Fine-grained Background Representation for Weakly Supervised Semantic Segmentation

Xu Yin, Woobin Im, Dongbo Min, Yuchi Huo, Fei Pan, Sung-Eui Yoon

***Abstract*—Generating reliable pseudo masks from image-level labels is challenging in the weakly supervised semantic segmentation (WSSS) task due to the lack of spatial information. Prevalent class activation map (CAM)-based solutions are challenged to discriminate the foreground (FG) objects from the suspicious background (BG) pixels (a.k.a. co-occurring) and learn the integral object regions. This paper proposes a simple fine-grained background representation (FBR) method to discover and represent diverse BG semantics and address the co-occurring problems. We abandon using the class prototype or pixel-level features for BG representation. Instead, we develop a novel primitive, negative region of interest (NROI), to capture the fine-grained BG semantic information and conduct the pixel-to-NROI contrast to distinguish the confusing BG pixels. We also present an active sampling strategy to mine the FG negatives on-the-fly, enabling efficient pixel-to-pixel intra-foreground contrastive learning to activate the entire object region. Thanks to the simplicity of design and convenience in use, our proposed method can be seamlessly plugged into various models, yielding new state-of-the-art results under various WSSS settings across benchmarks. Leveraging solely image-level (I) labels as supervision, our method achieves 73.2 mIoU and 45.6 mIoU segmentation results on Pascal Voc and MS COCO test sets, respectively. Furthermore, by incorporating saliency maps as an additional supervision signal (I+S), we attain 74.9 mIoU on Pascal Voc test set. Concurrently, our FBR approach demonstrates meaningful performance gains in weakly-supervised instance segmentation (WSIS) tasks, showcasing its robustness and strong generalization capabilities across diverse domains.**

***Index Terms*—Contrastive learning, Fine-grained background representation, Weakly supervised image segmentation.**

I. Introduction

FULLY-supervised semantic segmentation [29], [49] requires a pixel-annotated training set, which is costly and time-consuming to create. Thus, great efforts have been put into weakly supervised semantic segmentation to reduce the annotation cost by leveraging less expensive yet weakly spatially-informative supervision signals, such as image tags [13], bounding boxes [38], and scribbles [14], [17]. Current studies [23], [25], [26] usually start from generating class activation maps [6] by training a classification network to build the seeds and then utilize refinement techniques [1], [2], for generating reliable pseudo masks, which are finally used to train the segmentation model [37], [44]. In this work, we concentrate on the image-level weakly supervised semantic segmentation, where only the images' class labels are given; we aim to enhance the quality of seeds and, thus, the segmentation results.

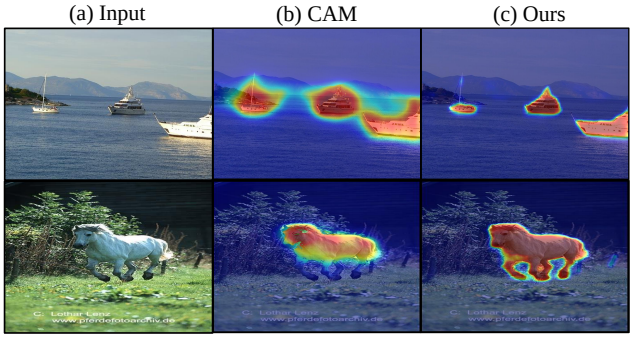


Fig. 1: (a) Input (b) Class activation maps via AMN [25] (c) Refined class activation maps with our method (on AMN). In the 1st row (b), class activation maps mistake the lake (co-occurring background semantic) as the boat; in the 2nd row (b), the horse is not completely activated.

First, we illustrate two common problems of class activation maps (shown in Fig. 1): co-occurring background semantics and incomplete object region. The former refers to the disturbance from the confusing image background information for the foreground classification [24], [26]. The background semantic frequently appearing with the target foreground object carries suspicious information and results in ambiguous recognition, e.g., boat and lake (in (a)). The latter problem showcases (in (b)) that class activation maps prefer highlighting the discriminative semantic region instead of the entire object part [26]. This issue manifests that the classifier uses less context-dependent information to explain its prediction due to the need for an in-depth understanding of class-level properties. These two issues motivate us to isolate the confusing background semantics from the classification process and learn more discriminative foreground features.

In this work, we analyze and conclude that the co-occurring problem of class activation maps resulted from disregarding the image background region in the classification training. However, existing background representation and feature sep-

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2023-00208506(2024)) and was partially supported by the National Key R&D Program of China (No. 2024YDLN0011) and NSFC (No. 62441205) (Corresponding author: Sung-eui Yoon.)

Xu Yin and Woobin Im are with the School of Computing, Korea Advanced Institute of Science and Technology, Daejeon 34141, South Korea (E-mail: yinofsgvr@kaist.ac.kr and iwbn@kaist.ac.kr).

Dongbo Min is the Faculty of the Department of Computer Science and Engineering, Ewha Womans University, Seoul 03760, South Korea (E-mail: dbmin@ewha.ac.kr).

Yuchi Huo is with the State Key Lab of CAD and CG, Zhejiang University, China and Zhejiang Lab, China 310058 (E-mail: huo.yuchi.sc@gmail.com).

Fei Pan is with the School of Computer Science Engineering, University of Michigan. Email: feipan@umich.edu).

Sung-eui Yoon is with the Faculty of School of Computing, Korea Advanced Institute of Science and Technology, Deajeon 34141, South Korea (E-mail: sungeui@gmail.com).

Copyright © 2024 IEEE. Personal use of this material is permitted. However, permission to use this material for any other purposes must be obtained from the IEEE by sending an email to pubs-permissions@ieee.org.

TABLE I: We summarize related methods with informative keywords and highlight the differences between them and ours.

Method	Publication Venue	Keywords
AdvCAM [23]	CVPR'21	attribute maps, adversarial attack, image perturbations
ToCo [36]	CVPR'23	token contrast, intermediate feature supervision, local-to-global consistency
C^2AM [47]	CVPR'22	class-agnostic activation map, contrastive learning, foreground-background representations
PPC [10]	CVPR'22	prototype learning, pixel-level supervision, cross-view semantic consistency
Ours	–	negative-region-of-interest, activate negative sampling, foreground-background contrastive learning

aration methods [6], [12] do not consider the foreground-background semantic discrepancy and thus fail to capture fine-granulate background semantics. To address these limitations, we propose a fine-grained background representation (FBR) method built upon contrastive learning. We recognize that the image background region contains a vast amount of task-irrelevant information and lacks descriptive semantics. Therefore, we model the background independently from the foreground. For every foreground class, we define a learnable prototype that encodes its essential features. In order to represent the image background, we specifically develop a primitive, negative-region-of-interest (NROI), capable of capturing the fruitful and diverse semantics of backgrounds. To address the co-occurring issue, we implement Fore-to-background contrast in a pixel-to-NROI manner to decouple the foreground objects effectively from the confusing background semantics and thus overcome the problem. Additionally, we tailor an active sampling method for weakly supervised semantic segmentation that selects negatives based on foreground semantic relationships and conduct intra-foreground contrastive learning. This way, we learn compacted foreground class features and obtain completer object regions.

In summary, our main contributions are three-fold:

- We propose a simple FBR method to address the co-occurring and incomplete object region problem for weakly supervised semantic segmentation.
- We propose a fine-grained background primitive, dubbed NROI, to represent image background effectively and implement the fore-to-background contrastive learning to enhance class activation maps' ability to distinguish co-occurring background cues. Also, we introduce an active method to sample efficient foreground negatives and conduct intra-foreground contrastive learning to activate integral object regions.
- Extensive experiments and evaluations in weakly supervised semantic and instance segmentation demonstrate our FBR approach can be used in different applications and is generalizable to various baseline architectures. In particular, our method achieves new state-of-the-art weakly supervised semantic segmentation performances on Pascal Voc 2012 and MS COCO 2014 test sets.

II. Related Work

Image-level weakly supervised semantic segmentation. This task aims to generate reliable pseudo masks from class labels to guide the segmentation practice. The main trend in this field is to produce complete object masks with class activation maps, yielding high-quality pseudo labels.

Current studies [16], [34] usually resort to getting more object information involved in the classification or using powerful backbone architectures, e.g., ViT [9], to get more object regions activated. For example, RIB [22] presents a novel pooling method to reduce the information bottleneck during classification and forces classifiers to identify less discriminative regions of target class objects. ToCo [36] proposes contrasting the patch representations from different layers and the local-global features of the class token to scale up the activated region of target class objects.

Despite the advanced results achieved by these methods, some confusing background semantics are inevitably activated when expanding the object region, introducing noise pixels around semantic boundaries. In this work, we explicitly model the image foreground and background semantics and implement the fore-to-background and intra-foreground contrastive learning to suppress the effect of ambiguous background information, thus learning accurate object regions.

Contrastive learning. The goal of this approach is to learn a similarity function in common feature space to pull views of similar data (positive) closer while pushing views of dissimilar ones (negative) apart. This way, we label each pixel based on the similarity measure and activate accurate object regions.

For instance, C^2AM [47] proposes the fore-background contrastive learning to generate the class-agnostic class activation maps. PPC [10] employs a prototype, namely the typical class features, to serve as the positive to perform the contrastive learning; besides, PPC adopts the "hardness" computation in [40] to draw hard negatives.

Nevertheless, most solutions [10], [31], [42] treat the image background equally as the foreground classes and optimize them jointly, ignoring the fact that the background content covers diverse object semantics. This nature makes the image background difficult to represent with a learnable prototype or features, resulting in inferior optimization effects.

By contrast, we model the image background with the proposed novel primitive, negative-region-of-interst, to denote its complex semantics and perform the fore-to-background contrastive learning. This way, we decouple the relationship between diverse background semantics and the foreground classes, addressing the co-occurring problem. Besides, we present an active sampling strategy to select the negative samples of foreground classes on the fly. Therefore, we accurately discriminate the intra-foreground relationships and learn integral object regions.

Technical comparison with related works:: Table I compares our method with the aforementioned studies. We summarize each study with three keywords to introduce their main novelties and methodologies. To compare, our main contributions rely on the proposed negative-region-of-interest for diverse background semantic representation and the effective active sample method to facilitate the learning of foreground class prototypes. In Sec. IV, we conduct extensive experiments to verify the effectiveness of our method and demonstrate its benefits from different perspectives.

III. METHODOLOGY

In this part, we first clarify an underlying motivation of fine-grained background representation in weakly supervised semantic segmentation (WSSS). Next, we tailor the class activation maps (CAM)-based contrastive learning and propose optimizing the fore-to-background (FB) and the intra-foreground (IF) relationships.

A. Motivation

In the image-level WSSS, we are given a training set $\mathcal{D}$, with data tuple $(\mathbf{x}, \mathcal{Y}) \in \mathcal{D}$, where each image $\mathbf{x}$ is associated with a class label $\mathcal{Y} = (y_1, y_2, ..., y_C)$; $y_c = 1$ denotes the presence of the foreground (FG) class c ($1 \leq c \leq C$) in $\mathbf{x}$ and 0 otherwise. While in the semantic segmentation task, we aim to learn a discriminative model (parameterized by neural networks) to approximate the conditional distribution $p(\mathbf{y}|\mathbf{x})$, where $\mathbf{y} \in \mathbb{R}^{(C+1)\times H\times W}$ ($H \times W$ denotes the spatial size) is the ground-truth semantic label that contains C FG classes and a background (BG) class.

Existing WSSS solutions [1], [2], [8], [13] utilize CAM (denoted with $\hat{\mathbf{y}} \in \mathbb{R}^{(C+1)\times H\times W}$) to approximate $p(\mathbf{y}|\mathbf{x})$ by learning a reliable semantic feature $f \in \mathbb{R}^{L\times H\times W}$ (L denotes the feature dimension) with the classification loss L_{cls}:

$$L_{cls} = -\frac{1}{C}\sum_{c=1}^{C}[y_c \log \sigma(\hat{s}_c) + (1-y_c)\log(1-\sigma(\hat{s}_c))], \quad (1)$$

where σ is softmax function and $\hat{s}_c$ is the classification score. We define CAM as a function that projects each pixel i's feature f_i with the parameter θ (the weights of the classifier) to the semantic label space $\hat{\mathbf{y}}_i \in \mathbb{R}^{C+1}$:

$$CAM(f_i, \theta) = \hat{\mathbf{y}}_i, \quad (2)$$

where $\hat{\mathbf{y}}_i$ is then normalized to a categorical distribution.

It is worth noting that L_{cls} essentially learns $p(y_c|\mathbf{x}_i)$, i.e., the probability of the pixel $\mathbf{x}_i$ being assigned to the FG semantic space, excluding the BG. In the CAM generation step, we argue that the entire BG region is treated as a virtual class that is ignored, resulting in f_i being vulnerable to the BG semantics, particularly co-occurring ones, and hence leading to classification ambiguity. This observation indicates the necessity of BG-oriented modeling. Also, with informative BG representations, we can further decouple the semantic correlation between the target objects [6], [26] and their nearby BG to better approximate the true $p(\mathbf{y}_i|\mathbf{x}_i)$.

Our work is the first attempt to address the aforementioned limitations of CAMs by the fine-grained BG representation. Unlike existing approaches [6], [31], [33] either using pixel features or prototypes to give an abstract BG representation, our key novelty resides in explicitly modeling the image BG with a novel fine-grained primitive and performing FB contrast to eliminate the BG confusion (Sec. IV-B). Besides, we design an active negative sampling method to implement effective IF contrast, thus learning the compacted FG features to activate the complete object masks. We enhance CAMs and obtain more reliable seeds by optimizing these two relationships.

B. Our Method

1) Contrastive learning setup for WSSS: To begin with, we follow the pipeline in [19] to generate the seed $\mathcal{H} \in \mathbb{R}^{H\times W}$, except replacing the Global Average Pooling (GAP) layer with the Thresholded Average Pooling (TAP) layer [4], which averages only the above-threshold pixels in the semantic feature f (introduced and ablated in the Supplementary). Additionally, we add a nonlinear projection head [7], [31], φ_{fg}, encoding f into a D-dimensional representation, $Z_{fg} \in \mathbb{R}^{D\times H\times W}$, with the same spatial resolution as $\mathcal{H}$ (shown in Fig. 2).

FG prototype assignment: According to the spatial location, we assign $\mathcal{H}$'s FG label information to the pixels in Z_{fg}. Following the setting in [10], [15], for each FG class c in the batch, we choose the top N pixels with the high CAM scores and compute its prototype, p_c, as the weighted average of the pixel-level representation.

$$p_c = \frac{\sum_{i\in\pi_c}\hat{\mathbf{y}}_{c,i}Z_{fg,i}}{\sum_{i\in\pi_c}\hat{\mathbf{y}}_{c,i}}, \quad (3)$$

where $\hat{\mathbf{y}}_c \in \mathbb{R}^{H\times W}$ denotes class c's activation map, and π_c is the spatial coordinate set of the top N pixels in $\hat{\mathbf{y}}_c$.

Query computation: We build the query set, Z_c^q, for each appeared c in the batch. Rather than querying all FG pixels, we consider CAM score as the certainty measure to determine Z_c^q adaptively, enabling the contrastive loss focus on uncertain (below the threshold β, set to 0.4) pixels in $\mathcal{H}$:

$$Z_c^q = \mathbb{1}[\mathcal{H} = c] \cdot \mathbb{1}[\hat{\mathbf{y}}_c < \beta] Z_{fg}. \quad (4)$$

Prototype-based Contrastive Learning (PCL): The standard contrastive loss [7], [27] functions by encouraging the query $q \in Z_c^q$ to be similar to its positive keys and dissimilar to the negative key $z^n \in Z^n$. In this work, we employ the estimated FG prototypes $p_c \in P$ as the positives and express the contrastive loss L_{pcl} with:

$$\begin{aligned} L_{pcl} &= PCL(P, Z^q, Z^n) \\ &= \sum_{p_c\in P}\sum_{q\in Z_c^q} -\log\frac{e^{(q\cdot p_c/\tau)}}{e^{(q\cdot p_c/\tau)} + \sum_{z^n} e^{(q\cdot z^n/\tau)}}, \end{aligned} \quad (5)$$

where $P = \{p_c\}_{c=1}^{C}$ and $Z^q = \{Z_c^q\}_{c=1}^{C}$ are the collections of prototypes and queries, τ and '$\cdot$' denote the temperature and dot product. We instantiate Z^n from BG and FG regions and follow the function form in Eq. 5 to respectively optimize the FB and IF contrastive relationships; loss functions are expressed in Eq. 7 and Eq. 9.

2) Fore-to-background (FB) Contrast: In Sec. III-A, we conclude that the conditional BG distribution $p(\hat{\mathbf{y}}_{C+1}|\mathbf{x})$ can not be optimized by L_{cls}, and hence f has weak BG description ability. Besides, image BG, unlike FG, does not have specific semantics and contains massive task-unrelated information; a single prototype, like Eq. 3, is incapable of covering its high variance (discussed in Sec. IV-C). Driven by these two concerns, we propose a fine-grained primitive, termed negative regions of interest (NROI), to comprehensively model the image BG that has a mixture of diverse semantics.

NROI for BG representation: Unlike existing approaches [7], [33], [47] that represent the FG and the BG semantics in a

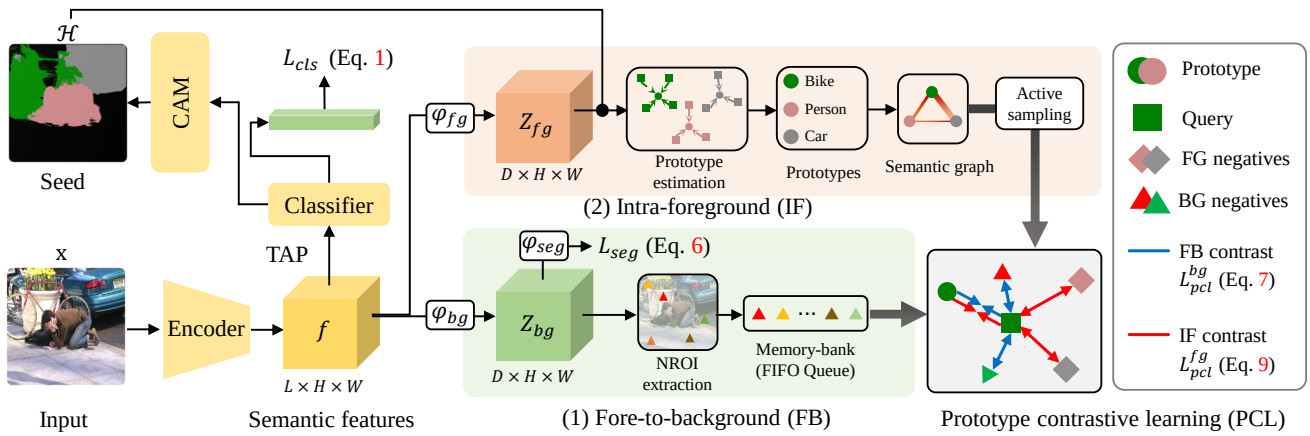


Fig. 2: Architecture overview. A standard feature encoder trained with the classification loss L_{cls} (with TAP [4]) takes an input image $\mathbf{x}$ and generates the seed $\mathcal{H}$. We consider that image BG has a different semantic granularity from FG and add two projection heads, φ_{fg} and φ_{bg}, model BG independently from FG to capture diverse BG information, and optimize two contrastive relationships: (1) fore-to-background and (2) intra-foreground. (1) enhances the semantic features f in representing BG semantics with the proposed fine-grained primitive, namely NROIs. We compute FG prototypes and store NROIs in a memory bank. Besides, the auxiliary BG segmentation loss L_{seg} is introduced. In (2), we present an active sampling strategy built upon the semantic graph to draw the FG negatives. The contrastive losses L_{pcl}^{bg} for (1) and L_{pcl}^{fg} for (2) pull the query closer to its prototype but push far from the FG and the BG negative keys, respectively.

common space, we model the image BG independently to distinguish it from the FG well. Formally, we add another projection head, φ_{bg} (shown in Fig. 2), parallel with φ_{fg}, to look for the reliable BG representation from a different mapping: $\varphi_{bg} : f \rightarrow Z_{bg}, Z_{bg} \in \mathbb{R}^{D \times H \times W}$.

A naive brute-force method (Fig. 3 (a)) is to use all BG features in Z_{bg} to perform the optimization, i.e., pixel-to-pixel contrast, which would be time-consuming and computationally expensive. Also, the large-scale variation and high-complexity nature of the intra-BG region in training set $\mathcal{D}$ challenge us to develop an efficient representation method to denote its content. To this end, assuming that the image BG composes multiple semantics, we explore discovering fine-grained semantics and effectively denote them using NROI. Specifically, a group of K NROIs $\{z_{bg}^k\}_{k=1}^K$ is used for BG description, where k is the NROI indices with respect to the input $\mathbf{x}$.

We perform online clustering [51] for NROI determination. We get the BG features (with the spatial information in $\mathcal{H}$) from the masked Z_{bg} and map them to K clusters with K-means. Intuitively, clustering [27], [51] imposes an inductive bias [51] that image BG consists of multiple semantics, and it thereby enables the model to discover the discriminative pixel groups, i.e., semantics. Hence, the NROIs $\{z_{bg}^k\}_{k=1}^K$ of each image, defined as the cluster centroids, are the typical representations of the BG semantics (in Fig. 3 (b)).

BG memory bank: We construct a queue-based memory bank [3], Z_{bg}^n, to store NROIs, and set it as fixed storage for spatial and computational efficiency. As shown in Fig. 2, the bank is updated at each training step with the extracted NROIs, and then we use the BG negative keys randomly sampled from Z_{bg}^n to contrast against FG queries.

Auxiliary BG segmentation. Unlike prior studies [7], [42], FBR adopts two projection heads to perform the contrastive learning, which brings a risk of a homogeneous representation between Z_{fg} and Z_{bg}. To avoid this trivial case, we formulate a learning objective to the BG representation Z_{bg}, distinguishing it from Z_{fg} and enhancing its BG discrimination ability. Specifically, we introduce binary segmentation as an auxiliary task, empirically considering pixels in $\mathcal{H}$ with a low-summed FG activation value [2] (smaller than 0.05) as the pseudo BG labels (termed M), and feed Z_{bg} (after batch normalization) into the BG predictor, φ_{seg}:

$$L_{seg} = BCE(\varphi_{seg}(Z_{bg}), M), \tag{6}$$

where BCE is a binary cross entropy loss.

Pixel-to-NROI contrast: With the query sets (Eq. 4) and the BG memory bank, we give the FB contrastive loss as:

$$L_{pcl}^{bg} = PCL(P, Z^q, Z_{bg}^n). \tag{7}$$

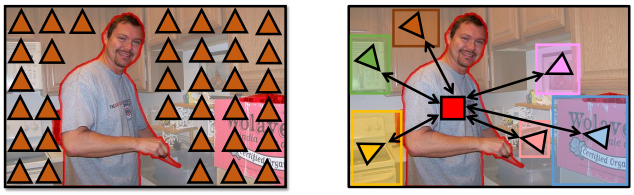
This pixel-to-NROI contrast maximizes the agreement between the queries and their belonging prototype while minimizing the agreement with the BG semantic, i.e., NROIs.

3) Intra-foreground (IF) contrastive learning: This part presents an active negative sampling method to select FG negative keys and conduct effective IF contrast.

Active negative sampling: We first define the query class c's full negative set, z_c^n, which contains all FG pixels that do not belong to c: i,e., $z_c^n = \mathbb{1}[(\mathcal{H} \neq c) \cap (\mathcal{H} \neq C+1)]Z_{fg}$. However, contrasting the query against all samples in z_c^n is computationally costly. Moreover, contrastive learning may be ineffective or even degenerate an overall performance due to the implausible label information of the seed $\mathcal{H}$.

Inspired by the recent semi-supervised semantic segmentation study [31], we propose actively drawing negatives from z_c^n and optimizing only with the selected samples to overcome the above limitations. For each batch, we compute a graph $G \in \mathbb{R}^{C \times C}$, where nodes and edges stand for the occurring classes and their relative semantic relations:

$$G[i,j] = Sim(p_i, p_j), \text{where } i,j \leq C, \text{and } i \neq j. \tag{8}$$



(a) brute-force contrast (b) NROI-based contrast (**ours**)

Fig. 3: Conceptual illustration of negative-region-of-interest (NROI) for the FB contrast. The brute-force strategy (a) exhaustively compares FG queries (the red cropped part) with all BG pixels (triangles), which requires expensive computational resources and is susceptible to implausible labels. By contrast (b), we propose recognizing the fine-grained BG semantic, i.e., NROI. This example's NROIs (marked with different colors) contain the washing machine, closet, etc. In training, we implement FB contrastive learning by comparing queries (the red rectangle) against NROIs.

Here, Sim is the cosine similarity. Unlike [31], we exclude the BG class from the graph and use the semantic distance between FG prototypes, rather than between the mean features, to measure the pair-wise relationship $G[i,j]$.

For each query class c, we turn its relationships in G against negative classes into a distribution by softmax: $\frac{\exp(G[c,i])}{\sum_{j\leq C, j\neq c}\exp(G[c,j])}$. We sample keys of each negative class i from z_c^n based on the distribution. Intuitively, this step performs a non-uniform sampling on z_c^n, drawing more samples from the classes that are semantically similar to c, while drawing fewer from dissimilar ones. It enables the classifier to learn compacted FG features and an accurate decision boundary by improving the discrimination ability regarding the confusing, negative classes.

Pixel-to-prototype contrast: Similar to Eq. 7, we formulate the IF contrastive loss as:

$$L_{pcl}^{fg} = PCL(P, Z^q, Z_{fg}^n), \tag{9}$$

where $Z_{fg}^n = \{z_c^n\}_{c=1}^C$. This pixel-to-prototype contrast learns compacted FG features by pulling the queries close to their prototype and pushing different classes far away.

C. Overall objective

As illustrated in Fig. 2, our FBR method can be integrated into existing WSSS solutions to obtain better seeds. We add two projection heads φ_{fg} and φ_{bg} after the encoder, mapping the semantic feature f into two high-dimensional representations Z_{fg} and Z_{bg}, and then implement the FB and the IF contrastive learning. The overall loss is expressed:

$$L = L_{cls} + L_{pcl} + L_{seg}, \tag{10}$$

where $L_{pcl} = \lambda_1 \cdot L_{pcl}^{bg} + \lambda_2 \cdot L_{pcl}^{fg}$, jointly optimizing these two constrative relationships using corresponding loss weights λ_1 and λ_2. Note that φ_{fg} and φ_{bg} are only applied during training and discarded in the inference phase.

IV. Experiments

We first ablate FBR to test its effectiveness and apply it to existing models to get state-of-the-art (SOTA) results.

A. Setup

Datasets & evaluation metrics. We experiment on two benchmarks, Pascal Voc 2012 [11] and MS COCO 2014 [30]. The former contains 20 object classes, with 10,582 images for training, 1,449 images for validation, and 1,456 for testing. MS COCO 2014 has 80 labeled classes, 80,781 training and 40,321 validation images. We evaluate the generated pseudo labels and the segmentation results with their ground-truth segmentation labels. Both experiments are evaluated with mean intersection over union (mIoU). Besides, we further explore the effectiveness of our method on weakly supervised instance segmentation [28], [38] (WSIS, with box-level annotations). We conduct experiments on MS COCO 2017 dataset, which has 115K images for training and 5K evaluation images. During inference, we report AP, AP_{50}, AP_{75} (averaged precision over different IoU thresholds) for the instance segmentation performance evaluation.

Implementation details. We set PPC [10] as the baseline and perform ablation experiments on the proposed FB and IF contrastive learning to test their effectiveness and compare them with peer studies. Next, we add the full method to various WSSS models [10], [25], [43] and experiment on the aforementioned datasets to show our FBR's generality and superiority. In training, we set the cluster number K to 8. We implement the projection head φ_{fg} and φ_{bg} with a 1×1 convolutional layer followed by ReLU and set D to 128. We return 256 negative keys for every query and set τ in Eq. 7 and Eq. 9 to 0.5 and 0.1. For semantic segmentation practice, we adopt DeepLab-v2-ResNet101 [5] as the backbone. In box-level WSIS, we deploy FBR on BoxInst [38], set the feature dimensionality D to 64, and keep other hyperparameters unchanged. To reduce the computational overhead, all images are resized to have their shorter side in the range [480, 640] during training. More details are reported in the Supplementary.

B. Overall results

In this section, we compare the proposed FBR method with existing WSSS studies regarding the accuracy of generated pseudo labels and yielding semantic segmentation results. Besides, we extend FBR to WSIS tasks to show its benefits.

Results of pseudo labels: In this part, we evaluate the quality of the pseudo-segmentation masks. We add FBR on three representative baselines, PPC [10] that use saliency map (I+S) in training, SEAM [43] and AMN [25] (I) that do not use, to manifest our approach's generality. Note that SEAM and PPC require cross-view inputs, while AMN does not. Table II reports the comparison performance. In Seed results, FBR achieves 7.2%/1.7%/0.7% mIoU improvements on SEAM/PPC/AMN. These gains are almost maintained after employing CRF (Conditional Random Field), IRN [1], or PSA [2] to refine the seeds and obtain the pseudo segmentation labels (**Mask**). Overall, our method achieves SOTA performance on Pascal Voc 2012 train set, surpassing the best-known WSSS study by 1.8% mIoU (75.9% $vs.$ 74.1%).

Fig. 4 compares the qualitative results of CAMs. Note that FBR greatly helps the baselines activate more object regions (the dog and cat example on AMN [25] column) and enhances

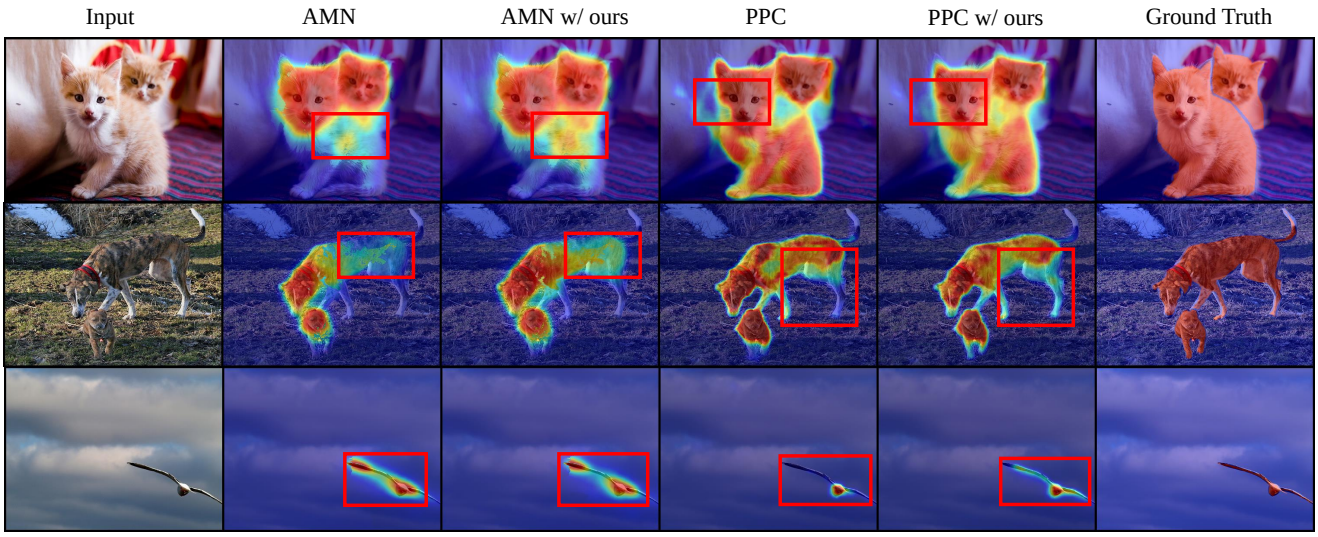


Fig. 4: Example results of CAMs on Pascal Voc 2012 train set. From left to right: input images, results of AMN, results of AMN w/ ours, results of PPC, results of PPC w/ ours and the ground truth. The red boxes highlight the refined details.

Fig. 5: Qualitative semantic segmentation results. The left figures are results from Pascal Voc 2012 val set, and the right ones are from MS COCO 2014 val set. (a) Input images, (b) Ours, (c) Ground truth.

their background discrimination ability and thus learns more accurate object boundaries (see the dog example on PPC).

Results of semantic segmentation. We conduct semantic segmentation practices with DeepLabV2-ResNet101 [5] and follow the training settings of existing implementations [1] [2]. Table III reports the mIoU results on Pascal Voc 2012 validation set and the test set, and compares our FBR method with recent WSSS studies. Training under the same setting, we improve SEAM and AMN by 3.6% and 1.7% mIoU on the test set of Pascal Voc, yielding the SOTA performance in the image-level (I) setting (72.4%→ 73.2% mIoU [3]).

Besides, PPC [10] equipped with our FBR method achieves 74.2% mIoU and 74.9% mIoU [4] on Pascal Voc 2012 Val and test sets, exceeding all (I+S) WSSS studies. Moreover, FBR achieves 45.6% mIoU on MS COCO 2014 val set (Table IV), obtaining the SOTA result. Fig. 5 presents example segmentation results on both benchmarks.

[1]https://github.com/YudeWang/deeplabv3plus-pytorch.git
[2]https://github.com/kazuto1011/deeplab-pytorch.git
[3]http://host.robots.ox.ac.uk:8080/anonymous/30LARO.html
[4]http://host.robots.ox.ac.uk:8080/anonymous/BHSCOK.html

TABLE II: mIoU (%) comparison of the seeds, seeds w/ CRF, and pseudo masks (**Mask**) on Pascal Voc 2012 train set. Throughout the paper, bolded and underlined represent the SOTA and the second-best SOTA results, respectively.

Method	Seed	+CRF	Mask
Refine with PSA [2]:			
SEAM CVPR '20 [43]	55.4	56.8	63.6
Ours (SEAM-based)	62.6$_{+7.2}$	65.3$_{+8.5}$	69.9$_{+6.3}$
RIB NeurIPS '21 [22]	56.5	62.9	68.6
EPS CVPR '21 [26]	69.4	71.4	71.6
RCA CVPR '22 [52]	-	-	<u>74.1</u>
PPC CVPR '22 [10]	<u>70.5</u>	<u>73.3</u>	73.3
Ours (PPC-based)	**72.2**$_{+1.7}$	**75.5**$_{+2.2}$	**75.9**$_{+2.6}$
Refine with IRN [1]:			
MCT CVPR '22 [48]	61.7	-	69.1
CLIMS CVPR '22 [46]	56.6	-	70.5
W-OoD CVPR '22 [24]	59.1	65.5	72.1
AMN CVPR '22 [25]	<u>62.2</u>	-	72.2
ACR CVPR '23 [20]	60.9	65.9	72.3
BECO CVPR '23 [34]		-	<u>73.0</u>
Ours (AMN-based)	**62.9**$_{+0.7}$	-	**73.1**$_{+0.1}$

Fig. 6: Qualitative instance segmentation results on MS COCO 2017 val set. Note that example pairs are from BoxInst [38] (left images) and ours (BoxInst w/ FBR, right images); we see that FBR substantially fines the mask predictions.

TABLE III: Segmentation results (mIoU %) comparisons with other SOTA studies on Pascal Voc 2012 validation (**Val.**) and test (**Test**) set. The supervisions (**Sup.**) used in the training include the image-level labels (**I**) and salience maps (**S**).

Method	Sup.	Backbone	Val.	Test
SEAM CVPR '20 [43]	I	ResNet38	64.5	65.7
ReCAM CVPR '22 [8]	I	ResNet101	68.4	68.2
SIPE CVPR '22 [6]	I	ResNet101	68.8	69.7
W-OoD CVPR '22 [24]	I	ResNet38	70.7	70.1
AMN CVPR '22 [25]	I	ResNet101	70.7	70.6
VIT-PCM ECCV '22 [35]	I	ResNet101	70.3	70.9
SBCE ECCV '22 [45]	I	ResNet101	70.0	71.3
AEFT ECCV '22 [50]	I	ResNet38	70.9	71.7
BECO CVPR '23 [34]	I	ResNet38	72.1	71.8
ToCo CVPR '23 [36]	I	VIT-B [9]	71.1	72.2
ACR CVPR '23 [20]	I	ResNet38	**72.4**	72.4
Ours (SEAM-based)	I	ResNet38	68.9	69.3
Ours (AMN-based)	I	ResNet101	71.8	**73.2**
MCT CVPR '22 [48]	I+S	ResNet38	71.9	71.6
L2G CVPR '22 [16]	I+S	ResNet101	72.1	71.7
ReCAM CVPR '22 [8]	I+S	ResNet101	71.8	72.2
RCA CVPR '22 [52]	I+S	ResNet38	72.2	72.8
SBCE ECCV '22 [45]	I+S	ResNet101	71.8	73.4
PPC CVPR '22 [10]	I+S	ResNet101	72.6	73.6
Ours (PPC-based)	I+S	ResNet101	**74.2**	**74.9**

TABLE IV: Semantic segmentation (mIoU %) comparisons with other WSSS studies on COCO 2014 validation set.

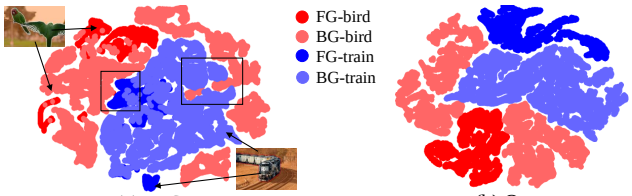
Method	Sup.	Backbone	Val.
SIPE CVPR '22 [6]	I	ResNet38	43.6
RIB NeurIPS '21 [22]	I+S	ResNet101	43.8
L2G CVPR '22 [16]	I+S	ResNet101	44.2
AMN CVPR '22 [25]	I	ResNet101	44.7
AEFT ECCV '22 [50]	I	ResNet38	44.8
ReCAM CVPR '22 [8]	I	ResNet101	45.0
VIT-PCM ECCV '22 [35]	I	ViT-B/16 [9]	45.0
BECO CVPR '23 [34]	I	ResNet101	45.1
ACR CVPR '23 [20]	I	ResNet38	45.3
Ours (AMN-based)	I	ResNet101	**45.6**

TABLE V: Instance segmentation comparisons (AP %) with other WSIS studies (box-level) on COCO 2017 validation set.

Method	Backbone	AP	AP_{50}	AP_{75}
DiscoBox [21] ICCV '21	ResNet50	30.2	52.1	30.7
Box2Mask-C [28] ECCV '22	ResNet101	33.5	56.9	34.2
BoxInst [38] CVPR '21	ResNet50	30.9	53.3	31.2
BoxInst [38] CVPR '21	ResNet101	32.1	55.3	32.4
Ours (BoxInst-based)	ResNet50	**32.9**	56.6	33.4
Ours (BoxInst-based)	ResNet101	**34.1**	57.7	34.9

TABLE VI: Ablation study (on PPC [10]) in terms of seed generation on Pascal Voc 2012 train set. † denotes excluding the background in the training. ‡ denotes adopting the auxiliary background segmentation on †. **FB:** fore-to-background contrast. **IF:** intra-foreground contrast.

Baseline	FB	IF	mIoU(%)
✓			73.3
✓†			73.6+(0.3)
✓†	✓		74.7+(1.4)
✓‡	✓		75.0+(1.7)
✓		✓	73.9+(0.6)
✓‡	✓	✓	**75.5**+(2.2)



(a) PPC (b) Ours

Fig. 7: Background-foreground feature visualization via t-SNE [39]. (a) foreground features are confused with background information (the cropped regions). (b) FB contrast learns compacted background features by fine-grained recognition and well separates background and foreground features.

Results of instance segmentation. In Table V, we test FBR on BoxInst [38] and report weakly supervised instance segmentation results on MS COCO 2017. Training under the same settings, FBR improves the baseline model by 2.0% AP no matter the backbone models, surpassing Box2Mask [28] by 0.6% (33.5%→34.1%). Fig. 6 shows the example results.

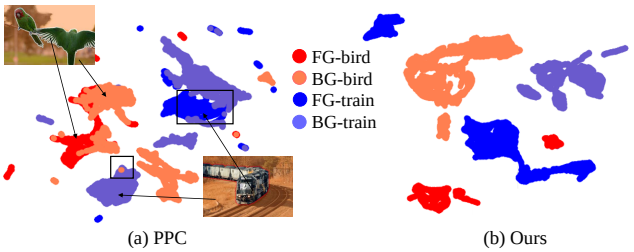
TABLE VII: NROI evaluation. **Left:** w/o and w/ φ_{bg} mean that background features to obtain NROIs are from Z_{fg} and Z_{bg}, respectively. **Right:** Strategy comparison. Ours: pixel-to-NROI. Brute-force: pixel-to-pixel.

Method	mIoU(%)
w/o φ_{bg}	74.0
w/ φ_{bg}	**74.7**

Method	mIoU(%)
pixel-to-pixel	74.5
pixel-to-NROI	**75.0**

C. Ablation Study

Below we test the effectiveness of each component and design. All results are the averages over five runs.



(a) PPC (b) Ours

Fig. 8: UMAP [41] visualization for background-foreground feature comparison.

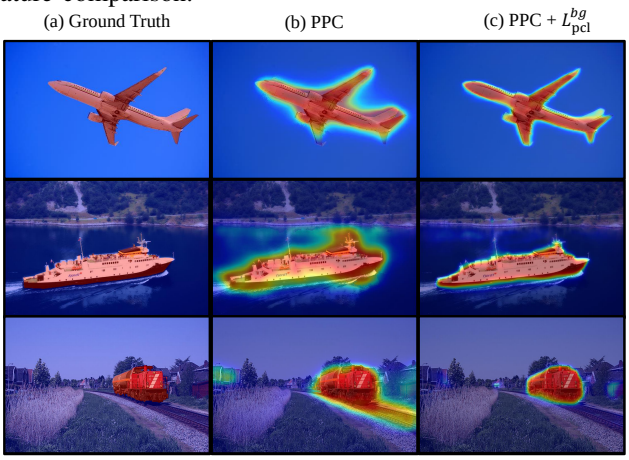


Fig. 9: Effect of FB contrast (i.e., L_{pcl}^{bg}) on PPC [10].

Overall ablation result. In Table VI, we ablate the FB and the IF contrastive learning in sequence. For the baseline PPC [10] that estimates a general background prototype to proceed with contrastive learning, the 2nd row (excluding background) shows 0.3% mIoU gap (73.3%→73.6%). This confirms our assumption that a single prototype is incapable of describing the image background and may even negatively influence foreground classes. The proposed FB and IF contrastive learning contribute 1.4% mIoU (73.3%→74.7%) and 0.6% mIoU improvements (73.3%→73.9%). Besides, the auxiliary segmentation loss L_{seg} facilitates better FB and brings an extra 0.3% mIoU gain (74.7%→75.0%). Overall, our FBR approach improves the baseline by 2.2% mIoU.

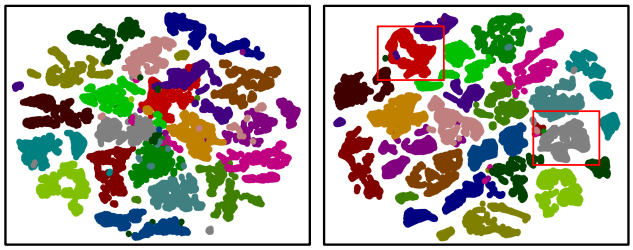
Effect of NROI. We ablate the design and method, and analyze the improvement source to assess the effect:

- Projection head φ_{bg}: Unlike existing studies implementing contrastive learning [7], [42] in a common space, we consider the semantic discrepancy between the image foreground and background, i.e., the background has finer semantic granularities. We argue that one representation space is insufficient to generate two different primitives, i.e., NROI and foreground prototype. φ_{bg} maps the background semantics to an independent space from the foreground classes, enabling expressive projection and better semantic representation. In Table VII (Left), we compare the single-head design, namely using one projection head φ_{fg} to represent both foreground and background, with ours; the 0.7% mIoU gain (74.0% →74.7%) verifies the benefit of our design.
- Contrast strategy: We compare our pixel-to-NROI contrast with the brute-force strategy (Fig. 3 (a)) that puts all background features of Z_{bg} into the memory bank to compute L_{pcl}^{bg}, a.k.a. pixel-to-pixel. To make a fair comparison, we set the bank size to 100**K** (causing longer sampling time) in the brute-force case, yet set ours to

TABLE VIII: Ablations of the clustering (**Clusters**) and memory bank (**Memo.**). **Left:** We compare mIoUs on Pascal Voc train set when setting K (the number of clusters) with different values. **Right:** we evaluate the effect of memory bank size.

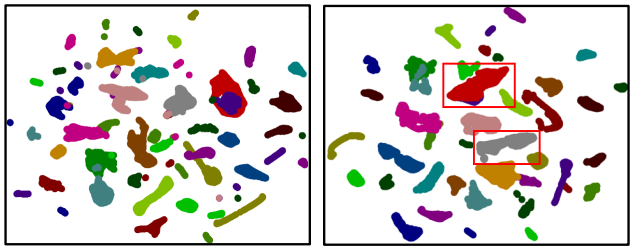
Clusters	mIoU (%)
K=4	74.6
K=6	74.8
K=8	**75.0**
K=16	74.6

Memo.	mIoU (%)
3×10^4	74.5
5×10^4	**75.0**
8×10^4	74.9
10^5	74.6



(1) PPC (2) Ours

(A) Category-level feature visualization via t-SNE.



(1) PPC (2) Ours

(B) Category-level feature visualization via UMAP.

Fig. 10: Visualization of semantic features via (A) t-SNE and (B) UMAP. **Left**: PPC [26]. **Right**: PPC w/ L_{pcl}^{fg}.

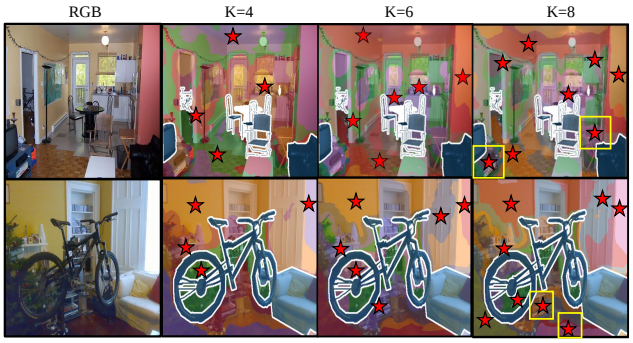


Fig. 11: Example results of cluster maps and NROIs. We visualize the clusters and NROIs in different cases. **From left to right:** RGB images and cluster maps (when K=4/6/8). In each cluster, we use the pixel closest (measured with the square distance computed along the feature channel) to the centroid as NROI (marked with the star); the white cropped area is the foreground. The yellow boxes highlight the semantics that are newly recognized with the increasing K.

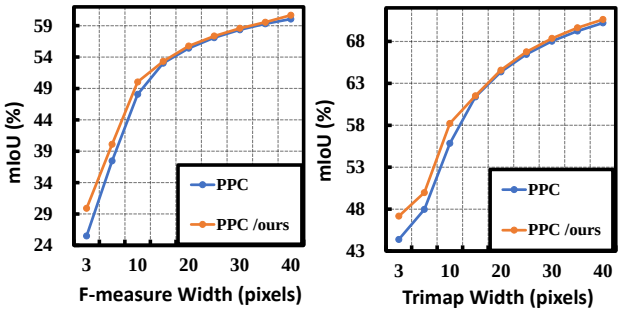


Fig. 12: Contour evaluation of seeds on Pascal Voc 2012 train set. The figures show how F-measure and Trimap mIoU results vary with the pixel width of the ground-truth boundary region.

50**K**. In Table VII (Right), ours outperforms the pixel-to-pixel contrast by 0.5% mIoU, showing NROI's benefits against the common pixel feature-based representation.

- Source of performance gain: The improvements obtained through NROIs mainly stem from the fine-grained background semantic recognition. In Fig. 7 and Fig. 8, we respectively employ t-SNE [39] and UMAP [41] for background-foreground feature analysis. Detailly, we compare features of PPC [10] (from its f_{proj}) before and after using FB contrast. Although the projected manifolds by these two methods have different shapes due to the difference in dimension reduction technique, we get consistent findings: in (a) of both figures, the scatted background features result in the spurious semantic correlation; foreground features are contaminated with confusing intra- and inter-image background information, e.g., the overlapped part in the boxes. In contrast, (b) adopts NROIs to capture the fine-grained background semantics and begets a structured background feature space. Furthermore, contrastive optimization suppresses the suspicious background cues and distinguishes them from the foreground, effectively avoiding confusion. Additionally, we select three representatives of co-occurring background semantics (sky, lake, and railroad) and compare CAMs in Fig. 9. FB contrast resists background disturbances and obtains more reliable results.

Hyperparameter discussion. In this part, we discuss the effect of NROI's hyperparameters:

- Number of clusters: In Table VIII (Left), we ablate the cluster number K and report the seed accuracy. Intuitively, increasing K will return finer-grain background semantics, but make NROIs less meaningful and increase the computational cost. Based on the comparison result, we set $K = 8$ throughout the paper. Besides, we visualize the clustering maps and NROIs in Fig. 11. We see that clustering groups the image background and fine-grained semantics are identified with the increasing K. For instance, the "TV cabinet" and the "table" (the 1st row) and the "book" and "floor" (the 2nd row). As for the extracted NROIs, i.e., the cluster centroids, indicating the most typical representation of the individual semantic group, usually located in the object-central region, capturing meaningful background semantics, with which we can model background content explicitly.

TABLE IX: Effects of loss weight λ_1 and λ_2 on mIoU (%) results of the seeds, seeds w/ CRF, and pseudo masks (**Mask**) on Pascal Voc 2012 train set. λ_1 and λ_2 are used to balance the L_{pcl}^{bg} and L_{pcl}^{fg} in L_{pcl} (Eq. 10), respectively.

λ_1	λ_2	Seed	+CRF	Mask
0.05	0.01	71.3	74.0	74.7
	0.05	71.6	74.8	75.1
	0.10	72.1	74.8	75.4
0.10	0.01	**72.2**	**75.5**	**75.9**
	0.05	71.7	74.7	75.3
	0.10	71.7	74.7	75.2
0.20	0.01	71.5	74.2	74.9
	0.05	70.9	73.7	74.4
	0.10	70.8	73.5	74.1

TABLE X: Negative sampling method comparison [10], [40] (mIoU (%)) in terms of seeds (**Seed**) and seeds with CRF (**+CRF**) on Pascal Voc 2012 train set.

Method	Seed	+CRF
PPC (hard, original)	70.5	73.3
PPC (active, ours)	**71.1**$_{+0.6}$	**73.9**$_{+0.6}$

- Size of memory bank: In Table VIII (Right), we investigate the effect of the memory bank size. The larger bank can store more NROIs yet also indicates a longer sampling time. Therefore, we set the bank size to 50**K**, 200**K** for Pascal Voc 2012 and MS COCO 2014.
- Loss weights: In Table IX, we conduct ablation studies on the loss weights λ_1 and λ_2 to evaluate the effects of the two components of the contrastive loss L_{pcl} (in Eq. 10): L_{pcl}^{bg} and L_{pcl}^{fg}. A higher value of λ_1 suggests that L_{pcl} would focus more on the fore-to-background relationship, and vice versa. Our results indicate that FBR performs best when λ_1 and λ_2 reach 0.10 and 0.01, respectively. Besides, the comparison results exhibit that our FBR method is more sensitive to λ_1 than λ_2, showing the importance of fore-to-background contrastive learning.

Effect of active sampling. In this part, we evaluate our active method and analyze its effectiveness:

- Active sampling $vs.$ hard sampling: We compare the active method against the hard sampling [10], [40] that selects negative keys based on the "hardness" computation. In Table X, after replacing the hard sampling (adopted in PPC) with ours, the baseline is improved by 0.6% mIoU. Besides, we additionally experiment with both sampling

TABLE XI: Evaluate the sampling methods under different query settings. The hard sampling experiments are implemented with PPC [10]. **naive query:** all pixels are queries. **hard query:** half hard queries and half random queries [10], [40]. **adaptive query:** select queries based on Eq. 4.

Method	Seed	+CRF
baseline (EPS [26])	69.5	71.4
+hard sampling (naive query)	70.4	73.2
+hard sampling (hard query)	70.5	73.3
+hard sampling (adaptive query)	**70.5**$_{+(1.0)}$	**73.4**$_{+(2.0)}$
+active sampling (naive query)	70.8	73.5
+active sampling (hard query)	71.0	73.7
+active sampling (adaptive query)	**71.1**$_{+(1.6)}$	**73.9**$_{+(2.5)}$

methods on EPS and report extensive comparison results (in Table XI) under different query settings: naive query, hard query, and adaptive query. We observe that the active method consistently improves the baseline and always outperforms the hard sampling method, showing its robustness. Particularly, when proceeding with IF contrast (i.e., L_{pcl}^{fg}) with the adaptive query setting (in Eq. 4), our method improves EPS by 1.6% mIoU on seed and 2.5% mIoU on masks (against the maximum 1.0%/2.0% mIoU $\uparrow$ achieved by the hard sampling method [10]). More importantly, our active sampling draws the negatives on the fly, without the hardness calculation and the sorting, resulting in a lower computation cost.

- Activate complete object regions: Considering that background is present in nearly every image and the inaccuracies in seed results, our active approach excludes the background class during negative sampling. This allows us to concentrate on optimizing relationships within the foreground and thus learn more discriminative class features. When combined with IF contrast, we prevent interference from suspicious background pixel features (false positives) in the sampling, particularly those in the transition area between the foreground and background. This way, we learn more accurate object contours. In Fig. 10, we visualize the foreground feature space via t-SNE [39] and UMAP [41]. After the sampling method replacement, the cropped class features become more compacted, e.g., "car" and "chair". This observation presents our IF contrast's effectiveness in learning discriminative FG features. Besides, we employ F-measure [32] and Trimap [5], [49] to evaluate the object contour quality of the generated seeds. Given a pixel width, both metrics assess the alignment degree between the prediction and its ground truth within a narrow band region from the true semantic boundary. In Fig. 12, our obtained performance gain on PPC [10] mainly in the region near the object contour (pixel width ≤ 10), indicating our IF contrast's effects in learning complete object regions.

Compared with existing approaches [10], [31], FBR does not require cross-view inputs and incurs no inference overhead. Our results (Table VI & Fig. 7) verify that the major improvements come from the proposed NROI-based FB contrastive learning, significantly improving CAMs' ability in background discrimination and thus avoiding classification ambiguity. Our FBR has high versatility (across different baselines) thanks to its simplicity in design and modularized techniques; the advanced segmentation results manifest FBR's usefulness.

Limitations. Despite achieving SOTA results in various WSSS tasks, our approach encounters challenges in two cases: 1) when the image background is visually too similar to the foreground classes and 2) when the foreground classes have irregular object contours. As shown in Fig. 13, the background buildings are misclassified as a part of "train" in (a) upon confusing appearances. Besides, our method can not accurately predict the boundary of "potted plants" (b) due to their complex shapes. A plausible explanation for these failures is that our method focuses on correcting low-confidence regions

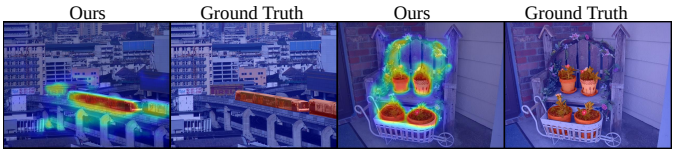


Fig. 13: Failure cases. We provide CAMs of (a) "train" and (b) "potted plants" and compare our FBR with the ground truth.

in CAMs, while these challenging regions already exhibit high activation scores during the classification process and are thus ignored. This problem can be mitigated with a powerful backbone, e.g., ViT, to obtain more precise seed predictions.

Future works. An intriguing direction is to integrate our approach with foundation models like SAM [18]. Leveraging SAM's object mask predictions, our FBR could accurately label the background/foreground semantics, generating pseudo-labels with high-quality object contours. These pseudo-labels, enriched with precise boundary information, could benefit segmentation practices in other domains, such as medical or remote sense images. However, an adaptive clustering method would be required to accurately determine the number of background semantics, which we leave for future development.

V. Conclusion

This paper proposed a simple fine-grained background representation method, FBR, to address the co-occurring background problem and learn integral object masks in weakly supervised semantic segmentation (WSSS). Our method designs a new background primitive and an active sampling method to perform the fore-to-background and intra-foreground contrastive learning. Extensive experiments on Pascal Voc and MS COCO demonstrated the good merits of FBR in generating pseudo masks, achieving new state-of-art performances in WSSS, and also benefiting the instance-level segmentation.

References

[1] J. Ahn, S. Cho, and S. Kwak, "Weakly supervised learning of instance segmentation with inter-pixel relations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Long Beach, US, 2019, pp. 2209–2218.

[2] J. Ahn and S. Kwak, "Learning pixel-level semantic affinity with image-level supervision for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Salt Lake City, US, 2018, pp. 4981–4990.

[3] I. Alonso, A. Sabater, D. Ferstl, L. Montesano, and A. C. Murillo, "Semi-supervised semantic segmentation with pixel-level contrastive learning from a class-wise memory bank," in *Proc. IEEE Int. Conf. Ccompu. Vis.*, Virtual/Online, 2021, pp. 8219–8228.

[4] W. Bae, J. Noh, and G. Kim, "Rethinking class activation mapping for weakly supervised object localization," in *Eur. Conf. Comput. Vis.* Glasgow, UK: Springer, 2020, pp. 618–634.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, and K. e. a. Murphy, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2017.

[6] Q. Chen, L. Yang, J. Lai, and X. Xie, "Self-supervised image-specific prototype exploration for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, USA, 2022, pp. 4278–4288.

[7] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. on Machine Learning*. PMLR, 2020, pp. 1597–1607.

[8] Z. Chen and T. e. a. Wang, "Class re-activation maps for weakly-supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 969–978.

[9] A. Dosovitskiy, L. Beyer, and A. e. a. Kolesnikov, "An image is worth 16x16 words: Transformers for image recogntion at scale," 2020. [Online]. Available: https://arxiv.org/abs/2010.11929

[10] Y. Du, Z. Fu, Q. Liu, and Y. Wang, "Weakly supervised semantic segmentation by pixel-to-prototype contrast," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4320–4329.

[11] M. Everingham and J. Winn, "The pascal visual object classes challenge 2012 (voc2012) development kit," *Patt. Anal., Statist. Modelling and Computat. Learning, Tech. Rep*, vol. 8, 2011.

[12] J. Fan and Z. Zhang, "Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Seattle, US, 2020, pp. 4283–4292.

[13] J. Fan, Z. Zhang, T. Tan, C. Song, and J. Xiao, "Cian: Cross-image affinity net for weakly supervised semantic segmentation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, New York, USA, 2020, pp. 10 762–10 769.

[14] S. Hong, S. Kwak, and B. Han, "Weakly supervised learning with deep convolutional neural networks for semantic segmentation: Understanding semantic layout of images with minimum human supervision," *IEEE Sig. Proc. Magaz.*, vol. 34, no. 6, pp. 39–49, 2017.

[15] H. Hu, J. Cui, and L. Wang, "Region-aware contrastive learning for semantic segmentation," in *Int. Conf. Comput. Vis.*, Virtual/Online, 2021, pp. 16 291–16 301.

[16] P.-T. Jiang, Y. Yang, Q. Hou, and Y. Wei, "L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 16 886–16 896.

[17] T.-W. Ke, J.-J. Hwang, and S. X. Yu, "Universal weakly supervised segmentation by pixel-to-segment contrastive learning," 2021. [Online]. Available: https://arxiv.org/abs/2105.00957

[18] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, "Segment anything," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.

[19] A. Kolesnikov and C. H. Lampert, "Seed, expand and constrain: Three principles for weakly-supervised image segmentation," in *Eur. Conf. Comput. Vis.* Amsterdam, US: Springer, 2016, pp. 695–711.

[20] 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*, 2023, pp. 11 329–11 339.

[21] S. Lan and Z. e. a. Yu, "Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision," in *Int. Conf. Comput. Vis.*, Virtual/Online, 2021, pp. 3406–3416.

[22] J. Lee, J. Choi, J. Mok, and S. Yoon, "Reducing information bottleneck for weakly supervised semantic segmentation," *AAdv. Neural Inform. Process. Syst.*, vol. 34, pp. 27 408–27 421, 2021.

[23] J. Lee and E. Kim, "Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Virtual/Online, 2021, pp. 4071–4080.

[24] J. Lee, S. J. Oh, S. Yun, J. Choe, E. Kim, and S. Yoon, "Weakly supervised semantic segmentation using out-of-distribution data," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 16 897–16 906.

[25] M. Lee, D. Kim, and H. Shim, "Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4330–4339.

[26] S. Lee and M. Lee, "Railroad is not a train: Saliency as pseudo-pixel supervision for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Virtual/Online, 2021, pp. 5495–5505.

[27] J. Li, P. Zhou, and C. e. a. Xiong, "Prototypical contrastive learning of unsupervised representations," 2020. [Online]. Available: https://arxiv.org/abs/2005.04966

[28] W. Li, W. Liu, J. Zhu, M. Cui, X.-S. Hua, and L. Zhang, "Box-supervised instance segmentation with level set evolution," in *Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, 2022, pp. 1–18.

[29] F. Lin, Z. Liang, S. Wu, J. He, K. Chen, and S. Tian, "Structtoken : Rethinking semantic segmentation with structural prior," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[30] 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 *Eur. Conf. Comput. Vis.* Zurich, Switzerland: Springer, 2014, pp. 740–755.

[31] S. Liu, S. Zhi, E. Johns, and A. J. Davison, "Bootstrapping semantic segmentation with regional contrast," 2021. [Online]. Available: https://arxiv.org/abs/2104.04465

[32] F. Perazzi, J. Pont-Tuset, B. McWilliams, and L. e. a. Van Gool, "A benchmark dataset and evaluation methodology for video object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Las Vegas, Nevada, 2016, pp. 724–732.

[33] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," 2020. [Online]. Available: https://arxiv.org/abs/2010.04592

[34] S. Rong, B. Tu, Z. Wang, and J. Li, "Boundary-enhanced co-training for weakly supervised semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 19 574–19 584.

[35] S. Rossetti, D. Zappia, M. Sanzari, M. Schaerf, and F. Pirri, "Max pooling with vision transformers reconciles class and shape in weakly supervised semantic segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2210.1740

[36] 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*, 2023, pp. 3093–3102.

[37] Y. Sun, L. Su, S. Yuan, and H. Meng, "Danet: Dual-branch activation network for small object instance segmentation of ship images," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[38] Z. Tian, C. Shen, X. Wang, and H. Chen, "Boxinst: High-performance instance segmentation with box annotations," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Virtual/Online, 2021, pp. 5443–5452.

[39] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Jour. of mach. learn. research*, vol. 9, no. 11, 2008.

[40] W. Wang and T. e. a. Zhou, "Exploring cross-image pixel contrast for semantic segmentation," in *Int. Conf. Comput. Vis.*, Virtual/Online, 2021, pp. 7303–7313.

[41] Y. Wang, H. Huang, and C. e. a. Rudin, "Understanding how dimension reduction tools work: an empirical approach to deciphering t-sne, umap, trimap, and pacmap for data visualization," *Jour. of Mach. Learn. Research*, vol. 22, no. 1, pp. 9129–9201, 2021.

[42] Y. Wang, H. Wang, and Y. e. a. Shen, "Semi-supervised semantic segmentation using unreliable pseudo-labels," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4248–4257.

[43] Y. Wang, J. Zhang, M. Kan, S. Shan, and X. Chen, "Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, Seattle, US, 2020, pp. 12 275–12 284.

[44] J. Wu, H. Fan, Z. Li, G.-H. Liu, and S. Lin, "Information transfer in semi-supervised semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[45] T. Wu, G. Gao, J. Huang, and X. e. a. Wei, "Adaptive spatial-bce loss for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer, 2022, pp. 199–216.

[46] J. Xie and X. e. a. Hou, "Clims: Cross language image matching for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4483–4492.

[47] J. Xie, J. Xiang, J. Chen, X. Hou, X. Zhao, and L. Shen, "C2am: Contrastive learning of class-agnostic activation map for weakly supervised object localization and semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 989–998.

[48] L. Xu and W. Ouyang, "Multi-class token transformer for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4310–4319.

[49] X. Yin, D. Min, Y. Huo, and S.-E. Yoon, "Contour-aware equipotential earning for semantic segmentation," *IEEE Trans. Multimedia*, pp. 1–11, 2022.

[50] S.-H. Yoon and H. e. a. Kweon, "Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation," in *Eur. Conf. Comput. Vis.* Tel Aviv, Israel: Springer Nature Switzerland Cham, 2022, pp. 326–344.

[51] T. Zhou, W. Wang, E. Konukoglu, and L. Van Gool, "Rethinking semantic segmentation: A prototype view," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 2582–2593.

[52] T. Zhou and M. e. a. Zhang, "Regional semantic contrast and aggregation for weakly supervised semantic segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, New Orleans, US, 2022, pp. 4299–4309.

Xu Yin received the B.S degree in information security from Chongqing university of posts and telecommunications, Chongqing, China, in 2017, and the M.S degree from Computer Engineering Department, Inha University, Incheon, South Korea, in 2020. He is currently working toward the Ph.D degree in computer science, at Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea.

Woobin Im received MS Degree in School of Computing at Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2018. He is currently a Ph.D. candidate in Computer Science at KAIST, South Korea, where he is advised by Professor Sung-Eui Yoon. His research focuses on solving computer vision problems using various machine learning techniques, with a particular interest in video analysis, optical flow, spacetime representations, neural radiance fields, and generative models.

Dongbo Min received the BS, MS, and PhD degrees from the School of Electrical and Electronic Engineering, Yonsei University, Seoul, South Korea, in 2003, 2005, and 2009, respectively. From 2009 to 2010, he was a post-doctoral researcher with Mitsubishi Electric Research Laboratories (MERL), Cambridge, Massachusetts. From 2010 to 2015, he was with the Advanced Digital Sciences Center (ADSC), Singapore. From 2015 to 2018, he was an assistant professor in the Department of Computer Science and Engineering, Chungnam National University, Daejeon, South Korea. Since 2018, he has been in the Department of Computer Science and Engineering, Ewha Womans University, Seoul, South Korea. His current research interests include computer vision, deep learning, video processing, and continuous/discrete optimization. He is a senior member of the IEEE.

Yuchi Huo is a "Hundred Talent Program" researcher in State Key Lab of CAD&CG, Zhejiang University. His research interests are in neural rendering, 3D computational vision, and optical neural networks, which are aiming for the realization of next-generation neural rendering pipeline and physical-neural computation.

Fei Pan received his BS degree at the School of Telecommunications Engineering, Xidian University in China. He received his MS and Ph.D degree both at KAIST, Korea. He is currently working as a Research Fellow at the School of Computer Science Engineering, University of Michigan. His current research focuses on high-level computer vision models such as detection and segmentation and their adaptability and generalization to novel domains. Besides, he is also interested in generative models including diffusion models and large-scale vision and language models.

Sung-eui Yoon is a professor at Korea Advanced Institute of Science and Technology (KAIST). He received the B.S. and M.S. degrees in computer science from Seoul National University in 1999 and 2001, respectively. He received his Ph.D. degree in computer science from the University of North Carolina at Chapel Hill in 2005. He was a postdoctoral scholar at Lawrence Livermore National Laboratory, USA. His research interests include rendering, image search, and motion planning spanning graphics, vision, and robotics. He has published more than 70 technical papers in top journals and conference related to graphics, vision, and robotics. He also gave numerous tutorials on ray tracing, collision detection, image search, and sound source localization in premier conferences like ACM SIGGRAPH, IEEE Visualization, CVPR, and ICRA. He served as conf. co-chair and paper co-chair for ACM I3D 2012 and 2013 respectively. At 2008, he published a monograph on real-time massive model rendering with other three co-authors. Recently, he also published an online book on Rendering at 2018. Some of his papers received a test-of-time award, a distinguished paper award, and a few invitations to IEEE Trans. on Visualization and Graphics. He is currently senior members of IEEE and ACM.