GLSC: GLOBAL–LOCAL SEMANTIC–STRUCTURAL COLLABORATION FOR SEMI-SUPERVISED REMOTE SENSING SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models and self-supervised vision foundation models provide complementary priors for semi-supervised remote-sensing segmentation: CLIP offers category-level semantic knowledge, whereas DINOv3 preserves fine-grained spatial structures. However, effectively coordinating these heterogeneous priors at both category and spatial levels remains challenging. We propose GLSC, a global–local semantic–structural collaboration framework that strengthens the interaction between CLIP and DINOv3 through two complementary modules. First, Shared-Query Response Fusion (SQRF) uses CLIP text prototypes and learnable DINOv3 queries to project the two visual streams into category-conditioned response spaces and fuses their responses while preserving an explicit category dimension. Second, Dual-Context Response Refinement (DCRR) refines the fused responses by modeling intra-category spatial dependencies under visual-feature guidance and inter-category relationships under query guidance. The refined responses are subsequently aggregated and reintegrated into the original visual representations through residual enhancement. During training, an asymmetric auxiliary guidance loss encourages DINOv3 to learn from CLIP, while gradients from this auxiliary pathway are stopped at the CLIP branch. Experiments on Potsdam, LoveDA, and WHDLD under four labeling ratios show that GLSC outperforms the reported Co2S results in all 12 dataset–ratio settings, achieving an average absolute improvement of 1.27 mIoU points and a maximum gain of 2.11 points on LoveDA with $1/8$ labeled data.

1 INTRODUCTION

Semi-supervised semantic segmentation reduces the dependence on dense pixel-level annotations by combining a small labeled set with a larger pool of unlabeled images. A dominant approach is consistency regularization, ranging from teacher-generated targets and feature perturbations to confidence-filtered weak-to-strong supervision (Tarvainen & Valpola, 2017; Ouali et al., 2020; Sohn et al., 2020; Yang et al., 2023). This setting is particularly relevant to remote sensing, where dense annotation is costly and models must distinguish semantically similar land-cover categories while preserving boundaries, elongated regions, and small objects.

Pretrained foundation models provide complementary priors for these requirements. CLIP learns transferable category semantics through image–text alignment (Radford et al., 2021), and subsequent dense-prediction studies show that its representations can produce pixel–text scores and class-conditioned responses (Zhou et al., 2022; Rao et al., 2022; Xu et al., 2022; Cho et al., 2024). DINOv3, by contrast, inherits the localized visual representations of self-supervised ViTs and provides structure-sensitive features for dense prediction (Caron et al., 2021; Amir et al., 2022; Siméoni et al., 2025). Co2S combines these priors through a heterogeneous CLIP–DINOv3 dual-student framework, explicit and implicit category guidance, and confidence-driven mutual supervision (Zhou et al., 2025). However, its cross-branch coordination is mainly performed through final predictions, leaving the category-conditioned intermediate representations of the two branches independently formed. Directly aligning their native features is also undesirable because the encoders are pretrained with different objectives. The resulting question is how to establish a common category-aware interface for semantic–structural interaction without forcing native-feature alignment.

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.

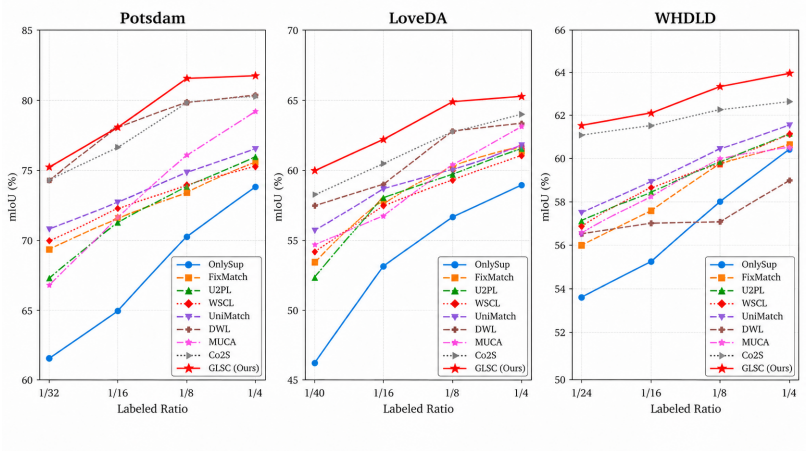

Figure 1: Performance across labeled-data ratios on Potsdam, LoveDA, and WHDLD. The curves use the values in Table 1; exact numerical comparisons are reported in Section 4.

We propose **Global–Local Semantic–Structural Collaboration (GLSC)** to organize this interaction in a shared category-response space. First, **Shared-Query Response Fusion (SQRF)** combines CLIP text prototypes and learnable DINOv3 queries to construct shared category queries. These queries project both visual streams into comparable category-conditioned responses, which are fused while preserving the category dimension. Second, **Dual-Context Response Refinement (DCRR)** progressively refines the fused responses. Visual-Guided Spatial Refinement (VSR) introduces cross-branch visual context to model intra-category spatial dependencies, whereas Semantic-Guided Category Refinement (SCR) reuses the shared queries to capture inter-category relationships. The spatial-to-category order first stabilizes local structures and then resolves semantic ambiguity. The refined responses are aggregated and reintegrated into the branch features through residual enhancement.

During training, **Asymmetric Collaborative Guidance (ACG)** uses reliable CLIP predictions as stop-gradient semantic targets for DINOv3 and suppresses the reverse direction to avoid contaminating the semantic anchor. ACG is an auxiliary objective rather than an additional architectural module. GLSC retains the original semi-supervised objectives and uses the collaborative pathway only during training. At evaluation, predictions are produced solely by the native CLIP branch, without prediction fusion or post-hoc branch selection. Figure 1 summarizes the results across four labeled-data ratios on three benchmarks.

Our contributions are summarized as follows. **(1)** We introduce GLSC, which coordinates heterogeneous foundation-model priors through category-conditioned responses instead of direct native-feature alignment. **(2)** We develop SQRF and DCRR to establish shared category responses and sequentially refine their spatial structures and category relationships, together with ACG for controlled CLIP-to-DINOv3 guidance. **(3)** Across 12 dataset–ratio settings, GLSC numerically exceeds the published Co2S results by an average of 1.27 mIoU points while retaining single-branch CLIP inference without additional test-time computation.

2 Related Work

Semi-supervised semantic segmentation. Modern semi-supervised segmentation leverages unlabeled data primarily through pseudo-labeling and consistency regularization, spanning Mean Teacher parameter ensembling (Tarvainen & Valpola, 2017), multi-decoder feature perturbation (Ouali et al., 2020), confidence-filtered weak-to-strong supervision (Sohn et al., 2020), cross-pseudo supervision (Chen et al., 2021), and contrastive learning from unreliable labels (Wang et al., 2022). In remote sensing, recent frameworks adapt these paradigms to address complex land-cover distributions, boundary ambiguity, and scale variations (Sun et al., 2020; Lu et al., 2023; Huang et al., 2024; Wang et al., 2025). However, standard dual-branch architectures implicitly rely on homogeneous encoders or shared feature spaces. When incorporating heterogeneous foundation models, late-stage prediction swapping (Zhou et al., 2025) leaves intermediate representations completely decoupled, whereas direct native-feature alignment risks destroying pretrained feature manifolds. In contrast, GLSC

introduces a category-conditioned response space, enabling explicit semantic–structural interaction without forcing element-wise alignment between incompatible native visual representations.

Vision-language priors for dense prediction. Vision-language models like CLIP (Radford et al., 2021) capture open-vocabulary semantic concepts through image–text alignment. To adapt these global priors to dense prediction, prior works construct pixel-level visual–text responses (Zhou et al., 2022; Rao et al., 2022), mine emergent semantic groupings (Xu et al., 2022), or aggregate similarity scores via cost volumes (Cho et al., 2024). While demonstrating strong category reasoning, these frameworks operate primarily within a single vision-language stream, where fine-grained spatial boundaries remain sub-optimal due to noisy text matching and the absence of structural visual priors. GLSC repurposes CLIP text prototypes not merely as static classifiers, but as part of a shared category-query interface. This interface projects both the vision-language stream and a self-supervised visual branch into a unified response space for joint context refinement.

Self-supervised visual priors and heterogeneous collaboration. Self-supervised vision transformers, such as DINO (Caron et al., 2021; Amir et al., 2022) and DINOv3 (Siméoni et al., 2025), provide fine-grained, structure-sensitive features that complement vision-language models. Although recent efforts explore dual-student learning between CLIP and DINOv3 (Zhou et al., 2025), their cross-branch collaboration is largely restricted to prediction probabilities. GLSC fundamentally restructures this collaboration mechanism: (1) Shared-Query Response Fusion (SQRF) projects incompatible native feature spaces into a comparable category-response space via shared category queries; (2) Dual-Context Response Refinement (DCRR) explicitly decouples spatial boundary stabilization from inter-category semantic reasoning in a progressive order; and (3) Asymmetric Collaborative Guidance (ACG) enforces unidirectional, gradient-stopped semantic supervision (CLIP $\rightarrow$ DINOv3) to systematically prevent structural noise from corrupting the text-anchored semantic space.

3 METHOD

3.1 GLSC OVERVIEW AND PROBLEM FORMULATION

We consider labeled data $\mathcal{D}_l = \{(\mathbf{x}_l, \mathbf{y}_l)\}$ and unlabeled data $\mathcal{D}_u = \{\mathbf{x}_u\}$ under weak-to-strong semi-supervised segmentation, with batch size B and K semantic categories. As illustrated in Figure 2, GLSC uses CLIP as the semantic anchor and evaluation branch, while DINOv3 provides a structure-sensitive auxiliary stream during training. The two heterogeneous encoders interact through category-conditioned responses rather than through an explicit alignment loss between their original feature spaces.

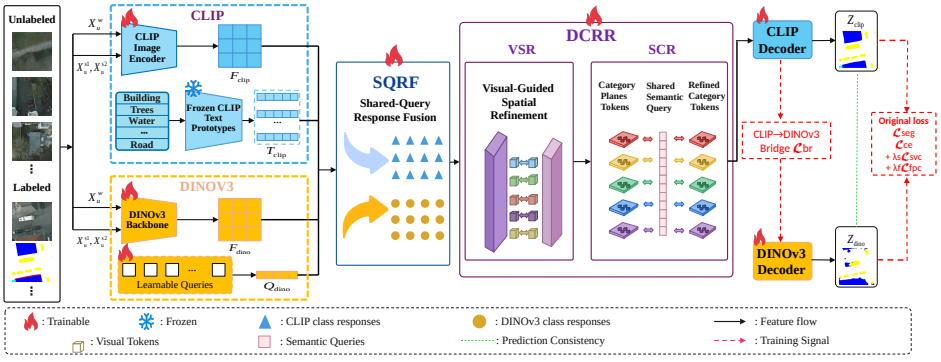

Figure 2: GLSC training framework. CLIP provides stop-gradient semantic guidance and remains the sole evaluation branch. SQRF, DCRR, and DINOv3 serve as training auxiliaries. Although the general two-branch interface is illustrated, CLIP-side collaborative supervision is disabled in the reported configuration.

For an image $\mathbf{x}$, the two encoders produce branch-specific visual features. These features are projected to a common interaction width $D = 256$, and the DINOv3 feature map is bilinearly aligned to the

CLIP spatial resolution:

$$\begin{aligned} \mathbf{F}_c &= E_c(\mathbf{x}), \quad \widetilde{\mathbf{F}}_c = \psi_c(\mathbf{F}_c), \\ \mathbf{F}_d &= E_d(\mathbf{x}), \quad \widetilde{\mathbf{F}}_d = \mathrm{Resize}_{H_c,W_c}\left(\psi_d(\mathbf{F}_d)\right), \end{aligned} \tag{1}$$

where $\widetilde{\mathbf{F}}_c, \widetilde{\mathbf{F}}_d \in \mathbb{R}^{B\times D\times H_c\times W_c}$.

CLIP features entering the collaborative pathway are detached from its gradient computation. Consequently, the GLSC-specific objectives optimize the DINOv3 branch and collaborative modules, whereas the native CLIP branch continues to be optimized by its original semi-supervised objective.

GLSC contains two architectural components. Shared-Query Response Fusion (SQRF) establishes a category-conditioned interface between the two branches. Dual-Context Response Refinement (DCRR) subsequently refines the fused responses through Visual-Guided Spatial Refinement (VSR) and Semantic-Guided Category Refinement (SCR). During training, Asymmetric Collaborative Guidance (ACG) provides an additional CLIP-to-DINOv3 semantic objective. ACG is an optimization mechanism rather than a third architectural module.

3.2 Shared-Query Response Fusion and Dual-Context Refinement

Figure 3 presents the response-fusion and sequential-refinement pathway.

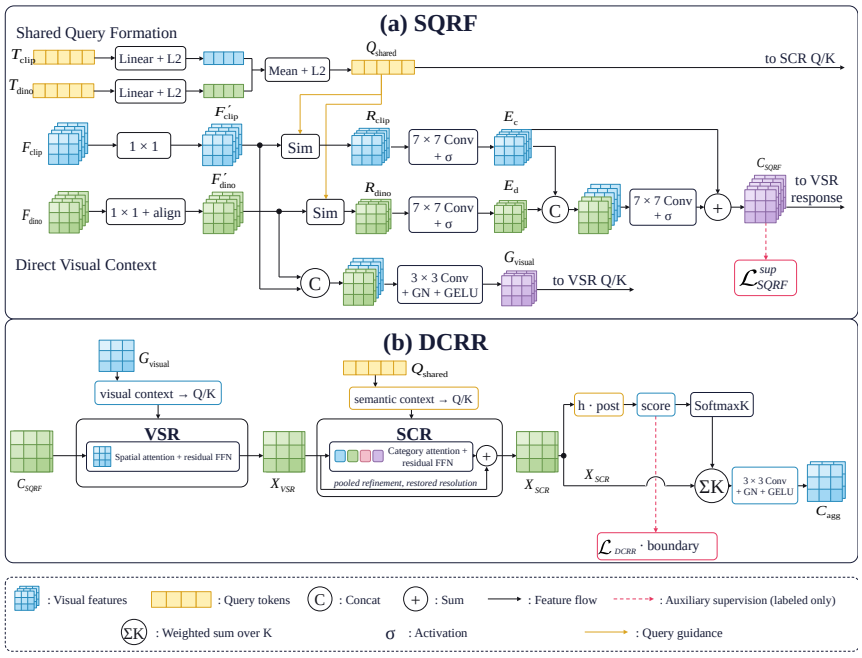

Figure 3: SQRF and DCRR. (a) Shared queries generate category-conditioned responses for residual fusion. (b) VSR and SCR sequentially refine spatial and category contexts before category-weighted aggregation. In the reported configuration, the aggregated collaborative response is supervised only through the DINOv3-side auxiliary pathway; dashed red arrows denote labeled auxiliary supervision.

Shared-Query Response Fusion. The frozen CLIP text encoder provides category prototypes $\mathbf{T}_c$, while the DINOv3 branch maintains learnable category queries $\mathbf{Q}_d$. Independent linear projections followed by L_2 normalization produce the shared category-query interface:

$$\begin{aligned} \mathbf{T}'_c &= \mathrm{Norm}\left(\psi_c^q(\mathbf{T}_c)\right), \\ \mathbf{Q}'_d &= \mathrm{Norm}\left(\psi_d^q(\mathbf{Q}_d^B)\right), \\ \mathbf{Q}_{\mathrm{shared}} &= \mathrm{Norm}\left[\frac{1}{2}\left(\mathbf{T}'_c + \mathbf{Q}'_d\right)\right], \end{aligned} \tag{2}$$

where $\mathbf{Q}_d^B$ denotes the batch-expanded DINOv3 queries.

For branch $r \in \{c, d\}$, the response of spatial location i to category k is computed using the same shared query. Each response map is then embedded independently using a 7×7 convolution and sigmoid activation:

$$\begin{aligned} \mathbf{R}_r(i,k) &= \left\langle \mathrm{Norm}\left(\widetilde{\mathbf{F}}_r(i)\right), \mathrm{Norm}\left(\mathbf{Q}_{\mathrm{shared}}(k)\right) \right\rangle, \\ \mathbf{E}_r^{(k)} &= \sigma\left(\phi_{7\times7}^r\left(\mathbf{R}_r^{(k)}\right)\right). \end{aligned} \tag{3}$$

Here, $\mathbf{R}_r \in \mathbb{R}^{B\times K\times H_c\times W_c}$ and $\mathbf{E}_r^{(k)} \in \mathbb{R}^{B\times D\times H_c\times W_c}$.

The two branches are concatenated only after their representations have been converted into category-conditioned responses. For each category, SQRF adds a fused response correction to the CLIP response:

$$\begin{aligned} \mathbf{C}_{\mathrm{sqrf}}^{(k)} &= \mathbf{E}_c^{(k)} + \sigma\left(\phi_{7\times7}^f\left[\mathbf{E}_c^{(k)}; \mathbf{E}_d^{(k)}\right]\right), \\ \mathbf{Z}_{\mathrm{sqrf}}^{(k)} &= \phi_{\mathrm{score}}\left(\mathbf{C}_{\mathrm{sqrf}}^{(k)}\right). \end{aligned} \tag{4}$$

The resulting response tensor retains an explicit category dimension. SQRF therefore enables category-preserving interaction without enforcing element-wise alignment between the original CLIP and DINOv3 feature spaces.

Visual-Guided Spatial Refinement. Although SQRF establishes category-level interaction, dense prediction also depends on local continuity and boundary structure. VSR first constructs direct visual context from the projected CLIP and DINOv3 features:

$$\begin{aligned} \mathbf{G}_{\mathrm{visual}} &= \mathrm{GELU}\left[\mathrm{GN}\left(\phi_{3\times3}^v\left[\widetilde{\mathbf{F}}_c; \widetilde{\mathbf{F}}_d\right]\right)\right], \\ \mathbf{C}_{\mathrm{vsr}}^{(k)} &= \mathbf{C}_{\mathrm{sqrf}}^{(k)} + \mathrm{VAttn}\left(\mathbf{C}_{\mathrm{sqrf}}^{(k)}, \mathbf{G}_{\mathrm{visual}}\right). \end{aligned} \tag{5}$$

Here, VAttn denotes category-wise eight-head spatial attention followed by a residual feed-forward network. For each category, the attention query retains the full spatial resolution, whereas keys and values are constructed from $2\times$ spatially pooled response and guidance tokens. This asymmetric resolution reduces the attention cost while preserving full-resolution response updates.

Semantic-Guided Category Refinement. SCR subsequently introduces shared category semantics. A $2\times$ spatial average-pooling operation first reduces the VSR responses. At each pooled spatial location, category attention is performed along the K-category dimension. Its query and key projections combine the response tokens with $\mathbf{Q}_{\mathrm{shared}}$, while its values are derived from the response tokens alone. The complete refinement is summarized as

$$\mathbf{C}_{\mathrm{dcrr}} = \mathbf{C}_{\mathrm{vsr}} + \mathrm{Resize}_{H_c,W_c}\left[\mathrm{CAttn}\left(\mathrm{Pool}_2\left(\mathbf{C}_{\mathrm{vsr}}\right), \mathbf{Q}_{\mathrm{shared}}\right)\right], \tag{6}$$

where CAttn denotes eight-head category attention followed by a residual feed-forward network. VSR and SCR therefore refine the collaborative responses sequentially: VSR first stabilizes intra-category spatial structure, and SCR then models inter-category semantic relationships.

The refined responses are converted into spatially varying category weights and aggregated into a collaborative representation. The general architecture defines residual interfaces for both branches:

$$\begin{aligned} \mathbf{Z}_{\mathrm{dcrr}}^{(k)} &= \phi_{\mathrm{score}}\left(\mathbf{C}_{\mathrm{dcrr}}^{(k)}\right), \\ \mathbf{W}_{\mathrm{cls}} &= \mathrm{Softmax}_{\mathrm{cls}}\left(\mathbf{Z}_{\mathrm{dcrr}}\right), \\ \mathbf{F}_{\mathrm{proxy}}(b,:,i) &= \sum_{k=1}^{K} \mathbf{W}_{\mathrm{cls}}(b,k,i)\mathbf{C}_{\mathrm{dcrr}}(b,:,k,i), \\ \mathbf{C}_{\mathrm{agg}} &= \mathrm{GELU}\left[\mathrm{GN}\left(\phi_{3\times3}^a\left(\mathbf{F}_{\mathrm{proxy}}\right)\right)\right], \\ \Delta_c &= \phi_{1\times1}^c\left(\mathbf{C}_{\mathrm{agg}}\right), \\ \Delta_d &= \phi_{1\times1}^d\left(\mathrm{Resize}_{H_d,W_d}\left(\mathbf{C}_{\mathrm{agg}}\right)\right), \\ \mathbf{F}_r' &= \mathbf{F}_r + \tanh(g_r)\Delta_r, \qquad r \in \{c, d\}. \end{aligned} \tag{7}$$

The general interface initializes $g_c = 0$ and $g_d = 0.05$. In the reported configuration, however, the CLIP-side supervised, unsupervised, and stabilization weights are all set to zero. No GLSC-specific objective depends on $\mathbf{F}'_c$, and the enhanced CLIP output is not used for evaluation. Consequently, only the DINOv3-side residual correction is active during collaborative optimization. The two-branch expression in Equation equation 7 is retained to specify the general architectural interface illustrated in Figure 2.

3.3 Asymmetric Collaborative Guidance and Learning Objective

For view v, the native predictions are $\mathbf{Z}_r^v = D_r(\mathbf{F}_r^v)$ for $r \in \{c, d\}$, while the active refinement-enhanced prediction is $\mathbf{Z}_d^{\mathrm{enh},v} = D_d(\mathbf{F}_d'^v)$. We write $\mathbf{P}_r^v = \mathrm{Softmax}(\mathbf{Z}_r^v)$ and $\mathbf{P}_d^{\mathrm{enh},v} = \mathrm{Softmax}(\mathbf{Z}_d^{\mathrm{enh},v})$.

To express the guidance objectives compactly, we define a normalized masked discrepancy:

$$\mathcal{D}_M(\mathbf{P}, \mathbf{Q}) = \frac{\sum_i M_i \left\|\mathbf{P}_i - \mathbf{Q}_i\right\|_2^2}{K \max\left(\sum_i M_i, 1\right)}. \tag{8}$$

For an unlabeled weak view, let $c_{c,i} = \max_k P_{c,i,k}^w$ and $c_{d,i} = \max_k P_{d,i,k}^w$. Reliable CLIP pixels are selected using $M_{c,i} = \mathbf{1}[c_{c,i} \geq \tau]$, while active CLIP-to-DINOv3 pixels use

$$M_i^{C\to D} = \mathbf{1}[c_{c,i} \geq \tau]\, \mathbf{1}\left[(c_{d,i} < \tau) \vee (c_{c,i} \geq c_{d,i})\right],$$

where $\tau = 0.95$.

ACG combines refinement-level supervision of the enhanced DINOv3 prediction with directional stabilization of its native prediction:

$$\begin{aligned}\mathcal{L}_{\mathrm{ACG}} = r_{\mathrm{ref}}(t)\Big[&0.1\ \mathrm{CE}\left(\mathbf{Z}_d^{\mathrm{enh},l}, \mathbf{Y}_l\right) \\ &+ 0.025\, \mathcal{D}_{M_c}\left(\mathbf{P}_d^{\mathrm{enh},w}, \mathrm{sg}\left(\mathbf{P}_c^w\right)\right)\Big] \\ &+ r_s(t)\, \mathcal{D}_{M^{C\to D}}\left(\mathbf{P}_d^w, \mathrm{sg}\left(\mathbf{P}_c^w\right)\right).\end{aligned} \tag{9}$$

Here, $\mathrm{sg}(\cdot)$ denotes stop-gradient. The reverse DINOv3-to-CLIP direction is assigned zero weight. The refinement warm-up $r_{\mathrm{ref}}(t)$ and stabilization schedule $r_s(t)$ are detailed in Appendix D.

The SQRF score head receives labeled category-response supervision. DCRR additionally receives a class-agnostic boundary objective. For class probabilities $\mathbf{P}$, the local morphological boundary response and the two auxiliary losses are

$$\begin{aligned}\mathcal{B}(\mathbf{P}) &= \max_k \left[\mathrm{MaxPool}_3(\mathbf{P}_k) - \mathrm{MinPool}_3(\mathbf{P}_k)\right], \\ \mathcal{L}_{\mathrm{SQRF}}^{\mathrm{sup}} &= \mathrm{CE}\left(\mathbf{Z}_{\mathrm{sqrf}}, \mathbf{Y}\right), \\ \mathcal{L}_{\mathrm{DCRR}}^{\mathrm{sup}} &= \mathrm{BCE}_\Omega\Big(\mathrm{clip}_{[10^{-4},1-10^{-4}]}[\mathcal{B}\left(\mathrm{Softmax}(\mathbf{Z}_{\mathrm{dcrr}})\right)], \\ &\qquad \mathcal{B}\left(\mathrm{OneHot}_\Omega(\mathbf{Y})\right)\Big),\end{aligned} \tag{10}$$

where $\Omega = \{i : Y_i \neq 255\}$ denotes valid labeled pixels. Ignored pixels are zeroed before boundary extraction and excluded from the BCE mean.

The original Co2S objective remains active for both native branches. For $r \in \{c, d\}$,

$$\mathcal{L}_r^{\mathrm{Co2S}} = \frac{1}{2}\left(\mathcal{L}_r^{\mathrm{sup}} + 0.25\mathcal{L}_r^{s1} + 0.25\mathcal{L}_r^{s2} + 0.5\mathcal{L}_r^{\mathrm{fp}}\right),$$

where $\mathcal{L}_r^{s1}$ and $\mathcal{L}_r^{s2}$ denote weak-to-strong consistency losses for the two strongly augmented views, and $\mathcal{L}_r^{\mathrm{fp}}$ denotes feature-perturbation consistency. The complete training objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{total}} = \mathcal{L}_c^{\mathrm{Co2S}} + \mathcal{L}_d^{\mathrm{Co2S}} + \mathcal{L}_{\mathrm{ACG}} \\ + r_{\mathrm{aux}}(e)\left(0.01\mathcal{L}_{\mathrm{SQRF}}^{\mathrm{sup}} + 0.005\mathcal{L}_{\mathrm{DCRR}}^{\mathrm{sup}}\right).\end{aligned} \tag{11}$$

The auxiliary schedule $r_{\mathrm{aux}}(e)$ starts at epoch 20 and increases linearly from zero to one over 15 epochs. The complete loss schedules and architectural hyperparameters are summarized in Appendix D.

At evaluation, the DINOv3 branch and collaborative pathway are removed, and predictions are produced solely by the native CLIP branch. The reported configuration therefore preserves the original CLIP inference path and introduces no additional test-time computation. Importantly, the GLSC-specific objectives do not directly update the CLIP encoder or decoder; the reported predictions correspond to the native CLIP output obtained under the complete training configuration rather than an enhanced or fused output.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets and protocol. We evaluate GLSC on the ISPRS Potsdam benchmark (ISPRS Working Group III/4, n.d.), LoveDA (Wang et al., 2021), and WHDLD (Shao et al., 2020). We follow the dataset construction, train–validation partitions, and labeled-data splits adopted by Co2S (Zhou et al., 2025). Performance is measured using mean Intersection-over-Union (mIoU, %).

Unless otherwise specified, all GLSC results are produced by the native CLIP branch. DINOv3 and the collaborative pathway are used only during training; no branch selection or prediction fusion is applied at evaluation.

Implementation details. All GLSC configurations are trained for 80 epochs using AdamW with an initial learning rate of 10^{-4}, weight decay of 10^{-2}, and polynomial decay (power 0.9). Learning-rate multipliers are 0.01, 10, and 2 for the visual backbones, segmentation heads, and collaborative pathway, respectively. The CLIP text encoder remains frozen.

Random scaling is sampled from $[0.5, 2.0]$. The crop size is 512×512 for Potsdam and LoveDA and 256×256 for WHDLD. The per-GPU batch size is 1 for Potsdam and LoveDA and 2 for WHDLD. The confidence threshold is set to 0.95. Evaluation follows the native sliding-window protocol with a stride of 426. The category-response width is 256, and eight attention heads are used. SQRF employs a 7×7 kernel, while the spatial-reduction ratio in VSR and the pooling factor in SCR are both set to 2. The detailed schedules of the auxiliary losses are provided in Appendix D.

4.2 COMPARISON WITH EXISTING METHODS

We compare GLSC with FixMatch (Sohn et al., 2020), U2PL (Wang et al., 2022), 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., 2025). The results of these competing methods are published values compiled under the Co2S evaluation protocol; we do not rerun the corresponding methods. Table 1 provides the primary quantitative comparison, and Figure 1 visualizes the same results across increasing labeled-data ratios.

Overall comparison. As shown in Table 1, the reported GLSC results are higher than the published Co2S results in all 12 dataset–ratio settings. The numerical margins range from 0.80 to 1.64 mIoU points on Potsdam, from 1.22 to 2.11 points on LoveDA, and from 0.39 to 1.29 points on WHDLD. Averaged across the 12 settings, the margin over Co2S is 1.27 points, with the largest difference of 2.11 points obtained on LoveDA under the $1/8$ labeled setting.

GLSC obtains the highest reported value in 11 of the 12 settings; DWL is highest on Potsdam at $1/16$. Because the competing methods were not rerun with identical implementations and seeds, these margins are comparisons with published results rather than variance-controlled estimates.

Table 1: Segmentation results (mIoU, %) at different labeled-data ratios. GLSC uses the native CLIP output of the complete training configuration; baselines are published Co2S-protocol values. Bold and underlined denote the best and second-best results, including ties.

Method	Potsdam				LoveDA				WHDLD			
Labeled ratio	1/32	1/16	1/8	1/4	1/40	1/16	1/8	1/4	1/24	1/16	1/8	1/4
OnlySup	61.5	64.9	70.2	73.8	45.9	53.1	56.6	58.9	53.6	55.2	58.0	60.5
FixMatch	69.4	71.5	73.5	75.7	53.3	57.9	60.1	61.5	56.0	57.6	59.8	60.8
U2PL	67.2	71.3	73.8	75.9	52.3	58.0	59.6	61.5	57.1	58.3	59.9	61.1
WSCL	69.9	72.3	73.9	75.4	54.1	57.4	59.4	61.1	56.8	58.6	59.8	61.1
UniMatch	70.7	72.6	74.8	76.4	55.5	58.6	60.0	61.7	57.4	58.8	60.4	61.5
DWL	74.2	**78.3**	79.8	80.3	57.4	59.0	62.8	63.4	56.5	57.0	57.1	58.9
MUCA	66.9	71.5	76.0	79.2	54.6	56.7	60.2	63.1	56.5	58.2	60.0	60.5
Co2S	74.3	76.6	79.8	80.2	58.2	60.4	62.7	64.0	61.1	61.5	62.2	62.6
GLSC (Ours)	**75.10**	78.00	**81.44**	**81.62**	**59.87**	**62.03**	**64.81**	**65.22**	**61.49**	**62.03**	**63.33**	**63.89**

Class-specific error analysis. Figure 4 examines selected building, car, and tree crops on Potsdam at 1/8 labels. Relative to Co2S, GLSC reduces building/tree false negatives and car false positives. These maps are qualitative; the dataset-level comparison remains in Table 1.

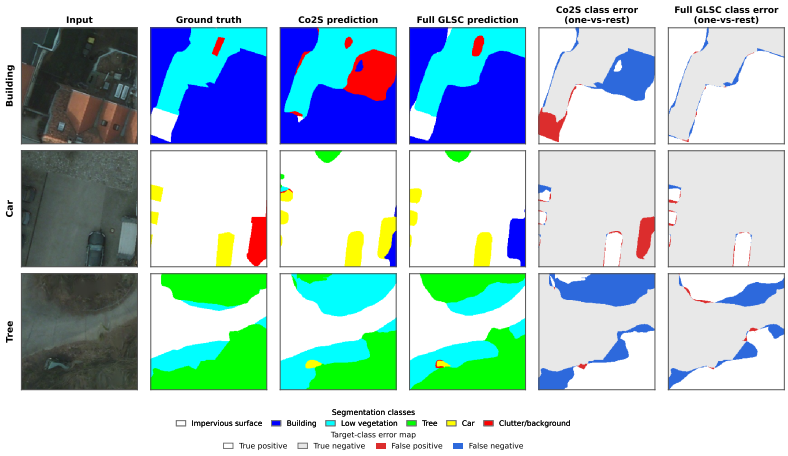

Figure 4: Class-specific errors on Potsdam at 1/8 labels. Rows show building, car, and tree crops. Red/blue denote false positives/negatives, and white/light gray denote true positives/negatives. Full GLSC is the native CLIP output.

Qualitative comparison. Figure 5 provides one case per benchmark. GLSC reduces LoveDA fragmentation, suppresses a spurious Potsdam region, and better preserves the WHDLD target shape; the selected cases are illustrative.

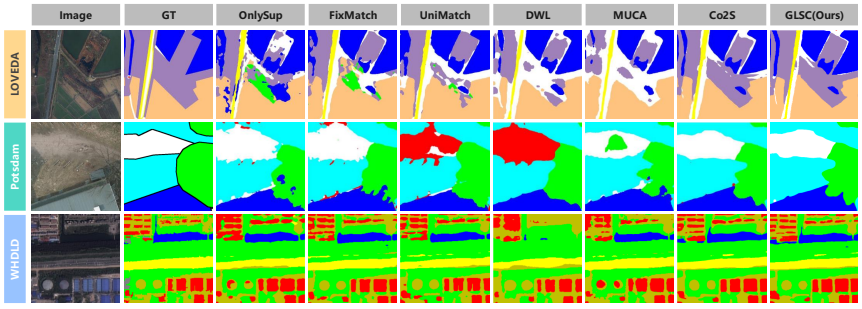

Figure 5: Qualitative comparison on LoveDA, Potsdam, and WHDLD (top to bottom). Columns show the input, ground truth, six baselines, and the full GLSC native CLIP output.

4.3 Configuration and Mechanism Analysis

Configuration-level comparison. Table 2 compares independently trained GLSC configurations on LoveDA at $1/16$, all evaluated through the native CLIP branch. Co2S is included only as a published external reference, so its difference from GLSC Base is not an isolated estimate of ACG's contribution.

Table 2: Configuration-level comparison on LoveDA with $1/16$ labeled data. All GLSC entries correspond to independently trained configurations evaluated using the native CLIP branch. The Co2S entry is a published reference rather than a matched ablation run.

Configuration	ACG	SQRF	DCRR	mIoU (%)
Co2S (published)	×	×	×	60.40
GLSC Base (ACG)	✓	×	×	61.47
GLSC Base + SQRF	✓	✓	×	61.62
GLSC Base + DCRR	✓	×	✓	61.57
GLSC (Full)	✓	✓	✓	**62.03**

SQRF and DCRR individually yield 61.62 and 61.57 mIoU, compared with 61.47 for GLSC Base; their combination reaches 62.03 mIoU. These are single-run, joint-training outcomes rather than variance-controlled component effects. The collaborative pathway uses stop-gradient CLIP references, while the evaluated native CLIP branch remains optimized by the original segmentation objectives.

Internal decomposition of DCRR. Table 3 reports the VSR/SCR decomposition. The individual variants obtain 61.58 and 61.68 mIoU, while their sequential use reaches 62.03 mIoU ($+0.41$ over SQRF). These single-run results support only configuration-level complementarity. Appendix C visualizes the associated response changes as a mechanistic illustration.

Table 3: Configuration-level decomposition of DCRR on LoveDA with $1/16$ labeled data. All entries are independently trained and evaluated using the native CLIP branch.

Configuration	SQRF	VSR	SCR	mIoU (%)
SQRF baseline	✓	×	×	61.62
SQRF + VSR	✓	✓	×	61.58
SQRF + SCR	✓	×	✓	61.68
Full GLSC	✓	✓	✓	**62.03**

5 Conclusion

We presented GLSC, a training-time collaboration framework for semi-supervised remote-sensing segmentation. SQRF maps heterogeneous CLIP and DINOv3 features into a shared category-response space, DCRR refines spatial and inter-category contexts, and ACG supplies one-way, stop-gradient semantic guidance. At evaluation, the DINOv3 and collaborative pathways are removed, retaining only the native CLIP branch without additional test-time computation. Across 12 dataset–ratio settings, GLSC achieves an average numerical margin of 1.27 mIoU points over the published Co2S results and the highest reported value in 11 settings. Configuration comparisons and response visualizations illustrate the observed component behavior, although the single-run differences are not variance-controlled effects. GLSC thus organizes complementary semantic and structural priors during training while preserving single-branch inference.

AI USE STATEMENT

Generative AI tools were used for language polishing, structural organization, and LaTeX drafting. They were not used to generate experimental measurements. The authors verified the technical descriptions, equations, citations, and reported values against the available source code, publications, and experiment records and take responsibility for the manuscript.

REPRODUCIBILITY STATEMENT

The paper specifies the branch roles, tensor operations, gradient directions, active losses, and GLSC-specific schedules. Sec. 4 records the data protocol, optimization settings, crop sizes, batch sizes, and evaluation rule, while Appendix D lists the active collaboration configuration and architectural hyperparameters. All reported primary results use the native CLIP branch; no fused prediction or post-hoc branch selection is used.

REFERENCES

Shir Amir, Yossi Gandelsman, Shai Bagon, and Tali Dekel. Deep ViT features as dense visual descriptors. *ECCV Workshop on What Is Motion For?*, 2022.

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.

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.

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.

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.

ISPRS Working Group III/4. 2d semantic labeling contest: Potsdam. https://www.isprs.org/resources/datasets/benchmarks/UrbanSemLab/2d-sem-label-potsdam.aspx, n.d. Accessed 2026-09-24.

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.

Yassine Ouali, Céline Hudelot, and Myriam Tami. Semi-supervised semantic segmentation with cross-consistency training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12674–12684, 2020.

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*, pp. 8748–8763, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18082–18091, 2022.

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.

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.

Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Xian Sun, Aijun Shi, Hai Huang, and Helmut Mayer. BAS^4Net: Boundary-aware semi-supervised semantic segmentation network for very high resolution remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 13:5398–5413, 2020. doi: 10.1109/JSTARS.2020.3021098.

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.

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*, 2021.

Shanwen Wang, Xin Sun, Changrui Chen, Danfeng Hong, and Jungong Han. Semi-supervised semantic segmentation for remote sensing images via multiscale 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, 2022.

Jiarui Xu, Shalini De Mello, Sifei Liu, Wonmin Byeon, Thomas Breuel, Jan Kautz, and Xiaolong Wang. GroupViT: Semantic segmentation emerges from text supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18134–18144, 2022.

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.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 696–712, 2022. doi: 10.1007/978-3-031-19815-1_40.

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, 2025.

A Class-wise Potsdam Comparison

Figure 6 complements the localized error maps in Figure 4 with full-set class-wise IoU differences. All six categories improve numerically, led by Clutter and Car.

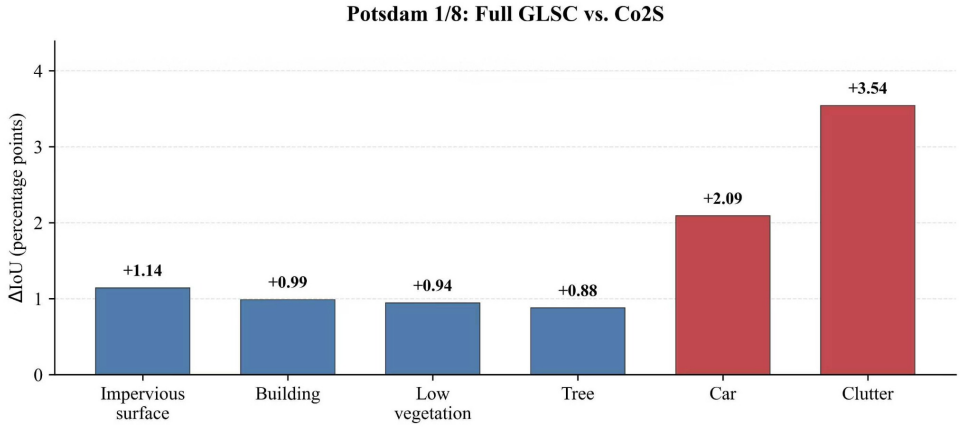

Figure 6: Class-wise IoU differences between Full GLSC and Co2S on Potsdam at $1/8$ labels; positive values favor Full GLSC.

B Class-wise LoveDA Changes

Figure 7 reports class-IoU changes on LoveDA at $1/16$ relative to GLSC Base. Full GLSC gains 0.56 mIoU points, with larger increases for Barren ($+4.24$) and Road ($+0.98$) but decreases in four classes. These single-run patterns are descriptive rather than causal effects.

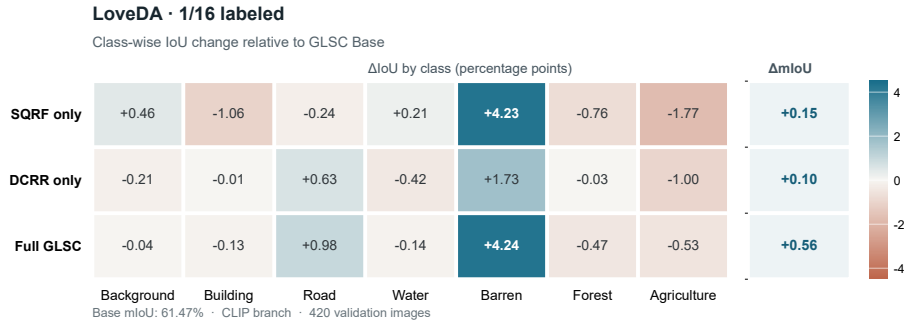

Figure 7: LoveDA $1/16$ class-wise IoU changes from GLSC Base; the right column reports the corresponding mIoU differences in percentage points.

C Mechanism Response Visualization

Figure 8 illustrates a Potsdam road sample. The upper row relates raw CLIP/DINOv3 responses to SQRF and the native CLIP prediction; the lower row magnifies the dashed region before and after VSR and SCR. The responses illustrate localized spatial and class-margin changes, not a fused inference output or independent dataset-level evidence.

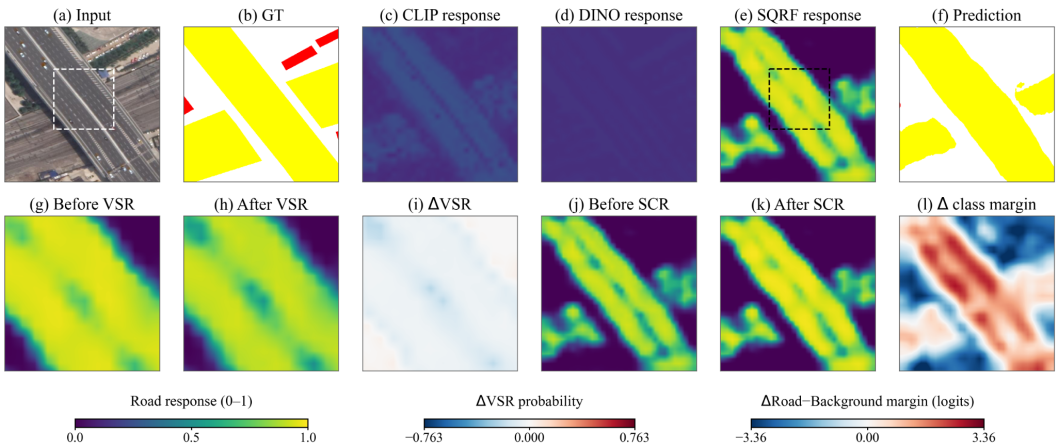

Figure 8: Mechanism visualization on a Potsdam road sample. (a–f) Input, ground truth, raw CLIP/DINOv3 responses, SQRF response, and native CLIP prediction. (g–l) Enlarged responses before/after VSR and SCR, with the VSR probability difference and road–background logit-margin change. The two difference maps use separate color scales.

D Additional Implementation Details

Schedules and active configuration. ACG refinement uses a five-epoch linear warmup with labeled/unlabeled weights $0.1/0.025$. Raw DINOv3 stabilization rises from 0 to 0.5 by cosine scheduling over the first 20% of training, then stays constant; reverse stabilization has zero weight. SQRF/DCRR auxiliary losses start at epoch 20 and ramp linearly to $0.01/0.005$ over 15 epochs. CLIP features are detached, whereas DINOv3 features remain differentiable. Only the enhanced DINOv3 branch receives branch-level collaborative supervision. Evaluation disables the collaborative pathway and uses native CLIP predictions.

Architecture. The interaction width is $D = 256$, with eight attention heads and one DCRR layer. SQRF embedding/fusion kernels are 7×7; VSR spatial reduction and SCR pooling both use factor 2. Attention/projection dropout is zero. Residual gates start at $g_c = 0$ and $g_d = 0.05$.