results reported in the Co2S comparison tables. All Co2S and S^{2}Prompt rows use the same data splits, trainable parameter set, training budget, and evaluation protocol. "Selected" applies Equation 12 independently within each system; C/D mark the branch chosen from unrounded validation scores.

Method	WHDLD				LoveDA				Potsdam			
	1/24	1/16	1/8	1/4	1/40	1/16	1/8	1/4	1/32	1/16	1/8	1/4
OnlySup	53.6	55.2	58.0	60.5	45.9	53.1	56.6	58.9	61.5	64.9	70.2	73.8
FixMatch	56.0	57.6	59.8	60.8	53.3	57.9	60.1	61.5	69.4	71.5	73.5	75.7
U2PL	57.1	58.3	59.9	61.1	52.3	58.0	59.6	61.5	67.2	71.3	73.8	75.9
WSCL	56.8	58.6	59.8	61.1	54.1	57.4	59.4	61.1	69.9	72.3	73.9	75.4
UniMatch	57.4	58.8	60.4	61.5	55.5	58.6	60.0	61.7	70.7	72.6	74.8	76.4
DWL	56.5	57.0	57.1	58.9	57.4	59.0	62.8	63.4	74.2	78.3	79.8	80.3
MUCA	56.5	58.2	60.0	60.5	54.6	56.7	60.2	63.1	66.9	71.5	76.0	79.2
Co2S-CLIP	60.9	61.1	62.2	62.3	57.8	60.4	62.5	63.9	74.0	76.1	79.9	80.0
Co2S-DINO	44.7	47.4	47.9	48.1	34.4	34.6	37.1	43.4	39.6	42.3	46.8	49.3
Co2S-Selected	60.9	61.1	62.2	62.3	57.8	60.4	62.5	63.9	74.0	76.1	79.9	80.0
S^{2}Prompt–CLIP	61.6	62.0	62.9	63.3	59.4	61.7	63.5	64.9	75.2	79.0	80.6	80.7
S^{2}Prompt–DINO	61.4	61.8	62.6	62.7	59.4	61.7	63.5	65.0	74.9	78.7	80.4	80.4
S^{2}Prompt–Selected	61.6^{C}	62.0^{C}	62.9^{C}	63.3^{C}	59.4^{D}	61.7^{C}	63.5^{D}	65.0^{D}	75.2^{C}	79.0^{C}	80.6^{C}	80.7^{C}

Table 2: Core controlled ablation on designated low-label protocols. Validation mIoU (%) always evaluates the DINO decoder: unprompted for A0 and conditioned for A1/A4/A6. Prompted variants match the two-channel projection and training budget. Values are rounded to one decimal.

Variant	Prompt	Graph	WHDLD 1/24	LoveDA 1/40	Potsdam 1/32	Avg.	Δ vs. A0
Co2S (A0)	None	None	44.7	34.4	39.6	39.6	0.0
Raw prior (A1)	$[\widetilde{P}_g, 0]$	None	59.3	56.9	72.1	62.8	+23.2
Random affinity (A4)	$[\widetilde{P}_g, P_{\mathrm{rand}}]$	Random-k	59.5	57.1	72.4	63.0	+23.4
S^{2}Prompt (A6)	$[\widetilde{P}_g, P_g^{\mathrm{ref}}]$	DINO-k	61.4	59.4	74.9	65.2	+25.7

Fully matched branch comparison. Without the proposed conditioning, Co2S-DINO trails its matched Co2S-CLIP counterpart by 13.7–34.4 mIoU across the 12 protocols (23.8 on average). S^{2}Prompt–DINO exceeds Co2S-DINO by 14.4–36.4 points, averaging 24.7. The dataset means are 15.1 on WHDLD, 25.0 on LoveDA, and 34.1 on Potsdam. It also exceeds the strongest non-S^{2}Prompt entry in every column by 0.1–1.6 points, with a mean margin of 0.7. The DINO backbone is present in both matched rows; the difference is CAP conditioning, whose components are tested by the fixed-DINO controls in Table 2. Figure 2 gives the corresponding branch-level results. Panel (a) reports branch-best scores across all 12 protocols, panel (b) reports the matched CLIP-to-CLIP gains, and panel (c) plots best-so-far validation trajectories for the lowest-label setting of each dataset. At the respective best validation checkpoints, panel (d) shows positive S^{2}Prompt–DINO gains in all 76 class–protocol pairs, ranging from 7.7 to 49.6 mIoU.

Branch complementarity and selection. Consistent with the branch-best comparison in Figure 2(a), within S^{2}Prompt, DINO differs from CLIP by only -0.6 to $+0.1$ mIoU, compared with the 13.7–34.4 point deficit before conditioning. CLIP is selected on all WHDLD and Potsdam protocols and on LoveDA-1/16, whereas DINO is selected on LoveDA-1/40, 1/8, and 1/4. At the displayed precision, the two branches tie on LoveDA-1/40, 1/16, and 1/8, while DINO is ahead by 0.1 point at 1/4; selection is determined from the unrounded scores. Thus, the selector changes the output only for the three LoveDA protocols in which CAP-DINO has the higher unrounded score. The fixed-DINO comparisons in Table 2 measure CAP conditioning independently of branch selection. CAP-improved DINO predictions can still affect CLIP through the existing confidence-based

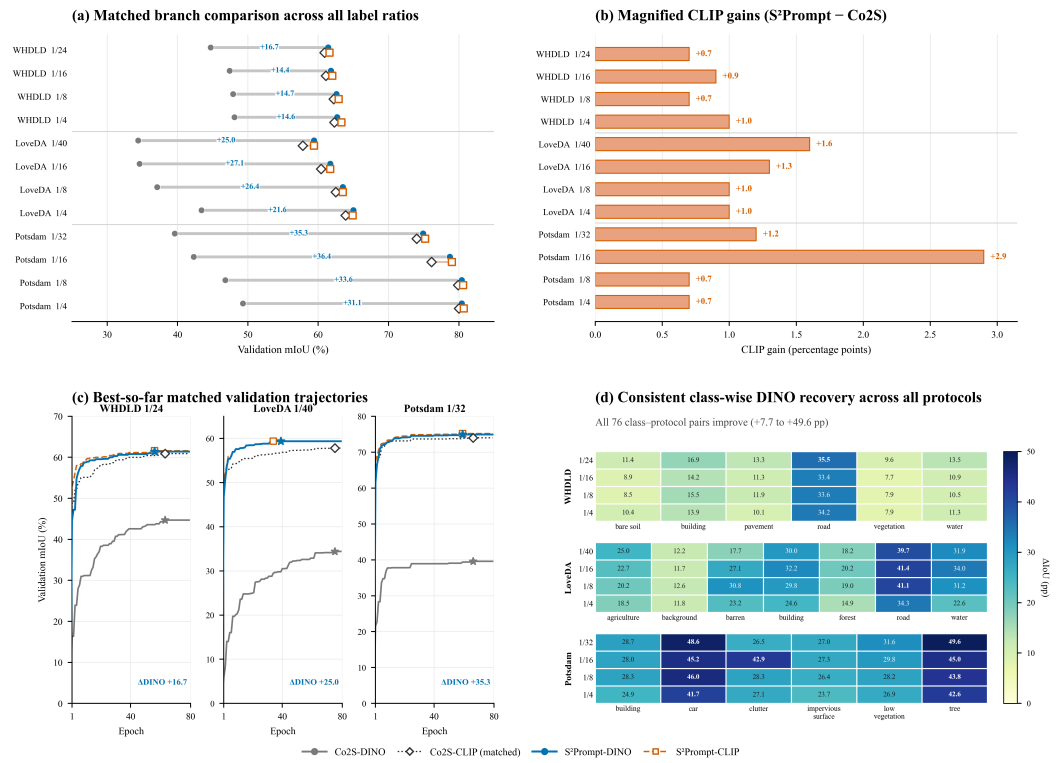

Figure 2: Branch-wise performance across 12 low-label protocols. (a) Best validation mIoU of the two matched Co2S branches and the two S^{2}Prompt branches; connector labels denote the gain from Co2S-DINO to S^{2}Prompt–DINO. (b) Matched gains from Co2S-CLIP to S^{2}Prompt–CLIP. (c) Best-so-far validation trajectories of all four matched branches under the lowest-label protocol of each dataset. (d) Per-class gains from Co2S-DINO to S^{2}Prompt–DINO.

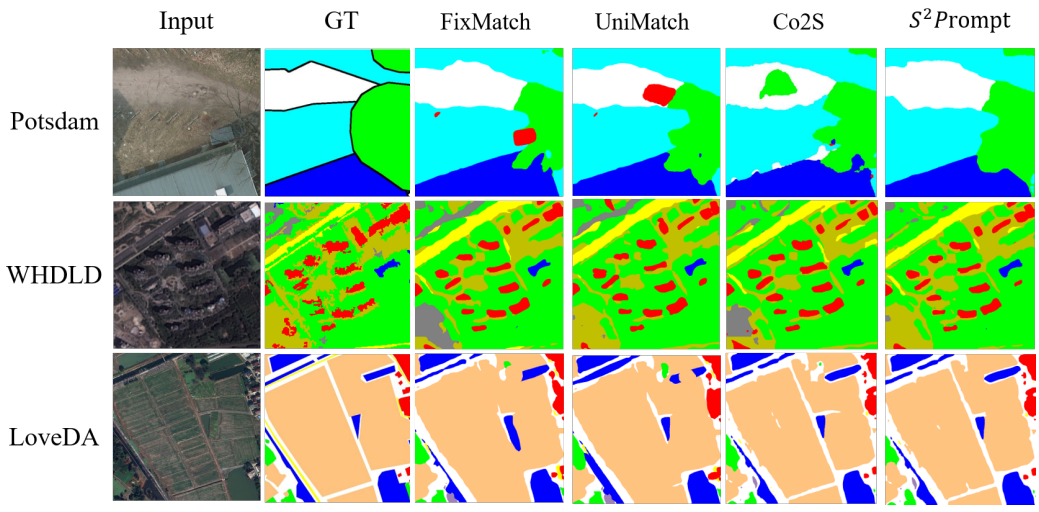

Figure 3: Qualitative comparisons on Potsdam, WHDLD, and LoveDA (top to bottom). Columns show the input image, ground truth, FixMatch, UniMatch, Co2S, and S^{2}Prompt (left to right). Colors follow the dataset-specific class palettes.

mutual-teaching objective, even though the prompting path is detached. We therefore treat the CLIP changes as a joint-training outcome.

Qualitative comparison. Figure 3 highlights three complementary error patterns. On Potsdam, FixMatch and UniMatch introduce conspicuous red false-positive islands, while Co2S produces an additional green region inside the large white area and small artifacts near class boundaries. S^{2}Prompt suppresses these errors and recovers the dominant white, cyan, green, and blue regions with boundaries that more closely follow the ground truth. On WHDLD, it retains more of the elongated red objects and thin yellow structures while avoiding the large gray intrusion visible in Co2S, resulting in a cleaner organization of densely interleaved classes. On LoveDA, FixMatch and UniMatch fragment narrow white regions and distort several blue structures; S^{2}Prompt reconstructs the main diagonal and vertical white paths and preserves the elongated blue objects with fewer breaks. These examples complement the aggregate mIoU results: the improvement is expressed not only as larger correctly labelled regions, but also as fewer isolated class errors and better preservation of class-aligned geometry.

4.3 Ablation Studies

The core ablation uses one designated low-label protocol per dataset. To isolate semantic prompting from structural propagation, all prompted variants use the same two-channel projection and training budget. The raw-prior control uses $[\tilde{P}_g, 0]$. The random-affinity control samples k neighbours without replacement from the same valid local window, uses the same padding mask and temperature, and holds the sampled neighbour indices fixed within each run. Therefore, the parameter-matched random graph isolates DINO-derived affinity from parameter count and generic local smoothing.

Controlled decomposition. The matched A6–A0 comparison improves the DINO decoder by 16.7, 25.0, and 35.3 mIoU on WHDLD-1/24, LoveDA-1/40, and Potsdam-1/32, respectively, for a three-dataset mean gain of 25.7. A1 shows that injecting the raw CLIP prior already contributes 14.6, 22.5, and 32.5 points on these protocols (23.2 on average). Replacing the structural channel with a parameter-matched random local graph changes this result by only +0.2 on average (A4 versus A1), whereas the DINO-affinity graph provides a further +1.9, +2.3, and +2.5 points on the three datasets (A6 versus A4). These controls show that the semantic prior is the dominant source of recovery, but that the additional improvement requires genuine local DINO affinity rather than arbitrary local smoothing.

5 Conclusion

We introduced S^{2}Prompt to make complementary CLIP semantics and DINOv3 structure interact explicitly in low-label remote-sensing segmentation. Rather than exchanging only confidence-filtered pseudo-labels, S^{2}Prompt transports the full, detached CLIP class distribution over a sparse local DINO graph and injects both the raw and refined priors into the DINO decoder through a zero-initialized, gradient-isolated residual prompt. Across 12 protocols, the selected predictor achieves the highest validation mIoU among all compared methods. Under fully matched settings, it improves over Co2S by 0.7–2.9 mIoU, while prompting raises the DINO branch by 14.4–36.4 points and reduces its gap to CLIP from 13.7–34.4 points to between -0.6 and $+0.1$. Fixed-DINO ablations further show that the raw semantic prior contributes 23.2 of the 25.7-point mean gain on three designated protocols, whereas learned DINO affinity adds 2.2 points over a parameter-matched random graph. These results support a specific conclusion: most of the recovery comes from exposing the structural student to soft class semantics, while local feature affinity provides a smaller but measurable gain by controlling where that evidence moves. The approach remains limited by the quality of the CLIP seed and by spurious links across repeated textures; selecting CAP-DINO also requires paired CLIP and DINO inference.

Reproducibility Statement

Section 3 specifies the information path, equations, gradient boundaries, and default hyperparameters. Section 4 reports the datasets and splits, optimization settings, checkpoint-selection rule, evaluation metrics, and controlled comparisons. The appendix provides focused design ablations, branch diagnostics, hyperparameter sensitivity, and the full dataset-specific implementation settings.

AI USE STATEMENT

Generative AI tools were used only for grammar correction and wording refinement. They were not used to generate research ideas, methods, analyses, results, or citations. The authors reviewed all AI-assisted edits and take full responsibility for the final content of this work.

REFERENCES

Jiwoon Ahn and Suha Kwak. Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4981–4990, 2018. doi: 10.1109/CVPR.2018.00523.

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. doi: 10.1109/ICCV48922.2021.00951.

Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2613–2622, 2021. doi: 10.1109/CVPR46437.2021.00264.

Ziyang Gong, Zhixiang Wei, Di Wang, Xiaoxing Hu, Xianzheng Ma, Hongruixuan Chen, Yuru Jia, Yupeng Deng, Zhenming Ji, Xiangwei Zhu, Xue Yang, Naoto Yokoya, Jing Zhang, Bo Du, Junchi Yan, and Liangpei Zhang. Crossearth: Geospatial vision foundation model for domain generalizable remote sensing semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(5):5147–5164, 2026. doi: 10.1109/TPAMI.2025.3649001.

Wei Huang, Yilei Shi, Zhitong Xiong, and Xiao Xiang Zhu. Decouple and weight semi-supervised semantic segmentation of remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 212:13–26, 2024. doi: 10.1016/j.isprsjprs.2024.04.010.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023. doi: 10.1109/ICCV51070.2023.00371.

Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. ProxyCLIP: Proxy attention improves CLIP for open-vocabulary segmentation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Jiahao Li, Bin Sun, Shutao Li, and Xudong Kang. Semisupervised semantic segmentation of remote sensing images with consistency self-training. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–11, 2022. doi: 10.1109/TGRS.2021.3134277.

Xiangtai Li, Hao He, Xia Li, Duo Li, Guangliang Cheng, Jianping Shi, Lubin Weng, Yunhai Tong, and Zhouchen Lin. Pointflow: Flowing semantics through points for aerial image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4217–4226, 2021. doi: 10.1109/CVPR46437.2021.00420.

Xiaoqiang Lu, Licheng Jiao, Lingling Li, Fang Liu, Xu Liu, Shuyuan Yang, Zhixi Feng, and Puhua Chen. Weak-to-strong consistency learning for semisupervised image segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2023.3272552.

Chengcan Qian, Dong Nie, Geng Chen, Daoqiang Zhang, and Xuyun Wen. Simple-ViLMedSAM: Simple text prompts meet vision-language models for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 30042–30052, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Franz Rottensteiner, Gunho Sohn, Jaewook Jung, Markus Gerke, Caroline Baillard, Sandrine Benitez, and Uwe Breitkopf. The ISPRS benchmark on urban object classification and 3d building reconstruction. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, I-3:293–298, 2012. doi: 10.5194/isprsannals-I-3-293-2012.

Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16846–16855, 2022. doi: 10.1109/CVPR52688.2022.01634.

Zhenfeng Shao, Weixun Zhou, Xueqing Deng, Maoding Zhang, and Qimin Cheng. Multilabel remote sensing image retrieval based on fully convolutional network. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:318–328, 2020. doi: 10.1109/JSTARS.2019.2961634.

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. doi: 10.1109/ICCV51701.2025.02180.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, 2025. URL https://arxiv.org/abs/2508.10104.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A. Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. FixMatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Lin Sun, Jiale Cao, Jin Xie, Xiaoheng Jiang, and Yanwei Pang. Cliper: Hierarchically improving spatial representation of clip for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23199–23209, 2025. doi: 10.1109/ICCV51701.2025.02154.

Xian Sun, Peijin Wang, Wanxuan Lu, Zicong Zhu, Xiaonan Lu, Qibin He, Junxi Li, Xuee Rong, Zhujun Yang, Hao Chang, Qinglin He, Guang Yang, Ruiping Wang, Jiwen Lu, and Kun Fu. RingMo: A remote sensing foundation model with masked image modeling. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–22, 2023. doi: 10.1109/TGRS.2022.3194732.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Di Wang, Qiming Zhang, Yufei Xu, Jing Zhang, Bo Du, Dacheng Tao, and Liangpei Zhang. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023. doi: 10.1109/TGRS.2022.3222818.

Jia-Xin Wang, Si-Bao Chen, Chris H. Q. Ding, Jin Tang, and Bin Luo. Ranpaste: Paste consistency and pseudo label for semisupervised remote sensing image semantic segmentation. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–16, 2022a. doi: 10.1109/TGRS.2021.3102026.

Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. LoveDA: A remote sensing land-cover dataset for domain adaptive semantic segmentation. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, pp. 1–16, 2021.

Shanwen Wang, Xin Sun, Changrui Chen, Danfeng Hong, and Jungong Han. Semi-supervised semantic segmentation for remote sensing images via multi-scale uncertainty consistency and cross-teacher-student attention. *IEEE Transactions on Geoscience and Remote Sensing*, 63:1–15, 2025. doi: 10.1109/TGRS.2025.3585489.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4248–4257, 2022b. doi: 10.1109/CVPR52688.2022.00421.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *Proceedings of the European Conference on Computer Vision*, 2024.

Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. ST++: Make self-training work better for semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4268–4277, 2022. doi: 10.1109/CVPR52688.2022.00423.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7236–7246, 2023. doi: 10.1109/CVPR52729.2023.00699.

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. doi: 10.1109/ICCV51701.2025.02288.

Yi Zhou, Xuechao Zou, Shun Zhang, Kai Li, Shiying Wang, Jingming Chen, Congyan Lang, Tengfei Cao, Pin Tao, and Yuanchun Shi. Toward stable semi-supervised remote sensing segmentation via co-guidance and co-fusion. *IEEE Transactions on Geoscience and Remote Sensing*, 64:1–14, 2026. doi: 10.1109/TGRS.2026.3680315.