Feature-level and Spatial-level Activation Expansion for Weakly-Supervised Semantic Segmentation

Junsu Choi[1,2*] Jin-Seop Lee[1*] Noo-ri Kim[1] SuHyun Yoon[1,3] Jee-Hyong Lee[1,†]
[1]Sungkyunkwan University, Suwon, South Korea
[2]Samsung Electronics, Suwon, South Korea
[3]Robotics Lab, Hyundai Motor Company, Uiwang, South Korea
junsu2.choi@samsung.com {wlstjq0602, pd99j, john}@skku.edu suhyunyoon@hyundai.com

Abstract

Weakly-supervised Semantic Segmentation (WSSS) aims to provide a precise semantic segmentation results without expensive pixel-wise segmentation labels. With the supervision gap between classification and segmentation, Image-level WSSS mainly relies on Class Activation Maps (CAMs) from the classification model to emulate the pixel-wise annotations. However, CAMs often fail to cover the entire object region because classification models tend to focus on narrow discriminative regions in an object. Towards accurate CAM coverage, Existing WSSS methods have tried to boost feature representation learning or impose consistency regularization to the classification models, but still there are limitation in activating non-discriminative area, where the focus of the models is weak. To tackle this issue, we propose FSAE framework, which provides explicit supervision of non-discriminative area, encouraging the CAMs to activate on various object features. We leverage weak-strong consistency with pseudo-label expansion strategy for reliable supervision and enhance learning of non-discriminative object boundaries. Specifically, we use strong perturbation to make challenging inference target, and focus on generating reliable pixel-wise supervision signal for broad object regions. Extensive experiments on the WSSS benchmark datasets show that our method boosts initial seed quality and segmentation performance by large margin, achieving new state-of-the-art performance on benchmark WSSS datasets. Our public code is available at https://github.com/obeychoi0120/FSAE.

1. Introduction

Semantic segmentation is widely used across various fields in the computer vision. It aims to classify pixel-wise labels of input images accurately. However, it requires expensive and laborious pixel-wise annotations for training semantic segmentation models. Weakly-Supervised Semantic Segmentation (WSSS) has been proposed to reduce these labeling costs. In WSSS, models are trained with relatively low-cost labels such as scribble [18,45], bounding box [22], and image-level class labels [4, 10, 21, 23]. The image-level WSSS methods are the most widely studied because of the high efficiency from the lowest labeling cost.

(a) (b) (c)

Figure 1. Visualization examples of CAM results. (a) Base model (PPC w/EPS [10]), (b) Ours, and (c) Ground truth. Our method activates more accurate object areas and produces clearer CAM results in the boundary regions.

Most image-level WSSS methods consist of two stages. In the first stage, they train the classification model and generate a pseudo-mask. Then, in the second stage, the semantic segmentation model is trained using the pseudo-mask as the ground-truth segmentation label. The most widely used technique for generating pseudo-masks in the first stage is refining class activation maps (CAMs) from the classification model. Therefore, CAMs are served as an initial seed for the overall process, so the high-quality CAMs which cover the entire object region are essential for achieving high segmentation performances.

However, CAM results extracted from classification models tend to highlight only a limited object region. Since

*Equal contribution
†Corresponding author

the model primarily focuses on regions that are most useful for distinguishing class labels, it can generate good pseudo-masks for only some object regions that are useful for classification. It struggles to generate pseudo-masks for others that are not useful for classification. This leads to inaccurate CAM results, and causes critical performance degradation in WSSS. To generate good CAM results, existing approaches learn pixel-wise feature representations, not only to learn sample-wise feature. They apply consistency regularization [8, 29, 31, 40, 48] and self-supervised learning [4, 10, 36] to the WSSS model. However, these approaches focus on learning pixel-wise representations, not on learning which class each pixel-wise representation belongs to.

In the segmentation task, the model needs to learn to assign each pixel-wise representation to its corresponding class. However, since existing methods are based on representation learning, they can only provide indirect guidance in learning whether a pixel is foreground or background. As a result, the CAM outputs still fail to generate clearly in boundary regions as shown in Fig. 1a. To learn pixel-wise representations with direct supervision, the model should effectively leverage reliable pixel-wise labels and extend them to boundary regions with different properties (e.g., texture, color, etc.) compared to interior regions.

In this paper, we propose a **F**eature-level and **S**patial-level **A**ctivation **E**xpansion (FSAE) framework to expand activation across the entire object for WSSS models. To train pixel-wise representations with direct supervision, FSAE extract reliable pixel-wise labels and use them to effectively expand the activation regions to the boundary areas. FSAE consists of two components, feature-level activation expansion (FAE) and spatial-level activation expansion (SAE). In FAE, we extract reliable pixel-wise labels through threshold-based pseudo-labeling, and then, train the model with these labels and pixel-wise feature representations. This allows the learning of weakly activated regions using reliable pixel-wise labels, which are propagated from highly activated regions. In SAE, we utilize dilation operations to effectively enhance the learning of non-discriminative boundary regions, where relying solely on augmentation is insufficient. This strengthens the learning of boundary regions, leading to generate clearer CAM results, as shown in Fig. 1b. To demonstrate our proposed method, we conducted experiments on various benchmark datasets, resulting in more precise CAMs and improved segmentation performance.

2. Related Works

2.1. Image-level WSSS

In contrast to fully-supervised semantic segmentation, which requires pixel-level annotations, WSSS uses 'weak' labels such as bounding box [22, 30], scribbles [18, 45], or lmage-level class labels [16, 24, 25, 27].

In image-level WSSS, the 2-stage methods using Class Activation Maps (CAMs) have been the most widely studied. In the first stage, an image classification model is trained with class labels to generate CAMs for each image. The CAMs are used as an initial seed and refined to pseudo-masks using post-processing techniques such as IRN [1] or standard dense-CRF. In the second stage, the off-the-shelf segmentation model is trained using these pseudo-masks as pixel-level labels. As the segmentation performance of 2-stage WSSS is highly dependent by the quality of initial seed, most studies focus on generating high-quality CAMs. However, inherently, CAMs tend to focus on the most discriminative area of the object, thus exhibiting poor object coverage.

To address these issues, most existing WSSS methods encourage the network to focus on non-discriminative regions of the object in order to increase CAM coverage. Adversarial erasing methods [15, 20, 37] erases highly-activated CAM regions, forcing the model to focus on other non-discriminative area and expand CAM activations. However, as there is no strict guideline for when to stop erasing, these methods typically suffer from over-erasing problem. These methods also entail multiple feed-forward passes of the image, resulting in high computational costs. Other approaches have attempted to reduce the supervision gap by generating additional supervision. One way is utilizing saliency maps generated from off-the-shelf saliency model [47] as an explicit supervision [6, 24, 25]. The saliency map highlights the salient regions of an object, providing relatively accurate boundary information. So, the saliency maps can be used for providing cues to discriminate object from background. Another way generates self-supervision signal via contrastive learning [5, 14, 19], which performs metric learning by constructing multiple views so that positive samples from same images get closer in feature space while negative samples from different images get farther. some studies [4, 10, 49] performed contrastive learning of representative feature embedding, known as prototypes, boosting activations of object regions similar to the prototype and suppress those are not. In addition, there are semantic mining techniques to share informations of object location [26, 35].

Aforementioned methods are indirect ways for expanding CAM activations. Different from these approaches, our work propagates explicit and reliable pixel-wise labels, for direct supervision for the model.

2.2. Consistency Regularization in WSSS

Consistency regularization is a widely used technique in the field of semi-supervised learning for model generalization. The technique is to apply perturbations to input or

network and impose consistency on the semantics or distribution of the output so that the prediction remains stable against various perturbations. The type of perturbations can be image augmentation [34, 42, 43] or network perturbations [7]. ST++ [43] applies strong data augmentation to unlabeled image, and CPS [7] applies consistency regularization to the prediction of the same input image by two differently initialized networks with different perturbations applied. These methods aim to provide additional self-supervision to the network through consistency regularization. As the technique improves significantly performance of semi-supervised semantic segmentation, recent WSSS studies brings the concept of the consistency regularization. CLIP-ES [29] utilized high-quality segmentation masks extracted from CLIP to enhance WSSS performance, and USAGE [31] introduced consistency regularization between seed areas generated from different views. Also, LPCAM and SFC [8, 48] proposed prototype-based learning methods, and MCTformer [40] introduced a method to extract class-specific CAM information for training. MARS [17] leverages semantically consistent features learned through USS to eliminate biased objects.

3. Proposed Method

To expand the narrow CAM activation, our FSAE framework generates an pixel-wise self-supervision signal for explicit learning of broad object regions. Overview is provided in Fig. 2. Our method consists of two main components: Feature-level Activation Expansion (FAE) and Spatial-level Activation Expansion (SAE). Classification models can explicitly learn various object features by FAE, and SAE enriches FAE by promoting learning of non-discriminative boundary area. We first review the conventional way of generating CAMs, and delve into details of the each component.

3.1. Preliminary: Obtaining CAMs

CAMs identify the most contributed regions for classification of an image. An image $x \in \mathbb{R}^{3\times H\times W}$ is passed to a feature extractor f to obtain a feature map $F \in \mathbb{R}^{D\times H\times W}$. The feature map is passed to global average pooling (GAP) layer and fully-connected layer with weight $w \in \mathbb{R}^{C\times D}$. C is the number of classes and D is the channel dimension. By Multiplying the classifier weight $w_{c,d}$ to the feature map F, we can derive the class score s_c for class c:

$$s_c = \frac{1}{HW}\sum_{d}^{D} w_{c,d} \sum_{i} F_{d,i} \tag{1}$$

and class activation map M_c for class c is as follows:

$$M_c = ReLU(\sum_{d}^{D} w_{c,d} F_{d,:}) \tag{2}$$

Recent methods [29, 32, 46] compute the CAMs directly from the feature map f without using a fully-connected layer. This method is proven theoretically equivalent to the traditional CAM calculation, but more simpler. Following this way, we use the class score map $S \in \mathbb{R}^{C\times H\times W}$ in our proposed method.

In Eq. (2), the weight w is optimized for discriminative classification. Consequently, CAMs also activate narrow part of the object. Thus, for segmentation task, additional modifications are necessary to expand the activation towards the entire region of the object.

3.2. Feature-level Activation Expansion (FAE)

Image classification models often concentrate on the most discriminative regions of an object. To broaden their focus, it is essential to learn various visual features, including the color/texture/shape of the object's region, only with precise supervision.

Thus, we propose a novel method based on consistency regularization with simple prior: if the network are learned to generate consistent predictions across wide variety of visual features in the object, the CAMs would be able to activate more integral object area. This motivates us to utilize weak-strong consistency, thus the pixel-wise prediction of the weak perturbed view is used as a online pseudo-label for the corresponding region of the strong perturbed view.

3.2.1 Generating Confident Online Pseudo-label

We aim to generate an explicit and accurate online pseudo-label using the pixel-wise CAM prediction for self-supervision. We first construct two views with a single image with different strength of perturbation. A weak transform $A_w(\cdot)$ is applied to the original image x to create the source view x_s, and a strong transform $A_s(\cdot)$ is applied to the source view to create the challenging target view x_t.

The types of weak transforms applied to the source view x_s are random resize, random crop, and color jittering. While in the target view x_t, gaussian blur and RandAugment [9] are additionally applied. Considering the property of the segmentation task, we exclude geometric transforms such as rotate, shear, and translate in RandAugment, using only color-space transforms. Our framework use the class prediction map $P_s \in \mathbb{R}^{H\times W}$ from the source score map S_s as online pseudo-labels to supervise target score map $S_t \in \mathbb{R}^{C\times H\times W}$.

$$x_s = A_w(x),\ x_t = A_s(x_s) \tag{3}$$

$$S_s = CAM(x_s),\ S_t = CAM(x_t) \tag{4}$$

$$P_s = \text{argmax}(S_s) \tag{5}$$

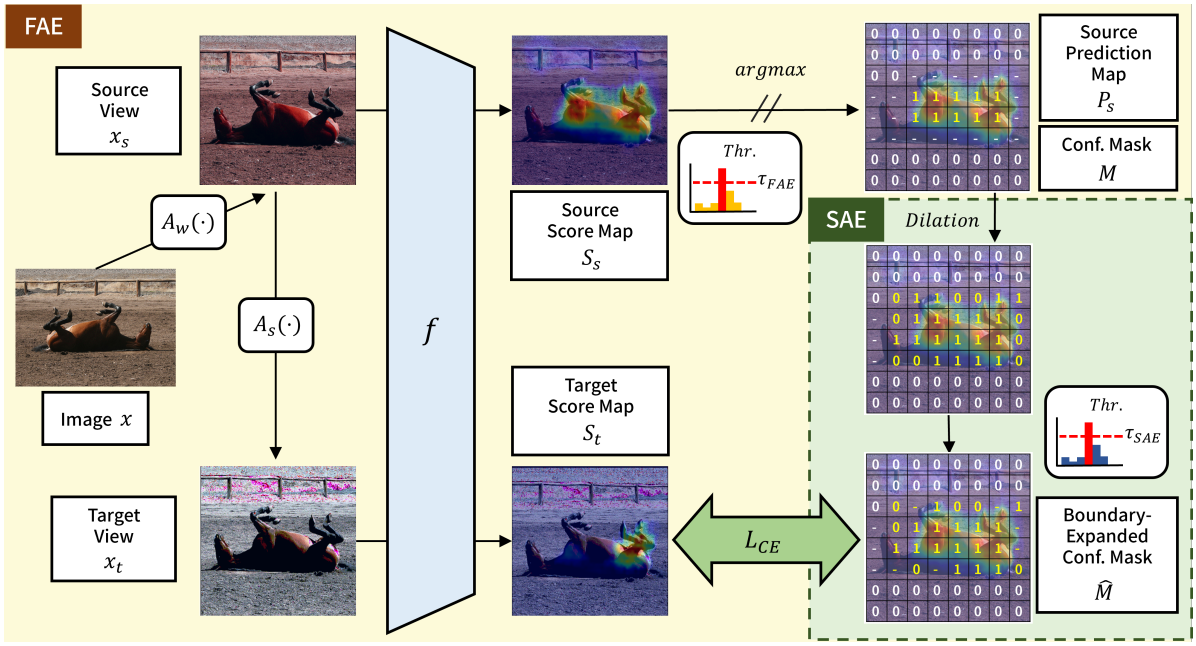


Figure 2. **Our FSAE framework**. Our method propagates the reliable pixel-wise pseudo-labels to strongly perturbed target view and enforces consistent prediction across wide variety of visual features.

Figure 3. Online pseudo-label regions (red) and actual object boundary area (white).

However, we do not use all pixels of P_s as online pseudo-labels, because it may contain inaccurate predictions and harm the training. Therefore, we create a high-confidence mask $M \in \mathbb{R}^{H \times W}$ for each pixel in P_s with confidence threshold τ_{FAE} to filter out unreliable pixel predictions. Consequently, high-confidence online pseudo-labels are generated by the element-wise multiplication of source prediction map P_s and high-confidence mask M. M can be formulated as follows:

$$M_i = \begin{cases} 1 & \text{if} \quad \text{softmax}(S_s)_i \geq \tau_{FAE} \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

where i denotes i-th pixel in source view S_s high-confidence mask M. The FAE facilitates the network to make consistent predictions on various object visual features, and the high-confidence pseudo-label area is gradually expanded by weak-strong consistency.

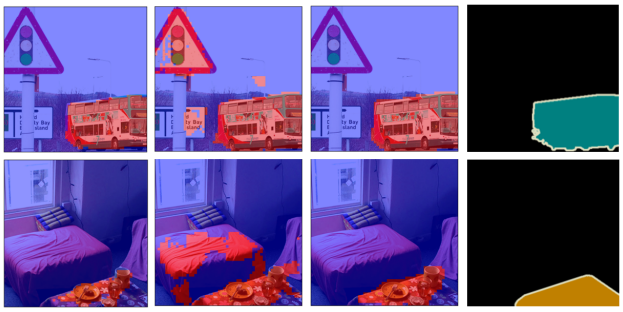
3.3. Spatial-level Activation Expansion (SAE)

The FAE helps the network to generate accurate and consistent predictions for a wide variety of visual features of the object. Nevertheless, the pseudo-label area is relatively discriminative and the complete coverage of the ambiguous low-confidence regions is not guaranteed. Notably, in Fig. 3, the low-confidence regions (blue) around the pseudo-label area are often located at the actual object boundaries, and this because of the absent of explicit information about object boundary region in WSSS.

Therefore, it would be beneficial if the network could properly learn these low-confidence regions with proper labels. To do so, we spatially expand the pseudo-label area towards the object boundary, by applying a standard *dilation* operation with kernel size k to the high-confidence mask M. This is based on two priors: 1) The pixel-wise predictions of the high-confidence pseudo-label area is accurate, and 2) adjacent pixels would have similar semantics, considering the characteristic of image data.

3.3.1 Boundary Expansion of Online Pseudo-label Region

During expansion of the pseudo-label area, noisy prediction be involved. Therefore, We set another confidence threshold τ_{SAE} to the confidence mask of added area M_{bdry}, filter out the unreliable predictions. The procedure is different with setting the first confidence threshold τ_{FAE} with lower value. As seen in Fig. 4, SAE can prevent confus-



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

Figure 4. (a) Pseudo-label area (red) with high threshold, (b) Pseudo-label area with low threshold, (c) Dilation of (a), and (d) GT.

ing areas involve in training by expanding the pseudo-label region near the boundary, whereas naively lowering τ_{FAE} includes noisy predictions.

$$M_{bdry,i} = \begin{cases} 1 & \text{if} \quad (\text{Dilation}(M) - M)_i = 1 \\ & \text{and} \quad \text{softmax}(S_s)_i \geq \tau_{SAE} \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

As a result, the expanded confidence mask $\hat{M}$ for generating online pseudo-labels is a union of initial confidence mask M and M_{bdry}:

$$\hat{M} = M \cup M_{bdry} \tag{8}$$

Finally, we derive the proposed Feature-level and Spatial-level Activation Expansion (FSAE) loss as the cross-entropy loss between target score map S_t and online pseudo label P_s for area corresponding to expanded confidence mask $\hat{M}_i$. $i \in \mathbb{I}$ denotes the each pixel of the CAM, and λ denotes the proportion of our loss.

$$\mathcal{L}_{FSAE} = \lambda \cdot \frac{1}{N} \sum_{n=1}^{N} \sum_{i \in \mathbb{I}} \hat{M}_i \mathcal{L}_{CE}(S_t, P_s) \tag{9}$$

To extract L_{FSAE}, we only use representations from feature maps, allowing our proposed method to be easily incorporated into other base models. Our approach is to combine the proposed L_{FSAE} with the objectives of existing methods and optimize them as total objectives.

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets and Evaluation Metrics

We conduct our experiment on the most popular benchmarks, PASCAL VOC 2012 [11] and MS COCO 2014 [28]. PASCAL VOC 2012 consists of 20 foreground classes and 1 background class for 1,464/1,449/1,456 images in training/validation/test set. Following the previous WSSS works [10, 21, 24, 40], we use the augmented train set [13] with 10,582 images for training. COCO 2014 consists of 80 foreground classes and 1 background class, for 82,081 and 40,137 in training/validation set. For all experiments, the mean Intersection-over-Union (mIoU) is used as the evaluation metric.

4.1.2 Base model

To validate the effectiveness of our proposed FSAE, we attach our method to the powerful WSSS model PPC$_{w/EPS}$, SIPE [4], and MCTformer [40]. PPC performs contrastive learning between prototypes, enriching representation learning via intra-view contrast and impose semantic consistency via inter-view contrast. EPS uses saliency maps explicitly as additional supervision, improving initial seed quality by ignoring co-occuring pixels that are not relevant to corresponding category. SIPE uses image-specific prototypes for refining CAM activations. MCTformer is a transformer-based WSSS method, which leverages multiple class tokens to generate more class-specific attention maps. Our FSAE can be incorporated seamlessly to these base models. Note that we report the reproduced results on our computing environment, with respect to the base models.

4.1.3 Implementation Details in Classification

We set the ratio $\lambda = 1.0$ of our proposed loss L_{FSAE}. The type of augmentation for source view is random rescale, random crop and color jitter. Additionally, gaussian blur and 3 random non-geometric transforms out of 9 Randaugment [9] transforms are added to source view to construct target view. The confidence threshold τ_{FAE} for initial confidence mask M_i is set to 0.95, and τ_{SAE} for expanded area M_{bdry} is set to 0.8. The kernel size k for *dilation* operation is set to 3. We use standard dense-CRF postprocessing to generate pseudo mask from initial seed.

The model-specific setting adjustments on VOC 2012 are as follows. In PPC$_{w/EPS}$, model is trained with batch size 8, learning rate 0.01 for 40k iterations. In SIPE, model is trained with batch size 16, learning rate 0.05 for 10 epochs. In case of MCTformer, we do not apply FSAE in the first 10 epochs.

On COCO 2014, we applied exponential moving average (EMA) of factor 0.999 in pseudo-label generation procedure in case of PPC$_{w/EPS}$ for inference stability. For the others, we follow the default hyperparameter setting of each base models. All experiments are conducted using a single NVIDIA RTX 3090.

Table 1. Quality of initial seeds and pseudo-masks (mIoU), which are refined by dense-CRF or by IRN [1]. Evaluated on PASCAL VOC 2012 train set. Net_{CAM} denotes the backbone of CAM generation network. 'R' denotes ResNet. * indicates reproduced results.

Method	Net_{CAM}	Seed	Mask
IRN [1] CVPR'19	R50	49.5	66.3
S-BCE [38] ECCV'22	R38	65.3	66.3
AMN [23] CVPR'22	R50	62.1	72.2
LPCAM$_{\text{w/AMN}}$ [8] CVPR'23	R50	65.3	72.7
CLIP-ES [29] CVPR'23	ViT-B	70.8	75.0
ToCo [41] CVPR'23	ViT-B	-	72.2
USAGE [31] ICCV'23	DeiT-S	67.7	72.8
SFC [48] AAAI'24	R38	64.7	73.7
SIPE [4] CVPR'22	R38	58.6	64.7
+ FSAE		**61.1**$_{+2.5}$	**70.5**$_{+5.8}$
PPC$_{\text{w/EPS}}$ [10] CVPR'22	R38	70.5	73.3
+ FSAE		**74.5**$_{+4.0}$	**77.0**$_{+3.7}$
MCTformer * [40] CVPR'22	DeiT-S	61.5	67.9
+ FSAE		**62.4**$_{+0.9}$	**68.8**$_{+0.9}$

4.1.4 Implementation Details in Segmentation

For semantic segmentation, DeepLabV1 [2] and DeeplabV2 [3] with ResNet-101 backbone are used for fair comparison. During inference, we use multi-scale and flip operations following previous studies [4, 10, 32], and standard dense-CRF to postprocess final segmentation mask.

4.2. Comparison with State-of-the-art Methods

4.2.1 Quality of Initial Seed and Pseudo-mask

First, we verify the effectiveness of our method by evaluating quality of both the initial seed and pseudo mask on PASCAL VOC 2012 train set. During CAM inference, we use multi-scale and flip inference following previous works. Table 1 compares initial seed quality between ours and previous works. The experimental results show that our proposed method outperforms other existing methods by a large margin. Incorporated with PPC$_{\text{w/EPS}}$, our method achieves 74.5% and 77.0% mIoU in initial seed and pseudo-mask, with +4.0%p and +3.7%p performance gain each. With SIPE, 61.1% and 70.5% mIoU were recorded with +2.5%p and +5.8%p gain. With MCTformer, 62.4% and 68.8% mIoU were recorded with +0.9%p and +0.9%p gain.

4.2.2 Segmentation Performance

For a fair comparison, We follow the training setting of each base model, and report the reproduced segmentation results. The comparison results are summarized in Tab. 2 and Tab. 3 for PASCAL VOC 2012 val/test set and Tab. 4 for COCO 2014 val set. The qualitative segmentation results in Fig. 5 shows our superiority. When incorporated into various baseline models, our method consistently improves segmentation performance across both DeepLabv1 and DeepLabv2 on multiple datasets, including PASCAL VOC 2012 and COCO 2014.

These results demonstrate that our proposed method can be effectively integrated into diverse base models, such as PPC, SIPE, and MCTformer. Furthermore, our method is effective regardless of the underlying backbones, including ResNet and Transformers, or the varying training objectives, such as consistency regularization and self-supervised learning.

Table 2. Segmentation results (mIoU) on PASCAL VOC 2012 using DeepLabV2. Sup. means the type of supervision. I: image-level labels, S: saliency maps, L: language supervision. Net_{Seg} denotes the backbone of segmentation network. * indicates reproduced results.

Method	Sup.	Net_{Seg}	Val	Test
IRN [1]CVPR'19	I	R50	63.5	64.8
S-BCE [38] ECCV'22	I	R38	68.5	69.7
AMN [23]CVPR'22	I	R101	69.5	69.6
LPCAM$_{\text{w/AMN}}$ [8]CVPR'23	I	R101	70.1	70.4
ToCo [41] CVPR'23	I	ViT-B	69.8	70.5
MCTformer * [40]CVPR'22	I	R38	69.0	69.8
+ FSAE	I		**69.8**$_{+0.8}$	**70.5**$_{+0.7}$
SIPE [4]CVPR'22	I	R101	68.8	69.7
+ FSAE	I		**69.9**$_{+1.1}$	**71.2**$_{+1.5}$
CLIMS [39]CVPR'22	I+L	R101	69.3	68.7
CLIP-ES [29]CVPR'23	I+L	R101	71.1	71.4
EPS [24]CVPR'21	I+S	R101	70.9	70.8
L2G [16]CVPR'22	I+S	R101	72.1	71.7
RCA [49]CVPR'22	I+S	R38	72.2	72.8
PPC$_{\text{w/EPS}}$ [10] CVPR'22	I+S	R101	72.6	73.6
+ FSAE			**74.4**$_{+1.8}$	**75.0**$_{+1.4}$

Table 3. Segmentation results (mIoU) on PASCAL VOC 2012 using DeepLabV1. † indicates the reproduced performance due to the lack of reported experimental results.

Method	Sup.	Net_{Seg}	Val	Test
ICD [12]CVPR'20	I	R101	67.8	68.0
AEFT [44]ECCV'22	I	R38	70.9	71.7
SIPE † [4]CVPR'22	I	R101	68.0	69.9
+ FSAE	I	R101	**70.3**$_{+2.3}$	**72.6**$_{+2.7}$
EPS [24]CVPR'21	I+S	R101	71.0	71.8
L2G [16]CVPR'22	I+S	R101	72.0	73.0
PPC$_{\text{w/EPS}}$ [10]CVPR'22	I+S	R101	72.3	73.5
+ FSAE	I+S	R101	**73.7**$_{+1.4}$	**75.3**$_{+1.8}$

Figure 5. Visualization examples of segmentation results. (a) Images, (b) PPC$_{w/EPS}$, (c) Our FSAE, and (d) GT.

Table 4. Segmentation results (mIoU) on COCO 2014 using DeepLabV2. † indicates reproduced performance due to the lack of reported experimental results. * indicates reproduced results.

Method	**Sup.**	**Net$_{Seg}$**	**COCO Val**
OC-CSE [20] CVPR'21	I	R38	36.4
AFA [33] CVPR'22	I	MiT-B1	38.9
SIPE * [4]CVPR'22	I	R101	39.3
+ FSAE			**39.5**$_{+0.2}$
EPS [24] CVPR'21	I+S	R101	35.7
PPC$_{w/EPS}$ † [10] CVPR'22	I+S	R101	33.7
+ FSAE			**35.4**$_{+1.7}$

Table 5. Comparison of False Positive Rate (FPR) and False Negative Rate (FNR) of pseudo-mask on PASCAL VOC 2012 train set.

Method	**FPR ↓**	**FNR ↓**
SIPE CVPR'22	19.4	12.7
+ Ours	**17.9** -1.5	**11.4** -1.3
PPC$_{w/EPS}$ CVPR'22	14.4	17.7
+ Ours	**12.0** -2.4	**15.1** -2.6
MCTformer CVPR'22	16.0	16.7
+ Ours	**15.4** -0.6	**16.6** -0.1

4.3. Ablation Studies

4.3.1 Effectiveness of Our Proposed Method

Table 5 demonstrates the effectiveness of our method. FPR(False Positive Rate) and FNR(False Negative Rate) represent the proportion of background regions predicted as objects and object regions predicted as background, respectively. Lower values indicate that the CAM results are well-aligned with the actual objects. As shown in Tab. 5, adding our method to the baseline results in lower FPR and FNR values. This demonstrates that our approach effectively expands CAMs for object and boundary regions by leveraging reliable pixel-wise labels.

Table 6. Initial seed quality with different component settings on PASCAL VOC 2012 *train* set. Evaluated in mIoU(%).

Method	PPC w/EPS	SIPE	MCTformer
Base	70.5	58.6	61.5
+ FAE	73.7	60.9	62.2
+ SAE	**74.5**	**61.1**	**62.4**

4.3.2 Effects of Each Components

We analyze the performance of PPC$_{w/EPS}$ with FSAE, and demonstrate the effectiveness of each component of our method. In Tab. 6, FAE significantly improves initial seed quality, and SAE further boosts FAE performance. Especially in Fig. 6, Our integral methods (FAE+SAE) provides more broad and accurate online pseudo-label in the training phase, compared to FAE only.

4.3.3 Ablation on Hyperparameters in FAE

We analyze the effect of hyperparameters in FAE. First, in Tab. 7, the initial seed quality gradually decreases as the confident threshold τ_{FAE} deviates from 0.95. It means pseudo-labeling with low τ_{FAE} values allow the network to learn broader regions during the initial training phase, but is more likely to learn the noise from the unstable prediction, resulting in performance degradation. In contrast,

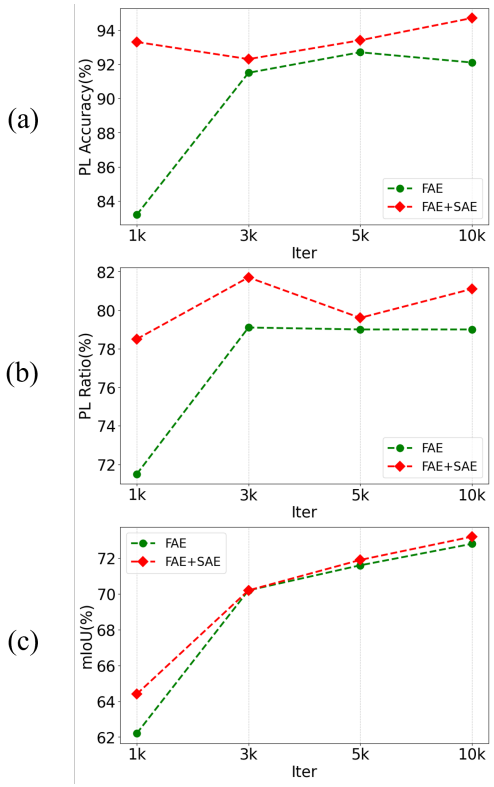


Figure 6. Accuracy and ratio of online pseudo-label in early training stage with actual initial seed quality. (a) Online pseudo-label accuracy, (b) Ratio of the online pseudo-label region, and (c) mIoU of initial seed.

Table 7. Effect of confident threshold τ_{FAE} in FAE, evalutated by initial seed quality in PASCAL VOC 2012 *train* set. Ablations are conducted with no SAE.

τ_{FAE}	0.80	0.85	0.90	**0.95**	0.99
mIoU(%)	70.0	72.8	73.5	**73.7**	73.3

setting too strict τ_{FAE} did not perform well because of the overly restricted pseudo-label region. Second, The number of Randaugment transforms controls the magnitude of perturbation applied to the target view x_t. Tab. 8 shows the initial seed quality with adjusting number of transforms. A higher value did not ensure superior performance, and the model showed fair performance even when applying just single transform. We set the number of RandAugment to 3, with all other hyperparameters fixed to in-paper settings.

4.3.4 Ablation on Hyperparameters in SAE

We evaluate the boundary expansion of pseudo-label region. The magnitude is determined by dilation kernel size k and restrained by second confident threshold τ_{SAE}. We provide

Table 8. Initial seed and pseudo-mask quality (mIoU) of PASCAL VOC 2012 *train* set according to the number of RandAugment.

# RandAug	1	**3**	5	7
Seed (%)	73.8	**74.5**	74.3	73.4
Mask (%)	75.6	**77.0**	76.9	76.0

Table 9. Effect of dilation kernel size k, with τ_{FAE} fixed to 0.95 and τ_{SAE} fixed to 0.8. Initial seed quality by mIoU (%) in PASCAL VOC 2012 *train* set.

k	None	**3**	5	7
mIoU(%)	73.7	**74.3**	74.2	73.9

Table 10. Effect of confident threshold τ_{SAE} in SAE, with τ_{FAE} fixed to 0.95. Initial seed quality in PASCAL VOC 2012 *train* set.

τ_{SAE}	0.0	0.6	0.7	**0.8**	0.9
mIoU(%)	74.2	74.2	74.3	**74.5**	74.4

ablation of k and τ_{SAE} with other hyperparameters fixed with in-paper settings. As evaluated in Tab. 9, the optimal value of dilation is 3. The large kernel size exceeding 3 degrades the performance, which mainly caused by excessive expansion. Also, as shown in Tab. 10, the performance was not sensitive to τ_{SAE}. It shows that the SAE can be implemented without significant adjustments.

5. Conclusion

We proposed the FSAE framework for weakly-supervised semantic segmentation, which were designed to generate and propagate explicit and accurate online pseudo-labels for pixel-wise representations. Leveraging weak-strong consistency and pseudo-label expansion strategies, the framework prompted the network to extract information from more extensive regions. Our method demonstrated its effectiveness on the PASCAL VOC 2012 and COCO 2014 datasets, consistently showing performance improvements across various baselines.

Acknowledgments. This work was partly supported by ICT Creative Consilience Program through the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (RS-2020-II201821, 33.3%), cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City (33.3%), and Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2024-00360227, Developing Multimodal Generative AI Talent for Industrial Convergence, 33.3%).

References

[1] Jiwoon Ahn, Sunghyun Cho, and Suha Kwak. Weakly supervised learning of instance segmentation with inter-pixel relations. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2, 6

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Semantic image segmentation with deep convolutional nets and fully connected crfs. arXiv preprint arXiv:1412.7062, 2014. 6

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE transactions on pattern analysis and machine intelligence, 40(4):834–848, 2017. 6

[4] Qi Chen, Lingxiao Yang, Jian-Huang Lai, and Xiaohua 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 (CVPR), pages 4288–4298, June 2022. 1, 2, 5, 6, 7

[5] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In International conference on machine learning, pages 1597–1607. PMLR, 2020. 2

[6] Tao Chen, Yazhou Yao, Lei Zhang, Qiong Wang, Guosen Xie, and Fumin Shen. Saliency guided inter-and intra-class relation constraints for weakly supervised semantic segmentation. IEEE Transactions on Multimedia, 2022. 2

[7] Xiaokang Chen, Yuhui Yuan, Gang Zeng, and Jingdong Wang. Semi-supervised semantic segmentation with cross pseudo supervision. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2613–2622, 2021. 3

[8] Zhaozheng Chen and Qianru Sun. Extracting class activation maps from non-discriminative features as well. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 3135–3144, 2023. 2, 3, 6

[9] Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops, pages 702–703, 2020. 3, 5

[10] Ye Du, Zehua Fu, Qingjie Liu, and Yunhong Wang. Weakly supervised semantic segmentation by pixel-to-prototype contrast. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2022. 1, 2, 5, 6, 7

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. International journal of computer vision, 88(2):303–338, 2010. 5

[12] Junsong Fan, Zhaoxiang Zhang, Chunfeng Song, and Tieniu Tan. Learning integral objects with intra-class discriminator for weakly-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4283–4292, 2020. 6

[13] Bharath Hariharan, Pablo Arbeláez, Lubomir Bourdev, Subhransu Maji, and Jitendra Malik. Semantic contours from inverse detectors. In 2011 international conference on computer vision, pages 991–998. IEEE, 2011. 5

[14] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 9729–9738, 2020. 2

[15] Qibin Hou, PengTao Jiang, Yunchao Wei, and Ming-Ming Cheng. Self-erasing network for integral object attention. Advances in Neural Information Processing Systems, 31, 2018. 2

[16] Peng-Tao Jiang, Yuqi Yang, Qibin Hou, and Yunchao Wei. L2g: A simple local-to-global knowledge transfer framework for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16886–16896, 2022. 2, 6

[17] Sanghyun Jo, In-Jae Yu, and Kyungsu Kim. Mars: Model-agnostic biased object removal without additional supervision for weakly-supervised semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 614–623, 2023. 3

[18] Tsung-Wei Ke, Jyh-Jing Hwang, and Stella X Yu. Universal weakly supervised segmentation by pixel-to-segment contrastive learning. arXiv preprint arXiv:2105.00957, 2021. 1, 2

[19] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. Advances in neural information processing systems, 33:18661–18673, 2020. 2

[20] Hyeokjun Kweon, Sung-Hoon Yoon, Hyeonseong Kim, Daehee Park, and Kuk-Jin Yoon. Unlocking the potential of ordinary classifier: Class-specific adversarial erasing framework for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF international conference on computer vision, pages 6994–7003, 2021. 2, 7

[21] Jungbeom Lee, Eunji Kim, and Sungroh Yoon. Anti-adversarially manipulated attributions for weakly and semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4071–4080, 2021. 1, 5

[22] Jungbeom Lee, Jihun Yi, Chaehun Shin, and Sungroh Yoon. Bbam: Bounding box attribution map for weakly supervised semantic and instance segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 2643–2652, 2021. 1, 2

[23] Minhyun Lee, Dongseob Kim, and Hyunjung Shim. Threshold matters in wsss: Manipulating the activation for the robust and accurate segmentation model against thresholds. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4330–4339, 2022. 1, 6

[24] Seungho Lee, Minhyun Lee, Jongwuk Lee, and Hyunjung Shim. Railroad is not a train: Saliency as pseudo-pixel

supervision for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 5495–5505, June 2021. 2, 5, 6, 7

[25] Jing Li, Junsong Fan, and Zhaoxiang Zhang. Towards noiseless object contours for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16856–16865, 2022. 2

[26] Xueyi Li, Tianfei Zhou, Jianwu Li, Yi Zhou, and Zhaoxiang Zhang. Group-wise semantic mining for weakly supervised semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 1984–1992, 2021. 2

[27] Yi Li, Yiqun Duan, Zhanghui Kuang, Yimin Chen, Wayne Zhang, and Xiaomeng Li. Uncertainty estimation via response scaling for pseudo-mask noise mitigation in weakly-supervised semantic segmentation. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, pages 1447–1455, 2022. 2

[28] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV, 2014. 5

[29] Yuqi Lin, Minghao Chen, Wenxiao Wang, Boxi Wu, Ke Li, Binbin Lin, Haifeng Liu, and Xiaofei 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, pages 15305–15314, 2023. 2, 3, 6

[30] Youngmin Oh, Beomjun Kim, and Bumsub Ham. Background-aware pooling and noise-aware loss for weakly-supervised semantic segmentation. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6913–6922, 2021. 2

[31] Zelin Peng, Guanchun Wang, Lingxi Xie, Dongsheng Jiang, Wei Shen, and Qi Tian. Usage: A unified seed area generation paradigm for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 624–634, 2023. 2, 3, 6

[32] Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 19574–19584, 2023. 3, 6

[33] Lixiang Ru, Yibing Zhan, Baosheng Yu, and Bo Du. Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 16846–16855, 2022. 7

[34] Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin A Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. Advances in neural information processing systems, 33:596–608, 2020. 3

[35] Guolei Sun, Wenguan Wang, Jifeng Dai, and Luc Van Gool. Mining cross-image semantics for weakly supervised semantic segmentation. In European conference on computer vision, pages 347–365. Springer, 2020. 2

[36] Yude Wang, Jie Zhang, Meina Kan, Shiguang Shan, and Xilin Chen. Self-supervised equivariant attention mechanism for weakly supervised semantic segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020. 2

[37] Yunchao Wei, Jiashi Feng, Xiaodan Liang, Ming-Ming Cheng, Yao Zhao, and Shuicheng 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, pages 1568–1576, 2017. 2

[38] Tong Wu, Guangyu Gao, Junshi Huang, Xiaolin Wei, Xiaoming Wei, and Chi Harold Liu. Adaptive spatial-bce loss for weakly supervised semantic segmentation. In European Conference on Computer Vision, pages 199–216. Springer, 2022. 6

[39] Jinheng Xie, Xianxu Hou, Kai Ye, and Linlin Shen. Clims: Cross language image matching for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4483–4492, 2022. 6

[40] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4310–4319, 2022. 2, 3, 5, 6

[41] Lian Xu, Wanli Ouyang, Mohammed Bennamoun, Farid Boussaid, and Dan Xu. Multi-class token transformer for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 4310–4319, June 2022. 6

[42] Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 7236–7246, 2023. 3

[43] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. St++: Make self-training work better for semi-supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4268–4277, 2022. 3

[44] Sung-Hoon Yoon, Hyeokjun Kweon, Jegyeong Cho, Shinjeong Kim, and Kuk-Jin Yoon. Adversarial erasing framework via triplet with gated pyramid pooling layer for weakly supervised semantic segmentation. In European Conference on Computer Vision, pages 326–344. Springer, 2022. 6

[45] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 12546–12555, 2020. 1, 2

[46] Xiaolin Zhang, Yunchao Wei, Jiashi Feng, Yi Yang, and Thomas S Huang. Adversarial complementary learning for weakly supervised object localization. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1325–1334, 2018. 3

[47] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 3085–3094, 2019. 2

[48] Xinqiao Zhao, Feilong Tang, Xiaoyang Wang, and Jimin Xiao. Sfc: Shared feature calibration in weakly supervised semantic segmentation. arXiv preprint arXiv:2401.11719, 2024. 2, 3, 6

[49] Tianfei Zhou, Meijie Zhang, Fang Zhao, and Jianwu Li. Regional semantic contrast and aggregation for weakly supervised semantic segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 4299–4309, 2022. 2, 6