Improved Weakly Supervised Semantic Segmentation with a Relationally Optimized Prototype Memory Bank Framework

Xiang Zhong, Qiudan Zhang, Jianmin Jiang *

College of Computer Science & Software Engineering,
Shenzhen University, Shenzhen 518060, China
zhongxiang2023@email.szu.edu.cn, qiudanzhang@szu.edu.cn,
jianmin.jiang@szu.edu.cn

Abstract

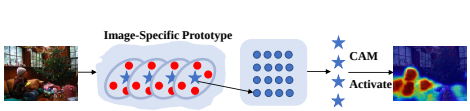
Confronted by the existing prototype-based methods that often fail to capture complex intra-class variations and inter-class semantic relationships, weakly supervised semantic segmentation (WSSS) faces persistent challenges due to the inherent incompleteness and boundary inaccuracies of Class Activation Maps (CAMs). To tackle these critical limitations, we propose a Relationally Optimized Prototype Memory Bank (RO-PMB) framework in this paper, which builds upon multi-prototype representations and uses Graph Neural Networks (GNNs) to organize globally maintained prototypes into a unified graph for propagating positive visual and contextual associations. Specifically, our Prototype Relationship Diagram Construction (PRDC) module leverages GNNs for contextual message passing over a learned prototype graph, enriching global prototypes with crucial associative knowledge. As a result, a subsequent context-aware refinement process is sustained to inject structured semantic information into local image-specific prototypes, thereby generating CAMs with improved completeness and boundary accuracy. Extensive experiments validate that, compared with the existing representative state-of-the-art in multi-stage WSSS baselines, our proposed RO-PMB achieves competitive performance on a range of challenging benchmarks, such as PASCAL VOC and MS COCO.

1 Introduction

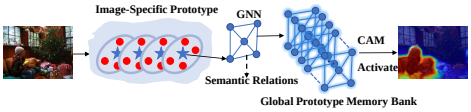
Weakly supervised semantic segmentation (WSSS) aims to alleviate the substantial annotation burden of pixel-level labels by leveraging weaker forms of supervision, such as image-level labels. In this challenging task, Class Activation Maps (CAMs) [1] have emerged as a cornerstone, providing initial coarse localization cues for objects within an image. These cues are subsequently refined into pseudo-labels for training a segmentation model [2]. Despite their widespread adoption, conventional CAMs primarily focus on the most discriminative parts of objects, leading to inherent incompleteness and sparse spatial supervision [3, 4]. This limitation results in insufficient guidance for the segmentation model, thereby hindering its performances. To address the issue of incomplete CAMs, prototype-based methods [5–7] have garnered increasing attention. These approaches typically construct prototypes for each class by pooling features from activated regions. Subsequently, these prototype-derived CAMs, generated via cosine similarity between pixel features and prototypes, tend to be more complete than the existing classifier-based CAMs, leading to improved segmentation performances.

*Corresponding author.

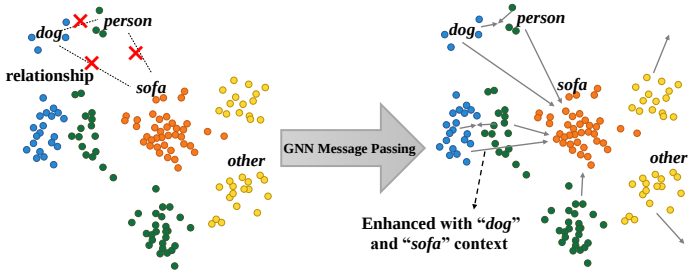
40th Conference on Neural Information Processing Systems (NeurIPS 2026).



(a) Single-Prototype Formulation.



(b) Our Prototype-Based Method.



(c) Comparison of Relationship Diagrams Before and After GNN.

Figure 1: Comparative illustration between a single-prototype formulation and our proposed approach for WSSS. Our method explicitly models positive supportive associations among global prototypes via a GNN, leading to context-aware prototype activation for CAM generation.

However, the existing prototype-based methods [6, 8] represent each category using a single prototype within an image, as conceptually illustrated in Fig. 1a. This simplifying assumption struggles to capture the inherent diversity of attributes and the substantial intra-class variations prevalent in real-world datasets [9]. For instance, a single "*sofa*" prototype may only activate its most salient features, failing to encompass its entire presentation or diverse appearances. Consequently, even features belonging to the same category but diverging from the single prototype may be only partially activated or entirely ignored. This limitation leads to residual incompleteness in prototype-based CAM predictions, which is particularly pronounced when handling objects with complex appearances or diverse scenes [10]. Such incomplete activation directly propagates to the generated pseudo-labels, ultimately constraining the segmentation model's ability to learn robust and comprehensive object boundaries.

To overcome this critical limitation, we propose a novel approach for maintaining a representative and dynamically adaptive set of prototypes that fully encapsulates the visual feature manifold of a category, thereby capturing diverse features within and across images. Inspired by the momentum-contrast (MoCo) framework [11], which constructs large-scale, consistent dynamic dictionaries, we discovered that image-specific prototypes generated directly from individual image features often exhibit instability. This instability arises from minor local variations and prevents the consistent capture of core category semantics in the absence of global supervision and smoothing mechanisms. By drawing parallels with MoCo's memory mechanism, we construct a category-level global prototype set. This set encompasses richer semantic variations within each category, extending beyond what a single image can express, and serves as a robust dictionary for querying and refining image-specific prototypes. Furthermore, we ensure that prototypes within this global set are stable and consistent, representing the core semantics of their respective categories while evolving smoothly during training to avoid significant fluctuations.

In this paper, we introduce a novel WSSS framework named Relationally Optimized Prototype Memory Bank (RO-PMB), which focuses on graph-based interaction among globally maintained prototypes. This framework resolves the aforementioned constraints by maintaining dynamically updated, relation-aware global prototype memory banks for each category. Unlike the existing approaches that rely on a single static prototype or simple feature similarity-based aggregation, RO-PMB's core innovation lies in organizing globally maintained prototypes from all categories into a unified relational graph for positive supportive interaction, as conceptually illustrated in Fig. 1b. Within this graph structure, we introduce Graph Neural Networks (GNNs) for message passing (illustrated in Fig. 1c), enabling each global prototype in the memory bank to fuse information from its semantic neighbors and become context-aware.

Specifically, for each input image, we first generate a series of image-specific prototypes to capture local feature variability. These local prototypes are then queried and receive optimizations from the relationship optimized global prototypes in the memory bank, effectively infusing rich structured semantic global knowledge into local representations. Concurrently, the original global prototypes in

the memory bank undergo momentum updates via newly input image-specific prototypes, ensuring the repository maintains dynamic adaptability and accurately reflects the evolving data distribution. This continuous interaction between image-specific prototypes and relation-enhanced global prototypes, coupled with deep semantic association modeling via GNNs, enables RO-PMB to generate significantly more complete, finely-bounded, and semantically coherent CAMs. These high-quality CAMs provide an unprecedented augmented supervision mechanism for semantic segmentation tasks, promising substantial improvements in model segmentation performance.

Our contributions are highlighted as follows:

- We propose a Relationally Optimized Prototype Memory Bank (RO-PMB), a framework that organizes globally maintained prototypes from all categories into a unified graph for relation-aware interaction. The multi-prototype memory provides the representation basis for bidirectional local–global refinement.
- We introduce the Prototype Relationship Diagram Construction (PRDC) module, which models positive visual and contextual associations among global prototypes as a graph structure. By employing Graph Neural Networks (GNNs) for contextual message passing, PRDC enables prototypes to encode rich associative knowledge, providing a structured solution for robustly capturing intra-class diversity and inter-class semantic associations.
- Extensive experiments on challenging WSSS benchmarks (PASCAL VOC 2012 and MS COCO 2014) demonstrate that our RO-PMB framework consistently generates more complete, coherent, and accurate CAMs, delivering new state-of-the-art performances.

2 Related Work

2.1 Class Activation Maps

While CAMs [1] are foundational for WSSS, providing localization cues, early methods [12–14] suffer from incompleteness and sparsity, focusing only on discriminative parts. To mitigate these limitations, various CAM refinement strategies have been developed. A prominent approach involves pixel-level contextual mining, often implemented through techniques such as adversarial erasing [15], self-supervised training [16], and contrastive representation learning [17] etc. Another significant research direction focuses on affinity-based propagation, where an initial sparse CAM is expanded to encompass entire objects by leveraging feature similarity. While these methods markedly improve CAM completeness, they are still limited by predefined rules and struggle with complex intra-class variances.

2.2 Prototype Learning for WSSS

Aiming at overcoming CAM limitations by constructing class prototypes from activated features, earlier methods like CPD [7] and PSDPM [5] typically learn a single prototype per class, either globally or on a per-image basis, and subsequently expand CAMs using these prototypes. SFC [6] further employs prototype-based consistency to calibrate shared features. Beyond single-prototype formulations, LPCAM [18] clusters class features into multiple local prototypes, while SMC [19] maintains multiple foreground and background prototypes in memory and adaptively selects prototype classifiers for each image. MuP-VSS [20] instead uses learnable class-token queries for prototype–patch interaction. Multiple prototypes and prototype memory banks have therefore already been explored. Our focus is whether globally maintained prototypes from different categories can be explicitly organized and jointly refined through graph-based interaction before image-specific localization. RO-PMB addresses this through positive relational propagation followed by bidirectional local–global refinement.

2.3 Graph Neural Networks in Segmentation

Having demonstrated remarkable capabilities in modeling relationships and propagating information within structured data, GNNs [21, 22] have been effectively applied to various computer vision tasks, including object detection [23], point cloud classification [24], and semantic segmentation [25]. For instance, some works utilize GNNs to construct super-pixel graphs on image features, facilitating contextual information propagation among spatially adjacent regions [26]. Others employ GNNs to

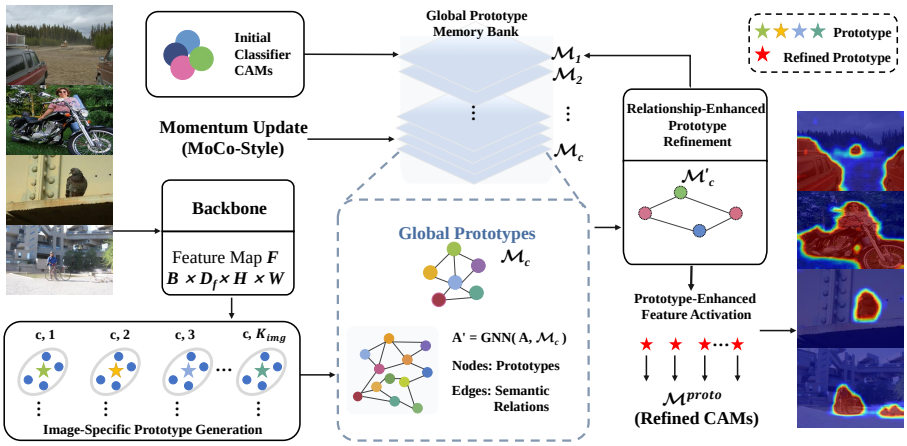


Figure 2: Overview of our proposed RO-PMB framework for WSSS, which integrates a Global Prototype Memory Bank with GNNs to explicitly model positive inter-prototype associations, enabling relationship optimized prototype refinement for generating robust and context-aware CAMs.

model relationships between object instances for panoptic segmentation [27] or to refine segmentation masks by capturing long-range dependencies [28]. However, their potential to explicitly model and reason about semantic relationships among class prototypes within WSSS remains largely unexplored, representing a significant opportunity for our work to enhance global semantic context utilization.

3 The Proposed Method

The overall architecture of our Relationally Optimized Prototype Memory Bank (RO-PMB) framework is illustrated in Fig. 2, which is designed to leverage explicit semantic relationships among global prototypes to generate more accurate and complete CAMs for WSSS. Prototype construction provides the representation basis, whereas PRDC constitutes the primary methodological contribution through graph-based interaction among globally maintained prototypes.

3.1 Image-Specific Prototype Generation

For an input image, we first extract its high-level feature map $F \in \mathbb{R}^{D_f \times H \times W}$ using a backbone encoder, where D_f, H, and W represent the feature dimension, height, and width, respectively. Concurrently, a classifier module generates initial CAMs. These CAMs highlight discriminative regions for each category, but as mentioned in the introduction, they are typically incomplete.

To capture image-specific feature distributions, we generate a set of prototypes for each class present in the image. For each class $c \in \{1, \ldots, C\}$, we identify activated regions in F, where C is the number of foreground classes. Within these activated regions, we apply a K-means clustering mechanism to group similar pixel features into K_{img} image-specific sub-clusters. For each sub-cluster k, a local prototype $\mathbf{P}_{c,k} \in \mathbb{R}^{D_f}$ is computed, typically as the mean of its constituent features. These image-specific prototypes collectively represent the diverse attributes of class c within the current image as follows:

$$\mathbf{P}_c = \text{Proj}(\{\mathbf{P}_{c,1}, \mathbf{P}_{c,2}, \ldots, \mathbf{P}_{c,K_{img}}\}), \tag{1}$$

where K_{img} is the number of image-specific prototypes for class c. Proj denotes mapping the dimension of $\mathbf{P}_c$ to D_p. These initial prototypes are local and may fluctuate due to variations in image appearance, necessitating further injection of global knowledge.

3.2 Global Prototype Memory Bank Management

The Relationship Optimized Prototype Memory Bank $\mathcal{M}$ provides the representation basis for PRDC and maintains a set of globally representative prototypes for each class across the entire dataset. $\mathcal{M}$ consists of C individual memory banks $\mathcal{M}_c$, one for each foreground class c. Each $\mathcal{M}_c$ stores K_g prototype vectors, $\mathbf{m}_{c,k_g} \in \mathbb{R}^{D_p}$ as follows:

$$\mathcal{M}_c = \{\mathbf{m}_{c,1}, \mathbf{m}_{c,2}, \ldots, \mathbf{m}_{c,K_g}\}. \tag{2}$$

Initially, these global prototypes can be initialized with the averaged features from a pre-trained model. This stage involves two substeps: Refinement and Update, which are designed to enhance local prototypes using knowledge from the global repository and ensure continuous evolution.

3.3 Prototype Relationship Diagram Construction

As shown in Fig. 2, we integrated a single-layer GCN [29] as GNN module using global prototypes $\mathcal{M}_c \in \mathbb{R}^{(C \cdot K_g) \times D_p}$ as nodes and learned the relationships between them to form the proposed RO-PMB. Specifically, we treat the set $\{\mathbf{m}_{c,k_g}\}_{c=1...C,k_g=1...K_g}$ of all foreground category global prototypes as a node set $\mathcal{V} = \bigcup_c \mathcal{M}_c$ in a graph. We construct a global prototype relationship graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where $\mathcal{E}$ represents the edges between prototypes. Specifically, we construct the graph's adjacency matrix $\mathbf{A} \in \mathbb{R}^{(C \cdot K_g) \times (C \cdot K_g)}$ by computing the cosine similarity between all pairs of global prototypes.

To bridge the gap between low-level visual similarity and high-level semantic associations (e.g., the co-occurrence of "*dog*" and "*grass*"), we redefine the adjacency matrix A by incorporating a Semantic Co-occurrence Prior (SCP). Specifically, we pre-calculate a statistical matrix $\mathbf{Q} \in \mathbb{R}^{C \times C}$ from the training set labels, where we use $\mathbf{Q}_{ij}$ to represent the conditional probability that class j occurs given the presence of class i. Correspondingly, the refined relationship weight A_{ij} can be formulated as:

$$A_{ij} = \text{ReLU}\left(\frac{m_i \cdot m_j}{||m_i|| \cdot ||m_j||}\right) \cdot \mathbf{Q}_{i,j}, \tag{3}$$

where $m_i, m_j \in \mathcal{V}$ are two global prototype vectors. This hybrid formulation ensures that the GNN message passing is guided not only by visual consistency but also by reliable contextual constraints, enabling a "*dog*" prototype to actively refine its representation through the close neighbor "*grass*". A visualization representation of the matrix $\mathbf{Q}$ is presented in Appendix C. To obtain sparse graphs, we apply thresholding to similarity scores. Here we employ the ReLU activation function to ensure non-negative similarity values. Negative cosine similarities are treated as the absence of message passing in the current unsigned GCN, rather than as inhibitory semantic relations. Explicit inhibitory reasoning would require a dedicated signed-graph formulation. Under the current single-layer GCN architecture, to maintain model simplicity and efficiency, we opt to focus on positive semantic connections through ReLU activation.

On graph $\mathcal{G}$, we employ GNNs as the prototype relationship inference module. It takes the global prototype $\mathcal{M}_c$ and the adjacency matrix $\mathbf{A}$ as inputs. Through message passing, it updates the prototype representations to fuse contextual information from the graph structure. For detailed principles, please refer to Appendix B. The global prototype $\mathcal{M}'_c$ can be obtained as follows:

$$\mathcal{M}'_c = \hat{\mathbf{A}} \mathcal{M}_c \mathbf{W}_G. \tag{4}$$

Here, $\hat{\mathbf{A}} = \mathbf{D}^{-\frac{1}{2}} \mathbf{A} \mathbf{D}^{-\frac{1}{2}}$ denotes the normalized adjacency matrix (including self-loops), $\mathbf{D}$ is the degree matrix, and $\mathbf{W}_G$ represents the learnable weights of the GNN layer. After the GNN layers, each global prototype $\mathbf{m}'_i$ integrates information from its neighbors, becoming more context-aware. These relationship optimized global prototypes $\mathcal{M}'_c$ will be used for refinement.

3.4 Prototype Refinement

For each image-specific prototype $\mathbf{P}_{c,k}$ generated in Eq. (1), we refine it using the relationship optimized global prototype $\mathcal{M}'_c$. The goal is to find how well each image-specific prototype aligns with the globally learned representations. Specifically, we first compute the cosine similarity between $\mathbf{P}_{c,k}$ and all relationship optimized global prototypes in $\mathcal{M}'_c$ as follows:

$$S(\mathbf{P}_{c,k}, \mathbf{m}'_{c,k_g}) = \frac{\mathbf{P}_{c,k} \cdot \mathbf{m}'_{c,k_g}}{||\mathbf{P}_{c,k}|| \cdot ||\mathbf{m}'_{c,k_g}||}, \tag{5}$$

where $\mathbf{m}'_{c,k_g}$ denotes the k_gth relationship optimized global prototype for category c. Based on these similarities, we generate a refined prototype $\mathbf{R}_{c,k}$ for each image-specific prototype. This refined prototype is a weighted sum of the relationship optimized global prototypes $\mathcal{M}'_c$, where

weights are determined by the similarity scores. This process effectively smooths out potential noise in image-specific prototypes while infusing them with rich global category knowledge, thereby transferring relation-enhanced global information to image-specific prototypes. Its formula is as follows:

$$\mathbf{R}_{c,k} = \sum_{k_g=1}^{K_g} \alpha_{k_g} \mathbf{m}'_{c,k_g}, \tag{6}$$

where α_{k_g} are the normalized similarity weights, obtained by applying a softmax function over the similarities as follows:

$$\alpha_{k_g} = \frac{\exp(S_{k_g}(\mathbf{P}_{c,k}, \mathbf{m}'_{c,k_g})/\tau)}{\sum_{k'_g=1}^{K_g} \exp(S_{k'_g}(\mathbf{P}_{c,k}, \mathbf{m}'_{c,k'_g})/\tau)}, \tag{7}$$

where τ is a temperature parameter controlling the sharpness of the distribution. In this way, $\mathbf{R}_{c,k}$ not only incorporates global diversity within categories but also indirectly integrates semantic contextual information across categories through GNN message passing, thereby achieving greater robustness and comprehensiveness.

3.5 Global Prototype Memory Bank Update

To ensure that the global memory bank continuously learns from current image-specific prototypes and remains representative, we update its prototypes using a momentum-based mechanism. For each image-specific prototype $\mathbf{P}_{c,k}$, we identify its most similar global prototype in $\mathcal{M}'_c$ as follows:

$$k_g^* = \arg\max_{k_g} S_{k_g}(\mathbf{P}_{c,k}, \mathbf{m}'_{c,k_g}). \tag{8}$$

The selected global prototype $\mathbf{m}_{c,k_g^*}$ is then updated by incorporating the information from the current image-specific prototype $\mathbf{P}_{c,k}$. This update mechanism resembles the key queue update in MoCo [11], ensuring the stability and consistency of the global prototype as follows:

$$\mathbf{m}_{c,k_g^*} \leftarrow \eta \cdot \mathbf{m}_{c,k_g^*} + (1-\eta) \cdot \mathbf{P}_{c,k}, \tag{9}$$

where $\eta \in [0, 1)$ is the momentum coefficient. A high η ensures slow and stable updates, preventing individual noisy samples from overly influencing the global prototypes, while allowing the memory bank to adapt gradually over the course of training. All prototype vectors are L2-normalized after each update. In this way, the memory bank can accumulate and refine category variants from different images, thereby constructing a more comprehensive and stable prototype collection.

3.6 Prototype-Enhanced Feature Activation

Instead of using the raw image-specific prototypes $\mathbf{P}_c$, we leverage the refined prototypes $\mathbf{R}_c = \{\mathbf{R}_{c,1}, \ldots, \mathbf{R}_{c,K_{img}}\}$ from Eq. (6) to generate enhanced CAMs. For each pixel feature $f_{xy} \in \mathbb{R}^{D_f}$ at spatial location (x, y) in the feature map F, and for each class c, we calculate its activation score with respect to all refined prototypes of that class as follows:

$$\text{Act}_{c,k}(x, y) = \text{ReLU}(f_{xy} \cdot \mathbf{R}_{c,k}). \tag{10}$$

These individual activation maps for each refined prototype are then aggregated to form the final prototype-enhanced CAM for class c, denoted as $\mathcal{M}_c^{proto}$ as follows:

$$\mathcal{M}_c^{proto}(x, y) = \sum_{k=1}^{K_{img}} \text{Act}_{c,k}(x, y). \tag{11}$$

The prototype-enhanced CAMs, $\mathcal{M}^{proto} = \{\mathcal{M}_1^{proto}, \ldots, \mathcal{M}_C^{proto}\}$, are then used as high-quality pseudo-labels for training a fully supervised segmentation model and are also directly utilized for weak supervision loss. By utilizing refined prototypes, $\mathcal{M}^{proto}$ is expected to be more complete and accurate than CAMs derived from single prototypes or classifier weights alone, effectively capturing the full extent of foreground objects and mitigating intra-class variations, thereby providing stronger supervisory signals. For network training details, refer to Appendix A.3.

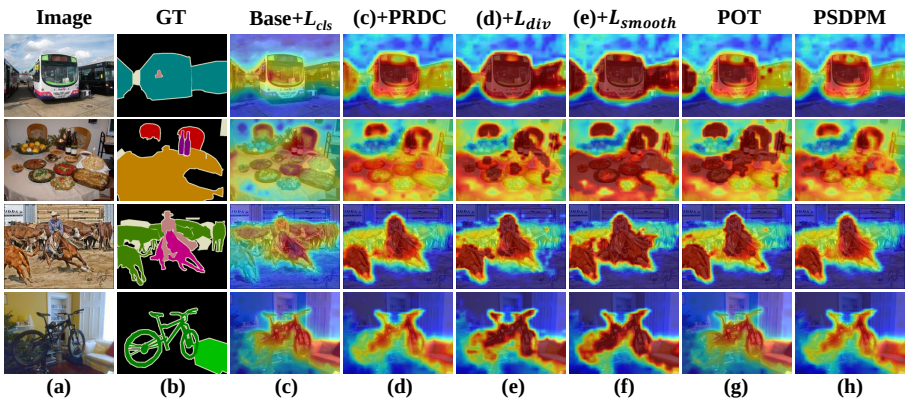


Figure 3: Comparative illustration of qualitative results, where (a) Original image. (b) Ground Truth. (c)-(f) CAM visualizations from ablative modules of RO-PMB. (f)-(h) Qualitative comparisons of (f) RO-PMB and recent prototype-based methods, including (g) POT [10], and (h) PSDPM [5].

4 Experiments

4.1 Datasets

Following the settings reported in the literature published over recent years [13, 30, 31], we conduct extensive experiments on two widely used datasets: PASCAL VOC 2012 [32] and MS COCO 2014 [33]. Following [10], we use ResNet-50 [34] as the CAM generation backbone and CLIP-ES [35] as the offline module for generating the initial CAMs. After the pseudo-labels are generated, a WideResNet-38 [36] segmentation network is trained for evaluating the final semantic segmentation. Following previous approaches [5, 37, 10], DenseCRF [38] is used as the post-processing method to generate pseudo labels. We adopt the mean Intersection-over-Union (mIoU) metric as the evaluation criterion, consistent with the mainstream practices in this field [37, 30].

4.2 Comparative Experiments

In this section, we present a comprehensive evaluation of our proposed RO-PMB framework and compare its performance with several state-of-the-art WSSS methods. Tab. 1 summarizes the performance of CAMs and fine segmentation masks on PASCAL VOC.

Table 1: CAM and pseudo-mask mIoU (%) on the PASCAL VOC 2012 training set. " / " indicates an unreported result.

Method	CAMs	Mask
ACR [39]$_{\text{CVPR' 23}}$	60.3	72.3
FPR [40]$_{\text{ICCV' 23}}$	63.8	/
USAGE[41]$_{\text{ICCV' 23}}$	67.7	72.8
MCT+ [31]$_{\text{TPAMI' 24}}$	68.7	72.7
PSDPM[5]$_{\text{CVPR' 24}}$	/	77.3
DuPL[42]$_{\text{CVPR' 24}}$	/	76.0
KTSE[43]$_{\text{ECCV' 24}}$	67.0	73.8
FFR[37]$_{\text{CVPR' 25}}$	/	76.4
MuP-VSS[20]$_{\text{CVPR' 25}}$	71.7	74.1
Hanna et al. [44]$_{\text{ICCV' 25}}$	/	74.5
POT [10]$_{\text{CVPR' 25}}$	75.0	79.3
RO-PMB(Our)	**76.6**	**80.4**

As shown in Tab. 1, our proposed RO-PMB framework significantly advances the state-of-the-art in weakly-supervised semantic segmentation. The proposed RO-PMB achieves the highest mIoU scores for both CAMs at 76.6% and segmentation masks at 80.4% on the PASCAL VOC dataset. Notably, our method outperforms the previous leading approach, POT [10], by 1.6 in CAMs (**visual comparisons shown in Fig. 3 (f)-(h)**) and 1.1 percentage points in masks. Such superior performances validate the effectiveness of our proposed framework in explicitly modeling semantic relationships between

global prototypes via GNNs to generate more coherent and accurate CAMs, leading to an efficient production of higher quality segmentation masks.

Table 2: Performance comparison of WSSS methods on PASCAL VOC and MS COCO, where Sup. indicates supervision type ($\mathcal{I}$ for image-level, $\mathcal{S}$ for saliency maps and $\mathcal{L}$ for CLIP-text).

Model	Backbone	Sup.	VOC		COCO
			val	test	val
Single-stage WSSS methods					
TSCD[45]$_{\text{AAAI' 23}}$	MIT-B1	$\mathcal{I}$	67.3	67.5	40.1
ToCo[46]$_{\text{CVPR' 23}}$	ViT-B	$\mathcal{I}$	71.1	72.2	42.3
DIAL[47]$_{\text{ECCV' 24}}$	ViT-B	$\mathcal{I}$	74.5	74.9	44.4
DuPL[42]$_{\text{CVPR' 24}}$	ViT-B	$\mathcal{I}$	73.3	72.8	44.6
PFSR[48]$_{\text{AAAI' 24}}$	ViT-B	$\mathcal{I}$	75.7	75.0	45.8
WeCLIP[30]$_{\text{CVPR' 24}}$	ViT-B	$\mathcal{I}+\mathcal{L}$	76.4	77.2	47.1
Multi-stage WSSS methods					
RIB[49]$_{\text{NeurIPS' 21}}$	RN50	$\mathcal{I}$	68.3	68.6	43.8
LPCAM[18]$_{\text{CVPR' 23}}$	RN38	$\mathcal{I}$	72.6	72.4	42.8
USAGE[41]$_{\text{ICCV' 23}}$	RN38	$\mathcal{I}$	71.9	72.8	44.3
PCSS[50]$_{\text{ECCV' 24}}$	RN38	$\mathcal{I}$	73.2	73.0	45.7
PSDPM[5]$_{\text{CVPR' 24}}$	RN50	$\mathcal{I}$	74.1	74.9	47.2
KTSE[43]$_{\text{ECCV' 24}}$	RN101	$\mathcal{I}$	73.0	72.9	45.9
MCT+[31]$_{\text{TPAMI' 24}}$	RN38	$\mathcal{I}$	74.0	73.6	45.2
SMC[19]$_{\text{CAAI TRIT' 25}}$	RN50	$\mathcal{I}+\mathcal{S}$	71.3	72.2	/
FFR[37]$_{\text{CVPR' 25}}$	ViT-B	$\mathcal{I}$	76.0	75.5	46.8
MuP-VSS[20]$_{\text{CVPR' 25}}$	RN38	$\mathcal{I}$	73.6	74.7	46.6
PCRE[51]$_{\text{CVPR' 25}}$	ViT-B	$\mathcal{I}$	75.5	75.9	47.2
POT [10] $_{\text{CVPR' 25}}$	RN50	$\mathcal{I}$	76.1	76.7	47.9
Our	RN38	$\mathcal{I}$	**77.6**	**77.8**	**48.8**

In Tab. 2, further experimental results are listed, which illustrate that our proposed method outperforms a number of existing state-of-the-art methods. Specifically, on PASCAL VOC, our method reaches 77.6% on the validation set and 77.8% on the test set, improves segmentation performance. This represents an improvement of 1.5 on validation and 1.1 percentage points on test compared to the previous best multi-stage method. In addition, on the more challenging MS COCO validation set, our method achieves a remarkable 48.8% mIoU, surpassing the existing best at 47.9% achieved by POT [10]. These results demonstrate the effectiveness and generalization ability of our framework. By refining the CAM generation and refinement, our method consistently promotes the best performing rates.

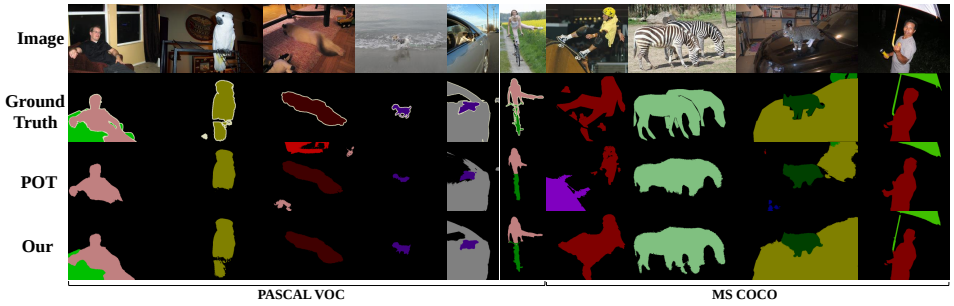


Figure 4: Qualitative illustration of segmentation samples on the validation set of PASCAL VOC and MS COCO.

Fig. 4 illustrates a number of segmentation samples in comparison with the representative existing state-of-the-art method POT. As seen, our proposed RO-PMB model significantly improves the completeness and boundary accuracy of segmentation on both PASCAL VOC and MS COCO. RO-PMB not only overcomes those common issues, including incomplete activations and background confusions, but also enables more accurate and coherent segmentation of objects in complex scenes.

4.3 Ablation Analysis

To fully validate the effectiveness of all the key components in our proposed RO-PMB, we have carried out extensive ablation experiments, all of which are conducted on the PASCAL VOC.

Impact of Individual Components: Fig. 3 (c)-(f) illustrates the ablation experiment results on CAM quality and further details are listed in Tab. 3. As seen in Tab. 3, while the baseline achieves 67.5% using only classification loss $\mathcal{L}_{cls}$, introducing the local-global consistency loss $\mathcal{L}_{loc-glo}$ immediately boosts its performance to 70.1%, demonstrating the initial benefit of aligning image-specific prototypes with global knowledge.

Table 3: Ablation results for individual components of our proposed RO-PMB, where PRDC denotes the Prototype Relationship Diagram Construction Module.

Base	PRDC	$\mathcal{L}_{div}$	$\mathcal{L}_{smooth}$	CAMs
$\mathcal{L}_{cls}$				67.5
$\mathcal{L}_{cls}$ +$\mathcal{L}_{loc-glo}$				70.1
$\mathcal{L}_{cls}$ + $\mathcal{L}_{loc-glo}$	✓			74.2
$\mathcal{L}_{cls}$ + $\mathcal{L}_{loc-glo}$	✓	✓		75.1
$\mathcal{L}_{cls}$ + $\mathcal{L}_{loc-glo}$	✓	✓	✓	76.6

When PRDC is activated, integrating the GNN-based relational querying and refinement, the result is significantly improved to 74.2%. This substantial improvement highlights the critical role of explicitly modeling inter-prototype relationships and refining prototypes with contextual information for generating more robust and accurate CAMs. After $\mathcal{L}_{div}$ is added, the value of mIoU is further increased to 75.1%, validating that encouraging prototypes within the memory bank to capture distinct visual patterns helps prevent prototype collapse and contributes to a more comprehensive representation of each category. Finally, by introducing $\mathcal{L}_{smooth}$, our full RO-PMB model achieves its best mIoU, demonstrating that regularizing the GNN to ensure that connected prototypes have similar features is crucial for learning meaningful graph structures.

Beyond validating individual components, we further investigate the sensitivity of our RO-PMB to key hyper-parameters. The corresponding results are listed in Tab. 4. This analysis is performed on our model to identify optimal settings and understand their impact on both CAMs and final segmentation, details of which are discussed below.

Table 4: Value of parameter η.

η	CAM	Seg.
0	*fail*	*fail*
0.9	75.9	76.4
0.99	76.6	77.6
0.999	76.1	76.8

Figure 5: Impact of K_g.

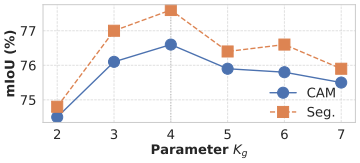
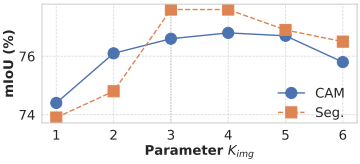


Figure 6: Impact of K_{img}.



Impact of Momentum Coefficient (η): The momentum coefficient η in Eq. (9) is crucial for updating global prototypes. As shown in Tab. 4, using a very high momentum like $\eta = 0.999$ results in suboptimal performance. While $\eta = 0$ fails to leverage historical information, $\eta = 0.999$ makes the prototypes too rigid to adapt to new data. By ensuring a stable yet adaptable global memory bank, a momentum of $\eta = \mathbf{0.99}$ strikes the best balance.

Compatibility with Different Prototype Constructions: The proposed PRDC module is designed to model relational interactions among prototypes. Therefore, its effectiveness should not be limited to the specific prototype construction strategy adopted in RO-PMB. To verify this point, we evaluate whether PRDC can consistently improve other multi-prototype representations under a matched setting. Specifically, we replace the prototype construction module of RO-PMB with two representative prototype-based alternatives: LPCAM-style prototypes and SMC-style memory prototypes. For each prototype source, we keep the same graph construction, contextual prior, GNN propagation, and prototype refinement procedure as in PRDC. Only the prototype generation strategy is changed.

As shown in Tab. 5, introducing PRDC consistently improves all three prototype constructions. In particular, LPCAM-style and SMC-style prototypes obtain improvements of 2.5 and 2.8 percentage points, respectively, after applying relational propagation. This demonstrates that the proposed

Table 5: Compatibility of PRDC with different prototype constructions.

Prototype source	PRDC	CAM mIoU	Gain
LPCAM-style prototypes		69.0	–
LPCAM-style prototypes	✓	71.5	+2.5
SMC-style prototypes		69.6	–
SMC-style prototypes	✓	72.4	+2.8
RO-PMB prototypes		70.1	–
RO-PMB prototypes	✓	**74.2**	**+4.1**

graph-based interaction is complementary to existing multi-prototype representations rather than being tightly coupled with the prototype construction mechanism of RO-PMB.

Impact of Global Prototypes per Class (K_g): Representing the number of global prototypes per class, K_g determines the granularity of our global memory bank. The results in Fig. 5(a) reveal that having too few prototypes ($K_g = 2$) limits the diversity of learned visual patterns, yet having an excessive number ($K_g = 7$) can introduce redundancy and make learning harder. Our experiments indicate that $K_g = \mathbf{4}$ provides the optimal balance, effectively capturing the intra-class variations without introducing excessive complexity.

Impact of the Number of Image-Specific Prototypes (K_{img}): As shown in Fig. 6, model performance improves as K_{img} increases, with particularly significant gains observed between values 2 and 3. Although setting $K_{img} = 4$ achieves a slightly higher segmentation accuracy, this performance increment is marginal. Crucially, this comes with a disproportionate increase in computational burden, specifically in the K-means clustering process and the subsequent prototype query computations. Considering the diminishing marginal gains in performance and computational efficiency, we ultimately selected $K_{img} = 3$.

Effect of Relation Transformation: The prototype graph uses cosine similarity to characterize the visual affinity between prototype nodes. In Eq. (3), we apply ReLU to retain positively supported relations while discarding negative similarities. To examine whether this choice is preferable to directly preserving or transforming negative similarities, we compare four relation transformations under the same training and evaluation setting.

Table 6: Ablation of different transformations of prototype cosine similarity for constructing the visual relation graph. All variants exclude the gain brought by L_{div} and L_{smooth}.

Relation transformation	CAM mIoU (%)
Raw cosine	72.8
Absolute cosine, $\lvert\cos\rvert$	71.9
Shifted cosine, $(1+\cos)/2$	73.5
ReLU cosine	**74.2**

As seen in Tab. 6, using the raw cosine similarity reduces CAM mIoU from 74.2% to 72.8%. In this formulation, positive and negative similarities are propagated through the same message-passing channel, although a negative cosine value is not necessarily a calibrated inhibitory relation. Thus, retaining negative values does not automatically provide meaningful signed relational reasoning.

5 Conclusion

In this paper, we have described a novel RO-PMB framework that builds upon multi-prototype representations and propagates positive visual and contextual associations among global prototypes by constructing a global relationship graph and applying a GNN. Through graph-based prototype interaction and bidirectional local–global refinement, our proposed RO-PMB supports relation-aware localization, generating more complete and boundary-accurate context-aware CAMs, and thus significantly enhances the effectiveness of WSSS. Extensive experiments on two mainstream datasets validate the superiority of our proposed framework, promoting a new avenue for advanced semantic understanding and research across areas relevant to WSSS.

References

[1] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921–2929, 2016.

[2] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[3] K. Sun, H. Shi, Z. Zhang, and Y. Huang, "Ecs-net: Improving weakly supervised semantic segmentation by using connections between class activation maps," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7283–7292, 2021.

[4] B. Zhang, J. Xiao, Y. Wei, and Y. Zhao, "Credible dual-expert learning for weakly supervised semantic segmentation," *International Journal of Computer Vision*, vol. 131, no. 8, pp. 1892–1908, 2023.

[5] X. Zhao, Z. Yang, T. Dai, B. Zhang, and J. Xiao, "Psdpm: Prototype-based secondary discriminative pixels mining for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3437–3446, 2024.

[6] X. Zhao, F. Tang, X. Wang, and J. Xiao, "Sfc: Shared feature calibration in weakly supervised semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, pp. 7525–7533, 2024.

[7] Q. Chen, L. Yang, J.-H. Lai, and X. Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4288–4298, 2022.

[8] J. Chen, R. Cong, Y. Luo, H. Ip, and S. Kwong, "Saving 100x storage: Prototype replay for reconstructing training sample distribution in class-incremental semantic segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 35988–35999, 2023.

[9] F. Tang, Z. Xu, Z. Qu, W. Feng, X. Jiang, and Z. Ge, "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.

[10] J. Wang, T. Dai, B. Zhang, S. Yu, E. G. Lim, and J. Xiao, "Pot: Prototypical optimal transport for weakly supervised semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 15055–15064, 2025.

[11] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[12] M. Lin, G. Li, S. Xu, Y. Hao, and S. Zhang, "Decoupling foreground and background with siamese vit networks for weakly-supervised semantic segmentation," *Neurocomputing*, vol. 610, p. 128540, 2024.

[13] X. Zhong, Q. Zhang, and J. Jiang, "Semantic driven vit and dual-level token fusion for weakly supervised semantic segmentation," *Neurocomputing*, p. 131204, 2025.

[14] K. Liu, F. Zhan, J. Zhang, M. Xu, Y. Yu, A. El Saddik, C. Theobalt, E. Xing, and S. Lu, "Weakly supervised 3d open-vocabulary segmentation," *Advances in Neural Information Processing Systems*, vol. 36, pp. 53433–53456, 2023.

[15] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1568–1576, 2017.

[16] J. Li, Z. Jie, X. Wang, Y. Zhou, L. Ma, and J. Jiang, "Weakly supervised semantic segmentation via self-supervised destruction learning," *Neurocomputing*, vol. 561, p. 126821, 2023.

[17] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4320–4329, 2022.

[18] Z. Chen and Q. Sun, "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.

[19] Z. Guo, D. Wu, C. Gao, and N. Sang, "Selective multiple classifiers for weakly supervised semantic segmentation," *CAAI Transactions on Intelligence Technology*, vol. 10, no. 6, pp. 1688–1702, 2025.

[20] S. Duan, X. Yang, and N. Wang, "Multi-label prototype visual spatial search for weakly supervised semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30241–30250, 2025.

[21] Y. Wu, D. Lian, Y. Xu, L. Wu, and E. Chen, "Graph convolutional networks with markov random field reasoning for social spammer detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, pp. 1054–1061, 2020.

[22] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," *Advances in neural information processing systems*, vol. 30, 2017.

[23] Z. Tang, Y. Liu, and Y. Shang, "A new gnn-based object detection method for multiple small objects in aerial images," in *2023 IEEE/ACIS 23rd International Conference on Computer and Information Science (ICIS)*, pp. 14–19, IEEE, 2023.

[24] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[25] Z. Du, M. Ren, W. Chu, and N. Chen, "Mvf-gnn: Multi-view fusion with gnn for 3d semantic segmentation," *IEEE Robotics and Automation Letters*, 2025.

[26] H. Jung, S. Y. Park, S. Yang, and J. Kim, "Superpixel-based graph convolutional network for semantic segmentation," 2021.

[27] S. Kachole, Y. Alkendi, F. B. Naeini, D. Makris, and Y. Zweiri, "Asynchronous events-based panoptic segmentation using graph mixer neural network," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4083–4092, 2023.

[28] Q. Liu, M. Kampffmeyer, R. Jenssen, and A.-B. Salberg, "Self-constructing graph neural networks to model long-range pixel dependencies for semantic segmentation of remote sensing images," *International Journal of Remote Sensing*, vol. 42, no. 16, pp. 6184–6208, 2021.

[29] X. He, K. Deng, X. Wang, Y. Li, Y. Zhang, and M. Wang, "Lightgcn: Simplifying and powering graph convolution network for recommendation," in *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, pp. 639–648, 2020.

[30] B. Zhang, S. Yu, Y. Wei, Y. Zhao, and J. Xiao, "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.

[31] L. Xu, M. Bennamoun, F. Boussaid, H. Laga, W. Ouyang, and D. Xu, "Mctformer+: Multi-class token transformer for weakly supervised semantic segmentation," *IEEE transactions on pattern analysis and machine intelligence*, 2024.

[32] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, pp. 303–338, 2010.

[33] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755, Springer, 2014.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[35] Y. Lin, M. Chen, W. Wang, B. Wu, K. Li, B. Lin, H. Liu, and X. He, "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.

[36] S. Zagoruyko and N. Komodakis, "Wide residual networks," *arXiv preprint arXiv:1605.07146*, 2016.

[37] Z. Yang, X. Zhao, X. Wang, Q. Zhang, and J. Xiao, "Ffr: Frequency feature rectification for weakly supervised semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 30261–30270, 2025.

[38] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," *Advances in neural information processing systems*, vol. 24, 2011.

[39] H. Kweon, S.-H. Yoon, and K.-J. Yoon, "Weakly supervised semantic segmentation via adversarial learning of classifier and reconstructor," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11329–11339, 2023.

[40] L. Chen, C. Lei, R. Li, S. Li, Z. Zhang, and L. Zhang, "Fpr: False positive rectification for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1108–1118, 2023.

[41] Z. Peng, G. Wang, L. Xie, D. Jiang, W. Shen, and Q. Tian, "Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 624–634, 2023.

[42] Y. Wu, X. Ye, K. Yang, J. Li, and X. Li, "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.

[43] T. Chen, X. Jiang, G. Pei, Z. Sun, Y. Wang, and Y. Yao, "Knowledge transfer with simulated inter-image erasing for weakly supervised semantic segmentation," in *European Conference on Computer Vision*, pp. 441–458, Springer, 2024.

[44] J. Hanna and D. Borth, "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.

[45] R. Xu, C. Wang, J. Sun, S. Xu, W. Meng, and X. Zhang, "Self correspondence distillation for end-to-end weakly-supervised semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, pp. 3045–3053, 2023.

[46] L. Ru, H. Zheng, Y. Zhan, and B. Du, "Token contrast for weakly-supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3093–3102, 2023.

[47] S. Jang, J. Yun, J. Kwon, E. Lee, and Y. Kim, "Dial: Dense image-text alignment for weakly supervised semantic segmentation," in *European Conference on Computer Vision*, pp. 248–266, Springer, 2024.

[48] J. He, L. Cheng, C. Fang, Z. Feng, T. Mu, and M. Song, "Progressive feature self-reinforcement for weakly supervised semantic segmentation," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, pp. 2085–2093, 2024.

[49] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," *Advances in neural information processing systems*, vol. 34, pp. 27408–27421, 2021.

[50] H. Kwon, J. Jeong, S.-H. Yoon, and K.-J. Yoon, "Phase concentration and shortcut suppression for weakly supervised semantic segmentation," in *European Conference on Computer Vision*, pp. 293–312, Springer, 2024.

[51] X. Xu, P. Zhang, W. Huang, Y. Shen, H. Chen, J. Lin, W. Li, G. He, J. Xie, and S. Lin, "Weakly supervised semantic segmentation via progressive confidence region expansion," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9829–9838, 2025.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Our abstract and introduction accurately reflect the contributions and scope of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please refer to Appendix G.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not include any theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Appendix A lists the necessary information in addition to our content pages.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Code will be released for reproducing the main experimental results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Appendix A lists the necessary information in addition to our content pages.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We clearly state that the variability originates from different random seeds for model initialization and data shuffling. The error bars are explicitly calculated by directly computing the standard deviation over 3–5 repeated experimental runs.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer:[Yes]

   Justification: Appendix A lists the necessary information in addition to our content pages.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in this paper conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This work has no societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This research does not release nor produce any data or models that have a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators and original owners of assets used in the paper have been appropriately cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release a codebase and our datasets alongside proper documentation and experiment scripts.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This research did not utilize human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This research did not utilize human subjects or crowdsourcing.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs were not an important, original, or non-standard component of the core methods of this paper.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Training Details

A.1 Hyper-Parameters

All experiments share the following base settings unless otherwise noted.

- Optimizer: AdamW
- Learning rate: $1e^{-4}$
- Weight decay: $1e^{-6}$
- Devices: 4
- Crop size: 512 × 512
- τ in Eq. (7): 0.1
- K_g: 4
- K_{img}: 3
- η in Eq. (9): 0.99
- δ_{div} in Eq. (14): 0.7
- λ_1: 0.5
- λ_2: 0.2
- λ_3: 0.3
- Batch size: 4
- Epochs: 45

A.2 Computational Resources

We conducted experiments using multiple GPU clusters equipped with NVIDIA A100 GPUs.

A.3 Network Training

The training objective of our proposed RO-PMB framework aims to jointly optimize image classification capability and the prototype learning process, thereby generating high-quality class activation maps. The total loss function $\mathcal{L}_{total}$ consists of four components: classification loss $\mathcal{L}_{cls}$, local-global consistency loss $\mathcal{L}_{loc-glo}$, prototype diversity loss $\mathcal{L}_{div}$ and graph smoothness loss $\mathcal{L}_{smooth}$. Together, these losses promote stable and effective prototype learning. Specifically, the standard multi-label classification loss applied to global average pooling (GAP) features derived from $\mathcal{M}^{proto}$ ensures the model correctly predicts the categories present in an image.

$$\mathcal{L}_{cls} = \text{MultiLabelSoftMarginLoss}(\text{GAP}(\mathcal{M}^{proto}), \mathbf{Y}), \tag{12}$$

where $\mathbf{Y}$ is the ground truth label at the image level.

The local-global consistency loss encourages the image-specific prototype $\mathbf{P}_{c,k}$ to remain consistent with its refined version $\mathbf{R}_{c,k}$, while allowing global knowledge to make corrections.

$$\mathcal{L}_{loc-glo} = \sum_{c,k} ||\mathbf{P}_{c,k} - \mathbf{R}_{c,k}||_2^2. \tag{13}$$

Prototype diversity loss penalizes similarity among global prototypes within the same category, encouraging them to learn distinct visual patterns and preventing prototype collapse. It is worth noting that $\mathcal{L}_{div}$ is applied to the original global prototype $\mathbf{m}_{c,k_g}$ rather than the GNN-enhanced $\mathbf{m}'_{c,k_g}$. This is to ensure the diversity of the base prototypes before GNN processing.

$$\mathcal{L}_{div} = \sum_{c} \sum_{k_g \neq k'_g} \text{ReLU}(\mathbf{m}_{c,k_g} \cdot \mathbf{m}_{c,k'_g} - \delta_{div}), \tag{14}$$

where δ_{div} is a threshold used to ensure that prototypes maintain a specific distance between each other.

To ensure the validity of prototype relationships learned by GNNs, we introduce a graph smoothness loss $\mathcal{L}_{smooth}$ as a structural regularization. Its purpose is to encourage prototypes connected through the graph structure to possess similar representations, thereby ensuring that nodes learned by the GNN remain close to each other in the feature space.

$$\mathcal{L}_{smooth} = \sum_{(i,j)\in\mathcal{E}} \mathbf{A}_{ij} ||\mathbf{m}'_i - \mathbf{m}'_j||_2^2, \tag{15}$$

where $\mathcal{E}$ represents the set of edges in the graph. The total loss function can be expressed as:

$$\mathcal{L}_{total} = \mathcal{L}_{cls} + \lambda_1 \mathcal{L}_{loc-glo} + \lambda_2 \mathcal{L}_{div} + \lambda_3 \mathcal{L}_{smooth}. \tag{16}$$

where $\lambda_1, \lambda_2, \lambda_3$ are weight parameters that balance the losses.

B Why Does a GNN Enhance Prototype Contextual Awareness Through Message Passing?

The core of GNN lies in its message-passing mechanism, which enables each global prototype in the graph to gather information from its direct neighbors and update itself by integrating this information with its own features. The single-layer GCN aggregates information from directly connected prototype nodes in one message-passing step. Category prototypes themselves represent visual patterns of a specific category. Through GNN's message passing, a prototype not only retains its own features but also absorbs characteristics from semantically related prototypes of other categories. For each edge in the graph, node *j* generates a message for its neighboring node *i*, and this message is derived from node *j*'s features through a learnable transformation function as described in Eq. (4).

The GNNs explicitly leverage the global prototype relationship graph we construct. This graph captures semantic connections between prototypes through edges defined by similarity. When the "*person*" prototype receives messages from the "*sofa*" prototype, it aggregates positive supportive information from a contextually related category. After GNN processing, each relationship optimized global prototype is no longer isolated but integrates contextual information from its semantic network. These relation-enhanced global prototypes are then used to refine image-specific prototypes.

Ultimately, when these relationship optimized prototypes are used to generate CAMs, they better activate pixels belonging to the same semantic context region. This results in CAMs that are spatially more coherent with more complete boundaries, avoiding the fragmented activation that might occur when relying solely on local features.

C Visual Analysis

To provide a more intuitive understanding of the SCP, we visualize the pre-calculated statistical matrix **Q** in Fig. 7. As seen, this matrix captures strong semantic dependencies that transcend pure visual similarity. For instance, there is a high conditional probability between "*diningtable*" and "*chair*" (0.68), "*sofa*" and "*tvmonitor*" (0.13). Further, the "*person*" category exhibits a broad correlation with various objects like "*bicycle*" (0.56) and "*horse*" (0.47), reflecting typical real-world scenes.

To empirically verify this enhancement, we visualize the learned prototype relationship in their Heatmaps in Fig. 8. As seen, our PRDC module captures a structured semantic manifold with two distinct characteristics. First, a prominent block-diagonal structure appears along the main diagonal, indicating that the GNN reinforces intra-class consistency across diverse prototypes of the same category. Second, significant off-diagonal activations emerge between contextually related classes (such as "*person*" and "*sofa*", "*bird*" and "*sky*"). These patterns illustrate within-category and cross-category prototype associations.

This structured connectivity facilitates the "infusion" of associative knowledge. For instance, when a "*person*" prototype interacts with a "*sofa*" prototype, it aggregates contextual features through their positive graph association. Consequently, the relationship optimized prototypes incorporate positive contextual information from connected prototypes. This mechanism effectively addresses

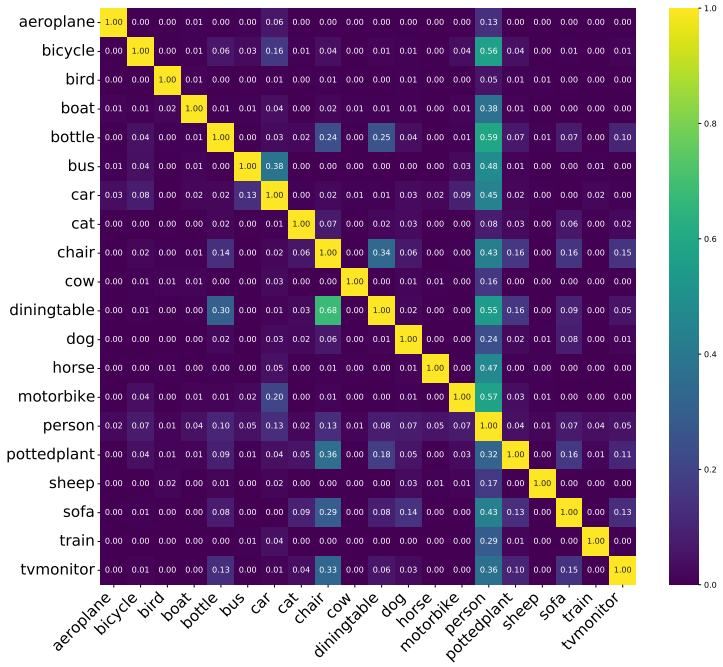


Figure 7: Illustration of the Statistical Class Co-occurrence Prior Matrix **Q** on PASCAL VOC 2012, where each cell (i, j) represents the normalized conditional probability that class j occurs in an image given the presence of class i. This matrix serves as a semantic prior to guide the message passing in our PRDC module.

Table 7: Ablation results on the impact of the hyperparameter δ_{div}, where optimal settings are highlighted in gray, and Seg. denotes the final segmentation result.

δ_{div}	CAM	Seg.
0.5	74.9	76.1
0.6	75.4	76.3
0.7	76.6	77.6
0.8	76.1	76.9
0.9	75.8	76.5

the discriminative bias of traditional CAMs, where only the most salient parts are captured yet less discriminative parts are ignored.

By leveraging these context-aware prototypes for feature activation, our framework generates CAMs that are spatially more coherent and more consistent in boundary accuracy. Via transitioning from isolated local feature matching to global semantic network reasoning, our proposed RO-PMB not only avoids fragmented activations but also provides high-quality pseudo-labels that are crucial for high-performing weakly supervised semantic segmentations.

D Further Ablation Analysis

Impact of Prototype Diversity Margin (δ_{div}): The diversity margin δ_{div} in our prototype diversity loss $\mathcal{L}_{div}$ regulates the separation between prototypes. The ablation results given in Tab. 7 demonstrate that an inadequate margin ($\delta_{div} = 0.5$) may not sufficiently encourage diversity, whilst a very large margin ($\delta_{div} = 0.9$) could force an unnatural separation. As a result, $\delta_{div} = \mathbf{0.7}$ represents a moderate margin and delivers the best performance, which suggests that encouraging a proper level of distinctiveness among prototypes is vital for comprehensive class representation.

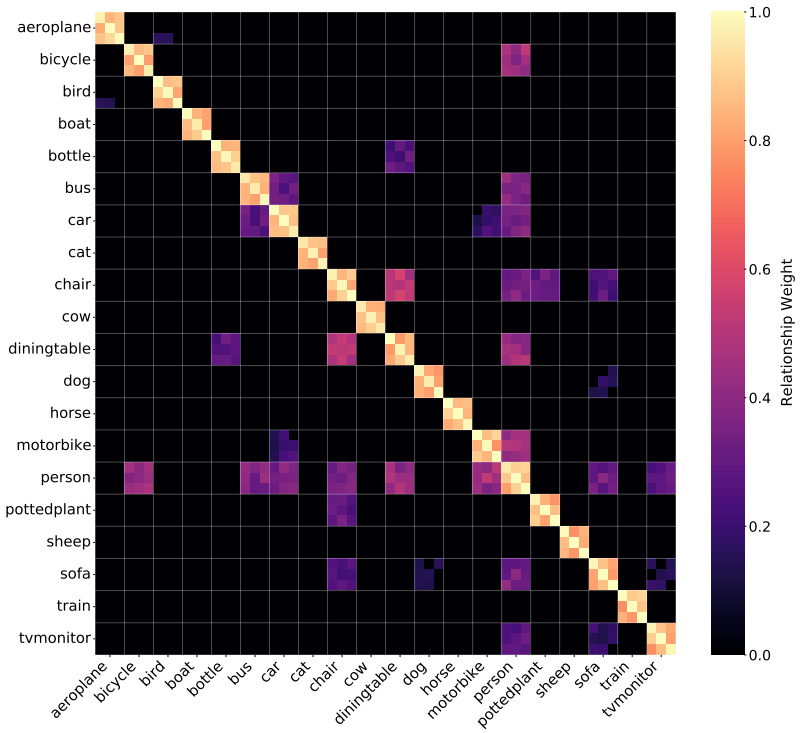


Figure 8: Visualization of the learned prototype relationship heatmap. Lighter colors indicate stronger semantic associations.

Loss Weight Parameters: Fig. 9 presents the results of an ablation study on the CAM performance of the RO-PMB framework, examining the three key loss weights λ_1, λ_2, and λ_3. Specifically, for the local-global consistency loss weight λ_1, the model achieves its best performance at λ_1=0.5. Values that are either too high or too low weaken the refinement effect. This indicates that moderate local-global information fusion is crucial for stable prototype learning. The prototype diversity loss weight λ_2 achieves optimal performance at 0.2, revealing that moderately encouraging prototype diversity prevents prototype collapse, while excessive diversity penalties may distract prototype learning. Finally, the graph smoothness loss weight λ_3 performs optimally at 0.3, confirming that moderate graph structure regularization enables GNNs to learn more meaningful and smooth prototype relationships, thereby enhancing the effectiveness of relationship optimized refinement. These results collectively validate the rationality of the selected weight parameters and emphasize the importance of synergistic interactions among auxiliary losses within the RO-PMB framework to achieve optimal performances for weakly supervised semantic segmentation.

E Analysis of Model Complexity

Tab. 8 provides a direct comparison of computational costs between our RO-PMB framework and the state-of-the-art POT baseline on PASCAL VOC, using identical hardware. Despite incorporating novel graph-based relational modeling, RO-PMB maintains remarkable computational efficiency. The model parameters and GPU memory consumption exhibit only slight increases, attributable to our implementation of multiple complexity-reduction strategies.

Table 8: Training cost comparison on PASCAL VOC, where all methods are run on NVIDIA A100-PCIE-40GB GPUs.

Method	Training time(h)	Inference time	Parameters (M)	Peak GPU memory (GB)
POT	29h	0.24s	28.6M	6.1G
Our	32h	0.26s	29.2M	7.9G

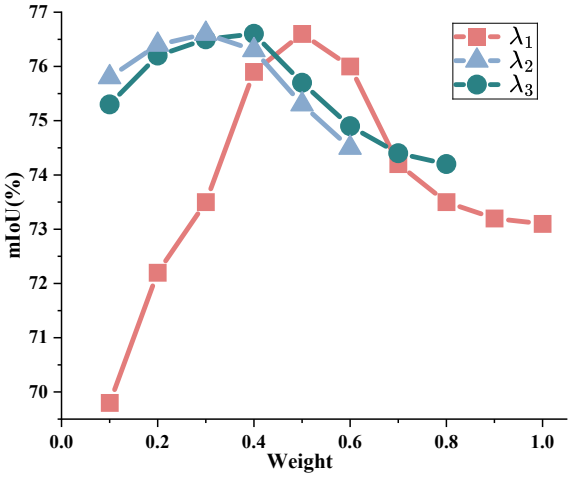


Figure 9: Ablation results of loss weights for RO-PMB.

(i) We strictly limit the number of global prototypes K_g per category within a small range, thereby avoiding an explosive increase in prototype graph construction complexity $O(N_P^2)$. **(ii)**When constructing the global prototype relationship graph, we employ a K-nearest neighbor strategy to ensure graph sparsity, effectively reducing the GNN's computational complexity from a potential $O(N_P^2 \cdot D_p^2)$ to near-linear $O(N_P \cdot k \cdot D_p + N_P \cdot D_p^2)$. **(iii)**The GNN module itself adopts a lightweight single-layer GCN architecture, avoiding the additional computational burden associated with deep GNNs.

Regarding time overhead, both training and inference times show modest, proportionate increases compared to the baseline. The additional training time arises from operations such as per-image K-means clustering, global graph construction, GNN message passing, and momentum updates. Similarly, the slightly extended inference time is due to the image-specific prototype generation and querying. Crucially, these increases are well within acceptable limits, demonstrating that our complexity control measures are highly effective.

F Additional Comparison with Existing Prototype-Based WSSS Methods

Existing prototype-based WSSS methods have demonstrated that prototype representations can alleviate the limited object coverage of CAMs and capture diverse visual patterns. In particular, LPCAM [18] clusters class-wise local features into multiple prototypes to represent different local semantic patterns, while SMC [19] maintains multiple foreground and background prototypes in a memory bank and adaptively selects suitable prototypes for individual images. MuP-VSS [20] further employs image-conditioned class query tokens and optimizes them through cross-image and cross-prototype consistency.

Table 9: Comparison with representative prototype-based WSSS methods. The comparison focuses on prototype organization and relational interaction.

Method	Prototype organization	Multiple modes per class	Cross-image mechanism	Cross-category relation modeling
LPCAM [18]	Class-wise local feature clustering	✓	Offline class-level clustering	/
SMC [19]	Foreground/background prototype memory	✓	Prototype memory bank	/
MuP-VSS [20]	Image-conditioned class query tokens	No explicit $K > 1$ modes	Contrastive consistency	Contrastive discrimination, without a prototype graph
RO-PMB	Local and global prototype sets	✓	Momentum-updated memory	Explicit graph construction and GNN propagation

As summarized in Tab. 9, existing methods mainly use prototypes as visual representatives, memory elements, or image-conditioned queries. Their cross-image mechanisms are implemented through offline clustering, prototype memory, or contrastive consistency, rather than explicit graph-based information propagation among prototypes. In contrast, RO-PMB maintains local and global prototype sets and constructs a unified relational graph over global prototypes from different semantic categories. A graph neural network is then used to propagate contextual information among these prototypes before the relation-enhanced global representations are used for image-specific prototype refinement.

G Limitations and Future Work

Limitations. Our current RO-PMB employs a single-layer GCN and ReLU-activated cosine similarity for prototype relationship modeling, prioritizing computational efficiency. This design inherently limits its capabilities in two aspects, namely: (i) ignored Negative Associations, where setting negative similarities to zero precludes the use of potentially valuable mutually exclusive or irrelevant semantic information between prototypes. (ii) Local Relationship Dependency, in which the single-layer GCN primarily captures one-hop neighborhood relationships, limiting its capacity for deep, multi-hop semantic reasoning and comprehensive modeling of intricate dependency networks.

Future Work. Given the aforementioned limitations, our future research will focus on two key directions: (i) We plan to explore advanced GNN architectures that can explicitly handle signed edge weights, enabling the comprehensive integration of both positive synergistic relationships and negative mutually exclusive semantic relationships; (ii) We will investigate the use of multi-layer GNNs, potentially augmented with attention mechanisms, to capture and aggregate deeper and more complex multi-hop prototype relationships. This aims to foster more global and granular semantic reasoning, thereby enhancing context awareness for improved segmentation accuracy and coherence in those challenging scenarios faced by WSSS.