ModuSeg: Decoupling Object Discovery and Semantic Retrieval for Training-Free Weakly Supervised Segmentation

Qingze He[1], Fagui Liu[1,2⋆], Dengke Zhang[1], Qingmao Wei[1,2], and Quan Tang[2⋆]

[1] South China University of Technology, Guangzhou, China
[2] Pengcheng Laboratory, Shenzhen, China
202230430064@mail.scut.edu.cn, fgliu@scut.edu.cn, tangq@pcl.ac.cn

Abstract. Weakly supervised semantic segmentation aims to achieve pixel-level predictions using image-level labels. Existing methods typically entangle semantic recognition and object localization, which often leads models to focus exclusively on sparse discriminative regions. Although foundation models show immense potential, many approaches still follow the tightly coupled optimization paradigm, struggling to effectively alleviate pseudo-label noise and often relying on time-consuming multi-stage retraining or unstable end-to-end joint optimization. To address the above challenges, we present ModuSeg, a training-free weakly supervised semantic segmentation framework centered on explicitly decoupling object discovery and semantic assignment. Specifically, we integrate a general mask proposer to extract geometric proposals with reliable boundaries, while leveraging semantic foundation models to construct an offline feature bank, transforming segmentation into a non-parametric feature retrieval process. Furthermore, we propose semantic boundary purification and soft-masked feature aggregation strategies to effectively mitigate boundary ambiguity and quantization errors, thereby extracting high-quality category prototypes. Extensive experiments demonstrate that the proposed decoupled architecture better preserves fine boundaries without parameter fine-tuning and achieves highly competitive performance on standard benchmark datasets. Code is available at https://github.com/Autumnair007/ModuSeg.

Keywords: Weakly Supervised Semantic Segmentation · Training-Free Framework · Retrieval-Augmented Segmentation

1 Introduction

Weakly supervised semantic segmentation (WSSS) aims to achieve pixel-level predictions using image-level labels [1, 16], thereby substantially reducing dense annotation costs. Existing mainstream methods typically rely on class activation mapping [43] to provide initial localization cues. However, these methods

⋆ Corresponding authors.

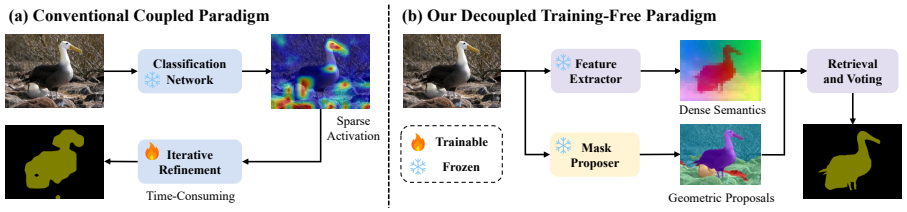

Fig. 1: Comparison of paradigms. Unlike conventional coupled methods that rely on iterative refinement, our decoupled approach directly assigns semantics to geometric proposals via a training-free retrieval mechanism.

inherently couple semantic recognition and object localization within a single classification objective. Consequently, as classification networks naturally tend to focus on the most discriminative local regions [4,29,42], the generated activation maps often cover only sparse object fragments. To rectify these incomplete initial regions, traditional paradigms typically rely on time-consuming multi-stage network retraining [21] or struggle with the instability of complex end-to-end joint training. This coupled design increases training complexity and restricts the geometric precision of the final segmentation masks.

The recent advancement of visual foundation models [18] provides a new opportunity to address the above dilemma. Class-agnostic segmentation models have demonstrated exceptional boundary perception capabilities, while vision-language and self-supervised models have learned highly expressive dense semantic descriptions from massive data [45]. Despite their potential, many existing weakly supervised segmentation approaches still follow the conventional path of coupled optimization, attempting to integrate foundation model priors through network fine-tuning or end-to-end distillation [2, 13, 37]. Such approaches often struggle to circumvent the interference of inherent noise in weakly supervised pseudo-labels, underutilizing the true capabilities of foundation models.

To address the limitations of conventional coupled paradigms illustrated in Fig. 1 (a), we propose ModuSeg, a novel training-free weakly supervised semantic segmentation framework shown in Fig. 1 (b). The core insight lies in the explicit decoupling of object discovery and semantic assignment. We delegate the localization task to an off-the-shelf, high-performance general mask proposer, which provides geometric proposals with reliable boundaries. Simultaneously, we utilize semantic foundation models to construct an offline feature bank. By transforming the segmentation task into a non-parametric feature retrieval process tailored for geometric proposals, our method better preserves the fine physical boundaries of objects while assigning semantics. This decoupling strategy facilitates a highly modular paradigm, enabling our approach to bypass time-consuming retraining and seamlessly integrate continuously evolving feature extractors.

To mitigate the structural alignment errors and boundary noise inherent in weakly supervised pseudo-labels, we further design two targeted modules to enhance the feature bank construction. We propose a semantic boundary purifi-

cation strategy to alleviate feature mixing by stripping away ambiguous edge regions, thereby improving the semantic purity of the extracted features. Furthermore, we introduce a soft-masked feature aggregation mechanism to pool visual embeddings via soft weights, which alleviates the noise caused by spatial hard quantization. These modules operate in conjunction to extract discriminative semantic representations for the subsequent retrieval process.

The main contributions of this work are summarized as follows:

- We propose the explicit decoupling of object discovery and semantic retrieval, diverging from traditional approaches constrained by coupled optimization (whether multi-stage or end-to-end), and constructing an efficient, training-free modular segmentation framework.
- We introduce the semantic boundary purification and soft-masked feature aggregation modules during the feature bank construction stage. These modules mitigate boundary ambiguity and quantization errors, improving the quality of feature prototypes.
- ModuSeg achieves highly competitive performance on the primary PASCAL VOC and MS COCO benchmarks, while auxiliary evaluations on ADE20K and Cityscapes further support generalizability across diverse datasets.

2 Related Work

2.1 Weakly Supervised Semantic Segmentation

The fundamental challenge in WSSS lies in the inherent coupling between semantic classification and object localization. Traditional approaches rely on Class Activation Maps (CAMs) [43], which inevitably focus on sparse discriminative object parts rather than the complete object extent. To rectify these incomplete parts, earlier methods such as AffinityNet [1] require complex iterative refinement and computationally expensive retraining of a final segmentation network.

Recent advancements leverage stronger priors from Vision Transformers (ViT), Vision-Language Models (VLMs) and class-agnostic segmentation models. Methods such as ToCo [22] and MCTformer+ [32] utilize self-attention mechanisms to encourage integral object activation. Meanwhile, CLIP-based approaches have emerged as a dominant paradigm. WeCLIP [40] and WeakCLIP [47] freeze the CLIP backbone to extract dense knowledge. SAM-based methods such as SEPL [3] and S2C [10] exploit class-agnostic mask priors to improve CAM or pseudo-label quality. However, these methods rely on a coupled optimization paradigm, forcing models to learn dense predictions from inherently noisy pseudo-supervision. In contrast, our work decouples object discovery and semantic retrieval by formulating segmentation as a training-free proposal-retrieval process.

2.2 Vision Foundation Models

The evolution of foundation models suggests a new philosophy in which specialized models achieve superior performance in their respective decoupled domains.

On the semantic front, dense vision foundation models such as DINOv2/v3 [14, 24] and C-RADIOv4 [19] provide robust pixel-wise semantic descriptors that surpass supervised baselines. On the geometric front, class-agnostic models such as EntitySeg [17] and SAM 2 [20] have effectively addressed the localization problem by providing high-fidelity boundaries independent of semantic categories.

Open-Vocabulary Semantic Segmentation (OVSS) assigns pixel labels using textual category descriptions, including categories unseen during training, covering training-free [23, 35, 41] and supervised [5, 11] methods. In contrast, standard WSSS learns from image-level labels for fixed category sets. ModuSeg follows this WSSS setting: image-level labels constrain feature-bank construction rather than define open-vocabulary queries. While OVSS models provide strong semantic priors, directly using their predictions for WSSS risks hallucinating absent categories (*i.e.*, false positives). Therefore, ModuSeg does not pursue open-vocabulary generalization; it integrates pre-trained foundation models into an image-level constrained WSSS framework, where proposal-level retrieval assigns semantics from the feature bank.

2.3 Prototype Learning

Prototype learning reformulates segmentation by representing categories as feature centroids rather than parametric weights, employing nearest-neighbor retrieval or feature reconstruction for dense prediction [26, 46]. In WSSS, this paradigm is adapted to bridge the modality gap. For instance, SSR [2] aligns visual features with learnable text prototypes to reduce class confusion, while ExCEL [37] enriches text prototypes via LLMs for fine-grained patch alignment. However, these approaches typically rely on intricate optimization processes for mask refinement. In contrast, we formulate segmentation as a retrieval task. By leveraging a purified feature bank derived from frozen foundation models, our method ensures robust semantic assignment without parameter tuning.

3 Methodology

3.1 Method Overview

As illustrated in Fig. 2, ModuSeg decouples object discovery from classification to achieve training-free semantic segmentation. The pipeline consists of: (1) Robust feature bank construction. We utilize image-level supervision to generate initial masks, which are processed via semantic boundary purification to mitigate feature mixing at object edges. These refined regions are converted into prototype vectors using soft-masked feature aggregation and filtered to enhance intra-class compactness. (2) Class-agnostic object discovery. During inference, we utilize a frozen, class-agnostic mask proposer to extract geometric proposals across diverse object categories. (3) Retrieval-augmented semantic assignment. We extract query embeddings for each proposal via soft masking and perform Top-K similarity retrieval against the offline bank. A majority voting strategy determines the class label, followed by confidence-priority rasterization to resolve spatial conflicts and produce the final segmentation.

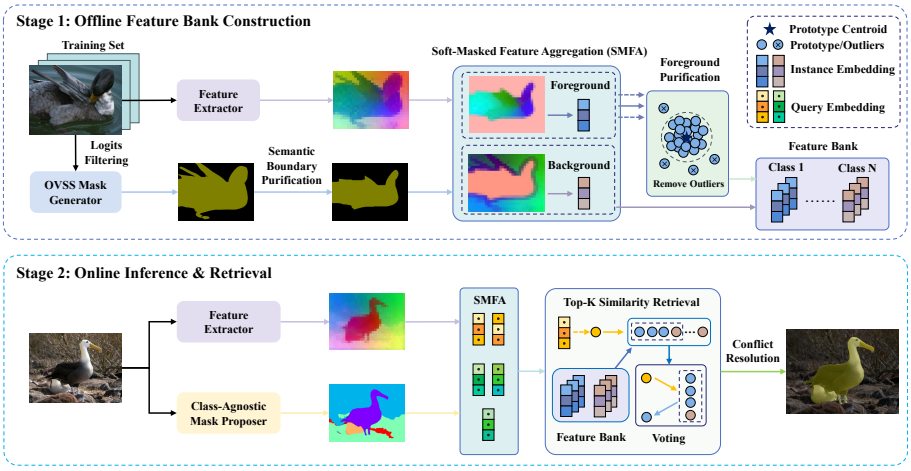

Fig. 2: Overview of our proposed ModuSeg. The framework operates in two stages: (1) Offline Feature Bank Construction, where we refine semantic boundaries to extract representative prototypes via soft-masked aggregation; and (2) Online Inference & Retrieval, which assigns semantics to class-agnostic proposals through a retrieval-based voting mechanism, effectively decoupling localization from classification.

3.2 Stage 1: Constructing a Robust Feature Bank

The efficacy of a retrieval-based segmentation framework hinges on the representational quality of its feature prototypes. In this stage, we focus on extracting semantically pure and discriminative representations from the training images to construct a high-fidelity feature bank.

Mask Generation with Semantic Boundary Purification To initialize the feature bank, we utilize image-level supervision to generate candidate regions. Let $\mathcal{D} = \{(I_i, \mathcal{Y}_i)\}_{i=1}^{N}$ denote the training dataset, where I_i denotes the i-th image, $\mathcal{Y}_i$ is the set of ground-truth class labels and N denotes the total number of images. We employ CorrCLIP [41] as our base mask generator $\mathcal{G}_{mask}$.

Image-Level Logits Filtering. While CorrCLIP provides a strong baseline, applying it directly can introduce false positives (*i.e.*, hallucinated classes). To ensure semantic fidelity, we modify the generation process by incorporating an image-level logits filtering mechanism. Specifically, before the final prediction, we mask the class logits corresponding to categories not present in $\mathcal{Y}_i$ by setting their values to $-\infty$. This constraint forces the model to assign pixels only to the classes actually present in the image or to the background. Formally, we generate the initial map S_i as follows:

$$S_i = \mathcal{G}_{mask}(I_i, \mathcal{Y}_i). \tag{1}$$

Here, $\mathcal{G}_{mask}$ denotes the mask generator conditioned on the image-level labels $\mathcal{Y}_i$, ensuring that S_i contains only valid categories.

Semantic Boundary Purification (SBP). Even with semantically correct labels, extracting features using ViT encounters two challenges regarding boundary fidelity: (1) **Structural Patch-Pixel Misalignment**, where the fixed-grid processing of ViT (*e.g.*, 16×16 patches) inevitably leads to patches at boundaries encoding mixed foreground-background representations; and (2) **Inherent Mask Noise**, as pseudo-masks generated from weak supervision often exhibit uncertainty and irregularity along object edges.

To address these challenges simultaneously, we propose the Semantic Boundary Purification (SBP) strategy. We first derive the raw binary mask for class c from the initial map S_i as $m_c = \mathbb{I}(S_i = c)$, where $\mathbb{I}(\cdot)$ denotes the indicator function. Instead of utilizing this raw mask directly, we apply t iterations of strict morphological erosion to explicitly discard ambiguous boundary regions:

$$m_c^{pure} = \text{Erode}^{(t)}(m_c, B_k). \tag{2}$$

In this equation, m_c^{pure} denotes the resulting purified mask, B_k represents a structuring element with a kernel size of k, and t denotes the number of iterations. This operation serves a twofold purpose: it eliminates feature mixing by ensuring that the extracted patches fall strictly within the object interior, and it filters mask noise by pruning the unreliable edge pixels. By sacrificing geometric completeness at the boundaries, SBP ensures that the subsequent feature extraction focuses solely on the semantic core of the object, thereby maximizing the purity of the feature bank.

Modular Feature Extraction Given the purified masks m_c^{pure}, we extract their corresponding visual embeddings. A naive approach is to downsample the mask to the feature grid size using nearest-neighbor interpolation. However, this hard quantization exacerbates spatial quantization noise, assigning binary labels to patches that may only be partially covered by the mask.

To achieve a more precise representation, we employ Soft-Masked Feature Aggregation (SMFA). Let $F \in \mathbb{R}^{h \times w \times D}$ be the feature map extracted from the frozen backbone Φ, where h, w, and D denote the height, width, and channel dimension of the feature grid, respectively. We project the high-resolution purified mask m_c^{pure} onto the low-resolution feature grid via area-based interpolation, resulting in a soft assignment map $W \in [0,1]^{h \times w}$. The value $W_{x,y}$ quantifies the semantic contribution of patch (x, y) to the object class c.

The instance-specific prototype vector $v_c \in \mathbb{R}^D$ is computed as a weighted aggregate:

$$v_c = \text{Normalize}\left(\frac{\sum_{x,y} W_{x,y} \cdot F_{x,y}}{\sum_{x,y} W_{x,y} + \epsilon}\right). \tag{3}$$

Here, ϵ is a small constant to prevent division by zero. This soft-weighting mechanism ensures that patches with higher mask coverage contribute more to the final prototype, while marginal patches are naturally suppressed. This design effectively decouples mask generation from feature extraction, rendering our framework model-agnostic and compatible with various ViT architectures.

Prototype-based Feature Purification Since pseudo-masks are inherently noisy, the initial feature bank inevitably contains outliers—features corresponding to misaligned regions or ambiguous visual patterns. To enhance intra-class compactness, we introduce a prototype-based purification mechanism.

The Cluster Assumption. Our strategy relies on the Cluster Assumption: in the embedding space, reliable features of the same semantic category tend to form a dense, unimodal cluster, whereas noisy samples are sparsely distributed in the periphery.

Selective Outlier Rejection. For each semantic class c, we aggregate all extracted feature vectors $\mathcal{V}_c$ to compute a global class prototype μ_c, which serves as the robust centroid of the distribution: $\mu_c = \frac{1}{|\mathcal{V}_c|} \sum_{v \in \mathcal{V}_c} v$.

We then measure the semantic deviation of each instance v via its Euclidean distance from the prototype $d_v = \|v - \mu_c\|_2$. To refine the feature bank, we discard the top $\alpha\%$ of samples with the largest deviations:

$$\mathcal{V}_c^{clean} = \left\{ v \in \mathcal{V}_c \;\middle|\; d_v \leq \text{Quantile}\left(1 - \frac{\alpha}{100}, \{d_{v'} \mid v' \in \mathcal{V}_c\}\right) \right\}. \tag{4}$$

Notably, this purification is applied exclusively to foreground classes. While foreground objects generally exhibit high intra-class similarity (*e.g.*, "cars" share common visual traits), the "background" class is inherently multimodal, encompassing diverse entities such as sky, roads, and vegetation. Filtering background features based on a single centroid would reduce the diversity of the negative samples, impairing the model's ability to distinguish objects from complex environments. Therefore, we retain all background embeddings to maintain a comprehensive representation of the visual context.

3.3 Stage 2: Inference via Semantic Retrieval

Unlike traditional WSSS paradigms that require a resource-intensive final segmentation network retraining on pseudo-labels, our approach explicitly decouples the task into two independent, training-free components: Class-Agnostic Localization and Retrieval-based Classification. This non-parametric design ensures high-quality boundary delineation using off-the-shelf geometric priors.

Localization via Class-Agnostic Object Discovery Traditional WSSS methods often struggle with "boundary ambiguity" or "over-activation" because classification networks lack explicit geometric constraints during the localization phase. To address these geometric deficiencies, we utilize specialized object discovery models as a Class-Agnostic Mask Proposer.

Formally, given a test image I_{test}, we employ a generalized segmentation model Ψ to extract a set of potential object proposals $\mathcal{P}$:

$$\mathcal{P} = \Psi(I_{\text{test}}) = \{(m_p, s_p^{obj})\}_{p=1}^{M}. \tag{5}$$

Here, M is the total number of proposals, and $m_p \in \{0,1\}^{H \times W}$ denotes the binary mask of the p-th proposal, where H and W are scalars representing the

height and width of the input image, respectively. $s_p^{obj} \in [0,1]$ represents its objectness score, indicating the confidence of being a valid object.

Training-Inference Asymmetry. A critical distinction in our framework lies in handling mask boundaries across stages. While we apply SBP during the feature bank construction (Sec. 3.2) to ensure feature purity, we utilize raw, uneroded proposals during inference. This asymmetry is intentional: training prioritizes precision, whereas inference prioritizes completeness by using raw proposals to preserve fine-grained geometric details that SBP otherwise discards. Furthermore, this design accommodates the inherent limitations of class-agnostic proposers. Unlike pristine ground-truth annotations, they typically over-partition scenes into fragmented entities based on low-level cues (as diagnosed in Sec. 4.4). Consequently, eroding these already fragmented inference proposals severely diminishes their representational capacity and degrades matching accuracy, even when they are strictly used for feature retrieval. Thus, relying on raw proposals during inference is essential to prevent exacerbating the semantic gap.

To ensure computational efficiency, we filter out low-quality proposals. We discard candidates with an objectness score below a threshold τ_{obj}:

$$\mathcal{P}_{valid} = \{m_p \mid (m_p, s_p^{obj}) \in \mathcal{P}, s_p^{obj} \geq \tau_{obj}\}. \tag{6}$$

Retrieval-Augmented Semantic Assignment Once a valid geometric proposal $m_p \in \mathcal{P}_{valid}$ is obtained, the subsequent task is to assign a semantic label to it. We address this semantic assignment via a non-parametric retrieval process from our offline feature bank $\mathcal{B} = \bigcup_c \mathcal{V}_c^{clean}$.

Query Extraction. For each proposal m_p, we extract its visual embedding q_p using the same frozen backbone Φ used in Stage 1. To ensure consistency between training and inference representations, we apply the same SMFA strategy (Sec. 3.2) to pool the feature maps within the region defined by m_p.

Top-K Similarity Retrieval. Rather than relying on a single nearest neighbor, we identify the K most similar visual embeddings within the feature bank $\mathcal{B}$ to provide a robust semantic context for q_p. Let $\mathcal{N}_K(q_p) = \{(v_j, y_j)\}_{j=1}^K$ denote the set of retrieved Top-K candidates, where v_j is the feature vector and y_j is its corresponding class label. The similarity between the query q_p and each retrieved feature v_j is quantified using cosine similarity:

$$\text{sim}(q_p, v_j) = \frac{q_p \cdot v_j}{\|q_p\|\|v_j\|}. \tag{7}$$

Hierarchical Semantic Voting. To determine the class label $\hat{c}_p$, we propose a hierarchical voting strategy that prioritizes neighborhood consensus over raw similarity magnitude. This strategy mitigates the influence of outliers that may possess high similarity scores despite being semantically irrelevant.

First, we calculate the vote count for each class c present in the neighborhood:

$$\text{Vote}(c) = \sum_{(v_j, y_j) \in \mathcal{N}_K(q_p)} \mathbb{I}(y_j = c). \tag{8}$$

We assign class labels by majority vote $\hat{c}_p = \arg\max_c \text{Vote}(c)$. In cases where multiple classes receive the same number of votes, we break ties by selecting the class with the highest cumulative similarity among the supporting samples. This strategy effectively treats classification as a lexicographical optimization problem: first prioritizing consensus count, and then similarity magnitude.

Confidence Estimation. To facilitate the subsequent conflict resolution step (Sec. 3.3), we compute a semantic confidence score s_p^{sem} for the proposal. This score is defined as the average cosine similarity of the neighbors belonging to the winning class $\hat{c}_p$:

$$s_p^{sem} = \frac{1}{\text{Vote}(\hat{c}_p)} \sum_{(v_j, y_j) \in \mathcal{N}_K(q_p), y_j = \hat{c}_p} \text{sim}(q_p, v_j). \tag{9}$$

This metric effectively quantifies the reliability of the prediction by measuring the density of the feature manifold around the query.

Conflict Resolution and Rasterization Given the dense generation of object proposals, the candidate set inherently contains overlapping proposals. We employ a two-step strategy to consolidate these candidates into a non-overlapping semantic segmentation map.

First, we apply Class-Specific Non-Maximum Suppression (NMS) to eliminate redundant detections. Specifically, we perform NMS independently for each predicted class $\hat{c}_p$ and within each group using an Intersection-over-Union (IoU) threshold τ_{nms}. This process removes duplicate predictions while preserving valid spatial overlaps between distinct categories (*e.g.*, a person riding a bicycle).

Second, to resolve pixel-level conflicts between categories, we employ confidence priority rasterization. We rank the remaining candidates in descending order based on their semantic confidence scores s_p^{sem} (derived in Sec. 3.3). The final semantic map S_{final} is constructed by sequentially assigning pixels to the highest-ranking candidate:

$$S_{final}(u, v) = \hat{c}_p \quad \text{if } (u, v) \in m_p \text{ and } S_{final}(u, v) = \emptyset. \tag{10}$$

Here, $S_{final}(u, v) = \emptyset$ denotes that the pixel at (u, v) is currently unassigned. This "first-come-first-served" mechanism ensures that high-confidence predictions take precedence, handling occlusions and resolving boundary ambiguities.

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We primarily evaluate ModuSeg on two standard WSSS benchmarks: PASCAL VOC 2012 [8] (10,582 train, 1,449 val, 1,456 test images) and MS COCO 2014 [12] (82,081 train, 40,137 val images). Accordingly, our main experimental results are reported on VOC and COCO. ADE20K

Table 1: Segmentation comparison with state-of-the-art methods on VOC and COCO. Train: ✓=Required, ✗=Training-Free.

Method	Train	VOC (mIoU) Val	VOC (mIoU) Test	COCO (mIoU) Val
Single-stage methods				
ToCo (CVPR'23) [22]	✓	71.1	72.2	42.3
DuPL (CVPR'24) [30]	✓	73.3	72.8	44.6
WeCLIP (CVPR'24) [40]	✓	76.4	77.2	47.1
PCRE (CVPR'25) [33]	✓	75.5	75.9	47.2
MoRe (AAAI'25) [38]	✓	76.4	75.0	47.4
ExCEL (CVPR'25) [37]	✓	78.4	78.5	50.3
TokenMasking (ICCV'25) [9]	✓	72.7	73.5	43.2
SSR (AAAI'26) [2]	✓	79.5	79.6	50.6
Multi-stage methods				
CLIP-ES (CVPR'23) [13]	✓	73.8	73.9	45.4
MCTformer+ (TPAMI'24) [32]	✓	74.0	73.6	45.2
CPAL (CVPR'24) [25]	✓	74.5	74.7	46.8
S2C (CVPR'24) [10]	✓	78.2	77.5	49.8
WeakCLIP (IJCV'25) [47]	✓	74.0	73.8	47.4
POT (CVPR'25) [28]	✓	76.1	76.7	47.9
OTPL (ICCV'25) [27]	✓	76.4	76.8	47.6
VPL (AAAI'25) [34]	✓	79.3	79.0	49.8
ModuSeg (Ours)	✗	**86.3**	**86.6**	**56.7**

[44] and Cityscapes [6] are used for generalization analysis. We adopt mean Intersection-over-Union (mIoU) as the primary metric.

Implementation Details. Our framework is implemented in PyTorch [15] using the pre-trained C-RADIOv4-SO400M [19] backbone and CorrCLIP [41] for pseudo-mask generation. To purify the feature bank, we perform morphological erosion using a 3×3 kernel for 20 iterations and discard the 25% of features with the largest prototype distances. During inference, we utilize EntitySeg [17] for mask proposals and set the retrieval neighbor count to $K = 25$. For the retrieval backend, we employ the FAISS library [7], adopting a deterministic flat index for PASCAL VOC to ensure reproducibility, while utilizing a GPU-accelerated inverted file index (IVF) for MS COCO to accelerate inference on the large-scale dataset. More details are provided in the supplementary material.

4.2 Comparisons with State-of-the-Art Methods

Performance of Semantic Segmentation. Tab. 1 demonstrates the superior performance of ModuSeg. Recent state-of-the-art methods leverage foundation models for dense prediction, including VLM-based approaches such as WeCLIP and ExCEL, and SAM-based methods like S2C that employ SAM to refine CAMs during training. However, these coupled optimization strategies

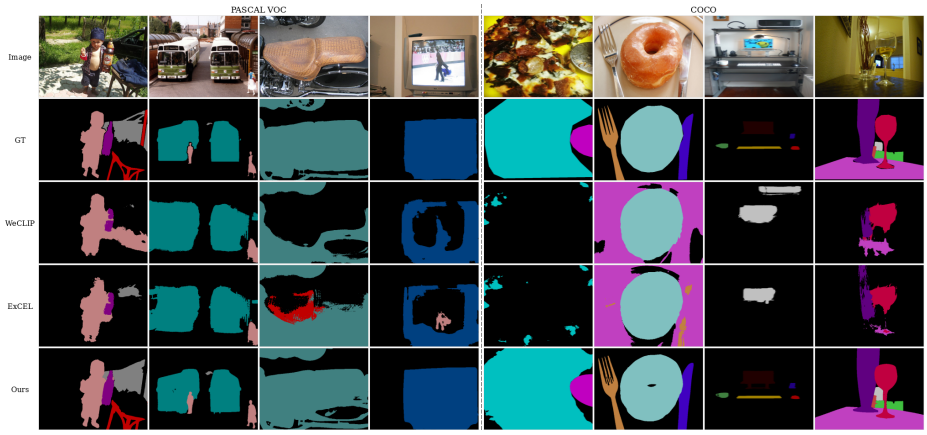

Fig. 3: Qualitative comparison on VOC and COCO. ModuSeg demonstrates superior object completeness and boundary adherence compared to WeCLIP and ExCEL. Our method successfully delineates fine-grained details while eliminating the semantic inconsistencies and noise observed in competing approaches.

often suffer from noisy activation seeds or the coarse spatial resolution inherent in CLIP features. By decoupling object discovery from semantic retrieval, ModuSeg achieves **86.3%** and **86.6%** mIoU on VOC val and test sets, and **56.7%** on COCO, surpassing the previous best-performing methods by 6.8%, 7.0% and 6.1%, respectively. These results indicate that combining pretrained geometric and semantic priors through a decoupled proposal-retrieval pipeline is highly effective for WSSS. Qualitative comparisons in Fig. 3 further substantiate that our paradigm produces integral masks with precise boundaries, effectively suppressing the fragmentation and noisy artifacts inherent to baseline methods. Additional qualitative visualizations and interpretability analyses are provided in the supplementary material.

Evaluation of Initial Seed Quality. Tab. 2 assesses the quality of initial seed on the VOC training set. By integrating image-level label filtering into the CorrCLIP backbone, our approach achieves **78.8%** mIoU. This result establishes a new benchmark for seed generation, outperforming ExCEL by 0.8% and effectively suppressing hallucinated categories during feature bank construction.

4.3 Ablation Studies

Effectiveness of Key Components. All ablations use validation sets unless otherwise specified. Tab. 3 presents the quantitative ablation results. The baseline with C-RADIOv4 and EntitySeg achieves 84.3% mIoU. Incorporating Soft-Masked Feature Aggregation (SMFA) alleviates boundary ambiguity during feature extraction and improves performance to 84.6%. Semantic Boundary Purification (SBP) yields a substantial gain of 0.9% by eroding uncertain regions to suppress feature mixing. The combination of both modules produces a

Table 2: Comparison of initial seed quality (CAMs or pseudo-masks) on VOC Train set. Since ModuSeg generates masks directly, we evaluate our generated pseudo-masks against the CAM seeds of state-of-the-art methods. **Sup.** denotes supervision: Image-level ($\mathcal{I}$), CLIP ($\mathcal{L}$), DINO ($\mathcal{D}$). **Arch.**: ViT (V), ResNet (R), DeiT (D).

Method	Sup.	Arch.	VOC Train (mIoU)
CLIP-ES (CVPR'23) [13]	$\mathcal{I}+\mathcal{L}$	V	70.8
ToCo (CVPR'23) [22]	$\mathcal{I}$	V	73.6
CPAL (CVPR'24) [25]	$\mathcal{I}+\mathcal{L}$	R	71.9
CTI (CVPR'24) [39]	$\mathcal{I}$	D	69.5
SeCo (CVPR'24) [36]	$\mathcal{I}$	V	74.8
DuPL (CVPR'24) [30]	$\mathcal{I}$	V	76.0
WeCLIP (CVPR'24) [40]	$\mathcal{I}+\mathcal{L}$	V	75.4
WeakCLIP (IJCV'25) [47]	$\mathcal{I}+\mathcal{L}$	V	61.7
VPL (AAAI'25) [34]	$\mathcal{I}+\mathcal{L}$	V	77.8
ExCEL (CVPR'25) [37]	$\mathcal{I}+\mathcal{L}$	V	78.0
SSR (AAAI'26) [2]	$\mathcal{I}+\mathcal{L}+\mathcal{D}$	V	78.7
ModuSeg (Ours)	$\mathcal{I}+\mathcal{L}+\mathcal{D}$	V	**78.8**

Table 3: Ablation study of key components on VOC val set. SBP and SMFA demonstrate synergistic effects for robust prototype learning.

Method	SBP	SMFA	mIoU
Baseline			84.3
w/ SMFA		✓	84.6
w/ SBP	✓		85.2
ModuSeg	✓	✓	**86.3**

synergistic effect, reaching a peak of 86.3% mIoU. This confirms that purified boundaries paired with soft aggregation are essential for robust prototype learning. A per-category SBP analysis is provided in the supplementary material.

Impact of Feature Bank Construction Strategies and Backbone Scalability. Tab. 4 (Left) indicates that filtering is crucial for maintaining prototype integrity. Without it, background noise degrades mask quality for CorrCLIP and Mask Adapter. Our logits-based filtering removes this noise and improves mask mIoU. Tab. 4 (Right) validates the scalability of ModuSeg. Performance scales consistently with backbone strength from DINOv2 to C-RADIOv4, indicating that powerful pre-trained foundation representations are an important source of the gains and that our decoupled design leverages stronger backbones.

Scalability and Generalizability. Tab. 5 demonstrates that ModuSeg's decoupled architecture can derive downstream segmentation gains from suitable pre-trained components while maintaining strong generalizability across different base models. For mask proposals, Tab. 5 (Left) compares class-agnostic SAM 2 against EntitySeg. While SAM 2 excels at geometric delineation, EntitySeg

Table 4: Impact of feature bank construction strategies and backbone scalability. *Left*: Effect of logits-based filtering on pseudo-mask mIoU over the training sets. *Right*: Performance scaling with backbone strength.

Model	Filtering	mIoU (%) VOC	COCO
CorrCLIP	w/o Filter	68.7	42.5
	w/ Filter	**78.8**	**54.6**
Mask Adapter	w/o Filter	76.6	55.1
	w/ Filter	**82.4**	**60.2**

Backbone	Size	mIoU (%) VOC	COCO
DINOv2	ViT-B	82.9	51.0
DINOv3	ViT-B	84.7	53.5
DINOv3	ViT-L	85.3	55.8
C-RADIOv4	SO400M	**86.3**	**56.7**

Table 5: Scalability and generalizability of ModuSeg. *Left*: Comparison of mask proposal sources. *Right*: Consistent plug-and-play gains across different models and datasets. "Direct" denotes zero-shot inference using the base model, while "+ Ours" indicates using the base model for pseudo-mask generation within our framework. ADE and City. denote ADE20K and Cityscapes, respectively.

Source	mIoU VOC	COCO
SAM 2	82.6	54.8
EntitySeg	**86.3**	**56.7**

Model	Method	VOC	COCO	ADE	City.
CorrCLIP	Direct	74.8	45.0	28.6	51.6
	+ **Ours**	**86.3**	**56.7**	**44.7**	**66.0**
Mask Adapter	Direct	81.2	55.4	38.2	55.0
	+ **Ours**	**86.6**	**59.3**	**45.6**	**64.2**

demonstrates superior instance-level consistency, serving as a more suitable pretrained geometric prior for our retrieval framework. Tab. 5 (Right) highlights ModuSeg's robustness across base models and datasets. The consistent gains on ADE20K and Cityscapes further support its generalizability across diverse datasets. ModuSeg also improves both correlation-based CorrCLIP and supervised Mask Adapter. Although Mask Adapter often yields stronger performance, our default configuration uses CorrCLIP to maintain a fully training-free pipeline with competitive results.

4.4 Deep Analysis

Hyper-parameter Analysis. We provide comprehensive ablation studies of key hyper-parameters—including the Top-K retrieved neighbors, feature outlier removal ratio, and mask erosion iterations—in the supplementary material.

Performance Bounds. Tab. 6 (Left) reveals that substituting the pseudo-mask bank with ground truth annotations yields only marginal gains on VOC. This suggests that the quality of our retrieved features is comparable to that of supervised counterparts, implying that the feature bank is not the primary bottleneck. Conversely, employing ground truth during the inference segmentation stage substantially boosts performance to 95.7%. This significant gap indicates that the limitation lies primarily in the class-agnostic segmenter, which lacks

Table 6: Internal property analysis. *Left*: The oracle analysis of bank quality and inference segmentation. *Right*: Data efficiency with samples per category (N).

Setting	Bank Source	Infer Source	mIoU (%) VOC	mIoU (%) COCO
Ours	Pseudo	Entity	**86.3**	**56.7**
Oracle Bank	GT	Entity	86.7	62.5
Oracle Seg	Pseudo	GT	95.7	82.1
Oracle Sys	GT	GT	95.7	84.9

N-Shot (Img/Cls)	mIoU (%) VOC	mIoU (%) COCO
5	10.7	14.9
10	46.2	29.3
20	69.8	41.4
50	80.2	52.3
100	85.7	54.7
Full	**86.3**	**56.7**

Table 7: Training efficiency comparison on VOC train set. Type $\mathcal{M}$ denotes multi-stage methods and $\mathcal{S}$ denotes single-stage methods. All models are evaluated on RTX 3090.

Method	Type	Time (min)	GPU (G)	Val (mIoU)
CLIMS (CVPR'22) [31]	$\mathcal{M}$	1068	18.0	70.4
CLIP-ES (CVPR'23) [13]	$\mathcal{M}$	420	12.0	72.2
MCTformer+ (TPAMI'24) [32]	$\mathcal{M}$	1496	18.0	74.0
SeCo (CVPR'24) [36]	$\mathcal{S}$	407	17.6	74.0
WeCLIP (CVPR'24) [40]	$\mathcal{S}$	270	6.2	76.4
POT (CVPR'25) [28]	$\mathcal{M}$	678	-	76.1
ModuSeg	$\mathcal{M}$	**84**	**5.3**	**86.3**

specific semantic guidance. Specifically, the segmenter often over-partitions complex backgrounds into discrete entities based on low-level geometric and textural cues. This behavior creates a substantial "semantic gap" compared to human-annotated ground truth, thereby hindering perfect alignment. Fully supervised counterparts are discussed in the supplementary material.

Data Efficiency. We further evaluate the data requirements in Tab. 6 (Right). As shown, the model performance improves significantly as the number of samples (N) increases. Notably, the setting with only 50 shots per category retains approximately 93% of the full-data performance on VOC. This demonstrates that our framework is highly data-efficient, effectively capturing the semantic distribution of categories even with limited support examples.

Training Efficiency Comparison. Tab. 7 compares the computational costs for model training. Unlike optimization-based methods like CLIMS and MCTformer+ that necessitate extensive GPU training, ModuSeg operates in a training-free manner without backpropagation. Consequently, ModuSeg achieves the lowest measured training time and the highest validation mIoU among the compared methods, highlighting the superior efficiency-performance trade-off of our retrieval-based paradigm. Online inference latency is further reported in the supplementary material.

5 Conclusion

In this paper, we propose ModuSeg, a training-free framework that explicitly decouples object discovery from semantic retrieval for weakly supervised semantic segmentation. By delegating localization to general mask proposals and constructing a robust offline feature bank, our approach circumvents the complexity of coupled optimization paradigms. Furthermore, our semantic boundary purification and soft-masked feature aggregation strategies effectively mitigate structural noise and quantization errors inherent in pseudo-labels. As demonstrated, ModuSeg achieves competitive performance and superior boundary adherence without parameter fine-tuning, offering a valuable perspective for leveraging foundation models in WSSS.

Acknowledgements

This work was partially supported by the Guangdong Major Project of Basic and Applied Basic Research (2019B030302002), the National Natural Science Foundation of China (U24B20151), the Science and Technology Project of Guangdong Province (2021B1111600001), and the Major Key Project of PCL under Grants PCL2025A13 and PCL2025A08.

References

1. Ahn, J., Kwak, S.: Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4981–4990 (2018)
2. Bi, X., Xiao, D., Fan, J., Xiao, B.: Ssr: Semantic and spatial rectification for clip-based weakly supervised segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 2444–2452 (2026)
3. Chen, T., Mai, Z., Li, R., Chao, W.l.: Segment anything model (sam) enhanced pseudo labels for weakly supervised semantic segmentation. arXiv preprint arXiv:2305.05803 (2023)
4. Chen, Z., Sun, Q.: Extracting class activation maps from non-discriminative features as well. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3135–3144 (2023)
5. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: 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)
6. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
7. Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.E., Lomeli, M., Hosseini, L., Jégou, H.: The faiss library. IEEE Transactions on Big Data (2025)

8. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
9. Hanna, J., Borth, D.: Know your attention maps: Class-specific token masking for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23763–23772 (2025)
10. Kweon, H., Yoon, K.J.: From sam to cams: Exploring segment anything model for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19499–19509 (2024)
11. Li, Y., Cheng, T., Feng, B., Liu, W., Wang, X.: Mask-adapter: The devil is in the masks for open-vocabulary segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 14998–15008 (2025)
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014)
13. Lin, Y., Chen, M., Wang, W., Wu, B., Li, K., Lin, B., Liu, H., He, X.: Clip is also an efficient segmenter: A text-driven approach for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15305–15314 (2023)
14. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
15. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
16. Pinheiro, P.O., Collobert, R.: From image-level to pixel-level labeling with convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1713–1721 (2015)
17. Qi, L., Kuen, J., Shen, T., Gu, J., Li, W., Guo, W., Jia, J., Lin, Z., Yang, M.H.: High quality entity segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4047–4056 (October 2023)
18. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
19. Ranzinger, M., Heinrich, G., McCarthy, C., Kautz, J., Tao, A., Catanzaro, B., Molchanov, P.: C-radiov4 (tech report). arXiv preprint arXiv:2601.17237 (2026)
20. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
21. Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16846–16855 (2022)
22. Ru, L., Zheng, H., Zhan, Y., Du, B.: Token contrast for weakly-supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3093–3102 (2023)
23. Shi, Y., Dong, M., Xu, C.: Harnessing vision foundation models for high-performance, training-free open vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 23487–23497 (2025)

24. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
25. Tang, F., Xu, Z., Qu, Z., Feng, W., Jiang, X., Ge, Z.: Hunting attributes: Context prototype-aware learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3324–3334 (2024)
26. Tang, Q., Liu, C., Liu, F., Jiang, J., Zhang, B., Chen, C.P., Han, K., Wang, Y.: Rethinking feature reconstruction via category prototype in semantic segmentation. IEEE Transactions on Image Processing **34**, 1036–1047 (2025)
27. Wang, J., Dai, T., Zhang, B., Yu, S., Lim, E.G., Xiao, J.: Class token as proxy: Optimal transport-assisted proxy learning for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21645–21654 (2025)
28. Wang, J., Dai, T., Zhang, B., Yu, S., Lim, E.G., Xiao, J.: Pot: Prototypical optimal transport for weakly supervised semantic segmentation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 15055–15064 (2025)
29. Wu, Y., Li, X., Li, J., Yang, K., Zhu, P., Zhang, S.: Dino is also a semantic guider: Exploiting class-aware affinity for weakly supervised semantic segmentation. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 1389–1397 (2024)
30. Wu, Y., Ye, X., Yang, K., Li, J., Li, X.: Dupl: Dual student with trustworthy progressive learning for robust weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3534–3543 (2024)
31. Xie, J., Hou, X., Ye, K., Shen, L.: Clims: Cross language image matching for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4483–4492 (2022)
32. Xu, L., Bennamoun, M., Boussaid, F., Laga, H., Ouyang, W., Xu, D.: Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation. IEEE transactions on pattern analysis and machine intelligence **46**(12), 8380–8395 (2024)
33. Xu, X., Zhang, P., Huang, W., Shen, Y., Chen, H., Lin, J., Li, W., He, G., Xie, J., Lin, S.: Weakly supervised semantic segmentation via progressive confidence region expansion. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9829–9838 (2025)
34. Xu, Z., Tang, F., Chen, Z., Su, Y., Zhao, Z., Zhang, G., Su, J., Ge, Z.: Toward modality gap: Vision prototype learning for weakly-supervised semantic segmentation with clip. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9023–9031 (2025)
35. Xuan, X., Deng, Z., Ma, K.L.: Reme: A data-centric framework for training-free open-vocabulary segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20954–20965 (2025)
36. Yang, Z., Fu, K., Duan, M., Qu, L., Wang, S., Song, Z.: Separate and conquer: Decoupling co-occurrence via decomposition and representation for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3606–3615 (2024)
37. Yang, Z., Meng, Y., Fu, K., Tang, F., Wang, S., Song, Z.: Exploring clip's dense knowledge for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20223–20232 (2025)

38. Yang, Z., Meng, Y., Fu, K., Wang, S., Song, Z.: More: Class patch attention needs regularization for weakly supervised semantic segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 9400–9408 (2025)
39. Yoon, S.H., Kwon, H., Kim, H., Yoon, K.J.: Class tokens infusion for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3595–3605 (2024)
40. Zhang, B., Yu, S., Wei, Y., Zhao, Y., Xiao, J.: Frozen clip: A strong backbone for weakly supervised semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3796–3806 (2024)
41. Zhang, D., Liu, F., Tang, Q.: Corrclip: Reconstructing patch correlations in clip for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 24677–24687 (2025)
42. Zhang, D., Tang, Q., Liu, F., Mei, H., Chen, C.P.: Exploring token-level augmentation in vision transformer for semi-supervised semantic segmentation. IEEE Signal Processing Letters (2025)
43. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2921–2929 (2016)
44. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ade20k dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 633–641 (2017)
45. Zhou, C., Loy, C.C., Dai, B.: Extract free dense labels from clip. In: European conference on computer vision. pp. 696–712. Springer (2022)
46. Zhou, T., Wang, W., Konukoglu, E., Van Gool, L.: Rethinking semantic segmentation: A prototype view. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2582–2593 (2022)
47. Zhu, L., Wang, X., Feng, J., Cheng, T., Li, Y., Jiang, B., Zhang, D., Han, J.: Weakclip: Adapting clip for weakly-supervised semantic segmentation. International Journal of Computer Vision **133**(3), 1085–1105 (2025)