Leveraging Pretrained Priors for Weakly Supervised Semantic Segmentation of Remote Sensing Images

Xin Li[1], Nicola Genzano[1], Marco Gianinetto[1], Marco Scaioni[*1]

[1] Department of Architecture, Built Environment and Construction Engineering, Politecnico di Milano
via Giuseppe Ponzio 31 Milano, Italy - (xin1.li, nicola.genzano, marco.gianinetto, marco.scaioni)@polimi.it

Keywords: Remote Sensing, Weakly Supervised, Semantic Segmentation, Large Language Model, Pre-trained Model.

Abstract

Semantic segmentation of remote sensing imagery (RSI) is essential for urban mapping, land-use monitoring, and many other domains. However, pixel-level annotation is expensive, making weakly supervised semantic segmentation (WSSS) that relies on image-level labels an attractive alternative. Pre-trained models provide strong priors from large-scale learned representations, making them beneficial for WSSS. However, when kept frozen, they often produce sparse and misaligned class activation maps (CAMs) due to domain gaps and static inference. We propose a lightweight and efficient framework that integrates CLIP and DINO foundation models to address three challenges: (i) semantic misalignment between generic text prompts and RSI-specific visuals; (ii) static CAM quality; and (iii) incomplete object coverage. Our design includes: (1) a *Textual Prototype-Aware Enrichment* (TPE) module that builds an RS-specific knowledge base using large language model (LLM)-generated descriptions to enrich text prompts; (2) a *Unified Semantic Relation Mining* (USR) module that fuses learnable adapter features with CLIP attention and DINO affinity for online CAM refinement; and (3) a *Visual Prototype-Aware Enrichment* (VPE) module, which maintains momentum visual prototypes to complete regions and sharpen boundaries. By freezing the CLIP and DINO backbones and optimizing only lightweight adapter and decoder modules, the proposed framework reduces the number of trainable parameters while achieving competitive performance. Experimental on iSAID and ISPRS Potsdam datasets demonstrate the effectiveness of the proposed framework, achieving 38.01% mIoU on iSAID dataset and 47.01% mIoU with 66.89% overall accuracy on Potsdam dataset.

1. Introduction

Semantic segmentation of remote sensing imagery (RSI), which aims to assign a semantic label to every pixel in the image, supports high-impact applications such as urban mapping (Zhang et al., 2018), land-use monitoring (Maulik and Chakraborty, 2017), precision agriculture (Omia et al., 2023), and environmental management (Andries et al., 2022) among others. However, creating large-scale, pixel-accurate annotations is time-consuming, expensive, and requires expertise in the specific domain, particularly for images characterized by complex spatial structures, large-scale variations, and fine linear features such as infrastructure and rivers. These challenges limit the scalability of fully supervised approaches and motivate image-level Weakly Supervised Semantic Segmentation (WSSS) for RSI, which learns from cheaper image-level labels.

A typical WSSS pipeline starts from the generation of Class Activation Maps (CAMs) (Zhou et al., 2016), refines them into pseudo labels (Ru et al., 2022), and finally trains a segmentation model, as shown in Fig. 1(a). Nevertheless, CAMs often highlight only the most discriminative parts, limiting segmentation completeness. This issue is challenging in RSI, where inter-class similarity is high and intra-class appearance varies due to spatial and contextual factors. Recently, priors from pretrained models have shown great potential for addressing these limitations. Vision–language models such as CLIP (Radford et al., 2021) provide rich semantic priors from large-scale image–text alignment, while self-supervised models such as DINO (Caron et al., 2021) capture strong spatial grouping without annotations. However, directly applying these models to RSI is non-trivial. Both CLIP and DINO are pretrained on natural imagery and generic textual descriptions, resulting in a substantial domain gap when transferred to aerial or satellite imagery. Keeping these pretrained backbones frozen leads to CAMs that cannot adapt or improve during training, whereas fully fine-tuning them is computationally expensive and may affect their generalization capability. Moreover, generic prompts fail to represent the distinctive semantic characteristics of RS scenes. Collectively, these challenges limit the effectiveness of semantic priors in WSSS.

To address these limitations, we propose a framework that strategically composes pretrained models for WSSS of RSI (see Fig. 1(b)). The proposed framework integrates frozen DINO and CLIP backbones with a lightweight feature adapter and decoder. The pretrained backbones remain fixed throughout training, while only the task-specific lightweight modules are optimized. To improve the quality of pseudo labels, we design three complementary modules: (1) the Textual Prototype-Aware Enrichment (TPE) module addresses semantic misalignment by leveraging descriptions generated by the Large Language Model (LLM) to build a comprehensive knowledge base of RS-specific attributes, which are dynamically obtained to enrich prompts and improve image-text alignment; (2) the Unified Semantic Relation Mining (USR) module focuses on the limitations of static CAMs by fusing learnable adapter features with frozen CLIP attention and DINO affinity graphs, online refining the pseudo labels through semantic diffusion; and (3) the Visual Prototype-Aware Enrichment (VPE) module attempts to tackle incomplete coverage by maintaining prototypes and updating them with momentum, producing prototype-aware CAMs that complete object regions and refine boundaries.

By keeping CLIP and DINO frozen while training only a lightweight transformer decoder, our method preserves the benefits of large-scale pretraining while achieving computational efficiency. The training objective combines per-pixel cross-entropy

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-163-2026

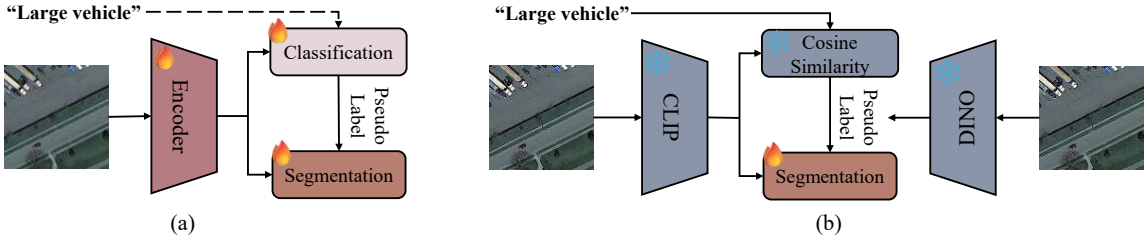


Figure 1. Comparisons between approaches for WSSS of RSI: (a) previous methods typically adopt a trainable encoder followed by classification and segmentation stages, both requiring supervision; (b) our proposed method integrates frozen CLIP and DINO backbones into a single-stage pipeline, where only the lightweight segmentation decoder is trainable.

on refined pseudo labels and binary supervision for spatial affinities. On the other hand, our proposed modules can also create a positive feedback loop: refined pseudo labels provide more accurate supervision to train the decoder, and the trained decoder builds more reliable feature relationships to generate accurate pseudo labels. Finally, we evaluated our method on the iSAID and Potsdam datasets and achieved competitive results.

2. Related Work

2.1 Image-level WSSS for RSI

WSSS for RSI alleviates the dependence on fully annotated datasets by learning from weaker and cheaper supervision. Existing approaches are generally categorized into (1) the image-level supervision, using only category labels per image, and (2) the spatial-level supervision, leveraging additional localization cues (e.g., boxes or points). In this work, we focus on the image-level setting, which is the most widely adopted due to its low annotation cost. Image-level WSSS is challenging because related labels provide no spatial cues, create a large gap between weak supervision and dense pixel-wise prediction. Most methods therefore exploit Class Activation Maps (CAMs) (Zhou et al., 2016) from classification networks as initial seeds for pseudo label generation. However, CAMs typically highlight only the most discriminative regions, yielding incomplete and noisy pseudo masks and limiting the quality of downstream segmentation. Accordingly, a major research line aims to improve CAM quality via multi-scale localization (Zhang et al., 2019), confidence and adversarial-based refinement (Li et al., 2021, Fang et al., 2022), coarse-to-fine label generation (Cao and Huang, 2022), and affinity-based consistency enhancement (Yan et al., 2022, Yan et al., 2023). More recently, multi-class RSI settings have been explored with transformer-based designs and stronger token-level constraints to improve the localization ability (Zhou et al., 2022a, Hu et al., 2024). Beyond CAM refinement, complementary strategies have also been studied, including saliency-based localization (Zeng et al., 2023), affinity and correlation learning (Qiao et al., 2023), and pseudo label filtering or noise-aware refinement (Li et al., 2023, Lu et al., 2024). Despite these advances, producing complete and accurate pseudo masks from image-level labels remains a key challenge.

2.2 Pretrained Priors

Pretrained models provide powerful priors that can substantially help weakly supervised learning. Two categories of pretrained priors are relevant: (i) vision–language models, which enable cross-modal semantic alignment; and (ii) self-supervised models, which provide rich structural representations.

2.2.1 Vision-Language Pretrained Models CLIP (Radford et al., 2021) learns aligned image–text representations and has been explored in WSSS to improve CAM quality. CLIMS (Xie et al., 2022) introduces CLIP-based contrastive guidance, while CLIP-ES (Lin et al., 2023) derives localization cues from CLIP features via GradCAM and attention affinity; both largely rely on hand-crafted text templates. In remote sensing, RSRefSeg (Chen et al., 2025) leverages CLIP-derived referring expressions to activate the Segment Anything Model in a two-stage pipeline; however, a substantial domain gap remains between natural images and RSI, which limits localization robustness.

2.2.2 Self-supervised Pretrained Models Self-supervised learning (SSL) acquires visual representations from unlabeled data via pretext objectives that enforce structural or semantic consistency (Jing and Tian, 2021). In SSL approaches, DINO (Caron et al., 2021) and DINOv2 (Oquab et al., 2024) exhibit emergent class-agnostic grouping at the patch level, a unique property rarely observed in supervised convolutional neural networks or vision transformers. This capability enables DINO to discover salient regions without labels, and prior work has utilized it for unsupervised object discovery via graph construction over patch tokens (Siméoni et al., 2021). Recent studies have exploited self-supervised features for segmentation by combining DINO embeddings with decoders or attention modules. Yet, these features are often class-agnostic and lack adaptation to the complex spatial patterns of RSI. Our work combines vision-language semantics and self-supervised structural priors with class-aware cues from weak supervision, aiming to achieve more complete and accurate segmentation results for RSI.

3. Methodology

3.1 Overview

Our end-to-end framework is illustrated in five steps (see Fig. 2):

(1) **Frozen CLIP features.** The remote sensing image is processed by the frozen CLIP image encoder to extract multi-level visual features, while textual features are obtained through the frozen text encoder.

(2) **Enriched textual semantics through the TPE module.** For each class in RSI, we use a large language model to generate fine-grained descriptions that cover spatial and contextual properties. These descriptions are encoded by CLIP to select and aggregate relevant prototypes, yielding the enriched text representations.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-163-2026 | © Author(s) 2026. CC BY 4.0 License.

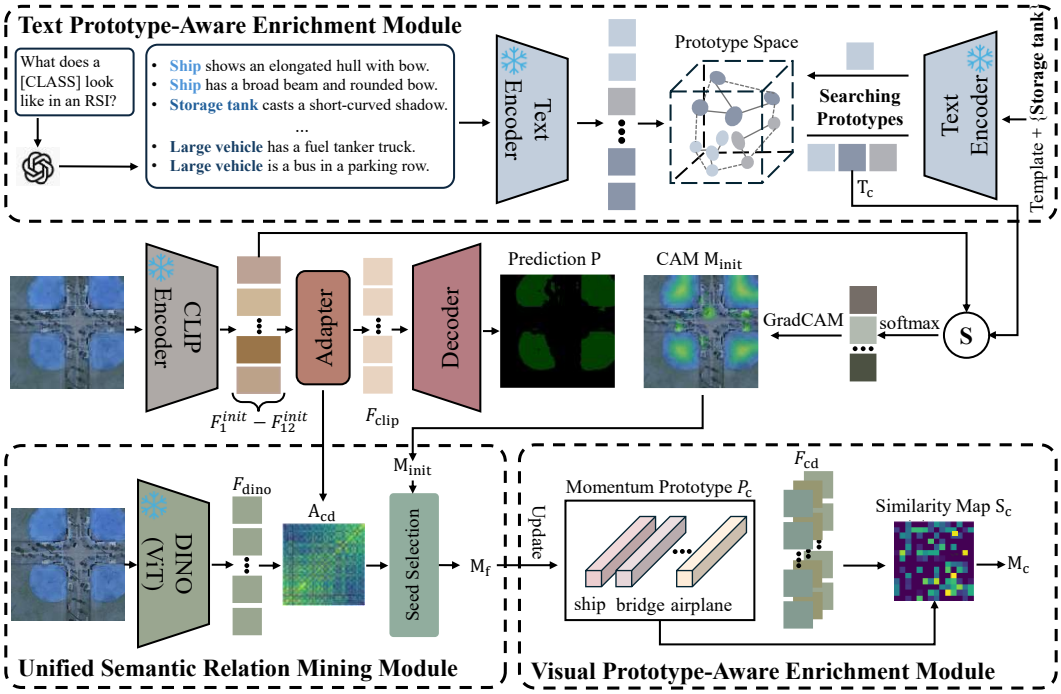


Figure 2. The pipeline of our method.

(3) **Dense and learnable prediction.** Initial CAMs are obtained in the shared image–text space by patch-level cosine similarity between CLIP tokens and enriched text embeddings. In parallel, intermediate CLIP features are fed into a lightweight adapter and decoder to produce dense segmentation logits.

(4) **CAM Refinement with USR and VPE.** The USR module fuses two complementary affinity maps to refine initial CAMs via semantic diffusion. The VPE module maintains momentum-updated visual prototypes aiming to complete object regions and sharpen boundaries.

(5) **Pseudo labels and training.** The refined CAMs are post-processed into pseudo labels for supervision, enabling the adapter–decoder to self-improve during training.

3.2 Textual Prototype-Aware Enrichment Module

RSI presents unique challenges due to the domain gap between natural images (on which CLIP was trained) and nadir imagery. Standard global prompts, such as “an overhead satellite image of [CLASS],” provide limited fine-grained semantic information crucial for pixel-level understanding. To address this limitation, we propose the *Textual Prototype-Aware Enrichment* (TPE) module inspired by DenseCLIP (Rao et al., 2022), which augments global text features with class-specific textual prototypes from a contextual knowledge base.

3.2.1 Prototype-knowledge base Construction We leverage LLM, GPT-4, to generate comprehensive class descriptions that capture the visual characteristics of objects in RSI. For each class c in the label space $c \in y = \{1, 2, \ldots, C\}$ (excluding the background class), we construct the following instruction for GPT-4: *“Generate n detailed descriptions for [CLASS] as it appears in satellite/aerial imagery. Focus on distinguishing visual properties, including spatial patterns, geometric shapes, size variations, and contextual relationships with surrounding land cover types. Each description should follow the format: ’an overhead satellite image of [CLASS]. It exhibits + descriptive context.’”* Here, *[CLASS]* corresponds to the semantic label c. We use GPT-4 in a *text-only* manner: the LLM only takes the class name and the structured instruction prompt as input and does not access any image content. The generated descriptions are fixed and are only used to construct the textual prototype knowledge base. This instruction yields n detailed descriptions for each class, which are encoded using CLIP’s text encoder to form a knowledge base:

$$T = \{\Phi(e_{c,k}) \mid c = 1, \ldots, C;\ k = 1, \ldots, n\}, \qquad (1)$$

where $e_{c,k}$ is the k-th GPT-generated description of the c-th class, $\Phi(\cdot)$ is CLIP’s text encoder, and $n \times C$ is the total number of descriptions. This knowledge base gathers descriptive properties for the entire dataset.

3.2.2 Prototype Clustering Instead of collapsing all class descriptions into a single text vector, we treat prompt construction as a prototype-discovery problem. We cluster the knowledge base into a set of generalized textual prototypes so that each class can map to multiple prototypes, capturing the substantial intra-class diversity in RSI. The clustered prototypes encode shared contextual and attribute-level cues across categories and help compensate for missing information in target-class recognition. We apply K-means clustering (Lloyd, 1982) on the prototype knowledge base to obtain a set of prototype centroids:

$$W = \mathrm{KMeans}(T, B) = a_i{}_{i=1}^{B}, \quad a_i \in \mathbb{R}^d, \qquad (2)$$

where B is the number of centroids. Each centroid a_i serves as a textual prototype that represents a distinct semantic feature shared across classes.

3.2.3 Prototype-Aware Enrichment With the prototype set W, we first send the global template, for example, $g(c) =$ *“an overhead satellite image of [CLASS]”*, into CLIP’s text encoder

to generate a global text embedding $t_c \in \mathbb{R}^d$, where d is the channel dimension. Then t_c is used to query its most relevant prototypes within W. To exclude irrelevant prototypes, we further propose selecting TopK prototype neighbors based on the similarity scores with t_c:

$$W_c = \{a_j : j \in \underset{1\le j\le B}{\arg\,\mathrm{TopK}}\ t_c^\top a_j\}. \tag{3}$$

We aggregate the selected prototypes using similarity-weighted averaging and add the result as complementary knowledge to the base embedding. The enriched text representation for class c is

$$T_c = t_c + \lambda \sum_{j=1}^{K} \mathrm{softmax}\big(t_c^\top W_c\big)_j\, a_j, \tag{4}$$

where λ is a balance factor and $\mathrm{softmax}(\cdot)$ is taken over the K similarity scores $t_c^\top W_c$.

3.3 Feature Adapter and Decoder

Given an RSI $I \in \mathbb{R}^{3\times H\times W}$, where H and W denote the image height and width, respectively, the CLIP ViT image encoder with N transformer blocks produces N intermediate outputs. For the l-th block ($l = 1,\dots,N$), we remove the class token and reshape the remaining patch embeddings into an initial feature map $\{F_{init}^l\}_{l=1}^N$.

3.3.1 Feature adapter To unify the representations across different levels, each F_{init}^l is projected into a common embedding space via a lightweight two-layer MLP (equivalently, two 1×1 convolutions applied per spatial location), producing F_{new}^l:

$$F_{\text{new}}^l = W_{fc}^1\Big(\sigma\Big(W_{fc}^2 F_{\text{init}}^l\Big)\Big), \tag{5}$$

where W_{fc}^1 and W_{fc}^2 are learnable 1×1 projections, and $\sigma(\cdot)$ is a point-wise nonlinearity. This step harmonizes channel dimensions across scales while preserving the spatial resolution, which is particularly important for small or elongated targets. All aligned feature maps are concatenated along the channel dimension and fused through a convolution to produce an aggregated multi-level feature representation:

$$F_{\text{clip}} = \mathrm{Conv}\big(\mathrm{Concat}[F_{\text{new}}^1, F_{\text{new}}^2, \dots, F_{\text{new}}^N]\big), \tag{6}$$

with $F_{\text{clip}} \in \mathbb{R}^{d\times h\times w}$.

3.3.2 Decoder Subsequently, a lightweight decoder consisting of D sequential multi-head transformer blocks refines F_{clip}:

$$\tilde{F} = \varphi(F_{\text{clip}}), \tag{7}$$

where $\varphi(\cdot)$ stacks D blocks, each composed of a multi-head self-attention module and a feed-forward network with residual connections and normalization layers (Dosovitskiy et al., 2021). Finally, class logits are generated by a 1×1 convolution and are upsampled to the original image resolution:

$$P = \mathrm{Upsample}\left(\mathrm{Conv}(\tilde{F})\right), \tag{8}$$

where $P \in \mathbb{R}^{(C+1)\times H\times W}$, $C+1$ is the number of classes (including background), and $\mathrm{Upsample}(\cdot)$ is bilinear interpolation. This feature adapter and decoder pipeline aggregates complementary semantics from different depths of the CLIP encoder and transforms frozen CLIP representations into dense, learnable, and multi-level features. This design enhances the model's robustness to scale variations and object diversity.

3.4 Initial CAM Generation

Unlike conventional WSSS approaches that rely on classification networks to generate CAMs, we directly use CLIP's shared image–text embedding space. Let the enriched text embeddings be $T = [T_1,\dots,T_C] \in \mathbb{R}^{d\times C}$, we extract visual features $F_{init}^N \in \mathbb{R}^{d\times h\times w}$ from the final transformer block of the image encoder, with h and w representing spatial sizes. For simplicity, we denote it as F, and reshape it into a token matrix $\bar{F} \in \mathbb{R}^{d\times hw}$. For each class c, $\hat{T}_c = \frac{T_c}{\|T_c\|}$, and $\hat{F} = \frac{\bar{F}}{\|\bar{F}\|}$. We compute the initial CAM by patch-wise cosine similarity followed by per-class min–max normalization:

$$M_{\text{init}}^c = \mathrm{Norm}\big(\hat{F}^\top \hat{T}_c\big). \tag{9}$$

where $\mathrm{Norm}(\cdot)$ linearly rescales each class map to $[0,1]$ using its spatial min and max (computed over hw locations). The resulting activation maps $\{M_{\text{init}}^c\}_{c=1}^C$ are stacked with a background channel to form the initial CAM $M_{\text{init}} \in \mathbb{R}^{(C+1)\times hw}$.

3.5 CAM Refinement

To provide supervision for the prediction P in Eq. 8, we generate pixel-wise pseudo labels from the initial CAMs produced by the frozen backbone. Since the backbone remains fixed, these CAMs are static and cannot be improved during training. Noisy pseudo labels may therefore misguide the optimization process. To address this issue, we introduce two online CAM refinement modules: the Unified Semantic Relation Mining (USR) module and the Visual Prototype-Aware Enrichment (VPE) module to dynamically update CAMs and then refine the CAMs to improve the quality of pseudo labels. Note that CLIP's image–text alignment is learned mainly at the global image level; thus, its initial CAMs can lack fine-grained localization and rich textual detail. With the CLIP encoder frozen, these limitations would remain if left uncorrected. By refining CAMs online via USR and VPE, we mitigate these issues and adapt the supervision to the task-specific spatial distribution in RSI.

3.5.1 Unified Semantic Relation Mining Module We aim to refine the initial CAM M_{init} by exploiting feature relationships. However, the attention maps produced by the CLIP image encoder cannot be directly adopted for this purpose, since they remain unchanged during training. In contrast, the feature adapter is continuously updated, and we therefore use its evolving features to construct feature relationships that help select informative attention values from the CLIP image encoder, preserving beneficial prior knowledge while suppressing noisy relationships. Furthermore, we introduce attention maps from DINO as complementary self-supervised priors, providing additional structural cues and enhancing the robustness of the unified relationship modeling.

With more reliable feature relationships, the CAM quality can be dynamically enhanced. For the RSI I, we first extract the initial DINO features $F_{\text{init}}^{\text{dino}}$. These features are processed through a feature adapter (see Sect. 3.3) with its own dimensional setting to obtain two refined feature maps F_{dino}. To integrate DINO and CLIP representations, we concatenate their adapted feature maps along the channel dimension:

$$F_{cd} = \mathrm{Concat}(F_{\text{clip}}, F_{\text{dino}}), \tag{10}$$

where $F_{cd} \in \mathbb{R}^{2d\times h\times w}$. The combined feature tensor F_{cd} is then reshaped into token sequences to compute a joint affinity map, capturing both vision–language semantics and self-supervised structural cues:

$$A_{cd} = \sigma\big(F_{cd}^{\top} F_{cd}\big), \tag{11}$$

where $\sigma(\cdot)$ denotes the element-wise sigmoid function. This cross-source affinity matrix serves as a unified relational representation that bridges CLIP's semantic alignment with DINO's spatial consistency. We extract all multi-head attention maps from the frozen CLIP image encoder, denoted as $\{A_s^l\}_{l=1}^{N}$, each attention map is $A_s^l \in \mathbb{R}^{hw\times hw}$. For each A_s^l, we use A_{cd} to evaluate its quality as follows:

$$S_l = \sum_{i=1}^{hw}\sum_{j=1}^{hw} \big|\, A_{cd}(i,j) - A_s^l(i,j) \,\big|. \tag{12}$$

The score S_l measures the deviation of each attention map from the reference. We then compute a binary filter G_l for each attention map according to:

$$G_l = \begin{cases} 1, & \text{if } S_l < \dfrac{1}{N-N_0+1}\sum_{l=N_0}^{N} S_l, \\ 0, & \text{otherwise,} \end{cases} \tag{13}$$

where $G_l \in \mathbb{R}^{1\times 1}$ is expanded to $G_l^e \in \mathbb{R}^{hw\times hw}$ for subsequent computation. The average of all S_l values serves as the threshold: if S_l is below this value, the corresponding attention map is regarded as reliable ($G_l = 1$); otherwise, it is filtered out. This procedure preserves high-quality attention maps while discarding weak or noisy ones. A_{cd} serves as a learnable reference affinity, as it is computed from the continuously updated adapter features (and DINO cues). We measure the discrepancy S_l between each frozen CLIP attention map A_s^l and A_{cd}; a smaller S_l indicates that A_s^l is more consistent with the current task-specific relations. We therefore keep only the attention maps with S_l below the mean discrepancy (Eq. 13), and discard the remaining ones as noisy or mismatched. Here N_0 denotes the starting layer index for computing the mean threshold (we use deeper layers as they encode more semantic relations). Next, we integrate the selected attention maps with the reference affinity map A_{cd} to construct the refining map:

$$R = A_{cd} \cdot \frac{1}{N_m}\sum_{l=N_0}^{N} G_l^e A_s^l, \tag{14}$$

where N_m is the number of valid attention maps, *i.e.*, $N_m = \sum_{l=N_0}^{N} G_l$. Following previous works (Ru et al., 2022), we generate the refined CAM as:

$$M_f^c = \left(\frac{R_{\text{nor}} + R_{\text{nor}}^{\top}}{2}\right)^{\alpha} \cdot M_{\text{init}}^c, \tag{15}$$

where c denotes the specific class, $M_{\text{init}}^c \in \mathbb{R}^{1\times hw}$ is the initial CAM reshaped into vector form, and R_{nor} is obtained from R via row and column normalization. The hyperparameter α controls the diffusion strength. From Eq. 15, the symmetrized normalized affinity is raised to α to control the diffusion strength (larger α yields stronger propagation along reliable relations), producing a refined CAM M_f with improved spatial coherence.

3.5.2 Visual Prototype-Aware Enrichment Module After generating the refined CAM M_f, it can be utilized as semantic guidance for WSSS, similar to CAM. However, M_f is susceptible to erroneous initial CAM seed selection owing to the over-activation issue, i.e., misidentifying background patches as target objects. This situation leads to imprecise propagation on the class-agnostic affinity graph and produces noise, which, in turn, degrades the final segmentation performance. To tackle this issue, we introduce the *momentum prototype* for each class to calibrate M_f. The motivation is that the momentum prototypes, updated by tokens across the entire dataset, can serve as more robust class-representative tokens compared to those from individual image instances. Consequently, we compute the similarity between image tokens and these prototypes to suppress unreliable regions. Also, the CAM values can be regarded as confidence scores for selecting reliable initial seeds corresponding to their labeled classes. However, some regions are simultaneously activated across multiple classes, leading to ambiguous or erroneous high-confidence activations that may produce incorrect seeds for later stages.

Seed selection. For CAM M_f, we first create a class-wise mask

$$I_{\text{cam}} = \arg\max_{c}(M_f), \tag{16}$$

which ensures that each spatial location is exclusively assigned to a single class during seed selection. Then, the CAM seeds for class c are defined as the top-$k\%$ of nonzero activations in the corresponding channel:

$$Q_c = \text{Top}_k\big(M_f \odot \mathbb{I}(I_{\text{cam}} = c)\big), \tag{17}$$

where Q_c denotes the set of positions selected as seeds for class c, and $\odot$ represents the Hadamard product. In this selection process, the hyperparameter k controls the proportion of patches retained as seeds. A smaller k corresponds to a higher confidence in the selected initial seeds.

Prototype-aware CAM. Then, the momentum prototypes P_c are updated using the tokens corresponding to the selected seeds Q_c. For each class c, its prototype P_c is updated as:

$$P_c \leftarrow \tau \cdot P_c + (1-\tau)\cdot \frac{1}{|Q_c|}\sum_{i\in Q_c} f_i, \tag{18}$$

where f_i denotes features in the patch token of the selected seed at position i, τ is the momentum coefficient for prototype updating, and $|\cdot|$ represents the cardinality of the collection. Given an image labeled as class c with its resized feature map $\hat{\mathbf{F}}_{cd} \in \mathbb{R}^{2d\times hw}$ and prototype $\mathbf{P}_c \in \mathbb{R}^{1\times 2d}$, we compute the similarity map $\mathbf{S}_c \in \mathbb{R}^{hw}$, where the similarity at patch position i is

$$S_c(i) \;=\; \max\left(\frac{\mathbf{P}_c\,\hat{\mathbf{F}}_{cd}(:,i)}{\|\mathbf{P}_c\|_2\,\big\|\hat{\mathbf{F}}_{cd}(:,i)\big\|_2},\; 0\right). \tag{19}$$

Finally, the similarity map S_c is reshaped to $h \times w$, and the prototype-aware CAM of class c is obtained as:

$$M_c = \text{Norm}\big(M_f^c \odot S_c^2\big). \tag{20}$$

We stack the class-wise maps M_c and feed them into an online post-processing module, i.e., the pixel-adaptive refinement module (Ru et al., 2022), to generate final online pseudo labels $M_p \in \mathbb{R}^{H\times W}$. After this process, the pseudo label serves

as not only the complementary supervision for the segmentation decoder but also the semantic guidance for affinity learning within the self-attention mechanism.

3.6 Loss Function

We use the segmentation losses of CAM and pseudo labels $\mathcal{L}_{\text{seg}}$ combined with an auxiliary affinity learning loss $\mathcal{L}_{\text{aff}}$ (Ru et al., 2022) to transfer the knowledge from refined CAM to the semantic affinity relations within the multi-head self-attention mechanism. The cross-entropy loss is used for $\mathcal{L}_{\text{seg}}$. The overall optimization objective is defined as:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{seg}} + \lambda_2 \mathcal{L}_{\text{aff}}, \tag{21}$$

where λ_1, λ_2 are the balancing coefficients that rescale the contributions of different learning objectives. Overall, we use TPE to improve the class separability, USR to enforce the structural consistency, and VPE to calibrate the over-activation. Jointly they produce cleaner multi-class CAMs and stronger pseudo labels.

4. Experiments

4.1 Datasets Descriptions

We evaluate the proposed framework on two complementary remote sensing benchmarks: iSAID and ISPRS Potsdam (see the next paragraphs). The iSAID dataset contains small, dense, and elongated objects in complex urban scenes, while the Potsdam dataset provides ultra–high-resolution land-cover labels having clean boundaries and structured regions. Together, they assess both object coverage and boundary accuracy.

iSAID dataset. The iSAID dataset (Waqas Zamir et al., 2019) is designed for instance and semantic segmentation of high-resolution aerial imagery. It contains 2,806 images covering 15 object categories. Following standard protocols, we randomly crop each image into 512×512 tiles, resulting in 7,500 training tiles, 1,653 validation tiles, and 1,315 test tiles.

ISPRS Potsdam dataset. It comprises 38 orthorectified aerial images captured over Potsdam, Germany, each with a resolution of 5 cm and a size of $6{,}000 \times 6{,}000$ pixels (ISPRS 2D Semantic Labeling Contest, 2018). It includes six land-cover classes. We follow the standard split with 24 images for training and 14 for testing, and crop images into non-overlapping 512×512 patches, yielding 3,456 training tiles and 2,016 test tiles. Its fine spatial detail and urban context make it a strong benchmark for assessing fine-grained structure recovery.

4.2 Experimental setup

We employ two frozen pretrained backbones: CLIP with the ViT-B/16 architecture (Dosovitskiy et al., 2021) and DINOv2-ViT-B/14 (Oquab et al., 2024). During training, input images are cropped to 320×320 for CLIP and 308×308 for DINO. The model is trained for 40,000 iterations using the AdamW optimizer with a constant batch size of 4 on a single NVIDIA RTX 3090 GPU. The weight decay is set to 1×10^{-3}. The initial learning rate is 2×10^{-3} and follows a polynomial decay schedule. The decoder consists of three transformer encoder layers with eight attention heads to generate the final feature map. During inference, a multi-scale testing strategy is applied to generate multiple CAMs, which are fused by averaging to obtain the final pseudo labels. Finally, dense CRF (Krähenbühl and Koltun, 2011) is used as a post-processing step for boundary refinement.

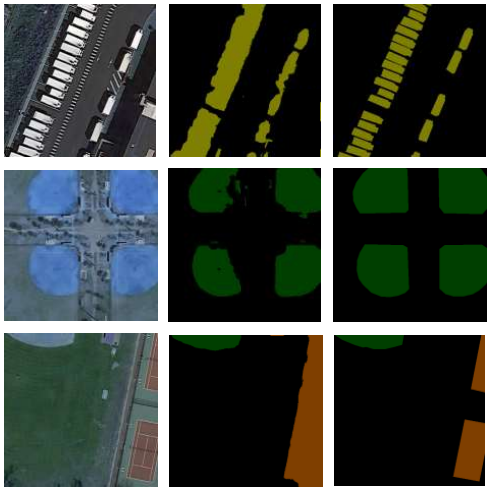


Figure 3. Visualization of the iSAID dataset results: (a) input image; (b) our method; and (c) ground truth.

4.3 Comparison With SOTA Methods

In this section, we present the results of state-of-the-art methods to facilitate a comprehensive performance comparison. Following standard evaluation protocols for semantic segmentation (Hu et al., 2024), we use the mean Intersection-over-Union (mIoU) as the main metric and also report the F1 score and overall accuracy (OA) for a comprehensive assessment.

iSAID dataset. As shown in Table 1, our method achieves an mIoU of 38.01%, outperforming CTFA (Hu et al., 2024) and other baselines. The largest improvements occur in small or elongated object categories such as *soccer-ball field, bridge, ship,* and *helicopter*, which align with our design for better object completeness and boundary accuracy. However, we observe slight performance drops for background and large-scale classes (e.g., *ground track field, storage tank*). Overall, our framework achieves a new SOTA mIoU under the same training setting, particularly excelling in fine-grained object categories. Fig. 3 shows more complete activations in crowded scenes.

ISPRS Potsdam dataset. Table 2 reports the comparison results on the Potsdam dataset. Our method attains the highest mIoU and OA, slightly outperforming CTFA while performing comparably on F1. Per-class results show notable gains for *building, tree,* and *clutter* classes, reflecting improved structural region recovery and noise suppression. Performance is slightly lower for *impervious surface, low vegetation,* and *car*. This may be attributed to conservative diffusion in homogeneous regions and ambiguity between visually similar classes. Overall, the results suggest that our method has good generalization and robustness across diverse urban scenes, with improved boundary quality.

4.4 Ablation Study

We conduct an ablation study on iSAID to evaluate TPE, USR, and VPE in Table 3. Adding TPE or USR alone gives modest but consistent improvements over the baseline. It may be because the improved text–image alignment and relation-based online refinement help suppress part-level and noisy activations,

Table 1. Per-class IoU (%) and mIoU on iSAID dataset.

Model	BG	GTF	SBF	SV	SH	BR	BC	BD	RA	PL	TC	LV	ST	HA	SP	HC	mIoU
S2EPC (Zhou et al., 2022b) [TGRS, 2022]	86.11	38.30	44.10	11.48	17.10	9.71	**53.30**	36.60	**34.59**	27.10	56.74	30.20	49.10	11.59	19.6	15.86	36.79
CISM (Zhou et al., 2023)[RS, 2023]	83.10	33.50	44.60	11.80	19.50	8.40	48.30	34.40	29.20	30.5	52.60	31.90	44.10	10.80	17.60	14.30	32.20
OME (Li et al., 2023)[TGRS, 2023]	**94.62**	36.60	52.72	13.12	22.53	13.12	32.49	9.87	31.15	**48.38**	**66.22**	**54.73**	43.03	31.31	8.17	16.99	35.94
SLRNet (Pan et al., 2022) [IJCV, 2022]	74.66	38.13	30.69	2.18	14.10	7.36	33.92	24.40	23.95	16.46	44.11	39.23	30.21	8.20	12.56	12.46	25.79
AFA (Ru et al., 2022)[CVPR, 2022]	75.17	31.50	40.13	4.76	8.68	9.93	48.86	13.79	24.29	25.93	49.59	33.95	39.53	32.46	32.22	25.03	30.58
CTFA (Hu et al., 2024) [TGRS, 2024]	90.20	**55.97**	40.18	**14.87**	28.73	14.97	42.12	38.21	20.15	37.72	50.67	35.72	**48.15**	**32.85**	32.82	22.10	37.84
Our	83.40	49.94	**53.55**	11.96	**29.61**	**18.76**	49.21	**40.96**	17.16	39.26	47.97	29.75	45.83	25.41	**38.02**	**27.40**	**38.01**

Table 2. Per-class IoU (%) and summary metrics on Potsdam dataset.

Model	Surface	Building	Low Veg.	Tree	Car	Clutter	mIoU	OA	F1
SLRNet (Pan et al., 2022)[IJCV, 2022]	51.20	45.93	35.86	33.61	39.63	4.63	35.14	54.12	48.12
AFA (Ru et al., 2022)[CVPR, 2022]	58.60	48.94	45.93	27.40	42.13	16.68	39.95	60.15	51.21
CTFA (Hu et al., 2024)[TGRS, 2024]	**58.75**	50.58	**51.20**	50.35	**50.82**	17.96	46.61	66.23	**57.12**
Ours	55.40	**52.43**	50.85	**52.67**	50.41	**20.31**	**47.01**	**66.89**	56.70

Table 3. Ablation study of each module on iSAID dataset.

Method	Modules			mIoU (%)
	TPE	USR	VPE	
Baseline				32.80
+ TPE	✓			33.28
+ USR		✓		33.40
+ VPE			✓	36.54
+ TPE + USR	✓	✓		33.95
+ TPE + VPE	✓		✓	36.99
+ USR + VPE		✓	✓	37.67
TPE + USR + VPE	✓	✓	✓	**38.01**

but do not fully address incomplete object coverage. In contrast, VPE brings a much larger gain, indicating that using visual prototypes provides strong dataset-level priors to complete regions and sharpen boundaries. Combining modules further improves the performance, with USR and VPE achieving better results than TPE and VPE, implying that affinity-guided diffusion is most effective when calibrated by reliable visual prototypes. Finally, enabling all three modules reaches 38.01% mIoU, confirming their complementary roles in enhancing pseudo label completeness and spatial coherence.

5. Conclusion

We proposed a weakly supervised framework that leverages frozen pretrained priors from CLIP and DINO for the semantic segmentation of remote sensing imagery. Without fine-tuning the large pretrained backbones, our method uses a prototype-aware module to enrich textual features and a lightweight adapter–decoder structure to capture visual representations efficiently. Two CAM refinement modules enhance pseudo label quality by combining learnable feature relations and DINO-based spatial correlations. Our approach achieves competitive accuracy while substantially reducing the number of trainable parameters. These results suggest that composing frozen vision–language and self-supervised priors is a promising direction for label-efficient semantic segmentation in remote sensing imagery.

Acknowledgments

Financial support from the programme of the China Scholarships Council (Grant No. 202406830041) is acknowledged. The research activities carried out in this paper are also supported by project TNE23-00012, ‘Green & Pink for Sustainable Education’ (GPSEducation), funded by Sub-Measure T4 ‘Transnational Initiatives in Education’, Investment 3.4 ‘University Teaching and Advanced Skills’ of the National Recovery and Resilience Plan, Mission 4 ‘Education and Research’, Component 1 ‘Strengthening the Offering of Education Services: From Nursery Schools to University’, for the promotion and implementation of transnational educational initiatives (TNE). The project is funded by the European Union, NextGenerationEU (CUP D74G23000280006). We gratefully thank the authors of iSAID and the ISPRS benchmark organizers for providing the iSAID and Potsdam datasets used in this work. We also thank the three anonymous reviewers for their constructive comments, which helped improve this paper.

References

Andries, A., Morse, S., Murphy, R. J., Lynch, J., Woolliams, E. R., 2022. Using data from earth observation to support sustainable development indicators: An analysis of the literature and challenges for the future. *Sustainability*, 14(3), 1191.

Cao, Y., Huang, X., 2022. A coarse-to-fine weakly supervised learning method for green plastic cover segmentation using high-resolution remote sensing images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 188, 157–176.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A., 2021. Emerging properties in self-supervised vision transformers. *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, IEEE, 9630–9640.

Chen, K., Zhang, J., Liu, C., Zou, Z., Shi, Z., 2025. RSRefSeg: Referring Remote Sensing Image Segmentation with Foundation Models. arXiv 2025. *arXiv preprint arXiv:2501.06809*.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Proc. Int. Conf. Learn. Represent. (ICLR)*.

Fang, F., Zheng, D., Li, S., Liu, Y., Zeng, L., Zhang, J., Wan, B., 2022. Improved pseudomasks generation for weakly supervised building extraction from high-resolution remote sensing imagery. *IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens.*, 15, 1629–1642.

Hu, Z., Gao, J., Yuan, Y., Li, X., 2024. Contrastive Tokens and Label Activation for Remote Sensing Weakly Supervised Semantic Segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 62, 1–11.

ISPRS 2D Semantic Labeling Contest, 2018. ISPRS 2D Semantic Labeling Contest. http://www2.isprs.org.

Jing, L., Tian, Y., 2021. Self-Supervised Visual Feature Learning With Deep Neural Networks: A Survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11), 4037–4058.

Krähenbühl, P., Koltun, V., 2011. Efficient inference in fully connected crfs with gaussian edge potentials. *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 109–117.

Li, R., Zheng, S., Zhang, C., Duan, C., Wang, L., Atkinson, P. M., 2021. ABCNet: Attentive bilateral contextual network for efficient semantic segmentation of Fine-Resolution remotely sensed imagery. *ISPRS journal of photogrammetry and remote sensing*, 181, 84–98.

Li, Z., Zhang, X., Xiao, P., 2023. One model is enough: Toward multiclass weakly supervised remote sensing image semantic segmentation. *IEEE Trans. Geosci. Remote. Sens.*, 61, 1–13.

Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X., 2023. CLIP is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 15305–15314.

Lloyd, S. P., 1982. Least squares quantization in PCM. *IEEE Trans. Inf. Theory*, 28(2), 129–136.

Lu, X., Jiang, Z., Zhang, H., 2024. Weakly Supervised Remote Sensing Image Semantic Segmentation With Pseudo-Label Noise Suppression. *IEEE Trans. Geosci. Remote. Sens.*, 62, 1–12.

Maulik, U., Chakraborty, D., 2017. Remote Sensing Image Classification: A survey of support-vector-machine-based advanced techniques. *IEEE Geosci. Remote Sens. Mag.*, 5(1), 33-52.

Omia, E., Bae, H., Park, E., Kim, M. S., Baek, I., Kabenge, I., Cho, B.-K., 2023. Remote Sensing in Field Crop Monitoring: A Comprehensive Review of Sensor Systems, Data Analyses and Recent Advances. *Remote Sensing*, 15(2).

Oquab, M., Darcet, T., Moutakanni, T., Vo, H. V., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., Assran, M., Ballas, N., Galuba, W., Howes, R., Huang, P., Li, S., Misra, I., Rabbat, M., Sharma, V., Synnaeve, G., Xu, H., Jégou, H., Mairal, J., Labatut, P., Joulin, A., Bojanowski, P., 2024. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.*, 2024.

Pan, J., Zhu, P., Zhang, K., Cao, B., Wang, Y., Zhang, D., Han, J., Hu, Q., 2022. Learning Self-supervised Low-Rank Network for Single-Stage Weakly and Semi-supervised Semantic Segmentation. *Int. J. Comput. Vis.*, 130(5), 1181–1195.

Qiao, W., Shen, L., Wang, J., Yang, X., Li, Z., 2023. A weakly supervised semantic segmentation approach for damaged building extraction from postearthquake high-resolution remote-sensing images. *IEEE Geosci. Remote. Sens. Lett.*, 20, 1–5.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I., 2021. Learning transferable visual models from natural language supervision. M. Meila, T. Zhang (eds), *Proc. Int. Conf. on Machine Learning (ICML)*, 139, 8748–8763.

Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J., 2022. Denseclip: Language-guided dense prediction with context-aware prompting. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 18082–18091.

Ru, L., Zhan, Y., Yu, B., Du, B., 2022. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), CVPR 2022*, 16825–16834.

Siméoni, O., Puy, G., Vo, H. V., Roburin, S., Gidaris, S., Bursuc, A., Pérez, P., Marlet, R., Ponce, J., 2021. Localizing objects with self-supervised transformers and no labels. *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 310.

Waqas Zamir, S., Arora, A., Gupta, A., Khan, S., Sun, G., Shahbaz Khan, F., Zhu, F., Shao, L., Xia, G.-S., Bai, X., 2019. isaid: A large-scale dataset for instance segmentation in aerial images. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 28–37.

Xie, J., Hou, X., Ye, K., Shen, L., 2022. CLIMS: cross language image matching for weakly supervised semantic segmentation. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 4473–4482.

Yan, X., Shen, L., Pan, J., Wang, J., Chen, C., Li, Z., 2023. ALNet: Auxiliary learning-based network for weakly supervised building extraction from high-resolution remote sensing images. *IEEE Trans. Geosci. Remote. Sens.*, 61, 1–16.

Yan, X., Shen, L., Wang, J., Wang, Y., Li, Z., Xu, Z., 2022. PANet: Pixelwise affinity network for weakly supervised building extraction from high-resolution remote sensing images. *IEEE Geosci. Remote. Sens. Lett.*, 19, 1–5.

Zeng, X., Wang, T., Dong, Z., Zhang, X., Gu, Y., 2023. Superpixel consistency saliency map generation for weakly supervised semantic segmentation of remote sensing images. *IEEE Trans. Geosci. Remote. Sens.*, 61, 1–16.

Zhang, L., Ma, J., Lv, X., Chen, D., 2019. Hierarchical weakly supervised learning for residential area semantic segmentation in remote sensing images. *IEEE Geosci. Remote. Sens. Lett.*, 17(1), 117–121.

Zhang, P., Ke, Y., Zhang, Z., Wang, M., Li, P., Zhang, S., 2018. Urban Land Use and Land Cover Classification Using Novel Deep Learning Models Based on High Spatial Resolution Satellite Imagery. *Sensors*, 18(11).

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A., 2016. Learning deep features for discriminative localization. *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2921–2929.

Zhou, R., Yuan, Z., Rong, X., Ma, W., Sun, X., Fu, K., Zhang, W., 2023. Weakly Supervised Semantic Segmentation in Aerial Imagery via Cross-Image Semantic Mining. *Remote. Sens.*, 15(4), 986.

Zhou, R., Zhang, W., Yuan, Z., Rong, X., Liu, W., Fu, K., Sun, X., 2022a. Weakly supervised semantic segmentation in aerial imagery via explicit pixel-level constraints. *IEEE Trans. Geosci. Remote. Sens.*, 60, 1–17.

Zhou, R., Zhang, W., Yuan, Z., Rong, X., Liu, W., Fu, K., Sun, X., 2022b. Weakly Supervised Semantic Segmentation in Aerial Imagery via Explicit Pixel-Level Constraints. *IEEE Trans. Geosci. Remote. Sens.*, 60, 1–17.

Appendix

A. Additional Qualitative Results

A.1 CAM Refinement Visualization

To better illustrate the effectiveness of the proposed refinement strategy, we visualize the refinement pipeline in Fig. 4, including the original image, the initial CAM generated by CLIP, the refined CAM after applying the USR and VPE modules, and the final segmentation prediction.

As shown in Fig. 4(b), the initial CAM produced by CLIP mainly highlights the most discriminative regions. However, these activations often fail to cover the full spatial extent of the target objects and tend to be biased toward dominant foreground cues, resulting in incomplete or partially incorrect responses. This limitation becomes more evident for small objects or low-contrast regions, where the activation maps are typically sparse and fragmented.

After applying the proposed refinement modules, the CAMs become noticeably more spatially coherent (Fig. 4(c)). The refined cam exhibits improved regional continuity and more comprehensive object coverage. In particular, the semantic propagation mechanism in the USR module alleviates isolated activations and enhances intra-object consistency, while the prototype-guided enhancement in the VPE module promotes compact and structurally complete responses while suppressing noise.

Finally, the segmentation predictions supervised by the refined CAMs (Fig. 4(d)) demonstrate improved object completeness and clearer boundary structures compared with predictions from the initial activations. Although challenging scenarios, such as densely distributed small objects, may still introduce minor ambiguities, the overall spatial consistency and regional integrity are substantially improved.

Thus, these comparisons indicate that the refinement process effectively transforms sparse and discriminative initial activations into more reliable supervisory signals for weakly supervised segmentation.

A.2 Segmentation Quality Comparisons

In addition, we provide qualitative comparisons on cropped regions containing complex object contours and closely adjacent structures to further examine the influence of the refinement modules on segmentation quality in Fig. 5. These examples highlight local characteristics that may not be fully reflected by global metrics such as mIoU, including boundary delineation, small-object separation, and regional completeness.

Compared with the baseline segmentation results, which only use CLIP, the predictions refined by USR and VPE generally exhibit smoother and more regular object boundaries, with fewer jagged edges and reduced scattered misclassified pixels near object contours, as illustrated in Fig. 5(c) and (d).

Beyond boundary improvements, the refinement process also enhances semantic region completeness. Object interiors become more contiguous and semantically consistent, and previously missing or weakly activated regions are more frequently recovered, leading to more complete object extents (Fig. 5(a), (b), and (e)). Meanwhile, background noise and spurious activations are effectively suppressed, resulting in more structurally coherent segmentation masks.

A.3 Scene-Level Qualitative Analysis

We further show scene-level qualitative results on full remote sensing images from Potsdam dataset in Fig. 6. As shown in Fig. 6, although local inconsistencies remain across cropped regions, the proposed method still preserves the main semantic structure of the scene. In particular, spatially continuous categories such as roads can be reasonably recovered at the full-image level.

These errors are mainly caused by the loss of contextual information after cropping, especially near crop boundaries, where semantically similar regions may be assigned different labels. Meanwhile, weakly supervised semantic segmentation for remote sensing imagery remains highly challenging, and current methods still achieve relatively limited mIoU. These observations indicate that full-scene weakly supervised segmentation remains challenging, particularly near crop boundaries where contextual information is incomplete.

B. Additional Quantitative Analyses

B.1 Computational Complexity and Inference Efficiency

Our framework is designed to leverage pretrained priors while maintaining computational efficiency during both training and inference. Specifically, the CLIP and DINO backbones remain frozen in our model, while only a lightweight feature adapter and transformer decoder are optimized. This design significantly reduces the optimization cost compared with approaches that require training or fine-tuning large backbone networks.

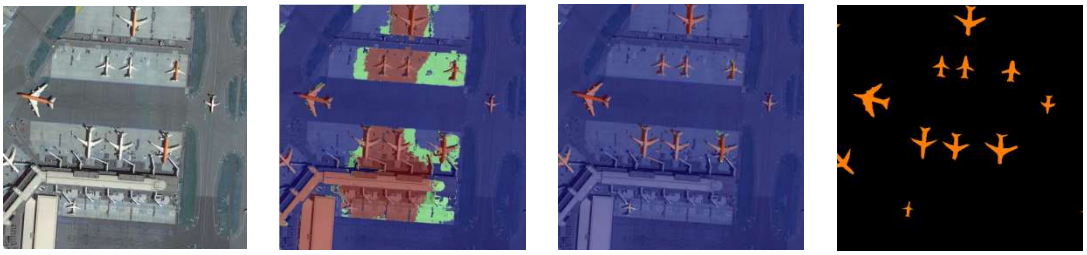
In our implementation, model complexity is reported using the number of trainable parameters, total parameters, and FLOPs. The parameter count is obtained by summing the element counts of all learnable tensors in the network. Trainable parameters refer to those updated during optimization, whereas total parameters include both trainable modules and frozen pretrained components. FLOPs are computed using `fvcore` by profiling a single forward pass of the deployed network under a fixed input resolution and reporting the total operation count in giga-FLOPs (G). For fairness, only neural network forward computations are included in the FLOPs calculation, while data loading, resizing, augmentation, and post-processing steps are excluded.

Table 4. Model complexity comparison on iSAID Dataset.

Method	Total Params (M)	Trainable Params (M)	FLOPs (G)
CTFA (Hu et al., 2024)	111.50	104.79	979.32
Ours	177.89	16.45	194.32

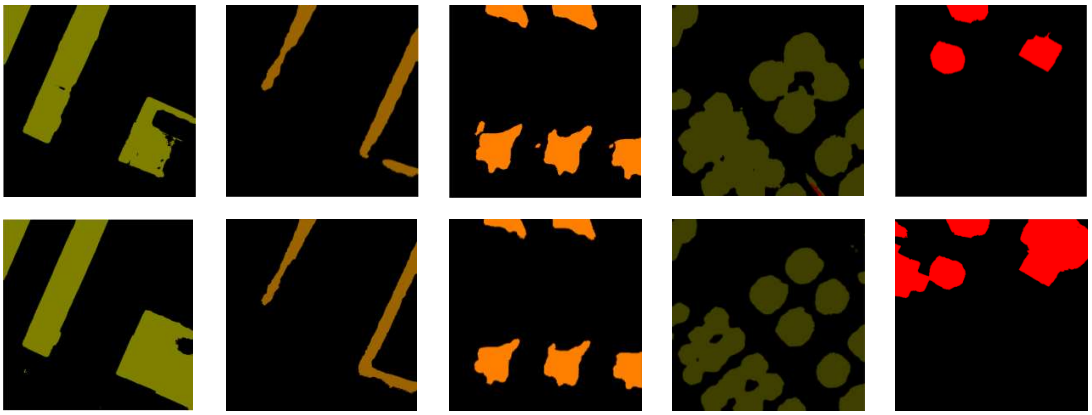
Compared with CTFA (Hu et al., 2024) in Table 4, our approach is substantially more parameter-efficient during training. CTFA jointly optimizes a large task-specific backbone and multiple auxiliary branches, which leads to a considerably larger number of trainable parameters. In contrast, we keep the pretrained CLIP and DINO backbones frozen and only update lightweight components such as the feature adapter and transformer decoder. As a result, the number of trainable parameters is reduced to nearly 10% of that of CTFA, while the total parameter count may remain comparable due to the inclusion of frozen pretrained priors. In addition, our method requires fewer FLOPs per forward pass, indicating improved computational efficiency at inference time. This design reduces optimizer states and gradient storage, thereby lowering the training memory footprint and improving overall training efficiency.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-163-2026 | © Author(s) 2026. CC BY 4.0 License.



(a) (b) (c) (d)

Figure 4. CAM refinement visualization. (a) Original image. (b) Initial CLIP-based CAM. (c) Refined CAM after applying USR and VPE. (d) Final segmentation prediction.



(a) (b) (c) (d) (e)

Figure 5. Segmentation quality comparisons. The top row shows the segmentation results without refinement, while the bottom row presents the results after applying the USR and VPE modules.

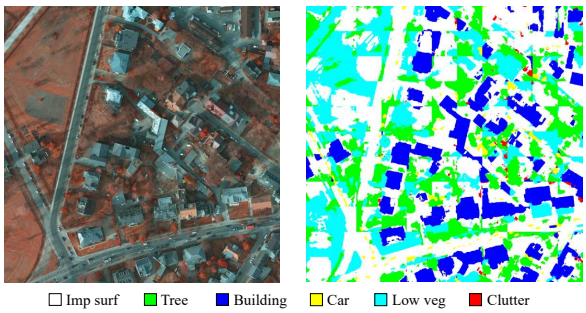


Figure 6. Scene-level qualitative results on full remote sensing images from the Potsdam dataset. From left to right: the input image and our prediction.

Consequently, our framework is particularly well-suited for scenarios with limited training resources.

B.2 Hyperparameter Sensitivity

We further analyze the sensitivity of key hyperparameters involved in the CAM refinement process. Unless otherwise specified, all experiments follow the same configuration as in the main paper, and performance is reported in terms of mIoU on the iSAID validation set.

Sensitivity to the number of clusters in K-means. We vary the number of clusters B used in the K-means-based prototype construction in TPE module. As shown in Table 5, the performance varies only slightly (within 0.56 mIoU) when $B \in \{4, 6, 8, 10\}$, indicating that the proposed method is relatively insensitive to the clustering granularity. Based on this observation, we adopt $B = 8$ as the default setting in all other experiments.

Table 5. Sensitivity to the number of clusters B in K-means.

B	4	6	8	10
mIoU	37.45	37.79	38.01	37.83

Table 6. Sensitivity to the diffusion strength α in the USR module.

α	1	2	3	4
mIoU	37.76	37.97	38.01	37.93

Sensitivity to the refinement steps in USR. We further investigate the influence of the diffusion strength α in Eq. 15. As shown in Table 6, the performance remains stable across $\alpha \in \{1, 2, 3, 4\}$, with a variation of less than 0.25 mIoU. The best performance is obtained at a moderate diffusion strength ($\alpha = 3$). When α becomes larger, slightly stronger propagation along the affinity relations may introduce mild over-smoothing effects. Therefore, we adopt $\alpha = 3$ as the default configuration.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-3-2026-163-2026 | © Author(s) 2026. CC BY 4.0 License.